Thèse de doctorat de l'établissement
Université Bourgogne Franche-Comté

École Doctorale n° 554 Environnements - Santé

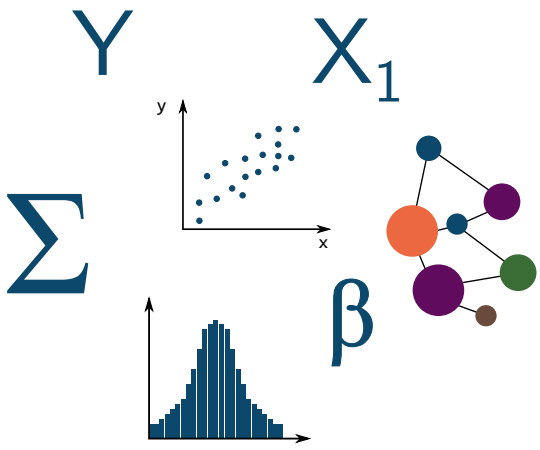
Spécialité : Biologie des populations et écologie

Par

Paul Savary

Utilisation conjointe de graphes génétiques et paysagers pour l'analyse de la connectivité écologique des habitats

Annexes B - Méthodes d'analyse statistique



ARP-Astrance, 75008 Paris, France
UMR 6049 ThéMA, Université Bourgogne Franche-Comté - CNRS, 25000 Besançon, France
UMR 6282 Biogéosciences, Université Bourgogne Franche-Comté - CNRS, 21000 Dijon, France

Table des matières

Annexe B1

Utilisation du principe d'indépendance conditionnelle pour la construction de graphes génétiques

Le principe d'indépendance conditionnelle repose sur le fait que deux événements ou deux variables sont conditionnellement indépendants s'ils sont statistiquement indépendants après avoir pris en compte un troisième événement ou une troisième variable (Magwene, 2001). Un graphe d'indépendance est un graphe qui représente les relations d'indépendance conditionnelle entre un ensemble de variables (Magwene, 2001).

En génétique des populations, ce type de graphe a pour la première fois été utilisé par Dyer et Nason (2004). Dans ce cas, les "variables" sont des populations et la série des valeurs de chaque "variable" est la série de fréquences alléliques caractérisant la population. Créer un graphe d'indépendance génétique consiste à identifier les paires de populations qui peuvent être considérées indépendantes une fois que toutes leurs relations avec les autres populations ont été prises en compte. Nous présentons ici les bases mathématiques et les étapes de calcul permettant la construction de ce type de graphe (Figure 1).

Soit $\mathbf{Y}$ un ensemble de p variables suivant une distribution normale multivariée : $\mathbf{Y} = \{y_1, y_2, \ldots, y_p\}$. Les trois postulats suivants sont équivalents (Krzanowski et Marriott, 1995, in Magwene, 2001) :

— Les variables y_1 et y_2 sont indépendantes, conditionnellement à $\mathbf{Y}_K$, avec $\mathbf{Y}_K$ tout sous-ensemble de $\mathbf{Y}$ n'incluant ni y_1 ni y_2.

— La corrélation partielle entre y_1 et y_2 est nulle : $\rho_{ij.\{K\}} = 0$

— Soit $\mathbf{C}$ la matrice de covariance de l'ensemble des variables $\mathbf{Y}$, alors l'élément π_{ij} de la matrice de covariance inverse $\mathbf{\Pi} = \mathbf{C}^{-1}$ (aussi appelée matrice de précision) est nul.

Ainsi, pour évaluer les relations d'indépendance conditionnelle d'un ensemble de populations, il faut tout d'abord calculer une matrice de corrélation partielle ou de précision à partir de données génétiques. En génétique des populations, les génotypes pour plusieurs *loci* des individus des différentes populations sont codés sous la forme d'une matrice ayant autant de colonnes que d'allèles et de lignes que d'individus (Figure 1A). L'absence d'un allèle est codée par un 0. La présence d'une ou deux copies d'un allèle dans le génotype d'un individu est codée par un 0.5 ou un 1, respectivement. Si ces données sont codées avec les valeurs 0, 1 et 2, comme dans les travaux de Fortuna *et al.* (2009) et Smouse et Peakall (1999), cela n'affecte pas le calcul.

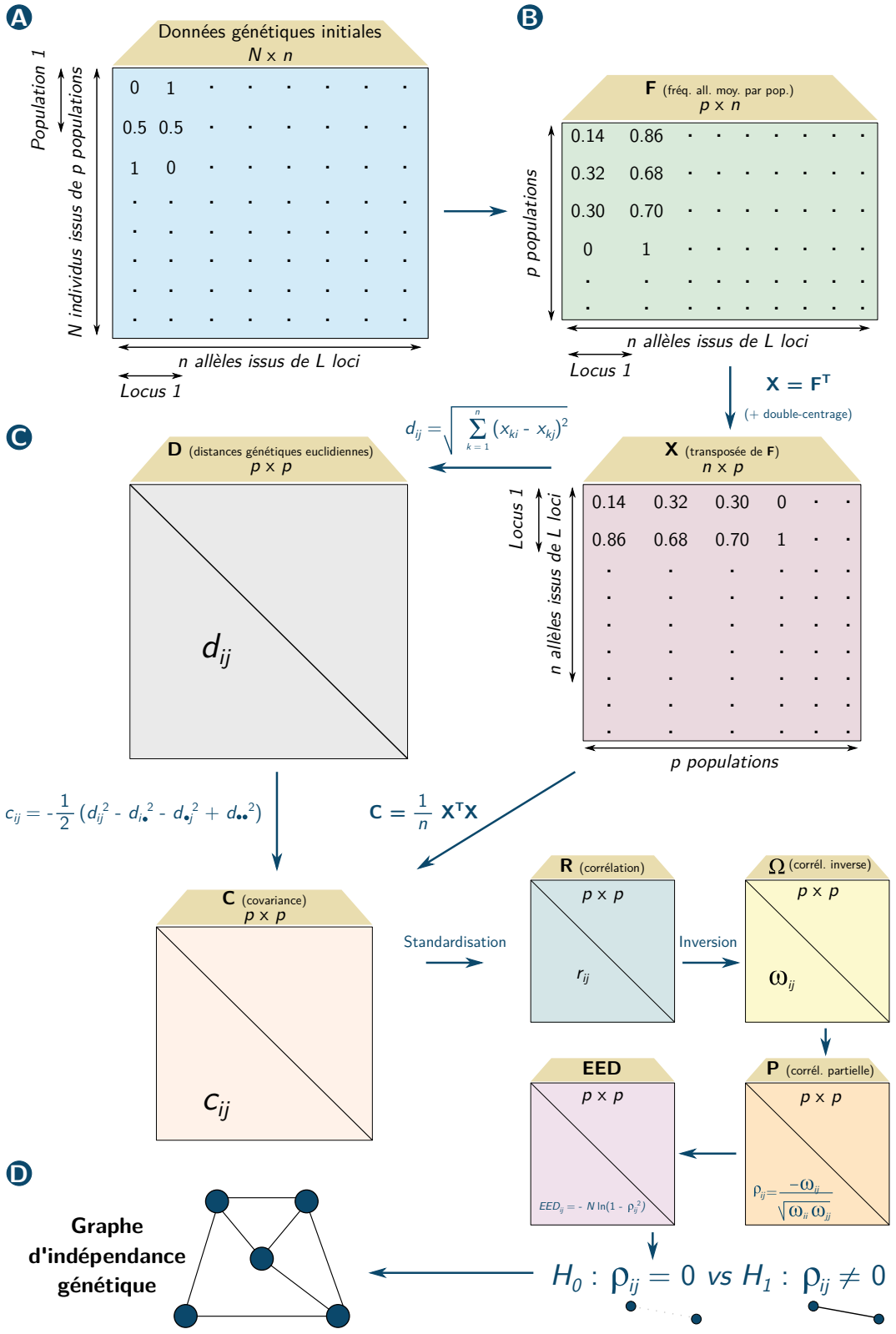


FIGURE 1 – Construction d'un graphe génétique à l'aide du principe d'indépendance conditionnelle. À partir des données génétiques initiales (A), on calcule les fréquences alléliques par population (B). (C) On calcule la matrice de covariance à partir de laquelle plusieurs étapes de calcul permettent d'aboutir au critère d'*Edge Exclusion Deviance*. (D) C'est à partir de ce critère qu'on teste la significativité de la corrélation partielle entre populations pour déterminer la présence ou non de chacun des liens du graphe d'indépendance génétique.

Dans un premier temps, les fréquences alléliques moyennes par population sont calculées. Ces fréquences sont les éléments d'une matrice **F** comptant autant de colonnes qu'il y a d'allèles et autant de lignes qu'il y a de populations (Figure 1B). Ces fréquences alléliques forment les séries de valeurs caractérisant chaque population, considérées comme les variables lors de la construction du graphe d'indépendance génétique. L'étape suivante consiste à calculer la covariance entre les populations (entre les lignes de **F** donc). Dyer et Nason (2004) calculent cette covariance en commençant par calculer une matrice de distance génétique euclidienne, suivant les travaux de Gower (1966) ayant démontré la dualité qui existe entre la distance et la covariance.

Pour cela, la matrice **F** des fréquences alléliques moyennes par population doit être centrée à la fois par lignes et par colonnes pour que le calcul de la covariance à partir des distances génétiques lors des étapes ultérieures soit correct. Néanmoins, dans ce cas particulier, cette étape n'est pas obligatoire car (i) les sommes des valeurs par ligne sont égales au nombre de *loci* car la somme des fréquences alléliques vaut 1 pour chaque *locus*, et (ii) car le fait de centrer par colonnes n'affecte pas les distances euclidiennes entre les populations (lignes), compte tenu de la définition d'une distance euclidienne. Sans ce "double-centrage", la matrice de covariance entre populations calculée à partir des distances génétiques est équivalente à la matrice de covariance entre les colonnes de la transposée **X** de la matrice doublement centrée **F** des fréquences alléliques. Nous démontrons cela par la suite. Nous démontrons également pourquoi la covariance doit être calculée à partir de distances au carré et non pas de distances simples, d'un point de vue strictement mathématique, en nous appuyant sur les travaux d'Everitt et Hothorn (2011) (page 107), de Gower (1966) et Smouse et Peakall (1999) (équation 13) (Figure 1C).

La distance génétique euclidienne d_{ij} entre les populations i et j est calculée à partir de la transposée **X** de la matrice **F** des fréquences alléliques. **X** est de dimension $n \times p$, avec n le nombre d'allèles et p le nombre de populations. La distance génétique est calculée avec la formule suivante :

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2} \tag{1.1}$$

La covariance c_{ij} entre les variables/populations i et j se calcule avec la formule suivante :

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}) \tag{1.2}$$

Comme **F** a été centrée à la fois par lignes et par colonnes, $\bar{x_i} = \bar{x_j} = 0$. Ainsi, la covariance entre les variables/populations i et j est simplement :

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ki} x_{kj} \tag{1.3}$$

Et par conséquent :

$$\begin{aligned} c_{ii} &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}^2 \\ c_{jj} &= \frac{1}{n}\sum_{k=1}^{n} x_{kj}^2 \end{aligned} \tag{1.4}$$

Il s'ensuit que la matrice de covariance $\mathbf{C}$ est :

$$\mathbf{C} = \frac{1}{n}\mathbf{X^T X} \tag{1.5}$$

tel que $\mathbf{X^T}$ est de taille $p \times n$, $\mathbf{X}$ de taille $n \times p$ et $\mathbf{C}$ de taille $p \times p$.

La somme des éléments de chaque ligne de $\mathbf{C}$ vaut ainsi :

$$\begin{aligned} \sum_{j=1}^{p} c_{ij} &= \sum_{j=1}^{p} \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= \frac{1}{n}\left[(\sum_{k=1}^{n} x_{ki}x_{k1}) + (\sum_{k=1}^{n} x_{ki}x_{k2}) + \ldots + (\sum_{k=1}^{n} x_{ki}x_{kp})\right] \\ &= \frac{1}{n}\left[(x_{1i}x_{11} + x_{2i}x_{21} + \ldots + x_{ni}x_{n1}) + \ldots + (x_{1i}x_{1p} + x_{2i}x_{2p} + \ldots + x_{ni}x_{np})\right] \\ &= \frac{1}{n}\left[x_{1i} \times (\sum_{j=1}^{p} x_{1j}) + x_{2i} \times (\sum_{j=1}^{p} x_{2j}) + \ldots + x_{ni} \times (\sum_{j=1}^{p} x_{nj})\right] \\ &= \frac{1}{n}\left[x_{1i} \times 0 + x_{2i} \times 0 + \ldots + x_{ni} \times 0\right] \\ &= 0 \end{aligned} \tag{1.6}$$

car les sommes des lignes de $\mathbf{X}$ sont nulles étant donné que $\mathbf{F}$ a été centrée par lignes et par colonnes.

La trace T de $\mathbf{C}$ est :

$$T = \sum_{i=1}^{p} c_{ii} \tag{1.7}$$

Exprimons d_{ij}^2 en fonction des éléments de $\mathbf{C}$:

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{n} (x_{ki} - x_{kj})^2 \\ &= \sum_{k=1}^{n} (x_{ki}^2 - 2x_{ki}x_{kj} + x_{kj}^2) \\ &= \sum_{k=1}^{n} x_{ki}^2 + \sum_{k=1}^{n} x_{kj}^2 - 2\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \end{aligned} \tag{1.8}$$

On a alors :

$$\begin{aligned}\sum_{i=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\ &= n \times (\sum_{i=1}^{p} c_{ii} + \sum_{i=1}^{p} c_{jj} - 2\sum_{i=1}^{p} c_{ij})\end{aligned} \tag{1.9}$$

Vu que $\sum_{j=1}^{p} c_{ij} = 0$ et que $\mathbf{C}$ est une matrice symétrique, $\sum_{i=1}^{p} c_{ij} = 0$. On a alors :

$$\begin{aligned}\sum_{i=1}^{p} d_{ij}^2 &= n \times (T + pc_{jj} - 2 \times 0) \\ &= n \times (T + pc_{jj}) \\ \sum_{j=1}^{p} d_{ij}^2 &= n \times (T + pc_{ii})\end{aligned} \tag{1.10}$$

On peut alors calculer $\sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2$:

$$\begin{aligned}\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p}\sum_{j=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\ &= n \times (\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ii} + \sum_{i=1}^{p}\sum_{j=1}^{p} c_{jj} - 2\sum_{i=1}^{p}\sum_{j=1}^{p} c_{ij}) \\ &= n \times (pT + pT - 2 \times 0) \\ &= n \times 2pT\end{aligned} \tag{1.11}$$

On calcule ensuite $d_{i\bullet}^2$, $d_{\bullet j}^2$ et $d_{\bullet\bullet}^2$:

$$\begin{aligned}d_{i\bullet}^2 &= \frac{1}{p}\sum_{j=1}^{p} d_{ij}^2 \\ &= \frac{1}{p} \times n \times (T + pc_{ii}) \\ &= n \times (\frac{T}{p} + c_{ii}) \\ &= n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{ii}) \\ d_{\bullet j}^2 &= n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{jj}) \\ d_{\bullet\bullet}^2 &= \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} d_{ij}^2 \\ &= \frac{n}{p^2} \times 2pT \\ &= 2 \times \frac{n}{p}\sum_{i=1}^{p} c_{ii}\end{aligned} \tag{1.12}$$

Il découle de la formule utilisée pour calculer la distance euclidienne que :

$$
\begin{aligned}
d_{ij}^2 &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \\
c_{ij} &= -\frac{1}{2}(\frac{d_{ij}^2}{n} - c_{ii} - c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} - n \times c_{jj} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}) \\
&= -\frac{1}{2n}\left[d_{ij}^2 - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{ii}) - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{jj}) + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}\right] \\
&= -\frac{1}{2n}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)
\end{aligned}
\tag{1.13}
$$

Ainsi, pour se conformer à la définition de la covariance, c_{ij} doit être calculée à partir de distances génétiques au carré bien que Dyer et Nason (2004) utilisent la formule suivante dans le package `popgraph` :

$$c_{ij} = -\frac{1}{2}(d_{ij} - d_{i\bullet} - d_{\bullet j} + d_{\bullet\bullet}) \tag{1.14}$$

La division par n dans l'équation (1.13) n'a aucune influence sur les étapes de calcul ultérieures car la matrice de covariance $\mathbf{C}$ est ensuite standardisée pour obtenir la matrice de corrélation $\mathbf{R}$. Cette matrice de corrélation est inversée pour obtenir la matrice de corrélation inverse $\mathbf{\Omega}$, qui est standardisée à son tour. Les éléments non-diagonaux ω_{ij} de $\mathbf{\Omega}$ sont multipliés par -1 pour obtenir la matrice de corrélation partielle $\mathbf{P}$ tel que (Magwene, 2001) :

$$\rho_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{1.15}$$

Enfin, pour déterminer si les populations i et j sont indépendantes conditionnellement à toutes les autres populations, il faut tester si chaque élément ρ_{ij} est significativement différent de 0. Pour cela, le critère d'*Edge Exclusion Deviance* (EED) est calculé selon la méthode de Whittaker (2009) tel que :

$$EED = -N \ln(1 - \rho_{ij}^2) \tag{1.16}$$

avec N le nombre total d'observations (nombre total d'individus ici, suivant Dyer et Nason (2004)).

Lorsque nous avons adapté cette méthode pour l'appliquer, nous avons considéré qu'un graphe d'indépendance génétique devait inclure des liens entre les populations qui sont corrélées positivement pour représenter des flux génétiques directs entre ces populations. Par conséquent, nous avons converti les éléments négatifs de $\mathbf{P}$ en 0 avant de calculer l'EED, alors que ce n'était pas le cas dans la méthode originale de Dyer et Nason (2004). Sans cela, une corrélation partielle négative conduisait à la même valeur d'EED qu'une corrélation partielle positive de même valeur absolue.

L'EED a une distribution du χ^2 asymptotique à 1 degré de liberté (Whittaker, 2009). Cette propriété permet de tester la significativité de chaque valeur d'EED et donc de tester l'hypothèse nulle $H_0 : \rho_{ij} = 0$ contre l'hypothèse alternative $H_1 : \rho_{ij} \neq 0$ (Figure 1D). Si H_0 est rejetée, alors il y a un

lien entre les populations i et j sur le graphe d'indépendance génétique.

Un seuil de significativité de 0.05 est communément utilisé lors du test de la significativité de l'EED, sans que les p-valeurs ne soient ajustées dans la méthode originale de Dyer et Nason (2004). Cependant, lors de notre utilisation de cette méthode, nous avons ajusté les p-valeurs avec la méthode séquentielle d'Holm (1979) pour limiter le risque d'erreur de type I. En effet, $\frac{p(p-1)}{2}$ tests sont réalisés pour construire un graphe, ce qui augmente ce risque d'erreur et pourrait surestimer le nombre de liens dans le graphe.

N.B. : Bien que nous n'ayons pas inclus ces détails dans cette présentation des bases théoriques de la méthode de construction des graphes génétiques à partir du principe d'indépendance conditionnelle, la méthode `popgraph` telle qu'elle est mise en œuvre dans le package éponyme comporte d'autres différences importantes avec les étapes décrites ici. Par exemple, la matrice des fréquences alléliques initiales est soumise à une ACP et à plusieurs SVD (*Singular Value Decomposition*) sans que l'utilité de ces étapes n'ait pu être élucidée. Par ailleurs, la façon dont les étapes de calcul sont décrites dans l'article de Dyer et Nason (2004) ne coïncide pas avec l'ordre de ces étapes dans le code de la fonction `popgraph`. Une des conséquences majeures de ce détail est que la méthode mise en œuvre par Fortuna *et al.* (2009) en adaptant directement la méthode décrite dans l'article de Dyer et Nason (2004) dans un autre langage (MATLAB), avec quelques autres ajustements, diffère de façon très importante de la méthode `popgraph` et des bases théoriques décrites ici.

Annexe B2

La régression *Partial Least Squares* (PLS-R)

1 Introduction

La régression *Partial Least Squares* (PLS-R) est une généralisation de la régression linéaire multiple (Wold *et al.*, 2001) et constitue une alternative à la régression en composantes principales lorsque les variables explicatives $\mathbf{X}$ sont fortement colinéaires ou corrélées. C'est également une extension de l'Analyse en Composantes Principales (ACP). En effet, elle est également basée sur une analyse factorielle mais contrairement à l'ACP, elle considère simultanément les variables issues de deux matrices $\mathbf{X}$ et $\mathbf{Y}$ pour réaliser cette factorisation (Long, 2013). Elle est souvent utilisée lorsqu'il y a davantage de variables explicatives que d'observations et permet aussi de modéliser simultanément plusieurs variables réponses $\mathbf{Y}$. Dans le chapitre 3 de cette thèse, nous n'avons pas présenté les résultats obtenus en modélisant plusieurs indices génétiques simultanément car ils étaient redondants avec ceux obtenus en les modélisant séparément. Ici, nous décrirons la façon dont la PLS-R s'applique à plusieurs variables réponses dans la mesure où le fait de n'en modéliser qu'une seule n'est qu'un cas particulier de cette approche plus globale.

2 Principe et expression du modèle

L'objectif de la PLS-R, comme de toute régression, est d'obtenir une expression de la forme suivante (Figure 2) :

$$\mathbf{Y} = \mathbf{XB} + \epsilon$$

telle que l'erreur ϵ soit minime. La PLS-R permet cela de façon indirecte dans la mesure où les matrices $\mathbf{X}$ et $\mathbf{Y}$ sont toutes deux soumises à une factorisation dans un premier temps. Elles sont donc exprimées en fonction de facteurs, ou composantes. L'objectif de cette étape, qui fait l'intérêt de la PLS-R, est de maximiser la corrélation entre les facteurs exprimant $\mathbf{X}$ et ceux exprimant $\mathbf{Y}$. C'est en quelque sorte comme si on réalisait une ACP "supervisée" des deux matrices de manière à ce que les composantes de l'une expliquent au mieux celles de l'autre. La solution optimale peut être obtenue en utilisant l'algorithme NIPALS, qui dépasse l'objet de ce document et est décrit par Tenenhaus (1998). Le plus souvent, les variables sont transformées avant les calculs, par exemple en les centrant et en les réduisant, ou en leur appliquant un logarithme.

Soit N le nombre d'individus, K le nombre de variables explicatives et M le nombre de variables réponses, les factorisations de la matrice $\mathbf{X}$ de dimension $N \times K$ et de la matrice $\mathbf{Y}$ de dimension $N \times M$ permettent d'obtenir les relations suivantes :

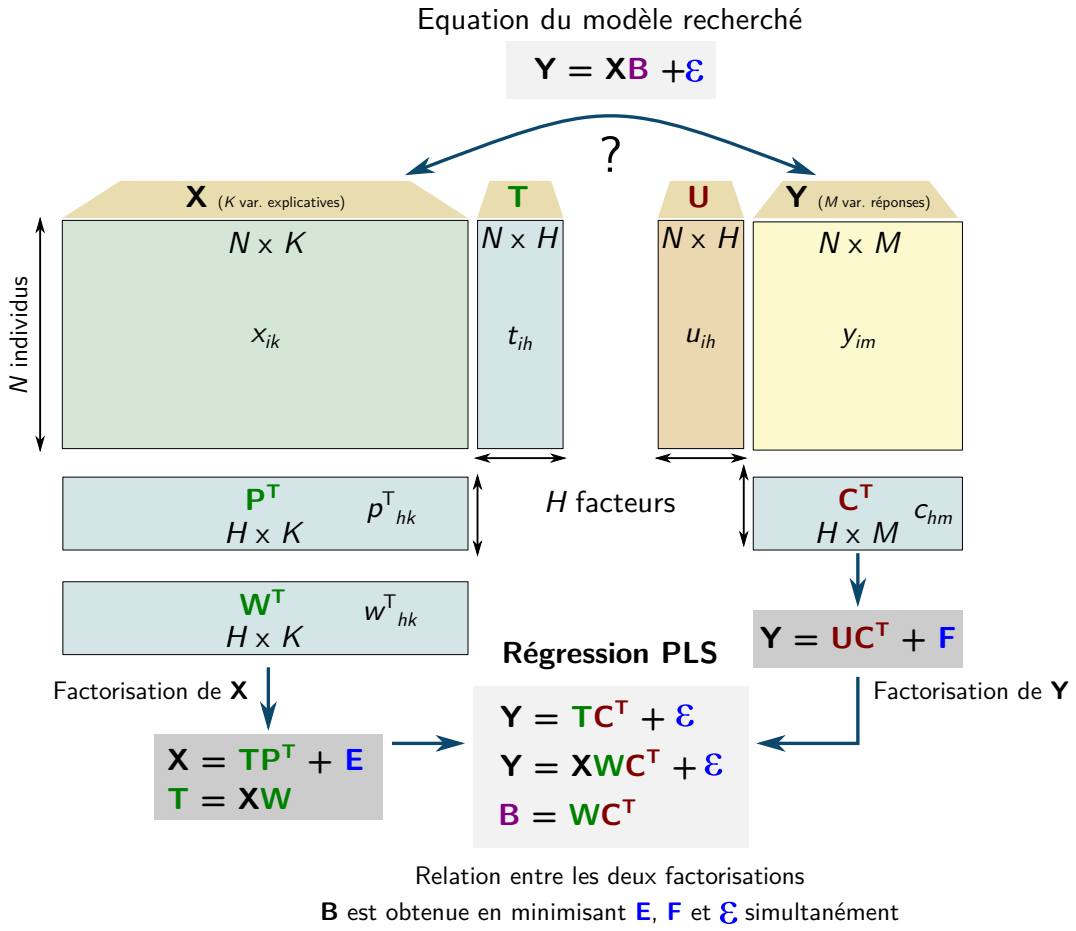


FIGURE 2 – Principe de la régression PLS, adapté à partir de Roy *et al.* (2015)

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{C}^{\mathbf{T}} + \mathbf{F}$$

La matrice $\mathbf{X}$ est donc liée aux variables factorielles de la matrice $\mathbf{T}$ $(N \times H)$ par les coefficients de la matrice $\mathbf{P}^{\mathbf{T}}$ $(H \times K)$ [1] mais ces deux dernières ne permettent pas d'exprimer $\mathbf{X}$ $(N \times K)$ parfaitement dans la mesure où il y a un terme résiduel $\mathbf{E}$ $(N \times K)$ dans la relation. C'est à partir de la matrice $\mathbf{W}$ $(K \times H)$, qui tient compte des résidus, qu'on peut relier $\mathbf{X}$ à $\mathbf{T}$:

$$\mathbf{T} = \mathbf{X}\mathbf{W}$$

Nous verrons par la suite comment le nombre H de facteurs considérés est déterminé.

Ainsi, les variables factorielles $t_1, t_2, \ldots, t_h$ de $\mathbf{T}$ sont des combinaisons linéaires des colonnes initiales de $\mathbf{X}$ et des coefficients de la matrice $\mathbf{W}$ (*weightings*) :

$$t_{ih} = \sum_{k=1}^{K} w_{kh} x_{ik}$$

Les facteurs de $\mathbf{T}$ et $\mathbf{U}$ sont choisis de façon à maximiser leurs corrélations. On a alors :

1. $\mathbf{P}^{\mathbf{T}}$ est la transposée de $\mathbf{P}$.

$$u_1 = r_1 t_1$$

$$u_2 = r_2 t_2$$

$$etc.$$

tels que r_1 et r_2 sont maximisés. De plus, les facteurs obtenus sont orthogonaux par construction (Tobias *et al.*, 1995). Cela explique pourquoi cette méthode constitue une solution lorsque des variables explicatives sont colinéaires ou corrélées.

Ainsi, les éléments des deux factorisations peuvent être mis en relation pour constituer la relation de régression PLS :

$$\mathbf{Y} = \mathbf{T}\mathbf{C}^\mathbf{T} + \epsilon$$

Par construction, cette relation permet de minimiser ϵ. On peut faire apparaître dans cette expression la matrice $\mathbf{X}$ et ainsi comprendre comment la matrice $\mathbf{B}$ a été obtenue :

$$\mathbf{Y} = \mathbf{X}\mathbf{W}\mathbf{C}^\mathbf{T} + \epsilon$$

$$\mathbf{B} = \mathbf{W}\mathbf{C}^\mathbf{T}$$

3 Évaluation de la qualité du modèle et choix du nombre de facteurs considérés

Nous nous placerons dans l'exemple où la matrice $\mathbf{Y}$ ne contient qu'une seule variable y pour présenter le raisonnement adopté dans les PLS-R réalisées au chapitre 3 de cette thèse. D'après les équations précédentes, la variable y est exprimée comme une combinaison linéaire des variables $t_1, t_2, \ldots, t_h$ issues de la matrice $\mathbf{T}$, elle-même liée à la matrice $\mathbf{X}$ par les coefficients de la matrice $\mathbf{W}$. On a donc :

$$\hat{y} = c_1 t_1 + c_2 t_2 + \ldots + c_H t_H$$

avec $c_1, c_2, \ldots, c_H$ les coefficients de la régression issus de la matrice $\mathbf{C}$ et $t_1, t_2, \ldots, t_H$ les facteurs issus de $\mathbf{T}$, tels que :

$$t_1 = w_{11} x_1 + w_{12} x_2 + \ldots + w_{1p} x_p$$

On a donc :

$$\begin{aligned} y = {} & c_1 w_{11} x_1 + c_1 w_{12} x_2 + \ldots + c_1 w_{1p} x_p + \\ & c_2 w_{21} x_1 + c_2 w_{22} x_2 + \ldots + c_2 w_{2p} x_p + \\ & \ldots \\ & c_H w_{H1} x_1 + c_H w_{H2} x_2 + \ldots + c_H w_{Hp} x_p \end{aligned} \tag{2.1}$$

La question du nombre H de facteurs à considérer se pose alors. Pour le déterminer, pour chaque valeur h, un modèle à h facteurs est calculé, soit (i) à partir de toutes les observations, soit (ii) à partir

d'un sous-ensemble d'observations. Une observation peut être laissée de côté (*Leave One Out cross Validation*) ou un bloc entier représentant une proportion $\frac{1}{k}$ des observations (*k-fold cross validation*). À partir de ces modèles, les valeurs de y sont prédites. On obtient :

— $\hat{y}_{hi}$, prédiction de y_i à partir du modèle à h facteurs basé sur toutes les observations,

— $\hat{y}_{h(-i)}$, prédiction de y_i à partir du modèle à h facteurs calibré sans intégrer l'observation i.

La qualité des modèles est évaluée selon deux critères :

$$RSS_h = \sum_{i=1}^{N} (y_i - \hat{y}_{hi})^2$$

et

$$PRESS_h = \sum_{i=1}^{N} (y_i - \hat{y}_{h(-i)})^2$$

qu'on appelle respectivement la somme des carrés résiduels (*Residual Sum of Squares*, RSS) et la somme des carrés des erreurs de prédiction (*PRediction Error Sum of Squares*, PRESS). Considérer un facteur supplémentaire est pertinent si :

$$\sqrt{PRESS_h} \leq 0.95 \times \sqrt{RSS_{h-1}}$$

ce qui signifie qu'en ajoutant ce facteur, l'erreur de prédiction est inférieure à 90.25 % de la valeur de l'erreur du modèle obtenu sans ajouter ce facteur :

$$PRESS_h \leq 0.9025 \times RSS_{h-1}$$

On retrouve :

$$\frac{PRESS_h}{RSS_{h-1}} \leq 0.9025$$

et

$$1 - \frac{PRESS_h}{RSS_{h-1}} \geq 0.0975$$

Le critère Q^2 est égal à :

$$Q^2 = 1 - \frac{PRESS_h}{RSS_{h-1}}$$

La valeur de Q^2 est calculée pour chaque facteur h ajouté à un modèle. On considère qu'un facteur a un effet significatif sur le modèle s'il améliore la prédiction de y et donc d'après ce critère si $Q^2 \geq 0.0975$ (Tenenhaus, 1998).

4 Évaluation de la significativité des effets des variables explicatives

Les valeurs w_{kh} indiquent le poids qu'a chaque variable explicative k sur le facteur h. Elles permettent de comprendre quelles variables explicatives influencent les variables réponses et l'ampleur de cette influence. Ces contributions permettent de projeter les variables explicatives et les variables réponses dans l'espace constitué par les facteurs $t_1, t_2, \ldots, t_H$ pour comprendre les relations entre les variables. Par ailleurs, la significativité des poids w_{kh} des variables explicatives peut être validée par *bootstrap* (Pérez-Rodríguez *et al.*, 2018). Pour cela, le jeu de données est échantillonné aléatoirement et avec remise un grand nombre de fois et à chaque fois les poids w_{kh} sont estimés. Si l'intervalle

contenant 95 % des valeurs (2.5-97.5 %) n'inclut pas 0, alors le poids est significativement différent de 0 et on considère que la variable k a un effet significatif sur le facteur h. Si ce facteur a lui-même un effet significatif sur y, alors on pourra interpréter la relation entre la variable explicative k et la variable réponse y.

Annexe B3

Les modèles de distribution d'espèces

1 Introduction

De manière générale, une espèce s'observe dans un type de milieu particulier, qui lui est propre. C'est de ce constat que découle le concept de niche écologique. Si on confronte les observations de la présence ou de l'absence d'une espèce aux conditions environnementales dans lesquelles on a réalisé ces observations, on peut comprendre quelles sont les conditions nécessaires à la présence de cette espèce. Ces conditions sont identifiées d'une façon d'autant plus fiable et précise que l'on a réalisé un grand nombre d'observations, aussi bien de présences que d'absences, dans des conditions variées et qu'on dispose d'un nombre important de variables environnementales pouvant expliquer la présence de l'espèce. Si on dispose d'une carte de ces conditions environnementales à l'échelle de tout un territoire, on peut cartographier les zones au niveau desquelles la présence de l'espèce sera la plus probable à partir des relations identifiées.

Un modèle de distribution d'espèce (*Species Distribution Model*[1], noté SDM par la suite) est précisément un modèle qui relie la présence ou l'absence d'une espèce à des variables environnementales. Il est fréquemment utilisé aujourd'hui car les données d'observation sont de plus en plus nombreuses, les données environnementales (spatiales) de plus en plus accessibles et de nombreuses méthodes de modélisation permettent de s'adapter à chaque contexte (Guisan et Zimmermann, 2000). Leur utilisation permet de prédire l'effet du changement climatique sur la distribution des espèces ou l'expansion d'espèces invasives, entre autres (Fletcher et Fortin, 2018).

Le choix de la méthode de modélisation dépend essentiellement (i) des données d'observation dont on dispose (présence *vs* présence/absence *vs* abondance) et de la forme de la relation existant entre les variables environnementales et les observations (linéaire, non-linéaire, etc.).

Les données de présence sont très nombreuses aujourd'hui, notamment grâce aux programmes de science participative. On peut se contenter de ces observations pour modéliser la distribution d'une espèce avec les méthodes les plus simples. Néanmoins, il est préférable de confronter les conditions environnementales de ces points à d'autres conditions environnementales pour comprendre ce qui détermine la présence de l'espèce. Pour cela, on peut générer aléatoirement des points de *background* (ou

1. On parle aussi d'*Ecological Niche Model* (ENM) ou d'*Habitat Suitability Model* (HSM). Nous retenons ici le terme de SDM car c'est la distribution spatiale d'une espèce dans un contexte donné qui est modélisée. On ne peut pas savoir si cela coïncide véritablement avec sa niche écologique fondamentale ou réalisée, d'autant plus que le concept de niche en lui-même fait débat. Nous éviterons donc ce terme.

de "pseudo-absence"). L'algorithme Maxent est largement utilisé dans ce cas de figure (Elith *et al.*, 2011 ; Phillips *et al.*, 2006).

Sans données concernant l'absence de l'espèce, on ne peut pas connaître la prévalence de l'espèce. En d'autres termes, on ne sait pas quel est le ratio entre le nombre de points de présence et le nombre de points d'absence. Les SDM ne fournissent donc dans ce cas qu'une probabilité de présence relative, la véritable probabilité de présence de l'espèce étant proportionnelle à sa prévalence. De plus, lorsqu'on ne dispose pas de points d'absence, les potentiels biais liés à la distribution spatiale des observations sont difficilement évitables. Prenons l'exemple où les observateurs ne s'éloignent jamais de plus de 100 m des routes pour faciliter leurs déplacements sur le terrain. Si les points de présence sont confrontés à des points de pseudo-absence choisis aléatoirement dans toute la zone d'étude, il est probable que la proximité à la route soit identifiée comme un facteur favorable à la présence de l'espèce. Si à la fois les points de présence et d'absence sont situés à moins de 100 m des routes, ce qui est le cas si les observateurs relèvent également l'absence de l'espèce, ce biais sera évité.

Pour ces différentes raisons, lorsqu'on dispose de données de présence/absence, leur analyse à l'aide d'un modèle approprié est préférable à l'utilisation de l'algorithme Maxent à partir des points de présence uniquement (Guillera-Arroita *et al.*, 2014). Lorsqu'on dispose de ces données et qu'il y a une relation monotone entre les variables environnementales et la présence de l'espèce, la régression logistique est une des méthodes les plus utilisées. C'est celle que nous avons choisie pour réaliser le SDM de la Paruline caféiette dans le cadre de cette thèse et que nous décrivons par la suite. Nous mentionnons à titre indicatif l'existence des modèles suivants :

— *Envelope models* (BIOCLIM par exemple)

— *Generalised Additive Models* (GAM) : adaptés aux relations non-linéaires et non-monotones

— *Classification Trees* ou *Random Forests* : adaptés aux relations non-linéaires et non-monotones et à la prise en compte de variables quantitatives et qualitatives simultanément

— Voir Fletcher et Fortin (2018) et Guisan *et al.* (2017) pour une description de ces méthodes et d'autres.

2 Création d'un SDM à partir d'une régression logistique

La régression logistique fait partie de la grande famille des *Generalised Linear Models* (GLM). Elle permet de modéliser une variable réponse y binaire (0, 1) à partir d'un ensemble de p variables quantitatives et/ou qualitatives $x_1, x_2, \ldots, x_p$ renseignant sur les conditions environnementales au niveau de chaque point d'observation. Plus précisément, l'objectif est de modéliser la probabilité de présence sachant les conditions environnementales $P(y = 1|x_1, x_2, \ldots, x_p) = p(x)$ à partir des conditions environnementales $x_1, x_2, \ldots, x_p$. Par définition, $p(x)$ est bornée par 0 et 1. L'utilisation d'une relation de la forme $p(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ serait inadéquate car elle pourrait dépasser ces bornes. Une solution consiste à exprimer $p(x)$ de la façon suivante :

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Ce ratio est compris entre 0 et 1. La variation de $p(x)$ en réponse à l'augmentation de x_1, si elle ne dépend que de x_1, a une forme "en S". En transformant quelque peu cette relation, et en lui appliquant un logarithme, on tombe sur les expressions suivantes :

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

$$\text{logit}(p(x)) = \log(\frac{p(x)}{1-p(x)}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

La fonction $\log(\frac{x}{1-x})$ s'appelle la fonction "logit". $\text{logit}(p(x))$ prend des valeurs continues dépassant l'intervalle $[0, 1]$ quand $p(x)$ est comprise entre 0 et 1. On va modéliser $\text{logit}(p(x))$ en fonction des variables $x_1, x_2, \ldots, x_p$. On retombe donc sur l'expression d'un modèle de régression linéaire multiple. C'est le principe des modèles linéaires généralisés (GLM). Ici, la "fonction de lien" permettant de retrouver un modèle de cette forme est la fonction "logit".

À l'aide d'une méthode le plus souvent basée sur la maximisation de la vraisemblance du modèle, on estime un ensemble de coefficients $\beta_0, \beta_1, \ldots, \beta_p$ qui relient $\text{logit}(p(x))$ aux variables $x_1, x_2, \ldots, x_p$. Suite à l'estimation du modèle , on peut calculer la valeur $\text{logit}(p(x))$ puis $p(x)$ pour l'ensemble des observations, et étendre ce calcul à l'ensemble des points pour lesquels on dispose des valeurs des variables $x_1, x_2, \ldots, x_p$. C'est cela qui permet de cartographier la distribution de l'espèce.

Notons qu'en amont de la modélisation, il est particulièrement important de s'assurer que :

— Il n'y a pas de biais dans la distribution spatiale des points considérés.

— Les variables explicatives considérées ne sont pas fortement corrélées ou colinéaires.

3 Qualité du modèle et choix d'un seuil de prédiction

À l'issue de la modélisation, on peut se poser deux questions :

— Le modèle obtenu explique-t-il bien la distribution des observations de présence et d'absence ?

— À partir de quelle probabilité de présence peut-on prédire que l'espèce sera présente ?

Nous allons voir que ces deux questions sont liées et y répondre.

Observation/Prédiction	**Présences prédites**	**Absences prédites**
Présences observées	Vrais positifs (TP)	Faux négatifs (FN)
Absences observées	Faux positifs (TP)	Vrais négatifs (TN)

Table 1 – Matrice de confusion

On dispose pour chaque point d'une prédiction de probabilité de présence comprise entre 0 et 1 et suivant un gradient continu alors que la variable y initiale était binaire (0, 1). Pour confronter les prédictions aux observations et évaluer la qualité du modèle, il est donc nécessaire de les "binariser". On choisit pour cela une valeur seuil et si la probabilité de présence est supérieure à ce seuil, on prédit la présence de l'espèce ($\hat{y} = 1$). Pour un seuil donné, on obtient deux séries de 0 et de 1, l'une observée et l'autre prédite. Avec un modèle parfait, ces deux séries de valeurs seraient identiques. Mais, c'est très rare en pratique. Un modèle peut prédire la présence d'une espèce là où elle était absente, ou au

contraire son absence là où elle était présente. Il s'agit d'un faux positif et d'un faux négatif, respectivement. La qualité globale du modèle dépendra du nombre d'erreurs de chaque type auxquelles il conduit. Pour quantifier cela, on réalise une matrice de confusion qui contient les nombres de vrais positifs (TP), faux négatifs (FN), faux positifs (FP) et vrais négatifs (TN) (Table 1).

À partir de la matrice de confusion, on peut calculer différents indices pour caractériser la performance du modèle. Parmi eux :

$$\text{Spécificité } = \frac{TN}{TN + FP}$$

c'est-à-dire le taux de prédiction d'une absence lorsqu'on a observé une absence. On retrouve :

$$1 - \text{ Spécificité } = \frac{FP}{TN + FP} = \text{ Taux de faux positifs (FDR)}$$

On calcule également :

$$\text{Sensibilité } = \frac{TP}{FN + TP} = \text{ Taux de vrais positifs (TPR)}$$

c'est-à-dire le taux de prédiction d'une présence lorsqu'on a observé une présence (taux de vrais positifs, TPR).

La valeur de ces indices calculés en "binarisant" les probabilités de présence avec une valeur seuil ne nous renseigne que sur la qualité de la prédiction pour ce seuil particulier. On ne connaît donc pas la qualité du modèle global ni le seuil optimal. Pour cela, on va répéter l'opération en choisissant itérativement des valeurs seuils comprises entre 0 et 1. On obtient une série de valeurs de FDR et de TPR, dont on représente la relation (Figure 3). La courbe obtenue s'appelle la courbe ROC (*Receiver Operating Characteristic*). Plus l'aire sous cette courbe est proche de 1, sa valeur maximale, meilleur est le modèle. Cette aire correspond à l'indice AUC (*Area Under the Curve*), souvent utilisé pour mesurer la qualité d'un SDM. Il est par construction indépendant du seuil choisi.

Pour identifier le seuil optimal au-delà duquel on peut prédire la présence de l'espèce, on utilise également l'indice de corrélation de Matthews (1975) car il tient compte de toutes les valeurs de la matrice de confusion :

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Il varie entre -1 et 1. Son utilisation est plus fréquente en informatique et en *machine learning* qu'en écologie (Baldi *et al.*, 2000). La valeur seuil maximisant cet indice (ou un autre) pourra être utilisée pour cartographier les zones d'habitat de l'espèce étudiée, comme nous l'avons fait pour la Paruline caféiette au chapitre 4 de cette thèse.

Comme tout modèle statistique, un SDM peut être sur-paramétré (*overfitting*). Il expliquera très bien les données utilisées pour le calibrer mais aura une faible valeur prédictive si on l'applique dans un autre contexte. Pour éviter cela, on peut calculer ces mêmes indices en considérant un ensemble d'observations non incluses lors de la calibration du modèle (données de validation). C'est en faisant cela que nous avons pu nous assurer que les coefficients associés aux variables environnementales n'étaient

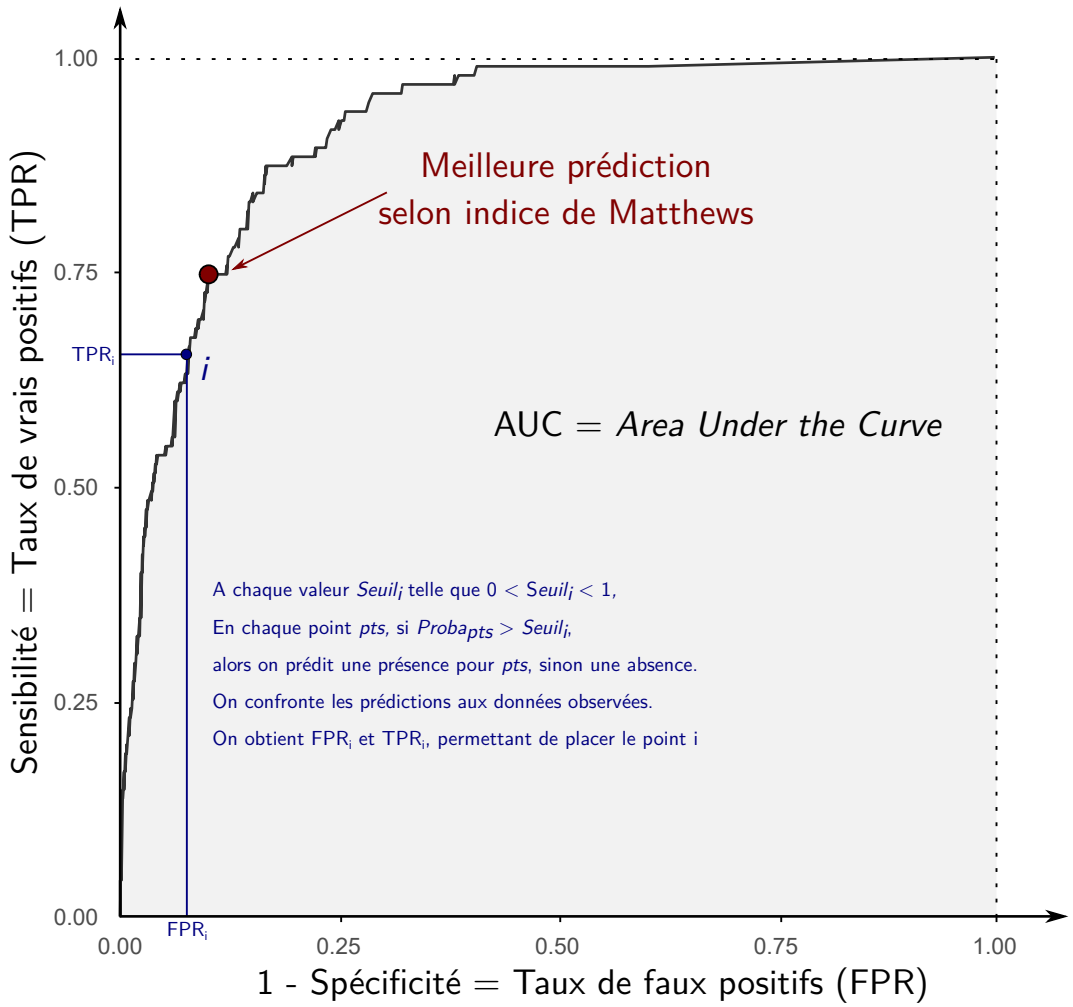


FIGURE 3 – Évaluation de la qualité d'un modèle de distribution d'espèce à l'aide de la courbe ROC (*Receiver Operating Characteristic*) et de l'AUC (*Area Under the Curve*). Le modèle fournit pour chaque point d'observation une estimation de probabilité de présence $Proba_{pts}$ entre 0 et 1. En choisissant itérativement des valeurs $Seuil_i$, on considère que si $Proba_{pts} > Seuil_i$, alors le point est classé comme un point de présence, et sinon comme une absence. On confronte ces résultats aux observations et on calcule les taux de vrais positifs (TPR) et de faux positifs (FPR) correspondants. Cela permet de placer le point i correspondant au $Seuil_i$. L'ensemble de ces points forment la courbe ROC. L'aire sous la courbe vaut au maximum 1 et plus elle se rapproche de 1, meilleur est le modèle.

pas biaisés par le jeu de données considéré dans le chapitre 4, et mettre en évidence l'importance du ré-échantillonnage des points d'observation de Basse-Terre.

Enfin, pour limiter le risque d'*overfitting* lors de la calibration du modèle, il est possible d'appliquer une régularisation aux coefficients du modèle. Pour ce faire, le package `glmnet` (Friedman *et al.*, 2021) permet d'utiliser la méthode *elastic net*, intermédiaire entre les méthodes *lasso* et *ridge* (Hastie *et al.*, 2017). Nous l'avons utilisée pour réaliser le SDM de la Paruline caféiette mais (i) l'amélioration des performances du modèle était négligeable et (ii) son utilisation ne reflétait pas la façon dont les SDM sont classiquement construits en amont de la construction d'un graphe paysager. Nous n'avons donc pas régularisé les coefficients de ce SDM.

Annexe B4

Les arbres de régression

1 Introduction

Au sein des méthodes statistiques basées sur les arbres, on distingue les arbres de régression et les arbres de classification. Tous deux sont parfois appelés "arbres de décision". Les premiers constituent des alternatives aux méthodes de régression, linéaires ou non, utilisées pour modéliser des variables réponses quantitatives. Les seconds constituent des alternatives aux méthodes de modélisation de variables qualitatives, telles que la régression logistique binaire ou multinomiale ou l'analyse discriminante, entre autres.

Nous avons modélisé des variables quantitatives avec des arbres de régression dans les chapitres 5 et 6 de cette thèse. Nous décrirons donc les bases de cette méthode. Nous n'aborderons pas le *bagging*, le *boosting* et les *random forests*, extensions de ces méthodes basées sur des approches de *machine learning*. Elles sont davantage adaptées à la prédiction des valeurs de la variable réponse qu'à l'interprétation de l'influence qu'ont des variables explicatives sur cette variable, ce qui ne correspondait pas aux objectifs poursuivis dans les chapitres 5 et 6. Ces méthodes sont décrites dans l'ouvrage de James *et al.* (2013), sur lequel nous nous basons dans les sections qui suivent.

2 Principe de construction d'un arbre de régression

Dans un modèle de régression linéaire, le lien entre la variable réponse Y et une variable explicative X_1 est de la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

avec β_0 et β_1 deux constantes et ϵ les résidus non expliqués par le modèle, égaux à $Y - \hat{Y}$. Le modèle se base donc sur une relation de proportionnalité entre Y et X_1.

Le principe d'un arbre de régression est très différent dans la mesure où il ne repose pas sur une relation de proportionnalité entre les variables. En effet, l'espace constitué par les variables explicatives $X_1, X_2, \ldots, X_p$ va être scindé en J régions $(R_1, R_2, \ldots, R_J)$ distinctes et non-chevauchantes en fonction des valeurs de ces différentes variables. Chaque région R_j constituera une feuille de l'arbre. La valeur $\hat{Y}_{R_j}$ prédite par le modèle pour les observations appartenant à la région R_j sera simplement égale à la moyenne des valeurs de Y dans cette région. Les critères permettant de scinder l'espace des variables $X_1, X_2, \ldots, X_p$ en J régions ou feuilles seront représentés sous la forme d'un arbre, rendant

le modèle facilement interprétable (Figure 4).

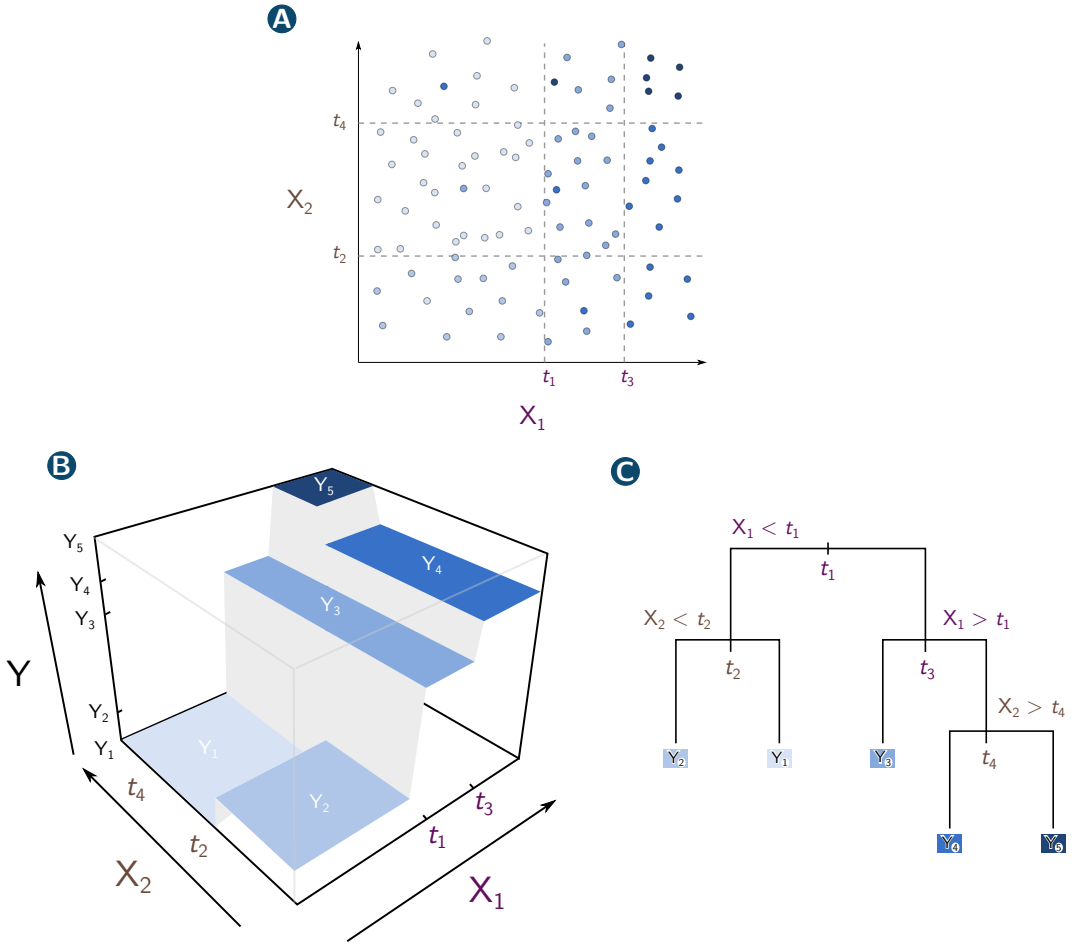


FIGURE 4 – Principe des arbres de régression, illustration adaptée à partir de Hastie *et al.* (2017). (A) Données d'entrée. Plus la couleur des points est foncée, plus la variable réponse Y prend une valeur forte. Le lien entre Y et les variables explicatives X_1 et X_2 n'est pas linéaire, ce qui justifie l'utilisation d'un arbre de régression pour modéliser Y. (B) Séparation optimale de l'espace formé par les variables Y, X_1 et X_2 en plusieurs régions (feuilles de l'arbre) au niveau desquelles Y prend des valeurs relativement homogènes en fonction des classes de valeurs respectives des variables X_1 et X_2. (C) Arbre de régression représentant la séparation optimale en feuilles en fonction des classes de valeurs des variables X_1 et X_2. En bas de l'arbre, on trouve pour chaque feuille la valeur moyenne de Y pour les observations appartenant à la feuille. Ces valeurs se retrouvent également en (B). Chaque séparation d'une branche de l'arbre en deux branches sépare les observations selon la classe à laquelle appartient la valeur qu'elles prennent pour les variables X_1 ou X_2.

La qualité du modèle s'évalue à partir de la somme des carrés des erreurs (SCE), calculée de la façon suivante :

$$SCE = \sum_{j=1}^{J} \sum_{i \in R_j} (Y_i - \hat{Y}_{R_j})^2$$

avec Y_i la valeur de Y pour l'observation i, $\hat{Y}_{R_j}$ la valeur de Y prédite pour l'observation i appartenant à la région R_j, c'est-à-dire la moyenne de Y au sein de cette région. Le meilleur modèle est identifié en cherchant les régions $R_1, R_2, \ldots, R_J$ qui permettent de minimiser la SCE.

Comme il y a une infinité de façons de scinder l'espace des p variables explicatives en J régions distinctes et non chevauchantes, ces régions vont être identifiées de façon récursive et binaire (*recursive binary splitting*). L'arbre est représenté la tête en bas (feuilles en bas) et construit de haut en bas,

donc du tronc (au niveau duquel toutes les observations appartiennent à la même région unique) vers les feuilles. Les branches partant du tronc se séparent au niveau de nœuds internes en d'autres branches jusqu'aux feuilles terminales. Chaque nœud interne donne naissance à deux régions selon un critère basé sur une des variables explicatives X. Par exemple, le critère donnant naissance aux deux premières régions R_1 et R_2 dépend de la variable X_j et d'une valeur seuil s :

$$R_1(j,s) = \{X|X_j < s\} \text{ et } R_2(j,s) = \{X|X_j \geq s\}$$

ce qui signifie qu'une observation i appartient à R_1 si $X_{ij} < s$ et à R_2 sinon. Pour construire l'arbre, on cherche les valeurs j et s qui minimisent la valeur suivante :

$$\sum_{i:X_i \in R_1(j,s)} (Y_i - \hat{Y}_{R_1})^2 + \sum_{i:X_i \in R_2(j,s)} (Y_i - \hat{Y}_{R_2})^2$$

Le résultat est obtenu assez rapidement. Cette étape est répétée, ce qui scinde à nouveau une région en deux nouvelles régions de façon à faire diminuer la valeur totale de SCE. L'arbre est construit ainsi. La figure 4 illustre le résultat que l'on peut obtenir à l'issue de ces étapes.

3 Élagage d'un arbre de régression

En pratique, il est toujours possible de créer de nouvelles régions et d'obtenir un arbre avec plus de branches et de feuilles terminales. Cela a plusieurs inconvénients. Tout d'abord, l'arbre obtenu est difficile à interpréter. De plus, les feuilles terminales incluent peu d'observations ce qui peut conduire à sur-interpréter les résultats obtenus. Enfin, l'arbre a une valeur prédictive faible s'il est trop complexe car dans ce cas, il est sur-paramétré (*overfitting*). Il existe plusieurs façons d'éviter cela.

Tout d'abord, on peut arrêter la construction de l'arbre lorsque la diminution de la SCE résultant de la création de nouvelles régions devient très faible. L'inconvénient est que même si un ajout donné fait peu diminuer la SCE, cela ne signifie pas que le suivant ne la fera pas diminuer davantage. Pour cette raison, on construit en général un très grand arbre T_0 qui est ensuite élagué pour réduire son nombre de branches et de feuilles. La construction de T_0 s'arrête lorsqu'une feuille contient un nombre d'observations inférieur à un certain seuil. Augmenter cette valeur seuil est d'ailleurs en soi une façon de réduire la taille de l'arbre. C'est ainsi que nous avons procédé au chapitre 6, car nous voulions qu'il y ait au minimum 40 observations par feuille pour pouvoir tester la significativité de la différence de ces valeurs par rapport à 0.

Pour élaguer T_0, on pourrait faire une validation croisée en scindant les données en deux jeux de données, de calibration d'une part et de validation d'autre part. On estimerait alors l'erreur de prédiction des données de validation associée à chaque sous-arbre. Néanmoins, cette approche serait extrêmement longue. Il existe une alternative basée sur un principe de "coût complexité" (*cost complexity pruning*). Nous l'avons utilisée au chapitre 5. Elle consiste à évaluer la qualité des sous-arbres en calculant leur SCE et en la pénalisant par le nombre de feuilles qu'ils contiennent à l'aide d'un paramètre α. On cherche ainsi à minimiser pour chaque sous-arbre $T \subset T_0$:

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha|T|$$

avec $|T|$ le nombre de feuilles du sous-arbre T, R_m la région correspondant à la $m^{\text{ième}}$ feuille terminale et $\hat{y}_{R_m}$ la valeur prédite par le sous-arbre pour la feuille R_m. On considère donc chaque observation i et la feuille R_m correspondante. Si $\alpha = 0$, c'est l'arbre T_0 qui permet de minimiser cette somme. À mesure que α augmente, ce sont des arbres plus petits qui permettent de la minimiser puisque le nombre de feuilles $|T|$ est pénalisé plus fortement. L'avantage de cette méthode est que lorsque α augmente, les branches sont élaguées d'une façon prévisible et les sous-arbres sont emboîtés. On peut donc facilement obtenir une séquence de sous-arbres en fonction de α. On se base alors sur une validation croisée pour déterminer la valeur α optimale et on identifie le sous-arbre correspondant.

Annexe B5

Les modèles gravitaires

1 Introduction

Les modèles gravitaires, traduction française des *gravity models*, s'inspirent de la loi de la gravitation universelle de Newton. Selon celle-ci, la force d'attraction F_{ij} qui s'exerce entre deux planètes i et j de masses m_i et m_j séparées par une distance d_{ij} a pour norme (en Newton) (Figure 5A) :

$$F_{ij} = F_{ji} = G \times \frac{m_i m_j}{d_{ij}^2}$$

tel que G est la constante universelle de gravitation.

Cette formule peut également s'exprimer de la façon suivante :

$$F_{ij} = F_{ji} = G \times m_i^o \times m_j^p \times d_{ij}^{2q}$$

avec $o = p = 1$ et $q = -1$.

Les modèles gravitaires s'inspirent de cette loi pour modéliser un flux ou l'intensité d'une interaction entre deux entités spatialement distinctes en fonction de la distance entre elles et des caractéristiques propres à chacune des entités. L'objectif est alors de déterminer les valeurs des exposants (o, p, q dans l'exemple). Ils ont été utilisés dans de nombreux domaines, en économie (marketing spatial, flux monétaires) ou en géographie (estimation de flux de migrants, réseaux de transport) en particulier. Ils n'ont été utilisés en écologie qu'après 2000, avec des applications à la dispersion de maladies (Ferrari *et al.*, 2006 ; Xia *et al.*, 2004) et à la connectivité (Bossenbroek *et al.*, 2001, 2007 ; Kong *et al.*, 2010).

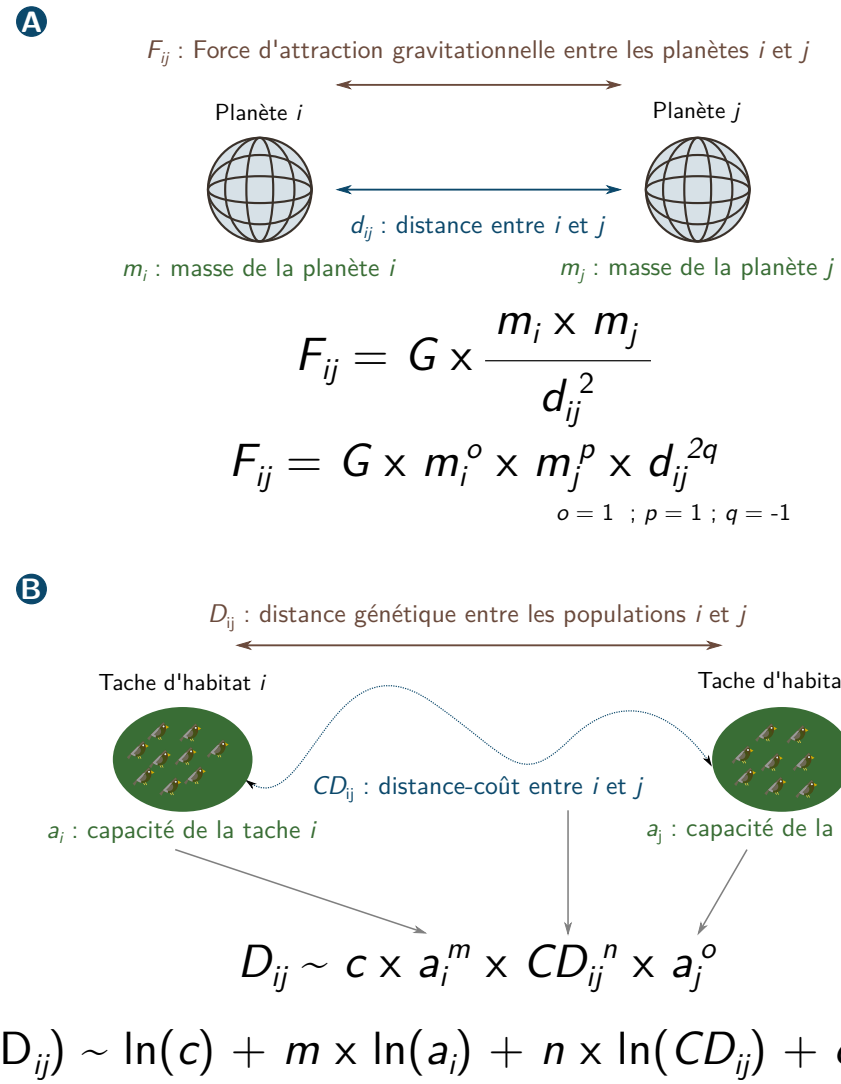


FIGURE 5 – Principe de l'utilisation des modèles gravitaires en génétique du paysage. (A) Loi de la gravitation universelle de Newton, permettant de calculer la force d'attraction gravitationnelle qui s'exerce entre deux planètes i et j en fonction de leurs masses respectives, de la distance qui les sépare et de la constante de gravitation G. (B) Application d'une loi d'interaction similaire à la loi de la gravitation pour expliquer la distance génétique entre deux populations i et j à partir des capacités de leurs taches d'habitat respectives et de la distance-coût qui les sépare. Le modèle est d'abord exprimé de façon similaire à la seconde notation de la loi de la gravitation universelle (A), puis en lui appliquant un logarithme, pour passer d'un modèle multiplicatif à un modèle additif ayant la forme d'une régression linéaire multiple.

2 Utilisation des modèles gravitaires en génétique du paysage

Les travaux de Murphy et al. (2010)

Murphy *et al.* (2010) ont introduit ce type de modèle en génétique du paysage dans le cadre d'une étude des déterminants de la structure génétique de populations de grenouilles (*Rana luteiventris*) occupant un réseau de lacs d'altitude. Ces auteurs ont montré grâce à ces modèles que la différenciation génétique entre les populations était influencée positivement par la distance entre les lacs, la présence de poissons dans les lacs et la topographie du paysage qui les sépare. Pour cela, ils ont réalisé un modèle de la forme suivante :

$$T_{ij} \sim k v_i^{\mu} w_j^{\alpha} c_{ij}^{-\beta}$$

tel que :

— T_{ij} est une estimation du flux génétique ($1-$ D_{PS}) entre les populations i et j.

— c_{ij} est une variable de résistance au flux entre les populations. Cette variable pouvait tenir compte de l'*elevation ratio relief*, du ratio entre surfaces de prairies et de forêts, entre autres. Elle est calculée dans une zone tampon (*buffer*) de largeur fixe autour du chemin reliant les populations en ligne droite.

— w_j est la distance géodésique euclidienne entre les sites.

— v_i est la variable caractérisant la population d'origine. Les variables prises en compte étaient notamment l'altitude ou la présence de poissons.

— k, α, μ et β sont les paramètres du modèle à estimer.

Cette expression du modèle n'est qu'un cas particulier car il en existe plusieurs (Anderson, 1979 ; Fotheringham et O'Kelly, 1989).

Utilisations ultérieures et généralisation de l'approche

Lors des utilisations de modèles gravitaires qui ont fait suite à celle de Murphy *et al.* (2010) en génétique du paysage (DiLeo *et al.*, 2014 ; Watts *et al.*, 2015 ; Zero *et al.*, 2017), d'autres expressions ont été utilisées. Ce qui les distingue en particulier est l'inclusion de variables relatives aux populations d'origine ou de destination, ou aux deux. Dans le cadre de cette thèse, nous avons inclus des variables relatives aux deux populations/taches d'habitat entre lesquelles nous avons modélisé la différenciation génétique, à l'instar de la loi de la gravitation universelle (Figure 5B). Notre modèle était donc de la forme suivante :

$$D_{ij} \sim c \times a_i^m \times CD_{ij}^n \times a_j^o$$

avec D_{ij} la distance génétique entre les populations i et j, c, m, n et o des constantes, a_i la capacité (ou l'effectif) de la population i et CD_{ij} la distance-coût entre les taches i et j occupées par les populations i et j. Notre modèle différait donc également de celui de Murphy *et al.* (2010) par le fait qu'au lieu d'inclure une variable relative à la distance géodésique euclidienne entre les populations et une autre relative à la résistance au flux entre ces populations, nous n'avons inclus qu'une seule variable représentant la distance-coût entre les populations. Cela se justifiait par le fait que nous utilisions les modèles gravitaires dans l'objectif d'optimiser des scénarios de coût. Par ailleurs, nous avons modélisé la distance D_{ij} estimée à l'aide du D_{PS} tandis que Murphy *et al.* (2010) modélisaient la variable T_{ij} égale à $1-$ D_{PS}.

L'estimation des paramètres du modèle ne peut pas se faire à partir des expressions des modèles multiplicatifs. Le produit est converti en une somme par transformation logarithmique. Dans le cas du modèle de Murphy *et al.* (2010) et dans notre cas, on obtient alors respectivement (Figure 5B) :

$$\ln T_{ij} \sim \ln k + \mu \ln v_i + \alpha \ln w_j - \beta \ln c_{ij}$$

et

$$\ln D_{ij} \sim \ln c + m \ln a_i + n \ln CD_{ij} + o \ln a_j$$

Ces modèles s'apparentent alors à des régressions linéaires multiples et peuvent être estimés à l'aide de la méthode des moindres carrés ordinaires (*Ordinary Least-Squares*) par exemple. Néanmoins, ils ne respectent pas la condition d'indépendance des observations car plusieurs flux et distances sont

relatifs à une même population. Une des solutions apportées à ce problème dans le cadre de l'utilisation de modèles gravitaires consiste à faire varier la constante k (ou c) en fonction de la population d'origine ou de destination. Le modèle est alors dit *singly constrained* (simplement contraint). C'est la solution utilisée par Murphy *et al.* (2010). Ces auteurs mentionnent le fait que le modèle est parfois non contraint (*unconstrained*) ou doublement contraint (*doubly constrained*), et les avantages et inconvénients associés à chaque option (Murphy *et al.*, 2010). Ces différentes options permettent également de fixer la somme totale des flux dans le cadre du paramétrage du modèle gravitaire. Ce point dépasse notre application de ces modèles. Nous renvoyons vers Bossenbroek *et al.* (2001) pour une méthode plus détaillée intégrant un contrôle des flux totaux.

Bien que nous n'ayons pas utilisé cette terminologie, la méthode que nous avons utilisée permet également de tenir compte de la non-indépendance des observations en estimant une valeur c différente selon une des deux populations impliquées. En effet, nous avons calibré les modèles gravitaires en utilisant des modèles linéaires mixtes de type MLPE (Clarke *et al.*, 2002 ; Van Strien *et al.*, 2012). Nous avons choisi ce type de modèle pour les raisons suivantes :

— Des travaux ont montré qu'ils étaient adaptés à la modélisation de distances génétiques (Shirk *et al.*, 2017).

— Ils intègrent un effet aléatoire relatif à une des deux populations de chaque paire. Ainsi, une valeur c est calculée pour chaque population (*random intercept model*), comme dans l'application de Murphy *et al.* (2010).

— Ils rendent possible l'élagage des matrices de distances génétiques, à condition de faire permuter certaines populations pour que l'estimation de la covariance des résidus soit possible [1].

Contrairement à Murphy *et al.* (2010), nous n'avons pas comparé les modèles selon un critère d'AIC mais en utilisant le R^2_β développé par Edwards *et al.* (2008) pour l'évaluation de l'ajustement des modèles mixtes. Cela vient en partie du fait que cet indicateur permet de comparer des modèles n'incluant pas le même nombre d'observations, ce qui a justifié son choix initialement. Finalement, les modèles comparés incluaient le même nombre d'observations et de variables. Pour cette raison, l'utilisation de l'AIC, critère d'ajustement pénalisé par la complexité du modèle, plutôt que le R^2_β ne se serait pas justifiée.

Ainsi, les modifications de la méthode de Murphy *et al.* (2010) que nous avons réalisées ne sont pas censées affecter la validité de nos modèles. Dans les deux cas, les modèles prenaient la forme d'une régression linéaire multiple intégrant des effets aléatoires de type *random intercept*.

1. Nous avons discuté de ce point avec Maarten van Strien, à l'origine de l'utilisation de ce type de modèle en génétique du paysage. Il nous a fourni un script permettant leur utilisation pour calibrer des modèles gravitaires à partir de matrices élaguées.

Bibliographie

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review*, 69(1):106–116.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. et Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics*, 16(5):412–424.

Bossenbroek, J. M., Johnson, L. E., Peters, B. et Lodge, D. M. (2007). Forecasting the expansion of zebra mussels in the United States. *Conservation Biology*, 21(3):800–810.

Bossenbroek, J. M., Kraft, C. E. et Nekola, J. C. (2001). Prediction of long-distance dispersal using gravity models : zebra mussel invasion of inland lakes. *Ecological Applications*, 11(6):1778–1788.

Clarke, R. T., Rothery, P. et Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices : estimating gene flow with distance. *Journal of agricultural biological and environmental statistics*, 7(3):361–372.

DiLeo, M. F., Siu, J. C., Rhodes, M. K., López-Villalobos, A., Redwine, A., Ksiazek, K. et Dyer, R. J. (2014). The gravity of pollination : integrating at-site features into spatial analysis of contemporary pollen movement. *Molecular Ecology*, 23(16):3973–3982.

Dyer, R. J. et Nason, J. D. (2004). Population graphs : the graph theoretic shape of genetic structure. *Molecular Ecology*, 13(7):1713–1727.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F. et Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, 27(29):6137–6157.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E. et Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17(1):43–57.

Everitt, B. et Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Ferrari, M. J., Bjørnstad, O. N., Partain, J. L. et Antonovics, J. (2006). A gravity model for the spread of a pollinator-borne plant pathogen. *The American Naturalist*, 168(3):294–303.

Fletcher, R. et Fortin, M.-J. (2018). *Spatial ecology and conservation modeling*. Springer.

Fortuna, M. A., Albaladejo, R. G., Fernández, L., Aparicio, A. et Bascompte, J. (2009). Networks of spatial genetic variation across species. *Proceedings of the National Academy of Sciences*, 106(45):19044–19049.

Fotheringham, A. S. et O'Kelly, M. E. (1989). *Spatial interaction models : formulations and applications*, volume 1. Kluwer Academic Publishers Dordrecht.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N. et Qian, J. (2021). Package 'glmnet'. *CRAN R Repositary*.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Guillera-Arroita, G., Lahoz-Monfort, J. J. et Elith, J. (2014). Maxent is not a presence–absence method : a comment on Thibaud et al. *Methods in Ecology and Evolution*, 5(11):1192–1197.

Guisan, A., Thuiller, W. et Zimmermann, N. E. (2017). *Habitat suitability and distribution models : with applications in R*. Cambridge University Press.

Guisan, A. et Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186.

Hastie, T., Tibshirani, R. et Friedman, J. (2017). *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, volume 2. Springer.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70.

James, G., Witten, D., Hastie, T. et Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kong, F., Yin, H., Nakagoshi, N. et Zong, Y. (2010). Urban green space network development for biodiversity conservation : Identification based on graph theory and gravity modeling. *Landscape and Urban Planning*, 95(1-2):16–27.

Krzanowski, W. et Marriott, F. (1995). *Multivariate Analysis vol. 2 : Classification, Covariance Structures, and Repeated Measurements*. London : Arnold.

Long, F. H. (2013). Multivariate analysis for metabolomics and proteomics data. *In Proteomic and Metabolomic Approaches to Biomarker Discovery*, pages 299–311. Elsevier.

Magwene, P. M. (2001). New tools for studying integration and modularity. *Evolution*, 55(9):1734–1745.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Murphy, M. A., Dezzani, R., Pilliod, D. S. et Storfer, A. (2010). Landscape genetics of high mountain frog metapopulations. *Molecular Ecology*, 19(17):3634–3649.

Pérez-Rodríguez, A., Khimoun, A., Ollivier, A., Eraud, C., Faivre, B. et Garnier, S. (2018). Habitat fragmentation, not habitat loss, drives the prevalence of blood parasites in a Caribbean passerine. *Ecography*, 41(11):1835–1849.

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

Roy, K., Kar, S. et Das, R. N. (2015). Statistical methods in QSAR/QSPR. *In A primer on QSAR/QSPR modeling*, pages 37–59. Springer.

Shirk, A. J., Landguth, E. L. et Cushman, S. A. (2017). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Molecular Ecology Resources*, 18(1):55–67.

Smouse, P. E. et Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82(5):561–573.

Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Editions TECHNIP.

Tobias, R. D. *et al.* (1995). An introduction to partial least squares regression. *In Proceedings of the twentieth annual SAS users group international conference*, pages 1250–1257. SAS Institute Inc Cary.

Van Strien, M. J., Keller, D. et Holderegger, R. (2012). A new analytical approach to landscape genetic modelling : least-cost transect analysis and linear mixed models. *Molecular Ecology*, 21(16):4010–4023.

Watts, A. G., Schlichting, P. E., Billerman, S. M., Jesmer, B. R., Micheletti, S., Fortin, M.-J., Funk, W. C., Hapeman, P., Muths, E. et Murphy, M. A. (2015). How spatio-temporal habitat connectivity affects amphibian genetic structure ? *Frontiers in Genetics*, 6:275.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Wold, S., Sjöström, M. et Eriksson, L. (2001). PLS-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Xia, Y., Bjørnstad, O. N. et Grenfell, B. T. (2004). Measles metapopulation dynamics : a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–281.

Zero, V. H., Barocas, A., Jochimsen, D. M., Pelletier, A., Giroux-Bougard, X., Trumbo, D. R., Castillo, J. A., Evans Mack, D., Linnell, M. A., Pigg, R. M. *et al.* (2017). Complementary network-based approaches for exploring genetic structure and functional connectivity in two vulnerable, endemic ground squirrels. *Frontiers in Genetics*, 8:81.