Journée des stagiaires 2014–2015

Jeudi 2 juillet 2015

Table des matières

1	Étienne Auclair	2
2	Romain Alexandre	3
3	Matthieu Pluntz	4
4	Rolande Kpekou-Tossou	5
5	Franck Boizard	6
6	Éloi Navarro	7
7	Sara Maqrot	8
8	Amélia Landre	9
9	Lina Hadj Nassar	10
10	Augustin Chini	11
11	Sara Hardi	12
12	Prune Pegot-Espagnet	13
13	Abdallah Mechati	14
11	Andrea Martelletti	15

Apprentissage de réseaux écologiques par réseau bayésien dynamique

Etienne Auclair Encadrement : Régis Sabbadin, Nathalie Peyrard Stage du 01/04/2015 au 30/09/2015

Objectif

Un réseau écologique décrit les relations existantes entre des espèces vivantes. Il peut s'agir de relations trophiques, mais aussi parasitiques, compétitives, facilitatrices etc. Dans certains cas, les relations peuvent être mal connues. Le but de ce stage est de développer une méthode permettant d'apprendre un réseau écologique mal connu à partir de séries temporelles de présence/absence d'espèces.

Modélisation

Un réseau écologique peut être vu comme un graphe orienté dont un nœud représente une espèce et un arc représente une interaction entre 2 espèces. Pour distinguer les relations possibles, les arcs sont labellisés afin de décrire le type d'interaction. Le label décrit le fait qu'une espèce soit une proie $(p)^1$, un facilitateur $(f)^2$ ou ait une influence négative $(-)^3$ sur une autre.

Des séries temporelles de présence/absence d'espèces permettent d'apprendre un tel réseau. La dynamique des espèces est ainsi décrite par un réseau bayésien dynamique : la probabilité de présence ou d'absence d'une espèce à un moment donné dépend uniquement de la présence ou de l'absence au moment précédent des espèces en interaction avec elle ⁴.

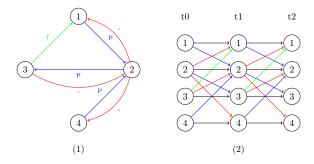


FIGURE 1 - (1) Réseau écologique sous forme de graphe. (2) Réseau bayésien dynamique associé

Un réseau bayésien décrit habituellement des probabilités conditionnelles sous forme de table. Dans le cadre du stage, le modèle est défini par un vecteur de paramètres θ , composé de probabilités de réussite de chaque interaction, d'une probabilité de recolonisation (indépendante des interactions) et d'une pénalité selon l'état de protection de la zone.

Méthodes

Des méthodes à base de score (ex : BIC) permettent d'apprendre la structure d'un réseau bayésien dynamique à l'aide d'estimation de paramètres (ex : par maximum de vraisemblance). Il faut adapter ces méthodes à un problème paramétré et à un graphe dont les arcs sont labellisés. Les étapes du stage sont les suivantes :

- Estimation des paramètres : À graphe fixé G, estimer les paramètres θ par maximum de vraisemblance. Calculer un score BIC à partir de ces estimations.
 - Écriture du maximum de vraisemblance (Effectué) : Exprimer la vraisemblance du modèle et un système d'équations à annuler.
 - Implémentation (en cours): Simuler des données et résoudre le système d'équation sous matlab.
- Amélioration de la structure : Rechercher un graphe G1 dont le score BIC est supérieur à celui de G avec les paramètres θ .
 - Algorithme d'apprentissage : Développer un algorithme itératif pour l'inférence de la structure, qui alterne les étapes d'estimation et d'amélioration.

L'algorithme sera validé par des données simulées et réelles.

Données

Jennifer Caselle et Laura Dee (Univ. of California, santa Barbara) ont fourni pour ce stage des données d'observation d'espèces marines dans des forêts de Kelp au large de la Californie. 163 espèces vivantes regroupées dans un réseau écologique connu sont observées pendant 14 ans sur 4 zones différentes.

^{1.} Une espèce a besoin de proies pour survivre, excepté si elle est en bas la chaîne alimentaire. Ex : Plantes, Algues, Plancton.

^{2.} Il s'agit d'une relation bénéfique, par exemple, le fait qu'une espèce végétale serve d'habitat à une espèce animale.

^{3.} Il peut s'agir d'un prédateur, d'un parasite ou d'un compétiteur.

^{4.} D'autres paramètres comme l'état de protection de la zone sont également pris en compte.

Analyse du compromis rendement-biodiversité sur un cas d'étude : système culture/adventices/pollinisateurs.

Romain ALEXANDRE, M2 IMAT Université Paul Sabatier.

Stage du 01/04/2015 au 30/09/2015

La plupart des familles d'insectes pollinisateurs sont victimes depuis les années 1920 d'un effondrement populationnel voire d'extinction d'espèces. Cela est dû, en partie, à la destruction des habitats naturels ou semi-naturels et à la perte de leurs ressources naturelles (ressources en fleurs). Avec une occupation de 53,2% de la surface de la France métropolitaine, l'agriculture pourrait jouer un rôle important dans la conservation de la biodiversité. La France, étant le premier pays agricole de l'Union européenne avec 18% du produit agricole et agro-alimentaire européen, verrait, si la présence d'adventices (nécessaire à la survie des pollinisateurs) augmentait dans les cultures, son rendement chuté par la nuisibilité des adventices. Il est important d'apporter une solution de compromis pouvant satisfaire différents acteurs ayant leur propre objectif, à savoir : les agriculteurs (maximiser le rendement des cultures), les apiculteurs (maximiser le rendement en miel) et les citoyens (maximiser la biodiversité).

Dans le cadre d'un cas d'étude de culture/adventice/pollinisateur et du modèle proposé lors de la thèse menée par J. Radoszycki ¹ (INRA MIAT), nous proposons, lors de ce stage, de comparer des stratégies de gestion sur la rotation des cultures d'un parcellaire (ensemble de parcelles) obtenues par optimisation ou par expertise, en terme du compromis au niveau du paysage.

En s'appuyant sur le premier modèle PDMF³ des dynamiques des adventices et pollinisateurs ainsi que sur les modèles de récompenses établies par J. Radoszycki, nous souhaitons vérifier que les résultats attendus concernant les comportements moyens des dynamiques des adventices et pollinisateurs pour des stratégies de gestion données (mêmes cultures sur tout le paysage, choix aléatoire-uniforme sur les cultures, répartion uniforme des cultures) soient du même ordre que les comportements connus par expertise. Pour cela, en plus de m'avoir permis de m'approprier le modèle, j'ai effectué une série d'analyses par réalisation de simulation concernant les comportements moyens des dynamiques des adventices/pollinisateurs. Après validation du modèle, nous proposons de mesurer l'importance de la prise en compte des interactions spatiales dans les dynamiques des adventices et des pollinisateurs, dans la recherche de stratégies de gestion permettant d'obtenir un compromis au niveau paysage entre les différents services (rendement, miel, biodiv). Si les dépendances spatiales sont ignorées alors qu'elles existent (propagation par le vent ou par les pollinisateurs) est-ce que nous arriverons à convevoir des stratégies de même qualité que si nous les prenons en compte (ce que nous faisons actuellement). Si la réponse est oui, nous pourrons travailler avec un modèle plus simple,sans dynamiques spatiales et traiter des problèmes de taille plus importante.

Le travail effectué:

- S'approprier le vocabulaire technique du sujet
- Comprendre les dynamiques des adventices, des pollinisateurs et les récompenses du modèle
- S'approprier le modèle par réalisation de simulation et étude du comportement moyen des dynamiques
- Construction des différentes questions écologiques pertinentes (en collaboration avec S. Gaba (écologue INRA Dijon)).

Le travail en cours:

— Proposition d'un modèle de type PDM des dynamiques adventices/pollinisateurs et d'un modèle de récompense sans la prise en compte des interactions spatiales.

Poursuite possible durant le stage :

— Etudier à quelles échelles il est plus intéressant de prendre des décisions (parcelles, ensemble de parcelles ou paysage) pour obtenir un compromis au niveau paysage entre les différents services.

Encadrants: Nathalie Peyrard; Régis Sabbadin.

^{1.} Thèse de Julia Radoszycki (2012/2015) : Résolution de processus décisionnels de Markov à espace d'état et d'action factorisés - Application en agroécologie.

Modélisation par chaîne de Markov cachée de la dynamique des espèces adventices dans un agrosystème

Matthieu Pluntz (ENSAE, Paris)

Les adventices sont les mauvaises herbes qui se trouvent dans les parcelles cultivées. Leur influence sur l'agrosystème dont elles font partie est multiple, car elles causent une perte de rendement agricole tout en jouant un rôle clé dans la biodiversité du milieu agricole. La question de la dynamique des populations d'adventices est donc fondamentale pour la compréhension des agrosystèmes.

Nous considérons des adventices dont le cycle de vie est annuel et qui se perpétuent en émettant des graines qui s'installent dans le sol, dont la germination donne naissance l'année suivante à une nouvelle génération de plantes. Ces graines ont la capacité de survivre plusieurs années dans le sol avant de germer. Ce phénomène s'appelle la banque de graines. La dynamique des populations d'adventices dans une parcelle agricole repose donc sur plusieurs mécanismes qui interagissent entre eux : reproduction annuelle, banque de graines, et colonisation par des graines venues de l'extérieur de la parcelle. Pour comprendre la dynamique d'une espèce, nous cherchons à quantifier le rôle de ces différents phénomènes dans son cycle de vie en estimant les paramètres respectifs qui les déterminent.

Les seules données disponibles pour réaliser cette estimation portent sur les plantes présentes en surface : il est très coûteux d'observer la banque de graines. Les modèles à chaîne de Markov cachée ou HMM (*Hidden Markov Models*) sont particulièrement adaptés pour décrire cette situation.

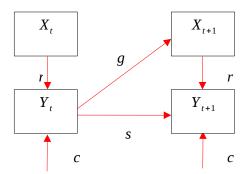


Fig 1: HMM représentant la présence ou absence de l'espèce en surface $X_t \in \{0,1\}$ (variable observable) et de graines $Y_t \in \{0,1\}$ (variable non observable). Les paramètres à estimer sont les probabilités de reproduction r, colonisation c, germination g et survie de la banque de graines g.

Notre travail porte donc d'une part sur la construction de modèles HMM avec une composante spatiale en s'appuyant sur les modèles déjà proposés^{1,2}, d'autre part sur l'estimation des paramètres à l'aide de méthodes adaptées au contexte d'un HMM. Il s'agit notamment de l'algorithme EM de maximisation itérative de la vraisemblance, où l'étape E s'appuie sur l'algorithme Forward-Backward de calcul de la loi conditionnelle des variables cachées. Pour des modèles plus complexes, des variantes variationnelles de l'algorithme EM ou des méthodes de Monte-Carlo dans un contexte bayésien seront utilisées.

Encadrants: P-O. Cheptou (CEFE, Montpellier) et N.Peyrard

date: 10 mai – 30 octobre 2015

¹Borgy,B., Reboud, X., Peyrard, N., Sabbadin, R., Gaba, S., Dynamics of weeds in the soil seed bank: a Hidden MarkovModel to estimate life history traits from standing plant time series, en revision (PlosOne).

²Fréville, H., Choquet, R., Pradel, R., Cheptou, P-O., Inferring seed bank from hidden Markov models: new insights into metapopulation dynamics in plants, Journal of Ecology, 2013.

Stage Analyse par simulation de l'interaction climat/rendement

Rolande KPEKOU 23 juin 2015

Stage encadré par Victor Picheny & Nathalie Villa-Vialaneix

Résumé du stage Des travaux récents en agronomie portent sur la modélisation de systèmes de culture à différentes échelles (paysage, parcelle, plante) tenant compte de conditions climatiques (ensoleillement, pluviométrie, etc.) qui sont typiquement représentées sous forme de séries temporelles. Ces conditions ayant une influence très forte sur le comportement de la majorité de ces systèmes, intégrer l'aléa climatique aux études d'analyse d'optimisation de rendement est une question primordiale. Cependant, la complexité des interactions plante/climat à l'œuvre ne permet pas d'étudier les systèmes de manière explicite. On a alors recours aux approches dites par simulation, où les relations entrées/sorties sont inférées à l'aide d'outils statistiques à partir d'un échantillon simulé.

Ce stage s'intéresse à l'étude de variétés de tournesol à l'aide d'un modèle de simulation dynamique (SUNFLO). L'objectif poursuivi est de comprendre la manière dont le climat influence le rendement du tournesol. Il s'agit précisément de mettre en valeur les motifs climatiques les plus influents pour le rendement du tournesol.

Pour atteindre cet objectif, nous disposons d'un jeu de données composé de diverses séries climatiques annuelles, d'une part, et de divers ensembles de traits phénotypiques du tournesol, d'autre part. Ces deux catégories de variables sont utilisées comme des paramètres d'entrée du simulateur dynamique (SUNFLO) pour le calcul du rendement. Les données climatiques contiennent 5 variables : la température minimale, la température maximale, l'évapotranspiration, l'ensoleillement et les précipitations. Elles correspondent à 190 relevés climatiques sur cinq stations (Avignon, Blagnac, Dijon, Poitiers et Reims). Nous avons ensuite généré un plan d'espériences en croisant les 190 simulations climatiques par 1000 types de plantes décrites au travers de 8 caractéristiques phénotypes.

Le problème peut se formaliser sous la forme d'une question dans laquelle on cherche à comprendre l'influence d'entrées X (le climat et les traits phénotypiques) sur une sortie Y (le rendement). Le modèle SUNFLO permet de calculer explicitement les sorties pour une entrée donnée mais les résultats ne sont pas facilement interpretables. On a alors recours à une approche dite par méta-modélisation : ce type d'approche consiste à construire un modèle statistique permettant d'approcher le modèle SUNFLO avec des temps de calcul moindres tout en conservant de bonnes performances prédictives. Elle repose sur la sélection de variables d'entrée influentes, et sur l'approximation de la relation entre les variables de sortie et ces variables influentes. Plusieurs types de méta-modèles sont utilisés dans la littérature. Dans cette étude, nous utilisons le modèle linéaire dans un premier temps et dans un second temps, une approche plus flexible, les forêts aléatoires. Afin d'identifier les variables les plus influentes et les périodes les plus sensibles, une analyse de sensibilité ou une recherche des variables contribuant le plus à la qualité de la prédiction est faite sur le méta-modèle. Dans le cas du modèle linéaire, des indices de sensibilité ont été calculés et dans le cas des forêts aléatoires, l'importance des variables a été utilisée.

Des méta-variables correspondant à des résumés (moyenne, écart-type, maximum) ont été d'abord générées sur des intervalles de temps réguliers (mois, 2 semaines, une semaine). Ensuite, plusieurs modèles linéaires (sur des résumés mensuels, bihebdomadaires et hebdomadaires) ont été effectués pour expliquer le rendement. Les premiers résultats indiquent que le modèle linéaire sur des données hebdomadaires explique 88% des variabilités du rendement. Des forêts aléatoires on été générées à partir des 186 méta-variables utilisées pour le modèle linéaire. Le calcul de l'importance des variables laisse croire que les variables climatiques exercent une forte influence sur le rendement à partir de la 20ième semaine de la période de culture (avril-septembre), sans laisser apparaître de motifs climatiques clairs. La principale difficulté rencontrée est la corrélation entre les variables, surtout pour l'analyse de sensibilité. Pour la suite, nous essayerons d'approfondir les travaux sur les forêts aléatoires en calculant des importances groupées. Nous envisageons aussi faire un modèle linéaire sur une ACP fonctionnelle.

Méthodes d'analyse de sensibilité de modèles pour entrées climatiques

Stagiaire: Franck Boizard

Encadrement: Robert Faivre, Ronan Trépos

Contexte:

La plateforme RECORD (Rénovation et CoorDination de la modélisation de cultures pour la gestion des agro écosystèmes) de modélisation et de simulation a pour objectif d'aider à la conception et à l'évaluation des systèmes de production. Les modèles biophysiques développés sont dynamiques, fonction des conditions climatiques et se caractérisent par un grand nombre de paramètres relatifs aux différents processus pris en compte dans le système. L'analyse de sensibilité de ces modèles à leurs facteurs d'entrée est une phase indispensable pour juger de leur qualité.

Objectifs du stage :

Ce stage a pour objectif la mise au point de méthodes génériques d'analyse de sensibilité basée sur la décomposition de la variance permettant de traiter des entrées temporelles comme la température, le rayonnement, les précipitations. Il s'agira de mettre au point différentes méthodes permettant de caractériser finement la sensibilité d'un modèle à des entrées de type « séries climatiques ».

Les méthodes usuelles d'analyse de sensibilité sont faites pour des entrées réelles. Les données climatique sont des données fonctionnels. Il est donc obligatoire de procéder à un prétraitement des données climatiques. L'approche que l'on a privilégiée pour ce stage est la constitution d'une classification des séries climatiques avant de procéder à une analyse de sensibilité du modèle avec cette typologie de climat et les autres facteurs du modèle.

Grâce à une distance adaptée, cette méthodologie permet de répondre aux problématiques liées aux séries climatique. On obtient un typologie de climat robuste et automatique. L'analyse de sensibilité permettra non seulement d'identifier l'importance de l'impact global du climat et des autres paramètre du modèle, mais aussi les interractions des différents types de climats.

La mise en œuvre de la méthodologie sera effectuée sur le modèle SUNFLO (modèle numérique permettant d'estimer le rendement d'une parcelle de tournesol en fonction de divers paramètres génétiques et environnementaux) disponible sur la plateforme RECORD.

Résumé de stage :

Optimisation de stratégies d'échantillonnage des stress biotiques au sein des agroécosystèmes.

Eloi Navarro - UMR AGIR - encadré par Jean-Noël Aubertot et Michel Goulard

Mon stage se déroule au sein de l'UMR AGIR sur le centre de l'INRA à Auzeville, il s'inscrit dans le cadre du projet CASIMIR (*CAractérisation SIMplifiée des pressions blotiques et des Régulation biologiques*) qui est un projet Pour et Sur le Plan ECOPHYTO (PSPE).

Au sein du plan ECOPHYTO, des agriculteurs volontaires mettent en placent des méthodes de gestion des bio-agressions utilisant moins de produits phytosanitaires dans 1900 fermes rassemblées au sein réseau FERMES du dispositif DEPHY. Le projet CASIMIR vise à proposer un soutien méthodologique quant à la caractérisation des évolutions des stress biotiques au sein du réseau. Plus précisément, les ingénieurs chargés du suivi des 1900 fermes doivent être pourvus de protocoles leur permettant, avec des ressources limitées, de rassembler des informations fiables sur les agroécosystèmes, et en particulier les stress biotiques à des fins de suivi et de recherche scientifique.

L'optimisation des stratégies d'échantillonnage participe de la bonne qualité des protocoles appliqués par les ingénieurs lors de leurs visites dans les parcelles agricoles. Les objectifs du stage sont de mener à bien une recherche bibliographique poussée sur l'échantillonnage pour les pressions biotiques, puis de mettre en forme les connaissances recueillies pour produire un outil d'aide à la décision, et enfin de tester l'outil sur des jeux de données disponibles.

La synthèse bibliographique a permis d'obtenir une vue d'ensemble sur les méthodes d'échantillonnage, de faire ressortir les contraintes déterminantes et de faire un recueil des bonnes pratiques. Les grandes lignes d'un outil d'aide à la décision ont ensuite pu être mises en place, elles doivent maintenant être formalisées.

Méthode approchée en programmation mathématique Application à la conception de vergers maraîchers

Sara MAOROT

Thématique: programmation linéaire, heuristique, optimisation combinatoire, agronomie.

Equipe d'accueil : Statistique et Algorithme pour la biologie.

Laboratoire d'accueil : Mathématique et Informatique Appliquées de Toulouse, Institut National

de Recherche Agronomique.

Afin de proposer des systèmes de production agricoles écologiques répondant aux enjeux actuels auxquels l'agriculture est confrontée, les systèmes vergers maraîchers semblent particulièrement prometteurs car ils combinent des productions à la fois diversifiées et relativement intensifiées. Le problème de la conception d'un tel système consiste à définir l'emplacement et le choix des arbres fruitiers ainsi que la rotation des cultures maraîchères sur un horizon temporel de plusieurs années. La grande taille de ce problème nécessite la mise au point de méthodes d'optimisation approchées. Pour cela, il existe un algorithme appelé *In-the-middle* ou *Wedelin* (Wedelin, 1995) qui a l'avantage de pouvoir s'appliquer à des problèmes de très grande taille (plusieurs millions de variables et des dizaines de milliers de contraintes pour le problème de rotations d'équipages par exemple) en donnant des solutions approchées souvent de bonne qualité.

Les travaux de stage consistent à développer l'algorithme In-the-middle au cadre de la programmation linéaire en nombres entiers, choisir une stratégie de réglage de ses paramètres et évaluer sa qualité sur une version simplifiée du problème de conception de vergers maraîchers.

Bibliographie

Wedelin D. An algorithm for large scale 0-1 integer programming with application to airline crew scheduling. Annals of Operations Research, 57, 1995

Gestion forestière et décision d'assurance en univers risqué

Amélia Landre

Encadrement : Stéphane Couture, Régis Sabbadin, Marie-Josée Cros

Résumé

La gestion forestiere est une activite a risque etant donnee qu'elle depend notamment des variations climatiques et ce d'autant plus avec le contexte actuel de changement climatique. Il s'agit egalement d'une activite sur de longues periodes. Le proprietaire forestier doit alors appliquer des actions de gestion prennant en compte ces facteurs an d'optimiser son activite. Parmis ces choix de gestion, en plus de la recolte, il peut soit souscrire une assurance soit realiser une epargne. An de comprendre les dierentes choix du proprietaire, nous avons realise trois modeles sous forme de Processus Decisionnels de Markov: le premier correspond a un modele simple de gestion forestiere, le second correspond a une extension du premier qui inclut la possibilite d'assurance et le troisieme est un modele de gestion forestiere incluant la possibilite d'epargne. Nous avons teste, a l'aide de ces modeles, l'eet du niveau de risque de tempête, du coût de l'assurance et de la preference vis-a-vis du risque du proprietaire sur les choix de gestion. Les resultats nous montrent que le niveau de risque de tempête n'a que tres peu d'impact sur les choix de gestion alors que le coût de l'assurance et la preference vis-a-vis du risque modient ces choix. La peur du risque induit une recolte precoce et l'augmentation du coût de l'assurance diminue la propension a s'assurer. Enfin, le modele de gestion forestiere avec la decision d'assurance montre que le proprietaire ne s'assure pas lorsqu'il a peur du risque. Ce resultat inattendu peut provenir du choix de modelisation realise dans cette etude. elle ouvre des perspectives, notamment vers un modele de gestion incluant a la fois l'assurance et l'epargne.

RÉSUMÉ DE STAGE

Analyse de sensibilité des modèles agronomiques à l'estimation de la réserve en eau des sols

Lina Hadj Nassar 2 Juillet 2015

Les sols, en tant que réservoirs d'eau verte, sont essentiels dans les écosystèmes agricoles. Leur fonctionnement hydrique est décrit dans les modèles agronomiques par leur Réserve Utile (RU), i.e. la quantité d'eau disponible pour la croissance de la plante. Cette propriété hydrique peut être estimée selon différentes méthodes, chacune introduisant une incertitude liée aux erreurs de mesure.

L'objectif du projet RUEdesSOLS est, entre autres, d'évaluer l'impact du choix de la méthode et de l'incertitude associée sur les sorties des modèles (rendement, qualité de la production, besoin en eau, etc.), afin de proposer une méthodologie pour l'estimation de la RU qui puisse être utilisée dans un large champ de conditions pédoclimatiques.

Mon stage consiste à analyser d'abord la relation entre l'incertitude sur la Réserve Utile et l'incertitude des sorties du modèle agronomique AqYield. À l'aide du logiciel R, je vais mettre en oeuvre des solutions efficaces par simulation pour la propagation d'incertitudes et développer des stratégies algorithmiques pour trouver des seuils. Parmi les sorties du modèle, celles qui nous intéressent sont des éléments du bilan hydrique, à savoir : le rendement, la somme des drainages, la somme des évaporations, la somme des transpirations et la satisfaction hydrique sur un an de culture. Le graphe ci-dessous montre une superposition de courbes de la somme des drainages sur un an pour différentes valeurs de RU.

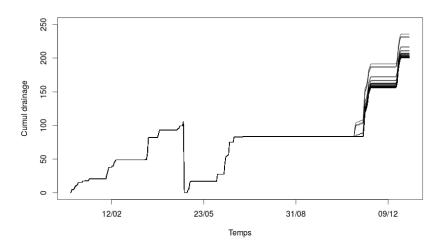


Figure 1: Superposition des sommes des drainages pour différentes valeurs de RU

Une fois l'analyse d'incertitude réalisée, je m'intéresserai à l'influence de la variable RU sur les sorties relativement à d'autres variables d'entrée (sol, itinéraire technique, plante, climat), grâce à une analyse de sensibilité.

Enfin, je réfléchirai au cas de données spatiales pour la RU, c'est-à-dire quand la parcelle n'est pas homogène. Il faudra proposer différentes méthodes pour traiter ce type de données, puis voir l'influence sur les résultats.

CHINI Augustin Tutrice: RAYNAL Hélène

Résumé stage de développement et intégration d'un modèle de simulation du taux de carbone dans le sol

Pour commencer le résumé de mon sujet de stage, je vais détailler les objectifs qui ont été fixés : la première partie de mon stage consiste à créer un modèle de simulation indiquant l'évolution du taux de carbone dans le sol forestier pour la plate-forme RECORD. Une fois que le modèle créé il a été stocké sur un serveur de Web Service -nommé eRECORD- qui a pour but d'offrir la consultation et modification des modèles hébergés directement par le biais de requêtes HTTP, ainsi à partir de n'importe quel plate-forme les modèle de simulation serons consultables. Une fois cette partie terminée j'ai créé un module Joomla (qui peut être assimilé à une petite partie d'un site s'intégrant à un gros site internet, plus précisément, une page web et tout le code qui lui est lié) qui a pour but de grâce à du PHP créé une requête HTTP et interroger le simulateur que j'ai créé sur la concentration de carbone dans le sol forestier. Ensuite le serveur eRECORD répond à cette requête en envoyant les résultats de simulation. Grâce à ces résultats le site met à disposition un tableau de résultats numériques exportable en CSV si cela est souhaité, mais aussi des courbes d'évolution dynamique du taux de carbone dans le sol.

Pour finir d'autres modèles d'indication du taux de carbone dans le sol qui ont déjà été créés (le modèle STICS qui est un modèle d'indication du taux de carbone dans le sol des grandes cultures et un autre dans les sols de prairies). Je vais alors devoir là aussi les héberger sur le serveur de Web service eRECORD et en suite créé le module Joomla correspondant pour les rendre disponible sur le site principal.

Je vais maintenant détailler l'utilisation future des outils créés lors de mon stage. Le site internet Joomla dans lequel je vais intégrer mon module de simulation, et le site ANAEE France, c'est un projet qui fait partie du programme "investissement d'avenir", et est subventionné par l'Agence National de la recherche, il est le fruit de la collaboration de l'INRA, du CNRS et de l'Université Joseph Fourier et est présent dans plusieurs pays au sein de l'Europe. Ce projet propose de comprendre, étudier et prédire la dynamique de la biodiversité et des écosystèmes dans un contexte de changement global (induis par l'impact de l'homme sur la planète), dans le but de faire émerger une bio-économie basée sur la connaissance du vivant et sur de nouvelles valorisations des ressources biologiques renouvelables. Le projet ANAEE offre des services en matière d'expérimentations mais aussi en matière de système d'information. Ainsi un ensemble de bases de données provenant de différents types d'expérimentations.

Pour conclure à terme, les outils que j'aurais créés dans le cadre de mon stage serons des outils d'aides à la décision qui serons utilisé par des scientifiques dans le cadre de recherches, par des professeurs pour créer des supports de cours mais aussi pour des décideurs dans différents secteurs qui grâce aux prévisions du modèle de simulation du taux de carbone pourra les conforter dans leur prise de décisions.

Sara HARDY

Encadrantes: Hélène CHIAPELLO, Annick MOISAN

<u>Titre</u>: Développement et mise en oeuvre d'un pipeline d'analyse du transcriptome de différentes souches du champignon phytopathogène *Magnaporthe oryzae*.

Résumé:

L'objectif de cette étude est de mettre en place un pipeline d'analyse des données de RNASeq de différentes souches du champignon phytopathogène *Magnaporthe oryzae (M. oryzae)* en condition d'infection dans le but de (1) **Mettre en évidence de nouveaux transcrits** de *M. oryzae*, exprimés en condition d'infection afin d'en faire l'annotation fonctionnelle. (2) analyser les gènes différentiellement exprimés entre les souches infectant leur hôte spécifique d'une part et un hôte universel (l'orge) d'autre part.

Le pipeline d'analyse de données RNAseq a permis la mise en évidence de nouveaux transcrits jusqu'alors non annotés dans le génome de la souche de référence de *M. oryzae*. Ces nouveaux transcrits ont été caractérisés (taille, expression, redondance, répartition génomique, comparaison de séquences). Ils possèdent en général une taille plus petite et un niveau d'expression plus élevé sur leur plante hôte que sur l'orge. La recherche de fonctions a montré que l'on retrouvait parmi ces nouveaux transcrits des parties de séquences répétées et de rétrotransposons.

L'analyse de l'expression de gènes différentiellement exprimés entre une souche infectant son hôte spécifique et une souche infectant un hôte universel a été réalisée avec succès pour un des échantillons : la souche FR13/riz comparée à la souche FR13/orge. En revanche, pour les autres échantillons qui ne possèdent pas de réplicats, seule une première recherche simple de gènes candidats différentiellement exprimés a été effectuée. Les résultats devront être interprétés avec prudence et validés expérimentalement.

Résumé du stage « Déterminisme de la résistance à la septoriose et à la fusariose du blé tendre »

Prune Pegot Espagnet

Encadrante: Brigitte Mangin

Cet exposé présente une étude du déterminisme génétique chez le blé tendre de la résistance à deux maladies : la septoriose et la fusariose. Ces deux maladies affectent le développement du blé tendre, provoquant une perte de rendement et de la qualité des grains. Or le blé tendre est la céréale la plus cultivée au monde et est une part essentielle du régime alimentaire de nombreuses populations, il est donc important que les sélectionneurs puissent produire de nouvelles variétés durablement résistantes à ces deux maladies.

Les données disponibles consistent en cinq populations d'haploïdes doublés, obtenues par croisement d'un même parent sensible avec cinq parents présentant des résistances partielles pour ces maladies. Chaque population est composée d'une centaine d'individus génotypés avec environ 420 000 SNP et phénotypés pour différents caractères sur trois sites en 2014.

Après une première étape de nettoyage des données, les marqueurs de chaque population sont cartographiés. Une carte consensus des cinq populations est alors réalisée. Dans le même temps, les données de phénotypage sont corrigées vis-à-vis d'un possible effet bloc. Deux méthodes de détection de QTL sont alors utilisées : la première étant une l'analyse de liaison, la seconde étant une étude d'association. Les résultats sont ensuite commentés et une discussion concernant les travaux à venir clôt cet exposé.

Reconstruction exacte de réseau bayésien. à partir d'observations complètes

Abdallah MECHATI Encadrant : Simon de Givry

Thématique: programmation linéaire, heuristique, optimisation combinatoire, agronomie.

Équipe d'accueil : Statistique et Algorithme pour la biologie.

Laboratoire d'accueil : Mathématique et Informatique Appliquées de Toulouse, Institut National de

Recherche Agronomique.

Le problème étudié est celui de la reconstruction automatique de la structure d'un réseau bayésien a partir d'observations. On suppose les variables aléatoires entièrement observées pour un ensemble de réalisations et la difficulté est de retrouver la structure qui maximise la vraisemblance pénalisée des observations.

Il s'agit d'un problème d'optimisation combinatoire NP-dur avec un espace de recherche constitué d'un nombre exponentiel de graphes dirigés sans circuit.

Récemment d'important gains de performance ont été obtenus en limitant le nombre de parents potentiels et en exploitant la taille de l'échantillon pour pré-évaluer un ensemble de configurations possibles de parents pour chaque variable aléatoire qui soit de taille raisonnable. Cette technique dite de cache a ouvert la voie a plusieurs approches d'optimisation dont la programmation dynamique (Silander et al, 2006), la programmation linéaire en nombre entiers (Barlett et Cussens, 2013) et la logique propositionnelle (Berg et al, 2014).

On se basant sur l'article de (Cussens, 2012) on a créé un programme avec le langage C a comme but la création des fichiers WCNF qui contiennent des clauses *Family* pour les liens de parenté entre les variable, les clauses *Ordoring* pour l'ordre de chaque variable, et des clauses pour interdire les cycles dans le réseau, afin de les passer par des solveur (Toulbar2, Eva500a, Akmaxsat, MaxHS) et trouver les résultats optimum de chaque réseau,

Pour réduire le nombre des clauses dans les fichier et amélioré la vitesse des solveurs on a travaillé avec les composantes fortement connexes en écartant les variables qui ne sont pas relier dans le réseau qui a réduit énormément le nombre des clause qui interdit les cycles

MARTELLETTI Andrea

MASTER 2 DROIT DE L'IMMATERIEL ET DES TECHNOLOGIES DE L'INFORMATION, PARCOURS DROIT ET INFORMATIQUE Année universitaire 2014/2015



Période du 02/04/2015 au 30/09/2015 (6 mois)

Tuteur de stage : Mme GANDON Nathalie

THEME GLOBAL DU STAGE:

Mise en place d'un site d'information sur la propriété intellectuelle numérique incluant la création de l'information structurée sur l'open data et les données personnelles.

THEME TRAITE JUSQU'EN JUILLET:

Guide juridique sur l'ouverture des données de la recherche (Open data).

RESUME:

« La recherche publique a pour objectifs le développement et le progrès de la recherche dans tous les domaines de la connaissance ; la valorisation des résultats de la recherche ; le partage et la diffusion des connaissances scientifiques ; le développement d'une capacité d'expertise et la formation à la recherche et par la recherche. »¹ Pour satisfaire ces objectifs, et sous l'impulsion de la Commission européenne, les établissements publics et semi-publics sont invités à répondre aux enjeux de l' « Open science » et du « Big data » en contribuant activement aux progrès scientifiques et à l'innovation au travers du partage et de la réutilisation des données issues de ses recherches.

Ce guide a pour vocation d'accompagner les agents des établissements concernés dans leur démarche d'ouverture des données issues de la recherche. Il ne présente pas un panorama exhaustif de la réglementation applicable mais tente de répondre aux questions pratiques les plus courantes auxquelles les agents pourront être confrontés.

¹ Article L112-1 du Code de la recherche