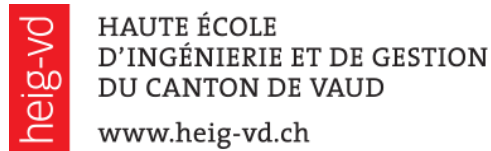


Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud
University of Applied Sciences Western Switzerland



Amélioration de la productivité de cultures tropicales par des méthodes d'apprentissage automatique

Caractériser et prédire la qualité des cafés colombiens

Thibault SCHOWING
Travail de Bachelor
10 juillet 2017

Professeur responsable : Carlos Andrès PEÑA
Superviseur (CIAT) : Sylvain DELERCE
Superviseur (CIAT) : Daniel JIMENEZ

Remerciements

Hugo Sylvain Daniel Osana
CIAT
PENA

Résumé

Résumé

Table des matières

1	Introduction	2
1.1	Question de recherche	2
1.2	Contexte du projet	2
2	À propos des données	3
2.1	Extraction, description et contextualisation des données . . .	3
2.1.1	Le système SICA	3
2.1.2	Données gustatives	3
2.1.3	Données climatiques	5
2.1.4	Données de sols	7
3	Méthodes de modélisation	9
3.1	Rappel des objectifs	9
3.2	Apprentissage supervisé	10
3.2.1	Random Forest	10
3.2.2	Partial Least Square (PLS)	11
3.2.3	Multi Block PLS	12
3.3	Apprentissage non-supervisé	14
3.3.1	SOM	14
3.4	Optimisation	16
3.4.1	Boosting	16
3.4.2	Cross-Validation	17
4	Analyse des données	18
4.1	Data Mining - Analyse exploratoire	18
4.1.1	Pre-processing	18
4.1.2	Analyse exploratoire	20
4.2	Résultats des analyses	40
4.2.1	Perceptron Multi-couches	40
4.2.2	Random Forest	40
4.2.3	Partial Least Squares	40
5	Discussion des résultats	41
6	Conclusion	42

1 Introduction

1.1 Question de recherche

A partir de données sur le climat, la qualité du sol et les pratiques culturelles, est-il possible d'expliquer et de prédire les différents traits de la qualité en bouche des cafés du département de Risaralda ?

1.2 Contexte du projet

Le sujet de ce Travail de Bachelor a été proposé par le « *Centro Internacional de Agricultura Tropical* » (CIAT) qui travaille dans le but d'améliorer la productivité et la gestion de l'agriculture en zone tropicale, et dont les bureaux se trouvent à Cali, en Colombie.

À 200 kilomètres de Cali, le comité des caféiculteurs de Risaralda souhaite pouvoir expliquer les différents traits de la qualité en bouche des cafés produits dans les différents secteurs de leur département. La filière café colombienne est en effet en concurrence avec d'autres pays exportateurs sur le marché international, et un des avantages comparatifs de la Colombie est que ses terroirs produisent des cafés de qualité et de caractères affirmés. Il est donc stratégique pour la fédération des caféiculteurs de Colombie d'être en mesure de faire valoir ces spécificités pour aller chercher la valeur ajoutée associée aux produits démarqués du lot.

Ce projet a pour but de trouver des méthodes de modélisation afin d'identifier les caractéristiques du café spécifiques à chaque secteur de la région en se basant sur des analyses gustatives, des données climatiques et géographiques, et d'autres données de pratiques culturelles.

Dans un premier temps, l'objectif est de catégoriser les cafés en tentant de trouver des tendances gustatives par rapport aux conditions de culture. Dans un second temps, il faudra pouvoir prédire la qualité en bouche des cafés par rapport aux conditions environnementales.

Le but de cette collaboration sur le long terme est de permettre au département de Risaralda de mettre en valeur la diversité de ses cafés, principalement à des fins de promotion auprès des acheteurs.

2 À propos des données

2.1 Extraction, description et contextualisation des données

2.1.1 Le système SICA

Le système SICA, pour *Sistema de Información Cafetera*, est un système géré par la Fédération Nationale des Caféculteurs (FNC), permettant d'identifier chaque parcelle de production de café en Colombie. C'est un système d'information d'envergure nationale, accessible via internet permettant de mettre à jour, consulter, analyser, modéliser et visualiser les données géospatiales sur les producteurs et les fermes de beaucoup de caféculteurs du pays. C'est l'outil d'information stratégique pour la conception, le développement, la cartographie et le suivi des politiques de compétitivité et de la durabilité du café colombien [2]. Chaque ferme possède un identifiant SICA, qui sera utilisé dans ce travail comme identifiant unique pour définir un café. Il est important car c'est ce numéro qui permet, via les services de la FNC, d'avoir un identifiant unique pour chaque parcelle et d'y associer des informations la concernant.

2.1.2 Données gustatives

Les données gustatives sont très relatives aux sens et à la perception de chaque gouteur. Cependant, la SCAA, *Speciality Coffee Association of America*, dispose d'un système de notation basé sur des hypothèses communautaires reconnues ce qui permet d'avoir une certaine régularité dans les données de dégustations. Les cafés sont notés sur 100 points répartis sur plusieurs critères : parfum/arôme, saveur, arrière-goût, acidité, corps, équilibre, douceur, clean-cup (absence de défauts marqués), uniformité et évaluation personnelle du testeur. Chacun de ces critères est noté sur 10 mais aussi par des termes qualitatifs. Par exemple, la saveur, c'est-à-dire la combinaison de l'odeur et du goût, la première impression qu'on a en goûtant le café, peut être notée 7/10 et "Caramel".

Le premier échantillon de données reçu contenait toutes ces informations de manière uniforme mais il s'est avéré que la partie mandante n'avait pas pu uniformiser la totalité des données brutes dans les délais. Ainsi, les données finalement reçues variaient beaucoup d'un document à l'autre, d'une part dans les données de dégustations présentes et dans le type de document mais aussi dans les méta-données permettant d'identifier précisément de

quelle café il s'agissait. Il a donc fallut effectuer un tri et ne garder que la masse qu'il était possible d'utiliser. Les critères permettant de garder une dégustation ou non sont les suivants : Identification possible du café grâce au numéro SICA ou au numéro d'identité du caféiculteur, présence des défauts physiques du café, présence des caractéristiques gustatives de manière uniforme. La FNC a été sollicitée afin de compléter les données une fois celles-ci triées afin d'y ajouter les numéros SICA ou les numéros d'identité manquants, et d'y ajouter les coordonnées de chaque parcelle sous la forme de référence spatiale EPSG :3116 en suite converties en coordonnées GPS classiques degrés-décimaux.

Traitement du café Pour avoir une vision globale, voici un petit résumé sur la production du café dans une des fermes du département de Risaralda. Cette ferme ne reflète pas la production de toutes les fermes du département cependant elle fait partie des meilleures plantation du secteur.

Lorsque les grains de café sont mûrs, ils sont récoltés à la main puis amené dans une grande cuve sous laquelle se trouvent les différentes machines permettant de traiter la baie afin d'en extraire le grain. La première de ces machine c'est la dépulpeuse qui permet d'enlever la partie charnue du grain. La pulpe est récupérée en contre-bas et le grain continue son chemin dans deux directions possibles. Si la ferme en est équipée, une machine appelée *desmucilagizador* va enlever la matière gluante entourant le grain, appelé *miel* ou en français *mucilage*, en le lavant. Si la ferme n'est pas équipé de cette machine, les grains vont être déversés dans une cuve où un processus de fermentation va être lancé variant entre une dizaine d'heures à plusieurs jours ce qui aura pour effet de laver le mucilage des grains. Une fois les grains lavés, ils seront séchés soit à l'air en utilisant la chaleur du soleil dans des grandes terrasse à café, ce processus prends environ dix jours, soit dans des machines à air chaud, plus onéreuse mais permettant de sécher de grandes quantités de grains en quelques heures. Une fois les grains séchés, ils sont vendus et l'étape suivant consiste à retirer de manière industrielle les grains endommagés car un seul grain peut rendre une tasse imbuvable. Des machines analyse les grains et éliminent ceux dont la densité ou la couleur n'est pas normale. [4]

Les différentes méthodes de préparation du café ont chacune leurs avantages économiques, écologiques ou gustatifs. La taille de l'arbre après un certain temps peut par exemple se faire de plusieurs manière affectant grandement le rendement. La complexité chimique de la fermentation peut apporter certains arômes tout comme un séchage rapide à l'air chaud peut en enlever.

2.1.3 Données climatiques

Les données climatiques comprennent les températures maximales, minimales et moyennes, la variation de température pendant la journée (DTR) et les quantités de précipitations. Les moyennes de ces mesures ont été calculée pour chaque mois et extrapolées sur une grande partie du territoire (à partir de stations météorologiques), permettant ainsi d'accéder aux mesures selon l'emplacement désiré à environ 500 mètres près.

En prenant par exemple les données de température maximale pour le mois de janvier 2011, en affectant pour chaque valeur une couleur, nous pouvons visualiser les données sous la forme d'une image comme sur la figure 2.1.

Les données climatiques sont données de 2011 à 2016. il faudra cependant faire attention au fait qu'un café dégusté en février 2011 a poussé bien plus tôt. Les processus de récolte, de nettoyage, de fermentation, de séchage et de torréfaction du grain prennent du temps. Ce temps a dû être pris en compte afin de sélectionner les bonnes données et a été fixé à 10 mois .

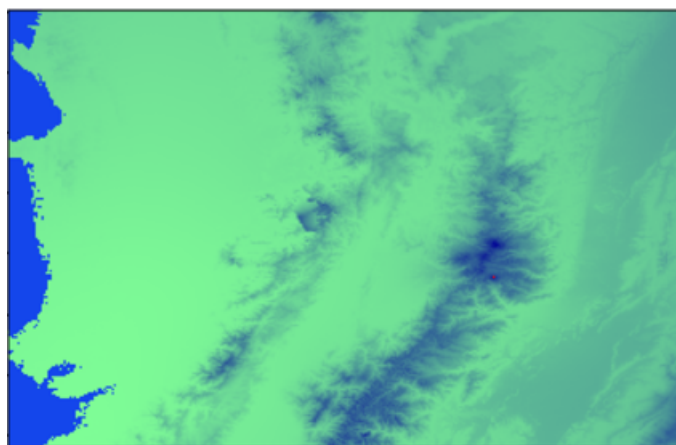


FIGURE 2.1 – Mise sous forme graphique du tableau des température maximales pour le mois de janvier 2011

Contexte climatique Colombien La Colombie se trouvant à proximité de l'équateur, on y trouve que deux saisons : l'été ou saison sèche (de décembre à janvier et de juillet à août) puis l'hiver ou saison des pluies (d'avril à mai et de octobre à novembre). Le relief du pays ainsi que sa taille, font varier le climat de chaud et humide pour la partie amazonienne et la région des caraïbes, désertique pour la région de Guajira tout au nord et glacial pour les zones en haute altitude à plus de 3000 mètres. Le département de Risaralda se trouve dans le centre de la Colombie dans la région de l'Axe du café et jouit de conditions climatiques, géographiques et

géologiques idéales pour la culture du café. Les températures oscillent entre 8 et 24 degrés mais un phénomène appelé *El Niño* perturbe régulièrement le climat à l'échelle du continent.

El Niño El Niño désigne un phénomène climatique qui se caractérise par une augmentation des températures de l'eau dans l'est du Pacifique sud due à une perturbation dans la circulation atmosphérique entre les pôles et l'équateur. Ces perturbations déplacent les zones de précipitations, modifient les routes des cyclones ou typhons provoquent à certains endroits de fortes précipitations et à d'autres de longues périodes de sécheresse. Même dans les zones tempérées, les périodes El Niño changent les habitudes climatiques. Durant l'été austral 2015-2016 s'est produit un des épisodes El Niño le plus fort jamais enregistré [6]. Si une grande partie de l'Amérique du Sud a été victime de fortes précipitations, la Colombie, elle, a subi une longue période de sécheresse et l'Europe a connu des records de chaleur. Sur la figure 2.2 on peut observer les différents pics correspondants à l'intensité du phénomène ainsi que pour son opposé, La Niña.

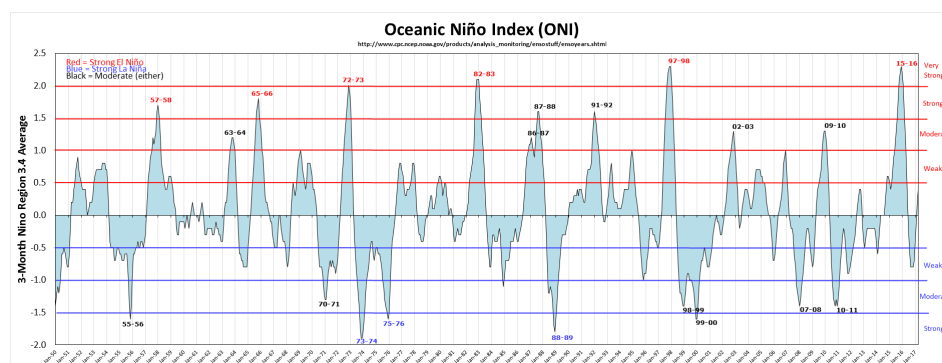


FIGURE 2.2 – Intensité du phénomène El Niño au cours des ans

Source : <http://ggweather.com/enso/oni.htm>

Impact du réchauffement climatique Outre les phases d'El Niño, il est nécessaire de rappeler que le climat mondial se réchauffe et que des conséquences se font ressentir. Le Centre du Commerce International [1] nous donne un aperçu des conséquences que ce réchauffement pourrait avoir pour la Colombie. "Les coûts de production sont susceptibles d'augmenter en raison des nouvelles conditions climatiques favorisant la prolifération des insectes, invasions et microbes pathogènes, et perturbant à l'équilibre naturel entre certains parasites et leurs prédateurs naturels. Les maladies se développeront vers de nouvelles zones. Les besoins en eau peuvent augmenter en raison de températures plus élevées causant plus d'évaporation, forçant de nombreux agriculteurs à recourir à l'irrigation. Dans certaines régions, les agriculteurs voudront transférer leur production de café à de

plus hautes altitudes afin de chercher d'un meilleur environnement.” (Guide de l'Exportateur de Café, CCI, 2011 [4])

2.1.4 Données de sols

Les données de qualité de sol sont subdivisées en profils. Chaque profil est séparé en une ou plusieurs couches d'une certaine profondeur dont sont renseignées les caractéristiques comme le pH, la texture ou encore le taux de matière organique. Les différentes textures sont présentées sur la figure 2.3. Afin d'avoir des données uniformes, les moyennes sur les 3 premières couches jusqu'à 1 mètre de profondeur ont été réalisées pour le pH et le niveau de matière organique alors que pour la texture, la somme des variable binaire a été effectuée.

Les données proviennent d'un GIS (Geographical Information System), d'où il a été possible de croiser les données point par point afin d'extraire le profil de sol correspondant à un set de coordonnées GPS. Malheureusement les profils ne contiennent que pH, matière organique et texture. Des données très hétérogènes contenaient d'autres informations sur la composition chimique du sol mais leur structure et leur répartition irrégulière dans la zone de travail ont forcé à abandonner leur utilisation par manque de temps et de ressources.

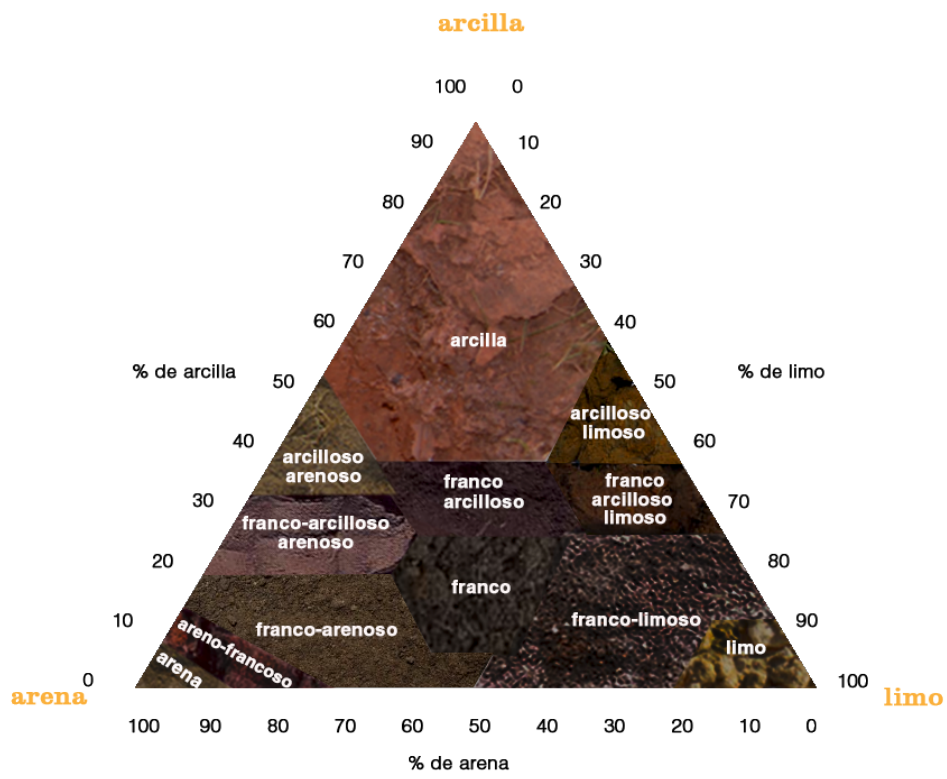


FIGURE 2.3 – Triangle représentant les différentes textures de sols
 Source : <http://www.construnatura.com/esp/articulo/agricultura-ecologica/el-suelo-como-fuente-de-vida-propiedades-ii->

3 Méthodes de modélisation

3.1 Rappel des objectifs

Ce projet a deux principaux objectifs. Le premier est de trouver s'il existe différents groupes de café ayant des relations entre les conditions de culture et les caractéristiques physiques ou sensorielles. On cherche donc dans cette première partie à caractériser les cafés. On peut ici parler de clustering. Le second objectif serait de prédire les caractéristiques physiques ou sensorielles à partir des données sur les conditions de culture. Nous avons donc ici plusieurs possibilités de manières d'agir. Par exemple, si le clustering a réussi à diviser les cafés en différentes classes, on cherchera à prédire dans quelle classe se situe un nouveau café. Plus spécifiquement, on pourra se concentrer sur certains attributs du café, par exemple l'acidité, afin d'estimer quelle sera la note attribuée.

3.2 Apprentissage supervisé

Le but de l'apprentissage supervisé est d'expliquer des sorties (outputs) à partir d'entrées (inputs). Des règles sont calculées à partir de données d'apprentissage selon différents modèles. Par la suite, le modèle est utilisé pour catégoriser des nouvelles données. On essaiera ici d'expliquer les données gustatives du café ou ses défauts physiques à l'aide des données climatiques et de sols.

3.2.1 Random Forest

La méthode Random Forest, ou *forêts d'arbres décisionnels* en français, fait partie des méthodes ensemblistes [3], qui utilisent la combinaison de plusieurs modèles de base, d'apprentissage automatique. Elle combine les concepts de sous-espaces aléatoires et de bagging.

Le bagging, ou *bootstrap agregation*, consiste à sous-échantillonner (ou ré-échantillonner au hasard avec doublons) le set d'entraînement et de faire générer à l'algorithme voulu un modèle pour chaque sous-échantillon. On utilise le bagging pour réduire la variance de la fonction de prédiction estimée. Le bagging semble bien fonctionner pour les procédures avec une grande variance et un petit biais, comme les arbres de décision. [5]

Random Forest effectue donc un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents [7].

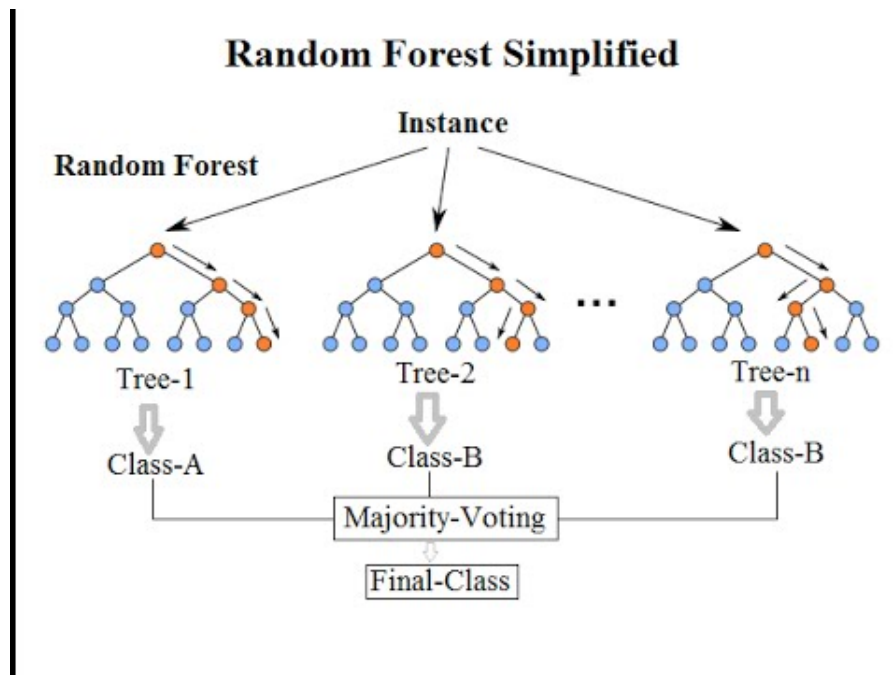


FIGURE 3.1 – Schéma simple du fonctionnement de Random Forest.
Source : <https://www.youtube.com/watch?v=ajTc5y3OqSQ>

3.2.2 Partial Least Square (PLS)

PLS, originalement pour *Partial Least Squares regression* puis plus récemment pour *Projection to Latent Structures* est une méthode qui combine des propriétés de la PCA ainsi que de multiples régressions linéaires. Au lieu de trouver un hyperplan de la variance maximale, entre les variables dépendantes et indépendantes, cette méthode va trouver un modèle de régression linéaire en projetant les variables indépendantes et dépendantes dans un nouvel espace. Ce sont les variables latentes. Cette méthode est particulièrement utile lorsqu'il est nécessaire de prédire un jeu de variables dépendantes à partir d'un très grand jeu de variables indépendantes.

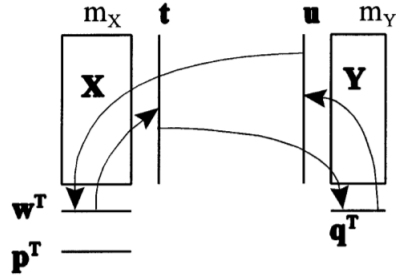


FIGURE 3.2 – Méthode PLS. X est représenté par son score t et Y par u . Une première estimation de U est multipliée à travers X pour obtenir une approximation du poid ω_t . Le poid est normalisé pour être de longueur 1 et remultiplié à travers X pour produire t . A partir de t et de Y , le poid q^T est obtenu ce qui donne un nouveau vecteur u . Cette opération est répétée jusqu'à la convergence de t . [8]

3.2.3 Multi Block PLS

La PLS multi block est une extension de la méthode PLS qui sépare les variables indépendantes en plusieurs blocks afin de leur donner une plus grande interprétabilité et plus d'informations sur la structure générale des données. Dans le cadre de ce projet, on peut imaginer séparer les données climatiques des données de sol par exemple. L'exécution est très similaire à la méthode PLS classique.

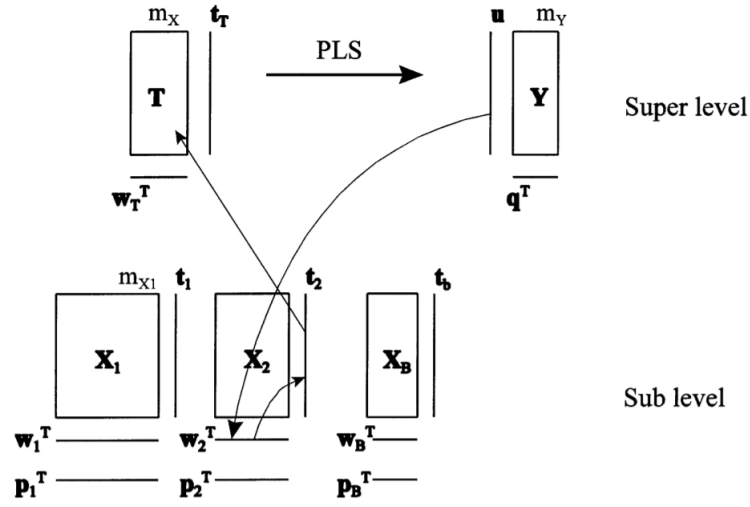


FIGURE 3.3 – Méthode MBPLS. Un score de départ u est régressé sur tous les blocs X_b pour donner les poids variables du bloc w_b^T . Les poids des variables de blocs sont normalisés à la longueur un et multipliés par les blocs pour donner les scores de blocs t_b . Les scores de blocs sont combinés dans le super bloc \mathbf{T} . Un cycle PLS entre \mathbf{T} et \mathbf{Y} est effectué pour donner le poids supérieur \mathbf{w}_T^T , qui est également normalisé à la longueur un, et le super score t_T . L'opération est répétée jusqu'à la convergence de t_T . [8]

3.3 Apprentissage non-supervisé

Contrairement à l'apprentissage supervisé, l'apprentissage non-supervisé tente de trouver des groupes dans des données hétérogènes. Le but est d'extraire des connaissances à partir de ces données. Comme mentionné dans la partie 3.1, notre but est de découvrir différents groupes de café identifiables.

3.3.1 SOM

Les différentes classes gustatives d'un café peuvent être considérées comme des entrées afin vérifier s'il est possible de regrouper différents cafés qui se distingueraient. Afin de trouver les différentes catégories de café, nous testerons les capacités de l'algorithme SOM (pour Self Organizing Map ou Cartes Auto Adaptatives en français) qui utilise un réseau de neurones pour étudier la répartition des données dans un espace de grande dimension.

Un bel exemple de SOM est celui de la carte de la pauvreté mondiale réalisé par le *Department of Computer Science and Engineering* de l'université *Helsinki University of Technology*.

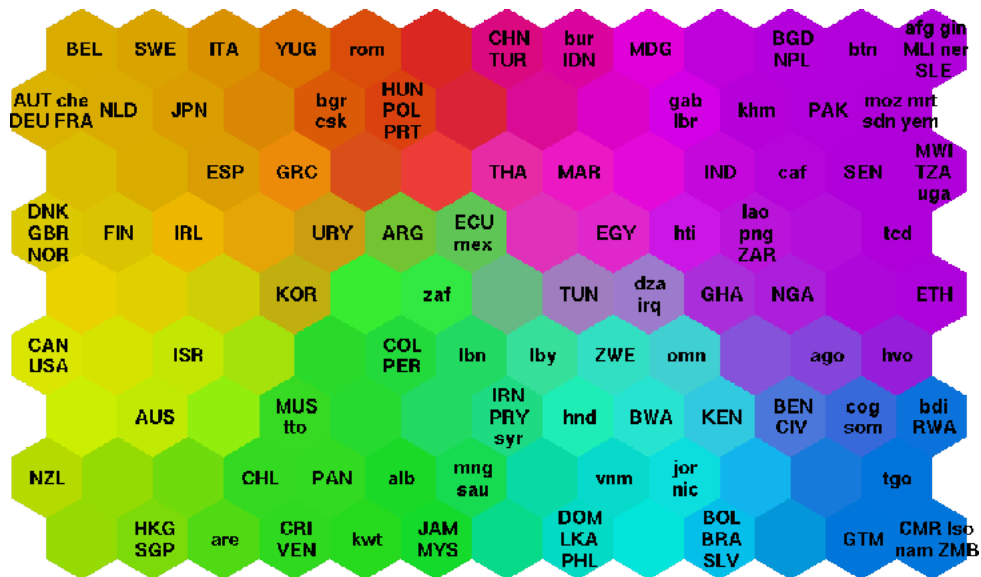


FIGURE 3.4 – Pays organisés en SOM d'après des indicateurs de pauvreté.
Source : <http://www.cis.hut.fi/research/som-research/worldmap.html>

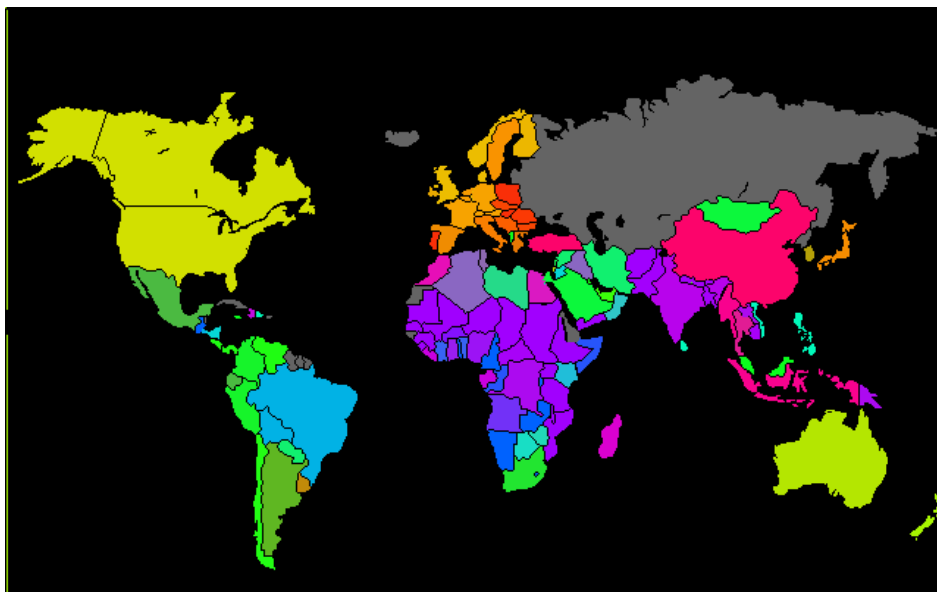


FIGURE 3.5 – Pays correspondants à la carte SOM de la figure 3.4
Source : <http://www.cis.hut.fi/research/som-research/worldmap.html>

3.4 Optimisation

3.4.1 Boosting

Le principe du boosting est quelque peu différent du bagging (voir section 3.2.1). Les différents classificateurs sont pondérés de manière à ce qu'à chaque prédiction, les classificateurs ayant prédit correctement auront un poids plus fort que ceux dont la prédiction est incorrecte.

Adaboost est un algorithme de boosting qui s'appuie sur ce principe, avec un paramètre de mise à jour adaptatif permettant de donner plus d'importance aux valeurs difficiles à prédire, donc en boostant les classificateurs qui réussissent quand d'autres ont échoué. Des variantes permettent de l'étendre à la classification multiclassées. Adaboost s'appuie sur des classificateurs existants et cherche à leur affecter les bons poids vis à vis de leurs performances [3].

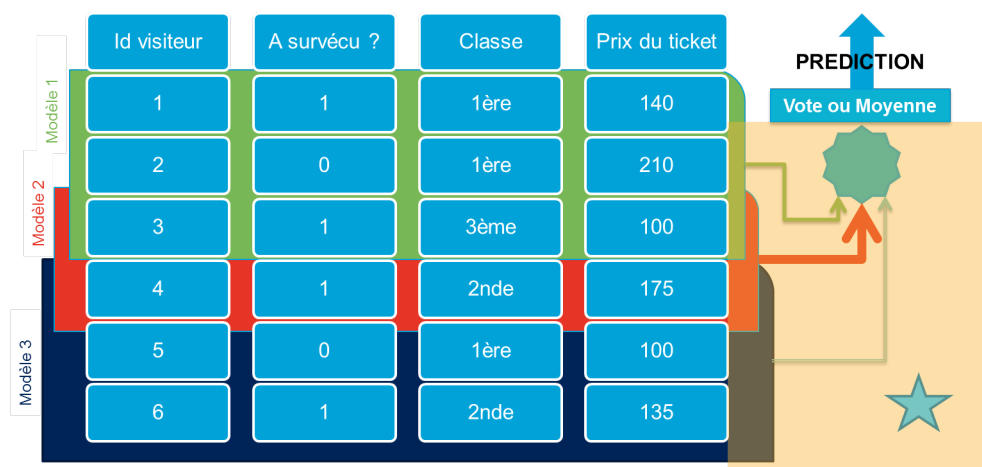


FIGURE 3.6 – Schéma du fonctionnement du Boosting

Source : <http://www.cis.hut.fi/research/som-research/worldmap.html>

Gradient Boosting est une méthode de machine learning utilisée pour les problèmes de classification et de régression. Elle fait aussi partie des méthodes ensemblistes, et est utilisée majoritairement avec des arbres de décision. L'idée est encore d'agréger plusieurs classificateurs ensemble mais cette fois en les créant itérativement.

Le classifieur du gradient boosting est donc au final paramétré par les poids de pondération des différents mini-classifieurs, ainsi que par les paramètres des fonctions utilisées. Il s'agit donc d'explorer un espace de fonctions simples par une descente de gradient sur l'erreur [3].

3.4.2 Cross-Validation

Contrairement au bagging qui est utilisé pour réduire l'overfitting en entraînant plusieurs modèles sur des données ré échantillonnées (avec répétition) puis en construisant un modèle sur la moyenne de ces modèles, la cross-validation est utilisée pour tester la fiabilité d'un modèle en se basant sur un échantillonnage des données d'entraînement et de test. Il existe plusieurs méthodes : « holdout method », « k-fold cross-validation » et « leave-one-out cross-validation ».

La première consiste à diviser le set de données en deux et en utilisant une partie pour entraîner le modèle puis une autre pour le tester. L'erreur est estimée en calculant un score de performance avec une méthode comme MSE (Erreur Quadratique Moyenne ou *Mean Square Error*).

Étant donné que les données sont souvent trop peu nombreuses pour se permettre de laisser tomber dès le départ une partie des données, la k-fold cross-validation devient utile. On divise le set en k échantillons puis on en sélectionne un comme étant le set de test puis les k-1 autres comme étant le set d'entraînement. On répète l'opération en sélectionnant chaque fois un échantillon différent pour le test. Le score de performance est calculé en réalisant la moyenne des scores des k validations effectuées. La méthode « Leave-one out » utilise le même principe mais en ne laissant qu'une seule entrée en dehors du set d'entraînement à chaque tour [5].

4 Analyse des données

4.1 Data Mining - Analyse exploratoire

4.1.1 Pre-processing

Une grande partie de l'étape de préprocessing a été réalisée durant l'extraction des données. Il a en effet fallu extraire les données d'une manière uniforme et cohérente dès le départ afin de ne pas se retrouver avec des variables présentes uniquement dans certaines parties du set de données ou avec des variables incohérentes comme cela a été le cas de certains cafés qui avaient comme note de dégustation plus de cent points sur cent par exemple. Des éliminations ou des corrections ont été réalisées de manière automatisées ou manuelles afin de supprimer les erreurs. On notera parmi les corrections importantes le remplacement de virgules par des points, quelques erreurs de frappes (68.5 points sur 10 au lieu de 6.85 par exemple) ou encore des utilisations d'unités différentes. Le dataset résultant est décrit dans l'Annexe A de ce projet. Une fois cette étape de nettoyage réalisée, les premières informations ont pu être extraites des données.

Le schéma 4.1 résume rapidement les différentes élimination d'observations au cours des étapes de construction du set de données.

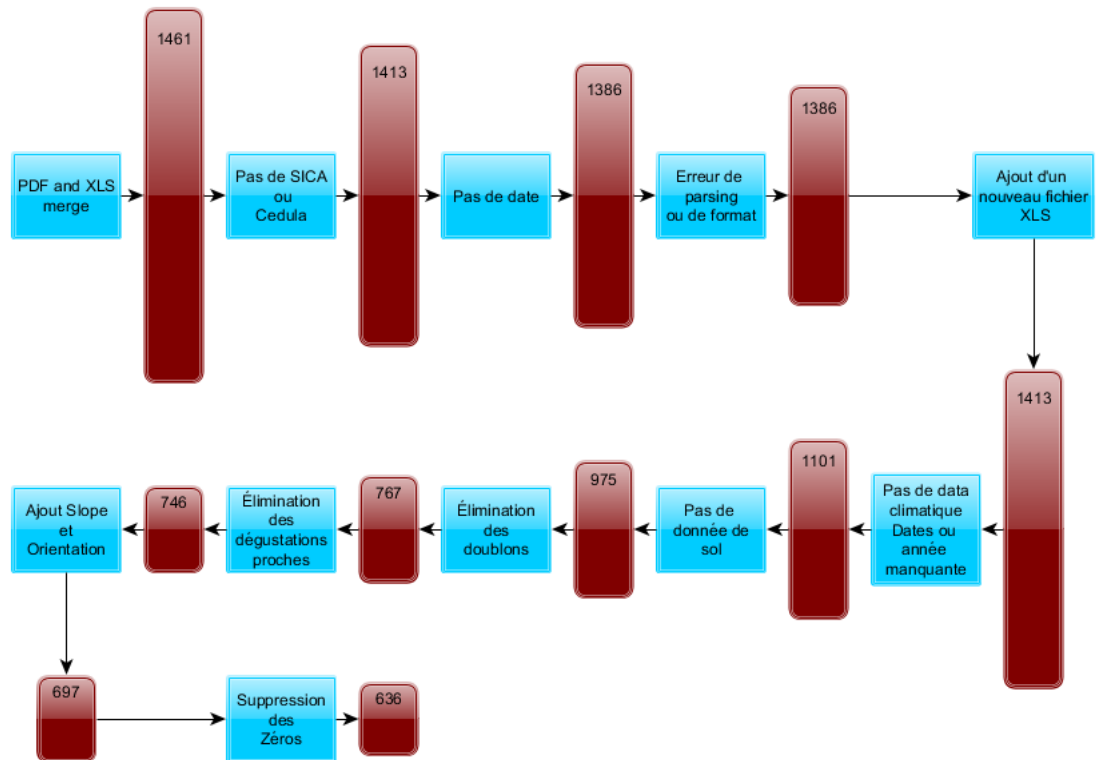


FIGURE 4.1 – Étape de construction du set de données et pertes d’observations.

Variables dépendantes Les variables de sorties, ou dépendantes, sont composées des dix analyses de dégustation et éventuellement de la totalité des défauts physiques des grains. Sauf pour les défauts physiques des grains, qui sont éliminés avant la dégustation, les variables de sorties sont très corrélées entre elles et il a été discuté avec le responsable du comité de sélectionner les variables les plus importantes. Ces variables sont *Acidez*, *Dulzor* et *Puntaje Total* ainsi que la catégorie décrite dans le paragraphe suivant qui a été ajouté au set de données.

Catégories Les cafés colombiens sont réputés excellents mais sur les 100 points attribuable lors des dégustations qu’est-ce que cela représente ? Le tableau 4.1 nous donne une bonne idée de l’échelle représentée. Ces catégories seront utilisées dans le set de données pour tenter de réaliser une classification.

TABLE 4.1 – Catégories de cafés d’après le nombre de points
Source : <http://www.scaa.org/?page=resources&d=cupping-protocols>

Total Score	Quality Classification	Specialty or not	Category
90-100	Outstanding	Specialty	1
85-89.99	Excellent	Specialty	2
80-84.99	Very Good	Specialty	3
<80.0	Below Specialty Quality	Not Specialty	4

Élimination des résultats avec zéro points Certain cafés du dataset ont zéro points en sortie pour chacune des 10 catégories notées. Il a été décidé de ne pas prendre en compte ces cafés lors des calculs de prédiction ou de clustering car la qualité du sol ou du climat ne peut justifier une telle baisse de qualité à elle seule dans une région réputée propice et qu’un défaut de traitement du grain, un mauvais tri avant la dégustation ou autre facteur externe doit en être la cause. Cependant, afin de se faire une idée d’éventuelles causes de cette qualité médiocre, les cafés sus-mentionnés ont été gardés pour la réalisation des cartes SOM au chapitre 4.1.2.

4.1.2 Analyse exploratoire

Emplacement des fermes

L’emplacement des différents cafés par rapport aux nombre de points est présenté sur la figure 4.2. La catégorie 1 (non représentée) correspond aux cafés avec plus de 90 points, la 2 aux cafés avec plus de 85, la 3 aux cafés avec plus de 80 et la 4 aux cafés en dessous de 80, ne correspondant donc pas à la qualité ”Specialty”. On peut y voir que l’emplacement dans le département n’as pas d’incidence sur les résultats, la répartition des différentes classes étant très uniforme.

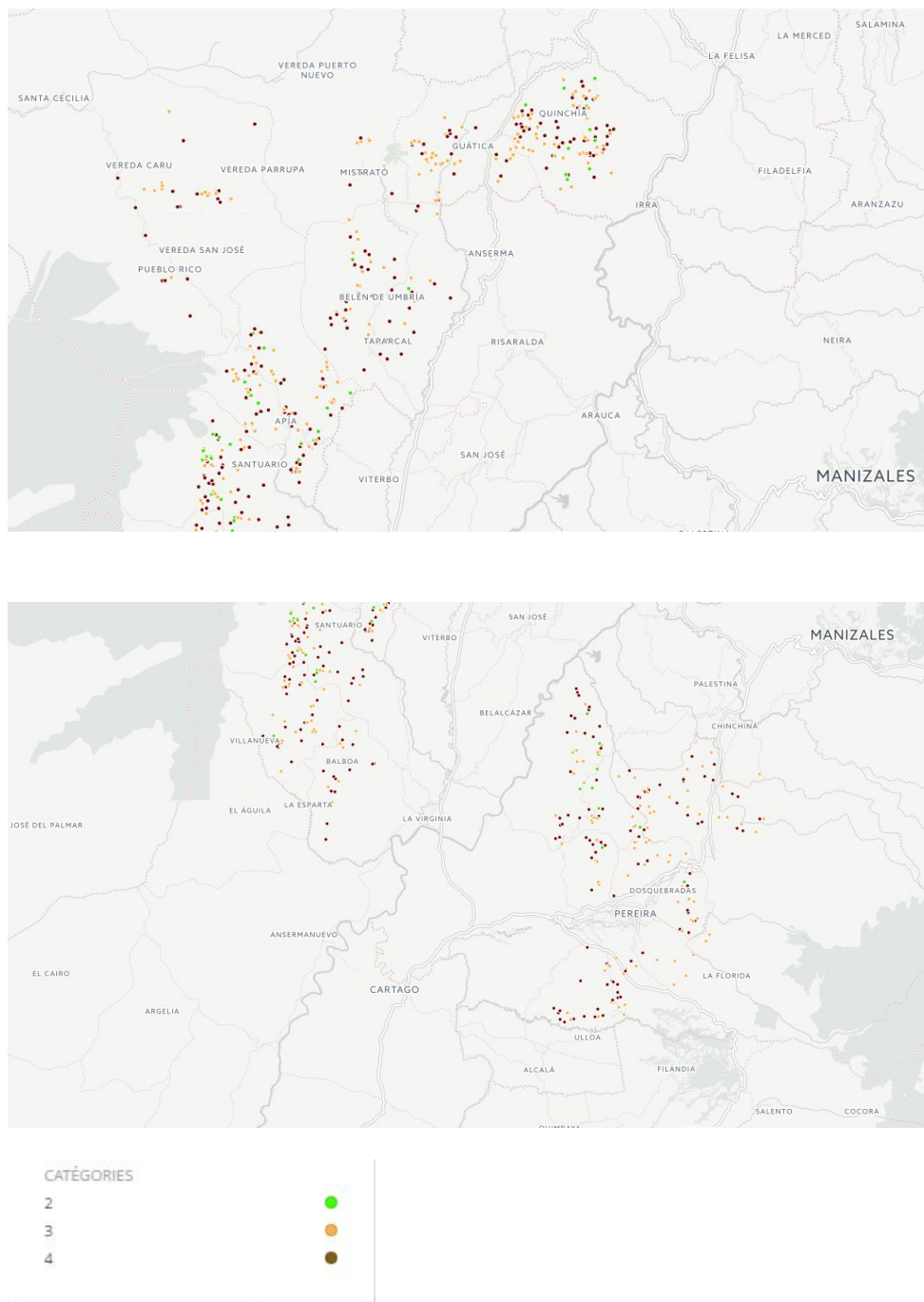


FIGURE 4.2 – Emplacement des fermes avec coloration selon le nombre de points attribués au café.

Corrélations entre variables

Afin d'avoir une bonne vue d'ensemble sur les variables et leurs liens, les matrices de corrélation ont été calculées pour toutes les variables. Premièrement la corrélation entre les différentes sorties. Sur la figure 4.3 on peut observer que les défauts physiques des grains ne sont que très peu liés entre eux ou avec les résultats de dégustation. On remarque cependant que les données gustatives du café sont fortement liées entre elles.

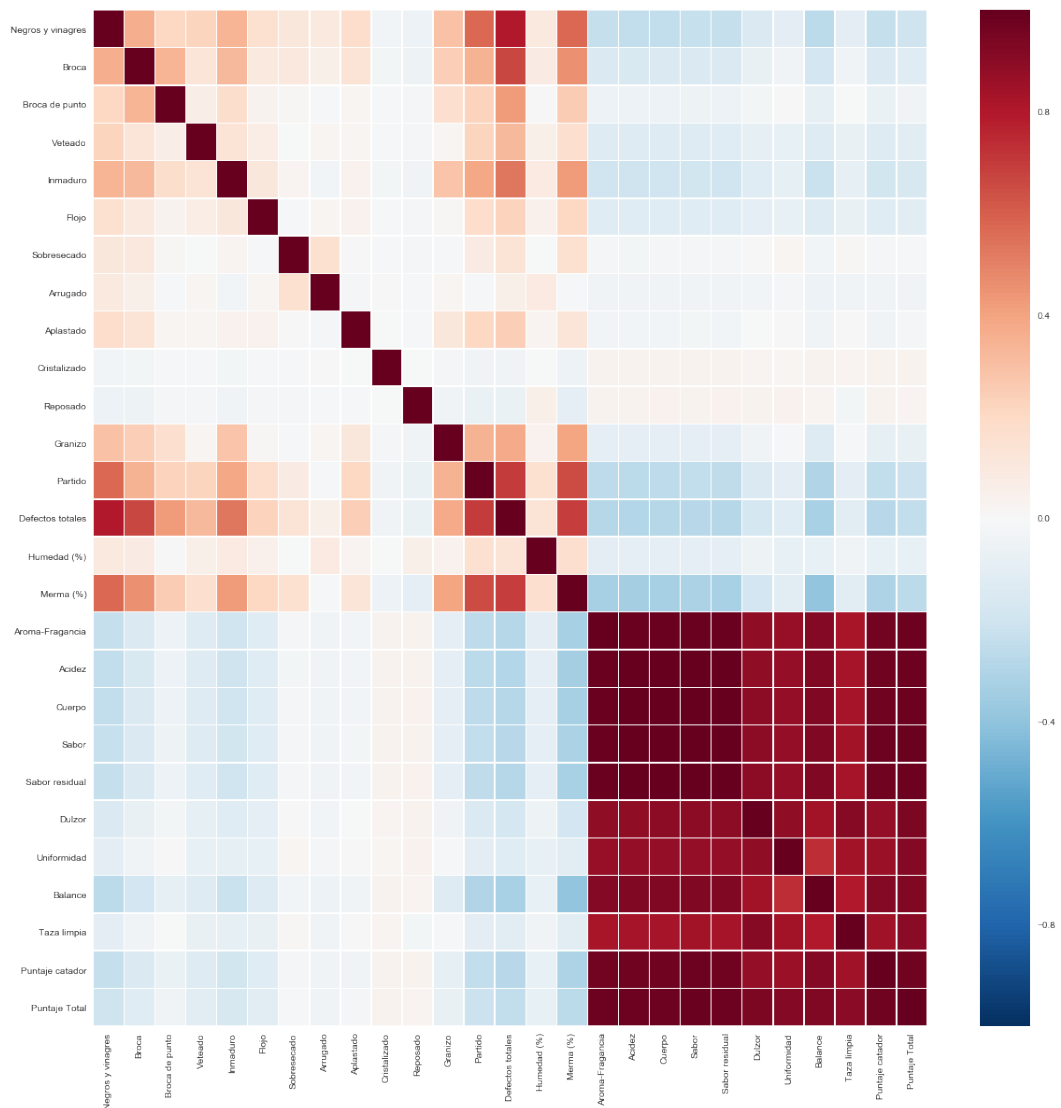


FIGURE 4.3 – Matrice de corrélation entre les différentes sorties.

Sur la figure 4.4 on peut observer les corrélations entre toutes les variables.

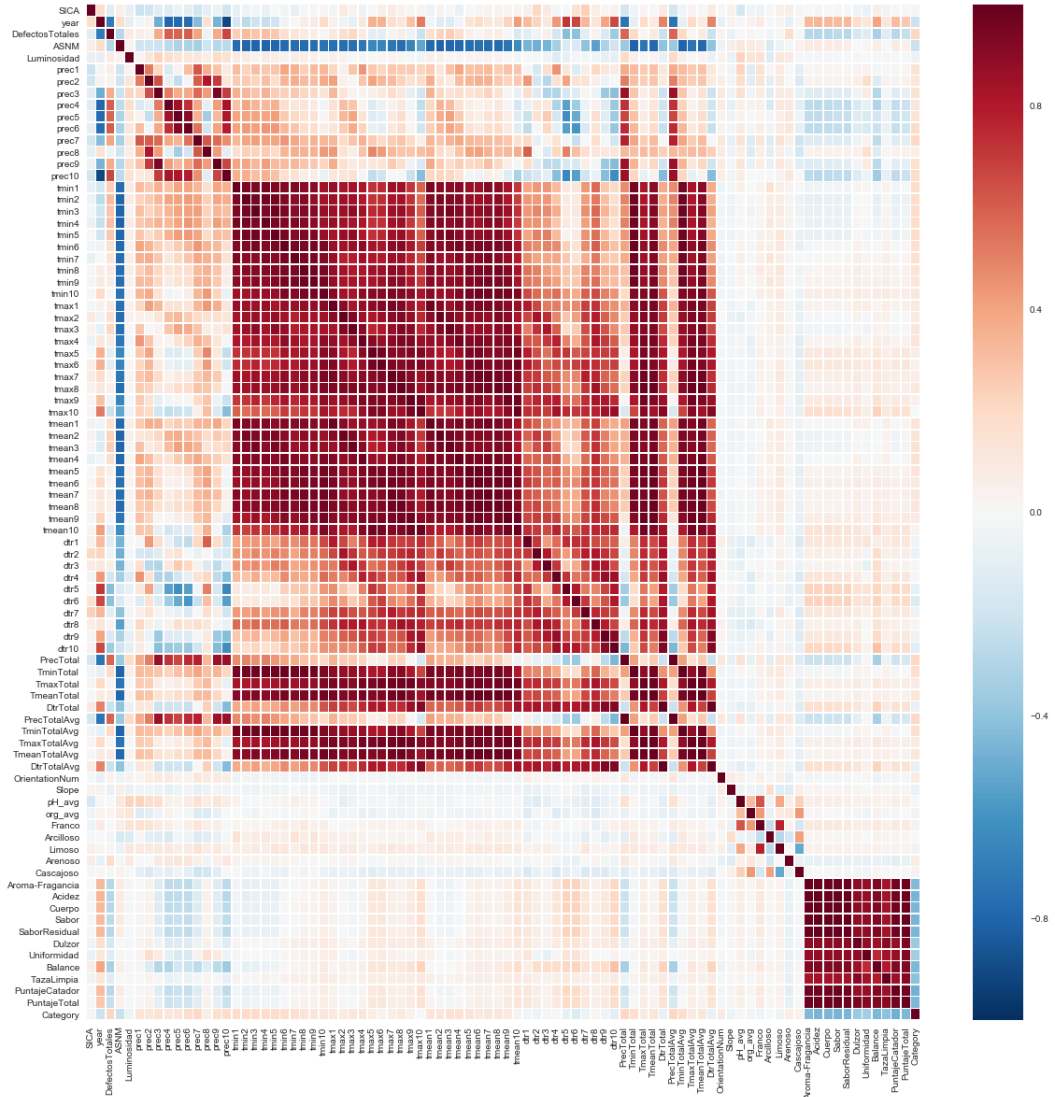


FIGURE 4.4 – Matrice de corrélation entre toutes les variables.

On observe que les données de sol sont parfois corrélées entre-elles mais presque pas du tout avec les données climatiques. On peut donc déjà soulever que le climat n'a pas ou peu d'influence sur la texture, le pH ou le taux de matière organique du sol des zones étudiées. Les précipitations ont une influence importante sur les défauts physiques des grains, en revanche nous n'avons aucune variable qui a une corrélation marquée avec les points totaux et donc la qualité du café.

Principal Component Analysis (PCA) et Clustering

La PCA, pour Analyse en Composantes Principales en français, est une méthode qui consiste à transformer un jeu de variables corrélées en nouvelles variables dé-corrélées les unes des autres. Ces nouvelles variables sont appelées composantes principales et permettent de rendre l'information moins redondante. Pour faire plus simple, l'utilité de la Composante Principale est de réduire le nombre de variables tout en gardant un maximum d'information. La figure 4.5 montre une représentation graphique de la composante principale.

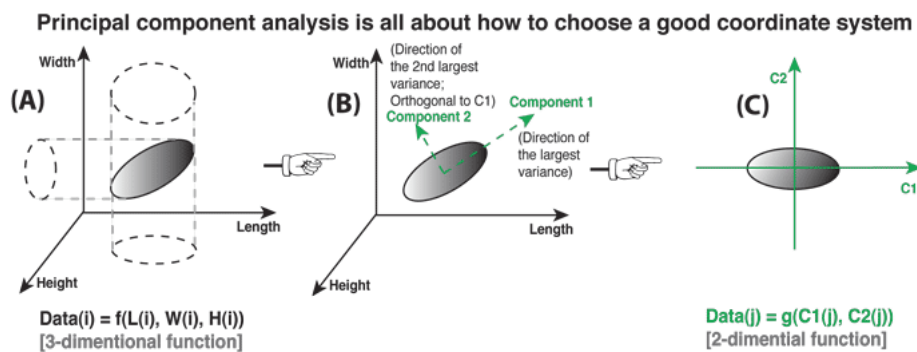


FIGURE 4.5 – Description de l'Analyse en Composante Principale. (A) Description d'un objet simple de manière compliquée (trois dimensions pour par exemple une ellipse en papier) (B) Trouver des nouvelles variables (axes de coordonnées) orthogonaux l'un à l'autre qui pointent dans les directions de la plus grande variance (C) Utiliser les nouvelles variables (axes) pour décrire l'objet d'une manière plus simple.

Résultats de la PCA L'analyse sur une version compacte des variables a donné les résultats présentés dans le tableau 4.2 et sur la figure 4.6. Une première analyse avait été effectuée sur le set de données complet, c'est-à-dire avec la totalité des données climatiques et non les moyennes, et les résultats se sont avérés similaires mais plus difficilement lisibles. Il a donc été choisi de résumer les variables pour réaliser la PCA et le clustering. La PCA du tableau ?? ne comprend pas les cafés avec zéro points.

	PC1	PC2	PC3	PC4	PC5	PC6
ASNM	0.4125	0.0766	-0.0360	0.1766	-0.1337	0.0397
Luminosidad	-0.0169	0.2320	0.1532	-0.3178	-0.2836	0.0157
PrecTotalAvg	-0.1714	0.1424	0.0222	-0.6767	-0.0667	0.0926
TminTotalAvg	-0.4589	-0.0115	0.1086	-0.1232	0.0013	0.0809
TmaxTotalAvg	-0.4745	-0.0334	0.0296	0.1501	-0.0253	-0.0329
TmeanTotalAvg	-0.4803	-0.0252	0.0631	0.0409	-0.0149	0.0133
DtrTotalAvg	-0.3323	-0.0517	-0.0884	0.4712	-0.0529	-0.1769
OrientationNum	0.0623	0.0592	-0.0886	-0.1839	0.3797	-0.5662
Slope	0.0657	-0.1842	0.1179	0.1047	0.0434	0.6905
pH_avg	0.0135	0.3892	0.3844	-0.0168	0.1244	0.0679
org_avg	0.0493	0.1781	0.5036	0.2176	-0.1639	-0.1841
Franco	-0.0157	0.5453	0.1441	0.1921	0.1676	0.1108
Arcilloso	-0.0180	-0.3026	0.2993	-0.0844	0.4532	0.1751
Limoso	-0.0754	0.5018	-0.2016	0.1095	0.2726	0.1712
Arenoso	-0.0210	0.0645	0.0180	0.0048	-0.6303	-0.0156
Cascajoso	0.0760	-0.2236	0.6130	0.0095	0.0109	-0.2053

TABLE 4.2 – Tableau des rotations des six premiers composants de la PCA avec mise en évidence des variables les plus importantes par composante.

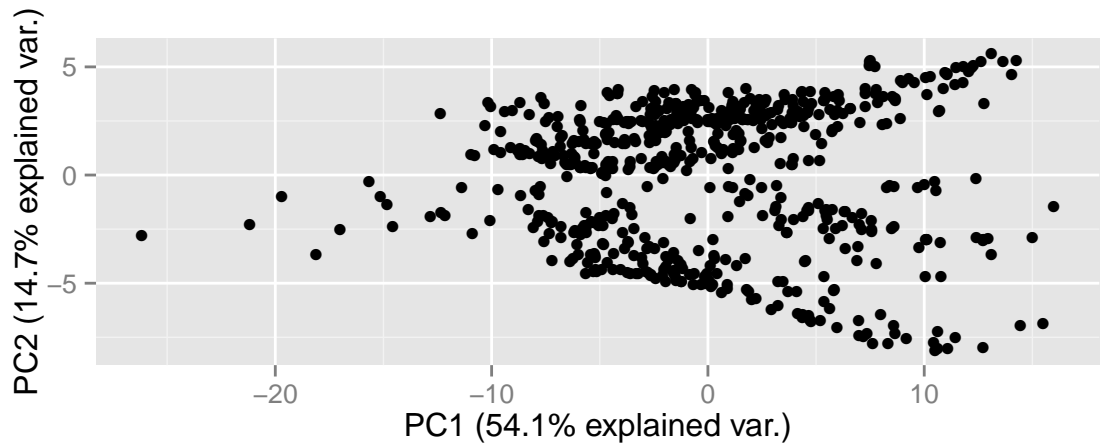


FIGURE 4.6 – Résultats de la PCA sous forme graphique. Réalisé avec la totalité des variables

Analyse des composantes Le tableaux 4.2 montre l'importance des variables dans les différentes composantes de la PCA réalisée avec un jeu de variables simplifié pour plus de lisibilité, après vérification que l'importance des types de variables des deux tableaux était similaire.

La première composante met en évidence les températures par rapport à l'altitude alors que la deuxième et la troisième mettent en évidence principalement les caractéristiques du sol. La quatrième montre une dé-corrélation entre les précipitations moyennes et les DTR et la cinquième composante montre une corrélation entre la texture argileuse du sol et son orientation à l'opposé à d'un sol sablonneux. La sixième composante met en évidence une relation entre l'orientation et les sols rocaillieux à l'opposé des sols pentus.

Clustering Afin de vérifier la présence éventuelle de groupes d'individus parmi la population de café, nous réalisons un HCPC, pour *Hierarchical Clustering on Principal Components*, à l'aide de la PCA. La figure 4.7 nous montre l'arbre hiérarchique créé ainsi que le nombre de cluster proposé.

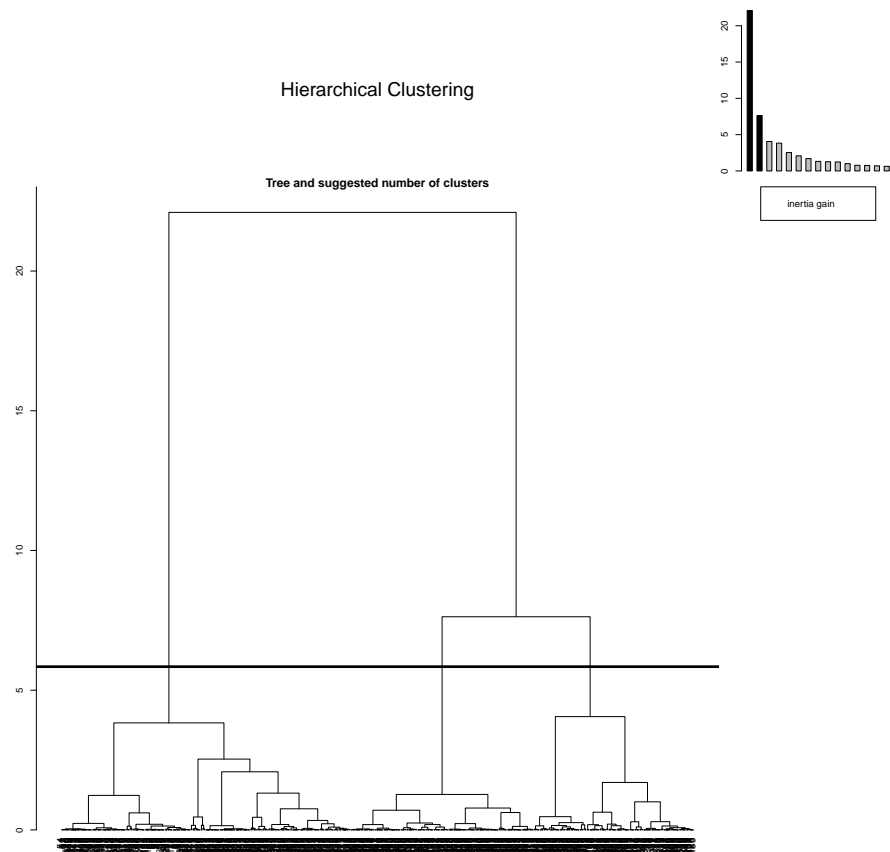


FIGURE 4.7 – HCPC et proposition de nombre de cluster

Hierarchical clustering on the factor map

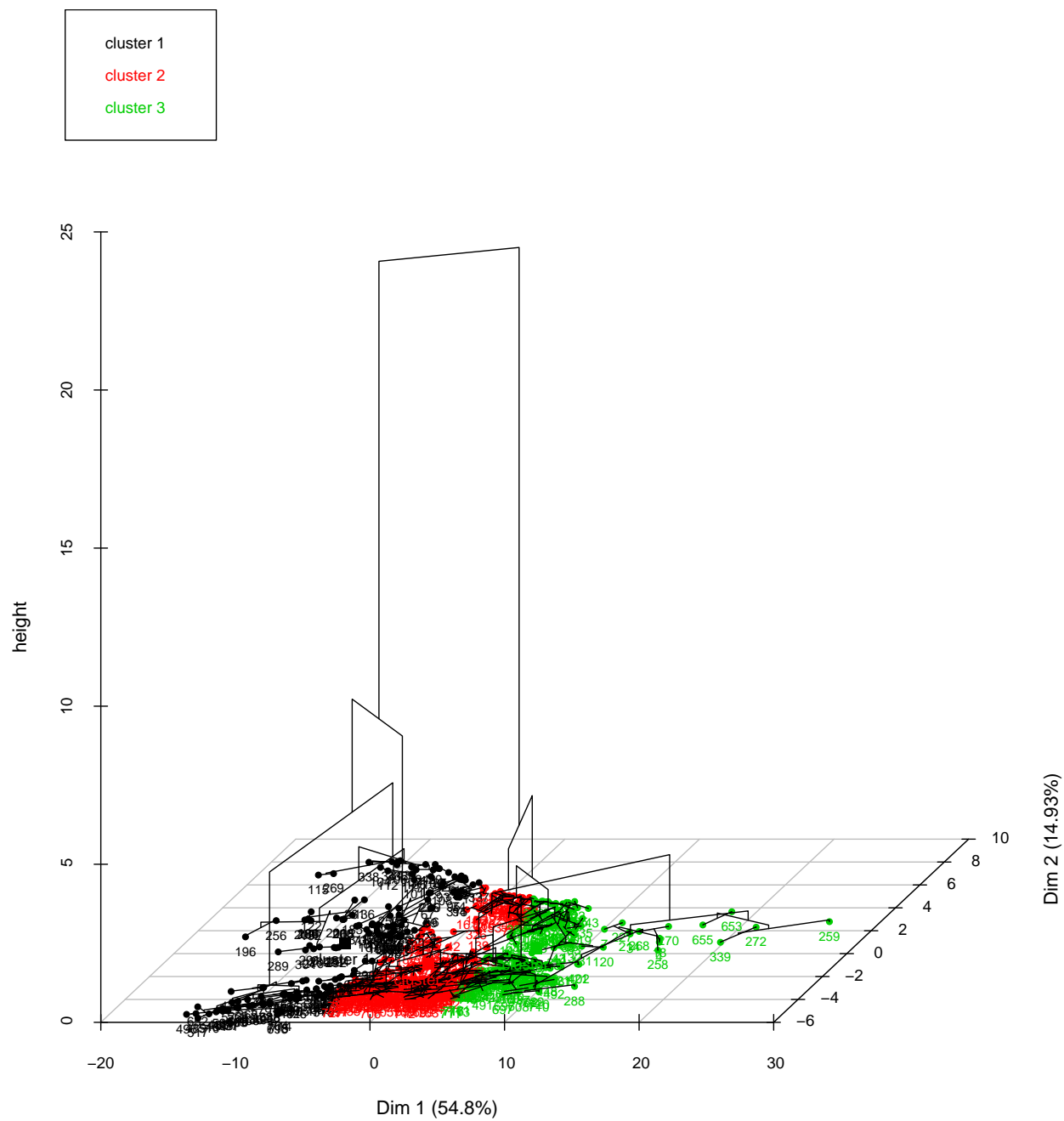


FIGURE 4.8 – HCPC arbre 3D

La figure 4.9 nous montre les sauts d'inertie du dendrogramme. On peut y voir qu'entre 2 et 3 clusters nous avons un saut assez grand puis à nouveau entre 4 et 5 avant une stabilisation.

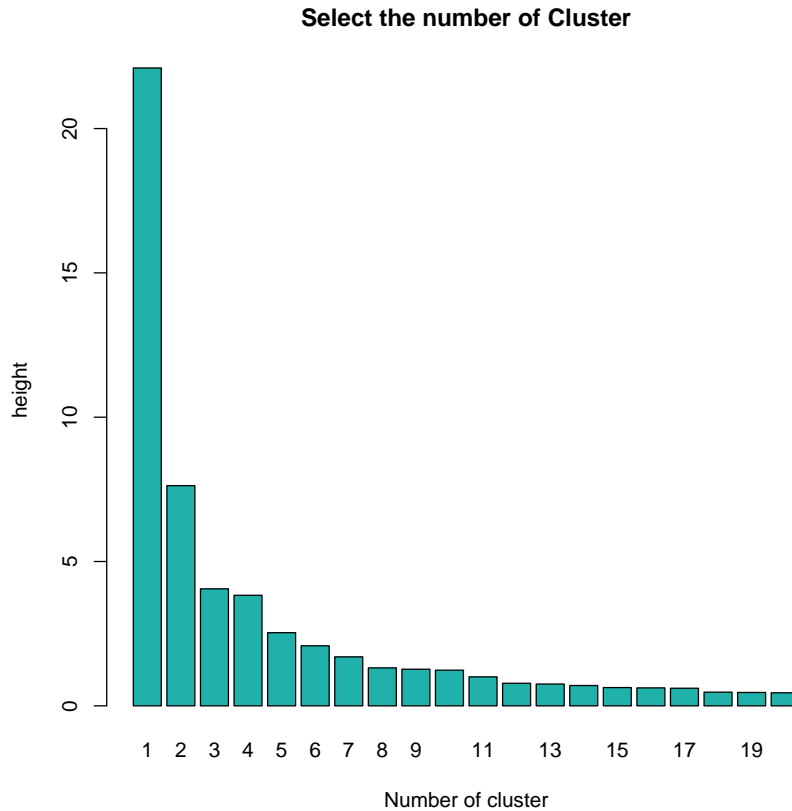


FIGURE 4.9 – Saut d'inertie du dendrogramme

L'ajout de clusters au set de données serait utile du moment que les clusters permettent significativement de séparer les types de cafés, au sens variables dépendantes du terme. Visuellement on devrait pouvoir observer une différence entre le nombre de café dans chaque cluster et les variables de sorties. Cependant ce n'est pas le cas comme on peut le voir ci-dessous :

	1	2	3
cat 2	0	28	21
cat 3	72	98	146
cat 4	44	70	157

TABLE 4.3 – HCPC avec trois clusters comparés à la sortie catégories

	1	2	3
5	0	3	0
5.5	0	1	0
6	3	6	18
6.25	1	1	3
6.5	5	8	35
6.75	1	0	7
7	39	50	102
7.25	13	13	24
7.5	53	40	76
7.75	1	6	8
8	0	59	47
8.25	0	8	4
8.5	0	1	0

TABLE 4.4 – HCPC avec trois clusters comparés à la sortie Acidez

	1	2	3		1	2	3		1	2	3
42	0	0	1	74.875	0	0	1	81.375	1	0	1
49	0	1	2	75	1	2	2	81.5	3	3	8
50	0	3	0	75.25	1	0	4	81.625	1	0	1
52.5	0	0	1	75.375	0	0	1	81.75	1	1	9
58	0	0	1	75.5	1	3	2	82	7	4	13
58.5	0	1	0	75.75	0	1	2	82.25	7	2	4
59	0	2	0	76	2	2	8	82.375	2	0	2
59.5	0	1	0	76.125	0	0	2	82.5	29	24	19
60	0	3	2	76.25	0	0	2	82.75	5	3	4
60.5	0	2	0	76.375	1	0	0	83	0	10	14
61	1	0	1	76.5	4	3	6	83.25	0	3	5
61.75	0	1	0	76.75	1	0	4	83.5	1	7	6
62.5	0	0	1	77	4	3	8	83.75	0	3	3
63	0	1	0	77.25	0	1	2	84	0	5	8
64	0	0	1	77.5	1	0	3	84.25	0	5	2
65.5	0	0	1	77.75	1	2	5	84.5	0	8	5
66	0	1	1	77.875	1	0	0	84.75	0	6	1
67	0	1	4	78	7	4	3	85	0	7	6
68	1	0	2	78.25	0	1	1	85.25	0	4	0
68.75	0	1	0	78.5	1	1	7	85.5	0	1	2
69	0	1	1	78.75	0	0	2	85.75	0	3	4
69.5	0	0	1	79	13	10	41	86	0	4	1
69.75	0	0	1	79.25	1	3	5	86.25	0	0	1
70	0	0	1	79.45	0	0	1	86.5	0	3	2
70.75	1	0	0	79.5	0	2	4	86.75	0	1	3
71	1	2	3	79.52	0	0	1	87	0	1	2
71.375	0	0	1	79.625	0	0	2	87.25	0	2	0
71.75	0	1	0	79.75	0	1	1	87.5	0	1	0
72	0	1	2	80	6	7	8	87.75	0	1	0
72.5	0	1	0	80.25	0	1	4				
72.75	0	0	1	80.375	0	0	1				
73	0	5	3	80.5	2	1	8				
73.25	0	1	0	80.625	1	0	0				
73.5	0	0	2	80.75	1	2	8				
74	0	0	2	81	4	0	6				
74.75	0	1	1	81.25	1	3	6				

TABLE 4.5 – Tableau des clusters pour la sortie Puntaje Total

	Eta2	P-value
TmaxTotalAvg	0.79323282	8.808042e-216
TmaxTotal	0.79323282	8.808042e-216
tmean6	0.79308656	1.101237e-215
tmax6	0.77941800	6.561977e-207
TmeanTotalAvg	0.77802701	4.779121e-206
TmeanTotal	0.77802701	4.779121e-206
tmax7	0.77433881	8.706622e-204
tmean10	0.77341542	3.162373e-203
tmean5	0.77255468	1.047399e-202
tmean7	0.77146031	4.770308e-202
tmean9	0.76659843	3.682110e-199
tmax5	0.75947622	4.888995e-195
tmean8	0.75943990	5.127819e-195
tmax8	0.75506157	1.527633e-192
tmax9	0.73797359	2.717166e-183
tmin10	0.73685898	1.038321e-182
tmean4	0.72591249	4.041247e-177
tmax1	0.72445394	2.159972e-176
tmax4	0.72115391	9.274507e-175
tmax10	0.71764667	4.803959e-173
tmean1	0.71497543	9.398029e-172
tmin6	0.70461972	7.372790e-167
TminTotalAvg	0.68401310	1.306441e-157
TminTotal	0.68401310	1.306441e-157
tmean2	0.67984980	8.149567e-156
tmin8	0.67467087	1.293430e-153
tmax2	0.66718704	1.700396e-150
tmin9	0.66714064	1.776914e-150
tmean3	0.66366823	4.707493e-149
tmin7	0.65549751	9.209679e-146
tmax3	0.64170621	2.220623e-140
tmin5	0.63104980	2.317687e-136

	Eta2	P-value
tmin3	0.61739273	2.231189e-131
tmin4	0.61250703	1.225099e-129
tmin1	0.61199847	1.853488e-129
tmin2	0.60842700	3.343356e-128
DtrTotalAvg	0.55204696	9.200736e-110
DtrTotal	0.55204696	9.200736e-110
prec10	0.52838295	1.045148e-102
PrecTotalAvg	0.51318493	2.320750e-98
PrecTotal	0.51318493	2.320750e-98
ASNM	0.50548203	3.287540e-96
prec6	0.48690365	3.716337e-91
dtr8	0.47172566	3.664625e-87
prec4	0.46275569	7.423006e-85
dtr1	0.45506364	6.575620e-83
prec5	0.43486454	6.354576e-78
dtr7	0.43333703	1.488559e-77
dtr10	0.42305017	4.330758e-75
dtr2	0.41092218	3.056574e-72
dtr4	0.39912083	1.589035e-69
dtr5	0.39851482	2.183452e-69
dtr6	0.38517429	2.199730e-66
dtr9	0.38032942	2.611213e-65
dtr3	0.32744102	4.212873e-54
prec7	0.27997934	8.894746e-45
prec9	0.27175535	3.172161e-43
prec3	0.26360964	1.050396e-41
prec1	0.12139964	1.224259e-17
prec8	0.11382522	1.787707e-16
prec2	0.10476930	4.266965e-15
Luminosidad	0.03438122	6.133050e-05
Arcilloso	0.02674672	6.590146e-04
OrientationNum	0.02251785	2.402284e-03
Limoso	0.01764046	1.041358e-02
pH_avg	0.01665250	1.395976e-02
Cascajoso	0.01553506	1.940861e-02
Franco	0.01291747	4.161129e-02

TABLE 4.6 – Importance des variables lors de la réalisation des clusters

Analyse des résultats de clustering

Self-Organizing Map

U-Matrix, répartition des cafés par classes et composants La carte auto-organisatrice a été réalisée en incluant toutes les variables afin de voir où se placent les variables dépendantes par rapport aux variables indépendantes et de permettre d'observer d'éventuels clusters.

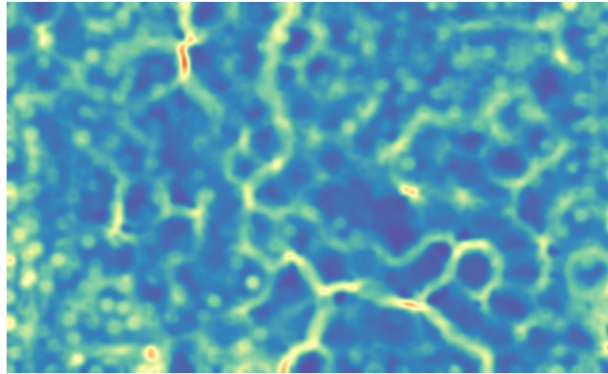


FIGURE 4.10 – U-Matrix de la carte SOM

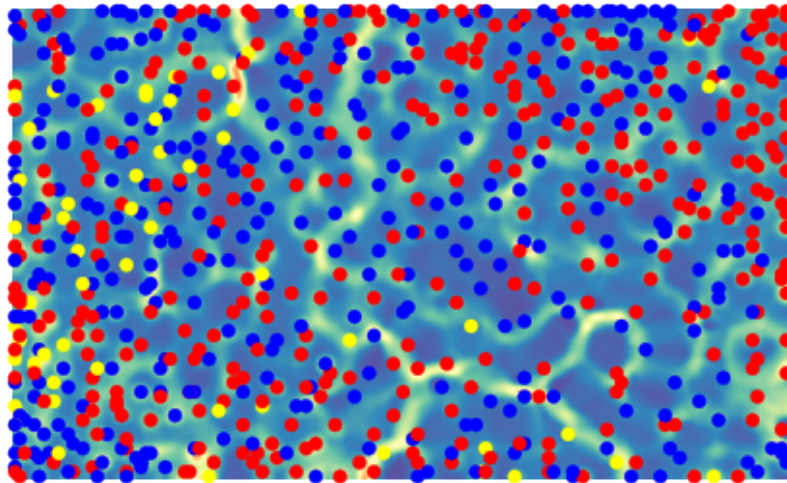


FIGURE 4.11 – U-Matrix avec les points. En jaune les cafés de catégorie 2, en bleu catégorie 3 et en rouge catégorie 4. Les catégories sont expliquées au point 4.1.2

FIGURE 4.12 – Composants de qualité - Variables de sortie

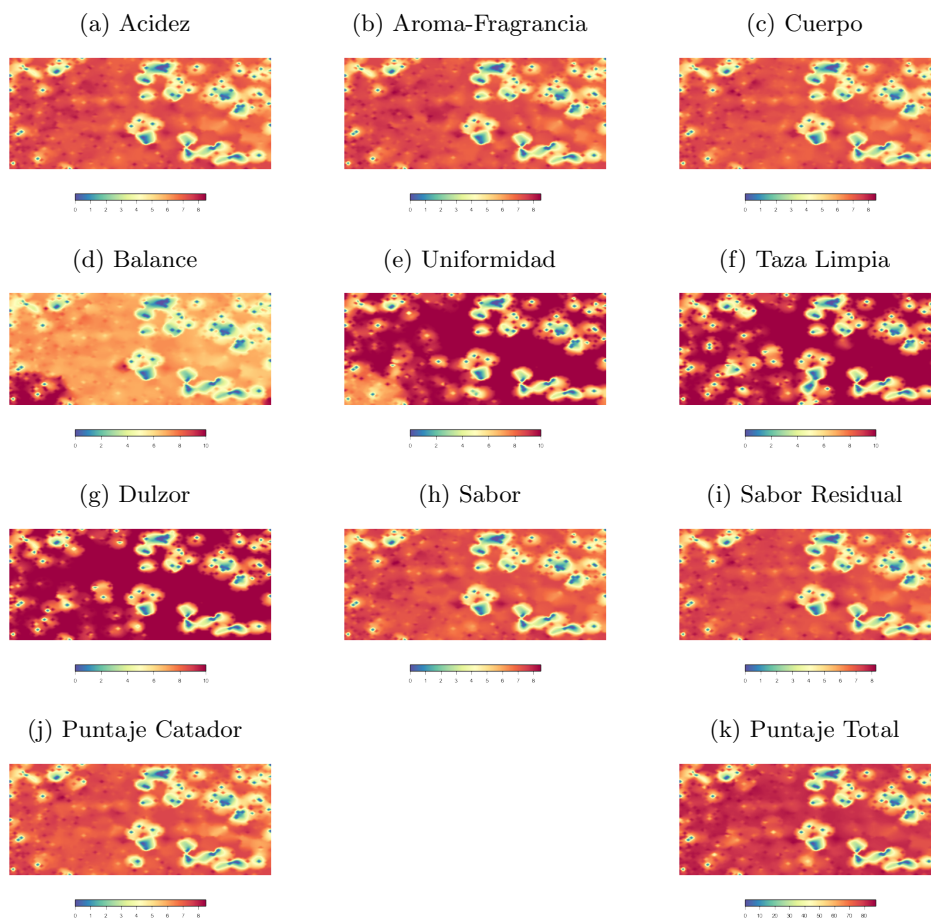


FIGURE 4.13 – Précipitations

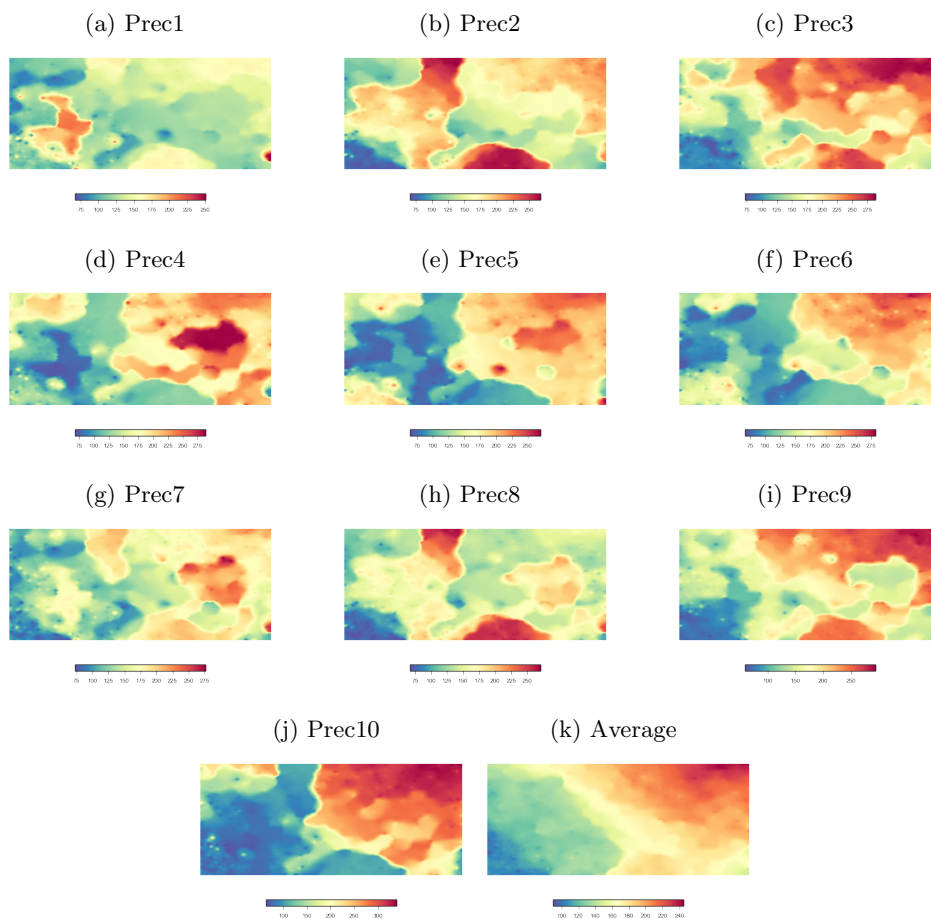


FIGURE 4.14 – SOM - Autres données

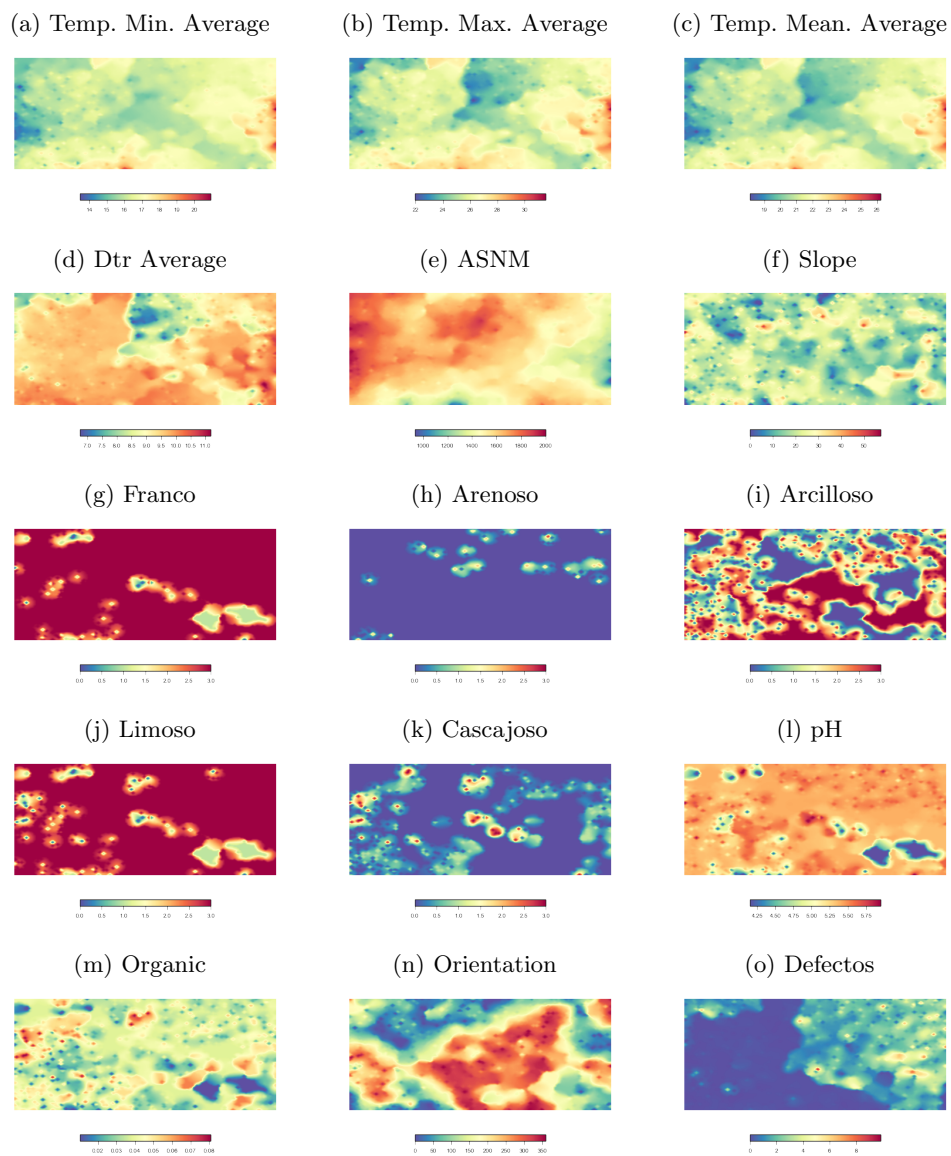
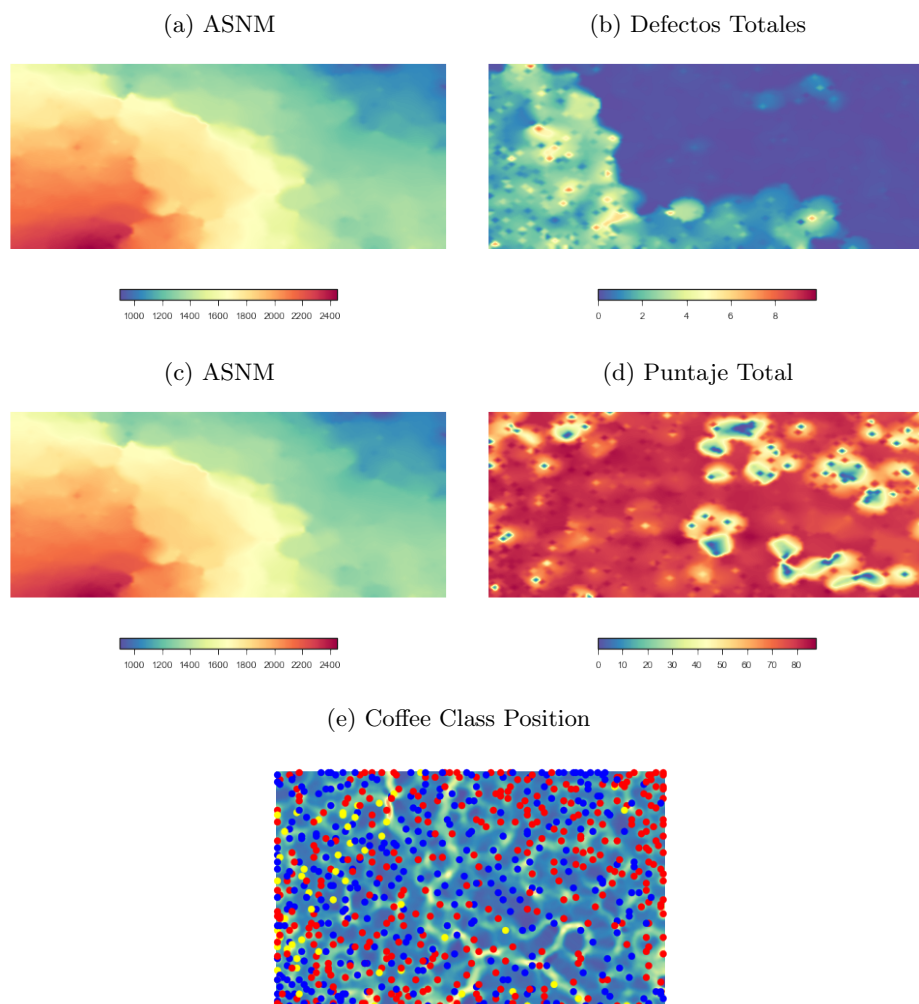


FIGURE 4.15 – Altitude et défauts lors d’une seconde exécution du réseau de neurones



Analyse des cartes On peut observer que les régions avec beaucoup de précipitations sont aussi celles qui ont le plus de cafés avec peu de points. Lors d'une seconde exécution du réseau on a pu observer une relation proche entre les défauts et l'altitude comme montré sur la figure 4.15. Cette relation n'était pas clairement visible lors de la première exécution. Le fait est déjà connu que le café qui pousse en altitude est de meilleure qualité cependant ici on remarque que les cafés cultivés aux altitudes les plus hautes sont les plus sujets à une grande quantité de défauts. Les défauts physiques n'interviennent que peu dans la qualité finale du café mais peuvent avoir une influence en cas de mauvais tri. On observe sur la figure ?? que la zone en bas à gauche, là où l'altitude est la plus haute, correspond à une zone possédant peu de cafés ayant eu le score 0 et beaucoup ayant eu des scores entre 85 et 90. Cependant, des cafés ayant un nombre de points de moins de 80 points, en rouge ou compris entre 89 et 85, en bleu, sont présents plus ou moins de manière homogène sur la totalité de la carte, indépendamment de l'altitude.

4.2 Résultats des analyses

Le but de cette section est d'analyser la possibilité de prédiction de la qualité des cafés à l'aide des données climatiques et de qualité de sols à disposition. La qualité du café fait références aux variables dépendantes citées dans la section 4.1.1.

4.2.1 Perceptron Multi-couches

Pendiente - pending

4.2.2 Random Forest

Méthode utilisée

Résultats obtenus

4.2.3 Partial Least Squares

5 Discussion des résultats

6 Conclusion

Bibliographie

- [1] Centre du commerce international, <http://www.intracen.org>.
- [2] Sistema information cafetera, www.federaciondecafeteros.org.
- [3] Issam El Alaoui. Les méthodes ensemblistes pour algorithmes de machine learning. *Octo blog*, 2014.
- [4] Centre du commerce international. *Guide de l'exportateur de café. - 3^{ème} éd.* Centre du commerce international, Unité des publications, Palais des Nations, Genève, Switzerland. Phone : +41-22 730 01 11, Fax : +41-22 733 44 39, Email : itcreg@intracen.org, URL : <http://www.intracen.org>, 3 edition, 2011.
- [5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction*. Springer, 2 edition, 2009.
- [6] OMMpublic.wmo.int. D'une intensité exceptionnelle, l'épisode el niño a amorcé son déclin, mais ses effets perdurent. 2016.
- [7] Leo Breiman Statistics and Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.
- [8] Johan A. Westerhuis, Theodora Kourti, and John F. MacGregor. Analysis of multiblock and hierarchical pca and pls models. *Journal of Chemometrics*, 12(5) :301–321, 1998.

Table des figures

2.1	Mise sous forme graphique du tableau des température maximales pour le mois de janvier 2011	5
2.2	Intensité du phénomène El Niño au cours des ans Source : http://ggweather.com/enso/oni.htm	6
2.3	Triangle représentant les différentes textures de sols Source : http://www.construnatura.com/esp/articulo/agricultura-ecol-gica/el-suelo-como-fuente-de-vida-propiedades-ii-	8
3.1	Schéma simple du fonctionnement de Random Forest. Source : https://www.youtube.com/watch?v=ajTc5y3OqSQ .	11
3.2	Méthode PLS. X est représenté par son score t et Y par u. Une première estimation de U est multipliée à travers X pour obtenir une approximation du poid ω_t . Le poid est normalisé pour être de longueur 1 et remultiplié à travers X pour produire t. A partir de t et de Y, le poid q^T est obtenu ce qui donne un nouveau vecteur u. Cette opération est répétée jusqu'à la convergence de t. [8]	12
3.3	Méthode MBPLS. Un score de départ u est régressé sur tous les blocs X_b pour donner les poids variables du bloc w_b^T . Les poids des variables de blocs sont normalisés à la longueur un et multipliés par les blocs pour donner les scores de blocs t_b . Les scores de blocs sont combinés dans le super bloc T. Un cycle PLS entre T et Y est effectué pour donner le poids supérieur W_T^T , qui est également normalisé à la longueur un, et le super score t_T . L'opération est répétée jusqu'à la convergence de t_T . [8]	13
3.4	Pays organisés en SOM d'après des indicateurs de pauvreté. Source : http://www.cis.hut.fi/research/som-research/worldmap.html	14
3.5	Pays correspondants à la carte SOM de la figure 3.4 Source : http://www.cis.hut.fi/research/som-research/worldmap.html	15
3.6	Schéma du fonctionnement du Boosting Source : http://www.cis.hut.fi/research/som-research/worldmap.html	16
4.1	Étape de construction du set de données et pertes d'observations.	19
4.2	Emplacement des fermes avec coloration selon le nombre de points attribués au café.	21
4.3	Matrice de corrélation entre les différentes sorties.	22
4.4	Matrice de corrélation entre toutes les variables.	23

4.5	Description de l'Analyse en Composante Principale. (A) Description d'un objet simple de manière compliquée (trois dimensions pour par exemple une ellipse en papier) (B) Trouver des nouvelles variables (axes de coordonnées) orthogonaux l'un à l'autre qui pointent dans les directions de la plus grande variance (C) Utiliser les nouvelles variables (axes) pour décrire l'objet d'une manière plus simple.	24
4.6	Résultats de la PCA sous forme graphique. Réalisé avec la totalité des variables	25
4.7	HCPC et proposition de nombre de cluster	27
4.8	HCPC arbre 3D	28
4.9	Saut d'inertie du dendrogramme	29
4.10	U-Matrix de la carte SOM	34
4.11	U-Matrix avec les points. En jaune les cafés de catégorie 2, en bleu catégorie 3 et en rouge catégorie 4. Les catégories sont expliquées au point 4.1.2	34
4.12	Composants de qualité - Variables de sortie	35
4.13	Précipitations	36
4.14	SOM - Autres données	37
4.15	Altitude et défauts lors d'une seconde exécution du réseau de neurones	38

Liste des tableaux

4.1	Catégories de cafés d'après le nombre de points Source : http://www.scaa.org/?page=resources&d=cupping-protocols	20
4.2	Tableau des rotations des six premiers composants de la PCA avec mise en évidence des variables les plus importantes par composante.	25
4.3	HCPC avec trois clusters comparés à la sortie catégories . . .	29
4.4	HCPC avec trois clusters comparés à la sortie Acidez	30
4.5	Tableau des clusters pour la sortie Puntaje Total	31
4.6	Importance des variables lors de la réalisation des clusters . .	33