Thèse en science des données pour l'agronomie

Modélisation et visualisation des liens entre cinétiques de variables agro-environnementales et qualité des produits dans une approche parcimonieuse Bayésienne

Centre INRA de MONTPELLIER

Ecole doctorale Information Structures Systèmes (I2S Montpellier)

Directeurs de thèse :

Bénédicte Fontez (UMR MISTEA) benedicte.fontez@supagro.fr

Nadine Hilgert (UMR MISTEA) nadine.hilgert@inra.fr

Thierry Simoneau (UMR LEPSE)

Etablissement : INRA Montpellier

Pour postuler, envoyer : CV, lettre de motivation, moyenne des notes du M2, classement du M2, dès que possible à Nadine Hilgert. Démarrage prévu avant fin 2017.

Financement : 50% Institut de Convergence #DigitAg – 50% INRA

Compétences demandées pour le doctorant :

Le/la candidat/e devra posséder des compétences en statistiques de niveau Master 2 et un goût réel et prononcé pour les applications en agronomie. Il/elle devra être capable de proposer, formaliser puis mettre en œuvre des méthodes d'analyses statistiques innovantes et adaptées au contexte d'application. Pour cela, il/elle doit être familiarisé avec le langage de programmation du logiciel statistique R et les packages utilisés en data science. Il/elle devra posséder des compétences et un attrait particulier pour les statistiques appliquées. Il/elle devra être capable de rédiger des articles/communications scientifiques mais également de proposer des articles/présentations vulgarisées. Il/Elle devra faire preuve de rigueur et d'autonomie. La capacité à dialoguer et travailler avec des chercheurs d'autres domaines (statistiques, agronomie, biologie ...) sera un plus.

Résumé

L'agriculture est caractérisée par un savoir-faire important et ancestral dans les pratiques. Par exemple, dans la filière « Vigne et Vin », les décisions à la vigne reposent essentiellement sur des approches construites sur l'expertise (qualitatives). Face aux enjeux actuels de compétitivité, les acteurs de filières agricoles sont en forte demande d'outils quantitatifs de conseil et d'aide à la décision. L'objectif de cette thèse est de proposer une méthode pour extraire de la connaissance de

grandes masses de données hétérogènes et incertaines issues de processus temporels, et de développer des outils pour expliquer ou prédire la qualité d'un produit ou d'une production. L'intégration de toutes les informations disponibles (capteurs, expertises, données observées ou issues de modèles) en tenant compte de la fiabilité associée à chaque source est un enjeu qui nécessite de rénover les outils statistiques d'analyse des données pour tenir compte de l'incertitude et de les coupler avec les outils et approches informatiques. Le doctorant sera en charge du développement des méthodes et de leur application notamment à la filière « Vigne et Vin », en lien avec des partenaires privés (entreprises de conseils de la filière vigne et vin : ITK, Fruition Sciences) et publics (UMR LEPSE, UMR SPO, Institut Français du Vin).

Modelling and viewing relations between agrienvironmental time courses and product quality using a parsimonious Bayesian approach.

Résumé Anglais

Traditional knowledge plays an important role in agricultural practices. For instance, in the vine and wine food chain, decisions taken in vineyards mainly rely on expert knowledge-based approaches. Confronted with new challenges, stakeholders in agricultural production chains need advanced quantitative-based decision support tools. The aims of this PhD are i) to propose a knowledge discovery method to deal with big data from time courses, ii) to explain and predict product quality. Data integration should deal with high resolution data from sensors or agronomic models, low resolution observations and expert knowledge. It requires taking into account the reliability of all sources and data uncertainties. This calls for a coupling between informatics and data analysis, and constitutes the core of the PhD. The main application concerns the vine and wine food chain, in close relation with industrial partners (consulting professionals: ITK, Fruition Sciences, technical institute IFV) and public research laboratories (Joint Units LEPSE and SPO).

Question scientifique abordée :

La question de recherche proposée au candidat est celle de la mise au point de méthodes d'extraction d'information interprétable avec un modèle variables latentes pour données temporelles hétérogènes et incertaines. Il s'agira d'intégrer toutes les incertitudes et de choisir l'espace de dimension réduite qui permet d'extraire l'information pertinente et sa fiabilité, pour expliquer ou prédire la qualité d'un produit ou d'une production.

Démarche :

Une application principale reprendra les données du projet européen Innovine, dont le LEPSE et SPO sont partenaires. Les données acquises au cours de trois expérimentations conduites au vignoble permettent de mettre en relation de façon dynamique pendant la phase de maturation, les caractéristiques micro-environnementales (température de la baie et rayonnement intercepté mesuré en continu) et compositionnelles (°Brix, acidité, teneur en anthocyanes) de la baie dans une large gamme de conditions.

Le doctorant sera en charge du développement de méthodes et de leur application en agronomie. Il sera supervisé par l'ensemble des partenaires du projet, MISTEA pour la partie méthodologique, et LEPSE et SPO pour la partie agronomie – écophysiologie de la vigne. Les partenaires privés régionaux (entreprises de conseils : ITK, Fruition Sciences) ou publics (Institut Français du Vin) interviendront pour partager leur expertise et leurs données.

Les données de cinétiques (température du micro-climat, accumulation des sucres et anthocyanes dans les baies...) sont hétérogènes et incertaines (capteurs, expertises, données observées ou issues

de modèles). Cela nécessite de rénover les outils statistiques d'analyse des données pour tenir compte de l'incertitude et de les coupler avec les outils et approches informatiques de visualisation. On souhaite extraire de la connaissance pour expliquer ou prédire la qualité d'un produit ou d'une production. Nous envisageons une réduction de la dimension par une approche de type modèle à facteurs latents (Asmann et al 2014, dans une version probabiliste Tipping et Bishop (1999), dans une version Bayésienne Rowe (2000), ou dans un cadre parcimonieux West (2003)). Cette approche, utilisée en agronomie, permet de visualiser les proximités entre individus et de construire des indices synthétiques des variables de départ (ou composantes principales). Mais, l'incertitude sur la position d'un individu et sur l'espace réduit où les données seront projetées est en revanche moins étudiée et comprise. Or, la prise en compte de l'incertitude sur les individus est primordiale en agronomie où les données sont issues de capteurs (fiabilité) ou reconstruites à partir de modèles et d'expertises. De même, l'incertitude sur les indices synthétiques et leur signification doivent être prises en compte pour éviter des interprétations erronées.

Des premiers travaux existent pour traiter l'incertitude sur les individus avec les travaux de Diday - Billard sur les données d'intervalle. L'approche décrite dans Billard *et al.* (2009) fournit des visualisations grossières et, pour pallier ce défaut, des pondérations sur les individus ou les variables sont proposées en fonction de leur incertitude. Nos premiers travaux (G. Weinrott *et al.*, 2016) montrent les limites et les dangers de ces pondérations sur des données agronomiques. Un certain nombre de points sont encore à travailler comme : l'estimation de la dimension de l'espace réduit, le fait que les facteurs sont connus à une rotation près.

Références bibliographiques

Asmann, C., Boysen-Hogrefe, J., & Pape, M. (2014). Bayesian analysis of dynamic factor models: an ex-post approach towards the rotation problem. Kiel Institute, working paper n°1902.

Billard,L., Douzal-Chouakria,A. & Diday, E. (2009). Symbolic principal component for interval-valued observations. HAL.

Tipping, M.E., and Bishop, C.M. (1999). Probabilistic principal component analysis. Journal of the royal statistical society, 61, 611-622.

Rowe, D. B. (2000). Bayesian factor analysis model with generalized prior information. Social science working paper, 1099.

Weinrott G., Fontez B., Hilgert N., Holmes S. (2016). Modèle Bayésien à facteurs latents pour l'analyse de données fonctionnelles, 48ème Journées de Statistique (SFdS), Montpellier.

West, M. (2003). Bayesian factor regression models in the « large p, small n » paradigm. Bayesian statistics, 7.