OGM Observatoire des Galliformes de Montagne

Office National de la Chasse et de la Faune Sauvage

Modélisation des zones d'hivernage du Tétras-Lyre dans les Alpes françaises du Nord

Programme ALCOTRA "Galliformes Alpins"

Clément Callenge - Tian Simiao

4 Avril 2012

alcotra

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Direction régionale de l'Environnement, de l'Aménagement et du Logement
RHÔNE-ALPES

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Direction régionale de l'Environnement, de l'Aménagement et du Logement
PROVENCE ALPES-CÔTE D'AZUR

Table des matières

1 Introduction **3**

2 Résumé de l'approche de modélisation choisie **6**

3 Etablissement des données **7**

3.1 Données raster 7

3.2 Données surfaciques vecteurs 10

3.3 Extraction des quatre zones prospectées 10

3.4 Données de crottiers 10

4 Stratégie de modélisation **14**

4.1 Problème posé par les orthophotographies 14

4.2 Formalisation du dispositif d'échantillonnage 16

4.3 Deux stratégies de modélisation 17

4.4 Présences et points de contexte 18

4.5 Les variables sélectionnées 19

5 Les méthodes utilisées **20**

5.1 Les distances de Mahalanobis 20

5.2 L'analyse factorielle des distances de Mahalanobis 22

5.2.1 Augmenter le biais de la prédiction pour diminuer l'erreur de prédiction 22

5.2.2 Diminuer la dimension de l'espace pour diminuer l'erreur de prédiction : la MADIFA 24

5.3 La régression logistique "complète" 26

5.4 Une régression logistique pas à pas (*stepwise*) 27

5.5 Les forêts d'arbres de décision 28

6 La validation des modèles **30**

6.1 Une modélisation en théorie en trois étapes 30

6.2 L'étape de validation 31

6.3 Les critères pour mesurer le pouvoir prédictif du modèle 33

6.3.1 Trois critères choisis 33

6.3.2 Le critère de Boyce *et al.* (2002) continu 33

6.3.3 La corrélation bisériale ponctuelle 34

6.3.4 Le critère AUC 34

6.4 Etape de test du modèle : qualité de prédiction 38

7 Résultats **39**

7.1 Analyse exploratoire préliminaire 39

7.1.1 La composition environnementale des zones d'étude . . . 39
7.1.2 Analyse K-select de la sélection de l'habitat par zone d'étude . . . 39
7.2 Modélisation prédictive pour le département de Haute Savoie . . . 41
7.3 Analyse sur toutes les Alpes du Nord . . . 45
7.4 Calibration et validation interne . . . 45
7.5 Validation externe des modèles . . . 46

8 Habitat ou conformité ? 50
8.1 Définir un seuil . . . 50
8.2 Deux risques . . . 50
8.3 Identification du seuil . . . 51
8.4 Prédiction des habitats . . . 51
8.5 Construction de la carte . . . 53

9 Discussion 54
9.1 Synthèse . . . 54
9.2 A propos des méthodes utilisées . . . 56
9.3 Le problème de l'autocorrélation spatiale . . . 57
9.4 Les problèmes de la prédiction statistique . . . 58

Remerciements

Les données sur lesquelles repose la modélisation mise en œuvre dans ce rapport résultent des prospections réalisées dans le cadre de la convention de recherche ONCFS/FDC38 N° 2009/20/6171. Nous tenons à remercier Estelle LAUER (Fédération départementale des chasseurs de l'Isère), coordinatrice, de les avoir mises à notre disposition ainsi que toutes les personnes qui ont contribué aux relevés de terrain : Philippe AULIAC (Fédération départementale des chasseurs de Savoie), Sylvain CAVALLINI (Office national de la chasse et de la faune sauvage), François DRILLAT (Office national des forêts), Emmanuel JOLY (Fédération départementale des chasseurs de Savoie), Yann MAGNANI (Office national de la chasse et de la faune sauvage), Maurice PANTALONI (bénévole), Marie-Odile RE (Office national de la chasse et de la faune sauvage), Anne-Marie RECOLLIN-BELON (bénévole) et Florian RODANEL (Fédération départementale des chasseurs de l'Isère).

Nous remercions également Pierre EYMARD-BIRON (Parc naturel régional du Vercors), François-Xavier GIRARDOT (Office national des forêts) et Lise WLERICK (Office national des forêts) de nous avoir communiqué les données utilisées pour la validation externe des modèles.

Merci enfin à Julien ARDIN (Office national de la chasse et de la faune sauvage) et Pascale COLLARD (Office national de la chasse et de la faune sauvage) pour leur appui à la constitution des fichiers de données.

1 Introduction

Dans un contexte de conservation du tétras-lyre dans les Alpes, l'Observatoire des galliformes de montagne se pose la question de la prédiction de la localisation des habitats d'hivernage de cette espèce. Lorsque les conditions hivernales sont trop rudes, cet animal se construit un "igloo" dans la neige, dans lequel il s'abrite pendant la nuit, et en sort au petit matin pour se nourrir. Il se creuse ensuite un nouvel igloo pour y demeurer pendant la journée, et n'en ressort que le soir, à nouveau pour se nourrir. Lorsque la neige fond, ces sites d'hivernage deviennent facilement disponibles à l'observation, car les fécès déposées par l'animal dans l'igloo contrastent fortement avec la neige environnante – on parle alors de "crottiers" de tétras-lyre.

L'une des sources principales de perturbation du tétras-lyre dans les Alpes étant les sports d'hiver (skieurs, randonneurs en raquette, remontées mécaniques, etc.), il est important de pouvoir déterminer quelles sont les zones les plus utilisées par le tétras-lyre en hiver – en particulier pour l'établissement des igloos. Une telle connaissance permettrait d'identifier les habitats critiques pour cette espèce, et donc de préserver ces habitats dans le cadre d'une politique de conservation du tétras-lyre. Dans un premier temps, une carte des potentialités du milieu permettrait de parer à l'urgence pour l'ensemble des massifs et d'optimiser l'effort de prospection, pour déterminer dans un second temps l'emplacement réel et l'abondance des crottiers.

Dans le cadre d'une convention de recherche entre l'ONCFS et la FDC 38, destinée à mettre au point une méthode de diagnostic standardisées des habitats d'hivernage des prospection ont été réalisées 2010 et 2011 sur les massifs montagneux de Flaine (département 74), Giettaz (74) et les Saisies (73 et 74), et en 2011 sur le massif du Collet d'Allevard (38), afin de localiser ces crottiers (figure 1). Dans ce rapport, nous souhaitons nous servir de ces données pour construire une carte prédictive des sites d'hivernage du tétras-lyre sur l'ensemble des Alpes du nord (département de la Drôme, Isère, Savoie et Haute-Savoie). En d'autres termes, nous souhaiterions pouvoir nous servir de variables environnementales externes (altitude, pente, etc.) pour construire un modèle permettant cette prédiction.

Dans un rapport précédent, effectué dans le cadre de la convention de recherche précitée, nous avons pu montrer une certaine stabilité spatiale de l'emplacement des zones d'hivernage d'une année sur l'autre. En effet, en comparant la position des sites d'hivernage en 2010 (année assez enneigée) et en

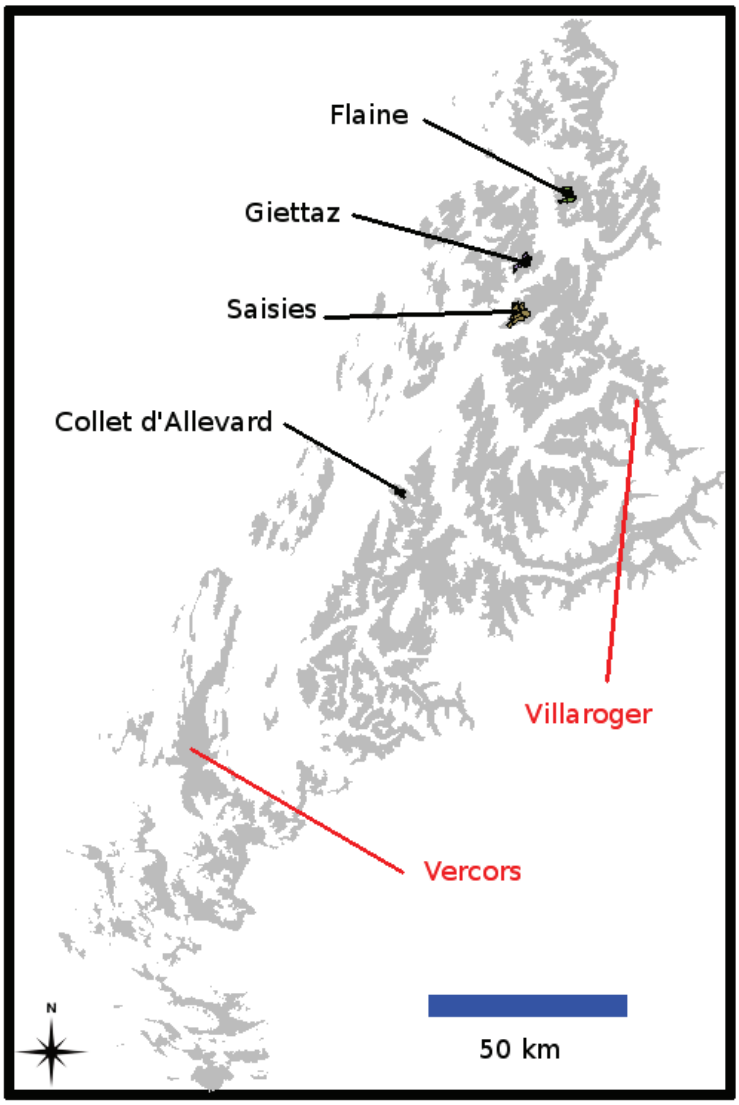

FIGURE 1 – Localisation dans les Alpes du nord des zones d'études sur lesquelles des prospections ont été menées à la recherche de crottiers (source : ONCFS/FDC 38). Les zones indiquées en rouges correspondent aux réserves naturelle dans lesquelles des observations occasionnelles, réalisées par les gestionnaires, ont été utilisées comme validation externe des modèles.

2011 (année très peu enneigée), nous avons pu montrer que malgré des conditions environnementales très différentes, l'emplacement des zones d'hivernage était approximativement le même. Ceci justifie ce travail de modélisation (en effet, si l'emplacement des crottiers était très variable d'une année à la suivante, alors l'environnement "stable" – altitude, pente, etc. – ne jouerait qu'un rôle mineur dans le déterminisme de l'emplacement des crottiers, et une modélisation de ces emplacements en fonction de cet environnement n'aurait pas grand sens).

Nous avons décomposé le travail de modélisation en deux étapes. Dans un premier temps, il nous a fallu établir les données. Cette étape d'établissement des données consiste tout d'abord en la sélection de variables environnementales pertinentes pour la modélisation. Cette sélection est probablement l'étape la plus difficile de la modélisation, comme l'ont souligné plusieurs auteurs (e.g. voir Guisan et Zimmermann, 2000, pour une revue). Nous avons sélectionné une liste de variables que nous supposons influencer l'établissement des sites d'hivernage :

- Des variables décrivant le paysage, dérivées des orthophotographies de l'Institut géographique national (IGN) ;
- Des variables décrivant la végétation, dérivées de la cartographie des zones forestières de l'Inventaire forestier national (IFN) ;
- Des variables descriptives du relief, dérivées des modèles numériques de terrain de l'IGN.

Un important travail d'établissement des données a ensuite pris place. Nous disposions, pour l'ensemble de la zone d'intérêt (départements de la Drôme, de l'Isère, de la Savoie et de la Haute-Savoie) d'un quadrillage fourni par la communauté européenne composé de quadrats de 100 mètres × 100 mètres. Ce quadrillage, qui doit également servir de base pour l'établissement du diagnostic des habitats d'hivernage, définit la résolution de notre étude. Notre objectif est d'identifier les quadrats de cette grille les plus utilisés pour l'établissement de zones d'hivernage de tétras-lyre. Ainsi, nous avons calculé la valeur des variables environnementales sélectionnées pour chacun des quadrats de la zone d'intérêt.

La deuxième étape du travail a été la modélisation proprement dite. Nous disposions, pour les quatre zones alpines déjà citées, des limites de la zone prospectée ainsi que de l'emplacement des crottiers de tétras-lyre détectés lors de ces prospections. Deux de ces zones étaient situées en haute savoie (Flaine et Giettaz), une de ces zones était à cheval entre la Savoie et la Haute-Savoie (les Saisies), et la dernière était située en Isère. Nous avons alors utilisé plusieurs approches pour modéliser la conformité à l'habitat dans chaque quadrat [1]. Il est important de garder en tête que l'objectif de la modélisation est avant tout prédictif : nous ne cherchons pas ici à mettre en évidence les variables importantes pour l'établissement d'une zone d'hivernage, mais à identifier ces zones d'hivernage. Les méthodes que nous avons utilisées tiennent compte de cet objectif. Ainsi, nous avons utilisé des méthodes optimisées pour la prédiction, mais pour lesquelles le pouvoir explicatif est faible : les distances de Mahalanobis (Clark *et al.*, 1993), l'analyse factorielle des distances de Mahalanobis (MADIFA, Calenge *et al.*, 2008), la régression logistique (McCullagh et Nelder, 1989), une régression logistique avec sélection des variables par un algorithme "pas à pas" (*stepwise regression*, Venables et Ripley, 2002, p. 175) et une forêt d'arbres décisionnels (Breiman, 2001).

Dans ce rapport, nous présentons le détail de cette modélisation.

1. Nous avons choisi d'utiliser l'expression *conformité à l'habitat* pour traduire l'expression anglaise "*habitat suitability*", afin d'éviter l'abus de langage consistant à désigner ce concept par le terme "qualité de l'habitat". Cette conformité reflète la probabilité de présence d'une espèce dans une zone, ce qui ne correspond pas nécessairement à la qualité de l'habitat : en effet, pour certaines espèces animales, une forte abondance d'une espèce dans une zone ne signifie pas nécessairement que cette zone est de bonne qualité pour l'espèce : les individus de certaines espèces peuvent être peu nombreux dans les "bons" habitats – e.g. ceux dans lesquels la survie des individus de l'espèce est maximale – et très nombreux dans les "mauvais" si un petit nombre d'individus dominant se partagent les bons habitats et que le plus grand nombre se concentre dans les habitats périphériques (van Horne, 1983). Bien que cela soit peu probable dans le cas du tétras-lyre, nous avons malgré tout décidé de conserver le terme "conformité à l'habitat", car plus approprié.

2 Résumé de l'approche de modélisation choisie

Dans cette section, nous résumons les différentes étapes de la modélisation que nous avons mise en œuvre. Nous donnons ce résumé afin d'économiser au lecteur pressé la lecture de la section 3, qui présente l'étape d'établissement des données, de la section 4, qui présente les stratégies de modélisation, de la section 5, qui décrit le principe des méthodes de modélisation utilisées, et de la section 6, qui décrit les étapes de validation des modèles prédictifs. Le lecteur pressé pourra donc passer directement à la section 7, qui présente les résultats de cette modélisation. Le lecteur désirant au contraire plus de détails sur les choix effectués pour cette modélisation pourra se reporter aux sections suivantes.

Structure des données : Notre structure de données est résumée dans la figure 8. Nous disposons de données collectées sur 4 zones d'étude (Flaine et Giettaz en Haute Savoie, les Saisies à cheval sur les départements de Savoie et de Haute-Savoie, Collet d'Allevard en Isère). Pour chaque zone, nous disposions de la composition environnementale (cf. tableau 1 pour la liste des variables environnementales disponibles) pour chaque crottier détecté lors des opérations de prospection, ainsi que pour chaque point de contexte (un point de contexte correspondant au centre d'un quadrat appartenant à une des zones d'étude) ;

Méthodes utilisées pour la modélisation : Nous avons utilisé cinq méthodes distinctes de prédiction de la conformité à l'habitat d'hivernage du tétras-lyre : les distances de Mahalanobis, la MADIFA, la régression logistique, une régression logistique avec sélection des variables par un algorithme pas à pas, et une forêt d'arbres décisionnels.

Evaluation de l'importance des orthophotographies pour la prédiction : comme nous ne disposions pas d'informations homogènes concernant le paysage dans tous les départements des Alpes du nord (orthophotographies pas relevées au même moment dans tous les départements, cf. section 4.1), nous avons dans un premier temps concentré notre modélisation sur le département de Haute Savoie sur lequel nous disposions de l'information la plus abondante (3 zones d'étude : Flaine, Giettaz et la partie du site des Saisies située en Haute Savoie). Pour ce département, nous avons procédé de la façon suivante. Pour chaque méthode de modélisation, nous avons ajusté un modèle basé sur toutes les variables prédictrices (incluant les variables dérivées des orthophotographies) et un modèle seulement sur les variables de végétation et de relief (en omettant donc les variables dérivées des orthophotographies). Nous avons donc ajusté 10 modèles (5 méthodes de modélisation $\times$ 2 modèles). La liste des variables considérées est donnée dans le tableau 2. Nous avons évalué l'efficacité des différents modèles prédictifs par validation croisée (i.e., (i) en calibrant le modèle sur deux des sites et en validant la prédiction sur le troisième, et (ii) en répétant l'opération (i) pour chacun des sites à tour de rôle). L'efficacité de la prédiction (cf. section 6.3) a été mesurée à l'aide de l'AUC, de l'indice de Boyce *et al.* (2002) et du coefficient de corrélation bisériale ponctuelle. Le seul objectif de cette partie du travail était de déterminer si l'ajout des variables dérivées des orthophotographies dans le modèle permettait une amélioration substantielle des capacités de prédiction des modèles, ou si l'on pouvait se passer de cette variable dans un modèle plus général.

Modélisation à l'échelle des Alpes du nord : Nous avons ensuite considéré les données collectées sur toutes les zones d'étude. Pour chaque méthode de modélisation, nous avons ajusté un modèle basé sur toutes les variables prédictrices à l'exception des variables dérivées des orthophotographies. Nous avons évalué l'efficacité des différents modèles prédictifs par validation croisée (i.e., (i) en calibrant le modèle sur trois des sites et en validant la prédiction sur le quatrième, et (ii) en répétant l'opération (i) pour chacun des sites à tour de rôle). L'efficacité de la prédiction (cf. section 6.3) a été mesurée à l'aide de l'AUC, de l'indice de Boyce *et al.* (2002) et du coefficient de corrélation bisériale ponctuelle. Nous avons également mesuré l'efficacité des différents modèles prédictifs par une étape de validation externe des modèles, en utilisant des observations occasionnelles collectées dans la réserve naturelle de Villaroger et dans le parc naturel du Vercors.

3 Etablissement des données

3.1 Données raster

Nous disposons de 1374 orthophotographies de 5×5 km, à résolution d'un mètre (i.e. chaque pixel couvre un mètre carré), couvrant les départements de la Drôme, de l'Isère, de la Savoie et de la Haute-Savoie. Nous avons importé chacune de ces orthophotographies dans le système d'information géographique GRASS 6.4, et nous avons décomposé chaque orthophotographie en 3 sous-cartes représentant les intensités de rouge, vert et bleu sur chaque zone.

Les orthophotographies sont stockées au format ECW, un format raster compressé. Lorsque ces cartes sont décompressées, chacune occupe sur le disque dur de l'ordinateur une taille de l'ordre de 300 Mo. Etant donné le nombre d'orthophotographies dont nous disposons (1374), il a été nécessaire de rééchantillonner ces orthophotographies de façon à réduire l'espace disque occupé par chaque carte. Nous avons choisi de rééchantillonner chaque orthophotographie de façon à disposer pour chacune d'une carte raster à la résolution de 20 mètres (rééchantillonnage effectué pour chaque sous-carte décrivant les niveaux d'intensité de rouge, vert et bleu). Pour chaque couleur (rouge, bleu, vert), nous avons ensuite accolé ("patché") les 1374 cartes obtenues de façon à disposer d'une unique carte pour la totalité des Alpes du nord. Ceci est illustré sur la figure 2.

Nous disposons par ailleurs d'un maillage de la France établi par la communauté européenne, et sur lequel nous effectuerons nos prédictions. La totalité des Alpes du nord est quadrillée selon des mailles de 100×100 mètres (figure 3).

Nous avons croisé ce maillage avec les cartes dérivées des orthophotographies. Pour chacune des mailles du quadrillage nous avons calculé les valeurs suivantes :

- ⋆ valeur moyenne du niveau de rouge dans la maille ;
- ⋆ écart-type du niveau de rouge dans la maille ;
- ⋆ valeur moyenne du niveau de vert dans la maille ;
- ⋆ écart-type du niveau de vert dans la maille ;
- ⋆ valeur moyenne du niveau de bleu dans la maille ;
- ⋆ écart-type du niveau de bleu dans la maille ;

Ces valeurs synthétisent les structures du paysage observées dans chaque quadrat. Nous disposons donc, pour chaque maille, de variables caractérisant les orthophotographies sur cette zone.

Par ailleurs, nous disposions en outre du modèle numérique de terrain à la résolution de 25 mètres pour chacun des quatre départements de la zone étudiée. Après avoir accolé ("patché") ces quatre cartes pour n'en former qu'une seule, nous avons pu déduire une carte des pentes et une carte de l'exposition sur l'ensemble de la zone. Puis, comme pour les orthophotographies, nous avons calculé les variables suivantes pour chacune des mailles de la grille européenne :

- ⋆ valeur moyenne de l'altitude dans chaque maille
- ⋆ écart-type de l'altitude dans chaque maille
- ⋆ valeur moyenne de la pente dans chaque maille
- ⋆ écart-type de la pente dans chaque maille
- ⋆ proportion de la maille recouvert par des expositions est
- ⋆ proportion de la maille recouvert par des expositions ouest
- ⋆ proportion de la maille recouvert par des expositions nord
- ⋆ proportion de la maille recouvert par des expositions sud

Deux semaines de calcul ont été nécessaires pour conduire les calculs décrit dans cette section sur un PC équipé d'un processeur dual core (2×2.4 GHz), fourni par l'OGM à la cellule d'appui à l'analyse

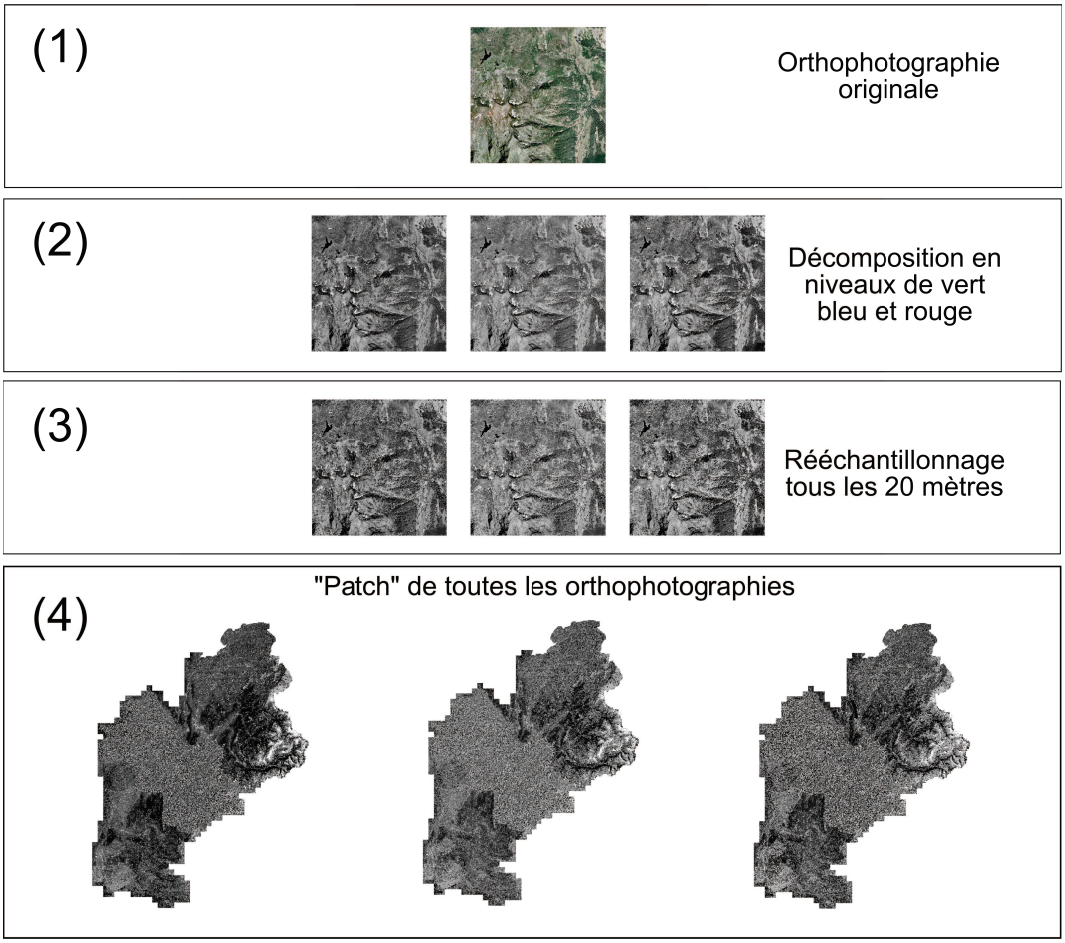

FIGURE 2 – Pré-traitement opéré sur les orthophotographies disponibles pour les 4 départements étudiés : chaque orthophotographie est tout d'abord décomposée en 3 cartes reflétant les niveaux de rouge, vert et bleu de l'orthophotographie originale. Ces trois cartes sont ensuite rééchantillonnées de façon à disposer de cartes à la résolution de 20 mètres. Pour chaque couleur (rouge, bleu, vert), l'ensemble des cartes rééchantillonnées est ensuite accolé en une seule carte couvrant l'ensemble de la zone étudiée

FIGURE 3 – Maillage européen disponible pour la zone d'intérêt. Chaque quadrat est un carré de 100 mètres × 100 mètres.

de données de l'ONCFS. Ces calculs ont été effectués à l'aide du système d'information géographique GRASS 6.4 (GRASS Development Team, 2008).

3.2 Données surfaciques vecteurs

Nous disposions par ailleurs d'une carte décrivant les types forestiers tels que décrits par l'Inventaire Forestier National (IFN). La carte décrit 67 types de végétation sur l'ensemble des Alpes du nord. Le nombre de types de végétation étant trop important pour permettre la construction d'un modèle prédictif, nous avons dû effectuer un reclassement des 67 types IFN initiaux en 8 types plus généraux. Les nouvelles classes sont les suivantes :

- ⋆ Forêts fermées de conifères
- ⋆ Forêts fermées de feuillus
- ⋆ Forêts ouvertes de conifères
- ⋆ Forêts ouvertes de feuillus
- ⋆ Forêts fermées mixtes
- ⋆ Landes
- ⋆ Formations herbacées
- ⋆ Autres

Nous avons alors calculé la proportion de la surface de chaque maille de la grille européenne recouverte par chacun des types de milieu ainsi défini. Pour ce, nous avons procédé de la façon suivante : nous avons accolé ("patché") les 4 cartes vectorielles décrivant les types IFN ainsi définis de façon à n'en former qu'une seule pour l'ensemble de la zone d'intérêt. Nous avons ensuite construit une base de données PostgreSQL/PostGIS (`http://www.postgis.fr/`) contenant la carte de l'IFN ainsi que le maillage européen. En effet, PostGIS offre des fonctions de traitement de cartes vecteur beaucoup plus efficaces que les autres systèmes d'information géographique (ArcGIS, qGIS, GRASS, etc.), ce qui a motivé le choix de ce système de base de données pour effectuer les traitements. Puis nous avons calculé l'intersection de ces deux cartes (maillage et carte IFN ; ce qui a pris seulement 6 heures de calcul sur l'ensemble des Alpes du nord), et nous avons calculé la surface de chaque polygone (intersection maille/type IFN). Nous avons pu en déduire la proportion de chaque maille couverte par chaque type IFN.

3.3 Extraction des quatre zones prospectées

Nous avons ensuite extrait les quadrats du maillage tombant dans les quatre zones prospectées dans des fichiers de forme : Flaine, Giettaz, les Saisies, et Collet d'Allevard. Nous avons ensuite importé ces quatre fichiers dans le logiciel R 2.13.0 (R Development Core Team, 2011). Les cartes des altitudes moyennes sur les quatre zones sont présentées sur la figure 4. A titre d'illustration, nous présentons également les cartes des différentes variables disponibles sur le site de Giettaz sur la figure 5. La liste des variables disponibles pour la modélisation est présentée dans le tableau 1.

3.4 Données de crottiers

Nous avons enfin importé les données de crottiers dans le logiciel R. Ces données correspondent à la localisation des crottiers détectés lors des opérations de prospection. Nous présentons ces données de crottiers sur une carte de chaque zone sur la figure 6.

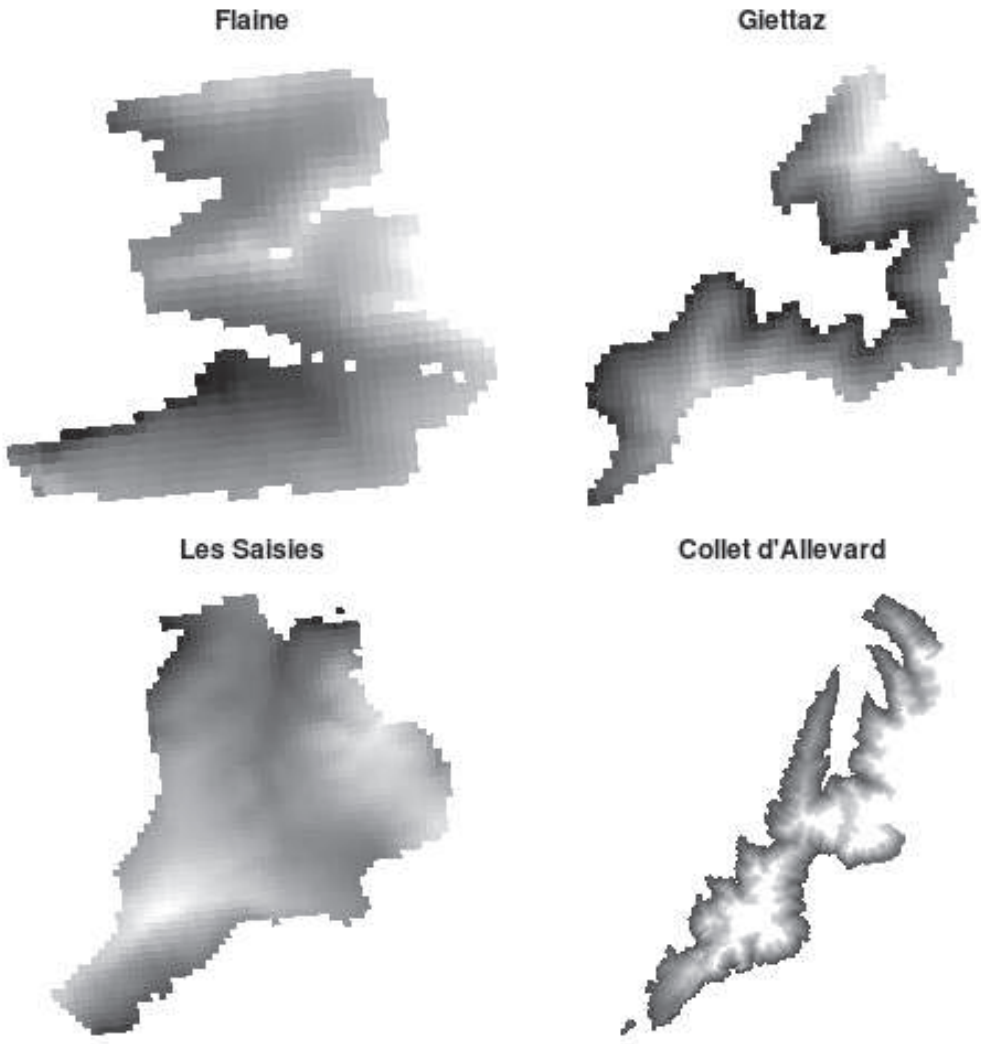

FIGURE 4 – Altitude moyenne dans chaque quadrat des quatre zones prospectées par l'OGM.

FIGURE 5 – Cartes des valeurs des différentes variables environnementales pour chaque quadrat de la zone Giettaz (voir tableau 1 pour la signification des noms des variables).

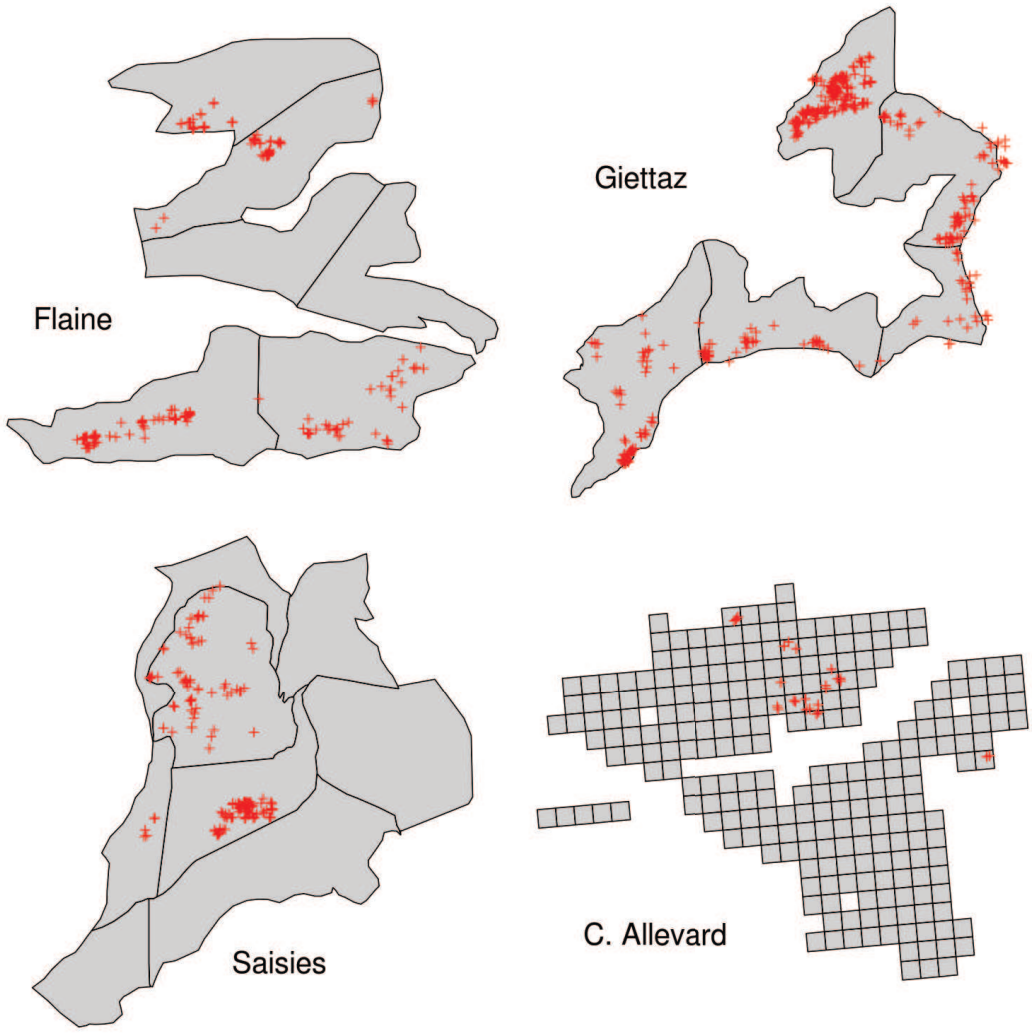

FIGURE 6 – Distribution des crottiers de tétras-lyre identifiés lors des opérations de prospection de 2010 et 2011 sur chacun des quatre sites d'étude.

TABLE 1 – Liste des variables disponibles pour la modélisation prédictive des zones d'hivernage du tétras-lyre. Toutes les variables ne seront pas effectivement utilisées dans la modélisation. Cf. tableau 2 pour la liste des variables effectivement utilisées, et la section 4.5 pour une justification de ces choix.

Acronyme utilisé	Variable
OTGMOY	Niveau moyen (MOY) de vert (Green) dans le quadrat (varie entre 0 et 256)
OTGSD	Ecart-type (SD) de vert (Green) dans le quadrat
OTRMOY	Niveau moyen (MOY) de rouge (Red) dans le quadrat (varie entre 0 et 256)
OTRSD	Ecart-type (SD) de rouge (Red) dans le quadrat
OTBMOY	Niveau moyen (MOY) de bleu (Blue) dans le quadrat (varie entre 0 et 256)
OTBSD	Ecart-type (SD) de bleu (Blue) dans le quadrat
ALTIMOY	Altitude moyenne dans le quadrat
ALTISD	Ecart-type des valeurs d'altitude dans le quadrat
SLOPEMOY	Pente moyenne dans le quadrat
SLOPESD	Ecart-type des valeurs de pentes dans le quadrat
ASPSUDMOY2	Proportion du quadrat recouvert par des expositions sud
ASPESTMOY2	Proportion du quadrat recouvert par des expositions est
ASPNORMOY2	Proportion du quadrat recouvert par des expositions nord
ASPOUEMOY2	Proportion du quadrat recouvert par des expositions ouest
FO_FE_DE_C	Proportion de forêts fermées de conifères dans le quadrat
FO_FE_DE_F	Proportion de forêts fermées de feuillus dans le quadrat
FO_FE_MI	Proportion de forêts fermées mixtes dans le quadrat
FO_OU_DE_C	Proportion de forêts ouvertes de conifères dans le quadrat
FO_OU_DE_F	Proportion de forêts ouvertes de feuillus dans le quadrat
FO_HE	Proportion de formations herbacées dans le quadrat
LA	Proportion de landes dans le quadrat
AU	Proportion d'autres types de végétation dans le quadrat

4 Stratégie de modélisation

4.1 Problème posé par les orthophotographies

La première spécificité de nos données est liée aux orthophotographies. Nous disposons d'orthophotographies pour les quatre départements de notre zone d'intérêt : la Drôme, l'Isère, la Savoie et la Haute-Savoie. Cependant, ces orthophotographies n'ont pas été prises au même moment. En effet, les orthophotographies de Haute-Savoie ont été prises en 2004, celles de Savoie et de la Drôme ont été prises en 2006, et celles de l'Isère ont été prises en 2003. Ceci va avoir des conséquences sur notre modélisation.

En effet, considérons les orthophotographies présentées figure 7. Visuellement, nous avons l'impression que les couleurs des orthophotographies sont plus vives en Savoie, et plus ternes en Isère. En effet, ces orthophotographies ne sont pas toutes prises dans les mêmes conditions de luminosité, ni même avec le même matériel. Or, nous souhaitons utiliser les moyennes et écart-types des niveaux de vert, bleu, et rouge de ces orthophotographies pour prédire la présence de zones d'hivernage utilisées par le tétras-lyre. Nous comprenons dès lors qu'une même combinaison de niveaux de vert, bleu et rouge n'aura pas la même signification dans les différents départements (puisqu'un même habitat paraîtra plus terne en Isère qu'en Haute Savoie). Ceci apparaît également très clairement sur la figure 2(4), qui présente les niveaux moyens de rouge, vert et bleu dans ces orthophotographies sur l'ensemble de la zone étudiée : il semble y avoir des ruptures de continuité des valeurs au niveau des frontières de département.

Le principe de la modélisation prédictive consiste à utiliser les données collectées dans les zones prospectées pour ajuster un modèle prédictif, i.e. estimer la forme de la relation entre les variables prédictrices (variables environnementales) et la variable réponse (nombre de crottiers). Sous certaines hypothèses que nous verrons plus loin, le modèle ainsi ajusté peut ensuite être utilisé pour prédire la variable

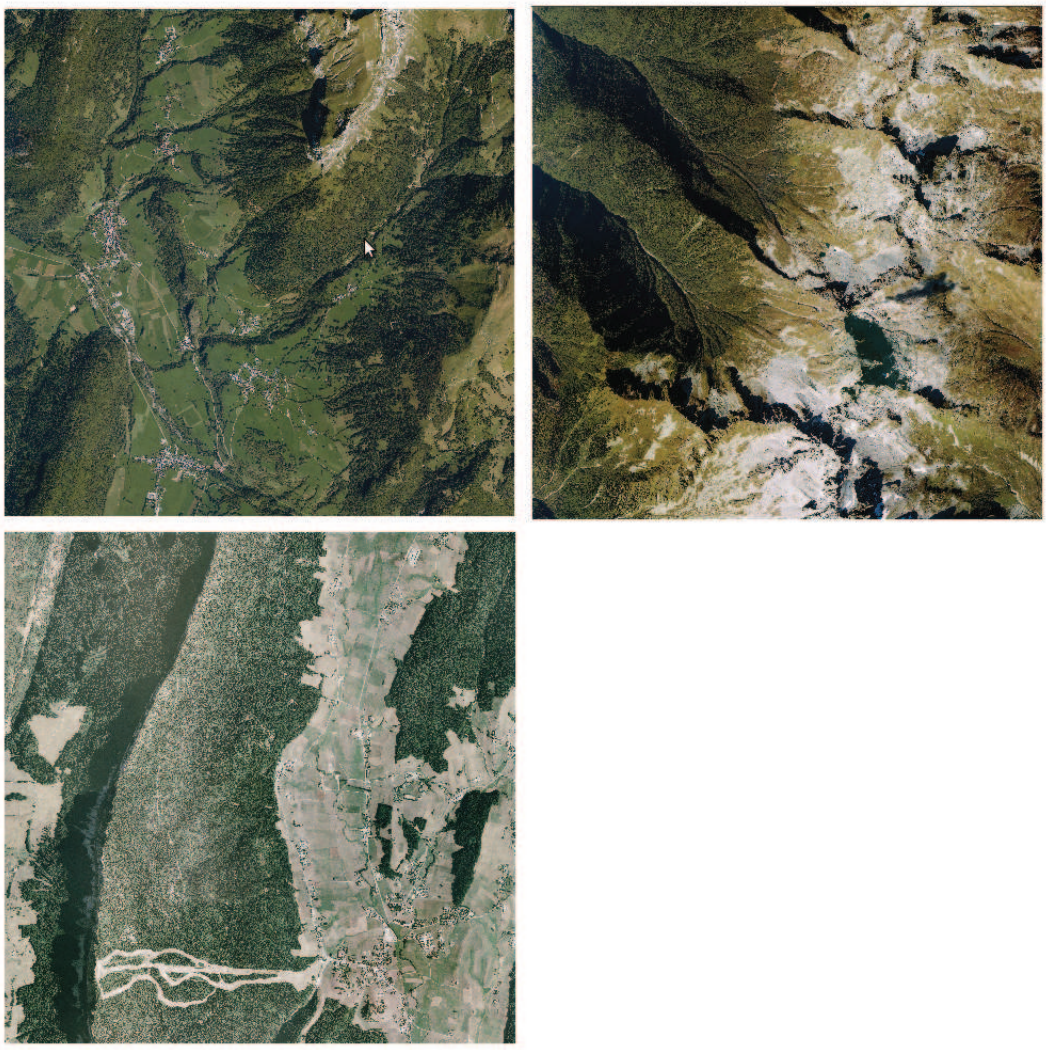

FIGURE 7 – Trois exemples d'orthophotographies : l'orthophotographie représentée en haut à gauche a été prise en Haute-Savoie en 2004 (comme toutes les orthophotographies de ce département). L'orthophotographie en haut à droite a été prise en Savoie en 2006 (comme toutes les orthophotographies de ce département). L'orthophotographie en bas à gauche a été prise en Isère en 2003 (comme toutes les orthophotographies de ce département). Bien que les zones représentées soient différentes, nous comprenons que ces cartes ne seront pas comparables dans une modélisation. L'appareillage utilisé, les conditions de luminosité, et peut-être même l'heure de la journée étaient différentes d'une orthophotographie à l'autre.

réponse sur une nouvelle zone (ou à une autre période), à condition que l'on dispose des mêmes variables prédictrices sur cette nouvelle zone. Or c'est là que nous rencontrons un problème : la signification des variables liées aux orthophotographies varie d'un département à l'autre. Ainsi, si l'on ajuste un modèle en se basant sur les zones d'étude de Flaine et de Giettaz (Haute-Savoie), nous ne pourrons pas nous en servir pour prédire l'habitat en Savoie.

Ceci va très fortement influencer notre stratégie de modélisation : dans un premier temps, nous nous concentrerons sur le département de Haute-Savoie, pour lequel nous disposons des données les plus abondantes. Nous ajusterons alors un modèle prédictif pour ce département en nous servant des sites d'étude de ce département (Flaine, Giettaz, et la moitié du site des Saisies), et de toutes les variables prédictrices. Puis, toujours sur ce département, nous ajusterons un autre modèle prédictif, mais cette fois sans les variables dérivées des orthophotographies. Nous comparerons alors le pouvoir prédictif des deux modèles, afin de déterminer si la suppression des variables dérivées des orthophotographies conduit à une perte substantielle de ce pouvoir. Nous verrons plus loin que ce n'est pas le cas. Nous pourrons alors nous servir des données collectées sur toutes les zones d'étude pour ajuster un modèle que nous pourrons utiliser pour prédire la conformité à l'habitat d'hivernage du tétras-lyre sur l'ensemble des Alpes du nord. Ce dernier modèle ne sera construit que sur les variables de végétation et de relief.

4.2 Formalisation du dispositif d'échantillonnage

Hirzel *et al.* (2002) indique que les études visant à estimer la conformité à l'habitat sont soumises à la difficulté d'identifier les absences réelles de l'espèce. En effet, une localisation donnée peut être classée comme "absence" si :

- **L'espèce n'a pas pu être détectée, même si elle y était présente** (probabilité de détection inférieure à 1). Dans notre étude, les crottiers sont identifiés à la suite de sorties conduites au moment de la fonte des neiges. Bien sûr, des igloos creusés au début de l'hiver apparaîtront uniquement à la fin du printemps (lorsque l'essentiel de la neige tombée ultérieurement aura fondu), alors que des igloos creusés à la fin de l'hiver seront les premiers à apparaître (car situés au "sommet" des différentes couches de neige tombées pendant l'hiver). Ainsi, pour être sûr d'identifier toutes les zones d'hivernage dans une zone d'étude, il est nécessaire de parcourir très fréquemment tous les quadrats d'une zone, de façon à ne pas manquer une zone de crottiers au moment de la fonte des neiges. En pratique, les prospections n'ont duré que deux à trois jours par massif. Il est donc délicat de supposer que tous les sites d'hivernage ont été identifiés sur toutes les zones d'étude : en effet, certaines zones sont très vastes. Ainsi, certains des quadrats, identifiés comme non-utilisés par le tétras-lyre, contenaient probablement des crottiers non-identifiés.

- **L'espèce n'était pas présente à ce point au moment de l'étude, même si la zone constitue un habitat pour l'espèce**. Dans ce cas, le problème vient de ce que la variable d'intérêt (l'habitat potentiel de l'espèce) ne correspond pas à la variable mesurée (présence de l'espèce).

- **La localisation constitue un habitat potentiel pour l'espèce, mais elle est inoccupée pour des raisons historiques**. Ce cas de figure ne nous concernera pas, le tétras-lyre pouvant potentiellement occuper la totalité des zones prospectées (aucune frontière physique n'empêche le déplacement des individus de l'espèce).

- **La localisation ne constitue pas un habitat pour l'espèce** ;

Ainsi, nous devrons prendre en compte, dans notre modélisation, qu'il sera difficile d'identifier les zones ne constituant pas des habitats pour l'espèce (nous pouvons certifier la présence de l'espèce là où nous l'avons localisée, mais nous ne pouvons pas certifier son absence ailleurs).

Nous devons alors formaliser le dispositif d'échantillonnage, afin de clarifier le statut des données, et l'approche qu'il nous faudra utiliser pour les modéliser sans données d'absence. Nous travaillerons à

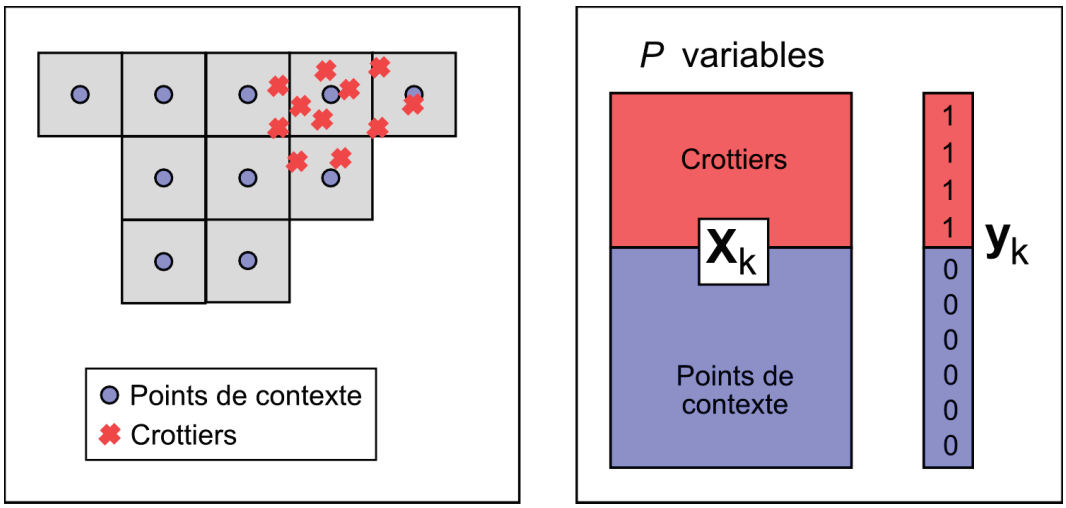

FIGURE 8 – Structure de données disponibles pour la modélisation de la conformité à l'habitat d'hivernage du tétras-lyre. Une zone d'étude k donnée est constituée de N_d^k quadrats appartenant au maillage européen. La zone d'étude est parcourue et N_u^k crottiers sont identifiés sur la zone. Par ailleurs, nous échantillonnons N_d^k points de contexte correspondant aux centres des quadrats de la zone d'étude. Ces données sont stockées sous la forme d'une matrice $\mathbf{X}_k$ et d'un vecteur $\mathbf{y}_k$ (cf. texte).

l'échelle de l'occurrence de crottier, et non à l'échelle du quadrat. Notre unité d'échantillonnage est donc une unité ponctuelle (le point crottier) et non une unité surfacique (le quadrat). Nous supposons que les zones d'études utilisées pour la modélisation (i.e. pour la modélisation couvrant le département de Haute-Savoie, les sites de Flaine, Giettaz et la moitié du site des Saisies située en Haute-Savoie ; pour la modélisation nord alpine, les 4 zones d'étude) ont été parcourues par des observateurs avec un effort de prospection inconnu mais uniforme dans l'espace, et qu'un certain nombre de crottiers a été détecté. Notre approche ne suppose donc pas que tous les crottiers présents sur ces zones ont été détectés : nous supposons qu'une fraction inconnue f des crottiers présents sur la zone sont détectés. L'ensemble des crottiers dont nous disposions est supposé être obtenu suite à un échantillonnage aléatoire simple des crottiers effectivement présents sur les zones. Nous supposons ensuite que les zones d'étude ont été parcourue, et que la composition environnementale a été mesurée à chaque point d'une grille de points placée sur la zone (ces points correspondant aux centres des quadrats de la grille européenne). Nous appellerons ces points *points de contexte environnemental* ou plus simplement *points de contexte* dans la suite du document (cf. section 4.4). Notre jeu de données final correspond donc à l'ensemble des points crottiers et des points de contexte collectés au cours de cette étude.

La structure de nos données est donc la suivante : Nous disposons de K zones d'étude (3 dans le cas de la modélisation en Haute-Savoie, 4 pour la totalité des Alpes du nord). Pour chaque zone, nous disposons de N_u^k points crottiers et de N_d^k points de contexte (figure 8). Le nombre total de points est noté $N^k = N_u^k + N_d^k$. Chaque point (crottier ou de contexte) est caractérisé par P variables environnementales. Pour une zone d'étude donnée, les données sont stockées sous la forme d'une matrice $\mathbf{X}_k$ contenant les valeurs de P variables environnementales (en colonnes) dans les N^k points de l'échantillon (en lignes), et d'un vecteur $\mathbf{y}_k$ de longueur N^k. Le i-ème élément de $\mathbf{y}_k$ prend la valeur 1 lorsque le i-ème point de l'échantillon est un point crottier, et la valeur 0 lorsque le i-ème point de l'échantillon est un point de contexte.

4.3 Deux stratégies de modélisation

Le type de données dont nous disposons est extrêmement courant dans la littérature. Il s'agit de données de type "présence seule" (*presence only*), car si nous pouvons certifier la présence d'un crottier aux emplacements où ils ont été détectés, nous ne pouvons pas certifier l'absence de crottier là où ils

n'ont pas été détectés.

Nous pouvons distinguer deux grandes stratégies utilisées dans la littérature pour modéliser la conformité à l'habitat avec ce type de données. La première stratégie consiste à utiliser des méthodes spécifiquement conçues pour ce type de données (distances de Mahalanobis, ENFA, MADIFA, etc.), qui ne nécessitent pas la connaissance des zones dans lesquelles l'espèce était absente (Clark *et al.*, 1993; Hirzel *et al.*, 2002; Rotenberry *et al.*, 2006; Calenge *et al.*, 2008). La seconde stratégie consiste à modifier légèrement la structure des données disponibles afin de pouvoir utiliser des méthodes statistiques à la fois plus classiques et plus flexibles (e.g. régression logistique, modèle additif généralisé) reposant sur la connaissance de la composition environnementale non seulement des zones de présence, mais aussi des zones d'absence (Guisan et Zimmermann, 2000).

Dans ce rapport, nous utiliserons les deux stratégies. En effet, notre objectif étant prédictif, il est totalement justifié d'utiliser plusieurs stratégies de modélisation, et de choisir celle dont le pouvoir prédictif est le plus fort (Friedman *et al.*, 2008). Nous décrirons dans la section 5 les outils que nous utiliserons pour modéliser la conformité à l'habitat en s'appuyant sur la première stratégie (distances de Mahalanobis et MADIFA). Nous décrivons plus précisément les raisonnements qui sous-tendent la seconde stratégie dans la section suivante.

4.4 Présences et points de contexte

De nombreux auteurs ont pu construire des cartes de conformité à l'habitat à partir de données de type présence seule, en procédant de la façon suivante (e.g. Engler *et al.*, 2004; VanDerWal *et al.*, 2009) :

- ⋆ Ils génèrent des *pseudo-absences* sur la zone d'étude. En général, les pseudo-absences sont des points tirés au sort sur la zone d'étude (cf. plus bas) ;
- ⋆ Ils construisent une variable y prenant la valeur 1 lorsqu'un point de l'échantillon est une occurrence de l'espèce, et la valeur 0 lorsqu'un point de l'échantillon est une pseudo-absence ;
- ⋆ Ils utilisent des méthodes de modélisation prévues pour les données de type présence/absence (e.g. régression logistique), pour modéliser la probabilité que $y = 1$. La probabilité prédite par ce modèle est supposée refléter la probabilité de présence de l'espèce.

C'est l'approche que nous avons suivie dans cette étude, bien que nous n'ayons pas appelé ces points pseudo-absences, mais points de contexte (*background information*). En effet, Phillips *et al.* (2009) notent : *Some modelers think of the background samples as implied absences : partly because the word "pseudo-absences" gives that impression. However, the intention in providing a background sample is not to pretend that the species is absent at the selected sites, but to provide a sample of the set of conditions available to it in the region. The critical step in selection of background data is to develop a clear understanding of the factors shaping the geographic distribution of presence records. Two key elements are the actual distribution of the species and the distribution of survey effort.* Dans notre cas de figure, comme nous supposions un effort de prospection uniforme, nous avons sélectionné ces points de contexte selon un échantillonnage systématique. En outre, nous verrons plus bas que notre approche de modélisation suppose que la probabilité qu'un point de la zone d'étude soit inclus dans l'échantillon des points de contexte est uniforme dans l'espace, ce qui justifie un tel échantillonnage.

Ordinairement, les approches de modélisation de la conformité à l'habitat consistent donc à modéliser une variable réponse y binaire (prenant la valeur 0 lorsqu'il s'agit d'un point de contexte, et 1 lorsqu'il s'agit d'une occurrence de l'espèce) en fonction des variables environnementales. Ces approches permettent de modéliser la probabilité que la variable y prenne la valeur 1 en fonction des variables environnementales. Cependant la probabilité que $y = 1$ n'est pas identique à la probabilité de présence de l'espèce lorsque l'on travaille avec des données de type présence/points de contexte. En effet, Pearce

et Boyce (2006) indiquent : *When using presence-only data it is generally not possible to calculate probabilities of presence ; instead we aim to predict the relative likelihood of presence. There are two reasons for this : (a) separate samples of presence and pseudo-absence data have been selected where sampling fractions are not known, and (b) the pseudo-absence data contains an unknown number of presences, and is thus a contaminated sample of absences.* Ainsi, selon ces auteurs, en modélisant la probabilité que $y = 1$, nous modélisons une vraisemblance relative. Pour un point donné, la probabilité que $y = 1$ ne correspond pas à la probabilité de présence de l'espèce. En revanche, selon ces auteurs, pour deux points donnés A et B, si $P(y_A = 1) < P(y_B = 1)$ alors la probabilité de présence réelle de l'espèce sur le point A est inférieure à celle sur le point B. Cela permet donc la construction de cartes de conformité à l'habitat, la probabilité prédite par le modèle pouvant être considérée comme un indice de la probabilité réelle.

Pearce et Boyce (2006) démontrent que l'on peut utiliser la régression logistique, normalement prévue pour des données de type "présence/absence", pour modéliser la conformité à l'habitat avec des données de type "présence seule". Cependant, nous n'avons pas trouvé de preuves de ces affirmations concernant d'autres méthodes classiquement utilisées sur des données de type présence/absence dans la littérature. En conséquence, pour le présent rapport, nous avons démontré cette propriété plus généralement. En d'autres termes, nous montrons que lorsque nous disposons de données du type points de contexte/présence, et que le statut d'un point est caractérisé par une variable binaire y (prenant la valeur 0 lorsque le point est un point de contexte, et 1 si le point est une occurrence de l'espèce), alors la probabilité que $y = 1$ est une fonction croissante de la probabilité de présence de l'espèce lorsque l'effort de prospection est uniforme. Ceci justifie l'utilisation de méthodes reposant sur des données de type présence/absence pour la construction de cartes de conformité à l'habitat. Cette démonstration est présentée en annexe de ce rapport. Cette démonstration s'appuie fortement sur l'hypothèse que la probabilité qu'un point de contexte fasse partie de l'échantillon ne dépend pas de la valeur des variables environnementales à ce point. Ceci justifie donc notre choix d'inclure dans l'échantillon des points de contexte *tous* les centres de quadrats [2].

4.5 Les variables sélectionnées

Dans toute optique de modélisation, il convient de s'assurer que les variables prédictrices ne soient pas trop corrélées entre elles (Guisan *et al.*, 2002). En effet, l'ajout à un modèle d'une nouvelle variable très corrélée à une variable déjà dans le modèle n'apporte pas beaucoup d'information utile à la prédiction, mais implique l'estimation de paramètres supplémentaires, et accroît la dimension du modèle : nous aurons besoin d'estimer un plus grand nombre de paramètres avec presque la même quantité d'information, ce qui augmentera la variance associée au modèle (cf. section 5.2.1 pour plus de précisions sur ces notions de variance associée au modèle, et de dimension du modèle).

Nous avons donc supprimé certaines variables de notre jeu de variables prédictrices, afin de limiter les corrélations entre ces variables. Ainsi, la corrélation entre niveaux moyen de rouge, de vert et de bleu dans les orthophotographies était généralement très élevée (coefficient de corrélation de Pearson toujours supérieur à 0.9), ce qui nous a conduit à ne conserver qu'une seule de ces variables comme variable prédictrice : le niveau moyen de vert. De même, les écarts-types des niveaux de vert, de rouge et de bleu sur les orthophotographies étaient très corrélés (coefficient de corrélation de Pearson toujours supérieur à 0.75). Nous n'avons donc conservé qu'une seule de ces variables : l'écart-type du niveau de vert. Enfin, nous avons supprimé la proportion de végétation de type "Autres types de végétation" dans un quadrat des variables prédictrices, car ce type de végétation était assez rare, et était même absent d'une des zones d'étude (collet d'Allevard), ce qui excluait une estimation correcte de l'effet de ce type de végétation dans un modèle prédictif [3]. En résumé, les variables dont nous nous servirons pour la modélisation sont

2. Nous pourrions supposer intuitivement que ne choisir comme points de contexte que les centres des quadrats détectés comme inoccupés permettrait d'accroître le pouvoir prédictif du modèle, car cela permettrait une meilleure discrimination entre occupation et non occupation. Cependant, les sites détectés comme inoccupés montrent probablement une composition environnementale qui diffère de la composition de la zone d'étude ; la probabilité d'inclusion d'un point de contexte serait alors dépendante de la valeurs des variables environnementales, et cette hypothèse ne serait plus respectée. Il est alors préférable d'inclure comme point de contexte les centres de tous les quadrats

3. En outre, "Autres types de végétation" n'a aucun sens écologique. Il ne s'agit pas à proprement parler d'un type de végétation. Il s'agit d'une classe hétérogène regroupant tous les types de végétation qui ne rentrent pas dans les "cases" que nous avons prédéfinies en section 3.2. "Autres types de végétation" peut donc regrouper des types de végétation très

présentées dans le tableau 2.

TABLE 2 – Liste des variables effectivement utilisées pour la modélisation prédictive des zones d'hivernage du tétras-lyre

Acronyme utilisé	Variable
OTGMOY	Niveau moyen (MOY) de vert (Green) dans le quadrat (varie entre 0 et 256)
OTGSD	Ecart-type (SD) de vert (Green) dans le quadrat
ALTIMOY	Altitude moyenne dans le quadrat
ALTISD	Ecart-type des valeurs d'altitude dans le quadrat
SLOPEMOY	Pente moyenne dans le quadrat
SLOPESD	Ecart-type des valeurs de pentes dans le quadrat
ASPSUDMOY2	Proportion du quadrat recouvert par des expositions sud
ASPESTMOY2	Proportion du quadrat recouvert par des expositions est
ASPNORMOY2	Proportion du quadrat recouvert par des expositions nord
ASPOUEMOY2	Proportion du quadrat recouvert par des expositions ouest
FO_FE_DE_C	Proportion de forêts fermées de conifères dans le quadrat
FO_FE_DE_F	Proportion de forêts fermées de feuillus dans le quadrat
FO_FE_MI	Proportion de forêts fermées mixtes dans le quadrat
FO_OU_DE_C	Proportion de forêts ouvertes de conifères dans le quadrat
FO_OU_DE_F	Proportion de forêts ouvertes de feuillus dans le quadrat
FO_HE	Proportion de formations herbacées dans le quadrat
LA	Proportion de landes dans le quadrat

5 Les méthodes utilisées

Nous décrivons à présent les méthodes utilisées dans ce rapport pour permettre la modélisation des habitats d'hivernage du tétras-lyre. Nous rappelons ici que notre objectif est ici uniquement *prédictif*, c'est à dire que nous ne cherchons pas à expliquer quelles sont les variables qui sont le plus importantes dans la sélection des zones d'hivernage, mais uniquement à identifier ces zones. Nous choisirons donc nos outils en conséquence.

5.1 Les distances de Mahalanobis

Les distances de Mahalanobis ont été introduites par CLARK *et al.* (1993) comme méthode de prédiction de la conformité à l'habitat en écologie, et ont été très utilisées depuis pour construire ce type de carte (e.g. KNICK et DYER, 1997; FARBER et KADMON, 2003). Le principal avantage de cette approche est qu'elle ne s'appuie pas sur les points de contexte pour définir la conformité à l'habitat : elle ne s'appuie que sur la distribution des crottiers. Nous pouvons donc modéliser la conformité à l'habitat avec cette approche sans avoir besoin de points de contexte.

Le principe de cette méthode repose sur le concept de niche écologique tel que formalisé par HUTCHINSON (1957). Cet auteur définit la niche écologique d'une espèce comme l'hypervolume, dans l'espace multidimensionnel défini par les variables environnementales, dans lequel une espèce donnée peut maintenir une population viable [4]. Nous décrivons ci-dessous comment ce concept peut être utilisé pour

conformes à l'habitat comme des types très peu conformes.

4. Le concept de niche écologique est au centre de nombreux débats en écologie (e.g. CHASE et LEIBOLD, 2003). La formalisation de ce concept par HUTCHINSON (1957) est très utile pour permettre le développement de méthodes statistiques d'analyse de la distribution spatiale des espèces (CALENGE, 2005), bien que dans la plupart des cas, on ne dispose d'aucune information sur la viabilité des populations. Les auteurs de méthodes statistiques utilisent donc le terme *niche écologique* afin de décrire l'hypervolume utilisé par l'espèce par abus de langage (HIRZEL *et al.*, 2002; CALENGE et BASILLE, 2008), ce qui est souvent reproché par les écologues (obs. pers.). Dans notre étude, nous ne travaillerons pas sur la niche écologique *per se*, car nous ne disposons pas non plus d'information sur la viabilité des populations étudiées. Nous travaillons sur une

prédire les caractéristiques environnementales conformes aux besoins de l'espèce (cf. figure 9).

Chacun des points dont nous disposons (points de contexte ou crottiers) est décrit par un certain nombre de variables environnementales (altitude, pente, etc.). Chacune de ces variables définit une dimension dans un espace multidimensionnel que nous appellerons *espace écologique*. Pour chaque point, la valeur prise par une variable environnementale définit une coordonnée sur la dimension correspondante dans l'espace écologique. Ainsi, à chaque point dans l'espace géographique, on peut associer un point dans l'espace écologique.

Concentrons nous à présent sur les points crottiers. Si l'on travaille sur une zone d'étude k caractérisée par N_u^k crottiers identifiés, ces N_u^k crottiers définissent un nuage de N_u^k points dans l'espace écologique (en rouge sur la figure 9). Nous considérons que ces points sont *utilisés* par l'espèce, et que la distribution de l'ensemble de ces points définit l'habitat d'hivernage de l'espèce dans cet espace (une forme de "niche" d'hivernage).

Le principe des distances de Mahalanobis comme outil de prédiction de la conformité à l'habitat est le suivant. Si nous supposons que la distribution des points utilisés dans l'espace écologique peut être décrit par une loi normale multivariée, alors le centre de gravité du nuage de points utilisé (i.e., le point défini dans l'espace écologique par les moyennes des valeurs prises par les variables environnementales par les points utilisés, en jaune sur la figure 9) représente l'habitat le plus conforme possible, et plus on s'éloigne de ces conditions de conformité idéales, moins l'environnement est conforme à l'habitat. Si l'on dispose d'un nouveau point (par exemple, provenant d'une nouvelle zone non étudiée, e.g. les points A, B et C bleus sur la figure 9), la distance *mesurée dans l'espace écologique* entre ce point et le centre de gravité du nuage de points utilisé reflétera l'éloignement des conditions environnementales dans ce point à cet environnement "idéalement conforme", à cet habitat typique. Cette distance, qui peut être calculée pour tout nouveau point, est donc inversement proportionnelle à la conformité à l'habitat. Pour toute nouvelle zone, nous pouvons alors construire des cartes géographiques de cette distance mesurée dans l'espace écologique, qui seront alors des cartes de conformité à l'habitat.

Clark *et al.* (1993) soulignent que l'on peut améliorer les indices de conformité à l'habitat en remplaçant la distance euclidienne classique par la distance de Mahalanobis. Ce type d'approche permet en effet d'intégrer les structures de corrélation du nuage de points utilisés (i.e. la forme non sphérique de ce nuage de points) lors de la mesure de la distance (les différents niveaux de gris sur la figure 9 représentent des classes d'égale distance de Mahalanobis dans cet espace).

En pratique, soit $\boldsymbol{\mu}$ le vecteur contenant les coordonnées dans l'espace écologique du centre de gravité de la niche calculées à partir de l'échantillon de crottiers identifiés sur les zones d'étude (ce vecteur est obtenu en calculant, pour chaque variable environnementale, la moyenne des valeurs de la variable aux points crottiers). Nous pouvons également calculer la matrice $\boldsymbol{\Sigma}$ de variances-covariances du nuage de points utilisés. A l'intersection de la ligne i et de la colonne j de cette matrice, on trouve la covariance, calculée uniquement sur les points crottiers, entre la variable environnementale i et la variable environnementale j.

Sous l'hypothèse que le nuage de points utilisés suit une loi normale multivariée, le vecteur $\boldsymbol{\mu}$ et la matrice $\boldsymbol{\Sigma}$ suffisent à caractériser le nuage de points utilisés [5]. Ainsi, si nous cherchons à cartographier la conformité à l'habitat sur une nouvelle zone, nous pourrons nous servir de ces deux éléments pour calculer la distance de Mahalanobis entre chaque point de la nouvelle zone et le centre de gravité du nuage de

petite fraction de l'habitat d'hivernage (i.e. les zones favorables à l'établissement de crottiers). Nous pourrons utiliser le terme "niche d'hivernage" dans ce document pour désigner cette petite fraction, mais nous sommes conscients qu'il s'agit d'un abus de langage (que nous jugeons sans gravité)

5. En pratique, comme le notent Clark *et al.* (1993), nous n'aurons pas besoin de supposer la normalité de la niche. En effet, si le nuage de points utilisés montre une distribution unimodale, cela indique qu'il y a une combinaison de valeurs des variables environnementales pour laquelle l'environnement est "idéalement conforme" à l'habitat, et plus on s'éloigne de cette combinaison, moins l'environnement est conforme. Si la niche n'est pas trop asymétrique, alors le vecteur $\boldsymbol{\mu}$ représente *grossièrement* la position de cet environnement idéalement conforme, et la distance de Mahalanobis restera un bon outil pour mesurer l'éloignement d'un nouveau point à cet habitat

points utilisé. Pour un point donné appartenant à cette nouvelle zone, notons $\mathbf{x}$ le P-vecteur donnant la valeur de chacune des P variables environnementales à ce point. La distance de Mahalanobis entre ce point et le centre de gravité du nuage de points utilisés est calculée par [6] :

$$D^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

En calculant cette distance pour tous les quadrats de la nouvelle zone, nous pouvons en déduire une carte de conformité à l'habitat sur cette nouvelle zone. Clark *et al.* (1993) indiquent qu'il est possible de transformer la carte des distances de Mahalanobis en carte de probabilité de présence de l'espèce. En effet, sous l'hypothèse que la niche d'hivernage a une forme normale multivariée, alors les distances de Mahalanobis entre les sites occupés par l'espèce et l'habitat optimum suivent une loi du χ^2 à P degrés de liberté. Il est alors possible de construire une carte de probabilité de présence en recherchant, pour chaque distance de Mahalanobis prédite, la probabilité d'observer cette distance pour un point tiré au sort dans la niche d'hivernage de l'espèce. Comme nous ne souhaitons pas nous appuyer trop fortement sur l'hypothèse de normalité multivariée de la niche (probablement fausse dans notre étude), cette utilisation de la loi du χ^2 doit surtout être vue comme un moyen d'échelonner la conformité à l'habitat prédite entre 0 et 1 (Alors que les distances de Mahalanobis en elles-mêmes n'ont aucune borne supérieure Clark *et al.*, 1993). La probabilité ainsi renvoyée ne doit donc pas être considérée comme une probabilité de présence, mais comme un indice de conformité à l'habitat compris entre 0 et 1.

5.2 L'analyse factorielle des distances de Mahalanobis

5.2.1 Augmenter le biais de la prédiction pour diminuer l'erreur de prédiction

Comme dans toute approche de modélisation, la modélisation prédictive doit trouver un bon compromis entre la justesse et la précision des estimations. Nous décrivons rapidement ces deux concepts dans cette section.

Dans toute approche de modélisation, nous supposons qu'un mécanisme inconnu $g(.)$ est à l'origine des données :

$$y_i = g(\mathbf{x}_i) + \epsilon_i$$

, avec ϵ_i un résidu aléatoire. Ce mécanisme dépend de variables prédictrices $\mathbf{x}_i$, et génère les valeurs de la variable réponse y_i dans l'unité statistique i (dans notre étude, un mécanisme inconnu génère la présence et la détection des points crottiers). Ce mécanisme étant inconnu, l'objectif de la modélisation consiste à *approcher* ce mécanisme à l'aide d'un modèle $f(\mathbf{x}_i)$ (dans le cas des distances de Mahalanobis, ce modèle est, on l'a vu, la loi normale multivariée), qui dépend d'un certain nombre de coefficients (dans notre cas, le vecteur des moyennes $\boldsymbol{\mu}$ et la matrice de variances-covariances $\boldsymbol{\Sigma}$), lesquels déterminent la façon dont les variables prédictrices sont combinées calculer la variable réponse. Comme nous sommes ordinairement incapables de fixer à priori des valeurs pertinentes pour ces coefficients, nous devons collecter des données générées par le mécanisme (dans notre étude, des points crottiers sur des zones restreintes), qui nous permettront de *calibrer* au mieux le modèle f.

Cette calibration s'effectue à l'aide de l'estimation des paramètres du modèle. Par exemple, dans le cas des distances de Mahalanobis, chaque composante du vecteur $\boldsymbol{\mu}$ est calculée comme la moyenne de la variable environnementale correspondante, calculée uniquement sur les points crottiers (on parle d'estimation *plug-in*). Bien sûr, le mécanisme inconnu comporte une composante aléatoire. Si nous devions recueillir un nouveau jeu de données sur la même zone d'étude, il y a peu de chances que nous retrouvions exactement le même nombre de crottiers, ni que ces crottiers soient localisés exactement au même endroit. Ainsi, si l'on devait recueillir un nouveau jeu de données, et procéder à une nouvelle estimation des coefficients, les valeurs des coefficients estimés seraient différentes, puisque cette estimation serait basée sur des données qui ne seraient pas les mêmes. Et donc les prédictions obtenues à partir de ce second

6. A proprement parler, cette équation permet de calculer le *carré* de la distance de Mahalanobis. Dans ce document, nous utilisons le terme "distance de Mahalanobis" de façon elliptique pour désigner D^2.

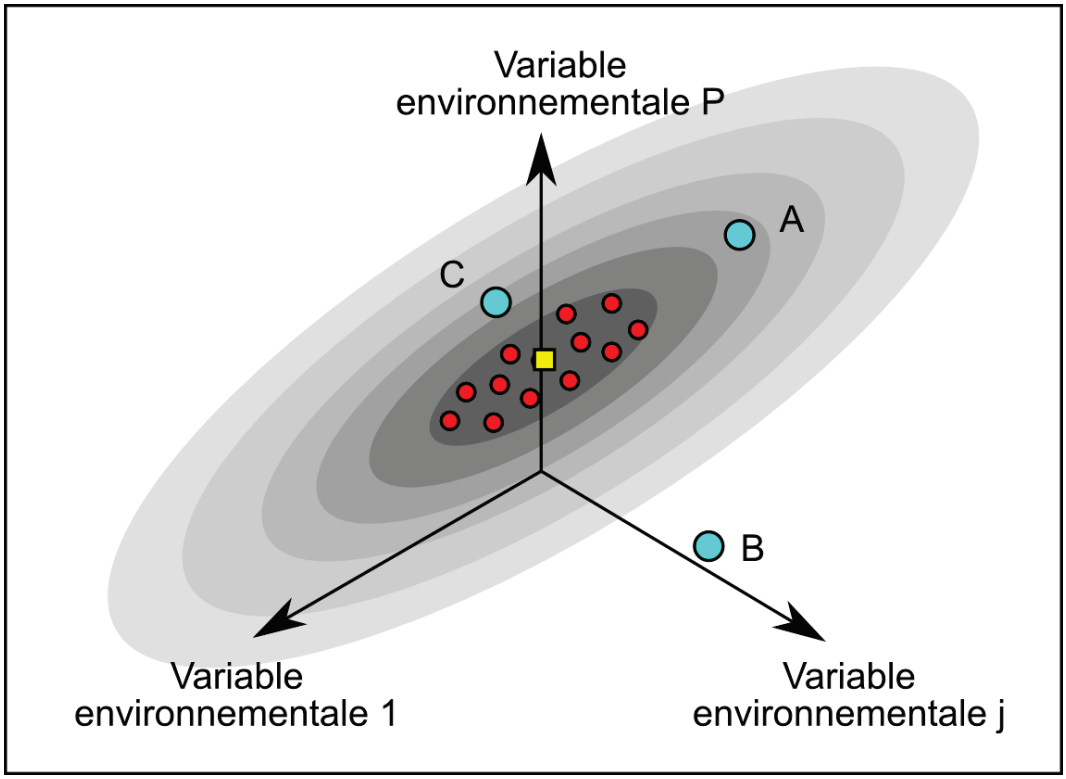

FIGURE 9 – Calcul des distances de Mahalanobis dans l'espace écologique. La valeur de P variables environnementales est mesurée pour chaque crottier identifié sur une zone d'étude. Dans le cas présent, nous illustrons le principe de la méthode avec $P = 3$ variables, mais le principe de cette méthode reste le même lorsque P augmente. Chacune de ces variables définit une dimension dans un espace multidimensionnel (espace écologique). Pour chaque crottier, la valeur d'une variable environnementale donnée définit une coordonnée sur la dimension correspondante. Donc à chaque crottier, on peut associer un point dans le repère cartésien défini par les variables environnementales (en rouge sur la figure). L'ensemble des crottiers définit un nuage de points que nous avons appelé la *niche d'hivernage* du tétras-lyre sur cette zone. Le centre de gravité de ce nuage de point définit les conditions environnementales *idéalement conformes*. Considérons les conditions environnementales mesurées à un point A situé sur une nouvelle zone. Ces conditions définissent un point A dans l'espace écologique. Plus la distance, mesurée dans l'espace écologique, entre le centre de gravité G de la niche (en jaune) et ce point est grande, et moins l'environnement est conforme à l'habitat de l'espèce. Les distances de Mahalanobis permettent d'intégrer les structures de corrélation dans la mesure de la conformité à l'habitat. En effet, considérons le point B. Mathématiquement la distance euclidienne "classique" entre B et le centre de gravité G est identique à la distance euclidienne entre A et G. Pourtant, le lecteur comprendra bien que la conformité à l'habitat est probablement plus faible en B qu'en A, la niche étant plus étroite lorsqu'on s'éloigne de G en direction de B que lorsqu'on s'éloigne de G en direction de A. Les niveaux de gris sur cette figure représentent différentes classes de distances de Mahalanobis, qui montre comment les structures de corrélation de la niche sont prises en compte dans le calcul de ces distances. Ainsi, les points C et A se trouvent à égale distance de Mahalanobis de G.

modèle calibré ne seraient pas les mêmes que celles obtenues à partir du premier.

Imaginons que nous collections une infinité de jeux de données en suivant exactement le même dispositif d'échantillonnage, pour chacun de ces jeux de données r, nous notons $\hat{f}_r(.)$ le modèle calibré associé. Pour une combinaison de valeurs de variables environnementales $\mathbf{x}$ donnée, le biais associé à un modèle f est :

$$\text{Biais} = \left(E(\hat{f}_r(\mathbf{x})) - g(\mathbf{x})\right)^2$$

avec $E(\hat{f}_r(\mathbf{x}))$ la moyenne des prédictions des modèles calibrés r (on parle d'*espérance*, que l'on note $E(.)$). La variance associée à un modèle r donné est :

$$\text{Variance} = E\left(\left(\hat{f}_r(\mathbf{x}) - E(\hat{f}_r(\mathbf{x}))\right)^2\right)$$

Ainsi, la variance mesure la dispersion des prédictions autour de la prédiction moyenne qui serait obtenue sur une infinité de jeux de données, et le biais mesure l'écart entre cette prédiction moyenne et la valeur réelle de la variable à prédire. Si nous disposons d'une nouvelle unité statistique i (e.g. un point provenant d'une nouvelle zone), mais que la valeur de la variable réponse y_i est inconnue, *l'erreur de prédiction* associée au modèle r pour cette unité peut être mesurée par la *fonction de perte* :

$$L = (y_i - \hat{f}_r(\mathbf{x_i}))^2$$

avec $\mathbf{x}_i$ le vecteur contenant les valeurs des variables prédictrices pour cette unité statistique. Cette erreur de prédiction augmente avec la variance associée au modèle, son biais, et la variance de la composante aléatoire ϵ_i du mécanisme ayant généré les données. Comme nous n'avons aucun contrôle sur cette dernière source d'erreur (puisque partie intégrante du mécanisme ayant généré les données), les seules composantes sur lesquelles nous pouvons agir pour diminuer l'erreur de prédiction sont le biais et la variance associée au modèle (Section 2.9 dans Friedman *et al.*, 2008). Ainsi, dans un contexte de prédiction, il peut être avantageux d'augmenter le biais associé à un modèle si cela permet de diminuer l'erreur de prédiction en diminuant la variance associée au modèle (Friedman *et al.*, 2008).

Or, la variance d'un modèle sera d'autant plus grande que le nombre de paramètres à estimer (la *dimension* du modèle) sera grand. Inversement, plus le nombre de paramètres d'un modèle sera grand et plus le biais sera faible. Il est donc nécessaire de choisir le nombre de paramètres qui permettra d'optimiser le compromis biais-variance.

5.2.2 Diminuer la dimension de l'espace pour diminuer l'erreur de prédiction : la MADIFA

Dans le contexte de l'approche basée sur les distances de Mahalanobis, les paramètres du modèle sont les composantes du vecteur $\boldsymbol{\mu}$ décrivant le centre de gravité du nuage de points utilisés, et les composantes de la matrice de variances-covariances $\boldsymbol{\Sigma}$ décrivant les structures de corrélation de ce nuage [7]. En conséquence, plus le nombre de variables environnementales considérées est important, et plus le nombre de paramètres à estimer sera important. Il peut alors être intéressant de diminuer le nombre de variables environnementales à prendre en compte dans le modèle (i.e. de modifier la structure du modèle f) de façon à faire diminuer l'erreur de prédiction globale, même si cela se traduit par une augmentation du biais. Mais comment sélectionner les variables environnementales à prendre en compte dans cette modélisation ?

Calenge *et al.* (2008) proposent une approche qui permet de diminuer la dimension du modèle sous-jacent à l'approche des distances de Mahalanobis sans avoir à diminuer le nombre de variables environnementales prises en compte. Cette approche s'appuie sur les points de contexte échantillonnés selon

7. Le modèle en lui-même correspond à la loi normale multivariée supposée décrire le nuage de points utilisé.

les recommandations de la section 4.4.

Si l'on dispose de P variables environnementales, l'espace multidimensionnel associé (l'espace écologique) est de dimension P. Calenge *et al.* (2008) proposent de rechercher, dans cet espace écologique, un nombre M plus réduit de directions dans lesquelles la distance de Mahalanobis moyenne entre les points de contexte et le centre de gravité des points crottiers est en moyenne la plus importante. Ces directions, ou *axes principaux*, correspondent aux directions dans lesquelles la sélection de l'habitat par l'espèce est la plus forte sur les zones étudiées. Ce sont des variables environnementales de synthèse, qui correspondent à des combinaisons linéaires des variables de départ, et qui représentent en général des facteurs limitants pour l'espèce. En diminuant le nombre de variables de synthèse (i.e. en choissant $M << P$), nous augmentons le biais associé aux prédictions, mais de façon négligeable en comparaison de la diminution de variance associée. Au final, cela permet d'obtenir des prédictions plus proches de la réalité.

Nous pouvons alors nous appuyer sur un nombre plus restreint d'axes biologiquement significatifs, et ne calculer ces distances qu'en ne nous concentrant sur ce nombre réduit d'axes. Cette analyse consistant à rechercher un nombre restreint d'axes sur lesquels la distance de Mahalanobis moyenne est maximale s'appelle la MADIFA (*Mahalanobis distance factor analysis*).

En pratique, l'approche est la suivante. Nous disposons d'une matrice $\mathbf{X}^*$ contenant la valeur des P variables environnementales (colonnes) à chacun des N_u points crottiers (lignes). Le centrage et la réduction de cette matrice nous permettent d'obtenir une matrice $\mathbf{X}$ (i.e. la moyenne de chaque variable de $\mathbf{X}$ est nulle et l'écart-type de chaque variable est égal à 1). Si l'on note x^*_{ij} la valeur de la j-ème variable environnementale du tableau $\mathbf{X}^*$, μ^*_j et σ^*_j respectivement la moyenne et l'écart-type de cette variable calculés sur les points crottiers, cette opération de centrage-réduction est effectuée par :

$$x_{ij} = \frac{x^*_{ij} - \mu^*_j}{\sigma^*_j}$$

et cette opération est effectuée pour tous les points crottiers i et toutes les variables j. Nous notons $\boldsymbol{\mu}^*_{\mathbf{X}}$ et $\boldsymbol{\sigma}^*_{\mathbf{X}}$ les vecteurs contenant respectivement les moyennes et les écart-types des variables de $\mathbf{X}^*$. Le centrage et la réduction basée sur ces vecteurs est notée :

$$\mathbf{X} = g(\mathbf{X}^*, \boldsymbol{\mu}^*_{\mathbf{X}}, \boldsymbol{\sigma}^*_{\mathbf{X}})$$

Soit $\mathbf{Y}^*$ la matrice contenant la valeur des P variables environnementales (colonnes) à chacun des N_d points de contexte (en ligne) de notre étude. Cette matrice est transformée exactement de la même façon que le tableau $\mathbf{X}^*$, c'est à dire :

$$\mathbf{Y} = g(\mathbf{Y}^*, \boldsymbol{\mu}^*_{\mathbf{X}}, \boldsymbol{\sigma}^*_{\mathbf{X}})$$

c'est à dire en utilisant les vecteurs $\boldsymbol{\mu}^*_{\mathbf{X}}$ et $\boldsymbol{\sigma}^*_{\mathbf{X}}$ calculés à partir du tableau $\mathbf{X}^*$. Cela permet d'assurer que ces deux tableaux seront comparables. Nous calculons alors la matrice $\boldsymbol{\Sigma}$ de covariance de $\mathbf{X}$:

$$\boldsymbol{\Sigma} = \mathbf{X}^t \mathbf{D}_u \mathbf{X}$$

avec $\mathbf{D}_u$ une matrice diagonale contenant $1/N_u$ sur la diagonale. Nous diagonalisons alors la matrice $\boldsymbol{\Sigma}$, c'est à dire que nous recherchons la matrice orthonormée de vecteurs propres $\mathbf{A}$ ainsi que la matrice diagonale des valeurs propres $\boldsymbol{\Lambda}$ telles que :

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^t$$

Nous calculons ensuite la matrice $\mathbf{G}$:

$$\mathbf{G} = \boldsymbol{\Lambda}^{-1/2}\mathbf{A}^t\mathbf{Y}^t\mathbf{D}\mathbf{Y}\mathbf{A}\boldsymbol{\Lambda}^{-1/2}$$

avec $\mathbf{D}$ une matrice diagonale contenant $1/N_d$ sur la diagonale. Et nous diagonalisons la matrice $\mathbf{G}$, c'est à dire que nous recherchons la matrice orthonormée de vecteurs propres $\mathbf{V}$ ainsi que la matrice diagonale des valeurs propres $\mathbf{\Theta}$ telles que :

$$\mathbf{G} = \mathbf{V}\mathbf{\Theta}\mathbf{V}^t$$

Nous définissons alors la matrice $\mathbf{B}$:

$$\mathbf{B} = \mathbf{A}\mathbf{\Lambda}^{-1/2}\mathbf{V}$$

Les premières colonnes de cette matrice sont des vecteurs qui définissent les directions de l'espace écologique dans lesquelles la distances de Mahalanobis moyenne entre les points de contexte et la niche sont maximales (ces directions sont orthogonales pour la métrique $\mathbf{\Sigma}^{-1}$, cf. Calenge *et al.*, 2008, pour plus de précisions). Nous pouvons alors définir la matrice $\mathbf{B}_M$ qui contient les M premières colonnes de la matrice $\mathbf{B}$. C'est cette matrice $\mathbf{B}_M$ qui nous servira pour nos prédictions. En effet, si l'on dispose d'un nouveau point provenant d'une nouvelle zone, pour lequel nous cherchons à estimer la conformité à l'habitat, alors si $\mathbf{h}^*$ est un vecteur contenant la valeur des variables environnementales mesurées à ce point, nous obtiendrons une approximation de la distance de Mahalanobis à l'environnement "idéalement conforme" (*reduced-rank Mahalanobis distances*) par :

$$D^2(\mathbf{h}^*) \approx D^2_a(\mathbf{h}^*) = g(\mathbf{h}^{*t}, \boldsymbol{\mu}^*_{\mathbf{X}}, \boldsymbol{\sigma}^*_{\mathbf{X}})^t \mathbf{B}_M \mathbf{B}^t_M g(\mathbf{h}^{*t}, \boldsymbol{\mu}^*_{\mathbf{X}}, \boldsymbol{\sigma}^*_{\mathbf{X}})$$

Le nombre M de variables de synthèse sélectionnées pour cette prédiction est choisi en examinant le diagramme des valeurs propres de l'analyse (contenues sur la diagonale de la matrice $\mathbf{\Theta}$), et en recherchant une "cassure" dans cette décroissance (dans notre étude, nous conserverons les trois premiers axes de l'analyse). Le lecteur souhaitant plus de détails sur cette analyse, et sur les raisons qui font de cette analyse un préliminaire optimal à une estimation de la conformité à l'habitat par les distances de Mahalanobis est encouragé à se reporter à l'article de Calenge *et al.* (2008).

Notons que comme pour les distances de Mahalanobis classiques, sous l'hypothèse que la niche d'hivernage a une forme normale multivariée, alors les distances de Mahalanobis approchées entre les sites occupés par l'espèce et l'habitat optimum suivent une loi du χ^2 à M degrés de liberté (cf. Rotenberry *et al.*, 2006). Il est alors possible de construire une carte de probabilité de présence en recherchant, pour chaque distance de Mahalanobis approchée prédite, la probabilité d'observer cette distance pour un point tiré au sort dans la niche d'hivernage de l'espèce. Comme pour les distances de Mahalanobis classiques, cette utilisation de la loi du χ^2 doit surtout être vue comme un moyen d'échelonner la conformité à l'habitat prédite entre 0 et 1 (Alors que les distances de Mahalanobis approchées en elles-mêmes n'ont aucune borne supérieure Clark *et al.*, 1993). La probabilité ainsi renvoyée ne doit donc pas être considérée comme une probabilité de présence, mais comme un indice de conformité à l'habitat compris entre 0 et 1.

5.3 La régression logistique "complète"

Nous avons montré en annexe qu'il était légitime d'utiliser des méthodes "classiques" pour modéliser la probabilité que $y = 1$ avec des données d'occurrence d'espèces complétées par des points de contexte, et que cette probabilité reflétait la probabilité de présence de l'espèce. Nous allons donc utiliser la régression logistique classique pour modéliser cette probabilité (McCullagh et Nelder, 1989). En d'autres termes, si l'on note $x_1, ..., x_P$ les P variables environnementales mesurées à un point, nous supposons que la probabilité de présence de l'espèce à ce point est reflétée par :

$$P(y = 1|\mathbf{x}) = \frac{\exp\{b_0 + b_1 x_1 + b_2 x_2 + ... + b_P x_p\}}{1 + \exp\{b_0 + b_1 x_1 + b_2 x_2 + ... + b_P x_p\}} \quad (1)$$

avec $b_0, b_1, ..., b_P$ des coefficients à estimer à partir des données par la méthode du maximum de vraisemblance. Ce type de modèle est extrêmement commun, disponible dans tous les logiciels de statistiques, et

donc très facile à ajuster.

Notons que l'équation 1 ne permet de modéliser que des relations monotones entre les variables prédictrices et la probabilité recherchée [8]. En d'autres termes, si le coefficient b_j est positif, alors plus la variable x_j sera importante (e.g. l'altitude), et plus la probabilité recherchée sera importante. Inversement, si le coefficient b_j est négatif, plus la variable x_j sera importante et moins la probabilité recherchée le sera. Or, pour plusieurs variables quantitatives (e.g. altitude, pente), une telle relation monotone est peu probable. En effet, il est plus vraisemblable de supposer qu'il existe un optimum de pentes, un optimum d'altitude, etc. Il est possible de modéliser un tel optimum en intégrant le carré de ces variables dans le modèle. Par exemple, l'effet de l'altitude sera :

$$b_1 \times \text{altitude} + b_2 \times \text{altitude}^2$$

Dans notre étude, nous avons ajouté un effet quadratique pour les variables quantitatives suivantes : altitude, pente et niveau moyen de vert dans les orthophotographies. Nous avons utilisé la fonction `glm` du logiciel R pour permettre l'ajustement de ce modèle.

5.4 Une régression logistique pas à pas (*stepwise*)

La régression logistique implémentée dans la section précédente implique d'ajuster un coefficient b_j par variable environnementale j ajoutée au modèle. Il est toutefois possible que certaines des variables environnementales sélectionnées ne jouent pas un grand rôle dans la prédiction de la conformité à l'habitat. Dans ce cas, il pourrait être intéressant de ne conserver dans le modèle que les variables ayant une influence réelle sur la conformité. En procédant ainsi, nous diminuerions la dimension du modèle, et bien que le biais associé au modèle en serait augmenté, cette augmentation serait minime en comparaison de la diminution de variance associée (cf. section 5.2.1 pour plus de détails sur ces notions de biais et de variance associées au modèle). La question se pose alors de savoir comment sélectionner les variables à inclure dans le modèle.

Il existe de nombreuses approches développées dans la littérature pour permettre cette réduction de la dimension d'un modèle (cf. Friedman *et al.*, 2008, pour une revue). Nous avons choisi une approche pas à pas (*stepwise*) basée sur le critère d'Akaiké (AIC) pour sélectionner ces variables (Venables et Ripley, 2002, p. 175).

Le critère d'Akaiké (AIC) a été introduit par Akaike (1973) comme estimation de la distance de Kullback-Leibler entre un modèle et la réalité [9]. Cet auteur démontre que, lorsque l'on souhaite comparer plusieurs modèles emboîtés ajustés par le maximum de vraisemblance (e.g un modèle A contenant 15 variables et un modèle B contenant seulement une partie – par exemple 6 – de ces variables), que ces modèles ne sont pas trop éloignés de la réalité, et que l'échantillon sur lequel le modèle est ajusté est très grand, alors une *estimation* de la distance relative de Kullback-Leibler entre un modèle et la réalité peut être obtenue par :

$$\text{AIC} = -2 \log L + 2 \times P$$

avec L la vraisemblance de l'échantillon (i.e., la probabilité d'obtenir l'échantillon sous l'hypothèse que le modèle est vrai). L'AIC nous donne un moyen de comparer plusieurs modèles emboités, le plus proche de la réalité étant celui qui est caractérisé par l'AIC en moyenne le plus faible [10].

8. une fonction monotone de x est une fonction toujours croissante ou toujours décroissante de x

9. La distance de Kullback-Leibler entre un modèle $f(x|\theta)$ – avec θ un vecteur de paramètres caractérisant le modèle et x les variables de ce modèle – et la réalité $g(x)$ est obtenue par (Burnham et Anderson, 1998, p. 36) :

$$\int_{-\infty}^{+\infty} g(x) \log \frac{g((x)}{f(x|\theta)}$$

10. Notons que l'AIC est une estimation de la distance de Kullback-Leibler, et non la distance elle-même. Comme toute estimation, elle est soumise aux fluctuations d'échantillonnage, et est donc caractérisée par une imprécision. Ainsi, si

Il est alors possible de développer une approche “pas à pas” automatisée pour la sélection des variables dans le modèle qui permet la comparaison de modèles emboîtés. Nous décrivons l'algorithme ci-dessous (Venables et Ripley, 2002, p. 175) :

1. Nous commençons par construire le modèle le plus complet (celui que nous avons ajusté dans la section précédente), prédisant la conformité à l'habitat en fonction des P variables environnementales. Nous calculons l'AIC associé à ce modèle. Ce modèle constitue le modèle “parent” ;

2. Pour chaque variable j du modèle parent, nous reconstruisons un modèle comprenant toutes les variables du modèle parent à l'exception de la variable j. Nous calculons alors les AIC associés à chacun des P modèles “descendants” ainsi construits ;

3. Nous sélectionnons alors le modèle présentant le plus petit AIC. S'il s'agit du modèle parent, alors l'algorithme se termine, et le modèle sélectionné est le modèle parent. Si l'AIC le plus faible est obtenu pour l'un des modèles descendants, nous remplaçons le modèle parent par ce modèle, et nous recommençons à l'étape 2.

Nous utilisons alors le modèle final comme modèle prédictif (voir section 9.2 pour une discussion approfondie de cette approche). Nous avons utilisé la fonction `step` du logiciel R pour l'application de cet algorithme.

5.5 Les forêts d'arbres de décision

Nous allons également utiliser une autre méthode s'appuyant sur des données de type présence/absence prédire la conformité à l'habitat. Plus précisément, nous modéliserons la probabilité que $y = 1$ à l'aide de forêts d'arbres décisionnels, aussi appelées forêts aléatoires ou *random forests*[11] (Breiman, 2001; Cutler *et al.*, 2007), de façon à prédire la probabilité relative de présence en utilisant les points de contexte.

Les forêts aléatoires reposent sur une autre méthodologie appelée “arbres de classification”, couramment utilisée pour la prédiction statistique (Classification and regression trees, ou CART ; Breiman *et al.*, 1993; De'ath et Fabricius, 2000). Nous devons donc tout d'abord décrire le principe de cette méthodologie pour comprendre celui des forêts aléatoires. Nous disposons d'une variable y_i à prédire (prenant la valeur 1 si le point i est un point crottier ou 0 si ce point est un point de contexte), et d'un ensemble de variables prédictrices (la composition environnementale). L'ajustement d'un arbre de classification consiste à rechercher, variable par variable, le “point de coupure” (*cutpoint*) optimum, c'est à dire qui aboutira à la meilleure séparation des points crottiers et des pseudo-absences. Par exemple, la première étape de l'algorithme pourra identifier deux groupes caractérisés par le point de coupure “altitude = 2000 mètres” : cela signifiera que nous définirons deux groupes “points localisés à une altitude inférieure à 2000 mètres” et “points localisés à une altitude supérieure à 2000 mètres”. Puis, au sein de ces deux groupes, nous réitérerons l'opération, en recherchant un autre point de coupure optimum, qui subdivisera l'un des deux groupes en deux sous-groupes qui maximisent la séparation des deux types de points au sein du groupe. Par exemple, la deuxième étape de l'algorithme pourrait identifier un deuxième point de coupure “pente $>$ 10%” au sein du groupe “altitude $<$ 2000 mètres”, définissant deux sous-groupes au sein de ce groupes. L'opération est réitérée jusqu'à obtention d'un arbre dont les feuilles (i.e. les noeuds terminaux) correspondent à des classes “pures” (ne contenant qu'un seul type de points : crottier ou contexte). Les points de coupure optimums sont déterminés à l'aide d'un critère mesurant l'homogénéité des classes (il s'agit en général de l'indice de diversité de Simpson).

Une fois l'arbre construit, il faut alors procéder à une étape *d'élaguage* (suppression de certaines parties de l'arbre), de façon à atteindre un compromis entre le biais et l'imprécision de l'arbre comme

plusieurs modèles sont caractérisés par des AIC très peu différents (différences entre les AIC inférieure à 2), alors il sera difficile de déterminer quel modèle est le plus proche de la réalité.

11. Nous utiliserons indistinctement les trois termes.

modèle prédictif. Nous ne décrivons pas le principe de cet élaguage, car les arbres de classification ne sont pas la méthodologie qui nous intéresse ici.

Le principal intérêt de cette méthodologie est sa simplicité d'interprétation. Malheureusement, cette méthodologie est également très instable. Friedman *et al.* (2008, section 9.2.4) soulignent : *One major problem with trees is their high variance. Often a small change in the data can result in a very different series of splits, making interpretation somewhat precarious. The major reason for this instability is the hierarchical nature of the process : the effect of an error in the top split is propagated down to all of the splits below it (...). It is the price to be paid for estimating a simple, tree-based structure from the data.* Ainsi, que l'on supprime une unité statistique et la topologie de l'arbre peut changer du tout au tout... sa topologie, mais pas nécessairement son pouvoir prédictif ! en effet, deux arbres très différents peuvent renvoyer des prédictions similaires lorsque les variables prédictrices sont corrélées entre elles (Breiman *et al.*, 1993).

Les forêts aléatoires tirent parti de cette instabilité. Le principe de cette méthodologie consiste à construire des jeux de données légèrement perturbés à partir des données originales. Nous construisons alors un arbre de classification pour chaque jeu de données. Nous disposons alors d'un grand nombre d'arbres de classification (une "forêt"), chacun représentant un point de vue sur les données. Plus précisément, ces perturbations sont :

- Des perturbations des données, utilisant le *bagging* (*Bootstrap aggregating*) : nous allons prendre un grand nombre d'échantillons bootstrap (i.e. si nous disposons de N points dans notre jeu de données de calibration original, un échantillon bootstrap est obtenu en tirant au sort *avec remise* N points de ce jeu de données - donc certains points seront répétés plusieurs fois dans notre échantillon, et d'autres en seront absents). Nous ajusterons alors un arbre par échantillon bootstrap (les différents arbres de classification forment alors la forêt).

- des perturbations de l'algorithme de construction de l'arbre. En théorie, lorsque l'on construit un arbre de classification, on recherche à chaque étape, parmi *toutes* les variables prédictrices, le meilleur point de coupure. Lors de l'ajustement d'une forêt aléatoire, nous allons modifier ce comportement : nous imposerons que l'algorithme recherche, à chaque noeud de chaque arbre, le meilleur point de coupure *en ne considérant qu'un nombre restreint de variables prédictrices.*

Chaque arbre de la forêt aléatoire représente un point de vue sur les données. Les perturbations ajoutées assurent que chaque point de vue est différent des autres (La corrélation entre les arbres est minimale, Breiman, 2001). Il est alors possible d'utiliser la forêt pour faire des prédictions : chaque arbre de la forêt prédit une classe pour tout nouveau point à évaluer (présence ou pseudo-absence de crottiers). Nous faisons alors *voter* les arbres : si une majorité d'arbres prédit que le point est dans la catégorie des "présences", nous prédirons de même.

Notons que la construction d'une forêt aléatoire ne nécessite pas d'étape d'élaguage lors de la construction des arbres : chaque arbre est construit jusqu'à l'obtention de classes "pures".

Liaw et Wiener (2002) indiquent, au sujet des forêts aléatoires : "*it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values*". C'est cette caractéristique qui rend la forêt aléatoire attractive dans notre contexte.

Nous devons toutefois faire attention à l'impact négatif de la différence entre le nombre de points crottiers et le nombre de points de contexte. En effet, imaginons un cas extrême, dans lequel nous disposons de 30 points crottiers et de 999970 points de contexte. Une telle différence entre les effectifs de crottiers et de points de contexte va jouer très fortement sur l'algorithme : celui-ci est conçu pour minimiser l'erreur de prédiction *globale* de la forêt. Or, une forêt aléatoire prédisant systématiquement $y = 0$ quelles que soient les valeurs des variables environnementales ne conduirait pas à une erreur de prédic-

tion globale élevée avec de telles données. En fait, une telle forêt ne se tromperait que dans 0.03% des cas. C'est que dans ce cas, les erreurs d'omission (un point crottier incorrectement classé comme point de contexte) et de commission (un point de contexte incorrectement classé comme point crottier) n'ont pas le même poids. Il est alors nécessaire de modifier l'algorithme sous-jacent aux forêts aléatoires pour donner le même poids à ces deux types d'erreur. En pratique, si l'on dispose de N_u points crottiers et de N_d points de contexte, avec $N_u < N_d$, nous devrons modifier l'algorithme de bootstrap de la façon suivante. Pour chaque arbre de la forêt l'algorithme de bootstrap est remplacé par l'algorithme suivant :

- ⋆ N_u points crottiers sont tirés au sort avec remise parmi les N_u points crottiers disponibles ;
- ⋆ N_u points de contextes sont tirés au sort avec remise parmi les N_d points de contexte disponibles ;
- ⋆ Le jeu de données utilisé pour la construction de l'arbre est effectuée sur ces $2 \times N_u$ points échantillonnés.

Cet algorithme permet de donner le même poids aux erreurs d'omission et de commission, et constitue d'après Chen *et al.* (2004) l'une des méthodes les plus efficaces pour la prédiction basée sur des jeux de données non équilibrés (cette approche est alors appelée *Balanced random forest*). Nous avons ajusté ce type de modèle avec la fonction `randomForest` appartenant au package du même nom (Liaw et Wiener, 2002). Notre forêt contenait 500 arbres, et pour chaque noeud de chacun des arbres ajustés, nous avons tiré au sort 4 variables parmi les variables environnementales disponibles [12].

6 La validation des modèles

6.1 Une modélisation en théorie en trois étapes

En théorie, une stratégie de modélisation prédictive doit se dérouler en trois étapes (Friedman *et al.*, 2008, Section 7.2) :

- ⋆ Etape de *calibration* des modèles : cette étape consiste à estimer les coefficients de chacun des modèles sélectionnés à partir des données. Nous appliquerons donc la méthode des distances de Mahalanobis, la MADIFA, les fonctions de sélection des ressources, et les forêts aléatoires pour construire ces modèles ;
- ⋆ Etape de sélection des modèles ou *validation* : cette étape consiste à comparer les différents modèles ajustés lors de la première étape sur la base de leur pouvoir prédictif, et à choisir le meilleur ;
- ⋆ Etape de *test* (*model assessment*) : cette étape consiste à estimer le pouvoir prédictif du meilleur modèle sélectionné à l'étape précédente (i.e. l'erreur de généralisation ou *generalization error*)

Un premier point essentiel à comprendre lorsque l'on aborde la question de la mesure du pouvoir prédictif d'un modèle est qu'il ne faut jamais utiliser les données qui ont servi à calibrer le modèle (que l'on appelle *jeu de données de calibration*) pour évaluer son pouvoir prédictif (Friedman *et al.*, 2008, cf. en particulier la section 7.2). En effet, cela conduit à surestimer ce pouvoir [13]. Il est donc nécessaire d'utiliser des données différentes à chacune des étapes de la construction du modèle.

12. Conformément aux recommandations de Breiman (2002), qui recommande de fixer cette valeur égale à $\sqrt{P}$, avec P le nombre de variables explicatives (dans notre étude, $P = 18$ pour la modélisation restreinte à la Haute Savoie, et $P = 16$ pour la modélisation générale

13. En outre, certaines méthodes comme les *random forests* sont par construction caractérisées par un taux d'erreur nul si l'on ne considère que le jeu de données de calibration. En effet, comme chaque arbre de la forêt est construit de telle façon que les feuilles de l'arbre sont caractérisées par une composition pure (dans une feuille donnée, toutes les observations appartiennent à la même classe – soit point crottier, soit pseudo-absence), pour une observation donnée du jeu de données de calibration, tous les arbres assignent la classe correcte. Donc 100% des observations du jeu de données de calibration sont bien classées, et ce, même si la forêt aléatoire a un très faible pouvoir prédictif (voir la discussion à l'adresse suivante : `http://r.789695.n4.nabble.com/Random-Forest-AUC-td3006649.html`)

En théorie, la voie royale pour permettre cette construction consiste donc à recueillir trois jeux de données différents pour mener à bien les trois étapes de cette construction : un premier jeu de données, appelé *jeu de données de calibration*, servirait à calibrer le modèle ; un second jeu de données, appelé *jeu de données de validation* servirait à sélectionner le meilleur modèle. Un dernier jeu de données, appelé *jeu de données de test* serait alors utilisé pour évaluer le pouvoir prédictif du meilleur modèle. En pratique, chacun des modèles ajustés sur les données de calibration est utilisé pour prédire la variable réponse dans le jeu de données de validation. Un critère mesurant la similarité entre la prédiction et les observations dans ce jeu de données sert alors de moyen de mesurer le pouvoir prédictif. Lorsque plusieurs modèles prédictifs sont comparés, le modèle pour lequel le critère est le plus élevé est alors choisi. Enfin, ce meilleur modèle est appliqué sur les données de test afin de mesurer la similarité entre les prédictions et les observations de ce jeu de données, le plus souvent à l'aide du même critère. Bien sûr, si cette solution est intéressante sur un plan statistique, elle a un coût important et ne sera pas applicable dans notre cas de figure [14].

Dans cette étude, nous nous servirons du jeu de données de validation pour procéder à l'étape de test, ne disposant pas de données séparées nous permettant de mener à bien une étape de test indépendante. En revanche, nous procéderons à une étape de validation assez poussée, que nous décrivons dans la section suivante.

6.2 L'étape de validation

L'étape de validation sera effectuée en deux étapes distinctes. Nous décrivons ces deux étapes dans cette section :

Validation interne : De nombreux auteurs proposent, pour passer l'étape de sélection des modèles, de mesurer le pouvoir prédictif du modèle à l'aide d'une approche par validation croisée (Guisan et Zimmermann, 2000; Friedman *et al.*, 2008; Boyce *et al.*, 2002). Le principe de cette approche est le suivant :

1. Nous découpons le jeu de données initial en K groupes, chaque groupe contenant une partie des observations collectées dans l'étude. Les morceaux sont numérotés de 1 à Q. Nous définissons l'indice $q = 1$.

2. Nous mettons le groupe d'observations q de côté, et nous ajustons chaque modèle à l'aide des données constituées par les Q-1 groupes restants, poolés pour faire un jeu de données unique. Nous utilisons alors les modèles ajustés pour prédire la variable réponse dans le morceau q mis de côté. Les prédictions sont ainsi effectuées sur un jeu de données qui n'a pas servi à ajuster le modèle.

3. Nous incrémentons q de un, et nous répétons l'étape 2 jusqu'à ce que nous disposions, pour chacune des observations de chacun des groupes, de prédictions effectuées à l'aide d'un modèle ajusté sur d'autres données.

4. Pour l'ensemble du jeu de données, nous disposons d'une variable réponse et d'une prédiction. Nous utilisons un critère mesurant la similarité entre ces deux variables, qui reflétera le pouvoir prédictif du modèle.

Le plus souvent, les Q groupes sont de taille égale, et les observations du jeu de données sont assignées de façon aléatoire à chacun des groupes. Dans ce cas de figure, cette approche s'appelle la *K-fold*

14. Le lecteur pourrait objecter que nous disposons de plusieurs zones d'étude, et que certaines d'entre elles pourraient être mises de côté pour constituer des jeux de données de validation et de test. Cependant, le pouvoir prédictif d'un modèle sera d'autant plus fort que le volume des données ayant servi à l'ajuster est important. Nous ne disposons pas d'un grand volume de données (car les surfaces prospectées sont relativement faibles au regard de la surface d'intérêt – les Alpes du nord). Notre objectif étant d'utiliser le modèle construit pour prédire la conformité à l'habitat sur l'ensemble des Alpes du nord, nous avons désespérément besoin de toutes les zones d'études pour calibrer le modèle. Nous verrons d'ailleurs dans les résultats que plus les contextes environnementaux utilisés pour calibrer les modèles sont diversifiés, et plus notre modèle sera généralisable.

cross validation (Friedman *et al.*, 2008, section 7.10.1). Cependant, dans notre étude, nous utiliserons une approche différente.

En effet, nous disposons de K zones d'études séparées (3 pour la modélisation restreinte à la haute savoie, 4 pour la modélisation sur l'ensemble des Alpes du nord). Ces K zones d'étude définissent une partition naturelle du jeu de données dont nous disposons. Nous nous servirons donc de cette partition naturelle pour définir les groupes dans l'étape de validation croisée. Par exemple, pour la modélisation restreinte à la Haute Savoie, nous procéderons ainsi :

- ⋆ Nous mettrons les données collectées à Flaine de côté et nous ajusterons nos modèles prédictifs sur l'ensemble des données de Giettaz et des Saisies. Nous utiliserons chacun des modèles pour prédire la probabilité que chaque point du jeu de données de Flaine soit un point crottier et non un point de contexte ;
- ⋆ Nous mettrons les données collectées à Giettaz de côté et nous ajusterons nos modèles prédictifs sur l'ensemble des données de Flaine et des Saisies. Nous utiliserons chacun des modèles pour prédire la probabilité que chaque point du jeu de données de Giettaz soit un point crottier et non un point de contexte. ;
- ⋆ Nous mettrons les données collectées aux Saisies de côté et nous ajusterons nos modèles prédictifs sur l'ensemble des données de Flaine et de Giettaz. Nous utiliserons chacun des modèles pour prédire la probabilité que chaque point du jeu de données des Saisies soit un point crottier et non un point de contexte ;

Nous utiliserons alors un critère pour mesurer l'accord entre la prédiction sur tout les quadrats et la réalité. Nous examinerons la valeur de ces critères afin d'évaluer la capacité de chacun des modèles à prédire la conformité à l'habitat dans des zones non utilisées pour calibrer les modèles.

Validation Externe : cette seconde étape de validation ne sera effectuée que pour la modélisation de l'ensemble des Alpes du nord. L'OGM dispose par ailleurs de données collectées dans deux autres zones d'étude distribuées sur les Alpes du nord (cf. figure 1) : le parc naturel du Vercors (département de la Drôme) et la réserve naturelle de Villaroger (Savoie). Ces jeux de données ne sont pas de même nature que ceux utilisés pour ajuster les données. Il s'agit d'*observations occasionnelles* collectées de 1998 à 2010 sur ces zones d'étude par les partenaires de l'OGM. Il n'y a donc pas de dispositif d'échantillonnage bien défini (il s'agit d'un échantillonnage opportuniste), ni de limites bien définies aux zones prospectées. Il n'est donc pas opportun d'inclure ces données dans le jeu de données servant à calibrer les modèles de prédiction [15]. En revanche, nous pouvons toujours regarder si les zones dans lesquelles des crottiers ont été détectés par les observateurs correspondent bien à des zones jugées favorables par les modèles. Ceci nous fournira donc une certaine validation externe des modèles ajustés. Concrètement, nous définirons des pseudo-limites à ces zones d'étude à partir du semis de crottiers détectés sur ces zones. Ces limites seront définies à l'aide du polygone convexe minimum entourant les localisations de chaque zone. Pour chacune des zones, nous identifierons les quadrats de la grille européenne recouvert partiellement ou totalement par le polygone convexe. Nous en déduirons un échantillon de points de contexte (centres de ces quadrats) pour lesquels la composition environnementale est connue. Par ailleurs, nous déterminerons également la composition environnementale à chacun des points crottiers dans ces zones. Pour chacune de ces zones de validation, nous disposerons donc d'une structure de données similaire à celle que nous avons utilisée pour la calibration des modèles (figure 8). Nous utiliserons donc chacun de nos modèles (distances de Mahalanobis, MADIFA, régression logistique complète et pas à pas, et forêt d'arbres décisionnels) pour prédire la conformité à l'habitat à chacun des points de ces zones. Puis, nous utiliserons un critère pour mesurer l'accord entre la prédiction de ces modèles et la réalité. Nous décrivons dans la section suivante les trois critères que nous avons utilisés pour sélectionner nos modèles.

15. En effet, la distribution observée des crottiers dépend non seulement de la distribution réelle des crottiers, mais aussi de la distribution des observateurs, laquelle est probablement plus dense à proximité des zones plus facile d'accès.

6.3 Les critères pour mesurer le pouvoir prédictif du modèle

6.3.1 Trois critères choisis

Nous abordons maintenant la question de la mesure de l'accord entre la prédiction d'un modèle sur une zone et la réalité. Le choix d'un critère mesurant cet accord est crucial dans une approche de prédiction, car la comparaison des différents modèles ajustés repose très fortement dessus (Friedman *et al.*, 2008). Il est alors préférable de ne pas se focaliser sur un seul critère, mais d'en utiliser plusieurs, afin de pouvoir juger de la fiabilité de nos décisions. C'est ce que nous ferons dans cette étude, en nous servant de trois critères distincts.

De nombreux critères ont été développés dans la littérature pour mesurer l'accord entre la prédiction de la probabilité de présence d'un évènement et des données de type présence/absence (cf. Fielding et Bell, 1997, pour une revue). La difficulté de notre étude est que nous ne disposons pas de ce type de données, mais de données de type présence/points de contexte. Or, il existe assez peu de critères développés pour mesurer l'accord entre prédiction et réalité spécifiquement pour ce type de données. Le plus utilisé est probablement le critère de Boyce *et al.* (2002), ainsi que son amélioration par Hirzel *et al.* (2006). Nous décrirons ce critère dans la section suivante, et nous nous en servirons pour sélectionner nos modèles.

Il est toutefois important de noter que certaines approches conçues pour les données de type présence/absence sont valides pour le type de données dont nous disposons (Phillips *et al.*, 2006). Nous allons décrire dans les prochaines sections deux de ces approches, le critère AUC (*area under the curve*) et le coefficient de corrélation bisériale ponctuelle. Nous utiliserons également ces deux critères pour nous aider dans le processus de sélection de modèles. Nous pourrons ainsi juger si trois critères différents nous conduisent à sélectionner les mêmes modèles dans les étapes de validation des modèles.

6.3.2 Le critère de Boyce *et al.* (2002) continu

Nous décrivons ici le principe de l'indice de Boyce *et al.* (2002), ainsi que son amélioration par Hirzel *et al.* (2006).

Imaginons que nous disposions d'un modèle prédictif de la présence d'une espèce. Nous disposons, pour chaque unité i d'un pool de N unités appartenant au jeu de données de validation, d'une prédiction de la probabilité relative de présence $\hat{c}_i$ de l'espèce d'après le modèle, ainsi que d'une variable y_i indiquant le statut de cette unité ($y_i = 1$ lorsque l'espèce est présente dans l'unité i et 0 s'il s'agit d'un point de contexte). Nous pouvons découper la variable prédite en G classes de valeurs prédites (e.g. classe 1 : $0 \leq \hat{c}_i < 0.1$, classe 2 : $0.1 \leq \hat{c}_i < 0.15$, etc.). Au sein de chaque classe g, nous pouvons : (i) calculer la valeur prédite "centrale" c_g (e.g. pour la classe $0 \leq \hat{c}_i < 0.1$, la valeur centrale sera de $c_g = 0.05$), (ii) calculer la proportion p_g de points du jeu de validation pour lesquels $y_i = 1$. Boyce *et al.* (2002) propose ensuite de calculer le coefficient de corrélation de Spearman entre les c_g et les p_g. Lorsque le modèle prédictif permet une bonne prédiction de la probabilité de présence, cette corrélation est proche de 1, ce qui permet un jugement de la qualité du modèle.

Hirzel *et al.* (2006) souligne toutefois que : (i) rien ne permet de guider le choix du nombre de classes B, (ii) l'indice de Boyce *et al.* (2002) est très sensible aux limites de classes. Ces auteurs proposent alors une amélioration de cet indice, en résolvant le second de ces problèmes. L'idée est d'utiliser une fenêtre glissante d'une certaine taille w (*moving window* ; e.g. de taille 0.1). Nous positionnons la fenêtre sur l'intervalle $[0, w]$, et nous calculons c_g et p_g au sein de cette fenêtre. Nous déplaçons alors progressivement cette fenêtre le long de l'intervalle $[0, 1]$, et nous recalculons à chaque fois c_g et p_g pour chaque nouveau positionnement de la fenêtre. Nous calculons enfin le coefficient de Spearman pour mesurer la corrélation entre ces deux variables.

Nous avons appliqué cette approche, après une petite modification. En effet, plutôt que de fixer une taille de fenêtre w constante, nous avons utilisé une fenêtre glissante de taille variable, mais contenant

un nombre constant de points de validation dans la fenêtre. Cela permet de s'assurer que la proportion p_g est toujours calculée sur le même nombre de points (et donc que l'estimation de cette proportion est toujours soumise aux mêmes fluctuations d'échantillonnage). En effet, si l'on fixait une taille de fenêtre constante, il y aurait un risque que la fenêtre, à un moment ou à un autre, ne contienne qu'un seul point, et la proportion p_i ne pourrait prendre que les valeurs 0 ou 1, ce qui donnerait une grande variabilité à l'estimation du coefficient de Spearman (pers. obs.). Pour notre étude, nous avons défini la fenêtre de telle façon à ce qu'elle contienne 20% du nombre total de points du jeu de validation, de façon à assurer une estimation suffisamment précise de la proportion p_i dans chaque fenêtre.

6.3.3 La corrélation bisériale ponctuelle

La corrélation bisériale ponctuelle (*point biserial correlation*, Lev, 1949) est une approche fréquemment utilisée pour comparer des modèles prédictifs de la présence d'espèce (Phillips *et al.*, 2009; Elith *et al.*, 2006). Si l'on reprend les notations de la section précédente, cette approche consiste simplement à calculer le coefficient de corrélation de Pearson entre la prédiction $\hat{c}_i$ et la variable réponse y_i. En théorie, tout comme le coefficient de corrélation classique, le coefficient de corrélation bisériale ponctuelle prend des valeurs comprises entre -1 (pire prédiction possible) et 1 (prédiction parfaite). Cependant, un modèle prédictif peut conduire à une séparation parfaite des points de contexte et des présences, sans que ce coefficient prenne une valeur égale à 1 (Cox et Wermuth, 1992). Un exemple est donné figure 10.

Ainsi, nous devrons éviter d'interpréter l'éloignement à 1 des valeurs de coefficients de corrélation bisériale ponctuelle obtenus comme des preuves de mauvais ajustement. Certes, plus le coefficient de corrélation ponctuelle bisérielle est positif, et meilleure est la relation entre prédiction et réalité. Mais nous ne connaissons pas à priori la valeur de ce coefficient pour laquelle une prédiction parfaite est obtenue. Nous nous servirons donc de ce coefficient uniquement comme critère de comparaison de modèles, conformément aux recommandations de plusieurs auteurs (e.g. Phillips *et al.*, 2009).

6.3.4 Le critère AUC

Comme nous l'avons noté dans les sections précédentes, l'AUC (*area under curve*) est une approche à l'origine développée pour mesurer la qualité de prédiction de modèles ajustés sur des données de type présence/absence. Cependant, l'utilisation de cette approche avec des données de type présence/points de contexte est légitime (cf. infra). Dans cette section, nous décrivons tout d'abord le principe de l'AUC cette approche pour le cas des données de présence/absence, puis nous justifierons son utilisation sur des données de présence/points de contexte.

Imaginons que nous disposions d'un modèle prédictif de la présence/absence d'une espèce. Nous disposons, pour chaque unité i d'un pool de N unités appartenant au jeu de données de validation, d'une prédiction de la probabilité de présence $\hat{c}_i$ de l'espèce d'après le modèle, ainsi que d'une variable y_i indiquant le statut de cette unité ($y_i = 1$ lorsque l'espèce est présente dans l'unité i et 0 sinon). Nous pouvons transformer la probabilité prédite $\hat{c}_i$ en prédiction de classe, en définissant une valeur seuil c_s (Hosmer et Lemeshow, 2000, p. 156). Par exemple, nous pourrions fixer le seuil à 0.5. Alors, si la probabilité de présence de l'espèce, prédite par le modèle, est de 0.6, nous considérons l'espèce présente dans l'unité. Si elle est de 0.4, nous la considérons absente.

Donc, pour un seuil donné, nous sommes en mesure de construire un **classifieur** permettant de classer chaque observation dans une classe (présence ou absence de l'espèce). A un classifieur donné, nous sommes en mesure d'associer une **matrice de confusion**, qui confronte les prédictions et les observations (Tableau 3).

Cette matrice joue un rôle très important dans la mesure de la capacité de prédiction du classifieur : Chacune des cases de cette matrice donne le nombre d'unités pour lesquelles le classifieur *prédit* une présence ou une absence, et pour lesquels on *observe* une présence ou une absence. Cette matrice con-

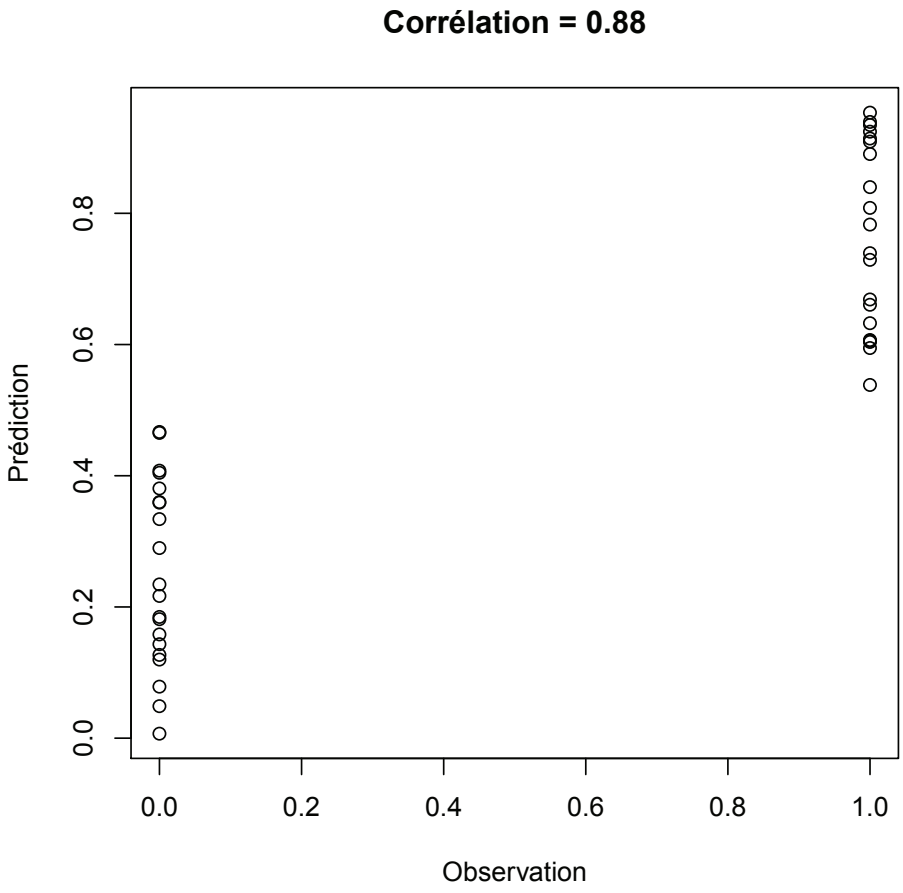

FIGURE 10 – Exemple de cas de figure où un modèle sépare parfaitement les présences (Observation=1) et les absences (Observation=0) et où la corrélation bisériale ponctuelle n'est pas égale à 1.

TABLE 3 – Matrice de confusion permettant de juger de la qualité de la prédiction d'un modèle. Les nombres a, b, c, et d correspondent au nombre d'unités statistiques du jeu de données de validation pour chaque combinaison de valeurs prédites et de valeurs observées

		Observé	
		Présence	Absence
Prédit	Présence	a	b
	Absence	c	d

fronte donc les observations aux prédictions. La prédiction idéale serait la prédiction pour laquelle on observerait un accord parfait entre prédiction et observation, c'est à dire la prédiction pour laquelle les nombres $b = c = 0$ dans la matrice de confusion (aucune erreur de prédiction).

Toutefois, une prédiction parfaite est rarement possible, et on observe en général des erreurs de prédiction. Deux types d'erreurs sont en général distingués (Fielding et Bell, 1997) :

- ⋆ Les *faux positifs* : ce sont des unités pour lesquelles on prédit la présence de l'espèce alors qu'elle n'est pas présente ;
- ⋆ Les *faux négatifs* : ce sont des unités pour lesquelles on prédit l'absence de l'espèce alors qu'elle est présente.

Selon la problématique étudiée, ces deux types d'erreurs n'auront pas le même poids. L'une des questions que nous devrons résoudre sera de définir le risque acceptable pour chacun de ces deux types d'erreur. Pour ce, nous devons introduire deux concepts.

Nous définissons la **sensitivité du classifieur** e comme la proportion d'unités pour lesquelles l'espèce est prédite présente parmi les unités où l'espèce est effectivement présente (Fielding et Bell, 1997). Autrement dit :

$$e = \frac{a}{a + c}$$

Nous définissons la **spécificité du classifieur** p comme la proportion d'unités pour lesquelles l'espèce est prédite absente parmi les unités où l'espèce est effectivement absente (Fielding et Bell, 1997). Autrement dit :

$$p = \frac{b}{b + d}$$

Les valeurs de sensitivité et de spécificité dépendent fortement de la valeur du seuil utilisé pour construire la matrice de confusion. Ainsi, si l'on définissait un seuil $s = 0$ alors la présence de l'espèce serait prédite dans toutes les unités, quelle que soit leur composition environnementale. Dans ce cas, la sensitivité serait maximale ($e = 1$), et la spécificité minimale ($p = 0$). Inversement, si l'on posait $s = 1$ alors l'espèce serait prédite absente de toutes les unités, quelle que soit leur surface. La sensitivité serait alors minimale ($e = 0$) et la spécificité maximale ($p = 1$).

Ces éléments peuvent être utilisés pour construire une mesure de qualité globale de prédiction par le modèle (étape préliminaire indispensable à la définition du classifieur). En effet, nous pouvons faire varier le seuil s utilisé pour construire la matrice de confusion de façon continue entre 0 et 1. Pour chaque valeur de c, nous disposons alors de deux mesures (la sensitivité et la spécificité). Nous pouvons tracer une courbe mettant en relation la sensitivité (en ordonnée) en fonction de 1-spécificité (en abscisse). On appelle cette courbe la courbe ROC (*Receiver Operating Characteristic*, Hosmer et Lemeshow, 2000, p. 162).

La courbe ROC obtenue peut prendre différentes formes, dont nous donnons des exemples sur la figure 11 :

- ⋆ La courbe verte correspond au cas où le modèle discrimine parfaitement entre les présences et les absences. Dans ce cas, à l'exception des cas extrêmes $c = 0$ et $c = 1$, la sensitivité et la spécificité sont toujours maximales : le modèle ne se trompe jamais ;
- ⋆ La courbe orange correspond au cas exactement inverse : à l'exception des cas extrêmes $c = 0$ et $c = 1$, la sensitivité et la spécificité sont toujours minimales : le modèle se trompe toujours ;
- ⋆ La courbe bleue correspond au cas dans lequel la prédiction est totalement aléatoire. Alors la

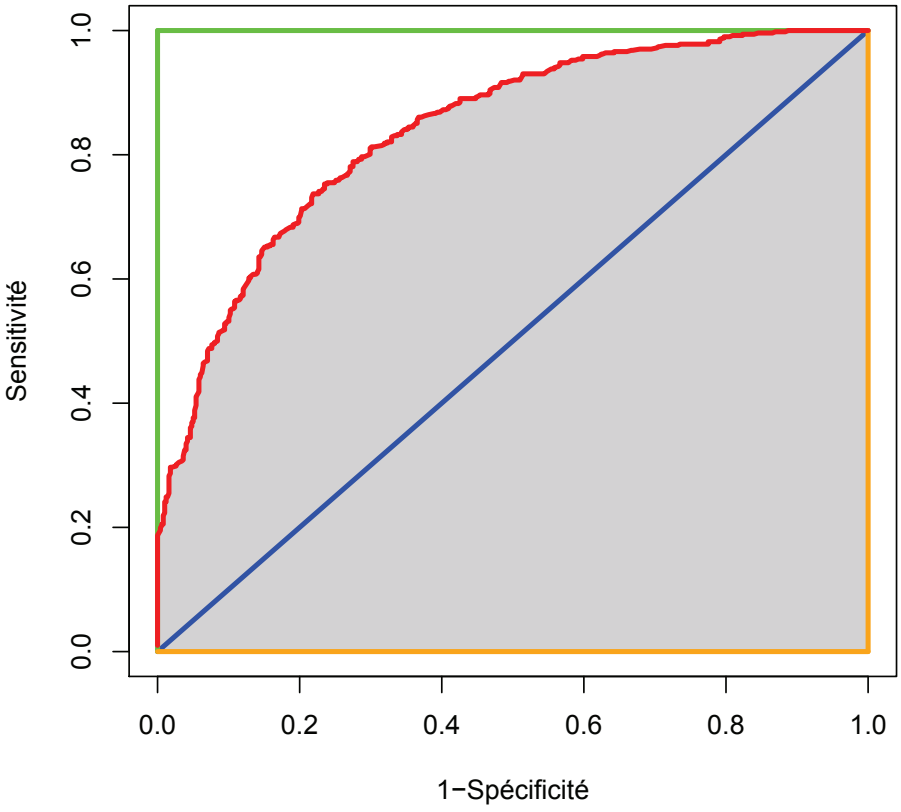

FIGURE 11 – Principe de construction de la courbe ROC (*Receiver Operating Characteristic*

sensitivité est égale à 1-Spécificité.

- ⋆ La courbe rouge correspond aux cas le plus fréquent. Cette courbe se situe en général quelque part entre la courbe du modèle totalement aléatoire et la courbe du modèle parfait.

Etant donnés les éléments donnés ci-dessus, il est possible de calculer un indice de la qualité de prédiction *globale* du modèle. En effet, la surface sous la courbe ROC (en gris sur la figure pour la courbe rouge), comprise entre 0 et 1, donne une mesure de la qualité de prédiction. Cette surface est appelée AUC (**Area Under the Curve**). Lorsque l'AUC est inférieur à 0.5, alors le modèle prédit plus mal que si l'on tirait les prédictions à pile ou face. Et plus l'AUC se rapproche de 1, meilleure est la prédiction.

Remarque : Hosmer et Lemeshow (2000, p. 163) et Fielding et Bell (1997) indiquent qu'il existe une autre interprétation possible à l'AUC. Imaginons que l'on tire au sort une unité A dans un ensemble d'unités où l'espèce est présente. Imaginons que l'on tire au sort une autre unité B dans un ensemble d'unités où l'espèce est absente. alors *l'AUC mesure la probabilité que la probabilité de présence de l'espèce prédite par le modèle pour l'unité A soit supérieure à la probabilité de présence de l'espèce prédite par le modèle pour l'unité B.*

Bien sûr dans notre cas de figure, nous ne disposons pas de données de type présence/absence. Cependant, Phillips *et al.* (2006) expliquent pourquoi cette approche reste valide dans le contexte de modèles ajustés sur des données de type présence/points de contexte : "*with presence-only data, the maximum achievable AUC is less than 1 (...). If the species' distribution covers a fraction a of the pixels, then the maximum achievable AUC can be shown to be exactly 1 - a/2. Unfortunately, we typically do not know the value of a, so we cannot say how close to optimal a given AUC value is. Nevertheless, we can still use standard methods to determine statistical significance of the AUC, and to distinguish between the predictive power of different classifiers. We note that random prediction still corresponds to an AUC of 0.5*". En effet, comme nous supposons que toute présence détectée sur le terrain correspond bien à une présence réelle, nous pouvons garantir que la sensitivité est estimée sans biais dans notre étude. En revanche, les points de contexte peuvent correspondre à des points de présence comme à des points d'absence. Ainsi, la spécificité estimée est nécessairement inférieure à la spécificité réelle d'un modèle, et donc même avec un modèle parfait, l'AUC sera inférieur à 1. Nous ne devrons donc pas interpréter les valeurs de l'AUC dans l'absolu. Nous ne pourrons nous en servir qu'à des fins de comparaison de différents modèles ajustés pour la même zone.

6.4 Etape de test du modèle : qualité de prédiction

Les critères présentés dans la section précédente permettent d'identifier, dans un pool de modèles, celui dont le pouvoir prédictif est le meilleur. En revanche, il est également utile de pouvoir juger de la qualité d'ajustement et de prédiction (*goodness of fit*) des modèles. Par exemple, ces critères peuvent nous indiquer qu'un modèle A est meilleur prédicteur qu'un modèle B, mais il est possible que le modèle A soit mauvais prédicteur, et le modèle B, très mauvais prédicteur. Dans ce cas, il serait absurde de recommander l'utilisation du modèle A. Peut-être qu'aucun des modèles comparés n'est suffisamment bon pour une utilisation pratique par des gestionnaires. Nous avons donc besoin de pouvoir juger de la qualité de l'ajustement dans l'absolu. Nous en arrivons donc à l'étape de test du modèle.

La difficulté de la mesure de la qualité de prédiction vient du type de données que nous sommes amenés à traiter. En effet, s'il existe beaucoup de critères développés pour les modèles basés sur des données de type présence/absence, il en existe beaucoup moins pour les données de type présence seule. Ainsi, l'AUC est fréquemment utilisé comme mesure de qualité de l'ajustement lorsqu'il est utilisé sur des données de présence/absence. Ainsi, (Swets, 1988) indique : "*Values of A [AUC] between 0.50 and 0.70 or so represent a rather low accuracy – the true-positive proportion is not much greater than the false-positive proportion anywhere along the curve. Values of A between about 0.70 and 0.90 represent accuracies that are useful for some purposes, and higher values represent a rather high accuracy*". Cependant, ces recommandations ne sont pas valables dans un contexte d'analyse basée sur des données de type "présence seule". En effet, nous avons vu dans la section 6.3.4 que la valeur maximale possible pour

ce critère dépendait de la fréquence de l'espèce sur la zone étudiée, ainsi que de la pression d'échantillonnage. Il n'est donc pas possible d'utiliser ce critère pour mesurer la qualité de l'ajustement, dans la mesure où la valeur de la borne supérieure de l'AUC (prédiction parfaite) est inconnue. L'AUC ne peut servir qu'à comparer des modèles entre eux ; pas à donner une mesure absolue de la qualité de la prédiction.

Par ailleurs, nous l'avons vu, le coefficient de corrélation bisériale ponctuelle ne peut pas non plus être utilisé comme mesure de la qualité de l'ajustement. En effet, ce coefficient peut prendre des valeurs très inférieures à 1, même lorsque la séparation des classes est excellents (Cox et Wermuth, 1992). Ainsi, comme pour l'AUC, nous ne connaissons pas la valeur minimale pour laquelle la séparation des classes (présence/absence) est parfaite. Et comme l'AUC, ce critère ne pourra être utilisé que pour comparer les modèles entre eux.

En réalité, l'indice de Boyce *et al.* (2002) est le seul des critères utilisés à pouvoir servir de mesure de qualité de l'ajustement. En effet, ce critère est compris entre -1 (prédiction inverse) et 1 (prédiction parfaite), et permet de juger, dans l'absolu, de la qualité de la prédiction. C'est ce critère qui nous servira de mesure de la qualité de l'ajustement. Nous examinerons aussi les semis de crottiers sur les cartes prédites par les méthodes utilisées, afin de pouvoir juger visuellement de cette qualité de prédiction.

7 Résultats

7.1 Analyse exploratoire préliminaire

7.1.1 La composition environnementale des zones d'étude

Dans un premier temps, nous avons examiné la composition environnementale sur les zones d'étude disponibles, afin de mieux comprendre les contextes dans lesquels les données ont été recueillies. Pour le moment, nous ne nous intéressons donc pas à la localisation des crottiers.

Nous avons effectué une analyse en composantes principales interclasses des 18 variables environnementales (Dolédec et Chessel, 1987). L'objectif de cette analyse est de trouver les directions, dans l'espace écologique, sur lesquelles la composition environnementale moyenne diffère le plus entre les zones d'étude. Elle permet donc d'identifier, du point de vue de la composition environnementale, les différences entre les zones d'étude.

Les résultats de cette analyse sont présentés figure 12. Cette analyse nous montre que Flaine et Collet d'Allevard sont, du point de vue de la composition environnementale, assez similaire (haute altitude, fortes pentes, riches en landes). Le site des Saisies est au contraire situé à une altitude plus faible, avec des pentes plus faibles, dans un environnement plus forestier. Le site de Giettaz est intermédiaire sur ce gradient.

7.1.2 Analyse K-select de la sélection de l'habitat par zone d'étude

Nous explorons à présent la sélection de l'habitat par le tétras-lyre, en recherchant en quoi l'utilisation de l'espace par cette espèce diffère de la composition moyenne des zones d'étude. Nous disposons de quatre zones d'étude : les zones de Flaine (184 crottiers détectés pour 1019 quadrats prospectés), Giettaz (396 crottiers détectés pour 775 quadrats prospectés), des Saisies (180 crottiers détectés pour 2000 quadrats prospectés) et du Collet d'Allevard (25 crottiers détectés pour 250 quadrats prospectés). Nous avons effectué une analyse K-select (Calenge *et al.*, 2005) afin d'identifier les directions dans lesquelles la sélection de l'habitat était la plus forte sur chacun des sites, et afin d'identifier les différences de sélection de l'habitat entre les sites. Cette analyse consiste à rechercher, dans l'espace écologique, les directions sur lesquelles la somme des carrés des différences sur chaque zone d'étude entre l'utilisation moyenne de l'environnement et sa disponibilité moyenne (on parle de *marginalité*) est la plus grande. Nous ne détaillons pas le principe de cette analyse (le lecteur intéressé pourra se reporter aux articles de Calenge

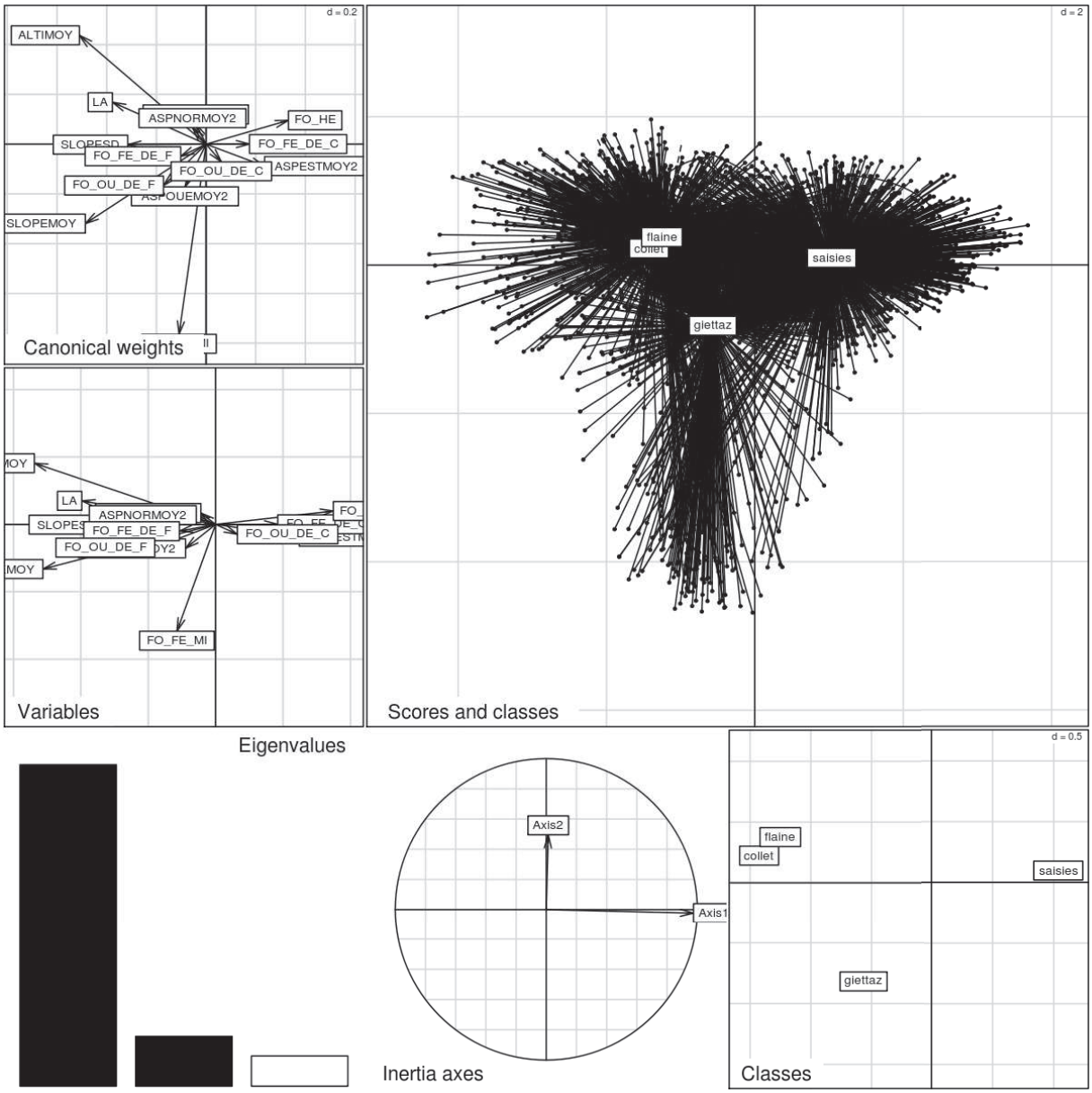

FIGURE 12 – Résultats de l'analyse inter-classes effectuée pour identifier les différences de composition environnementale entre les zones d'étude (les variables dérivées des orthophotographies ne sont pas prises en compte dans cette analyse). Le graphique en bas à gauche, intitulé "Eigenvalues", représente les valeurs propres de l'analyse. Nous focaliserons notre interprétation sur un seul axe, une nette cassure dans la décroissance étant observée après la première valeur propre. Les graphiques intitulés "Variables" et "Canonical weights" sont deux moyens d'interpréter la signification des axes. Dans les deux cas, nous identifions une opposition entre les zones présentant des fortes altitudes et des fortes pentes, ainsi qu'une abondance de landes importante aux zones aux caractéristiques contraires. Le graphique principal "Scores and classes" présente une "photographie" de l'espace écologique sur les deux premiers axes de l'analyse. Chaque point correspond à un quadrat de notre jeu de données, et les quadrats d'un même site sont reliés entre eux par les branches d'une étoile. Ce graphique (ainsi que le graphique intitulé "Classes", qui présente uniquement les coordonnées des zones d'étude sur les axes principaux) nous permet de placer chaque zone sur ce gradient altitudinal : alors que Flaine et Collet d'Allevard sont des zones situées à forte altitude (et de fortes pentes, et donc une présence abondante de landes), le site des Saisies est situé à une altitude relativement faible (et est donc plus riche en forêt de conifères, avec des pentes plus faibles), et le site de Giettaz présente des altitudes et des pentes intermédiaires. Le graphique "Inertia axes" présente la corrélation entre les axes de l'analyse inter-classes et les axes de l'analyse en composantes principales classique du jeu de données (sans contraintes). Il nous révèle que la structure exprimée par le premier axe de l'analyse séparant au mieux les sites d'étude est aussi la structure environnementale principale exprimée par les données.

et al., 2005; Calenge, 2007), et nous présentons directement les résultats de ces analyses. Notons que nous avons supprimé les variables dérivées des orthophotographies pour cette analyse. Cependant, une analyse K-select effectuée sur les zones d'étude situées dans le département de Haute Savoie, et incluant ces variables, renvoie les mêmes résultats.

Cette analyse montre que la sélection de l'habitat par le tétras-lyre varie en fonction de la zone considérée (figure 13). Dans les sites de Flaine et des Saisies, les crottiers de tétras-lyre sont le plus souvent identifiés dans les zones orientées au nord. Dans les sites de Giettaz et du Collet d'Allevard, les crottiers de tétras-lyre sont le plus souvent identifiés dans les zones riches en forêt de feuillus. Cependant, l'examen des scores des quadrats sur le premier plan de l'analyse K-select indique que l'évitement des zones orientées au sud est une caractéristique commune à tous les sites (figure 14). L'exposition des versants jouera probablement un rôle essentiel dans la prédiction de la localisation des zones d'hivernage.

7.2 Modélisation prédictive pour le département de Haute Savoie

Dans cette section, nous ajustons des modèles prédictifs à partir des données collectées sur les trois sites haut-savoyards. Deux types de modèles sont ajustés : (i) des modèles basés sur toutes les variables disponibles, (ii) des modèles basés sur toutes les variables en excluant les variables dérivées des orthophotographies (niveau moyen de vert et écart-type du niveau de vert). Pour chaque type de modèle, nous avons utilisé les cinq méthodes (distances de Mahalanobis, MADIFA, régression logistique complète, régression logistique pas à pas, et forêt d'arbres décisionnels), et nous avons utilisé l'approche de validation croisée décrite dans la section 6.2 pour mesurer le pouvoir prédictif des modèles incluant ou excluant les variables dérivées des orthophotographies, afin de déterminer si l'on pouvait exclure ces variables d'une modélisation plus générale. Notons que les variables prédictrices ont toutes subi une transformation racine carrée, afin de stabiliser leur distribution (qui était assez asymétrique sans cette transformation).

Nous présentons dans les tableaux 4, 5 et 6 respectivement les valeurs d'AUC, de critère de Boyce *et al.* (2002) et de corrélation bisériale ponctuelle calculées pour chaque type de modèle, avec ou sans les variables dérivées des orthophotographies. Nous constatons tout d'abord qu'il y a accord entre les différents critères concernant la mesure du pouvoir prédictif des modèles. Ainsi, quel que soit le critère considéré, il semble que la régression logistique soit la plus efficace, suivie par les forêts aléatoires, la MADIFA et enfin les distances de Mahalanobis. Notons que l'algorithme de sélection de variable pas à pas ne semble pas modifier beaucoup le modèle de régression logistique (un examen plus détaillé montre des différences très peu marquées entre les prédictions des deux modèles).

Nous constatons en outre que l'ajout des variables dérivées des orthophotographies n'apporte pas grand chose à la prédiction, et ce, quel que soit le critère considéré. La suppression de ces variables permet d'améliorer le pouvoir prédictif de certains modèles (e.g. distances de Mahalanobis, MADIFA d'après le critère de Boyce *et al.* (2002)), et diminue très légèrement le pouvoir de prédiction de la forêt aléatoire et des régressions logistiques. Nous pourrons donc nous passer de ces variables pour permettre la construction d'un modèle plus général valable pour l'ensemble des Alpes du nord. Cela nous permettra en outre, d'inclure un plus grand nombre de zones d'études (sud du site des Saisies et site du Collet d'Allevard), ce qui nous permettra de calibrer le modèle en prenant en compte une plus grande diversité de conditions environnementales.

Une évaluation visuelle de la qualité des prédictions est présentée sur la figure 15. Nous pouvons constater qu'en effet, les régressions logistiques semblent ici mieux prédire les zones dans lesquelles le tétras place ses sites d'hivernage. Les distances de Mahalanobis et la MADIFA semblent trop optimistes : bien que dans la plupart des cas, les zones dans lesquelles on trouve des crottiers sont effectivement prédites comme favorables, beaucoup de zones favorables ne contiennent aucun crottier détecté. La forêt aléatoire semble avoir un pouvoir prédictif intermédiaire entre ces deux approches.

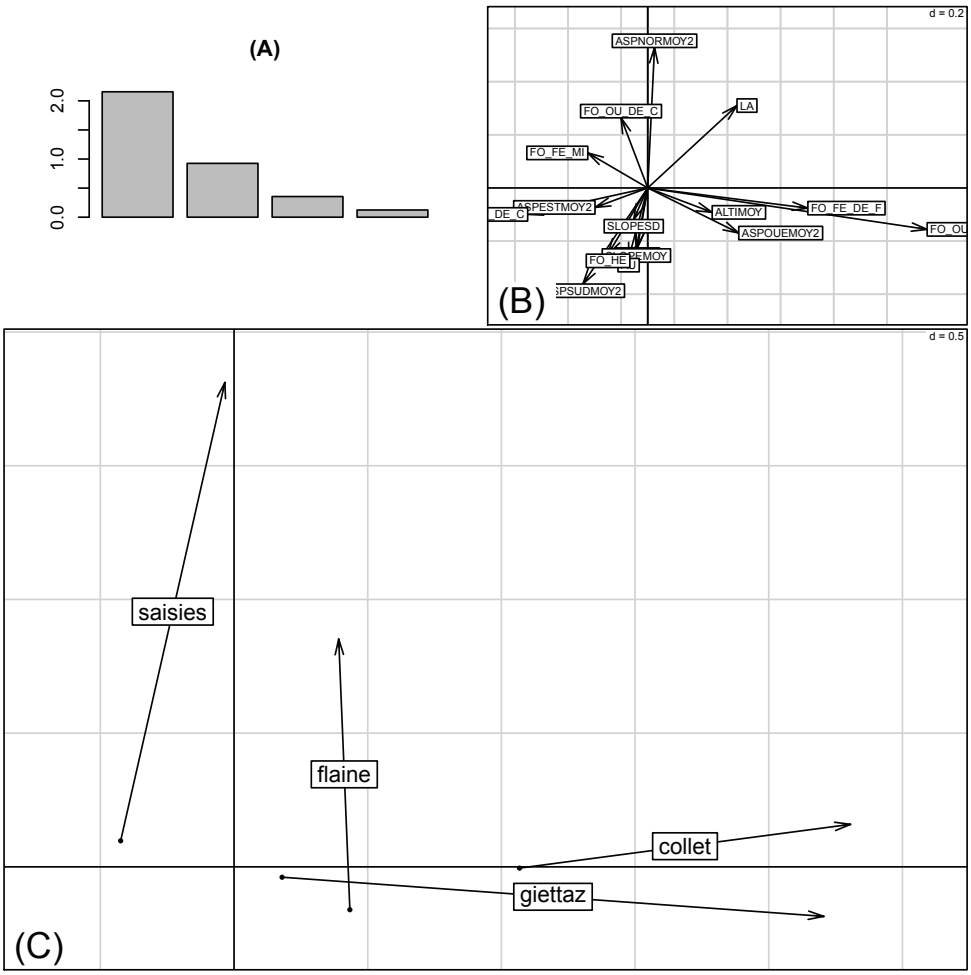

FIGURE 13 – Résultats de l'analyse K-select effectuée pour identifier les similarités et différences de la sélection de l'habitat entre les zones d'étude : (A) valeurs propres associées à cette analyse (i.e., quantité de marginalité moyenne expliquée par chaque axe). Nous concentrons notre analyse sur deux axes ; (B) scores des variables environnementales sur les deux premiers axes de l'analyse. Le premier axe oppose, du côté des valeurs positives, les zones riches en forêts de feuillus (ouvertes ou fermées) aux autres zones (valeurs négatives). Le deuxième axe oppose les zones orientées nord (valeurs positives) aux autres zones (valeurs négatives) ; (C) projection du vecteur de marginalité associé à chaque zone sur le plan formé par les deux premiers axes de l'analyse K-select. La base de chaque vecteur représente les conditions environnementales disponibles en moyenne sur la zone, et l'extrémité représente les conditions environnementales utilisées en moyenne par le tétras-lyre pour l'établissement des sites d'hivernage. Il apparaît ainsi clairement que le tétras recherche les zones riches en forêts de feuillus dans les sites de Giettaz et du Collet d'Allevard, et les zones orientées au nord à Flaine et aux Saisies.

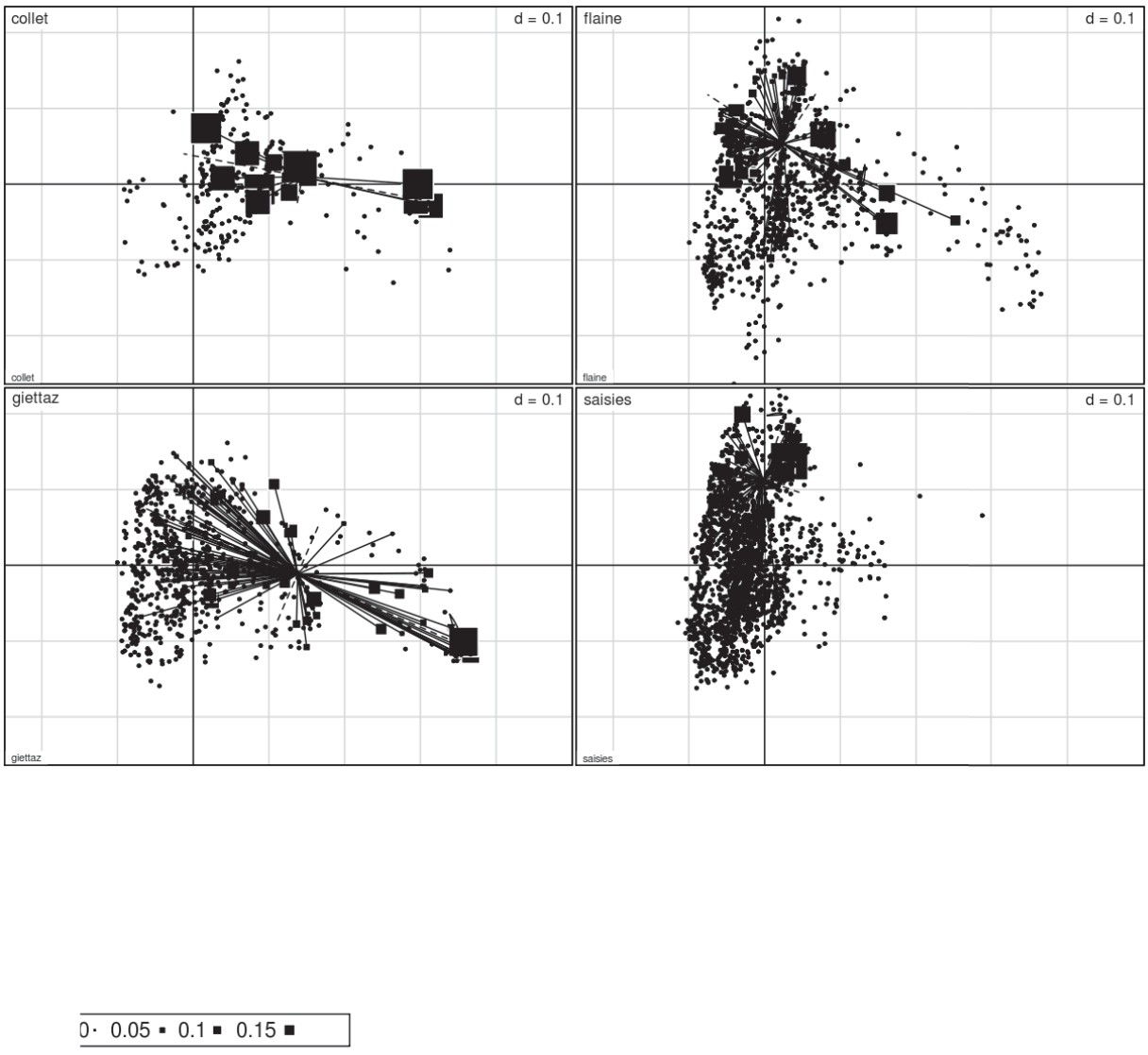

FIGURE 14 – Représentation de la distribution des quadrats sur les deux premiers axes de l'analyse K-select effectuée pour identifier les similarités et différences de la sélection de l'habitat entre les zones d'étude. Pour chaque site, sont représentés : les quadrats (points), et l'importance de leur utilisation par le tétras-lyre (carrés noirs dont la surface est proportionnelle au nombre de crottiers détectés). Notez que les points localisés dans la partie en bas à droite de ce graphique ne sont presque jamais utilisés, et ce, quelle que soit la zone d'étude (ces points correspondent aux orientations sud, cf. figure 13). Ainsi, les versants orientés au sud sont rarement utilisés par le tétras-lyre pour l'établissement de sites d'hivernage.

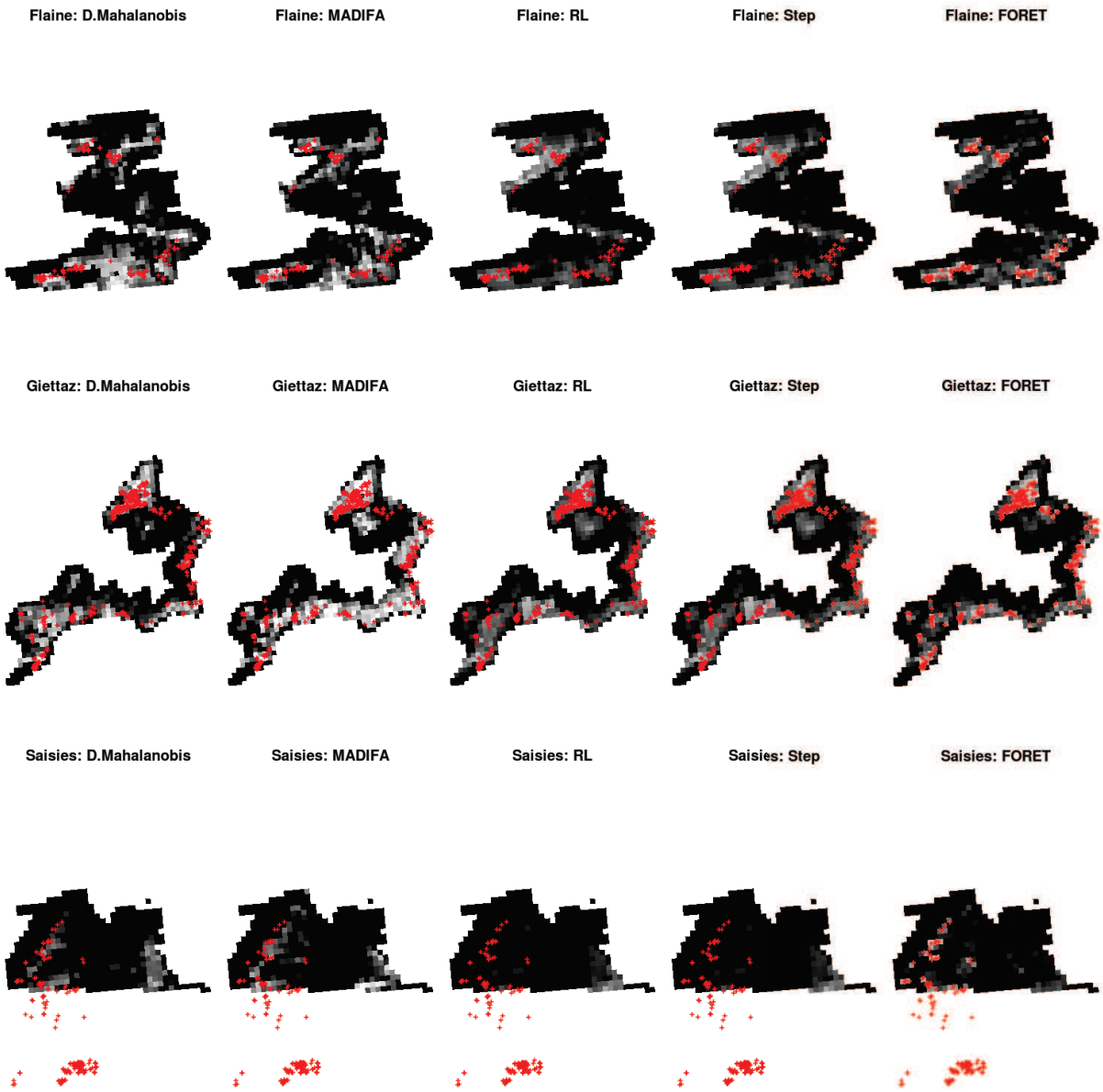

FIGURE 15 – Cartes de conformité à l'habitat d'hivernage du tétras-lyre obtenues pour les sites d'étude du département de haute savoie à l'aide des distances de Mahalanobis, de la MADIFA, de la régression logistique complète (RL) et pas à pas (Step) et d'une forêt d'arbres décisionnels. Les points rouges indiquent la localisation des sites d'hivernage identifiés. Notons que seule la partie du site des Saisies située en Haute Savoie est prédite ici.

TABLE 4 – Valeurs d'AUC calculées après avoir poolé les prédictions sur chacune des trois zones de Haute Savoie. Pour une zone donnée, les prédictions ont été calculées à l'aide d'un modèle calibré sur les deux autres zones. Ces valeurs sont présentées pour chacun des modèles, avec et sans les variables dérivées des orthophotographies.

	Mahalanobis	MADIFA	Régression complete	pas à pas	Random Forest
Avec Ortho	0.51	0.54	0.77	0.77	0.67
Sans Ortho	0.53	0.54	0.73	0.73	0.65

TABLE 5 – Valeurs de l'indice de BOYCE *et al.* (2002) calculées après avoir poolé les prédictions sur chacune des trois zones de Haute Savoie. Pour une zone donnée, les prédictions ont été calculées à l'aide d'un modèle calibré sur les deux autres zones. Ces valeurs sont présentées pour chacun des modèles, avec et sans les variables dérivées des orthophotographies.

	Mahalanobis	MADIFA	Régression complète	pas à pas	Random Forest
Avec Ortho	0.38	0.46	0.95	0.95	0.91
Sans Ortho	0.62	0.66	0.95	0.96	0.80

7.3 Analyse sur toutes les Alpes du Nord

7.4 Calibration et validation interne

Nous avons ensuite calibré les modèles prédictifs à l'aide des données collectées sur les quatre zones d'étude, en nous concentrant sur les variables de relief et de végétation. Les valeurs d'AUC, de critère de BOYCE *et al.* (2002), et de coefficient de corrélation bisériale ponctuelle pour les différents modèles ajustés sont présentés dans le tableau 7.

Nous pouvons constater que tous les critères semblent en accord : la MADIFA et les régressions logistiques donnent les meilleurs résultats (avec un très léger avantage pour la MADIFA), suivi par la forêt aléatoire et par les distances de Mahalanobis. Notons également que le critère de BOYCE *et al.* (2002) est, pour les trois meilleures méthodes, assez important (très proche de 1), ce qui indique une qualité de prédiction exceptionnelle. Notons enfin que l'utilisation de l'algorithme de sélection de variables pas à pas ne semble pas améliorer les capacités prédictives de la régression logistique (il y a de très légères différences : la régression pas à pas est caractérisée par un AUC très légèrement plus grand que la régression complète, mais la différence d'AUC se situe au troisième chiffre après la virgule. Le même phénomène est observé pour les deux autres critères).

Il peut sembler surprenant que la MADIFA renvoie de si bons résultats dans cette modélisation globale, alors que ses performances étaient relativement faibles lorsque l'on ne se concentrait que sur le département de Haute-Savoie. En réalité, nous l'avons vu dans la section 7.1.2, il y a une certaine variabilité de la sélection de l'habitat entre les zones, et l'inclusion d'un plus grand nombre de zones d'étude pour la calibration du modèle global a permis la prise en compte d'une plus grande diversité de conditions environnementales. C'est cette prise en compte qui a permis de disposer d'une vision plus complète de la niche d'hivernage, en élargissant le champs des choix possibles du tétras, et donc d'extraire, grâce à la MADIFA, les principaux facteurs limitants gouvernant l'établissement des places d'hivernage quelle que soit la zone considérée. On voit là qu'il était préférable de perdre les variables dérivées des orthophotographies pour gagner davantage de zones d'étude dans le jeu de données de calibration.

La figure 16 montre que la qualité d'ajustement de la MADIFA et des régressions logistiques est assez bonne sur la plupart des zones d'étude : on ne trouve les crottiers que dans les zones les plus claires sur ces cartes, et rarement dans les zones les plus sombres. Notons toutefois que l'ajustement semble plus mauvais au Collet d'Allevard. L'indice de BOYCE *et al.* (2002) prend une valeur très proche de 1 pour ces deux méthodes, soulignant la très bonne qualité des prédictions.

TABLE 6 – Valeurs du coefficient de corrélation bisériale ponctuelle, calculées après avoir poolé les prédictions sur chacune des trois zones de Haute Savoie. Pour une zone donnée, les prédictions ont été calculées à l'aide d'un modèle calibré sur les deux autres zones. Ces valeurs sont présentées pour chacun des modèles, avec et sans les variables dérivées des orthophotographies.

	Mahalanobis	MADIFA	Régression complète	pas à pas	Random Forest
Avec Ortho	0.02	0.06	0.38	0.38	0.25
Sans Ortho	0.04	0.06	0.33	0.33	0.21

TABLE 7 – Valeurs d'AUC, de l'indice de BOYCE *et al.* (2002) et du coefficient de corrélation bisériale ponctuelle calculées après avoir poolé les prédictions sur chacune des quatre zones des Alpes du nord – pour une zone donnée, les prédictions ont été calculées à l'aide d'un modèle calibré sur les deux autres zones.

	Mahalanobis	MADIFA	Régression complète	pas à pas	Random Forest
AUC	0.66	0.75	0.72	0.72	0.68
Indice de Boyce et al.	0.87	0.99	0.97	0.97	0.96
Corrélation	0.21	0.32	0.28	0.28	0.23

Ainsi, pour le moment, les résultats obtenus nous conduiraient à sélectionner soit la régression logistique (complète ou pas à pas : les deux approches renvoient des prédictions très corrélées), soit la MADIFA comme modèle prédictif de la conformité à l'habitat (avec une très légère préférence pour cette dernière, étant données ses performances très légèrement meilleures). **A ce stade, nous pouvons donc éliminer les distances de Mahalanobis et la forêt aléatoire des modèles possibles**.

7.5 Validation externe des modèles

Nous avons vu dans les sections précédentes que les deux meilleurs modèles, si l'on se base sur l'étape de validation interne (i.e., la validation croisée), sont la MADIFA et la régression logistique. Ces deux méthodes semblent présenter une capacité de prédiction similaire, ce qui rend difficile le choix de l'une ou l'autre de ces méthodes. Nous appuierons ce choix sur les résultats de l'étape de validation externe, i.e. en s'appuyant sur les observations occasionnelles collectées dans la réserve naturelle de Villaroger (715 observations) et dans le parc naturel du Vercors (287 observations).

Le tableau 8 présente les valeurs d'AUC, du critère de BOYCE *et al.* (2002), et du coefficient de corrélation bisériale ponctuelle pour les cinq méthodes et pour chacun de ces deux sites. Tous ces critères donnent le même diagnostic : les régressions logistiques et la MADIFA semblent prédire avec la même efficacité dans le parc naturel du Vercors. En revanche, la MADIFA est beaucoup plus efficace dans la réserve naturelle de Villaroger (les régressions logistiques semblent donner des prédictions totalement absurdes dans cette réserve). Notons également que la qualité de la prédiction par la MADIFA, mesurée par le critère de BOYCE *et al.* (2002) est extrêmement forte pour ces deux zones (> 0.90).

Il semble donc que la MADIFA soit le modèle le plus approprié. C'est donc le modèle que nous recommanderons pour la prédiction des habitats d'hivernage du tétras-lyre. En effet, bien que la capacité de prédiction des régressions logistiques soit aussi bonne que celle de la MADIFA sur la plupart des terrains d'études pour lesquels nous disposons de données, la MADIFA prédit mieux sur la zone de Villaroger. Une évaluation visuelle de l'efficacité des deux méthodes est présentée figure 17 pour le Vercors et figure 18 pour Villaroger.

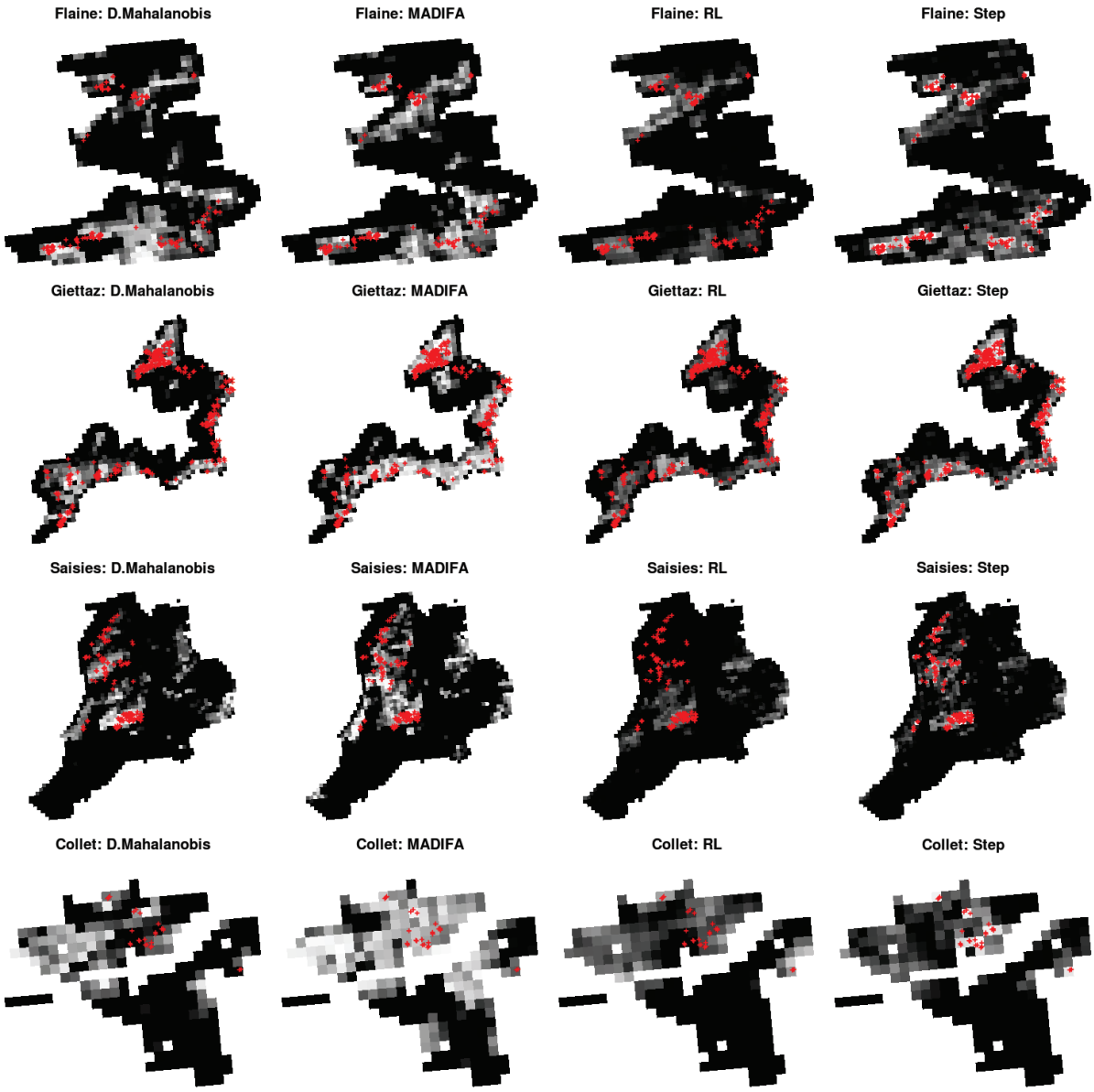

FIGURE 16 – Cartes de conformité à l'habitat d'hivernage du tétras-lyre obtenues pour tous les sites d'étude à l'aide des distances de Mahalanobis, de la MADIFA, de la régression logistique complète (RL) et pas à pas (Step), et d'une forêt d'arbres décisionnels. Les points rouges indiquent la localisation des sites d'hivernage identifiés.

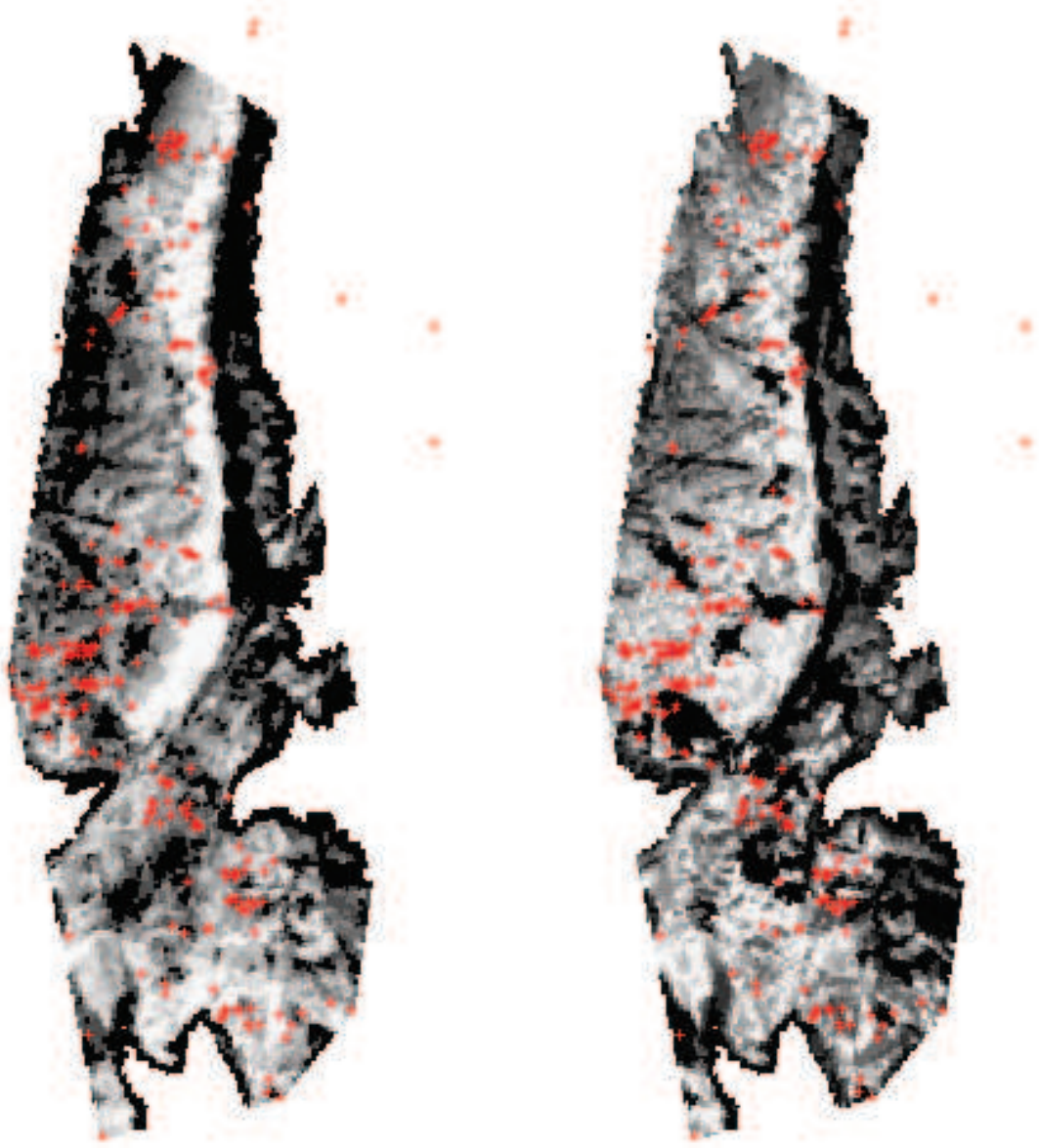

FIGURE 17 – Cartographie de la conformité à l'habitat dans le parc naturel du Vercors prédite par la régression logistique complète (à gauche) et par la MADIFA (à droite). Pour plus de clarté, les valeurs de conformité ont été obtenues en calculant le rang des prédictions sur la zone (le rang correspond au numéro d'ordre d'une prédiction lorsque toutes les prédictions d'une zone sont ordonnées par ordre croissant). Les points rouges correspondent aux localisations des observations occasionnelles de crottiers.

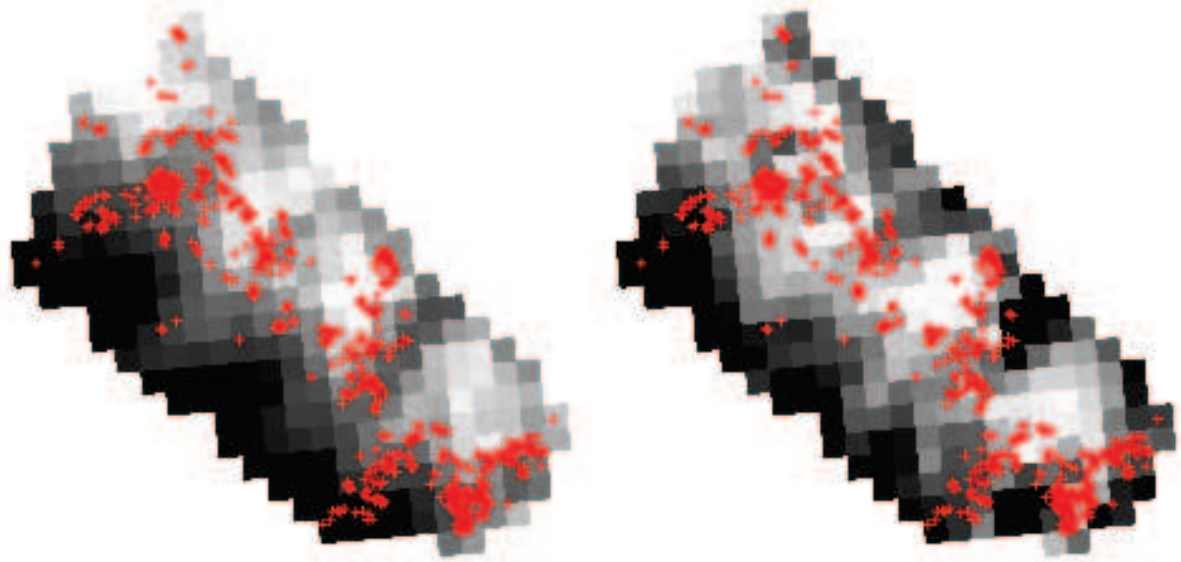

FIGURE 18 – Cartographie de la conformité à l'habitat dans la réserve naturelle de Villaroger prédite par la régression logistique complète (à gauche) et par la MADIFA (à droite). Pour plus de clarté, les valeurs de conformité ont été obtenues en calculant le rang des prédictions sur la zone (le rang correspond au numéro d'ordre d'une prédiction lorsque toutes les prédictions d'une zone sont ordonnées par ordre croissant). Les points rouges correspondent aux localisations des observations occasionnelles de crottiers.

Table 8 – Valeurs d'AUC, du critère de Boyce *et al.* (2002) et du coefficient de corrélation bisériale ponctuelle, calculées pour les deux sites de validation externe (Réserve naturelle de Villaroger et Parc naturel du Vercors), pour chacune des cinq méthodes.

	Mahalanobis	MADIFA	Régression complète	pas à pas	Random Forest
Villaroger : AUC	0.55	0.63	0.56	0.56	0.60
Villaroger : Boyce	0.16	0.96	-0.10	-0.09	0.84
Villaroger : corrélation	0.09	0.20	0.09	0.09	0.15
Vercors : AUC	0.63	0.65	0.63	0.62	0.63
Vercors : Boyce	0.90	0.94	0.96	0.95	0.95
Vercors : corrélation	0.06	0.07	0.06	0.06	0.06

8 Habitat ou conformité ?

8.1 Définir un seuil

Ainsi, les résultats précédents nous conduiraient à sélectionner la MADIFA comme méthode de prédiction de la conformité à l'habitat. Cette méthode renvoie donc un indice reflétant la probabilité de présence de crottiers en un point. Nous nous posons à présent la question : est-il possible de transformer cette prédiction de la conformité à l'habitat en prédiction de l'habitat ?

En d'autres termes, comment transformer une variable continue reflétant la probabilité de présence de crottier en variable binaire indiquant si une zone est un habitat ou non ? cette question fait appel aux éléments données dans la section 6.3.4. En effet, pour ce faire, il est nécessaire de définir un seuil s : si la conformité prédite à un point est inférieure à ce seuil, nous prédirons que ce point n'est pas un habitat. Si la conformité prédite est supérieure à s, nous prédirons que la zone est un habitat. Bien sûr, la principale critique que l'on pourrait faire à une telle approche est que l'on perd toute l'information sur l'incertitude de la prédiction : par exemple, si l'on fixe un seuil à $s = 0.5$, une zone dans laquelle la conformité prédite est de 0.51 sera classée habitat au même titre qu'une zone dans laquelle la conformité prédite est de 1. Pourtant cette dernière aura beaucoup plus de chances d'être un habitat que cette première. Mais un tel découpage a aussi ses avantages, car, une fois défini, il facilite le travail du gestionnaire, qui peut identifier aisément les habitats.

8.2 Deux risques

Nous déplaçons le problème : pour transformer une prédiction de la conformité et prédiction de l'habitat, il est nécessaire de définir un seuil. Le choix de ce seuil doit se faire en ayant connaissance des deux types de risque que nous avons déjà introduit dans la section 6.3.4. D'une part, un premier risque est d'exclure de la catégorie des habitats prédits des zones qui sont en réalité des habitats (modèle trop restrictif). D'autre part, un second risque est d'inclure dans cette catégorie des zones qui ne sont pas des habitats. Ainsi, un cas extrême consisterait à définir comme habitat toute zone quelle que soit sa composition environnementale. Nous serions alors sûr d'inclure tous les habitats, mais une telle carte serait inutile car elle inclurait également tous les non-habitats. Un autre cas extrême consisterait à exclure toutes les zones de cette catégorie, ce qui conduirait à une autre carte inutile.

Il est donc nécessaire de choisir ce seuil de façon à atteindre le meilleur compromis entre ces deux types de risque. Le choix de ce compromis n'est pas purement scientifique : il dépend des deux types de risques que le lecteur de la carte est prêt à prendre. Les conséquences de cette décision sont surtout politiques, et c'est sur cette base que le choix du seuil doit se faire. L'aspect scientifique de ce choix ne tient que dans l'information qui est donnée au décisionnaire concernant les deux risques pris. C'est à dire, pour un seuil donné, nous devrions être capable de donner au décisionnaire les deux risques associés : probabilité qu'une localisation exclue soit un habitat, et probabilité qu'une localisation inclue n'en soit pas un.

Mais c'est là que réside la difficulté dans notre étude : il est impossible de chiffrer le second type de risque, dans la mesure où nous sommes incapables de déterminer quelles zones ne sont pas des habitats, même sur les zones où des données ont été collectées (cf. section 4.2). Nous ne pourrons donc pas contrôler le second type de risque (pour la même raison que nous ne pouvons pas utiliser l'AUC comme mesure de qualité d'ajustement, cf. section 6.4).

En revanche, nous sommes capables d'évaluer le premier de ces risques, en examinant sur les données ayant servi à calibrer le modèle la proportion de crottiers pour lesquels la valeur de conformité prédite par le modèle est inférieure à s (une approche similaire a été utilisée par Hirzel *et al.*, 2002, pour calculer ce type de risque).

Nous pourrons nous faire une vague idée du compromis atteint pour un seuil donné s en calculant la proportion des points crottiers pour lesquels la conformité prédite est inférieure à s (qui mesure le premier de ces risques), et en calculant la proportion des points de contexte pour lesquels la conformité prédite est supérieure à s (qui reflétera très indirectement et imparfaitement le second). Nous atteindrons donc un bon compromis lorsque ces deux valeurs seront relativement faibles.

8.3 Identification du seuil

Nous nous proposons de fixer un seuil s tel que les habitats prédits contiendront 90% des crottiers du jeu de données de calibration. En ce qui concerne les prédictions de la MADIFA, ce seuil est fixé à $s = 0.08$. La figure 19 présentent la distribution lissée des valeurs de conformité prédites par la MADIFA pour les crottiers et les points de contexte. Il apparaît qu'en fixant $s = 0.08$, 90% des crottiers tombent dans les habitats (ce qui est normal, ce seuil a été choisi ainsi). En outre, les habitats d'hivernage prédits représentent 38% de la surface totale des zones de calibration. Ainsi, nous identifions qu'à peu près un tiers des zones étudiées contiennent 90% des crottiers.

8.4 Prédiction des habitats

Dans la section précédente, nous avons fixé un seuil s délimitant les zones considérées comme habitat et les zones considérées comme non-habitat, de telle façon que 90% des crottiers détectés dans les zones de calibration du modèle soient localisés dans les habitats prédits. Nous pouvons nous interroger sur la capacité du modèle à prédire les habitats dans des zones non-utilisées pour la calibration. Nous présentons dans le tableau 9 la proportion des observations occasionnelles de crottiers situés dans des zones prédites comme habitat par la MADIFA. A titre d'information, nous donnons également cette proportion pour la régression logistique complète (les habitats étant définis pour cette méthode de la même façon que dans la section précédente, i.e. en définissant un seuil au delà duquel 90% des crottiers du jeu de données de calibration sont détectés). Nous pouvons constater que la proportion de crottiers détectés dans les zones prédites comme habitat dans le Vercors et à Villaroger est bien inférieure à 90%. En effet, un tiers des crottiers détectés à Villaroger tombe dans des zones prédites par la MADIFA comme non-habitat, et cette proportion est encore plus faible pour le massif du Vercors.

Table 9 – Proportion des crottiers identifiés dans le Vercors et à Villaroger localisés dans les habitats prédits par la MADIFA et par la régression logistique complète.

	MADIFA	Régression
Villaroger	0.68	0.10
Vercors	0.58	0.31

Deux explications sont possibles pour expliquer ces faibles résultats. Tout d'abord, nous devons rappeler que nous travaillons ici sur des observations occasionnelles : nous ne pouvons pas garantir que la

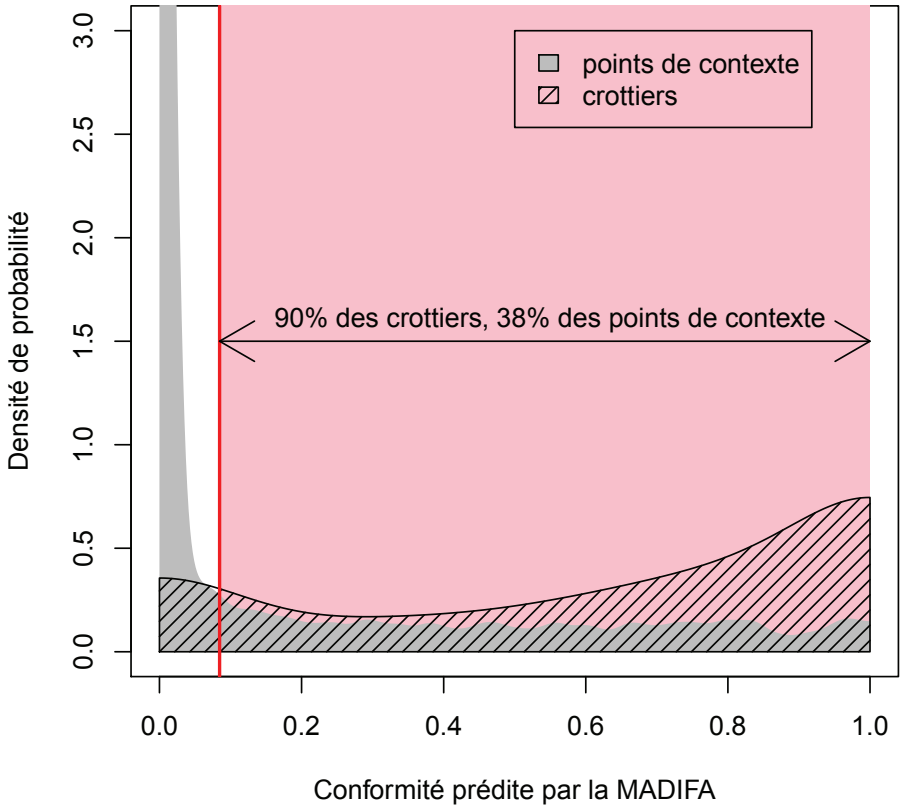

FIGURE 19 – Identification du seuil s délimitant les habitats et les non-habitats. Le seuil s est choisi de telle façon que 90% des points crottiers présente une conformité à l'habitat prédite par la MADIFA supérieure à s. Chacun des crottiers du jeu de données de calibration est caractérisé par une valeur pour chacune des variables environnementales. Donc pour chaque crottier, on peut calculer une valeur de conformité de l'habitat. Donc, l'ensemble des crottiers définit une distribution de valeurs de conformité. Cette distribution est présentée en hachuré sur la figure ci-dessus (il s'agit de la distribution lissée à l'aide de la méthode du noyau, WAND et JONES, 1995). Nous recherchons donc la valeur de conformité s pour laquelle 10% de la distribution des crottiers présente une conformité inférieure. Dans le cas présent, le seuil est choisi égal à $s = 0.08$. Nous présentons également la distribution des valeurs de conformité calculée pour les points de contexte (en gris). Ainsi, on voit que cette distribution présente un très fort pic (qui dépasse les limites du graphique) pour les très faibles valeurs de conformité. Ainsi, seulement 38% des points de contexte présentent une conformité prédite supérieure à s.

distribution spatiale de l'effort d'échantillonnage était uniforme. Il est donc possible que les zones prédites comme habitat aient été moins prospectées que les non-habitats [16]. Dans ce cas, la proportion réelle de crottiers dans les quadrats classés comme habitat est probablement supérieure à celle que nous avons calculée dans le tableau 9.

Mais il y a une autre explication possible, qui est une limite commune à toutes les méthodes de prédiction statistique. Nous avons ajusté un modèle en nous appuyant sur des données collectées essentiellement en Haute Savoie (une faible proportion des données est collectée en Savoie et en Isère), dans un contexte environnemental donné. Or, nous nous en servons pour prédire l'habitat en Savoie et dans la Drôme, dans des contextes très différents. Nous pourrions donc nous attendre à ce que la validité des modèles construits soit remise en question pour ces nouvelles zones.

Pourtant, **la validité du modèle *prédisant la conformité à l'habitat* n'est pas remise en question dans ces nouvelles zones** : il est même assez surprenant de voir à quel point la corrélation entre la conformité prédite par la MADIFA et le nombre de crottiers détectés dans ces nouvelles zones est importante (critères de Boyce *et al.* (2002) tous supérieurs à 0.9 dans le tableau 8), ce qui illustre le bon pouvoir prédictif de notre modèle pour *la conformité*, même sur ces nouvelles zones : plus la conformité prédite d'une zone est grande, et plus le nombre de crottiers qui y est trouvé est important.

Mais ce n'est pas le modèle prédisant la conformité qui est remis en question ; ce sont les zones classées comme habitat et non-habitat après application d'un seuil s de conformité. Ceci peut s'expliquer : la composition environnementale sur ces nouvelles zones est différente de celle des zones de calibration. Dans ces nouvelles zones, le tétras-lyre est peut-être moins exigeant concernant les zones acceptables (e.g., à cause d'un climat moins rude, ou d'une pression de dérangement par les skieurs plus faibles).

Nous avons défini comme seuil de conformité la valeur de prédiction au delà de laquelle on trouvait 90% des crottiers *dans les zones de calibration*. Mais si l'on considère une population de tétras-lyre localisée dans une zone au climat hivernal moins rude que les zones de calibration, ou dans une zone dans laquelle la pression de dérangement par les skieurs est plus faible, il est probable que le tétras-lyre se montrera moins exigeant quant au choix de ses sites d'hivernage, et donc que l'étendue des valeurs de conformité acceptables sera *"décalée vers le bas"*, c'est à dire vers des valeurs de conformité plus faibles. En d'autres termes, **il n'est peut-être pas légitime d'utiliser le même seuil s de conformité pour distinguer les habitats et les non-habitats dans toutes les zones Alpines**.

Autrement dit, si nous pouvons valider l'utilisation de l'indice de conformité prédit par la MADIFA, il est plus délicat de trouver un seuil unique valable partout dans toutes les Alpes du nord, permettant de distinguer les habitats et les non-habitats.

8.5 Construction de la carte

Nous avons donc choisi de construire deux cartes des Alpes du nord :

- ⋆ Une carte présentant la conformité "brute" à l'habitat sur les Alpes du nord, telle que prédite par la MADIFA ;
- ⋆ Une carte présentant les habitats définis pour différents seuils s (et non pas seulement le seuil tel que les zones pour lesquelles la prédiction est supérieure à s contiennent 90% des crottiers du jeu de données de calibration). Cette variation du seuil distinguant les habitats et les non-habitats permettra au lecteur de la carte de choisir le seuil adéquat selon la zone étudiée ;

16. Rappelons que nous n'avons aucune information sur la distribution de cet effort d'échantillonnage

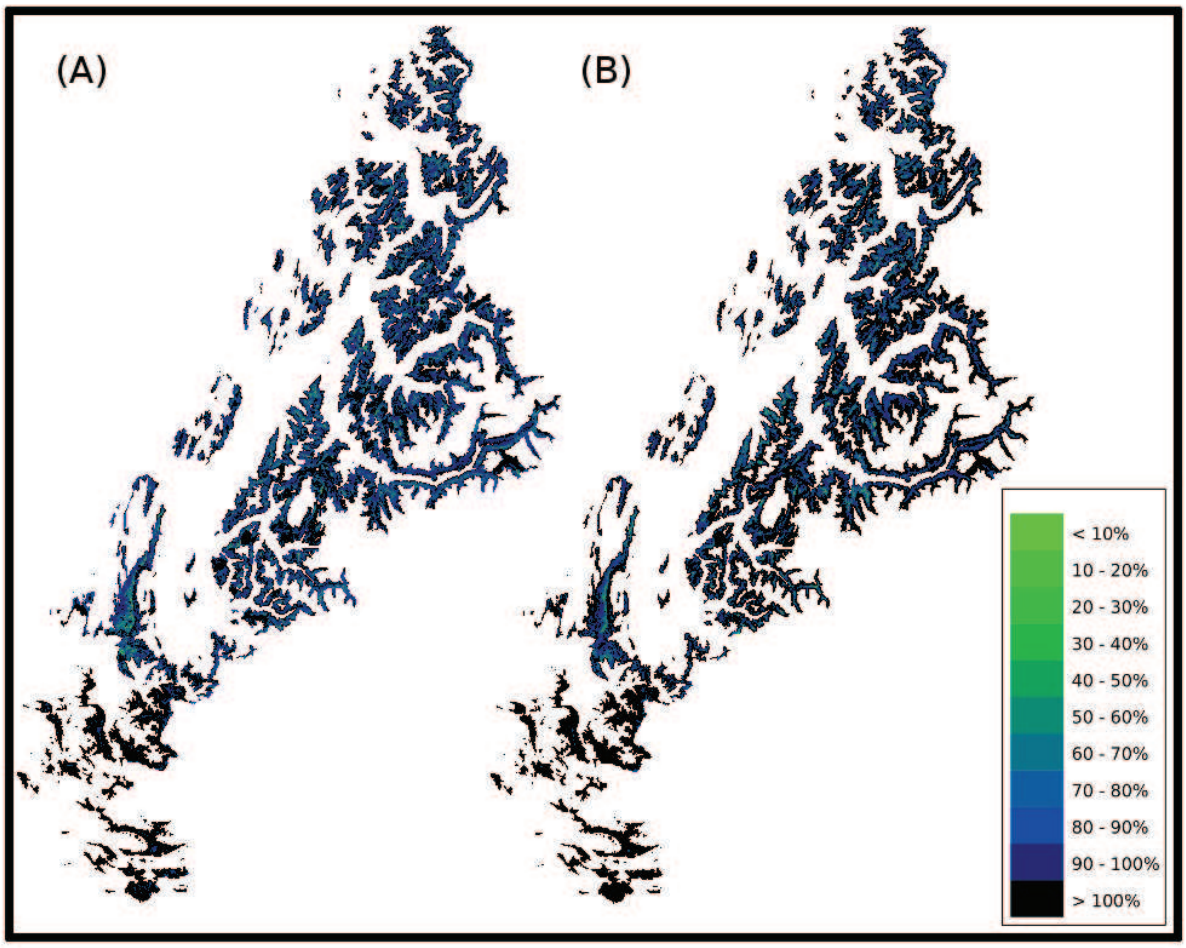

FIGURE 20 – Cartes de conformité à l'habitat d'hivernage du tétras-lyre obtenues à l'aide de : (A) la MADIFA, (B) une régression logistique complète (donnée pour information). Les couleurs représentent différents seuils possibles pour la définition des types d'habitat. Un seuil s est défini de telle façon que l'ensemble des points du jeu de données de calibration pour lesquels la prédiction du modèle est supérieure à s contient $X\%$ des crottiers, avec X variant de 10 à 100% (cf. texte).

La figure 20 présente cette dernière carte sur les Alpes du nord. La figure 21 présente un zoom de cette carte sur le parc du Vercors, ainsi que les observations occasionnelles de crottiers dans cette zone.

Le fichier de formes (*shapefile*) contenant ces deux cartes prédites est disponible auprès de l'auteur de ce rapport sur simple demande.

9 Discussion

9.1 Synthèse

Dans ce rapport, nous avons construit une carte de prédiction de la conformité des Alpes du nord à l'habitat d'hivernage du tétras-lyre. Nous avons tout d'abord évalué s'il était nécessaire d'inclure, pour cette modélisation, les variables dérivées des orthophotographies parmi les variables prédictrices, en ajustant un modèle sur le département de Haute Savoie. Il s'est avéré qu'il n'était pas nécessaire d'inclure ces variables. Nous avons donc pu inclure aux données de calibration deux autres zones d'études situées dans d'autres départements, pour permettre une modélisation de la conformité à l'habitat à l'échelle des Alpes du Nord, basée uniquement sur des variables de végétation dérivées de la cartographie de l'inven-

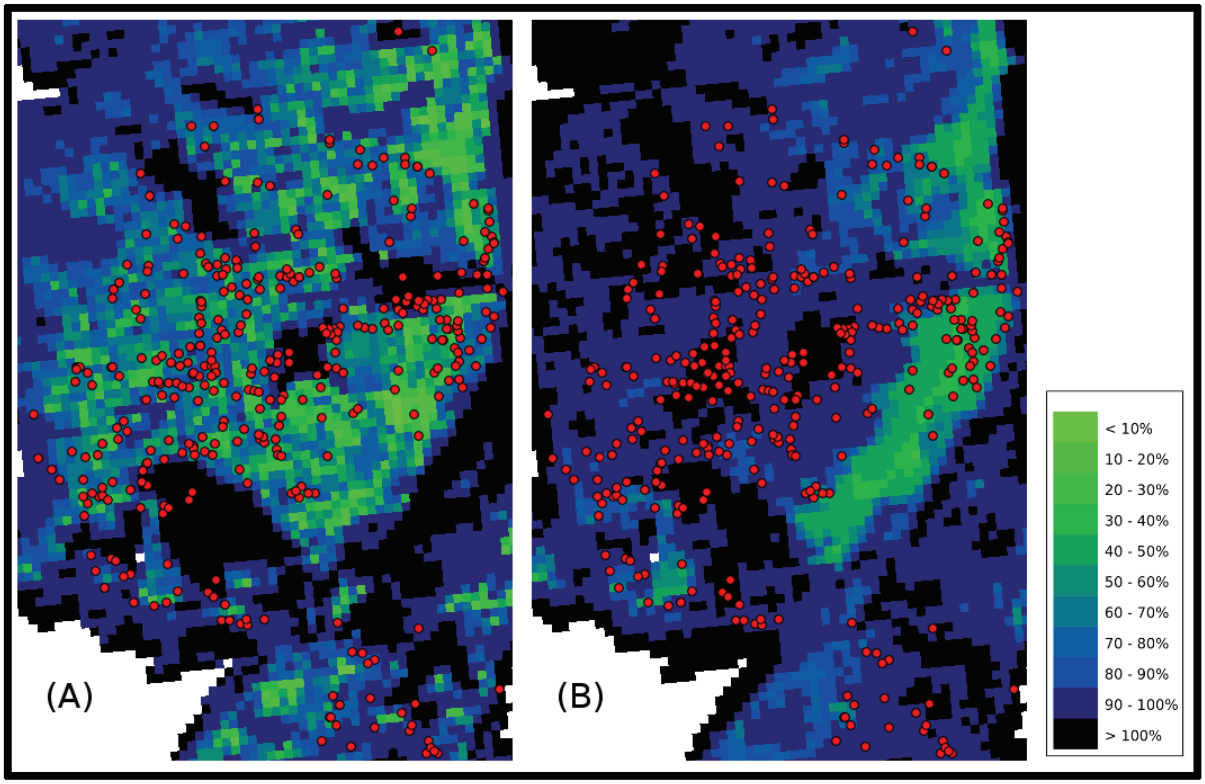

FIGURE 21 – Cartes de conformité à l'habitat d'hivernage du tétras-lyre sur le Parc naturel du Vercors obtenues à l'aide de : (A) la MADIFA, et (pour information), (B) une régression logistique complète. Les couleurs représentent différents seuils possibles pour la définition des types d'habitat. Un seuil s est défini de telle façon que l'ensemble des points du jeu de données de calibration pour lesquels la prédiction du modèle est supérieure à s contient $X\%$ des crottiers, avec X variant de 10 à 100% (cf. texte).

taire forestier national et sur des variables décrivant le relief (altitude, pente, exposition). Nous avons comparé cinq méthodes de modélisation (distances de Mahalanobis, analyse factorielle des distances de Mahalanobis [MADIFA], régression logistique "complète", régression logistique pas à pas, et forêt d'arbres décisionnels), et trois des méthodes (régression logistique complète et pas à pas, et MADIFA) ont montré un pouvoir de prédiction plus important que les autres. En s'appuyant sur des observations occasionnelles collectées dans deux autres zones d'étude (Parc du Vercors et Réserve naturelle de Villaroger), nous avons alors sélectionné la carte de prédiction renvoyée par la MADIFA, qui semblait bien prédire les zones de crottier.

Nous avons ensuite défini un seuil de conformité prédite s afin de distinguer les habitats et les non-habitat sur une zone. Nous avons choisi le seuil de conformité le plus élevé pour lequel 90% des crottiers présents sur les zones de calibration tombaient dans les zones prédites comme habitat. Cependant, nous avons vu que les habitats prédits par cette approche contenaient seulement les deux-tiers des crottiers à Villaroger et dans le Vercors. Bien que nous ayons conscience que ces crottiers sont des observations occasionnelles collectées sans dispositif d'échantillonnage fixé (et donc pour lesquelles nous ne pouvons garantir un effort d'échantillonnage spatialement uniforme), nous avons supposé qu'une partie de ces mauvaises prédictions était causée par un seuil s inadéquat. En effet, si la conformité est correctement prédite dans cette zone (on trouve bien plus de crottiers dans les zones prédites comme plus conformes), le seuil unique permettant le découpage en deux groupes de zones (habitat et non habitat) est probablement inadéquat, et devrait varier d'une zone à l'autre. La carte finale est présentée figure 20.

9.2 A propos des méthodes utilisées

Dans cette étude, nous avons choisi d'utiliser cinq méthodes différentes pour construire notre modèle prédictif.

L'approche basée sur les distances de Mahalanobis "brutes" s'est révélée extrêmement décevante. Nous avons donné dans la section 5.2.1 des éléments d'explication pour ce faible pouvoir prédictif : le calcul de la distance de Mahalanobis entre un point et l'environnement idéalement conforme nécessite l'estimation de $P+P\times(P-1)/2$ paramètres (le vecteur de moyenne et la matrice de variances-covariances, soit 136 paramètres dans notre étude) à partir des données, avec P le nombre de variables incluses dans le modèle (16 dans notre étude). Ce nombre de paramètres étant très grand, la variance associée à l'estimation des distances de Mahalanobis est en conséquence très forte. Cette approche illustre très bien qu'il est parfois préférable d'augmenter le biais associé à un modèle pour en diminuer l'erreur globale de prédiction. Ainsi, en travaillant sur une partie de l'espace écologique, la MADIFA conduit à des prédictions beaucoup plus fiables.

La régression logistique s'est également révélée une méthode de prédiction assez efficace (bien que moins efficace que la MADIFA). Notons que l'algorithme de sélection de variables pas à pas n'a pas permis d'améliorer sensiblement la précision des prédictions. Il y a une bonne raison à cela : cet algorithme repose sur la comparaison de modèles basés sur l'AIC. Or, nous l'avons indiqué (cf. note en bas de page 10), l'AIC n'est pas une mesure de la distance entre le modèle et la réalité, mais une estimation de cette distance, donc soumise aux fluctuations d'échantillonnage. Or, l'algorithme de sélection pas à pas ne tient pas compte de cette incertitude. L'effet d'une erreur au début de l'algorithme se propage au fur et à mesure de son évolution. Cela se traduit par une forte variance associée au modèle sélectionné (Guisan *et al.*, 2002; Guisan et Zimmermann, 2000). C'est à dire que la précision gagnée grâce à l'estimation d'un plus petit nombre de paramètres est perdue à cause de l'augmentation de l'incertitude associée à la sélection de modèles. Cette forte variance associée à l'algorithme à conduit certains statisticiens à considérer ce type d'approche comme une "expédition de pêche" (e.g., Johnson, 1981). Il n'est donc pas surprenant que le pouvoir prédictif du modèle sélectionné par la régression pas à pas ne soit pas sensiblement meilleur que la régression "complète" (laquelle, d'ailleurs, ne contenait que des variables environnementales sélectionnées avec soin : le modèle complet nous paraissait réaliste).

Notons qu'il existe des méthodes plus efficaces que la régression pas à pas pour réduire la dimension des modèles de régression. Ainsi, les régressions Ridge et Lasso (Friedman *et al.*, 2008) permettent

de définir des contraintes sur les coefficients b_j du modèle de régression. Ces contraintes (appelées *régularisation*) permettent de réduire la dimension du modèle sans conduire à une trop grande incertitude sur la sélection du modèle. Le temps imparti à cette étude étant limité, et la mise en œuvre de ce type d'approche pouvant être complexe, nous n'avons pas pu étudier ces possibilités dans cette étude. Notons toutefois qu'il pourrait être intéressant d'utiliser ces approches dans une future étude.

9.3 Le problème de l'autocorrélation spatiale

Dans leur article synthétisant les problèmes des approches couramment utilisées pour modéliser la conformité à l'habitat, Guisan *et al.* (2006) indiquent :

Even though current models of species' distributions are often said to be 'spatial', in most cases they are only partially spatial. The species–environment relationship is often fitted without explicit consideration of the neighbouring spatial context, for example without taking spatial autocorrelation or dispersal into account. (...) Predictions generated by models fitted within environmental space are then projected back onto the geographical space.

La prise en compte de la dimension spatiale permettrait probablement une amélioration de nos modèles. Mais il faut tout d'abord réfléchir aux causes de cette autocorrélation spatiale, et aux conséquences de sa non-prise en compte dans nos modèles, afin de savoir si le surcroît de travail impliqué par son intégration dans les modèles permettrait une amélioration substantielle ou non du pouvoir prédictif. En effet, la prise en compte de la dimension spatiale dans ce type de modèle est assez complexe, dépend de la source et de l'échelle à laquelle se produit l'autocorrélation (d'agit-il de gradients à grande échelle ou de phénomènes locaux ?, cf. Guisan *et al.*, 2006) et ne peut pas être appliquée avec toutes les méthodes que nous avons utilisé ici (e.g., difficile d'intégrer la dimension spatiale avec les distances de Mahalanobis ou la MADIFA).

Lorsque l'on examine la figure 6, nous constatons la présence d'une très forte structure spatiale dans la distribution des crottiers identifiés sur la zone : ceux-ci ne sont pas distribués aléatoirement sur les zones prospectées, mais sont regroupés par "paquets" (*clusters*). Plusieurs auteurs soulignent deux sources possibles pour les structures observées dans un semis d'occurrences d'espèces (Legendre, 1993; Lennon, 1999) :

- ⋆ la distribution de ces crottiers est influencée par des variables environnementales qui sont elles même structurées spatialement.
- ⋆ la distribution de ces crottiers est influencée par le comportement de l'espèce en lui-même, qui implique un processus contagieux au niveau de la distribution des occurrences de l'espèce ;

Les modèles utilisés pour prédire la conformité à l'habitat ne permettent la prise en compte que de la première source de structuration spatiale. Par exemple, nous avons vu dans la section 7.1.2 que la présence de crottiers était favorisée par la présence de zones orientées au nord. Si, sur une zone donnée, ce type d'exposition n'est présent que sous la forme de "patchs" (i.e. pas aléatoirement distribué sur la zone, mais présent uniquement à certains endroits restreints de la zone), alors il est parfaitement logique d'observer la même structuration au niveau de la distribution des crottiers.

La seconde source de structuration spatiale, i.e. des processus biotiques liés à l'espèce elle-même, sont toutefois possibles. Bien sûr, il y a peu de chances pour que le tétras-lyre tende à rechercher ses conspécifiques pour l'établissement de ses zones d'hivernage (si tel était le cas, alors la présence d'un crottier de tétras-lyre à un endroit tendrait à "attirer" les autres tétras-lyre, et ce regroupement purement social expliquerait la présence des clusters observés dans la distribution des crottiers). En effet, un crottier est basiquement un igloo creusé dans la neige (donc peu visible), dans lequel l'animal est seul (donc pas d'interactions sociales). Nous ne voyons donc aucune raison pour que de tels processus sociaux soient la cause de ces regroupements observés des crottiers. En revanche, comme un même individu creuse plusieurs

crottiers le long de l'hiver, et que ces individus sont sédentaires, il est possible qu'un même animal creuse ses crottiers à proximité les uns des autres. En effet, le domaine vital hivernal d'un tétras-lyre couvre une surface de l'ordre de 40 hectares (Yann Magnani, com. pers.), c'est à dire une surface relativement faible par rapport à la surface de nos zones d'étude. Cette sédentarité, non prise en compte dans notre modèle, pourrait expliquer une partie du *clustering* observé pour la distribution des crottiers.

En réalité, le processus de modélisation que nous avons mis en œuvre suppose que la seule explication de la structuration observée au niveau du semis de points constitué par les crottiers vient de la structuration spatiale des variables *utilisées dans le modèle*[17]. Nous ignorons donc les causes de structuration liées aux variables environnementales non incluses dans le modèle (parce que non disponibles) et les causes liées à l'espèce en elle-même (e.g. la sédentarité des individus). La prise en compte de cette autocorrélation permettrait d'accroître le pouvoir prédictif des modèles que nous construisons, en rendant ces modèles biologiquement plus réalistes.

Pourtant, nous ne pensons pas que le surcroît de travail impliqué par une telle réflexion serait réellement bénéfique : nous devons garder en tête notre objectif. Nous souhaitions construire une carte prédisant assez précisément la conformité à l'habitat. Nous avons vu, en examinant les figures 15, 17 et 18, et en quantifiant le pouvoir prédictif de la MADIFA aux étapes de validation interne et externe, que ce pouvoir prédictif était excellent (indice de Boyce *et al.* (2002) toujours supérieur à 0.9, tableau 7). Ainsi, s'il est possible que la prise en compte de la dimension spatiale dans les modèles prédictifs (en tous cas, dans ceux où cette prise en compte est possible, comme la régression logistique) ait amélioré le pouvoir prédictif du modèle, cette augmentation n'aurait pas été réellement substantielle. Nous devons garder en tête que l'objectif d'un modèle prédictif est de prédire correctement, et tous nos examens semblent indiquer que nous l'avons atteint.

9.4 Les problèmes de la prédiction statistique

Nous devons enfin mettre en garde le lecteur sur le principe de la prédiction statistique. Nous reproduisons dans cette partie un passage tiré de la thèse de Calenge (2005), qui est de circonstance ici :

> Pourtant, même lorsque les modèles sont validés sur des données indépendantes, l'inférence peut être un problème. Une illustration de ce problème est donnée dans les deux études de Knick et Dyer (1997) et de Knick et Rotenberry (1998) sur le lièvre de Californie (*Lepus californicus*). Knick et Dyer (1997) ont construit une carte de qualité de l'habitat du lièvre de Californie à l'aide des distances de Mahalanobis sur un jeu de localisations recueillies de 1987 à 1989, et de 1992 à 1993. Puis, le modèle construit à été validé par le calcul des prédictions sur un jeu de localisations indépendant recueillies sur la même zone de 1993 à 1995. Ces localisations étaient toujours détectées dans des zones prédites comme de haute qualité. Ce modèle avait été ajusté dans le cadre de la conservation de l'aigle royal (*aquila chysaetos*) dans L'Idaho, dont la principale proie est le lièvre de Californie. En maximisant la qualité de l'habitat pour le lièvre, on maximise du même coup celle du rapace. Mais leur zone d'étude est soumise à de fréquents incendies, qui se traduisent par une réduction de l'embroussaillement de la zone. Le feu peut donc devenir un moyen de gestion des deux espèces. Par conséquent, Knick et Rotenberry (1998) se sont posés la question de savoir s'il était possible de se servir de leur modèle pour prédire les réponses de la distribution des lièvres à différents scénarios de gestion, en utilisant des simulations. Ces auteurs ont répondu par la négative ; les simulations du modèle aboutissaient à des prédictions absurdes. **Lorsque la configuration de l'habitat change sur une zone, la validité du modèle peut être remise en question**. La question la plus importante est donc parfaitement exprimée par Johnson (1981) :
>
> "*What is the universe to which our results are to pertain ? If it is a single study area, in a single year, with the measurements we have observed, there is no problem. If we want to generalize, let us be careful. Our study area must be representative of the area we want to extrapolate to, similarly the year and the habitat.*".

Nous disposons ici d'un modèle ajusté sur certaines zones d'étude, et nous l'utilisons pour prédire la conformité à l'habitat sur des zones externes. Certes, nous avons cherché à collecter des données provenant de zones d'études aux conditions variées (cf. section 7.1.1). Nous avons encore augmenté la diversité des conditions environnementales auxquelles le tétras-lyre pouvait être soumis en utilisant les données col-

17. Notons que si la structuration spatiale est causée par une variable environnementale qui n'est pas incluse dans le modèle, alors le modèle ne rendra pas compte de toutes les structures spatiales du semis de crottier, et la prédiction pourrait être mauvaise

lectées lors de deux années aux conditions climatiques très différentes (une année très enneigée, et une année peu enneigée). Nous avons vérifié la validité de notre modèle sur des zones très éloignées (Vercors, Villaroger), en utilisant des données collectées lors d'années encore différentes. Nous avons donc tout mis en œuvre pour assurer la plus grande robustesse possible de notre modèle.

Mais les zones utilisées pour calibrer le modèle sont de petite taille par rapport à la zone sur laquelle le modèle est appliqué (cf. figure 1). Bien que la validation externe du modèle nous donne une certaine confiance dans ses prédictions, il est malgré tout possible que sur l'ensemble des Alpes du nord, il existe des zones au contexte environnemental particulier, fondamentalement différent des zones étudiées ici. Ainsi, les prédictions obtenues pour l'extrême sud des Alpes du nord nous paraissent suspectes : on ne trouve aucun habitat prédit au sud du Vercors (figure 16). Ne disposant d'aucune donnée pour étudier le pouvoir prédictif de notre modèle dans cette zone, et le doute étant trop grand, nous ne pourrons donc valider l'utilisation de ce modèle pour les zones situées au sud du Vercors.

Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag.

Boyce, M., Vernier, P., Nielsen, S., et Schmiegelow, F. (2002). Evaluating resource selection functions. *Ecological modelling*, 157 : 281–300.

Breiman, L. (2001). Random Forests. *Machine learning*, 45 : 5–32.

Breiman, L. (2002). *Manual On Setting Up, Using, And Understanding Random Forests V3.1.*

Breiman, L., Friedman, J., Olshen, R., et Stone, C. (1993). *Classification and regression trees.* Chapman & Hall.

Burnham, K. et Anderson, D. (1998). *Model selection and inference.* Springer, Berlin.

Calenge, C. (2005). *Des outils statistiques pour l'analyse des semis de points dans l'espace écologique.* PhD thesis, Université Claude Bernard Lyon 1.

Calenge, C. (2007). Exploring Habitat Selection by Wildlife with adehabitat. *Journal of Statistical Software*, 22.

Calenge, C. et Basille, M. (2008). A general framework for the statistical exploration of the ecological niche. *Journal of Theoretical Biology*, 252 : 674–685.

Calenge, C., Darmon, G., Basille, M., Loison, A., et Jullien, J. (2008). The factorial decomposition of the Mahalanobis distances in habitat selection studies. *Ecology*, 89 : 555–566.

Calenge, C., Dufour, A., et Maillard, D. (2005). K-select analysis : a new method to analyse habitat selection in radio-tracking studies. *Ecological Modelling*, 186 : 143–153.

Chase, J. et Leibold, M. (2003). *Ecological niches. Linking class and contemporary approaches.* The University of Chicago Press.

Chen, C., Liaw, A., et Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data. Technical report, University of Berkeley.

Clark, J., Dunn, J., et Smith, K. (1993). A multivariate model of female black bear habitat use for a geographic information system. *Journal of Wildlife Management*, 57(3) : 519–526.

Cox, D. et Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. *American Statistician*, pages 1–4.

Cutler, D., Edwards Jr, T., Beard, K., Cutler, A., Hess, K., Gibson, J., et Lawler, J. (2007). Random forests for classification in ecology. *Ecology*, 88(11) : 2783–2792.

De'ath, G. et Fabricius, K. (2000). Classification and regression trees : a powerful yet simple technique for ecological data analysis. *Ecology*, 81 : 3178–3192.

Dolédec, S. et Chessel, D. (1987). Rythmes saisonniers et composantes stationnelles en milieu aquatique I- Description d'un plan d'observations complet par projection de variables. *Acta Œcologica, Œcologia Generalis*, 8 : 403–426.

Elith, J., Graham, C., Anderson, R., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R., Huettmann, F., Leathwick, J., Lehmann, A., Li, J., Lohmann, L., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC. Overton, J., Peterson, A., Phillips, S., Richardson, K., Scachetti-Pereira, R., Schapire, R., Soberon, J., Williams, S., Wisz, M., et Zimmermann, N. (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, 29 : 129–151.

Engler, R., Guisan, A., et Rechsteiner, L. (2004). An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41(2) : 263–274.

Farber, O. et Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological modelling*, 160 : 115–130.

Fielding, A. et Bell, J. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1) : 38–49.

Friedman, J., Hastie, T., et Tibshirani, R. (2008). *The elements of statistical learning. Data Mining, Inference and Prediction. Second Edition.*

GRASS Development Team (2008). *Geographic Resources Analysis Support System (GRASS GIS) Software*. Open Source Geospatial Foundation.

Guisan, A., Edwards, T., et Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions : setting the scene. *Ecological modelling*, 157 : 89–100.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., McC. Overton, J., Aspinall, R., et Hastie, T. (2006). Making better biogeographical predictions of species' distributions. *Journal of Applied Ecology*, 43 : 386–392.

Guisan, A. et Zimmermann, N. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135 : 147–186.

Hirzel, A., Hausser, J., Chessel, D., et Perrin, N. (2002). Ecological-niche factor analysis : How to compute habitat suitability maps without absence data ? *Ecology*, 83(7) : 2027–2036.

Hirzel, A., Le Lay, G., Helfer, V., Randin, C., et Guisan, A. (2006). Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2) : 142–152.

Hosmer, D. et Lemeshow, S. (2000). *Applied logistic regression. Second Edition.* John Wiley & Sons.

Hutchinson, G. (1957). Concluding remarks. In *Cold Spring Harbour Symposium*, volume 22, pages 415–427. Quantitative Biology.

Johnson, D. (1981). The use and misuse of statistics in wildlife habitat studies. In Capen, D., editor, *The use of multivariate statistics in studies of wildlife habitat*, pages 11–19. USDA Forest Service.

Knick, S. et Dyer, D. (1997). Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho. *Journal of Wildlife Management*, 61(1) : 75–85.

Knick, S. et Rotenberry, J. (1998). Limitations to mapping habitat use areas in changing landscapes using the Mahalanobis distance statistic. *Journal of Agricultural, Biological, and Environmental Statistics*, 3(3) : 311–322.

Legendre, P. (1993). Spatial autocorrelation : trouble or new paradigm ? *Ecology*, 74(6) : 1659–1673.

Lennon, J. (1999). Resource selection functions : taking space seriously ? *Trends in Ecology & Evolution*, 14(10) : 399–400.

Lev, J. (1949). The point biserial coefficient of correlation. *The Annals of Mathematical Statistics*, 20(1) : 125–126.

Liaw, A. et Wiener, M. (2002). Classification and regression using randomForest. *R News*, 2 : 18–22.

McCullagh, P. et Nelder, J. (1989). *Generalized linear models. Second Edition.* Chapman & Hall, London.

Pearce, J. et Boyce, M. (2006). Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, 43 : 405–412.

Phillips, S., Anderson, R., et Schapire, R. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4) : 231–259.

Phillips, S., Dudík, M., Elith, J., Graham, C., Lehmann, A., Leathwick, J., et Ferrier, S. (2009). Sample selection bias and presence-only distribution models : implications for background and pseudo-absence data. *Ecological Applications*, 19(1) : 181–197.

R Development Core Team (2011). *R : A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rotenberry, J., Preston, K., et Knick, S. (2006). GIS-based niche modeling for mapping species habitat. *Ecology*, 87 : 1458–1464.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science*, 240(4857) : 1285.

van Horne, B. (1983). Density as a misleading indicator of habitat quality. *Journal of Wildlife Management*, 47 : 893–901.

VanDerWal, J., Shoo, L., Graham, C., et Williams, S. (2009). Selecting pseudo-absence data for presence-only distribution modeling : How far should you stray from what you know ? *ecological modelling*, 220(4) : 589–594.

Venables, W. et Ripley, B. (2002). *Modern applied statistics with S-Plus. Fourth Edition.* Springer, Berlin.

Wand, M. et Jones, M. (1995). *Kernel smoothing.* Chapman & Hall/CRC.

Annexe : relation entre la probabilité de présence réelle d'une espèce et la probabilité de présence modélisée avec des données de type "points de contextes/occurrences de l'espèce"

Nous démontrons dans cette annexe qu'il est possible d'utiliser des méthodes prévues pour des données de type "présence/absence", pour modéliser la probabilité de présence d'un phénomène à partir de données de type "présence/points de contexte".

Nous supposons le dispositif suivant : nous supposons que l'ensemble de la zone d'étude a été prospectée avec une intensité de recherche constante, et que lorsqu'un point crottier était localisé, la valeur des variables environnementales était mesurée. Soit N_u le nombre de crottiers identifiés lors de ces opérations. Nous supposons par ailleurs que N_d points ont ensuite été placés aléatoirement sur la zone d'étude, et que les caractéristiques environnementales ont été mesurées à chaque point. Ces N_d points représentent des points de contexte. L'échantillon sur lequel nous travaillons est donc constitué des $N = N_d + N_u$ points crottiers et points de contexte.

Nous travaillons à présent avec un nombre infini d'unités statistiques (une surface – la zone d'étude – est constituée d'une infinité de points). Soit $\mathbf{m}$ le vecteur de longueur 2 contenant les coordonnées spatiales d'un point de la zone d'étude. Nous définissons une variable aléatoire $s_{\mathbf{m}}$ prenant la valeur 1 lorsque le point $\mathbf{m}$ appartient à l'échantillon et 0 sinon. Soit $y_{\mathbf{m}} = 1$ si le point $\mathbf{m}$ correspond à un crottier *identifié comme tel au moment des opérations de prospection*, et $y_{\mathbf{m}} = 0$ quand le point $\mathbf{m}$ correspond à un point de contexte dans notre échantillon.

Comme nous ne disposons pas de la totalité des crottiers, il est possible que certains points de contexte correspondent en réalité à des crottiers. Comme précédemment, nous pouvons définir une variable $z_{\mathbf{m}}$ définissant le statut réel du point $\mathbf{m}$: il prendra la valeur 1 si le point $\mathbf{m}$ est effectivement un point crottier dans la réalité, et la valeur 0 sinon. Soit $\mathbf{x_m}$ le vecteur contenant les valeurs des variables environnementales au point $\mathbf{m}$.

Nous posons les hypothèses suivantes :

$$
\begin{aligned}
P(s_{\mathbf{m}} = 1 | y_{\mathbf{m}} = 1) &= 1 && (2)\\
P(s_{\mathbf{m}} = 1 | y_{\mathbf{m}} = 0, \mathbf{x_m}) &= P(s_{\mathbf{m}} = 1 | y_{\mathbf{m}} = 0) && (3)\\
P(y_{\mathbf{m}} = 1 | z_{\mathbf{m}} = 1, \mathbf{x_m}) &= P(y_{\mathbf{m}} = 1 | z_{\mathbf{m}} = 1) && (4)\\
P(y_{\mathbf{m}} = 1 | z_{\mathbf{m}} = 0, \mathbf{x}_m) &= 0 && (5)
\end{aligned}
$$

Les hypothèses 2 et 3 sont vérifiées par construction : elles indiquent : (i) que dès qu'un crottier est identifié sur la zone d'étude, il appartient à l'échantillon (il s'agit d'une relation d'implication) ; (ii) que les points de contexte sont générées totalement aléatoirement et que la probabilité qu'un point de la zone d'étude où l'on n'a pas localisé de crottier appartienne à l'échantillon ne dépend pas de l'environnement. L'hypothèse 4 indique que la probabilité de détecter un crottier ne dépend pas de la composition environnementale à l'emplacement du crottier (détectabilité identique d'un habitat à l'autre). L'hypothèse 5 indique qu'il est impossible d'identifier un crottier à un point où un crottier est absent, et ce quelles que soient les conditions environnementales (pas d'identification incorrecte).

Les données dont nous disposons nous permettent alors de modéliser $P(y_{\mathbf{m}} = 1 | s_{\mathbf{m}} = 1, \mathbf{x_m})$, c'est à dire la probabilité qu'un point de l'échantillon soit un crottier et non un point de contexte en fonction des variables environnementales. Tout d'abord, notons que :

$$
P(y_{\mathbf{m}} = 1 | s_{\mathbf{m}} = 1, \mathbf{x_m}) = \frac{P(y_{\mathbf{m}} = 1, s_{\mathbf{m}} = 1 | \mathbf{x_m})}{P(s_{\mathbf{m}} = 1 | \mathbf{x_m})} \quad (6)
$$

La probabilité qu'un point de la zone d'étude appartienne à l'échantillon et qu'il s'agisse d'un point

crottier est le produit de la probabilité que ce point soit un point crottier, de la probabilité qu'un point crottier soit identifié comme tel lors de la prospection, et de la probabilité qu'un point crottier identifié comme tel soit inclus dans notre échantillon, c'est à dire :

$$\begin{aligned} P(y_{\mathbf{m}}=1, s_{\mathbf{m}}=1|\mathbf{x_m}) &= P(z_{\mathbf{m}}=1|\mathbf{x_m})P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1,\mathbf{x_m})P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=1,z_{\mathbf{m}}=1,\mathbf{x}_m) \\ &= P(z_{\mathbf{m}}=1|\mathbf{x_m})P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1) \end{aligned} \tag{7}$$

cette dernière simplification est la conséquence des hypothèses 2 et 4. Par ailleurs, nous pouvons noter que :

$$\begin{aligned} P(s_{\mathbf{m}}=1|\mathbf{x_m}) &= P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=1,\mathbf{x_m})P(y_{\mathbf{m}}=1|\mathbf{x_m}) + P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0,\mathbf{x_m})P(y_{\mathbf{m}}=0|\mathbf{x_m}) \\ &= P(y_{\mathbf{m}}=1|\mathbf{x_m}) + P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0)P(y_{\mathbf{m}}=0|\mathbf{x_m}) \end{aligned} \tag{8}$$

cette dernière simplification est la conséquence des hypothèses 2 et 3. Enfin, nous pouvons noter que :

$$\begin{aligned} P(y_{\mathbf{m}}=1|\mathbf{x_m}) &= P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1,\mathbf{x_m})P(z_{\mathbf{m}}=1|\mathbf{x_m}) + P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=0,\mathbf{x_m})P(z_{\mathbf{m}}=0|\mathbf{x_m}) \\ &= P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)P(z_{\mathbf{m}}=1|\mathbf{x_m}) \end{aligned} \tag{9}$$

cette dernière simplification est la conséquence des hypothèses 4 et 5. Nous pouvons alors replacer l'équation 9 dans l'équation 8, et les équations 8 et 7 dans l'équation 6, ce qui donne :

$$\begin{aligned} &P(y_{\mathbf{m}}=1|s_{\mathbf{m}}=1,\mathbf{x_m}) = \\ &\frac{P(z_{\mathbf{m}}=1|\mathbf{x_m})P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)}{P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)P(z_{\mathbf{m}}=1|\mathbf{x_m}) + P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0)(1-P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)P(z_{\mathbf{m}}=1|\mathbf{x_m}))} \\ &= \frac{P(z_{\mathbf{m}}=1|\mathbf{x_m})P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)}{P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)P(z_{\mathbf{m}}=1|\mathbf{x_m}) + P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0) - P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0)P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)P(z_{\mathbf{m}}=1|\mathbf{x_m})} \\ &= \frac{P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)}{P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1) + \frac{P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0)}{P(z_{\mathbf{m}}=1|\mathbf{x_m})} - P(s_{\mathbf{m}}=1|y_{\mathbf{m}}=0)P(y_{\mathbf{m}}=1|z_{\mathbf{m}}=1)} \\ &= \frac{a}{b+\frac{c}{P(z_{\mathbf{m}}=1|\mathbf{x_m})}} \end{aligned} \tag{10}$$

avec a, b, c des constantes inconnues. Nous démontrons avec cette dernière équation qu'en modélisant, en fonction des variables environnementales, la probabilité qu'un point de l'échantillon soit un crottier et non un point de contexte est une fonction monotone croissante de la probabilité qu'un point de la zone d'étude soit un point crottier. Ainsi, si nous considérons deux points $\mathbf{m}_1$ et $\mathbf{m}_2$, alors si $P(y_{\mathbf{m}_1}=1|s_{\mathbf{m}_1}=1,\mathbf{x}_{\mathbf{m}_1}) < P(y_{\mathbf{m}_2}=1|s_{\mathbf{m}_2}=1,\mathbf{x}_{\mathbf{m}_2})$, nous avons la garantie que $P(z_{\mathbf{m}_1}=1|\mathbf{x}_{\mathbf{m}_1}) < P(z_{\mathbf{m}_2}=1|\mathbf{x}_{\mathbf{m}_2})$.