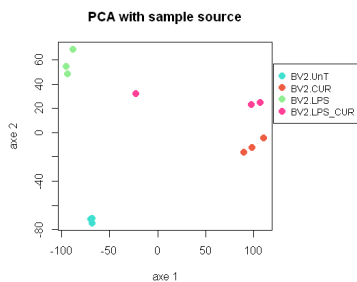


# Fiche de synthèse

Internship at the *Rowett Institute of Nutrition and health*  
with BioSS (*Biomathematics and Statistics Scotland*) to Aberdeen, Scotland.

**« A meta-analysis of publicly available microarray data  
to find effects of phytochemicals on gene expression. »**

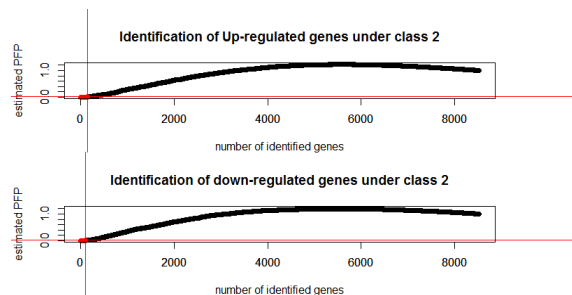
Nutrition has a significant impact on lifelong health and wellbeing. Plant derived chemicals (phytochemicals) have been identified as major contributors to the beneficial effects of fruit and vegetables. A number of high throughput analysis, like proteomics and microarray, have been carried out to understand the mechanistic basis for the effects of phytochemicals on cells. The analysis of microarray experiments is an important part of biological researchers on nutrition and health. Therefore they need the help of statisticians to deepen their researches because microarrays produce big data sets, that are not easy to interpret.



Moreover, a single data set gives only a limited picture but that looking at many datasets can give a more complete one.

There are many datasets from the literature that we can bring into a common analysis format and then analyse for significant similarities and differences. These results will improve understanding of the impacts of diet rich in fruit and vegetables on lifelong health. The project has been done

around that biological questions and methods have been implemented to find some results. The analysis of microarray experiments requires many implemented methods using linear models and empirical Bayesian methods. Here, we used principally a method called "Limma" to find the best differentially expressed genes in the microarray data. Moreover, the most important thing is this project in the "meta-analysis" which means that we want to find best differentially expressed genes across all experiments. We have therefore implemented some other functions which allow us to combine all experiments and apply other methods like "RankProduct" or "P-value combination" to get some results for the "meta-analysis".



## Modélisation de la fidélité aux enseignes



Presque tous les foyers français se rendent régulièrement dans les enseignes de grande distribution pour y effectuer leurs achats. Ces enseignes cherchent donc toutes à attirer le plus de clients possible pour avoir les meilleures performances par rapport à leurs concurrents. Dans ce contexte, il est facile de se demander comment les enseignes peuvent faire pour améliorer leurs performances.

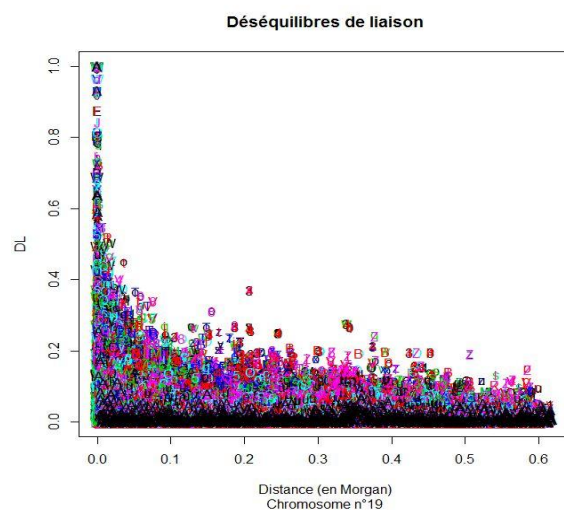
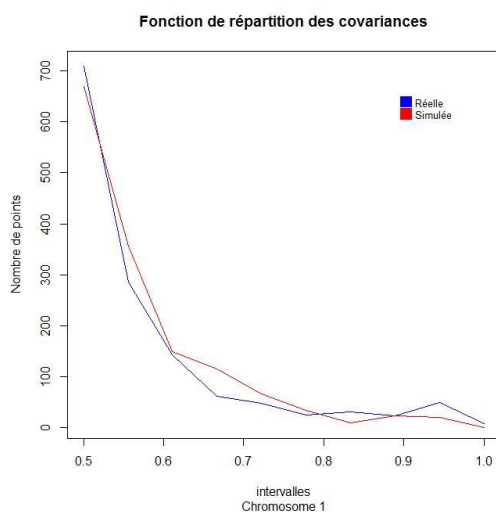
Le but de ce stage était donc d'essayer de répondre à cette question. En se focalisant dans un premier temps sur les rayons permettant de fidéliser les clients aux enseignes, on mettra en place des régressions PLS sur des données issues du panel de Kantar Worldpanel. Puis dans un second temps, on essaiera de comprendre comment les consommateurs choisissent leur magasin de grande distribution principal, en utilisant l'approche PLS sur des données qualitatives.

À la fin de cette étude, on notera donc que ce sont les rayons des produits de base, comme les produits frais ou les boissons, qui permettent de fidéliser les consommateurs. On remarquera aussi que ce qui pousse les consommateurs à fréquenter un magasin, est la proximité de celui-ci et l'attachement qu'ils lui portent, tandis que lors du processus de choix de magasin principal, c'est l'expérience en magasin qui entre en compte, notamment par le fait que le magasin soit agréable à fréquenter.



## Résumé :

L'amélioration des animaux d'élevage est une préoccupation majeure des éleveurs qui cherchent à sélectionner les meilleurs reproducteurs afin d'obtenir les descendants les plus performants et les mieux adaptés aux conditions d'élevage d'aujourd'hui et de demain. Dans les années 1990, les chercheurs ont commencé à utiliser l'information moléculaire sous forme de marqueurs ADN, pouvant être disponibles sur n'importe quel individu dès la naissance et permettant d'en prédire la valeur génétique, qui est liée à la « performance » pour certains caractères d'intérêt. La prise en compte de l'information moléculaire est particulièrement avantageuse sur des caractères difficilement mesurables, des caractères qui nécessitent la mort de l'animal (qualité de la viande) ou mesurables sur les femelles uniquement (production laitière) et les caractères peu héréditaires ou exprimés tardivement. Les méthodes de prédiction génomique permettent d'effectuer une telle sélection. Le but de ce stage était d'estimer l'efficacité de ces méthodes et de simuler des données génomiques dont l'acquisition est coûteuse. L'efficacité des méthodes de prédiction génomique peut être estimée en calculant le coefficient de détermination entre la valeur génétique prédite d'un individu (prévision faite à l'aide de modèles linéaires) et sa valeur réelle. Pour la simulation de données génomiques, j'ai d'abord étudié la répartition des déséquilibres de liaisons, qui caractérisent une association préférentielle entre deux allèles d'un même gène et qui sont égaux à la corrélation carrée entre des marqueurs binaires. J'ai pu constater que ceux-ci suivent une loi normale généralisée, dont les paramètres évoluent en fonction de la distance entre les marqueurs. Ainsi, après avoir simulé des données de marqueurs binaires indépendantes, j'ai pu simuler des variables suivant une loi normale généralisée afin de créer une matrice de variance-covariance pour les marqueurs en adéquation avec la réalité. Finalement, j'ai pu corrélérer les marqueurs en utilisant la transformée de Cholesky de la matrice de variance-covariance créée précédemment.



## Stage de 4<sup>ème</sup> année GMM (MMS) chez MEDIAMOBILE

### Sujet : Corrélations spatio-temporelles d'arcs de vitesses et prédiction court-terme

Mediamobile est une filiale du Groupe Télévision De France opératrice de services d'information trafic et de mobilité en temps réel.

Les données fournies proviennent de sources différentes comme Floating Mobile Data (FMD) ou principalement de Floating Car Data (FCD) fournies par les voitures qui disposent d'un GPS. Ceci dit, il peut arriver qu'il n'y ait pas de données à transmettre aux clients à cause d'un manque de circulation ou alors à cause de coupures causées par des problèmes techniques. Ceci donne lieu à des données manquantes que l'on doit essayer de compléter. En présence de données, il est aussi intéressant de pouvoir prévoir les vitesses à court terme pour prévoir des événements anormaux comme les bouchons.

L'objectif du stage est de répondre aux problématiques précédentes en élaborant une méthodologie de lissage à noyaux pour les vitesses observées sur des portions d'un réseau routier. Nous supposons que le réseau est formé d'arêtes sur lesquelles nous observons des vitesses de véhicules. Ces vitesses sont des moyennes toutes les 3 minutes, obtenues par moyennisation de vitesses FCD.

La méthode de complétion (et prédiction court terme) repose sur **une régression à noyaux** avec un noyau spatio-temporel

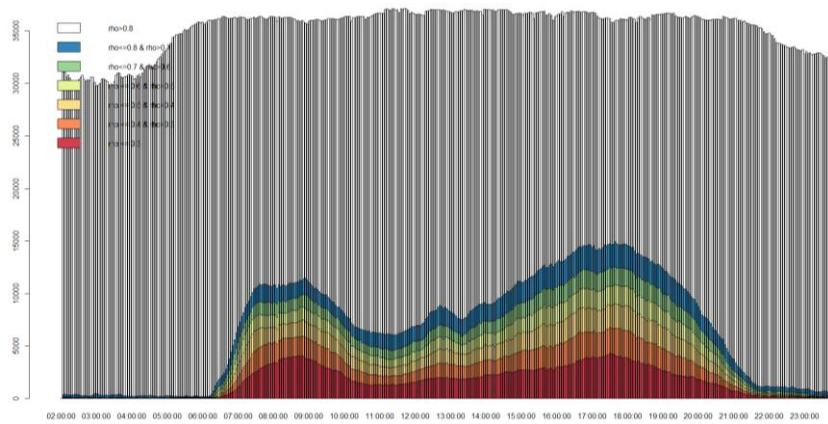
$$V(x, t) = \frac{\sum_{\{i,j=1\}}^n K\left(\frac{x-x_i}{h_x}, \frac{t-t_j}{h_t}\right) V(x_i, t_j)}{\sum_{\{i,j=1\}}^n K\left(\frac{x-x_i}{h_x}, \frac{t-t_j}{h_t}\right)}$$

Une fois la zone spatio-temporelle de l'étude est sélectionnée (on favorise des zones perturbées à faibles vitesses pour augmenter la réactivité du modèle final), il va nous falloir estimer correctement les paramètres de fenêtres  $h_t$  et  $h_x$  :

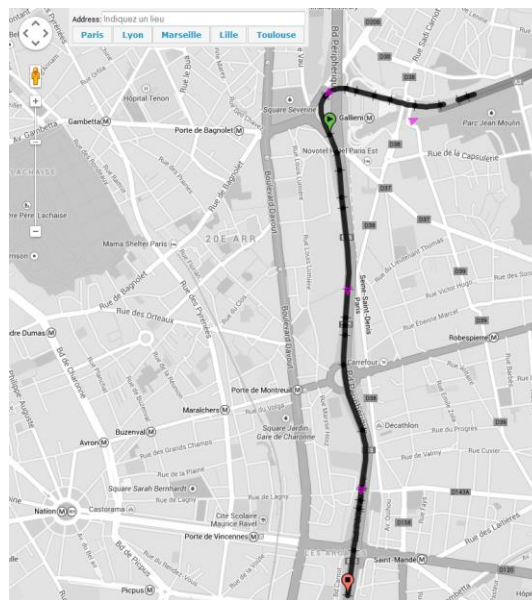
Pour un arc cible choisi, nous déterminerons un voisinage d'influence spatio-temporel en utilisant les corrélations de Spearman entre les variables de vitesses de chaque arc.

Nous choisirons des plages horaires où les coefficients de corrélations sont temporellement stables pour définir le voisinage spatio-temporel d'influence.

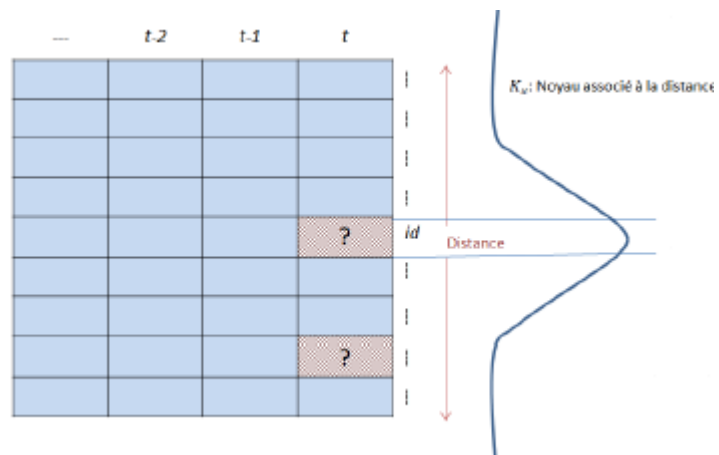
Ainsi, nous pourrions procéder à la régression non paramétrique, en veillant à prendre un noyau temporel asymétrique. Les résultats finaux montrent que le modèle convient aux cas de faibles vitesses (bouchons) mais des pistes d'améliorations sont proposer pour pouvoir généraliser le modèle à d'autres zones spatiales.



Répartition des perturbations de vitesses sur une journée



Résultat des corrélations de Spearman (Arcs corrélés avec l'arc cible en flèche verte)



Principe de la régression par noyaux

# Sujet du stage :

## Détection d'événements inhabituels sur des séries temporelles dans le cadre d'un système de surveillance épidémiologique syndromique

Stage à l'Institut de Veille Sanitaire (InVS)

Thomas Chauvet, département GMM, option MMS

La surveillance syndromique est un enjeu majeur pour l'institut de veille sanitaire. Ceci consiste à analyser des données épidémiologiques issues des services de santé, afin de détecter au plus vite des événements inhabituels. Le but est, par exemple, d'endiguer le plus rapidement une épidémie.

La nécessité de mettre en place un système de surveillance syndromique est apparue en France suite à la canicule de 2003 qui a conduit à un afflux massif de patients, le plus souvent âgés, dans les services d'urgence sans que cela ait pu être détecté assez tôt, voir anticipé, par la surveillance sanitaire.

La surveillance syndromique est basée sur le suivi quotidien d'indicateurs de santé afin de détecter une variation inhabituelle des indicateurs et d'en limiter les impacts. L'objectif du stage était de mettre en place des méthodes statistiques permettant la détection d'anomalies sur des séries temporelles de façon automatique. Le nombre de données étant important, il est nécessaire que cette analyse s'effectue en routine (quasi temp-réel) par des algorithmes performants. L'automatisation permet de surveiller plus d'indicateurs mais nécessite une intervention humaine en cas d'alarme.

Mon travail a consisté à effectuer une revue de la littérature afin de déterminer les méthodes applicables à ce projet. Ensuite, une étude exploratoire m'a permis de mieux comprendre les données disponibles. Enfin, j'ai pu concevoir différents algorithmes statistiques permettant la détection de ces événements inhabituels. Par la suite, j'ai pu évaluer la performance des différents algorithmes afin d'identifier les plus performants suivant la série étudiée. Pour finir, j'ai pu concevoir une application afin d'utiliser ces différentes méthodes à travers une interface ergonomique (un premier exemple de l'application sur des données simulées est disponible sur le lien suivant : [https://thomasc.shinyapps.io/Detection\\_simulee/](https://thomasc.shinyapps.io/Detection_simulee/))

Ce stage s'inscrit dans le cadre du projet SurSaUD<sup>®</sup> (Surveillance Sanitaire des Urgences et des Décès<sup>1</sup>) de l'InVS.

---

<sup>1</sup><http://www.invs.sante.fr/Espace-professionnels/Surveillance-syndromique-SurSaUD-R>

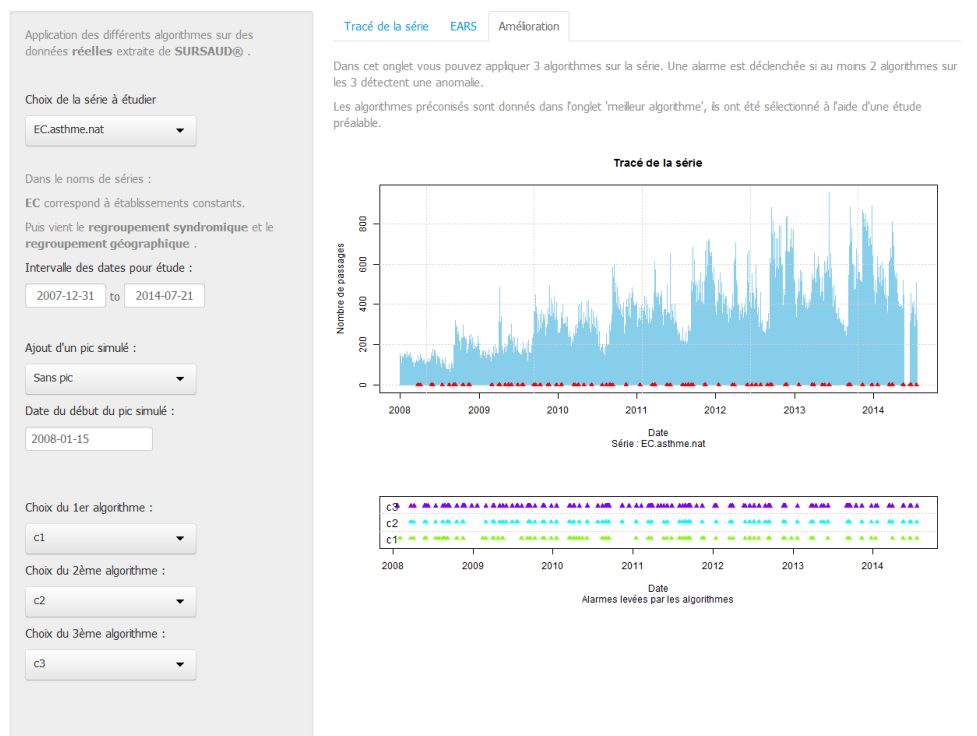


Figure 1: Exemple d'analyse de série temporelles (nombre de passages aux urgence au cours du temps suivant un regroupement syndromique) - Application développée avec R à l'aide de Shiny

## SIMULATION ET INFERENCE DANS LES EQUATIONS DIFFERENTIELLES STOCHASTIQUES

$$dX(t) = \mu(t, X(t)) dt + \sigma(t, X(t)) dW(t)$$

Les équations différentielles stochastiques sont des objets encrés dans les mathématiques modernes qui trouvent des applications aussi bien en biologie qu'en électronique. Leurs champs d'application principaux sont les mathématiques financières. Ainsi les modèles retenus en guise d'exemple dans ce rapport sont largement utilisés en économétrie, il s'agit des modèles d'Ornstein-Uhlenbeck (aussi nommé Vasicek), du modèle Gaussien géométrique (ou Modèle de Black-Scholes) et du modèle de Cox-Ingersoll-Ross (CIR).

Si il est envisageable de résoudre théoriquement une équation différentielle stochastique, comme cela sera fait dans le chapitre III, il est bien plus rapide de simuler les solutions via des schémas numériques (Schéma d'Euler et Milstein) ou des méthodes de linéarisation locale (Méthode d'Ozaki et Ozaki-Shoji). Ces simulations doivent cependant être le plus proche possible de la solution exacte et stable.

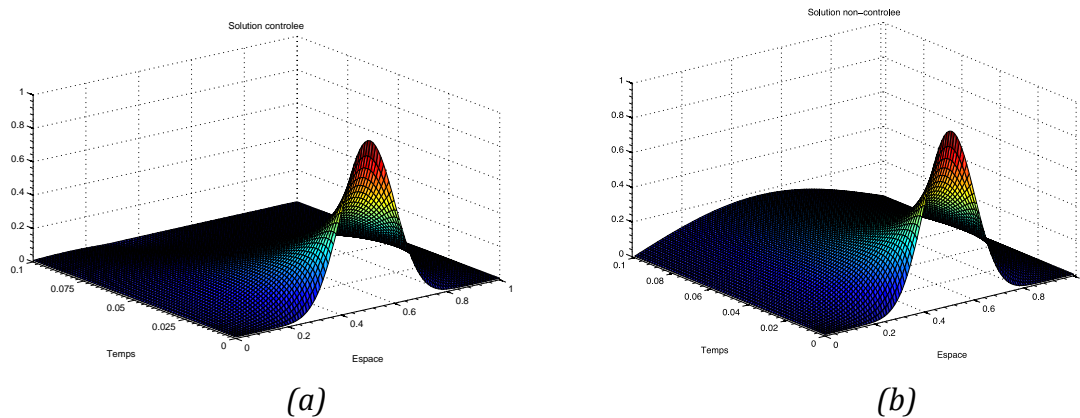
L'estimation de paramètres dans les modèles d'équation différentielle stochastique pose des problèmes récurrents dans les études des paramètres de processus dépendants. L'estimation par maximisation de la vraisemblance reste alors particulièrement efficace mais il existe une alternative intéressante qui est la méthode généralisée des moments.

Ce rapport présente un chapitre qui concerne l'estimation des points de changement de régime, modification brusque de la volatilité. Ainsi, un ensemble de programmes Matlab applicatifs est fourni, avec une documentation détaillée pour illustrer tous les concepts théoriques étudiés dans ce rapport.



## Analyse Numérique des Equations aux Dérivées Partielles et Contrôlabilité

Le but du stage est d'étudier, à l'aide d'articles scientifiques, la méthode de G. Lebeau et L. Robbiano permettant de démontrer la contrôlabilité à zéro de l'équation de la chaleur dans le cas continu et semi-discret. Plus précisément, cette méthode prouve qu'il existe une fonction de contrôle telle que la solution du système soit nulle à un temps  $T$  fixé. Cependant, la méthode n'étant pas efficace numériquement, le cas discret s'effectue à l'aide d'un algorithme de contrôle réécrit sous un problème d'optimisation avec la méthode HUM (Hilbert Uniqueness Method). Différents schémas de discrétisation sont mis en place (Différences Finies, Eléments Finis et Volumes Finis).



Figures : (a) Solution contrôlée - (b) Solution non contrôlée

Ce stage s'est déroulé au laboratoire de Mathématiques d'Orléans avec Mr. Jérôme Le Rousseau.

## Fiche de synthèse de stage de 4ème année GMM-MMN : Vincent Michaud

Mon stage s'est déroulé dans le département de météorologie de l'Université de Reading. J'étais intégré au sein de l'équipe de recherche CPOM (Centre for Polar Observation and Modelling). Mon stage portait sur la simulation de la banquise de l'Arctique. L'incertitude que l'on a lors de simulations de prédiction de l'état de la banquise est due à la représentation de sa physique dans le modèle, à la nature chaotique du climat (l'effet papillon) et aux différents futurs scénarios de taux de rejets de CO<sub>2</sub> dans l'atmosphère. L'objectif du stage était de déterminer l'importance de l'incertitude due à la représentation physique du comportement de la banquise dans le modèle comparée à l'incertitude due à la nature chaotique du climat. Tout d'abord, j'ai effectué une optimisation de modèles de la banquise, avec différentes représentations physiques du comportement de la banquise, pour déterminer quel modèle correspond le mieux aux observations satellites que l'on a de ses caractéristiques (son aire et son épaisseur). Ensuite, j'ai utilisé ce modèle pour faire des projections dans le futur et déterminer l'année de disparition de la banquise de l'Arctique. En modifiant la physique utilisée dans le modèle (en rajoutant ou en enlevant des modules modélisant de la physique non prise en considération ou modélisée de manière trop grossière), j'ai déterminé à quel point l'année de disparition de la banquise variait. Cela m'a donné une mesure de l'incertitude due à la représentation physique du comportement de la banquise. J'ai enfin comparé cette variation à une étude déjà effectuée donnant les variations d'années de disparition de la banquise dues non seulement à la représentation physique du comportement de la banquise mais aussi à la nature chaotique du climat.

