كلية العلوم و التقنيات فاس
ⵜⴰⵖⵙⵓⵍⵜ ⵏ ⵜⵙⴰⵡⴰⵍⵜ ⴷ ⵜⵓⵙⵙⵙⵜ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵡⴰⵍⵜ ⵓⵙⴰⵙ ⵎⵓⵃⵎⴰⴷ ⵏ ⵄⴰⴱⴷⵍⵍⴰ
Université Sidi Mohamed Ben Abdellah

Département d'Informatique

Projet de fin d'études

Master Sciences et Techniques
Systèmes Intelligents & Réseaux

Développement des méthodes métaheuristiques pour la sélection de variables pour les données spectrales

Lieu du stage : MAScIR

Réalisé par :

- Boubker Ayoub

Encadré par :

- Pr. Mrabti Fatiha
- Pr. Ben Abbou Rachid
- Mr. Rabie Reda

Soutenu le 14.07.2021 devant le jury composé de :

- Pr. Boushaba Adelali	Faculté des Sciences et Techniques de Fès	(Président)
- Pr. Benabbou Adil	Faculté des Sciences et Techniques de Fès	(Examinateur)
- Pr. Mrabti Fatiha	Faculté des Sciences et Techniques de Fès	(Encadrante)
- Pr. Ben Abbou Rachid	Faculté des Sciences et Techniques de Fès	(Encadrant)

Année Universitaire 2020 – 2021

Dédicace

A Mes chers parents, puisque rien au monde ne pourrait compenser les énormes sacrifices que vous déployez pour guider mes pas, et vos encouragements continus, qui me poussent à entreprendre mes années avec sérénité et confiance en soi. Mes amis et à tous ceux qui me sont chers, pour vos soutiens, votre compréhension et vos dévouements continus. Mes professeurs à La FSTF.

Remerciement

Tout d’abord, j’adresse mes remerciements à mes enseignants, et mes encadrants Pr MRABTI FATIHA et Pr BEN ABBOU Rachid de la Faculté des Sciences et Techniques Université Sidi Mohamed Ben Abdellah de Fès pour leurs aides, leurs encouragements, et leurs soutiens tout au long de mon cursus académique et mon stage.

Je tiens aussi à remercier mon encadrant à la fondation MAScIR, Mr RABIE Reda pour ses conseils, ses collaborations actives et aussi pour sa présence.

Mes remerciements également aux membres de jury d’avoir accepté de juger ce travail.

Mes vifs remerciements à toutes les personnes qui ont contribué de près ou de loin au succès de ce travail.

Résumé

La croissance des plantes aussi que leur rendement dépend étroitement de la qualité de certaines matières chimiques et organiques. Pour cette raison et depuis longtemps, les agriculteurs effectuaient des analyses dans les laboratoires pour mesurer les quantités de ces matières dans leurs plantes. Par contre ces analyses nécessitent beaucoup de temps. Une nouvelle alternative qui a été expérimentée ces dernières années consiste en l'estimation des taux des matières organiques et chimiques dans les plantes en se basant sur la spectroscopie infrarouge. Ces données spectrales sont d'une grande dimension, et présentent un ensemble des problèmes : corrélation, colinéarité, etc. Ce qui influence les performances des modèles de prédiction. Le projet consiste à faire une étude et conception d'un modèle hybride impliquant une combinaison d'une méthode d'optimisation et d'un modèle de prédiction. La construction du modèle hybride est basée sur l'approche enveloppe des méthodes de sélection de variables où le modèle de prédiction va permettre de tester les différents sous-ensembles de variables générés itérativement par les méthodes d'optimisation. Comme méthodes d'optimisation, nous avons utilisé en premier temps une méthode à base des algorithmes génétiques et une autre à base du recuit simulé, et enfin une méthode à base de colonie de fourmis. Or comme modèle de prédiction, nous avons utilisé la régression des moindres carrés partiels pour estimer les quantités du Carbone et d'Azote dans les tomates cerises. Les résultats présentés dans ce rapport ne sont qu'une implémentation des prototypes en attendant les résultats finaux d'après MAScIR afin de les comparer avec les résultats des travaux des autres membres du projet.

Mots clés : sélection de variables, spectroscopie, GA, SA, Colonie de fourmis

Abstract

The growth of plants as well as their yield depends closely on the quality of certain chemical and organic materials. For this reason and for a long time, the farmers carried out analyses in laboratories to measure the quantities of these materials in their plants. However, these analyses are very time consuming. A new alternative that has been tested in recent years is the estimation of the organic and chemical content of plants based on infrared spectroscopy. These spectral data are of a large dimension, and present a set of problems: correlation, collinearity, etc. This influences the performance of the prediction models. The project consists of a study and design of a hybrid model involving a combination of an optimization method and a prediction model. The construction of the hybrid model is based on the envelope approach of the variable selection methods where the prediction model will allow to test the different subsets of variables generated iteratively by the optimization methods. As optimization methods, we used first a method based on genetic algorithms and another based on simulated annealing, and finally a method based on ant colony. As a prediction model, we used partial least squares regression to estimate the amounts of Carbon and Nitrogen in cherry tomatoes. The results presented in this report are only an implementation of the prototypes while waiting for the final results from MAScIR in order to compare them with the results of the work of other members of the project.

Keywords: variable selection, spectroscopy, GA, SA, Ant colony

Tables des Matières

Listes des tableaux

Listes des figures

Liste des acronymes et abréviations

Acronyme	Signification
ACP	Analyse par composantes principales
AIC	Akaike Information Criterion
Ant Colony	Colonie de fourmis
GA	Genetics Algorithms
GPU	Graphics Processing Unit
IA	Intelligence Artificielle
MAScIR	Moroccan foundation for advanced science innovation and research
MSE	Mean Square Error
NIPALS	Non linear Iterative Partial Least Squares
NIR	Near Infra-Red
PLSR	Partiel least squares regression
RMSE	Root Mean Square Error
R^2	R square
SA	Simulated Annealing

Introduction générale

Le domaine de l'intelligence artificielle (IA) a connu au cours des dernières années une croissance fulgurante, autant au niveau de la recherche qu'au niveau de l'application dans différents domaines tels : le transport, l'agriculture, l'industrie, etc. Son objectif est de permettre aux dispositifs intelligents de prendre conscience de l'environnement. À cet égard, une machine utilisant l'IA exécute des tâches en imitant l'intelligence humaine.

Conscients des bénéfices que peuvent leur apporter les outils de l'intelligence artificielle, un nombre croissant d'entreprises voient en IA un moyen pour améliorer leur compétitivité. C'est le cas de MAScIR (Moroccan foundation for advanced science, innovation and research) qui a pris l'initiative d'implémenter les outils de l'intelligence artificielle dans le domaine clé pour le développement du Maroc : l'agriculture.

La croissance des plantes aussi que leur rendement dépend étroitement de la qualité de certaines matières chimiques et organiques. Pour cette raison et depuis longtemps, les agriculteurs effectuaient des analyses régulières de leurs plantes.

Les techniques traditionnelles d'analyse réalisaient au niveau du laboratoire, prenaient beaucoup de temps, et demandaient des ressources financières importantes. À cet égard, les chimistes se sont orientés vers le développement de techniques de plus en plus rapides et qui demandent un minimum de préparation des échantillons. Parmi les techniques modernes les plus prometteuses pour des analyses rapides directement sur site sont les méthodes de la spectroscopie infrarouge.

L'estimation de la quantité des matières chimiques et organiques dans les plantes peut être réalisée par l'utilisation des modèles de prédiction entrainés avec les données spectrales. Mais l'efficacité de ces outils de l'intelligence artificielle dépendent de leur bonne application, et en particulier de la qualité des données.

Le problème qui s'imposait c'est que les données spectrales contiennent souvent des milliers de variables. En effet il est presque rare que toutes les variables de l'ensemble de données soient utiles pour construire les modèles de prédiction. Effectivement les variables dans les données spectrales révèlent un haut niveau de corrélation, de colinéarité et de non pertinences, donc l'intégration de variables sans intérêt peut induire un bruit dans le modèle ce qui influence considérablement les performances obtenues [1].

Par conséquent, afin de tirer le meilleur parti de ce grand nombre de variables dans les données, l'utilisation de méthodes de sélection de variables est nécessaire. Ce processus va permettre la réduction de la dimension des données et donc les temps d'exécution, de même va permettre la réduction du sur-apprentissage et l'amélioration de la précision de prédiction.

La sélection de variables dans les données spectrales revient à sélectionner un sous-ensemble des longueurs d'onde ou des fréquences considérées par le processus comme pertinentes.

Notre travail consiste à :

- Filtrer les fréquences des spectres enregistrés et identifier celles qui peuvent donner les meilleures prédictions des taux de matières en se basant sur les méthodes d'optimisation : les algorithmes génétiques, le recuit simulé et les colonies de fourmis.
- Utiliser les algorithmes de prédictions pour estimer les taux du Carbone et Azote dans les tomates cerises en se basant sur la régression des moindres carrés partiels.

L'utilisation : des algorithmes génétiques, du recuit simulé et du colonie de fourmis était une contrainte technique imposée par MAScIR afin de les comparer avec d'autre méthodes de sélection de variables implémentées par d'autre membres du projet.

Le présent rapport s'articule autour de quatre chapitres, une introduction générale et une conclusion :

- Chapitre I : contient une vue globale sur l'organisme d'accueil : la fondation MAScIR, suivis par une présentation de la problématique et la solution proposée.
- Chapitre II : est dédié à l'état de l'art pour présenter les différents algorithmes utilisés.
- Chapitre III : est dédié à l'étude des données en présentant les analyses et les prétraitements effectués.
- Chapitre IV : contient l'implémentation des méthodes proposées et les résultats.

Chapitre I : Contexte général du projet

Introduction

Comme Objectif de développement de l'agriculture au Maroc, la fondation MAScIR essaye d'implémenter des nouvelles technologies de l'IA et de la spectroscopie infrarouge pour l'estimation des matières chimiques et organiques dans les plantes en se basant sur les composants Carbone et Azote dans les tomates cerises.

Dans cette partie, l'intérêt est porté au contexte général du projet. Nous commençons par présenter l'organisme d'accueil MAScIR, son fonctionnement, ses rôles et sa structure, ensuite nous abordons le projet en précisant le cadre général et ses objectifs par présenter la problématique et la solution proposée et finalement nous exposons la démarche suivie pour son déroulement.

1. Organisme d'accueil

1.1 Présentation de MAScIR

MAScIR est un organisme de recherche à caractère scientifique et technologique à but non lucratif en vue d'accompagner le développement du Maroc et participer au développement d'une nouvelle économie du savoir.

Fondée en 2007 par le Gouvernement Marocain, MAScIR est voué à la recherche en nanotechnologie, en biotechnologie, en technologie numérique, en microélectronique, en énergie et en environnement.

MAScIR est un contributeur clé du développement de la science et la technologie au Maroc, grâce à son investissement dans un capital humain, de chercheurs et ingénieurs qui œuvrent dans des domaines aussi innovants que complémentaires.

La mission de MAScIR est de promouvoir l'excellence dans la recherche et le développement des technologies au MAROC dans le but de générer de la valeur, des emplois à travers l'intégration dans le marché des technologies avancées dans le but de produire de la propriété intellectuelle et de créer des spin-offs.

MAScIR compte aujourd'hui 3 pôles :

- MAScIR MicroElectronics and Packaging
- MAScIR BioTechnology
- Nano Technology

1.2 Département Microélectronique et Packaging

MAScIR MicroElectronics est un centre d'innovation et de développement des technologies

dans le milieu microélectronique. Il se concentre sur le micro packaging, l'ingénierie, les tests de simulation, le design, la qualification, le prototypage de produits micro-électroniques et les systèmes embarqués. Parmi les projets traités par MAScIR MicroElectronics, nous citons :

- Design et Micro packaging CSP et PILR,
- Tests de fiabilité sur les packages
- Circuits embarqués sur une application wafer level camera fabriquée au Maroc par Nemotek Technologie.

MAScIR MicroElectronics fournit des services pour des clients industriels, mais elle développe aussi son propre business dans les domaines suivants :

- L'intégration et la miniaturisation des systèmes microélectroniques.
- L'analyse de fiabilité et défaillance des produits.
- Modélisation des systèmes complexes.
- Prototypage et industrialisation des produits innovants.
- Industrialisation des idées et résultats académiques.

L'organigramme du département Microélectronique est présenté dans la figure 1:

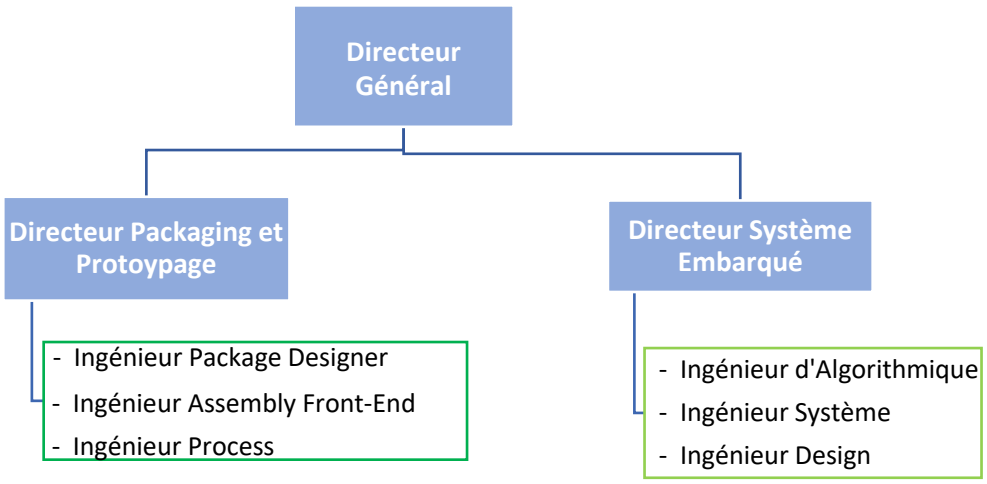

Figure 1 : Organigramme du département Microélectronique

2. Problématique

Les domaines : l'intelligence artificielle, la science de données et la chimiométrie représentent un centre d'intérêt pour les différentes équipes de chercheurs au sein de MAScIR.

La croissance des plantes aussi que leur rendement dépend étroitement de la qualité de certaines matières chimiques et organiques. Pour cette raison et depuis longtemps, les

agriculteurs effectuaient des analyses régulières de leurs plantes. Ces techniques traditionnelles d'analyse réalisaient au niveau du laboratoire, prenaient beaucoup de temps, et demandaient des ressources financières importantes.

Dans ces thématiques, le projet s'intéresse au développement d'un analyseur qui permettra d'estimer la quantité des composants Carbone et Azote dans les tomates cerises en utilisant des techniques modernes, rapides et directement sur site en se basant sur la spectroscopie infrarouge et les outils de l'intelligence artificielle.

L'utilisation des modèles de prédiction et les mesures par spectre proche infrarouge pour l'estimation des matières chimique permettent souvent de réaliser des économies liées à l'amélioration du contrôle et de la qualité des produits et peuvent fournir des résultats beaucoup plus rapidement que les analyses traditionnelles en laboratoire.

Cependant les mesures par spectre proche infrarouge peuvent révéler des problèmes potentiels tels : colinéarité, corrélation, etc. Donc il est nécessaire de promouvoir des actions correctives afin de dépasser ces problèmes en améliorant les performances des modèles de prédiction.

La spectroscopie infrarouge s'appuie sur l'absorption du rayonnement électromagnétique aux longueurs d'onde dans la gamme de 780 à 2 500 nm ce qui rend les informations contenues contiennent un grand nombre de variables par rapport au nombre des échantillons (figure 2) ce qui influence les performances des modèles de prédiction.

	N	350	351	352	353	354	355	356	357	358	...	2491
1	0.335550	0.685202	0.703378	0.707244	0.700635	0.710577	0.712560	0.717379	0.726715	0.725688	...	0.652016
2	1.390083	1.211069	1.220758	1.177340	1.153564	1.207712	1.223311	1.230584	1.240578	1.209556	...	0.631960
3	4.606212	1.141793	1.118914	1.150225	1.198876	1.146297	1.200844	1.236696	1.208364	1.247691	...	0.966450
4	2.614611	1.088999	1.105354	1.099355	1.091881	1.109881	1.113219	1.107186	1.103721	1.113831	...	0.817366
5	1.167868	0.816191	0.822812	0.830625	0.834704	0.833480	0.838764	0.844735	0.846843	0.844858	...	0.380068
...	...	...	...	...	...	...	...	...	...	...	...	..
466	1.701524	0.712458	0.717163	0.722574	0.733401	0.752346	0.764106	0.768740	0.774272	0.796959	...	0.447097
467	3.895315	0.838513	0.833610	0.856102	0.888355	0.868371	0.880126	0.894986	0.892708	0.880606	...	0.401016
468	1.650202	0.686904	0.688760	0.695874	0.703717	0.697182	0.712483	0.722287	0.718949	0.729298	...	0.485390
469	2.787011	0.958575	0.978741	0.949793	0.923068	0.960016	0.960689	0.957064	0.973580	1.004691	...	0.534545
470	2.715643	0.892809	0.892730	0.883102	0.875398	0.905165	0.918359	0.939804	0.970134	0.959413	...	0.551340

470 rows × 2152 columns

Figure 2 : Une partie de la BD d'Azote

Ces variables représentent les longueurs d'onde ou les fréquences dans les données spectrales révèlent une forte corrélation et une multi-colinéarité entre ces variables et ça revient à l'existence des interférences entre les ondes. Une carte thermique des coefficients de corrélation de Pearson [2] est présentée dans la figure 3 pour mesurer l'intensité de liaison linéaire entre chaque couple de variables.

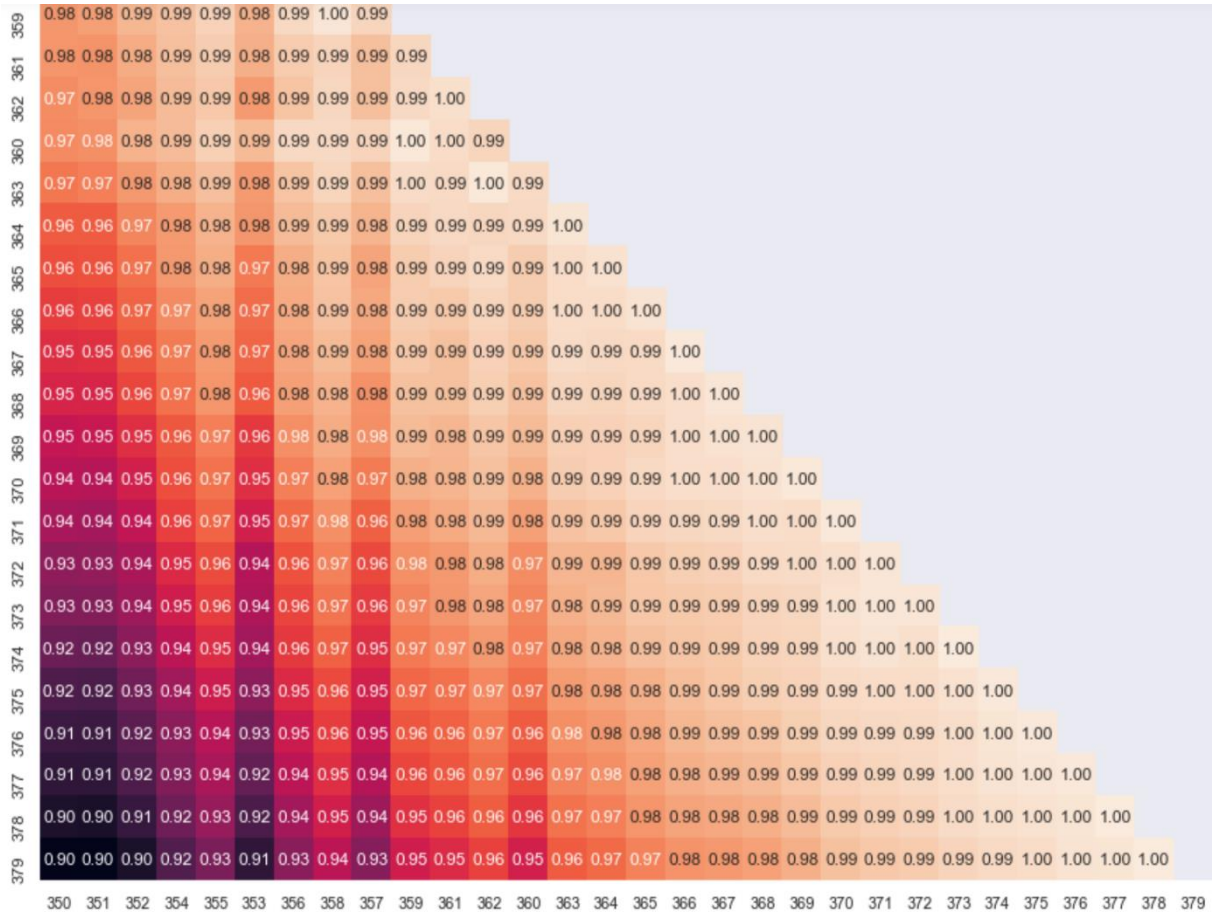

Figure 3 : Carte thermique des coefficients de corrélation de Pearson

D'après la figure 3, les coefficients de corrélations sont proches de 1 ce qui explique la forte corrélation entre les variables. Ces résultats montrent l'existence de la redondance des informations et des variables dans les données.

Dans la spectroscopie infrarouge, l'identification des groupements chimiques se fait pour chaque élément chimique en se basant sur le principe que chaque groupement absorbe la lumière différemment en fonction de sa longueur d'onde. Puisque le spectre proche infrarouge présente les informations de toutes les longueurs d'onde dans la gamme de 780 à 2 500 nm, les longueurs d'ondes où l'élément chimique à estimer n'absorbe pas de lumière n'offrent aucune information utile. Ce qui signifie l'exitance des variables non pertinentes dans les données. Pour vérifier ce constat, nous avons effectué un test statistique où nous avons vérifié la signification des coefficients d'un modèle de prédiction basé sur la régression.

Dans ce test, l'hypothèse H_0 suppose que le coefficient lié à la variable n'est pas significatif et ne présente aucune importance au modèle de prédiction et peut affecter les performances du modèle de prédiction. La figure 4 représente les coefficients calculés du modèle de régression des moindres carrés partiels (Partial Least Squares Regression, PLSR) en utilisant les données d'Azote, et la figure 5 représente les valeurs du **p_value** du test statistique **JackKnife** [3] avec $\alpha = 0.05$.

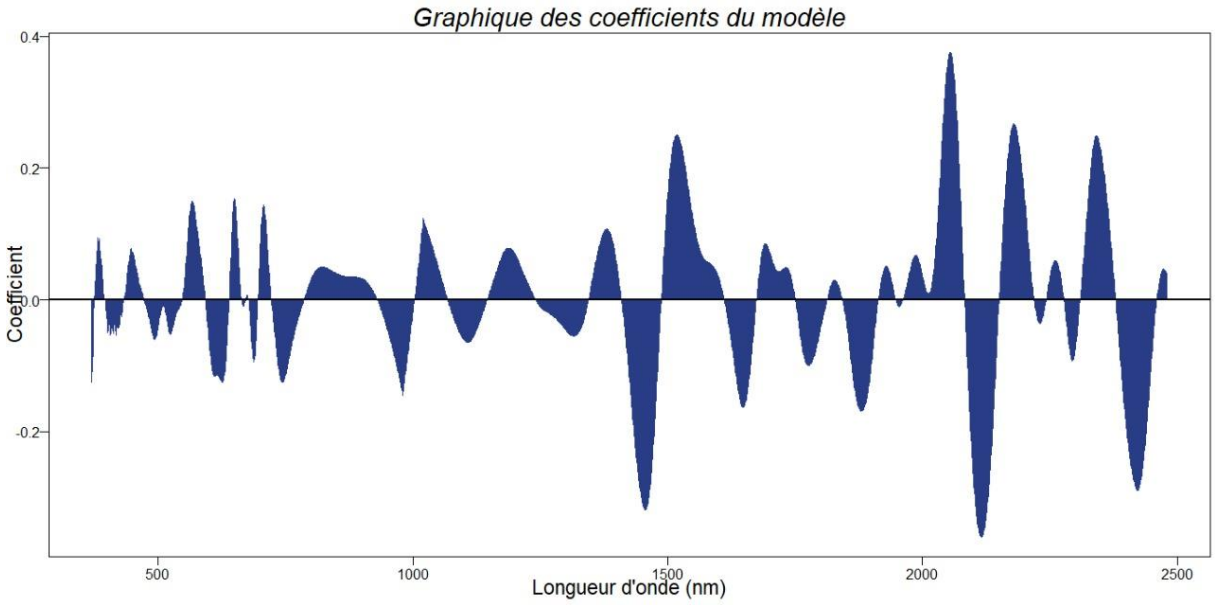

Figure 4 : Graphe représentant les coefficients du modèle de PLSR

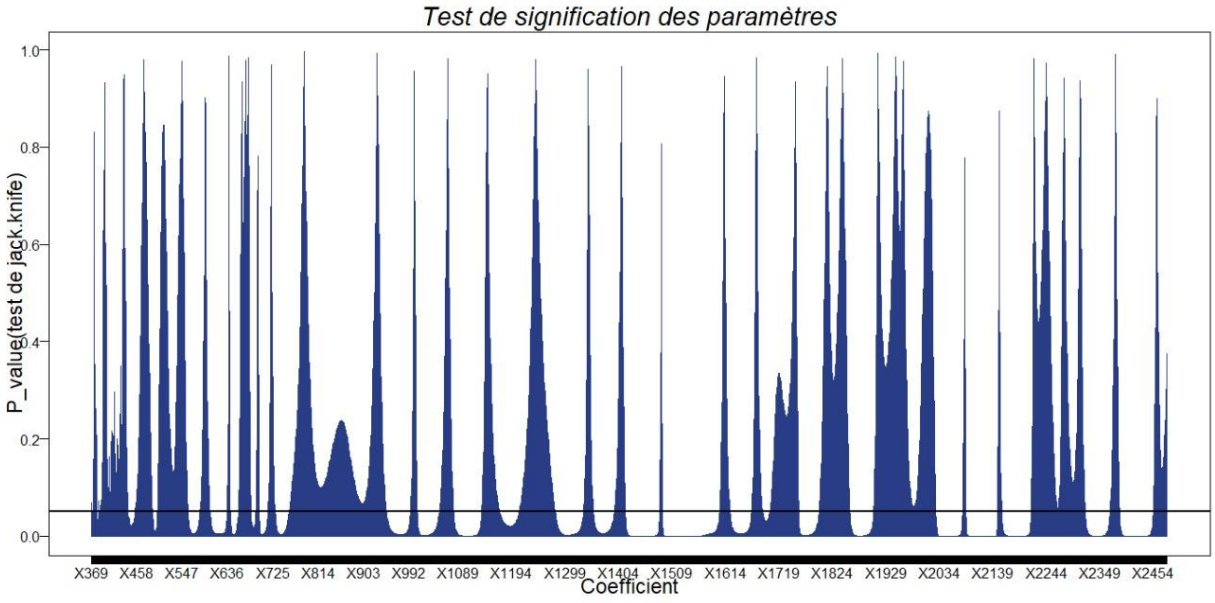

Figure 5 : Test de signification des coefficients du modèle PLSR

D'après la figure 5, les coefficients qui ont une **p_value** < 0.05 vérifie l'hypothèse H_0, ce qui montre l'existence des variables non pertinentes dans les données.

En général, la spectroscopie NIR est utilisée en combinaison avec des techniques multivariées pour une analyse qualitative ou quantitative. Le grand nombre de variables spectrales dans la plupart des ensembles de données rencontrées en chimiométrie spectrale rend souvent la prédiction d'une variable dépendante compliquée (l'élément chimique à prédire), cependant par l'utilisation de techniques de projection ou de sélection appropriées, le problème peut être minimisé [1].

Les données spectrales révèlent un ensemble de problèmes tels : la colinéarité, la corrélation, la grande dimension des variables par rapport aux nombres des échantillons et l'existence des variables non informatives dans ces données. Pour lutter contre ces problèmes, nous sommes amenés à développer des méthodes d'optimisation permettant la sélection de variables.

3. Solution

La sélection des variables ou sélection des longueurs d'onde, lorsqu'elle est appliquée aux données spectroscopiques, est une étape très importante de l'analyse des données, car l'élimination des variables non informatives, inutiles, bruyantes et redondantes ou corrélées produira une meilleure prédiction et des modèles plus simples [1] . Il est maintenant largement admis qu'une sélection de variables bien effectuée peut aboutir à des modèles ayant une plus grande capacité de prédiction [4].

Pour lutter contre ces problèmes tels : corrélation, colinéarité, etc. Nous avons proposé de développer un modèle hybride qui repose sur la combinaison des méthodes de sélection de variables pour filtrer les fréquences des spectres enregistrés et identifier celles qui peuvent donner les meilleures prédictions des taux de matières et des modèles de prédiction pour estimer la quantité des matières chimiques.

Une recherche exhaustive de tous les sous-ensembles peut être réalisée si le nombre de variables est faible mais avec ce grand nombre de variables, un algorithme exhaustif devient impossible à mettre en place. C'est pourquoi des méthodes d'optimisation et en particulier des métaheuristiques ont été proposées pour aborder ce problème, et permettre de trouver de très bons sous-ensembles à défaut de trouver le meilleur.

La phase d'analyse de données et de la recherche nous a permis de choisir la régression des moindres carrées partiels (PLSR) comme modèle de prédiction. Or l'utilisation : des algorithmes génétiques, du recuit simulé et de la colonie de fourmis était une contrainte technique imposée par MAScIR afin de les comparer à d'autres méthodes de sélection de variables implémentées par les autres membres du projet.

4. Planification du projet

Pour garantir le bon déroulement du projet nous avons dû appliquer une planification. Cette dernière va nous permettre de définir les travaux à réaliser, synchroniser les actions, diminuer des risques et tracer l'état d'avancement du projet en se basant sur le diagramme de Gantt.

Ce projet est divisé en 4 phases :

- Phase d'intégration et recueil des informations :

  Dans cette phase, nous avons étudié le sujet du projet en se familiarisant avec les

concepts de la spectroscopie et les méthodes de la sélection de variables, les caractéristiques et les outils existants dans ces champs.

- Phase d'analyse et traitement des données :

  Cette phase consiste à analyser les données et faire les prétraitements nécessaires afin de préparer les données.

- Conception du modèle :

  Cette étape consiste à la recherche et la comparaison des solutions existantes : les algorithmes pour la sélection de variables et les algorithmes d'estimation.

- Réalisation :

  Cette dernière partie du projet consiste à implémenter notre modèle hybride et voir les résultats obtenus.

Le figure 6 ci-dessous présente le diagramme de Gantt du projet :

Nom	Date de début	Date de fin
Phase d'intégration	22/02/2021	20/03/2021
Analyse du projet	22/02/2021	26/02/2021
Etat de l'art de la spectroscopie	27/02/2021	05/03/2021
Etudes sur les outils de traitements des données spectralles	01/03/2021	05/03/2021
Etat de l'art des méthodes de la sélection de variables	06/03/2021	20/03/2021
Etat de l'art des méthodes de régression	12/03/2021	20/03/2021
Phase d'analyse et traitement de données	21/03/2021	09/04/2021
Phase de Conception du modèle	10/04/2021	23/06/2021
Etudes des algorithmes de prédiction	10/04/2021	04/05/2021
Etude des algorithmes métaheuristiques	15/04/2021	23/06/2021
Phase de réalisation	25/05/2021	23/06/2021
Rédaction du rapport	10/05/2021	28/06/2021

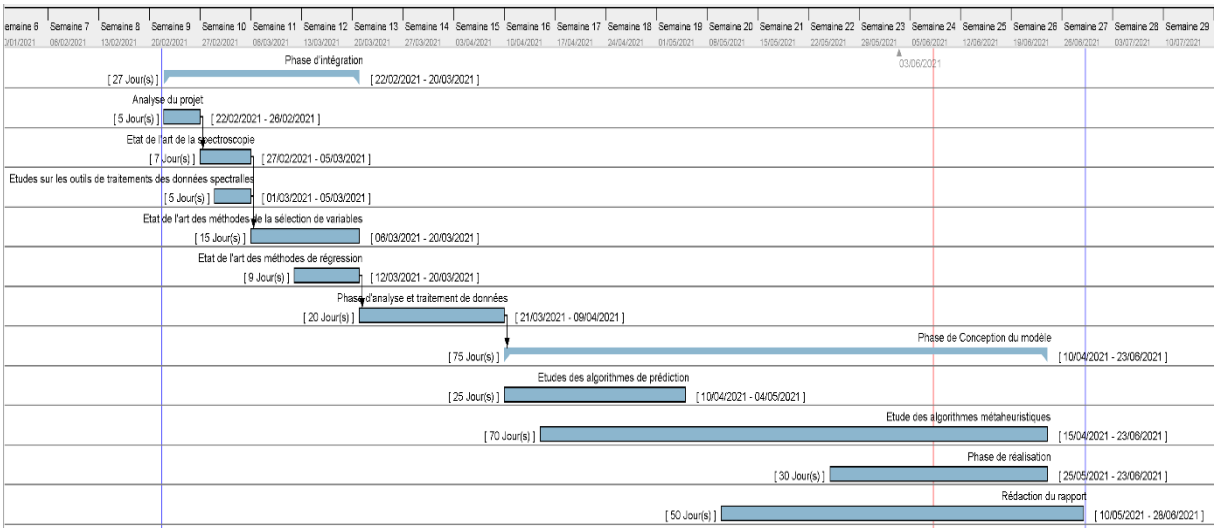

Figure 6 : Digramme de Gantt

Conclusion

La sélection de variables dans les données spectrales est une étape très importante de l'analyse des données pour l'identification et la suppression des variables non informatives, bruyantes et redondantes. Les méthodes d'optimisation sont un moyen efficace de traiter le problème de sélection de variables. Ces méthodes seront présentées dans le chapitre suivant.

Chapitre II : Etat de l'art de la sélection de variables et l'optimisation

Introduction

Depuis longtemps, les agriculteurs effectuaient des analyses dans les laboratoires pour mesurer les quantités des matières chimiques et organiques dans leurs plantes. Par contre ces analyses nécessitent beaucoup de temps. Une nouvelle alternative qui a été expérimentée ces dernières années consiste en l'estimation des taux des matières organiques et chimiques dans les plantes en se basant sur la spectroscopie infrarouge. Ces données spectrales sont d'une grande dimension, et présentent un ensemble des problèmes : corrélation, colinéarité, etc. Ce qui influence les performances des modèles de prédiction.

Notre objectif est de développer un modèle hybride impliquant une combinaison d'une méthode d'optimisation : les algorithmes génétiques, recuit simulé et colonie de fourmis, et d'un modèle de prédiction : PLSR afin d'estimer les quantités du Carbone et de l'Azote dans les tomates cerises en dépassant les problèmes imposés par les données spectrales en effectuant une sélection de variables.

1. Sélection de variables

Lors de la construction d'un modèle d'apprentissage automatique, il est presque rare que toutes les variables de l'ensemble de données soient utiles pour construire le modèle. Alors il est souhaitable de réduire le nombre de variables d'entrée à la fois pour réduire le coût de calcul de la modélisation et pour améliorer les performances du modèle.

1.1 Principe de la sélection de variables

La sélection de variables est un domaine de recherche actif en statistiques et science de données. Le principe général consiste à choisir un sous-ensemble de variables en éliminant les variables qui ont peu ou pas d'influence sur l'information que l'on souhaite prédire [5].

Ce processus est particulièrement adapté pour réduire la dimension des données et donc les temps d'exécution. Il permet aussi de réduire le sur-apprentissage et d'améliorer la précision de prédiction puisque l'intégration de variables sans intérêt peut induire un bruit dans le modèle. Enfin, le modèle final est généralement plus facilement interprétable avec un nombre restreint de variables [6].

Les données d'entrée du processus sont constituées par l'ensemble initial de variables qui forment l'espace de représentation et l'ensemble des données d'apprentissage du problème étudié. Le processus de sélection de variables se décompose de la manière suivante, figure 7 :

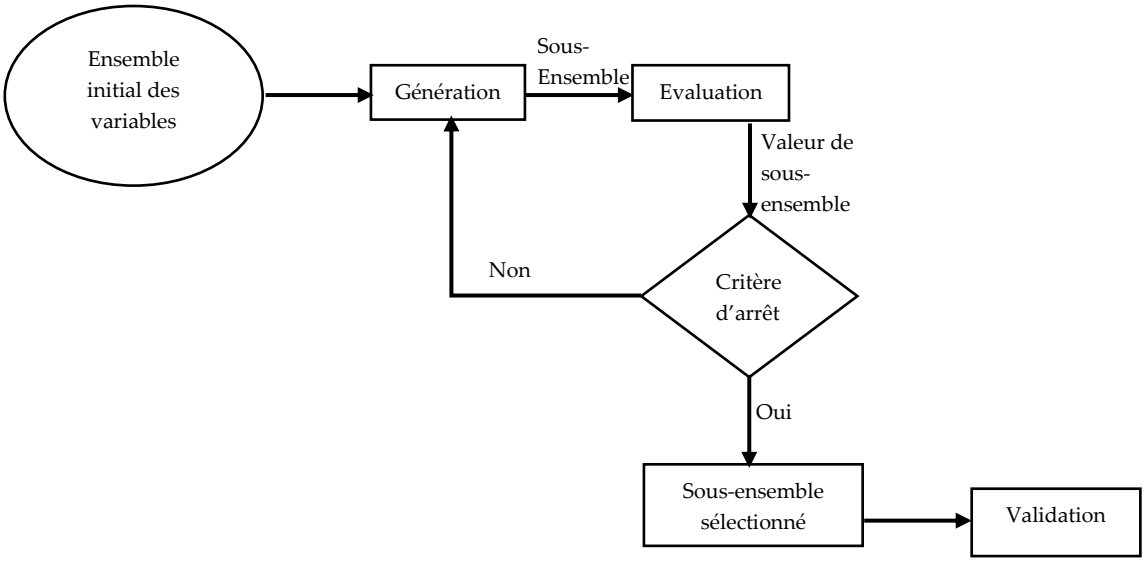

Figure 7 : Processus de sélection de variables

Les principaux enjeux et conséquences de la sélection de variables sont divers [5] tels :

- Permettre de déterminer les variables considérées comme pertinentes.
- Permettre de supprimer les variables redondantes et le bruit généré par certaines variables.
- Permettre la réduction de la taille de l'espace de représentation.
- Permettre la réduction du coût de calcul dans la phase d'apprentissage.

1.2 Catégories des algorithmes de sélection de variables

Les méthodes de sélection de variables sont classiquement classées en fonction de la manière dont elles combinent l'algorithme de sélection et la construction du modèle :

- **Les méthodes filtre** (filters methods) : Ces méthodes génèrent un sous-ensemble de variables qui est soumis à une fonction d'évaluation propre à chaque méthode filtre, et en fonction du résultat, le sous-ensemble est considéré ou non comme optimal [5] (figure 8).
- **Les approches enveloppes** (wrappers methods) : Ces méthodes utilisent l'algorithme d'apprentissage comme fonction d'évaluation. Elles permettent la génération itérative de sous-ensembles de variables. L'algorithme d'apprentissage va permettre de tester les différents sous-ensembles de variables générés. Il intervient donc au sein même du processus de sélection de variables [5] (figure 8).
- **Les approches intégrées** (embedded methods) : Ces méthodes sont intégrées à un modèle et lui sont spécifiques (les forêts de décision aléatoires) [5] (figure 8).

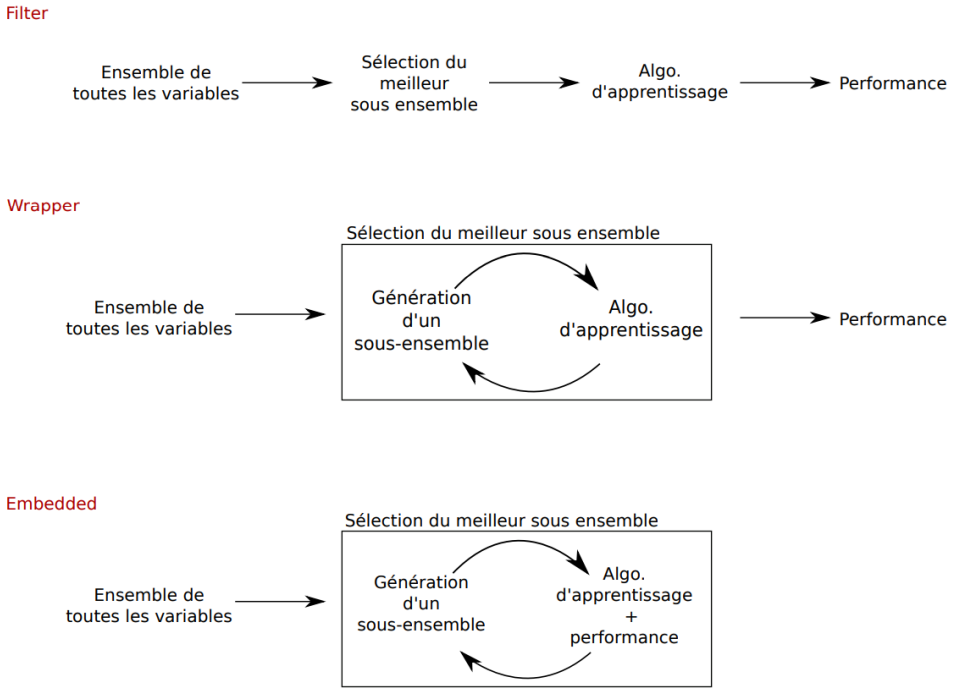

Figure 8 : Les approches filter, wrapper et embedded

1.3 Optimisation combinatoire pour la sélection de variables

Un problème d'optimisation combinatoire consiste à trouver la meilleure solution dans un ensemble discret dit ensemble des solutions réalisables. En général, cet ensemble est fini mais compte un très grand nombre d'éléments, et il est décrit de manière implicite, c'est-à-dire par une liste, relativement courte, de contraintes que doivent satisfaire les solutions réalisables.

Les méthodes d'optimisation sont un moyen efficace de traiter le problème de sélection de variables. En effet, rechercher un sous-ensemble de variables pertinentes, revient à se poser la question du meilleur sous-ensemble par rapport à un échantillon d'évaluation donné [7].

Une recherche exhaustive de tous les sous-ensembles peut être réalisée si le nombre de variables est faible. Cependant, dans des projet les jeux de données contiennent un nombre important de variables. Dans ce contexte de grande dimension, le problème est connu pour être NP-complet [8] dont un algorithme exhaustif devient impossible à mettre en place. C'est pourquoi des méthodes approchées et en particulier des méthodes métaheuristiques ont été proposées pour aborder ce problème, et permettre de trouver de très bons sous-ensembles à défaut de trouver le meilleur [6]. L'article [3] présente une utilisation des méthodes métaheuristiques pour faire une sélection de variables dans la spectroscopie infrarouge pour la caractérisation biologique du sol et des vers de terre.

Utiliser des méthodes d'optimisation pour la sélection de variables nécessite de définir comment combiner la méthode d'optimisation et le critère d'évaluation permettant de mesurer la qualité du modèle.

2. Les méthodes métaheuristique d'optimisation

De nombreux problèmes d'optimisation sont NP-complet [8] et que nous ne disposons pas d'un algorithme en temps polynomial pour le résoudre, ou de méthode classique efficace. Donc on considère des méthodes qui fournissent rapidement une solution réalisable, pas nécessairement optimale, pour un problème d'optimisation NP-difficile [8] comme les méthodes métaheuristiques.

Les métaheuristiques sont des algorithmes approximatifs et généralement non déterministes, permettant d'obtenir une solution de bonne qualité pour un problème d'optimisation difficile. L'objectif est d'explorer efficacement un espace de recherche afin d'obtenir une solution proche de la solution optimale. Un grand nombre de métaheuristiques existent et peuvent être regroupées en deux catégories : les métaheuristiques à base voisinage et les métaheuristiques à base de populations [6] (figure 9).

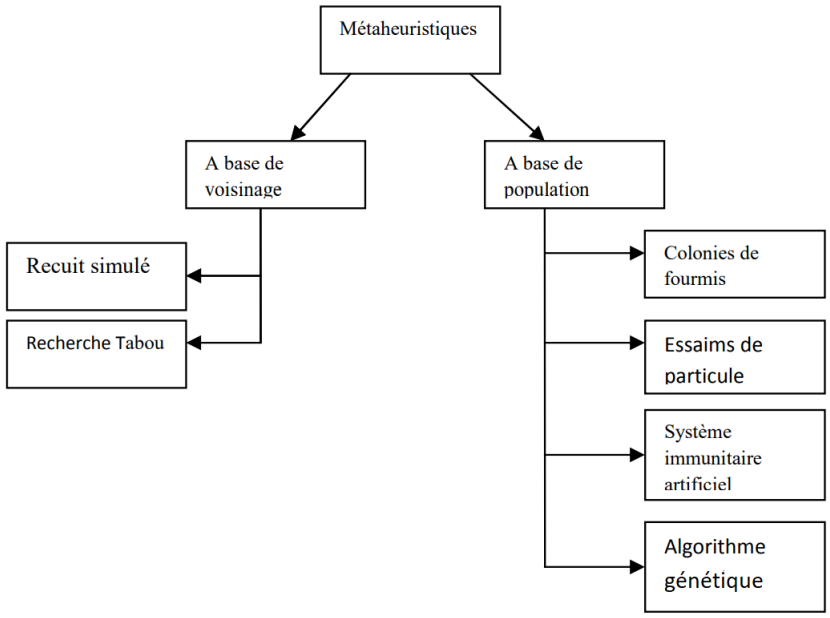

Figure 9 : Répartition des métaheuristiques

- **Métaheuristique à base voisinage**

Les métaheuristiques à base de voisinage sont les méthodes de recherche locale. Elles sont itératives : à partir d'une solution unique x_0 considérée comme point de départ, la recherche consiste à passer d'une solution à une solution voisine par déplacements successifs dans un voisinage constitué de l'ensemble des solutions [6]. C'est le principe adopté pour la recherche tabou et le recuit simulé.

- **Métaheuristique à base population**

Les méthodes de recherche à base de population travaillent sur un ensemble de points de

l'espace de recherche ou population de solutions. Ces métaheuristiques sont inspirées de la biologie et utilisent des phénomènes d'auto organisation. Parmi les métaheuristiques à base de populations, les algorithmes à base d'essaims sont basés sur le comportement en collectivité de certaines espèces comme les fourmis ou les abeilles [6].

2.1 Les algorithmes génétiques

Les algorithmes génétiques (Genetics Algorithms GA) ont été développés par John Holland, ses collègues et ses étudiants à l'Université du Michigan. Ce sont des programmes informatiques inspirés des processus de l'évolution biologique pour résoudre des problèmes et modéliser des systèmes évolutifs.

Les GA sont des algorithmes itératifs de recherche globale (figure 10) dont l'objectif est d'optimiser une fonction prédéfinie appelée fonction coût ou fonction « fitness » [9].

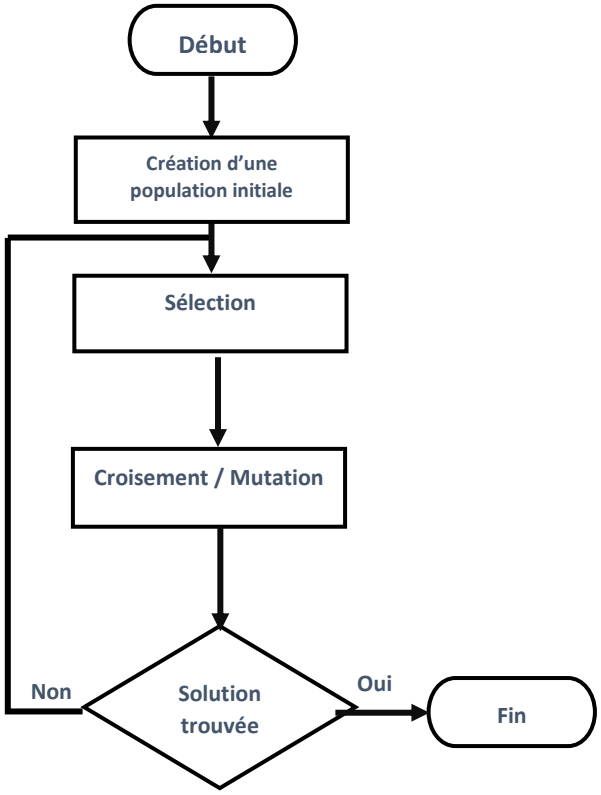

Figure 10 : Principe général des algorithmes génétiques

Selon Lerman et Ngouenet (1995) un algorithme génétique est défini par [9] :

- Environnement : l'espace de recherche.
- Population : un ensemble de chromosomes ou de points de l'espace de recherche.
- Individu / chromosome / séquence : représente une solution potentielle au problème.
- Fonction de fitness : la fonction que nous cherchons à maximiser ou minimiser.

Avant de passer vers la phase de la sélection, il faut choisir comment représenter les données

en spécifiant le codage à utiliser.

- **Codage**

Cette étape associe à chacun des points de l'espace d'état une structure de données. Elle se place généralement après une phase de modélisation mathématique du problème traité. Le choix du codage des données conditionne le succès des algorithmes génétiques. Les codages binaires ont été très employés à l'origine. Les codages réels sont désormais largement utilisés, notamment dans les domaines applicatifs, pour l'optimisation de problèmes à variables continues.

- **La Sélection**

Après avoir calculé la valeur d'aptitude de chaque individu de la population, un processus de sélection est utilisé pour déterminer quels individus de la population pourront se reproduire et créer la progéniture qui formera la génération suivante.

Ce processus est basé sur le score de fitness des individus. Ceux dont le score est le plus élevé ont plus de chance d'être choisis et de transmettre leur matériel génétique à la génération suivante, alors que les individus dont le score de fitness est faible peuvent encore être choisis, mais avec une probabilité faible. De cette façon, leur matériel génétique n'est pas complètement exclu [9].

Il existe plusieurs techniques de sélection. Voici les principales utilisées [9]:

- Un opérateur de sélection simple est la technique de la roulette pondérée où chaque individu d'une population occupe une zone de la roulette proportionnelle à la valeur de sa fonction de fitness. Pour la reproduction, les candidats sont sélectionnés avec une probabilité proportionnelle à leur fitness. Pour chaque sélection d'un individu, une simple rotation de la roulette donne le candidat sélectionné.
- Un autre opérateur de sélection est appelé sélection par tournoi. Cette technique utilise une sélection proportionnelle sur des paires d'individus, puis sélectionne parmi ces paires l'individu ayant le meilleur score de fitness.
- Sélection par rang, cette technique de sélection choisit toujours les individus ayant les meilleurs scores d'adaptation, le hasard n'intervient donc pas dans ce mode de sélection.
- Sélection uniforme, cette technique de sélection se fait de manière aléatoire, uniforme et sans intervention de la valeur de fitness.

- **Mutation**

Le but de l'opérateur de mutation est de rafraîchir périodiquement et aléatoirement la population, d'introduire de nouveaux modèles dans les chromosomes et d'encourager la recherche dans des zones inexplorées de l'espace de solution [9].

Il existe différents types de mutation, on cite :

- Mutation par permutation
- Mutation par déplacement : c'est le déplacement d'une séquence des chromosomes (figure 11) :

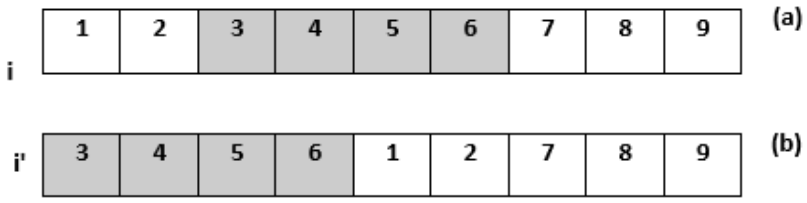

Figure 11: Opération de mutation par déplacement

- **Croisement**

L'opérateur de croisement permet la création de nouveaux individus selon un processus fort simple. Il permet donc l'échange d'information entre les chromosomes (individus). Tout d'abord, deux individus, qui forment alors un couple, sont tirés au sein de la nouvelle population issue de la reproduction. Puis un ou plusieurs sites de croisement s est tiré aléatoirement. Enfin, selon une probabilité Pc que le croisement s'effectue, les segments finaux des deux parents sont alors échangés autour de ce site [9].

On trouve deux différents types de croisement :

- Le croisement à un point, si le génotype est une chaîne binaire de longueur n. Le croisement à un point place un point de croisement au hasard. Un enfant prend une section avant le point de croisement d'un parent et prend l'autre section après le point de croisement de l'autre parent puis recombine les deux sections pour former une nouvelle chaîne binaire. L'autre enfant se construit inversement (figure 12) [9].

Figure 12 : Opération de croisement en un point

- Le croisement à deux points, place deux points de croisement au hasard, et prend une section entre les points d'un parent et les autres sections en dehors des points de l'autre parent puis les recombine (figure 13) [9].

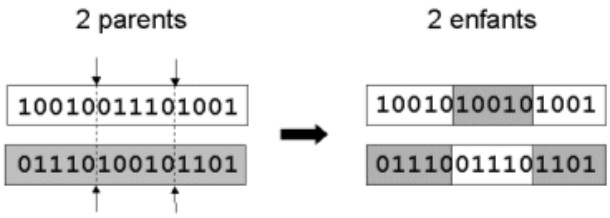

Figure 13 : Opération de croisement en deux points

Il existe plusieurs applications des algorithmes génétiques dans le domaine de la sélection de variables [10].

2.2 Le Recuit simulé

L'algorithme de recuit simulé (Simulated Annealing, SA) est issu de la mécanique statistique. S. Kirkpatrick et ses collègues [11] en 1983 ont proposé un algorithme basé sur l'analogie entre le recuit des solides et la résolution de problèmes d'optimisation combinatoire.

Le recuit est le processus métallurgique qui consiste à chauffer un solide puis à le refroidir lentement jusqu'à ce qu'il cristallise. Les atomes de ce matériau ont une énergie élevée à très haute température. Cela donne aux atomes une grande liberté dans leur capacité à se restructurer. Lorsque la température est réduite, l'énergie de ces atomes diminue, jusqu'à atteindre un état d'énergie minimale. Dans un contexte d'optimisation, SA cherche à émuler ce processus [12].

Le SA commence à une température très élevée où les valeurs d'entrée peuvent varier considérablement. Au fur et à mesure que l'algorithme progresse, la température diminue. Cela restreint le degré de variation des entrées. Ceci conduit souvent l'algorithme à une meilleure solution, tout comme un métal obtient une meilleure structure cristalline grâce au processus de recuit.

Ceci permet à l'algorithme d'accepter non seulement les meilleures solutions mais aussi les moins bonnes avec une probabilité donnée [12].

La principale caractéristique de SA est sa capacité à sortir de l'optimum local en se basant sur la règle d'acceptation d'une solution candidate. Si la solution actuelle (f_{new}) a une valeur de fonction objective plus petite (en supposant une minimisation) que celle de l'ancienne solution (f_{old}), alors la solution actuelle est acceptée. Sinon, la solution actuelle peut également être acceptée si la valeur donnée par la distribution de Boltzmann :

$$e^{-\frac{f_{new}-f_{old}}{T}} \qquad (1)$$

est supérieure à un nombre aléatoire uniforme dans l'intervalle [0,1], où T est le paramètre de contrôle de la température [12].

Le principe de l'algorithme SA est représenté par l'organigramme de la figure 14 :

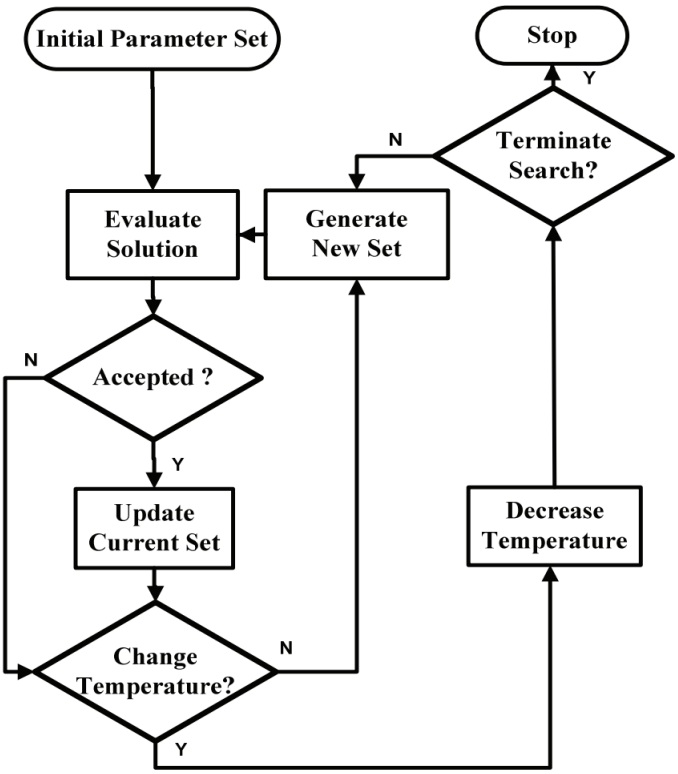

Figure 14 : Principe du Recuit simulé

Le SA se base sur de différents paramètres de contrôles, on définit :

- **Configuration initiale**

Initialement, on part d'une configuration aléatoire ou choisie astucieusement en fonction du problème. Le SA est une méthode qui améliore cette solution initiale. Plus l'estimation initiale est proche de l'optimum global, plus le processus d'optimisation sera rapide [12].

- **Température initiale**

Le paramètre de contrôle 'température' doit être soigneusement défini car il contrôle la règle d'acceptation définie par (2). T_{max} (T_0) doit être suffisamment grand pour permettre à l'algorithme de sortir d'un minimum local, mais suffisamment petit pour ne pas s'éloigner d'un minimum global. La valeur de T_{max} doit être définie dans une approche basée sur l'application, car elle est liée à l'ampleur des valeurs de la fonction valeurs de la fonction objectif [12].

- **Décroissance de la température**

Il existe plusieurs méthodes de décroissance de la température. Le programme de refroidissement le plus courant est la règle géométrique pour la variation de la température (figure 13) :

$$T_{i+1} = sT_i \quad \text{avec} \quad s < 1. \qquad \textbf{(2)}$$

Un autre paramètre est le nombre d'itérations à chaque température, qui est souvent lié à la taille de l'espace de recherche ou avec la taille du voisinage. Ce nombre d'itérations peut même être constant ou, au contraire, être fonction de la température ou basé sur le retour d'information du processus [12].

- **Mécanisme de perturbation**

Le mécanisme de perturbation permet de créer de nouvelles solutions à partir de la solution actuelle. En d'autres termes, il s'agit d'une méthode permettant d'explorer le voisinage de la solution actuelle en créant de petits changements dans la solution actuelle [12].

- **Fonction objective**

Le coût ou la fonction objective est une expression qui, dans certaines applications, relie les paramètres à une propriété (distance, coût, etc.) que l'on souhaite minimiser ou maximiser [12].

- **Arrêt de système**

Ceci signale la fin du traitement ou l'arrêt de l'algorithme. Il correspond au fait qu'aucune autre transformation n'est acceptable. Pour ce faire, Il existe plusieurs méthodes pour contrôler la fin de l'algorithme. Quelques exemples de critères sont :

- Nombre maximal d'itérations.
- Valeur minimale de la température, de la fonction objective ou du taux d'acceptance.

L'un des inconvénients du SA est le réglage des paramètres de l'algorithme pour maximiser les performances. Une étude été faite sur l'effet des paramètres en variant ces derniers pour faire une sélection des variables et les comparants dans le cadre de l'analyse en composantes principales et de l'analyse discriminante [13].

Le recuit simulé est une puissante méthode de recherche stochastique applicable dans différents problèmes qui se posent dans diverses disciplines comme la sélection de variables [14].

2.3 La Colonie de fourmis

Colonie de Fourmis (Ant Colony) est une méta-heuristique dont le comportement est basé sur celui des fourmis réelles.

- **Principe**

Les fourmis sont capables de résoudre collectivement des problèmes complexes, comme trouver le plus court chemin entre deux points dans un environnement accidenté. Pour cela, elles communiquent entre elles de façon locale et indirecte, grâce à une hormone volatile,

appelée phéromone : au cours de leur progression, les fourmis déposent une trace de phéromone ; elles choisissent ensuite leur chemin de façon aléatoire, selon une probabilité dépendant de la quantité de phéromone précédemment déposée. Ce mécanisme, qui permet aux fourmis de résoudre collectivement des problèmes complexes, est à l'origine des algorithmes à base de fourmis artificielles.

Ces algorithmes ont été initialement proposés comme une approche multi-agents pour résoudre des problèmes d'optimisation combinatoire. L'idée est de représenter le problème à résoudre sous la forme de la recherche d'un meilleur chemin dans un graphe, puis d'utiliser des fourmis artificielles pour rechercher de bons chemins dans ce graphe. Le comportement des fourmis artificielles est inspiré des fourmis réelles : elles déposent des traces de phéromone sur les composants du graphe et elles choisissent leurs chemins relativement aux traces de phéromone précédemment déposées ; ces traces sont évaporées au cours du temps. Intuitivement, cette communication indirecte fournit une information sur la qualité des chemins empruntés afin d'attirer les fourmis, dans les itérations futures, vers les zones correspondantes de l'espace de recherche [15].

La figure 15 présente un processus de décision de fourmis choisissant leurs déplacements. Lorsque les fourmis se rencontrent au point de décision A, certaines choisissent un côté et d'autres choisissent l'autre côté au hasard. Supposons que ces fourmis rampent à la même vitesse, celles qui choisissent le côté court arrivent au point de décision B plus rapidement que celles qui choisissent le côté long. Les fourmis qui choisissent par hasard le côté court sont les premières à atteindre le nid. Le côté court reçoit donc la phéromone plus tôt que le côté long et ce fait augmente la probabilité que d'autres fourmis le choisissent plutôt que le côté long. En conséquence, la quantité de phéromone est laissée avec une vitesse plus élevée dans le côté court que dans le côté long parce que plus de fourmis choisissent le côté court que le côté long. Le nombre de lignes brisées dans la figure 15 est un rapport direct avec le nombre de fourmis environ. Le système de colonies de fourmis artificielles est conçu à partir du principe du système de colonies de fourmis pour résoudre des problèmes d'optimisation. La phéromone est la clé de la prise de décision des fourmis [16].

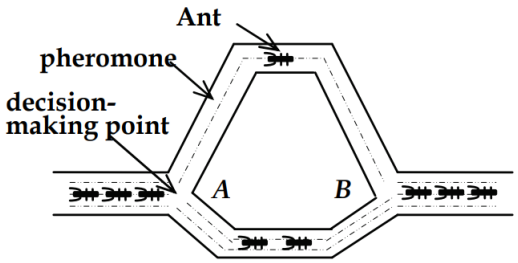

Figure 15 : Processus de la prise de décision chez les fourmis

Ces caractéristiques du comportement des fourmis artificielles définissent la "métaheuristique d'optimisation par une colonie de fourmis" ou "Ant Colony Optimization (ACO)

metaheuristic".

Les algorithmes de colonies de fourmis possèdent plusieurs caractéristiques intéressantes, mentionnons notamment :

- La flexibilité : une colonie de fourmis est capable de s'adapter à des modifications de 1' environnement.
- La robustesse : une colonie est apte à maintenir son activité si quelques individus sont défaillants.
- La décentralisation : une colonie n'obéit pas à une autorité centralisée.
- L'auto-organisation : une colonie trouve elle-même une solution, qui n'est pas connue à l'avance.

Le premier algorithme qui s'inspire de cette analogie a été proposé en 1996 par Colorni, Dorigo et Maniezzo [17]. Le but initial de cet algorithme était de résoudre le problème du voyageur de commerce. Si l'on considère un problème de voyageur de commerce à N villes, chaque fourmi k parcourt le graphe et construit un trajet de longueur n = |N|. Pour chaque fourmi, le trajet d'une ville i à une ville j dépend de :

- La liste des villes déjà visitées, qui définit les mouvements possibles à chaque pas, quand la fourmi k est sur la ville i: J_i^k;

- L'inverse de la distance entre les villes $\eta_{ij} = \frac{1}{d_{ij}}$, appelée visibilité. Cette information est utilisée pour diriger les fourmis vers des villes proches et ainsi, éviter de trop longs déplacements.

- La quantité de phéromone déposée sur l'arête reliant deux villes τ_{ij} , appelée intensité de la piste. Cette quantité définit l'attractivité d'une piste, et elle est modifiée après le passage d'une fourmi. C'est la pseudo-mémoire du système [18].

La règle de déplacement est la suivante :

$$\boldsymbol{p_{ij}^k(t)} = \begin{cases} \dfrac{(\tau_{ij}(t))^{\alpha}(\eta_{ij})^{\beta}}{\sum_{l \in J_i^k}(\tau_{il}(t))^{\alpha}(\eta_{il})^{\beta}} & \textbf{si} \quad \boldsymbol{j \in J_i^k} \\ \mathbf{0} & \textbf{si} \quad \boldsymbol{j \notin J_i^k} \end{cases} \qquad \textbf{(3)}$$

où α et β sont deux paramètres contrôlant l'importance relative de l'intensité et de la visibilité. Après un tour complet, chaque fourmi dépose une quantité de phéromone $\Delta\tau_{ij}^k(t)$ sur l'ensemble de son parcours. Cette quantité dépend de la qualité de la solution trouvée, et elle est définie par :

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \frac{Q}{L^{k}(t)} & \text{si} \quad (i,j) \in T^{k}(t) \\ 0 & \text{si} \quad (i,j) \notin T^{k}(t) \end{cases} \qquad (4)$$

où $T^{k}(t)$ est le trajet effectué par la fourmi k à l'itération t, $L^{k}(t)$ est la longueur de $T^{k}(t)$ et Q est un paramètre fixé. Enfin, il est nécessaire d'introduire un processus d'évaporation des phéromones. En effet, pour éviter de se faire piéger dans des solutions sous optimales, il est nécessaire qu'une fourmi "oublie" les mauvaises solutions. La règle de mise à jour est donc :

$$\tau_{ij}(t+1) = (1-\rho) \cdot \tau_{ij}(t) + \Delta\tau_{ij}(t) \qquad (5)$$

où $\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t)$ et m est le nombre de fourmis. Depuis le développement de la première méthode basée sur l'analogie de colonies de fourmis, cette méthode a été étendue à la résolution de plusieurs problèmes d'optimisation, discrets et continus.

En fait, il existe maintenant un nombre considérable d'applications de ces algorithmes où des performances de classe mondiale sont obtenues. Citons par exemple les applications des algorithmes de colonie de fourmis à des problèmes tels que l'ordonnancement séquentiel, l'ordonnancement [19], l'équilibrage des chaînes de montage, les TSP probabilistes [20], le routage par commutation de paquets dans des réseaux de type Internet [21], etc.

Les méthodes d'optimisation présentées : les algorithmes génétiques, le recuit simulé et la colonie de fourmis présentent des avantages et des inconvénients (Tableau 1).

Tableau 1 : Avantages et inconvénients des GA, SA, Ant Colony

	Avantages	**Inconvénients**
Algorithmes génétiques	• Faculté d'adaptation, réactivité et prise en compte de l'environnement. • Possibilité de traiter des espaces de recherche important. • Relativité de la qualité de la solution selon le degré de précision.	• Nécessitent plus de calculs. • Pas assuré que la solution trouvée est la meilleure. • Problèmes des optimums locaux.

Recuit Simulé	• Possibilité de traiter des systèmes et des fonctions de coût arbitraires • Garantie de la découverte d'une solution optimale statistiquement, et donne généralement une "bonne" solution. • Facile à coder, même pour les problèmes complexes.	• Temps de calcul pour trouver la solution optimale. • Difficulté de déterminer le programme de refroidissement approprié. • Nombre important des paramètres à régler.
Colonie de fourmis	• Possibilité de chercher parmi une population en parallèle. • Adaptabilité aux changements tels que de nouvelles distances. • Garantie de la convergence	• Difficulté de l'analyse théorique. • Changement de la distribution de probabilité par itération. • Temps de convergence est incertain

3. Régression des moindres carrés partiels

La régression des moindres carrés partiels (Partial least Square Regression PLSR) a été inventée en 1983 par Herman Wold. PLSR est née dans le domaine des sciences sociales mais est devenue largement utilisée en chimie (aujourd'hui connue sous le nom de chimiométrie) grâce au fils d'Herman, Svante Wold.

La PLSR est la combinaison de l'algorithme NIPALS (Non linear Iterative Partial Least Squares) développé par Herman Wold pour l'analyse en composantes principales et de l'approche PLS proposée par Herman Wold [22].

La PLSR s'avère concrètement une méthode efficace qui justifie son emploi très répandu mais présente le défaut de ne pas se prêter à une analyse statistique traditionnelle qui exhiberait les lois de ses estimateurs. Elle est ainsi restée une marge des approches traditionnelles de la Statistique mathématique [24].

Différentes versions de PLSR sont proposées en fonction de l'objectif poursuivi :

- **PLS1 :** Une variable cible Y quantitative est à expliquer, modéliser, prévoir par p variables explicatives quantitatives X^j.

- **PLS2** Version canonique : Mettre en relation un ensemble de q variables quantitatives Y^k et un ensemble de p variables quantitatives X^j.
- **PLS2** Version régression : Chercher à expliquer, modéliser un ensemble de q variables Y^k par un ensemble de p variables explicatives quantitatives X^j.
- **PLS-DA** Version discriminante : Cas particulier du cas précédent. La variable Y qualitative à q classes est remplacée par q variables indicatrices (dummy variables) de ces classes.

- **Modèle mathématique**

L'objectif du projet est de mesurer la quantité des éléments Carbone et Azote dans les tomates cerises. Puisqu'une seule variable cible à prédire, la version PLS utilisée est la version PLS1.

Régression PLS1

Une variable cible **Y** quantitative est à expliquer, modéliser, prévoir par p variables explicatives quantitatives $\mathbf{X}^j$. Comme pour la régression sur composantes principales, le principe est de rechercher un modèle de régression linéaire sur un ensemble de composantes orthogonales construites à partir de combinaisons linéaires des p variables explicatives centrées $\mathbf{X}^j$. Dans le cas de la PLS, la construction des composantes est optimisée pour que celles-ci soient les plus liées à la variable Y à prédire au sens de la covariance empirique, alors que les composantes principales ne visent qu'à extraire une part de variance maximale sans tenir compte d'une variable cible [24].

Soit $\mathbf{X}_{(n\times p)}$ la matrice des variables explicatives centrées avec n pouvant être inférieur à p. On cherche une matrice **U** de coefficients ou pondérations (loading vectors) définissant les r composantes Ξ_h (ou variables latentes) par combinaisons linéaires des variables X^j :

$$\boldsymbol{\Xi} = \mathbf{XU} \qquad \textbf{(6)}$$

La matrice U est solution du problème suivant :

$$\begin{aligned} \text{Pour } h = 1, \dots, r, \qquad \mathbf{u}_h &= \arg \max_{\mathbf{u}} \operatorname{Cov}(Y, \Xi_h)^2 \\ &= \arg \max_{\mathbf{u}} \mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{u} \\ \text{Avec } \mathbf{u}_h'\mathbf{u}_h &= 1 \\ \text{et } \boldsymbol{\xi}_h'\boldsymbol{\xi}_h = \mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{u} &= 0, \qquad \text{pour } \ell = 1 \dots, h-1. \end{aligned}$$

La matrice **U** est obtenue par la démarche itérative de l'algorithme 1 ;

Algorithm 1 Régression PLS1

$\mathbf{X}$ matrice des variables explicatives centrées,
Calcul de la matrice $\mathbf{U}$ des coefficients.
for $h = 1$ à r **do**
  $\mathbf{u}_h = \frac{\mathbf{X}'\mathbf{Y}}{\|\mathbf{X}'\mathbf{Y}\|}$,
  $\boldsymbol{\xi}_h = \mathbf{X}\mathbf{u}_h$
  Déflation de $\mathbf{X}$: $\mathbf{X} = \mathbf{X} - \xi_h \xi_h' \mathbf{X}$
end for

Figure 16 : Algorithme du PLS1

Ensuite il suffit de calculer la régression de **Y** et la matrice **U** qui contienne les r variables ξ_h centrées, appelées variables latentes. Le choix du nombre de composantes r est optimisé par validation croisée [24].

La formule de régression correspondante peut être facilement écrite comme une fonction des variables initiales **X :**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}_a \qquad (7)$$

Avec :

$$\widehat{\boldsymbol{\beta}}_a = \mathbf{A}_a(\mathbf{A}_a^t \mathbf{S}\mathbf{A}_a)^{-1}\mathbf{A}_a^t \mathbf{s}$$
$$\mathbf{A}_a = [\mathbf{s}|\mathbf{S}\mathbf{s}| \ldots | \mathbf{S}^{a-1}\mathbf{s}],$$
$$\mathbf{S} = \mathbf{X}^t\mathbf{X}$$
$$\mathbf{s} = \mathbf{X}^t\mathbf{Y}$$

La PLSR est une méthode ancienne largement utilisée, notamment en chimiométrie dans l'agro-alimentaire [23].

Conclusion

L'utilisation des méthodes : les algorithmes génétiques, recuit simulé et colonie de fourmis, est une contrainte technique par MAScIR, pourtant le choix de l'utilisation du PLSR a été après une phase d'analyse de données. Les étapes d'analyse et de prétraitements de données établies afin de choisir le PLSR sont présentées dans le chapitre suivant.

Chapitre III : Analyses et prétraitement des données spectrales

Introduction

Dans le chapitre précédant, nous avons présenté les méthodes d'optimisation : les algorithmes génétiques, le recuit simulé et colonie de fourmi et l'algorithme de prédiction que nous allons utiliser. Avant de passer vers la phase de l'entrainement, les données spectrales doivent passer par des différentes phases :

- Une phase des prétraitements spectraux pour la correction des déformations des spectres enregistrés.
- Une analyse descriptive et statistique pour choisir le modèle de prédiction adéquat.
- Enfin des prétraitements pour la détection des valeurs aberrantes.

1. Spectroscopie

1.1 Définition

La spectroscopie infrarouge repose sur le principe que chaque groupement chimique absorbe la lumière différemment en fonction de sa longueur d'onde. Les bandes d'absorption (zone où la lumière est absorbée) permettent donc d'identifier les groupements atomiques. Beer (1729) et Lambert (1760) ont ainsi proposé d'observer l'atténuation d'un faisceau de la lumière afin de prédire la concentration d'un composant selon l'expression suivante :

$$\boldsymbol{A_\lambda} = \mathrm{s}_\lambda * \boldsymbol{l} * \boldsymbol{c} \qquad \textbf{(8)}$$

Avec :

- $\boldsymbol{A_\lambda}$ **:** C'est l'Absorbance à une longueur d'onde λ donnée.
- s_λ **:** C'est le coefficient d'extinction molaire.
- $\boldsymbol{l}$ **:** la longueur du trajet optique dans l'échantillon.
- $\boldsymbol{c}$ **:** la concentration de la solution.

La première application analytique de NIR était en 1962 par Hart et Norris [19]. Ils ont fondé un modèle linéaire sur la loi de longueur d'onde unique [25].

Il existe de différents types d'analyse dans la spectroscopie infrarouge :

- Analyse par transmission : le passage d'un rayonnement électromagnétique à travers un milieu.
- Analyse par réflexion : le passage par lequel un rayonnement électromagnétique est renvoyé soit à la limite entre les deux milieux (surface de réflexion) ou à l'intérieur d'un milieu (volume de réflexion)

L'analyse du spectre proche infrarouge propose plusieurs avantages tels que :

- Analyse rapide multi composant, en temps réel.
- Coût d'analyse moins élevé.
- L'échantillon nécessite peu ou pas de préparation
- Possibilité d'analyse de produits toxiques ou dangereux à distance (+ de 500m en utilisant des fibres optiques)

1.2 Prétraitements spectraux

Les conditions idéales de la spectrométrie infrarouge correspondent à une mesure d'une solution transparente et peu concentrée. Ces conditions se rencontrent au laboratoire, avec des échantillons préparés. Dans ces conditions, la concentration d'un composant d'intérêt est directement reliée à son absorbance [25].

En réalité la loi de Beer-Lambert jamais utilisée dans les conditions idéales car les spectres enregistrés sont influencés par plusieurs phénomènes à savoir les concentrations des autres composants présents dans l'échantillon, la température, la granulométrie de l´échantillon, etc. Par conséquence ceci cause des déformations des spectres. Dans la figure 17 nous présentons les spectres d'Azote avant leur correction.

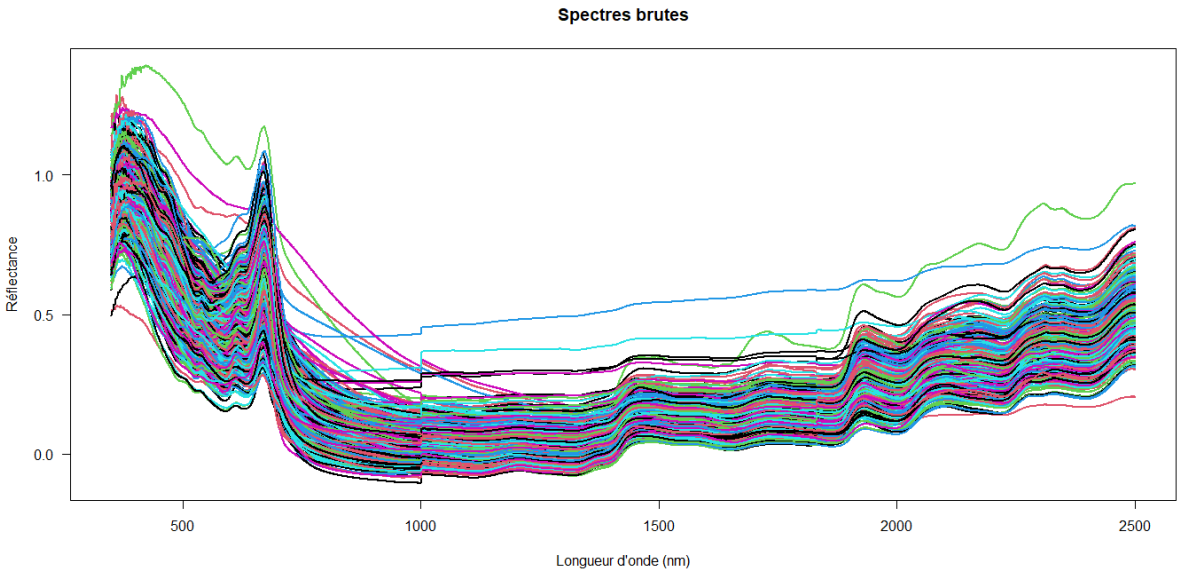

Figure 17 : Spectres brutes d'Azote

Toutes ces déformations de spectres entraînent des déformations des vecteurs (les spectres mesurés sont digitalisés en vecteurs de dimension P) et donc peuvent :

- Entraîner une déformation de l'espace décrit par les spectres servant à l'apprentissage, et finalement mener à une mauvaise estimation du modèle.

- Entraîner des erreurs dans les résultats de prédictions des modèles.

Pour éviter ceci, les prétraitements tentent d´éliminer les déformations subies par les spectres (figure 18) et ainsi se rapprocher de la seule contribution du paramètre recherché.

Les prétraitements spectraux sont réalisés par les autres membres participants au projets qui font leur étude dans la chimie.

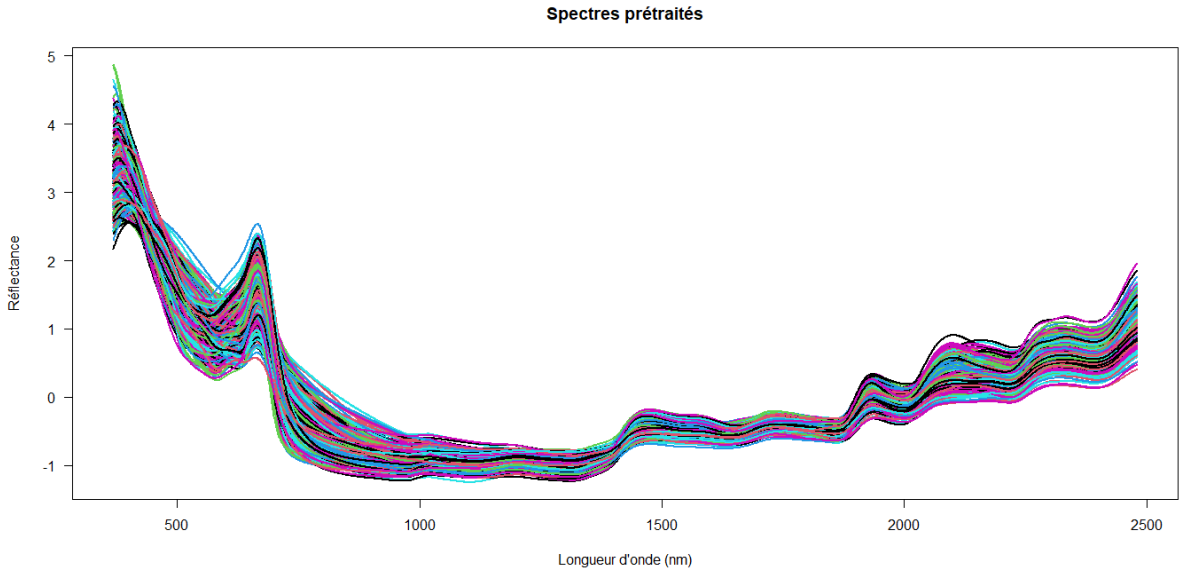

Figure 18 : Spectres d'Azote prétraités

2. Analyses et prétraitement

L'acquisition et la construction de données des spectres des deux éléments chimiques (Azote et Carbone) étaient faites par MAScIR. De même le calcul des données de sortie été réalisé au laboratoire puisque nous allions effectuer un apprentissage supervisé.

Les données des spectres que nous allons utiliser contiennent les spectres qui sont déjà passés par les prétraitements spectraux.

Pour déterminer le modèle de prédiction adéquat à utiliser sur les données nous avons été amenés à appliquer les tests de normalité : graphiques et statistiques.

2.1 Test graphique de normalité des données spectrales

2.1.1 Les coefficients d'asymétrie et d'aplatissement

La loi normale est caractérisée par un coefficient d'asymétrie et un coefficient d'aplatissement nuls [26]. Calculer ces indicateurs donnera une idée sur la distribution des données.

Nous commençons par présenter les histogrammes de fréquences [26] des mesures de données

d'Azote et du Carbone et leurs distributions (figure 21).

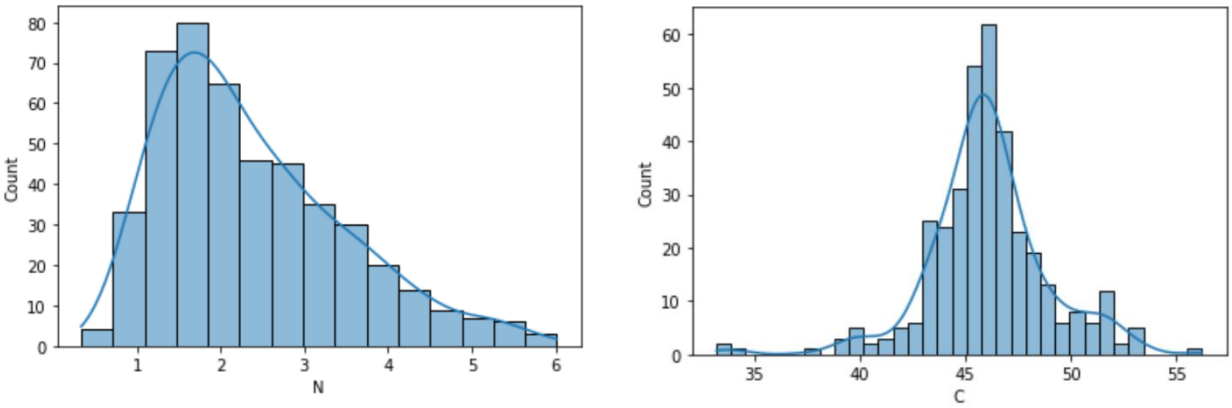

Figure 19 : Histogramme d'Azote (N) et du Carbone (C)

D'après les histogrammes, on constate une asymétrie droite pour l'Azote, et un aplatissement positif pour le Carbone, ce qui confirme les résultats des coefficients d'asymétrie trouvés (Tableau 2). Donc les distributions de ces variables n'ont pas la forme d'une cloche où on peut conclure alors que les distributions des données ne sont pas normales.

Tableau 2 : Coefficients d'asymétrie et d'aplatissement

	Coefficient d'asymétrie	***Coefficient d'aplatissement***
	Sans unité	
Azote	0.866	0.267
Carbone	-0.331	3.042

2.1.2 Test par droite d'Henry

Parmi les techniques graphiques, pour tester la normalité des données, nous avons utilisé le test par droite d'Henry [26].

Ce test permet de comparer les distributions de deux ensembles de données. Si les données sont compatibles avec la loi normale, alors les points (x(i), x*(i)) forment une droite, dite droite de Henry, alignés sur la diagonale principale.

Puisque la droite de Henry n'est pas alignée sur la diagonale principale (figure 22), les données des deux éléments (Azote et Carbone) ne suivent pas alors une distribution normale.

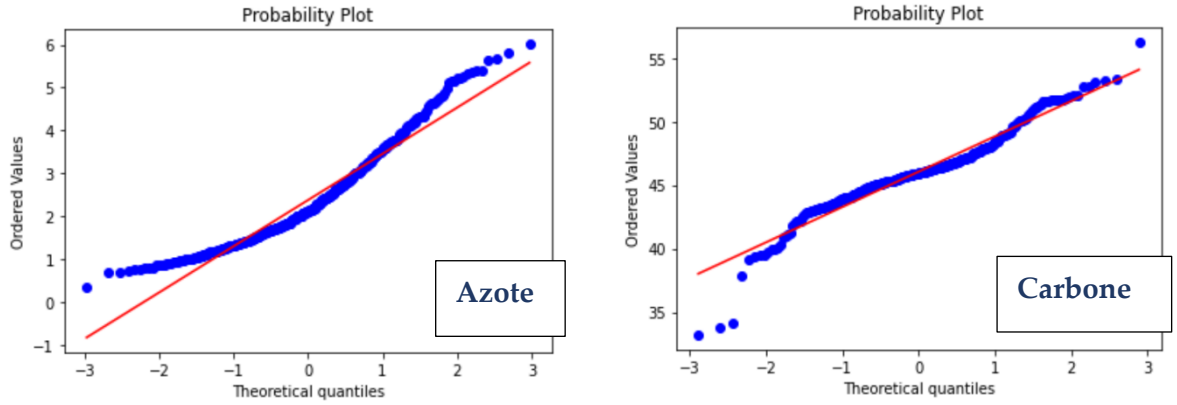

Figure 20 : Test de normalité par la droite de Henry

2.2 Test statistique de normalité des données spectrales

Le test de Shapiro-Wilk [26] est un test très populaire basé sur la statistique W :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i\left(x_{(n-i+1)}-x_i\right)\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (9)$$

Où :

- x_i : Correspond à la série des données triées.
- $\left[\frac{n}{2}\right]$: Partie entière du rapport $\frac{n}{2}$
- a_i : sont des constantes générées à partir de la moyenne et de la matrice de variance covariance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques.

La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générés à partir de la loi normale et les quantiles empiriques obtenus à partir des données. Plus W est élevée, plus la compatibilité avec la loi normale est crédible. La région critique, rejet de la normalité, s'écrit :

$$W < W\text{crit.}$$

Les valeurs seuils **Wcrit** pour différents risques α et effectifs **n** (nombres des échantillons) sont lues dans la table de Shapiro-Wilk.

D'après les résultats des tests (Tableau 3), les valeurs du ***p-value*** sont inférieures à 0.05, ce qui signifie le rejet de l'hypothèse H_0 et l'acceptation de H_1 de la non normalité des données.

L'information qu'on peut conclure à partir de ces tests, c'est qu'on ne peut pas utiliser des

modèles de prédiction qui exigent des données d'une distribution normale. Et puisque les données révèlent un grand taux de corrélation, la PLSR est un choix parfait comme modèle de prédiction.

Tableau 3 : Test de normalité par Shapiro-Wilk

	Statistique ***W***	***p-value***
	Sans unité	
Azote	0.940	$2.02*e^{-10}$
Carbone	0.944	$7.15*e^{-13}$

2.3 Détection des valeurs aberrantes

Lorsqu'une observation diffère de la majorité des données, ou est suffisamment improbable selon le modèle de probabilité supposé des données, elle est considérée comme une valeur aberrante. Les valeurs aberrantes, présentes dans les données, peuvent être le résultat : d'un défaut de l'instrument d'acquisition des spectres, une perturbation du processus de calcul des valeurs de sortie dans le laboratoire.

La présence de valeurs aberrantes dans un ensemble de données peut considérablement compromettre l'analyse et tout résultat ultérieur basé sur ces données. Lors du traitement des données spectroscopiques, qui consistent en des mesures à plusieurs longueurs d'onde, il faut envisager la possibilité de l'apparition de valeurs aberrantes.

Il existe plusieurs méthodes pour détecter les valeurs aberrantes. Certaines méthodes graphiques comme les boîtes à moustaches, des méthodes statistiques comme le z-score ou d'autres à base d'apprentissage non supervisé tel que ACP.

Pour déterminer les valeurs aberrantes dans les données, nous avons utilisé l'analyse par les composantes principales (ACP) [27]. Cette méthode statistique descriptive consiste à transformer des variables liées entre elles (corrélées) en de nouvelles variables non corrélées. Elle est utile pour la réduction de la dimensionalité et la détection des valeurs aberrantes.

Avec ACP nous avons utilisé trois composantes principales qui ont pu expliquer jusqu'à **88%** de la variance des données. En traçant ces composantes principales dans un graphe 3-D et en ajoutant une sphère seront notre moyen pour détecter les valeurs aberrantes (figure 23).

Cette sphère représente une région de confiance statistique basée sur le test **T^2 d'Hotelling** [27] avec des limites de confiance de **95%.**

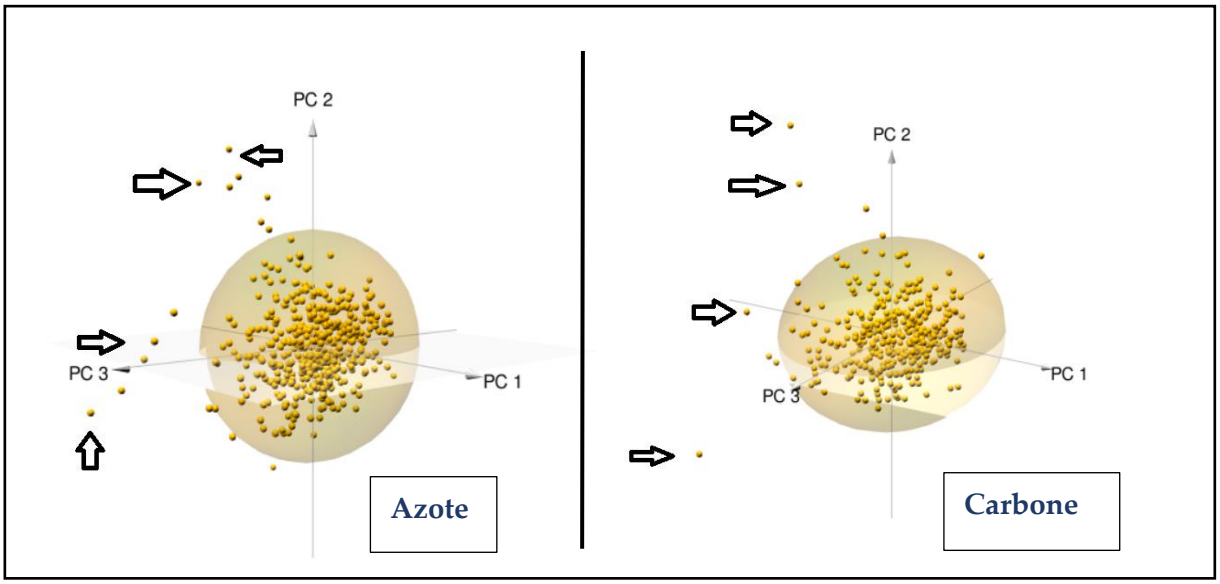

Figure 21 : Représentation des données par ACP et Hotelling T^2

Après la détection des valeurs aberrantes, il faut absolument réagir avant de continuer vers la phase de modélisation. Puisque nous ne pouvons pas demander de refaire les calculs et corriger ces valeurs, la solution la plus simple dans notre cas est la suppression. Cette suppression de ces valeurs est logique d'après l'utilisation du test statistique **T^2 d'Hotelling** avec un des limites de confiances de **95%.**

Lors de la suppression des valeurs aberrantes de données, nous avons éliminé 15 échantillons des spectres d'Azote et 20 échantillons des spectres du Carbone. Les valeurs supprimées ont des valeurs des composantes principales très grandes, ce qui signifie la grande distance avec le centre de la sphère (figure 24).

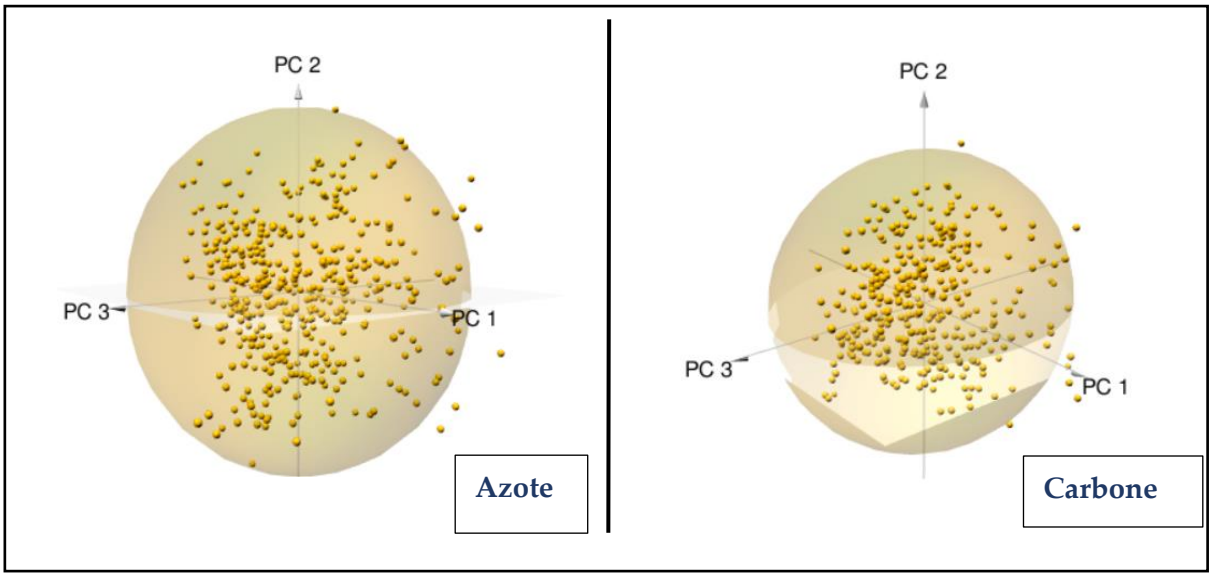

Figure 22 : Représentation des données par ACP et Hotelling T^2 après la suppression des valeurs aberrantes

Conclusion

Une analyse descriptive et statistique nous a permis de choisir le modèle adéquat pour la prédiction. Les prétraitements spectraux établies par les autres membres participants dans le projet sont pour la correction des spectres or la détection des valeurs aberrantes dans les données est pour la vérification de l'homogénéité de données.

L'implémentation et les résultats du modèle hybride combinant les méthodes les méthodes métaheuristiques et le modèle de prédiction PLSR sont présentés dans le chapitre suivant.

Chapitre IV : Implémentation et résultats

Introduction

Dans ce dernier chapitre, nous allons implémenter les différents algorithmes que nous avons introduit dans le chapitre II en utilisant les données décrites dans le chapitre III.

L'objectif de notre projet est de développer un modèle hybride combine les méthodes d'optimisation : les algorithmes génétiques, le recuit simulé et les colonies de fourmis pour effectuer une sélection des variables pertinentes et non redondantes dans les données spectrales et l'algorithme de prédiction PLSR pour l'estimation des matières chimiques (Azote et Carbone) dans les tomates cerises. Nous commencerons par implémenter les algorithmes génétiques ensuite le recuit simulé et enfin les colonies de fourmis.

1. Comparaison des modèles

Pour comparer les performances du modèle hybride en fonction de la méthode d'optimisation utilisé dans le modèle, nous avons utilisé des différents indicateurs.

1.1 Coefficient de détermination

Le coefficient de détermination (noté R^2 ou r^2 et prononcé R Square) est un indicateur qui permet de juger la qualité d'une régression. Il mesure l'adéquation entre le modèle et les données observées ou encore à quel point l'équation de régression est adaptée pour décrire la distribution des points. Il varie généralement de 0 à 1, il est défini par :

$$\boldsymbol{R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}} \qquad \textbf{(10)}$$

Où :

- **n** : nombre de mesures
- $\boldsymbol{y_i}$: la valeur de la mesure de n^i
- $\boldsymbol{\hat{y}_i}$: la valeur prédite correspondante
- $\boldsymbol{\bar{y}_i}$: la moyenne des mesures

1.2 Coefficient de détermination ajusté

Le coefficient de détermination ajusté (noté $\bar{R}^2$ et prononcé rajusté R Square) est une mesure statistique qui montre la proportion de variation expliqué par la ligne de régression estimée. Il prend toujours une valeur comprise entre 0 et 1, Plus il est à 1, mieux l'équation de régression estimée correspond ou explique la relation entre X et Y. Il est étroitement lié au R^2 et défini par :

$$\boldsymbol{\bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right]} \qquad \textbf{(11)}$$

Avec :

- $\boldsymbol{R^2}$: coefficient de détermination
- **n** : nombre des mesures
- **k** : nombre des variables indépendantes

1.3 Erreur quadratique moyenne

L'erreur quadratique moyenne (Mean Squared Error, MSE) est une mesure de la qualité d'un estimateur. Il mesure la moyenne des carrés des erreurs, c'est-à-dire la différence quadratique moyenne les valeurs estimées et la valeur réelle. Il s'agit toujours d'une valeur positive avec une erreur décroissante à mesure que l'erreur se rapproche de zéro. Il est défini par :

$$\mathbf{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i}\right)^{\mathbf{2}} \qquad \textbf{(12)}$$

Avec :

- **n** : nombre des mesures
- $\boldsymbol{y_i}$: la valeur de mesure de n^0 i
- $\boldsymbol{\hat{y}_i}$: la valeur prédite correspondante

1.4 Racine de l'erreur quadratique moyenne

C'est la racine carrée du MSE (en anglais Root Mean Squared Erreur, noté RMSE), Il s'agit de la racine des différences entre les valeurs prédites et les valeurs observées. RMSE est toujours non négatif, une valeur 0 indiquerait un ajustement parfait aux données. Il est défini par :

$$\mathbf{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\mathbf{2}}} = \sqrt{\boldsymbol{MSE}} \qquad \textbf{(13)}$$

Avec :

- **n** : nombre des mesures
- $\boldsymbol{y_i}$: la valeur de mesure de n^0 i
- $\boldsymbol{\hat{y}_i}$: la valeur prédite correspondante

1.5 Critère d'information d'Akaike

C'est le critère d'information d'Akaike (en anglais Akaike Information Criterion, noté AIC) est un estimateur de l'erreur de prédiction et donc de la qualité relative des modèles statistiques pour un ensemble de données connus aprioris. Etant donné une collection de modèles pour les données, AIC estime la qualité de chaque modèle par rapport à chacun des autres modèles.

Il utilise le maximum de vraisemblance, mais en pénalisant les modèles comportant trop de variables, il est défini par :

$$\mathbf{AIC} = \boldsymbol{n}\mathbf{ln}\ \mathbf{MSE} + \mathbf{2}\boldsymbol{k} \qquad \mathbf{(14)}$$

Avec :

- **n** : nombre des mesures
- **k** : nombre des variables indépendantes
- **MSE** : Mean Squared Erreur

La comparaison des différentes implémentations et le critère d'évaluation des performances sont basés sur les résultats de ces indicateurs.

La construction de notre modèle hybride est basée sur l'approche enveloppe des méthodes de sélection de variables où le modèle de prédiction va permettre de tester les différents sous-ensembles de variables générés itérativement par les méthodes d'optimisation. Le critère de comparaison entre les sous-ensembles générés par les méthodes d'optimisation est basé sur l'indicateur RMSE où nous essayons de minimiser cette fonction objectif RMSE.

2. Implémentation des modèles

Le modèle hybride est une combinaison d'une méthode d'optimisation et le modèle de prédiction PLSR. Les méthodes d'optimisations ont comme objectif de faire une sélection de variables où chaque variable représente une longueur d'onde. Ces modèles vont présenter le sous-ensemble optimal des longueurs d'ondes qui minimise la valeur du RMSE résultante du PLSR.

L'implémentation des algorithmes d'optimisation et le modèle de prédiction est écrite en Python et à l'aide de la bibliothèque **Numpy** et la bibliothèque open source **Scikit-Learn**.

PLSR est une méthode qui projette l'ensemble de variables dans un nouvel espace en utilisant de novelles variables appelées variables latentes. Durant l'apprentissage et l'évaluation des sous-ensembles générés par les méthodes d'optimisation, nous avons réalisé un réglage du paramètre : nombre de variables latentes demandées par PLSR, où il varie de 1 à 21 variables latentes. Le PLSR retourne la valeur la plus adapté du RMSE en fonction de ce paramètre pour chaque candidat qui sera un critère de comparaison avec les autres candidats.

Nous avons ajouté lors de la description de ces modèles des fichiers logs pour historiser le processus et les résultats de chacun deux.

Pour la phase d'apprentissage, les données sont divisées en deux parties. Une partie contient 80% pour l'apprentissage et l'autre contient 20% pour le test.

L'évaluation des performances des modèles est réalisée en calculant les scores des indicateurs en utilisant des données d'apprentissage, de la validation croisée et du test.

Les 80% de données sont utilisées pour calculer les scores des indicateurs pour les données d'apprentissage et pour la validation croisée. Les scores d'apprentissage donnent une vue sur la capacité du modèle à apprendre les données d'apprentissage. Or les scores avec les données de la validation sont utilisés pour indiquer dans quelle mesure le modèle appris se généralise, et pour fournir une évaluation non biaisée de l'ajustement du modèle sur l'ensemble de données d'apprentissage. Nous avons utilisé ***K-fold*** comme méthode de validation tel que ***k=4***.

Après que le modèle est formé, les 20% de données sont utilisés pour calculer les scores en utilisant ces données de test afin de voir le comportement du modèle avec des données totalement inconnues.

La comparaison entre ces scores en utilisant les différents ensembles de données sont aussi utilisées vérifier le sur-apprentissage ou le sous-apprentissage.

Les résultats des implémentations du modèle hybride présentés dans ce rapport ne sont que des prototypes, nous avons réalisé une simulation en utilisant des petites valeurs des paramètres des algorithmes d'optimisation pour que les traitements ne prennent pas beaucoup de temps.

Les résultats finaux des implémentations en effectuant un réglage de paramètres et en utilisant des valeurs adéquates de ces paramètres ne sont pas disponibles et ceci revient au manque du matériel un niveau du MAScIR. Nous attendons alors ces résultats pour les interpréter et les analyser afin de déduire des conclusions sur les performances de ces modèles hybrides.

2.1 Application des algorithmes génétiques

Le premier algorithme que nous avons implémenté pour la sélection de variables est les algorithmes génétiques. Cette méthode est simple à implémenter, très efficace, et n'exige pas un grand nombre de paramètres à régler.

Les algorithmes génétiques sont combinés avec l'algorithme régression des moindres carrées partiels afin d'avoir un modèle hybride.

2.1.1 Modélisation par GA

Nous présentons la modélisation de l'algorithme :

- **Codage**

Le choix du codage était un codage réel, chaque gêne représente l'indice d'une longueur d'onde. Les valeurs constituant l'individu désignent les variables de données (les longueurs d'onde) et varient de 350 et 2500 (figure 25).

370	785	1654	486	2400	….	….	654	296	1275

Figure 23 : Exemple d'un individu

- **Fonction de Fitness**

La fonction de fitness choisie est l'indicateur RMSE résultant de l'algorithme PLSR.

L'évaluation des individus se fait par l'appel de la fonction de fitness RMSE laquelle l'algorithme essaye de minimiser : Chaque individu généré est passé à l'algorithme d'apprentissage PLSR qui retourne la valeur du RMSE correspondante à cet individu.

La séquence des longueurs d'ondes de l'individu ayant la valeur minimale du RMSE représente la solution optimale de notre expérience.

- **Génération de la population initiale**

Dans cette phase, **n** individus de même taille **s** constituent la population initiale. Ils sont générés d'une façon aléatoire et sans redondance des individus et des gênes dans chaque individu.

- **Sélection**

La phase de la sélection commence après l'initiation de la population. Dans cette étape l'algorithme sélectionne **n/2** individus. La méthode de la sélection que nous avons utilisée est la sélection uniforme où tous les individus ont la même probabilité d'occurrence

- **Croisement**

Les **n/2** individus sélectionnés (les parents) par la phase de la sélection passeront à la phase du croisement.

Dans cette phase nous prenons un couple de deux individus aléatoirement, et avec une probabilité Pc on applique le croisement en deux points. Les deux points sont générés aléatoirement.

- **Mutation**

La phase de mutation utilisera les nouveaux individus (les fils) générés par le croisement.

Pour chaque individu fils, et avec une probabilité Pm, la mutation par déplacement est réalisée en générant aléatoirement deux points différents, la suite des gênes entre ces deux points sera déplacée au début de l'individu.

- **Reproduction**

Dans la phase de reproduction, les individus parents générés par la phase de la sélection sont combinés avec les individus générés par la phase de mutation pour revenir à la taille **n** de la

population initiale.

2.1.2 Les étapes du GA

Voici les étapes de l'algorithme :

1. Initialisation :
    - Spécifier le nombre des itérations.
    - Générer une population initiale de ***n*** individus, et chaque individu contient ***s*** variables générées aléatoirement et sans redondance.
2. Evaluer les individus de la population
    - Pour j = 1 à n,
        - Estimer la valeur du RMSE obtenue par le PLSR de chaque individu Ij.
        - Vérifier les performances avec le RMSE minimum, et stocker le sous-ensemble correspondant de variables.
3. Sélection ***n*/2** individus par une sélection uniforme (les parents).
4. Faire le croisement jusqu'à avoir ***n*/2** nouveaux individus,
    - Choisir 2 individus aléatoirement de la population des parents.
    - Si ***r*** : une variable aléatoire uniforme > ***(1 – Pc)***,
        - Faire le croisement à deux points.
        - Ajouter les nouveaux deux individus à la population des fils.
5. Faire la mutation par déplacement pour chaque individu de la population des fils selon une probabilité ***Pm***.
6. Rassembler la population des parents et la population des fils pour avoir une nouvelle population d'une taille de ***n*** individus.
7. Si le nombre d'itérations est inférieur au nombre maximal d'itérations, passer à l'étape 2.

2.1.3 Résultats

Les paramètres utilisés pour l'exécution du modèle hybride GA et PLSR :

- Taille de la population initiale **(n) :** 100
- Taille de l'individu **(s)** : 200
- Nombre d'itérations : 500

- Probabilité du croisement **(Pc)** : 0.7
- Probabilité de la mutation **(Pm)** : 0.3

Nous présentons dans le tableau 4 les résultats d'apprentissage en utilisant le sous-ensemble de variables sélectionné par GA pour les données d'Azote et du Carbone.

Tableau 4 : Résultats du GA

		Carbone
Temps d'exécution		**11h, 48min**
Cross-Validation(cv=4)	R^2	**0.86**
	R^2 ajusté	**0.84**
	MSE	**0.97**
	RMSE	**0.98**
Train	R^2	**0.91**
	R^2 ajusté	**0.90**
	MSE	**0.64**
	RMSE	**0.80**
	AIC	**-64.83**
Test	R^2	**0.73**
	R^2 ajusté	**0.67**
	MSE	**1.18**
	RMSE	**1.08**
	AIC	**57.13**

D'après les résultats obtenus, et avec individu de 200 variables, le score du R^2 a atteint 73% et R^2ajusté 67%, ces variables ne sont pas satisfaisantes pas ceci revient aux paramètres et leur réglage. Les tests finaux contiennent un réglage des paramètres (Taille de l'individu et Taille de la population) et avec un nombre des itérations intéressant.

2.2 Application du recuit simulé

Le deuxième modèle que nous allons implémenter est le recuit simulé (SA). Cette méthode d'optimisation est simple à implémenter, et d'après l'état de l'art, elle a plus d'avantage de ne pas être bloquer dans les minimums locaux, et elle peut converger vers le minimum global rapidement mais le problème se pose au niveau configuration initiale et réglage des

paramètres.

Avec le SA, le processus démarre d'une configuration aléatoire ou déterministe, donc plus l'estimation initiale est proche de l'optimum global, plus le processus d'optimisation sera rapide et aussi un bon choix des paramètres donne des bons résultats. Le paramètre de la température initiale et le programme de refroidissement pour la décroissance de la température doivent être choisis soigneusement pour permettre l'algorithme de sortir d'un minimum local.

2.2.1 Modélisation par SA

La modélisation de l'algorithme recuit simulé est présentée comme suivant :

- **Codage**

Pour le SA, nous avons utilisé deux types de codage :

- Un codage binaire où les solutions sont représentées par un tableau de taille égale au nombre initial des variables. Chaque colonne désigne une variable et la valeur 1 signifie que cette variable appartient à l'ensemble optimal des longueurs d'onde. Avec ce type de codage, le nombre des longueurs d'onde (variables) de chaque solution est différent.
- Un codage réel où les solutions sont de taille fixe de variables. Ce codage est équivalent au codage utilisé dans les algorithmes génétiques.

- **Configuration initiale**

Initialement, SA part d'une configuration initiale et essaye d'améliorer cette solution. Pour chaque codage, nous avons initialisé SA d'une manière différente :

- Codage binaire : la première solution est générée aléatoirement où chaque colonne du tableau prend aléatoirement une valeur 0 ou 1.
- Codage réel : les longueurs d'onde sont générées d'une façon aléatoire et sans redondance.

- **Température initiale**

La température initiale ou alors la température maximale **(T_0)** doit être soigneusement défini, nous avons utilisé des différentes valeurs pour ce paramètre.

- **Décroissance de la température**

Pour ce paramètre nous avons utilisé le programme de refroidissement pour la décroissance de la température suivant : la méthode Géométrique : $\mathbf{T_{i+1} = sT_i}$ **.**

- **Nombre d'itérations**

Nous avons utilisé ce paramètre deux fois dans notre implémentation :

Nombre d'itérations à chaque température (N_iter_max) : est équivalent au nombre des solutions à évaluer à chaque température.

Nombre d'itérations maximale (N_restart_max) : pour changer la solution courante si l'algorithme n'a pas pu l'améliorer.

- **Mécanisme de perturbation**

Ce paramètre définit la méthode permettant d'explorer le voisinage des solutions actuelles.

Pour avoir une solution voisine, on change les valeurs d'un nombre fixe de colonnes, les colonnes à remplacer sont générer aléatoirement. Les valeurs utilisées pour ce paramètre sont en fonction du codage utilisé.

- **Arrêt de système**

Nous avons choisi comme paramètre pour l'arrêt du système c'est atteindre une température minimale $\mathbf{T_{min}}$.

2.2.2 Les étapes du SA

Voici les étapes de l'algorithme :

1. Initialisation :
    - Spécifier les paramètres du SA
    - Définir la Température par la Température maximale
    - Créer une solution initiale générée aléatoirement
2. While itération < ***N_iter_max***:
    - Estimer la valeur du RMSE obtenue par le PLSR de la solution actuelle.
    - Perturber la solution actuelle.
    - Générer une solution voisine candidate.
    - Estimer la valeur du RMSE obtenue par le PLSR de la solution voisine.
    - Si RMSE de la solution voisine est meilleur que la solution actuelle (amélioration de la solution actuelle)
        - Mettre à jour la solution actuelle.
        - Vérifier les performances avec le RMSE minimum, et stocker le sous-ensemble correspondant de variables.
    - Sinom :

- Calculer la probabilité de la règle d'acceptation avec :

$$e^{-\frac{RMSE_{\text{new}} - RMSE_{\text{old}}}{T}} \quad \textbf{(15)}$$

Où

***RMSE**new*: RMSE de la solution voisine

***RMSE**old* : RMSE de la solution actuelle

T : température actuelle

- Si ***r*** : une variable aléatoire uniforme > probabilité,
  - Rejeter la solution voisine.
- Sinon,
  - Accepter la solution voisine et mettre à jour la solution actuelle.

- Si le nombre maximal d'itérations (***N_restart_max***) est atteint et les performances ne sont pas améliorées.
  - Mettre à jour la solution actuelle par la solution ayant le minimum RMSE.

3. Diminuer la Température avec l'équation :

$$\mathbf{T_{i+1} = sT_i} \quad \text{avec} \quad s < 1. \quad \textbf{(16)}$$

4. Si la Température est supérieure à la température minimale, passer à l'étape 2.
5. Déterminer le sous-ensemble optimal de variables.

2.2.3 Résultats

Les paramètres que nous avons utilisé pour l'exécution du SA :

- Température initiale **(T_0)** : 100000.0
- Taille d'une solution : 200
- Décroissance de la température : $\mathbf{T_{i+1} = sT_i}$ avec **s** = 0.87
- Nombre d'itérations à chaque température **(N_iter_max)** : 50
- Nombre d'itérations maximal **(N_restart_max) :** 10
- Mécanisme de perturbation : 5
- Codage réel
- Arrêt de système **(T_{min})** : 1.0

Nous présentons dans le tableau 5 les résultats d'apprentissage en utilisant le sous-ensemble de variables sélectionné par SA pour les données d'Azote et du Carbone.

Tableau 5 : Résultats du SA

		Azote	Carbone
Temps d'exécution		**1h27min**	**1h24min**
Cross-Validation(cv=4)	R^2	**0.94**	**0.86**
	R^2 ajusté	**0.93**	**0.85**
	MSE	**0.06**	**0.91**
	RMSE	**0.25**	**0.95**
Train	R^2	**0.95**	**0.90**
	R^2 ajusté	**0.94**	**0.90**
	MSE	**0.05**	**0.64**
	RMSE	**0.23**	**0.80**
	AIC	**-871**	**-64**
Test	R^2	**0.94**	**0.74**
	R^2 ajusté	**0.93**	**0.69**
	MSE	**0.07**	**1.10**
	RMSE	**0.26**	**1.05**
	AIC	**-315**	**50**

D'après les résultats obtenus, et avec des solutions de 200 variables, le score du R^2 a atteint 94% et R^2 ajusté 93% pour Azote, par contre, les résultats du Carbone sont avec R^2 74% et R^2ajusté 69%. Ces variables ne sont pas satisfaisantes pas ceci revient aux paramètres et leur réglage. Le problème qui s'impose avec l'utilisation du recuit simulé, c'est pour avoir des bons résultats, il faut choisir les valeurs des paramètres saignement. SA demande un réglage en variant tous les paramètres. Les tests finaux contiennent un réglage de tous les paramètres.

2.3 Application de colonie de fourmi

Le dernier algorithme que nous avons implémenté, est l'algorithme de la colonie de fourmi. Nous avons implémenté une approche proposée par Ahmed Al-Ani dans l'article [28] basée sur l'algorithme de colonie de fourmis. Cet algorithme est simple à implémenter, et n'exige pas un grand nombre de paramètres à régler.

2.3.1 Modélisation par Colonie de fourmis

Pour un problème de sélection de variable en utilisant les algorithmes de colonies de fourmis peut être énoncé comme suit : étant donné l'ensemble original ***F***, de ***n*** variables, trouver le

sous-ensemble ***S***, qui compose de m variables ***(m < n, S ⊂ F)***, tel que le score RMSE du modèle PLSR soit minimisé.

La représentation de la sélection des variables exploitée par les fourmis artificielles comprend les éléments suivants :

- ***n*** variables qui constituent l'ensemble original, $F = \{f_1, ..., f_n\}$.
- Un certain nombre de fourmis artificielles pour rechercher dans l'espace des caractéristiques (***na*** fourmis).
- τ_i, l'intensité de la piste de phéromone associée à la variable f_i, qui reflète la connaissance antérieure de l'importance de f_i.
- Pour chaque fourmi ***j***, une liste qui contient le sous-ensemble de variables sélectionné, $S_j = \{s_1, \ldots, s_m\}$.

Lors de la première itération, chaque fourmi choisira aléatoirement un sous-ensemble de ***m*** variables. Seuls les ***k*** meilleurs sous-ensembles, ***k < na***, seront utilisés pour mettre à jour l'essai de phéromone et influencer les sous-ensembles de variables de l'itération suivante. Dans la deuxième itération et les suivantes, chaque fourmi commencera avec ***m - p*** variables qui sont choisies aléatoirement parmi les ***k*** meilleurs sous-ensembles sélectionnés précédemment, où ***p*** est un nombre entier compris entre ***1*** et ***m - 1***.

De cette façon, les caractéristiques qui constituent les ***k*** meilleurs sous-ensembles auront plus de chances d'être présentes dans les sous-ensembles de l'itération suivante. Cependant, il sera toujours possible pour chaque fourmi de considérer d'autres variables.

2.3.2 Les étapes de colonie de fourmis

Voici les étapes de l'algorithme :

1. Initialisation :
    - Définir $\tau_i = cc$ et $\Delta\tau_i = 0$, ***(i = 1, ..., n)***, où ***cc*** est une constante et $\Delta\tau_i$ est la quantité de changement de la quantité de phéromone pour la variable ***fi***.
    - Définir le nombre maximal d'itérations.
    - Définir ***k***, où les ***k*** meilleurs sous-ensembles influenceront les sous-ensembles de l'itération suivante.
    - Définir p, où ***m-p*** est le nombre de variables avec lesquelles chaque fourmi commencera dans la deuxième et la suivante.
2. Si la première itération,
    - Pour ***j = 1*** à ***na***,

- Attribuer aléatoirement un sous-ensemble de m variables à S_j.

- Aller à l'étape 4

3. Sélectionnez les p variables restantes pour chaque fourmi :

   - Pour $mm = m-p+1$ à m,

     - Pour $j = 1$ à na,

       - Étant donné le sous-ensemble S_j, choisir la variable f_i en utilisant la sélection roulette pondérée en se basant sur les probabilités cumulées : Pour chaque variable, sa probabilité est calculée avec :

$$p_i = \frac{\tau_i}{\sum_{i=1}^{n} \tau_i}, \ \forall i = 1,2, \dots, n \qquad (17)$$

       - Remplacer les variables dupliquées par des variables choisis aléatoirement.

4. Evaluer le sous-ensemble sélectionné de chaque fourmi en utilisant le PLSR :

   - Pour $j = 1$ à na,

     - Estimer la valeur du RMSE obtenue par le PLSR de chaque sous-ensemble Sj.

   - Trier les sous-ensembles en fonction de leur RMSE. Mettre à jour le RMSE minimum, et stocker le sous-ensemble correspondant de variables.

5. En utilisant les sous-ensembles de variables des k meilleures fourmis, mettre à jour l'intensité de phéromone et initialiser les sous-ensembles pour l'itération suivante :

   - Pour $j = 1$ à k, /* mise à jour des intensités de phéromone de chaque variable*/

$$\Delta \mathcal{T}_i = \begin{cases} \frac{\max_{g=1:k}(RMSE_g) - RMSE_j}{\max_{h=1:k}\left(\max_{g=1:k}(RMSE_g) - RMSE_h\right)} & \text{si } f_i \in S_j \\ 0 & \text{Autre} \end{cases} \qquad (16)$$

$$\mathcal{T}_i = (1-\rho) \cdot \mathcal{T}_i + \Delta \mathcal{T}_i \qquad (18)$$

     Où ρ est une constante, $1-\rho$ représente l'évaporation de phéromone.

   - Pour $j = 1$ à na,

     - A partir des variables des k meilleures fourmis, produire aléatoirement $m - p$ sous-ensemble de variables pour la fourmi j, à utiliser dans l'itération suivante, et le stocker dans S_j.

6. Si nombre d'itérations est inférieur au nombre maximal d'itérations, passer à l'étape 3.
7. Déterminer le sous-ensemble optimal de variables.

2.3.3 Résultats

Les paramètres utilisés pour l'exécution du modèle hybride Ant Colony et PLSR :

- Nombre des fourmis **(na) :** 50
- Taille des sous-ensembles **(m)** : 20
- Nombre des itérations : 50
- Tau d'évaporation **(ρ)** : 0.5
- Nombre de meilleurs sous-ensembles **(k)** : 40
- Nombre de variables qui divise chaque sous-ensemble ***(P) :*** 33

Nous présentons dans le tableau 6 les résultats d'apprentissage en utilisant le sous-ensemble de variables sélectionné par colonie de fourmis pour les données d'Azote et du Carbone.

Tableau 6 : Résultats d'algorithme de Colonie de fourmis

		Azote	Carbone
Temps d'exécution		**17min**	**18min**
Cross-Validation(cv=4)	R^2	**0.93**	**0.85**
	R^2 ajusté	**0.92**	**0.83**
	MSE	**0.07**	**1.04**
	RMSE	**0.28**	**1.02**
Train	R^2	**0.93**	**0.88**
	R^2 ajusté	**0.93**	**0.87**
	MSE	**0.07**	**0.84**
	RMSE	**0.26**	**0.91**
	AIC	**-789**	**-0.24**
Test	R^2	**0.93**	**0.68**
	R^2 ajusté	**0.92**	**0.61**
	MSE	**0.07**	**1.38**
	RMSE	**0.28**	**1.17**
	AIC	**-301**	**72**

D'après les résultats obtenus, et avec des fourmis d'une taille de 20 variables, le score du R^2 a

atteint 93% et R^2 ajusté 92% pour Azote, par contre, les résultats du Carbone sont avec R^2 68% et R^2ajusté 61%. Ces variables ne sont pas satisfaisantes pas ceci revient aux paramètres et leur réglage. Avant de déduire des conclusions, il faut utiliser des paramètres convenable (Taille des fourmis). Les tests finaux contiennent un réglage de tous les paramètres.

Conclusion

Au cours de ce chapitre, nous avons présenté les différentes méthodes que nous avons implémenté en présentant leur modélisation et les étapes de chaque algorithme.

Les résultats obtenus ne sont pas satisfaisants et ceci revient au réglage des paramètres. Nous ne pouvons pas comparer les performances des algorithmes implémentés et donc de déduire des conclusions. Les tests finaux contenant des paramètres convenables et un réglage de ces paramètres est en cours d'attente à cause du manque du matériel (serveurs).

Conclusion générale

L'utilisation de l'analyse traditionnelle pour analyser et extraire la quantité des matières chimiques dans les plantes prend beaucoup de temps et demande des ressources financières importantes. Ceci a poussé les chercheurs à utiliser des nouvelles méthodes comme la spectroscopie infrarouge. Donc pour analyser la quantité des matières chimiques, ils utilisent des modèles de prédiction en les entrainant sur les données spectrales afin d'estimer ces quantités. Le problème qui se pose avec les données spectrales, c'est qu'elles sont d'une grande dimension et contiennent un nombre important de variables non pertinentes, et elles révèlent un taux élevé de corrélation et de colinéarité, ce qui influence les performances des modèles de prédiction.

Notre projet de ce stage au sein de la fondation MAScIR était de proposer une solution à ces problèmes en utilisant les méthodes d'optimisation pour faire une sélection de variables. Nous avons proposé au cours de ce projet un modèle hybride basé sur l'approche des méthodes enveloppes, en combinant la régression des moindres carrées partiels et des méthodes métaheuristiques pour déterminer les longueurs d'ondes adéquates en utilisant les données d'Azote et Carbone dans les tomates cerises.

Cependant, ce projet a rencontré certaines difficultés qui ont eux aussi contribué à ajouter des contraintes de plus à notre projet. Pendant que nous étions entrains d'avancer sur notre étude, il y avait un problème au niveau matériel, MAScIR a bien ses propres serveurs et stations de travail, cependant, ils sont partagés entre les équipes, donc nous n'avions pas un accès direct et libre au serveur et nous ne pouvions pas charger le serveur pour une longue durée afin d'effectuer plus de tests en utilisant des valeurs convenables des paramètres des méthodes d'optimisation et en effectuant un réglage de ces paramètres.

Les résultats de l'implémentation du modèle hybride en utilisant des différentes méthodes d'optimisation seront comparés avec les résultats des autres réalisations des autres membres participants au projet. Cependant ces réalisations ne sont pas encore prêtes, donc comparer les travaux et déduire des conclusions sont reportés jusqu'à l'obtention des résultats finaux et les résultats des autres membres.

En perspectives, puisque le processus des méthodes d'optimisation sera très lent lors de la phase du réglage de paramètres, nous chercherons alors à optimiser ces processus et d'implémenter le parallélisme sur ces algorithmes en bénéficiant de la puissance GPU. De même nous chercherons aussi à utiliser de données avec des différents prétraitements spectraux.

Bibliographie

[1] B. Nadler et R. R. Coifman, « The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration », *J. Chemom.*, vol. 19, nº 2, p. 107-118, févr. 2005, doi: 10.1002/cem.915.

[2] R. Rakotomalala, « Analyse de corrélation Étude des dépendances - Variables quantitatives ». [En ligne]. Disponible sur: https://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf

[3] R. Moratelli, C. S. de Andreazzi, J. A. de Oliveira, et J. L. P. Cordeiro, « Current and potential distribution of Myotis simus (Chiroptera, Vespertilionidae) », *mammalia*, vol. 75, nº 3, janv. 2011, doi: 10.1515/mamm.2011.028.

[4] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, et M. Hanpin, « Variables selection methods in near-infrared spectroscopy », *Anal. Chim. Acta*, vol. 667, nº 1-2, p. 14-32, mai 2010, doi: 10.1016/j.aca.2010.03.048.

[5] Legrand, « La sélection de variables ». [En ligne]. Disponible sur: http://theses.univlyon2.fr/documents/lyon2/2004/legrand_g/pdfAmont/legrand_g_chapitre02.pdf

[6] J. Hamon, « Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale », p. 160.

[7] D. Corne, C. Dhaenens, et L. Jourdan, « Synergies between operations research and data mining: The emerging use of multi-objective approaches », *Eur. J. Oper. Res.*, vol. 221, nº 3, p. 469-479, sept. 2012, doi: 10.1016/j.ejor.2012.03.039.

[8] E. Amaldi et V. Kann, « On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems », *Theor. Comput. Sci.*, vol. 209, nº 1-2, p. 237-260, déc. 1998, doi: 10.1016/S0304-3975(97)00115-1.

[9] M. Melanie, « An Introduction to Genetic Algorithms », p. 162.

[10] D. Fouskakis et D. Draper, « Comparing Stochastic Optimization Methods for Variable Selection in Binary Outcome Prediction, With Application to Health Policy », *J. Am. Stat. Assoc.*, vol. 103, nº 484, p. 1367-1381, déc. 2008, doi: 10.1198/016214508000001048.

[11] « S. KIRKPATRICK, C.D. GELATT, M.P. VECCHI, « Optimisation by simulated annealing, Science», Vol. 220, n0 4598, 1983.pdf ».

[12] R. Chibante, *Simulated Annealing : Theory with Applications*.

[13] M. J. Brusco, « A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis », *Comput. Stat. Data Anal.*, vol. 77, p. 38-53, sept. 2014, doi: 10.1016/j.csda.2014.03.001.

[14] R. Meiri et J. Zahavi, « Using simulated annealing to optimize the feature selection problem in marketing applications », *Eur. J. Oper. Res.*, vol. 171, nº 3, p. 842-858, juin 2006, doi: 10.1016/j.ejor.2004.09.010.

[15] H. Hachimi, « HYBRIDATIONS D'ALGORITHMES MÉTAHEURISTIQUES EN OPTIMISATION GLOBALE ET LEURS APPLICATIONS ».

[16] A. Ostfeld, *Ant colony optimization: methods and applications*. Rijeka, Croatia: Intech, 2011.

[17] M. Dorigo, V. Maniezzo, et A. Colorni, « Ant system: optimization by a colony of cooperating agents », *IEEE Trans. Syst. Man Cybern. Part B Cybern.*, vol. 26, nº 1, p. 29-41, févr. 1996, doi: 10.1109/3477.484436.

[18] S. Voß, « J. Dreo, A. Petrowski, P. Siarry, E. Taillard: Metaheuristics for Hard Optimization: Springer, Berlin, ISBN 10 354023022X, 2006, 369 pages », *Math. Methods Oper. Res.*, vol. 66, nº 3, p. 557-558, nov. 2007, doi: 10.1007/s00186-007-0180-y.

[19] C. Blum, « Beam-ACO—hybridizing ant colony optimization with beam search: an application to open shop scheduling », *Comput. Oper. Res.*, vol. 32, nº 6, p. 1565-1591, juin 2005, doi: 10.1016/j.cor.2003.11.018.

[20] P. Balaprakash, M. Birattari, T. Stützle, Z. Yuan, et M. Dorigo, « Estimation-based ant colony optimization and local search for the probabilistic traveling salesman problem », *Swarm Intell.*, vol. 3, nº 3, p. 223-242, sept. 2009, doi: 10.1007/s11721-009-0031-y.

[21] G. Di Caro et M. Dorigo, « AntNet: Distributed Stigmergetic Control for Communications Networks », *J. Artif. Intell. Res.*, vol. 9, p. 317-365, déc. 1998, doi: 10.1613/jair.530.
[22] M. A. Bouhlel, « Optimisation auto-adaptative en environnement d'analyse multidisciplinaire via les modèles de krigeage combinés à la méthode PLS », Institut Supérieur de l'Aéronautique et de l'Espace. [En ligne]. Disponible sur: https://www.theses.fr/2016ESAE0002.pdf
[23] T. Udelhoven, C. Emmerling, et T. Jarmer, « Quantitative analysis of soil chemical properties with diffuse reflectance spectrometry and partial least-square regression: A feasibility study », p. 11.
[24] « Composantes principales et régressions PLS parcimonieuses ». Univérsité de Toulouse. [En ligne]. Disponible sur: https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-sparse-pls.pdf
[25] J.-M. Roger et M. Ecarnot, « Grain 5 : Prétraitements », p. 17.
[26] R. Rakotomalala, « Tests de normalité Techniques empiriques et tests statistiques ». [En ligne]. Disponible sur: https://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf
[27] K.-E. Häggblom, « Basics of Multivariate Modelling and Data Analysis », *Princ. Compon. Anal.*, p. 39.
[28] A. Al-Ani, « Feature Subset Selection Using Ant Colony Optimization », p. 6.

Développement des méthodes métaheuristiques pour la sélection de variables pour les données spectrales

Résumé

La croissance des plantes aussi que leur rendement dépend étroitement de la qualité de certaines matières chimiques et organiques. Pour cette raison et depuis longtemps, les agriculteurs effectuaient des analyses dans les laboratoires pour mesurer les quantités de ces matières dans leurs plantes. Par contre ces analyses nécessitent beaucoup de temps. Une nouvelle alternative qui a été expérimentée ces dernières années consiste en l'estimation des taux des matières organiques et chimiques dans les plantes en se basant sur la spectroscopie infrarouge. Ces données spectrales sont d'une grande dimension, et présentent un ensemble des problèmes : corrélation, colinéarité, etc. Ce qui influence les performances des modèles de prédiction. Le projet consiste à faire une étude et conception d'un modèle hybride impliquant une combinaison d'une méthode d'optimisation et d'un modèle de prédiction. La construction du modèle hybride est basée sur l'approche enveloppe des méthodes de sélection de variables où le modèle de prédiction va permettre de tester les différents sous-ensembles de variables générés itérativement par les méthodes d'optimisation. Comme méthodes d'optimisation, nous avons utilisé en premier temps une méthode à base des algorithmes génétiques et une autre à base du recuit simulé, et enfin une méthode à base de colonie de fourmis. Or comme modèle de prédiction, nous avons utilisé la régression des moindres carrés partiels pour estimer les quantités du Carbone et d'Azote dans les tomates cerises. Les résultats présentés dans ce rapport ne sont qu'une implémentation des prototypes en attendant les résultats finaux d'après MAScIR afin de les comparer avec les résultats des travaux des autres membres du projet.

Mots clés : sélection de variables, spectroscopie, GA, SA, Colonie de fourmis

Development of metaheuristic methods for variable selection for spectral data

Abstract

The growth of plants as well as their yield depends closely on the quality of certain chemical and organic materials. For this reason and for a long time, the farmers carried out analyses in laboratories to measure the quantities of these materials in their plants. However, these analyses are very time consuming. A new alternative that has been tested in recent years is the estimation of the organic and chemical content of plants based on infrared spectroscopy. These spectral data are of a large dimension, and present a set of problems: correlation, collinearity, etc. This influences the performance of the prediction models. The project consists of a study and design of a hybrid model involving a combination of an optimization method and a prediction model. The construction of the hybrid model is based on the envelope approach of the variable selection methods where the prediction model will allow to test the different subsets of variables generated iteratively by the optimization methods. As optimization methods, we used first a method based on genetic algorithms and another based on simulated annealing, and finally a method based on ant colony. As a prediction model, we used partial least squares regression to estimate the amounts of Carbon and Nitrogen in cherry tomatoes. The results presented in this report are only an implementation of the prototypes while waiting for the final results from MAScIR in order to compare them with the results of the work of other members of the project.

Keywords : variable selection, spectroscopy, GA, SA, Ant colony

Master Systèmes Intelligents & Réseaux
Département d'Informatique
Faculté des Sciences et Techniques de Fès

A.U. 2020 - 2021