Programme CarHAB

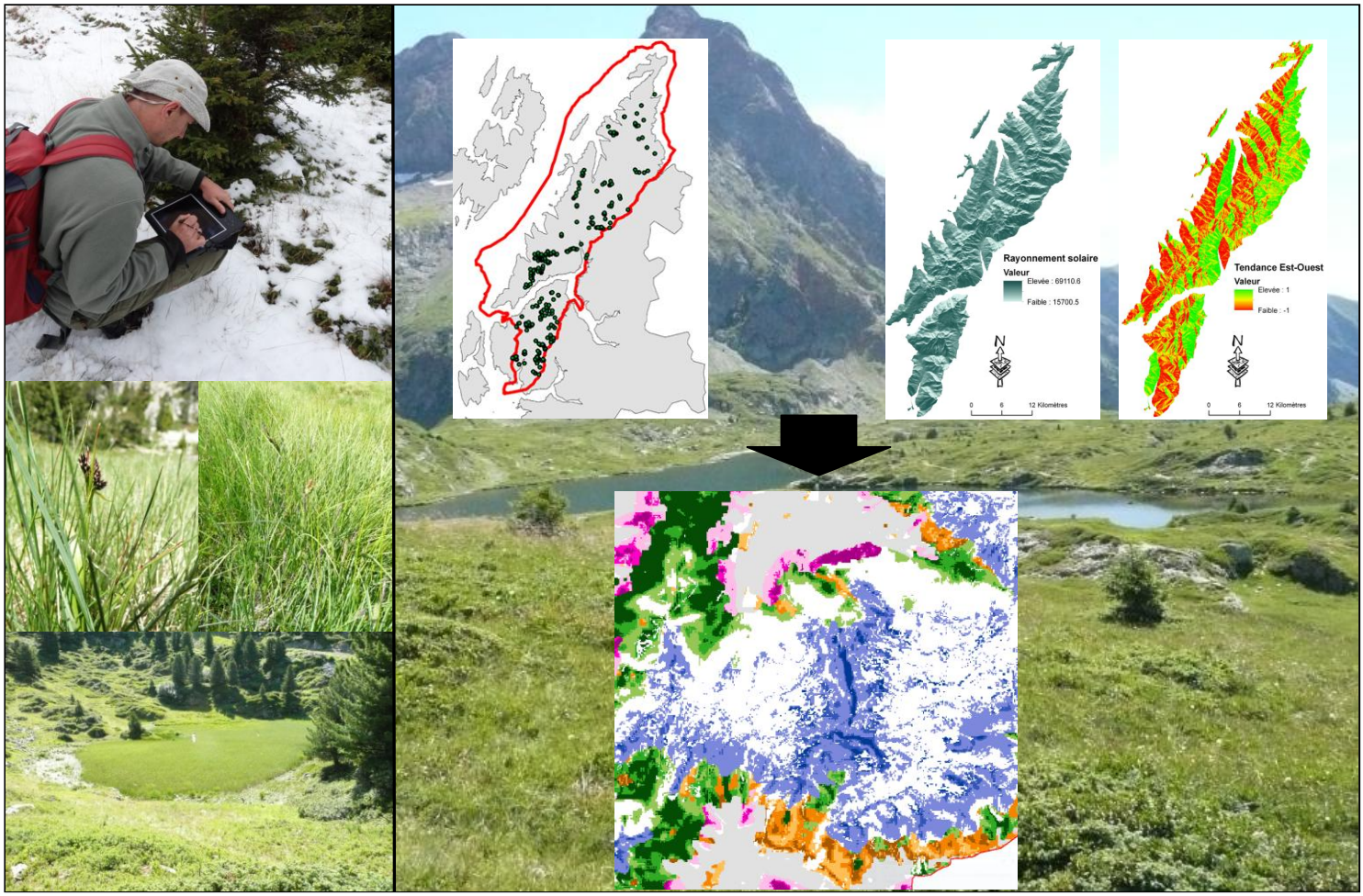
Volet transversal « Synthèse des expériences »

Synthèse bibliographique sur les expériences de modélisation de la végétation en Europe et en France

Octobre 2012

Pilotage du volet synthèse des expériences pour la modélisation : Irstea Grenoble (Sandra LUQUE) et FCBN (Jérome MILLET).

Chargée de mission : Mathilde REDON (FCBN/Irstea).



Contenu

I. Préambule

I.1 Objectif de la synthèse dans le cadre du programme CarHAB

Cette synthèse bibliographique sur les expériences de modélisation de la végétation en Europe et en France est élaborée dans le cadre du volet transversal « synthèse des expériences ». Ce volet a pour objectif d'établir un bilan des éléments méthodologiques structurants à prendre en compte dans le cadre du programme CarHAB. Il comprend quatre autres synthèses complémentaires : une synthèse des expériences européennes de cartographie de la végétation (pilotage SPN/MNHN), une synthèse des expériences françaises de cartographie de la végétation/mobilisation des données existantes (pilotage FCBN), une synthèse et analyse méthodologique des approches de cartographie de la végétation en Europe (pilotage : UBO) et une synthèse bibliographique sur les expériences de télédétection appliquée à la cartographie des habitats aux échelles européennes et françaises (pilotage : ISTHME et IRSTEA Montpellier).

Dans le cadre du programme CarHAB, la modélisation spatialisée de la distribution de la végétation a pour objectif de contribuer à la réalisation de la cartographie nationale de la végétation au travers de différentes actions : 1/ en ciblant sur les végétations présentes de façon très fragmentaires et difficiles à cartographier par des méthodes de télédétection, 2/ en facilitant la valorisation des jeux de données ponctuelles des CBNs, 3/ en orientant les prospections (secteurs sous-échantillonnés…) pour faciliter la réalisation terrain de la carte de végétation par les CBNs et 4/ en contribuant à l'alimentation du fond blanc par pré-remplissage des polygones issus de la segmentation effectuée dans le cadre du volet télédétection. La modélisation spatialisée de la distribution des espèces et des habitats est un domaine en plein essor depuis une trentaine d'années, en lien notamment avec le développement des outils d'observation de la surface de la Terre (photographies aériennes, images satellites, stations météorologiques…) et des outils informatiques. Il existe aujourd'hui une très grande diversité de méthodes et d'outils permettant de prédire la répartition spatiale des espèces et des habitats. **Afin de déterminer quels seront les outils, données et savoir-faire nécessaires pour atteindre les objectifs du volet modélisation du programme CarHAB, il est donc d'abord nécessaire de réaliser un état des lieux des méthodes employées et des expériences réussies pour prédire la distribution de la végétation. Compte-tenu des objectifs du programme, cette synthèse est centrée sur les études réalisées en Europe et en France**.

I.2 Définition de la modélisation de la végétation telle qu'abordée dans cette synthèse

La modélisation prédictive de la végétation peut être définie comme la prédiction de la distribution géographique de la composition de la végétation dans un paysage à partir de variables environnementales spatialisées (Franklin, 1995). Il existe différentes manières d'atteindre cet objectif. Nous avons choisi de centrer cette synthèse sur les modèles **spatialisés** permettant de **prédire la distribution géographique actuelle d'espèces, de groupes d'espèces ou d'habitats sur la base des relations quantitatives entre leur occurrence ou leur abondance et des variables environnementales spatialisées décrivant les conditions abiotiques du milieu**. Ces modèles sont

basés sur l'hypothèse que la distribution du syntaxon étudié peut être prédite à partir de données d'observation (généralement ponctuelles) et de variables environnementales spatialisées en continu sur l'ensemble de la zone d'étude, qui sont corrélées à la distribution de ce syntaxon.

Note : l'ensemble des méthodes de modélisation recensées dans cette synthèse peuvent être appliquées en théorie aussi bien à des espèces qu'à des groupes d'espèces (ex : alliances ou associations végétales) ou à des habitats. Par soucis de simplification du texte, nous parlons souvent de la modélisation de la distribution d'une espèce X dans la description des méthodes, mais ces approches sont tout à fait transposables à différents syntaxons ou à des habitats.

I.3 Précisions sur le contenu de la synthèse

Ce rapport présente une synthèse bibliographique des expériences de modélisation de la végétation en Europe et en France. La bibliographie consultée est majoritairement constituée d'articles de revues scientifiques internationales. Il existe une multitude des publications sur le sujet, un très grand nombre de méthodes et la plupart des expériences menées en Europe sont basées sur des approches développées dans d'autres régions. Nous avons donc choisi de donner d'abord un aperçu de la diversité des approches de modélisation existantes avec leurs principales caractéristiques. Nous rappelons également le contexte du développement de la modélisation spatialisée de la distribution des espèces et des habitats et le cadre théorique associé. Nous proposons ensuite un tour d'horizon de la modélisation de la végétation en Europe et en France, avec quelques exemples concrets accompagnés d'une bibliographie complémentaire qui ne peut cependant pas être exhaustive. Les informations rassemblées dans cette synthèse mettent en avant l'existence d'outils intéressants et donnent des pistes de réflexion pour orienter le choix d'une démarche modélisation adaptée dans le cadre du volet modélisation du programme CarHAB.

.

II. Introduction

II.1 Contexte général

L'accroissement des pressions sur les ressources naturelles et les changements globaux conduisent aujourd'hui à de profondes modifications de notre environnement, avec des conséquences importantes pour la préservation des habitats naturels et de la biodiversité. Les changements d'occupation des sols, la propagation des espèces envahissantes et la dégradation des habitats naturels sont les principales menaces actuellement en action. Les enjeux de cette perte de biodiversité pour la préservation des ressources naturelles et pour la pérennité du bien-être des sociétés humaines sont aujourd'hui mondialement reconnus, et la lutte contre la disparition de la biodiversité et la dégradation des habitats naturels est devenue un objectif international prioritaire (Scholes & Biggs 2004; Hanski 2005).

Dans ce contexte, les données géographiques sur la distribution des habitats et des espèces sont devenues indispensables pour les gouvernements et les organismes publics et privés qui doivent prendre des décisions dans les domaines de l'aménagement « durable » des territoires, de l'adaptation au changement climatique et de la conservation de la biodiversité (Jetz et al., 2012). En effet, l'amélioration de l'efficacité des programmes de conservation et de protection du patrimoine naturel ne peut se faire sans informations sur la localisation dans l'espace des espèces et des habitats naturels. Cependant, la réalisation d'inventaires d'espèces et d'habitats et la cartographie de leurs aires de répartition sur le terrain sont très coûteuses en temps et en moyens. Les données disponibles sont, de ce fait, souvent insuffisantes pour répondre aux besoins de la conservation (Polasky et al., 2000; Gurnell et al., 2002; Wilson et al., 2005). Face à cette difficulté majeure, les 30 dernières années ont été riches en innovations dans le domaine de la modélisation spatialisée pour prédire la distribution des espèces et de leurs habitats. Les premiers travaux concernant la modélisation prédictive de la distribution de la végétation remontent à la fin des années 70 (Franklin, 1995). L'essor des outils informatiques, des techniques de télédétection et de photo-interprétation s'est avéré déterminant pour le développement de ces nouvelles approches. Il existe aujourd'hui une grande diversité de méthodes performantes permettant de prédire la distribution géographique d'espèces ou d'habitats dans un territoire donné, à partir de variables environnementales spatialisées décrivant le milieu et de données d'observation.

II.2 Deux grand types d'approches de modélisation

Ces méthodes de modélisation peuvent être classées en **2 grandes catégories :** les approches « mécanistes » (dynamiques) et les approches « empiriques » (statiques), (Carpenter et al., 1993; Zimmermann and Kienast, 1999).

Les modèles *mécanistes* (ou dynamiques) ont généralement pour objectif de comparer la performance ou la survie d'un organisme au cours du temps ou dans différentes conditions environnementales. Ils réalisent des prédictions à partir d'informations détaillées et explicites sur les paramètres de performance des populations (ex : taux de reproduction ou de mortalité, phénologie, physiologie) et sur leur interaction avec leur environnement. Ces modèles demandent donc une

connaissance approfondie des traits de vie des espèces. Ils sont été peu utilisés pour la modélisation de la distribution spatiale des espèces (mais voir par exemple la méthode CLIMEX développée par Sutherst and Maywald (1985).

Les modèles *empiriques* sont basés sur la recherche de structures de corrélations entre des variables environnementales et la distribution du syntaxon ciblé dans une zone géographique donnée. Ils partent de l'hypothèse que les relations entre les patrons de distribution observés pour les espèces et leur environnement sont à l'équilibre et statiques. Ils nécessitent beaucoup moins de données que les modèles dynamiques. Ces modèles statiques sont, de ce fait, souvent considérés comme étant les plus adaptés lorsque l'objectif est de modéliser précisément des entités biologiques à de grandes échelles spatiales dans des conditions environnementales présentes (Guisan and Zimmermann, 2000) ; la grande majorité des approches utilisées pour la modélisation spatialisée de la distribution des espèces et des habitats appartiennent à cette catégorie. La suite de la synthèse bibliographique s'intéressera uniquement à cette deuxième catégorie de modèles.

III. Diversité des méthodes de modélisation prédictive spatialisée et cadre théorique associé

III.1 Une diversité d'objectifs

Le nombre de travaux portant sur la modélisation de la distribution des espèces et des habitats a augmenté de façon exponentielle au cours des dernières années. Une part importante de la littérature concerne le développement de nouvelles méthodes, leur comparaison et l'amélioration des algorithmes et des outils de validation correspondants (voir par exemple Anderson et al.(2003), Elith et al. (2006), Phillips and Dudik (2008), Tsoar et al.(2007), Guisan and Zimmermann (2000), Guisan et al. (1999), Zimmermann et al. (2010)). Les autres travaux présentent des cas concrets d'utilisation de ces méthodes dans le domaine de la conservation du patrimoine naturel, où de nombreux champs d'application ont été envisagés :

- Localisation et délimitation de nouveaux espaces protégés ou évaluation de l'efficacité des statuts de protection (Polasky et al., 2000; Wilson et al., 2005; Martínez et al., 2006; Godet et al., 2007; Thomaes et al., 2008),
- prévision et suivi de la progression des espèces invasives ou pathogènes (Jiménez-Valverde et al., 2011; Rodda et al., 2011; Ward, 2007),
- orientation des prospections naturalistes dans des secteurs mal connus (Debinski et al., 1999; Hernandez et al., 2008; Stabach et al., 2009; Rebelo and Jones, 2010),
- détermination de la continuité spatiale des habitats favorables à une espèce (corridors écologiques), (ex. Decout et al., 2012),
- détermination de la distribution d'espèces rares ou menacées pour faciliter leur protection (Luoto et al., 2002; Gibson et al., 2004; Pearson, et al., 2007; Kafley et al., 2009; Kumar and Stohlgren, 2009; Hu and Jiang, 2010),
- localisation de sites compatibles pour la réintroduction d'espèces (Martinez-Meyer et al., 2006),

- prédiction de l'évolution de la distribution des espèces dans le contexte du changement climatique (Huntley et al., 1995; Elith et al., 2010; Songer et al., 2012),
- amélioration des connaissances sur les relations entre les espèces et leur habitat (ex : avifaune (Hagan and Meehan, 2002; Mac Faden and Capen, 2002; Oja et al., 2005; Fearer et al., 2007), papillons (Fleishman et al., 2001) qui peuvent être utilisées pour le développement d'indicateurs indirects de biodiversité (ex : Kerr et al., 2001; Ewers et al., 2005),
- cartographie des habitats naturels et semi-naturels (Mücher et al., 2009).

La majorité de ces travaux s'intéressent à la distribution d'espèces animales, les études portant sur la distribution spatiale de la végétation étant un peu moins fréquentes, mais voir par exemple : (Huntley et al., 1995; Zimmermann and Kienast, 1999; Luoto et al., 2002; Coudun et al., 2006).

III.2 Théorie de la niche et relations espèces - environnement

Quelque soit l'objectif recherché ou le syntaxon étudié, les méthodes de modélisation sont souvent basées sur la *théorie de la niche* qui définit la niche écologique d'une espèce comme l'ensemble des facteurs biotiques et abiotiques nécessaires à sa survie dans un lieu donné (Grinnell, 1917; Hutchinson, 1959; Hirzel and Le Lay, 2008). De ce fait, chaque espèce est adaptée à un ensemble particulier de conditions biotiques et abiotiques, qui détermine sa persistance à long terme et conditionne sa coexistence avec d'autres espèces dans un environnement donné. Suivant cette théorie, il est alors possible de reconstituer la distribution des sites favorables à une espèce à partir d'informations sur les caractéristiques environnementales des localités qu'elle occupe (Hirzel and Le Lay, 2008).

Cette niche, observée, échantillonnée, constitue la niche « effective » de l'espèce, c'est-à-dire celle qu'elle occupe réellement et actuellement (Phillips et al., 2006) et qui inclut les interactions biotiques (ex : compétition, facilitation), (Guisan and Zimmermann, 2000). Cette niche « effective » s'oppose à la niche « fondamentale », qui représente l'ensemble des localités comportant des paramètres environnementaux favorables à la survie de l'espèce, mais pas forcément occupées par elle (Phillips et al., 2006). En effet, différentes raisons peuvent expliquer l'absence d'une espèce dans un habitat théoriquement favorable : i) l'espèce subit une compétition interspécifique à son désavantage sur le territoire considéré, ou la présence d'un prédateur majeur (Hirzel and Le Lay, 2008), ii) l'espèce est absente pour des raisons historiques (ex : une barrière géographique empêche la colonisation de la zone) (Hirzel et al., 2002), iii) la capacité de dispersion de l'espèce rend la zone considérée inaccessible directement (Hirzel and Le Lay, 2008), iv) l'espèce peut être éteinte dans la zone considérée (dans ce cas, il y a eu une présence ancienne) (Anderson et al., 2003). La niche « effective » peut alors être considérée comme un sous-échantillon de la niche fondamentale (Hirzel and Le Lay, 2008). Les conditions environnementales décrivant les localités où l'espèce est présente représentent alors un échantillon de la niche effective de l'espèce (Phillips et al., 2006). De ce fait, *les conditions environnementales au niveau des points d'observation d'une espèce représentent un échantillon plus ou moins représentatif de la niche effective de l'espèce* (nous verrons par la suite l'importance de la représentativité des points d'observation que cela implique pour la modélisation).

Ainsi, un modèle de probabilité de distribution d'une espèce basé sur la théorie de la niche correspond à la prédiction spatialisée de la présence ou de l'abondance de l'espèce à une position XY

dans l'espace considéré en se basant sur ses exigences écologiques et sur les caractéristiques de l'habitat (valeurs des variables environnementales) à cette position XY.

III.3 Une diversité de méthodes

La recherche de méthodes permettant de quantifier les relations entre les espèces et leur environnement et de les extrapoler dans l'espace a été alimentée par la réalisation d'importantes avancées en statistiques au cours des 30 dernières années. Plusieurs approches statistiques existantes telles que les techniques de classification ou les modèles de régression (Guisan et al., 2002) ont été adaptées aux particularités des données environnementales et d'observation, aboutissant à de nouvelles approches plus robustes. Dans le même temps, l'essor des outils informatiques a permis le développement d'un vaste panel d'algorithmes de modélisation basés sur des processus d'apprentissage (appelés dans la littérature « *machine learning methods* »). Parmi les approches les plus connues appartenant à cette deuxième catégorie, on peut citer l'enveloppe environnementale (Busby, 1991; Carpenter et al., 1993), les réseaux de neurones (Pearson et al., 2002) ou le maximum d'entropie (Moore and Hooper, 1975).

Ces deux grands types d'approches, « statistiques » et « d'apprentissage », sont basées sur des philosophies différentes. La modélisation avec des *approches statistiques* nécessite de déterminer *a priori* le meilleur type de modèle qui permettra d'expliquer le jeu de donnée considéré ; les paramètres du modèle (ex : significativité et importance de l'effet des variables) sont ensuite estimés à partir des données. Les *méthodes d'apprentissage*, au contraire, partent du principe que le processus qui sous-tend le jeu de donnée est trop complexe pour être déterminé *a priori*. Elles utilisent un algorithme pour comprendre la relation entre les prédicteurs (ex : variables environnementales) et la variable à expliquer (ex : occurrence d'une espèce), sur la base de l'analyse des variables en entrées et des sorties possibles du processus. Elles en déduisent ensuite le meilleur modèle permettant d'expliquer le jeu de donnée considéré (Elith et al., 2008; voir aussi Remm (2004) pour une introduction intéressante sur les méthodes d'apprentissage). Le tableau 1 présente une synthèse des méthodes de modélisation les plus couramment utilisées depuis les années 90 avec leurs principales caractéristiques. Il existe plusieurs autres synthèses décrivant et/ou comparant différentes méthodes de modélisation, qui peuvent également être consultées pour plus d'informations (Franklin, 1995; Guisan and Zimmermann, 2000; Elith et al., 2006; Tsoar et al., 2007; Tarkesh and Jetschke, 2012). Par exemple, la synthèse de Franklin (1995) recense 16 approches de modélisation utilisées spécifiquement pour la modélisation prédictive de la végétation entre 1986 et 1994. La synthèse de Guisan and Zimmermann (2000) est une référence pour la modélisation spatialisée de distributions et recense également les principaux types de méthodes de modélisation actuellement utilisés.

III.3.1 Un exemple d'approche statistique : modèles basés sur des régressions

Beaucoup de méthodes de modélisation sont basées sur des modèles de type régression, où l'objectif est d'expliquer un phénomène (variables à expliquer), à partir de plusieurs autres variables continues ou catégorielles. Le résultat peut ensuite être spatialisé à partir des valeurs des variables explicatives spatialisées. Dans le cas des GLMs (modèles linéaires généralisés), les plus courants, la niche de l'espèce est estimée selon l'hypothèse qu'il existe une relation entre la moyenne des

valeurs de la variable à expliquer et une combinaison linéaire des variables explicatives. L'ajustement de ces modèles nécessite de connaître préalablement les lois de distribution des variables explicatives. Lorsque la variable à expliquer est binaire (ex : présence/absence), c'est la régression logistique qui est la plus adaptée. Différentes adaptations des GLMs ont été réalisées afin de répondre à différentes situations rencontrées lors de la manipulation de données écologiques. Les GAMs (modèles additifs généralisés), par exemple, sont utilisés lorsque les relations entre variables ne sont pas linéaires et/ou lorsque les lois de distribution des variables explicatives sont complexes. D'autres approches comme les modèles de dissimilarité généralisés ou les régressions multivariées adaptatives, ont également été développées mais sont moins largement utilisées (cf. tableau 1).

III.3.2 Quelques exemples d'approches basées sur un processus d'apprentissage

III.3.2.1 Enveloppe environnementale

Le principe de l'enveloppe environnementale correspond à la représentation de la niche d'une espèce sous la forme d'une « enveloppe » multivariée rectilinéaire dont les limites sont fixées par les valeurs minimales et maximales des variables environnementales au niveau des points d'observation (Busby, 1991). Plusieurs tailles d'enveloppe peuvent être considérées, en fonction de si l'ensemble de la distribution des valeurs des variables au niveau des points d'observation est utilisée ou si on ne considère qu'un sous-ensemble de la distribution de ces valeurs (ex : valeurs comprises entre les quantiles 5 et 95 pour restreindre l'enveloppe) (Carpenter et al., 1993). Les sites considérés comme favorables à l'espèce correspondent alors à ceux où toutes les variables bioclimatiques tombent dans les limites déterminées par cette distribution.

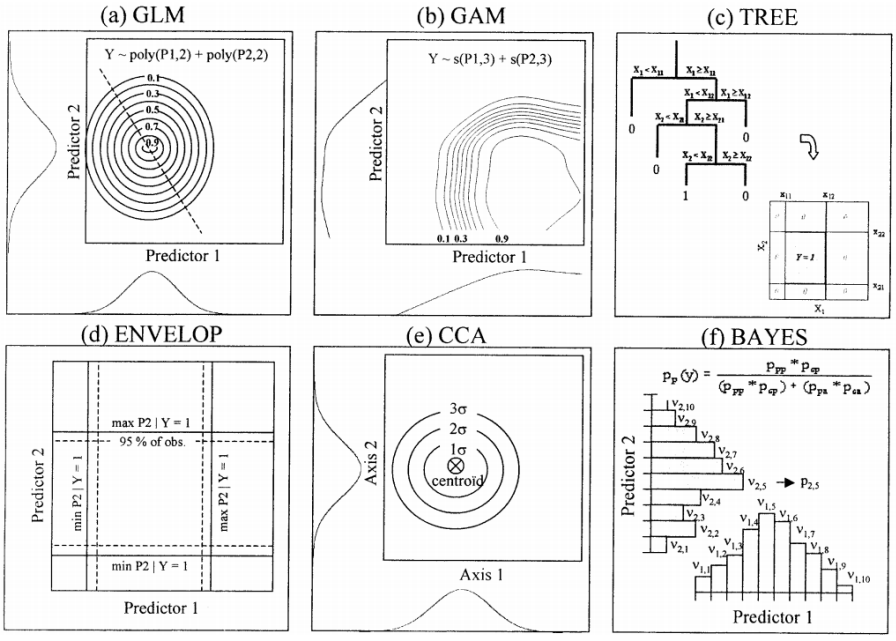


Figure 1. Exemples de courbes de réponse pour différentes approches de modélisation de la distribution des espèces (source : Guisan and Zimmermann, 2000) : (a) Modèle linéaire généralisé avec terme quadratique ; (b) modèle généralisé additif avec fonction de lissage ; (c) arbre de classification ; (d) enveloppe environnementale ; (e) analyse canonique des correspondances ; (f) statistiques bayésiennes (Aspinall, 1992).

III.3.2.2 Techniques de classification

Les techniques de classification comprennent une diversité de méthodes qui reposent sur un processus hiérarchique et itératif aboutissant à la partition d'un jeu de données initial en groupes ou en classes homogènes (Guisan and Zimmermann, 2000). Les arbres de classification (*classification trees*) et les arbres de régression (*regression trees*) sont les deux techniques de classification les plus connues dans le cadre de la modélisation de la distribution d'espèces ou d'habitats. Les premiers sont utilisés lorsque la variable à expliquer est catégorielle, les seconds quand elle est continue (ex : abondance) (De'ath and Fabricius, 2000). La création de groupes est basée sur une succession de règles de décision, qui permettent, à chaque étape, de départager le jeu de données en deux ou plusieurs groupes distincts. L'efficacité de chaque règle est déterminée en fonction du degré de « pureté » des groupes obtenus en divisant le jeu de données avec cette règle. Le même principe est appliqué de façon itérative, jusqu'à ce que les groupes soient suffisamment distincts en regard des objectifs fixés (Moore and Hooper, 1975; De'ath and Fabricius, 2000).

III.3.2.3 Le maximum d'entropie

La méthode du maximum d'entropie a été développée à l'origine pour réaliser des prédictions ou des interpolations à partir d'informations incomplètes dans différentes disciplines (Phillips et al., 2006). Elle a été récemment adaptée pour la modélisation de la distribution d'espèces à partir de données de présence, et son utilisation s'est rapidement répandue. Cette approche repose sur un algorithme qui permet d'estimer la distribution la plus probable de l'espèce en se basant sur le principe que la meilleure estimation d'une distribution inconnue est celle qui est la moins contraignante (avec le maximum d'entropie) pour l'espèce. Les « contraintes » sont définies en comparant la distribution des valeurs des variables environnementales aux points d'observation avec leur distribution pour un grand nombre de points pris au hasard dans la zone d'étude (absence ou présence de l'espèce possible) (Suárez-Seoane et al., 2008; Baldwin, 2009). Plus les valeurs des variables environnementales en un point donné dans la zone d'étude sont proches des conditions moyennes aux points d'observation par rapport aux conditions au niveau des points aléatoires, plus la probabilité d'occurrence prédite pour l'espèce est élevée (Phillips et al., 2006).

Une application en libre accès a été développée pour faciliter l'utilisation de la méthode (Phillips et al., 2005). Elle comprend une interface relativement accessible qui permet d'intégrer différents types de variables environnementales (quantitatives, qualitatives…), des jeux de points d'observation et/ou de test ainsi que de paramétrer les modèles. La mise en œuvre de la méthode aboutit à plusieurs résultats comprenant des statistiques et des informations qui permettent de juger de la pertinence des prédictions, ainsi qu'une représentation spatialisée des résultats qui peut être exportée dans un logiciel SIG pour être analysée.

Tableau 1. Synthèse des approches les plus couramment utilisées pour la modélisation spatialisée de la distribution des espèces, groupes d'espèces ou habitats, avec leurs principales caractéristiques. Logiciel : les exemples entre parenthèses correspondent à des logiciels « génériques » existants, Données : P = présence, P/A = présence/absence, Ab = abondance, Variables : Cont = continues, Cat = catégorielles, Inter = possibilité prise en compte interactions entre les variables.

Concept	Technique de modélisation	Nom du modèle	Logiciel ou application spécifique	Type de données d'observation	Types de variables environnementales	Exemples d'échelles d'application	Référence principale	Pays d'origine	Expériences européennes recensées?
Maximum d'entropie	Maximum d'entropie	Maxent	Maxent	P	Cont/Cat/Inter	Mondiale ; Continentale	Phillips et al., 2006	USA	Oui
Techniques d'ordination	Indice de distance géométrique moyenne	Ecological Niche Factor Analysis	Biomapper	P	Cont	Nationale ; Régionale ; Cantonale	Hirzel et al., 2002	Suisse	Oui
	CCA / RDA/ ACP	/	Non (ex : *R*[1] ; CANOCO[2])	P/A	Cont/Cat	Nationale	Ter Braak, 1986, 1987	Pays-Bas	Oui
Enveloppe environnementale	Réseaux de neurones	SPECIES	Stuttgart Neural Network Simulator	P/A	Cont/Cat	Européenne ; nationale	Pearson et al., 2002	Royaume-Uni	Oui
	Arbres de classification	BIOCLIM (+ adaptations : HABITAT ; SRE)	Bioclim	P, P/A	Cont	Continentale ; Nationale	Busby 1991, Nix 1986 ; Walker & Cocks, 1991	Australie	Non
	Indices de similarité multivariés (métrique de Gower)	DOMAIN	Domain	P, P/A	Cont	Nationale ; Régionale	Carpenter et al., 1993	Australie	Non
	"Boîtes" écologiques	Parallel-epiped model (PED)	Non	P/classes	Cat	Régionale	/	/	Oui
Distance de Mahalanobis	Mahalanobis distance	/	Non (ex : ArcView, MATLAB)	P	Cont/ binaires (0/1)	Nationale	Farber & Kadmon, 2003 ; Shao & Halpin, 1995	Israel, USA	Oui
Régression	Régression pondérée	/	Non	P/A	Cont	Européenne	/	/	Oui
	Modèles linéaires généralisés / Modèles additifs généralisés	GLMs / GAMs et adaptations : Multivariate adaptive regression splines (MARS)	Non (ex : R[1])	P/A	Cont/Cat/Inter	Régionale	ex : Guisan et al., 2002, Pearce & Ferrier, 2000 ; Friedman 1991	/	Oui
	Régression logistique								
Algorithme génétique	Algorithme génétique	GARP (Genetic Algorithm for Rule-set Production)	GARP Modelling System / Desktop GARP	P, P/A, Ab	Cont/Cat	Nationale	Stockwell & Noble, 1991	USA	Non
Techniques de classifications (règles de décision)	Analyse discriminante linéaire ou mixte	/	Non (ex : *R*[1])	P/A	Cont/Cat		Fisher, 1936	USA	Oui
	Combinaison d'arbres de régression et de "boosting"*	Boosted regression tree (BRT) / generalized boosted models	Non (ex : *R*[1])	P/A	Cont/Cat/Inter	Nationale	Ridgeway, 1999	USA	Oui
	Processus de classification dichotomique	Arbres de classification (CART + adaptations : SIMPLE, Random forest) ou arbres de régression	CART, S-Plus (RPART)	P, P/A, Ab	Cont/Cat/Inter	Européenne ; Régionale	Breiman et al 1984 ; Moore et al., 1991 ; Breiman, 2001	USA	Oui
Analyse multicritère	Croisement de couches pondérées ou non dans un SIG	/	Non, logiciel SIG	P/A/aucune	Cont/Cat	Régionale	ex : Moffet & Sarkar, 2006	?	Oui
Théorème de Bayes	Statistiques bayésiennes	/	Non (ex : ArcView, ArcWofE)	P/classes	Cont/Cat/Ordinales	Locale ; Nationale	Bonham-Carter et al., 1989 ; Aspinall, 1992	Canada	Oui
Géostatistiques	Krigeage	/	Non (ex : ArcGIS, *R*[1])	P	/	Nationale		/	Oui

*boosting = méthode permettant de combiner plusieurs modèles simples pour améliorer la performance des prédictions.

[1] = *R* development core team (2012)

[2] = ter Braak and Smilauer, 1998

III.3.3 Les analyses multicritères

Les analyses multicritères constituent une catégorie de méthodes à part, de type approche « experte ». Elles ont des champs d'application très diversifiés. Elles ont en particulier été largement développées dans le cas de la prise de décision impliquant différents acteurs aux objectifs différents, pour faciliter l'obtention d'un consensus (Moffett and Sarkar, 2006). Elles ont été relativement peu utilisées pour la modélisation de la distribution d'espèces ou d'habitats. Pourtant, elles peuvent constituer une manière très simple et directe de modéliser une distribution, bien qu'avec une précision spatiale moindre que les approches précédentes. Dans ce cas, la décision à prendre est de savoir si l'espèce est présente ou non en un lieu donné, éventuellement avec différents niveaux de fiabilité (ex : probabilité faible, moyenne, forte) (Store and Jokimäki, 2003). Chaque critère correspond à une couche SIG représentant une variable environnementale spatialisée, potentiellement structurante pour la distribution de l'espèce étudiée. Des connaissances d'expert permettent de définir pour chaque variable quelles sont les plages de valeurs favorables ou non à l'espèce. Ces critères peuvent être pondérés ou non en fonction de leur importance pour expliquer la présence de l'espèce. La modélisation se fait ensuite par croisement des couches géographiques dans un SIG, avec différentes techniques possibles de croisement des informations telles que l'addition de rasters booléens (Luque and Vainikainen, 2008), l'intersection de rasters reclassifiés (Mücher et al., 2009) ou des statistiques bayésiennes (Romero-Calcerrada and Luque, 2006). L'espèce est alors considérée comme présente dans les secteurs combinant l'ensemble des conditions qui lui sont favorables, on parle souvent d'indice d'habitat favorable (*habitat suitability index*) (Store and Kangas, 2001).

Nous avons présenté ici le fonctionnement de quelques approches de modélisation parmi les plus couramment utilisées. Le fonctionnement des autres méthodes recensées dans le tableau 1 peut être trouvé dans la littérature associée.

III.4 Vers des approches « consensus » multi-modèles

Face au vaste choix de méthodes de modélisation disponibles, il n'est pas toujours évident de déterminer celle qui est la plus appropriée pour atteindre les objectifs recherchés. La synthèse de (Guisan and Zimmermann, 2000) propose un tour d'horizon très complet de tous les facteurs qui peuvent intervenir dans le choix de la méthode de modélisation avec les hypothèses de travail associées. Ce choix peut reposer sur différents facteurs tels que l'objectif poursuivi, l'emprise géographique de la zone d'étude, la répartition des données d'observation disponibles et l'écologie de l'espèce étudiée (Segurado and Araujo, 2004; Luoto et al., 2005; Tsoar et al., 2007). La taille de l'échantillonnage, la nature de l'algorithme utilisé et ses hypothèses associées, ainsi que la région de l'étude déterminent également la fiabilité des prédictions (Wisz et al., 2008). Le schéma 1 synthétise les interactions entre les principaux facteurs qui peuvent orienter le choix d'une approche de modélisation. Certains facteurs sont plus contraignants que d'autres pour ce choix : par exemple, l'échelle d'application des approches de modélisation dépend surtout de l'objectif, de l'échelle des variables environnementales (étendue, résolution) et de la distribution des points d'observation disponibles ; elle est de ce fait assez indépendante du mécanisme de modélisation lui-même. De même, plusieurs approches différentes peuvent généralement être utilisées pour atteindre le même

objectif. Par contre, le type de données d'observation disponibles, leur nombre et leur répartition, peuvent davantage orienter le choix d'une approche de modélisation.

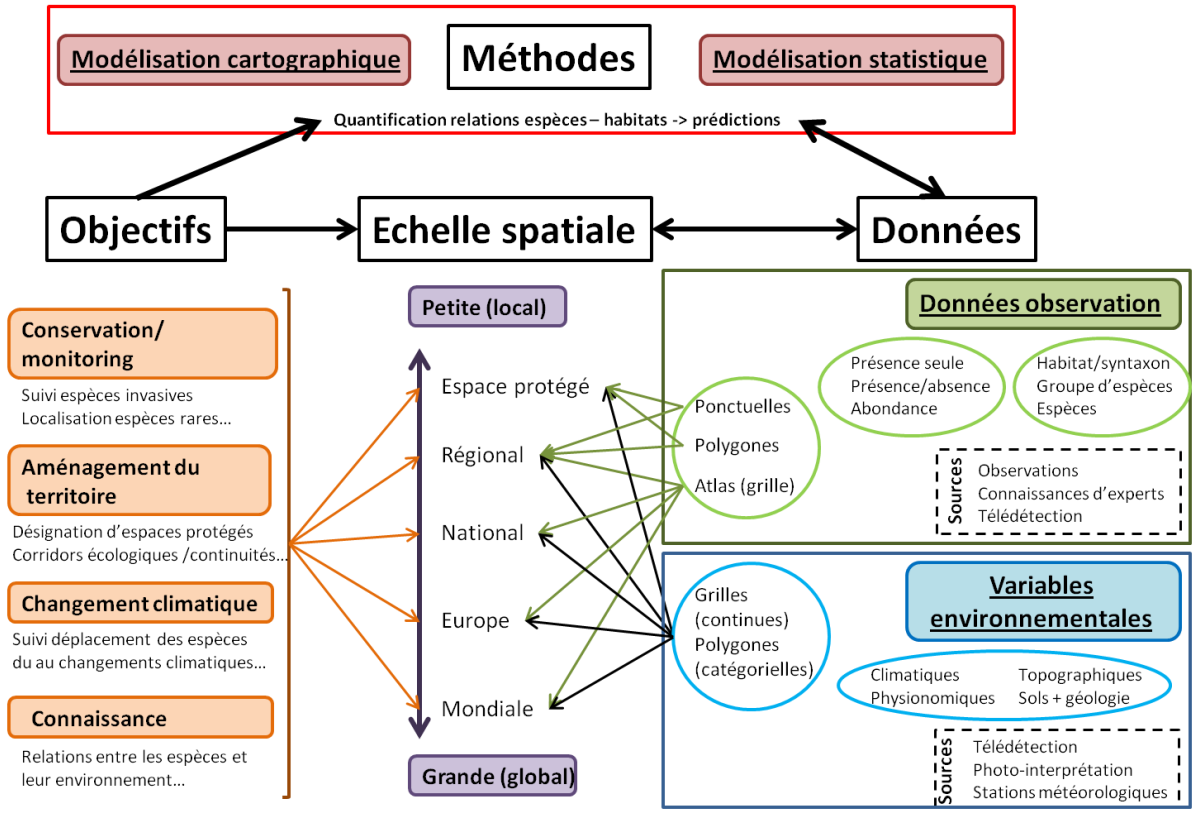


Figure 2. Les principaux facteurs déterminants dans le choix d'une démarche de modélisation.

Cette diversité des facteurs qui ont une influence possible sur la cohérence des résultats complexifie la sélection *a priori* de la méthode la plus appropriée pour atteindre l'objectif recherché. Dans ce contexte, une nouvelle « école » est apparue dans la modélisation prédictive spatialisée : au lieu de sélectionner une seule méthode et de l'appliquer une fois, il s'agit de multiplier les expériences de modélisation puis de tirer parti de l'ensemble des résultats pour aboutir à une prédiction plus robuste (Araújo and New, 2007). En effet, il est connu depuis longtemps que de combiner les résultats de différents modèles permet de diminuer fortement l'erreur moyenne associées aux prédictions. De plus, selon les objectifs recherchés et les jeux de données utilisés, chaque méthode possède ses avantages et ses inconvénients, aucune d'entre elles ne pouvant déterminer le modèle optimal dans toutes les situations (Robertson et al., 2003; Johnson and Gillingham, 2005). L'idée est donc de se rapprocher du meilleur modèle expliquant le système par l'intermédiaire d'un ensemble de « copies » légèrement différentes de ce même modèle (ex : variation des paramètres du modèle, des conditions initiales), qui apportent chacune une partie de l'information renseignant sur le processus réel sous-jacent au système étudié (Araújo and New, 2007).

III.4.1 Approches consensus par nature

Certaines approches de modélisation, fondées sur des processus stochastiques, sont basées par nature sur la comparaison de différents modèles, comme par exemple les BRTs (boosting regression trees), la méthode « Random Forests » (Elith et al., 2008), les réseaux de neurones et l'algorithme génétique GARP.

Dans le cas des BRTs, le processus de modélisation est itératif. A chaque itération, un sous-ensemble aléatoire du jeu de données initial est utilisé pour obtenir un modèle intermédiaire. Le modèle final correspond ensuite à une combinaison linéaire des *n* modèles intermédiaires issus des *n* itérations, qui correspond à un modèle de régression où chaque terme serait un modèle (Elith et al., 2008). Les réseaux de neurones incorporent également la notion de prédictions consensus puisque les modèles sont lancés plusieurs fois et les prédictions sont ensuite moyennées.

L'algorithme génétique GARP fonctionne selon un processus itératif dont la première étape est de sélectionner un type de modèle parmi plusieurs possibilités (enveloppes, régressions multinomiales, etc). A chaque itération, la technique de modélisation sélectionnée est testée, validée puis le résultat est conservé ou non pour aboutir au résultat consensuel final issu de la combinaison des résultats intermédiaires (Peterson and Cohoon, 1999; Stockwell and Peters, 1999). L'approche Maxent intègre aussi des processus itératifs dans l'algorithme de modélisation (Araújo and New, 2007).

L'intérêt de ces approches est que leurs prédictions sont issues de la combinaison de différents modèles construits à partir des mêmes données initiales. Cependant, pour une approche de modélisation donnée, la nature de l'algorithme utilisé est identique pour tous les modèles comparés.

III.4.2 Utilisation de plusieurs approches en parallèle

L'autre alternative est de faire tourner simultanément et séparément différentes méthodes de modélisation (chacune avec ou sans itérations) sur le même jeu de données et la même emprise géographique initiaux (Nabout et al., 2010). Il est ensuite possible soit de comparer les résultats pour sélectionner l'approche qui donne les résultats les plus cohérents au regard des objectifs fixés, soit de combiner les différents résultats en un seul modèle final (Thomaes et al., 2008). La combinaison des résultats peut se faire de différentes manières, par exemple en additionnant ou moyennant les valeurs prédites par les modèles intermédiaires (Marmion et al., 2009).

La plateforme BIOMOD (Thuiller, 2003; Thuiller et al., 2009; http://www.will.chez-alice.fr/Software.html), par exemple, a été développée spécifiquement dans cet objectif. Elle est en libre accès et a pour but de faciliter la mise en œuvre simultanée de différentes approches de modélisation et de sélectionner la méthode de modélisation la plus adaptée et la plus précise pour chaque espèce ou groupe d'espèce considérés. Elle permet de réaliser des prédictions de la distribution actuelle d'espèces et de les projeter dans le futur en fonction de scénarios de changement climatique. La dernière version de la plateforme est capable d'optimiser et de comparer neuf techniques de modélisation différentes et fournit plusieurs outils pour tester et comparer les modèles. La majorité de ces approches nécessite normalement des données de présence et d'absence, mais il est possible de ne travailler qu'avec des données de présence. Cette plateforme BIOMOD a déjà été largement utilisée, principalement en Europe, avec des objectifs variés comme

l'estimation de la distribution d'une espèce menacée dans des espaces protégés (Thomaes et al., 2008), la comparaison de différentes méthodes de combinaison de résultats dans le cas d'approches consensus (Marmion et al., 2009) ou la prédiction de la distribution d'espèces dans le futur (Thuiller, 2003).

III.5 Conclusion

Il existe donc une grande diversité de méthodes de modélisation, avec chacune leurs spécificités et différentes manières d'appréhender et de mettre en œuvre une démarche de modélisation. Nous allons maintenant passer en revue des expériences réussies de modélisation de la végétation en Europe et en France afin de déterminer quelles sont les méthodes les plus fréquemment utilisées, les caractéristiques des données utilisées et les facteurs les plus déterminants dans le choix des méthodes de modélisation.

IV. Expériences européennes et françaises de modélisation de la végétation

IV.1 Tour d'horizon de la modélisation en Europe

De nombreuses expériences de modélisation spatialisée de la distribution d'espèces, de groupes d'espèces et d'habitat naturels ont été menées en Europe (Tableau 1). Ces expériences sont caractérisées par une grande diversité d'approches de modélisation, d'objets d'études (espèces végétales et animales, communautés végétales, habitats…), d'objectifs et d'échelles d'application (échelle locale à européenne). Beaucoup de méthodes utilisées actuellement en Europe et en France ont été développées en Amérique du Nord (ex : Maxent, Phillips et al., 2006) ou en Australie (ex : BIOCLIM (Busby, 1991) ou DOMAIN (Carpenter et al., 1993)). Peu d'entre elles sont européennes (ex : SPECIES (Pearson et al., 2002) ; ENFA (Hirzel 2002)).

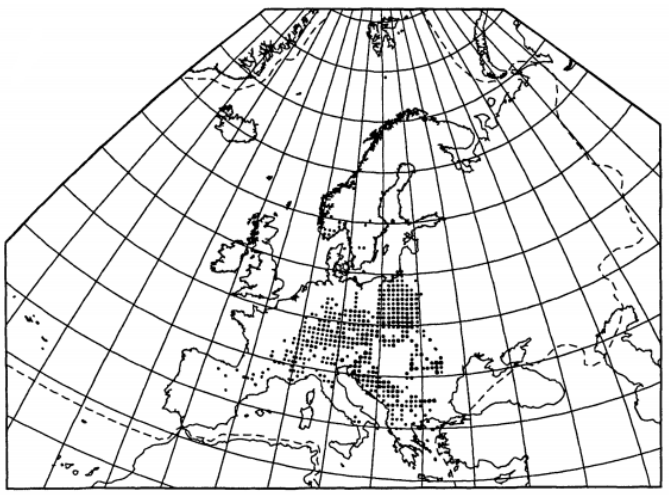
Nous avons centré les recherches sur les expériences de modélisation basées sur les méthodes recensées dans le Tableau 1 et conduisant à des prédictions **spatialisées**. Cependant, ce domaine est si vaste qu'il n'était ni possible ni pertinent dans le temps imparti de réaliser une synthèse exhaustive. Les travaux recensés donnent tout de même un bon aperçu de l'état actuel de la recherche et de la diversité des expériences dans le domaine.

IV.1.1 Quelques exemples d'expériences de modélisation à l'échelle du continent européen

Dans le cadre du projet CarHAB, les expériences *a priori* les plus intéressantes à étudier sont celles se déroulant à l'échelle nationale et régionale, correspondant à l'échelle à laquelle la modélisation de la végétation devra être mise en œuvre pour contribuer à la cartographie de la végétation à l'échelle nationale. Cependant, quelques travaux réalisés à l'échelle européenne (zone d'étude = Europe) méritent d'être mentionnés car ils reflètent des expériences réussies de modélisation à très grande échelle spatiale qui peuvent apporter des informations intéressantes pour la modélisation à plus petite échelle. Nous avons sélectionné trois études, que nous présentons ici sous forme d'encadrés.

IV.1.1.1 **Modélisation de la distribution potentielle de huit espèces de végétaux supérieurs à l'échelle européenne** (Huntley et al., 1995).

Objectif : tester si la distribution actuelle et future des végétaux supérieurs à l'échelle continentale est principalement déterminée par le macroclimat.



Données d'observation : données de présence-absence issues de *l'Atlas Florae Europaeae* (Jalas and Suominen, 1972).

Variables environnementales :
climatiques (température, somme des degrés jours supérieurs à 5°C, évapotranspiration…)

Résolution : 50 km.

Méthode de modélisation : régression pondérée localement. Les résultats sont spatialisés sous forme de grille.

Figure 3. Prédiction de la distribution du sapin pectiné (*Abies alba*) à l'échelle européenne.

IV.1.1.2 **Prédiction de la distribution de 61 espèces d'essences européennes avec la plateforme multi-modèles BIOMOD** (Thuiller, 2003).

Objectif : Comparer la performance de différentes approches de modélisation pour prédire la distribution présente d'essences européenne et projeter leur distribution dans le futur.

Données d'observations : données en présence/absence issues de l'*Atlas Florae Europaeae* (Jalas and Suominen, 1972).

Variables environnementales :

Données climatiques issues de la « Climatic Research Unit » (http://www.cru.uea.ac.uk/data).

Résolution : 50 km

Méthode : modèles linéaires généralisés, modèles additifs généralisés, réseaux de neurones et arbres de classification.

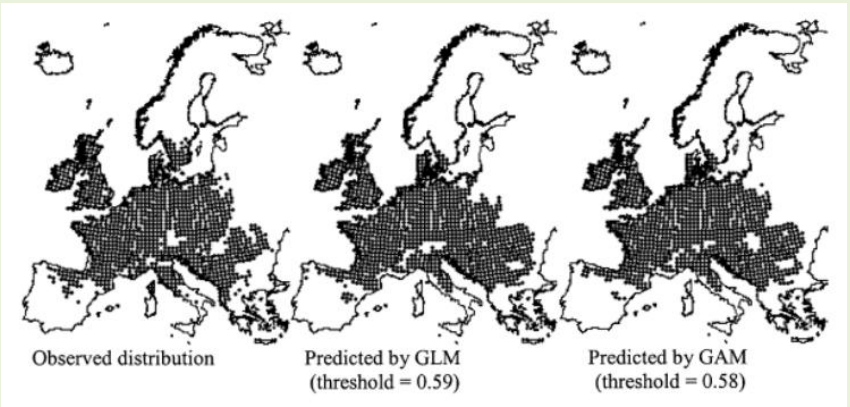


Figure 4. Distribution observée et prédite pour le Chêne sessile (*Quercus petraea*) à l'aide de deux approches de modélisation de la plateforme BIOMOD.

IV.1.1.3 Prédire la distribution de l'habitat 9150 « Hêtraies calcaires », avec différents niveaux de probabilité d'occurrence à l'échelle européenne (Mücher et al., 2009).

Objectif : connaître la distribution géographique et l'étendue des habitats de la Directive en contribuant à la réalisation d'une carte pan-européenne des habitats naturels.

Données :

- Espèces indicatrices de l'habitat issues de l'*Atlas Florae Europaeae* (Jalas and Suominen, 1972) et de la base de données sur la végétation naturelle en Europe (Bohn, 2003).
- Occupation du sol (Corine Land Cover, PELCOM et GLC2000 : voir (Mücher et al., 2009) pour plus de détails)
- Altitude (MNT global pour l'Europe : GTOPO30)
- Grands types de sols (Base de données européenne sur les sols et base de données sols de la FAO)
- Écorégions (carte des régions biogéographiques de l'Europe)

Résolution : 250 m

Méthode : Analyse multicritères. Les résultats sont présentés sous forme de grille.

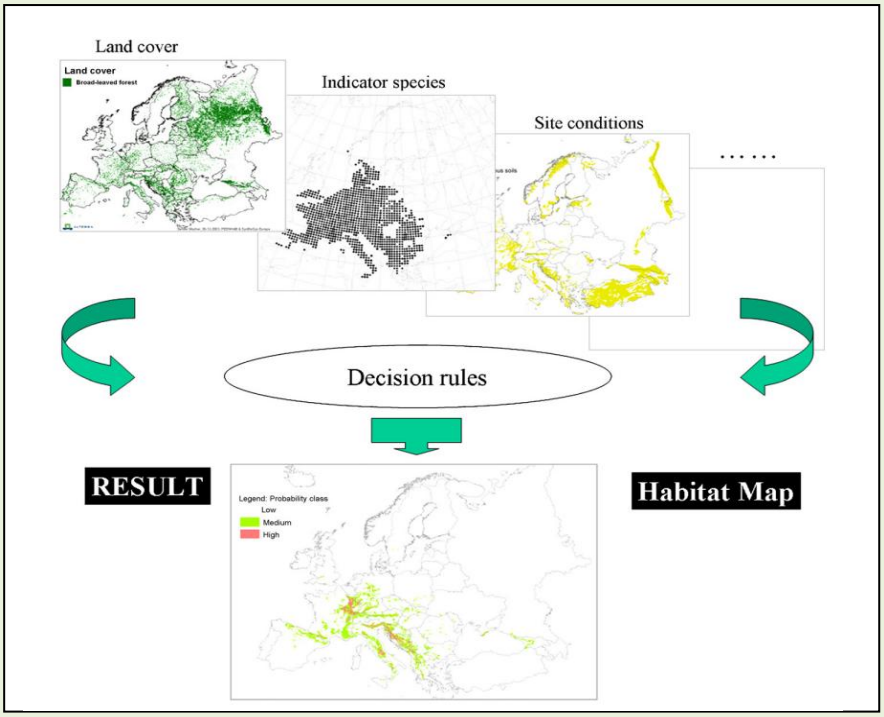


Figure 5. Démarche méthodologique pour la prédiction de la distribution de l'habitat « hêtraies calcaires » à l'échelle européenne (source : Mücher et al., 2009).

IV.1.2 Quelques exemples d'application dans divers pays européens

Après les travaux à l'échelle européenne, nous nous sommes intéressés aux expériences réalisées aux **échelles nationale, régionale (> 300 km²) ou locale**. Afin que ces études puissent alimenter efficacement les discussions dans le cadre du programme CarHAB, nous les avons classées dans la mesure du possible en fonction des grands types de milieux déterminés dans le projet : milieux ouverts d'altitude, milieux ouverts de plaine et milieux forestiers (Tableaux 2, 3 et 4).

Tableau 2. Exemples d'expériences de modélisation de la végétation en Europe à une échelle nationale.

Référence	Pays	Objet d'étude	Résolution	Méthode	Type de milieu
Martínez et al., 2006	Espagne	Lichens	10 km	ENFA	Milieux forestiers
Pearson et al., 2002	Royaume-Uni	23 espèces végétales	5 km	Réseaux de neurones	/
Thuiller, 2003	Portugal	4 essences méditerranéennes	10 km	Classification trees, GLMs, GAMs	Milieux forestiers
Garzón et al., 2006	Péninsule Ibérique (Espagne et Portugal)	1 espèce : *Pinus sylvestris*	1 km	Random forest, réseaux de neurones, arbres de classification et de régression	Milieux forestiers

Tableau 3. Exemples d'expériences de modélisation de la végétation en Europe à une échelle régionale.

Référence	Pays	Région	Surface	Objet d'étude	Objet modélisé	Résolution	Méthode(s)	Type de milieu
Hörsch, 2003	Suisse	Alpes de l'ouest	300 km²	2 alliances de forêts de montagne	Distribution géographique	25m	ACP et PED	Milieux forestiers
Luoto et al., 2002	Finlande	Sud Finlande	601 km²	370 plantes vasculaires	Distribution de la richesse spécifique	0.25 km	GLM	Milieux ouverts de plaine
Zimmermann and Kienast, 1999	Suisse	Alpes suisses	17500 km²	Espèces et communautés végétales dominantes de graminées subalpines	Distribution géographique	50 m	Régression logistique	Milieux ouverts d'altitude
Remm, 2004	Estonie	Sud-Est Estonie	Non spatialisé	Essences forestières	Composition forestière et distribution d'essences forestières	10 m	4 méthodes basées sur des processus d'apprentissage	Milieux forestiers
Münier et al., 2001	Danemark	centre Jutland	6082 km²	Communautés végétales	Distribution géographique	25 m	Analyse multicritères	Milieux ouverts de plaine
Green, 2005	Suède	Bordure Ouest Suède	1300 km²	Communautés végétales	Distribution géographique	300 m	Arbre de classification	Milieux ouverts et forestiers de montagne
Tarkesh and Jetschke, 2012	Allemagne	Thuringe	673.1 km²	Alliance "*Teucrio-Seslerietum*"	Distribution géographique	25 m	BIOCLIM, GARP, Maxent, MARS, Nonparametric multiplicative regression,Logistic regression tree	Milieux ouverts de plaine
Marmion et al., 2009	Finlande	Nord-Est Finlande	41750 km²?	28 plantes vasculaires menacées	Distribution géographique	mailles de 25 ha	Approche consensus (8 méthodes différentes et 5 types de combinaisons de résultats)	Milieux de plaine (boréal)

Tableau 4. Exemples d'expériences de modélisation de la végétation en Europe à une échelle locale.

Référence	Pays	Région	Surface	Objet d'étude	Objet modélisé	Résolution	Méthode	Type de milieu
Dirnböck et al., 2003	Autriche	Nord-est Alpes	60 km²	Communautés végétales	Distribution géographique	4 m	CCA	Milieux ouverts d'altitude
Gottfried et al., 1998	Autriche	Schrankogel	0.65 km²	Espèces et communautés végétales	Distribution géographique	1 m	CCA	Milieux ouverts d'altitude
Guisan et al., 1998	Suisse	Aletsch region	environ 22 km²	1 espèce (*Carex curvula* ssp. *curvula*)	distrubtion géographique de l'abondance	25 m	GLM	Milieux ouverts d'altitude
Skov and Svenning, 2003	Danemark	Jutland	1.65 km² * 2	60 espèces de sous bois	Distribution géographique et diversité	5 m	Analyse multicritère	Milieux forestiers

IV.1.2.1 Expériences de modélisation de la végétation en milieux ouverts d'altitude

Parmi les études recensées, plusieurs portent sur des milieux ouverts d'altitude et sur des communautés végétales proches de celles rencontrées sur la zone test 2012 Belledonne CORA. Trois études localisées dans les Alpes peuvent être retenues :

- **Modélisation de la distribution de l'espèce *Carex curvula ssp. curvula*** (espèce phare de l'association *Caricion curvulae*) **dans la région du Valais, Alpes Suisses** (Guisan et al., 1998). Les données utilisées sont des données d'observation en présence/absence et des variables environnementales majoritairement dérivées du MNT. Différents types de modèles linéaires généralisés sont utilisés pour prédire la distribution de l'espèce. La démarche générale adoptée est synthétisée dans la figure 6, qui illustre bien le principe général de la modélisation spatialisée supportée par SIG.

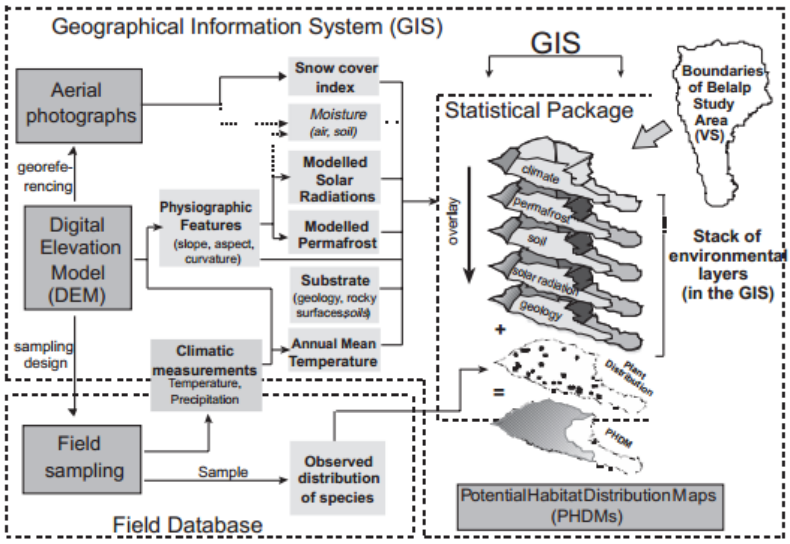


Figure 6. Démarche méthodologique pour la modélisation de la distribution du *Carex curvula* dans les Alpes Suisses (source : Guisan et al., 1998).

- **Prédiction de la distribution géographique d'espèces graminées et de communautés végétales dans les Alpes Suisses** (Zimmermann and Kienast, 1999). Plusieurs espèces appartiennent aux mêmes alliances que celles étudiées sur Belledonne. Les données utilisées sont des données d'observation en présence/absence issues de différentes sources, et un jeu de variables environnementales dérivées d'une base de données climatiques et du MNT (résolution 50 m). La méthode utilisée est la régression logistique. Leurs résultats montrent qu'en zone de montagne, les prédictions sont meilleures pour les communautés végétales que pour les espèces individuelles.

- **Modélisation des communautés végétales dominantes dans les Alpes autrichiennes** (Dirnböck et al., 2003). La modélisation est réalisée à l'aide d'une analyse canonique des correspondances (CCA). Le principal intérêt de cette étude est qu'elle combine une démarche de modélisation « classique » et un processus de segmentation qui permet d'obtenir une carte de végétation sous forme de polygones. Les résultats de la CCA sont superposés aux polygones issus de la segmentation ce qui permet de calculer la proportion de chaque type de végétation modélisé dans chaque polygone pour faciliter leur classification. Cette expérience est un témoignage intéressant à exploiter lors du croisement des résultats de la modélisation et du fond blanc physionomique pour les milieux ouverts d'altitude.

D'autres études comme celles de Gottfried et al. (1998) concernent des milieux ouverts d'altitude proches de ceux que l'on trouve en France.

IV.1.2.2 Expériences de modélisation en milieux forestiers

Plusieurs études concernent la modélisation d'essences forestières ou d'habitats forestiers en plaine ou en montagne. Nous pouvons retenir par exemple :

- **Prédiction de la distribution géographique des forêts de Pin sylvestre (*Pinus sylvestris*) dans la Péninsule Ibérique** (Garzón et al., 2006). Les auteurs comparent la performance de trois méthodes basées sur des processus d'apprentissage pour modéliser la distribution des pinèdes à pin sylvestre (arbres de classification et de régression, réseaux de neurones et Random Forest). Leurs modèles sont basés sur des variables topographiques et climatiques et sur des données d'observation issues d'une cartographie récente des forêts espagnoles. Cette étude représente une des premières expériences de modélisation prédictive spatialisée avec l'algorithme Random Forest (Breiman, 2001), qui s'avère être la méthode la plus performante parmi les approches testées et qui permet également de déterminer l'importance relative des variables impliquées.

- **Prédire la distribution de deux alliances forestières de Montagne (*Vaccinio-Piceion/Larici-Pinetum cembrae* et *Quercion pubescenti-petraeae*),** (Hörsch, 2003). L'originalité de cette étude est que les données d'observation utilisées sont obtenues par classification d'images satellites et de photo-aériennes infrarouges (avec zones d'entraînement terrain), qui permet de distinguer 52 alliances de végétation différentes, dont les deux citées ci-dessus qui sont retenues pour la modélisation. Les variables environnementales utilisées sont majoritairement dérivées du MNT. Ces données sont d'abord utilisées pour déterminer les variables les plus structurantes pour expliquer la distribution des deux alliances, puis leur distribution est modélisée à l'aide d'une méthode de type enveloppe environnementale (méthode Parallel-epiped model). Cette méthode de modélisation donne des résultats très performants et cohérents.

IV.1.2.3 Expériences de modélisation en milieux ouverts de basse altitude

Les études localisées en milieux ouverts de plaine semblent moins représentées que les études en milieux ouverts d'altitude et en milieux forestiers. Nous pouvons citer par exemple l'étude de Luoto et al. (2002) qui concerne les milieux agricoles. Leur objectif est de prédire la distribution de la richesse spécifique totale de la végétation et celle de 16 espèces rares à partir de variables topographiques, d'humidité du sol et de structure d'habitat (taille moyenne des patchs, diversité des habitats…). Ils utilisent des régressions linéaires multiples. Les résultats permettent notamment de localiser les « hotspots » de biodiversité végétale. La variable la plus importante pour expliquer la distribution de la richesse totale en espèces est la diversité des habitats. Pour les espèces rares, ce sont la taille des patchs (petite), la diversité des habitats et l'abondance (faible) en zones construites. Les auteurs indiquent par contre qu'une des limitations de leur étude est l'impossibilité de prendre en compte l'historique d'usage des sols (pâturage, fauche) qui empêchent de distinguer certaines prairies sub-naturelles intéressantes de secteurs fortement anthropisés. Cette conclusion est intéressante dans la mesure où il possible que nous soyons confrontés à la même problématique lors des tests de modélisation en milieux ouverts de plaine.

IV.1.3 Bilan sur les expériences de modélisation de la végétation en Europe

Les études recensées en Europe (hors France métropolitaine) ont été majoritairement réalisées à une échelle régionale (> 300 km²), (8 études). Les expériences aux échelles nationales et locales étant moins nombreuses (4 études dans chaque cas). La résolution spatiale des variables utilisées et des résultats est fortement liée à l'étendue modélisée, la résolution étant beaucoup plus fine (de l'ordre du mètre) pour les études locales que pour les études nationales (de l'ordre du km²). Quand la résolution devient très grossière (ex : 50 km² à l'échelle européenne), les résultats ont tendance à être représentés sous forme de grille, chaque « pixel » devenant une maille de surface importante.

Les trois types de milieux distingués dans le cadre du programme CarHAB sont représentés, avec un plus grand nombre d'études en milieux forestiers (7 études), par rapport aux milieux ouverts d'altitude (5), et aux milieux ouverts de plaine (4).

Une grande diversité d'approches de modélisation est représentée. Quelque soit le type de milieu, les méthodes les plus fréquemment utilisées sont des techniques d'ordination ou des régressions, la plupart des études étant basées sur des données d'observation en présence/absence.

Les données d'observation utilisées sont souvent issues de jeux de données existants, notamment pour les études portant sur de grandes étendues. Par exemple, les trois études recensées à l'échelle européenne utilisent les données de l'*Atlas Florae Europaeae* qui ont une résolution de 50 km. Au contraire, les travaux portant sur de petites surfaces sont plus souvent basés sur des inventaires spécifiques aux objectifs de l'étude. Par exemple, l'étude de Gottfried et al. (1998) est basée sur 1000 placettes d'1 m² où toutes les plantes vasculaires sont recensées. Les variables environnementales utilisées sont presque toujours des données climatiques et des données topographiques dérivées du MNT, peu d'étude intégrant d'autres types de variables comme des variables de sol.

Concernant les syntaxons étudiés, la majorité des études portent sur la modélisation de la distribution d'espèces, deux études concernent des alliances et cinq des communautés végétales définies de différentes manières.

IV.2 Expériences de modélisation de la végétation en France

Très peu d'expériences de modélisation de la végétation ont été menées en France, ou ont en tout cas fait l'objet de publications scientifiques ou de rapports diffusés. La majorité des études recensées dans le cadre de cette synthèse concernent des milieux forestiers (habitats, espèces ou communautés végétales) et ont été initiées dans le cadre de travaux menés par le Laboratoire d'Étude des Ressources FOrêt-Bois (LERFOB), Unité Mixte de Recherche AgroParisTech - INRA de Nancy. Les travaux de modélisation des habitats naturels engagés sur le département de la Seine-et-Marne par le CBNBP constituent une exception notable et sont déjà en cours de développement dans le cadre du programme CarHAB.

IV.2.1 La modélisation de la distribution des habitats forestiers à l'échelle nationale (Bertrand, 2012).

Ce travail de modélisation a été initié dans le cadre du rapportage Natura 2000 à l'horizon 2013. L'objectif est de développer une méthode qui permette d'estimer puis de suivre l'évolution de la surface de 19 habitats forestiers de la Directive. Deux approches sont envisagées pour la modélisation : une modélisation « directe » de la distribution des habitats et une modélisation indirecte à partir de la modélisation de la distribution d'espèces typiques.

La première approche est basée sur l'utilisation de modèles additifs généralisés (GAM), avec une sélection ascendante des variables. Les données d'observation (présence/absence) sont issues de la base de données EcoPlant, où les relevés de végétation sont rattachés à un habitat. Quinze variables environnementales à une résolution de 1 km et reflétant 4 facteurs structurants pour la végétation (température, hydromorphie, disponibilité en eau et nutrition) ont été retenues. Le résultat est une cartographie de la probabilité de présence des habitats à une résolution de 1km qui est ensuite généralisée à une résolution de 10 km (Figure 7).

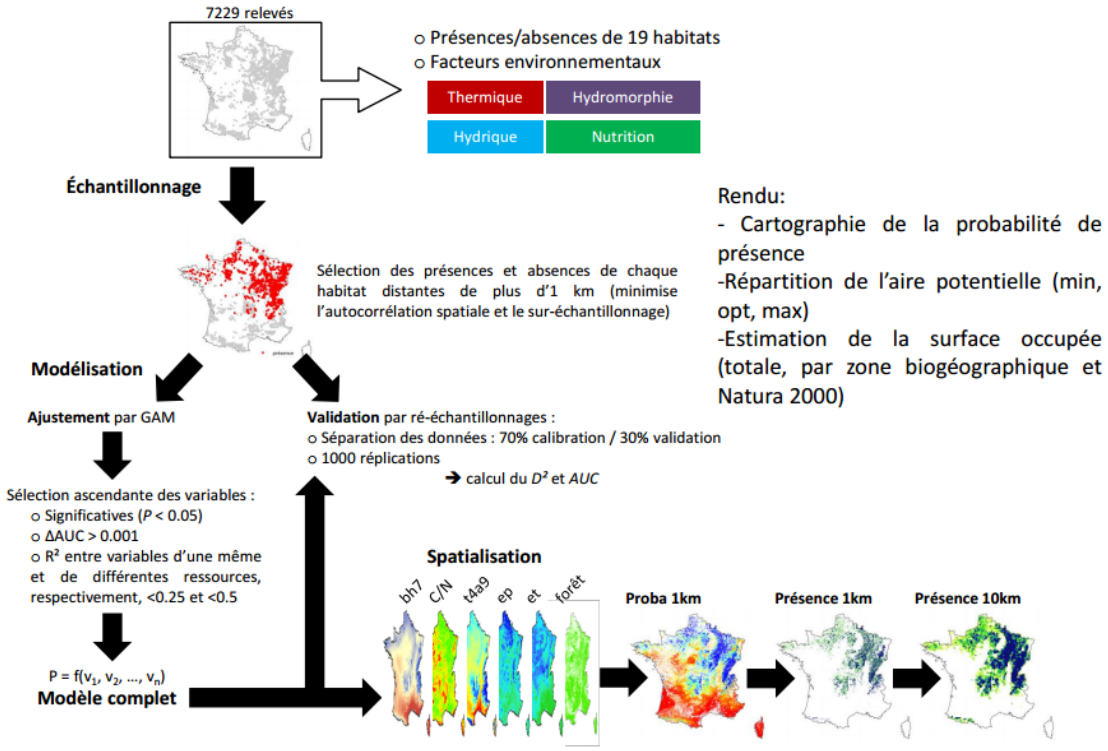


Figure 7. Exemple de méthode de modélisation "directe" de la distribution des habitats forestiers (Source : Bertrand, 2012).

La deuxième approche, « indirecte », consiste à modéliser la distribution d'espèces typiques (sélectionnées comme étant les plus importantes pour différentier les 19 habitats), puis d'assembler les communautés d'espèces typiques modélisées sur la base de connaissances d'expert afin de reconstituer la distribution des habitats. Des informations sur la biogéographie, la topographie et l'occupation du sol sont également utilisées pour l'assemblage des communautés. Les variables environnementales et la méthode de modélisation utilisées sont les mêmes que pour l'approche directe. L'habitat prédit en un lieu donné correspond à celui qui a le plus de probabilité d'être observé d'après la composition floristique prédite, la zone biogéographique, la topographie et l'occupation du sol.

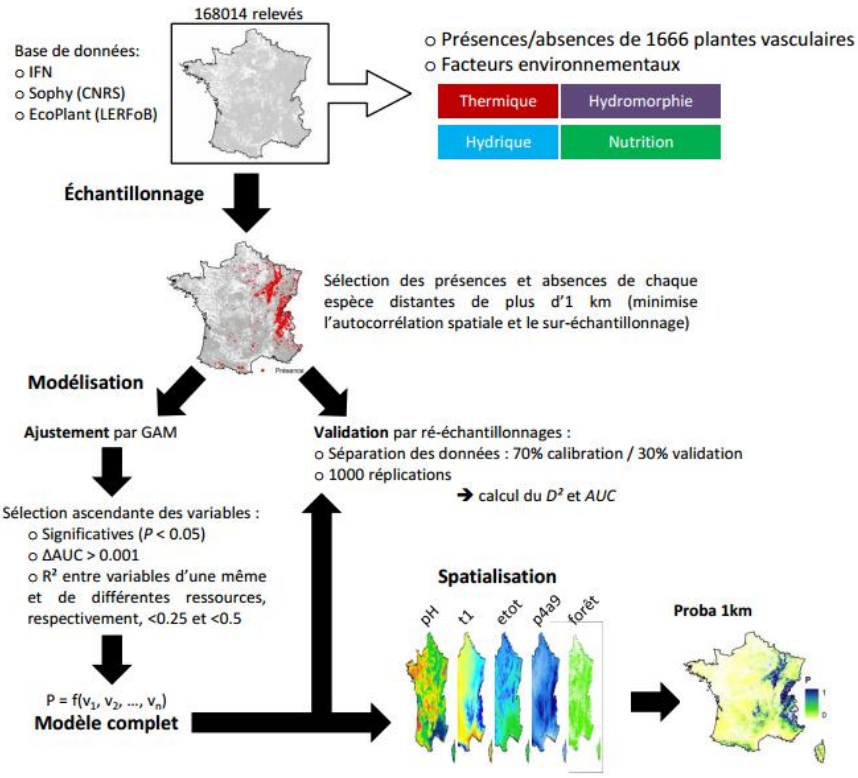


Figure 8. Exemple de méthode de modélisation "indirecte" de la distribution des habitats forestiers (Source : Bertrand, 2012).

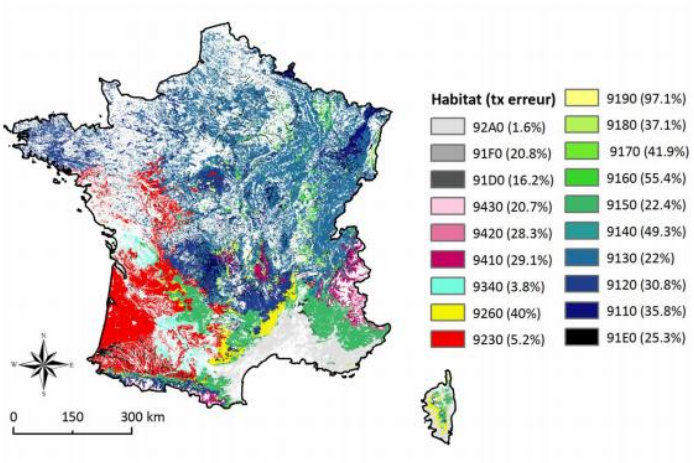


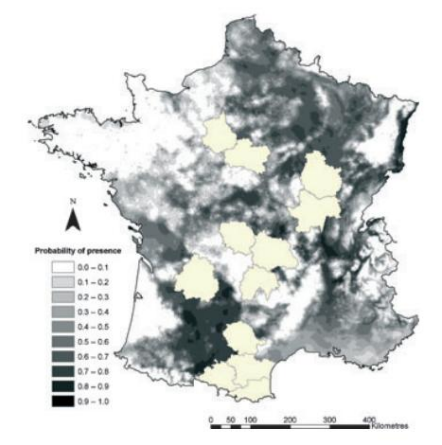
Figure 9. Résultat de la modélisation des habitats avec la méthode « indirecte ».

IV.2.2 Autres travaux du LERFOB, AgroParisTech

Ces travaux n'ont pas pour vocation première de tester ou de développer des méthodes de modélisation de la végétation, mais plutôt de tester l'importance relative de différents facteurs pour expliquer la distribution d'espèces ou d'habitats forestiers ciblés. Ils valent cependant la peine d'être cités car ils impliquent l'utilisation de méthodes de modélisation spatialisées avec des résultats à grande échelle spatiale (nationale, régionale).

IV.2.2.1 Modélisation de la distribution de l'érable champêtre (*Acer campestre*) à l'échelle nationale (Coudun et al., 2006)

Objectif : comparer des modèles climatiques et sol/climat pour la modélisation spatialisée de l'érable champêtre en France.



Données d'observation : données de présence/absence issues de la base de données EcoPlant (Gégout et al., 2005).

Variables explicatives :

- Climatiques basées sur le MNT et la base de données AURHELY de Météo France (Benichou and Le Breton, 1987)
- Édaphiques (pH, C/N…) obtenues par attribution de valeurs de variables décrivant le sol au niveau des relevés de végétation, en fonction de l'écologie des espèces recensées. Ces valeurs sont ensuite interpolées à une résolution de 1 km².

Résolution : 1 km

Méthode de modélisation : régression logistique

Figure 10. Prédiction de la distribution de la présence de l'érable champêtre en France, à partir de variables climatiques et édaphiques. Les zones blanches correspondent à des départements pour lesquels les données floristiques ne sont pas encore disponibles et pour lesquels les auteurs ne pouvaient donc pas obtenir de données de sol (Source : Coudun et al., 2006).

IV.2.2.2 Modélisation de la distribution de l'abondance de la myrtille (*Vaccinium myrtillus*) à l'échelle nationale (Coudun and Gégout, 2007)

Objectif : tester l'importance des données édaphiques par rapport aux données climatiques pour modéliser la distribution de l'abondance de la myrtille à l'échelle nationale.

Données d'observation : données d'abondance issues de la base de données EcoPlant pour la calibration des modèles et données de la base de données SOPHY (PHYtoSOciologie), (Brisse et al., 1995) pour la validation des résultats.

Variables explicatives :

- Climatiques issues de la base de données AURHELY
- Edaphiques (pH, ratio carbone/azote (C/N))

Résolution : 1km

Méthode de modélisation : Adaptation du modèle linéaire généralisé (GLM) pour pouvoir modéliser des classes d'abondance (« *proportional odds ordinal regressions* »).

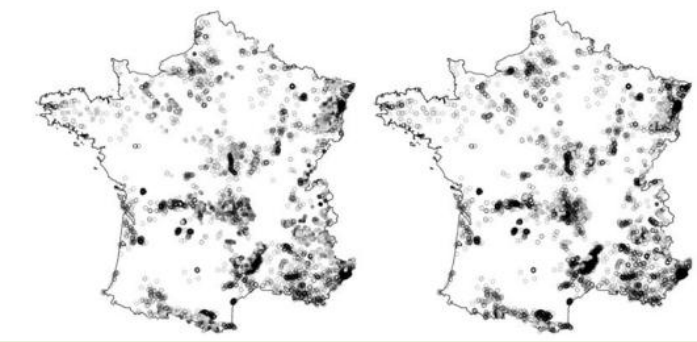


Figure 11. Résultat de la modélisation de la distribution de l'abondance de la myrtille (*Vaccinium myrtillus*) pour deux classes d'abondance. A gauche, probabilité de présence avec un recouvrement > 5% ; à droite, probabilité de présence avec un recouvrement > 50% (Source : Coudun and Gégout, 2007).

Modélisation de la distribution du Panicaut blanc des Alpes (*Eryngium spinalba*) dans le bassin-versant du Petit-Buëch (Marage et al., 2008)

Objectif : tester si la modélisation de la distribution d'une espèce endémique peut être améliorée par la prise en compte de facteurs anthropiques et tester si la distribution de l'espèce dépend d'un gradient de sècheresse.

Données d'observation : données de présence/absence et d'abondance récoltées par échantillonnage aléatoire stratifié.

Variables explicatives :

- Climatiques issues de la base de données AURHELY
- Topographiques issues du MNT à 50m

Résolution : 50 m

Méthode de modélisation : régression logistique et adaptation du modèle linéaire généralisé (GLM) pour pouvoir modéliser des classes d'abondance (« *proportional odds ordinal regressions* »).

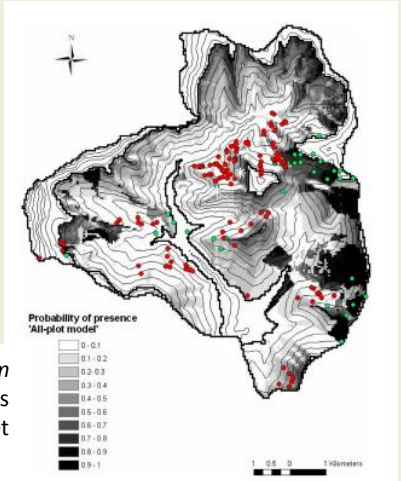


Figure 12. Distribution de l'habitat potentiel du Panicaut blanc des Alpes (*Eryngium spinalba*) dans le bassin-versant du Petit-Buëch (57 km², Hautes-Alpes). Les points colorés correspondent aux données de validation du modèle : points rouges pour les absences et verts pour les présences (source : Marage et al., 2008).

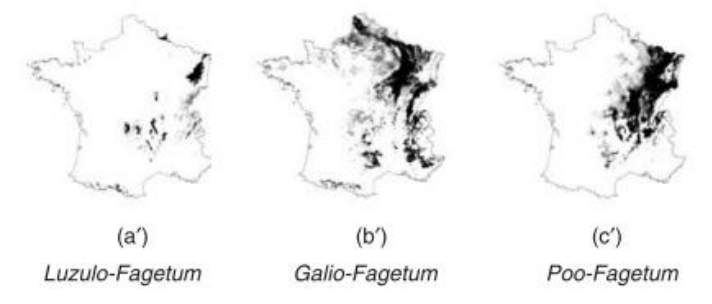
IV.2.2.3 Modélisation de la distribution de six types de hêtraies à l'échelle nationale (Marage and Gégout, 2009)

Objectif : Modéliser la distribution de six communautés végétales à l'échelle nationale, en regardant l'importance des données édaphiques pour expliquer la distribution prédite.

Données d'observation : Présence/absence des habitats : utilisation des Cahiers d'habitats Natura 2000 pour créer des points de relevés « habitats » en cherchant des correspondances entre la composition des relevés floristiques de la base de données EcoPlant et les six habitats étudiés.

Variables explicatives :

- Fonctionnelles (croissance végétation, production primaire, balance hydrique) et climatiques, dérivées de la base de données AURHELY
- Edaphiques (pH, ratio C/N), interpolées à partir des données flore et édaphiques de l'Inventaire Forestier National (IFN) : à chaque relevé, une valeur de pH et de C/N est attribuée selon les optimums recensés pour les espèces composant le relevé.



Résolution : 50 m

Méthode de modélisation : régression logistique

Figure 13. Probabilité de présence de trois types de hêtraies à l'échelle nationale (source : Marage and Gégout, 2009).

IV.2.3 Modélisation de la présence d'habitats naturels en Seine-et-Marne (CBNBP)

Ce travail de modélisation a pour objectif de renseigner des polygones du fond blanc de la Seine-et-Marne, l'Ecomos, pour lesquels il n'existe pas de données de végétation valides. Ce travail est basé sur deux jeux de données cartographiques : l'Ecomos et une carte des végétations naturelles résultant de six ans de prospections terrain et de photo-interprétation. Les analyses statistiques sont réalisées de façon aspatiale, ce qui distingue cette approche des autres expériences présentées jusqu'à présent. Toutefois, elle permet bien d'aboutir à une information géographique sur la végétation.

Les analyses de modélisation sont réalisées par petites régions naturelles, afin d'obtenir un résultat plus précis qu'au niveau du département. Quatre variables environnementales sont utilisées pour caractériser les polygones « végétation » d'une part et les polygones de l'Ecomos d'autre part. Il s'agit de l'altitude, la pente, l'exposition avec une résolution de 25 m et d'un indice de végétalisation, le NDVI, avec une résolution de 30m. Des analyses en composantes principales (ACP) sont réalisées pour projeter les types de végétation puis les polygones Ecomos dans un espace multi-varié. La correspondance entre les types de végétation et les polygones de l'Ecomos se fait en attribuant à un identifiant Ecomos donné l'identifiant du polygone végétation qui en est le plus proche dans l'espace multi-varié (conditions environnementales les plus proches). Le type de végétation est généralement renseigné au niveau alliance ; le niveau « classe » est utilisé lorsque le polygone de l'Ecomos comporte une trop grande diversité d'alliances. Le détail de la méthodologie développée est présenté dans un rapport du CBNBP (Rambaud and Azuelos, 2012).

IV.2.4 Bilan sur les expériences de modélisation de la végétation en France

A l'exception des travaux du CBNBP, toutes les études recensées sont basées sur des données ponctuelles de présence/absence issues des bases de données nationales EcoPlant et SOPHY.

La base de données EcoPlant concerne uniquement les espèces et données phytoécologiques forestières (Gégout et al., 2005 ; https://www2.nancy.inra.fr/unites/lerfob/ecologie-forestiere/bd/ecoplant.htm). La base de données SOPHY correspond à une compilation bibliographique de relevés phytosociologiques principalement réalisés entre 1965 et 2010, et couvre une grande diversité de milieux (Brisse et al., 1995 ; http://sophy.univ-cezanne.fr/sophy.htm). Ces deux bases de données pourraient fournir un bon complément d'information pour la modélisation et la cartographie de la végétation dans le cadre du programme CarHAB ; elles semblent cependant relever du domaine privé et leur accessibilité n'est pas évidente.

Du fait de la disponibilité de données en présence/absence, les méthodes de modélisation utilisées sont toutes de type régression, sauf l'approche du CBNBP qui est de type technique d'ordination.

Les variables environnementales utilisées sont majoritairement issues de bases de données nationales, disponibles en continu sur l'ensemble du territoire français. Il s'agit de la base de données AURHELY de Météo France (Benichou and Le Breton, 1987), du MNT (BD Alti) et de données édaphiques interpolées à partir des bases de données phytoécologiques de l'IFN ou d'EcoPLant, qui permettent toutes les deux de mettre en relation une composition floristique avec des caractéristiques du sol. Dans la majorité des études menées à une échelle nationale, la résolution des variables est de 1 km ; elle descend jusqu'à 25 m pour les études plus locales.

Concernant les syntaxons étudiés, ces travaux concernent principalement des espèces ou des habitats, seuls les travaux du CBNBP portant sur des syntaxons « intermédiaires » comme des alliances ou des classes.

V. Caractéristiques des données d'entrées des modèles : pistes de réflexion pour le choix d'une démarche de modélisation

Dans l'ensemble, l'analyse des travaux recensés montre que la modélisation prédictive spatialisée de la végétation nécessite de définir des objectifs et des échelles de travail précis. Cela est indispensable pour assurer un bon déroulement de la démarche de modélisation qui dépend notamment : 1/ du choix et de la disponibilité de données d'observation et de variables environnementales spatialisées à une échelle (étendue et résolution) adaptée aux objectifs poursuivis, 2/ du savoir-faire dans l'utilisation d'outils de traitement des données, de bureautique, de statistique et de SIG adaptés, et 3/ de la disponibilité d'experts locaux et nationaux sur l'écologie des syntaxons étudiés. En particulier, la plupart des études recensées sont basées sur des jeux de données pré-existants dans des bases de données, et de type présence/absence. D'autre part, les variables environnementales utilisées sont souvent de même type (climat, topographie, sol…), mais leur nombre et la nature des variables retenues par les modèles diffère fortement d'une étude à l'autre, en fonction des objectifs poursuivis. Ces caractéristiques ne sont pas anodines car la nature des données disponibles et la phase de préparation des données d'entrée des modèles est particulièrement déterminante pour la bonne marche du processus de modélisation.

V.1 Données d'observation

La très grande majorité des approches de modélisation prédictive spatialisée recensées dans cette synthèse sont basées sur des données d'observation ponctuelles géolocalisées (latitude et longitude connues). Ces données proviennent majoritairement d'inventaires déjà existants, et il peut être difficile de déterminer si l'échantillonnage disponible est suffisant et représentatif de la distribution de l'espèce étudiée. Or, le type de données disponibles (catégorielles ou continues, présence ou présence/absence), leur nombre et leur répartition peuvent fortement influencer la pertinence des prédictions et doivent donc être pris en compte dans le choix d'une approche de modélisation. Nous détaillons ici les implications sur la démarche de modélisation de deux principales caractéristiques des données d'observation : 1/ la taille et la représentativité de l'échantillonnage et 2/ la disponibilité en données d'absence.

V.1.1 La taille et la représentativité de l'échantillonnage

Le nombre de données d'observation disponibles et leur représentativité vis-à-vis de la niche écologique de l'espèce étudiée sont déterminants pour la précision des prédictions (Guisan and Zimmermann, 2000; Foody, 2011; Sánchez-Fernández et al., 2011). En particulier, la question de la taille de l'échantillon peut être une problématique essentielle dans le cas de la conservation d'espèces ou d'habitats rares, pour lesquels les données d'observation sont peu nombreuses et souvent agrégées dans l'espace (Kumar and Stohlgren, 2009).

La plupart des techniques de modélisation peuvent donner de bons résultats avec relativement peu de points d'observation, dans la mesure où ceux-ci reflètent bien l'ensemble des conditions environnementales dans lesquelles l'espèce ou l'habitat étudiés peuvent être présents (Stockwell

and Peters, 1999; Elith et al., 2006). Les études de Barbet-Massin et al. (2010), Thuiller et al. (2004) et de Sánchez-Fernández et al. (2011), par exemple, montrent qu'un manque de représentativité des observations affecte fortement la fiabilité des résultats de modélisation. En fait, le nombre de points d'observations nécessaires pour obtenir des résultats robustes dépend en partie de l'amplitude de la niche écologique de l'espèce étudiée : plus l'espèce aura une niche large, plus il sera difficile d'obtenir un échantillonnage qui couvre l'ensemble des conditions écologiques qui lui sont favorables (Stockwell and Peterson, 2002). Au contraire, les modèles seront souvent plus précis pour les espèces dont la niche est restreinte, et pour lesquelles un faible nombre de points peut suffire à représenter l'ensemble de la niche de l'espèce (Hernandez et al., 2006). Certaines approches de modélisation, toutefois, sont plus robustes que d'autres à la faible taille de l'échantillonnage. La méthode Maxent, par exemple, est une des plus robustes à cette contrainte (Baldwin, 2009) et plusieurs études ont obtenus de bons résultats avec très peu de points d'observation (ex : 5 points (Pearson, Raxworthy, Nakamura, and Peterson, 2007) ; 5, 10 et 15 points (Hernandez et al., 2006) ; 11 points (Kumar and Stohlgren, 2009)), (voir également Hernandez et al., 2008). L'algorithme GARP semble également bien fonctionner avec peu de données (Stockwell and Peterson, 2002; Wisz et al., 2008). Au contraire, la méthode BIOCLIM, la régression logistique, les modèles additifs généralisés et les modèles basés sur des approches de type *boosting*[1] semblent assez mal gérer cette contrainte (Stockwell and Peterson, 2002; Hernandez et al., 2006; Wisz et al., 2008).

Certains auteurs ont cherché à déterminer le nombre minimal de points d'observation nécessaires à l'obtention de résultats cohérents. Le travail de Stockwell (2002), par exemple, montre qu'à partir de 30 à 40 données la précision des prédictions se stabilise, pour quatre types de modèles incluant la régression logistique et l'algorithme GARP. Les résultats de Wisz et al. (2008) montrent également qu'il faut au moins 30 points d'observation pour obtenir des résultats cohérents avec 12 méthodes différentes. Ceci est toutefois mitigé par les résultats de Pearce (2000) qui trouvent qu'il faut au moins 250 observations pour produire un modèle qui commence à être assez précis avec des méthodes de type GLM et GAM, ce qui confirme que les approches de type régression gèrent plutôt mal les petits jeux de données.

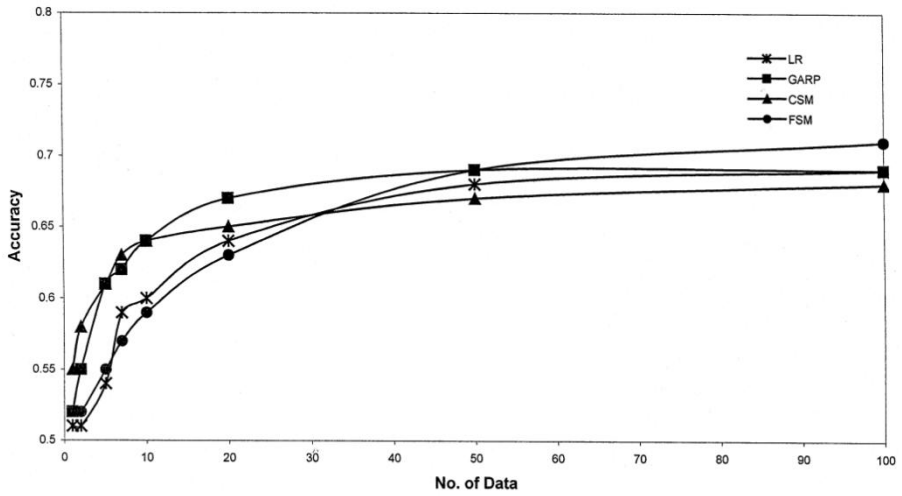


Fig. 2. General relationship between accuracy and sample size, across 103 species re-sampled for six sample sizes with the LR, FSM, CSM, and GARP methods.

[1] Voir Tableau 1.

Figure 14. Relation générale entre la précision des prédictions et la taille de l'échantillonnage, évaluée pour 103 espèces, 6 tailles d'échantillon et quatre approches de modélisation différentes (source : Stockwell and Peterson, 2002).

Ces études donnent donc des pistes de réponse à la question complexe de la taille minimale optimale de l'échantillonnage, mais il n'existe pas à ce jour de consensus sur le nombre minimal de points nécessaires pour la modélisation. Il s'agit surtout de veiller à la bonne représentativité des observations en fonction des connaissances sur l'écologie de l'espèce étudiée et de réaliser un complément d'échantillonnage dans les conditions peu représentées si besoin.

V.1.2 La disponibilité en données d'absence

Nous avons vu que la grande majorité des études recensées dans cette synthèse sont basées sur des données d'observation en présence/absence. La disponibilité en données d'absence est indispensable pour certaines approches de modélisation comme les méthodes de type régression, les réseaux de neurones et les techniques de classification (Tableau 1). Cependant, il est largement reconnu que lorsque des données d'absence sont disponibles, elles doivent être considérées avec précaution car de nombreux facteurs peuvent conduire à la non détection d'une espèce qui est en réalité présente (ex : espèces nocturnes, rares, méthode d'échantillonnage non adaptée, mauvaise détermination) (Hirzel et al., 2002; Hirzel and Le Lay, 2008; Lobo et al., 2010). De plus, lorsque les données d'observation utilisées proviennent de bases de données existantes (ex : herbiers, observations, inventaires), ce qui est souvent le cas, elles ont souvent une pression d'échantillonnage trop faible et leur fiabilité n'est généralement pas suffisante pour en déduire des données d'absence (Anderson et al., 2003; Chefaoui and Lobo, 2008). Or, l'utilisation de « fausses » données d'absence peut fortement altérer la fiabilité des modèles de qualité d'habitats (Chefaoui and Lobo, 2008).

Dans ce contexte, des méthodes de modélisation capables de réaliser des prédictions uniquement avec des données de présence ont été développées. Ces méthodes caractérisent la niche écologique de l'espèce en comparant les conditions écologiques au niveau des points d'observation (valeurs des variables environnementales) aux conditions au niveau de situations aléatoires (*background data*) ; ces dernières correspondant à l'hypothèse où l'espèce n'aurait pas de préférence quant aux caractéristiques environnementales de son habitat (Brotons et al., 2004; Phillips et al., 2009). De nombreuses méthodes fonctionnent selon ce principe ; par exemple l'ENFA (Ecological Niche Factor Analysis) (Hirzel et al., 2002; Chefaoui and Lobo, 2008; Poirazidis et al., 2010), DOMAIN (Carpenter et al., 1993), BIOCLIM (Hirzel et al., 2002), MaxEnt (Phillips et al., 2006) ou la distance de Mahalanobis (Farber and Kadmon, 2003), (Tableau 1). La synthèse d'Elith et al. (2006) propose une comparaison de la performance de 16 méthodes de modélisation utilisant des données en présence seule pour la prédiction spatialisée de distributions potentielles d'espèces.

Toutefois, plusieurs études suggèrent que lorsque les données disponibles le permettent, les méthodes de modélisation basées sur des données de présence et d'absence devraient être privilégiées par rapport aux approches en présence seule (Zaniewski et al., 2002; Wisz and Guisan, 2009; Tarkesh and Jetschke, 2012). Certaines situations semblent particulièrement justifier ce choix, comme par exemple lorsqu'on s'intéresse à la distribution d'espèces ou d'habitats avec une niche écologique étendue (Brotons et al., 2004). Cependant, il n'y a pas de consensus à ce sujet ; la méthode en présence seule Maxent, par exemple, donne régulièrement des résultats équivalents à ceux d'approches en présence-absence (Gormley et al., 2011; Tarkesh and Jetschke, 2012). D'autre

part, les méthodes en présence seule permettent toujours de s'affranchir des risques encourus à utiliser de fausses données d'absence.

V.2 Variables environnementales

Toutes les études recensées reposent sur des bases de données environnementales à grande échelle spatiale, souvent dérivées 1/ de bases de données climatiques européennes, nationales ou plus locales et 2/ d'un MNT à des échelles et résolutions variables, qui permet de calculer de nombreuses variables topographiques (pente, orientation, rayonnement solaire, positionnement topographique…). La résolution de ces données est généralement liée à l'étendue de la zone étudiée : faible résolution pour les études à l'échelle nationale (de l'ordre du km²), forte résolution pour les études locales (de l'ordre du m²). Le format des variables environnementales (catégorielles, continues, booléennes…) varie également d'une étude à l'autre et peut contribuer à déterminer le type de méthode de modélisation qui peut être utilisée (Tableau 1).

Le nombre de variables qui peuvent être utilisées pour les prédictions peut être très grand, certaines études intégrant près de 50 variables différentes dans l'analyse. Cependant, le choix d'un nombre raisonnable de variables (notion de parcimonie) qui soient pertinentes par rapport à l'objet étudié et l'objectif recherché est un facteur fondamental pour s'assurer de la bonne performance statistique et de la bonne cohérence écologique des modèles (Guisan and Zimmermann, 2000; Pearce and Ferrier, 2000). Plusieurs études montrent notamment que de diminuer le nombre de variables permet d'augmenter le pouvoir de prédiction des modèles et réduit fortement le temps de calcul nécessaire au modèle (Guisan and Zimmermann, 2000; Remm, 2004). Toutefois, la sélection des variables les plus pertinentes et l'élimination des variables inutiles n'est pas toujours évidente.

Le choix des variables environnementales explicatives passe généralement par une sélection réalisée *a priori* à l'aide de bonnes connaissances sur l'écologie de l'espèce étudiée (connaissances expertes ou étude bibliographique) puis est suivie par une étape de sélection statistique en amont ou lors de la mise en œuvre de l'algorithme de modélisation. Pour la phase de sélection statistique, la plupart des méthodes de modélisation permettent de comparer la performance de différents modèles ou d'inclure et d'exclure des variables avec des procédures pas à pas (ex : méthodes de type régression). Les paramètres des modèles ou les sorties de l'algorithme permettent alors de connaître l'importance relative de chaque variable, leur pourcentage de contribution au modèle ainsi que le sens et la significativité de leur effet sur la distribution prédite, ce qui permet de classer les variables explicatives en fonction de leur pertinence pour le modèle. A charge ensuite au modélisateur de ne garder que les plus pertinentes pour le modèle final. On peut distinguer deux principaux mécanismes de classement statistique des variables explicatives :

- La sélection *a priori* « forward selection » : le modèle initial ne contient aucune variable, puis les variables sont ajoutées une par une dans le modèle et on regarde l'évolution de sa performance à chaque nouvelle variable,
- La sélection *a posteriori* « backward selection » : toutes les variables sont initialement incluses dans le modèle, puis celles qui ne sont peu corrélées à la variable réponse sont éliminées une par une jusqu'à trouver la configuration qui assure la meilleure performance (Pearce and Ferrier, 2000).

On retrouve ces mécanismes dans la plupart des méthodes de modélisation, bien que peut-être de façon plus « explicite » pour les méthodes de modélisation de type régression.

Le choix de la méthode de modélisation peut alors dépendre de la facilité avec laquelle on a accès aux paramètres des modèles et à des informations sur l'importance relative des variables. Ces informations sont assez bien accessibles dans la plupart des approches de modélisation. Dans certains cas, cependant, l'accès à ces informations est plus difficile, comme pour la distance de Mahalanobis, par exemple, pour laquelle il n'existe pas de tests de significativité ou de magnitude pour déterminer le pouvoir explicatif des variables impliquées dans le modèle (Griffin et al., 2010). L'accès aux variables et aux paramètres structurants des modèles sont également compliqués dans le cas des réseaux de neurones (Garzón et al., 2006).

VI. Conclusions et implications pour la modélisation de la végétation dans le cadre du programme CarHAB

Cette synthèse bibliographique donne un aperçu de la diversité des méthodes disponibles pour la modélisation prédictive spatialisée de la végétation, avec leurs principales caractéristiques et contraintes de mise en œuvre. Une grande partie de ces méthodes a été appliquée en Europe et en France, avec de nombreux points communs quant à la nature des données d'observations et des variables environnementales utilisées. Ces expériences mettent également en avant l'existence de bases de données européennes et nationales qui pourraient s'avérer utiles dans le cadre du programme CarHAB.

Concernant les données d'observation, la majorité des études recensées utilisent des méthodes de modélisation adaptées au traitement de jeux de données en présence/absence. Cependant, dans le cadre du projet CarHAB, **la modélisation devra certainement s'orienter vers l'utilisation de jeux de données en présence seule**. En effet, d'une part, les jeux de données utilisés proviendront certainement des bases de données des Conservatoires botaniques ou d'autres organismes, où les relevés de végétation reflètent généralement la présence des espèces mais n'ont pas pour objectif de déterminer leur absence. D'autre part, il est déjà souvent hasardeux de déterminer des données d'absence au niveau de l'espèce (cf. paragraphe 4.1.1.2), cela paraît donc d'autant plus complexe au niveau d'un syntaxon qui peut être présent sous forme dégradée ou modifiée, ou en situation de transition selon les conditions du milieu.

Par ailleurs, les jeux de données d'observation doivent être **homogènes sur l'ensemble de la zone d'étude et représentatifs** de l'ensemble des conditions environnementales dans lesquelles le syntaxon étudié peut être observé. Cela demande de rassembler un grand nombre de données ce qui peut demander un gros travail d'harmonisation si les données proviennent de sources diverses (ex : de différents conservatoires botaniques). Les expériences de modélisation en France mettent en avant l'existence de bases de données phyto-écologique, les bases de données EcoPlant pour les milieux forestiers et Sophy, qui pourraient apporter un bon complément de données d'observation de la flore. Cependant, elles semblent relever du domaine privé et les modalités de leur accessibilité restent à définir. La base de données floristique de l'IFN peut aussi apporter un bon complément d'information, et permet également de mettre en relation une composition floristique avec des

caractéristiques édaphiques et des peuplements forestiers. Le seul bémol est la localisation approximative des points de relevés. Au niveau national, d'autres bases de données centralisées comme celle prévue dans le projet VegFrance pourront à l'avenir apporter un appui supplémentaire. Au niveau européen, il existe plusieurs bases de données comme l'*Atlas Florae Europaeae*, qui est très complète du point de vue de la diversité des espèces représentées mais sa résolution (50 km) est beaucoup trop grossière pour être utilisée comme support de cartographie des habitats à l'échelle nationale. Par contre, d'autres bases de données ou de métadonnées en cours de création, comme les projets SynBioSys et EVA (European Vegetation Archive), pourront éventuellement enrichir les informations disponibles dans les années à venir.

Concernant les variables environnementales, les études recensées montrent l'importance de disposer de jeux de données suffisamment complets (diversité de variable permettant de refléter l'ensemble des caractéristiques du milieu expliquant la distribution des syntaxons étudiés) et précis (résolution adaptée au type de milieu et de végétation étudié) pour modéliser la végétation. En particulier, **trois principaux types de données sont utilisés de façon récurrente pour la modélisation de la végétation : données climatiques, topographiques et édaphiques**. En France, ces variables peuvent s'appuyer sur des bases de données nationales telles que la base de données AURHELY de Météo France, le MNT issu de la BD alti et des bases de données phyto-écologiques comme celles de l'IFN ou EcoPlant. Pour permettre des prédictions cohérentes, ces données de base doivent **être disponibles** sur l'ensemble de la zone de modélisation, c'est-à-dire **sur l'ensemble du territoire national sans discontinuités, à une résolution cohérente avec les objectifs fixés**. Par exemple, la résolution est souvent de l'ordre du km² pour la modélisation à l'échelle nationale et concernant des habitats ou des syntaxons largement répandus (ex : grands types de formations forestières), elle peut par contre fortement augmenter pour les syntaxons localisés (ex : 50 m), et devrait être d'autant plus élevée que les milieux concernés sont hétérogènes et les syntaxons étudiés présents de façon ponctuelle.

Table des figures

Figures

Tableaux

Bibliographie

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modelling 162, 211–232.

Araújo, M.B., New, M., 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22, 42–47.

Aspinall, R., 1992. An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data. International journal of geographical information systems 6, 105–121.

Baldwin, R.A., 2009. Use of Maximum Entropy Modeling in Wildlife Research. Entropy 11, 854–866.

Barbet-Massin, M., Thuiller, W., Jiguet, F., 2010. How much do we overestimate future local extinction rates when restricting the range of occurrence data in climate suitability models? Ecography 33, 878–886.

Benichou, P., Le Breton, O., 1987. Prise en compte de la topographie pour la cartographie des champs pluviométriques statistiques. Météorologie 19, 23–24.

Bertrand, R., 2012. WP1-Modélisation des habitats forestiers par approche directe et indirecte., COPIL programme TEECH, rapportage européen 2013. LERFOB, AgroParisTech.

Breiman, L., 2001. Random Forest. Machine Learning 45, 5–32.

Brisse, H., De Ruffray, P., Grandjouan, G., Hoff, M., 1995. The phytosociological Database "SOPHY" Part I: Calibration of indicator plants Part II: Socio-ecological classification of the relevés. Annali di Botanica LIII, 177–190.

Brotons, L., Thuiller, W., Araujo, M.B., Hirzel, A.H., 2004. Presence-absence versuspresence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448.

Busby, J.R., 1991. BIOCLIM: A bioclimate analysis and prediction system, in: Margules, C.R. (Ed.), Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Australia, pp. 64–68.

Carpenter, G., Gillison, A.N., Winter, J., 1993. DOMAIN: a flexible modelling procedure for mapping potential distribution of plants and animals. Biodiversity and Conservation 2, 667–680.

Chefaoui, R.M., Lobo, J.M., 2008. Assessing the effects of pseudo-absences on predictive distribution model performance. Ecological Modelling 210, 478–486.

Coudun, C., Gégout, J.C., 2007. Quantitative prediction of the distribution and abundance of Vaccinium myrtillus with climatic and edaphic factors. Journal of Vegetation Science 18, 517–524.

Coudun, C., Gégout, J.C., Piedallu, C., Rameau, J.C., 2006. Soil nutritional factors improve models of plant species distribution: an illustration with Acer campestre (L.) in France. Journal of Biogeography 33, 1750–1763.

De'ath, G., Fabricius, K.E., 2000. Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. Ecology 81, 3178–3192.

Debinski, D.M., Kindscher, K., Jakubauskas, M.E., 1999. A remote sensing and GIS-based model of habitats and biodiversity in the Greater Yellowstone Ecosystem. International Journal of Remote Sensing 20, 3281–3291.

Decout, S., Manel, S., Miaud, C., Luque, S., 2012. Integrative approach for landscape-based graph connectivity analysis: a case study with the common frog (Rana temporaria) in human-dominated landscapes. Landscape Ecology 27, 267–279.

Dirnböck, T., Dullinger, S., Gottfried, M., Ginzler, C., Grabherr, G., 2003. Mapping alpine vegetation based on image analysis, topographic variables and Canonical Correspondence Analysis. Applied Vegetation Science 6, 85–96.

Elith, J., Graham, C.H., Anderson, R.P., Dudı´k, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. Overton, J., Peterson, A.T., Phillips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Sobero´n, J., Williams, S., Wisz, M.S., Zimmermann, N.E.,

2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.
Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution 1, 330–342.
Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77, 802–813.
Ewers, R.M., Didham, R.K., Wratten, S.D., Tylianakis, J.M., 2005. Remotely sensed landscape heterogeneity as a rapid tool for assessing local biodiversity value in a highly modified New Zealand landscape. Biodiversity and Conservation 14, 1469–1485.
Farber, O., Kadmon, R., 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecological Modelling 160, 115–130.
Fearer, T.M., Prisley, S.P., Stauffer, D.F., Keyser, P.D., 2007. A method for integrating the Breeding Bird Survey and Forest Inventory and Analysis databases to evaluate forest bird–habitat relationships at multiple spatial scales. Forest Ecology and Management 243, 128–143.
Fisher, R.A., 1936. "The Use of Multiple Measurements in Taxonomic Problems."
Fleishman, E., Mac Nally, R., Fay, J.P., Murphy, D.D., 2001. Modeling and predicting species occurrence using broad-scale environmental variables: an example with butterflies of the Great Basin. Conservation Biology 15, 1674–1685.
Foody, G.M., 2011. Impacts of imperfect reference data on the apparent accuracy of species presence-absence models and their predictions. Global Ecology and Biogeography 20, 498–508.
Franklin, J., 1995. Predictive vegetation mapping: geographic modelling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19, 474–499.
Friedman, J.H. 1991. "Multivariate Adaptive Regression Splines." The Annals Os Statistics 19, 1–141.
Garzón, M.B., Blazek, R., Neteler, M., Dios, R.S. de, Ollero, H.S., Furlanello, C., 2006. Predicting habitat suitability with machine learning models: The potential area of Pinus sylvestris L. in the Iberian Peninsula. Ecological Modelling 197, 383–393.
Gégout, J.C., Coudun, C., Bailly, G., Jabiol, B., 2005. EcoPlant: a forest site database linking floristic data with soil and climate variables. Journal of Vegetation Science 16, 257–260.
Gibson, L.A., Wilson, B.A., Cahill, D.M., Hill, J., 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. Journal of Applied Ecology 41, 213–223.
Godet, L., Devictor, V., Jiguet, F., 2007. Estimating relative population size included within protected areas. Biological Conservation 16, 2587–2598.
Gormley, A.M., Forsyth, D.M., Griffioen, P., Lindeman, M., Ramsey, D.S.L., Scroggie, M.P., Woodford, L., 2011. Using presence-only and presence-absence data to estimate the current and potential distributions of established invasive species. Journal of Applied Ecology 48, 25–34.
Gottfried, M., Pauli, H., Grabherr, G., 1998. Prediction of vegetation patterns at the limits of plant life: a new view of the alpine-nival ecotone. Arctic and Alpine Research 30, 207–221.
Green, J.A., 2005. An application of predictive vegetation mapping to mountain vegetation in Sweden.
Griffin, S.C., Taper, M.L., Hoffman, R., Mills, L.S., 2010. Ranking Mahalanobis Distance Models for Predictions of Occupancy From Presence-Only Data. Journal of Wildlife Management 74, 1112–1121.
Grinnell, J., 1917. The niche relationships of the California Thrasher. The Auk 34, 427–433.
Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological Modelling 157, 89–100.
Guisan, A., Theurillat, J.P., Kienast, F., 1998. Predicting the potential distribution of plant species in an alpine environment. Journal of Vegetation Science 9, 65–74.
Guisan, A., Weiss, S.B., Weiss, A.D., 1999. GLM versus CCA spatial modeling of plant species distribution. Plant Ecology 143, 107–122.
Guisan, A., Zimmermann, N.E., 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135, 147–186.

Gurnell, J., Clark, M.J., Lurz, P.W., Shirley, M.D., Rushton, S.P., 2002. Conserving red squirrels (Sciurus vulgaris): mapping and forecasting habitat suitability using Geographic Information Systems approach. Biological Conservation 105, 53–64.

Hagan, J.M., Meehan, A.L., 2002. The effectiveness of stand-level and landscape-level variables for explaining bird occurrence in an industrial forest. Forest Science 48, 231–242.

Hernandez, P.A., Franke, I., Herzog, S.K., Pacheco, V., Paniagua, L., Quintana, H.L., Soto, A., Swenson, J.J., Tovar, C., Valqui, T.H., Vargas, J., Young, B.E., 2008. Predicting species distributions in poorly-studied landscapes. Biodiversity and Conservation 17, 1353–1366.

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83, 2027–2036.

Hirzel, A.H., Le Lay, G., 2008. Habitat suitability modelling and niche theory. Journal of Applied Ecology 45, 1372–1381.

Hörsch, B., 2003. Modelling the spatial distribution of montane and subalpine forests in the central Alps using digital elevation models. Ecological Modelling 168, 267–282.

Hu, J., Jiang, Z., 2010. Predicting the potential distribution of the endangered Przewalski's gazelle. Journal of Zoology 282, 54–63.

Huntley, B., Berry, P.M., Cramer, W., Alison P. McDonald, 1995. Special Paper: Modelling Present and Potential Future Ranges of Some European Higher Plants Using Climate Response Surfaces. Journal of Biogeography 22, 967–1001.

Hutchinson, G.E., 1959. Homage to Santa Rosalia or why are there so many kinds of animals? American Naturalist 93, 145–159.

Jalas, J., Suominen, J., 1972. Atlas Florae Europaeae. Vol. 1-10._, The commitee for mapping the flora of Europe and Societas Biologica Fennica Vanamo. Jalas, J. & Suominen, J., Helsinki.

Jetz, W., McPherson, J.M., Guralnick, R.P., 2012. Integrating biodiversity distribution knowledge: toward a global map of life. Trends in Ecology & Evolution 27, 151–159.

Jiménez-Valverde, A., Peterson, A.T., Soberón, J., Overton, J.M., Aragón, P., Lobo, J.M., 2011. Use of niche models in invasive species risk assessments. Biological Invasions 13, 2785–2797.

Johnson, C.J., Gillingham, M.P., 2005. An evaluation of mapped species distribution models used for conservation planning. Environmental Conservation 32, 117–128.

Kafley, H., Khadka, M., Sharma, M., 2009. Habitat evaluation and suitability modeling of Rhinoceros Unicornis in Chitwan National Park, Nepal: A geospatial approach. Presented at the XIII World Forestry Congress, Buenos Aires, Argentina, 18 – 23 October 2009, p. 12.

Kerr, J.T., Southwood, T.R.E., Cihlar, J., 2001. Remotely sensed habitat diversity predicts butterfly species richness and community similarity in Canada. Proceedings of the National Academy of Sciences of the United States of America 98, 11365–11370.

Kumar, S., Stohlgren, T.J., 2009. Maxent modeling for predicting suitable habitat for threatened and endangered tree Canacomyrica monticola in New Caledonia. Journal of Ecology and natural Environment 1, 94–98.

Lobo, J.M., Jiménez-Valverde, A., Hortal, J., 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33, 103–114.

Luoto, M., Pöyry, J., Heikkinen, R.K., Saarinen, K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. Global Ecology and biogeography 14, 575–584.

Luoto, M., Toivonen, T., Heikkinen, R.K., 2002. Prediction of total and rare plant species richness in agricultural landscapes from satellite images and topographic data. Landscape Ecology 17, 195–217.

Luque, S., Vainikainen, N., 2008. Habitat quality assessment and modelling for forest biodiversity and sustainability, in: Lafortezza, R. (Ed.), Patterns and Processes in Forest Landscapes - Multiple Use and Sustainable Management. Springer, pp. 241–264.

Mac Faden, S.W., Capen, D.E., 2002. Avian habitat relationships at multiple scales in a New England Forest. Forest Science 48, 243–253.

Marage, D., Garraud, L., Rameau, J.C., 2008. The influence of management history on spatial prediction of Eryngium spinalba, an endangered endemic species. Applied Vegetation Science 11, 139–148.

Marage, D., Gégout, J.-C., 2009. Importance of soil nutrients in the distribution of forest communities on a large geographical scale. Global Ecology and Biogeography 18, 88–97.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R.K., Thuiller, W., 2009. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15, 59–69.

Martínez, I., Carreño, F., Escudero, A., Rubio, A., 2006. Are threatened lichen species well-protected in Spain? Effectiveness of a protected areas network. Biological Conservation 133, 500–511.

Moffett, A., Sarkar, S., 2006. Incorporating multiple criteria into the design of conservation area networks: a minireview with recommendations. Diversity and Distributions 12, 125–137.

Moore, N.W., Hooper, M.D., 1975. On the number of bird species in British woods. Biological Conservation 8, 239–250.

Mücher, C.A., Klijn, J.A., Bunce, R.G.H., Schaminée, J.H.J., Schaepman, M.E., 2009. Modelling the Spatial Distribution of Natura 2000 habitats across Europe. Landscape and Urban Planning 92, 148–159.

Münier, B., Nygaard, B., Ejrn\aes, R., Bruun, H.G., 2001. A biotope landscape model for prediction of semi-natural vegetation in Denmark. Ecological Modelling 139, 221–233.

Nabout, J.C., Soares, T.N., Diniz-Filho, J.A.F., De Marco Júnior, P., Telles, M.P.C., Naves, R.V., Chaves, L.J., 2010. Combining multiple models to predict the geographical distribution of the Baru tree (Dipteryx alata Vogel) in the Brazilian Cerrado. Brazilian Journal of Biology 70, 911–919.

Nix, H.A., 1986. A Biogeographic Analysis of Australian Elapid Snakes. Australian Flora and Fauna Series. Canberra: Australian Government.

Oja, T., Alamets, K., Pärnamets, H., 2005. Modelling bird habitat suitability based on landscape parameters at different scales. Ecological Indicators 5, 314–321.

Pearce, J., Ferrier, S., 2000. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. Ecological Modelling 128, 127–147.

Pearson, R.G., Dawson, T., Berry, P., Harrison, P., 2002. SPECIES: a spatial evaluation of climate impact on the envelope of species. Ecological Modelling 154, 289–300.

Pearson, R.G., Raxworthy, C.J., Nakamura, M., Peterson, A.T., 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102–117.

Pearson, R.G., Raxworthy, C.J., Nakamura, M., Townsend Peterson, A., 2007. ORIGINAL ARTICLE: Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34, 102–117.

Peterson, A.T., Cohoon, K.P., 1999. Sensitivity of distributional prediction algorithms to geographic data completeness. Ecological Modelling 117, 159–164.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190, 231–259.

Phillips, S.J., Anderson, R.P., Schapired, R., 2005. Maxent software for species distribution modeling. AT&T Labs-Research, Princeton University, Center for Biodiversity and Conservation, American Museum of Natural History.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19, 181–197.

Phillips, S.J., Dudik, M.2008, 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Poirazidis, K., Kati, V., Schindler, S., Triantakonstantis, D., Kalivas, D., Gatzogiannis, S., 2010. Landscape and biodiversity in Dadia – Lefkimi – Soufli Forest National Park, in: Greece, W. (Ed.), Biodiversity, Management and Conservation. WWF Greece, Athens, pp. 103–114.

Polasky, S., Camm, J.D., Solow, A.R., Csuti, B., White, D., Ding, R., 2000. Choosing reserve networks with incomplete species information. Biological Conservation 94, 1–10.

Rambaud, M., Azuelos, L., 2012. Proposition d'une méthode modélisant la présence d'habitats naturels en Seine-et-Marne. CBNBP, délégation Ile-de-France.

Rebelo, H., Jones, G., 2010. Ground validation of presence-only modelling with rare species: a case study on barbastelles Barbastella barbastellus (Chiroptera: Vespertilionidae). Journal of Applied Ecology 47, 410–420.

Remm, K., 2004. Case-based predictions for species and habitat mapping. Ecological Modelling 177, 259–281.

Ridgeway, G., 1999. "The State of Boosting." Computing Science and Statistics 31, 172–181.

Robertson, M.P., Peter, C.I., Villet, M.H., Ripley, B.S., 2003. Comparing models for predicting species' potential distributions: a case study using correlative and mechanistic predictive modelling techniques. Ecological Modelling 164, 153–167.

Romero-Calcerrada, R., Luque, S., 2006. Habitat quality assessment using Weights-of-Evidence based GIS modelling: The case of< i> Picoides tridactylus</i> as species indicator of the biodiversity value of the Finnish forest. Ecological Modelling 196, 62–76.

Sánchez-Fernández, D., Lobo, J.M., Hernández-Manrique, O.L., 2011. Species distribution models that do not incorporate global data misrepresent potential distributions: a case study using Iberian diving beetles. Diversity and Distributions 17, 163–171.

Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31, 1555–1568.

Skov, F., Svenning, J.-C., 2003. Predicting plant species richness in a managed forest. Forest Ecology and Management 180, 583–593.

Songer, M., Delion, M., Biggs, A., Huang, Q., 2012. Modeling Impacts of Climate Change on Giant Panda Habitat. International Journal of Ecology 2012.

Stabach, J.A., Laporte, N., Olupot, W., others, 2009. Modeling habitat suitability for Grey Crowned-cranes (*Balearica regulorum gibbericeps*) throughout Uganda. International Journal of Biodiversity and Conservation 1, 177–186.

Stockwell, D.R.B., Noble, I.R., 1991. "Induction of Sets of Rules from Animal Distribution Data: a Robust and Informative Method of Data Analysis." Mathematics and Computers in Simulations 32, 249–254.

Stockwell, D., Peters, D., 1999. The GARP modelling system: problems and solutions to automated spatial prediction. International Journal of Geographical Science 13, 143–158.

Stockwell, D.R.B., Peterson, A.T., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148, 1–13.

Store, R., Jokimäki, J., 2003. A GIS-based multi-scale approach to habitat suitability modeling. Ecological Modelling 169, 1–15.

Store, R., Kangas, J., 2001. Integrating spatial multi-criteria evaluation and expert knowledge for GIS-based habitat suitability modelling. Landscape and Urban Planning 55, 79–93.

Suárez-Seoane, S., García de la Morena, E.L., Morales Prieto, M.B., Osborne, P.E., de Juana, E., 2008. Maximum entropy niche-based modelling of seasonal changes in little bustard (Tetrax tetrax) distribution. Ecological Modelling 219, 17–29.

Sutherst, R.W., Maywald, G.F., 1985. A computerised system for matching climates in ecology. Agriculture, Ecosystems & Environment 13, 281–299.

Tarkesh, M., Jetschke, G., 2012. Comparison of six correlative models in predictive vegetation mapping on a local scale. Environmental and Ecological Statistics 19, 437–457.

ter Braak, C.J.F., Smilauer, P., 1998. CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4).

Thomaes, A., Kervyn, T., Maes, D., 2008. Applying species distribution modelling for the conservation of the threatened saproxylic Stag Beetle (Lucanus cervus). Biological Conservation 141, 1400–1410.

Thuiller, W., 2003. BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. Global Change Biology 9, 1353–1362.

Thuiller, W., Brotons, L., Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172.

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD - a platform for ensemble forecasting of species distributions. Ecography 32, 369–373.

Tsoar, A., Allouche, O., Steinitz, O., Rotem, D., Kadmon, R., 2007. A comparative evaluation of presence-only methods for modelling species distribution. Diversity and Distributions 13, 397–405.

Walker, P. A., Cocks, K. D., 1991. "HABITAT: A Procedure for Modelling a Disjoint Environmental Envelope for a Plant or Animal Species." Global Ecology and Biogeography Letters 1, 108–118.

Wilson, K.A., Westphal, M.I., Possingham, H.P., Elith, J., 2005. Sensitivity of conservation planning to different approaches to using predicted species distribution data. Biological Conservation 122, 99–112.

Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecology 9, 1–13.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group†, 2008. Effects of sample size on the performance of species distribution models. Diversity and Distributions 14, 763–773.

Zaniewski, A.E., Lehmann, A., Overton, J.M., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecological Modelling 157, 261–280.

Zimmermann, N.E., Edwards, T.C., Graham, C.H., Pearman, P.B., Svenning, J.-C., 2010. New trends in species distribution modelling. Ecography 33, 985–989.

Zimmermann, N.E., Kienast, F., 1999. Predictive mapping of alpine grasslands in Switzerland: Species versus community approach. Journal of Vegetation Science 10, 469–482.

Annexe 1 - Base de données bibliographique

L'ensemble des références citées dans ce rapport a été compilé dans une base de données bibliographique stockée en format RIS (jointe avec le rapport dans le dossier .zip), compatible avec la plupart des logiciels de gestion de bases de données bibliographiques dont Zotero, EndNote, Mendeley et ZabRef.

Elle comprend également des références complémentaires, toutes en lien avec la modélisation spatialisée. Ces références présentent essentiellement d'autres exemples d'application des méthodes de modélisation recensées dans cette synthèse et des études critiques sur différents aspects de la modélisation (données d'entrée, algorithmes, hypothèses écologiques sous-jacentes, domaine d'application, etc.).

Par ailleurs, Emilie Gentilini, documentaliste à Irstea Grenoble, a accepté de réaliser un comparatif des quatre logiciels de gestion bibliographique les plus utilisés à Irstea et dans la communauté scientifique : EndNote, JabRef, Mendeley et Zotero. L'objectif de cette comparaison est d'apporter des éléments pour aider à déterminer quel logiciel est le plus adapté aux besoins de différents organismes pour le partage de références dans le cadre d'un projet commun.

Nous la remercions pour le travail réalisé. Le comparatif est synthétisé dans le tableau suivant.

Comparatif de logiciels bibliographiques : EndNote, Zotero, Jabref, Mendeley

10/09/2012

Logiciel	EndNote	Zotero	JabRef	Mendeley
Développeur	Thomson Reuters	Center for History and New Media	JabRef developers	Mendeley
Dernière version stable	X6 (X3-X5 à Irstea)	3.0.8	2.8.1	1.6
Type de licence	Propriétaire / Payant	GPL / Gratuit*	GPL / Gratuit	Propriétaire / Gratuit*
Open Source	Non	Oui	Non	Oui
Système d'exploitation	Windows / Mac	Windows / Mac / Linux	Windows / Mac / Linux	Windows / Mac / Linux
Compatibilité avec navigateur Internet	Oui	Firefox uniquement	Non	Oui
Espace de stockage	1Go - 25k ref (X5)/ 10k ref (<X5)	100 Mo	-	500Mo perso / 500Mo partage
Gestions des références				
Organisation				
Nombre de bibliothèques	Illimité	1	Illimité	1
Groupes	Oui	Oui	Oui	Oui
Groupes dynamiques	Oui	Oui	Oui	Non
Gestion des doublons	Oui	Oui	Oui	Oui
Recherche				
Recherche par mots-clés / tags	Oui	Oui	Oui	Oui
Recherche multi-champs	Oui	Oui	Oui	5 champs (Titre, auteurs, année, revue, notes)
Recherche dans les PDF	à partir de la v. X5	Oui (s'ils sont indexés)	Non	Oui
Sauvegarde des recherches	Oui	Oui	Non	Non
Autres fonctionnalités				
Indexation	Oui	Oui	Oui	Oui
Gestion de liste de termes	Oui	Non	Oui	Pour les champs my tags, auteurs, revues, keywords
Modification / Création champs et types de document	Oui	Non	Oui	Non
Annotation dans les PDF	à partir de la v. X5	Non	Non	Oui
Import des références				
Formats supportés				
BibTeX	Non	Oui	Oui	Oui
RIS	Oui	Oui	Oui	Oui
EndNote / Refer / BibIX	Oui	Oui	Oui	Oui
Systèmes d'import				
Modification / Création de filtres et connexions	Oui	Non	Oui	Non
Connexion aux bdd et catalogues	Oui	Oui	Oui	Oui
Téléchargement des PDF automatique	Non	Oui (selon les éditeurs)	Non	Oui (selon les éditeurs)
Capture d'écran lors du téléchargement	Non	Oui	Non	Non
Extraction de métadonnées à partir de PDF	à partir de la v. X5	Oui	Oui	Oui
Export des références				
Intégration dans un traitement de texte				
Traitements de texte compatibles	Word / Open Office	Word / Open Office / LateX	Word / Open Office / LateX	Word / Open Office / LateX
Citation / Bibliographie / Note de bas de page	Oui / Oui / Oui	Oui / Oui / Non	Oui / Oui / Non	Oui / Oui / Oui
Styles bibliographiques				
Nombre de styles	plus de 5000	plus de 2000	?	environ 1200
Création / Modification de styles	Oui	Oui (CSL)	Oui	Oui (CSL)
Partage des références				
Groupes de références partagées sur le Web	Oui (via EndNote Web)	Oui	Non	Oui (5 groupes max avec 10 membres)
Gestion des droits lecture / écriture	Oui (via EndNote Web)	Oui	Non	Oui

* version payante pour augmenter les capacités

Quelques précisions sur la comparaison des logiciels par catégories :

Typologie de logiciel

Tous les logiciels fonctionnent sous Mac, Windows et Linux, sauf Endnote qui ne fonctionne pas sous Linux. EndNote est un logiciel sous licence payante et propriétaire dont le coût varie selon le nombre, et le type d'organisme acquéreur. JabRef est totalement gratuit et libre. Zotero et Mendeley proposent des versions gratuites de leur logiciel, mais il faut payer pour avoir plus d'options, notamment plus d'espace de stockage de référence. Une possibilité néanmoins avec Zotero est intéressante : il est en effet possible de créer un serveur WebDav pour avoir un espace personnel de taille plus importante que les 100Mb initialement prévu dans la version gratuite.

Gestion des références

Les quatre logiciels étudiés sont globalement équivalents sur l'ensemble de ces critères. Seul Zotero ne permet pas l'annotation de PDF et reste assez « rigide » concernant la gestion des listes de termes et le paramétrage des champs.

Import des références

Il existe toujours un format commun entre les logiciels pour les échanges (format RIS), et les fonctions d'import sont équivalentes bien que certaines sources fonctionnent mieux avec certains formats que d'autres.

Export des références

Parmi les différences notables pour l'export, on notera que le format LateX est disponible dans tous les logiciels sauf EndNote, et que les notes de bas de page ne sont pas possibles avec Zotero et JabRef. Le nombre de styles proposés par les quatre logiciels est important, et il est toujours possible de les modifier ou d'en ajouter.

Partage de références

JabRef est le seul logiciel de ce comparatif qui ne permet pas le partage de références. Les trois autres logiciels permettent de gérer les droits des utilisateurs en lecture ou écriture. La différence se fera au niveau des espaces de stockage que chacun autorise.

Critère « utilisateur »

Enfin, il est important de prendre en compte également des critères non techniques concernant la pratique des utilisateurs, et leurs contraintes personnelles. En effet, avant de choisir un logiciel, il faut s'intéresser à l'ergonomie du logiciel, la facilité de prise en main, la documentation existante, la stabilité, la maintenance, et la taille et la réactivité de la communauté. Des contraintes extérieures telles que le logiciel utilisé par les collaborateurs, l'envie de changer ou non, et des besoins différents, peuvent également aiguiller pour le choix d'un logiciel.