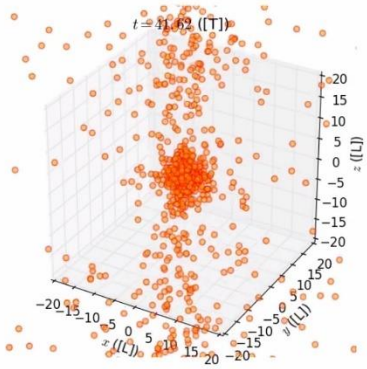
Dynamic Indicators of Gravitational Shocks on Globular Star Clusters

The study of globular cluster dynamics enables to spot whether or not it underwent a perturbation. While this can be used to detect if an observed cluster has recently been disturbed, the dynamical indicators described here can also apply to other stellar systems of different spatial scales.

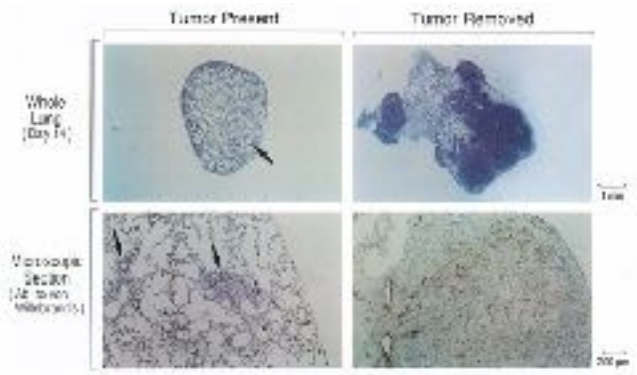
Numerical N-Body simulations of globular clusters within a galaxy gravitational field were conducted to find those indicators. Many functions were tested and compared to a baseline control simulation.



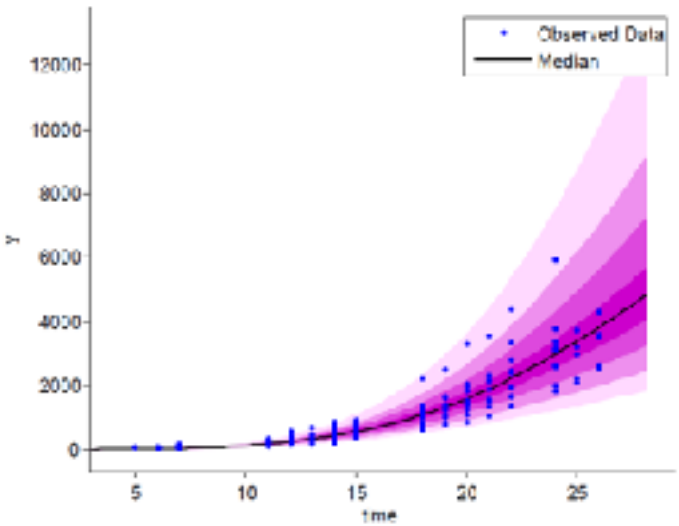
Six indicators are found to be relevant and efficient with gravitational shocks. Based on a confidence interval determined on the baseline simulation, their conjunction describes well the perturbations studied in this project.

Those indicators will be used on other stellar systems like globular clusters systems in compact groups of galaxies and should work well. The report presents the whole investigation process from the design of a simple but consistent N-Body simulator (in C language) to the tuning of those dynamical indicators.

The phenomenon of concomitant resistance, discovered since 1906, traduces the inhibitory effect from a first tumor on the growth of a distant tumor. The importance of the investigation on the concomitant resistance was found following the removal of the primary tumor which could lead to dramatic clinical consequences due to the suppression of this inhibition : the post-surgery metastatic acceleration. We report here on a study of a mathematical model representing the concomitant resistance between two tumors in the same organism.
First, the study involves a statistical analysis of the tumor growth in 10 mice with a population approach:the non-linear mixed effect model which is the most common tool to describe the global behavior of all individuals. The goal was to compare different software which implement the method, where the function NLME on R has the fastest execution time.
Second, the study allows the validation of the concomitant resistance mathematical model on independent data thanks to the obtaining of a highest goodness-of-fit and a good prediction.
This study not only informs on the validity of the model but also provides a non-monotony of the metastatic acceleration depending on the volume of the tumor at the day of excision.



O'Reilly, Folkman et al., Angiostatin: A Novel Angiogenesis Inhibitor That Mediates te Suppression of Metastases by a Lewis Lung Carcinoma, Cell 1994}



Prediction distribution of tumor growth with Gompertz model on Monolix

ANALYSE DE DONNEES APPLIQUEE AU DESIGN DE GUIDES D'ONDES DE SATELLITES TELECOM

Les satellites de télécommunications ont pour principale mission de transmettre des signaux électromagnétiques d'une région à une autre du globe, pour des applications telles que la téléphonie ou les programmes de radio et de télévision. Les guides d'ondes, dont le rôle est d'acheminer des signaux électromagnétiques, sont donc des éléments intrinsèques de ces types de satellites. Avec un nombre pouvant aller jusqu'à 2000 guides d'ondes par satellite, ces conduits de section rectangulaire ont la particularité de pouvoir supporter des puissances très élevées (centaines de watts) ou au contraire très faibles (quelques picowatts) et ce, avec un minimum de pertes.

Ainsi, en fonction de leur complexité géométrique, les guides d'ondes vont nécessiter ou non d'analyses mécaniques et thermo-élastiques, sous forme de simulations numériques. A l'heure actuelle, ce sont des ingénieurs expérimentés qui décident si un guide d'ondes donné requiert une analyse ou s'il est directement validé conforme. L'objectif de ce stage était de faciliter cette étape de décision, manuelle et particulièrement subjective, grâce à un algorithme, afin de réduire le coût que l'analyse peut engendrer et permettre un gain de temps considérable.

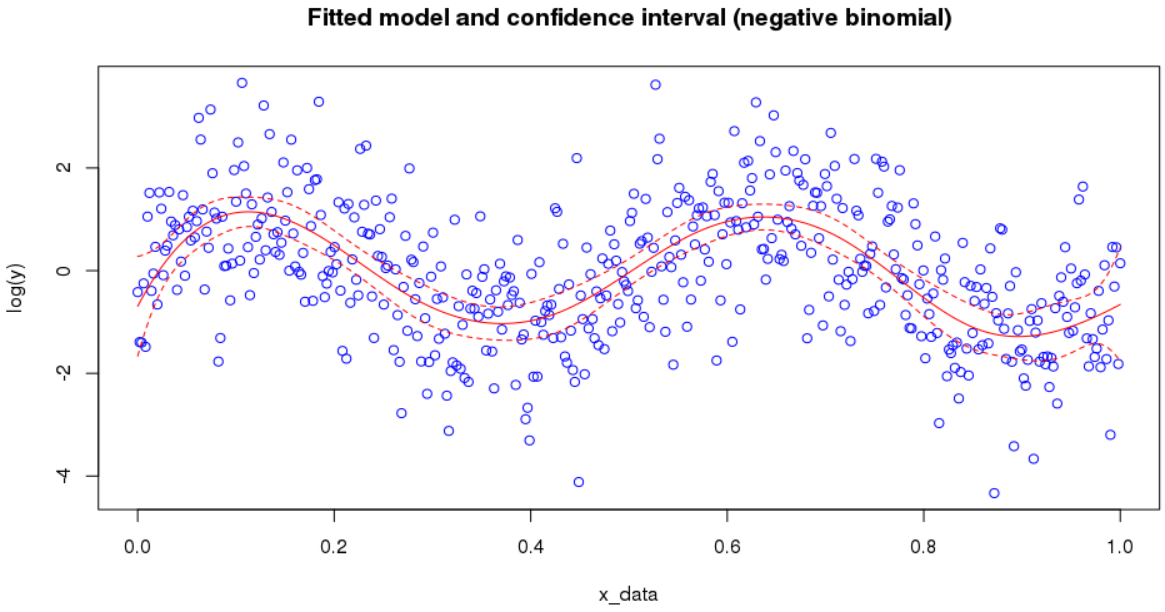
Pour cela, une première grosse étape de prétraitement a été effectuée dans le but d'obtenir un jeu de données propre et exploitable. Par la suite, différentes méthodes d'apprentissage supervisé de la librairie Scikit-Learn, telles que les arbres de classification, les SVM, les forêts aléatoires ou encore les réseaux de neurones, ont été mises en place. Après avoir défini les paramètres optimaux de chaque modèle au moyen d'analyses paramétriques, ces méthodes de machine learning ont alors pu être évaluées et comparées par des techniques dites de validation croisée. Finalement, le stage aura abouti à un outil de prédiction en Python fiable et offrant diverses perspectives d'amélioration possibles.

ALGORITHME POUR L'AJUSTEMENT DE LONGS MODELES ADDITIFS GENERALISES LONGITUDINAUX

En tant qu'étudiante en 4ème année à l'INSA de Toulouse, j'ai eu la chance d'effectuer un stage de trois mois au sein du laboratoire de recherches Gagneur Lab, entre le 4 Juillet et le 30 septembre 2016. Ce laboratoire universitaire, situé sur le campus de l'Université Technique de Munich (TUM, Allemagne), a pour but d'étudier la régulation génique et ses implications sur les maladies. Le projet sur lequel j'ai travaillé pendant mon stage s'intègre dans le développement d'un environnement de travail appelé GenoGAM. C'est un package permettant l'analyse statistique de données génomiques.

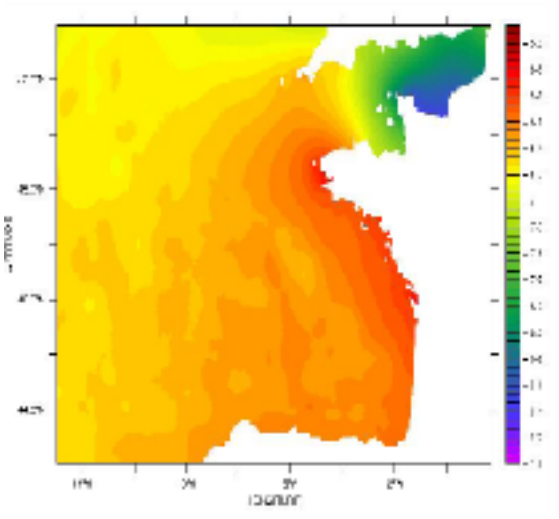
Dans le cadre de mon stage, il s'agit en particulier de quantifier les intéractions protéines-ADN grâce à un séquençage par immunoprecipitation de chromatine (ChIP-Seq). On cherche à modéliser les données collectées à l'aide de modèles additifs généralisés (GAM). Le package *mgcv*, implémenté en R par Simon Wood, propose déjà des fonctions qui fournissent de très bons résultats, mais présentent l'inconvénient d'avoir une complexité quadratique. Etant donné la dimension des systèmes qui nous intéressent (le génome humain par exemple), le temps d'exécution de ces fonctions est bien trop élevé. Mon travail consistait à concevoir un nouvel algorithme (en R) pour l'ajustement des données, à partir de GAM, avec une complexité sub-quadratique.

Etant donné un ensemble de positions génomiques x_i, et la valeur mesurée correspondante par ChIP-Seq y_i, on cherche le vecteur de paramètres $\boldsymbol{\beta}$ tel que $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}$, où $\mathbf{X}$ est la matrice de design. On minimise pour ce faire la log-vraisemblance à l'aide de la fonction *optim*. Ceci étant fait, on s'intéresse à l'intervalle de confiance correspondant. On veut donc calculer la matrice de covariance de $\boldsymbol{\beta}$, qui s'avère être l'opposé de l'inverse de la matrice hessienne de la log-vraisemblance. Plusieurs méthodes ont été implémentées, on retiendra en particulier l'utilisation de l'algorithme SPAI (Sparse Preconditionner Approximate Inverse) ainsi que de la fonction *solve,* fournie dans le package {Matrix}.

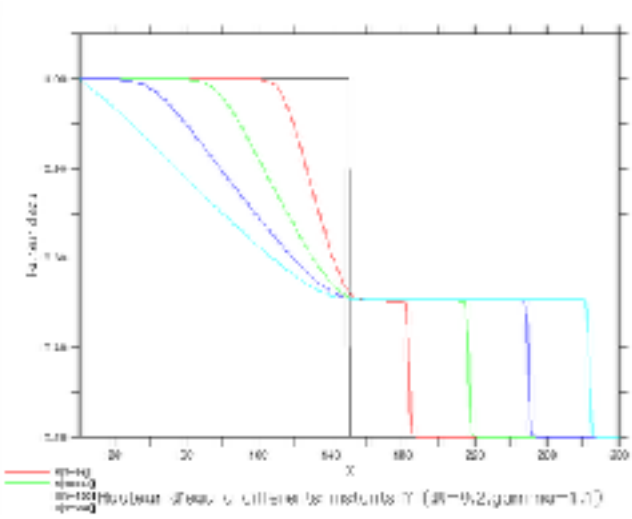


D'après les prévisions faites, l'algorithme final exécuté en parallèle sur 32 cœurs devrait présenter un temps d'exécution de cinq heures pour un génome humain au lieu de la trentaine d'heures nécessaires avec les programmes utilisant les fonctions de *mgcv*.

J'ai réalisé mon stage de 4ème année GMM-MMN au Service Hydrographique et Océanographique de la Marine (SHOM) dans la division Hydrographie, Océanographie, Météorologie (HOM) à Toulouse. Les prévisions océanographiques réalisées par le SHOM nécessitent des modèles sur lesquels se baser pour prévoir, aussi bien, les phénomènes de marée, de surcotes que de vagues. Le modèle plus particulièrement ciblé par ce stage était le modèle de surcotes qui présente des bruits au niveau des zones de découvrement. L'intérêt de mon stage était donc de tester un nouveau schéma pour pallier ces problèmes. L'intitulé exact était : « Modélisation de la dynamique océanique avec le modèle aux équations primitives HYCOM : Implémentation d'un schéma semi-implicite entropique sur un maillage décalé vitesse-pression pour le système Shallow-Water monodimensionnel ». Ce schéma entropique présente des qualités qui pourraient permettre d'effacer les problèmes d'instabilités près des fronts secs lors du calcul des surcotes, à un coût informatique raisonnable. J'ai donc réalisé une étude poussée de ce schéma avec de l'implémenter en Fortran avec une méthode volumes finis. Une montée en ordre a ensuite été réalisée afin d'améliorer les résultats obtenus et diminuer la diffusion tout en gardant les propriétés importantes du schéma considéré. Les résultats ainsi obtenus sont de bonne qualité pour des tests représentatifs.



Modélisation de la côte française avec le modèle aux équations primitives



Modélisation de rupture de barrage sur lit mouillé

Caractérisation de l'organisation de la matrice extracellulaire par imagerie numérique

J'ai réalisé mon stage au sein de l'équipe MORPHEME. Celle-ci est une équipe multi-disciplinaire (biologie/traitement du signal) et internationale. Elle fait partie des équipes de l'INRIA Sophia Antipolis-Méditerranée mais aussi des équipes de l'Institut de Biologie-Valrose (iBV).

Ce stage comportait deux phases: l'une était centrée sur du traitement du signal, nous voulions savoir s'il était possible de caractériser ceux que nous voyions dans les images par des nombres (des orientations, des épaisseurs...). L'autre phase concernait les tests permettant de confirmer ou d'infirmer si ce que nous avions extrait des images avait un sens, pour cela nous avons essayé de classifier les images à notre disposition en utlisant du machine learning.

Si on donne plus de détails, dans notre corps, les cellules vivent empêtrées dans une sorte de toile d'araignée, celle-ci est appelée matrice extracellulaire (MEC) et elle prodigue aux cellules un soutien structurel ainsi que biochimique. Des biologistes ont montré que le principal constituant de cette matrice, la fibronectine, existe sous quatre formes différentes différentiables par leur composition. Durant ce stage, nous avons cherché s'il était possible que ces différences de composition puissent aboutir á des différences de topographie au niveau de la MEC. S'il cela est avéré, il serait alors possible de deviner à quels variants appartient la MEC apparaissant sur tel cliché. Il s'agit d'un problème de classification d'images. Pour extraire des caractéristiques des images, nous avons utilisé deux transformations mathématiques: la Scale Invariant Feature Transform (SIFT) et la transformée en curvelets. Les résultats que nous avons obtenus montrent que nous sommes capables de construire des classifieurs très précis. Nous en tirons donc la conclusion que les informations que nous extrayons de l'image sont significatives et pourraient donc être utilisées pour des applications plus approfondies comme la construction d'un modèle de MEC par exemple. La figure 1 résume le travail que nous avons réalisé durant ce stage.

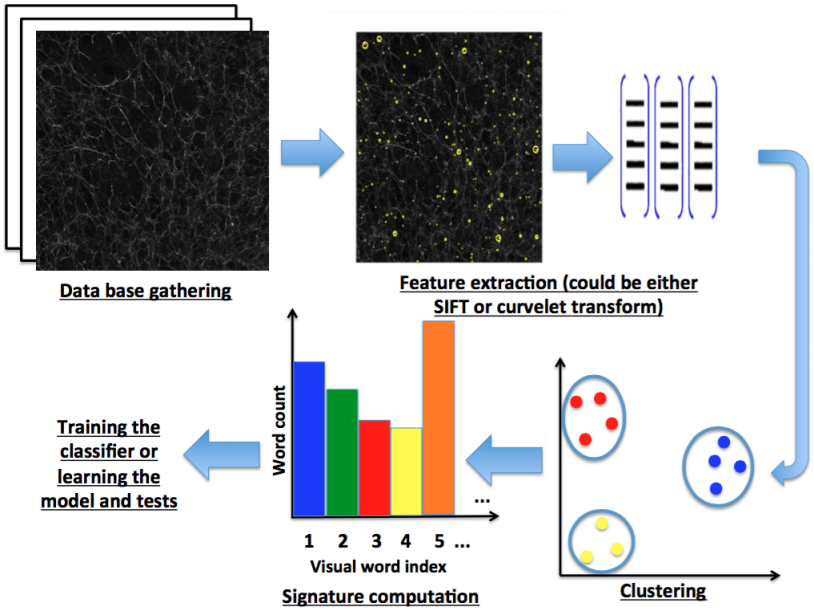


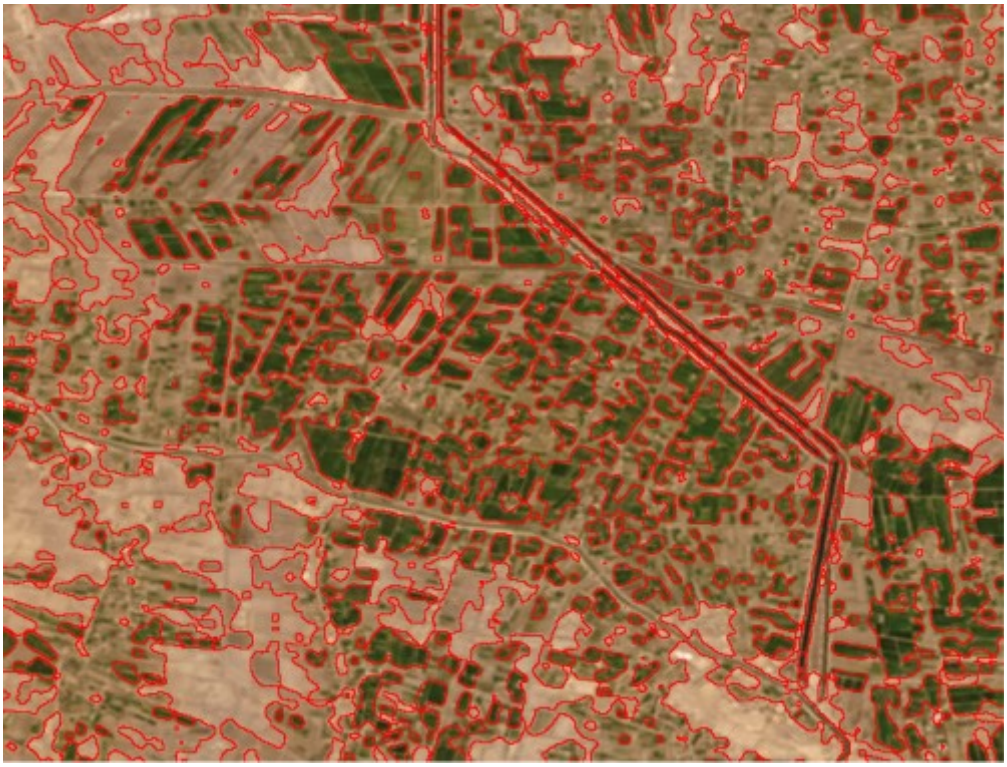
Figure 1: Algorithme de classification complet

Segmentation of Hyperspectral Images and Ensemble Clustering

The purpose of this internship was to develop a method for the segmentation of satellite or aerial images. This method had to be based upon the partition of color data with ensemble clustering. Ensemble clustering is a family of algorithms aimed at producing a consensus partition from several input partitions of a point cloud. We worked on satellite images from *Landsat* and *Sentinel* programs and aerial images acquired with *Aviris* spectrometer. Those images contain more than three channels: a dozen bands for satellite images and two hundreds for *Aviris* images. Those bands are chosen so as to distinguish easily the different features which can be found at the surface of the earth: water, forest, urban area, etc.

The mathematical part of this internship involved unsupervised clustering and dimensionality reduction. The computer science part of this internship involved programming with C++ language, parallel programming with OpenMP and the use of Qt and Eigen libraries.

This internship took place in Novosibirsk, the third largest city of Russia, at Akademgorodok. This district of Novosibirsk is dedicated to science and hosts the Novosibirsk State University (NSU) and several institutes belonging to the Siberian Branch of the Russian Academy of Sciences, where many internships are offered each year in biology, laser physics or mathematics for instance.



Pierre Fabre Dermo-Cosmétique
Centre de recherche sur la peau, Toulouse
Service Biophysique et Imagerie

Analyse statistique de données biologiques complexes en dermo-cosmétique

Depuis sa création au début des années 60, le groupe pharmaceutique Pierre Fabre ne cesse de se développer. Le Centre de recherche sur la Peau à l'Hôtel Dieu Saint-Jacques à Toulouse participe à ce développement en réalisant des études visant à comprendre la physiologie de la peau et les mécanismes impliqués dans les dérèglements cutanés.

Le monde de la Dermo-Cosmétique est aujourd'hui de plus en plus en quête de validations scientifiques. Parallèlement, l'informatique est de plus en plus performante et permet de traiter de gros volumes de données. Les études statistiques sur différents types de données, telles que les puces d'ADN ou encore les spectres RAMAN, sont mises en place afin de comprendre la structure et le fonctionnement de la peau et du cheveu et d'étudier l'efficacité des produits dermo-cosmétique.

Pendant les trois mois de stage que j'ai effectués, j'ai eu l'occasion de travailler sur trois missions différentes.

Ma première mission consistait à effectuer des analyses exploratoires et différentielles sur des données de PCR quantitative afin d'évaluer l'efficacité de deux produits dermo-cosmétiques sur l'acné.

Pour mon deuxième projet, j'ai réalisé une analyse exploratoire sur des données d'imagerie RAMAN afin de mettre en évidence des différences structurelles entre les différentes couches de la peau.

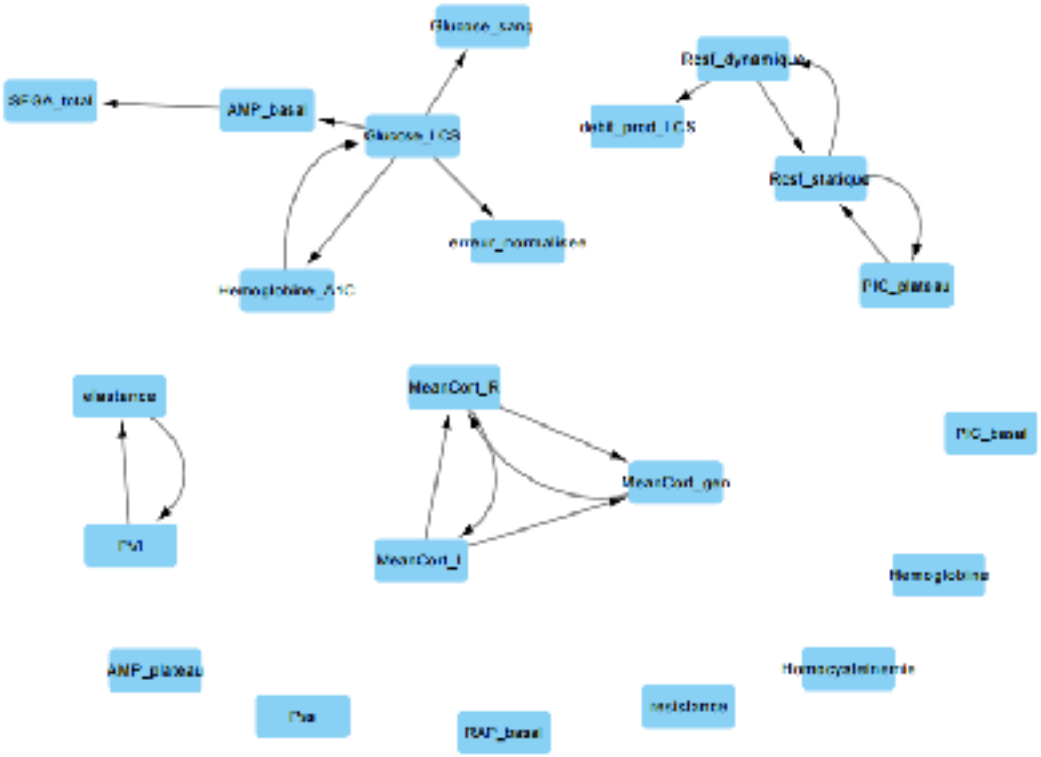
Enfin, la dernière étude sur laquelle j'ai travaillé traitait du vieillissement des cheveux chez la femme. L'objectif était de faire une interprétation biologique des résultats d'analyse différentielle obtenus sur des données de microarrays.

Analyse de données médicales
issues d'une étude du vieillissement du cerveau humain.

Le monde de la médecine est un domaine encore partiellement inexploré aujourd'hui, d'autant plus lorsqu'il touche au cerveau humain qui est une grande énigme que les chercheurs rêvent de résoudre. Les mathématiques sont un très bon outil afin de mieux comprendre certains aspects ou fonctionnement du vivant, et peuvent donc aider à la recherche médicale.

L'étude clinique sur le vieillissement du cerveau dirigée par le docteur Schimdt au CHU de Toulouse a permis le relevé d'un grand nombre de données diverses. Cette étude, menée sur un échantillon de 100 patients âgés et diagnostiqués malades résulte d'un énorme travail de récolte de données en amont. A l'aide d'outils issus de l'analyse de donnée, le but du projet est de mieux comprendre le fonctionnement du cerveau et son mécanisme de vieillissement, notamment avec le développement de maladies dites neurodégénératives. Ces maladies sont en grande augmentation depuis des décennies, dû au vieillissement de la population mais aussi au mode de vie moderne de l'aire industrielle.

L'étude menée ici est faite sur un échantillon de 100 patients relativement âgés et malades dont on a soutiré un grand nombre de données cliniques, biologiques et anatomiques. On a cherché, grâce à ces données, à trouver des similitudes parmi un maximum de variables afin d'idéalement reconnaître les quatre sous-groupes de maladies déjà diagnostiquées. Dans le cadre du début de ce projet je me suis concentrée sur trouver des liens qui ressortent entre les différentes variables, voire éventuellement des liens de causalité. Pour cela, j'ai adapté et implémenté différents algorithmes, sachant qu'il nous fallait des résultats les simples possibles à interpréter. Les résultats les plus parlants que j'ai obtenus sont sous formes de réseaux bayésiens, visible en exemple sur la figure ci-dessous.



J'ai réalisé mon stage au sein de l'équipe Connaissance Clients de la direction Marketing chez Engie.

Dans le cadre de l'installation des compteurs communicants « Linky » - qui relèvent la consommation en électricité à distance et les transfère directement au gestionnaire du réseau - dans les foyers français, un déluge de données est reçu par le groupe. Ce dernier se réserve donc le droit de les utiliser afin d'améliorer la connaissance de ses clients en explorant les habitudes de leur consommation et ainsi leur offrir des services pour mieux contrôler leur consommation.

Compteur communicant Linky

En partant de différentes informations concernant le client : qu'elles soient liées à la nature de leur contrat, à leurs caractéristiques géographiques, sociodémographiques ou même aux offres souscrites et au mode de paiement, le but de mon stage, dans un premier temps, était d'établir une classification des courbes de consommation électrique.

Le deuxième principal objectif consistait à mettre en œuvre un nouvel algorithme Self Organizing Map en adoptant cette fois-ci deux visions : le but de la première était de tracer une courbe de charge approximative représentante de chaque classe en se basant sur ces informations; inversement, celui de la seconde était de dresser des profils clients (âge, puissance souscrite, segment, type de l'offre, type de comptage …) en utilisant uniquement les données de consommation collectées à l'aide du compteur et de déterminer les variables les plus influentes sur les habitudes de consommation.

J'ai majoritairement utilisé le logiciel statistique R, mais j'ai également eu recours à SAS quelques fois pour la préparation des données. Une présentation PowerPoint a été réalisée à la fin de mon stage afin d'expliquer à mes collègues le travail que j'ai effectué.