Université Laval - 25-26 mai 2017

- Deux jours d'ateliers et de conférences
- Un colloque interdisciplinaire regroupant des intervenants d'une dizaine de disciplines
- À Québec! En français!

« Le premier grand colloque annuel interdisciplinaire et francophone dédié à R en Amérique du Nord! »

raquebec.ulaval.ca

Programme disponible en ligne

Jeudi 25 mai – Ateliers

	Salle 2326 – 2e étage	Le Cercle A – 4e étage	Le Cercle B – 4e étage
9h00	**Atelier d'introduction à R** (1re partie) Vincent Goulet	**Programmation avancée en R – Fonctions avancées et recherche reproductible** Marc Mazerolle	**Visualisation de données en R** Sophie Baillargeon
12h00	**Dîner** - Espace Jardin		
13h00	**Atelier d'introduction à R** (2e partie) Vincent Goulet	**Visualisation de données en R** Sophie Baillargeon	**Programmation avancée en R – Fonctions avancées et recherche reproductible** Marc Mazerolle
17h00	**Cocktail dînatoire** - Atrium		

Vendredi 26 mai – Conférences

8h30	**Mot de bienvenue** - Salle Hydro-Québec – 2e étage
8h45	Conférence principale **Extending R with C++: Motivation and Examples** Dirk Eddelbuettel
9h45	Pause - Atrium

	Salle Hydro-Québec – 3e étage **Aller plus loin avec R** Modérée par **Charles Fleury**	Le Cercle – 4e étage **Simulation et biostatistique** Modérée par **Alexandre Bureau**
10h00	**Le projet R dans Google Summer of Code** Toby Dylan Hocking	**Analyse de survie avec expositions qui varient dans le temps: simulation et modèles d'exposition cumulative pondérée** Marie-Pierre Sylvestre
10h30	**Innovation analytique: utilisation de R chez Desjardins Assurance** Guillaume Lepage	**Oecologia in silico : l'utilisation de simulations pour comprendre les processus écologiques** - Louis Donelle et Sylvie Clappe
11h00	**R en entreprise: statistiques, marketing et analyse géospatiale** François Pelletier	**Construction et validation d'un modèle de micro-simulation pour étudier la sensibilité du dépistage du cancer du sein par mammographie** - Nathalie Vandal
11h30	**Annonces de politiques monétaires: allier R et recherche reproductible pour une analyse linguistique longitudinale** William Sanger	**Sélection automatique de variables confondantes avec l'algorithme Bayesian causal effect estimation** Denis Talbot
12h00	**Dîner** - Espace Jardin	

	Salle Hydro-Québec – 3e étage **Analyses multivariée et bayésienne** Modérée par **Marc Mazerolle**	Le Cercle – 4e étage **Apprentissage automatique et analyse spatiale** Modérée par **Anne-Sophie Charest**
13h00	**Analyse des correspondances du taxi** Vartan Choulakian et Jacques Allard	**Couper des arbres avec R : combiner l'apprentissage automatique et les capacités géospatiales de R pour mesurer le risque d'inondation en assurance** - Frédérick Guillot et Étienne Larrivée-Hardy
13h30	**Un guide pratique du package vegan pour les analyses en écologies des communautés** - Guillaume Blanchet	**Introduction au deep learning avec MXNET** Jérémie Desgagné-Bouchard
14h00	**Utilisation des packages lavaan et psych de R appliqué à la validation psychométrique d'un questionnaire d'abus de substances (DAST)** - Charles-Édouard Giguère	**Données spatiales: sf, le nouveau spatial** Étienne Racine
14h30	**Comment évaluer l'évidence pour théories scientifiques avec R** Robert van Hulst	**Détection de menaces par classification de paquets de pare-feu** Mohamed Dahmane
15h00	Pause - Atrium	

	Salle Hydro-Québec – 3e étage **Développement avancé** Modérée par **Vincent Goulet**	Le Cercle – 4e étage **Analyse de réseaux** Modérée par **Arnaud Droit**
15h30	**Méthode de programmation de fonctions renvoyant des fonctions : application au package SPImficmcm et sa Shiny App** Molière Nguile-Makao	**Les défis de l'analyse des réseaux dynamiques : un exemple de la co-délinquance** Yanick Charette
16h00	**Modélisation financière à l'aide de la programmation orientée objet avec Rcpp** - Denis-Alexandre Trottier	**PathQuant, un package R pour l'annotation de données pan-génomiques combinées à la métabolomique** - Sandra Therrien-Laperrière
16h30	**Les Shiny App ou comment rendre les applications R plus accessibles** - Aurélien Nicosia	**alien: un package R pour modéliser les interactions entre espèces** - Steve Vissault
17h00	**Data.table: un incontournable!** Jean-Philippe Le Cavalier	
17h30	**Mot de la fin – Prix de présence – Salle Hydro-Québec**	

Université Laval - 25-26 mai 2017

- Deux jours d'ateliers et de conférences
- Un colloque interdisciplinaire regroupant des intervenants d'une dizaine de disciplines
- À Québec! En français!

raquebec.ulaval.ca

« Le premier grand colloque annuel interdisciplinaire et francophone dédié à R en Amérique du Nord! »

Formateurs

Sophie Baillargeon	Étudiante au doctorat en statistique et chargée de cours au Département de mathématiques et de statistique de l'Université Laval
Vincent Goulet	Professeur à l'École d'actuariat de l'Université Laval
Marc Mazerolle	Professeur de biologie de la conservation au Département des sciences du bois et de la forêt de l'Université Laval

Conférenciers

Dirk Eddelbuettel	**Conférencier principal** **Un des principaux contributeurs de R et auteur du livre « Seamless R and C++ Integration with Rcpp » dans la série « UseR! » de Springer**		
Jacques Allard	Professeur au Département de mathématiques et de statistique de la Faculté des sciences de l'Université de Moncton	**Guillaume Blanchet**	Stagiaire postdoctoral au département de biologie de l'Université de Sherbrooke
Yanick Charette	Professeur adjoint à l'École de service social de la Faculté des sciences sociales de l'Université Laval	**Sylvie Clappe**	Étudiante au doctorat à l'Université Claude Bernard Lyon 1
Jérémie Desgagné-Bouchard	Analyste en actuariat chez Intact Corporation Financière	**Vartan Choulakian**	Professeur au Département de mathématiques et de statistique de la Faculté des sciences de l'Université de Moncton
Mohamed Dahmane	Chercheur, Équipe Vision et imagerie au Centre de recherche informatique de Montréal	**Louis Donelle**	Étudiant à la maîtrise au département de biologie de l'Université Concordia
Charles-Édouard Giguère	Analyste statisticien au Centre de recherche de l'Institut universitaire en Santé Mentale de Montréal	**Frédérick Guillot**	Directeur recherche et analytique chez Co-operators et chargé de cours à l'École d'actuariat de l'Université Laval
Toby Dylan Hocking	Consultant et stagiaire postdoctoral en bioinformatique au Département de génétique humaine à l'Université McGill	**Jean-Philippe Le Cavalier**	Analyste principal en actuariat chez Promutuel Assurance et chargé de cours à l'École d'actuariat de l'Université Laval
Guillaume Lepage	Analyste de données scientifiques chez Desjardins Assurances	**Molière Nguile-Makao**	Prof. de recherche en biostatistique au Centre de recherche clinique et évaluative en oncologie (CRCEO) de Québec
Aurélien Nicosia	Statisticien chez GSK et étudiant au doctorat à l'Université Laval	**François Pelletier**	Analyste en bases de données actuarielles chez Desjardins Assurances
Etienne Racine	Analyste de données chez Intact Lab et étudiant au doctorat en foresterie à l'Université Laval	**Marie-Pierre Sylvestre**	Biostatisticienne et professeure adjointe au Département de médecine sociale et préventive de l'Université de Montréal
Denis Talbot	Professeur adjoint au Département de médecine sociale et préventive de l'Université Laval	**Sandra Therrien-Laperrière**	Étudiante à la maîtrise au Département de biochimie et médecine moléculaire de l'Université de Montréal.
Étienne Larrivée-Hardy	Analyste en recherche et innovation chez Co-operators et chargé de cours à l'École d'actuariat de l'Université Laval	**Denis-Alexandre Trottier**	Étudiant au doctorat en finance et assurance à la Faculté des sciences de l'administration de l'Université Laval
Nathalie Vandal	Statisticienne à la Direction de l'analyse et de l'évaluation des systèmes de soins et services de Institut national de santé publique du Québec	**Robert van Hulst**	Professeur au Département de biologie de l'Université Bishop's
Steve Vissault	Professionnel de recherche pour la Chaire de recherche du Canada en Écologie intégrative à l'Université de Sherbrooke	**William Sanger**	Directeur de projet au Centre interuniversitaire de recherche en analyse des organisations (CIRANO)

Partenaires Or

Université Laval – Faculté des sciences et de génie – Chaire d'actuariat

RStudio®

Centre de recherches mathématiques

Partenaires Argent

Université Laval – Faculté de foresterie, de géographie et de géomatique

Société Statistique du Canada / Statistical Society of Canada

Partenaires Bronze

cef – Centre d'étude de la forêt

Desjardins Assurances

Association des statisticiennes et statisticiens du Québec

PROMUTUEL ASSURANCE

Université Laval – Faculté des sciences sociales

Université Laval - 25-26 mai 2017

- Deux jours d'ateliers et de conférences
- Un colloque interdisciplinaire regroupant des intervenants d'une dizaine de disciplines
- À Québec! En français!

« Le premier grand colloque annuel interdisciplinaire et francophone dédié à R en Amérique du Nord! »

Résumés des conférences du vendredi

8h45 - Conférence principale

Extending R with C++: Motivation and Examples - Dirk Eddelbuettel

Le calcul numérique joue maintenant un rôle central en statistique, et ce, tant en recherche que dans les applications pratiques. À cette fin, R est maintenant reconnu comme la lingua franca de la statistique. Cela dit, comme le mentionne Chambers (2016), il existe toujours des raisons d'étendre le rayon d'action de R. Dans cette conférence, nous nous pencherons sur le package Rcpp, qui est devenu au fil du temps l'outil le plus fréquemment employé pour ainsi étendre R. Nous expliquerons les motivations derrière le package et nous expliquerons sommairement son fonctionnement. Nous étudierons divers cas, des plus simples où des fragments de code R se voient remplacés par du code C++ plus performant, à des cas plus complexes où l'on ajoute aux fonctionnalités de R en tirant profit de bibliothèques de calcul offrant des interfaces C/C++.

Aller plus loin avec R – Salle Hydro-Québec – 3e étage

10h00 - Le projet R dans Google Summer of Code - Toby Dylan Hocking

Google Summer of Code (GSOC) a pour but d'enseigner la programmation des logiciels libres aux étudiants des universités du monde entier. Le projet R participe depuis 2008. Je suis co-administrateur depuis 2012 et mentor depuis 2013. Chaque été, à l'aide des mentors bénévoles, une vingtaine d'étudiants travaillent sur des packages R. Dans cette présentation, je vais expliquer l'histoire de R dans GSOC et comment participer (en tant que mentor ou étudiant).

10h30 - Innovation analytique: utilisation de R chez Desjardins Assurance - Guillaume Lepage

L'utilisation de R au sein des grandes entreprises est relativement récente. Dans cette présentation, nous détaillerons comment R peut être utilisé au sein d'une organisation plus habituée aux solutions propriétaires. On présentera les différents concepts aidant à cette diffusion, que ce soit au niveau de la manipulation de données avec les packages du tidyverse, de la modélisation de volumes de données importants en utilisant des solutions info-nuagiques ou de la diffusion de résultats avec Shiny. L'objectif est de montrer qu'avec les nouveaux outils développés autour de R, il devient de plus en plus aisé d'imposer R comme une alternative performante ou un complément aux logiciels propriétaires, y compris au sein d'une industrie traditionnellement conservative comme l'assurance. On présentera des cas concrets de projets réalisés entièrement avec R.

11h00 - R en entreprise: statistiques, marketing et analyse géospatiale - François Pelletier

Cet atelier présentera un survol des usages de R dans le secteur financier et des assurances. Les domaines explorés seront les statistiques, le marketing et l'analyse géospatiale. R permet de partager l'expertise et de favoriser la collaboration entre les différents secteurs d'affaires, dont la gestion des risques, le marketing et les opérations. Il permet aussi une intégration rapide des nouveaux employés aux projets existants en ne nécessitant pas l'apprentissage d'une technologie propriétaire.

11h30 - Annonces de politiques monétaires: allier R et recherche reproductible pour une analyse linguistique longitudinale - William Sanger

Nous nous intéressons à l'utilisation d'algorithmes en R pour la caractérisation systématique des discours (203) et des réponses (3501) apportées par les trois présidents successifs de la Banque centrale européenne (BCE). Inaugurée en 1999, la BCE a pour mission de gérer la politique monétaire au sein de la zone euro. Avec la crise économique de 2008, l'objectif d'inflation qui prévalait auparavant va être remplacé par l'objectif d'absorption de la crise ainsi que de la relance de l'économie avec le dernier président Mario Draghi en particulier. Chaque mois, le Président de la BCE rapporte à la presse les discussions du Conseil des Gouverneurs. L'impact de ces discours a été illustré dans la littérature à plusieurs niveaux, notamment sur les marchés financiers. Ces présentations sont structurées en deux parties: une première décrivant les annonces de la BCE et une seconde où le Président répond aux questions des journalistes. De par l'ampleur des communications, le nombre de réponses apportées et la fréquence des discours, ancrer notre recherche dans le cadre de la recherche reproductible avec R s'avère essentiel. En effet, la base de données complète depuis 1998 équivaut à plus de 900 000 mots, et augmente chaque mois. À partir d'une plateforme de sciences de données utilisant R (Nüance-R), ces algorithmes nous permettent d'effectuer la collecte des données (gsheet), la structuration de la base de données (RTextTools, tm, tidyr), de déterminer la polarité et la teneur lexicale des différentes communications (sentiment, ggplot2) et l'intégration des résultats (rmarkdown) dans un processus reproductible.

Simulation et biostatistique – Le Cercle – 4e étage

10h00 - Analyse de survie avec expositions qui varient dans le temps: simulation et modèles d'exposition cumulative pondérée - Marie-Pierre Sylvestre

Les études épidémiologiques évaluent fréquemment l'effet d'expositions complexes dont le statut et l'intensité varient avec le temps. L'analyse de ces études pose un défi particulier, celui de modéliser l'association entre ces expositions complexes et le risque, particulièrement lorsque la pertinence étiologique des expositions prises lors de différentes périodes de temps est incertaine. Je présenterai deux ensembles de routines en R qui sont particulièrement utiles pour les études longitudinales. Tout d'abord, je présenterai le modèle d'exposition cumulative pondérée (Sylvestre et Abrahamowicz, 2009, package WCE, 2014). Ce modèle estime non seulement l'association entre les expositions qui varient dans le temps et le risque d'événements, mais produit aussi une courbe de pondération qui indique l'importance relative de chaque exposition encourue dans le passé sur la probabilité d'un événement. Ce modèle a été validé sur des données simulées et réelles. En deuxième lieu, je présenterai l'outil PermAlgo qui a servi à simuler des données de survie complexes pour valider WCE (Sylvestre et Abrahamowicz 2008). J'expliquerai comment PermAlgo peut être utilisé pour générer des jeux de données longitudinaux plausibles en recherche médicale. L'algorithme permutationnel est un des algorithmes de simulation de données de de survie les plus rapides et polyvalents présentement disponible sur R.

10h30 - Oecologia in silico : l'utilisation de simulations pour comprendre les processus écologiques - Louis Donelle et Sylvie Clappe

Durant les deux dernières décennies, la place des simulations par ordinateur en écologie a connu une progression fulgurante. Nous proposons ici un survol des méthodes de simulation les plus utilisées en écologie, ainsi que de leur implémentation et utilisation dans R, en partant des modèles « multi-équilibre » à l'échelle du paysage pour aller jusqu'aux modèles basés sur l'individu. Nous définirons également le domaine d'applicabilité et les types de questions auxquelles ces méthodes répondent le plus fréquemment. L'utilisation de simulations par ordinateur s'avère une approche intéressante pour comprendre les processus à l'œuvre dans les cas où la complexité des systèmes écologiques est telle que l'observation ou l'expérimentation ne sont pas en mesure de tester ou d'identifier les processus sous-jacents à un phénomène. Ces méthodes de simulations peuvent généralement s'adapter et être utiles à d'autres disciplines devant également composer avec la complexité de leurs systèmes d'étude. Outre leur réplicabilité et leur faible coût, les études in silico comportent également des avantages très importants. En effet, simuler un système permet de connaître et contrôler les processus que l'on souhaite étudier en modifiant les paramètres utilisés. Par ailleurs, la possibilité de connaître l'état de chaque variable à n'importe quel moment de la simulation favorise une étude et une compréhension plus détaillée du système. Ensuite, si les études in silico peuvent servir à mieux comprendre les processus sous-jacents à un phénomène, elles s'avèrent également très utiles pour développer et évaluer l'efficacité des méthodes statistiques.

11h00 - Construction et validation d'un modèle de micro-simulation pour étudier la sensibilité du dépistage du cancer du sein par mammographie Nathalie Vandal

La vraie sensibilité du dépistage par mammographie ne peut être directement mesurée puisque le nombre de femmes avec un cancer du sein en période préclinique au moment du dépistage, dans une population donnée, est inconnu. Afin d'étudier la vraie sensibilité, un modèle de micro-simulation a été construit dans R. Ce modèle consiste à bâtir une cohorte fictive de femmes pour lesquelles certaines seront atteintes d'un cancer du sein. Cette cohorte de femmes pourra ensuite être soumise à un ou plusieurs dépistages par mammographie. Certains paramètres du modèle peuvent être déterminés à partir de données existantes (par ex. l'incidence du cancer du sein, l'âge au décès), tandis que d'autres paramètres demeurent inconnus (par ex. la sensibilité du dépistage, la durée de la période préclinique du cancer) et doivent être déterminés par calibration. La calibration est un processus laborieux qui demande la réalisation d'un grand nombre de simulations et qui prend beaucoup de temps à réaliser. L'utilisation du package snowfall a permis de réduire considérablement les temps de calcul. Le choix des paramètres inconnus a été basé sur la comparaison de 66 indicateurs à des valeurs cibles dérivées des données du Programme québécois de dépistage du cancer du sein (PQDCS). L'utilisation de représentations graphiques et une analyse en composantes principales des meilleurs ensembles de paramètres a permis de raffiner le processus de calibration. Dans cette présentation, nous présenterons les principales étapes de la construction du modèle de micro-simulation, dont certaines embûches, ainsi que les solutions retenues.

11h30 - Sélection automatique de variables confondantes avec l'algorithme Bayesian causal effect estimation - Denis Talbot

L'estimation de l'effet d'une exposition à l'aide de données observationnelles nécessite habituellement le contrôle pour des variables confondantes. La sélection de ces variables peut cependant être une tâche difficile ; le contrôle pour trop de variables peut réduire la puissance statistique et un contrôle insuffisant peut engendrer un biais. L'utilisation de méthodes de sélection de variables basées sur les données peut être une solution à ce problème. Toutefois, la majorité des approches classiques de sélection de variables peut introduire un biais dans les estimations en plus de produire des intervalles de confiance avec un taux de couverture inférieur à celui désiré. L'algorithme Bayesian causal effect estimation (BCEE) est une approche bayésienne qui permet de tenir compte de l'incertitude associée à la sélection des variables confondantes et ainsi de produire des inférences appropriées. Intuitivement, BCEE cherche à favoriser les modèles effectuant un contrôle suffisant pour les variables confondantes associées simultanément à l'exposition et à l'issue. Afin d'améliorer la puissance statistique, BCEE vise également à éviter de contrôler pour les variables qui ne sont qu'uniquement associées à l'exposition. Cette présentation illustrera l'application de BCEE pour l'estimation de l'effet du tabagisme sur la tension artérielle systolique en utilisant des données du Framingham Heart Study. Les étapes importantes de l'analyse effectuée avec le package R BCEE seront présentées. Nous verrons que l'estimation de l'effet obtenue sans effectuer de contrôle est insensée alors que l'estimation produite en contrôlant pour toutes les variables est plausible, mais imprécise. BCEE permet d'obtenir un résultat vraisemblable et plus précis.

Université Laval - 25-26 mai 2017

- Deux jours d'ateliers et de conférences
- Un colloque interdisciplinaire regroupant des intervenants d'une dizaine de disciplines
- À Québec! En français!

« Le premier grand colloque annuel interdisciplinaire et francophone dédié à R en Amérique du Nord! »

Analyses multivariée et bayésienne – Salle Hydro-Québec – 3e étage

13h00 - Analyse des correspondances du taxi
Vartan Choulakian et Jacques Allard
Analyse des correspondances du taxi (TCA) est une méthode robuste d'analyse des correspondances (CA) pour visualiser un tableau de contingence. Nous présenterons la théorie mathématique sous-jacente et le package TCA-R, que nous avons développé récemment. Nous montrerons que, sur des tableaux de contingences éparses, TCA produit des cartes plus interprétables que CA.

13h30 - Un guide pratique du package vegan pour les analyses en écologies des communautés
Guillaume Blanchet
Le package R vegan offre plusieurs outils mathématiques et statistiques pour répondre à diverses questions en écologies des communautés. En plus des efforts importants qui ont été investis dans les ordinations simples et canoniques ainsi que les tests par permutations, les fers de lances de ce package, vegan offre la possibilité de faire de nombreuses autres analyses. vegan offre des outils permettant de faire de l'exploration de données, des estimations de richesses d'espèces (nombre d'espèces par site d'échantillonnage), du partitionnement de la diversité, de la modélisation spatiale en plus de proposer de nombreuses façons de construire des modèles nuls. De plus, un effort soutenu a été investi pour s'assurer que les nouvelles versions de vegan soient rétrocompatibles, permettant ainsi à l'utilisateur moyen de travailler avec vegan aisément même avec la version la plus récente du package. Dans cette présentation, j'introduirai vegan en utilisant une approche de questions/réponses permettant ainsi de décrire le type de questions écologiques qui peuvent être approchées par vegan en plus d'illustrer comment vegan peut être utilisé pour répondre à ces différentes questions. Toutes les problèmes étudiés dans cette présentation seront résolus en utilisant exclusivement les fonctions et données disponibles dans le package vegan. Pour conclure cette présentation, je vais discuter de ce qui est prévu pour le futur à moyen et long terme pour le package vegan.

14h00 - Utilisation des packages lavaan et psych de R appliqué à la validation psychométrique d'un questionnaire d'abus de substances (DAST)
Charles-Édouard Giguère
Les qualités psychométriques d'un instrument sont essentielles à étudier puisqu'elles nous confirment que les inférences obtenues en analysant les réponses à cet instrument sont valides et cohérentes d'un groupe à l'autre. Dans le présent projet, le Drug Abuse Screening Test (DAST) a été validé en utilisant l'environnement R et plus particulièrement les packages lavaan et psych sur les données de 912 participants de la banque Signature qui ont été recrutés à l'urgence psychiatrique de l'Institut universitaire en santé mentale de Montréal. Cette conférence vise à montrer comment à l'aide du package psych nous avons analysé la cohérence interne de l'instrument. Ce package permet d'aller au-delà d'une simple mesure d'alpha de Cronbach et offre toute une série d'indicateurs ainsi que la possibilité de calculer des intervalles de confiance asymptotiques ou obtenus par rééchantillonnage. Une analyse factorielle des items dichotomiques (lien probit) de l'instrument a aussi été effectuée en utilisant le package lavaan. Ce dernier permet également de réaliser des analyses multi-groupes nous ayant permis de tester l'invariance de l'instrument selon le sexe. Des analyses tests-retest ont été effectuées sur les patients ayant répondu à l'instrument dans un délai de 30 jours ou moins (r=0.86). Finalement, le DAST a montré une bonne qualité de prédiction des patients souffrant de troubles d'utilisation de substances évalués indépendamment par des psychiatres en faisant une analyse des courbes ROC. R s'est avéré un environnement très riche nous permettant de confirmer les excellentes propriétés du DAST dans le contexte d'une urgence psychiatrique.

14h30 - Comment évaluer l'évidence pour théories scientifiques avec R
Robert van Hulst
Les notions statistiques de base sont réputées d'être mal comprises par la majorité des scientifiques — et c'est vrai que beaucoup de ces notions en statistique conventionnelle ne sont pas du tout intuitives. Pour cette raison et aussi pour sa plus grande polyvalence, l'utilisation de la statistique bayésienne continue d'augmenter. Ces principes peuvent être enseignés sans trouble aux débutants parce qu'ils sont plus limpides que ceux de la statistique fréquentiste. Malheureusement, le système R de base ne contient presque pas de fonctions bayésiennes, bien qu'il existe beaucoup de packages bayésiens. J'en ai écrit un autre : l'évaluation d'évidence scientifique ou Évidence. L'approche fréquentiste étant plus orientée vers les tests d'hypothèses, l'approche bayésienne met plutôt l'accent sur la modélisation — comme d'ailleurs la statistique fréquentiste moderne. Au lieu de tests d'hypothèses, le package favorise des graphiques et la modélisation. Par exemple, le package contient des fonctions pour l'analyse de proportions, de tableaux de contingence et de modèles linéaires simples. Pour des modèles plus complexes utilisant MCMC, rstan et rstanarm sont utilisés. J'ai écrit un livre de cours pour accompagner ce package. Ces approches redonnent à l'analyste un rôle actif plutôt que celui d'un utilisateur d'un livre de recettes. Il est devenu urgent de moderniser notre enseignement des principes de l'inférence statistique pour la recherche scientifique. C'est mon espoir que ce genre d'efforts puissent aider à maintenir la place importante que R s'est acquise comme outil essentiel pour l'analyse statistique.

Apprentissage automatique et analyse spatiale – Le Cercle – 4e étage

13h00 - Couper des arbres avec R : combiner l'apprentissage automatique et les capacités géospatiales de R pour mesurer le risque d'inondation en assurance
Frédérick Guillot et Étienne Larrivée-Hardy
Nous présenterons un cas où les multiples capacités de R, des données massives sur l'environnement et des images satellites nous ont permis de développer un modèle de risque pour un produit d'assurance inondation. La présentation permet d'entrevoir l'étendue des possibilités de R pour l'industrie de l'assurance, notamment la facilité de combiner l'apprentissage automatique avec l'analyse géospatiale pour répondre à des problèmes appliqués.

13h30 - Introduction au deep learning avec MXNET
Jérémie Desgagné-Bouchard
Alors que de nombreux packages dédiés au deep learning ont émergé au cours des dernières années, peu d'options étaient offertes à l'environnement R. MXNET offre les fonctionnalités permettant le développement de modèles sophistiqués, qu'il s'agisse de problèmes de régression par Multi Layer Perceptron, de classification d'images par Convolutional Neural Network ou encore d'analyse séquentielle de texte par Recurrent Neural Network. La présentation se veut une introduction à l'architecture générale de MXNET, à ses principales composantes symboliques et aux utilitaires pour la construction des différentes familles de modèles et à leur réutilisation.

14h00 - Données spatiales: sf, le nouveau spatial
Étienne Racine
L'abondance de GPS rend les données spatiales de plus en plus abondantes. Le package sp supporte les données spatiales depuis le début des années 2000, avant l'introduction de standards d'interopérabilité. Ainsi, le support pour des géométries complexes ou multiples et le 3D ne sont pas supportées extensivement par sp puisqu'elles n'existaient pas. Le projet Simple Feature vise à introduire le format de données spatiales OGC qui sera plus compatible avec tidyverse et les autres outils comme PostGIS. Je présenterai le nouveau package sf et ses derniers développement par une étude de cas concrète.

14h30 - Détection de menaces par classification de paquets de pare-feu
Mohamed Dahmane
R rend accessible et simplifie l'application des algorithmes d'apprentissage dans plusieurs domaines. Cela s'applique également dans le domaine de la sécurité informatique. En effet, dans son livre Data Driven Security, J. Jays démontre comment adapter facilement en R des outils et des techniques éprouvés utilisés dans d'autres disciplines pour bâtir une adéquation adaptative et évolutive en sécurité des données. Étant donné que les approches conventionnelles ne parviennent pas à faire face au style adaptatif des cyberattaques, l'apprentissage machine présente un complément nécessaire à ces approches (pare-feu, outils d'authentification et réseaux privés virtuels). Par exemple, à partir du fichier journal du pare-feu, on sélectionne un ensemble de connexions avec lesquelles on construit facilement avec R un système permettant d'apprendre à reproduire l'ensemble des règles définies par le pare-feu sous forme de modèle arborescent (XGBoost de Chen, 2016). En mode fonctionnement, la réponse du pare-feu doit correspondre à la réponse du modèle sinon une alarme sera émise.

Université Laval - 25-26 mai 2017

- Deux jours d'ateliers et de conférences
- Un colloque interdisciplinaire regroupant des intervenants d'une dizaine de disciplines
- À Québec! En français!

« Le premier grand colloque annuel interdisciplinaire et francophone dédié à R en Amérique du Nord! »

Développement avancé – Salle Hydro-Québec – 3e étage

15h30 - Méthode de programmation de fonctions renvoyant des fonctions : application au package SPlmficmcm et sa Shiny App
Molière Nguile-Makao
Le langage de programmation R offre la possibilité de construire des fonctions qui renvoient des fonctions. Le principal avantage de cette technique est qu'elle permet d'avoir le contrôle sur les fonctions renvoyées (par exemple, évaluer une fonction de log-vraisemblance à différentes valeurs de ses paramètres pour effectuer des tests du rapport de vraisemblance). Nous avons utilisé cette technique pour développer le package SPmlficmcm disponible sur CRAN et sa Shiny App. Ledit package implémente une méthode d'estimation semi-paramétrique du maximum de vraisemblance proposée par Chen et al. (2012) ainsi que son extension proposée par Nguile-Makao et Bureau (2015) pour résoudre le problème d'analyse des effets d'interaction gène-environnement sur les risques de complications obstétriques dans un devis cas-témoins de couples mère-enfant prenant en compte dans l'estimation le lien parental entre la mère et son enfant. Le calcul de la log-vraisemblance se fait en deux étapes : 1) la construction et la résolution d'un système d'équation non-linéaire à partir des données, et 2) l'évaluation de la fonction de log-vraisemblance à partir de la solution du système non-linéaire et des données. Notre travail a été d'implémenter cette méthode en R avec la contrainte de contrôle sur le système non-linéaire et la fonction de log-vraisemblance. Programmer une fonction R qui renvoie la fonction de log-vraisemblance nous a permis de résoudre le problème à moindre coût de calcul.

16h00 - Modélisation financière à l'aide de la programmation orientée objet avec Rcpp
Denis-Alexandre Trottier
Notre conférence portera sur l'utilisation de la programmation orientée objet via Rcpp. Un exemple de modélisation financière où ces techniques sont très utiles sera présenté à titre de fil conducteur. Les participants seront néanmoins capables de suivre la présentation sans connaissance préalable en finance. L'utilisation de modèles à changement de régime est maintenant très répandue pour modéliser la dynamique des séries de rendements financiers. Il existe un nombre très élevé de spécifications possibles pour ce genre de modèle. En l'absence de techniques de programmation orientée objet, le programmeur doit fixer le nombre de régimes, le nombre de séries financières à modéliser, la spécification individuelle de chacune, ainsi que leur structure de dépendance. En pratique, il est souvent nécessaire de considérer un grand nombre de variations de ces différents aspects du modèle, par exemple lorsqu'on désire faire une étude comparative, ou bien lorsqu'on désire utiliser des techniques d'agrégation de différents modèles. L'exercice de programmation devient alors fastidieux, pouvant parfois nécessiter plusieurs milliers de lignes de code redondantes. Notre conférence illustrera comment ce problème peut être résolu par un programmeur R à l'aide de la programmation orientée objet avec Rcpp. Ces connaissances sont souvent très utiles en pratique, par exemple pour notre groupe de recherche dans le cadre d'un projet subventionné par l'Autorité des marchés financiers. Le présentateur est également co-auteur du package MSGARCH, qui utilise une approche similaire pour implémenter des modèles GARCH à changement de régime.

16h30 - Les Shiny App ou comment rendre les applications R plus accessibles
Aurélien Nicosia
Dans le cadre de mon travail au Service de consultation statistique, j'ai souvent affaire avec des gens qui veulent utiliser R pour faire leurs analyses, mais qui craignent de se lancer dans la programmation. Les Shiny App développées en R sont un bon moyen d'améliorer l'utilisation de tout l'écosystème généré autour de R dans une application web facile d'utilisation et partageable. Dans cette présentation, je vais présenter rapidement ce qu'est une Shiny App, les bases pour en créer une rapidement et je présenterai des cas concrets d'applications que j'ai créées qui fournissent des sorties aussi complètes que SAS le fait.

17h00 - Data.table: un incontournable!
Jean-Philippe Le Cavalier
La classe prédestinée à stocker des données structurées en R est le data.frame. Conformément à chaque classe développée dans le coeur de R, on ne peut accéder à un objet de classe data.frame directement en référence, on doit créer une copie de l'objet à chaque fois où on veut en modifier une partie et ensuite le réassigner. Introduite en 2006 par Matt Dowle, la classe data.table, provenant du package du même nom, est une extension du data.frame permettant de contourner cette obligation. En effet, un objet de classe data.table permet une modification de lui-même en référence, ce qui devient nécessaire lorsque le jeu de données utilisé s'approche de l'espace disponible en mémoire sur un poste de travail. De plus, un concept de clés, permettant d'effectuer une recherche par arbres binomiaux, a été introduit afin de rendre la jointure d'objets de classe data.table beaucoup plus efficace que la méthode utilisée d'emblée pour une jointure de deux objets de classe data.frame. Ma présentation vise à faire une introduction de la classe data.table à des utilisateurs qui ne l'ont jamais utilisée. Les concepts de base seront illustrés à l'aide d'exemples simples.

Analyse de réseaux – Le Cercle – 4e étage

15h30 - Les défis de l'analyse des réseaux dynamiques : un exemple de la co-délinquance
Yanick Charrette
Les fonctions R sont souvent gourmandes en mémoire vive. Ceci représente un défi important lors du traitement et de la modélisation de données massives. Lors de cette présentation, nous exposerons les stratégies utilisées pour manipuler un réseau social criminel dynamique de 181 615 individus arrêtés sur une période de huit ans. Nous aborderons premièrement l'opérationnalisation des mesures de la structure réseau supra dyadique, soit l'influence de la composition des interconnexions telles la centralité et la transitivité, réalisées à l'aide du package reshape2. Ensuite, nous aborderons l'analyse qui porte sur la récurrence temporelle des liens du réseau. Puisque les données sont censurées à droite, des modèles à risques proportionnels incluant des variables dynamiques furent privilégiés et réalisés à l'aide du package survival. Ces modèles considèrent la dépendance entre les observations (tant dyadique que temporelle). Étant donnée la taille des données dynamiques (10 809 140 points de mesures), un modèle utilisant l'ensemble de l'échantillon était trop exigeant en mémoire vive et en temps de calcul. Nous avons donc favorisé des itérations de sous-échantillonnage. Les modèles furent exécutés sur chacun des sous-échantillons en parallèle sur un super-ordinateur. Par la suite, les coefficients de chacun de ses modèles itérés furent regroupées (pooled) pour l'interprétation des résultats. Une discussion portera sur le fait qu'aucune stratégie optimale n'a pu être mise en place, sinon celle du compromis.

16h00 - PathQuant, un package R pour l'annotation de données pan-génomiques combinées à la métabolomique
Sandra Therrien-Laperrière
Avec l'avènement des sciences omiques dans le domaine de la biologie vient un besoin criant pour la création d'outils d'analyse des mégadonnées générées. Ceci s'avère le cas des études combinant la génomique à la métabolomique, nommées mGWAS, qui rapportent des associations gène(s)-métabolite(s). Notre objectif est de concevoir un outil permettant l'annotation quantitative des données mGWAS afin de faciliter leur interprétation en utilisant la distance réactionnelle (dr) comme métrique. Nous avons conçu un package R, PathQuant, d'après les standards de Bioconductor en exploitant le cadriciel RStudio et guidée par les tests unitaires (Runit) selon les bonnes pratiques de programmation en développement de logiciel. Le processus d'annotation comprend : 1) la création de graphes représentant la carte des voies métaboliques de la base de données Kyoto Encyclopedia of Genes and Genomes (KEGG), et 2) la classification, la cartographie et le calcul du dr pour chaque association gène-métabolite. L'annotation des données de l'étude mGWAS de Shin et al. (2014) a permis un tri rapide des 299 associations rapportées et révèle que 73% des associations impliquant un gène codant pour une enzyme sont à une distance inférieure à 5 des métabolites auxquels ils sont associés. PathQuant affiche des attributs de qualité logicielle dont la performance et la fiabilité pour une annotation rapide des données mGWAS. Ce package peut être modifié afin d'intégrer d'autres bases de données en vue de rehausser la valeur biologique de cette annotation et faire en sorte qu'il devienne un outil d'analyse incontournable en métabolomique.

16h30 - alien: un package R pour modéliser les interactions entre espèces
Steve Vissault
Depuis plusieurs années, les écologistes se sont penchés sur différentes techniques mathématiques et statistiques afin de mieux comprendre et prédire les relations qui lient les espèces entre elles. Certaines méthodes utilisent des propriétés physiologiques ou morphologiques des espèces et d'autres reposent sur des algorithmes d'apprentissage. Même si plusieurs de ces méthodes ont déjà été implémentées (parfois avec R), il n'existe actuellement aucun programme informatique rassemblant ces différentes méthodes. C'est pour combler cette lacune que nous avons développé le package R alien. Ce package R a pour but principal de standardiser l'utilisation des différentes méthodes présentement disponibles pour étudier les relations entre les espèces. En plus d'offrir un cadre d'implémentation flexible, il permettra d'inclure les potentiels nouveaux développements méthodologiques qui émergeront dans le futur. Le tout repose sur la définition d'un standard sur la structure d'entrée des données et de sortie des résultats des modèles. Dans cette présentation, j'illustrerai le fonctionnement de alien en présentant comment utiliser le package avec diverses structures de jeux de données. J'expliquerai aussi comment construire différents modèles permettant d'étudier les relations entre les espèces. Je conclurai la présentation en discutant des développements futurs que nous envisageons pour alien.