Classification nuageuse à partir d'observations de MSG4 et à l'aide de méthodes d'intelligence artificielle. [CEMS]

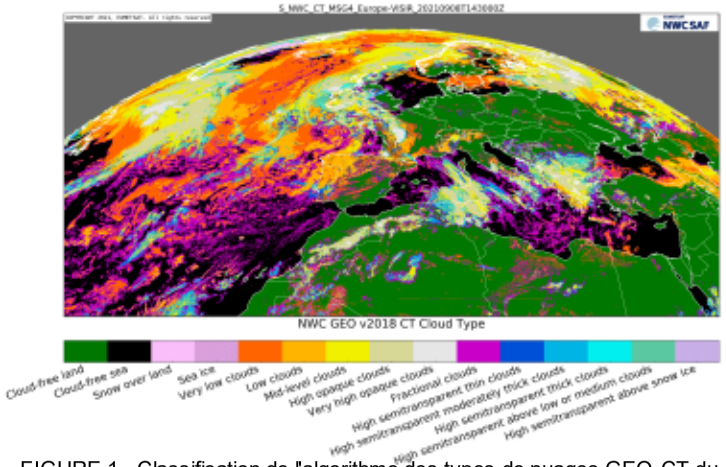


FIGURE 1 - Classification de l'algorithme des types de nuages GEO-CT du SAF-NWC.
(Source : https://www.nwcsaf.org/ARCHIVO_GIF/NRT/S_NWC_CT_MSG_Europe-VISIR_LAST.ct.gif)

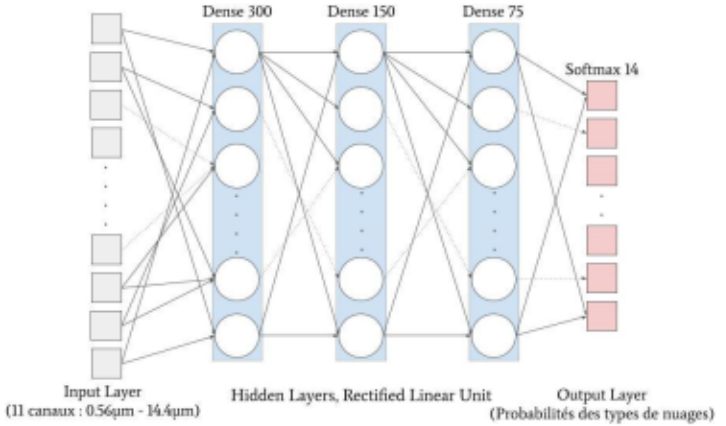


FIGURE 2 - Réseau de neurones développé pour classifier les pixels nuageux en 14 classes.

Dans le cadre du projet européen SAF-NWC, le CEMS/Météo-France a développé l'algorithme de classification des nuages GEO-CT utilisé en prévision à moyen terme. Ce dernier permet d'identifier automatiquement, avec fiabilité et rapidité, les types des nuages issus des observations du satellite MSG4. En prévision du lancement en 2023 du satellite MTG-I1, le CEMS s'est intéressé aux méthodes d'intelligence artificielle, qui ont récemment fait leurs preuves dans le domaine de la météorologie, comme potentielle alternative à l'algorithme de classification existant. Mon stage, effectué au sein de l'équipe Nuages du CEMS, a porté sur l'étude de ces méthodes pour la classification nuageuse à partir d'observations MSG4.

Dans un premier temps, après avoir analysé les données satellite acquises par MSG4 et simulées par RTTOV pour MSG4, j'ai étudié trois méthodes d'apprentissage supervisé : la régression logistique multinomiale, la forêt aléatoire et le perceptron multicouches. Les modèles implémentés, entraînés sur des observations MSG4 puis optimisés par validation croisée ont été évalués en comparaison avec GEO-CT. La forêt aléatoire et le réseau de neurones dépassent 90% de justesse mais n'atteignent pas le temps de prédiction requis.

Dans un second temps, mon rôle a été de déterminer si ces méthodes, entraînées sur des simulations RTTOV pour MSG4, étaient capables de classifier des observations MSG4. Les classifications de la forêt aléatoire et du réseau de neurones implémentés ont atteint cette fois-ci jusqu'à 80% de justesse. Les performances des modèles testés lors de ce stage n'atteignent les objectifs fixés, néanmoins les résultats de classification obtenus restent prometteurs. En effet, entraînés sur des simulations RTTOV pour MTG, ces modèles permettraient potentiellement de classifier des pixels nuageux des observations de MTG-I1.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

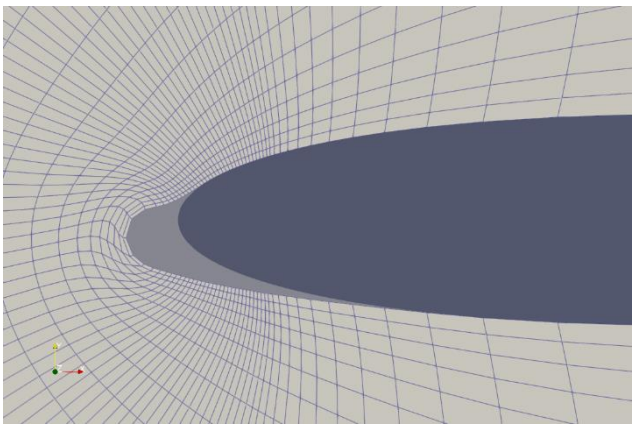
ONERA THE FRENCH AEROSPACE LAB

Utilisation de la déformation de maillage pour la prise en compte de l'accrétion de glace sur une aile d'avion.

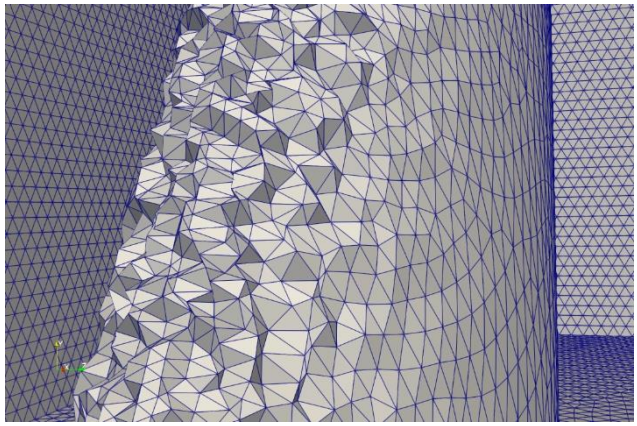
J'ai réalisé mon stage de 4ème année à l'Office National d'Études et de Recherches Aérospatiales (ONERA) sur le site de Toulouse.
L'objectif de ce stage était de mettre en place un outil de déformation de maillage (Quantum*) dans un code modélisant les problèmes tri-dimensionnels d'accrétion de glace.

En effet, en aéronautique le givrage est un problème majeur. Ce phénomène est étudié dans différents laboratoires pour pallier la variété de cas où il intervient. L'étude porte sur l'accrétion de givre pendant un vol.

Après avoir, dans un premier temps, mis en place cette méthode dans une chaîne de calculs traitant les problèmes en deux dimensions (IGLOO2D*), cette dernière a été implémentée dans une chaîne de calculs modélisant les problèmes en trois dimensions (IGLOO3D*). Pour obtenir de tels résultats, l'utilisation de la CFD General Notation System a été nécessaire. La manipulation des données modélisant le profil était ainsi un point central de l'étude.



Résultat obtenu sur un problème en deux dimensions avec une méthode pas-à-pas



Résultat obtenu sur un problème en trois dimensions avec de très petits déplacements

* : l'ensemble de ces codes sont des codes qui ont été développés par l'ONERA

Camille Latapie – 5ème année Mathématiques Appliquées

Use of data science and business intelligence to improve university student outcomes

Marie LE CHEVÈRE

2020-2021

I had the opportunity to do my fourth-year internship at Ulster University, Northern Ireland. Unfortunately, due to the pandemic situation, it was entirely remotely conducted. I was tutored by Mr Michael Callaghan, speaker at Ulster University, and worked with another French engineer student on the project.

My placement topic was to explore student attendance monitoring, and to provide a new solution to the current existing system in the University of Ulster, that was composed of two main technologies: GEL (Game Enhanced Learning) and Power BI, a business intelligence tool from Microsoft.

As the Ulster University uses Blackboard Learn, a learning management system (LMS), to connect students and teachers through courses by providing an easy-to-use platform, it was requested that we create a new attendance monitoring system. The idea was to link Blackboard Learn, which contains a database of all courses and students, to a tool taking attendance, and then to a business intelligence (BI) application to visualize and try to predict patterns in the attendance. Data can then be automatically refreshed and loaded into the BI dashboard without having to do it manually. This kind of analyses offers a possibility to understand how to improve university courses and overall students' successes.

I was able to use different tools and softwares during this internship, such as AWS EC2 (an Amazon Web Service instance), Blackboard, Qwickly (an attendance tool), Power BI, and communication tools, such as Slack and Trello.

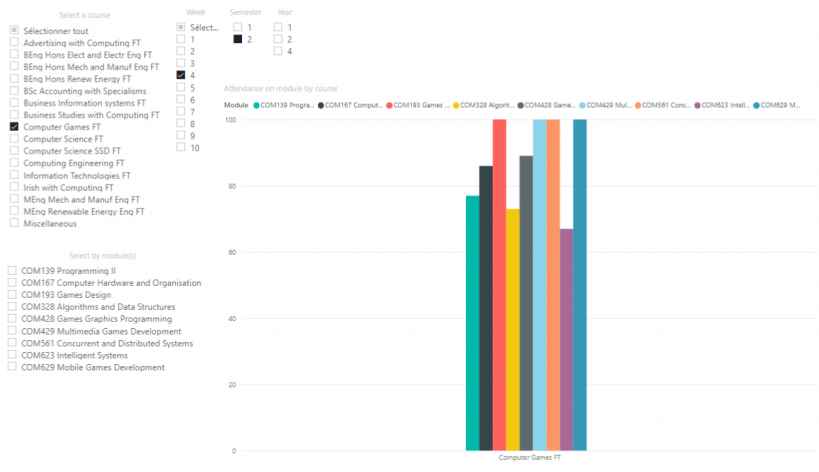
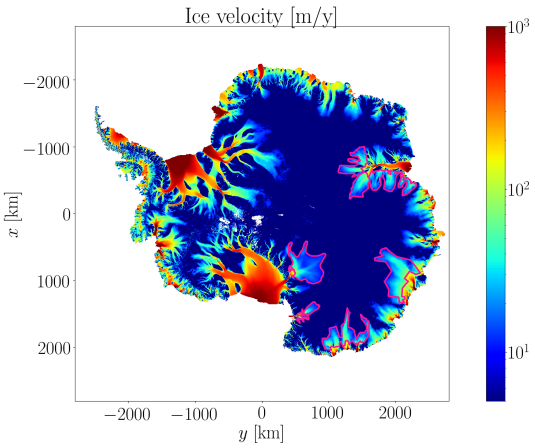


Figure: Power BI dashboard showing students' attendance for Computer Games course (Week 4, Semester 2.

Estimation de la topographie du lit rocheux antarctique

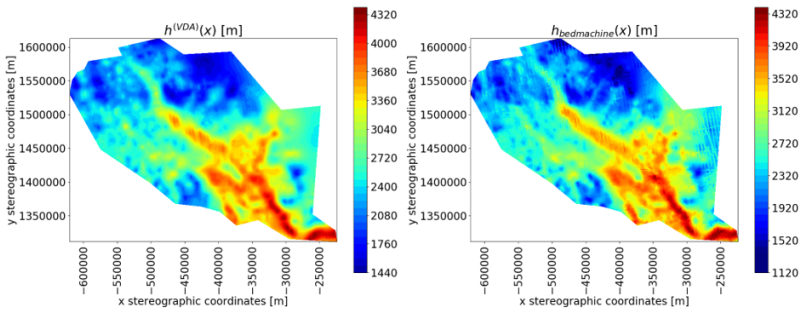
L'Antarctique est un vaste continent encore méconnu. Froid et sec, il est aujourd'hui décrit comme étant le plus important désert de notre planète. Néanmoins, cette hostile immensité n'est que la surface d'un véritable continent enseveli sous des kilomètres de glace. Mais alors, à quoi ressemblerait l'Antarctique sans sa calotte ? Qu'y a-t-il sous cette épaisse couche de glace ? La question peut sembler triviale au premier abord mais nécessite un intérêt bien plus poussé pour pouvoir y répondre.

En l'occurrence, l'Antarctique de l'ouest est aujourd'hui quadrillée de mesures aériennes permettant d'approcher la topographie de son lit rocheux. Cette densité de données s'explique notamment par la contribution principale de cette zone à la hausse du niveau des mers. Les glaciers y sont instables et se déplacent à des vitesses dépassant les 100 mètres par an. Ce n'est en revanche pas le cas de l'Antarctique de l'est, bien plus vaste mais aussi plus stable. La plupart de ses glaciers ne dépassent pas les 100 mètres parcourus par an. Y estimer l'épaisseur de glace nécessite donc des modèles supplémentaires à ceux utilisés en Antarctique de l'ouest.



Vitesse surfacique en Antarctique et zones considérées à l'est.

L'objectif du stage est d'estimer la topographie du lit rocheux en Antarctique de l'est. À ce jour, la cartographie admise comme référence mondiale est BedMachine. Cette dernière n'est néanmoins valide que pour des vitesses supérieures à 30 mètres par an. C'est donc dans l'intervalle de vitesses inférieur que l'introduction d'un nouveau modèle RU-SIA par Jérôme MONNIER et Jiamin ZHU peut apporter une information plus précise. Néanmoins, le modèle direct nécessite la topographie basale pour ensuite pouvoir délivre l'altimétrie surfacique. Or, c'est exactement l'inverse qui est nécessaire ici puisque des données satellitaires de surface sont disponibles. Il faut donc inverser le modèle pour qu'il puisse délivrer l'altitude basale, et donc l'épaisseur de glace. Cette inversion nécessite une assimilation variationnelle de données, une estimation d'un paramètre adimensionnel γ par réseau de neurones et une nouvelle assimilation variationnelle de données. Le résultat final visible ci-dessous est encourageant puisqu'il s'accorde plus finement aux données mesurées que la cartographie précédente BedMachine.



Cartographie finale (à gauche) et cartographie BedMachine (à droite).

Fiche synthèse :

Wrapping Python du code de calcul DassFlow2D

Contexte

DassFlow2D est un code de calcul codé en **Fortran** qui permet d'accomplir des tâches d'assimilation de données variationnelles. Le fait de le wrapper en **Python** permet d'enrichir le logiciel avec les fonctionnalités de haut niveau de Python (en l'occurrence, un minimiseur **scipy** a été appliqué aux codes wrappés).

Wrapper un code Fortran en Python est intéressant pour les raisons suivantes :

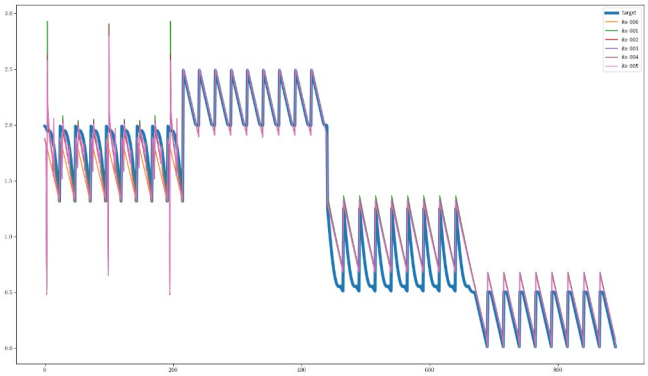
1. On peut conserver l'efficacité d'un code Fortran tout en l'appelant dans un code Python
2. On peut appliquer des librairies Python qui ont des fonctionnalités de haut niveau au code Fortran qui réalise des fonctionnalités de bas niveau.

Résultats :

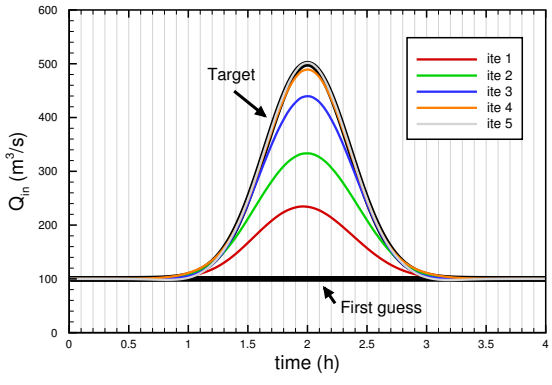
En me basant sur l'excellent travail du précédent stagiaire Ngo Nghi Truyen Huynh,

1. J'ai compris le fonctionnement de DassFlow2D (simulation directe, adjointe et la minimisation)
2. J'ai terminé le wrapping de DassFlow2D, pour ce, il a fallu :
   - 2.1 Généraliser la lecture du maillage et les conditions aux limites en wrappant les routines Fortran existantes
   - 2.2 Wrapper la routine permettant de lancer une simulation (directe, adjointe) et la minimisation
   - 2.3 Wrapper la routine permettant de lire/générer les observations pour les tâches d'assimilation de données.
   - 2.4 Conserver la fonctionnalité mpi de DassFlow2D
3. J'ai mis à jour la documentation de DassFlow2D
4. J'ai validé partiellement le wrappé (un problème de minimisation en cas de mpi a été détecté lors de la validation du wrappé) .

J'ai également réalisé des simulations directes et j'ai réalisé l'identification des paramètres d'entrée en appelant le minimiseur **m1qn3** (déjà existant) et **LBFGS** (ajouté pendant ce stage).



Voici un exemple de l'identification d'un paramètre, ici c'est la bathymétrie. L'identification fonctionne mal à cause d'un problème lié au calcul du gradient par rapport au paramètre de bathymétrie qui devait être investigué.



Voici un autre exemple de l'identification du paramètre. L'identification porte sur le débit en amont de la rivière et fonctionne correctement par rapport à la version précédente du code.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

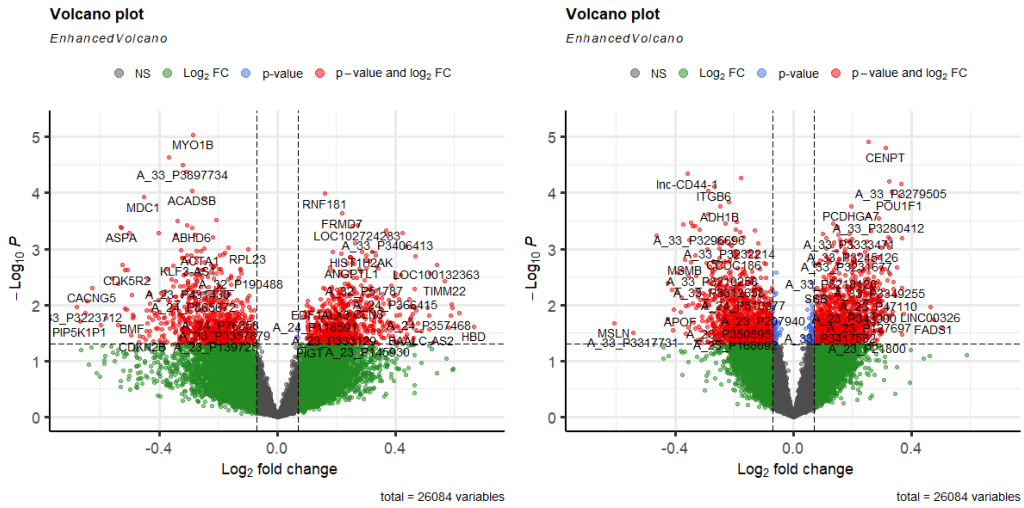
i2MC Institut des Maladies Métaboliques et Cardiovasculaires

Analyse de données biopuces – Recherche de gènes différentiellement exprimés au cours d'une intervention entre deux groupes d'individus obèses

L'obésité est un sujet important et d'actualité qui concerne la quasi-totalité de la planète. L'obésité, reconnue comme une maladie chronique par l'Organisation mondiale de la santé, est un excès de masse grasse et une modification du tissu adipeux. Cette maladie peut entraîner de nombreux problèmes de santé tels que le diabète, l'hypertension et des cancers pouvant réduire l'espérance de vie.

Pendant la quasi-totalité de mon stage au sein de l'Institut de Maladies Métaboliques et Cardiovasculaires (I2MC), j'ai travaillé sur le projet MONA (Metabolism Obesity Nutrition Age). Il s'agissait d'une étude comparative entre deux groupes d'hommes obèses, un groupe d'adultes jeunes (30 à 40 ans) et un groupe de seniors (60 à 70 ans). Les deux groupes ont participé à une intervention diététique de 8 semaines avec restrictions calorique modérée (déficit de 20% par rapport aux besoins énergétiques quotidiens) et un entraînement physique supervisé par un éducateur sportif : marche de 45 à 60 min, 5 fois par semaine. De nombreux échantillons de tissu adipeux et de muscles ont été utilisés pour quantifier l'expression des gènes à l'aide de la technologie biopuce.

L'objectif étant de rechercher des gènes différentiellement exprimés j'ai tout d'abord réalisé une analyse exploratoire des données en faisant une analyse en composantes principales afin de détecter les tendances, identifier les effets non désirés puis les supprimer. Ensuite, pour trouver les gènes différentiellement exprimés avant et après l'intervention j'ai réalisé des tests-t de Student des données appariées et pour trouver les gènes différentiellement exprimés entre jeunes et seniors avant intervention j'ai utilisé un package qui s'appelle limma.



Volcano plot de la différence d'expression des gènes selon l'intervention pour tous les patients. Gauche : Données du muscle / Droite : Données du tissu adipeux. L'axe des abscisses représente le Fold Change (FC) en log2. L'axe des ordonnées représente les valeurs en -log10. En bleu les gènes dont les p-valeurs sont significatives, en vert les gènes dont les FC sont significatifs et en rouge les gènes correspondant aux deux conditions précédentes. Les gènes se trouvant sur la partie positive du log2(FC) sont les gènes surexprimés chez les seniors et sur la partie négative, les sous exprimés chez les seniors.

Prédiction de dimensions et espaces tangents de variétés de haute dimension

Aldo MELLADO AGUILAR

L'objectif de ce projet a été d'estimer la dimension de variétés de haute dimension à partir d'un nuage de points, ainsi que trouver des approximations des plans tangents à ses points.

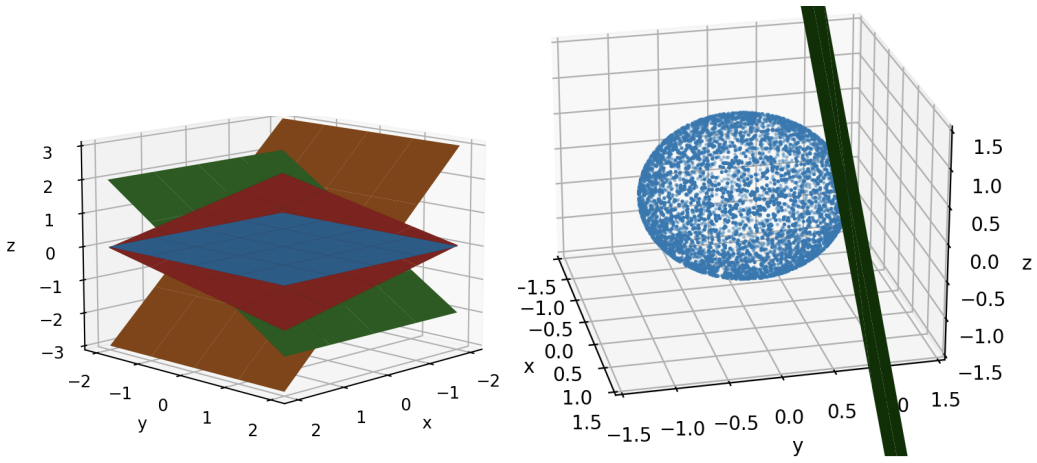
Une variété est une figure qui ressemble localement à un espace euclidien. Les exemples les plus simples sont les cercles et les sphères, qui sont des variétés de dimension 1 et 2 respectivement.

Dans un premier temps on a expérimenté avec des nuages de points provenant de cercles et sphères pour tester des algorithmes pour prédire leurs dimensions. Les méthodes SVD, la règle d'Oja et l'EM-ACP ont donné des bonnes résultats pour la prédiction de la dimension des variétés. Néanmoins, une fois qu'on a testé ces méthodes sur des variétés plus complexes, on n'a pas trouvé des bonnes résultats. Chaque méthode donnait une prédiction différente pour la même variété, mais l'algorithme qui donnait les meilleurs prédictions était la SVD.

Ensuite, pour faire l'estimation des espaces tangents aux points des variétés, on a travaillé avec les grassmanniennes, qui sont des espaces contenant des sous-espaces vectoriels linéaires d'un espace vectoriel fixe. On a étudié une méthode pour calculer la moyenne de la grassmannienne et on a aussi proposé un algorithme de point fixe pour le calcul de la médiane. Malheureusement cet algorithme ne convergeait pas vers le bon résultat.

Finalement, on a simulé les plans tangents aux variétés en utilisant la moyenne de la grassmannienne. On a testé cette méthode avec des cercles et sphères et les résultats ont été très proches des vraies valeurs.

Pour continuer ce projet, il serait intéressant de tester cette estimation de plans tangents dans des variétés plus complexes. De même, on pourrait continuer à chercher un bon algorithme pour trouver la médiane de la grassmannienne, qui pourrait servir à estimer des plans tangents des nuages des points avec des outliers.



INSTITUT de MATHEMATIQUES de TOULOUSE

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Fiche synthèse

Modélisation et analyse du déplacement en trois dimensions des microplastiques dans la baie de Marseille en utilisant Ichthyop

Eloïse Merlaud – 5GMM

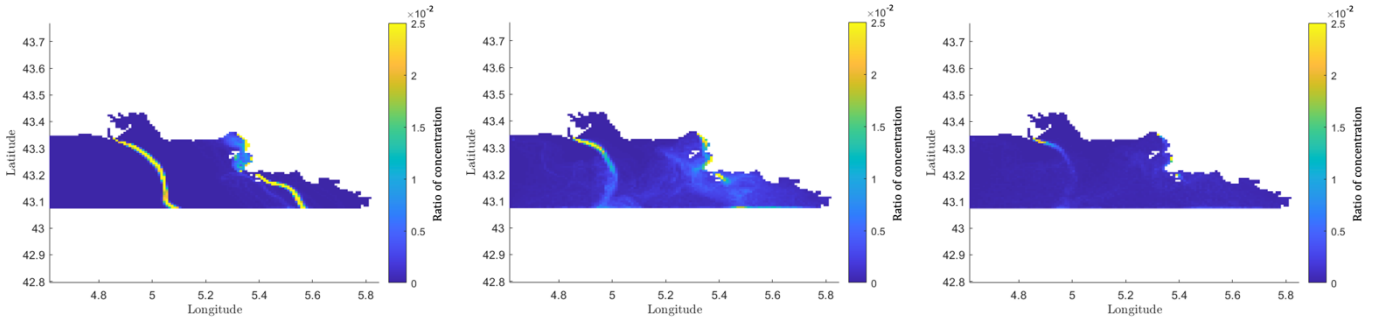
Tuteurs : Cristèle Chevalier et Nicolas Barrier

Mon stage de 4ième année s'est déroulé au sein de l'Institut Méditerranéen d'Océanologie de Marseille dont l'objectif principal consiste à étudier les écosystèmes et les écoulements marins pour mieux les comprendre et les protéger. Ce stage s'inscrivait dans un projet sur l'étude des microplastiques dans la Méditerranée, en particulier dans la baie de Marseille en collaboration avec deux équipes : la CEM (Chimie des Environnements Marins) et l'OPLC (Océanographie Physique, Littoral et Côtière).

Dans un premier temps, j'avais pour mission d'implémenter un modèle de déplacement des microplastiques dans la mer en trois dimensions. Pour cela, j'ai développé le logiciel Ichthyop. Ce-dernier permettait jusqu'à ce jour de modéliser le transport des planctons ou des plastiques en utilisant des équations d'advection et de dispersion mais en ne prenant pas en compte la dispersion verticale due aux vagues et à la turbulence. Ainsi, à partir d'un modèle lagrangien aléatoire de déplacement vertical, j'ai dû ajouter ce phénomène dans le logiciel pour avoir une modélisation de leur transport plus réaliste.

Dans un second temps, pour répondre à la demande de certains chercheurs en biologie du laboratoire, j'ai dû analyser le déplacement des particules de plastiques dans la Baie de Marseille. Pour étudier aux mieux ces flux, 4 sources de microplastiques ont été identifiées (le Rhône, l'Huveaune, les sorties d'épuration à la calanque de Cortiou et le port industriel de l'Estaque). A partir de données météorologiques et de courant dans ces zones, des simulations de déplacement de plastiques ont pu être faites sur toute l'années 2020. Elles ont notamment permis de mettre en avant différentes causes expliquant les flux de plastiques observés.

Par exemple, l'impact saisonnier et le vent semble jouer un impact considérable sur la direction des microplastiques. En effet, le Mistral, très présent en hiver notamment, modifie la direction des plastiques suivant les courants marins, pour les amener vers le sud de la Baie. Toutes ces conclusions pourront permettre d'adapter le traitement des eaux et de mieux prévoir les quantités de plastiques par endroit.



c. Sans dispersion verticale

b. Avec dispersion verticale

a. Avec dispersion verticale mais en ne comptant que les particules comprises entre 0 et 1 mètre

Figure 1 : Panaches des concentrations maximales de microplastiques relâchés par la surface (1000 microplastiques relâchés par site au premier jour de la simulation) pour une simulation débutant le 2 février 2020 pour 30 jours.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Extracting dispersion curves by Deep Learning to study underground structure

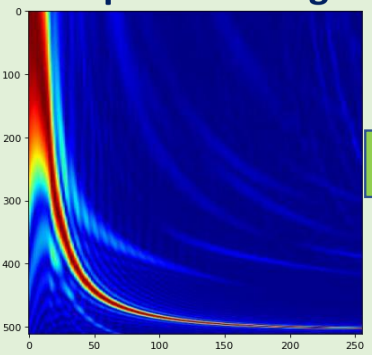
NGUYEN Hai Vy- GMM

During my internship at l'OMP, I build up a data-base by simulating different programs in geophysics. From there, I build and train a model to extract dispersion curves from dispersion images. Based on these dispersion curves, researchers can reconstruct the underground structure by the process of dispersion curve inversion. Therefore, extracting correct dispersion curves is extremely important.

The process of extracting dispersion curves is divided into two steps:

- Finding the energy-zones: image segmentation by DCNet which is a deep neural network with convolutional layers.
- Post-processing to fit dispersion curves: Clustering by DBSCAN. Rectifying curves by Gaussian filter. In each energy zone that we found, we can fit the dispersion curve by finding for each frequency the phase velocity at which the dispersion energy attains local maximum peak

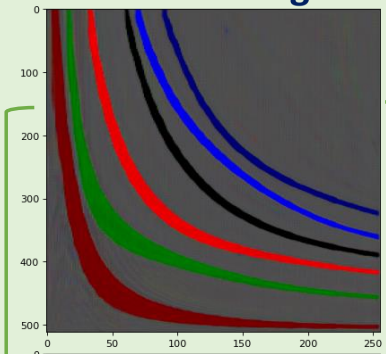
Dispersion image



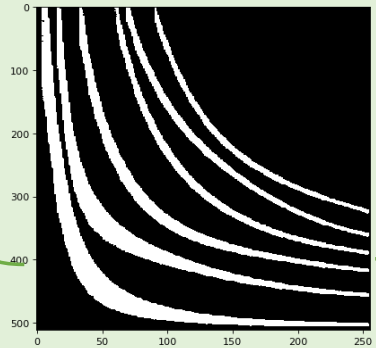
Deep Neural Net

Input → DCNet → Output

Embedding

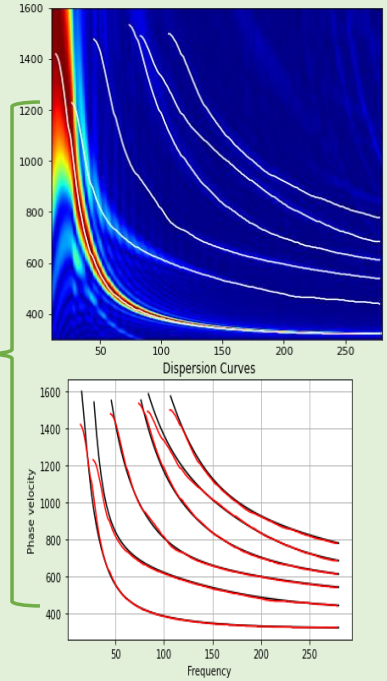


Segmentation



Post-processing → Output

Extracted Dispersion Curves



Comparing with theoretical curves by CPS

Observatoire Midi-Pyrénées OMP

G E T

Supervisors:
Roland MARTIN
Thomas OBERLIN
José DARROZES
Bastien PLAZOLLES

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

adelyce

Benchmark

Classification et comparaison de clients par l'intermédiaire d'indicateurs clés

Adelyce, Labège (31)

Mots-clés : analyse de données, statistiques, classification non supervisée, réduction de dimension, ACP, données publiques, visualisation de données, segmentation de clients.

Adelyce est une jeune entreprise dynamique créée en 2007 et localisée à Labège (Haute-Garonne). Cette société de conseil a créé une solution experte de pilotage financier de la masse salariale des collectivités territoriales françaises.

Ce stage s'est déroulé au sein du pôle Recherche & Développement (R&D) d'Adelyce. Il consistait en la contribution à Benchmark, un vaste projet d'explorations statistiques sur les clients et prospects d'Adelyce. Un des objectifs consistait en la comparaison de la masse salariale des clients, en créant des clusters homogènes, par le croisement de ces données avec des indicateurs liés entre autres à la géographie, l'attractivité et le dynamisme économique du territoire. Une exploration de données exogènes était alors nécessaire. Cette comparaison constituait un début de recherche pour la conception d'un nouveau module proposé par le produit d'aide au pilotage de la masse salariale. Des comparaisons sur la masse salariale, mais aussi sur le profil des agents ont été menées.

Un des autres grands objectifs était de segmenter une part du marché d'Adelyce : les communes françaises dont la population était comprise entre 1 500 et 10 000 habitants. Des méthodes de clustering ont pu être utilisées, et les variables d'intérêt étaient multiples telles que les charges de personnel, la masse salariale et plusieurs indicateurs de dynamisme économique.

Les techniques usuelles de clustering, de réduction de dimension et de modélisation statistique ont été particulièrement utilisées. Les implémentations de ces projets ont été réalisées en R, et les résultats ont pu être visualisés et partagés aux autres services de l'entreprise *via* des markdowns, comprenant des interfaces graphiques (Figure 1). Des restitutions sous la forme de cartes interactives se sont également avérées utiles (Figure 2).

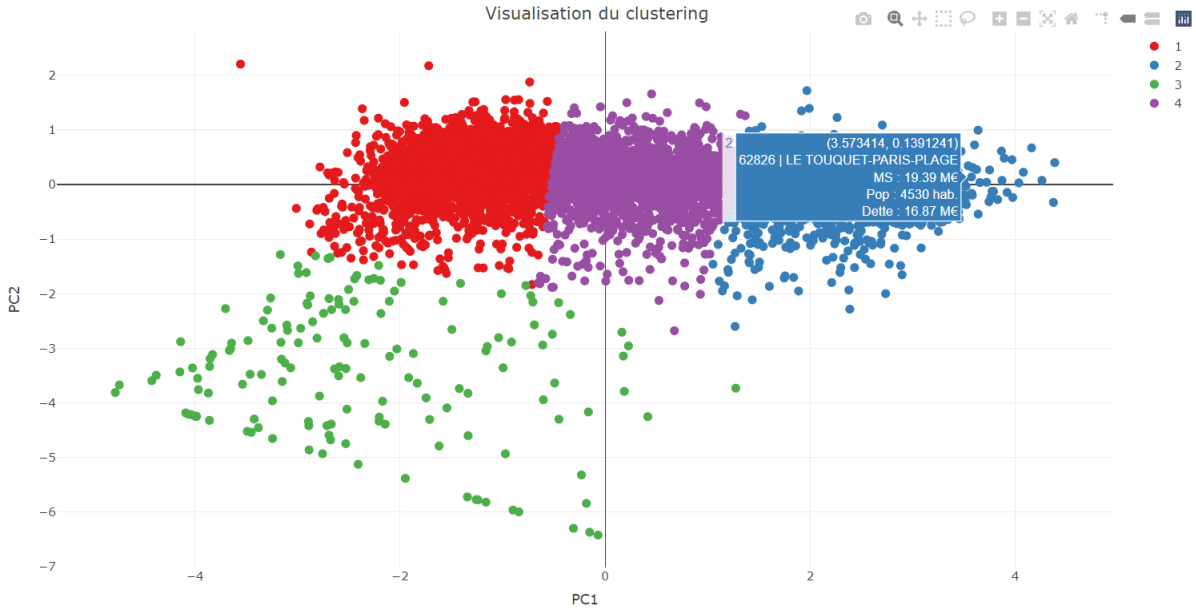


FIGURE 1 – Segmentation de clients - Visualisation d'un clustering K-means à l'aide de la librairie `plotly` de R.

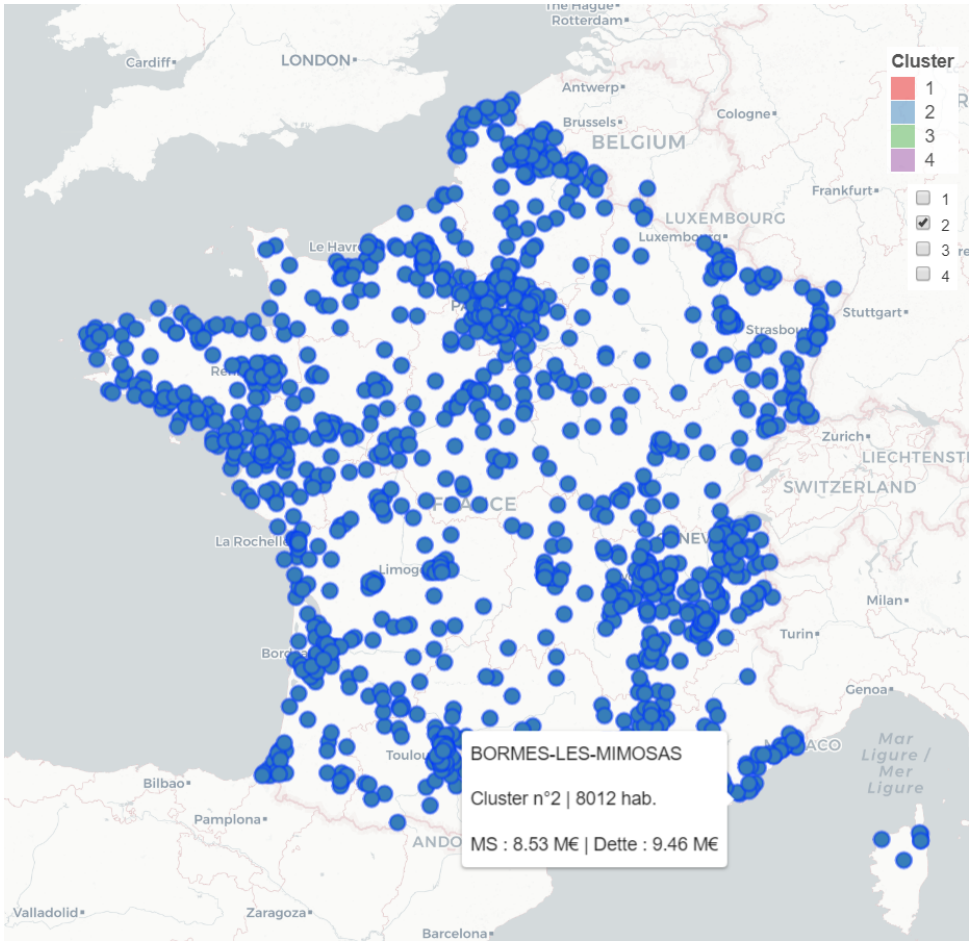


FIGURE 2 – Visualisation cartographique des communes appartenant à un cluster

Résumé de stage :

Etude pour de nouveaux modèles en Assurance.

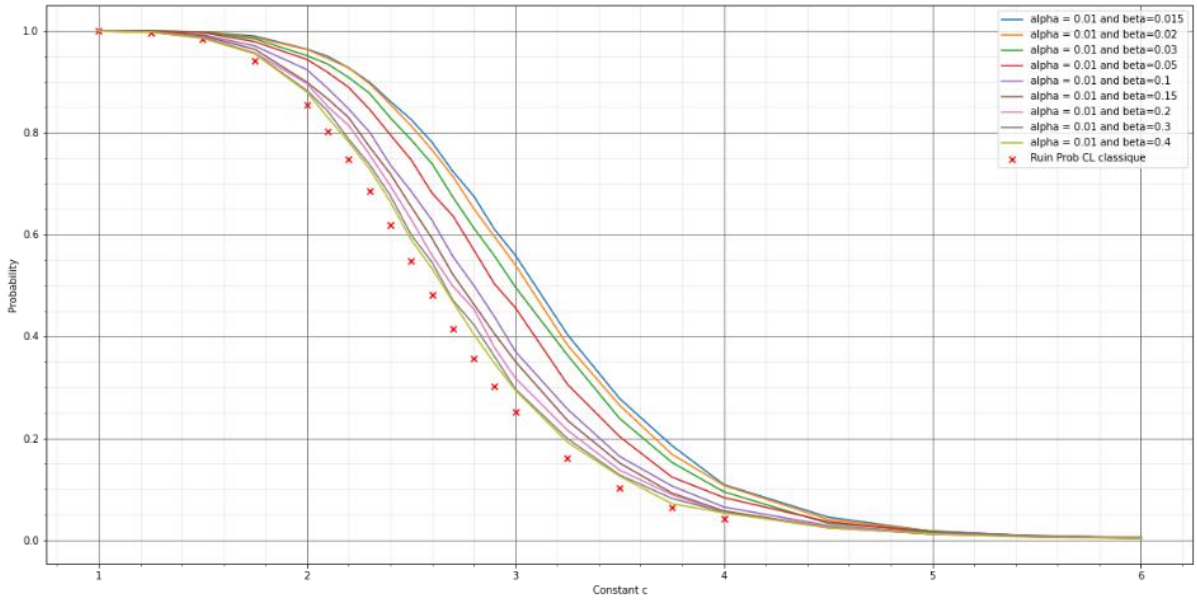
En industrie « conventionnelle », le cycle de production permet de donner aux objets manufacturés leur valeur. En d'autres termes, le prix de vente peut être connu avant l'achat. L'assurance quant à elle, est soumise à cycle de production inversé, l'assureur ne connaît pas en avance le montant d'un sinistre futur ! Mais alors comment faire pour tarifer un contrat d'assurance garantissant la solvabilité (être capable de rembourser ses assurés) de l'assureur ?

La modélisation statistique en assurance est le socle de base qui permet d'approximer le comportement des gens, des sinistres et de leurs survenances afin de pouvoir tarifer un contrat. Dans la grande majorité des modèles utilisés, la survenance des sinistres suit une distribution probabiliste qui lui est attribuée après une étude statistique sur les données d'années passées. Le montant des sinistres est lui aussi associé à sa distribution. Ce qui est important de remarquer ici, c'est que chaque domaine (survenance et montant des sinistres) est étudié chacun de leur côté.

Pendant mon stage, l'objectif était donc d'étudier un modèle qui prenne en compte une interaction entre les sinistres et leur temps de survenance. Imageons tout ça par une application du modèle en assurance santé : lorsqu'un gros sinistre survient (typiquement une intervention chirurgicale) il peut entrainer d'autres dépenses avec lui (une rééducation, une assistance à domicile etc…). On voit donc comment un sinistre peut interagir sur l'intensité (pensé comme probabilité de déclenchement) de survenance d'autres sinistres.

Pour étudier ce type de modèle, nous avons commencé par étudier les processus de Hawkes et comment est-ce qu'ils pourraient intervenir dans un processus de risque en assurance. La première étape fut donc d'étudier les processus de Hawkes qui ont la caractéristique d'être auto-excitants, c'est-à-dire que lorsqu'un « évènement » se produit son intensité est incrémentée automatiquement. Afin de mieux l'imaginer, ce processus a initialement été créé pour modéliser les secousses sismiques survenant après la première. L'étape suivante fut de l'intégrer dans un processus de risque, c'est-à-dire un processus qui modélise la « richesse » mathématique de l'assureur et de montrer numériquement la décroissance de la probabilité de ruine avec l'augmentation de la prime d'assurance.

Pendant ce stage, nous avons donc montré numériquement la décroissance de la probabilité de ruine d'un processus de risque ayant pour intensité un processus de Hawkes. Nous avons également entamé la démonstration théorique mais elle est toujours en cours d'étude.



INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

CICR

Aide à l'identification de migrants pour le Comité International de la Croix Rouge : Etude des biais d'algorithmes de reconnaissance faciale

J'ai effectué mon stage de 4ème année dans le cadre d'un partenariat entre le groupe INSA et le Comité International de la Croix-Rouge (CICR). J'ai participé à un programme d'identification de migrants noyés en Méditerranée. Cette identification était réalisée grâce à un logiciel et un algorithme de reconnaissance faciale. Je devais tenter de quantifier les biais que pouvaient avoir ces deux outils. Ce stage de recherche avait pour but d'identifier les paramètres qui pouvaient influencer la reconnaissance, comme la qualité des images, l'origine ethnique et le genre des personnes à reconnaître, l'alignement des visages, ou encore la présence des yeux et de la bouche. Il était également important de voir comment fonctionnait cette reconnaissance sur des défunts, en les comparant à des photos d'eux vivants.

La reconnaissance faciale n'a cependant pas fonctionné aussi efficacement que ce que l'on espérait, et encore moins sur les personnes décédées. J'ai donc également cherché des pistes d'amélioration, comme le surentraînement des algorithmes. Les résultats que j'ai obtenus ont permis de conclure sur certains aspects, mais il serait tout de même nécessaire de poursuivre les études pour voir les effets de certains paramètres que je n'ai pas pu observer.

Figure 1 : Exemple de reconnaissance erronée : les deux premiers visages représentent clairement la même personne et le logiciel de reconnaissance faciale indique qu'elles sont différentes, alors que le premier et le troisième visage appartiennent évidemment à deux femmes différentes et le logiciel indique que c'est la même (base de données LFW)

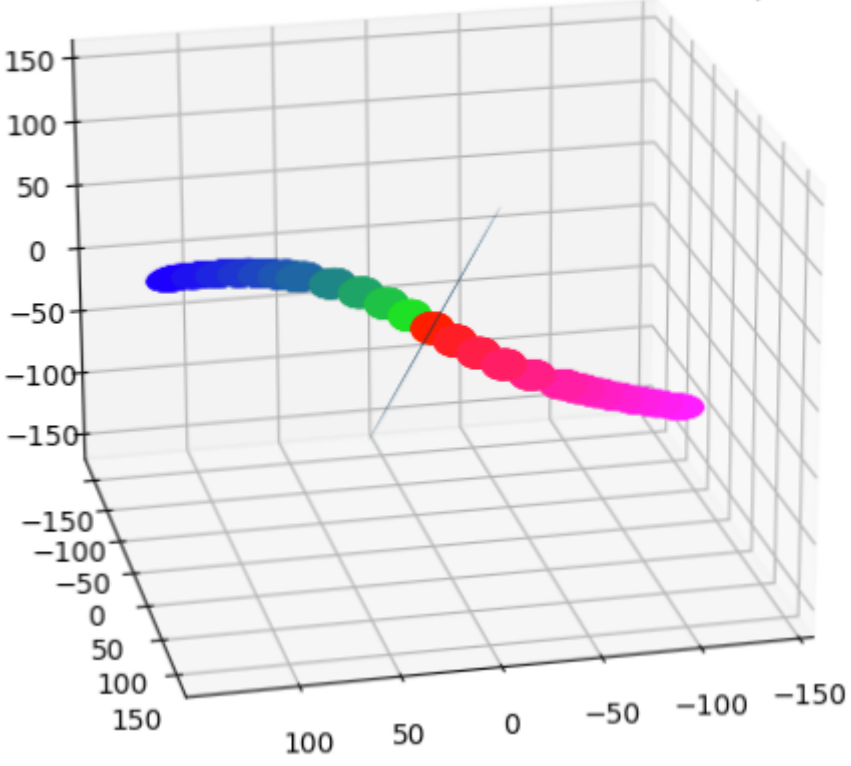
Tuteurs :
Charles DOSSAL (INSA)
Jose Pablo Baraybar (CICR)

Stagiaire :
Zoé Philippon

Fiche de synthèse :
Mise en place de critères sur la nature de rencontres spatiales

L'objectif principal de ce stage au LAAS-CNRS a été d'élaborer un tableau de bord d'indices permettant d'analyser la nature des rencontres spatiales afin de décider des hypothèses et des méthodes de calcul de la probabilité de collision. Le sujet comporte une réflexion théorique sur la nature des rencontres telles qu'elles sont décrites dans la littérature, l'analyse des critères de classification existant et leur généralisation éventuelle, la programmation sous Matlab de routines les calculant avec leur partie graphique.
Les premiers critères mis en avant sont les distances entre les deux objets en fonction du temps, la miss distance et la miss distance de Mahalanobis. Un autre critère utilisé est la probabilité de collision instantanée pour chaque instant de l'intervalle de temps considéré. Cependant, le critère visuel qui a demandé le plus de travail est la construction du Swept Volume. La construction du Swept Volume présenté dans ce rapport est faite selon deux méthodes. Le principe est d'exprimer les boules de collision sur plusieurs temps t et de les rétro-propager linéairement pour avoir une réunion d'ellipsoïdes.

Figure 1 : Représentation du Swept Volume
cas Alfano 7, intervalle de temps [-600,600] secondes
méthode linéaire sans incertitude sur la vitesse au temps de collision



Fiche de synthèse de
stage de 4ème année

Détection des signes de détresses de la population agricole

Younes Radi

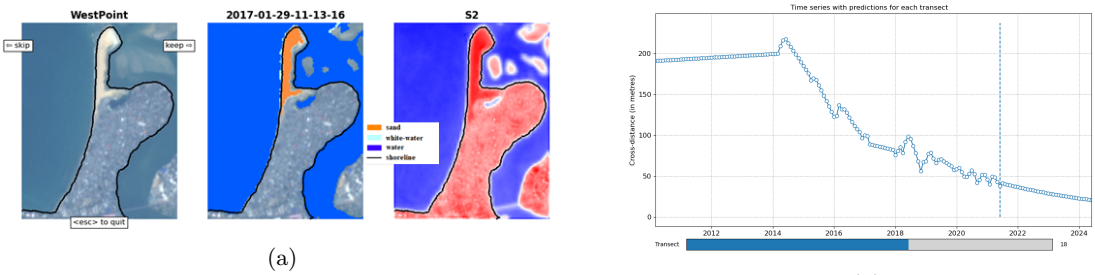
Les agriculteurs présentent un risque de décès par suicide plus élevé que l'ensemble de la population, qui s'est même accru en 2016, dernière année pour laquelle des données sont disponibles. Pour régler ce problème, l'IMSA souhaite trouver un dispositif pour assister les agents de la MSA dans la détection des signes de détresse de la population agricole et rurale. Le but du stage est de tester l'analyse de texte issue d'une base de données de la MSA pour détecter des signes de détresses.

Mes missions effectuées durant ce stage se résument en 3 axes. Le premier est le test des performances des services de transcription d'audio en texte pour potentiellement créer de nouvelles bases de données textuelles et détecter des signes de détresses. Le deuxième axe est l'anonymisation des données. Les données textuelles qui vont être traitées contiennent des informations personnelles des adhérents de la société. Pour être apte à traiter ces données, le RGPD recommande d'anonymiser les données pour le respect des droits et libertés des personnes. Et le dernier axe consiste en l'analyse de requêtes que les adhérents de la MSA envoient à cette dernière pour détecter des signes de détresses.

GIS and remote sensing data processing for change detection and environment degradation assessments

In Liberia's capital Monrovia, the West Point slum, located on a peninsula, is threatened by rising water levels due to climate change. The objective is then to develop a system to detect changes in the Liberian coastline that is able to predict the future evolution.

To address this issue, we developed a model that detects the position of the coast in previous years on satellite images and that uses deep learning to infer its future position. The detection of the coastline is obtained from a classification method that allows to associate each pixel of an image to a class among 'sand', 'white water', 'land' and 'water'. Of course, since the model is based on satellite images that were taken at different times, it is necessary to calibrate the measurements with a tides table. Regarding the prediction, it does not depend on any other data than the past position of the coastline. The future ones are computed by a Holt's linear trend model which is a forecasting method for time series. The previous shorelines detected on the satellite images need indeed to be discretized to build time series.



(a) (b)

Figure 1: (a) Land classification on S2 image of West Point (sand: orange, water: blue, white water: clear blue, shoreline: black) (b) Time series with prediction at the right of the vertical line

To visualize the results in the simplest way possible, a Google Earth Engine application seems to be quite appropriate and allows to alert quickly if there is a coastal shrinkage. It also makes visible where and how many people are threatened.

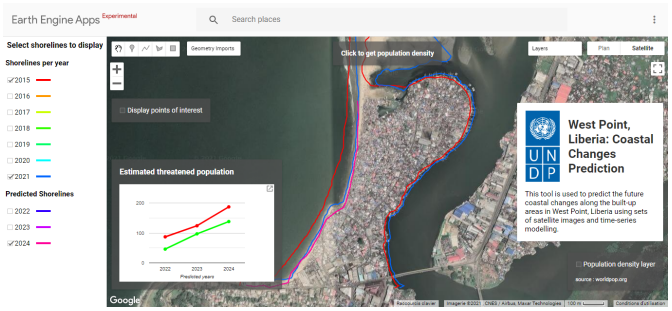


Figure 2: Google Earth Engine application

Gas Migration in Defective Cement

Folke Skrunes 5GMM

Unwanted fluid migration in oil and gas wells is a problem that has been diagnosed in a considerable number of producing and abandoned wells. This poses several risks for operators, including contaminated groundwater, decreased production, increased emissions of greenhouse gases and, in the worst case, major blowouts with catastrophic outcomes. Although the risk of leakage is lower for newer wells thanks to advances in cementation practices, once a well is subject to defective cement, it can be very costly to remediate.

Among others, the Norwegian Research Center (NORCE) is looking for ways to prevent gas leakage both in existing and newer wells. As part of this effort, we conducted what is known as a *tracer experiment* on a test section shown in Figure 1 in order to better understand the migration paths. The experiment consists of pumping a traceable substance, PTSA, through the system and measure the corresponding concentration of PTSA at the outlet, which is called the *breakthrough curve*.

Using theory from previous similar experiments, my main contribution was an attempt to fit the concentration curves to a class of probability distributions called *stable distributions*.

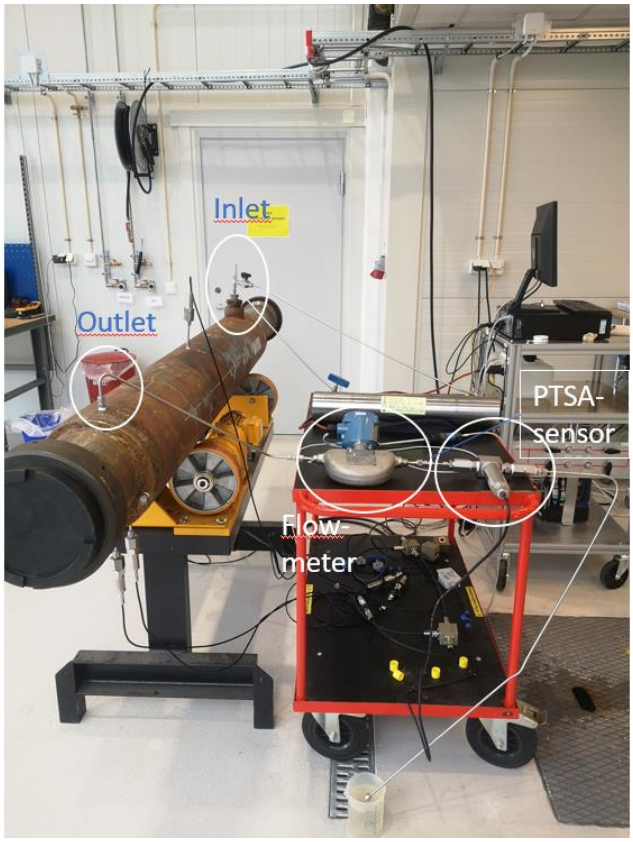


Figure 1

Analyse de la médiation et inférence causale en statistiques

Juan AYALA

Travail supervisé par M. Benoît LEPAGE, MCU-PH

Année scolaire 2020-2021

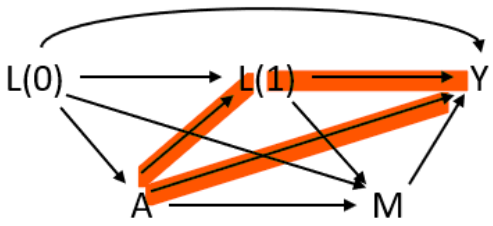
J'ai effectué mon stage de 4e année à la Faculté de médecine de Purpan à Toulouse, au sein de l'équipe EQUITY. Cette équipe travaille sur la santé des populations et sur l'épidémiologie du cours de la vie. Dans le cadre de l'analyse de la médiation et de l'inférence causale, mon tuteur et moi avons travaillé sur les estimations des effets directs et indirects randomisés d'une exposition sur une population. Ces effets sont définis respectivement par :

$rNDE = \mathbb{E}(Y_{A=1,\Gamma_{A=0|L(0)}}) - \mathbb{E}(Y_{A=0,\Gamma_{A=0|L(0)}})$ et $rNIE = \mathbb{E}(Y_{A=1,\Gamma_{A=1|L(0)}}) - \mathbb{E}(Y_{A=1,\Gamma_{A=0|L(0)}})$.

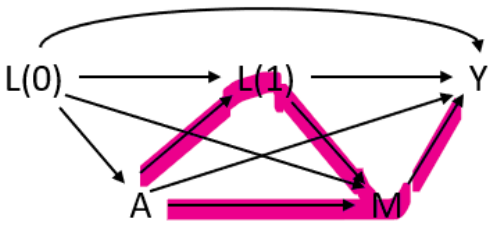
Dans les modèles causals ci-dessous, $Y_{A=a,\Gamma_{A=a^*}}$ correspond à la valeur contrefactuelle du critère de jugement Y sous un scénario fictif où l'exposition d'intérêt A prendrait la valeur $A = a$ pour toute la population, et où le médiateur M prendrait une valeur tirée au hasard dans la distribution attendue Γ de $M_{A=a^*}|L(0)$ sous le scénario contrefactuel $A = a^*$.

Nous avons d'abord étudié la performance de quelques estimateurs proposés pour estimer ces effets. Ces premiers sont par exemple la *g-computation* et la TMLE (*Targeted Maximum Likelihood Estimator*). Ensuite, en nous basant sur le travail de chercheurs en analyse de la médiation et en statistiques, nous avons mesuré ces effets sous différentes hypothèses et structures de bases de données.

Enfin, nous avons analysé l'apport de l'apprentissage automatique sur ces estimateurs avec des librairies de `R` telles que `SuperLearner` et `sl3`, dans le but d'améliorer leurs performances et ainsi réduire leur biais et leur sensibilité aux hypothèses testées.



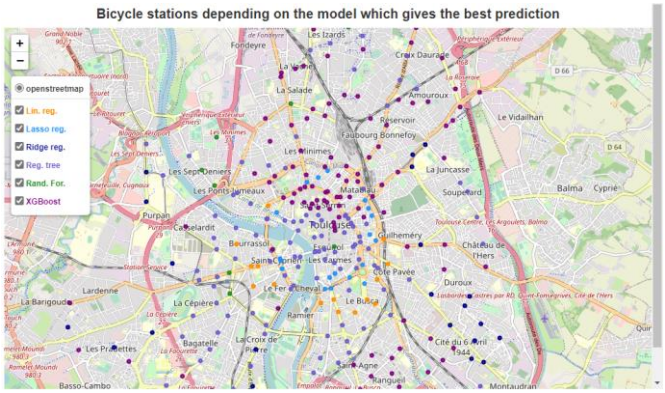
(a) Effet direct de A sur Y



(b) Effet indirect de A sur Y

Analyse de données - Création et exploitation d'une base de données sur la plateforme JENGA de CS

Le principal objectif du stage était de réaliser des cas d'utilisation d'analyse de données sur la plateforme JENGA de CS.



Le premier cas d'usage était la prédiction du nombre de vélos disponibles à une certaine station dans Toulouse. Des techniques d'apprentissage supervisé ont pu être mises en place, comme la régression linéaire multiple, la régression Lasso, la régression Ridge, les arbres de régression, les forêts aléatoires ou encore la méthode xgboost.

Le deuxième cas d'utilisation traitait de la génération de scénarios. L'étude de modèles stochastiques comme les chaînes de Markov cachées a permis la prédiction du cours des actions en bourse de deux sociétés. Ces mêmes modèles ont été appliqué à l'étude de l'évolution de la pandémie de la Covid-19. Finalement, une première approche aux modèles compartimentaux épidémiologiques a été effectuée.

Démarche globale du stage. La première étape a été de collecter des données provenant de diverses sources publiques. Ensuite, il a fallu agréger ces données selon les cas d'usage, les nettoyer et gérer le problème des données manquantes. Après une première analyse descriptive, différents phénomènes ont été modélisé selon les cas d'utilisation. Finalement, la dernière étape a consisté à intégrer tout le travail réalisé sur une plateforme en cours de développement par l'entreprise, et à l'exécuter dans cet environnement de type micro-services par le biais de containeurs Docker. Cette plateforme permettra à terme d'automatiser la création et la gestion de pipelines de données, tout en étant compatible avec diverses sources de données et en assurant la traçabilité des données.

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

adelyce

Benchmark

Classification et comparaison de clients par l'intermédiaire d'indicateurs clés

Adelyce, Labège (31)

Mots-clés : analyse de données, statistiques, classification non supervisée, réduction de dimension, ACP, données publiques, visualisation de données, segmentation de clients.

Adelyce est une jeune entreprise dynamique créée en 2007 et localisée à Labège (Haute-Garonne). Cette société de conseil a créé une solution experte de pilotage financier de la masse salariale des collectivités territoriales françaises.

Ce stage s'est déroulé au sein du pôle Recherche & Développement (R&D) d'Adelyce. Il consistait en la contribution à Benchmark, un vaste projet d'explorations statistiques sur les clients et prospects d'Adelyce. Un des objectifs consistait en la comparaison de la masse salariale des clients, en créant des clusters homogènes, par le croisement de ces données avec des indicateurs liés entre autres à la géographie, l'attractivité et le dynamisme économique du territoire. Une exploration de données exogènes était alors nécessaire. Cette comparaison constituait un début de recherche pour la conception d'un nouveau module proposé par le produit d'aide au pilotage de la masse salariale. Des comparaisons sur la masse salariale, mais aussi sur le profil des agents ont été menées.

Un des autres grands objectifs était de segmenter une part du marché d'Adelyce : les communes françaises dont la population était comprise entre 1 500 et 10 000 habitants. Des méthodes de clustering ont pu être utilisées, et les variables d'intérêt étaient multiples telles que les charges de personnel, la masse salariale et plusieurs indicateurs de dynamisme économique.

Les techniques usuelles de clustering, de réduction de dimension et de modélisation statistique ont été particulièrement utilisées. Les implémentations de ces projets ont été réalisées en R, et les résultats ont pu être visualisés et partagés aux autres services de l'entreprise *via* des markdowns, comprenant des interfaces graphiques (Figure 1). Des restitutions sous la forme de cartes interactives se sont également avérées utiles (Figure 2).

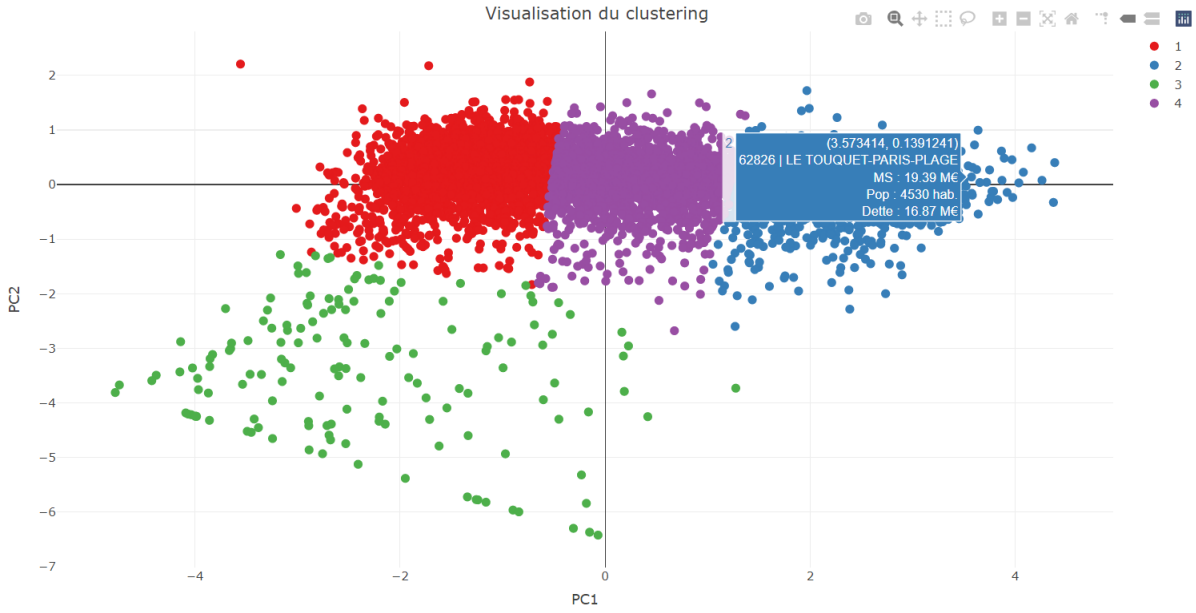


FIGURE 1 – Segmentation de clients - Visualisation d'un clustering K-means à l'aide de la librairie `plotly` de R.

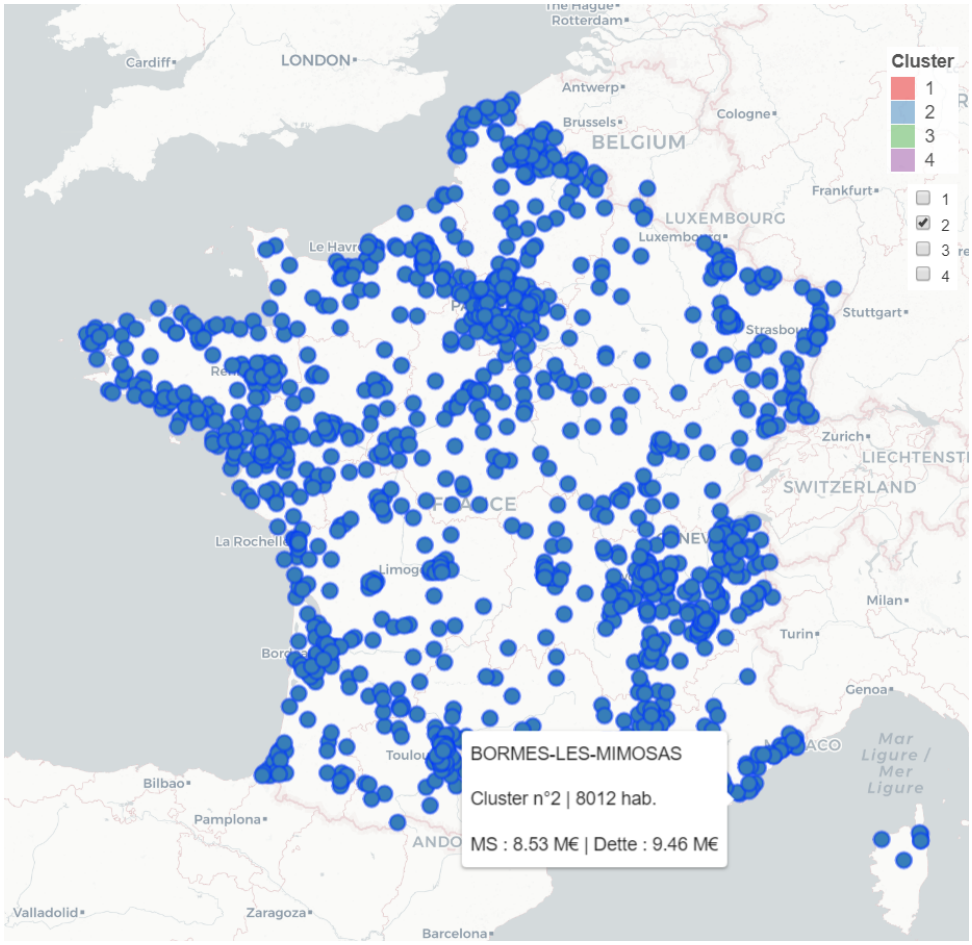


FIGURE 2 – Visualisation cartographique des communes appartenant à un cluster

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

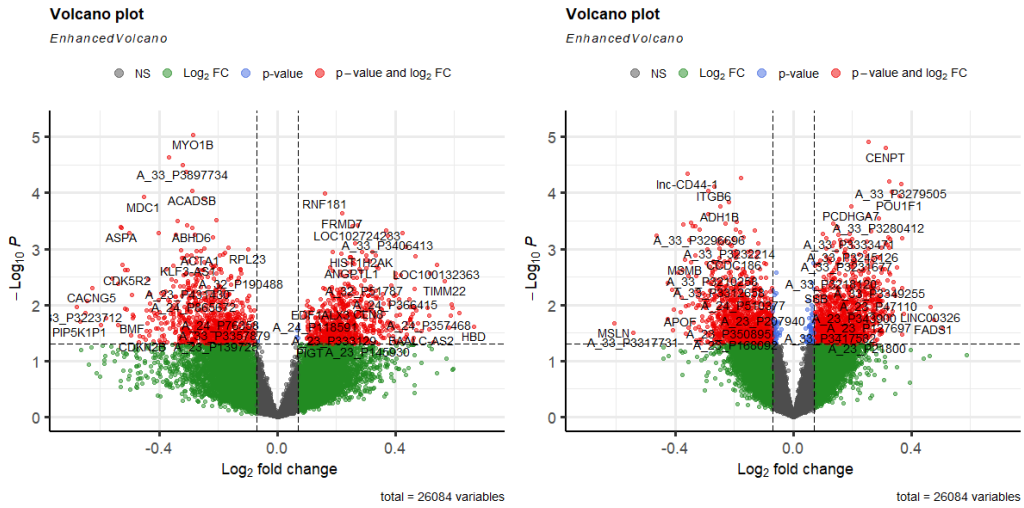
i2MC Institut des Maladies Métaboliques et Cardiovasculaires

Analyse de données biopuces – Recherche de gènes différentiellement exprimés au cours d'une intervention entre deux groupes d'individus obèses

L'obésité est un sujet important et d'actualité qui concerne la quasi-totalité de la planète. L'obésité, reconnue comme une maladie chronique par l'Organisation mondiale de la santé, est un excès de masse grasse et une modification du tissu adipeux. Cette maladie peut entraîner de nombreux problèmes de santé tels que le diabète, l'hypertension et des cancers pouvant réduire l'espérance de vie.

Pendant la quasi-totalité de mon stage au sein de l'Institut de Maladies Métaboliques et Cardiovasculaires (I2MC), j'ai travaillé sur le projet MONA (Metabolism Obesity Nutrition Age). Il s'agissait d'une étude comparative entre deux groupes d'hommes obèses, un groupe d'adultes jeunes (30 à 40 ans) et un groupe de seniors (60 à 70 ans). Les deux groupes ont participé à une intervention diététique de 8 semaines avec restrictions calorique modérée (déficit de 20% par rapport aux besoins énergétiques quotidiens) et un entraînement physique supervisé par un éducateur sportif : marche de 45 à 60 min, 5 fois par semaine. De nombreux échantillons de tissu adipeux et de muscles ont été utilisés pour quantifier l'expression des gènes à l'aide de la technologie biopuce.

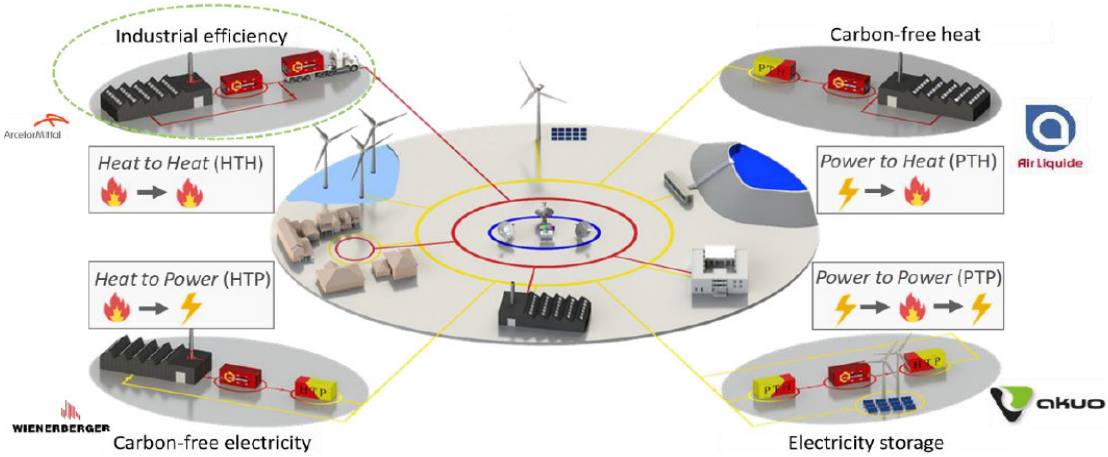
L'objectif étant de rechercher des gènes différentiellement exprimés j'ai tout d'abord réalisé une analyse exploratoire des données en faisant une analyse en composantes principales afin de détecter les tendances, identifier les effets non désirés puis les supprimer. Ensuite, pour trouver les gènes différentiellement exprimés avant et après l'intervention j'ai réalisé des tests-t de Student des données appariées et pour trouver les gènes différentiellement exprimés entre jeunes et seniors avant intervention j'ai utilisé un package qui s'appelle limma.



Volcano plot de la différence d'expression des gènes selon l'intervention pour tous les patients. Gauche : Données du muscle / Droite : Données du tissu adipeux. L'axe des abscisses représente le Fold Change (FC) en log2. L'axe des ordonnées représente les valeurs en -log10. En bleu les gènes dont les p-valeurs sont significatives, en vert les gènes dont les FC sont significatifs et en rouge les gènes correspondant aux deux conditions précédentes. Les gènes se trouvant sur la partie positive du log2(FC) sont les gènes surexprimés chez les seniors et sur la partie négative, les sous exprimés chez les seniors.

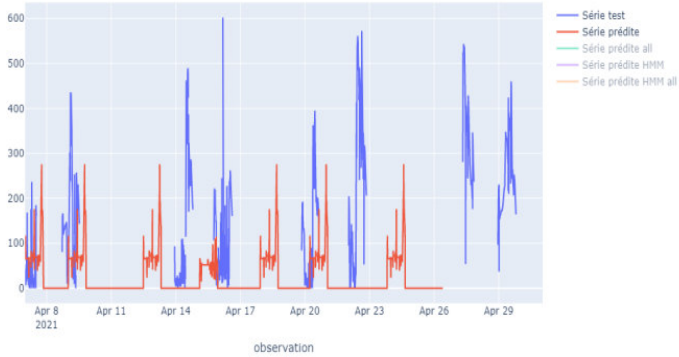
Fiche synthèse: MODÉLISATION ET OPTIMISATION D'UN RÉSEAU MULTI-ÉNERGIES DANS UNE VILLE BAS CARBONE

Au cours de mon stage, j'ai dû fournir des modèles de prédiction pour des séries temporelles de puissances des chaleur fatales dans le milieu industriel, tant dans un contexte de production que de consommation. Ces modèles permettraient de générer des scénarii prédisant les évolutions de puissance thermiques produites ou consommées à des fins de contrôle d'un réseau multi-énergies.

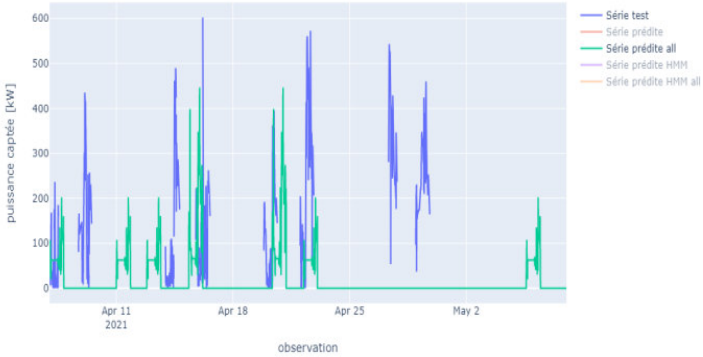


Les prévisions issus de ces modèles seraient données en entrée à un algorithme ayant pour but de fournir les bonnes commandes sur les stockages thermiques et électriques sur la base de ces données recréées. Ces prédictions substitueraient alors les séries temporelles de puissance sur les chaleurs fatales qui sont à ce jour encore peu accessibles dans le milieu industriel thermique.

Pour répondre à la problématique de prédiction dans le cas de production industrielle, un des modèles implémenté est basé sur une simulation de chaine de Markov discrète sur des profils moyens obtenu par clustering (CMT). Les méthodes de prédictions implémentées peuvent être utilisées pour de la planification sur des temps de prévisions sur un trimestre par exemple pour visualiser des possibles scénarii de production.



(a) Une simulation CMT



(b) Une simulation CMT données élargies

Au cours de mon stage, on a pu voir que les données d'apprentissages utilisées constituent une part importante de la performance des modèles de prédictions. Ainsi de manière plus globale, développer l'accès et le traitement de ces données serait un levier d'amélioration des prédictions de ressources énergétiques, thermiques ou pour d'autres sources. A ce jour, le développement de modèles de prédiction dans ce domaine est une question ouverte et des recherches sont faites dans ce sens.

Analyste d'affaires en service RH

Jiawen WU

Business Analysts (BA) sont responsables de l'analyse des exigences et de la numérisation des processus d'affaires qui étaient auparavant réalisés manuellement. Un processus complet de flux d'exigences devrait rassembler à ceci : L'utilisateur, le RH, et au cours de son travail quotidien, il estime que certains processus doivent être optimisés. Il communique donc ses exigences au BA. Après avoir compris les exigences, le BA doit d'abord juger rationnellement si cette exigence est raisonnable et nécessaire. Si les réponses sont positives, nous aideront l'utilisateur à mettre en œuvre l'exigence et à rédiger un document d'exigences clair et compréhensible. Le développement produit le code sur la base de ce document. Le BA assure le suivi et les teste, et enfin livré un processus correct à l'utilisateur.

Le BA est l'intermédiaire entre les RH et les développeurs. Ils doivent non seulement connaître la technologie, mais aussi comprendre l'affaires.

Pools de demande			Développement					Test			Disposition	
A planifier	BA	Planifié	Nom	A développer	En cours	Auto-test	Ajustements	Test	BA Test	UAT	A disposer	Disposé
Analysé	ZHAO		ZHENG									
	QIAN		WANG	2		3		4				
	SUN		FENG									
Non analysé	LI		CHEN									
	ZHOU		CHU									
	WU	1	WEI						5	6	7	8
Accord d'équipe			Amélioration de la qualité des travaux					Obstacles problématiques				
								9				

Nous gérons une exigence en cours de développement en utilisant un tableau d'affichage dans le diagramme ci-dessus.

La description du poste d'analyste d'affaires est la suivante :

- Améliorer l'expérience utilisateur du site web global de l'entreprise, participer à la planification des règles commerciales, promouvoir le système en ligne et l'optimisation continue.
- Collaboration interdépartementale, interface avec la commercialisation, le marketing et les besoins des clients.
- Conception de documents sur les exigences du produit, conception de prototypes, bonne planification et mise en œuvre des exigences.
- Test quotidien du site web et soutien opérationnel.