Distribution spatiale des arbres *Yucca brevifolia* en fonction de variables environnementales

Duruz Solange*, Jollivet Céline*, Kaufmann Anaëlle* et Pellaud Claire* (Groupe 1)

Introduction

Yucca brevifolia (aussi connu sous le nom de Joshua tree) est une espèce d'arbre principalement présente dans désert du Mojave dans le Sud-Ouest de la Californie (Smith, 1983). Il s'agit d'un arbre d'une quinzaines de mètres, adapté aux conditions désertiques et dont la durée de vie est d'environ 300 ans (Gilliland, 2006).

Cette espèce d'arbre peut être pollinisée par deux espèces de mouches (*Tegeticula synthetica* et *Tegeticula anthitetica*), dont la principale différence est la longueur de l'abdomen (Pellmyr, 2003). Ainsi, la longueur de l'ovule de la fleur est également différente chez les arbres selon le pollinisateur, donnant lieu à deux variétés de *Yucca brevifolia* (variété *brevifolia* et *jaegeriana*). Pour la suite de cet article, la variété d'arbre sera désignée par le nom de son pollinisateur. Ce phénomène de mutualisme et de possible co-évolution, concept important en biologie, a attiré l'attention de nombreux scientifiques, faisant ainsi de *Y. brevifolia* une espèce largement étudiée (voir par exemple Baker, 1986).

La distribution spatiale de ces deux variétés est en fait largement compartimentée. Il a été identifié à plusieurs reprises que les arbres pollinisés par l'espèce de mouche *T. synthetica* (variété d'arbre *brevifolia*) occupent la partie Ouest du désert du Mojave alors que ceux pollinisés par *T. antithetica* (variété *jaegeriana)* se sont développés sur la partie Est. Les deux variétés se trouvent tout de même en contact dans une zone relativement restreinte au Nord du désert, où les arbres peuvent potentiellement être pollinisés par les deux types de pollinisateurs. Les deux zones sont séparées par une surface dont l'altitude est plus basse, formant ainsi une barrière naturelle à la propagation des deux variétés. Par soucis de lisibilité, cette région de basse altitude sera appelée canyon pour la suite de cet article, bien qu'il ne s'agisse pas à proprement dit d'un canyon. Rowlands et al. (1982) ont par ailleurs remarqué que la végétation et le climat diffèrent légèrement entre ces deux parties du désert.

Les conditions climatiques sont d'ailleurs un facteur important concernant la distribution spatiale des espèces en général et plus particulièrement de l'espèce qui nous intéresse, *Y. brevifolia*. D'une part, chaque espèce a une niche écologique lui correspondant. Le concept de niches est souvent utilisé en écologie et est défini comme étant l'ensemble des étendues de valeurs de facteurs environnementaux pour lesquels une espèce est adaptée (Hooper, 2008). D'autre part, un organisme vivant dans des conditions environnementales peu favorables à son développement peut aussi s'adapter à cet environnement et se différencier par le processus de sélection naturelle exposé par Darwin (Darwin, 1859). Cependant, une mutation peut également apparaître par un processus aléatoire non-conditionné par des variables environnementales (Lande, 1976). Ainsi, Godsoe et al. (2009a) cherchent à déterminer si les deux variétés de *Y. brevifolia* se sont différenciées sous l'effet de la sélection naturelle.

Dans cet article, nous étudierons l'influence des paramètres environnementaux sur la répartition spatiale de cette espèce. Le but n'est pas de déterminer si *Y. brevifolia* est soumis à un phénomène de sélection naturelle mais d'offrir une analyse de la distribution des arbres en

*Les auteurs ont contribué de manière égale à l'élaboration de ce travail.

fonction des variables environnementales. Nous chercherons donc à identifier si les niches écologiques des deux variétés sont similaires.

Données

Un relevé quasi-exhaustif des arbres Joshua dans la zone d'intérêt a été fourni par Godsoe et al. (2009a et b). Un total de 5765 arbres représentés spatialement par un point sont à disposition. Pour chaque arbre relevé, les coordonnées ainsi que l'espèce de mouche pollinisatrice sont données. L'arbre a alors un attribut nommé *synthetica, antithetica* ou *contact* (pollinisé par les deux espèces de mouches).

Le modèle numérique de terrain a été extrait par l'outil en ligne DEM explorer pour la zone du relevé avec une résolution de 1 arc seconde (~30m sur la zone) (Han, 2012).

Les données environnementales sont basées sur une période de trente ans entre 1961 et 1990, et ont une résolution spatiale de 10 minutes. Ces données raster comportent des informations sur les précipitations (la quantité en mm/mois et le coefficient de variation saisonnier), le nombre de jours par mois avec des précipitations supérieures à 0.1 mm, la température moyenne en °C, l'intervalle moyen des températures journalières en °C, l'humidité relative en pourcentage, le pourcentage d'ensoleillement maximum possible sur la durée du jour, le nombre de jour par mois où le sol est gelé et la vitesse du vent à 10m du sol en m/s. Les moyennes annuelles des variables ont été utilisées. (Climate Research Unit, 2002)

Méthodes

Prétraitement

Premièrement et après l'extraction des différentes données, la pente et l'orientation ont été calculées à partir du modèle numérique de terrain. Ceci a été fait grâce au logiciel de SIG SAGA GIS[1], en utilisant le module *Terrain Analysis – Local Morphometry*.

Sur un autre logiciel de SIG, Quantum GIS[2], une grille de 2378 éléments de 0.1° (~11km) a été créée grâce à l'outil *Vector grid*. Les onze facteurs environnementaux ainsi que le nombre d'arbres de chaque variété ont été extraits pour chaque élément de la grille à l'aide de l'outil *Transfer Height* et de requêtes SQL dans Manifold System[3] (également un logiciel SIG). A noter que les quelques arbres étant pollinisés par les deux mouches (*contact* dans le jeu de données) n'ont pas été pris en compte lorsque le nombre d'arbres par variété a été compté.

Afin de faciliter les analyses prévues, la zone étudiée a été divisée en deux parties (Est et Ouest) séparées par le canyon mentionné dans l'introduction. Ce découpage a été fait à la main, en suivant approximativement les limites du MNT et en s'assurant que chaque variété d'arbre soit uniquement présente sur une seule des parties.

A la fin de ces étapes, une grille avec les attributs suivants était à disposition : onze variables environnementaux (altitude, pente, orientation, température, ...), trois variables espèces binaires (présence/absence de *synthetica*, présence/absence d'*antithetica*, présence/absence d'un arbre Joshua quelle que soit sa variété), une variable binaire de localisation (1 si l'élément de grille est à l'Ouest du canyon 0 autrement).

1 http://www.saga-gis.org/

2 http://www.qgis.org/

3 http://www.manifold.net/

Analyse

Pour ce projet, il a été choisi d'effectuer des analyses basées sur la méthode de la régression logistique. La régression logistique est un modèle de prédiction qui mesure le degré de relation d'une variable dépendante catégorielle (p. ex. variable binaire de présence/absence d'arbre; variable trois catégories : présence de *synthetica*, présence d'*antithetica* et absence de *synthetica* et *antithetica*) et de variables indépendantes explicatives (les variables environnementales dans notre cas). Le résultat d'une régression logistique est une probabilité d'être d'une certaine catégorie comparée à la probabilité de ne pas être de cette catégorie. (Wikipedia, 2014b).

La régression logistique est une méthode paramétrique se basant sur les mêmes principes que la régression linéaire ordinaire. Comme la variable dépendante est catégorielle, on utilise une transformation appelée transformation logit π(w) pour obtenir une variable dépendante continue C(x), le logit (Rikotamalala R., 2013):

où

$$C(x) = ln\left(\frac{\pi(w)}{1 - \pi(w)}\right)$$

w : *un individu*

$\pi(w) \in [0,1]$

$\frac{\pi(w)}{1 - \pi(w)}$: *ratio de la probabilité d'être d'une certaine catégorie sur la probabilité de ne pas être de cette catégorie*

Ce dernier ratio est le résultat de la régression logistique. Une régression linéaire est alors effectuée sur le logit. On obtient (R. Rikotamalala, 2013):

$$C(x) = a_0 + a_1 X_1 + \cdots + a_j X_j$$

La qualité de la régression peut ensuite être estimée, notamment grâce à la valeur p du test de Student. Le test de Student teste la nullité d'un coefficient de régression. La valeur p donne la probabilité d'avoir un test de Student plus petit avec un modèle aléatoire, donc d'avoir un coefficient de régression moins significatif avec un modèle aléatoire. Si la valeur p est inférieure à un seuil choisi (dans notre cas 0.01), alors la régression est significative. (Wikipedia, 2014d)

Les régressions logistiques sont classées en deux grands groupes. Les régressions logistiques univariées (ou simples) et multivariées. Pour la première catégorie, la fonction de régression estime la probabilité d'avoir une certaine catégorie binaire en fonction d'une variable indépendante alors que pour la deuxième, la fonction de régression estime la probabilité d'avoir une certaine catégorie parmi plusieurs catégories en fonction de plusieurs variables indépendantes. (Wikipedia, 2014b)

Pour chaque analyse effectuée, la construction du modèle a été effectuée selon deux cas, se différenciant par la variable dépendante choisie pour construire le modèle logistique :

1) $\begin{cases} 2, & si\ présence\ de\ synthetica \\ 1, & si\ présence\ d'antithetica \\ 0, & si\ absence\ de\ synthetica\ et\ antithetica \end{cases}$ *sur les éléments de grille*

2) a. $\begin{cases} 1, & si\ présence\ de\ synthetica \\ 0, & si\ absence\ de\ synthetica \end{cases}$ *sur les éléments de grille à l'Ouest du canyon*

b. $\begin{cases} 1, & si\ présence\ de\ antithetica \\ 0, & si\ absence\ de\ antithetica \end{cases}$ *sur les éléments de grille à l'Est du canyon*

Ainsi, le modèle 1 cherche à discriminer entre la probabilité d'avoir une ou l'autre des espèces, alors que le modèle 2 essaie de montrer si les conditions sont favorables à l'implantation d'une espèce donnée sur la partie opposée du canyon. Ce dernier modèle permet aussi de comparer si les fortes probabilités sur la partie opposée correspondent aux endroits où l'autre espèce s'est développée.

Premièrement, dans le but de discriminer les variables environnementales influençant majoritairement la variable dépendante de chaque cas, une sélection des variables a été faite grâce au critère d'information d'Akaike sur un ajustement d'un modèle logistique entre la variable dépendante et les variables environnementales. Cette sélection a été faite à l'aide des fonction *step* et *multinom* du logiciel R[4]. Ce logiciel est utilisé principalement pour des analyses statistiques. Le critère mesure la qualité relative d'un modèle en fonction de la qualité de son ajustement et de sa complexité (Wikipedia, 2014a).

Deuxièmement, avec Matlab[5] (logiciel de calcul numérique), un modèle logistique multivarié est ajusté sur les variables environnementales pour les deux cas cités ci-dessus grâce aux fonctions *mnrfit* et *mnrval*. Le modèle est construit en se basant sur une fraction du nombre de pixels concernés par le modèle (1/3 pour le modèle 1 et 1/2 pour le modèle 2 pour s'assurer d'avoir suffisamment de pixels pour construire le modèle). Le modèle obtenu est ensuite appliqué à tous les pixels étudiés (y compris la partie opposée du canyon pour le modèle 2). Il en résulte une prédiction donnant la probabilité que la variable dépendante appartienne à une certaine catégorie pour chaque élément de grille.

Finalement, les prédictions obtenues sont visualisées sous forme de cartes de probabilité (créées avec Manifold System).

Contrôle

Pour les deux modèles retenus, un seuil pour décider à partir de quelle probabilité résultante une variable dépendante appartient à une certaine catégorie est ensuite calculé dans Matlab. Ce seuil permet par la suite d'effectuer un test d'exactitude de la prédiction de chaque modèle. Pour le modèle 2a et 2b, deux tests d'exactitude sont calculés : un pour le côté de l'espèce prédite et un autre sur l'autre côté. Ce dernier test n'est pas à proprement parlé un test d'exactitude mais permet de vérifier si les conditions favorables à la présence d'une espèce sont réunies où pousse l'autre espèce. Cette probabilité a été calculée à l'aide de l'analyse de la courbe ROC (Receiver Operating Curve). Cette analyse se base sur le calcul de la sensibilité et la spécificité du résultat (dont les calculs sont développés dans le tableau 1). Notons simplement que la sensibilité et la spécificité dépendent du seuil défini. On peut donc représenter graphiquement ces derniers pour différents seuils. Le seuil optimal peut être lu graphiquement à l'endroit où la sensibilité et la spécificité se croisent et correspond à la maximisation des prédictions correctes en minimisant les erreurs de prédictions. (Thuiller, 2003), (Wikipedia, 2014c), (SigmaPlot, 2014)

4 http://www.r-project.org/

5 http://www.mathworks.ch/products/matlab/

		Réalité		
		Positif réel (catégorie 1)	Négatif réel (catégorie 0)	
Prédiction	Positif prédit (Catégorie 1) (probabilité > seuil)	Vrai positif	Faux positif	Précision $\frac{\sum Vrai\ positif}{\sum Positif\ prédit}$
	Négatif prédit (Catégorie 0) (probabilité < seuil)	Faux négatif	Vrai négatif	Valeur prédite négative $\frac{\sum Vrai\ négatif}{\sum Négatif\ prédit}$
		Sensitivité = $\frac{\sum Vrai\ positif}{\sum Positif\ réel}$	Spécificité = $\frac{\sum Vrai\ négatif}{\sum Négatif\ réel}$	

Tableau 1: Calcul de la sensitivité et de la spécificité

Finalement, un indice d'exactitude des prédictions résultant des trois modèles est calculé dans Matlab. L'indice choisi correspond à la sensibilité (voir tableau 1). Il est donc dépendant du seuil choisi. Il a été calculé pour deux seuils différents pour chaque modèle :

1) Seuils obtenus à l'aide de l'analyse de courbe ROC optimisée pour chaque espèce
2) Seuil = 0.8, ce qui correspond à un seuil pour un bon résultat (Thuiller, 2003)

A noter que la sensibilité ne tient pas compte du nombre de faux positifs car en écologie l'absence est difficilement interprétable.

Une valeur p associée à chaque indice d'exactitude obtenu précédemment est aussi calculée avec Matlab. Cette valeur p nous indique la significativité de l'indice d'exactitude. Pour la calculer, on procède par la méthode de Monte-Carlo :

1) Calcul de la sensibilité de la prédiction obtenue en appliquant le modèle ajusté à une matrice explicative où chaque valeur des variables environnementales est permutée aléatoirement
2) Répétition de l'étape 1, 999 fois
3) la valeur p :

$$p = \frac{(\#\ sensiblité\ obtenue\ par\ permutation\ aléatoire > sensibilité\ réelle) + 1}{nombre\ de\ permutation\ aléatoire\ (= 999) + 1}$$

On considère le résultat comme significatif si $p < 0.01$.

Résultats

Comme dit précédemment deux modèles ont été retenus pour les analyses. La numérotation des modèles est donnée dans les méthodes.

Les résultats de la sélection par le critère d'Akaike pour les deux modèles retenus sont reportés dans le tableau 2.

Variables retenues après la séléction par le critère d'Akaike		
Modèle 1	**Modèle 2a (synthetica)**	**Modèle 2b (antithetica)**
Altitude	Altitude	Altitude
Orientation	Orientation	Pente
Pente	Pente	Température
Température	Température	Durée du jour
Durée du jour	Durée du jour	Humidité relative
Humidité relative	Humidité relative	Jours de pluie
Jours de pluie	Jours de pluie	CV précipitation
CV précipitation	Précipitation (mm/mois)	Jours où le sol est gelé
Précipitation (mm/mois)	Jours où le sol est gelé	Intervalle moyen de temperature
Jours où le sol est gelé	CV précipitation	Orientation
Intervalle moyen de temperature	Intervalle moyen de temperature	Précipitation (mm/mois)

Tableau 2: Les variables retenues après la sélection par le critère d'Akaike sont en vert. Les variables rejetées sont en rouge.

Pour le modèle 1 aucune variable n'a été éliminée. Pour le modèle 2, deux variables sont éliminées dans les deux cas.

Les tableaux 3 représente les valeurs trouvées pour la valeur p calculée avec le test de Student pour les deux modèles étudiés (permettant d'estimer la significativité de la régression logistique).

Valeur p (test de Student)	Modèle 1		Modèle 2a (synthetica)	Modèle 2b (antithetica)
Altitude	3.51E-02	2.20E-01	4.93E-02	2.82E-01
Orientation	7.60E-03	7.02E-01	4.34E-03	
Pente	6.55E-04	7.42E-02	1.28E-05	2.29E-02
Température	1.30E-07	2.21E-05	2.77E-14	3.34E-02
Durée du jour	1.03E-02	1.69E-03	2.04E-01	2.01E-03
Humidité relative	4.33E-04	4.22E-09	1.83E-10	1.29E-05
Jours de pluie	2.48E-10	6.06E-03	2.90E-13	8.64E-06
CV précipitation	6.08E-02	8.17E-03		1.81E-03
Précipitation (mm/mois)	1.48E-02	1.91E-01	8.26E-04	
Jours où le sol est gelé	1.20E-04	5.27E-05	1.32E-14	4.05E-01
Intervalle moyen de temperature	9.85E-04	3.86E-02		4.37E-02

Tableau 3: Valeur p moyennes calculée avec le test de Student. La régression est significative si la valeur p est inférieure au seuil de 0.01 et est mise en évidence en violet dans le tableau. A noter qu'il y a deux valeurs p pour le modèle 1, la première pour synthetica et la deuxième pour antithetica.

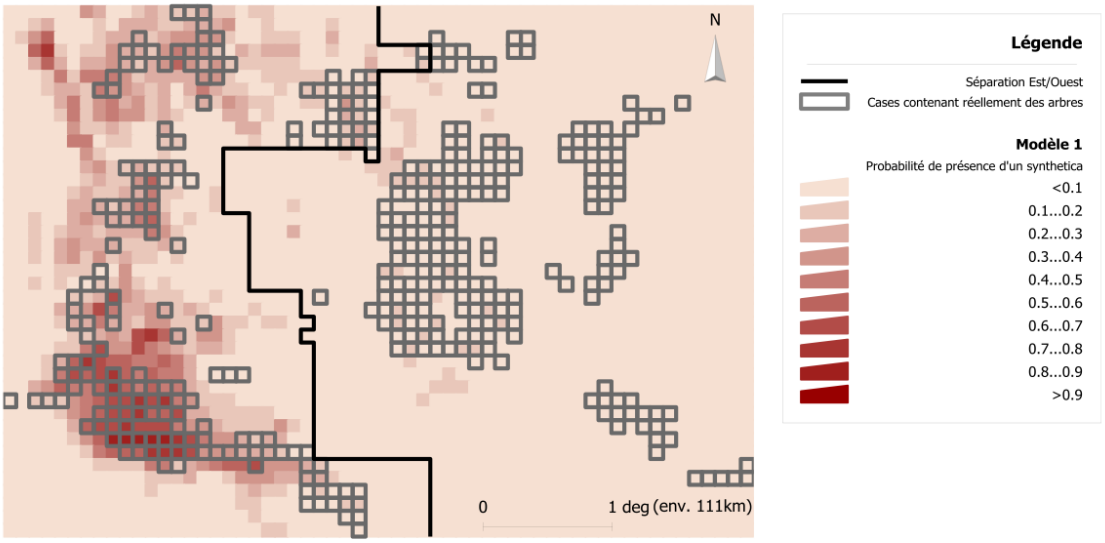
Pour les modèles 1 et 2a, la régression est significative pour une majorité des variables. Pour le modèle 2b, il existe autant de variable non significatives que de variables significatives.

Le modèle logistique multivarié ajusté sur les variables environnementales a permis d'obtenir des prédictions quant à la probabilité de présence des deux espèces. Les cartes de probabilité obtenues pour les deux modèles sont présentées dans les figures 1 et 2 ci-dessous.

Modèle 1 - Cartes de probabilité de présence d'un synthetica et d'un antithetica ou d'absence d'arbres

Les cartes suivantes ont été créées avec le modèle 1 qui se base sur la présence de syntethica, la présence d'antitethica ou l'absence de ces deux espèces sur toute la grille.

a. Probabilité de présence de synthetica sur le site étudié



b. Probabilité de présence d'antithetica sur le site étudié

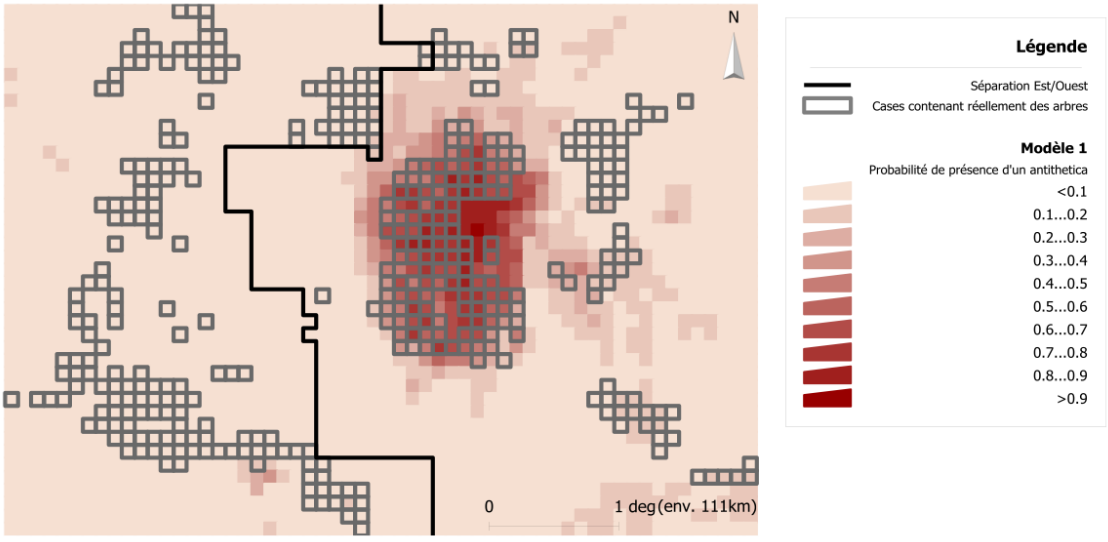
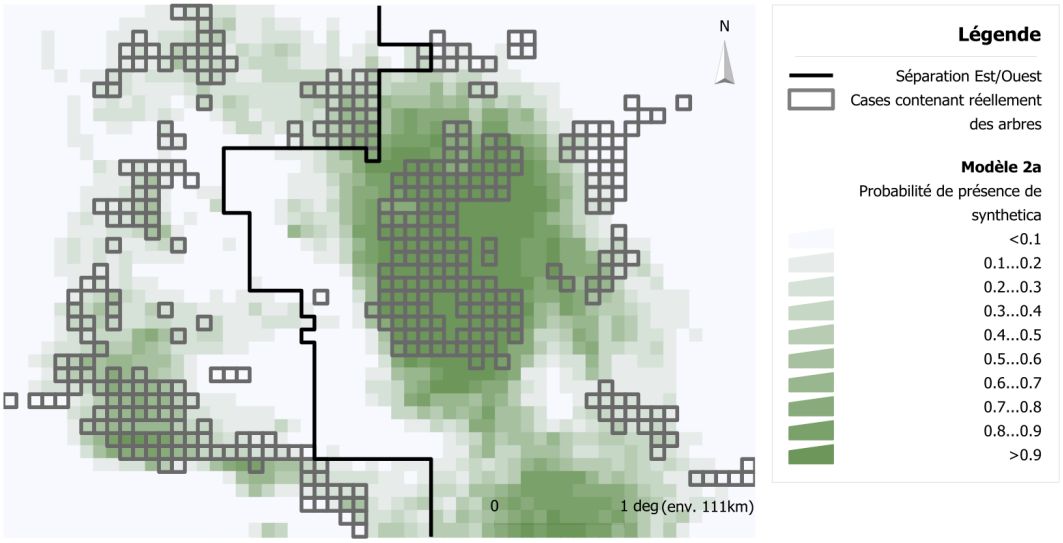


Figure 1 Les pixels dont le tour est gris correspondent aux pixels ayant un arbre. La ligne noire est la délimitation est-ouest (délimitant les espèces) mentionnée dans la partie méthode.

Sur la figure 1 on remarque que la probabilité de présence prédite des *synthetica* est plus forte à l'Ouest et correspond bien à la réalité, la probabilité de présence des *antithetica* est forte dans l'amas d'arbres du centre mais les *antithetica* qu'il y a autour sont peu prédits.

Modèle 2 - Cartes de probabilité de présence des synthetica ou des antithetica

a. Prédiction de la probabilité de présence de synthetica sur toute la surface d'étude selon un modèle basé sur la présence ou non de synthetica à l'Ouest du canyon



b. Prédiction de la probabilité de présence d'antithetica sur toute la surface d'étude selon un modèle basé sur la présence ou non d'antithetica à l'Est du canyon

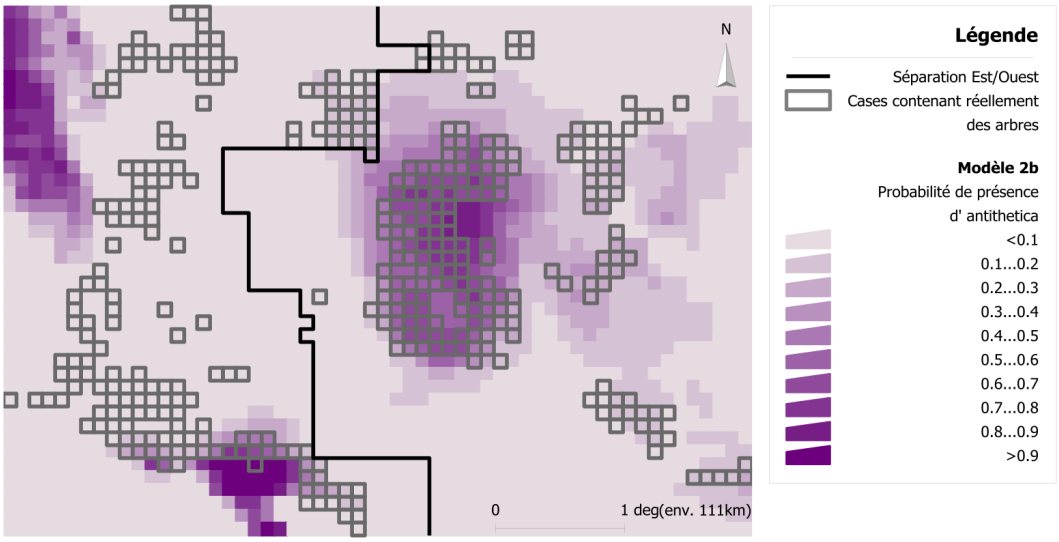


Figure 2 Carte de probabilité du modèle 2

Sur la figure 2.a. on voit que dans la partie Ouest les prévisions de présence des *synthetica* correspondent assez bien à la réalité (forte probabilité où il y a effectivement des arbres). Cependant, les probabilités les plus fortes de présence de *synthetica* sont majoritairement concentrées dans la partie où il y a des *antithetica* dans la réalité (côté Est). Il y a aussi une forte probabilité de présence au Sud de cette partie alors qu'il n'y a pas d'arbres en réalité.

Concernant la figure 2.b. les probabilités de présence des *antithetica* sur la partie Ouest sont en cohérence avec la réalité avec à nouveau une tache de forte probabilité dans le centre. Par contre, sur la partie Est, les *antithetica* ne sont pas prédits où les *synthetica* sont présents, mais à d'autres endroits.

L'analyse de la courbe ROC a donné les résultats suivants pour la détermination des seuils :

Moyenne des seuils obtenus avec l'analyse ROC				
Modèle 1	**Modèle 2 (synthetica)**		**Modèle 2 (antithetica)**	
	Ouest	Est	Est	Ouest
0.11	0.26	0.31	0.13	0.04

Tableau 4: Moyenne des seuils obtenus avec l'analyse ROC

Les seuils obtenus sont bas, particulièrement pour le modèle 2b-Ouest.

La méthode de Monte Carlo a donné les résultats suivants pour les seuils calculés :

	Valeur p calculée avec la méthode de Monte-Carlo					
	Seuils calculés			Seuil de 0.8		
	Sensibilité réelle	Sensibilité MC	Valeur p	Sensibilité réelle	Sensibilité MC	Valeur p
Modèle 1 - antitheticas	0.7512	0.3227	0.001	0.0498	0.1773	1
Modèle 1 - syntheticas	0.7784	0.3250	0.001	0.0162	0.1503	1
Modèle 1 - absence d'arbres	0.7642	0.3238	0.001	0.0337	0.1643	1
Modèle 2a (syn.) - Ouest	0.6378	0.4694	0.001	0.0270	0.4250	1
Modèle 2a (syn.) - Est	0.6468	0.4657	0.001	0.5771	0.4247	0.001
Modèle 2b (ant.) - Est	0.6965	0.4645	0.001	0.0149	0.3261	1
Modèle 2b (ant.) - Ouest	0.6432	0.4996	0.001	0.0054	0.3263	1

Tableau 5: Valeurs p calculées avec la sensibilité réelle et la sensibilité calculée par permutation avec la méthode de Monte-Carlo (MC). Les modèles avec une bonne significativité sont mis en évidence en violet dans le tableau.

Les valeurs p sont basses (0.001) pour les seuils calculés avec l'analyse ROC alors qu'elles sont hautes (1) pour pour un seuil de 0.8, à l'exception du modèle *2a. - Est* qui était aussi le modèle avec le seuil calculé le plus élevé (tableau 4).

Discussion

La sélection des variables environnementales par le critère d'Akaike permet d'éliminer peu de variables. Les variables éliminées étant celles qui n'apportent pas d'explication au modèle, cela montre que la plupart des variables sont importantes pour son élaboration.

Les résultats du premier modèle montrent que chacune des espèces est prédite avec une forte probabilité aux endroits où on la trouve réellement, c'est à dire *antithetica* à l'Est et *synthetica* à l'Ouest. Par contre, elles ne diffusent pas sur le côté opposé, certainement en partie à cause du fait que le modèle est construit en prenant en compte les absences de chaque espèce, et que la somme des probabilités de présence d'*antithetica*, présence *synthetica* et absence d'arbre est égale à un : dès qu'une espèce est prédite fortement à un endroit, l'autre ne peut pas l'être. Ceci tend à montrer que les conditions dans lesquelles poussent les deux variétés sont différentes ou du moins différentiables.

Comme mentionné dans les résultats pour le modèle 2a, l'espèce *synthetica* est prédite aux endroits où se trouvent réellement des arbres pollinisés par l'espèce *antithetica* (côté Est de la zone). Pour le modèle 2b, l'espèce *antithetica* est prédite sur des zones non peuplées du côté Ouest, et est peu prédite du côté Est. Ces résultats suggèrent la présence d'une barrière écologique entre les deux zones, qui n'a pas pu être franchie par les arbres, puisque les conditions environnementales semblent être favorables au développement de chaque variété sur les deux parties de la zone. Cependant la prédiction d'*antithetica* du côté Est où elle se situe réellement, ne correspond pas très bien à la réalité, ce qui laisse penser que les paramètres utilisés ne sont pas suffisants pour expliquer la répartition de cette variété. Ce manque de garantie sur les paramètres environnementaux utilisés peut aussi être un élément de réponse quant aux résultats obtenus : un paramètre limitant pour la croissance qui ne serait

pas pris en compte dans le modèle peut être à l'origine de la prédiction dans les zones opposées de chaque variété.

Le fait que l'espèce *synthetica* ne soit pas prédite par le modèle 2a sur la partie Nord-Ouest de la zone peut être expliqué par des conditions environnementales peu favorables à *synthetica* (mais favorables à *antithetica*). Il est aussi possible que les absences de *synthetica* qui aident à construire le modèle ne permettent pas la prédiction d'arbre dans cette région, alors que la niche écologique pour cette espèce est peut être existante, mais simplement encore vide (les facteurs environnementaux sont favorables mais la zone n'a pas encore été atteinte par l'espèce). Ce résultat peut être confirmé par le fait que le seuil calculé avec la méthode ROC pour le modèle 2b-Ouest a un seuil extrêmement faible. Ceci est certainement dû au fait qu'il y a très peu de « vrais positifs » (prédiction de *antithetica* là où *synthetica* est présent).

On pourrait interpréter ces résultats de la manière suivante : les *synthetica* ont une niche écologique similaire à celle des *antithetica* puisque le modèle prédit des *synthetica* là où il y a un *antithetica*. Par contre, le contraire ne semble pas vrai (les conditions dans lesquelles poussent les *synthetica* ne sont pas favorables aux *antithetica*). Il est ainsi possible que la niche des *synthetica* soit plus large.

Les seuils déterminés par la méthode ROC sont très faibles comparés aux valeurs de seuils proposées dans la littérature pour de tels cas (0.8 à 0.9) (Thuiller, 2003). Ceci peut être dû au fait que notre modèle n'est pas idéal (les limites du modèle sont discutées plus bas). En parallèle, les valeurs p obtenues par les permutations du test de Monte Carlo montrent la faible significativité des modèles si le seuil choisi est trop élevé. Les valeurs p obtenues pour des seuils de l'ordre de ceux déterminés par la méthode ROC suggèrent alors une bonne significativité des modèles.

Les seuils obtenus par la méthode ROC montrent que les modèles construits sont peu sensibles aux changements de variables environnementales. En augmentant faiblement le seuil, le nombre de vrais positifs rapporté au nombre de positifs réels (la sensitivité) diminue fortement. En revanche la spécificité (nombre de vrai négatif rapporté au nombre de négatifs réels) diminue faiblement avec le seuil (pour un seuil de 0.5, la spécificité reste très élevée). Ainsi peu de positifs seraient obtenus par les prédictions, et garder un seuil bas permet de conserver une sensibilité relativement haute.

Le test d'exactitude correspond à la sensibilité, donc ne tient pas compte de l'absence. En effet, l'absence d'arbre est en réalité difficilement interprétable car elle peut provenir de différents facteurs. Premièrement, il est possible que la prédiction de la présence de l'arbre soit fausse. Cependant, il est également possible que les conditions soient réunies pour que l'arbre soit présent mais il n'est pas présent au moment des relevés ou encore que l'arbre ne soit pas dispersé régulièrement sur le territoire. C'est pour cela qu'il a été décidé de ne pas prendre en compte les faux positifs dans le test d'exactitude.

De nombreuses limites peuvent être formulées sur les modèles construits. Ainsi, les résultats formulés et les conclusions qui en découlent devraient être considérés avec précautions.

La résolution spatiale choisie jouant un rôle important dans la description de l'emplacement des arbres, il aurait pu être nécessaire de varier l'échelle des données, ou d'obtenir des données plus finement résolues. Il en va de même pour la résolution temporelle des données : nous avons travaillé ici avec des moyennes annuelles sur une période de trente ans. Or la longévité des arbres de l'espèce étudiée est nettement plus grande que cette période puisque ce type d'arbre peut atteindre jusqu'à trois cents ans. Ainsi, il est possible que les arbres en place au

moment du relevé soient plus anciens que le début de la période sur laquelle les variables environnementales ont été observées. Il est aussi possible que l'importance de certaines variables environnementales change selon le stade de croissance. Par ailleurs, les variables sélectionnées pour l'étude ne sont pas exhaustives et d'autres variables environnementales telles que le pH du sol ou le niveau de la nappe phréatique par exemple pourraient être des variables pertinentes. De plus les variables choisies pour construire les modèles peuvent présenter des corrélations entre elles, même si elles apportent chacune une partie explicative pour le modèle.

Le modèle de régression logistique utilisé repose sur un modèle linéaire, alors que le réponses écologiques (variables environnementales) ont certainement un comportement unimodale (Guisan, 2000). Ainsi même si une niche écologique est prédite, il est possible que l'optimum des conditions environnementales pour l'espèce soit dépassé, et donc que la niche n'aurait pas pu exister en réalité.

Pour la construction des modèles, l'absence d'un arbre est considérée comme une information, ce qui est justifiable comme le relevé des arbres est presque exhaustif. Cependant il aurait pu être plus adapté de ne travailler que sur les présences, puisque une absence ne montre pas forcement que la niche n'est pas adaptée au développement des arbres, mais peut aussi montrer que la niche est adaptée mais simplement encore vide, ou occupée par une autre espèce.

Comme les données ont été agrégées dans une grille régulière, il aurait pu être intéressant de travailler en termes d'abondance d'arbres, et pas de présence absence.

Conclusion

Basé sur une analyse des variables environnementales à disposition ainsi que d'un modèle de régression logistique, il semble que les variétés *brevifolia* et *jaegariana* (pollinisée par *synthetica* et *antithetica* respectivement) se sont développées dans des conditions environnementales différentes. En effet, lorsque le modèle est prédit selon trois catégories (présence *synthetica*, présence *antithetica* ou absence d'arbre), la réponse des deux espèces est différentiable. Il semblerait que la croissance d'arbres pollinisés par *synthetica* soit possible dans les mêmes niches que ceux pollinisés par *antithetica* mais que l'inverse ne soit pas vrai, et donc que la niche de *synthetica* soit plus large. Ces résultats ont pu être obtenus en construisant le modèle sur un côté puis en prédisant sur l'autre et en comparant si les prédictions d'arbres se trouvent là où l'autre variété est établie.

Cependant, la régression logistique multiple sur laquelle est basée notre analyse n'est pas forcément la meilleure manière de modéliser des niches écologiques et donc les conclusions de notre analyse peuvent être biaisées. Ainsi, il serait intéressant de refaire l'analyse avec d'autres modèles (tenant compte d'un optimum et n'ayant pas une réponse linéaire par exemple). Il serait aussi certainement pertinent d'inclure d'autres variables environnementales dans l'analyse. Néanmoins, le meilleur moyen de vérifier la possible croissance d'une des variétés dans une niche écologique de l'autre variété serait de faire des tests réels de plantation. Les *Y. brevifolia* ayant des croissances très lentes, ces tests ne sont en réalité pas possibles.

Références

Baker, H. G. (1986). Yuccas and yucca moths-a historical commentary.*Annals of the Missouri Botanical Garden*, 556-564.

Climatic Research Unit, 2002. Ten Minute Climatology, disponible à http://www.cru.uea.ac.uk/cru/data/hrg/tmc/

Darwin C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray.

Gilliland, K. D., Huntly, N. J., & Anderson, J. E. (2006). Age and population structure of Joshua trees (Yucca brevifolia) in the northwestern Mojave Desert. *Western North American Naturalist*, *66*(2), 202-208.

Godsoe W, Strand E, Smith CI, Yoder JB, Esque TC, Pellmyr O (2009a) Divergence in an obligate mutualism is not explained by divergent climatic factors. New Phytologist 183(3): 589–599. doi:10.1111/j.1469-8137.2009.02942.x

Godsoe W, Strand E, Smith CI, Yoder JB, Esque TC, Pellmyr O (2009b) Data from: Divergence in an obligate mutualism is not explained by divergent climatic factors. Dryad Digital Repository. doi:10.5061/dryad.6s67t

Guisan, A., & Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. Ecological modelling, 135(2), 147-186.

Han, W., Di, L., Zhao, P., Shao, Y., 2012. DEM Explorer: An online interoperable DEM data sharing and analysis system, Environmental Modelling & Software, 38, 101-107

Hooper, H. L., Connon, R., Callaghan, A., Fryer, G., Yarwood-Buchanan, S., Biggs, J., ... & Sibly, R. M. (2008). The ecological niche of Daphnia magna characterized using population growth rate. *Ecology*, *89*(4), 1015-1022.

Lande R. (1976). Natural selection and random genetic drift in phenotypic evolution. Evolution, 314-334.

Pellmyr, O., & Segraves, K. A. (2003). Pollinator divergence within an obligate mutualism: two yucca moth species (Lepidoptera; Prodoxidae: Tegeticula) on the Joshua tree (Yucca brevifolia; Agavaceae). *Annals of the Entomological Society of America*, *96*(6), 716-722.

Rakotomalala R. (2013). Pratique de la Régression Logission, Régression Logistique Binaire et Polytomique (Version 2.0), *Université Lumière Lyon 2*, (http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf)

Rowlands, P., Johnson, H., Ritter, E., & Endo, A. (1982). The Mojave Desert. *Reference handbook on the deserts of North America. Greenwood Press, Westport, CT*, 103-162.

Smith, S. D., Hartsock, T. L., & Nobel, P. S. (1983). Ecophysiology of Yucca brevifolia, an arborescent monocot of the Mojave sesert. *Oecologia*, *60*(1), 10-17.

SigmaPlot, Exact Graphs and Data Analysis, dernière visite : 07.01.2014, *ROC Curves Analysis*, (http://www.sigmaplot.com/products/sigmaplot/ROC_Curves_Analysis.pdf)

Thuiller, W., Vayreda, J., Pino, J., Sabate, S., Lavorel, S., & Gracia, C. (2003). Large‐scale environmental correlates of forest tree distributions in Catalonia (NE Spain). *Global Ecology and Biogeography*, *12*(4), 313-325.

Wikipedia, derinère modification 04.01.2014, dernière visite 07.01.2014a. *Akaike information criterion*, (http://en.wikipedia.org/wiki/Akaike_information_criterion)

Wikipedia, dernière modification 04.01.2014, dernière visite 07.01.2014b. *Logistic regression*, (http://en.wikipedia.org/wiki/Logistic_regression)

Wikipedia, dernière modification 30.12.2013, dernière visite 07.01.2014c. *Receiver Operation Caracteristic* (http://en.wikipedia.org/wiki/Logistic_regression)

Wikipedia, dernière modification 29.11.2013, dernière visite 07.01.2014d. *Test de Student*, (http://fr.wikipedia.org/wiki/Test_de_Student)

Contributions des auteurs

Le choix de l'hypothèse et de la méthode de travail a été fait en groupe. Les différentes étapes des calculs ainsi que l'interprétation des résultats obtenus ont également pour la plupart été faites en groupe. S.D. a rédigé l'introduction, et la conclusion. C.P. a rédigé la partie sur les données et la discussion. C.J. a rédigé les méthodes. Les figures présentées dans les résultats, ainsi que la rédaction de cette dernière partie ont été effectués par A.K. Une relecture a été effectuée en commun afin d'arriver à un consensus pour la version finale.