RÉVISION AUTOMATIQUE DE THÉORIES ÉCOLOGIQUES

par

Philippe Desjardins-Proulx

thèse présentée au Département de biologie en vue
de l'obtention du grade de docteur ès sciences (Ph.D.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, juillet 2018

Le 11 juillet 2018

le jury a accepté la thèse de Monsieur Philippe Desjardins-Proulx dans sa version finale.

Membres du jury

Professeur Dominique Gravel
Directeur de recherche
Département de biologie

Professeur Timothée Poisot
Co-directeur de recherche
Département de sciences biologiques
Université de Montréal

Professeur Pierre-Étiennes Jacques
Président rapporteur
Département de biologie

Professeur Alireza Tamaddoni Nezhad
Évaluateur externe
Department of Computer Science
University of Surrey

Professeur Shengrui Wang
Évaluateur interne
Département d'informatique

À tous ceux qui m’ont soutenu

SOMMAIRE

À l'origine, ce sont des difficultés en biologie évolutive qui ont motivé cette thèse. Après des décennies à tenter de trouver une théorie basée sur la sélection capable de prédire la diversité génomique, les théoriciens n'ont pas trouvé d'alternatives pratiques à la théorie neutre. Après avoir étudié la relation entre la spéciation et la diversité (Annexes A, B, C), j'ai conclu que l'approche traditionnelle pour construire des théories serait difficile à appliquer au problème de la biodiversité. Prenons par exemple le problème de la diversité génomique, la difficulté n'est pas que l'on ignore les mécanismes impliqués, mais qu'on ne réussit pas à construire de théories capable d'intégrer ces mécanismes. Les techniques en intelligence artificielle, à l'inverse, réussissent souvent à construire des modèles prédictifs efficaces justement là où les théories traditionnelles échouent. Malheureusement, les modèles bâtis par les intelligences artificielles ne sont pas clairs. Un réseau de neurones peut avoir jusqu'à un milliard de paramètres. L'objectif principal de ma thèse est d'étudier des algorithmes capable de réviser les théories écologiques. L'intégration d'idées venant de différentes branches de l'écologie est une tâche difficile. Le premier défi de ma thèse est de trouver sous quelle représentation formelle les théories écologiques doivent être encodées. Contrairement aux mathématiques, nos théories sont rarement exactes. Il y a à la fois de l'incertitude dans les données que l'on collecte, et un flou dans nos théories (on ne s'attend pas à que la théorie de niche fonctionne 100% du temps). Contrairement à la physique, où un petit nombre de forces dominent la théorie, l'écologie a un très grand nombre de théories. Le deuxième défi est de trouver comment ces théories peuvent être révisées automatiquement. Ici, le but est d'avoir la clarté des théories traditionnelles et la capacité des algorithmes en intelligence artificielle de trouver de puissants modèles prédictifs à partir de données. Les tests sont faits sur des données d'interactions d'espèces.

Mots-clés : Intelligence Artificielle ; apprentissage automatique ; écologie théorique ; intéraction des espèces.

REMERCIEMENTS

Je tiens à remercier mes deux directeurs, Dominique Gravel et Timothée Poisot, non seulement pour leur soutien, mais surtout pour les intéressantes, passionnées et vigoureuses discussions sur la théorie et l'écologie. Ma passion en apprentissage automatique est née durant mes études et c'est au cours de nos discussion que j'ai découvert le sujet de la révision automatique de théories scientifiques. Ce doctorat n'est que le début de mes recherches sur le sujet et je vous en suis reconnaissant.

Je remercie le CRSNG (Conseil de recherches en sciences naturelles et en génie du Canada), Microsoft, et NVIDIA d'avoir financé ce doctorat.

Je remercie mes collabateurs : James Rosindell, Ignasi Bartomeus, Idaline Laigle, et aussi les membres du laboratoire, Hedvig Nenzén, Steve Vissault, Guillaume Blanchet, pour leur aide et commentaires pertinents.

Finalement, je tiens à remercier de tout coeur mes parents, Sergine Desjardins et Rodrigue Proulx, pour leur soutien indéfectible pendant toutes ces années. Je vous aime beaucoup !

TABLE DES MATIÈRES

LISTE DES ABBRÉVIATIONS

AI : Artificial Intelligence

ML : Machine Learning

PGM : Probabilistic Graphical Model

SVM : Support Vector Machine

KNN : *K* Nearest Neighbour

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1

INTRODUCTION GÉNÉRALE

1.1 La grande obsession

Au début du vingtième siècle, Sewall Wright, Ronald Fisher, et J.B.S. Haldane développent une théorie de l'évolution, la synthèse moderne, basée principalement sur la sélection naturelle, la dérive génétique, et les mutations (Provine, 2001). Suite au progrès en génétique moléculaire, il devient enfin possible dans les années 60 de mesurer la diversité génétique de façon précise (Lewontin et Hubby, 1966). Les résultats sont problématiques pour les théoriciens : il y a beaucoup plus de diversités qu'on s'y attendait (Provine, 2001). La réponse, la théorie neutre, veut que la sélection soit un phénomène marginal (Kimura, 1968 ; King et Jukes, 1969 ; Kimura, 1983). Cette théorie a le mérite d'être simple à paramétriser mais des découvertes modernes en génomique des populations montrent que la sélection est commune (Gillespie, 2004 ; Begun et al., 2007). Malheureusement, nous avons peu d'alternatives théoriques viables au modèle neutre (Hahn, 2008).

Pourquoi la biodiversité est-elle si complexe d'un point de vue théorique ? Il y a plusieurs raisons, tant au niveau du génome qu'au niveau des populations. L'effet d'une mutation varie énormément d'un endroit à l'autre. Sur le même gène, une mutation peut grandement influencer la fonction du gène alors qu'une mutation à proximité aura peu d'influence. Cette complexité complique l'analyse des mutations, qui est souvent réduite à comparer les mutations où le nucléotide d'un codon est changé à celle qui ne change pas le nucléotide (mutations silencieuses) (Hahn, 2008). Cette simplification ignore la majorité des facteurs qui influencent l'effet d'une mutation sur le phénotype (Wilke, 2012).

Au niveau des populations, la sélection varie dans l'espace et le temps alors que l'environnement biotique et abiotique change. Loin d'être une force constante qui pousse certaines mutations vers la fixation ou l'extinction, la sélection change constamment en fonction de l'environnement. Une population a des proies, des compétiteurs, des prédateurs différents d'un endroit à l'autre, ce qui a pour effet de créer des pressions sélectives différentes dans chaque communauté. Ces pressions changent aussi avec le temps alors que la composition de la communauté se transforme (Bell, 2010).

Aux deux facteurs précédents doit s'ajouter le déséquilibre de liaison (Lynch, 2007), qui est de plus en plus considéré comme un facteur important pour comprendre la diversité du génome (Begun et al., 2007). Une mutation positive qui se trouve localisée près d'une mutation désavantageuse va favoriser la progression de cette dernière dans la population. Bref, non seulement la pression sélective sur les traits (et donc les gènes) varie dans le temps et l'espace, mais cette pression a des conséquences sur les loci rapprochés sur le chromosome. L'intégration du déséquilibre de la liaison dans une théorie de la biodiversité moléculaire est une priorité pour les théoriciens mais elle demeure difficile sans l'ajout de plusieurs paramètres (Hahn, 2008).

Dernièrement, la spéciation a une structure spatiale complexe (Gravilets, 2004). La majorité des évènements de spéciation surviennent lorsqu'une population est partiellement ou complètement isolée du reste de l'espèce. C'est la spéciation allopatrique. Rarement, des populations au même endroit deviennent isolées au niveau de la reproduction (spéciation sympatrique) (Coyne et Orr, 2004). J'ai consacré mes premiers articles à développer une version de la théorie neutre de la biodiversité qui intègre la notion d'isolation géographique (Annexes A, B, C). Ultimement, les nouvelles espèces peuvent occuper des niches différentes. La spéciation a donc un rôle important pour comprendre la biodiversité au niveau des communautés.

En bref, la biodiversité est difficile sur le plan théorique parce qu'elle implique beaucoup de facteurs et que nous avons rarement assez d'informations pour paramétriser des modèles plus sophistiqués que des théories neutres. Il est donc difficile de capturer l'essence de la biodiversité avec des modèles théoriques suffisamment simples. Hahn (2008) note : *the consequence of this is that we have tied ourselves into philosophical knots by using null models no one believes but are easily parameterized.* Si l'ambition des théoriciens de découvrir une grande théorie de la biodiversité est vaine, il faut explorer des alternatives. Une avenue possible serait d'utiliser des techniques en intelligence artificielle pour combiner, à l'intérieur d'une base de connaissance unifiée, plusieurs théories. Au lieu de chercher **une** théorie de la biodiversité, notre rôle serait alors de trouver comment combiner plusieurs théories. Pour cette thèse, je vais explorer diverses représentations de la connaissance et leur potentiel pour la synthèse d'idées en écologie, et comment exploiter ces représentations par des techniques d'apprentissages automatiques. J'utiliserai des données sur les interactions entre les espèces pour tester ces techniques.

1.2 Interaction des espèces

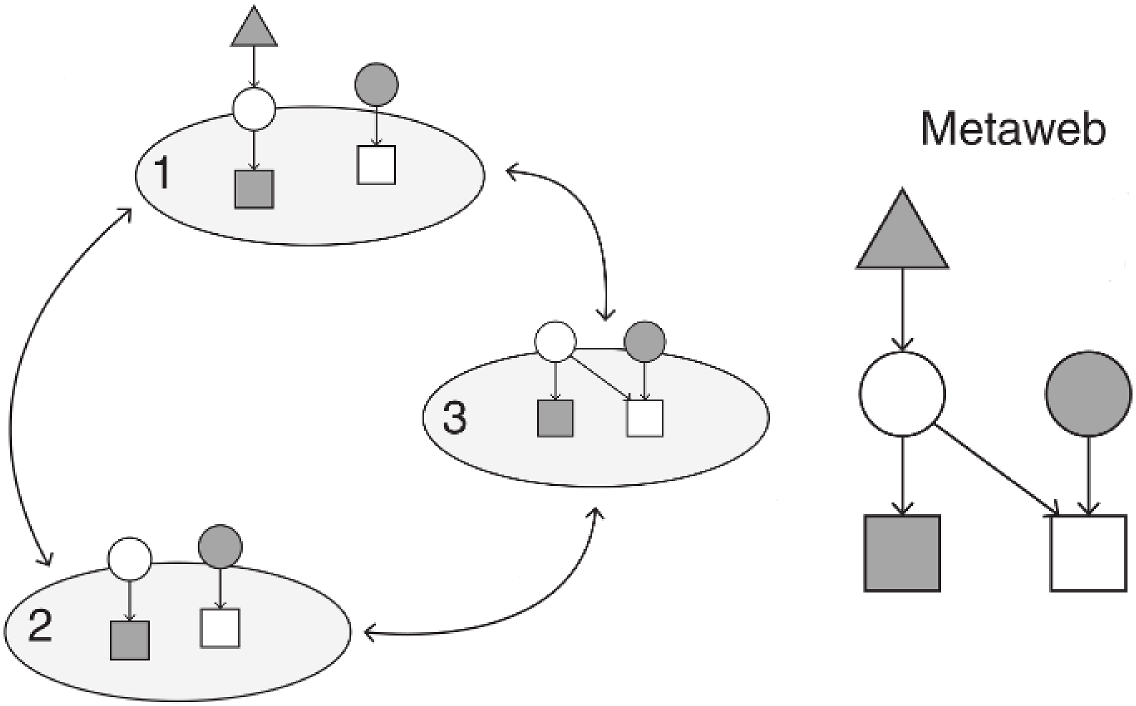
Toutes les données utilisées pour cette thèse se rapportent à l'interaction des espèces (Pimm, 1982) (Figure 1.1). Ce problème a été choisi pour deux raisons. Primo, l'interaction des espèces est une des raisons expliquant pourquoi la biodiversité est si difficile à modéliser (Bartomeus et al., 2016). Une théorie générale de la biodiversité devrait inclure de solides prédictions sur les interactions pour comprendre quelles forces influencent les populations. Secondo, plusieurs membres du laboratoire étudient les interactions, ce qui me donne accès à la fois à des données de qualité mais aussi à des gens capables de les interpréter.

Les données utilisées pour cette thèse proviennent de trois ensembles de données :

- Le premier, assemblé par Idaline Laigle, contient plus de 30 000 interactions observées parmi près de 900 espèces. Pour chaque espèce, nous avons les informations sur 28 traits (la taille, si l'espèce est sous la terre, sa taxonomie, etc.). Ces données sont détaillées à la sous-section 3.4.1 et à la table 3.2.
- Le second, utilisé dans les chapitres 4 et 5, couvre les interactions entre pollinisateurs et plantes. Il contient plus de 700 observations ainsi que plusieurs traits pour les pollinisateurs et les plantes. Il est décrit en détail aux sous-sections 4.3.1 et 5.4.1 ainsi qu'à la table 5.1.
- Le dernier a été assemblé par notre laboratoire et ses collaborateurs, il contient un réseau d'interactions tritrophiques avec des interactions entre parasites et insectes, et des interactions entre ces insectes et des saules. Il a été collecté sur plusieurs sites en Europe et contient des informations sur les précipitations et la température moyenne sur chacun de ces sites. Il est décrit à la section 6.7, la table 6.3, ainsi que dans l'article de Kopelke et al. (2017).

Le problème est intéressant d'un point de vue théorique dans le sens où il peut être approché de plusieurs façons. On peut vouloir prédire une interaction à partir des traits des espèces, en se basant sur les interactions d'une espèce similaire. On peut considérer les pairs d'espèces qui n'ont pas d'interactions comme une absence d'évidence (approche par recommandation) ou

Figure 1.1 – Le metaweb et la représentation des interactions



Le metaweb (à droite) et des réseaux d'interactions locaux (à gauche). Les espèces sont représentées par des formes de couleurs différentes. Deux espèces peuvent interagir dans une communauté locale et non dans une autre, comme par exemple le cercle blanc et le carré blanc, qui ont une interaction dans la communauté 3 mais pas dans la 1. Le metaweb est la collection de toutes les interactions locales. Il est souvent présumé que si une espèce interagit avec une autre dans une communauté, elle le fera aussi dans toutes les autres communautés, mais ce n'est pas le cas. La figure provient de Poisot et al. (2012).

une évidence d'absence (complétion de matrice). Ces données ont cependant des limites. Les interactions sont bivalentes, soit deux espèces ont une interaction, soit elles n'en n'ont pas. Les données ignorent aussi les interactions qui pourraient impliquer plus d'une espèce.

1.3 Deux questions

Ma thèse s'attaque à deux questions : quelle représentation des connaissances permet de représenter les théories en écologie ? Comment réviser automatiquement des théories écologiques formulées avec cette représentation ?

Ces deux questions mènent à deux branches de l'intelligence artificielle : la représentation des connaissances et l'apprentissage automatique (machine learning).

Pour être clair, la question ici n'est pas de savoir comment représenter les **données** écologiques, mais bien de représenter les **théories**. Par exemple, on voudrait pouvoir représenter des théories comme l'équation de Lotka-Volterra, le modèle de niche (Williams et Martinez, 2000) ou des théories probabilistiques comme la théorie neutre de Hubbell (2001).

1.4 Représentation des connaissances

Les théories en écologie se distinguent des modèles mathématiques par leur flexibilité : personne ne s'attend à ce qu'un modèle qui prédit la population d'une espèce soit exact, alors que les modèles en physique et les bases de théorèmes mathématiques s'accomodent facilement de la bivalence (tout est soit vrai, soit faux, sans nuance). Le chapitre 2 explique en détails en quoi la représentation des théories écologiques est un sujet difficile (Russell et Norvig, 2009).

1.5 Apprentissage automatique

Le second ingrédient est l'apprenttisage automatique : la capacité de bâtir des modèles (ou de réviser des modèles existants) automatiquement à partir de données. Il existe deux grandes branches : l'apprentissage supervisé (Figure 1.2) et l'apprentissage non supervisé (Figure 1.3). Ces deux approches seront utilisées lors de la thèse. Selon l'approche supervisée, les données $\mathcal{D}$ sont formées d'un vecteur de *features* $\mathbf{x}$ et d'une sortie y (ce qu'on veut prédire) :

$$\mathcal{D} = \{(\mathbf{x_i}, y_i)\}_{i=0}^{|\mathcal{D}|-1}. \tag{1.1}$$

Si y est un nombre entier, on a un problème de classification et si c'est un nombre réel, on a un problème de régression. La régression n'est pas explorée lors de cette thèse, mais les chapitres 3 et 5 traitent du problème d'interaction des espèces comme un problème de classification. $\mathbf{x}$ représente les traits de deux espèces et y dénote si ces espèces interagissent. Le chapitre

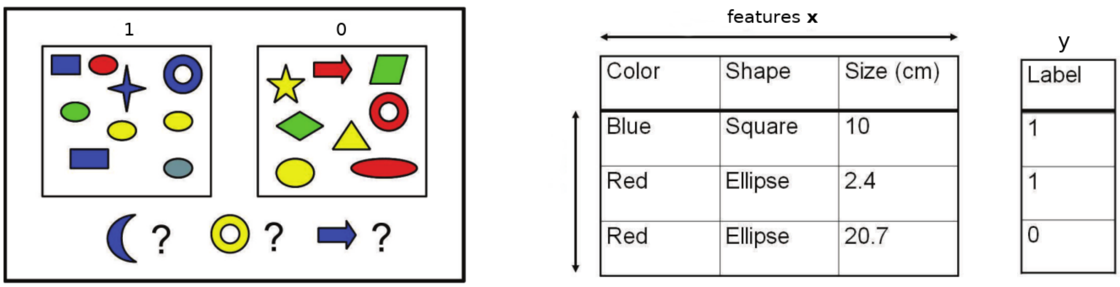
3 utilise plusieurs méthodes standards en apprentissage supervisé comme les *decision trees*, les *support vector machines* et les *random forests* (Murphy, 2012). Le chapitre 5 utilise une méthode que j'ai développé pour apprendre des règles simples en logique floue pour prédire si un pollinisateur et une plante interagissent.

L'approche non supervisée est plus difficile à décrire formellement. Elle implique de trouver des structures intéressantes à l'intérieur des données :

$$\mathcal{D} = \{(\mathbf{x_i})\}_{i=0}^{|\mathcal{D}|-1}. \tag{1.2}$$

Plusieurs méthodes non supervisées sont utilisées. On peut voir le problème des interactions comme un problème de complétion de matrice (chapitres 3 et 4) : la matrice représente les interactions, les trous dans la matrice représentent des paires d'espèces, et l'objectif est de remplir ces trous soit par une interaction ou une non-interaction. On peut aussi voir le problème comme un système de recommandations (chapitre 3), ou chaque espèce a un groupe de proies et on tente de suggérer à cette espèce des proies éventuelles, à la manière dont Netflix ou Amazon tente de prédire les items qui peuvent vous intéresser en se basant sur des utilisateurs qui vous ressemblent ou des items qui ressemblent à celui qui vous intéresse. Cette approche a l'avantage de considérer seulement les évidences positives et d'ignorer les absences d'interactions. Étant donné qu'il est difficile en pratique de confirmer que deux espèces n'ont pas d'interactions, cette méthode est particulièrement appropriée pour étudier les interactions. Finalement, le chapitre 6 explore l'apprentissage de règles relationelles, une autre branche de l'apprentissage non supervisé (Richardson et Domingos, 2006). Les approches relationelles sont expliquées brièvement au chapitre 2, et plus en détails au chapitre 6.

Figure 1.2 – Apprentissage supervisé

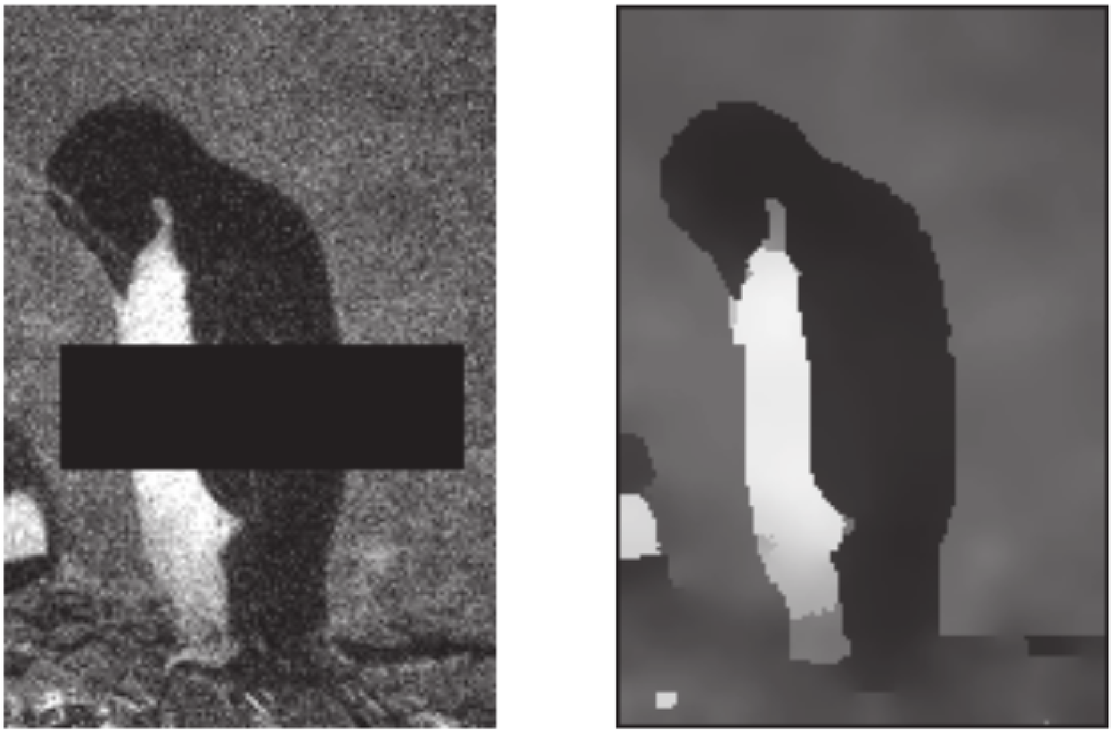


Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

(a) (b)

L'approche supervisée consiste à prédire un *label* y à partir des *features* $\mathbf{x}$. Figure de Murphy (2012).

Figure 1.3 – Apprentissage non supervisé



(a) (b)

L'approche non supervisée est plus difficile à définir, elle consiste à découvrir des structures intéressantes dans les données. Ici, on voit l'exemple de complétion de matrice, ici une image avec un trou. L'objectif est de trouver de bonnes valeurs pour remplir ce trou. Figure de Murphy (2012).

1.6 Le plan de la thèse

La thèse se sépare en sept chapitres : l'introduction, cinq articles, et une conclusion générale. Le chapitre 2 explore la relation entre l'intelligence artificielle et les théories scientifiques, en mettant l'emphase sur les théories en écologie et le besoin d'étudier des représentations plus flexibles que celles adoptées par les mathématiciens. Les chapitres 3 et 4 appliquent des algorithmes standards en intelligence artificielle au problème d'interactions des espèces. Bien que ces deux chapitres ne touchent pas directement à la question de révision des théories scientifiques, les leçons apprises lors de ces recherches m'ont servi lors de mes travaux sur la révision automatique. Le chapitre 5 étudie une nouvelle approche pour apprendre des règles avec la logique floue. Appliquée à des interactions pollinisateurs-plantes, cette technique nous permet de réviser une théorie de base simple pour arriver à un modèle à la fois plus clair et plus efficace que les techniques standards en intelligence artificielle. Le chapitre 6 explore la logique de Markov. Cette technique n'a pas bien fonctionnée, mais elle donne d'importantes pistes pour l'étude d'approches capables d'intégrer des théories mathématiques et la révision automatique. Ce chapitre termine avec une discussion sur les prochaines étapes pour progresser sur la question de révision automatique de théories écologiques. La conclusion générale boucle la thèse en retournant sur les deux questions posées dans cette introduction.

CHAPITRE 2

ECOLOGICAL SYNTHESIS AND ARTIFICIAL INTELLIGENCE

2.1 Description de l'article et contribution

Avant de pouvoir réviser des théories écologiques avec des techniques d'intelligence artificielle, il faut choisir une représentation de la connaissance pour les théories écologiques. Dans cette contribution, nous explorons diverses représentations de la connaissance et comment celles-ci peuvent servir à encoder les théories écologiques. Les mathématiciens ont depuis plusieurs décennies assemblé de larges bases de connaissances en utilisant une forme de logique (predicate logic). Malheureusement, cette forme de logique est trop inflexible pour l'écologie. On ne s'attend pas, contrairement aux mathématiciens et aux physiciens, à ce que nos théories soient toujours exactes. La complexité des écosystèmes nous force à accepter des théories imparfaites, mais utiles. Combiner ces théories dans un tout est d'autant plus complexe que relativement peu de travail a été fait pour assembler les théories inexactes. Cet article discute de différents systèmes de raisonnement allant des probabilités à la logique pure, ainsi que la logique floue. Nous expliquons pourquoi l'intégration de théories écologiques et la révision de celles-ci requièrent que l'on réfléchisse à la structure des connaissances écologiques.

J'ai conçu et fait la recherche de cet article de perspective. Dominique Gravel et Timothée Poisot m'ont assisté pour la révision. L'article a été publié sur bioRxiv. Cet article servira de base pour une revue de littérature sur l'union de l'intelligence artificielle et des théories écologiques qui sera soumis à un numéro spécial de *Frontiers in Ecology and Evolution*.

2.2 Ecological Synthesis

Artificial Intelligence presents an important paradigm shift for science. Science is traditionally founded on theories and models, most often formalized with mathematical formulas handcrafted by theoretical scientists and refined through experiments. Machine learning, an important branch of modern Artificial Intelligence, focuses on learning from data. This leads to a fundamentally different approach to model-building : we step back and focus on the design of algorithms capable of building models from data, but the models themselves are not designed by humans. This is even more true with deep learning, which requires little engineering by hand and is responsible for many of Artificial Intelligence's spectacular successes (LeCun et al., 2015). In contrast to logic systems, knowledge from a deep learning model is difficult to understand, reuse, and may involve up to a billion parameters (Coates et al., 2013). On the other hand, probabilistic machine learning techniques such as deep learning offer an opportunity to tackle large complex problems that are out of the reach of traditional theory-making. It is possible that the more intuition-like (LeCun et al., 2015) reasoning performed by deep learning systems is mostly incompatible with the logic formalism of mathematics. Yet recent studies have shown that deep learning can be useful to logic systems and vice versa. Success at unifying different paradigms of Artificial Intelligence from logic to probability theory offers unique opportunities to combine data-driven approaches with traditional theories. These advancements are susceptible to impact significantly theoretical work on ecology and evolution, allowing for the integration of data and theories in unified knowledge bases.

Ever since molecular biology allowed for precise measurements of molecular diversity, theoreticians have struggled to explain why populations are so diverse (Lewontin et Hubby, 1966 ; Gillespie, 2004 ; Hahn, 2008). Integrating selection proved hard enough for neutral theories to dominate in practice since few solid alternatives are easy enough to parameterize (Hahn, 2008). Theoretical ecology has had a similar struggle with similar consequences, Hubbell's neutral theory of biodiversity is almost exactly the same as Kimura's neutral theory of molecular evolution, with alleles being different species and point mutation being replaced by point speciation (Kimura, 1983 ; Hubbell, 2001).

Recent work in population genomics show the importance of linkage Begun et al. (2007) and point to a world where selection is a constant presence in the genome, not as a simple force pushing mutations toward extinction or fixation, but as a stochastic mechanism varying in time and space Bell (2010). In *Toward a Selection Theory of Molecular Evolution*, Matthew Hahn noted how population geneticists became dependant on neutral theories and have effectively "tied [themselves] into philosophical knots by using null models no one believes but are easily parameterized" Hahn (2008). Thus, interestingly, the reason a truly unified theory of biodiversity has proven elusive is not because we do not know the basic mechanisms involved, it is because we do not how to effectively build a theory simple enough to be usable.

Our thesis is that biodiversity is unlikely to simplify to a single elegant equation. In that case, we could turn to and tools from A.I. to help us integrate various ecological and evolutionary ideas into a single unified knowledge base. Instead of looking for a single theory, we would grow knowledge, similarly to how mathematicians have build large knowledge bases of theorems (Kaliszyk et Urban, 2015). However, ecology and evolution are not pure mathematics, we do not expect our ideas to be as precise as theorems. Thus, the question becomes : what is the correct knowledge representation for ideas in ecology and evolution ?

2.3 A.I. and knowledge representation

Science would greatly benefit from a unification of Artificial Intelligence with traditional mathematical theories. Modern research at the intersection of logic, probability theory, and fuzziness yielded rich representations increasingly capable of formalizing scientific knowledge. Such formal corpus could both include hand-crafted theories from Einstein's $e = mc^2$ to the Breeder's equation (Rice, 2004), but also harness modern A.I. algorithms for testing and learning. Comprehensive synthesis is difficult in fields like biology, which have not been reduced to a small set of formulas. For example, while we have a good idea of the underlying forces driving evolution, we struggle to build effective predictive models of molecular evolution (Hahn, 2008). This is likely because selection changes in time and space (Bell, 2010), which brings population, community, and ecosystem ecology into the mix. Ecology also has a porous frontier with evolution : speciation is a common theme in community ecology theory (Desjardins-Proulx et Gravel, 2012).

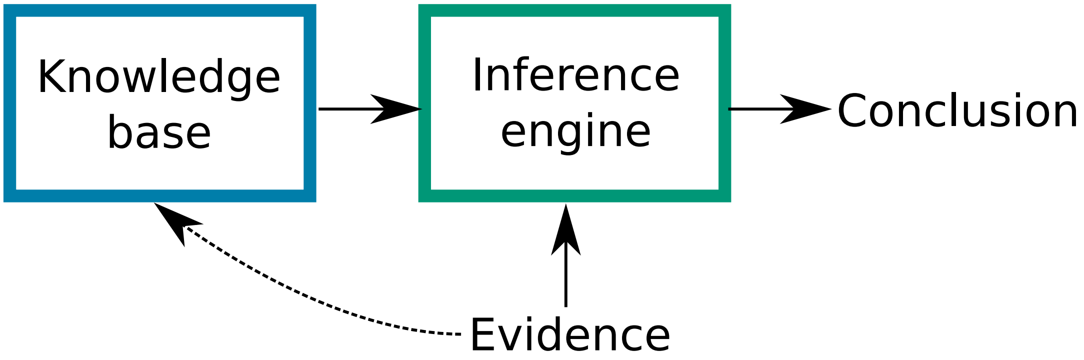
From a theoretical perspective, work to formalize scientific theories would reveal much about the nature of our theories. Surely, scientific theories require more flexibility than mathematical corpora of knowledge, which are based on pure logic. From a practical standpoint, a formal representation both offers ways to test large corpora of knowledge and extend it with A.I. techniques. This is arguably the killer feature of a formal representation of scientific knowledge : allowing A.I. algorithms to search for revisions, extensions, and discover new rules. This is not a new ambition. Generic techniques for rule discovery were well-established in the 1990s (Muggleton et de Raedt, 1994). Unfortunately, these techniques were based on pure logic, and purely probabilistic approaches to revision cannot handle mathematical theories. Recent experiences in linguistics has shown that building a knowledge base capable of handling several problems at the same time yielded better results than attacking the problem in isolation because of the problems' interconnectedness (Yoshikawa et al., 2009). Biology, as a complex field made of more-or-less arbitrary subfields, could gain important insights from unified approach to knowledge combining A.I. techniques with traditional mathematical theories.

2.4 A quick tour of knowledge representations

Deep learning is arguably the dominant approach in probabilistic machine learning, a branch of A.I. focused on learning models from data (Goodfellow et al., 2016). The idea of deep learning is to learn multiple levels of compositions. If we want to learn to classify images for instance, the first layer of the deep learning network will read the input, the next layer will capture lines, the next layer will capture contours based on lines, and then more complex shapes based on contours, and so on (Goodfellow et al., 2016). In short, the layers of the network begin with simple concepts, and then compose more complicated concepts from simpler ones (Bengio et al., 2013). Deep learning has been used to solve complex computer science problems like playing Go at the expert level (Silver et al., 2016), but it is also used for more traditional scientific problems like finding good candidate molecules for drugs, predicting how changes in the genotype affect the phenotype (Leung et al., 2016), or just recently to solving the quantum many-body problem (Carleo et Troyer, 2017).

In contrast, traditional scientific theories and models are mathematical, or logic-based. Einstein's $e = mc^2$ established a logical relationship between energy e, mass m, and the speed of light c. This mathematical knowledge can be reused : in any equation with energy, we could replace e with mc^2. This ability of mathematical theories to establish precise relationships between concepts, which can then be used as foundations for other theories, is fundamental to how science grows and forms an interconnected corpus of knowledge. Furthermore, these theories

Figure 2.1 – McCarthy's "Program with Common Sense" established the importance of separating knowledge from reasoning.



In his 1959's "Programs with Common Sense", McCarthy established the importance of separating knowledge and reasoning (McCarthy, 1959). The knowledge base stores knowledge, while the inference engine exploits it, along with evidence, to reach conclusions. McCarthy believed the knowledge base should store rules in some form of predicate logic, but the idea of separation of knowledge and reasoning holds with other representations as well (Bayesian networks, detailed in figure 2.2, are probabilistic knowledge bases (Darwiche, 2009)). The dotted line represents automatic theory revision, where evidence is used not to a answer query but to discover new knowledge or revise existing theories.

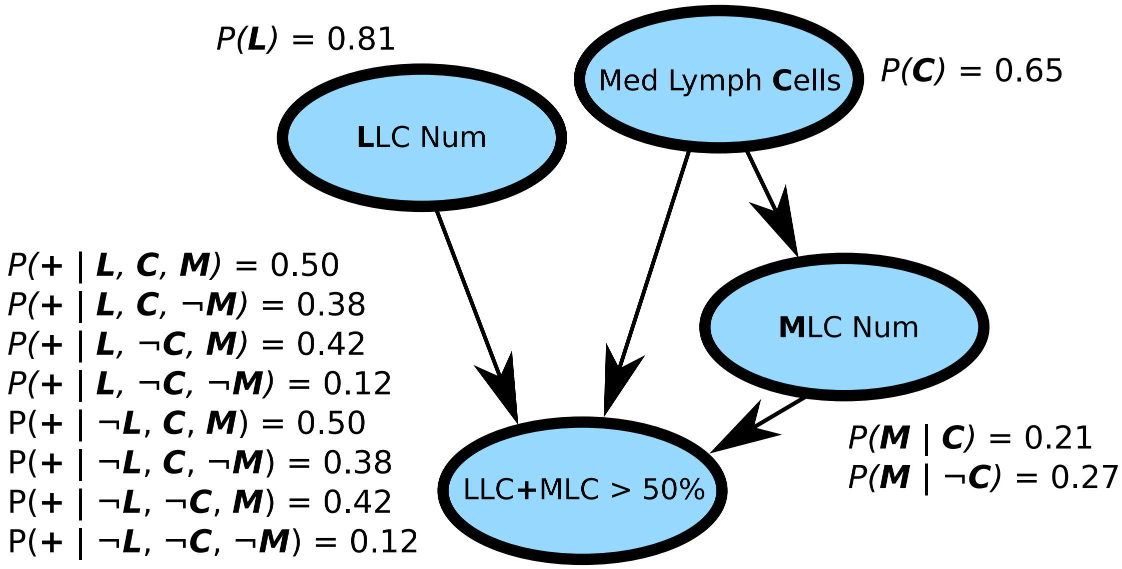
are compact and follow science's tradition of preferring theories as simple as possible. There are many different foundations for logic systems. Predicate logic is a good starting point : it is based on predicates, which are functions of terms to a truth value. For example, the predicate $PreyOn$ could take two species, a location, and return true if the first species preys on the second at that location, like $PreyOn(Wolverine, Squirrel, Quebec)$. Terms are either *constants* such as 1, π, or $Wolverine$, *variables* that range over constants, such as x or $species$, or *functions* that map terms to terms, such as additions, multiplication, integration, differentiation. In $e = mc^2$, the equal sign $=$ is the predicate, e and m are variables, c and 2 are constants, and there are two functions : the multiplication of m by c^2 and the the exponentiation of c by 2. The key point is that such formalism lets us describe compact theories and understand precisely how different concepts are related. Complex logic formulas are built by combining predicates with connectives such as negation $\neg$, "and" $\wedge$, "or" $\vee$, "implication" $\Rightarrow$. We could have a rule to say that predation is asymmetrical $s_x \neq s_y \wedge PreyOn(s_x, s_y, l) \Rightarrow \neg PreyOn(s_y, s_x, l)$, or define the classical Lotka-Volterra :

$$\frac{dx}{dt} = \alpha x - \beta xy \wedge \frac{dy}{dt} = \delta xy - \gamma y, \tag{2.1}$$

where x and y are the population sizes of the prey and the predator, respectively, α, β, δ, γ are constants, and the time differential d/dt, multiplication and subtraction are functions. Equality $(=)$ is the sole predicate in this formula. Both predicates are connected via $\wedge$ ("and"). Not all logic formulas have mathematical functions. Simple logic rules such as $Smoking(p) \Rightarrow Cancer(p)$ ("smoking causes cancer") are common in expert systems.

Artificial Intelligence researchers have long being interested in logic systems capable of scientific discoveries, or simply capable of storing scientific and medical knowledge in a single coherent system (Figure 2.1). DENDRAL, arguably the first expert system, could form hypotheses to help identify new molecules using its knowledge of chemistry (Lindsay et al., 1993). In the 1980s, MYCIN was used to diagnose blood infections (and did so more accurately than professionals) (Buchanan et Shortliffe, 1984). Both systems were based on logic, with MYCIN adding a "confidence factor" to its rules to model uncertainty. Other expert systems were based on probabilistic graphical models (Koller et Friedman, 2009), a field that unites graph theory with probability theory to model the conditional dependence structure of random variables (Koller et Friedman, 2009 ; Barber, 2012). For example, Munin had a network of more than 1000 nodes to analyze electromyographic data (et al., 1996), while PathFinder assisted medical professional for the diagnostic of lymph-node pathologies (Heckerman et Nathwani, 1992) (Figure 2.2). While these systems performed well, they are both too simple to store generic scientific knowledge and too static to truly unify Artificial Intelligence with scientific research. The ultimate goal is to have a representation rich enough to encode both logic-mathematical and probabilistic scientific knowledge.

Figure 2.2 – Reasoning in Bayesian networks.


P(**L**) = 0.81
LLC Num
Med Lymph **C**ells
P(**C**) = 0.65
MLC Num
LLC+MLC > 50%
P(+ | **L**, **C**, **M**) = 0.50
P(+ | **L**, **C**, ¬**M**) = 0.38
P(+ | **L**, ¬**C**, **M**) = 0.42
P(+ | **L**, ¬**C**, ¬**M**) = 0.12
P(+ | ¬**L**, **C**, **M**) = 0.50
P(+ | ¬**L**, **C**, ¬**M**) = 0.38
P(+ | ¬**L**, ¬**C**, **M**) = 0.42
P(+ | ¬**L**, ¬**C**, ¬**M**) = 0.12
P(**M** | **C**) = 0.21
P(**M** | ¬**C**) = 0.27

A Bayesian network with four binary variables and possible conditional probability tables. These four nodes were taken from PathFinder, a Bayesian network with more than 1000 nodes used to help diagnose blood infections (Heckerman et Nathwani, 1992). The nodes represent four variables related to blood cells and are denoted by a single character (in bold in the figure) : $C, M, L, +$. All variables are binary, and negation is denoted with $\neg$. Since $P(\neg x|\mathbf{y}) = 1 - P(x|\mathbf{y})$, we need only $2^{|Pa(x)|}$ parameters per nodes, with $|Pa(x)|$ being the number of parents of node x. The structure of Bayesian networks both highlights the conditional independence assumptions of the distribution and reduces the number of parameters for learning and inference. Example query : $P(L, \neg C, M, \neg +) = P(L)P(\neg C)P(M|\neg C)P(\neg + |L, \neg C, M) = 0.81 \times (1 - 0.65) \times 0.27 \times (1 - 0.42) = 0.044$. See (Darwiche, 2009) for a detailed treatment of Bayesian networks and (Koller et Friedman, 2009) for a more general reference on probabilistic graphical models.

2.5 Beyond monolithic systems

In terms of representation, expert systems generally used a simple logic system, not powerful enough to handle uncertainty, or purely probabilistic approaches unable to handle complex mathematical formulas. In terms of flexibility, the expert systems were hand-crafted by human experts. After the experts established either the logic formulas (for logic systems like DENDRAL) or probabilistic links (in the case of systems like Munin), the expert systems act as static knowledge bases, capable of answering queries but unable of discovering new rules and relationships. While no system has completely solved these problems yet, much energy has been put in unifying logic-based systems with probabilistic approaches (Getoor et al., 2007). Also, several algorithms have been developed to learn new logic rules (Muggleton et de Raedt, 1994), find the probabilistic structure in a domain with several variables (Yuan et Malone, 2013), and even transfer knowledge between tasks (Mihalkova et al., 2007). Together, these discoveries bring us closer to the possibility of flexible knowledge bases contributed both by human experts and Artificial Intelligence algorithms. This has been made possible in great part by efforts to unify three distinct languages : probability theory, predicate logic, and fuzzy logic (Figure 2.3).

The core idea behind unified logic/probabilistic languages is that formulas can be weighted, with higher values meaning we have greater certainty in the formula. In pure logic, it is impossible to violate a single formula. With weighted formulas, an assignment of concrete values to variables is only *less likely* if it violates formulas. The higher the weight of the formula violated, the less likely the assignment is. It is conjectured that all perfect numbers are even ($\forall x : Perfect(x) \Rightarrow Even(x)$), if we were to find a single odd perfect number, that formula would be refuted. It makes sense for mathematics but for many disciplines, such as biology, important principles are only expected to be true *most* of the times. To illustrate, in ecology, predators generally have a larger body weight than their preys, which can expressed in predicate logic as $PreyOn(predator, prey) \Rightarrow M(predator) > M(prey)$, with $M(x)$ being the body mass of x. This is obviously false for some assignments, for example $predator : greywolf$ and $prey : moose$. However, it is useful knowledge that underpins many ecological theories (Williams et Martinez, 2000). When our domain involves a great number of variables, we should expect useful rules and formulas that are not always true.

The idea of weighted formulas is not new. Markov logic, invented a decade ago, allows for logic formulas to be weighted (Richardson et Domingos, 2006 ; Domingos et Lowd, 2009). It supports algorithms to add weights to existing formulas given a data-set, learn new formulas or revise existing ones, and answer probabilistic queries. For example, Yoshikawa et al. used Markov logic to understand how events in a document were time-related (Yoshikawa et al., 2009). Their research is a good case study of interaction between traditional theory-making and artificial intelligence. The formulas they used as a starting point were well-established logic rules to understand temporal expressions. From there, they used Markov logic to weight the rules, adding enough flexibility to their system to beat the best approach of the time. Brouard et al. (Brouard et al., 2013) used Markov logic to understand gene regulatory network, noting how the resulting model provided clear insights, in contrast to more traditional machine learning techniques. Expert systems can afford to make important sacrifices to flexibility in exchange for a simple representation. Yet, a system capable of representing a large body of scientific knowledge will require a great deal of flexibility to accommodate various theories. While a step in the right direction, even Markov logic may not be powerful enough.

2.6 Case study : The niche model

To show some of the difficulties of representing scientific knowledge, we will build a small knowledge base for an established ecological theory : the niche model of trophic interactions (Williams et Martinez, 2000). The first iteration of the niche model posits that all species are described by a niche position N (their body size for instance) in the $[0, 1]$ interval, a diet D in the $[0, N]$ interval, and a range R such that a species preys on all species with a niche in the $[D - R/2, D + R/2]$ interval. We can represent these ideas with three formulas :

$$\forall x, y : \neg PreyOn(x, y), \tag{2.2a}$$

$$\forall x : D(x) < N(x), \tag{2.2b}$$

$$\forall x, y : PreyOn(x, y) \Leftrightarrow D(x) - R(x)/2 < N(y) \wedge N(y) < D(x) + R(x)/2, \tag{2.2c}$$

where $\forall$ reads *for all* and $\Leftrightarrow$ is logical equivalence (it is true if and only if both sides of the operator have the same truth value, so for example $False \Leftrightarrow False$ is true and $True \Leftrightarrow False$ is false). As pure logic, this knowledge base makes little sense. Formula 2.2a is obviously not true all the time. It is mostly true, since most pairs of species do not interact. We could also add that cannibalism is rare $\forall x : \neg PreyOn(x, x)$ and that predator-prey are generally asymmetrical $\forall x, y : PreyOn(x, y) \Rightarrow \neg PreyOn(y, x)$. In hybrid probabilistic/logic approaches like Markov logic, these formulas would have a weight that essentially defines a marginal probability (Domingos et Lowd, 2009 ; Jain, 2011). Formulas that are often wrong are assigned a lower weight but can still provide useful information about the system. The second formula says that the diet is smaller than the niche value. The last formula is the niche model : species x preys on y if and only if species y's niche is within the diet interval of x.

So far so good ! Using Markov logic networks and a data-set, we could learn a weight for each formula in the knowledge base. This step alone is useful and provide insights into which formulas hold best. With the resulting weighted knowledge base, we could make probabilistic queries and even attempt to revise the theory automatically. We could find, for example, that the second rule does not apply to parasites or some group and get a revised rule such as $\forall x : \neg Parasite(x) \Rightarrow D(x) < N(x)$. However, Markov logic networks struggle when the predicates cannot easily return a simple true-or-false truth values. For example, let's say we wanted to express the idea that when populations are small and have plenty of resources, they grow exponentially (Kot, 2001).

$$\forall x, l, t : SmallP(x, l, t) \text{ and } Resources(x, l, t) \Rightarrow P(x, l, t + 1) = G(x) \times P(x, l, t), \quad (2.3)$$

where $P(x, l, t)$ is the population size of species x in location l at time t, G is the rate of growth, $SmallP$ is whether the species has a small population and $Resources$ whether it has resources available. The problem with hybrid probabilistic/logic approach is that predicates do not capture the inherent vagueness well. We can establish an arbitrary cutoff for what a small population is, for example by saying that if it is less than 10% the average population size for the species, it is small. Similarly, resource availability is not a binary thing, there is a world of grey between starvation and satiety. Perhaps worst of all, the prediction that $P(x, l, t + 1) = G(x) \times P(x, l, t)$ is almost certainly never be exactly true. If we predict 94 rabbits and observe 93, the formula is false. Weighted formulas help us understand *how often a rule is true*, but in the end the formula has to give a binary truth value : true or false, there is no place for vagueness.

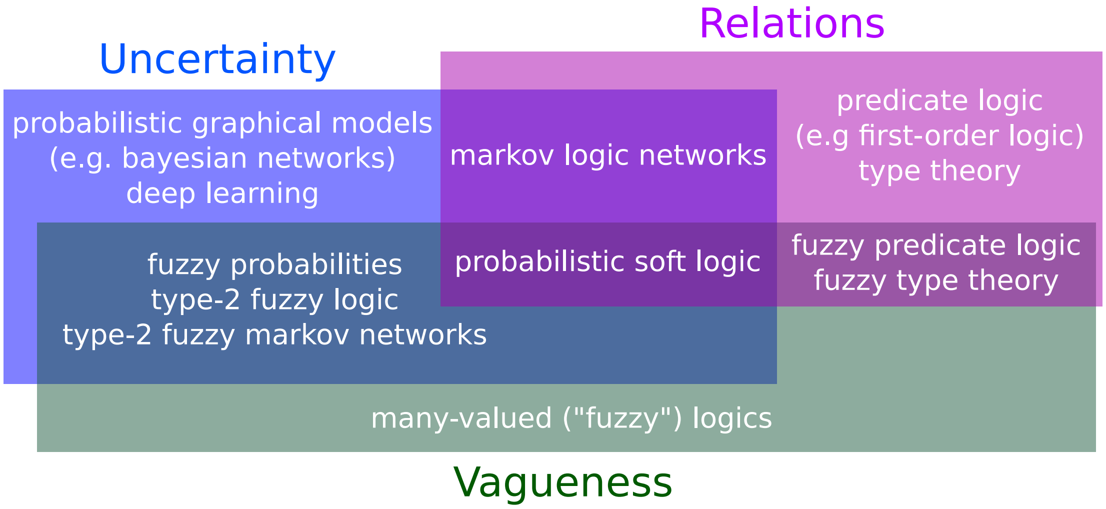
Fuzzy sets and many-valued ("fuzzy") logics were invented to handle vagueness (Zadeh, 1965 ; Hájek, 1998 ; Bergmann, 2008 ; Behounek et al., 2011). In practice, it simply means that predicates can return any value in the $[0, 1]$ closed interval instead of only true and false. It is used in both probabilistic soft logic (Kimmig et al., 2012 ; Bach et al., 2015) and deep learning approaches to predicate logic (Zahavy et al., 2016 ; Hu et al., 2016). For our formula 2.3, $SmallP$ could be defined as $1 - P(x, l, t)/P_{max}(x)$, where $P_{max}(x)$ is the largest observed population size for the species. $Resources$ could take into account how many preys are available, and $P(x, l, t + 1) = G(x) \times P(x, l, t)$ would return a truth value based on how close the observed population size is the predicted population size. Fuzzy logic then defines how operators such as and and $\Rightarrow$ behave with fuzzy values.

Both Markov logic networks and probabilistic soft logic define a probability distribution over logic formulas, but what about the large number of probabilistic models ? For example, the niche model has a probabilistic counter-part defined as (Williams et al., 2010) :

$$\forall x, y : PPreyOn(x, y) = \alpha \times \exp\left[-\left(\frac{N(y) - D(x)}{R(x)/2}\right)^2\right], \tag{2.4}$$

where $PPreyOn(x, y)$ is the probability that x preys on y. Again, this formula is problematic in Markov logic because we cannot easily force the equality into a binary true-or-false, but fuzziness can help model the nuance of probabilistic predictions. There are no algorithms yet to learn and revise rules in many-valued logic.

Figure 2.3 – Various reasoning languages and their ability to model uncertainty, vagueness, and relations.



The size of the rectangles has no meaning. **In the blue rectangle** : languages capable of handling uncertainty. Probabilistic graphical models combine probability theory with graph theory to represent complex distributions (Koller et Friedman, 2009). Deep learning is, strictly speaking, more general than its usual probabilistic interpretation, but it is arguably the most popular probabilistic Artificial Intelligence approach at the moment (Goodfellow et al., 2016). Alternatives to probability theory for reasoning about uncertainty include possibility theory and Dempster-Shafer belief functions, see (Halpern, 2003) for an extended discussion. **In the green rectangle** : Fuzzy logic extends standard logic by allowing truth values to be anywhere in the $[0,1]$ interval. Fuzziness models vagueness and is particularly popular in linguistics, engineering, and bioinformatics, where complex concepts and measures tend to be vague by nature. See (Kosko, 1990) for a detailed comparison of probability and fuzziness. **In the purple rectangle** : relations, as in : mathematical relations between objects. Even simple mathematical ideas, such as the notion that all natural numbers have a successor ($\forall x \exists y : y = x + 1$), requires relations. *Predicate* and *Relation* are synonymous in this context. Alone, these reasoning languages are not powerful enough to express scientific ideas. We must thus focus on what lies at their intersection. Type-2 Fuzzy Logic is a fast-expanding (Sadeghian et al., 2014) extension to fuzzy logic, which, in a nutshell, models uncertainty by considering the truth value itself to be fuzzy (Mendel et Bob John, 2002 ; Zeng et Liu, 2008). Markov logic networks (Richardson et Domingos, 2006 ; Domingos et Lowd, 2009) extends predicate logic with weights to unify probability theory with logic. Probabilistic soft logic (Kimmig et al., 2012 ; Bach et al., 2015) also has formulas with weights, but allow the predicates to be fuzzy, i.e. have truth values in the $[0,1]$ interval. Some recent deep learning studies also combine all three aspects (Garnelo et al., 2016 ; Hu et al., 2016).

2.7 Where's our unreasonably effective paradigm ?

Wigner's *Unreasonable Effectiveness of Mathematics in the Natural Sciences* led to important discussions about the relationship between physics and its main representation (Wigner, 1960 ; Hamming, 1980). The Mizar Mathematical Library and the Coq library (The Coq development team, 2004) host tens of thousands of mathematical propositions to help build and test new proofs. In complex domains with many variables, Halevy et al. argued for the *Unreasonable Effectiveness of Data* (Halevy et al., 2009), noting that simple algorithms, when fed large amount of data, would do wonder. High-dimensional problems like image imputation, where an algorithm has to fill missing parts from an image, require hundred of thousands of training images to be effective. Goodfellow et al. noted that roughly 10 000 data-points per possible labels were necessary to train deep neural networks (Goodfellow et al., 2016). These approaches are unsatisfactory for fields like biology where theories and principles are seldom exact. We cannot afford the pure logic-based knowledge representations favoured by mathematicians and physicists, and fitting a model to data is a different task than building a corpus of interconnected knowledge.

Fortunately, we do not need to choose between mathematical theories, probabilistic models, and learning. New inventions such as Markov logic networks and probabilistic soft logic are moving Artificial Intelligence toward rich representations capable of formalizing and even extending scientific theories. This is a great opportunity for synthesis. There are still problems : inference is often difficult in those rich representations. Recently, Garnelo et al. (Garnelo et al., 2016) designed a prototype to extract logic formulas from a deep learning system, while Hu et al. (Hu et al., 2016) created a framework to learn predicate logic rules using deep learning. Both studies used flexible fuzzy predicates and weighted formulas while exploiting deep learning' ability to model complex distributions via composition. The end result is a set of clear and concise weighted formulas supported by deep learning for scalable inference. The potential for science is important. Not only these new researches allow for deep learning to interact with traditional theories, but it opens many exciting possibilities, like the creation of large databases of scientific knowledge. The only thing stopping us from building a unified corpus of, say, ecological knowledge, is that normal pure-logic systems are too inflexible. They do not allow imperfect, partially-true theories, which are fundamental to many sciences. Recent developments in Artificial Intelligence make these corpora of scientific knowledge possible for complex domains, allowing us to combine a traditional approach to theory with the power of Artificial Intelligence.

It is tempting to present deep learning as a threat to traditional theories. Yet, there is a real possibility that the union of Artificial Intelligence techniques with mathematical theories is not only possible, but would help the integration of knowledge across various disciplines. Otherwise, short of discovering a small set of elegant theories, what is our plan to combine ideas from ecosystem ecology, community ecology, population ecology, and evolution ?

2.8 References

Bach, S., Broecheler, M., Huang, B., Getoor, L. (2015). Hinge-loss markov random fields and probabilistic soft logic *arXiv : 1505.04406*.

Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.

Begun, D., Holloway, A., Stevens, K., Hillier, L., Poh, Y., Hahn, M., Nista, P., Jones, C., Kern, A., Dewey, C., Pachter, L., Myers, E., Langley, C. (2007). Population genomics : Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLOS Biology *5*, e310.

Behounek, L., Cintula, P., Hájek, P. (2011). Introduction to mathematical fuzzy logic. In : Cintula, P., Hájek, P., Noguera, C. (Eds.), Handbook of Mathematical Fuzzy Logic volume 1. College Publications, London, Ch. 1, pp. 1–101.

Bell, G. (2010). Fluctuating selection : the perpetual renewal of adaptation in variable environments. Phil. Trans. R. Soc. B *365*, 87–97.

Bengio, Y., Courville, A., Vincent, P. (2013). Representation learning : A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence *35*, 1798–1828.

Bergmann, M. (2008). An introduction to many-valued and fuzzy logic. Cambridge University Press.

Brouard, C., Vrain, C., Dubois, J., Castel, D., D, M., d'Alche Buc, F. (2013). Learning a markov logic network for supervised gene regulatory network inference. BMC Bioinformatics *14*, 273.

Buchanan, B., Shortliffe, E. (1984). Rule-based Expert Systems : The Mycin experiments of the Stanford Heuristic Programming Project. Addison-Wesley.

Carleo, G., Troyer, M. (2017). Solving the quantum many-body problem with artificial neural networks. Science *355*, 602–606.

Coates, A., Huval, B., Wang, T., Ng, D. W. A., Catanzaro, B. (2013). Deep learning with COTS HPC systems. Journal of Machine Learning Research Workshop and Conference Proceedings *28*, 1337–1345.

Darwiche, A. (2009). Modeling and Reasoning with Bayesian Networks. Cambridge University Press.

Desjardins-Proulx, P., Gravel, D. (2012). A complex speciation-richness relationship in a simple neutral model. Ecology and Evolution *2*, 1781–1790.

Domingos, P., Lowd, D. (2009). Markov Logic : An Interface Layer for Artificial Intelligence. Morgan & Claypool Publishers.

et al., S. A. (1996). Evaluation of the diagnostic performance of the expert EMG assistant MUNIN. Electroencephalogr Clin Neurophysiol *101*, 129–144.

Garnelo, M., Arulkumaran, K., Shanahan, M. (2016). Towards deep symbolic reinforcement learning *arXiv :1609.05518v2*.

Getoor, L., Friedman, N., Koller, D., Pfeffer, A., Taskar, B. (2007). Probabilistic relational models. In : Getoor, L., Taskar, B. (Eds.), Introduction to Statistical Relational Learning. MIT Press.

Gillespie, J. H. (2004). Population Genetics : A Concise Guide, 2nd Edition. Hopkins Fulfillment Service.

Goodfellow, I., Bengio, Y., Courville, A. (2016). Deep Learning. MIT Press.

Hahn, M. (2008). Toward a selection theory of molecular evolution. Evolution *76*, 255–265.

Halevy, A., Norvig, P., Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems *24*, 8–12.

Halpern, J. (2003). Reasoning about Uncertainty. The MIT Press.

Hamming, R. (1980). The unreasonable effectiveness of mathematics. The American Mathematical Monthly *87*, 81–90.

Heckerman, D., Nathwani, B. (1992). An evaluation of the diagnostic accuracy of Pathfinder. Computers and Biomedical Research *25*, 56–74.

Hu, Z., Ma, X., Liu, Z., Hovy, E., Xing, E. (2016). Harnessing deep neural networks with logic rules *arXiv :1603.06318*.

Hubbell, S. P. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Vol. 32 of Monographs in Population Biology. Princeton University Press.

Hájek, P. (1998). Metamathematics of Fuzzy Logic. Springer Netherlands.

Jain, D. (2011). Knowledge Engineering with Markov Logic Networks : A Review. In : DKB 2011 : Proceedings of the Third Workshop on Dynamics of Knowledge and Belief.

Kaliszyk, C., Urban, J. (2015). Learning-assisted theorem proving with millions of lemmas. Journal of Symbolic Computation *69*, 109–128.

Kimmig, A., Bach, S., Broecheler, M., Huang, B., Getoor, L. (2012). A short introduction to probabilistic soft logic. In : Proceedings of the NIPS Workshop on Probabilistic Programming.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Koller, D., Friedman, N. (2009). Probabilistic Graphical Models. The MIT Press.

Kosko, B. (1990). Fuzziness vs probability. Int J General Systems *17*, 211–240.

Kot, M. (2001). Elements of Mathematical Ecology. Cambridge University Press.

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. Nature , 436–444.

Leung, M., Delong, A., Alipanahi, B., Frey, B. (2016). Machine learning in genomic medicine : A review of computational problems and data sets. Proceedings of the IEEE *104*, 176–197.

Lewontin, R., Hubby, J. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics *54*, 595–609.

Lindsay, R., Buchanan, B., Feigenbaum, E., Lederberg, J. (1993). Dendral : A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence *61*, 209–261.

The Coq development team (2004). The Coq proof assistant reference manual. LogiCal Project, version 8.0.
URL http://coq.inria.fr

McCarthy, J. (1959). Programs with common sense.

Mendel, J., Bob John, R. (2002). Type-2 fuzzy sets made simple. IEEE Transactions on Fuzzy Systems *10*, 117–127.

Mihalkova, L., Huynh, T., Mooney, R. (2007). Mapping and revising markov logic networks for transfer learning. In : Proceedings of the 22nd Conference on Artificial Intelligence. pp. 608–614.

Muggleton, S., de Raedt, L. (1994). Inductive logic programming : Theory and methods. The Journal of Logic Programming *19-20*, 629–679.

Rice, S. (2004). Evolutionary Theory : Mathematical and Conceptual Foundations. Sinauer.

Richardson, M., Domingos, P. (2006). Markov logic networks. Machine Learning *62*, 107–136.

Sadeghian, A., Mendel, J., Tahayori, H. (2014). Advances in Type-2 Fuzzy Sets and Systems. Springer.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489.

Wigner, E. (1960). The unreasonable effectiveness of mathematics in the natural sciences. Communications in Pure and Applied Mathematics *13*, 1–14.

Williams, R., Anandanadesan, A., Martinez, N. (2010). The probabilistic niche model reveals the niche structure and role of body size in a complex food web. PLOS One *5*, e12092.

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.

Yoshikawa, K., Riedel, S., Asahara, M., Matsumoto, Y. (2009). Jointly identifying temporal relations with markov logic.

Yuan, C., Malone, B. (2013). Learning optimal bayesian networks : A shortest path perspective. Journal of Artificial Intelligence Research *48*, 23–65.

Zadeh, L. (1965). Fuzzy sets. Information and Control *8*, 338–353.

Zahavy, T., Ben-Zrihem, N., Mannor, S. (2016). Graying the black box : Understanding DQNS *arXiv : 1602.02658*.

Zeng, J., Liu, Z.-Q. (2008). Type-2 fuzzy markov random fields and their application to handwritten chinese character recognition. IEEE Transactions on Fuzzy Systems *16*, 747–760.

CHAPITRE 3

ECOLOGICAL INTERACTIONS AND THE NETFLIX PROBLEM

3.1 Description de l'article et contribution

Cet article explore diverses techniques en apprentissage automatique (machine learning) appliquées au problème d'interaction des espèces. Deux approches sont étudiées. Dans la première, nous tentons de prédire une interaction entre un prédateur X et une proie Y en allant chercher les K espèces les plus similaires au prédateur X. La logique est que si les espèces similaires au prédateur X s'intéressent à Y, alors X s'intéressera aussi à Y. C'est la technique qu'utilise Netflix pour prédire quel contenu intéresse un utilisateur. Netflix tente d'aller chercher les utilisateurs qui vous ressemblent le plus et s'inspire du contenu qui les intéresse pour vous en proposer. Nous utilisons aussi une autre technique dite supervisée (supervised learning) où nous entraînons un modèle à partir des traits des prédateurs et des proies pour prédire une interactions.

Nos résultats montrent qu'il est possible de prédire efficacement les interactions à la fois avec la méthode Netflix et avec les traits.

La conception de cet article a été faite avec Dominique Gravel. Les données ont été assemblées par Idaline Laigle, et le manuscrit a été édité avec Idaline Laigle, Timothée Poisot et Dominique Gravel. L'article a été accepté pour publication dans PeerJ en 2017.

3.2 Abstract

Species interactions are a key component of ecosystems but we generally have an incomplete picture of who-eats-whom in a given community. Different techniques have been devised to predict species interactions using theoretical models or abundances. Here, we explore the K nearest neighbour approach, with a special emphasis on recommendation, along with a supervised machine learning technique. Recommenders are algorithms developed for companies like Netflix to predict whether a customer will like a product given the preferences of similar customers. These machine learning techniques are well-suited to study binary ecological interactions since they focus on positive-only data. By removing a prey from a predator, we find that recommenders can guess the missing prey around 50% of the times on the first try, with up to 881 possibilities. Traits do not improve significantly the results for the K nearest neighbour, although a simple test with a supervised learning approach (random forests) show we can predict interactions with high accuracy using only three traits per species. This result shows that binary interactions can be predicted without regard to the ecological community given only three variables : body mass and two variables for the species' phylogeny. These techniques are complementary, as recommenders can predict interactions in the absence of traits, using only information about other species' interactions, while supervised learning algorithms such as random forests base their predictions on traits only but do not exploit other species' interactions. Further work should focus on developing custom similarity measures specialized for ecology to improve the KNN algorithms and using richer data to capture indirect relationships between species.

3.3 Introduction

Species form complex networks of interactions and understanding these interactions is a major goal of ecology (Pimm, 1982). The problem of predicting whether two species will interact has been approached from various perspectives (Bartomeus et al., 2016 ; Morales-Castilla et al., 2015). Williams et Martinez (2000) for instance built a simple theoretical model capable of generating binary food webs sharing important features with real food webs (Gravel et al., 2013), while others have worked to predict interactions from species abundance data (Aderhold et al., 2012 ; Canard et al., 2014) or exploiting food web topology (Cohen, 1978 ; Staniczenko et al., 2010). Being able to predict with high enough accuracy whether two species will interact given simply two sets of attributes, or the preferences of similar species, would be of value to conservation and invasion biology, allowing us to build food webs with partial information about interactions and help us understand cascading effects caused by perturbations. However, the problem is made difficult by the small number of interactions relative to non-interactions and relationships that involve more than two species (Golubski et al., 2016).

In 2006, Netflix offered a prize to anyone who would improve their recommender system by more than 10%. It took three years before a team could claim the prize, and the efforts greatly helped advancing machine learning methods for recommenders (Murphy, 2012). Recommender systems try to predict the rating a user would give to an item, recommending them items they would like based on what similar users like (Aggarwal, 2016). Ecological interactions can also be described this way : we want to know how much a species would "like" a prey. Interactions are treated as binary variables, two species interact or they do not, but the same methods could be applied to interaction matrices with preferences. There are two different ways to see the problem of species interactions. In the positive-only case, a species has a set of preys, and we want to predict what other preys they might be interested in. This approach has the benefit of relying only on our most reliable information : positive (preferably observed) interactions. The other approach is to see binary interactions as a matrix filled with interactions (1s) and non-interactions (0s). Here, we want to predict the value of a specific missing entry (is species x_i consuming species x_j ?). For this chapter, we focus on the positive-only approach, which relies on a simple machine learning approach called the K nearest neighbour.

Statistical machine learning algorithms (Murphy, 2012) have proven to be reliable to build effective predictive models for complex data (the "unreasonable effectiveness of data" (Halevy et al., 2009)). The K nearest neighbour (KNN) algorithm is an effective and simple algorithm for recommendation, in this case finding good preys to a species with positive-only information. The technique is simple : for a given species, we find the K most similar species according to some distance measure, and use these K species to base a prediction. If all the K most similar species prey on species x, there is a good chance that our species has interest in x. In our case, similarity is simply computed using traits and known interactions, but more advanced techniques could be used with a larger set of networks. For example, it is possible to learn similarity measures instead of using a fixed scheme (Bellet et al., 2015). For this study, we use a data-set from Digel et al. (Digel et al., 2014), which contains 909 species, of which 881 are

Table 3.1 – Summary of the two methods used to predict interactions : the *K*NN algorithm and random forests.

Method	Input	Prediction
Recommender (*K*NN)	Set of traits & preys for each species	Recommend new preys
Supervised learning (RF)	Traits (binary and real-valued)	Interaction (1) or non-interaction (0)

The recommender uses the K nearest neighbour algorithm with the Tanimoto distance measure. The Tanimoto *K*NN makes a recommendation, while supervised learning with random forests (RF) predict either an interaction or a non-interaction.

involved in predator-prey relationships and 871 have at least one prey. The data comes from soil food webs and includes invertebrates, plants, bacteria, and fungi. In total, the data-set has 34 193 interactions. The data was complemented with information on 25 binary attributes (traits) for each species, plus their body mass and information on their phylogeny. We also compare our approach to a supervised learning method, random forests, which is used to predict interactions with only the species' traits.

A summary of the two methods used can be found in table 3.1. The approaches are not directly comparable. For example, the positive-only *K*NN recommends preys to a species. If we remove a prey from a species, ask the algorithm to recommend a prey, and check whether the prey will come up as the recommendation, there are up to 881 possibilities. On the other hand, the random forest predicts either an interaction or a non-interaction, a 50% chance of success by random. These approaches have different uses. Positive-only algorithms are interesting because we are rarely certain that two species do not interact. Also, the *K*NN approach uses information on what similar species do, while random forests only rely on traits.

We show the *K*NN is particularly effective at retrieving missing interactions in the positive-only case, succeeding 50% of the times at recommending the right species among 881 possibilities. With few traits, the random forests can achieve high accuracy ($\approx$ 98% for both interactions and non-interactions) without any information about other species in the community. Random forests require only three traits to be effective : body mass and two traits based on the species' phylogeny. Our results show that, with either three traits per species or partial knowledge of the interactions, it is possible to reconstruct a food web accurately. These approaches are complementary, covering both the case where traits are readily available and when only partial knowledge of the food web is known. Both techniques can be used to reconstruct a food web with different types of information.

3.4 Method

3.4.1 Data

The first data-set was obtained from the study of Digel et al. (Digel et al., 2014), who documented the presence and absence of interactions among 881 species from 48 forest soil food webs, details of which are provided in the original publication. 34 193 unique interactions were observed across the 48 food webs, and a total of 215 418 absence of interactions. In order to improve representation of interactions involving low trophic levels species that were not identified at the species level in the first data-set, we compiled a second data-set from a review of the literature. We selected all articles involving interactions of terrestrial invertebrate species for a total of 126 studies, across these, a total of 1 439 interactions were recorded between 648 species. Only 88 absences of interactions were found. We selected traits based on to their potential role in consumption interactions (table 3.2). For each species or taxa, these traits were documented based on a literature review or from visual assessment of pictures. In addition to these traits, we included two proxies for hard-to-measure traits : feeding guild and taxonomy. The traits were chosen for their potential relevance for species interactions and their availability, see (I Laigle et al., 2017) for greater details on the data-set.

3.4.2 *K*-nearest neighbour

Our recommender uses the *K*-nearest neighbour (*K*NN) algorithm (Murphy, 2012). The *K*NN algorithm is an instance-based method, it does not build a general internal model of the data but instead bases predictions on the K nearest (i.e. most similar) entries given some distance metrics. In the case of recommendation, each species is described by a set of traits and a set of preys, and the algorithm will recommend new preys to the species based on the preys of its K nearest neighbours. For example, if $K = 3$, we take the set of preys of the three most similar species to decide which prey to recommend. If species A is found twice and B once in the set of preys of the most similar species, we will recommend A first (assuming, of course, that the species does not already have this prey). See table 3.3 for a complete example of recommendation. In the "Netflix" problem, this is equivalent to recommending new TV series/movies to a user by searching for the users with the most similar taste and using what

Table 3.2 – The 26 binary traits and the three continuous traits used for the supervised learning algorithm.

Features	Abbr.	Description	n
AboveGroud	AG	Whether the species live above the ground.	538
Annelida	An	For species of the annelida phylum.	34
Arthropoda	Ar	For species of the arthropoda phylum.	813
Bacteria	Bc	For species of the bacteria domain.	1
BelowGround	BG	For species living below the ground.	464
Carnivore	Ca	For species eating other animals.	481
Crawls	Cr	Whether the species crawls.	184
Cyanobacteria	Cy	Member of the cyanobacteria phylum.	1
Detritivore	De	For species eating detribus.	355
Detritus	Ds	Whether the species can be classifying as a detritus.	2
Fungivore	Fg	For species eating fungi.	111
Fungi	Fu	Member of the fungi kingdom.	2
HasShell	HS	Whether the species has a shell.	274
Herbivore	He	For species eating plants.	130
Immobile	Im	For immobile species.	85
IsHard	IH	Whether the species has a though exterior (but not a shell).	418
Jumps	Ju	Whether the species can jump.	30
LongLegs	LL	For species with long legs.	59
Mollusca	Mo	Member of the mollusca phylum.	45
Nematoda	Ne	Member of the nematoda phylum.	5
Plantae	Pl	Member of the plant kinggom.	3
Protozoa	Pr	Member of the protozoa kingdom.	3
ShortLegs	SL	For species with short legs.	538
UsePoison	UP	Whether the species uses poison.	177
WebBuilder	WB	Whether the species builds webs.	89
Body mass	M	Natural logarithm of the body mass in grams	881
Ph_0	Ph_0	Coordinate on the first axis of a PCA of phylogenetic distances	881
Ph_1	Ph_1	Coordinate on the second axis of a PCA of phylogenetic distances	881

All traits are binary except for body mass, Ph_0, and Ph_1. We use taxonomy as a proxy of latent traits following (Mouquet et al., 2012). To do so, we used the R package *ape* to obtain taxonomic distances between the species, perform classical multidimensional scaling (or principal coordinates analysis) (Cox et Cox, 2001) on taxonomic distances, and use the scores of each species on the first two axes (Ph_0 and Ph_1) as taxonomy-based traits. These three real-valued variables are scaled to be in the $[0, 1)$ range. For the Tanimoto similarity index, these three continuous variables have to be converted to binary features. For each, we create four binary features of equal size ($n = 881/4$).

they liked as recommendation. It is also possible to tackle the reverse problem : Amazon uses item-based recommendations, in which case we are looking for similar items instead of similar users to base our recommendations (Aggarwal, 2016).

Choosing the right value for K is tricky. Low values give high importance to the most similar entries, while high values provide a larger set of examples. Fortunately, the most computationally intensive task is to compute the distances between all pairs, a step that is independent of K. As a consequence, once the distances are computed, we can quickly run the algorithm with different values of K.

Different distance measures can be used. We will use the Tanimoto coefficient for recommendations. The Tanimoto (or Jaccard) similarity measure is defined as the size of the intersection of two sets divided by their union, or :

$$tanimoto(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}, \tag{3.1}$$

Since it is a similarity measure in the $[0, 1]$ range, we can transform it into a distance function with $1 - tanimoto(\mathbf{x}, \mathbf{y})$. The distance function uses two types of information : the set of traits of the species (see table 3.2) and their set of preys. We define the distance function with traits as :

$$tanimoto_d(\mathbf{x}, \mathbf{y}, w_t) = w_t(1 - tanimoto(\mathbf{x_t}, \mathbf{y_t})) + (1 - w_t)(1 - tanimoto(\mathbf{x_i}, \mathbf{y_i})), \tag{3.2}$$

where w_t is the weight given to traits, $\mathbf{x_t}$ and $\mathbf{y_t}$ are the sets of traits for species x and y, and $\mathbf{x_i}$, $\mathbf{y_i}$ are their sets of preys. Thus, when $w_t = 0$, only interactions are used to compute the distance, and when $w_t = 1$, only traits are used. See table 3.3 for an example.

The data is the set of preys and binary traits for each species (Table 3.2). To test the approach, we randomly remove an interaction for each species and ask the algorithm to recommend up to 10 preys for the species with the missing interaction. Interactions are removed one-at-a-time and similarity is computed before the interaction is removed. The code for computing

Table 3.3 – Fictional example to illustrate recommendations with K nearest neighbour using the Tanimoto distance measure modified to include species traits.

Species ID	Traits	Preys	Most similar	Recommendations
0	$\{Ar, Ca\}$	$\{6, 42, 47\}$	$\{6, 28, 70\}$	$[812, 70, 72]$
6	$\{Ar, Ca\}$	$\{42, 47, 70, 72\}$		
28	$\{Ar, Ca\}$	$\{42, 47, 70, 812\}$		
70	$\{Ca\}$	$\{42, 47, 812\}$		
...	...	...		

We are trying to recommend a prey to species 0 given that the three most similar species are species 6, 28, and 70. For example, the distance from species 0 to species 70 would be $w_t 0.5 + (1 - w_t)2/4$. To find recommendations, the set of preys found in the $K = 3$ most similar entries is computed, in this case $\{812 = 2, 70 = 2, 72 = 1\}$, leading to the list of recommendations $[812, 70, 72]$. Because they are found most often in the K most similar species, candidates 812 and 70 will be suggested before 72. To test this approach, we remove a prey from a species and check whether the algorithm recommend the missing prey. Especially with low K, it's possible that no recommendations can be found, for example if the most similar species has the exact same preys.

similarities *after* the interaction is removed is available in the code repository, but it has little effect on the results while making the program much slower to run since the similarity matrix must be computed for each trial. We count how many recommendations are required to retrieve the missing interactions and compute the top1, top5, and top10 success rates, which are defined as the probabilities to retrieve the missing interaction with 1, 5, or 10 recommendations. We repeat this process 10 times for each species with at least 2 preys, totally 7200 attempts. We test all odd values of K from 1 to 19, and $w_t = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. We also divided species in groups according to the number of preys they have to see if it is easier to find the missing interaction for species with fewer preys.

3.4.3 Supervised learning

We also do a simple test with random forests to see if it is possible to predict interactions in this data-set using only the traits (Breiman, 2001). In this case, the random forests perform supervised learning : we are trying to predict y (interaction) from the vector of traits $\mathbf{x}$ by first learning a model on the training set, and testing the learned model on a testing set. We keep 5% of the data for testing. We perform grid search to find the optimal parameters for the random forests.

For our predictions, we count the number of true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). The score for predicting interactions ($Score_y$), non-interactions ($Score_{\neg y}$) and the accuracy are defined as

$$Score_y = \frac{tp}{tp + fp}, \tag{3.3}$$

$$Score_{\neg y} = \frac{tn}{tn + fn}, \tag{3.4}$$

$$Accuracy = \frac{Score_y 34193 + Score_{\neg y} 741968}{881^2}, \tag{3.5}$$

with 34193 and 741968 being the number of observed interactions and non-interactions in the 881 by 881 matrix. We then use the True Skill Statistics (TSS) to measure how accurate the random forest is, defined a

$$TSS = \frac{(tp \times tn) - (fp \times fn)}{(tp + fn)(fp + tn)}. \tag{3.6}$$

The TSS ranges from -1 to 1.

3.4.4 Code and Data

Since several machine learning algorithms depends on computing distances (or similarities) for all pairs, many data structures have been designed to compute them efficiently from kd-trees discovered more than thirty years ago (Friedman et al., 1977) to ball trees, metric skip lists, navigating nets (Izbicki et Shelton, 2015), and cover trees (Beygelzimer et al., 2006 ; Izbicki et Shelton, 2015). We use an exact but naive approach that works well with small data-sets. Since $distance(x, y) = 0$ if $x = y$ and $distance(x, y) = distance(y, x)$, our C++ implementation stores the distances in a lower triangular matrix without the diagonal, yielding $n(n-1)/2$ distances to compute. A linear scan is then used to find the most similar species. Computing the distance matrix and testing the predictions 7000 times for a set of parameters takes less than a second. We used Scikit for random forests (Pedregosa et al., 2011). The C++11 code for the KNN algorithm, Python scripts for random forests, and all data-sets used are available at https://github.com/PhDP/EcoInter (also stored on zenodo with a DOI : (Desjardins-Proulx, 2016)).

3.5 Results

3.5.1 Recommendation

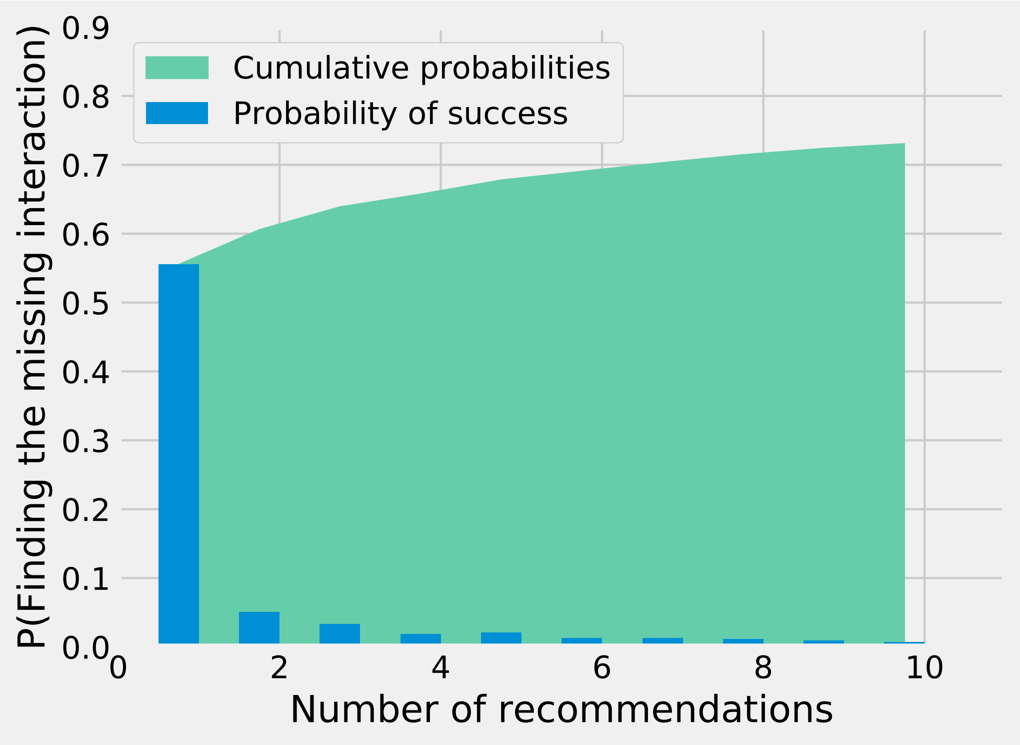
While matrix imputation has a 50% change of success by random, the Tanimoto KNN needs to pick the right prey among up to 881 possibilities. Yet, it succeeds on its first recommendation around 50% of the times. When the first recommendation fails, the next 9 recommendations only retrieve the right species around 15% of the times so the top5 and top10 success rates are fairly close to the top1 success rate (see figure 3.1). The Tanimoto measure is particularly effective for species with fewer preys, achieving more than 80% success rate for species with 10 or fewer preys (Figure 3.2).

The highest first-try success rates (the probability to pick the missing interaction on the first recommendation) are found with $K = 7$ and no weights to traits, and with $K = 17$ and a small weight of 0.2 to traits (Table 3.4). Overall, the value of K had little effect on predictive ability.

3.5.2 Supervised learning

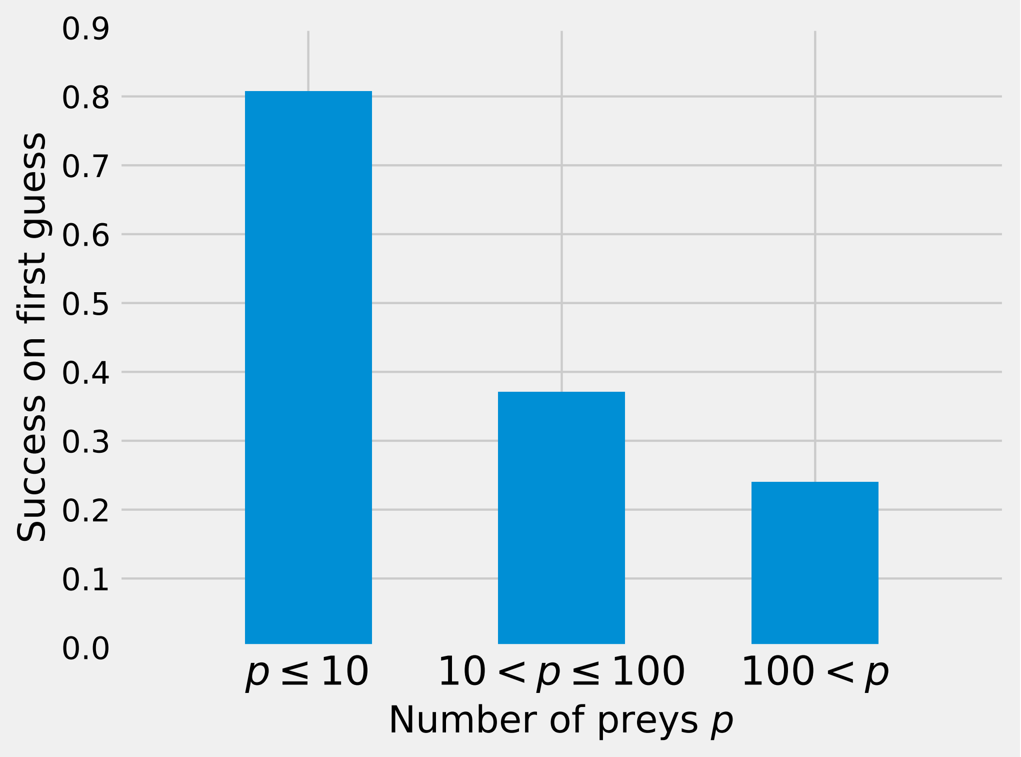
Random forests predict correctly 99.55% of the non-interactions and 96.81% of the interactions, for a TSS of 0.96. Much of this accuracy is due to the three real-valued traits (body mass, Ph_0, Ph_1). Without them, too many entries have the same feature vector $\mathbf{x}$, making it impossible for the algorithm to classify them correctly. Removing the binary traits has little effect on the model. With only body mass, Ph_0, Ph_1, the TSS of the random forests is 0.94.

Figure 3.1 – Finding the missing interaction with KNN/Tanimoto approach.



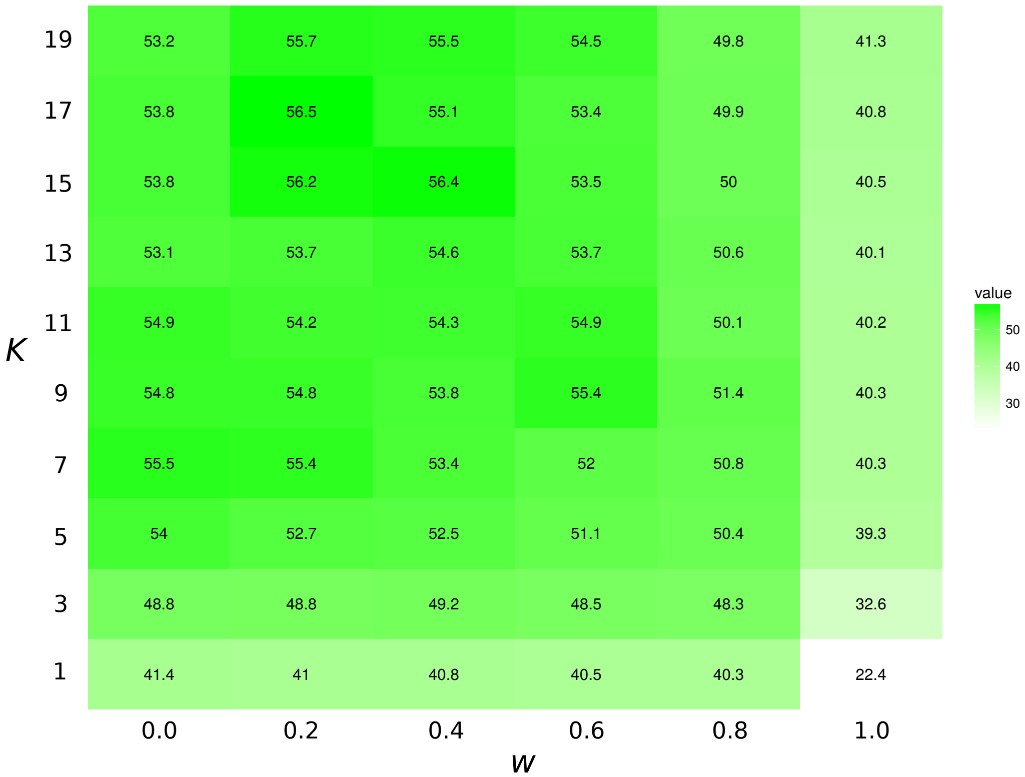
After removing a prey from a predator, we ask the KNN algorithms with Tanimoto measure to make 10 recommendations (from best to worst). The figure shows how many recommendations are required to retrieve the missing interaction. Most retrieved interactions are found with the first attempt. This data was generated with $K = 7$ and $w_t = 0$.

Figure 3.2 – Success on first guess with Tanimoto similarity as a function of the number of prey.



The KNN algorithm with Tanimoto similarity is more effective at predicting missing preys when the number of preys is small. This is probably in good part because there are more information available to the algorithm, since 473 species have 10 or fewer preys, 295 have between 10 and 100, 103 species have more than 100 preys.

Table 3.4 – Top1 success rates for the KNN/Tanimoto algorithm with various K and weights to traits.



K \ w	0.0	0.2	0.4	0.6	0.8	1.0
19	53.2	55.7	55.5	54.5	49.8	41.3
17	53.8	56.5	55.1	53.4	49.9	40.8
15	53.8	56.2	56.4	53.5	50	40.5
13	53.1	53.7	54.6	53.7	50.6	40.1
11	54.9	54.2	54.3	54.9	50.1	40.2
9	54.8	54.8	53.8	55.4	51.4	40.3
7	55.5	55.4	53.4	52	50.8	40.3
5	54	52.7	52.5	51.1	50.4	39.3
3	48.8	48.8	49.2	48.5	48.3	32.6
1	41.4	41	40.8	40.5	40.3	22.4

When $w_t = 0.0$, the algorithm will only use interactions to compute similarity between species. When $w_t = 1$, the algorithm will only consider the species' traits (see table 3.2). The value is the probability to retrieve the correct missing interaction with the first recommendation. For each entry, $n = 871$ (the number of species minus 10, the number of species with no preys). The best result is achieved with $K = 17$ and $w = 0.2$, although the results for most values of K and $w = [0.0, 0.2]$ are all fairly close. The success rate increases with K when only traits are considered ($w = 1$).

3.6 Discussion

We applied different machine learning techniques to the problem of predicting binary species interactions. Recommendation is arguably a better fit for binary species interactions, since it is essentially the same problem commercial recommenders such as Netflix face : given that a user like item i, what is the best way to select other items the user would like ? In this case, users are species, and the items are their preys, but the problem is the same. In both cases, we can have solid positive evidence (observed or implied interactions), but rarely have proofs of non-interactions. The approach yields strong results, with a top1 success rate above 50% in a food web with up to 881 possibilities. The approach could be used, for example, to reconstruct entire food webs using global database of interactions (Poelen et al., 2014). The method's effectiveness rely on nestedness : how much species cluster around the same set of preys in a food web (Guimaraes et Guimaraes, 2006). Thus, it should be less effective in food webs with more unique predators.

The KNN algorithm falls into the realm of unsupervised learning, where the goal is to find patterns in data (Murphy, 2012). The other class of machine learning algorithms, supervised learning, have the clearer goal of predicting a value y from a vector of features $\mathbf{x}$. For example, in supervised learning, we would try to predict an interaction y from the vector of traits $\mathbf{x}$, while a unsupervised approach can fill entries from an incomplete matrix regardless of what the entry is (interaction or trait). With a larger set of food webs, it may be possible to use an unsupervised algorithm, for example a neural network, to train a model for matrix imputation. Instead of recommending new preys, imputation would simply fill missing entries from a matrix (interaction or non-interactions).

Our random forests achieve a TSS of 0.96 using the binary traits, body mass, and the coordinates of the multidimensional scaling. This is consistent with previous research that has shown that ecological networks have relatively few dimensions (Eklof et al., 2013). A random forest can build effective predictive models by creating complex rules based on the traits, while the KNN algorithm relies on a simplistic distance metrics. However, the KNN approach has some advantages over supervised learning, namely the capacity to recommend preys using only the information from the other species' interactions. The solution to improve the KNN approach in ecology is likely to *learn* distance metrics (Bellet et al., 2015) instead of using a fixed formula. This would allow complex rules while maintaining the KNN's ability to exploit partial food web structures. Learning distance metrics is a promising avenue to improve our results. Much efforts on the Netflix prize focused on improving similarity measures (Toscher et Jahrer, 2008 ; Hong et Tsamis, 2006), and custom similarity metrics can be used to improve unsupervised classification algorithms (Bellet et al., 2015) by exploiting complex domain-specific rules. Maybe species with many preys, apex predators, or specialists behave differently enough to need different similarity measures. Learning distance metrics from data is a common way to improve methods based on a nearest neighbour search (Xing et al., 2003 ; Bellet et al., 2015), allowing the measure itself to be optimized. We only used the K nearest neighbour algorithm for unsupervised learning, but several other algorithms can be used to solve the "Netflix problem". For example : techniques based on linear programming, such as recent exact methods for matrix completion based on convex optimization (Candès et Recht, 2009) or low-rank matrix factorization. The latter method reduces a matrix to a multiplication between two smaller matrices, which can be used both to predict missing entries and to compress large matrices into small, more manageable matrices (Vanderbei, 2013). Given enough data, deep learning methods such as deep Boltzmann machines could also be used (Zhang, 2011). Deep learning revolutionized machine learning with neural networks made of layers capable of learning increasingly detailed representations of complex data (Hinton et al., 2006). Many of the most spectacular successes of machine learning use deep learning (Mnih et al., 2013). However, learning several neural layers to form a deep networks would require larger data-sets.

The low sensitivity to K in recommendations is interesting and makes the approach easier to use. This is caused by the fact that, as K grows, the set of species includes more and more unrelated species with widely different set of preys. If we increase K from k to $k + \delta$ for a recommendation, the species in δ range are not only less similar, but they are less likely to share preys among themselves. Since recommendations are based on how many times a prey is found in the K nearest species, the species in the δ range are unlikely to have as much weight as the first k species. Our KNN recommender is particularly easy to parametrize since it is neither sensible to the weight given to traits nor to the choice of K.

Our results have two limitations. It is possible that our food web was exceptionally simple, and that a food web with distinct structural properties would behave differently, especially if it has lower nestedness. The success of the KNN algorithms depends on local structure : how much can we learn from similar species. If each species has a unique set of preys, the KNN will struggle more. Also, a deeper issue is that real food webs are not binary structures. Species, populations, and individuals have different densities, prey more strongly on some resources than others, and have preferences. In a binary matrix, we can predict if two species will interact while completely ignoring the rest of the network, but real food webs involve complex indirect relationships (Wootton, 1994). It is unclear how much we can learn about ecosystems and species interactions from binary matrices, and our results show that binary interactions can be predicted without direct knowledge of the community, since we are able to effectively predict if two species interact given only three traits. Species interactions are better represented with a weighted hypergraph (Gao et al., 2012), which is well-suited to model relations with an arbitrary number of participants. The hyperedge would allow for complex indirect relationships to be included. Understanding these hypergraphs is outside the scope of the *K*NN algorithm but could be understood with modern techniques such as Markov logic (Richardson et Domingos, 2006).

Recommendation (*K*NN algorithm with Tanimoto distance) and supervised learning (random forests) are complementary techniques. Supervised learning is more useful when we have traits and no information about interactions, but it is useless without the traits. On the other hand, the recommender performs well without traits but requires at least partial information about interactions, although it might be possible to use the interactions from different food webs. We suggest more research could be done on developing better distance metrics for ecological interactions or learning these metrics from data.

3.7 References

Aderhold, A., Husmeier, D., Lennon, J., Beale, C., Smith, V. (2012). Hierarchical bayesian models in ecology : Reconstructing species interaction networks from non-homogeneous species abundance data. Ecological Informatics *11*, 55–64.

Aggarwal, C. (2016). Recommender Systems. Springer.

Bartomeus, I., Gravel, D., Tylianakis, J., Aizen, M., Dickie, I., Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. Functional Ecology *30*, 1894–1903.

Bellet, A., Habrard, A., Sebban, M. (2015). Metric Learning. Morgan & Claypool.

Beygelzimer, A., Kakade, S., Langford, J. (2006). Cover trees for nearest neighbor. In : Proceedings of the 23nd International Conference on Machine Learning.

Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.

Canard, E., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D., Gravel, D. (2014). Empirical evaluation of neutral interactions in host-parasite networks. American Naturalist9 *183*, 468–479.

Candès, E., Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics *9*, 717–772.

Cohen, J. (1978). Food webs and niche space. Princeton University Press.

Cox, T., Cox, M. (2001). Multidimensional Scaling. Chapman and Hall.

Desjardins-Proulx, P. (2016). https ://github.com/phdp/ecointer. http://doi.org/10.5281/zenodo.161602.

Digel, C., Curtsdotter, A., Riede, J., Klarner, B., Brose, U. (2014). Unravelling the complex structure of forest soil food webs : higher omnivory and more trophic levels. In : Oikos. Vol. 123. pp. 1157–1172.

Eklof, A., Jacob, U., Kopp, J., Bosch, J., Castro-Urgal, R., Chacoff, N., Dalsgaard, B., de Sassi, C., Galetti, M., Guimarães, P., Beatriz Lomáscolo, S., Martín González, A., Aurelio Pizo, M., Rader, R., Rodrigo, A., Tylianakis, J., Vázquez, D., Allesina, S. (2013). The dimensionality of ecological networks. Ecology Letters *16*, 577–583.

Friedman, J., Bentley, J., Finkel, R. (1977). An algorithm for finding best matches in logarithmic expected time. Transactions on Mathematical Software *3*, 209–226.

Gao, J., Zhao, Q., Ren, W., Swami, A., Ramanathan, R., Bar-Noy, A. (2012). Dynamic shortest path algorithms for hypergraphs. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks , 238–245.

Golubski, A., Westlund, E., Vandermeer, J., Pascual, M. (2016). Ecological networks over the edge : Hypergraph trait-mediated indirect interaction (tmii) structure. Trends in Ecology and Evolution *31*, 1083–1090.

Gravel, D., Poisot, T., Albouy, C., Velez, L., Mouillot, D. (2013). Inferring food web structure from predator–prey body size relationships. Methods in Ecology and Evolution *4*, 1083–1090.

Guimaraes, P., Guimaraes, P. (2006). Improving the analyses of nestedness for large sets of matrices. Environmental Modelling and Software *21*, 1512–1513.

Halevy, A., Norvig, P., Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems *24*, 8–12.

Hinton, G., Osindero, S., Teh, Y. (2006). A fast learning algorithm for deep belief nets. Neural computation *18*, 1527–1554.

Hong, T., Tsamis, D. (2006). Use of KNN for the Netflix Prize.

I Laigle, I., Aubin, I., Digel, C., Brose, U., Boulangeat, I., Gravel, D. (2017). Species traits as drivers of food web structure. In preparation .

Izbicki, M., Shelton, C. (2015). Faster cover trees. In : Proceedings of the 32nd International Conference on Machine Learning.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv .

Morales-Castilla, I., Matias, M., Gravel, D., Araújoemail, M. (2015). Inferring biotic interactions from proxies. Ecological Informatics *30*, 347–356.

Mouquet, N., Devictor, V., Meynard, C., Munoz, F., Bersier, L., Chave, J., Couteron, P., Dalecky, A., Fontaine, C., Gravel, D. (2012). Ecophylogenetics : advances and perspectives. Biological reviews *87*, 769–785.

Murphy, K. (2012). Machine Learning : A Probabilistic Perspective. The MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Pimm, S. (1982). Food Webs. Springer.

Poelen, J., Simons, J., Mungall, C. (2014). Global biotic interactions : An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics *24*, 148–159.

Richardson, M., Domingos, P. (2006). Markov logic networks. Machine Learning *62*, 107–136.

Staniczenko, P., Lewis, O., Jones, N., Reed-Tsochas, F. (2010). Structural dynamics and robustness of food webs. Ecology Letters *13*, 891–899.

Toscher, A., Jahrer, M. (2008). The BigChaos solution to the Netflix prize.

Vanderbei, R. (2013). Linear programming : Foundations and extensions .

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.

Wootton, J. (1994). The nature and consequences of indirect effects in ecological communities. Annual Review of Ecology and Systematics , 443–466.

Xing, E., Ng, A., Jordan, M., Russell, S. (2003). Distance metric learning with application to clustering with side-information. Advances in neural information processing systems *15*, 505–512.

Zhang, J. (2011). Deep transfer learning via restricted boltzmann machine for document classification. In : ICMLA : Machine Learning and Applications. Vol. 1. pp. 323–326.

CHAPITRE 4

THE K NEAREST NEIGHBOUR ALGORITHM WITH THRESHOLD TO PREDICT ECOLOGICAL INTERACTIONS

4.1 Description de l'article et contribution

Le KNN est une approche simple et efficace pour prédire les interactions sans traits. Cependant, cet algorithme tend à sous-prédire des interactions. Pour cette contribution, nous étudions une variante du KNN qui permet d'améliorer des prédictions sur les interactions. Nous testons notre approche sur des données d'interactions entre pollinisateurs et plantes. Notre approche réussit à bien prédire à la fois les interactions et les non-interactions pour ce jeu de données.

J'ai conçu cette recherche. Les données ont été assemblées par Nacho. Le manuscrit a été révisé par Nacho, Timothée Poisot et Dominique Gravel. L'article sera soumis sous sa forme actuelle à un journal généraliste.

Abstract

We present a variant to a popular machine learning method, the K nearest neighbour (KNN) algorithm, and use it to predict whether two species interact. The algorithm has several advantages over other methods : it requires no species traits and is simple to understand, implement, and interpret. Because ecological interactions tend to be rare relative to non-interactions, it is common to use measures of success like the true skill statistic (TSS) that will give equal weights to the predictions of interactions and non-interactions. These measure penalize more false negatives than false positives, which makes sense given the ecological significance of interactions. In this context, we add a threshold to the KNN algorithm to improve the prediction of interactions. We test our algorithm on a data-set of plant-pollinator interactions and found the method to achieve better TSS with low thresholds.

4.2 Introduction

A major goal of ecology is to understand how species interact (Pimm, 1982; Poisot et al., 2016b). Strong predictive models could help us understand how interaction networks evolve and transform with biotic and abiotic changes. Different methods have been devised to predict these interactions, from simple theoretical models (Williams et Martinez, 2000) to inference methods based on abundance data (Aderhold et al., 2012; Canard et al., 2014) and machine learning techniques (Desjardins-Proulx et al., 2017).

Here, we study a variant of a well-established method, the KNN algorithm (Cover et Hart, 1967). In a nutshell, the KNN algorithm works by finding the K most similar elements in a set and using them to generate a prediction, with K being an odd integer. If we want to know whether species x interacts with a, we must first find the K species most similar to x (in terms of number of interactions shared) and generate a prediction based on how many of these species interact with K. The standard method works by majority vote, so the prediction for a will be the most common value among the K nearest. For example, if three out of the five species most similar to x have an interaction with a, we will predict an interaction with a. The rationale of the methods is rooted in simple ecological theory, because plant-pollinator interaction networks are typically nested, a measure of how much species form clusters with similar interactions (Guimaraes et Guimaraes, 2006). The effectiveness of the KNN may also stands from the hypothesis that traits are ultimately constraining ecological interactions (Bartomeus et al., 2016), but that often they can't be measured. Because many traits involved in ecological interactions take times to evolve, it happens that often species sharing a given set of interactions will also be likely to share another set.

For this study, we test a variant of the KNN with an arbitrary threshold, allowing for the algorithm to predict an interaction even if fewer than half of the K nearest neighbours have the interaction. The idea of imposing a threshold for KNN can be counter-intuitive. For example, if only one of the three closest species have an interaction, we will predict an interaction if our threshold is of 1/3 or less. However, it means an interaction would also be erroneously predicted for those two species without an interaction. The logic behind the threshold is that we care more about predicting interactions than their relative occurrences. In nature, most pairs of species do not interact (i.e. mean connectance values for plant-pollinator networks is low (Vázquez et al., 2009)). Thus, a predictive model that would always predict non-interactions would have high accuracy. On the other hand, the standard statistics to compare presence/absence of interactions, like the True Skill Statistic, tend to give equal weights to predictions of presence and absence, despite the latter being a lot more common (Allouche et al., 2006). With these measures, it makes sense to risk more false positives to correctly capture more interactions.

We test our variant using detailed data from plant-pollinator interactions collected in a Mediterranean shrubland. We show that the KNN provides strong fit, with high accuracy for both predicting interactions and non-interactions. A lower threshold, where interactions are predicted when only one of the K nearest has the interaction, is more effective than the standard majority rule approach using the TSS as measure of performance. Our results show how a small twist on a well-known algorithm can help it perform better at predicting rare events.

4.3 Methods

4.3.1 Data

Plant-pollinator interactions were collected across 16 sites in the southwest of the Iberian peninsula (Huelva and Seville) from March to June 2015. All sites are dominated by Mediterranean open forests with a flowery understory including *Cistus sp.*, *Lavandula sp.*, *Rosmarinus officinalis*. Sampling was conducted on seven rounds per site along the season and each round consisted in three types of censuses. (i) A linear transect, in which all seen interactions were recorded along a 100 m transect during 30 minutes, (ii) focal 5 minutes census on the dominant plants were all visitors were recorded and (iii) opportunistic sampling of non previously recorded interactions. Those species that were unable to be identified in the field were collected and identified by professional taxonomists in the lab. There are 270 species of pollinators and 65 species of plants for a total of 14 580 possible interactions. Interaction between 739 pollinator-plant pairs have been observed (4.2% of all possible combinations).

4.3.2 Imputation with the K nearest neighbour algorithm

The KNN algorithm has two steps. In the first step, for all pairs of species, we compute a similarity index $\in [0,1]$. We use the Jaccard similarity coefficient, which measures similarity as the size of intersection divided by the size of the union of two sets :

$$jaccard(\mathbf{x},\mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}. \tag{4.1}$$

Here, we use the set of plants for each pollinator to compute their similarity. For example, if a species x pollinates plants $\{A,B,C\}$ and species y pollinates plants $\{A,C,D,E,F\}$, then :

$$\begin{aligned} jaccard(\mathbf{x},\mathbf{y}) &= \frac{|\{A,B,C\} \cap \{A,C,D,E,F\}|}{|\{A,B,C\} \cup \{A,C,D,E,F\}|} \qquad (4.2) \\ &= \frac{|\{A,C\}|}{|\{A,B,C,D,E,F\}|} = \frac{2}{6}. \qquad (4.3) \end{aligned}$$

This or other similarity indexes can be calculated over other values, for example on species traits (Desjardins-Proulx et al., 2017). Alternatively, measures learned from data can be also used (Bellet et al., 2015).

The resulting similarity matrix is used to find the K most similar species, with K normally being an odd integer. More advanced techniques such as cover trees can be used to find the K nearest species faster than a matrix (Beygelzimer et al., 2006; Izbicki et Shelton, 2015) but the small size of our data-set allowed us to simply scan all species to find all the K nearest neighbours. Since there are many species with the same similarity we sometime have to randomly pick in a set of species with the same similarity to get exactly K neighbours. This random aspect create small variations in the results. We tested all odd values of K in the $[1,11]$ range.

To predict whether pollinator x interacts with plant a, we count how many of the K species most similar to x have an interaction with a, divide that number by K to normalize, and predict an interaction if that number is higher than a threshold t. If the ratio is lower or equal to the threshold we predict a non-interaction. This is a deviation from the standard KNN algorithm, which relies on a majority vote approach : the prediction is based on the most common value found in the K nearest elements, which in this case would be equivalent to a threshold value of 0.5 (i.e. we predict an interaction only if more than 50% of the K closest species also have the interaction). Allowing the threshold to take other values than 0.5 means we can, for example, predict an interaction if only one of the five nearest species have the interaction. We tested all threshold in the $[0.0, 0.9]$ range (with 0.1 increments). Since we require the KNN ratio to be higher than the threshold to predict an interaction, a threshold of 0.0 means we require only one interaction in the K nearest to predict one. We do not test a threshold of 1 since it would require *more* than 100% of the K nearest to have an interaction.

4.3.3 Scoring

We use the true skill score (TSS) to compare results. It is defined as

$$TSS = \frac{(tp \times tn) - (fp \times fn)}{(tp + fn)(fp + tn)}, \tag{4.4}$$

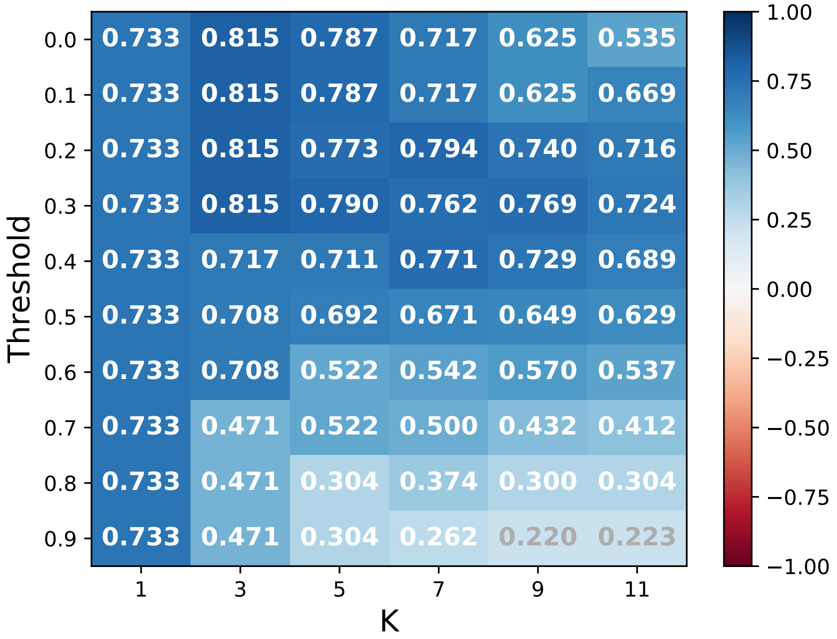
where tp, tn, fp, fn are the true positives, true negatives, false positives, and false negaties, respectively. The TSS ranges from -1 to 1.

4.4 Results

4.4.1 KNN

The KNN algorithm is capable of predicting most interactions and non-interactions, with a peak TSS just above 0.80 for $K = 3$ and thresholds in the $[0.0, 0.3]$ range. The standard threshold of 0.5 yields good but significantly worse results than lower thresholds, with TSS generally below 0.70 except for $K = 1$ (since there is only one species in the set of nearest neighbours, the threshold has no effect) and $K = 3$ (Figure 4.1).

Figure 4.1 – TSS scores for the KNN algorithm with K ranging from from 1 to 11 (odd integers only) and thresholds from 0 to 0.9.



Results with a threshold of 0.5 represents the standard KNN majority vote approach, where the prediction is based on the most common values found in the K nearest elements. The usual majority vote (threshold of 0.5) yields good results, but lower thresholds always give better predictions.

The ability to correctly predict an interaction is one of the trickiest problem given most species do not interact, making it difficult for algorithms to handle positive cases (Figure 4.3). However, the threshold does predictably worse on predicting non-interactions (Figure 4.4) and, since 95.6% of pollinator-pairs do not interact, the (Figure 4.2). Table 4.1 shows the full confusion matrices for $K = 3$ and thresholds of 0.0 and 0.5.

Figure 4.2 – Accuracy for the KNN algorithm with various values K and threshold, defined as the number of correct predictions divided by the total number of predictions.

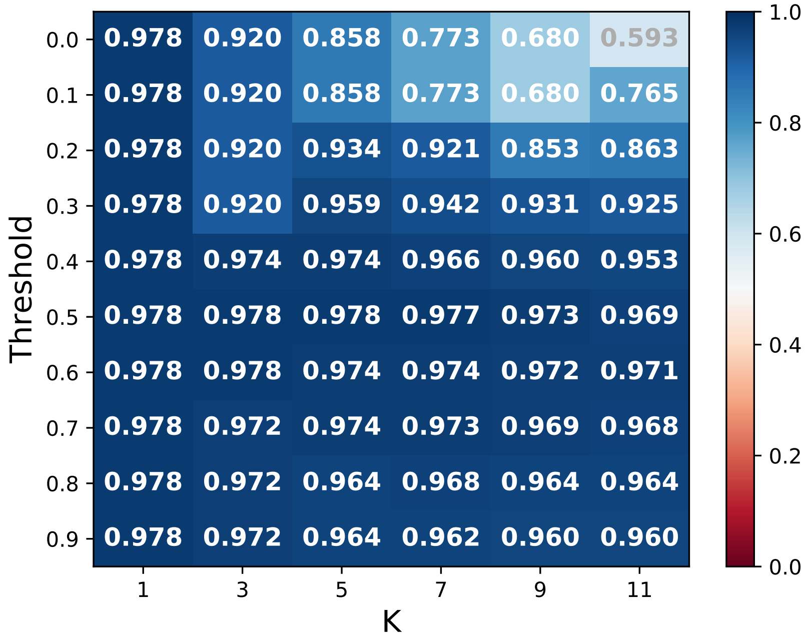
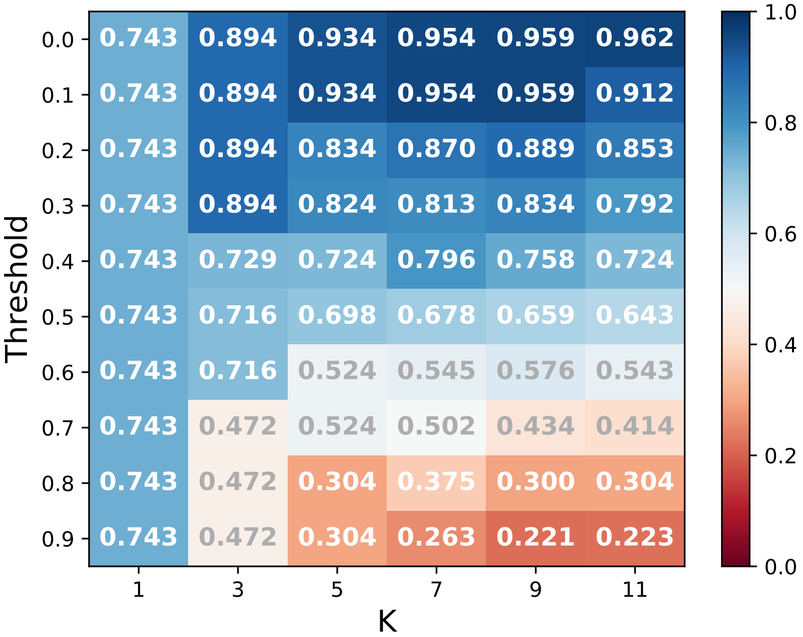


Table 4.1 – Full confusion matrix for KNN algorithm with $K = 3$ and threshold of 0.0 and 0.5 (majority vote).

Threshold = 0		*Threshold* = 0.5	
True pos. 662	*False pos.* 952	*True pos.* 533	*False pos.* 144
False neg. 77	*True neg.* 12889	*False neg.* 206	*True neg.* 13697

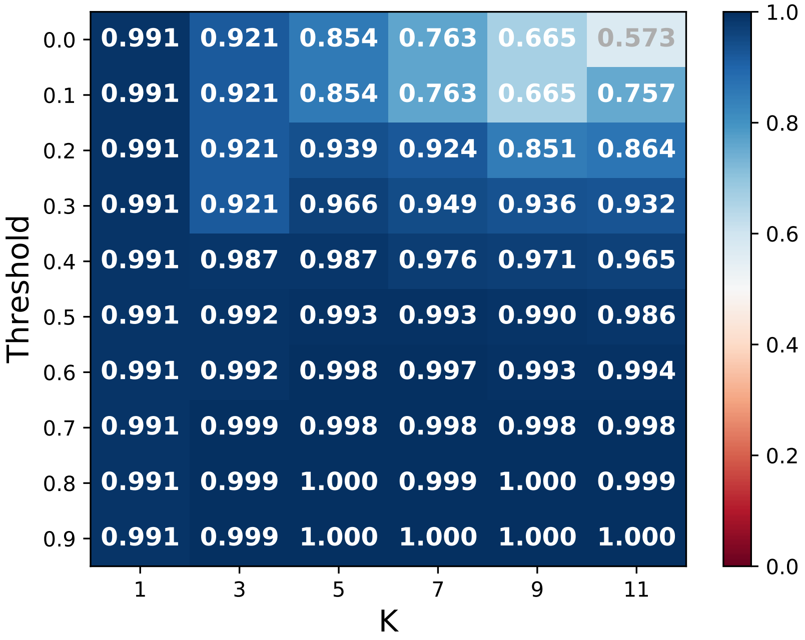
Full confusion matrix for $K = 3$ and threshold of 0.0 and 0.5 (majority vote). With the threshold of 0, witnessing a single interaction among the 3 nearest pollinators is enough to predict an interaction, while 2/3 interactions must be present for the majority vote approach to predict an interaction. We see how the lower threshold makes fewer good predictions (13 551 vs 14 230) but they are better distributed between true positives and true negatives. The confusion matrix on the left has a lower accuracy but a better TSS score than the majority vote on the right (0.815 vs 0.708). There are 270 species of pollinators and 65 species of plants for a total of 14 580 predictions.

Figure 4.3 – Heatmap for the true positive rate (also called recall or sensitivity) of the KNN with threshold.



True positive rate (also called recall or sensitivity) defined as $true\ positives/(true\ positives + false\ negatives)$. We can see the deleterious effect of higher thresholds on the ability to predict interactions.

Figure 4.4 – Heatmap for the true negative rate (also called recall or sensitivity) of the KNN with threshold.



Threshold \ K	1	3	5	7	9	11
0.0	0.991	0.921	0.854	0.763	0.665	0.573
0.1	0.991	0.921	0.854	0.763	0.665	0.757
0.2	0.991	0.921	0.939	0.924	0.851	0.864
0.3	0.991	0.921	0.966	0.949	0.936	0.932
0.4	0.991	0.987	0.987	0.976	0.971	0.965
0.5	0.991	0.992	0.993	0.993	0.990	0.986
0.6	0.991	0.992	0.998	0.997	0.993	0.994
0.7	0.991	0.999	0.998	0.998	0.998	0.998
0.8	0.991	0.999	1.000	0.999	1.000	0.999
0.9	0.991	0.999	1.000	1.000	1.000	1.000

True negative rate (or specificity) defined as $true\ negatives/(true\ negatives + false\ positives)$.

4.5 Discussion

We showed how a simple algorithm can, without using species traits, predict species interactions effectively. Eschewing the standard majority rule for KNN was beneficial, as the best results were found with $K = 3$ and threshold values in the $[0, 0.3]$ range. Interaction similarity among species may encapsulate latent information difficult to measure from traits and phylogenetic constraints. An alternative to the KNN would be to use supervised learning algorithms, such as random forests, to predict interactions from traits. Our results show clearly how the KNN approach can offer a solid alternative to supervised learning when traits are either unavailable or not informative enough (Desjardins-Proulx et al., 2017). The KNN's effectiveness depends on nestedness : how much species cluster around the same set of preys in a food web (Guimaraes et Guimaraes, 2006). It may also relate to ecosystem stability. Recent research suggests that clustering of species helps stabilize food webs (Johnson et al., 2016).

The KNN approach combined with large assemblage of species interactions extracted from mangal.io (Poisot et al., 2016a) or globi (Poelen et al., 2014) can be used to predict local realizations of interactions. This approach has been used to reconstruct food webs in a data-poor environment (Beauchesne et al., 2017). Further research on the use of KNN for predicting interactions should (i) study different measures of similarity, and (ii) look at probabilistic variants of the KNN. For this study we focused on the Jaccard similarity coefficient, but it is possible to use other similarity indexes or even to use to learn distance metrics. Metric learning is a growing branch of machine learning interested in how good similarity/distance measures can be learned from data (Bellet et al., 2015). For this chapter we used a measure based on interactions but we could imagine how more sophisticated measures could integrate traits, phylogeny, and interactions to estimate similarity. Such measure can be learned from data instead of being hand-crafted. As for probabilistic variants, Holmes and Adams (Holmes et Adams, 2002) proposed a probabilistic nearest neighbour algorithm (PNN). The PNN is both more flexible thanks to predictions being proper probabilities, but has also no parameters, with K being replaced by a smooth decrease of influence as similarity decreases.

4.6 References

Aderhold, A., Husmeier, D., Lennon, J., Beale, C., Smith, V. (2012). Hierarchical bayesian models in ecology : Reconstructing species interaction networks from non-homogeneous species abundance data. Ecological Informatics *11*, 55–64.

Allouche, O., Tsoar, A., Kadmon, R. (2006). Assessing the accuracy of species distribution models : prevalence, kappa and the true skill statistic (tss). Journal of Applied Ecology *43*, 1223–1232.

Bartomeus, I., Gravel, D., Tylianakis, J., Aizen, M., Dickie, I., Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. Functional Ecology *30*, 1894–1903.

Beauchesne, D., Desjardins-Proulx, P., Archambault, P., Gravel, D. (2017). Thinking outside the box-predicting biotic interactions in data-poor environments. Vie et Milieu *66*, 333–342.

Bellet, A., Habrard, A., Sebban, M. (2015). Metric Learning. Morgan & Claypool.

Beygelzimer, A., Kakade, S., Langford, J. (2006). Cover trees for nearest neighbor. In : Proceedings of the 23nd International Conference on Machine Learning.

Canard, E., Mouquet, N., Mouillot, D., Stanko, M., Miklisova, D., Gravel, D. (2014). Empirical evaluation of neutral interactions in host-parasite networks. American Naturalist9 *183*, 468–479.

Cover, T., Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory *13*, 21–27.

Desjardins-Proulx, P., Laigle, I., Poisot, T., Gravel, D. (2017). Ecological interactions and the netflix problem. PeerJ *5*, e3644.

Guimaraes, P., Guimaraes, P. (2006). Improving the analyses of nestedness for large sets of matrices. Environmental Modelling and Software *21*, 1512–1513.

Holmes, C., Adams, N. (2002). A probabilistic nearest neighbour method for statistical pattern recognition. R. Statist. Soc. B *64*, 295–306.

Izbicki, M., Shelton, C. (2015). Faster cover trees. In : Proceedings of the 32nd International Conference on Machine Learning.

Johnson, S., Domínguez-García, V., Donetti, L., Muñozb, M. (2016). Trophic coherence determines food-web stability. PNAS *111*, 17923–17928.

Pimm, S. (1982). Food Webs. Springer.

Poelen, J., Simons, J., Mungall, C. (2014). Global biotic interactions : An open infrastructure to share and analyze species-interaction datasets. Ecological Informatics *24*, 148–159.

Poisot, T., Baiser, B., Dunne, J., Kéfi, S., Massol, F., Mouquet, N., Romanuk, T., Stouffer, D., Wood, S., Gravel, D. (2016a). mangal - making ecological network analysis simple. Ecography *39*, 384–390.

Poisot, T., Stouffer, D., Kéfi, S. (2016b). Describe, understand and predict : why do we need networks in ecology ? Functional Ecology *30*, 1878–1882.

Vázquez, P., Blüthgen, N., Cagnolo, L., Chacoff, N. (2009). Uniting pattern and process in plant–animal mutualistic networks : a review. Annals of Botany *103*, 1445–1457.

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.

CHAPITRE 5

COMBINING ECOLOGICAL THEORIES WITH MACHINE LEARNING USING FUZZY LOGIC

5.1 Description de l'article et contribution

Dans cet article, nous créons une nouvelle approche basée sur la logique floue pour réviser une théorie écologique. Notre approche prend des règles simples et les révise en ajoutant de nouvelles règles en logique floue. Ces règles sont claires et ont une interprétation écologique facilement compréhensible. Notre approche performe mieux pour prédire les interactions entre pollinisateurs et plantes que les meilleures approches en apprentissage supervisé (RF). De plus, contrairement à ces approches qui résultent en des modèles incompréhensibles, notre algorithme de révision aboutit à de simples règles.

J'ai conçu la méthode et programmer l'algorithme. Nacho a assemblé les données. Nacho, Timothée Poisot et Dominique Gravel m'ont aidé à éditer l'article. Cet article sera soumis sous sa forme actuelle à PLOS Computational Biology.

5.2 Abstract

We design an approach based on fuzzy logic to revise ecological theories with new rules learned from data. Our initial hypothesis has two simple rules to predict an interaction between a pollinator and a plant : if similar pollinators have an interaction with the plant, predict an interaction, otherwise we predict a non-interaction. Using an algorithm to revise this hypothesis, we find new rules to improve our model. Unlike most machine learning algorithms, the rules we discover have clear ecological interpretations. For example, our algorithm discovers that plants with small flower width interact more with pollinator with small tongues. Using the True Skill Statistic (TSS) to measure success, we find that our initial model before revision did worse than supervised learning techniques such as random forests (TSS of 0.689 vs 0.795), but that our revised theory improved the initial model enough to surpass random forests (TSS of 0.885). Our approach shows that it is possible to revise ecological theories and hypotheses with clear rules. True unification of machine learning with traditional ecological theories could help build large databases of ecological knowledge capable of both storing rules in a clear format, and automatically discover new rules from data.

5.3 Introduction

Biodiversity proved a difficult subject for traditional mathematical theories. After trying to devise an effective theory of molecular evolution for more than half a century, a consensus has emerged that established mathematical theories fail to correctly predict the level of genomic diversity witnessed in nature (Begun et al., 2007 ; Hahn, 2008). The issue is often not lack of understanding of basic mechanisms but the difficulty to integrate them into a coherent theory. In particular, much of the difficulties in both ecology and evolution have their origin in the complexity of selection (Bell, 2010). Selection, far from being a fixed value, then to show complex spatio-temporal variations. Selection has both complex abiotic (the environment changes) and biotic roots (species interactions), making it difficult to model in a single theory. The temptation, in both evolution and ecology, has been to dodge this complexity and rely on neutral theories (Kimura, 1983 ; Hubbell, 2001). However, both suffer from predictable issues : selection is both widespread and complex, ignoring it lead to unreliable theories (Hahn, 2008 ; McGill et al., 2006 ; 2007 ; Desjardins-Proulx et Gravel, 2012). At the same time, machine learning has achieved impressive results in a great of number of domains by learning model from data (the "unreasonable effectiveness of data" (Halevy et al., 2009 ; Murphy, 2012)). Unfortunately, machine learning models tend to be difficult, if not impossible, to interpret. We suggest that, to solve ecology's struggle with the complexity of ecosystems, we need techniques capable of combining hand-crafted ecological theories and hypotheses with machine learning to benefit from the clarity of the former and the ability of the latter to handle the complexity of data.

Machine learning is a branch of Artificial Intelligence interested in the study of algorithms capable of learning models from data (Russell et Norvig, 2009). In supervised learning, the algorithm is fed a set of data-points with the correct answer, use them to learn a model, and then has its accuracy tested on new data points. A canonical example of supervised learning is random forests (Breiman, 2001), an ensemble approach that will learn a number of decision trees and use the mode of their predictions. For this study, we designed an approach based on fuzzy logic to revise ecological theories, adding rules to an initial knowledge base given a data-set. The ultimate goal is to be able to build large knowledge bases of ecological theories, revise them automatically, test them as new data become available, while maintaining the ability to represent our knowledge with clear rules. Our approach is based on fuzzy logic, an extension of standard logic. Standard logic is bivalent : a statement is either true or false. Fuzzy logic breaks that mold by allowing truth values to be anything from 0 (false) to 1 (true). Fuzzy logic can often be confused with probability theory since they both involve values in the $[0, 1]$ range. Probability deals with the uncertainty of events, while fuzziness deals with ambiguity. One can sometime simulate the other (Murphy, 2012), but they are fundamentally different concepts. See (Kosko, 1990) for an extended discussion of fuzziness versus probabilities. For example, we can say in fuzzy logic that a viroid is neither a living organism not inanimate, since it shares many but not all features of living organisms. More concretely, it allows us to use clear logic formulas with standard logic connectives (conjunction "and" denoted $\wedge$, disjunction "or" denoted $\vee$) but with truth values in the $[0, 1]$ range. We use fuzzy logic to build fuzzy knowledge bases, which are series of $If\ X\ then\ Y$ rules.

To test our algorithm, we use a data-set of 739 pollinator-plant interactions and 739 non-interactionns. To test our algorithm's ability to revise theory, we use the K nearest neighbour as starting point (KNN). In a nutshell, to predict whether a pollinator interacts with a plant, the KNN algorithm picks the K most similar pollinators and count how many interact with the plant. Divided by K, this ratio should predict interactions. We start with an initial knowledge base with two rules : if the KNN ratio is high, there is an interaction, if the KNN ratio is low, there is no interaction. Then, we use an algorithm to revise this theory, adding new rules and modifying them to improve their ability to predict interactions. We find our approach capable of improving the theory from an initial TSS of 0.69 on the testing data-set to a revised theory with a TSS of 0.885, well above the scores of standard supervised learning methods (random forests and support vector machines). The rules added automatically to the knowledge bases have clear ecological interpretations. Our results show that techniques based on fuzzy logic can help combine ecological theories with machine learning techniques to achieve both clarity and predictive effectiveness.

5.4 Method

5.4.1 Data

Plant-pollinator interactions were collected across 16 sites in the southwest of the iberian peninsula (Huelva and Seville) from March to June 2015. All sites are dominated by Mediterranean open forests with a flowery understory including *Cistus sp.*, *Lavandula sp.*, *Rosmarinus officinalis*. Sampling was conducted on seven rounds per site along the season and each round consisted in three types of censuses. (i) A linear transect, in which all seen interactions were recorded along a 100 m transect during 30 minutes, (ii) focal 5 minutes census on the dominant plants were all visitors were recorded and (iii) opportunistic sampling of non previously recorded interactions. Those species that were unable to be identified in the field were collected and identified by professional taxonomists in the lab. Plant and pollinator traits relating to morphological trait matching (Bartomeus et al., 2016) were measured in the lab. Those are flower width, and nectar tube depth and diameter for plants and intertegular span (a proxy of body size) and tongue length for pollinators (Cariveau et al., 2016).

The data-set has 277 species of pollinators, 65 species of plants. Most pollinator-plant interactions were observed more than once but we are only interested in whether two species interact. The data-set has observations for 739 pairs of pollinator-plant. We randomly pick 739 pollinator-plant pairs that are not interacting to build a data-set of 1478 entries, half of interactions, half of non-interactions. 90% of the entries are used for training and 10% is held for testing. Seven real-valued features are defined for each entry : plant's nectar tube dimension,

Table 5.1 – The input variables along with their abbreviation and range. All ranges are in millimetres except generalism and the KNN ratio.

Variable	Abbreviation	Range
Plant's nectar tube dimension	TubeDim	[0, 20]
Plant's nectar tube depth	TubeDepth	[0, 17.7]
Plant's flower width	FlowerWidth	[1.73, 65]
Pollinator's body size	PollBodySize	[0.29, 75.5]
Pollinator's tongue length	TongueLength	[0, 45.2]
Pollinator generalism	PollGen	[1, 8]
K nearest neighbour ratio	KNN	[0, 1]

plant's nectar tube depth, plant's flower width, an estimate of the pollinator's body size, the pollinator's tongue length, the number of families the pollinator is interacting with (a rough specialism-generalism scale), and lastly the KNN ratio (Table 5.1). The KNN ratio is the ratio of the number of the K nearest pollinator interacting with the plant, divided by K. To find the K most similar species, we use the Tanimoto (or Jaccard) similarity measure on the set of plants each pollinator is interacting with. The Tanimoto measure is defined as the size of the intersection divided by the size of the union :

$$tanimoto(\mathbf{x},\mathbf{y}) = \frac{|\mathbf{x} \cap \mathbf{y}|}{|\mathbf{x} \cup \mathbf{y}|}. \tag{5.1}$$

For example, if a species x pollinates plants $\{A,B,C\}$ and species y pollinates plants $\{A,C,D,E,F\}$, then :

$$\begin{aligned} tanimoto(\mathbf{x},\mathbf{y}) &= \frac{|\{A,B,C\} \cap \{A,C,D,E,F\}|}{|\{A,B,C\} \cup \{A,C,D,E,F\}|} & (5.2) \\ &= \frac{|\{A,C\}|}{|\{A,B,C,D,E,F\}|} = \frac{2}{6}. & (5.3) \end{aligned}$$

The KNN algorithm yields better results with $K = 5$ (Desjardins-Proulx et al., 2017a).

5.4.2 Measure of success

We use the true skill score (TSS) to compare algorithms (Allouche et al., 2006). It is defined as

$$TSS = \frac{(tp \times tn) - (fp \times fn)}{(tp + fn)(fp + tn)}, \tag{5.4}$$

where tp, tn, fp, fn are the number of true positives, true negatives, false positives, and false negatives, respectively. The TSS ranges from -1 to 1. The TSS is used to train and test supervised learning algorithms (e.g. random forests) and our fuzzy knowledge base.

5.4.3 Supervised learning

We test three supervised learning algorithms available in the scikit-learn Python package : decision trees, support vector machines, and random forests (Pedregosa et al., 2011). Grid-search was used to find the optimal parameters for random forests. These algorithms use the same data as our fuzzy logic method, with 90% of the data being used for training and 10% held for testing.

5.4.4 Fuzzy knowledge base

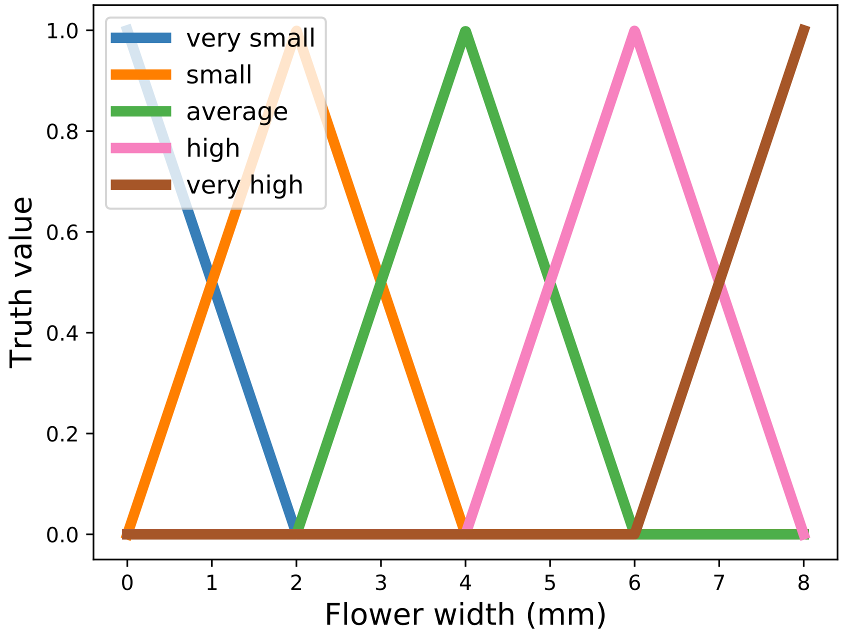
A knowledge base is a set of rules $K = \{R_0, R_1, ..., R_{|K|-1}\}$ of the form

$$\textbf{If } \textit{Antecedants} \textbf{ then } \textit{Consequents}. \tag{5.5}$$

Here the consequent is either I^- for a non-interaction or I^+ for an interaction. The antecedents are fuzzy variables of the form

$$InputVariable \text{ } is \text{ } LinguisticVariable \tag{5.6}$$

Figure 5.1 – Example of fuzzy sets.



An input (plant's flower width) and its five linguistic variables. The flower width ranges from 0 to 65. If a plant has a flower width of $22.6mm$, then the truth value of $FlowerWidth\ is\ verysmall$ would be 0.0, $FlowerWidth\ is\ small = 0.65$, $FlowerWidth\ is\ average = 0.35$, and $FlowerWidth\ is\ high = 0.0$. Triangles are popular in fuzzy logic because, for every input value, the sum of the fuzzy sets equals 1. There is no upper limits on the number of linguistic variables but all our tests were done with 2, 3, 5, or 7 per input variables.

and are joined together by conjunction ("and", $\wedge$). For example, the rule

$$\textbf{If } KNN\ is\ high \wedge PollGeneralism\ is\ low \textbf{ then } I^{+} \tag{5.7}$$

has two variables in the antecedent, KNN and $PollGeneralism$, respectively associated with the linguistic variables $high$ and low. See figure 5.1 for an explanation of how input and linguistic variables map to truth values. Conjunction $\wedge$ can take different definitions in fuzzy logic depending on the $T - norm$ used, see table 5.2.

We predict an interaction for a given input if the disjunction of the antecedents ($\vee$, see table

5.2 for definitions) for the rules with the I^+ consequent is higher than the disjunction of the antecedents with the I^- consequent. For example, given the three rules :

If $\textit{KNN is high} \land \textit{PollGen is low}$ **then** I^+
If $\textit{FlowerWidth is small} \ \land \textit{TongueLength is small}$ **then** I^+
If $\textit{KNN is low}$ **then** I^-

and input $KNN : 0.4, PollGeneralism : 1, FlowerWidth : 22.6, TongueLength : 9$, we get :

If $0.4 \land 0.5$ **then** I^+
If $0.65 \land 0.8$ **then** I^+
If 0.6 **then** I^-

We predict an interaction if the disjunction of the antecedents for the I^+ rules is higher than the disjunction of the antecedents for the I^- rules, in this case :

$$(0.4 \land 0.5) \lor (0.65 \land 0.8) > 0.6. \tag{5.8}$$

The result is close enough for the choice of norm to influence the result. The left-side of the inequality evaluates to 0.65 with the Gödel-Dummett norm, 0.616 with the Product norm, but only 0.54 with the Łukasiewicz norm. Thus, an interaction would be predict under the first two norms but not the last (table 5.2).

Table 5.2 – Definitions of conjunction ("and", $\wedge$) and disjunction ("or", $\vee$) following different norms.

Name	$x \wedge y \; (T - norm)$	$x \vee y \; (S - norm)$
Gödel-Dummett	$min(x, y)$	$max(x, y)$
Product	xy	$x + y - xy$
Łukasiewicz	$max(0, x + y - 1)$	$min(1, x + y)$

There are different ways to define fuzzy connectives so that they will evaluate to the same values as standard logic connectives when truth values are restricted to 0s and 1s. The results may only differ if there are truth values in the $(0, 1)$ range. For example, the simple formula $(x \wedge y) \vee (a \wedge b)$ with the truth values $\{x : 0.9, y : 0.7, a : 0.5, b : 0.8\}$ evaluates to 0.7 with the Gödel-Dummett norms, 0.778 with the product norms, and 0.9 under the Łukasiewicz norms. However, the same formula with truth values $\{x : 1, y : 1, a : 0, b : 1\}$ evaluates to 1 (true) for all norms. The $S - norm$ can be derived from the $T - norm$. By De Morgan's laws we have $x \vee y = \neg(\neg x \wedge \neg y)$. Negation in fuzzy logic can be defined as $\neg x = 1 - x$. If we wanted to derive the $S - norm$ from the product $T - norm$ we would get $1 - ((1 - x) \times (1 - y)) = x + y - xy$ (also called the probabilistic sum). These are three of the most common $T - norms$ used in fuzzy logic but there are others. All three norms were tested for all our experiments.

```
Data: Fuzzy knowledge base $k$, dataset $d$, time steps $t_{max}$
Result: Fuzzy knowledge base
$tss \leftarrow tss(k, d)$;
$t \leftarrow 0$;
while $t < t_{max}$ do
    $candidate \leftarrow generateRandomRule()$;
    $k \leftarrow k \cup \{candidate\}$;
    $newtss \leftarrow tss(k, d)$;
    if $newtss - tss > threshold(t)$ then
        $tss \leftarrow newtss$;
    else
        $k \leftarrow k - \{candidate\}$;
    end
    $t \leftarrow t + 1$;
end
```

Algorithm 1: Algorithm to learn new rules.

5.4.5 Revising theories

We begin with the initial theory that if the KNN ratio is high, then an interaction will occur, and if is low, the no interactions will occur. Our initial knowledge base is thus :

$$\textbf{If } KNN \textit{ is high } \textbf{then } I^{+}.$$
$$\textbf{If } KNN \textit{ is low } \textbf{then } I^{-}.$$

Then we apply an elementary algorithm to revise the theory. For 4000 time steps, we either generate a rule randomly, remove a rule, or try to modify an existing rule. The rules are kept in the knowledge base if they improve the TSS score on the training set above some threshold value that decreases with time :

$$threshold(t) = 0.15\alpha^{t}. \tag{5.9}$$

We test $\alpha \in \{0.95, 0.99, 0.995, 0.999, 0.9995, 0.9999\}$. With 0.95, the threshold decreases quickly so rules with small improvements are accepted early, while with 0.9999, the threshold decreases almost linearly. We start with a high threshold of 0.15, see algorithm 1.

It is possible to divide the range of the input variables in an arbitrary number of fuzzy sets (see figure 5.1 for an example with five fuzzy sets). We run the tests with 3, 5, 7, 9, and 11 fuzzy sets per input variables. In theory, more fuzzy sets could improve the performance by allowing more precision in the rules, but at the risk of overfitting and loss of clarity. The KNN input variable is always divided in two fuzzy sets (low, high) so the initial model remains the same across all simulations.

All tests are made with the three types of fuzzy logics described in table 5.2 : Gödel-Dummett, Product and Łukasiewicz.

5.5 Code

The C++ code also, along with the formatted data, is available in the *examples* folder of the ConjunctAI project (https://github.com/PhDP/ConjunctAI). The code is released under the permissive Apache 2 license and the data under the Creative-Commons CC-BY 4.0 license.

5.6 Results

Table 5.4 shows the TSS scores for the supervised learning algorithms, the initial fuzzy knowledge base, and the revised knowledge base. The initial KNN theory performs better than support vector machines but worse than decision trees and random forests. The revised fuzzy knowledge base beats all approaches and improves on the initial theory by roughly 0.2 on the TSS score. These results used the product norms for fuzzy logic.

Figure 5.2 shows how different norms performed with different number of fuzzy sets per input variable. Overall, the algorithm is highly resilient to changes in the $T - norm$, threshold function, in the probability of adding versus modifying rules, and in the number of fuzzy sets per input variables. The average TSS for all combinations of parameters remained in the [0.8, 0.9] range, above the established supervised learning algorithms. The best results were achieved with only three fuzzy sets per input variables, and a small but noticable decrease in performance is seen with more fuzzy sets per variables (Fig. 5.2). More fuzzy sets make the resulting rules more difficult to interpret and may lead to overspecific rules.

Table 5.5 shows a knowledge base after revision. Most of the rules have clear ecological explanations. Pollinators with small bodies can more easily enter inside the flower tubes and have fewer barriers. Plant flower width, a good proxy of flower size, may capture more generally how difficult it is to land in the flower and exploit it. Thus the rule that an interaction occurs for pollinators with a small tongue and flowers with a small flower width.

Half the generated rules had the interaction consequent (I^+) and half had the non-interaction (I^-). Despite this, all the new rules that made it into the knowledge base were for interaction. This is because the KNN theory under-predicts interactions and thus the algorithm has to compensate. The KNN model predicts correctly 98% of non-interactions but only 71% of

Table 5.3 – Full confusion matrix before and after theory revision with our fuzzy algorithm using the product norm.

Before revision		After revision	
True pos. 32%	*False pos.* 1%	*True pos.* 45%	*False pos.* 3%
False neg. 15%	*True neg.* 52%	*False neg.* 2%	*True neg.* 50%

Confusion matrix for a single run with the Product T-norm. The original model struggles to predict interactions and much of the gain in the TSS score is due to the revised model's ability to fix this weakness of the original model, which explains why almost all new rules cover interactions.

Table 5.4 – TSS scores for our fuzzy logic approach along with standard supervised learning algorithms.

Method	TSS
Support Vector Machine	0.564
Fuzzy logic initial model	0.659
Decision Tree	0.723
Random Forest	0.795
Fuzzy logic after revision	0.885

TSS scores for supervised learning algorithms, the initial fuzzy logic knowledge, and the knowledge base after theory revision. These results used the average of the three $T - norm$, three fuzzy sets per input variables, $\alpha = 0.95$ for the threshold function and no rule modification ($P_a(t) = 1$).

interactions. Table 5.3 shows the full confusion matrices before and after revision (Product norm).

Table 5.6 shows the 20 most frequent rules after revision.

The algorithm is also highly efficient. Our C++ implementation can process 5.7 million time steps under 2 minutes 20 seconds with an Intel Core i5-6600 CPU (four cores, 2015).

Figure 5.2 – Mean TSS on the testing data for the three $T - norm$ tested studied and 3, 5, 7, 9, or 11 fuzzy sets per input variables.

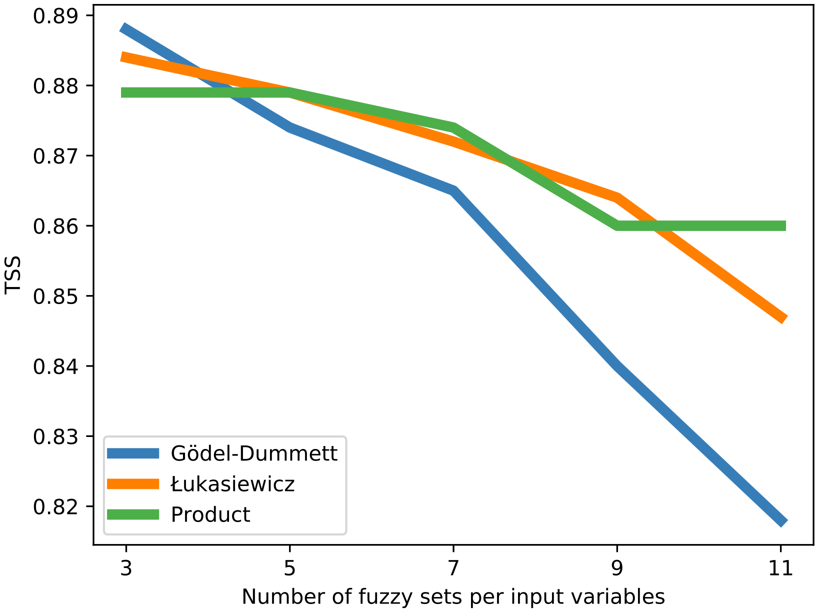


Table 5.5 – Example of a knowledge base after the theory revision algorithm revised the initial theory.

Rules
If NectarTubeDepth is average **then** I^+.
If PollBodySize is small **then** I^+.
If KNN is high **then** I^+.
If PollGen is small and NectarTubeDim is small **then** I^+.
If FlowerWidth is average **then** I^+.
If FlowerWidth is small and TongueLength is small **then** I^+.
If KNN is low **then** I^-.

Table 5.6 – Rules with the highest improvements in TSS when added to the initial knowledge base.

Δ	Rules
0.13	**If** BodySize is small **then** I^+
0.10	**If** TongueLength is small **then** I^+
0.10	**If** NectarTubeDim is small **then** I^+
0.09	**If** NectarTubeDepth is small **then** I^+
0.09	**If** BodySize is small **and** TongueLength is small **then** I^+
0.08	**If** TubeDim is small **and** BodySize is small **then** I^+
0.08	**If** TubeDepth is small **and** BodySize is small **then** I^+
0.08	**If** FlowerWidth is average **then** I^+.

5.7 Discussion

The K nearest neighbour provides a simple hypothesis for species interactions : if the species most similar to a species x interact with some species y, then x probably interact with y. However, this theory tend to underpredict interactions. Our fuzzy approach was able to fix this issue by finding rules specifically tailored to predict interactions. The algorithm was able to improve predictive abilities by adding rules with clear ecological interpretations, unlike state-of-the-art supervised learning algorithms such as random forests, which are highly effective but difficult to interpret. Furthermore, our algorithm is resilient to change in parameters and perform best with a small number of fuzzy sets per variable, making the rules easy to read and interpret. While our algorithm was able to even beat random forests for this particular problem, it is unlikely to be generally true. Random forests are well-established and tend to outperform most other machine learning techniques. The aim is to *approach* the performance of state-of-the-art algorithms while maintaining clarity and the ability to revise existing theories and hypotheses.

This study has two major limitations. First, it uses a rather naive theory to revision. While the KNN is often quite effective (Desjardins-Proulx et al., 2017b) it would be better to revise more established ecological theories such as the niche model (Williams et Martinez, 2000), selection theories of evolution (Gillespie, 1991), some form of the Lotka-Volterra equation, or even the neutral theory of biodiversity (Hubbell, 2001). A theory revision algorithm in a sufficiently rich representation should work with any ecological theory. Which leads to the second, and more subtle, limitation of our approach. Our algorithm is not relational, the input given to the fuzzy knowledge base is a vector. A data point is a point in a multidimensional space. However, rich mathematical representations such as predicate logic, which are used by mathematicians to store lemmas ("helping theorems" used to prove theorems), are relational. Data is represented in tables (the relations) with formulas defining the relationships between those tables. Theory revision of ecological theories should be explored in relational representations such as Markov logic (Richardson et Domingos, 2006), probabilistic soft logic (Kimmig et al., 2012 ; Bach et al., 2015), or fuzzy logics (Hájek, 1998).

There is a trend in machine learning toward methods capable of learning complex patterns at the cost of clarity. A recent deep learning network with a billion parameters is the perfect poster-child of this trend (Coates et al., 2013). While there are sometimes few alternative to these methods, for example for image and sound classification, we should work toward integration of scientific theories with machine learning to ensure that our methods remain clear to understand. This requires two steps : a deep reflection of the type of knowledge knowledge needed for ecological and evolutionary theories, and the design of algorithms to revise these theories. Ecology and evolution are fundamentally different than mathematics and physics. The former is happy with strict binary rules, which are expected to always hold true. The latter has a relatively small set of core mathematical equations which are, again, expected to always hold true. Uncertainty and ambiguity are everywhere in ecology. It offers a great opportunity for theoretical ecologists and machine learning specialists to develop representations and algorithms to handle ecosystem's complex rules, and combine fragmented theories into unifying knowledge bases.

5.8 References

Allouche, O., Tsoar, A., Kadmon, R. (2006). Assessing the accuracy of species distribution models : prevalence, kappa and the true skill statistic (tss). Journal of Applied Ecology *43*, 1223–1232.

Bach, S., Broecheler, M., Huang, B., Getoor, L. (2015). Hinge-loss markov random fields and probabilistic soft logic *arXiv : 1505.04406*.

Bartomeus, I., Gravel, D., Tylianakis, J., Aizen, M., Dickie, I., Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. Functional Ecology *30*, 1894–1903.

Begun, D., Holloway, A., Stevens, K., Hillier, L., Poh, Y., Hahn, M., Nista, P., Jones, C., Kern, A., Dewey, C., Pachter, L., Myers, E., Langley, C. (2007). Population genomics : Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLOS Biology *5*, e310.

Bell, G. (2010). Fluctuating selection : the perpetual renewal of adaptation in variable environments. Phil. Trans. R. Soc. B *365*, 87–97.

Breiman, L. (2001). Random forests. Machine Learning *45*, 5–32.

Cariveau, D., Nayak, G., Bartomeus, I., Zientek, J., Ascher, J., Gibbs, J., Winfree, R. (2016). The allometry of bee proboscis length and its uses in ecology. PLoS ONE *11*, e0151482.

Coates, A., Huval, B., Wang, T., Ng, D. W. A., Catanzaro, B. (2013). Deep learning with COTS HPC systems. Journal of Machine Learning Research Workshop and Conference Proceedings *28*, 1337–1345.

Desjardins-Proulx, P., Bartomeus, I., Poisot, T., Gravel, D. (2017a). The *knn* algorithm with threshold with application to species interaction.

Desjardins-Proulx, P., Gravel, D. (2012). How likely is speciation in neutral ecology ? The American Naturalist *179*, 137–144.

Desjardins-Proulx, P., Laigle, I., Poisot, T., Gravel, D. (2017b). Ecological interactions and the netflix problem. PeerJ *5*, e3644.

Gillespie, J. H. (1991). The Causes of Molecular Evolution. Oxford University Press, USA.

Hahn, M. (2008). Toward a selection theory of molecular evolution. Evolution *76*, 255–265.

Halevy, A., Norvig, P., Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems *24*, 8–12.

Hubbell, S. P. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Vol. 32 of Monographs in Population Biology. Princeton University Press.

Hájek, P. (1998). Metamathematics of Fuzzy Logic. Springer Netherlands.

Kimmig, A., Bach, S., Broecheler, M., Huang, B., Getoor, L. (2012). A short introduction to probabilistic soft logic. In : Proceedings of the NIPS Workshop on Probabilistic Programming.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kosko, B. (1990). Fuzziness vs probability. Int J General Systems *17*, 211–240.

McGill, B., Etienne, R., Gray, J., Alonso, D., Anderson, M., Benecha, H., Dornelas, M., Enquist, B., Green, J., He, F., Hurlbert, A., Magurran, A., Marquet, P., Maurer, B., Ostling, A., Soykan, C., Ugland, K., White, E. (2007). Species abundance distributions : moving beyond single prediction theories to integration within an ecological framework. Ecology Letters *10*, 995–1015.

McGill, B. J., Maurer, B. A., Weiser, M. D. (2006). Empirical evaluation of neutral theory. Ecology *87*, 1411–1423.

Murphy, K. (2012). Machine Learning : A Probabilistic Perspective. The MIT Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. Journal of Machine Learning Research *12*, 2825–2830.

Richardson, M., Domingos, P. (2006). Markov logic networks. Machine Learning *62*, 107–136.

Russell, S., Norvig, P. (2009). Artificial Intelligence : A Modern Approach, 3rd Edition. Prentice Hall.

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.

CHAPITRE 6

RELATIONAL APPROACHES TO ECOLOGICAL THEORIES

6.1 Description de l'article et contribution

Ce chapitre décrit une approche, la logique de Markov, qui permet d'unifier les probabilités avec le predicate logic, une forme avancée de la logique qui permet de représenter des équations mathématiques. Appliquée à un système tritrophique parasite -> insecte -> arbre, cette technique ne réussit pas à trouver de bonnes révisions. En fait, après plusieurs semaines de calcul, une seule nouvelle règle est découverte, et elle n'a pas vraiment de sens. Ce chapitre explique pourquoi la logique de Markov n'est probablement pas une bonne approche pour intégrer les théories en écologie et, s'inspirant du succès avec la logique floue, j'explore comment une approche relationnelle (basée sur le predicate logic) et floue pourrait résoudre les problèmes de la logique de Markov.

J'ai conçu cette recherche et fait les analyses sur les données. Timothée Poisot et Dominique m'ont aidé à éditer l'article.

6.2 Synthesis

Ecological theories are often formalized with mathematical equations using languages like predicate logic. When they are not, they can still be described more formally in some form of logic, especially if the logic is flexible (a major theme of this discussion). In the context of this chapter : *logic*, *relational*, and *mathematics* are mostly interchangeable. *Relational* refers to the fact that advanced logic systems like predicate logic organize data in tables, which are then linked together by relations. The concept will be more formally explainnated in the introduction to predicate logic. Formal representations offer quite a few advantages. They allow the combination of different theories into a single base, to rapidly check if a new idea conflicts with existing ones, to see if a data-set fits the theories, and to automatically revise the theories with machine learning algorithms. Traditional mathematical models tend to struggle in domains where machine learning thrives. Peter Norvig, director of research at Google, urge us to *stop expecting to find a simple theory, and instead embrace complexity, and use as much data as well as we can to help define (or estimate) the complex models we need for these complex domains*. Machine learning can build effective models in complex domains (Halevy et al., 2009 ; Murphy, 2012) but the resulting models are difficult, if not impossible, to interpret. Furthermore, they are generally incompatible with traditional mathematical theories. We'd like a relational representation capable of storing ecological theories and allowing us to automatically revise the theories given data. In short, we both want the clarity of mathematical theories along with the ability of machine learning algorithms to build solid predictive models from data.

To be clear, we are not looking for a representation for ecology itself. The debate is not about the nature of ecological data, but about the nature of ecological theories. Even if a probabilistic / statistical perspective would be enough to understand and analyze most ecological data-sets, it does not help us to represent mathematical ideas. That is : a relational representation is necessary to extract a more mechanistic understanding of data. Each knowledge representation has its limits (Table 6.1) and this is where things get complicated (and fascinating) for ecology. Mathematicians have built knowledge bases with millions of helper theorems to assist the discovery of new ideas (Kaliszyk et Urban, 2015). The Mizar Mathematical Library is a growing library of theorems, which are added after new candidate theorems are approved by the proof checker and peer-reviewed for style. Such library helps mathematicians deal with a growing body of knowledge and offers a concrete answer to the issue of knowledge synthesis. Mizar uses a language powerful enough to describe ecological theories, however, they cannot handle any form of ambiguity or uncertainty. In these pure logic representations, a statement is either true or false. Ecological theories require more flexibility. We do not expect our theories to be exactly true all the time, and some conflicts between theories is to be expected. At the same time, we do want the ability to represent mathematical ideas. Integration in ecology and evolution is difficult. Arguably the major roadblock for a selection theory of molecular evolution is our inability to handle the complex spatio-temporal variations of selection (Hahn, 2008 ; Bell, 2010). These spatio-temporal variations are caused by a number of abiotic and biotic changes in the environment, bringing ecology right into the equation for molecular evolution. Yet, efforts toward synthesis in ecology and evolution are on a case-by-case basis, we have no equivalent to Mizar, nor do even know how to represent ecological knowledge at scale.

In this chapter, we provide an introduction to relational languages such as predicate logic and discuss extensions based on probability theory and fuzzy logic. This text assumes basic notions of probability theory but no prior knowledge of predicate logic. We emphasize both the importance of thinking about representation as the first step toward integration of handcrafted theories and machine learning techniques. This is not the first time a relational approach to species interaction is being attempted. Bohan et al. (2011) and Tamaddoni-Nezhad et al. (2013) used a pure-logic approach to hypothesize species interactions given an initial handcrafted rule. Here, we focus on how to revise a theory in the context of ecological interactions and discuss issues with the approach along with possible solutions. Our research showns that fuzziness is important to theory revision (Desjardins-Proulx et al., 2017a) in ecology and we discuss how relational languages could be extended to include fuzziness.

Table 6.1 – Different languages can be distinguished by their ontological and epistemological commitments.

Language	Ontological commitment	Epistemological commitments
Propositional logic	Facts	$\{True, False\}$
Relational / Predicate logic	Facts, Objects, Relations	$\{True, False\}$
Probability theory	Facts	Belief $\in [0,1]$
Fuzzy logic	Facts with degree of truth	Interval value

The ontological commitment tells us what kind of objects exists within the language. The languages of propositional logic and probability theory only recognize facts (also called events), while relational languages (e.g. predicate logic) have objects and relations between them. The epistemological commitment tells us the kind of knowledge is allowed. Logic systems traditionally live in a binary true-or-false world, while probability theory allows a degree of truth for facts. There are several languages at the intersection of the basic languages here. For example, Markov logic unites predicate logic with probability theory (Richardson et Domingos, 2006), probabilistic soft logic unites fuzziness, probability theory, and predicate logic (Bach et al., 2015), while predicate fuzzy logic unites predicate logic with fuzziness (Hájek, 1998). Table adapted from Russel and Norvig (Russell et Norvig, 2009).

6.3 Relational thinking with predicate logic

In mathematics, a function f maps terms $\mathbf{X}$ (its domain) to other terms $\mathbf{Y}$ (its codomain) $f : \mathbf{X} \rightarrow \mathbf{Y}$. The number of arguments of a function, $|\mathbf{X}|$, is called its arity. A predicate is simply a function that maps 0 or more terms to a truth value : true or false, often denoted 0 and 1. A predicate is thus defined as $p : \mathbf{X} \rightarrow \{True, False\}$. Terms are either variables than range over a domain such as x or $city$, constants such as 42, $Manila$, π, or functions. By convention, variables always start with a lowercase letter. Variables have to be quantified either universally with $\forall$ (forall) or existentially with $\exists$ (exists). $\forall x : f(x)$ means $f(x)$ must hold true for all possible values of x, while $\exists x : f(x)$ means it must hold for at least one value of x. Quantifiers are often ignored and variables assumed to be universally quantified by default, such that $\forall x : x + 1 > x$ is simply written $x + 1 > x$. The type of predicate logic used in this text is called first-order logic, which allows only quantification over variables. Higher-order logics are needed to quantify over higher structures (e.g. sets, predicates).

Predicate logic builds formulas (or rules) by linking predicates using the unary connective $\neg$ (negation) or binary connectives (see table 5.2). For example, $PreyOn(s_x, s_y)$ is a predicate that maps two species to a truth value, in this case whether the first species preys on the second

Table 6.2 – List of common binary connectives and their truth table with T standing for True and F for False.

			Truth table			
Name	Common	Symbols	T x T	T x F	F x T	F x F
Conjunction	and	$\wedge$	T	F	F	F
Disjunction	or	$\vee$	T	T	T	F
Implication	implies	$\Rightarrow$	T	F	T	T
Material equivalence	iff	$\Leftrightarrow$	T	F	F	T
Exclusive disjunction	xor	$\veebar$	F	T	T	F

iff is read *if and only if*. Implication is one of the most common connective and may have surprising behavior. It will always return true when the left side is false. For example we may say $NonNegative(x) \Rightarrow \sqrt{x^2} = x$, and this formula holds for all real numbers, including negative ones. For example with $x = -1$, $NonNegative(-1)$ is False and $F \Rightarrow F$ returns True.

species. We could build more complex formulas from there, for example :

$$\forall s_x : \neg PreyOn(s_x, s_x). \tag{6.1a}$$

$$\forall s_x, s_y : PreyOn(s_x, s_y) \Rightarrow Larger(s_x, s_y). \tag{6.1b}$$

$$\forall s_x, s_y : PreyOn(s_x, s_y) \wedge \neg IsParasite(s_x) \Rightarrow Larger(s_x, s_y). \tag{6.1c}$$

A set of formulas like the one formed by the equations in 6.1 is called a knowledge base. The first formula says that species don't prey on themselves. The second says that predators are larger than their preys. The third refines the second formula by adding that predators are larger than their preys unless the predator is a parasite. None of these rules are expected to the true all the time, and we will cover ways to introduce uncertainty into logic, but for now we focus on pure logic.

Predicate logic is almost synonymous with mathematics and all databases of mathematical knowledge rely on it (or a form of type theory) for theorem proving. Thus, it is expressive enough to represent and manipulate all the mathematical theories in ecology and evolution. For example, an important axiom of real numbers is that, for every real number x, there is a real number y such that $y = x + 1$:

$$\forall x : Real(x) \Rightarrow \exists y : Real(y) \wedge y = x + 1. \tag{6.2}$$

Let's dissect this formula.

- We have four terms. x and y are variables ranging over numbers, 1 is a constant, and the function $+$ maps two numbers to a number.
- We have two predicates : $Real$ and $=$. $Real$ is a predicate with arity 1, it returns true for real numbers. $=$ has two arguments (left and right-hand sides), and returns true if both sides have the same value.
- We have two connectives, $\wedge$ ("and") and $\Rightarrow$ ("implies").
- We have two quantifiers. The first, $\forall$, tells us the entire formula should be true for all values of x. The second, $\exists$, tells us there is one value of y that satisfied $Real(y) \wedge y = x + 1$.

For an ecological twist, we could say that if species s_x is a predator found in location l, then there is a species s_y such that $PreyOnAt(s_x, s_y, l)$ is true :

$$\forall s_x, l : Predator(s_x) \wedge Presence(s_x, l) \Rightarrow \exists s_y : PreyOnAt(s_x, s_y, l). \tag{6.3}$$

Everything from the axioms of probability to the Lotka-Volterra equation can be expressed with predicate logic formulas :

$$\dot{x} = \alpha x - \beta xy \wedge \dot{y} = \delta xy - \gamma y. \tag{6.4}$$

Here equality $=$ is the only predicate, the time differential $\dot{x}$ and multiplications are functions, x and y are variables, and lastly $\alpha, \beta, \delta, \gamma$ are constants.

Like mathematics, we need to replace variables with actual values to get a value. $PreyOn(s_x, s_y)$ can be neither true nor false until we assign constants to the variables s_x and s_y. The process of replacing variables with constants is called **grounding**, and we talk of ground terms / predicates / formulas when no variables are present. Given the set of species $\{Manticore, ApisFlorea, Kirin\}$, we have 3^2 groundings for the predicate $PreyOn(s_x, s_y)$:

$$\begin{gathered}
PreyOn(Manticore, Manticore) \\
PreyOn(Manticore, Kirin) \\
PreyOn(Manticore, ApisFlorea) \\
PreyOn(ApisFlorea, Manticore) \\
PreyOn(ApisFlorea, Kirin) \\
PreyOn(ApisFlorea, ApisFlorea) \\
PreyOn(Kirin, Manticore) \\
PreyOn(Kirin, Kirin) \\
PreyOn(Kirin, ApisFlorea)
\end{gathered} \tag{6.5}$$

The explosion in the number of ground predicates and formulas is a major computational issue in logic. One strength of logic is its ability to connect several ideas with clear formulas. Let's use a simple knowledge base to see how inference works in logic :

$$IsHuman \Rightarrow IsMammal. \tag{6.6a}$$

$$IsMammal \Rightarrow IsVertebrate. \tag{6.6b}$$

This knowledge base has three predicates : $IsMammal$, $IsVertebrate$, and $IsHuman$, all without arguments. Since they have no variables, they are ground predicates. A **possible world** is a truth assignment to all ground predicates. For example,

$$\{IsMammal, IsVertebrate, \neg IsHuman\} \tag{6.7}$$

is a possible world. To simplify, we only write the name of the predicate in the possible world to say it is true and use $\neg$ to denote it is false, so the previous possible world could also be written more explicitely

$$\{IsMammal : True, IsVertebrate : True, IsHuman : False\}. \tag{6.8}$$

We say a possible world **satisfies** a formula (or a knowledge base) if the formula is true with its assignments. $\{IsMammal, IsVertebrate, \neg IsHuman\}$ satisfies the knowledge base 6.6 since the two formulas are true with this assignment ($F \Rightarrow T$ and $T \Rightarrow T$ are both true, see table for clarifications on implication it this is unclear).

The most basic inference problem is to determine whether a knowledge base KB **entails** a formula f, or $KB \models f$. We want to know if the knowledge base 6.6 entails the formula

$$IsHuman \Rightarrow IsVertebrate. \tag{6.9}$$

or

$$IsVertebrate \Rightarrow IsHuman. \tag{6.10}$$

Formally, the entailment $KB \models f$ means that for all possible worlds in which KB is true, f is true. More intuitively, it can be read as the formula *following from* the knowledge base. It can be verified with a brute approach by checking the 2^n possible worlds (n being the number of predicates). In this case, it can be verified that our knowledge base does entail formula 6.9 but not formula 6.10. It does not entail the latter since the possible world $\{IsVertebrate, \neg IsHuman, IsMammal\}$ satisfies the knowledge base but not formula 6.10. Entailment ensures new formulas are consistent with an existent body of knowledge.

Another important question is whether there is a possible world to satisfy a given knowledge base. This is the Propositional Satisfiability Problem, or just SAT, and is found everywhere from theorem proving to resolving software dependencies (Russell et Norvig, 2009 ; Harrison, 2009). The problem is difficult to solve since there are 2^n possible worlds given a knowledge base with

n predicates, but stochastic search algorithms like WalkSAT are fast enough in practice to find the correct solution even with close to a million symbols. In a nutshell, WalkSAT starts with a random truth assignment to the predicates (a possible world), picks an unsatisfied formula in the knowledge base and flips randomly the value of one of its predicate. If it fails to find a possible world that satisfies all formulas of the knowledge base after a given number of flips, the algorithm starts again with a new random possible world. An important generalization, the MaxSAT problem, tries to find the assignment that will satisfy the most, but not necessaily all, formulas and is related to MAP inference in unified probabilistic predicate logic.

6.3.1 Relational ?

A key word in the logic world is *relational*. The techniques we will explore to unify logic with probability are often called *statistical relational* and the concept is important to distinguish advanced logic systems with simpler ones. The predicate $PreyOn(species_x, species_y, location)$ can be seen as a table with three columns (the predator, the prey, the location). Formulas with several predicates, like $Larger(species_x, species_y) \Rightarrow PreyOn(species_x, species_y, location)$, link these tables together. This is quite different from statistics, where data is generally organized in a single table (see Table 6.1), and points are essentially reduced to a location in a multidimensional space. By separating data by predicates, relational representations can scale to large knowledge bases with thousands of different domains.

6.3.2 Clausal forms

Most, if not all, logic programs convert formulas into clauses for faster processing. It is important to understand this representation well since it is the most common way to write logic formulas. A clause is in the form :

$$B_0 \wedge B_1 \wedge ... \wedge B_{|\mathbf{B}|-1} \Rightarrow H_0 \vee H_1 \vee ... \vee H_{|\mathbf{H}|-1}, \tag{6.11}$$

where $\mathbf{B}$ is the set of predicates of the *body* and $\mathbf{H}$ is the set of predicates in the *head*. Is it also common to write clauses in the form $\mathbf{H} \leftarrow \mathbf{B}$. Clauses can intuitively been read as

"If **B** $then$ **H**" rules. All formulas can be converted into clauses without loss of generality, although some formulas will be converted to more than one clause (Harrison, 2009). For example, $X \Leftrightarrow Y$ will be converted to the two clauses $X \Rightarrow Y$ and $Y \Rightarrow X$. Clauses are implicitly joined by conjunctions ("and", $\wedge$). The technical term for this type of clause is CNF : conjunctive normal form. Many programs, such as Alchemy2 for Markov logic (Domingos et Lowd, 2009), will accept standard formulas as input but convert them to clauses for processing and output.

Several logic languages such as Prolog only accept clauses with at most one element in the head ($|\mathbf{H}| \leq 1$). It simplifies inference at the cost of expressiveness, since some formulas cannot be converted to clauses with at most one element in the head.

6.4 Probabilistic graphical models

Probabilistic graphical models combine graph theory with probability theory to represent complex probability distributions while exploiting independences (Koller et Friedman, 2009 ; Barber, 2012). There are primarily two motivations behind probabilistic graphical models. First, even for binary random variables, we need to learn $2^n - 1$ parameters for a distribution of n variables. This is unmanageable on many levels : it is computationally difficult to do inference with so many parameters, it requires a large amount of memory, and makes it difficult to learn parameters without an unreasonable volume of data (Koller et Friedman, 2009). Second, probabilistic graphical models provide important information about independences and the overall structure of the distribution.

For most probabilistic graphical models, the distribution is represented by a graph $G = (\mathbf{V}, \mathbf{A})$, where $\mathbf{V}$ is a set of vertices (the random variables) and $\mathbf{A}$ a set of directed or undirected edges (or arcs). Two important types of graphical models used for representation are Bayesian networks and Markov networks. We will focus on binary variables with the convention that $x = 0 \equiv \neg x \equiv false$ and $x = 1 \equiv x \equiv true$. Context should make clear when x refers to the variable or to $x = 1$.

Bayesian networks are directed acyclic graphs that represent the joint probability for a set of variables $\mathbf{X}$ as :

$$P(\mathbf{X}) = \mathbf{x}) = \prod_{x_i \in \mathbf{x}} P(x_i | Pa(x_i)), \tag{6.12}$$

where $Pa(x_i)$ is the set of parents of variable x_i. Because no cycles are allowed, the set $Pa(x_i)$ can only involve variables already seen on the left of x_i, so $P(a)P(b|a)p(c)$ is a valid Bayesian networks but not $P(a)P(b|c)P(c|b)$. See figure 6.1 for a detailed example. Bayesian network's inability to model cycles and symmetric relationship made them unpopular as a foundation for unified probabilistic-logic approaches (but see (Jaeger, 1997)). An alternative is the undirected Markov network, where the joint probability is defined as

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \prod_i \phi_i(\mathbf{x}), \tag{6.13}$$

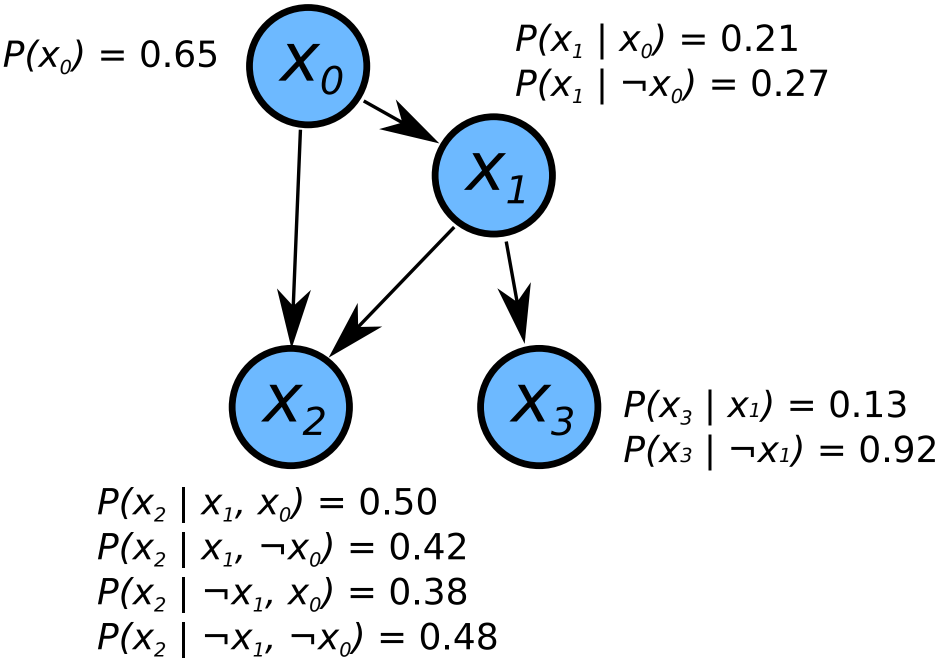
with the partition function Z being

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_i \phi_i(\mathbf{x}), \tag{6.14}$$

with $\mathbf{x} \in \mathcal{X}$ being all possible assignments to $\mathbf{x}$. Markov networks are bit less intuitive to understand. Since the edges are not directed, they don't represent simple conditional relationships of the form $P(x|y)$ like Bayesian networks. Instead, Markov networks have factors ϕ, in this case one for each maximal clique in the network. In graph theory, a clique is a fully connected subset of the graph, a maximal clique is a clique that is not part of a larger clique. Factors are not probabilities, they capture affinities between variables. To extract a probability, we must thus divide the product of the factors by Z, the sum of all possible products of factors. Figure 6.2 shows a complete example.

Most modern approaches to extend probabilistic graphical models with predicate logic rely on Markov networks (Richardson et Domingos, 2006 ; Bach et al., 2015), and ofen using a

Figure 6.1 – Inference in Bayesian networks.



A bayesian network with four binary variables. Bayesian networks are limited to directed acyclic graphs. Thus, it would be impossible to have a bidirectional relationship or an edge from x_3 to x_0. An example query : $P(\neg x_0, x_1, \neg x_2, x_3) = P(\neg x_0)P(x_1|\neg x_0)P(\neg x_2|\neg x_0, x_1)P(x_3|x_1)$. In this case $P(\neg x_0, x_1, \neg x_2, x_3) = (1-0.65) \times 0.27 \times (1-0.42) \times 0.13 = 0.05$.

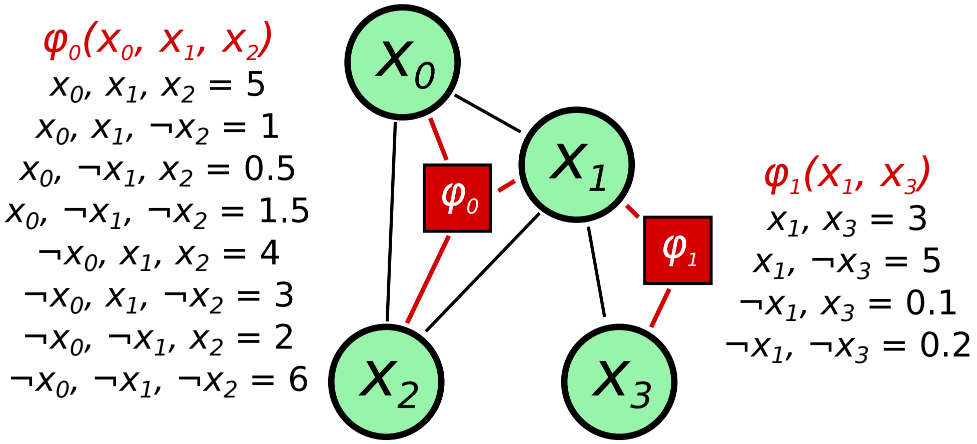
log-linear ("maximum entropy") model :

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z}\prod_k \phi_k(x_{\{k\}}) = \frac{1}{Z}\exp\left(\sum_i w_i f_i(\mathbf{x}_i)\right), \tag{6.15}$$

where w_i is the weight of factor i, $\mathbf{x}_i$ is the subset of $\mathbf{x}$ in factor i, and f_i is any real-valued function of the state. Equation 6.15 defines a structured log-linear model. This approach to Markov networks will be used for unifying probabilistic graphical models with predicate logic, using logic formulas instead of maximal cliques to define factors.

Given a set of random variables $\mathbf{W}$ and evidence $\mathbf{E} \subseteq \mathbf{W}$, the most common queries for probabilistic graphical models are the conditional probability and MAP queries. For conditional probability, we have a set of query variables $\mathbf{X} \subseteq \mathbf{W}$ and want to evaluate $P(\mathbf{X} = \mathbf{x}|\mathbf{E} = \mathbf{e})$.

Figure 6.2 – An example of Markov logic networks.


$\varphi_0(x_0, x_1, x_2)$
$x_0, x_1, x_2 = 5$
$x_0, x_1, \neg x_2 = 1$
$x_0, \neg x_1, x_2 = 0.5$
$x_0, \neg x_1, \neg x_2 = 1.5$
$\neg x_0, x_1, x_2 = 4$
$\neg x_0, x_1, \neg x_2 = 3$
$\neg x_0, \neg x_1, x_2 = 2$
$\neg x_0, \neg x_1, \neg x_2 = 6$
x_0
x_1
x_2
x_3
φ_0
φ_1
$\varphi_1(x_1, x_3)$
$x_1, x_3 = 3$
$x_1, \neg x_3 = 5$
$\neg x_1, x_3 = 0.1$
$\neg x_1, \neg x_3 = 0.2$

A Markov networks with four binary variables. Unlike Bayesian networks, Markov networks can form loops and the edges are undirected, but their interpretation is a bit trickier than Bayesian networks. A popular variant of Markov network is the maximal-clique network seen here. A clique is a fully connected subgraph. This network has two maximal cliques : one for vertices x_0, x_1, x_2 and one for x_1, x_3. In red : the related factor graph, with the red square vertices representing the factors in the maximal cliques. Here the partition function $Z = 107$, and thus $P(\neg x_0, x_1, x_2, \neg x_3) = \phi_0(\neg x_0, x_1, x_2)\phi_1(x_1, \neg x_3)/Z = 4 \times 5/107 = 0.1869$.

When $|\mathbf{E}| > 0$ we have a conditional probability, otherwise we have a joint probability query for $|\mathbf{X}| > 1$ and a marginal probability with $|\mathbf{X}| = 1$. MAP inference is the task of finding the most probable joint probability given evidence, or :

$$\mathrm{MAP}(\mathbf{w}|\mathbf{e}) = \mathrm{argmax}_w P(w, e). \tag{6.16}$$

MAP inference is arguably the most common inference task in practice since it provides an assignment (and thus a prediction) for all variables. Another key function of probabilistic graphical model is to serve as a foundation for learning. Early models were hand-crafted by experts but we can learn parameters from data, and modern algorithms can be used to learn the graphical structure (or revise existing one) (Friedman et al., 1999 ; Yuan et Malone, 2013).

6.5 Markov logic : The union of predicate logic with probability theory

A pure logic knowledge base can be seen as a set of hard constraints. If we have a formula saying $\neg PreyOn(s_x, s_x)$, it means cannibalism is impossible. Any possible world with, say, $PreyOn(Human, Human)$ violates the formula and is thus false. We want a framework where it is possible to define soft constraints such as a possible world that violates a formula would be less likely but not necessarily impossible. We also want to be able to ask probabilistic queries about our variables. Markov logic networks (Richardson et Domingos, 2006) do exactly this using weighted formulas that serve as template for Markov networks.

In Markov logic, a knowledge base is a set of tuple $KB = \{(f_i, w_i)\}_{i=1}^{N}$, with f being a formula in first-order predicate logic and w a weight $\in \mathbb{R}$ associated with it. The intuition is : the higher the weight associated with a formula, the greater the penalty for violating it (or alternatively : the less likely a possible world is). Before delving into the internal mechanics of Markov logic, let's begin with a simple example :

$$\neg PreyOn(s_x, s_x), 2.0. \tag{6.17a}$$

$$s_x \neq s_y \wedge PreyOn(s_x, s_y) \Rightarrow \neg PreyOn(s_y, s_x), 2.7. \tag{6.17b}$$

$$s_x \neq s_y \wedge PreyOn(s_x, s_y) \wedge \neg IsParasite(s_x) \Rightarrow Larger(s_x, s_y), 1.1. \tag{6.17c}$$

This Markov logic networks has three formulas. First : cannibalism is rare (with weight 2.0). In pure logic, this formula would mean : cannibalism is impossible, but we simply say it is rare. The second formula says if the species s_x and s_y are different ($s_x \neg s_y$), predation is assymetric. This formula has the highest weight. The last formula says that if species s_x and s_y are different, if s_x preys on s_y, and if s_x is not a parasite, then s_x should be larger than s_y. Implication ($\Rightarrow$) is often use to define formulas that could roughly be translated as $if a, b, cthend$. Implication is also a bit counter-intuitive, remember that if the left-side is false, then the formula is true (e.g. $False \wedge True \Rightarrow True$ is $False$, see table 5.2 for details).

Weights range from $-\infty$ to ∞. in short they can be any real number. A formula with an infinite weight cannot be unsatisfied without making the possible world impossible. Thus, "normal" non-probabilistic logic is a subset of Markov logic where all formulas have a positive infinite weight. Formulas with weights close to 0 have little effect on the probabilities, in short the cost of violating them is small. A formula with a negative weight is expected to be false. For the rest of this chapter, we will assume that all weights are positive real numbers (including zero), with no loss of generality since $(f, -w) \equiv (\neg f, w)$.

As in logic, we must ground the knowledge base with constants. Let's say we have only two species, the disgusting Snow Crab (represented by the letter C) and the gentle (but murderous) Bearded Seal (represented by the letter S). A possible world with these constants could be :

$$\begin{gathered}\{PreyOn(C,C), \neg PreyOn(C,S), PreyOn(S,C), \neg PreyOn(S,S), \neg IsParasite(S), \\ \neg IsParasite(C), \neg Larger(C,C), \neg Larger(C,S), Larger(S,C), \neg Larger(S,S)\}\end{gathered} \tag{6.18}$$

We simply write p and $\neg p$ to denote whether the ground predicate is true or false. Applied to our Markov logic networks, these ground predicates give the ground formulas :

$$\neg PreyOn(C,C) : False. \tag{6.19a}$$

$$\neg PreyOn(S,S) : True. \tag{6.19b}$$

$$C \neq S \wedge PreyOn(C,S) \Rightarrow \neg PreyOn(S,C) : True. \tag{6.19c}$$

$$S \neq C \wedge PreyOn(S,C) \Rightarrow \neg PreyOn(C,S) : True. \tag{6.19d}$$

$$C \neq S \wedge PreyOn(C,S) \wedge \neg IsParasite(C) \Rightarrow Larger(C,S) : True. \tag{6.19e}$$

$$S \neq C \wedge PreyOn(S,C) \wedge \neg IsParasite(S) \Rightarrow Larger(S,C) : True. \tag{6.19f}$$

The **cost** of a possible world is the sum of the weights of the unsatisfied formulas. Here, only formula 6.19a is unsatisfied under possible world . Since this formula has a weight of 2 in 6.17, we say the cost of this possible world is 2. We omitted ground formulas that are false because of the $s_x \neq s_y$ predicate in both the listing 6.5 and in figure 6.3, for example $C \neq C \wedge PreyOn(C,C) \Rightarrow \neg PreyOn(C,C)$. These ground formulas have factors that are always true regardless of the possible world.

A different possible world with a higher cost would be :

$$\begin{gathered}\{PreyOn(C,C), PreyOn(C,S), PreyOn(S,C), PreyOn(S,S), \neg IsParasite(S),\\ \neg IsParasite(C), \neg Larger(C,C), \neg Larger(C,S), \neg Larger(S,C), \neg Larger(S,S)\}.\end{gathered} \tag{6.20}$$

would leave most formulas in unsatisfied for a cost of 11.6 (to verify as an exercise). The last possible world also show something odd : we have both $\neg Larger(C,S)$ and $\neg Larger(S,S)$, which is impossible. Since Markov logic does allow hard constraints in the form of formulas with infinite weight, we could add to our knowledge base :

$$\neg Larger(s_x, s_x), \infty. \tag{6.21a}$$

$$s_x \neq s_y \wedge Larger(s_x, s_y) \Rightarrow \neg Larger(s_y, s_x), \infty. \tag{6.21b}$$

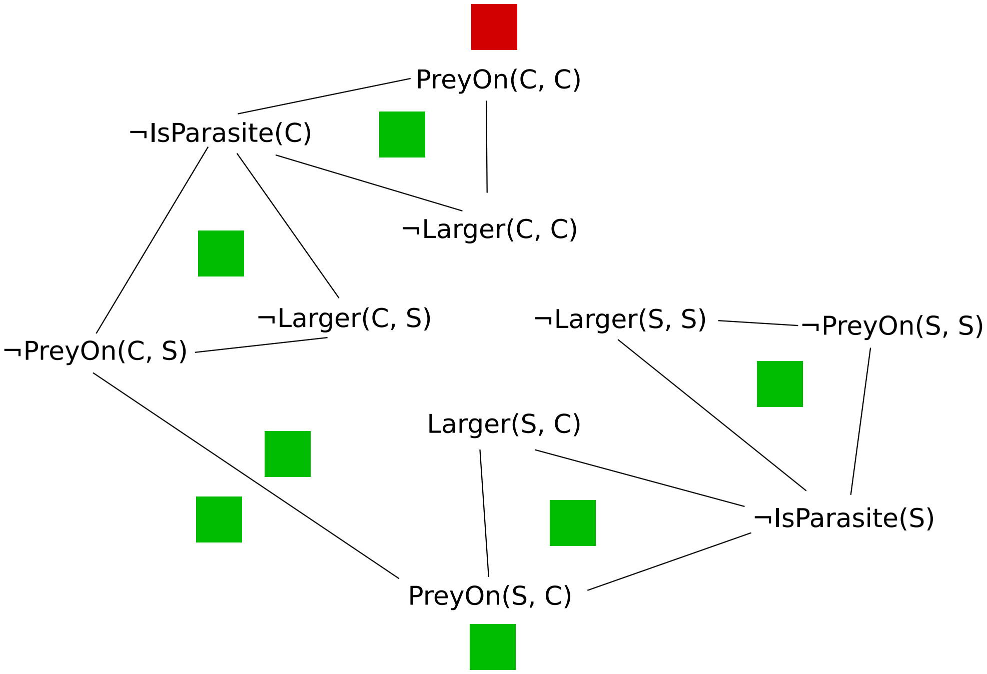
Finding the possible world with the highest joint probability (MAP inference) can be done by looking for the possible world with the lowest cost. The algorithm for MAP inference in Markov logic is similar to the SAT problem encountered in section 6.3, and a variant of WalkSAT, MaxWalkSAT, can be used. Roughly speaking : we start with a random possible world, compute its cost, pick a random predicate from an unsatisfied formula and flip its value and keep the new possible world if it has a lower cost. Random steps and restarts are used to prevent being stuck in a local minimum.

Figure 6.3 shows the Markov network for a grounded knowledge base. Markov logic networks act as template for log-linear Markov networks, but instead of having a factor for each maximal clique, the network has a factor for each grounded formula. Predicates form the vertices and are linked to other predicates if they are present in the same formula. The resulting network defines a probability distribution with the same formula as the log-linear Markov network :

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} \exp\left(\sum_i w_i f_i(x)\right), \tag{6.22}$$

where w_i is the weight of formula i and f_i is the indicator function for formula i (it equals 0 if the formula is false and 1 if it is true).

Figure 6.3 – Inference in Markov logic networks.



A Markov logic network is a set of weighted formulas that serves as a template for a log-linear Markov network. All predicates found in the same formula are connected with an edge, and each grounded formula has a factor (square vertices). Unlike a maximal-clique Markov network, some factors are not associated with a maximal clique. Here for example, the two factors from the formula $\neg PreyOn(s_x, s_x)$ are not maximal cliques. Factors are green if the formula evaluates to true given the possible world in 6.5 and red if it evaluates to false. The cost of a possible world in is the sum of the weights of the unsatisfied formulas. Here, the cost would only be 2 (the weight of the only unsatisfied rule : $\neg PreyOn(s_x, s_x)$). We are not showing formulas discarded by $s_x \neg s_y$, i.e. $C \neq C \wedge PreyOn(C, C) \Rightarrow \neg PreyOn(C, C)$.

6.6 Learning weights and theory revision in Markov logic

The first step to theory revision in Markov logic is to learn the weights of the initial formulas. Weights are learned by maximizing the weighted pseudo-log-likelihood of a database (our data-set) in regard to weights (see (Domingos et Lowd, 2009) for details). Second-order algorithms such as the L-BFGS (Nocedal et Wright, 2006) should converge quickly to optimal weights, especially if they start close to a solution (Domingos et Lowd, 2009). Unfortunately, we have very few guarantees that the weights are optimal both because the L-BFGS can only converge to local optima but also, more fundamentally, because it is unclear how optimizing the weighted pseudo-log-likelihood really approximate optimizing the log-likelihood. The log-likelihood is the log of equation 6.22, but it is too computationally expansive to optimize and is biased toward formulas with highest-arity predicates (those with many arguments).

Structure learning has been well-studied in Markov logic. Kok and Domingos' 2005 (Kok et Domingos, 2005) paper on structure learning Markov logic preceded the publication of the first complete description of Markov logic (Richardson et Domingos, 2006). At its core, theory revision in Markov logic involves generating a candidate formula, recomputing the weights, and checking whether a weighted pseudo-log-likelihood measure of a database has improved sufficiently to justify adding the formula to the knowledge base. Recent algorithms have improved mostly by better selecting which candidate rules for revision (Mihalkova et Mooney, 2007 ; Kok et Domingos, 2009 ; 2010).

6.7 Markov logic meets the Salix data-set : initial model

We explore theory revision with Markov logic using the Salix data-set (Kopelke et al., 2017). The Salix data-set has parasites, gallers (insects), and salix, forming a tritrophic ecological network ($Parasite \rightarrow Galler \rightarrow Salix$). Furthermore, we have partial phylogenetic information for the species, their presence/absence in a number of locations, interactions, and some environmental information on the locations (Table 6.3).

Table 6.3 – The 18 predicates of the Salix data-set with their two domains : species and location.

Predicate	Domain	Positives
$PreyOn(x,y)$	$species \times species$	1080
$PreyOnAt(x,y,l)$	$species \times species \times location$	4342
$PresenceAt(x,l)$	$species \times location$	3906
$IsParasite(x)$	$species$	126
$IsGaller(x)$	$species$	96
$IsSalix(x)$	$species$	52
$CloselyRelated(x,y)$	$species \times species$	27522
$ModerateCooccurrence(x,y)$	$species \times species$	508
$HighCooccurrence(x,y)$	$species \times species$	98
$HigherPhyloValue(x,y)$	$species \times species$	19366
$HighTemperature(l)$	$location$	186
$HighPrecipitation(l)$	$location$	186
$HighSpeciesRichness(l)$	$location$	208
$HighSalixRichness(l)$	$location$	374
$HighGallsRichness(l)$	$location$	238
$HighParasitoidRichess(l)$	$location$	187
$HighLinkDensity(l)$	$location$	203
$HighConnectance(l)$	$location$	175

There are 126 species of parasites interacting with 96 species of gallers, interacting with 52 species of Salix. Phylogenetic values for the HigherPhyloValue were computed using the Ape R package. Predicate starting with $High$ are based on an arbitrary cutoff around the average value, for example $HighPrecipitation$ is true for all of the 641 location where the precipitations were above average. In total there are 58 853 positive literals. We work under the the closed-world assumption that if a predicate is not positive, then it is negative.

We provide an initial model based on our understanding of the interactions and the variables available :

$$\neg PreyOn(x, x) \tag{6.23a}$$

$$PreyOnAt(x, y, l) \Rightarrow PreyOn(x, y) \tag{6.23b}$$

$$IsGaller(x) \wedge PreyOn(x, y) \Rightarrow IsSalix(y) \tag{6.23c}$$

$$IsParasitoid(x) \wedge PreyOn(x, y) \Rightarrow IsGaller(y) \tag{6.23d}$$

$$IsSalix(x) \Rightarrow \neg PreyOn(x, y) \tag{6.23e}$$

$$CloselyRelated(x, y) \wedge PreyOn(x, z) \Rightarrow PreyOn(y, z) \tag{6.23f}$$

$$HigherPhyloValue(x, y) \Rightarrow HighCooccurrence(x, y) \tag{6.23g}$$

$$PreyOn(x, y) \Leftrightarrow PreyOn(x, x) \tag{6.23h}$$

$$PreyOn(x, y) \Leftrightarrow HighCooccurrence(x, y) \tag{6.23i}$$

$$HigherPhyloValue(x, y) \Rightarrow PreyOn(x, y) \tag{6.23j}$$

6.8 Results

We ran the basic learning algorithm from Alchemy-2 (Richardson et Domingos, 2006). After three weeks, the algorithm found no substantial revisions. The resulting knowledge base in clausal form is :

$$PreyOnAt(x,y,z) \Rightarrow PreyOn(x,y), 5.20. \tag{6.24a}$$

$$HighCooccurrence(x,y) \Rightarrow PreyOn(x,y), 4.22. \tag{6.24b}$$

$$IsGaller(x) \wedge PreyOn(x,y) \Rightarrow IsSalix(y), 4.15. \tag{6.24c}$$

$$IsParasitoid(y) \wedge PreyOn(y,x) \Rightarrow IsGaller(x), 3.49. \tag{6.24d}$$

$$PreyOn(x,y) \Rightarrow HighCooccurrence(x,y), 1.57. \tag{6.24e}$$

$$PreyOn(x,y) \Rightarrow PreyOn(x,x), 1.52. \tag{6.24f}$$

$$HigherPhyloValue(x,y) \Rightarrow PreyOn(x,y), 1.13. \tag{6.24g}$$

$$HigherPhyloValue(x,y) \Rightarrow HighCooccurrence(x,y), -0.59. \tag{6.24h}$$

$$PreyOn(x,x) \Rightarrow PreyOn(x,y), 0.45. \tag{6.24i}$$

$$\neg IsSalix(x) \vee \neg PreyOn(x,y), 0.02. \tag{6.24j}$$

$$CloselyRelated(x,y) \wedge PreyOn(x,z) \Rightarrow PreyOn(y,z), 0.00. \tag{6.24k}$$

Markov logic adds rules for all 18 lone predicates, we do not show them here. None of the rules or weights are particularly interesting. Formula 6.24a tells us that if species x interacts with species y at some location, then the two species interact. This rule should have an infinite weight as it is always true (and part of how the predicates were defined), but Markov logic is prudent about infinite weights since violating a formula with an infinite weight yields a probability of 0. Formulas 6.24b and 6.24e tell us that species that interact tend to co-occur (which *does* make it easier for interaction).

The high weight of 6.24g shows that having a higher value in the phylogenetic clustering does somewhat predict interaction. Formula 6.24k says that if x is closely related to y and x preys on z, then we can predict that y will also prey on z. In short, closely related species should have similar preys. This formula has a weight of 0, meaning it is neither true nor false.

Perhaps the most interesting rule is formula 6.24i, which was added by revision and makes little sense. It says that if x is a cannibal ($PreyOn(x, x)$) then x preys on y. However, this formula showcase an issue with the idea of defining a probability distribution over formulas and the difficulty of knowledge engineering in this context (Jain, 2011). Our data-set has no cases of cannibalism, however, remember that $A \Rightarrow B$ is true if A is false. So while this formula makes little sense, it is true for all pairs of species x, y where x is not a cannibal. Pure logic algorithms use measures such as information gain to judge if a formula should be added (Quinlan, 1990). Information gain has been well studied in the context of learning decision trees (Quinlan, 1986) and here, since this formula is only true because the left-side of the formula is always false, the information gain is 0. However, it is easy for such formulas to creep in because of the approximate nature of optimizing the weighted pseudo-log-likelihood.

6.9 Discussion

An interesting twist may be that, as much as probabilistic thinking is important to ecology, defining a probabilistic distribution over formulas may not be. To understand why, let's consider the niche model (Williams et Martinez, 2000). The first iteration of the niche model posits that all species are described by a niche position N (their body size for instance) in the $[0,1]$ interval, a diet D in the $[0,N]$ interval, and a range R such that a species preys on all species with a niche in the $[D-R/2, D+R/2]$ interval. We can represent these ideas with two formulas :

$$\forall x : D(x) < N(x), \tag{6.25a}$$

$$\forall x,y : PreyOn(x,y) \Leftrightarrow D(x) - R(x)/2 < N(y) \wedge N(y) < D(x) + R(x)/2, \tag{6.25b}$$

A probabilistic version of the theory (Williams et al., 2010) yields a probability that x preys on y using a Gaussian distribution centered on the predator's diet :

$$\forall x,y : PPreyOn(x,y) = \alpha \times \exp\left[-\left(\frac{N(y)-D(x)}{R(x)/2}\right)^2\right], \tag{6.26}$$

This is a probabilistic model, yet it is quite problematic in Markov logic. Since predicates in Markov logic have to resolve to $true$ or $false$, this formula would have to be exactly true (whatever it means for a probabilistic model !). The same problem arise for deterministic equations like the Lotka-Volterra : these formulas are seldom exactly correct, and we shoud not expect them to be. By allowing predicates to take any value in the $[0,1]$ range with fuzzy logic, we gain enough flexibility to handle probabilistic theories at the level of the predicate, but also ambiguity in deterministic theories. It is important to emphasis where fuzzy logic and Markov logic differ since both intend to relax the rigid nature of pure logic. In fuzzy logic, ambuigity lies in the predicate themselves, which can take any value in the $[0,1]$ range. Also, evaluated formulas resolve to a truth value in the $[0,1]$ range. As an example, imagine a formula that says that species undergoe exponential growth $N(x,t+1) = G(x) \times N(x,t)$ only when the population is small ($SmallN$) and has resources :

$$SmallN(x) \wedge Resources(x) \Rightarrow N(x,t+1) = G(x) \times N(x,t). \tag{6.27}$$

In fuzzy logic, we can say $SmallN$ is mostly true (say 0.8), the species has a lot of resources

(0.9), and exponential growth predicts 95 rabbits when in reality there are 100 (truth value of 0.95). In fuzzy logic with the product T-norm, the left-side of the formula would resolve to $0.8 \times 0.9 = 0.72$ and then, since implication $x \Rightarrow y$ is defined as $1 - x + xy$ with the product T-norm, we would get a truth value of 0.964. In Markov logic, the formula would have a weight to determine the probability of the formula to be true, but we would need to set arbitrary cutoffs for when $SmallN$, $Ressource$, and exponential growth are true.

In Markov logic, we define a probability distribution over formulas, allowing those that are violated to be "less probable". However, both predicates and formulas resolve to $true, false$. Not only the flexibility lies in different places but it has different consequences for theory revision. Formulas in fuzzy logic are not weighted, they have the exact same form as standard logic rules, except we allow the predicates to take truth values in the $[0, 1]$ range. Adding a formula in Markov logic triggers weight optimization for the entire knowledge base. Adding formulas in fuzzy logic does not, and the new formulas could be evaluating using measures such as information gain (which can then be used in Bayesian approach to theory revision (Muggleton, 1994)).

It seems clear that we need fuzzy predicates to encode key nuances found in both ecological data and theories. Whether we need probabilistic interpretation of the formulas themselves is an open question, but it does seem difficult to scale such representation to a large number of formulas, whereas there are no equivalent issues with fuzzy logic. Fuzzy logic can also be extended to type-2 fuzzy logic to handle uncertainty (e.g. type-2 interval logic, see table 6.4). In Markov logic, data is in the form of true/false predicates. With fuzzy logic, we allow any values in the $[0, 1]$ range. With type-2 fuzzy logic, additional uncertainty in the predicate's truth value can be modelled, which would make it easier to integrate probabilistic theories with uncertainty.

Thanks to its ability to represent relations and mathematical models, Markov logic has the potential to offer a more mechanistic understanding of the data than simple probabilistic distributions. Yet, the results here are not really impressive. To be fair, the data is larger (more than 50k ground predicates) and more complex than most examples used in papers for theory revision, and it is also possible that no revisions were found because the initial model is already good. However, the big issue is that evaluating a single candidate took hours, often tens of hours for formulas with more than two predicates (the ones we're interested in). While better algorithms may improve the quality of the candidates, the core issue is that evaluating a

Table 6.4 – How evaluation works in type-2 interval fuzzy logic.

Name	Evaluation of Type-2 intervals (a,b)
Negation	$\neg(a,b) = (1-b, 1-a)$
Conjunction	$(a,b) \wedge (a',b') = (a \wedge a', b \wedge b')$
Disjunction	$(a,b) \vee (a',b') = (a \vee a', b \vee b')$

The interval approach is the simplest : the truth value is an interval from a to b ($a \leq b$ and $a, b \in [0,1]$). Type-2 interval fuzzy logic reduces to type-1 fuzzy logic (or just : fuzzy logic) when $a = b$. Type-2 interval fuzzy logic can be used to model uncertainty in the truth value. It could be useful to introduce uncertainty in fuzzy predicates based on probabilistic models. This approach is increasingly popular in engineering to model uncertainty in fuzzy rule-based systems (Mendel, 2017).

candidate takes a long time with no guarantees that the algorithm actually optimize the weights correctly. In contrast, a simple Python script was able to evaluate one million candidates (and successfully find revisions) using fuzzy logic on a different data-set (Desjardins-Proulx et al., 2017a). If Markov logic struggles to evaluate candidates for a single data-set, it is unlikely to realize the ambition of a unifying knowledge base for ecology. The core issue is that weights have to be computed for each new candidate.

There are two main issues with Markov logic. The first is bivalence. Standard predicate logic is bivalent : a predicate is either true or false. Markov logic allows probabilistic queries on predicates and formulas, but the predicates themselves are restricted to $true, false$. There are many ways in which this is limiting, and we will extend on this later on, but suffice to say that it forced us to reduce rich real-valued variables such as temperature to binary values. There is also evidence that real-valued traits, such as body size and phylogenetic distances, are particularly important to predict interactions (Desjardins-Proulx et al., 2017b).

The second problem is computational. Learning-assisted theorem proving is now possible with millions of lemmas (Kaliszyk et Urban, 2015). Here, it took hours to evaluate a single candidate for a problem with two domains (species, location) and 18 predicates. Even if the optimization algorithm could be made a hundred time faster to learn the weights of the formula, it is difficult to see how this could be used for a general ecological knowledge base spanning several subfields. The key issue is that, by defining a probability distribution over all formulas, we are forced to re-evaluate all the weights for every new formulas. It is also important to emphasize that the second-order optimization algorithm used (L-BFGS (Nocedal et Wright, 2006)) should converge quickly when the weights are near the optimum, and the weights *are* near the optimum when we try to evaluate a new candidate.

In this chapter, we described relational approaches based on predicate logic, with an emphasis on Markov logic. The test with Markov logic highlighted important issues with this framework but gave us insights into how theory revision could work in the context of ecology. Markov logic is an interesting extension of first-order logic, but our results were not impressive. For one, the only new rule we were able to learn made no sense, and while the weights do provide some interesting information, much simpler methods could have yielded the same insights (e.g. simple counting how often the rules are true). The process also took weeks to evaluate a few hundred candidate rules, whereas the fuzzy learning algorithm presented in the previous chapter can evaluate millions of rules in minutes. While these approached are not strictly comparable, it is difficult to image how Markov logic could scale to large databases of ecological knowledge.

More fundamentally, Markov logic fails to provide the type of flexibility we need to integrate ecological theories. Our previous results show that theory revision with fuzzy logic is both fast and effective (Desjardins-Proulx et al., 2017a). That particular fuzzy logic method was not relational, it used a more primitive form of fuzzy logic incapable of representing mathematical ideas. However, it shows that fuzzy logic has the key ingredients for effective theory revision and it is trivial to extend this approach to predicate fuzzy logic. Ultimately, a scalable knowledge base capable of handling mathematical theories would require a relational representation (or something similar). Further work should focus on type-1 or type-2 fuzzy predicate logic. There is no known algorithms to revise such theories, but techniques from pure-logic theory revision may be adapted (Muggleton et de Raedt, 1994).

6.10 References

Bach, S., Broecheler, M., Huang, B., Getoor, L. (2015). Hinge-loss markov random fields and probabilistic soft logic *arXiv : 1505.04406.*

Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press.

Bell, G. (2010). Fluctuating selection : the perpetual renewal of adaptation in variable environments. Phil. Trans. R. Soc. B *365*, 87–97.

Bohan, D., Caron-Lormier, G., Muggleton, S., Raybould, A., Tamaddoni-Nezhad, A. (2011). Automated discovery of food webs from ecological data using logic-based machine learning. PLoS ONE *6*, e29028.

Desjardins-Proulx, P., Bartomeus, I., Poisot, T., Gravel, D. (2017a). Combining ecological theories with machine learning using fuzzy logic.

Desjardins-Proulx, P., Laigle, I., Poisot, T., Gravel, D. (2017b). Ecological interactions and the netflix problem. PeerJ *5*, e3644.

Domingos, P., Lowd, D. (2009). Markov Logic : An Interface Layer for Artificial Intelligence. Morgan & Claypool Publishers.

Friedman, N., Nachman, I., Pe'er, D. (1999). Learning bayesian network structure from massive datasets : The sparse candidate algorithm. In : Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence.

Hahn, M. (2008). Toward a selection theory of molecular evolution. Evolution *76*, 255–265.

Halevy, A., Norvig, P., Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems *24*, 8–12.

Harrison, J. (2009). Handbook of Practical Logic and Automated Reasoning. Cambridge University Press.

Hájek, P. (1998). Metamathematics of Fuzzy Logic. Springer Netherlands.

Jaeger, M. (1997). Relational bayesian networks. In : Uncertainty in Artificial Intelligence.

Jain, D. (2011). Knowledge Engineering with Markov Logic Networks : A Review. In : DKB 2011 : Proceedings of the Third Workshop on Dynamics of Knowledge and Belief.

Kaliszyk, C., Urban, J. (2015). Learning-assisted theorem proving with millions of lemmas. Journal of Symbolic Computation *69*, 109–128.

Kok, S., Domingos, P. (2005). Learning the structure of markov logic networks. In : Proceedings of the 22nd international conference on Machine learning. pp. 441–448.

Kok, S., Domingos, P. (2009). Learning markov logic network structure via hypergraph lifting. In : Proceedings of the 26nd international conference on Machine learning.

Kok, S., Domingos, P. (2010). Learning markov logic networks using structural motifs. In : Proceedings of the 27nd international conference on Machine learning.

Koller, D., Friedman, N. (2009). Probabilistic Graphical Models. The MIT Press.

Kopelke, J., Nyman, T., Cazelles, K., Gravel, D., Vissault, S., Roslin, T. (2017). Food-web structure of willow-galling sawflies and their natural enemies across europe. Ecology *98*, 1730.

Mendel, J. (2017). Uncertain Rule-Based Fuzzy Systems : Introduction and New Directions, 2nd edition. Vol. 32. Springer.

Mihalkova, L., Mooney, R. (2007). Bottom-up learning of markov logic network structure. In : Proceedings of the 24nd International Conference on Machine Learning.

Muggleton, S. (1994). Bayesian inductive logic programming. In : Proceedings of the seventh annual conference on Computational learning theory. pp. 3–11.

Muggleton, S., de Raedt, L. (1994). Inductive logic programming : Theory and methods. The Journal of Logic Programming *19-20*, 629–679.

Murphy, K. (2012). Machine Learning : A Probabilistic Perspective. The MIT Press.

Nocedal, J., Wright, S. (2006). Numerical Optimization 2nd edition. Springer.

Quinlan, J. (1986). Induction of decision trees. Machine Learning *1*, 81–106.

Quinlan, J. (1990). Learning logical definitions from data. Machine Learning *5*, 239–266.

Richardson, M., Domingos, P. (2006). Markov logic networks. Machine Learning *62*, 107–136.

Russell, S., Norvig, P. (2009). Artificial Intelligence : A Modern Approach, 3rd Edition. Prentice Hall.

Tamaddoni-Nezhad, A., Milani, G., Raybould, A., Muggleton, S., Bohan, D. (2013). Construction and validation of food webs using logic-based machine learning and text mining. Advances in Ecological Research *49*, 225–289.

Williams, R., Anandanadesan, A., Martinez, N. (2010). The probabilistic niche model reveals the niche structure and role of body size in a complex food web. PLOS One *5*, e12092.

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.

Yuan, C., Malone, B. (2013). Learning optimal bayesian networks : A shortest path perspective. Journal of Artificial Intelligence Research *48*, 23–65.

CHAPITRE 7

DISCUSSION ET CONCLUSION GÉNÉRALE

7.1 Introduction

Cette thèse a exploré diverses techniques pour prédire les interactions des espèces, en mettant l'emphase sur l'apprentissage de règles claires et l'objectif à long terme d'assembler une base de connaissances unifiées pour les théories écologiques. Cette discussion fera, dans un premier temps, un bref retour sur les résultats principaux avant de revoir les deux questions posées dans l'introduction (section 1.3). Je terminerai par une discussion sur la suite des choses, comment progresser vers une base de connaissance unifiée en tenant compte des leçons apprises lors de ces travaux.

7.2 Résultats en bref

7.2.1 Méthodes classiques

Plusieurs techniques classiques supervisées et non supervisées ont été explorées dans les chapitres 3 et 4. Deux conclusions s'imposent. Primo, l'importance de traits aux valeurs continues ($\in \mathbb{R}$), en particulier la taille et la taxonomie. À partir de ces trois traits, il fut possible d'entraîner un *random forest* capable de prédire les interactions entre deux espèces avec grand succès. Le TSS de ce modèle était de 0.94. Ajouter les 25 traits binaires n'augmente le TSS qu'à 0.96.

La deuxième leçon des méthodes classiques est que les réseaux d'interactions écologiques ont un imbrication (*nestedness*) élévée. L'imbrication mesure à quel point des prédateurs tendent à partager les mêmes proies. Cet imbrication permet au *k-nearest neighbors* (KNN) de prédire avec succès les interactions d'une espèce en se basant seulement sur les proies de prédateurs similaires. Contrairement aux *random forests*, le KNN est une méthode simple à comprendre, programmer, et interpréter. Le succès du KNN est directement lié à l'imbrication, et donc son efficacité a une interprétation écologique simple.

7.2.2 La clarté par le flou

Le chapitre 5 explore un système simple que j'ai développé, basé sur la logique floue. Les règles ont la forme : **Si** X_0 est Y_0 et X_1 est Y_1 $\ldots$ **alors** (interaction ou non-interaction) X représente un trait, par exemple la taille du pollinisateur, et Y une variable linguistique (small, average, high, ...). J'ai développé un algorithme pour apprendre ces règles à partir de données. Contrairement aux méthodes classiques, ce modèle génère des règles simples, formulées en langue courante et donc faciles à lire et interpréter.

Le résultat montre que cette méthode est non seulement plus facile à interpréter mais donne de meilleures prédictions que les méthodes établies comment les *random forests*. En plus de leur capacité à prédire les interactions entre les pollinisateurs et les plantes, des discussions avec le Dr. Ignasi Bartomeus, le chercheur à l'origine de ces données, démontrent que les règles découvertes par mon approche ont du sens sur plan écologique. Malgré que le Dr. Bartomeus n'ait pas de formation en apprentissage automatique ou en informatique, il n'eut aucune difficulté à comprendre les règles. C'est une grande force de cette approche par logique floue : le résultat est clair.

Cette recherche confirme le potentiel de l'apprentissage de la logique floue en écologie et confirme aussi que les interactions entre pollinisateurs et plantes peuvent être prédites avec précision à partir de peu de traits reliés à la taille des pollinisateurs et à la forme des plantes.

7.2.3 Logique de Markov

Le chapitre 6 introduit la logique de Markov, une approche qui n'avait pas été explorée en écologie. La force de la logique de Markov est d'unifier la théorie des probabilités avec une forme de logique, la logique de premier ordre, capable d'intégrer des connaissances mathématiques avancées. Cette capacité est importante si l'on désire construire des bases de connaissances qui relient diverses théories écologiques, car plusieurs théories en écologie sont formulées avec des équations mathématiques (le modèle de Lotka-Volttera, la théorie de niche, etc). Aussi, l'intégration de modèles mathématiques permet d'automatiquement relier les connaissances écologiques. Par exemple, dans une base de connaissances de premier ordre, une formule mathématique qui estime une variable écologique sera directement liée à toutes les formules avec cette variable. Un exemple serait la théorie métabolique en écologie, qui prédit plusieurs variables à partir de formules de la forme $\alpha \times \beta^{\gamma}$ (West et al., 1997).

Pour cette étude nous avons établi un modèle de base, défini par un certain nombre de formules en logique de premier ordre, et nous avons testé un algorithme de révision sur ces formules avec les données de parasites-insectes-saules. Malheureusement, l'algorithmne n'a pas réussi à trouver des règles de bonne qualité. La cause semble être double. Dans un premier temps, il n'existe qu'une seule implémentation complète de la logique de Markov, elle ne roule pas en parallèle, et prend plusieurs douzaines d'heures pour évaluer des candidats. Dans un deuxième temps, la logique de Markov ne semble pas assez flexible. Le succès de la logique floue tient au fait que cette logique tolère des nuances, alors que la logique de Markov force les données à être bivalentes (vrai ou faux). Plusieurs des prédicats auraient bénéficié de la flexibilité de la logique floue (Figure 7.1). Ce résultat confirme la leçon apprise lors des chapitres précédents qui affirme l'importance de traits aux valeurs réelles ($\in \mathbb{R}$) comme la taille ou la promixité taxonomique.

Figure 7.1 – Prédicats flous et logique de Markov

Predicate	Domain	Interpretation
$PreyOn(x,y)$ 🔴	$species \times species$	Whethere x preys on y in some locations
$PreyOnAt(x,y,l)$ 🟠	$species \times species \times location$	Whether x preys on y in l.
$IsParasite(x)$ 🟢	$species$	Whether the species is a parasite
$IsPlant(x)$ 🟢	$species$	Whether the species is a plant
$IsHerbivore(x)$ 🟢	$species$	Whether the species is a herbivore
$CloselyRelated(x,y)$ 🔴	$species \times species$	Whether the species are in the same family
$Covariation(x,y)$ 🔴	$species \times species$	Presence/Absence correlation > 0?
$HigherPhyloValue(x,y)$ 🔴	$species \times species$	x has higher phylo value than y †
$HighTemperature(l)$ 🔴	$location$	Temperature higher than X
$HighPrecipitation(l)$ 🔴	$location$	Precipitation higher than X
$HighSalixRichness(l)$ 🔴	$location$	Salix richness higher than X
$HighGallsRichness(l)$ 🔴	$location$	Galls richness higher than X
$HighParasitoidRichess(l)$ 🔴	$location$	Parasitoid richness higher than X
$HighLinkDensity(l)$ 🔴	$location$	Local good web link density higher than X
$HighConnectance(l)$ 🔴	$location$	Local good web connectance higher than X

Liste des prédicats utilisés pour la logique de Markov au chapitre 6. La logique de Markov force les prédicats à être bivalents (vrai ou faux), mais plusieurs prédicats auraient bénéficié de plus de flexibilité. En vert, les prédicats qui fonctionnent bien en tant que prédicats bivalents. Une espèce est soit un parasite ou elle ne l'est pas (en tout cas pour ces données il n'y avait pas de nuance nécessaire). En orange, le prédit *PreyOnAt* fonctionne relativement bien en tant que prédicat bivalent mais aurait pu être flou, par exemple en ayant comme valeur de vérité l'intensité des interactions à cette location. En rouge, les prédicats qui auraient dû être flous. Forcé le prédicat sur la proximité taxonomique à être bivalent enlève beaucoup de nuances importantes aux données.

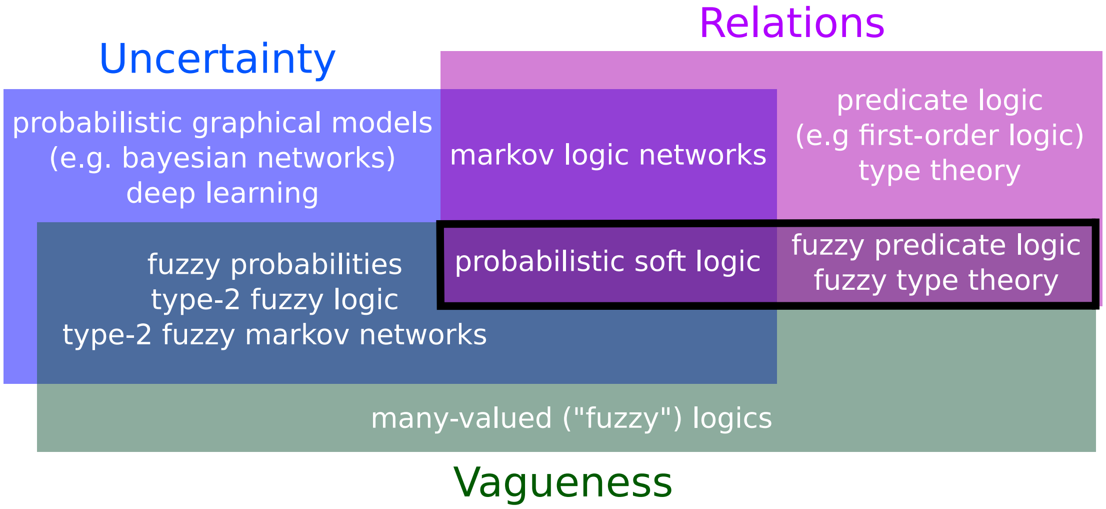
7.3 Conclusion : retour sur les deux questions

Le nombre de représentations de connaissances est grand et malheureusement, peu de ces représentations ont attiré l'intérêt des écologistes. Ma thèse montre l'importance de ces approches, et applique deux méthodes nouvelles (logique floue et logique de Markov) au problème de l'interaction des espèces. Lors de l'introduction, j'ai posé deux questions : quelle est la représentation de la connaissance la mieux adaptée à la synthèse des théories écologiques, et comment exploiter automatiquement ces représentations ? La seconde question est, au final, assez simple à répondre. Il existe plusieurs algorithmes simples qui permettent de raisonner à partir d'une base de connaissances. De loin, la première question est la plus difficile : quelle est la bonne représentation pour la synthèse d'idées écologiques ? La thèse apporte une réponse partielle : cette représentation doit être relationelle comme la logique de Markov pour permettre l'intégration d'équations mathématiques, et permettre les prédicats flous pour accomoder les nuances des données écologiques.

Sous sa forme actuelle, la logique floue explorée au chapitre 6 est imparfaite comme représentation pour l'écologie. Ces règles ne sont pas relationnelles comme la logique du premier ordre (first-order logic). Par conséquent, elles ne peuvent exprimer des idées mathématiques complexes, et donc elles ne peuvent réaliser l'ambition d'une base de connaissances unifiées. Des recherches futures devraient par contre prendre la logique floue comme point de départ. L'écologie a besoin des prédicats nuancés de la logique floue et il existe des extensions à la logique de premier ordre, bien qu'il n'y ait pas d'algorithmes pour apprendre ces règles. Il y a un clair intérêt en intelligence artificielle pour la logique floue relationelle et un besoin de flexibilité pour l'intégration de connaissances dans les domaines complexes (Kimmig et al., 2012 ; Bach et al., 2015 ; Garnelo et al., 2016 ; Hu et al., 2016).

Les représentations les plus prometteuses se trouvent donc à la frontière entre les méthodes relationnelles capables de représenter des formules mathématiques, et la logique floue, capable de capturer les nuances des données écologiques. La figure 7.2 met l'emphase sur ces approches. L'écologie est un domaine complexe, qui comprend énormément de théories allant de modèles sur la dynamique des populations à des modèles sur le flot d'énergie dans les écosystèmes. L'intégration de ces théories dans un tout cohérent devrait être une priorité pour les théoriciens, à la fois pour permettre un meilleur partage des connaisances et la cohérence des théories développées.

Figure 7.2 – Retour sur les représentations de la connaissance



La région encerclée met l'emphase sur les représentations capables de représenter les idées mathématiques (méthodes relationelles) et d'accommoder les nuances des prédicats écologiques (logique floue). Ces méthodes devraient être les prochaines à explorer pour la synthèse de théories écologiques.

ANNEXE A

HOW LIKELY IS SPECIATION IN NEUTRAL ECOLOGY ?

A.1 Abstract

Patterns of biodiversity predicted by the neutral theory rely on a simple phenomenological model of speciation. To further investigate the effect of speciation on neutral biodiversity, we analyze a spatially-explicit neutral model based on population genetics. We define the metacommunity as a system of populations exchanging migrants and we use this framework to introduce speciation with little or no gene flow (allopatric and parapatric speciation). We find that with realistic mutation rates, metacommunities driven only by neutral processes cannot support more than a few species. Adding natural selection in the population genetics of speciation increases the number of species in the metacommunities and generate patterns of species distribution similar to those predicted by Hubbell's neutral theory of biodiversity.

A.2 Introduction

How patterns of biodiversity arise through ecological and evolutionary processes is a central question in modern ecology (Johnson et Stinchcombe, 2007 ; Fussmann et al., 2007). According to Hubbell's neutral theory of biodiversity (NTB), patterns of biodiversity such as species-abundance distributions can be explainnated by the balance between speciation, dispersal and random extinction (Hubbell, 2001). The neutral theory provides a good fit to species distribution curves (Hubbell, 2001), and it has been extended in several ways (Haegeman et Etienne, 2009 ; Volkov et al., 2005 ; Rosindell et al., 2010). The neutral theory has been shown to be flexible enough to fit nearly any distribution (Chave et al., 2002), but it is often regarded as a valid starting point and an interesting null hypothesis for community ecology (Alonso et al., 2006).

While a lot has been said about the assumption of ecological equivalence (Abrams, 2001 ; Purves et Turnbull, 2010), much less attention has been given to the speciation mode (Etienne et al., 2007), which is sometime seen as the theory's weakest point (Kopp, 2010). In recent years, some studies improved the speciation model within neutral ecology (Etienne et al., 2007 ; Haegeman et Etienne, 2009 ; Rosindell et al., 2010). However, nothing has been done to relate the theory to population genetics and known models of speciation, despite the fact that, as Etienne et al. noted (Etienne et al., 2007), such a mechanistic model could eventually force us to reject neutrality. The neutral theory with point speciation has also been criticized for predicting too many rare species, too many young species (Ricklefs, 2003), and for assuming a direct relationship between abundance and speciation (Etienne et al., 2007).

In this article, we introduce a neutral theory of biodiversity with a speciation model derived from population genetics. We emphasize the role of allopatric and parapatric speciation. Speciation modes are most often distinguished according to the level of gene flow between the diverging populations. Allopatric speciation occurs when the new species originates from a geographically isolated population. By contrast, sympatric speciation is often defined as speciation without geographical isolation, in short, when the diverging populations share the same location. Lastly, parapatric speciation covers the middle ground between these two extremes (Gavrilets, 2003).

In the original neutral theory's formulation, Hubbell presented two models of speciation, point-speciation and random-fission speciation (Hubbell, 2001). Both are phenomenological individual based models. In the case of point-speciation, a newly recruited individual is selected at random and undergoes speciation. In the case of random-fission, the whole species is divided in two at random. The random-fission model is more realistic and does improve some predictions related to speciation and the number of rare species, but the resulting species abundance curves do not fit data as well as the point-speciation model (Etienne et Haegeman, 2011) In both cases, the probability of speciation of a given species is directly proportional to abundance and independent of dispersal. Hubbell associates the point-speciation model with sympatric speciation, and the random fission model with allopatric speciation (Hubbell, 2001). Some rare forms of sympatric speciation are indeed similar to the point speciation model, namely polyploid speciation, but most sympatric speciation events involve a population being divided in two by non-geographical factors (Coyne et Orr, 2004). Also, as neither models take gene flow into consideration, neither can distinguish sympatric and allopatric speciation events.

While theoretical models have shown sympatric speciation to be possible, empirical studies have uncovered only a very few solid cases (Bolnick et Fitzpatrick, 2007) and much of the theory is controversial (Spencer et Feldman, 2005 ; Barr et Wells, 2005 ; Gavrilets, 2004). Despite the growing acceptance of sympatric speciation as a plausible cause of speciation, most speciation events are still thought to occur with limited gene flow (Coyne et Orr, 2004 ; Gavrilets et Vose, 2005 ; Bolnick et Fitzpatrick, 2007 ; Fitzpatrick et al., 2008). Allopatric and parapatric speciation events are more common, but modelling them require some details about the spatial structure of the metacommunity. We chose to base our model on the most common forms of speciation despite the increased complexity of a spatially-explicit framework. We find that with realistic parameters, metacommunities cannot support more than a few species when the genetics of speciation is assumed to be neutral. We also considered a simple alternative pseudo-selection model by adding natural selection at the genetic level, but keeping the ecological equivalence assumption at the individual level. This approach shows that the rates of speciation typical of the NTB cannot be obtained without selection pushing mutations to fixation.

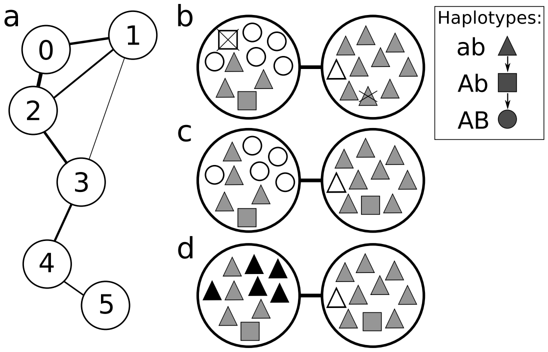


Figure A.1 – Allopatric speciation in a metacommunity. (a) The metacommunity as a graph of (local) communities. Each community is connected by dispersal to one or more communities. (b) A metacommunity with two communities connected by dispersal and 10 individuals per community ($n = 2$, $J = 10$). The metacommunity has two species (white and gray). Individuals are represented by different shapes to identify their haplotypes : triangles for ab, squares for Ab and circles for AB. (c) The two individuals selected to die in (b) are replaced, both by local replacements. In the community on the left, the empty spot is filled by a gray individual with ab. In the other community, the empty spot is filled by a gray individual who mutates from ab to Ab. (d) Because all white individuals carry AB in the left community, they speciate and are now identified by black triangles (ab).

A.3 Model

We model speciation with the Bateson-Dobzhansky-Muller model (BDM) in which reproductive isolation is the consequence of the accumulation of incompatible alleles (Bateson, 1909 ; Dobzhansky, 1937 ; Muller, 1942 ; Orr., 1996 ; Orr et Turelli., 2001). While the BDM model is simple, we have many empirical and theoretical reasons to think that speciation events often follow a similar scheme (Coyne et Orr, 2004 ; Gavrilets, 2003). We use a two-loci and two-alleles version of the model where sexual reproduction is ignored (Gavrilets, 2004, p. 131). Each local population starts with the ab haplotype fixed. The allele at the first locus, a, mutates to A, and the allele at the second locus, b, mutates to B. Both mutates at the same rate μ. We follow Gavrilets and ignore back mutations (Gavrilets, 2004). Back mutations have been shown to slow down speciation in this model, but not dramatically (Gavrilets, 2004, p. 131). Alleles a and B are incompatible, so the path from ab and AB can be seen as a process with three states :

$$ab \longrightarrow Ab \longrightarrow AB \tag{A.1}$$

Speciation occurs when all individuals in the local population carry the AB haplotype. Migration bring new individuals, always with the ab haplotype, at rate m. To integrate Gavrilet's BDM model in a metacommunity, we'll connect local communities composed of populations of one or more species. We assume the two loci are distinct for each population. Speciation is a complex process, but this simple model captures many important characteristics of speciation events that are ignored in the NTB. First, it that takes time. Most often, speciation is the result of a long process where a population diverge from the rest of the species to the point where a reproductive barrier stops current and future gene flow between the diverging populations (Coyne et Orr, 2004). Second, with a few exceptions, the starting population size of the new species is likely to have more than a single individual. Third, gene flow (migration) has a strong homogenizing effect that will inhibit speciation (Coyne et Orr, 2004 ; Gavrilets, 2004). Lastly, speciation occurs as a population of a given species diverge, most often in well-defined geographic areas (Avise, 2000 ; Coyne et Orr, 2004). None of these characteristics are present in the original neutral theory (Hubbell, 2001), although protracted speciation partially solve the first two problems by adding a parameter to account for the waiting time to speciation (Rosindell et al., 2010).

One of the difficulty with speciation in individual-based models is that it is often impossible to distinguish populations, but speciation is a population-process. Speciation in the consequence of divergences between populations. In the NTB and most of its variants, only two levels of organization are recongnized ; the individual and the species. Because these individuals are not grouped in populations, it is hard, if not impossible, to integrate speciation theory, which is deeply rooted in the idea of diverging populations (Coyne et Orr, 2004). To integrate Gavrilets' model in the NTB, we need a spatially-explicit model of the metacommunity that allows us to distinguish populations within species. Several approaches have been used to model the spatial structure of populations and local communities. Some are spatially-explicit at the level of the individual. In these models, the location of each individual is known, generally by using a grid (Rosindell et Cornell, 2007) or a graph (Lieberman et al., 2005). Another approach is to consider the position of populations, but ignore the exact position of the individuals within the populations. Again, this method has been used with grids (Gavrilets et Vose, 2005) and graphs (Minor et Urban, 2007 ; Economo et Keitt, 2008 ; Dale et Fortin, 2010). Because speciation is a population-process, we use the latter approach and model the metacommunity as a graph of n local communities (hereafter simply referred to as communities), where each community x can support a total of J_x individuals. Communities are connected together by dispersal (Leigh et al., 2004) and composed of one or more species (Economo et Keitt, 2008 ; 2010) (Fig. 1a). This spatial representation allows us to distinguish three levels of organization : species, populations (individuals of a given species in a given community), and individuals. Dispersal between two communities will always be low enough to assume that the individuals in the two communities can be defined as distinct populations (Berryman, 2002).

Each individual has a haplotype (either ab, Ab, AB) and we follow explicitly their dynamics in each community. As these haplotypes represent a path toward speciation in their community, it should be seen as a different pair of loci for each population. For example, if an individual of species i migrates from community x to community y, it will carry the ab haplotype in its new community, regardless of its haplotype in community x. In short, we assume the variation acquired in a given community has no effect on the mutations toward speciation in other communities. This assumption is not realistic in all situations, as both mutational-order and ecological speciation are know to be influenced by complex interactions between the diverging populations (Mani et Clarke, 1990 ; Schluter, 2009 ; Nosil et Flaxman, 2011 ; Coyne et Orr, 2004 ; Gavrilets, 2004). Integrating the effect of these divergences would require many more assumptions about the nature of speciation, and in most cases cannot be done without introducing the concept of niche (Schluter, 2000). We ignore much of the details of speciation in favor of a simple model that captures many of the most fundamental characteristics of speciation as a population-process. Because there is no niche differentiation, new mutations toward speciation are always allowed to appear regardless of the ecological context. As soon as all the individuals of a given species inside a local population carry the haplotype AB, they undergo speciation (Fig. 1d).

Metacommunity dynamics follow Hubbell's neutral model of biodiversity (Hubbell, 2001), for each unit of time t a disturbance randomly kill D individuals from the metacommunity (we considered $D = 2\%$). After the disturbance the metacommunity is refilled using post-disturbance abundance and a migration matrix $\mathbf{m}$, according to the following equation (Gravel et al., 2006) :

$$p(i,x) = \sum_{y} m(y,x) \frac{N(i,y)}{J_p(y)} \tag{A.2}$$

Where $p(i,x)$ is the probability of picking an individual of species i to fill an empty spot on community x. $m(y,x)$ is the probability that an individual will be picked from community y to migrate to community x, if $y \neq x$ it is a migration event, and if $y = x$ it is a local replacement event. $N(i, y)$ is the number of individual of species i in community y, i.e. : the population size. $J_p(y)$ is the post-disturbance size of community y ($J_p(y) = J_y - D_y$). This equation always ignore the haplotype of the individual, which is decided after a species has been picked to refill the empty spot. The rules for deciding the haplotype of the individual depends on the model (neutral or pseudo-selection).

We considered three different metacommunity shapes ; star, circle and complete. In the circular metacommunity, all communities are linked to two other communities to form a circle. In the star, a single central communities forms a link to all outer communities, which have no other links. In the complete metacommunity, all communities are linked to all the others. The migration matrix is built with a single parameter ω, which is the strength of the links between communities. To establish the strength of migration between communities we draw a number in the uniform distribution between 0 and ω (mean : $\omega/2$) for each link. After this number is drawn, the migration rate between two linked communities is found by dividing the number by the sum of all links plus one (for local replacement events). The probability that an individual killed in community x will be replaced by migration is, on average :

$$m_x = \frac{c\omega/2}{1 + c\omega/2} \approx c\omega/2 \tag{A.3}$$

With c being the number of communities linked to x. The 1 in the denominator stands for the weight given to local replacement events. Because the rate of migration is much smaller than 1, the rate of migration is approximately ω times the number of linked communities. The average migration rate for communities in circle metacommunities ($c = 2$) is roughly ω. For stars, it is $(n - 1)\omega$ for the central community and $\omega/2$ for the others. Lastly, for a complete metacommunity the rate of migration is $(n - 1)\omega$ for all communities. In this model, the strength of migration depends on the weight of the links between communities ω and the number of linked communities.

Within this framework we explore two models. In the neutral model, new mutations will be driven to extinction or fixation by drift and migration alone. We follow explicitly the number of individuals with each haplotype in each local community (Fig. 1). When a species is picked by equation A.2 for migration, it will carry the ab genotype. When a species is picked for a local replacement event the haplotype is selected according to the relative abundance of the haplotypes, with a probability μ of mutation fronm ab to Ab and from Ab to AB. Because even weak selection can dramatically reduce the time to speciation (Gavrilets, 2004), we also considered a pseudo-selection model with within-population selection. For this variant, we use a multiplicative fitness regime (Charlesworth et Charlesworth, 2011, p. 166), so the fitness of an individual with haplotype ab, Ab or AB is, respectively, 1, $1+s$ or $(1+s)^2$. One of the basic tenets of the NTB is ecological equivalence, so to keep the pseudo-selection model within the framework of neutral ecology, the probability to pick an individual from one species will still follow equation A.2 regardless of its internal genetic composition. The haplotypes are only considered after a species has been chosen. Then, the haplotype is chosen according to its abundance and fitness. In reality, if a population has many individuals with haplotypes Ab and AB, it should have an advantage over a population with only ab individuals, but this will break the ecological equivalence assumption of neutral biodiversity so we ignore it.

We explored the model by simulations using an implementation in ANSI C99. For all simulations, the number of local communities n is set to 10, and each community starts filled with a single species so there are n species at the beginning. We compared simulations with $J = 10^3$ to 10^6 individuals and found similar results. We thus use $J = 10^5$ for all simulations. The mutation rate μ for eukaryotes is generally between 10^{-4} and 10^{-6} (Drake et al., 1998 ; Gavrilets, 2004) and we set μ to the highest realistic value, $\mu = 10^{-4}$. The simulations ran for 2×10^7 time steps, and we recorded the average local and regional species richness over the last 10^5 steps.

A.4 Results

We found that for our neutral metacommunity model with realistic parameter values, regardless of its size, shape, and dispersal rate, the regional species richness never exceeds n, the number of local communities. Increasing the migration rate reduces the number of species supported by the metacommunity. For all values of ω, the number of species supported at equilibrium is equal or below 10, the initial number of species. Unsurprisingly, we find that reducing the average migration rate increases the speciation rate, but it also increases the number of extinctions. For $\omega \leq 10^{-5}$, the regional species richness remain at its original value of 10, while for

$\omega \geq 10^{-3}$, the entire metacommunity supports only a single species. We find a threshold migration rate around $\omega = 10^{-4}$ where the regional species richness increases suddenly. Around this value, the number of species varies between 1 and 10. When $\omega \leq 10^{-4}$, the communities are so isolated that they are dominated by a single species, often with a small number of individuals from one or two other species.

We studied the effect of increasing the mutation rate beyond realistic values. Keeping J at 10^5 and $\omega = 10^{-3}$, we ran simulations for several mutation rates. Even a tenfold increase in the mutation rate ($\mu = 10^{-3}$) has little effect on the equilibrium regional species richness. The metacommunity sustains higher diversity around a mutation rate of $\mu = 10^{-2}$. This mutation rate is well above the typical mutation rate (Drake et al., 1998; Gavrilets, 2004). This finding lends credit to the theory that the NTB requires unrealistically high speciation rates (Ricklefs, 2003).

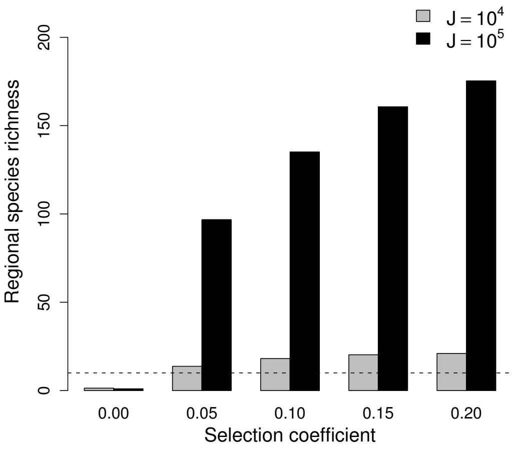


Figure A.2 – The number of species at equilibrium in the pseudo-selection model increases non-linearly with the selection rate. The number of species quickly increases, in part because selection pushes the alleles toward speciation, but also because it reduces the fitness of migrants. Selection diminishes the inhibiting effect of gene flow on speciation.

Adding selection has an important impact on the equilibrium species richness. We find approximately 100 species at the regional scale (10/community) with $s = 0.05$ (Fig. 2). In both metacommunities shaped like stars and circles, the regional species abundance distribution is lognormal-like with negative skewness (a long left tail) (Fig. 3). This regional distribution is similar to the NTB's distribution for local communities (Hubbell, 2001). There are fewer rare species than the regional distribution seen in the NTB, supporting the criticism of Ricklefs (Ricklefs, 2003), which argued that the NTB predicted too many rare species. In all simulations with $\omega = 10^{-3}$, the metacommunities with the complete shape could only support one species for the entire metacommunity.

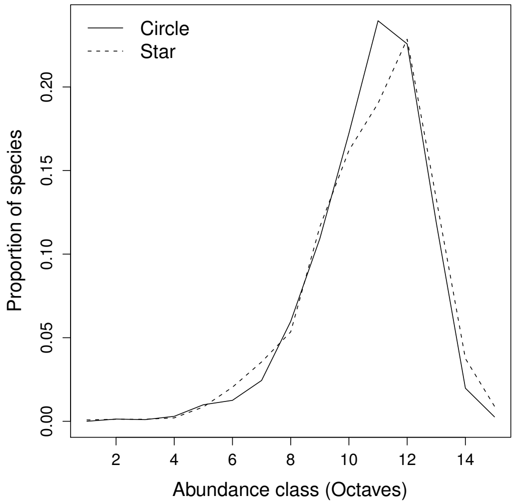


Figure A.3 – Regional species distribution in the pseudo-selection model. Both star and circle communities share a similar distribution. Simulations with the complete metacommunities (all communities linked together) ended with a single species because of the crushing effect of high migration rates.

A.5 Discussion

In this study, we developed a simple framework to study speciation as a population-process within neutral ecology. In our models, the speciation rate is not assumed to take any particular value, it is an emergent property of the system. It depends on selection, the mutation rate, and migration, which in turn is influenced by the shape of the metacommunity. Also, we made no assumptions about the relationship between abundance and speciation rate. Species with more individuals are likely to occupy more communities, so they will have more opportunities to speciate, but the relationship will depend on the shape of the metacommunity and the spatial distribution of the populations. Our goal was to examine the relationship between neutral ecology and a mechanistic model of speciation based on population genetics. Phenomenological models are not inherently inferior (McGill et Nekola, 2010) but they should be confronted to their mechanistic counterpart to determine if they can provide a good approximation of reality, under what conditions this approximation can hold, and what kind of assumptions are required to make it hold. Our assumptions deliberately made speciation easy to achieve. We used the BDM model with only two steps required to reach speciation and we ignored back mutations (Gavrilets, 2004). The mutation rate chosen was the highest plausible rate for eukaryotes (Drake et al., 1998 ; Kumar et Subramanian, 2002). There was always a mutation toward speciation available, arguably the most unrealistic assumption as the conditions for speciation are seldom common (Coyne et Orr, 2004). All these assumptions greatly favor speciation, yet the model still failed to produce metacommunities with many species unless selection is added or the mutation rate is set to impossible levels.

Adding selection to the NTB is not compatible with some key assumptions of neutral ecology, or at least would require new assumptions to be made to explainnat how natural selection can act at the genetic level but not at the ecological level. Speciation can be achieved easily if mutations toward speciation are given some positive selection coefficient. But new species, being the result of the accumulation of fitness-enhancing mutations, should have greater fitness, which would violate the NTB's ecological equivalence assumption. Zhou and Zhang's nearly neutral model showed that even small differences between species lead to markedly different species distributions (Zhou et Zhang, 2008). There is little doubt that selection plays an important role in speciation events (Coyne et Orr, 2004) and few neutral models of speciation have been developed (Nei et al., 1983). Speciation by drift alone is simply too slow (Turelli et al., 2001). A possible solution to the dilemma would be to assume a constant flow of positive mutations. As fitness would increase for all species, the effect of mutations toward speciation would be lost, preserving ecological equivalence. This solution assumes a "Red Queen" scenario where individuals have to run to keep in the same place (Van Valen, 1973), and is similar to the view of evolution defended by Ronald Fisher (Fisher, 1930). However, this assumption requires many additional assumptions about adaptation and the rate of evolution.

When comparing models, one aspect to consider is their complexity. Theoretical populations genetics is mostly based on mathematical models that are simple enough to be analytically tractable, which has lead to a tendency to ignore spatial complexity (Epperson et al., 2010). As allopatric and paratric speciation events rely on this spatial complexity, we have few theoretical models to study the effect of these forms of speciation on diversity. We chose to base our theory on the most common forms of speciation and introduced a simple method to model allopatric and parapatric speciation in complex spatial structures. While using graphs add a layer of complexity to neutral ecology, our approach fixes some of the problems of the point-speciation model without adding new parameters for speciation, as we replace the speciation rate v with the mutation rate μ. More importantly, this approach allows us to divide a species in populations, a fundamental unit in evolution. Ricklefs (Ricklefs, 2003) argued that new species under the point-speciation model would not be recognized as species, because those species have appeared instantaneously and are likely too similar. Rosindell et al. (Rosindell et al., 2010) improved the speciation model with protracted speciation. In this model, a new parameter is added to account for the waiting time to speciation. While this approach solves the problem of instantaneous speciation, it requires two free parameters for speciation and still offers no explanation for the speciation rate. One of the problems with a fixed speciation rate v is that speciation is directly influenced by ecological factors such as isolation and habitats. In particular, the effect of dispersal on speciation is ignored in most community models with speciation (Hubbell, 2001 ; Etienne et al., 2007 ; Haegeman et Etienne, 2009 ; Volkov et al., 2005 ; Rosindell et al., 2010), despite the fact that gene flow greatly reduces the probability of speciation (Coyne et Orr, 2004).

A.6 References

Abrams, P. A. (2001). A world without competition. Nature *412*, 858–859.

Alonso, D., Etienne, R. S., McKane, A. J. (2006). The merits of neutral theory. Trends in Ecology and Evolution *21*, 451–457.

Avise, J. C. (2000). Phylogeography. Harvard University Press.

Barr, M., Wells, C. (2005). Toposes, triples, and theories. Theory and Applications of Categories *1*, 1–289.

Bateson, W. (1909). Heredity and variations in modern lights. In : Seward, A. C. (Ed.), Darwin and Modern Science. Cambridge University Press, pp. 85–101.

Berryman, A. A. (2002). Population : a central concept for ecology ? Oikos *97*, 439–442.

Bolnick, D. I., Fitzpatrick, B. M. (2007). Sympatric speciation : models and empirical evidence. Annual Review of Ecology, Evolution, and Systematics *38*, 459–487.

Charlesworth, B., Charlesworth, D. (2011). Elements of evolutionary genetics. Roberts & Company Publishers.

Chave, J., Muller-Landau, H. C., Levine, S. A. (2002). Comparing classical community models : theoretical consequences for patterns of diversity. The American Naturalist *159*, 1–23.

Coyne, J. A., Orr, H. A. (2004). Speciation. Sinauer Associates.

Dale, M. R. T., Fortin, M.-J. (2010). From graphs to spatial graphs. Annual Review of Ecology, Evolution, and Systematics *41*, 21–38.

Dobzhansky, T. G. (1937). Genetics and the origin of species. Columbia University Press.

Drake, J. W., Charlesworth, B., Charlesworth, D., Crow, J. F. (1998). Rates of spontaneous mutation. Genetics *148*, 1667–1686.

Economo, E. P., Keitt, T. H. (2008). Species diversity in neutral metacommunities : a network approach. Ecology Letters *11*, 52–62.

Economo, E. P., Keitt, T. H. (2010). Network isolation and local diversity in neutral metacommunities. Oikos *10*, 1–9.

Epperson, B. K., McRae, B. H., Scribner, K., Cushman, S. A., Rosenberg, M. S., Fortin, M. J., James, P. M. A., Murphy, M., Manel, S., Legendre, P., Dale, M. R. T. (2010). Utility of computer simulations in landscape genetics. Molecular Ecology *19*, 3549–3564.

Etienne, R. S., Apol, M. E. F., Olf, H., Weissing, F. J. (2007). Modes of speciation and the neutral theory of biodiversity. Oikos *116*, 241–258.

Etienne, R. S., Haegeman, B. (2011). The neutral theory of biodiversity with random fission speciation. Theoretical Ecology *4*, 87–109.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Clarendon.

Fitzpatrick, B. M., Fordyce, J. A., Gavrilets, S. (2008). What, if anything, is sympatric speciation ? Journal of Evolutionary Biology *21*, 1452–1459.

Fussmann, F. G., Loreau, M., Abrams, P. A. (2007). Eco-evolutionary dynamics of communities and ecosystems. Functional Ecology *21*, 465–477.

Gavrilets, S. (2003). Perspective : models of speciation : what have we learned in 40 years ? Evolution *57*, 2197–2215.

Gavrilets, S. (2004). Fitness Landscapes and the Origin of Species. Princeton University Press.

Gavrilets, S., Vose, A. (2005). Dynamic patterns of adaptive radiation. Proceedings of the National Academy of Sciences of the United States of America *102*, 18040–18045.

Gravel, D., Canham, C. D., Beaudet, M., Messier, C. (2006). Reconciling niche and neutrality : The continuum hypothesis. Ecology Letters *9*, 399–409.

Haegeman, B., Etienne, R. S. (2009). Neutral models with generalised speciation. Bulletin of Mathematical Biology *71*, 1507–1519.

Hubbell, S. P. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Vol. 32 of Monographs in Population Biology. Princeton University Press.

Johnson, M. T. J., Stinchcombe, J. R. (2007). An emerging synthesis between community ecology and evolutionary biology. Trends in Ecology and Evolution *22*, 250–257.

Kopp, M. (2010). Speciation and the neutral theory of biodiversity : modes of speciation affect patterns of biodiversity in neutral communities. Bioessays *32*, 564–570.

Kumar, S., Subramanian, S. (2002). Mutation rates in mammalian genomes. Proceedings of the National Academy of Sciences of the United States of America *99*, 803–808.

Leigh, E. G., Davidar, P., Dick, C. W., Puyravaud, J.-P., Terborgh, J., Wright, S. J. (2004). Why do some tropical forests have so many species of trees ? Biotropica *36*, 447–473.

Lieberman, E., Hauert, C., Nowak, M. A. (2005). Evolutionary dynamics on graphs. Nature *433*, 312–316.

Mani, G. S., Clarke, B. (1990). Mutational order : a major stochastic process in evolution. Proc. R. Soc. London B *240*, 29–37.

McGill, B. J., Nekola, J. C. (2010). Mechanisms in macroecology : AWOL or purloined letter ? towards a pragmatic view of mechanism. Oikos *119*, 591–603.

Minor, E. S., Urban, D. L. (2007). Graph theory as a proxy for spatially explicit population models in conservation planning. Ecological Applications *17*, 1771–1782.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. Biological Symposia *6*, 71–125.

Nei, M., Maruyama, T., Wu, C.-I. (1983). Models of evolution of reproductive isolation. Genetics *105*, 557–579.

Nosil, P., Flaxman, S. M. (2011). Conditions for mutation-order speciation. Proc. R. Soc. London B *278*, 399–407.

Orr., H. A. (1996). Dobzhansky, Bateson and the genetics of speciation. Genetics *144*, 1331–1335.

Orr, H. A., Turelli., M. (2001). The evolution of post-zygotic isolation : accumulating Dobzhansky-Muller incompatibilities. Evolution *55*, 1085–1094.

Purves, D. W., Turnbull, L. A. (2010). Different but equal : the implausible assumption at the heart of neutral theory. Journal of Animal Ecology *79*, 1215–1225.

Ricklefs, R. E. (2003). A comment on hubbell's zero-sum ecological drift model. Oikos *100*, 185–192.

Rosindell, J. L., Cornell, S. J. (2007). Species-area relationships from a spatially explicit neutral model in an infinite landscape. Ecology Letters *10*, 586–595.

Rosindell, J. L., Cornell, S. J., Hubbell, S. P., Etienne, R. S. (2010). Protracted speciation revitalizes the neutral theory of biodiversity. Ecology Letters *13*, 716–727.

Schluter, D. (2000). The ecology of adaptive radiation. Oxford University Press.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. Science *323*, 737–741.

Spencer, H. G., Feldman, M. W. (2005). Adaptive dynamics, game theory and evolutionary population genetics. Journal of Evolutionary Biology *18*, 1191–1193.

Turelli, M., Barton, N. H., Coyne, J. A. (2001). Theory and speciation. Trends in Ecology and Evolution *16*, 330–343.

Van Valen, L. (1973). A new evolutionary law. Evol. Theory *1*, 1–30.

Volkov, I., Banavar, J. R., He, F., Hubbell, S. P., Maritan, A. (2005). Density dependence explains tree species abundance and diversity in tropical forests. Nature *438*, 658–661.

Zhou, S.-R., Zhang, D.-Y. (2008). A nearly neutral model of biodiversity. Ecology *89*, 248–258.

ANNEXE B

A COMPLEX SPECIATION-RICHNESS RELATIONSHIP IN A SIMPLE NEUTRAL MODEL

B.1 Abstract

Speciation is the “elephant in the room” of community ecology. As the ultimate source of biodiversity, its integration in ecology’s theoretical corpus is necessary to understand community assembly. Yet, speciation is often completely ignored or stripped of its spatial dimension. Recent approaches based on network theory have allowed ecologists to effectively model complex landscapes. In this study, we use this framework to model allopatric and parapatric speciation in networks of communities and focus on the relationship between speciation, richness, and the spatial structure of communities. We find a strong opposition between speciation and local richness, with speciation being more common in isolated communities and local richness being higher in more connected communities. Unlike previous models, we also find a transition to a positive relationship between speciation and local richness when dispersal is low and the number of communities is small. Also, we use several measures of centrality to characterize the effect of network structure on diversity. The degree, the simplest measure of centrality, is found to be the best predictor of local richness and speciation, although it loses some of its predictive power as connectivity grows. Our framework shows how a simple neutral model can be combined with network theory to reveal complex relationships between speciation, richness, and the spatial organization of populations.

B.2 Introduction

For a long time speciation was not part of community ecology's theoretical framework. MacArthur and Wilson's seminal work on island biogeography does mention speciation but their model and most of its inheritors ignored it completely (MacArthur et Wilson, 1967). This is surprising given speciation's central role : ultimately, all species appear through speciation events. The importance of speciation to understand patterns of diversity was noted by Wallace in the 1850s (Wallace, 1855) and played a key role in the modern synthesis of evolutionary biology (Mayr, 1942). Fortunately, ecologists are increasingly aware of the importance of speciation. Recently Vellend argued that, while a great number of processes shape communities, they can be grouped in four classes : drift, dispersal, selection, and speciation (Vellend et Orrock, 2009 ; Vellend, 2010). Recent theoretical models, such as those based on Hubbell's neutral theory (Hubbell, 2001) or the Webworld (Caldarelli et al., 1998), have also made speciation an important part of community ecology (Drossel et al., 2001 ; Etienne et al., 2007 ; Kopp, 2010 ; Rosindell et al., 2010 ; Melián et al., 2010 ; Rosindell et Phillimore, 2011 ; Etienne et Haegeman, 2011). In particular, the neutral theory covers three of the four classes of processes described by Vellend, leaving only selection untouched (Hubbell, 2001 ; Vellend, 2010) in favor of a more phenomenological (some would say pragmatic) description of community dynamics (Wennekes et al., 2012). However, whereas drift and dispersal are well integrated in the neutral theory, the treatment of speciation remains dubious (Etienne et al., 2007 ; Desjardins-Proulx et Gravel, 2012).

In community ecology, speciation is often reduced to a mutation leading instantaneously to a new species with a single individual. We refer to this modeling approach as "speciation-as-a-mutation". This is the approach of both the neural theory (Hubbell, 2001) and Webworld (Caldarelli et al., 1998). Because the processes determining the fate of mutations (gene flow, selection, drift) have no effect on the mutation rate, population geneticists can simply obtain an estimate from field data and plug the mutation rate in equations (Crow et Kimura, 1970 ; Vellend, 2003). Speciation is different. The speciation rate is an emergent property of selection, drift, and dispersal processes (Coyne et Orr, 2004). To put it another way : a mutation is a molecular phenomenon unaffected by allele dynamics so it can be treated independently. Speciation on the other hand is a population process influenced by the structure and dynamics of populations, so it cannot be treated as a fixed rate. In particular, as gene flow tends to inhibit divergence, the spatial organization of a species' populations and their level of isolation will determine the likelihood of speciation (Coyne et Orr, 2004; Rice, 2009). In this study we use networks of communities to move the neutral theory from “speciation-as-a-mutation” to “speciation-as-a-population-process” (Fig. 1).

Spatial patterns of diversity are notoriously hard to study theoretically. Part of the problem lies in the lack of effective analytical methods for nontrivial spatial models (Epperson et al., 2010). Network theory provides tools to study patterns of connections and allows us to model almost any kind of spatial structure (Dale et Fortin, 2010). Furthermore, theorists have developed algorithms to analyze various aspects of networks, making it an effective tool to extract information from highly complex structures (Newman, 2010). A network is simply defined by two sets : a set of vertices and a set of edges. In our case, the vertices represent local communities and the edges denote dispersal (Fig. B.1). The entire network forms the metacommunity. A spatial network combines the combinatorial properties of a network with a topological space in any number of dimensions (Kobayashi, 1994). They have been used, among other things, to study networks on maps (Sedgewick, 2002; Penrose, 2003; Dale et Fortin, 2010) and the three-dimensional structure of molecules (Shinjo et Taniyama, 2003).

Economo and Keitt pioneered the use of networks in theoretical community ecology (Economo et Keitt, 2008; 2010). They extended Hubbell’s neutral theory to networks and studied how different topologies influence diversity. However, they kept the “speciation-as-a-mutation” model and did not use the network to account for the influence of isolation and gene flow on speciation (Economo et Keitt, 2008; 2010). To introduce speciation in a realistic matter, we have to go beyond the “speciation-as-a-mutation” framework and treat it as a population-process inhibited by gene flow (B.1). Speciation modes are most often distinguished by their biogeography (Coyne et Orr, 2004). Allopatric speciation occurs when the new species originates from a geographically isolated population, sympatric speciation is often defined as speciation without geographical isolation, and finally, parapatric speciation covers the middle ground between these two extremes (Coyne et Orr, 2004). The relative importance of gene flow to speciation is still the subject of a hot debate (Nosil, 2008; Johannesson, 2010). Nonetheless, speciation with little or no gene flow is still thought to be more common (Coyne et Orr, 2004; Fitzpatrick et al., 2008; Bolnick et Fitzpatrick, 2007). To study the effect of speciation on diversity, we extended the framework of Economo and Keitt to a population-based speciation model (Desjardins-Proulx et Gravel, 2012). We show that treating speciation as a population-process inhibited by gene flow has profound consequences on the predicted patterns of diversity. We discover a complex relationship between speciation and local richness. When the number of local communities is small, we find a strong positive relationship : communities with more speciation events are also the ones with the highest richness. However, this relationship does not hold as the number of communities increases. Finally, we compare the effectiveness of different centrality measures as predictors of local richness and speciation.

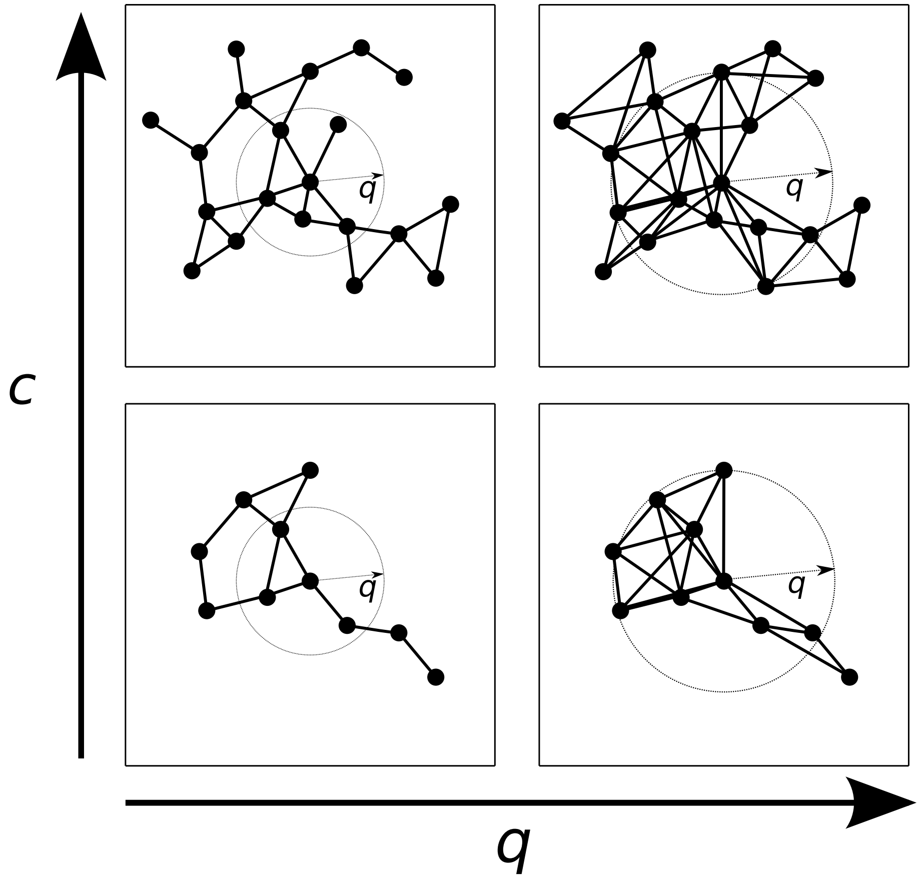


Figure B.1 – Four metacommunities represented by random geometric networks in the unit square ($x, y \in [0, 1]$). We define the metacommunity as the entire network of communities. Here the local communities (the vertices) are represented by black circles and the thick black lines (the edges) denote links by dispersal. Each of the c vertices has a position in two-dimensional space and is linked to all vertices within some threshold Euclidean distance q, which can be seen as the dispersal range of the species. A community is a set of populations of different species. We define a population as the entire set of individuals of a given species in a given vertex. As q increases, the number of links grow larger and local communities are less isolated. Similarly, the number of links per community also increases with c. While these networks are random, they exhibit locality, an important feature of real landscapes. Networks are well suited to distinguish populations within a species and thus to model speciation as a population process. Dispersal rates between connected communities are always low so we can assume the individuals of a given species in a given local community is a population in the strict sense (Berryman, 2002). Within each local community, the populations of one or more species fluctuate by drift and dispersal in the exact same way as Hubbell's neutral model (Hubbell, 2001). Unlike the models by Hubbell (Hubbell, 2001) and Economo (Economo et Keitt, 2008), we model speciation as a population process (Desjardins-Proulx et Gravel, 2012). Populations diverge through mutations and within-species selection and if a population accumulates enough divergence it undergoes speciation. In this model, each local community offers a possibility of speciation, which is inhibited by the homogenizing effect of gene flow. The rate of speciation is determined by the number of populations in the metacommunity and how much inhibiting gene flow is present. Thus, speciation is an emergent property of the metacommunity.

B.3 Methods

B.3.1 Metacommunity dynamics

Metacommunity dynamics is similar to Hubbell's neutral model of biodiversity (Hubbell, 2001). It can be described in three steps (Desjardins-Proulx et Gravel, 2012). 1 : For each time step an individual is selected and killed in each community, with all individuals having the same probability of being selected. 2 : The individuals selected in step 1 are replaced either by dispersal or by local replacement. The probability of dispersal from vertex x to vertex y is given by the dispersal matrix $\mathbf{m}$. In the case of dispersal from x to y ($x \neq y$), the new individual belong to species i with probability N_{ix}/J_x, with N_{ix} being the population size of species i in community x and J_x the size of the local community. Each individual belongs to a species and carry a genotype, either $a_x b_x$, $A_x b_x$, or $A_x B_x$, with x being the community. We assume that migrants carry no mutations at the focal loci for the population into which they move so the haplotype is ignored and the new individual will always carry $a_y b_y$. This assumption implies that each population has its own unique path to speciation. For local replacement events, the new individual will belong to species i with probability N_{ix}/J_x. However, the fitness of the haplotypes is used to determine the new individual's haplotype. One of the basic tenets of the neutral theory is ecological equivalence, so to introduce selection within the framework of neutral ecology the probability to pick an individual from one species has to ignore the internal genetic composition. After the species is selected, we select the haplotype using the fitness 1.0, $1+s$, $(1+s)^2$ for the three haplotypes respectively. When the haplotype is selected, $a_x b_x$ mutates to $A_x b_x$ and $A_x b_x$ to $A_x B_x$ with probability μ. 3 : Lastly, all populations with $A_x B_x$ fixed undergo speciation. The individuals of the new species will carry $a_x b_x$ and a new path toward speciation is open. A detailed description of the model can be found in Desjardins-Proulx and Gravel (Desjardins-Proulx et Gravel, 2012).

We study metacommunities with a varying number of vertices (local communities) c and threshold values q. We generate the random geometric networks by randomly placing the vertices in the unit square and connecting all vertices within some Euclidean distance q (Fig 1) (Sedgewick, 2002 ; Penrose, 2003). We generate random geometric networks until a connected one is found. This method might introduce a bias as many networks will be rejected when q is small but it is necessary because the presence of disconnected components makes the analysis of networks notoriously difficult (Newman, 2010). The number of communities c vary from 5 to 125 with a fixed metacommunity size of 100000 individuals (i.e. : we have 5 communities of 20 000 individuals, or 10 of 10 000, or 25 of 4 000, and so on). In a previous study we found that global diversity was optimal around $s = 0.15$ and $\omega = 5e - 4$ and we use these values unless otherwise noted (Desjardins-Proulx et Gravel, 2012). ω is a parameter used to create the dispersal matrix and can roughly be defined as the dispersal rate between two vertices (Desjardins-Proulx et Gravel, 2012). We also tried different values of s to test the solidly of our results ($s = \{0.05, 0.10, 0.20, 0.30, 0.40\}$). We set the mutation rate μ to $1e - 4$, a high but realistic value for eukaryotes (Drake et al., 1998 ; Gavrilets, 2004). See Desjardins-Proulx and Gravel for details on the effect of selection, ω, and the mutation rate on diversity (Desjardins-Proulx et Gravel, 2012). All simulations started with 20 species evenly distributed in the metacommunity and ran for 100 000 generations. The simulations were written in ANSI C99 and the code is available on github (https ://github.com/PhDP/origin).

B.3.2 Centrality measures

We explore the effect of five measures of centrality and importance on diversity (Newman, 2010). The first is the degree of the vertex, which is the number of edges starting from the vertex plus the number of edges going into the vertex divided by two. The second measure is eigen-centrality. It assigns scores to vertices so that connections to high-scoring nodes are more important than connections to low-scoring vertices. Closeness centrality is the average geodesic distance between the vertex and all other vertices. Unlike the degree, which is only affected by the neighbors, closeness centrality depends on the global structure of the network. Betweenness centrality is the number of shortest paths from all vertices to all others that pass through the vertex. In short, if we compute the shortest paths for all pairs of vertices, how many times a vertex is present in these paths determine its betweenness centrality. Lastly, communicability centrality is the sum of closed walks of all lengths starting and ending at the vertex. We used the Python library NetworkX to compute these centrality measures and the clustering coefficients (Hagberg et al., 2008).

B.4 Results

B.4.1 Local patterns of diversity

We first analyzed diversity on a vertex-by-vertex basis to understand the effect of network structure on local diversity. For all vertices we counted the number of speciation events, local richness at the end of the simulation, local and global extinctions events, and various measures of centrality. We compared metacommunities with 100 000 individuals divided into c local communities and used a threshold value of q to generate the random geometric network (see Fig. B.1). Species richness and the number of speciation events show strong positive correlations when c is small and strong negative correlations when c is large (Fig. B.2). An interesting trade-off occurs : communities that are more connected (high c) can host more species, which means more opportunities for speciation. On the other hand, they host more species because of greater dispersal, and greater dispersal means greater inhibiting gene flow for speciation (Fig. B.3). When c is fairly large, the communities with a greater number of links support more species than isolated communities, but the effect of gene slow is so strong that speciation is very difficult. Furthermore, because so many individuals come from migration events, it becomes hard for local populations to diverge enough to speciate. This result is coherent with known patterns of diversity and speciation (Diamond, 1972 ; Gavrilets et Losos, 2009). However, when c is small the total number of links, even in well-connected communities, is also small. The inhibiting effect of gene flow is still present but communities with more connections and more species will still witness more speciation events (Fig.B.4).

Then we investigated the relationship between local community diversity and several properties of the spatial network. In particular, we studied how different measures of centrality could be used to predict local richness and speciation events. The degree of a vertex (local community) is the crudest measure of centrality. While it gives information about how many communities are linked to the community of interest, it says nothing about the influence of the network's structure. Still, for all metacommunity sizes and all values of q, the degree shows the highest correlation with species richness and speciation (Fig. B.5). The correlation between the degree and species richness is close to one in many cases and performs poorly only for very high c and q, where the number of species in the metacommunity becomes small. Closeness centrality is a measure of the average geodesic distance (the length of the shortest path in the network (Dijkstra, 1959)) between a vertex and all other vertices. Closeness centrality's ability to predict local diversity is thus related to the metacommunity size and global structure. Closeness centrality is a worse predictor than the degree and achieves better result for small c (Fig. B.1). Closeness centrality is affected by all vertices, even those that are very far and unlikely to have any impact on the vertex. This is why closeness centrality performs better for small metacommunities : all communities are close so the measure cannot be biased by distant vertices. Eigen-centrality is an interesting alternative to the degree and closeness centrality. Unlike the first, it is affected by more than just neighbors but unlike the second far off vertices will have little effect on it. Overall, eigen-centrality is very similar to the degree but perform a little worst on larger communities, again showing the disproportionate effect of neighbors and the small effect of the overall structure of the network (Fig. B.1). Unsurprisingly, betweenness centrality performs rather poorly compared to the other measures (Table B.1). Local extinction is always strongly correlated with diversity ($r > 0.80$) and global extinction is strongly correlated with speciation ($r > 0.80$). Patterns of local and global extinctions closely follow the patterns of local richness and speciation, with a strong positive relationship with small c and a transition to a negative relationship as c increases.

B.4.2 Global patterns of diversity

We investigated the effect of the global network features such as the number of local communities c and the radius q used to generate the network(which determines the average number links between communities). We find that global species richness is strongly affected by both c and q (Fig. B.5). For low values of c, diversity increases slightly with q. However, q strongly reduces diversity for communities with high c. Because the dynamics is neutral and local communities with $n = 125$ can only support 800 individuals, a population is much less likely to stay long enough in a local community to undergo speciation. For $c = 5$ the species richness is roughly unaffected by q, increasing from 158 species with $q = 0.25$ to 182 with $q = 0.75$. This is not surprising since the average number of links is about equal (3.2 for $q = 0.25$ and 4.2 for $q = 0.75$). The effect become more pronounced as c grows larger. For $c = 25$ the average number of links increases markedly from 5 ($q = 0.25$) to 20 ($q = 0.75$), and species richness decreases from 186 to 125. The effect of q becomes evident at $c = 125$, where the average number of links jump from 20 to 100, and diversity crashes from 100 to only 4 species. In other words, in a neutral model with allopatric/parapatric speciation, the dispersal range will have a strong inhibiting effect on speciation especially if the metacommunities are divided into a great number of communities can only support a few individuals. These results highlight the complex role played by connectivity on diversity when the effect of gene flow on speciation is considered. On one hand connectivity inhibits speciation by increasing gene flow. On the other hand it promotes diversity by increasing dispersal. We show that an increase in connectivity will increase diversity if the metacommunity is divided into a few large communities and inhibit diversity as the number of communities increases.

We then took a closer look at the relationship between dispersal (threshold radius q) and diversity when the number of communities c is held constant. q increases the number of links and thus dispersal but its impact on the number of links depends on the number of communities and the formation of clusters. We studied the effect of dispersal on diversity by comparing metacommunities with the same number of communities c and generated with the same threshold value q. Because the networks are randomly generated, the average number of links per community will vary even if c and q are fixed. The correlation between the number of links and global diversity varies greatly. For example, with $n = 5$ the correlation is very strong (r = 0.85), regardless of the value of q : more links equals higher diversity. On the other hand the correlation is almost nonexistent for $c = 125$. As the metacommunity grows larger, other structural characteristics of networks play a greater role on diversity than the total number of links. Lastly, we explored the effect of clustering on species richness. Clustering is a measure of the tendency of vertices to form groups. High clustering will provide more opportunities of dispersal but also decreases the number of isolated vertices. The correlation between the clustering coefficient and species richness decreases with both c and q. The correlation is strongly positive for $c = 5$ with $r > 0.75$ for all values of q. The correlation decreased sharply with q for $c = 10$: from 0.60 to 0.32 and -0.36. With $c = 25$ the correlation coefficient is -0.11, -0.30, -0.58 for $q = 0.25, 0.50, 0.75$, respectively. It shows that, as c and q grows, clustering will tend to inhibit speciation enough to have a negative effect on diversity.

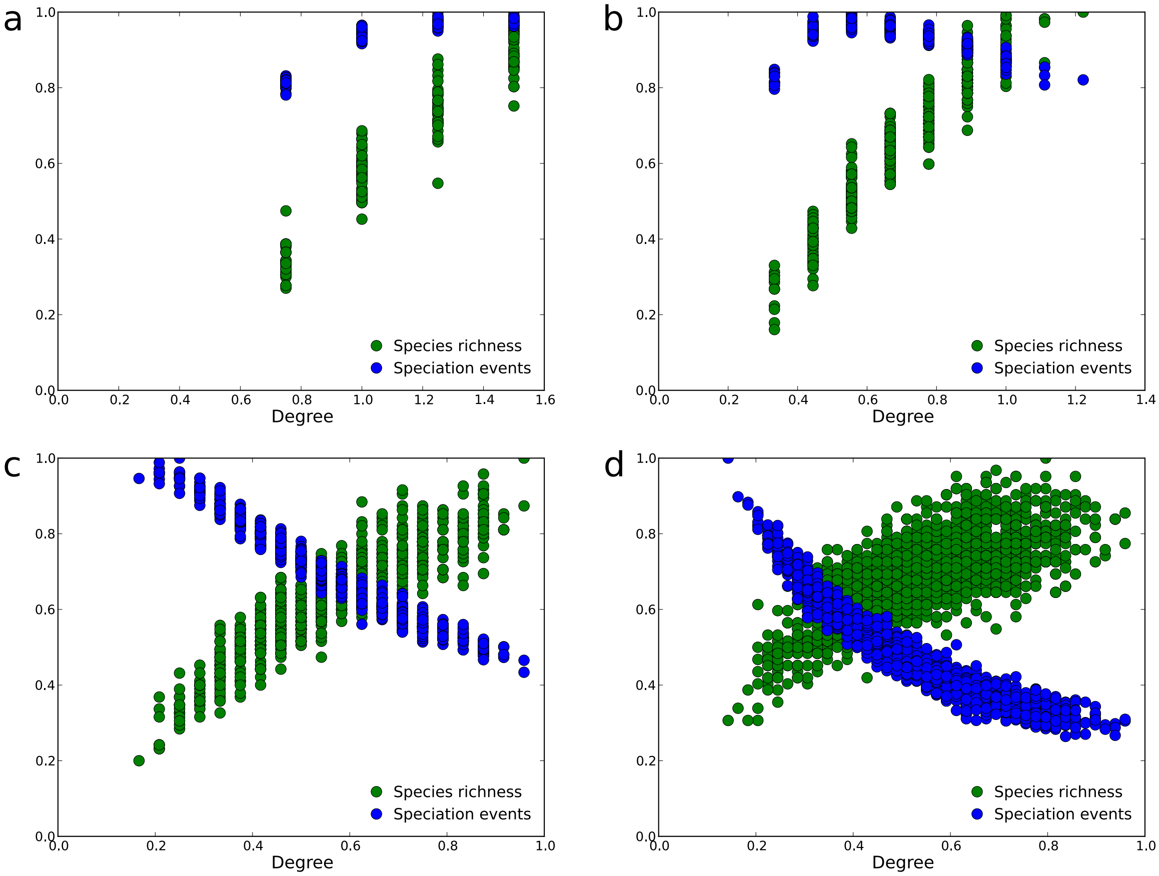


Figure B.2 – .

The relationship between degree centrality and local diversity/speciation events for local communities in metacommunities with 5 vertices (a), 10 vertices (b), 25 vertices (c) and 50 vertices (d). Species richness, the number of speciation events, and the vertex degree are all normalized. The points represent the vertices (local communities) at the end of the simulation and shows how much local richness and how many speciation events were found according to their degree centrality. The average degree is higher as the number of vertices in the metacommunity, denoted c, increases. In practice it means higher dispersal as c increases. When the number of vertices is small, speciation and richness are positively correlated. With $c = 10$ we observe the transition between a positive relationship to a negative one (b). It is the only plot where communities with an intermediate degree have the most speciation events. For larger metacommunities with higher dispersal (c & d), there is a clear negative relationship between speciation and local diversity. The models of Hubbell and Economo (Hubbell, 2001 ; Economo et Keitt, 2008) use a constant speciation rate per individual, so a similar plot would yield a flat line for speciation events (i.e. : it is unrelated to richness or dispersal). Selection was fixed at $s = 0.15$ and we used the threshold value $q = 0.50$ to generate the random networks. We ran 32 simulations for each figure, leading to $32 \times c$ data points per figure.

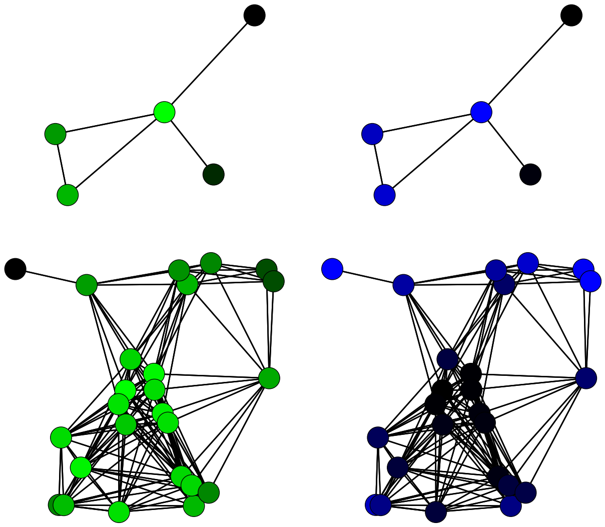


Figure B.3 – Local richness and diversity in networks of communities for two simulations. On top : a simulation with 5 communities and at the bottom : a simulation with 25 communities. On the left, communities change from black to green as local richness increases and on the right, communities change from black to blue as the number of speciation events increases. For the metacommunity composed of 5 communities, richness and speciation events are strongly correlated. The more connected communities support more species and more speciation events. With 25 communities the opposite is true. The communities with the most speciation events are far from the geodesic center, where local richness is higher.

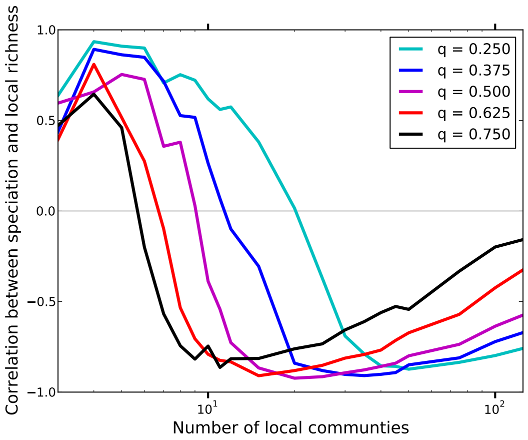


Figure B.4 – The relationship between speciation and diversity with an increasing number of local communities (total number of individuals in the metacommunity fixed at 100 000). Between $c = 3$ and $c = 8$, there is a positive relationship between them. As c approaches 10, the correlation drops and quickly becomes negative. It reaches a peak around $c = 25$, where the correlation starts to grow weaker. This relationship shows the various forces at play in our model and is a direct consequence of the "speciation-as-a-population-process" framework. Because the number of links per community is low with $c < 10$, communities with more species simply have more opportunities for speciation without being crushed by inhibiting gene flow. As c increases, a larger number of individuals are migrants. They not only inhibit speciation because of the gene flow but also take precious space. Populations fluctuate randomly and eventually become extinct. As the size of local communities grows smaller and the number of migrants grow, the chance for a population to stay in the community and accumulate diversity is severely reduced. These simulations used a within-species selection coefficient of $s = 0.15$ for the mutations leading to speciation. We tried different values $s = \{0.05, 0.10, 0.20, 0.30, 0.40\}$ and found similar patterns. With a purely neutral model ($s = 0.0$), the metacommunities could not support more than a few species (Desjardins-Proulx et Gravel, 2012).

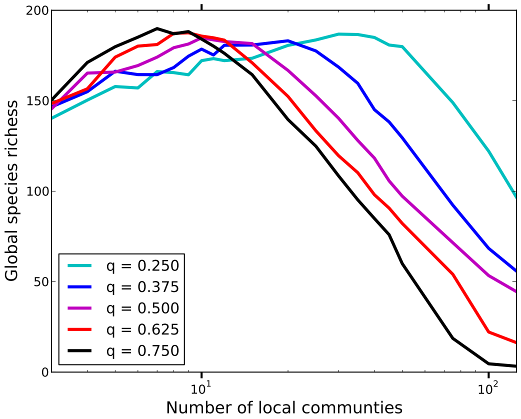


Figure B.5 – Relationship between global species richness and number of vertices c for various threshold distances q. Global diversity increases with higher connectivity and dispersal for 5 and 10 local communities. The effect of dispersal on global diversity changes with c. With 25 communities, richness at $q = 0.75$ is only 60% of the diversity with $q = 0.25$. With 50 communities, it is 30%, and with 125 communities : only 4%. Communities get smaller with increasing c and have more opportunities to form links. Thus, a much larger proportion of individual are migrants, making speciation very hard to achieve. Several metacommunities with $c = 125$ and $q = 0.75$ had a single species at the end of simulations.

Table B.1 – The relationship between different types of spatial centality and speciation/diversity rates.

With $q = 0.25$	5	10	25	50	125
Degree-Speciation	0.87	0.64	-0.49	-0.97	-0.94
Degree-Local diversity	0.97	0.95	0.92	0.88	0.78
Eigen-Speciation	0.87	0.53	-0.52	-0.76	-0.73
Eigen-Local diversity	0.97	0.84	0.69	0.68	0.63
Closeness-Speciation	0.81	0.53	-0.28	-0.73	-0.82
Closeness-Local diversity	0.95	0.85	0.57	0.73	0.71
Comm.-Speciation	0.81	0.59	-0.56	-0.80	-0.65
Comm.-Local diversity	0.92	0.90	0.82	0.69	0.57
Between.-Speciation	0.68	0.38	-0.22	-0.40	-0.57
Between.-Local diversity	0.82	0.70	0.38	0.41	0.53

With $q = 0.50$	5	10	25	50	125
Degree-Speciation	0.85	-0.56	-0.98	-0.95	-0.95
Degree-Local diversity	0.96	0.96	0.90	0.82	0.60
Eigen-Speciation	0.88	-0.50	-0.93	-0.93	-0.94
Eigen-Local diversity	0.95	0.92	0.86	0.79	0.59
Closeness-Speciation	0.79	-0.56	-0.95	-0.93	-0.94
Closeness-Local diversity	0.93	0.93	0.88	0.81	0.59
Comm.-Speciation	0.86	-0.56	-0.91	-0.88	-0.90
Comm.-Local diversity	0.95	0.92	0.83	0.76	0.56
Between.-Speciation	0.54	-0.57	-0.67	-0.74	-0.79
Between.-Local diversity	0.74	0.68	0.62	0.66	0.49

With $q = 0.75$	5	10	25	50	125
Degree-Speciation	0.51	-0.93	-0.95	-0.92	-0.85
Degree-Local diversity	0.94	0.90	0.77	0.59	0.11
Eigen-Speciation	0.56	-0.90	-0.95	-0.92	-0.86
Eigen-Local diversity	0.93	0.89	0.76	0.55	0.11
Closeness-Speciation	0.47	-0.94	-0.93	-0.87	-0.82
Closeness-Local diversity	0.94	0.89	0.75	0.54	0.11
Comm.-Speciation	0.44	-0.91	-0.94	-0.90	-0.84
Comm.-Local diversity	0.75	0.89	0.75	0.54	0.10
Between.-Speciation	0.29	-0.81	-0.80	-0.81	-0.73
Between.-Local diversity	0.84	0.70	0.64	0.48	0.10

Correlations between centrality measures and patterns of local richness. Degree-centrality outperforms closeness centrality for all combinations of q and c, degree-centrality outperforms eigen-centrality in 28 out of 30 cases, and eigen-centrality beats closeness centrality in about the same number of cases. Betweenness centrality has the worst performance. The results for $c = 125$ and $q = 0.75$ might seem puzzling but is actually simple to explainnat : at this point the metacommunity supports only a few species, often only one, because of small local community size and the crushing effect of gene flow. In short, the measures of centrality are bad at predicting diversity and speciation because there is very little of either.

B.5 Discussion

Community ecology is about drift, dispersal, speciation and selection (Vellend, 2010). Selection is a central component of community dynamics (Gravel et al., 2006) but if we want to understand its impact on spatial patterns of biodiversity we first need a solid reference template (Rosindell et al., 2011). Neutral models reveal the spatial structure of biodiversity expected from the combined effects of dispersal, drift, and speciation in the absence of selection. Despite its simplicity, our model generates clear predictions on the relationship between richness and speciation. It explainnats patterns of interest to both community ecologists and evolutionary biologists by integrating speciation as a population process. Our model predicts that isolation reduces diversity, a well-known pattern in biogeography (Diamond, 1972). It also predicts that isolation will stimulate speciation, a well-known pattern in speciation theory (Gavrilets et Losos, 2009). Rosindell and Phillimore used a very similar model and also found a negative relationship between speciation and local richness (Rosindell et Phillimore, 2011). The twist, in our model, is that this pattern is only true in some cases. When we model the metacommunity as a network with only a few communities, the most connected communities are both more diverse and witness more speciation events. In these cases, more diversity means more opportunities for speciation and the inhibiting effect of gene flow is not strong enough to counter the greater number of opportunities. Our models reveals a complex relationship between local richness, speciation, and isolation : one in which the spatial organization of communities and the strength of dispersal have the power to influence the speciation-richness relationship.

The neutral theory is often seen as weak when it comes to predictions related to speciation (Etienne et al., 2007 ; Etienne et Haegeman, 2011). In both Hubbell's model (Hubbell, 2001) and its spatially-explicit counterpart (Economo et Keitt, 2008 ; 2010), the rate of speciation is directly related to the number of individuals. Thus, it is not affected in one way or another by gene flow or the structure of the metacommunity. It is constant in a given community, regardless of the number of species present or the strength of dispersal. Our framework solves this problem by replacing the "speciation-as-a-mutation" approach with a speciation model based on populations (Desjardins-Proulx et Gravel, 2012). Our model predicts a negative relationship between diversity and speciation but it also predicts a positive relationship in some cases : namely when dispersal is weak. The exact relationship between diversity and speciation is complicated and not very well understood, especially from a mechanistic perspective. Emerson and Kolm did found a positive relationship between diversity and endemism, which could be interpreted as a rough index of speciation (Emerson et Kolm, 2005). They argue that diversity begets diversity : more species means more opportunities for other species to invade or speciate (Erwin, 2005). Our model does predict a positive relationship in some cases but for different reasons. Cadena et al. and Witt and Maliakal-Witt (Cadena et al., 2005 ; Witt et Maliakal-Witt, 2007) suggest that species richness and endemism are positively correlated because of a mutual dependence on life spans (Cadena et al., 2005). This is very similar to our model's prediction. When the metacommunity is divided into a few communities of large size, the gene flow is not strong enough to inhibit speciation and the greater local richness will offer more opportunities for speciation. The interesting twist in our framework is that this relationship will be reversed with greater gene flow and smaller communities. This prediction could be tested by comparing patterns of speciation and diversity in sets of islands of various sizes and connectivity.

The study's main limitation is arguably the neutral assumption. Our previous work suggests that neutral ecology is hard to reconcile with parapatric and allopatric speciation (Desjardins-Proulx et Gravel, 2012), requiring an uncomfortable compromise in the form of ecological equivalence at the species level and within-species selection. Some questions related to the relationship between speciation and richness will require the integration of adaptation and niches. For example, Emerson and Kolm's hypothesis that diversity stimulates speciation can only be tested in a trophic model (Gravel et al., 2011). Also, adaptive radiations and ecological speciation can easily define communities (Schluter, 2000 ; Gillespie, 2004) and, because of their explosive nature, they might be more sensitive to the overall structure of the metacommunity. Some network structures are known to inhibit selection while others will stimulate it (Nowak, 2006). In our model the degree is a better predictor of diversity and speciation than the more sophisticated measures of centrality. The degree does not take into account the overall structure of the metacommunity, it is only determined by the number of neighbors. The fact that it is a better predictor than eigen-centrality and closeness centrality, which are both influenced by the overall structure of the community, shows that the behavior of our model is mostly driven by small scale patterns. It remains to be seen if this result is an artifact of ecological equivalence or a feature of real communities.

Despite the limitations of neutrality, it provides a simple null hypothesis and could serve as a point of comparison with more realistic models of speciation in space. Theoretical community ecologists have mostly ignored the importance of space to model speciation. While recent developments have increased the visibility of speciation in community ecology, its treatment is often disconnected from speciation theory, making it difficult to unify ecology and evolution. Our approach to speciation is very simple but it captures many of the most important aspects of speciation : it is a population process often inhibited by gene flow and it cannot be treated as a simple mutation. We hope to have showed how network theory, a promising framework to study patterns of diversity in space (Dale et Fortin, 2010), can be used to integrate speciation in a more realistic matter and create a bridge between community ecology and speciation theory.

B.6 References

Berryman, A. A. (2002). Population : a central concept for ecology ? Oikos *97*, 439–442.

Bolnick, D. I., Fitzpatrick, B. M. (2007). Sympatric speciation : models and empirical evidence. Annual Review of Ecology, Evolution, and Systematics *38*, 459–487.

Cadena, C. D., Ricklefs, R. E., Jiménez, I., Bermingham, E. (2005). Is speciation driven by species diversity ? Nature *438*, E1–E2.

Caldarelli, G., Higgs, P. G., McKane, A. J. (1998). Modelling coevolution in multispecies communities. Journal of Theoretical Biology *193*, 345–358.

Coyne, J. A., Orr, H. A. (2004). Speciation. Sinauer Associates.

Crow, J. F., Kimura, M. (1970). Introduction to Population Genetics Theory. Harper & Row Publishers.

Dale, M. R. T., Fortin, M.-J. (2010). From graphs to spatial graphs. Annual Review of Ecology, Evolution, and Systematics *41*, 21–38.

Desjardins-Proulx, P., Gravel, D. (2012). How likely is speciation in neutral ecology ? The American Naturalist *179*, 137–144.

Diamond, J. M. (1972). Biogeographic kinetics - estimation of relaxation-times for avifaunas of southwest pacific islands. Proc. Natl. Acad. Sci. USA *69*, 3199–3203.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. Numer. Math. *1*, 269–271.

Drake, J. W., Charlesworth, B., Charlesworth, D., Crow, J. F. (1998). Rates of spontaneous mutation. Genetics *148*, 1667–1686.

Drossel, B., Higgs, P. G., McKane, A. J. (2001). The influence of predator-prey population dynamics on the long-term evolution of food web structure. Journal of Theoretical Biology *208*, 91–107.

Economo, E. P., Keitt, T. H. (2008). Species diversity in neutral metacommunities : a network approach. Ecology Letters *11*, 52–62.

Economo, E. P., Keitt, T. H. (2010). Network isolation and local diversity in neutral metacommunities. Oikos *10*, 1–9.

Emerson, B. C., Kolm, N. (2005). Species diversity can drive speciation. Nature *434*, 1015–1017.

Epperson, B. K., McRae, B. H., Scribner, K., Cushman, S. A., Rosenberg, M. S., Fortin, M. J., James, P. M. A., Murphy, M., Manel, S., Legendre, P., Dale, M. R. T. (2010). Utility of computer simulations in landscape genetics. Molecular Ecology *19*, 3549–3564.

Erwin, D. (2005). Seeds of diversity. Science *308*, 1752–1753.

Etienne, R. S., Apol, M. E. F., Olf, H., Weissing, F. J. (2007). Modes of speciation and the neutral theory of biodiversity. Oikos *116*, 241–258.

Etienne, R. S., Haegeman, B. (2011). The neutral theory of biodiversity with random fission speciation. Theoretical Ecology *4*, 87–109.

Fitzpatrick, B. M., Fordyce, J. A., Gavrilets, S. (2008). What, if anything, is sympatric speciation ? Journal of Evolutionary Biology *21*, 1452–1459.

Gavrilets, S. (2004). Fitness Landscapes and the Origin of Species. Princeton University Press.

Gavrilets, S., Losos, J. B. (2009). Adaptive radiation : contrasting theory with data. Science *323*, 732–737.

Gillespie, R. (2004). Community assembly through adaptive radiation in Hawaiian spiders. Science *303*, 356–359.

Gravel, D., Canham, C. D., Beaudet, M., Messier, C. (2006). Reconciling niche and neutrality : The continuum hypothesis. Ecology Letters *9*, 399–409.

Gravel, D., Massol, F., Canard, E., Mouillot, D., Mouquet, N. (2011). Trophic theory of island biogeography. Ecology Letters *14*, 1010–1016.

Hagberg, A., Schult, D., Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In : Varoquaux, G., Vaught, T., Millman, J. (Eds.), Proceedings of the 7th Python in Science Conference (SciPy2008). Pasadena, CA USA, pp. 11–15.

Hubbell, S. P. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Vol. 32 of Monographs in Population Biology. Princeton University Press.

Johannesson, K. (2010). Are we analyzing speciation without prejudice ? Ann. N.Y. Acad. Sci. *1206*, 143–149.

Kobayashi, K. (1994). On the spatial graph. Kodai Mathematical Journal *17*, 511–517.

Kopp, M. (2010). Speciation and the neutral theory of biodiversity : modes of speciation affect patterns of biodiversity in neutral communities. Bioessays *32*, 564–570.

MacArthur, R. H., Wilson, E. O. (1967). The theory of island biogeography. Princeton University Press.

Mayr, E. (1942). Systematics and the origin of species. Columbia University Press, New York.

Melián, C., Alonso, D., Vázquez, D. P., Regetz, J., Allesina, S. (2010). Frequency-dependent selection predicts patterns of radiations and biodiversity. PLOS Computational Biology *6*, e1000892.

Newman, M. (2010). Networks : An Introduction. Oxford University Press.

Nosil, P. (2008). Speciation with gene flow could be common. Molecular Ecology *17*, 2103–2106.

Nowak, M. A. (2006). Evolutionary Dynamics. Harvard University Press.

Penrose, M. (2003). Random Geometric Graphs. Oxford University Press.

Rice, S. H. (2009). A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution. BMC Evolutionary Biology *8*, 262.

Rosindell, J. L., Cornell, S. J., Hubbell, S. P., Etienne, R. S. (2010). Protracted speciation revitalizes the neutral theory of biodiversity. Ecology Letters *13*, 716–727.

Rosindell, J. L., Hubbell, S. P., Etienne, R. S. (2011). The unified neutral theory of biodiversity and biogeography at age ten. Trends in Ecology and Evolution *26*, 340–348.

Rosindell, J. L., Phillimore, A. B. (2011). A unified model of island biogeography sheds light on the zone of radiation. Ecology Letters *14*, 552–560.

Schluter, D. (2000). The ecology of adaptive radiation. Oxford University Press.

Sedgewick, R. (2002). Algorithms in C Part 5 : Graph Algorithms. Addison-Wesley.

Shinjo, R., Taniyama, K. (2003). Homology classification of spatial graphs by linking numbers and Simon invariants. Topology and its Applications *134*, 53–67.

Vellend, M. (2003). Island biogeography of genes and species. The American Naturalist *162*, 358–365.

Vellend, M. (2010). Conceptual synthesis in community ecology. The Quarterly Review of Biology *8*, 183–206.

Vellend, M., Orrock, J. L. (2009). Ecological and genetic models of diversity. In : Losos, J. B., Ricklefs, R. E. (Eds.), The Theory of Island Biogeography Revisited. Princeton University Press, pp. 439–462.

Wallace, A. R. (1855). On the law which has regulated the introduction of new species. Ann. Magazine Nat. Hist. Ser. 2 *16*, 184–196.

Wennekes, P. L., Rosindell, J. L., Etienne, R. S. (2012). The neutral – niche debate : A philosophical perspective. Acta Biotheoretica .

Witt, C. C., Maliakal-Witt, S. (2007). Why are diversity and endemism linked on islands ? Ecography *30*, 331–333.

ANNEXE C

A SIMPLE MODEL TO STUDY PHYLOGEOGRAPHIES AND SPECIATION PATTERNS IN SPACE

Abstract

In this working paper, we present a simple theoretical framework based on network theory to study how speciation, the process by which new species appear, shapes spatial patterns of diversity. We show that this framework can be expanded to account for different types of networks and interactions, and incorporates different modes of speciation.

C.1 Motivation

The peculiar spatial relationship between closely related species was among the first patterns of diversity used to infer evolution. As early as the 1850s, Alfred Wallace noted that the closest relatives were often observed in adjacent yet non-overlapping regions (Wallace, 1852 ; 1855). Wagner and Jordan later relied on a similar observation to argue for the importance of geography and isolation in the formation of new species (Coyne et Orr, 2004). And finally, Mayr developed a theory of allopatric speciation, a cornerstone of the modern synthesis, again using similar observations (Mayr, 1942 ; 1963). The relationship between phylogeny and geography has shaped our understanding of the origin of species (Coyne et Orr, 2004 ; Losos, 2009). It is also crucial to the development of a unified theory of community assembly (Rice, 2009 ; Mouquet et al., 2012). Yet, theory remains mostly silent about the subject. Few models can generate phylogeographies, and none can be used to study the effect of complex spatial structures (Barraclough et Vogler, 2000). This is surprising, not only because of the theoretical importance of phylogeography, but also because several phylogenetic methods use geography to infer patterns of speciation (Lynch, 1989 ; Barraclough et Vogler, 2000 ; Losos et Glor, 2003).

Part of the problem lies in the limitations of traditional mathematical methods : analytical solutions to spatially-explicit models are often only available for the most trivial cases (Epperson et al., 2010). Thus, we are left with no theoretical framework to study the patterns noted by Wallace, Wagner, and Mayr. In this document, we describe a very simple algorithm to generate phylogeographies in spatial networks. Our approach is inspired by metapopulation theory (Levine, 1969 ; Hanski, 1999 ; 2009) although the spatio-temporal scale is different : we're interested in the dynamics of populations at the regional scale during long periods. The model will be used to study phylogeographies in various spatial contexts and to develop better tools to understand the relationship between phylogeny and geography.

We use the term "phylogeography" in the general sense : it is the union of phylogenetics with geography. Our approach emphasizes how spatial patterns of speciation shape biodiversity. It cannot be used to study within-species variations, a major focus of phylogeography (Freeland et al., 2011). This is more consistent with the field known as comparative phylogeography.

C.2 Modeling the landscape

We model the landscape as a spatial network of communities. A network is a flexible mathematical object defined as a set of vertices V and a set of edges E, which are used to connect the vertices (Newman, 2010). Here, the vertices represent communities and the edges denote migration (Economo et Keitt, 2008 ; 2010 ; Desjardins-Proulx et Gravel, 2012b ;a). Spatial networks are simply networks in which vertices are embedded in a known topological space (Kobayashi, 1994), in our case a two-dimensional map. Thus, each community is represented by a vertex in the network and to a position on a map. Networks are increasingly common in ecology as they can be used to model complex structures and quantify the effect of clustering, connectivity, and isolation (Minor et Urban, 2007 ; 2008 ; Urban et al., 2009 ; Dale et Fortin, 2010). In particular, isolation is the most important factor in many speciation events (Coyne et Orr, 2004), making networks well-suited to study patterns of speciation in different contexts (Desjardins-Proulx et Gravel, 2012b ;a). The spatial network can be built in two ways. First, random geometric networks can be generated by randomly placing the vertices on a surface, normally the unit square, and linking all communities within some threshold distance (Penrose, 2003). This technique is used to test network algorithms applied to maps (Sedgewick, 2001). Second, a real map can be used as a template for a spatial network (Minor et Urban, 2007 ;

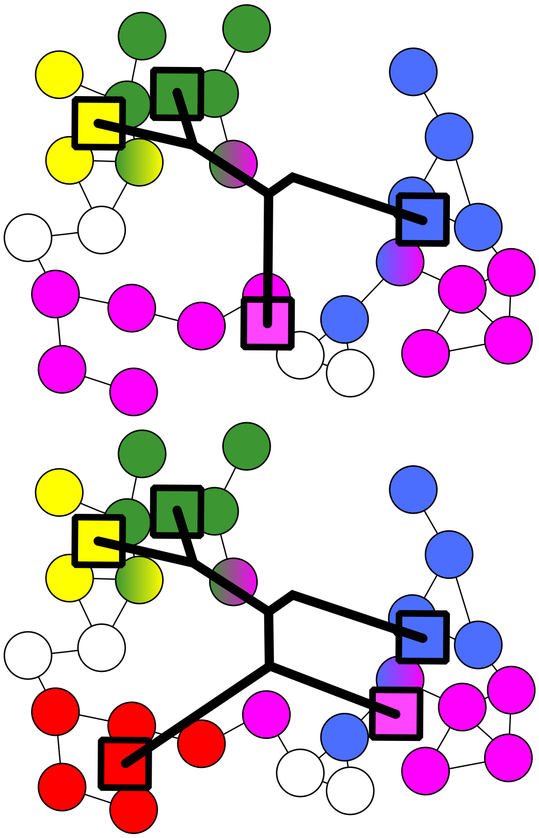


Figure C.1 – How the Wagner model handles speciation in space. Top : a phylogeography with four species (yellow, blue, green, pink). The populations are distributed in a spatial network, with each community (circles) hosting populations from 0 or more species. Empty communities are white and a gradient is used for communities with more than one species. The communities are connected by migration (thin black lines). Bottom : a speciation event. The pink species is divided in three groups of populations. Its leftmost group undergoes speciation and a connected subgroup now belongs to a new species (in red).

Dale et Fortin, 2010). This method offers the opportunity to generate predictions specific to a given spatial structure, and test the predictions of our algorithm against empirical data.

C.3 The model

A species is divided in populations which are distributed in a network of communities. A species is either present or absent in a community, we do not keep track of the number of individuals. Occupancy thus follows the standard colonization/extinction dynamics of metapopulation theory (Hanski, 1999). For each time step, all populations have the opportunity to colonize adjacent communities (the vertices connected by an edge in the network). The probability of a successful colonization of community x by species i is

$$c(i,x) = c_{max} \exp\left(-\aleph \sum_{j \in \{S_x \setminus i\}} \delta_{ij}^{-1} \right), \tag{C.1}$$

with $\{S_x \setminus i\}$ being the set of populations present in community x minus i, δ_{ij} is the time since species i and j's most recent common ancestor, c_{max} is the highest possible colonization rate and $\aleph$ a positive constant (with $\aleph \geq 0$). $\aleph$ describes the decline of the intensity of interactions with phylogenetic divergence. In short, a higher $\aleph$ makes it difficult for closely related species to coexist. c_{ix} is a very simple function derived from exponential decay. It is based on an old hypothesis by Darwin : closely related species are more likely to compete. It has recently received experimental support (Jiang et al., 2010; Violle et al., 2005). A strong assumption of trait conservatism underlies the model (Losos, 2008). At each time step, all populations have the same probability e of extinction. Speciation occurs in groups of populations. We define a group as a set of connected populations from the same species (Fig. C.1). Each group has a probability v of undergoing speciation. When speciation occurs in a group, a random subset of $[1, n]$ connected populations will speciate, with n being the number of populations in the original group (Fig. C.1).

C.4 Variations

The basic model can easily be extended to account for various types of interactions. In this section we briefly discuss a few extensions.

C.4.1 Allopatric speciation

Our model is mostly parapatric, with strictly allopatric speciation occurring only with probability $1/n$, with n being the size of the group to speciate. An alternative is to always force allopatric speciation by making the entire group speciate.

C.4.2 Sympatric speciation ?

With few solid cases of sympatric speciation, and many of them involving important allopatric/parapatric phases (Coyne et Orr, 2004 ; Bolnick et Fitzpatrick, 2007 ; Fitzpatrick et al., 2008), it is hard to decide how to do a phenomenological sympatric speciation model. Furthermore, the assumption of strong niche conservatism would be hard to maintain, as niche overlap between diverging populations is one of the hardest challenge for sympatric speciation. Nevertheless, if enough sympatric speciation events can be analyzed, our model could be modified to allow sympatric, parapatric, and allopatric speciation.

C.4.3 Variable α

$\aleph$ is fixed in the original model, but it could vary in time and space. For example, smaller regions could have higher $\aleph$ to account for a lower carrying capacity.

C.4.4 Variable extinction rates

The extinction rate could have the same form as the colonization rate and be affected by closely related species.

C.4.5 Variables v

The speciation rate could decrease with higher diversity (more niches are filled) or increase ("diversity begets diversity") (Erwin, 2005).

C.4.6 Growing food webs

The basic idea of using spatial networks and groups of connected populations for speciation could be used to model how complex food webs grow with speciation events. This integration would, however, require many new assumptions and a more sophisticated model for the colonization and extinction rates.

Integrating food web dynamics lead to some difficulties. For example, a trophic model would involve very different species with potentially different rates of dispersal. The threshold value r used to determine the realized links in the spatial network would have to be different for each group of species. For example, group-specific threshold values could be linked to the niche value (i.e. : smaller species have lower dispersal ranges). A connected random geometric networks could then be generated with the lowest threshold value, ensuring that all networks are fully connected.

C.4.7 Positive interactions

Positive interactions between closely related species are also possible, for example in plants. This variation can be achieved by making c_{ix} increase when related species are present.

C.5 Implementation

An open-source implementation is available on github : https://github.com/PhDP/wagner.

C.6 References

Barraclough, T. G., Vogler, A. P. (2000). Detecting the geographical pattern of speciation from species-level phylogenies. The American Naturalist *155*, 419–434.

Bolnick, D. I., Fitzpatrick, B. M. (2007). Sympatric speciation : models and empirical evidence. Annual Review of Ecology, Evolution, and Systematics *38*, 459–487.

Coyne, J. A., Orr, H. A. (2004). Speciation. Sinauer Associates.

Dale, M. R. T., Fortin, M.-J. (2010). From graphs to spatial graphs. Annual Review of Ecology, Evolution, and Systematics *41*, 21–38.

Desjardins-Proulx, P., Gravel, D. (2012a). A complex speciation-richness relationship in a simple neutral model. Ecology and Evolution *2*, 1781–1790.

Desjardins-Proulx, P., Gravel, D. (2012b). How likely is speciation in neutral ecology ? The American Naturalist *179*, 137–144.

Economo, E. P., Keitt, T. H. (2008). Species diversity in neutral metacommunities : a network approach. Ecology Letters *11*, 52–62.

Economo, E. P., Keitt, T. H. (2010). Network isolation and local diversity in neutral metacommunities. Oikos *10*, 1–9.

Epperson, B. K., McRae, B. H., Scribner, K., Cushman, S. A., Rosenberg, M. S., Fortin, M. J., James, P. M. A., Murphy, M., Manel, S., Legendre, P., Dale, M. R. T. (2010). Utility of computer simulations in landscape genetics. Molecular Ecology *19*, 3549–3564.

Erwin, D. (2005). Seeds of diversity. Science *308*, 1752–1753.

Fitzpatrick, B. M., Fordyce, J. A., Gavrilets, S. (2008). What, if anything, is sympatric speciation ? Journal of Evolutionary Biology *21*, 1452–1459.

Freeland, J., Kirk, H., Petersen, S. (2011). Molecular Ecology, 2nd Edition. Wiley-Blackwell.

Hanski, I. (1999). Metapopulation ecology. Oxford University Press.

Hanski, I. (2009). The theories of island biogeography and metapopulation dynamics. In : Losos, J. B., Ricklefs, R. E. (Eds.), The Theory of Island Biogeography Revisited. Princeton University Press, pp. 186–213.

Jiang, L., Tan, J., Pu, Z. (2010). An experimental test of darwin's naturalization hypothesis. The American Naturalist *175*, 415–423.

Kobayashi, K. (1994). On the spatial graph. Kodai Mathematical Journal *17*, 511–517.

Levine, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America *15*, 237–240.

Losos, J. B. (2008). Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecology Letters *11*, 995–1003.

Losos, J. B. (2009). Lizards in an Evolutionary Tree : Ecology and Adaptive Radiation of Anoles. University of California Press.

Losos, J. B., Glor, R. E. (2003). Phylogenetic comparative methods and the geography of speciation. Trends in Ecology and Evolution *18*, 220–227.

Lynch, J. D. (1989). The gauge of speciation : on the frequency of modes of speciation. In : Otte, D., Endler, J. A. (Eds.), Speciation and its consequences. Sinauer Associates, pp. 527–553.

Mayr, E. (1942). Systematics and the origin of species. Columbia University Press, New York.

Mayr, E. (1963). Animal Species and Evolution. Belknap, Cambridge, MA.

Minor, E. S., Urban, D. L. (2007). Graph theory as a proxy for spatially explicit population models in conservation planning. Ecological Applications *17*, 1771–1782.

Minor, E. S., Urban, D. L. (2008). A graph theory framework for evaluating landscape connectivity and conservation planning. Conservation Biology *22*, 297–307.

Mouquet, N., Devictor, V., Meynard, C., Munoz, F., Bersier, L., Chave, J., Couteron, P., Dalecky, A., Fontaine, C., Gravel, D. (2012). Ecophylogenetics : advances and perspectives. Biological reviews *87*, 769–785.

Newman, M. (2010). Networks : An Introduction. Oxford University Press.

Penrose, M. (2003). Random Geometric Graphs. Oxford University Press.

Rice, S. H. (2009). A stochastic version of the Price equation reveals the interplay of deterministic and stochastic processes in evolution. BMC Evolutionary Biology *8*, 262.

Sedgewick, R. (2001). Algorithms in C++ Part 5 : Graph Algorithms., 3rd Edition. Addison-Wesley Professional.

Urban, D. L., Minor, E. S., Treml, E. A., Schick, R. S. (2009). Graph models of habitat mosaics. Ecology Letters *12*, 260–273.

Violle, C., Nemergut, D. R., Pu1, Z., Jiang, L. (2005). Phylogenetic limiting similarity and competitive exclusion. Ecology Letters *14*, 782–787.

Wallace, A. R. (1852). On the monkeys of the amazon. Proc. Zool. Soc. Lond. *20*, 107–110.

Wallace, A. R. (1855). On the law which has regulated the introduction of new species. Ann. Magazine Nat. Hist. Ser. 2 *16*, 184–196.

BIBLIOGRAPHIE

Bach, S., Broecheler, M., Huang, B., Getoor, L. (2015). Hinge-loss markov random fields and probabilistic soft logic *arXiv : 1505.04406*.

Bartomeus, I., Gravel, D., Tylianakis, J., Aizen, M., Dickie, I., Bernard-Verdier, M. (2016). A common framework for identifying linkage rules across different types of interactions. Functional Ecology *30*, 1894–1903.

Begun, D., Holloway, A., Stevens, K., Hillier, L., Poh, Y., Hahn, M., Nista, P., Jones, C., Kern, A., Dewey, C., Pachter, L., Myers, E., Langley, C. (2007). Population genomics : Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLOS Biology *5*, e310.

Bell, G. (2010). Fluctuating selection : the perpetual renewal of adaptation in variable environments. Phil. Trans. R. Soc. B *365*, 87–97.

Coyne, J. A., Orr, H. A. (2004). Speciation. Sinauer Associates.

Garnelo, M., Arulkumaran, K., Shanahan, M. (2016). Towards deep symbolic reinforcement learning *arXiv :1609.05518v2*.

Gillespie, J. H. (2004). Population Genetics : A Concise Guide, 2nd Edition. Hopkins Fulfillment Service.

Gravilets, S. (2004). Fitness Landscapes and the Origin of Species. Princeton University Press.

Hahn, M. (2008). Toward a selection theory of molecular evolution. Evolution *76*, 255–265.

Hu, Z., Ma, X., Liu, Z., Hovy, E., Xing, E. (2016). Harnessing deep neural networks with logic rules *arXiv :1603.06318*.

Hubbell, S. P. (2001). The Unified Neutral Theory of Biodiversity and Biogeography. Vol. 32 of Monographs in Population Biology. Princeton University Press.

Kimmig, A., Bach, S., Broecheler, M., Huang, B., Getoor, L. (2012). A short introduction to probabilistic soft logic. In : Proceedings of the NIPS Workshop on Probabilistic Programming.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

King, J., Jukes, T. (1969). Non-darwinian evolution. Science *164*, 788–798.

Kopelke, J., Nyman, T., Cazelles, K., Gravel, D., Vissault, S., Roslin, T. (2017). Food-web structure of willow-galling sawflies and their natural enemies across europe. Ecology *98*, 1730.

Lewontin, R., Hubby, J. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. Genetics *54*, 595–609.

Lynch, M. (2007). The Origins of Genome Architecture. Sinaur Associates.

Murphy, K. (2012). Machine Learning : A Probabilistic Perspective. The MIT Press.

Pimm, S. (1982). Food Webs. Springer.

Poisot, T., Canard, E., Mouillot, D., Mouquet, N., Gravel, D. (2012). The dissimilarity of species interaction networks. Ecology Letters *15*.

Provine, W. (2001). The Origins of Theoretical Population Genetics. University Of Chicago Press.

Richardson, M., Domingos, P. (2006). Markov logic networks. Machine Learning *62*, 107–136.

Russell, S., Norvig, P. (2009). Artificial Intelligence : A Modern Approach, 3rd Edition. Prentice Hall.

West, G., Brown, J., Enquist, B. (1997). A general model for the origin of allometric scaling laws in biology. Science *276*, 122–126.

Wilke, C. O. (2012). Bringing molecules back into molecular evolution. PLOS Computational Biology *8*, e1002572.

Williams, R., Martinez, N. (2000). Simple rules yield complex food webs. Nature *404*, 180–183.