



HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD

www.heig-vd.ch

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

Analyse des relations entre les conditions de culture et la qualité
des cafés du département de Risaralda en Colombie

Collaboration avec le Centre International pour l'Agriculture Tropicale



International Center for Tropical Agriculture
Since 1967 *Science to cultivate change*

Thibault SCHOWING
Travail de Bachelor
18 juillet 2017

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

Remerciements

Je tiens à adresser mes remerciements à tous ceux qui m'ont accompagné dans la réalisation de ce projet, sans lesquels rien n'aurait été possible.

Un grand merci au professeur Carlos Andrés Peña, qui m'a donné l'opportunité de sortir des sentiers battus et de découvrir un environnement de travail exceptionnel à Cali en Colombie. C'est une expérience que je ne suis pas près d'oublier et qui va sans doute longtemps me suivre dans mon parcours professionnel et personnel.

Merci à Sylvain, Hugo et Daniel qui m'ont guidé à travers ce projet grâce à leur grande expérience dans le domaine. Grâce à vous j'ai appris énormément et ce bagage me sera d'une grande utilité dans le futur.

Merci à la famille et aux amis pour leur temps de relecture et leur patience.

Et enfin, un grand merci à tous les collègues et amis qui m'ont immédiatement accepté et avec qui j'ai vécu une grande aventure de deux mois, du Pacifique aux montagnes de l'axe du café.

Cahier des Charges

Objectifs Dans un premier temps, l'objectif est de catégoriser les différents 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.

Tâches

- Analyse du problème et planification des étapes du projet
- Analyse du contexte technique et scientifique ainsi que de l'état de l'art
- Décomposition du problème et conception de la solution
 1. Prise en main et analyse des données disponibles
 2. Analyse et sélection des méthodes de modélisation
 - (a) Méthodes pour la caractérisation
 - (b) Méthodes pour la prédiction
- Réalisation, implémentation et tests
 1. Implémentation des méthodes pour la caractérisation (par ex. Clustering, PCA, SOM, ...)
 2. Implémentation des méthodes pour la prédiction (par ex. Réseaux de neurones, Random Forest, Logique floue,...)
- Analyse des informations obtenues et discussion des résultats
- Document et présentation

Résumé

Ce travail de bachelor présente une analyse des relations entre les données environnementales et les résultats de dégustations des cafés du département de Risaralda, en Colombie. Le but principal est de permettre au comité des caféculteurs du département de caractériser leur cafés en relation avec les différentes conditions de culture locale, à des fins principalement de marketing. Pour se faire, des données de dégustations ont été rassemblées et associées aux conditions environnementales correspondantes afin de pouvoir y appliquer des méthodes de Machine Learning et d'en extraire des informations

Dans une première partie, l'analyse a été réalisée de manière non-supervisée en utilisant des cartes auto-organisatrices (SOM), l'analyse en composante principale (PCA) ou le clustering, afin de tenter de repérer des groupes se différenciant les uns des autres et de tenter d'expliquer ces différences. En suite, des méthodes d'apprentissage supervisé comme Random Forest (RF) ou Partial Least Squares (PLS) ont été utilisées afin d'explorer les capacités de prédiction à l'aide des données à disposition.

Les différents éléments obtenus ne permettent pas de prédire la qualité du café. Cependant, des zones à risques ont pu être détectées et des améliorations possibles dans la collecte de données ont été mises en avant et sont en cours de réalisation.

Table des matières

1	Introduction	1
1.1	Question de recherche	1
1.2	Contexte du projet	1
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 de pratiques culturelles et données générales	3
2.1.3	Données gustatives	3
2.1.4	Données climatiques	5
2.1.5	Données de sols	7
2.2	Quelques chiffres et informations sur les données	8
3	Méthodes de modélisation	15
3.1	Rappel des objectifs	15
3.2	Apprentissage supervisé	15
3.2.1	Random Forest	15
3.2.2	Partial Least Square (PLS)	16
3.2.3	Multi Block PLS	17
3.3	Apprentissage non-supervisé	19
3.3.1	SOM	19
3.4	Optimisation	21
3.4.1	Cross-Validation	21
4	Analyse des données	23
4.1	Data Mining - Analyse exploratoire	24
4.1.1	Pré-traitement	24
4.1.2	Analyse exploratoire et apprentissage non-supervisé	26
4.1.2.1	Corrélations entre variables	26
4.1.2.2	Principal Component Analysis (PCA)	29
4.1.2.3	Clustering	33
4.1.2.4	Self-Organizing Map	36
4.2	Apprentissage supervisé	44
4.2.1	Random Forest	44
4.2.1.1	Méthode utilisée	44
4.2.1.2	Résultats de classification	45
4.2.1.3	Régression sur la variable dépendante <i>Puntaje Total</i>	48
4.2.1.4	Régression sur la variable dépendante <i>Acidez</i>	51

4.2.1.5	Régression sur la variable dépendante <i>Dulzor</i>	54
4.2.2	Random Forest avec variables à haute variabilité	57
4.2.3	Partial Least Squares	60
4.2.3.1	Analyse des résultats	62
5	Discussion des résultats	65
6	Conclusion	69
A	Description des données	71
A.1	Dataset	71
A.2	Données de dégustation brutes	73
A.3	Données climatiques et géographiques brutes	73
B	Importances des variables par cluster	77
C	Autres informations	81

1 Introduction

1.1 Question de recherche

A partir de données sur le climat, la qualité du sol et les pratiques culturales, 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 caféculteurs 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 caféculteurs 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 culturales.

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.



International Center for Tropical Agriculture
Since 1967 *Science to cultivate change*

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 national, 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 [8]. 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 de pratiques culturelles et données générales

Par pratiques culturelles on entend ici les différentes façons de traiter la plante durant sa vie et le traitement des grains une fois récoltés. Malheureusement, aucune données de ce type n'a été fournies pour la réalisation de cette analyse. Les seules données à disposition sont des données de pratiques culturelles issues de la littérature ou d'expériences personnelles. [1] [2] [3] [6] [11]

2.1.3 Données gustatives

Les données gustatives sont très relatives aux sens et à la perception de chaque goutteur. Cependant, la SCAA, *Speciality Coffee Association of America*, dispose d'un système de notation [7] 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".

Un premier échantillon de trois cafés 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 et la tâche d'uniformisation nous a donc été confiée. 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 quel 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 caféculteur, 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 d'ensemble, voici un petit résumé concernant 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, et c'était une occasion pour l'équipe de s'informer directement sur le terrain.

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 *desmucilaginador* va enlever la matière gluante entourant le grain, appelé *miel* ou en français *mucilage*, en le lavant. Si la ferme n'est pas équipée 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 terrasses à café, ce processus prend environ dix jours, soit dans des machines à air chaud, plus onéreuses 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. [11]

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

2.1.4 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. Lors d'un précédent projet du CIAT, les moyennes de ces mesures ont été calculées pour chaque mois et extrapolées sur une grande partie du territoire à partir de données provenant de différentes stations météorologiques du pays, permettant ainsi d'accéder aux mesures selon l'emplacement désiré à environ 500 mètres près.

En prenant par exemple les données de températures maximales pour le mois de janvier 2011, en affectant aux valeurs une échelle de couleurs, 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 sans prendre en compte le mois précédent la dégustation, qui est estimé comme temps nécessaire à la récolte et au traitement du grain.¹

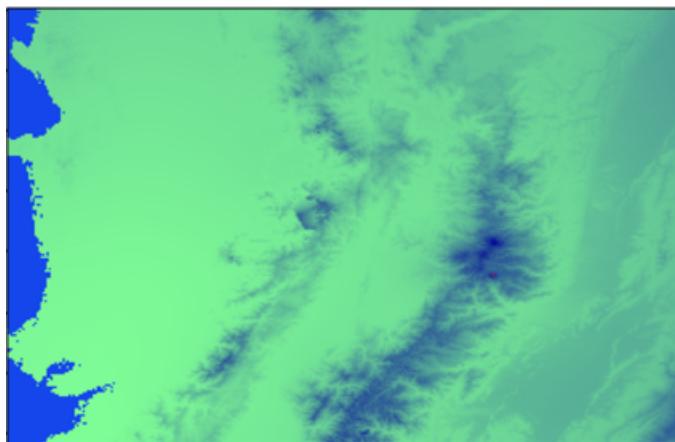


FIGURE 2.1 – Mise sous forme graphique du tableau des températures maximales pour le mois de janvier 2011

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

1. Ce seuil de 10 mois a été approuvé par M. Felipe Rincón, responsable au comité des caféculteurs de Risaralda.

nord ou le désert de Tatacoa au centre et glacial pour les zones en hautes altitudes 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. [3] [4] [10] Les températures oscillent entre 8 et 24 degrés mais le phénomène appelé *El Niño* perturbe régulièrement le climat à l'échelle du continent, voir même du monde.

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 [16] [12]. Si une grande partie de l'Amérique du Sud a été victime de fortes précipitations, la Colombie, elle, a subit 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.

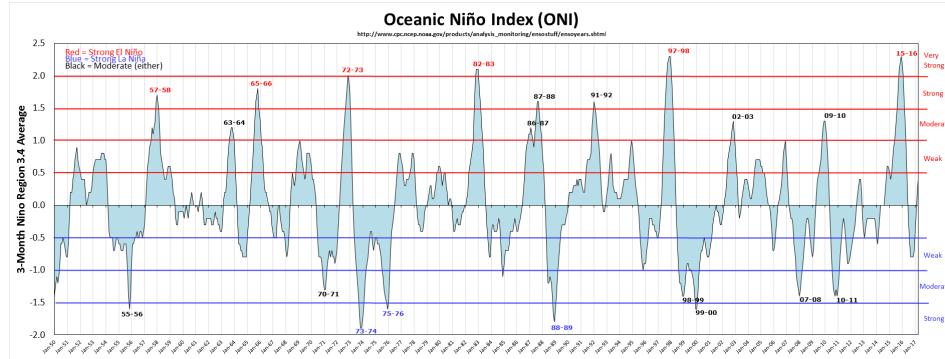


FIGURE 2.2 – Intensité du phénomène El Niño au cours des ans
Source : <http://ggweather.com/enso/oni.htm>

Impact du réchauffement climatique Outre les phases d'El Niño, il est nécessaire de rappeler que le climat mondiale se réchauffe et que des conséquences se font ressentir. Le Centre du Commerce International [2] 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 [11])

2.1.5 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 décrivant la texture des couches a été effectuée.

Les données proviennent d'un GIS², 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. D'autres données très hétérogènes (Système *Pinsat* par exemple) contenait d'autres informations sur la composition chimique du sol mais leur structures et leurs répartitions irrégulières dans la zone de travail ont forcé l'abandon de leur utilisation par manque de temps et de ressources.

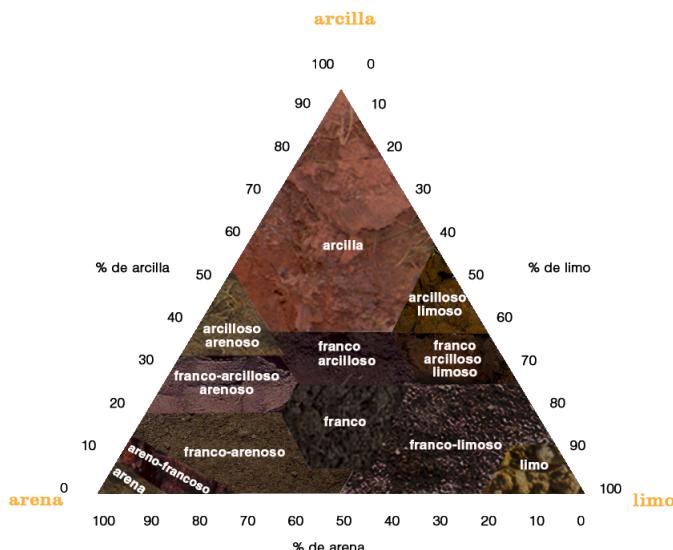


FIGURE 2.3 – Triangle représentant les différentes textures de sols

Source : <http://www.construnatura.com/esp/articulo/agricultura-ecologica/el-suelo-como-fuente-de-vida-propiedades-ii->

2. Geographical Information System

2.2 Quelques chiffres et informations sur les données

Emplacement des fermes L'emplacement des différents cafés par rapport aux nombre de points est présenté sur la figure 2.5. 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'a, à première vue, pas d'incidence sur les résultats, la répartition des différentes classes étant très uniforme. Les différentes catégories attribuées aux cafés sont expliquées plus en détail au point 4.1.

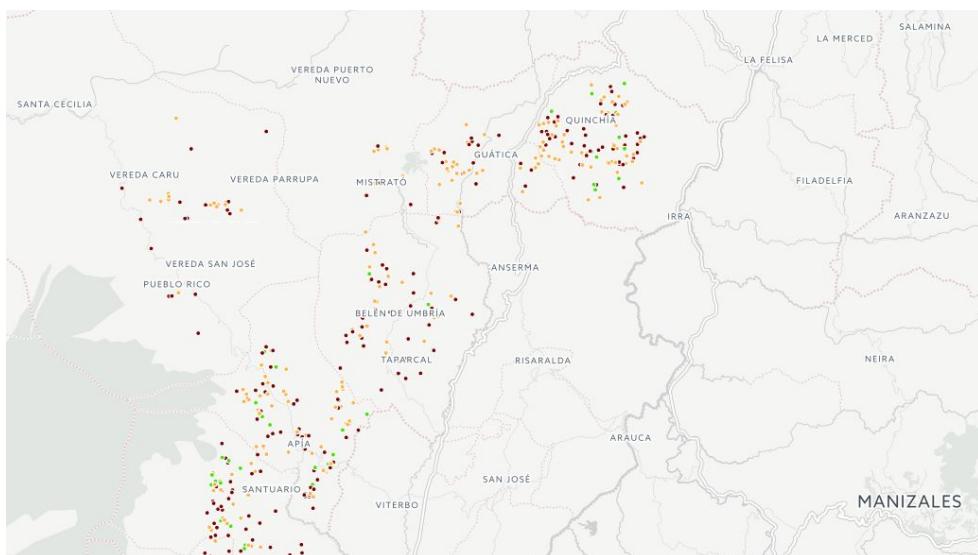


FIGURE 2.4 – Emplacement des fermes avec coloration selon le nombre de points attribués au café. Nord.

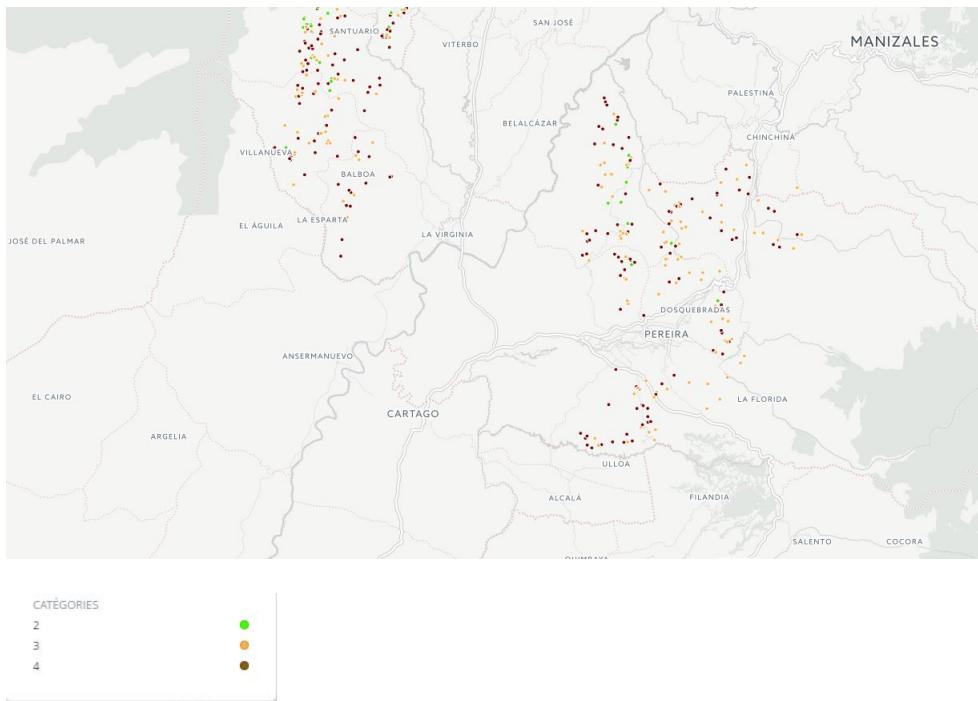


FIGURE 2.5 – Emplacement des fermes avec coloration selon le nombre de points attribués au café. Sud.

Variétés de café Les variétés suivantes sont représentées parmi les cafés présents dans les données :

- CATURRA 335
- COLOMBIA 217
- CASTILLO 185
- TIPICA 9

Les différentes catégories sont réparties comme sur la figure 2.6. On observe aucune particularité dans la répartition des variétés. Dans les tableaux 2.2 et 2.1 aucune différence entre catégorie n'est flagrante. Des analyses plus détaillées sont disponibles dans l'annexe C.

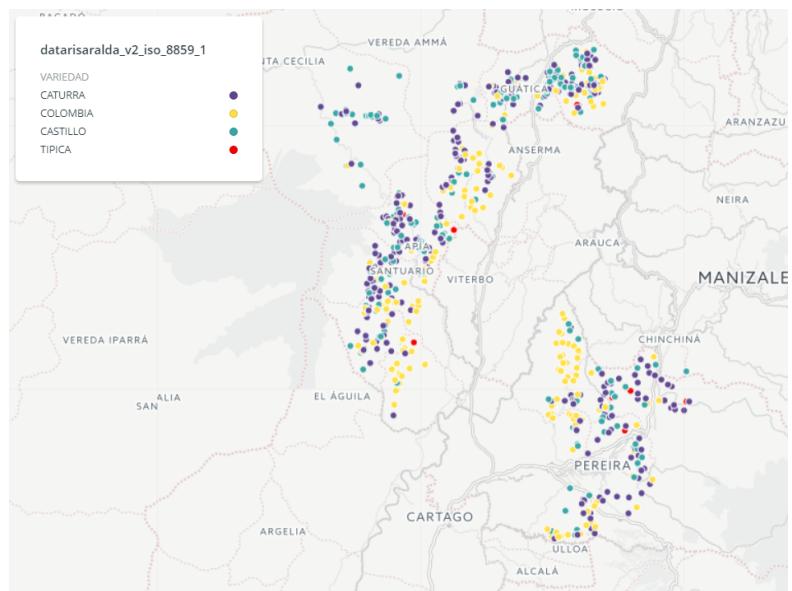


FIGURE 2.6 – Répartition des variétés de café dans le département.

TABLE 2.1 – Nombre de points aux dégustations par catégories

CASTILLO	count	185.000000
	mean	73.370270
	std	21.859713
	min	0.000000
	25%	77.000000
	50%	80.750000
	75%	82.500000
	max	86.750000
CATURRA	count	335.000000
	mean	71.638060
	std	24.449360
	min	0.000000
	25%	76.500000
	50%	80.250000
	75%	82.500000
	max	87.250000
COLOMBIA	count	217.000000
	mean	71.741221
	std	23.564094
	min	0.000000
	25%	75.750000
	50%	79.000000
	75%	82.500000
	max	87.750000
TIPICA	count	9.000000
	mean	81.138889
	std	3.187160
	min	78.000000
	25%	79.000000
	50%	80.000000
	75%	82.500000
	max	87.250000

TABLE 2.2 – Altitude par catégorie

CASTILLO	count	185.000000
	mean	1619.178378
	std	174.143106
	min	1047.000000
	25%	1536.000000
	50%	1637.000000
	75%	1744.000000
	max	2001.000000
CATURRA	count	335.000000
	mean	1641.564179
	std	153.884477
	min	935.000000
	25%	1552.500000
	50%	1644.000000
	75%	1750.000000
	max	1993.000000
COLOMBIA	count	217.000000
	mean	1544.004608
	std	149.512464
	min	1178.000000
	25%	1452.000000
	50%	1533.000000
	75%	1637.000000
	max	1957.000000
TIPICA	count	9.000000
	mean	1580.555556
	std	162.340777
	min	1344.000000
	25%	1500.000000
	50%	1596.000000
	75%	1639.000000
	max	1810.000000

Altitude et points³ Les plantations sont réparties entre une altitude de 935 mètres et de 2001 mètres alors que nombre de points totaux attribués aux cafés varie entre 0 et 87.75.

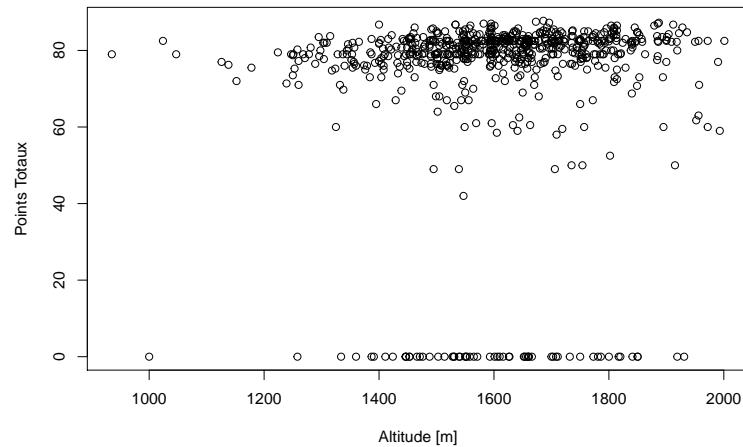


FIGURE 2.7 – Nombre de points totaux vs altitude

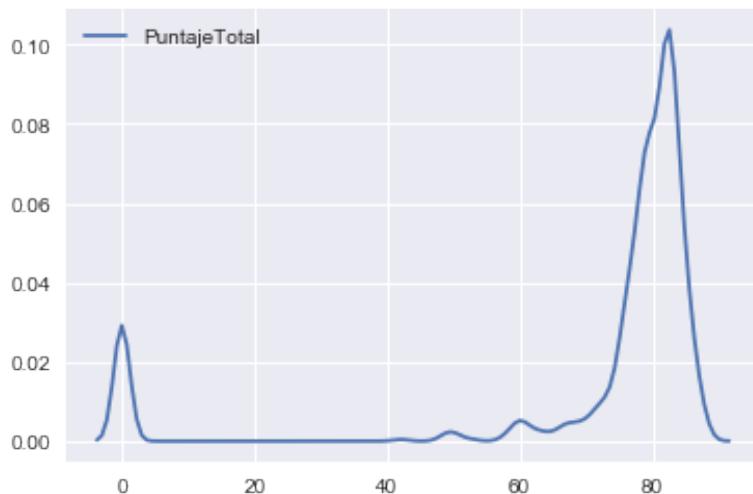


FIGURE 2.8 – Répartition des cafés par rapport au nombre de points totaux

3. Les sets de données utilisés ici peuvent différer, certaines données ont été extraites du set de base, contenant des données manquantes (pour certaines variables) et d'autres ont été extraits du set complet utilisé pour les calculs. Les données ne diffèrent que très peu entre ces sets et l'impact est minimal pour le calcul de simples moyennes à titre exploratoire.

Précipitations et températures Les deux années présentes dans les données, 2011 et 2016 sont deux années sensiblement différentes au niveau climatique, surtout au regard des précipitations.

TABLE 2.3 – Précipitations par année

2011	count	350.000000
	mean	192.985313
	std	28.732351
	min	108.813344
	25%	172.463001
	50%	199.349562
	75%	212.438373
	max	245.214644
2016	count	396.000000
	mean	135.710853
	std	24.468758
	min	89.579520
	25%	119.952117
	50%	131.346726
	75%	140.572487
	max	189.482009

TABLE 2.4 – Températures moyennes par année

2011	count	350.000000
	mean	20.969629
	std	1.195129
	min	18.106276
	25%	20.116997
	50%	20.880830
	75%	21.778835
	max	26.238706
2016	count	396.000000
	mean	21.169157
	std	1.080422
	min	18.105124
	25%	20.549152
	50%	21.305588
	75%	21.918895
	max	24.874828

Points totaux des dégustations Des différences sont visibles par année en ce qui concerne les moyennes des points. En effet, l'année 2011 possède beaucoup plus de cafés notés avec zéro points que l'année 2016 par exemple.

TABLE 2.5 – Points totaux par année

2011	count	330.000000
	mean	65.870833
	std	30.274299
	min	0.000000
	25%	76.000000
	50%	79.000000
	75%	82.000000
	max	85.500000
2016	count	367.000000
	mean	78.203597
	std	11.202238
	min	0.000000
	25%	77.375000
	50%	81.250000
	75%	83.750000
	max	87.750000



FIGURE 2.9 – Répartition des points totaux par année

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 cafés 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 est 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 prédire des nouvelles données. On essayera 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 [13], 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 aggregation*, 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. [14]

Random Forest effectue donc un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents [17].

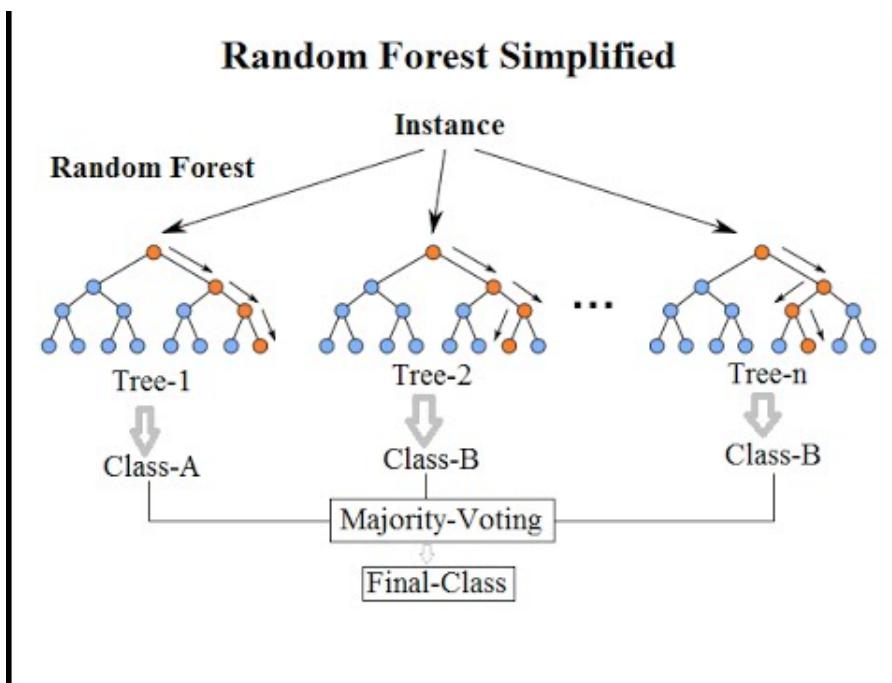


FIGURE 3.1 – Schéma simple du fonctionnement de Random Forest.
Source : <https://www.youtube.com/watch?v=ajTc5y3OqSQ>

3.2.2 Partial Least Square (PLS)

PLS, initialement 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.

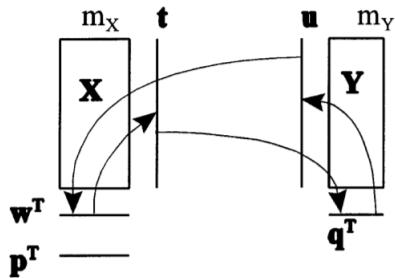


FIGURE 3.2 – Méthode PLS. X est représentée par son score t et Y par u. Une première estimation de U est multipliée à travers X pour obtenir une approximation du poids ω_t . Le poids est normalisé pour être de longueur 1 et re-multiplié à travers X pour produire t. A partir de t et de Y, le poids q^T est obtenu ce qui donne un nouveau vecteur u. Cette opération est répétée jusqu'à la convergence de t. [18]

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.

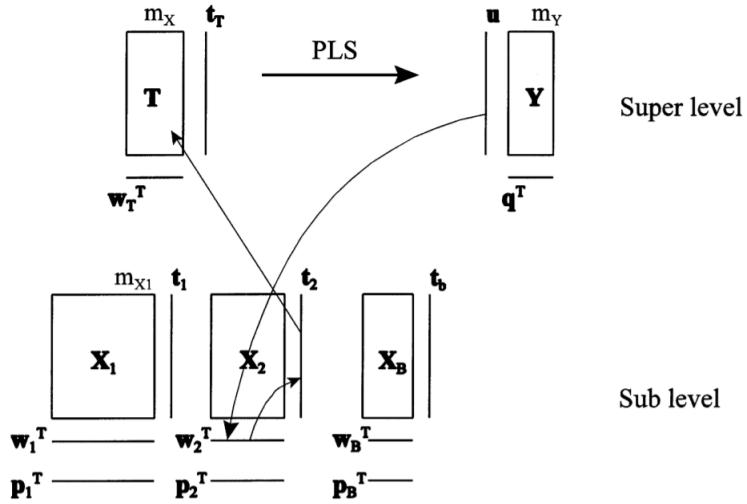


FIGURE 3.3 – Méthode MBPLS. Un score de départ u est régressé sur tous les blocs X_b pour donner les poids variables du bloc w_b^T . Les poids des variables de blocs sont normalisés à la longueur un et multipliés par les blocs pour donner les scores de blocs t_b . Les scores de blocs sont combinés dans le super bloc 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 . [18]

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 cafés identifiables.

3.3.1 SOM

Les cartes SOM, pour *Self Organizing Map* aussi appelées cartes de Kohonen du nom du statisticien ayant développé la méthode, sont des réseaux de neurones disposés en grille permettant de réduire dans un espace en deux dimensions des données ayant n dimensions.

La carte est composée de composants appelés noeuds ou neurones. Ces neurones sont associés à un vecteur de poids de dimension identique au données d'entrée (initialisé aléatoirement) et à une position dans l'espace de la carte. L'algorithme va en suite trouver quelle neurone est le plus proche à la donnée pour chaque observation et va rapprocher les données du neurone dans l'espace de la carte et modifier les poids du neurone afin d'accroître l'influence du neurone sur les prochaines observations proches [15].

On souhaite vérifier s'il est possible de regrouper des cafés qui se distinguaient et à l'aide des différents composants de la carte SOM, comprendre quelles variables ont une influence sur la qualité.

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*.

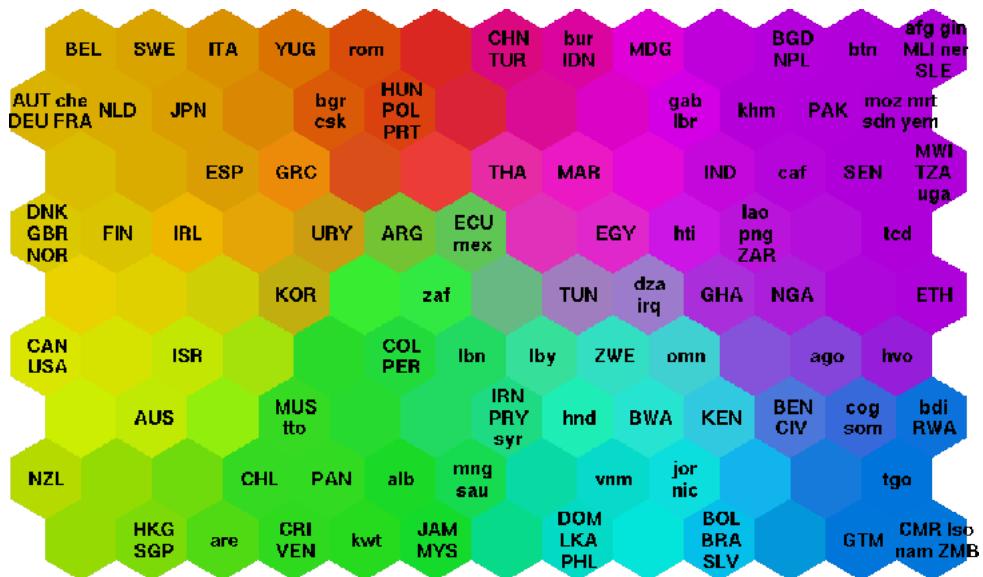


FIGURE 3.4 – Pays organisés en SOM d'après des indicateurs de pauvreté.
Source : <http://www.cis.hut.fi/research/som-research/worldmap.html>

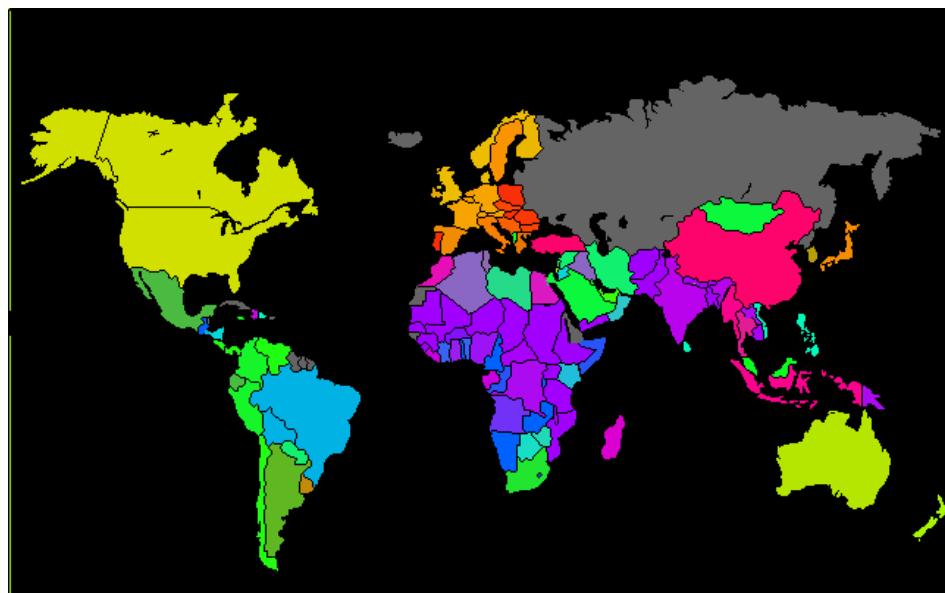


FIGURE 3.5 – Pays correspondants à la carte SOM de la figure 3.4
Source : <http://www.cis.hut.fi/research/som-research/worldmap.html>

3.4 Optimisation

3.4.1 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 [14].

4 Analyse des données

Contents

4.1 Data Mining - Analyse exploratoire	24
4.1.1 Pré-traitement	24
4.1.2 Analyse exploratoire et apprentissage non-supervisé	26
4.2 Apprentissage supervisé	44
4.2.1 Random Forest	44
4.2.2 Random Forest avec variables à haute variabilité	57
4.2.3 Partial Least Squares	60

4.1 Data Mining - Analyse exploratoire

4.1.1 Pré-traitement

Une grande partie de l'étape de pré-traitement a été réalisée durant l'extraction des données. Il a en effet fallu les extraire 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 et 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 selon les sources. Le dataset résultant est décrit dans l'Annexe A de ce projet. Une fois cette étape d'extraction et de nettoyage réalisée, les premières informations ont pu être extraites des données.

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

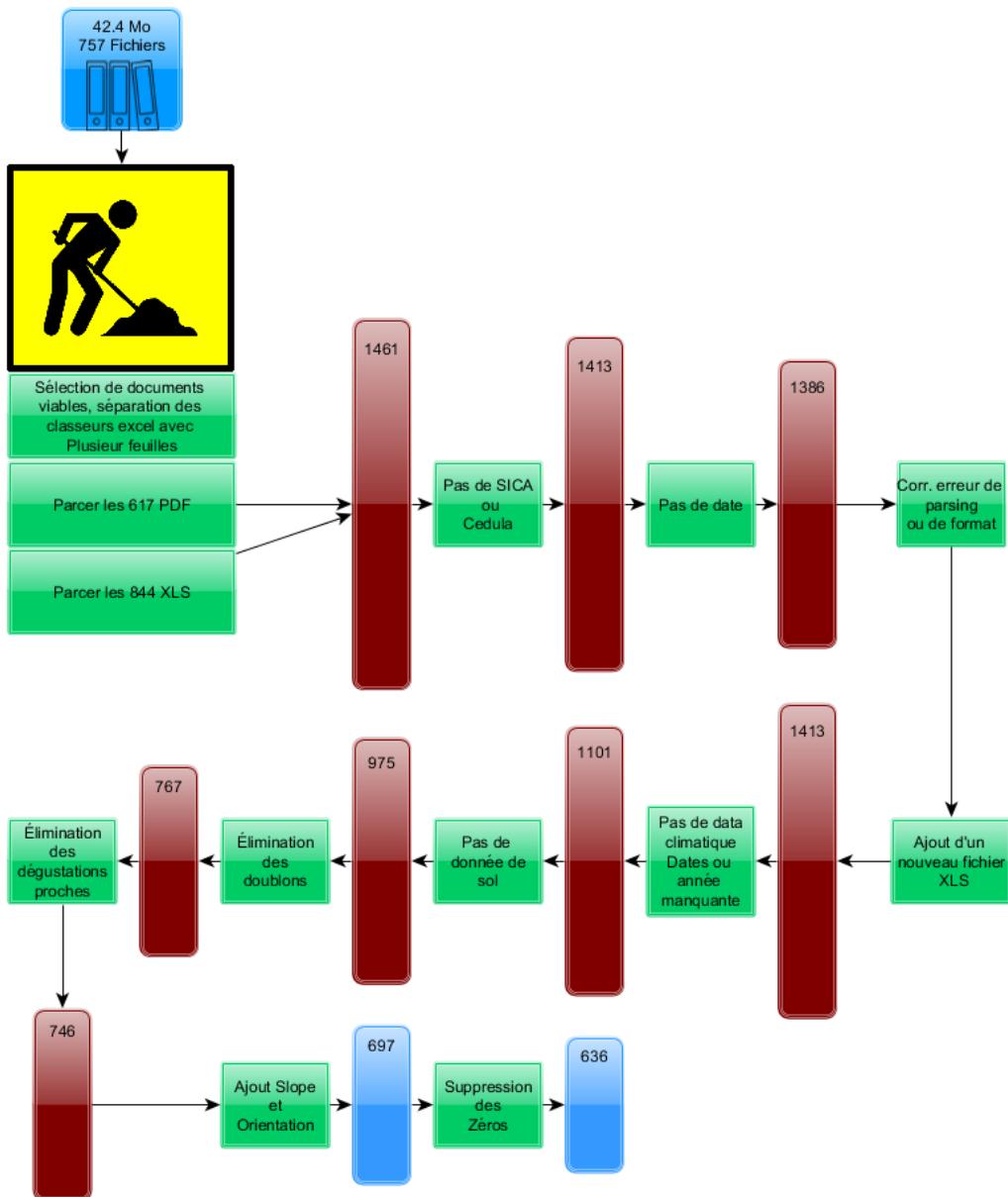


FIGURE 4.1 – Étapes de construction du set de données et pertes d'observations.

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 coordinateur de gestion au comité départemental des caféculteurs de Risaralda, M. Felipe Rincón, 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 attribuables lors des dégustations qu'est-ce que cela représente ? Le tableau 4.1 nous donne une bonne idée de l'échelle de notation. 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.

4.1.2 Analyse exploratoire et apprentissage non-supervisé

4.1.2.1 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 4.2 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.

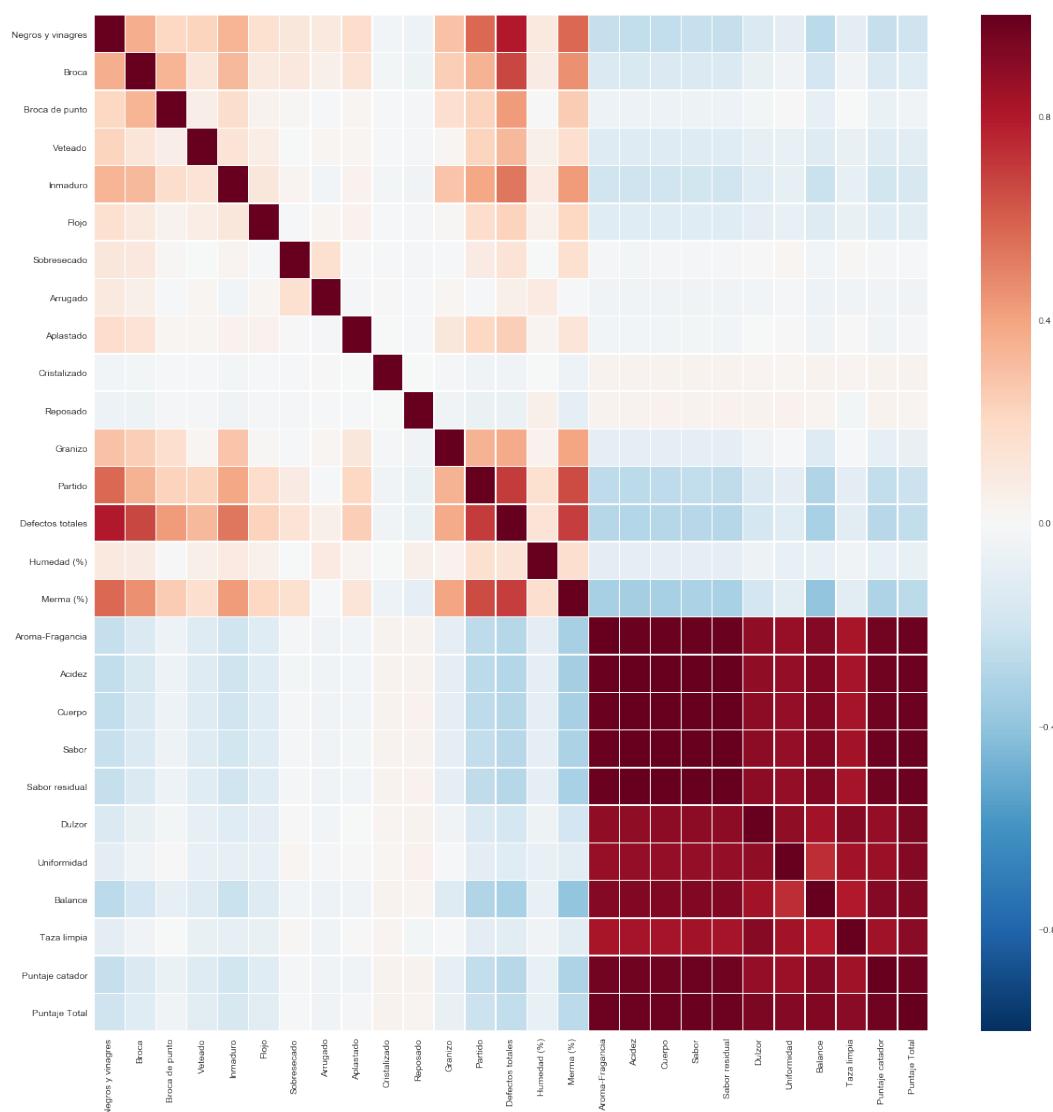


FIGURE 4.2 – Matrice de corrélation entre les différentes sorties.

Sur la figure 4.3 on peut observer les corrélations entre toutes les variables.

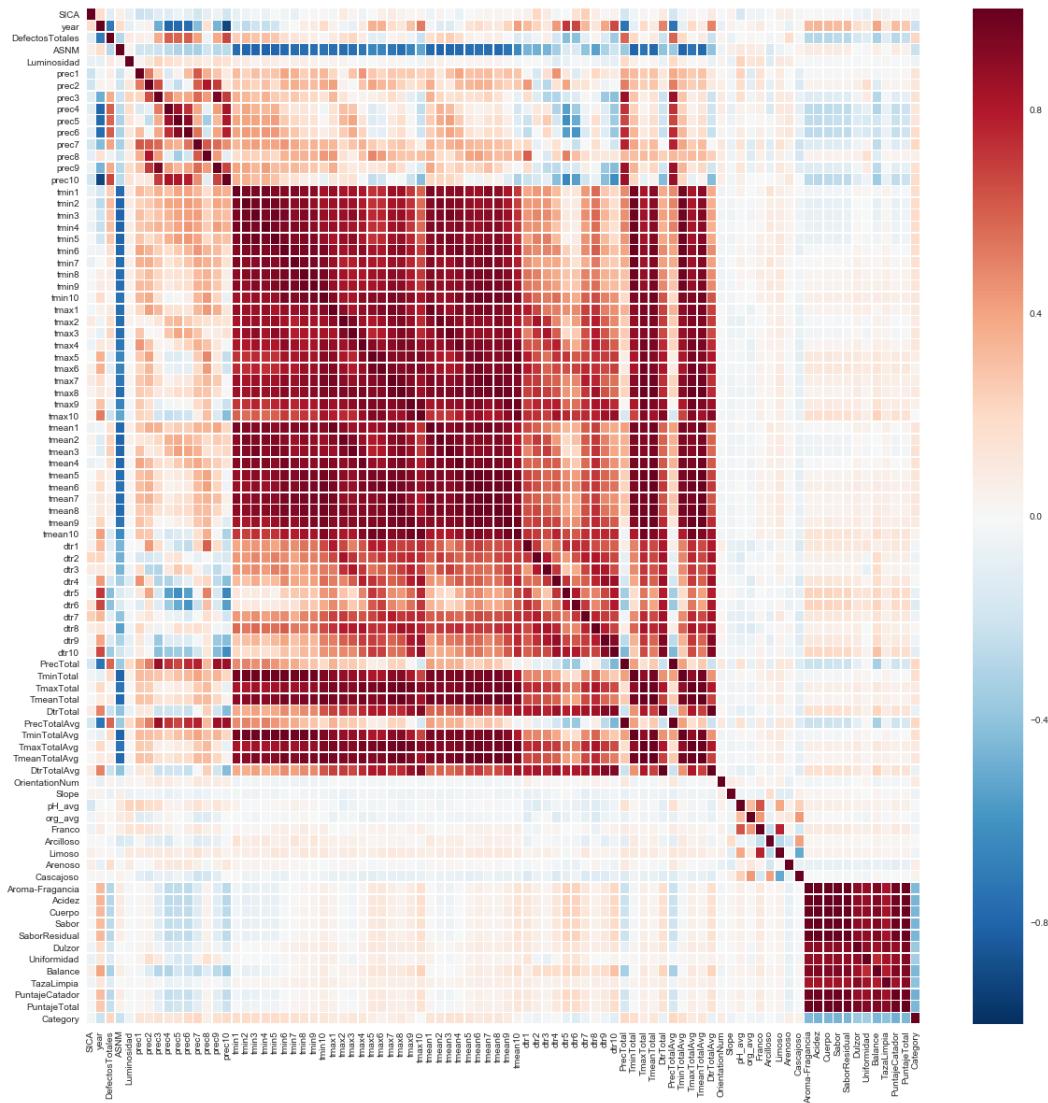


FIGURE 4.3 – Matrice de corrélation entre toutes les variables.

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 visible sur les défauts physiques des grains (3^{ème} position sur la matrice), en revanche nous n'avons aucune variable qui a une corrélation marquée avec les points totaux et donc la qualité du café.

4.1.2.2 Principal Component Analysis (PCA)

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-correlées 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 4.4 montre une représentation graphique de la composante principale.

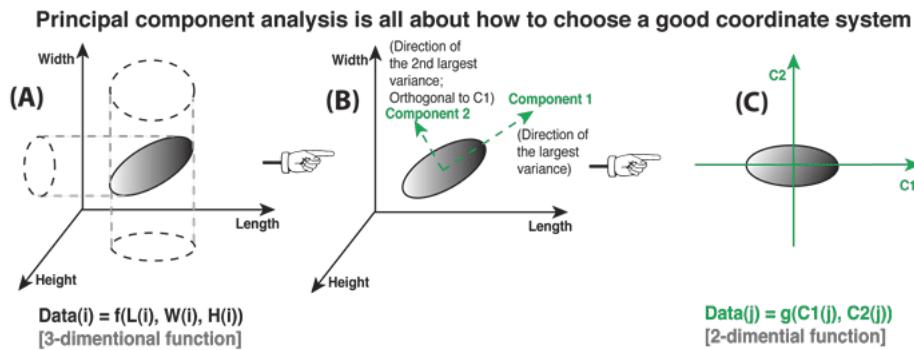


FIGURE 4.4 – Description de l’Analyse en Composante Principale. (A) Description d’un objet simple de manière compliquée (trois dimensions pour par exemple une ellipse en papier) (B) Trouver des nouvelles variables (axes de coordonnées) orthogonaux l’un à l’autre qui pointent dans les directions de la 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 4.5. 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 4.2 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.

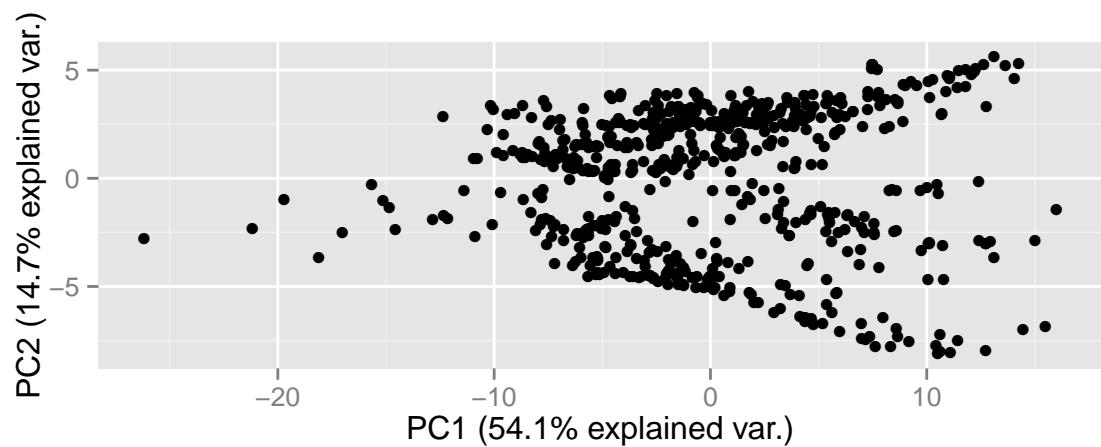


FIGURE 4.5 – Résultats de la PCA sous forme graphique. Réalisé avec la totalité des variables

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 et les sols rocailleux à l'opposé des sols pentus.

PCA avec prcomp Afin de visualiser au mieux les données dans la PCA, une autre PCA a été réalisée à l'aide d'un autre outil (prcomp) et chaque point a été coloré selon son année (figure 4.6) ou selon sa catégorie (figure 4.7).

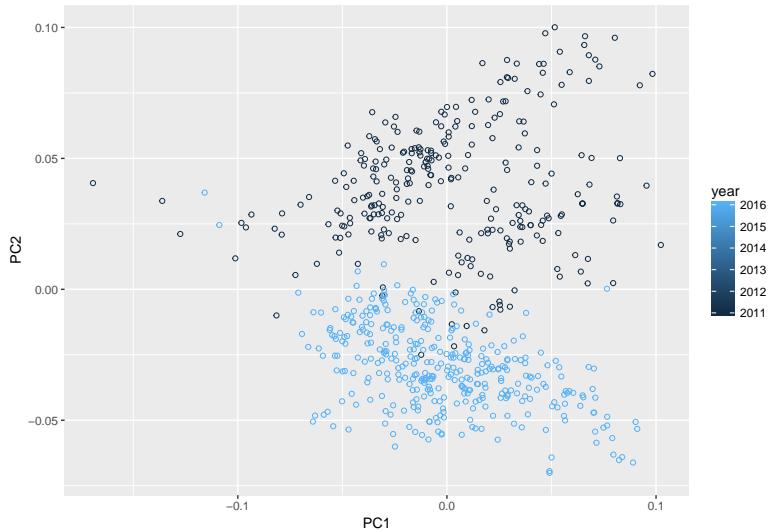


FIGURE 4.6 – PCA avec Prcomp : Coloration par année

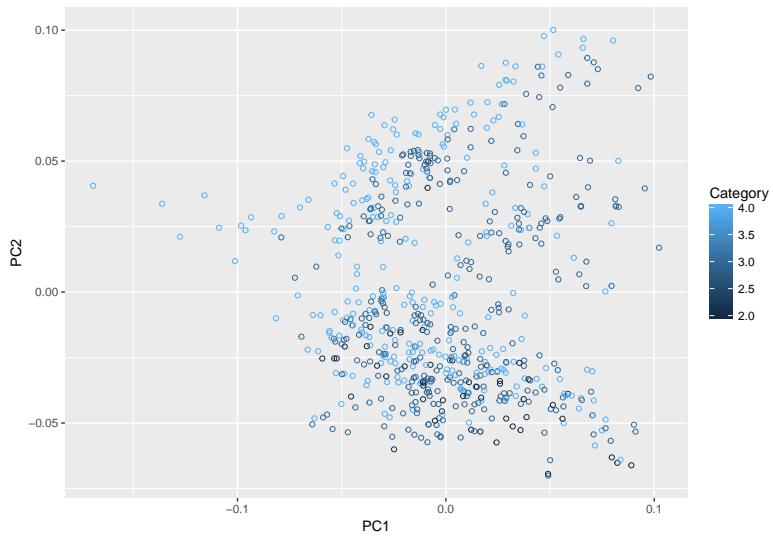


FIGURE 4.7 – PCA avec Prcomp : Coloration par catégorie

On peut observer que les deux années sont bien différenciées. Les premiers composants de la PCA montrent l'importance des variables climatiques et les différences entre années sont visibles sur ces graphiques. Les catégories sont aussi différemment réparties. En retournant sur les moyennes de chaque année (Voir description des données section 2.5), on peut voir que l'année 2011 possède en effet moins de café en nombre mais plus de cafés ayant reçu la note de zéro. On peut aussi voir que les conditions climatiques sont différentes entre les deux années, surtout en termes de précipitations. La section 4.1.2.4 montre les relations entre le climat et les défauts totaux, qui peuvent influencer la qualité finale du café.

4.1.2.3 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 4.8 nous montre l'arbre hiérarchique créé ainsi que le nombre de cluster proposé.

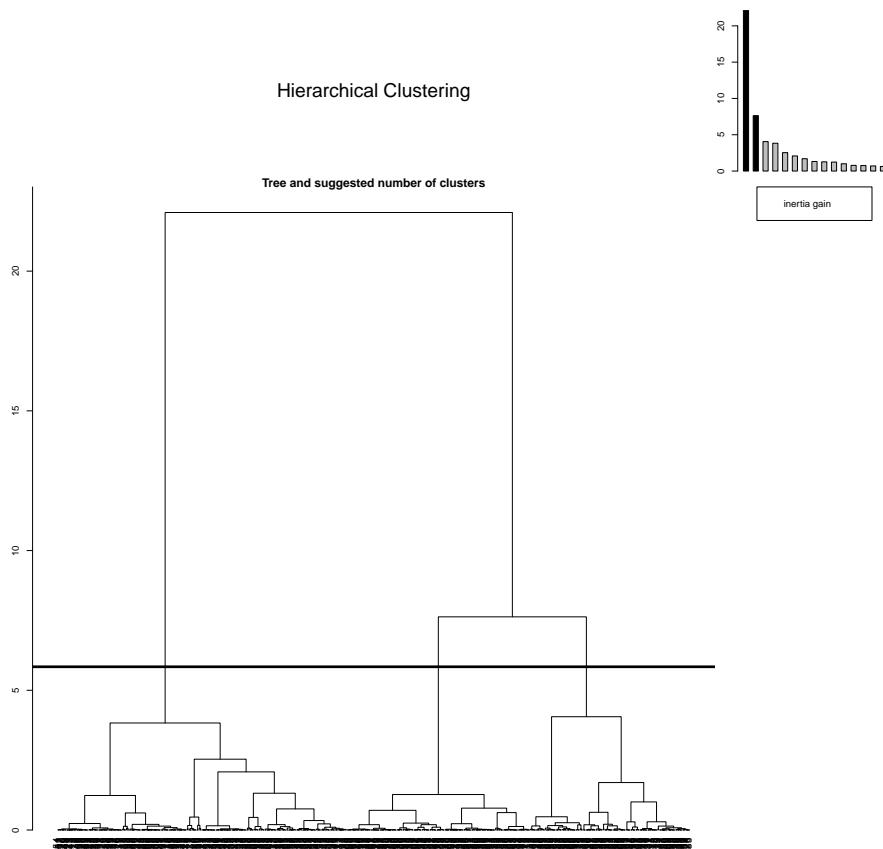


FIGURE 4.8 – HCPC et proposition de nombre de cluster

La figure 4.9 montre une représentation du HCPC en 3D. On reconnaît la forme de la PCA mais les clusters sont plutôt décevants car le découpage ne fait pas ressortir les groupes éventuellement visibles. Comme on l'a vu précédemment, avec la PCA et les différentes moyennes, les deux années sont très différentes et on aurait pu s'attendre à un clustering plus efficace.

Hierarchical clustering on the factor map

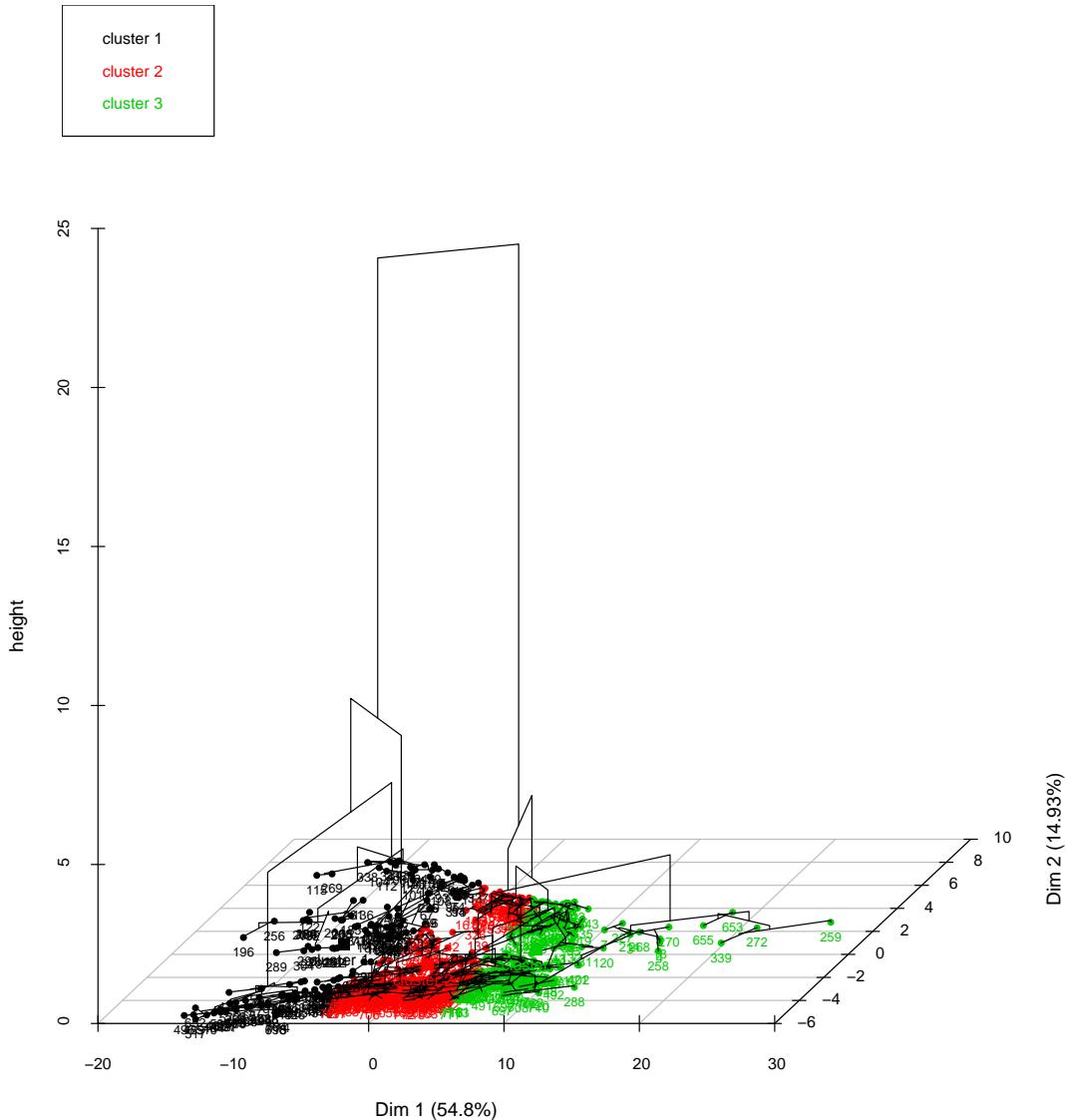


FIGURE 4.9 – HCPC arbre 3D

La figure 4.10 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.

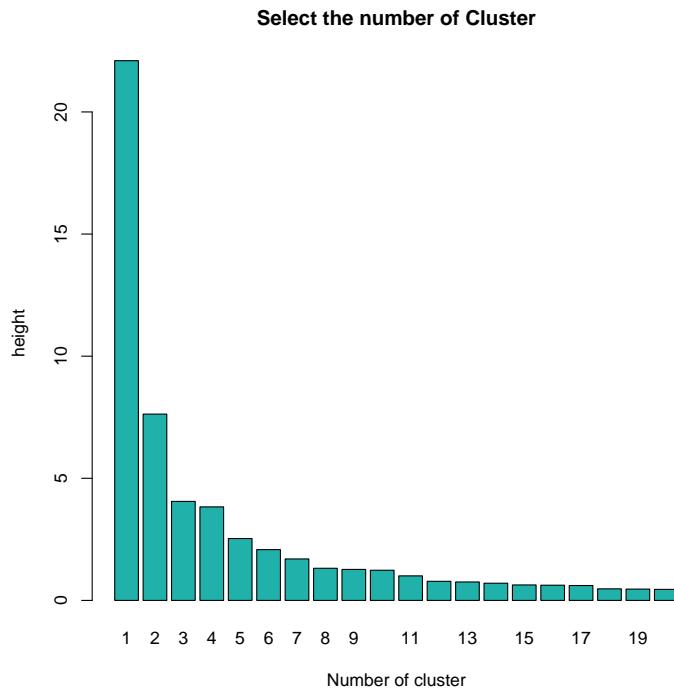


FIGURE 4.10 – Saut d'inertie du dendrogramme

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

Les autres résultats ainsi que l'importance des variables pour la génération des clusters se trouvent dans l'annexe B.

Les variables participent presque toutes activement à la séparation en parties cependant la séparation n'apporte rien au set de données puis-qu'elle ne sépare en rien les différentes catégories de cafés.

4.1.2.4 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.

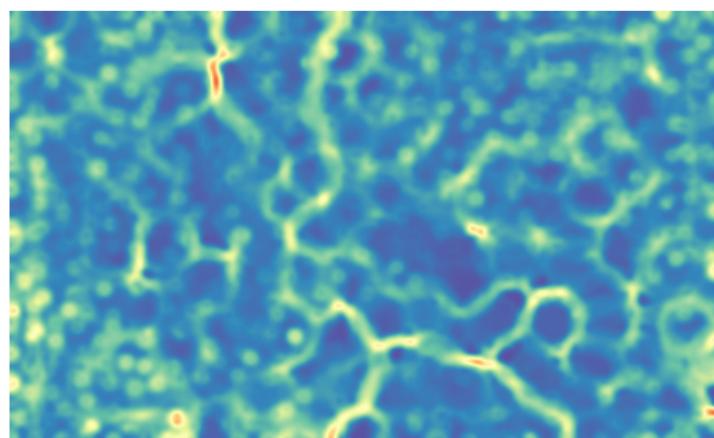


FIGURE 4.11 – U-Matrix de la carte SOM

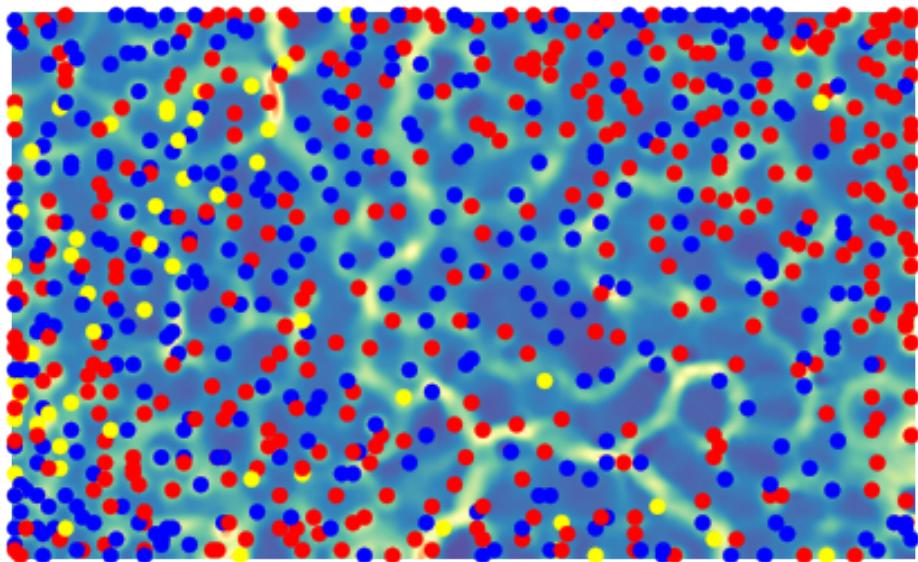


FIGURE 4.12 – U-Matrix avec les points. En jaune les cafés de catégorie 2, en bleu catégorie 3 et en rouge catégorie 4. Pour rappel, les catégories sont expliquées au point 2.2

FIGURE 4.13 – Composants de qualité - Variables de sortie

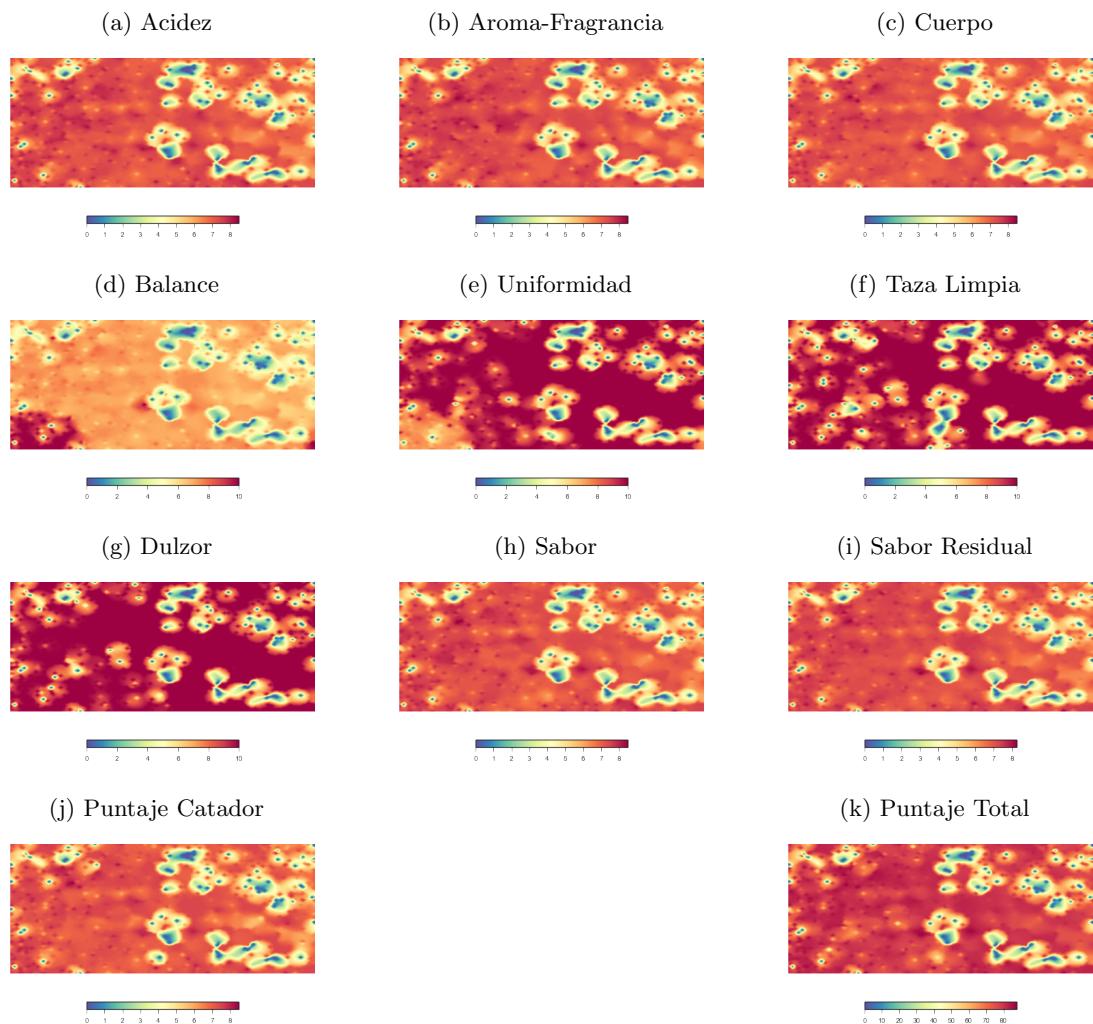


FIGURE 4.14 – Précipitations

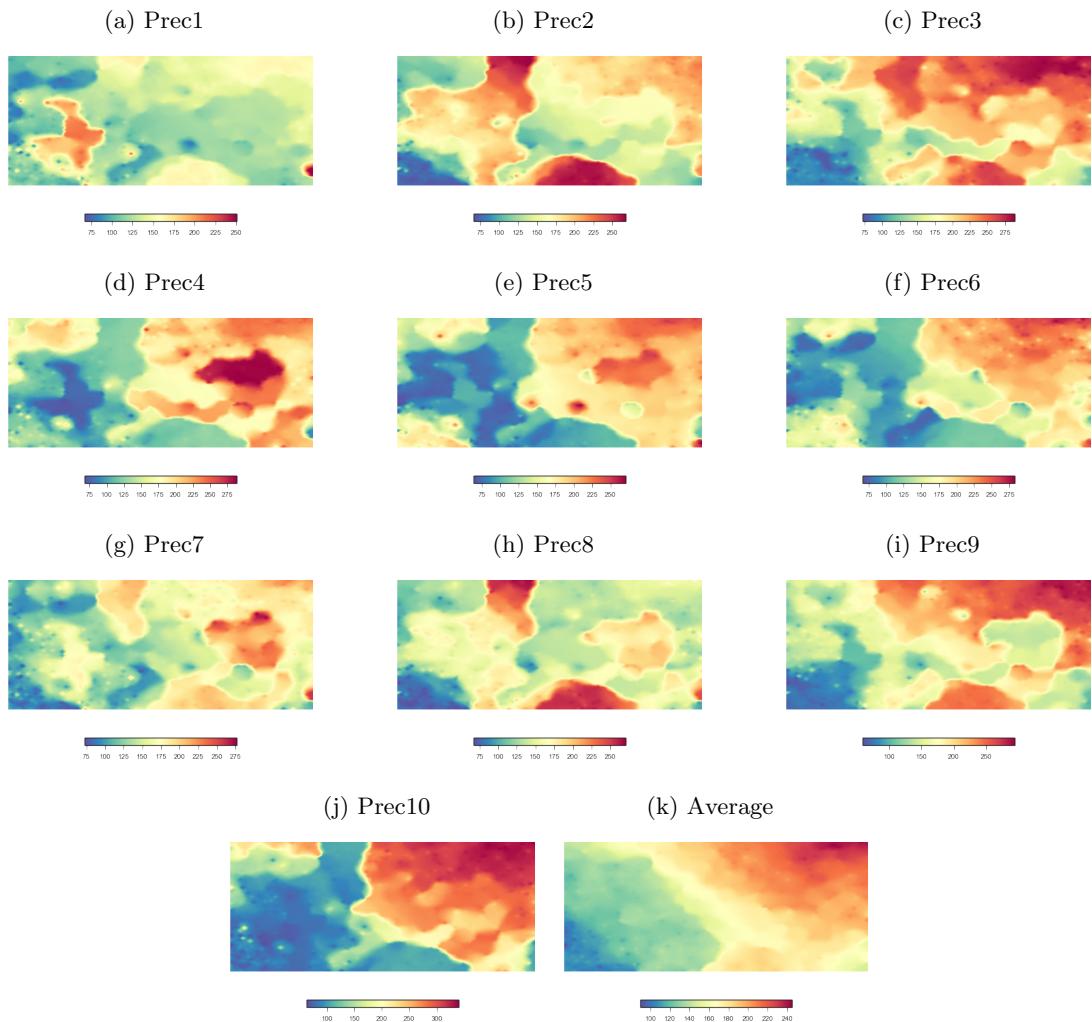


FIGURE 4.15 – SOM - Autres données

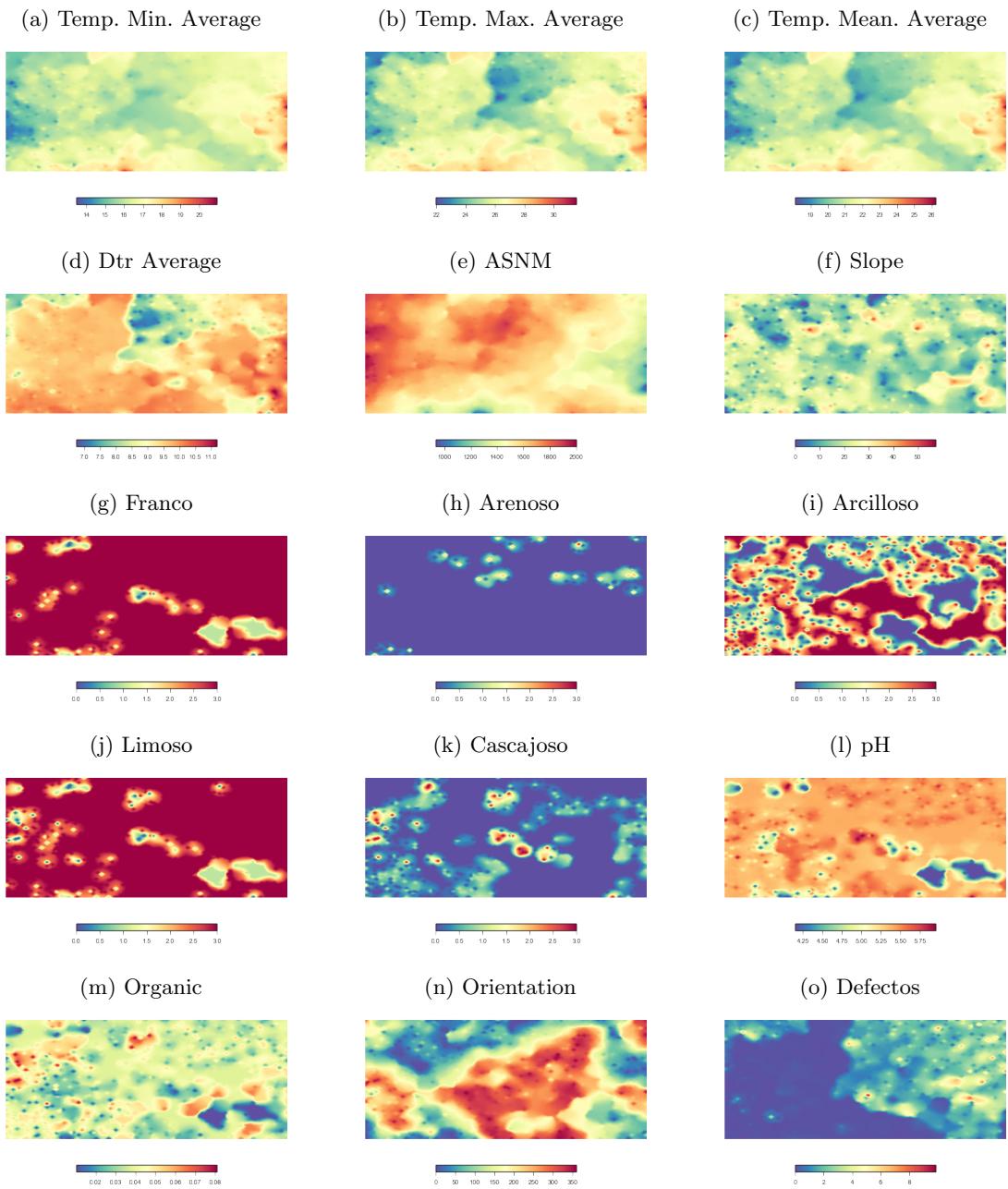
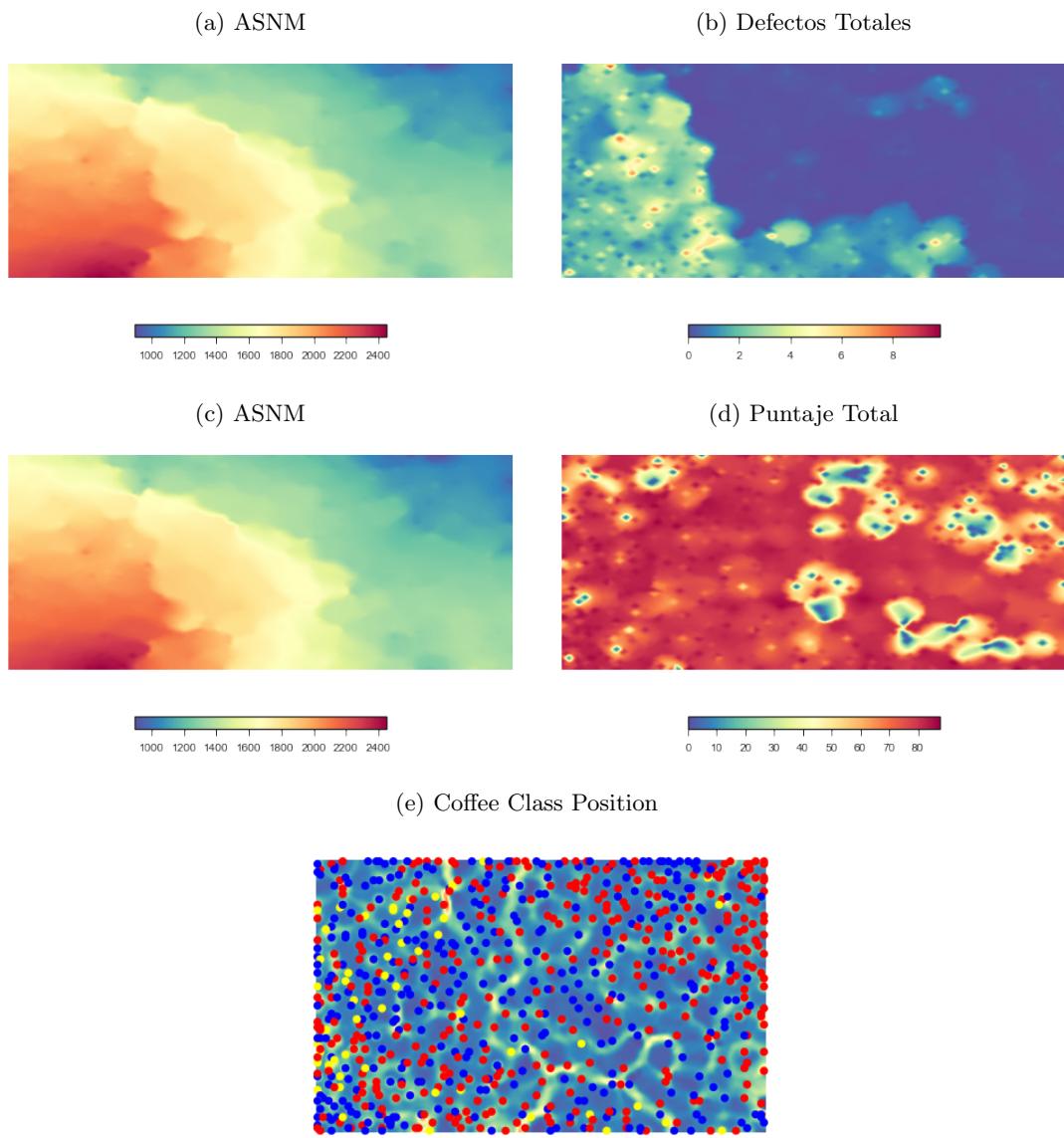


FIGURE 4.16 – Altitude, points totaux et défauts lors d'une seconde exécution du réseau de neurones

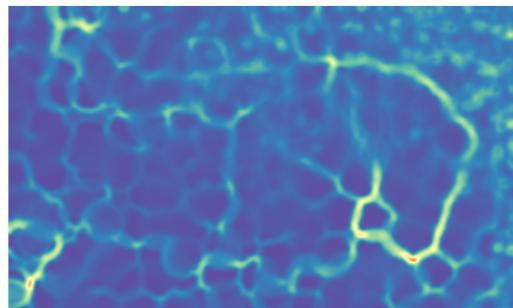


Afin de mieux visualiser la répartition des cafés d'année en année sur la carte SOM, une autre exécution a été réalisée¹ et les années affichées.

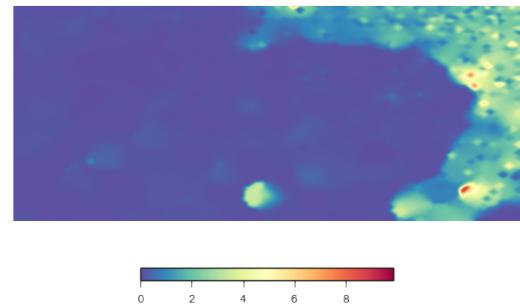
1. Les cartes SOM sont initialisées aléatoirement et les exécutions diffèrent les unes des autres. Il est difficile de sauvegarder l'environnement pour ensuite réaliser d'autres opérations sur les cartes d'une part et de l'autre chaque exécution peut amener une information supplémentaire..

FIGURE 4.17 – Répartition par année et mise en avant des composants intéressants de la SOM

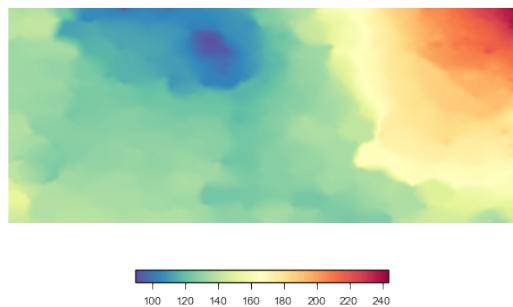
(a) Umatrix



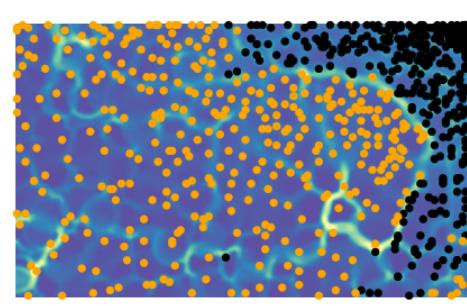
(b) Defectos Totales



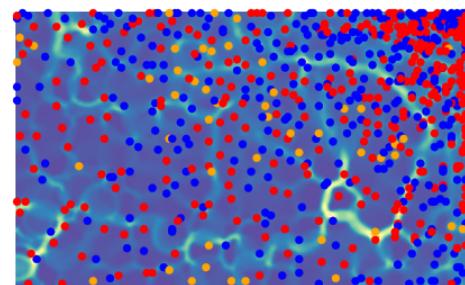
(c) Precipitations



(d) Year, 2011 (Noir), 2016 (Orange)



(e) Coffee Class Position, idem à 4.12



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 4.16. 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 4.16 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. Sur la figure 4.17 on peut observer que les fortes précipitations sont liées au nombre de défauts mais aussi fortement à l'année. Presque aucun des café de l'année 2011 n'a eu de score compris dans la catégorie 2.

4.2 Apprentissage supervisé

Le but de cette section est d'analyser la possibilité de prédiction de la qualité des cafés² à l'aide des données climatiques et de qualité de sols à disposition. Les méthodes Random Forest et Partial Least Squares seront utilisées.

4.2.1 Random Forest

Cette section présente comment la méthode Random Forest a été utilisée et les résultats produits. Les unités pour mesurer les erreurs sont la RMSE (Root Mean Squared Error), plus elle est proche de zéro, moins les points sont éloigné de la droite estimée lors d'une régression linéaire et R-Squared ou R^2 , un autre indice d'erreur pour lequel un bon résultat est proche de 1 et un mauvais proche de 0.

4.2.1.1 Méthode utilisée

Afin de se faire la meilleure idée possible des possibilités de prédiction, quatre variables de sorties ont été testées. Le nombre total de points (Puntaje Total), l'acidité (Acidez), la douceur (Dulzor) et la catégorie, obtenue d'après le tableau à la section 4.1.

La méthode Random Forest du package Caret a été utilisée avec une K-Fold Cross Validation ayant comme paramètre K = 10 (nombre de *folds*), répétée 3 fois.

```
RFmodelCategory=train (x_cat ,y_cat ,method="rf" ,
tunegrid=tunegrid_cat ,
trControl=trainControl(method="repeatedcv" ,
number=10,repeats=3),
tuneLength=10,importance = TRUE, proximity=T)
```

Listing 4.1 – Fonction d'entraînement et de test du modèle avec Random Forest

2. La qualité du café fait référence aux variables dépendantes citées dans la section 4.1.1.

4.2.1.2 Résultats de classification

En première partie, la tentative de classification avec comme variable de sortie la catégorie. Nous avons pour le modèle :

- 490 échantillons³
- 72 variables indépendantes
- 3 classes ("2", "3", "4")

Re-échantillonnage des résultats parmi les paramètres de réglage :

TABLE 4.4 – Réglage du modèle

mtry	Accuracy	Kappa
2	0.4995912	0.08462730
9	0.5117305	0.10653423
17	0.5062583	0.09657805
25	0.5109109	0.10608506
33	0.5075520	0.10148125
40	0.5014028	0.08705477
48	0.5088592	0.10056043
56	0.5095934	0.10397803
64	0.5081794	0.10110968
72	0.5061636	0.09829896

On observe qu'avec un *mtry*⁴ de 9, l'*Accuracy* est la meilleure .

3. La totalité des échantillons ont été utilisés ici, y compris ceux possédant 0 points.

4. À chaque split de l'arbre, l'algorithme va chercher *mtry* variables dans le set. Chaque nouveau split sera composé de *mtry* variables sélectionnées aléatoirement dans le set de données principal

TABLE 4.5 – Les vingt variables les plus importantes par catégorie (sur 72) pour la classification

	2	3	4
ASNM	60.754	38.042	100.00
prec9	64.442	8.963	90.27
tmin5	1.302	30.008	87.63
dtr7	86.974	69.378	54.59
tmean1	12.041	42.173	86.82
prec10	29.188	30.604	82.33
tmin4	16.840	56.842	79.80
tmax5	8.655	51.687	76.56
tmin1	21.357	56.763	75.76
tmax4	29.270	71.631	26.10
tmean2	12.710	42.159	71.51
tmean5	18.301	37.089	70.87
dtr5	16.198	70.822	39.11
tmax8	9.814	50.375	70.72
dtr8	10.785	43.304	70.20
tmax6	18.386	68.704	51.16
prec6	40.057	25.298	68.17
DtrTotal	13.864	67.471	36.17
Cascajoso	42.708	54.120	65.70
tmax10	6.866	45.774	65.59

Enfin après avoir construit le modèle, on peut le valider avec nos données de test. Nous obtenons les résultats suivants :

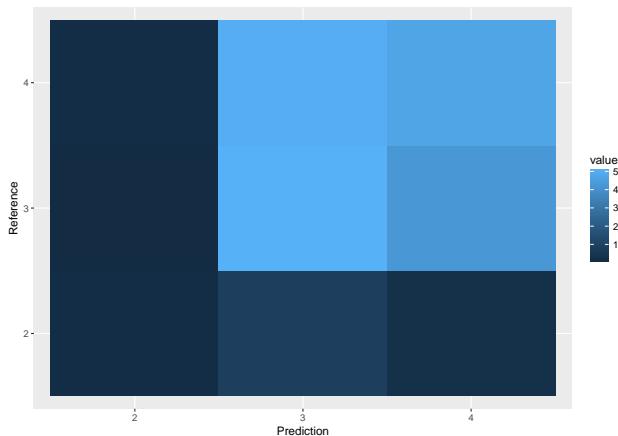


FIGURE 4.18 – Matrice de confusion de la classification avec Random Forest

		Reference		
Prediction	2	3	4	
2	2	1	2	
3	9	50	50	
4	3	43	47	

Overall Statistics

Accuracy : 0.4783
 95% CI : (0.4085 , 0.5486)
 No Information Rate : 0.4783
 P–Value [Acc > NIR] : 0.52732

Kappa : 0.0416
 Mcnemar's Test P–Value : 0.06796

Statistics by Class:

	Class: 2	Class: 3	Class: 4
Sensitivity	0.142857	0.5319	0.4747
Specificity	0.984456	0.4779	0.5741
Pos Pred Value	0.400000	0.4587	0.5054
Neg Pred Value	0.940594	0.5510	0.5439
Prevalence	0.067633	0.4541	0.4783
Detection Rate	0.009662	0.2415	0.2271
Detection Prevalence	0.024155	0.5266	0.4493
Balanced Accuracy	0.563657	0.5049	0.5244

Listing 4.2 – Test du modèle de classification

4.2.1.3 Régression sur la variable dépendante *Puntaje Total*

Notre modèle contient :

- 448 échantillons
- 72 variables indépendantes

TABLE 4.6 – Réglage du modèle

mtry	RMSE	Rsquared
2	6.033466	0.03056280
9	6.063304	0.03686526
17	6.042679	0.04446260
25	6.056445	0.04747083
33	6.064986	0.05062423
40	6.052046	0.05279225
48	6.076724	0.05268180
56	6.084703	0.05124400
64	6.066546	0.05294682
72	6.079099	0.05567081

On observe qu'avec un *mtry* de 2, la RMSE est la plus petite. Cependant, une fois testé le modèle donne l'erreur suivante : RMSE = 7.1726 et Rsquared = 0.00431.

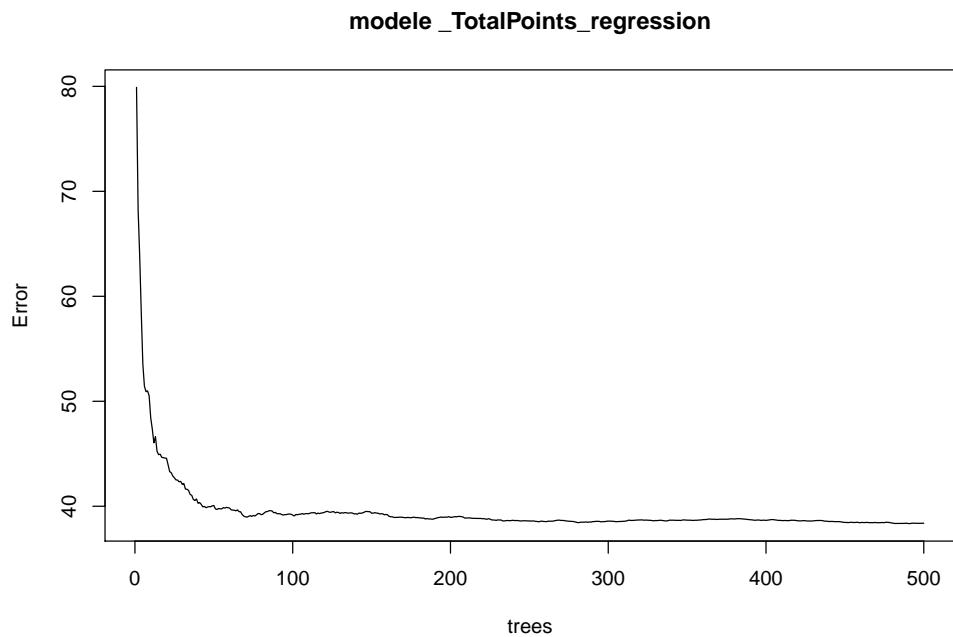


FIGURE 4.19 – Performances de la régression pour la variable *Puntaje Total*

Call:
randomForest(x = x, y = y, mtry = param\$mtry,
importance = TRUE,
proximity = .3, tuneGrid = .1)

Type of random forest: regression
Number of trees: 500
No. of variables tried at each split: 2

Mean of squared residuals: 38.38491
% Var explained: -9.59

Listing 4.3 – Test du modèle de classification

TABLE 4.7 – Les 20 variables les plus importantes du modèle.

	Overall
tmean2	100.00
tmax6	99.28
tmean4	96.71
tmax2	94.71
tmax5	92.77
PrecTotalAvg	90.81
tmax4	88.96
dtr4	88.89
tmin2	87.82
DtrTotal	87.43
tmin3	87.25
tmin5	86.97
tmean1	86.93
tmean8	85.72
tmax8	85.34
TminTotalAvg	84.71
dtr2	84.71
tmean10	84.42
tmin1	83.93
prec10	83.85

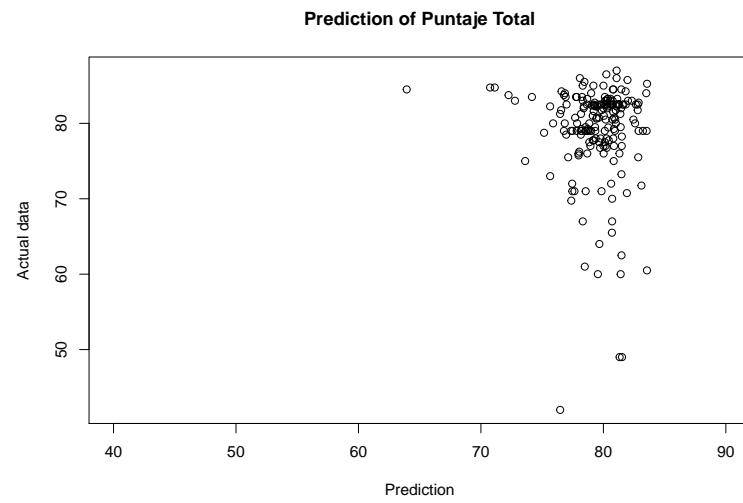


FIGURE 4.20 – Comparaison des prédictions et des valeurs actuelles pour la variable *Puntaje Total*

4.2.1.4 Régression sur la variable dépendante *Acidez*

Notre modèle contient :

- 448 échantillons
- 72 variables indépendantes

TABLE 4.8 – Réglage du modèle

mtry	RMSE	Rsquared
2	0.5523984	0.07780696
9	0.5587745	0.07205180
17	0.5591882	0.07332069
25	0.5614777	0.07128568
33	0.5601199	0.07188263
40	0.5602361	0.07391410
48	0.5608331	0.07075389
56	0.5599010	0.07333161
64	0.5604395	0.07236162
72	0.5607297	0.07082403

On observe qu'avec un *mtry* de 2, la RMSE est la plus petite. Cependant, une fois entraîné le modèle donne l'erreur suivante : RMSE = 0.5587 et Rsquared = 0.0261.

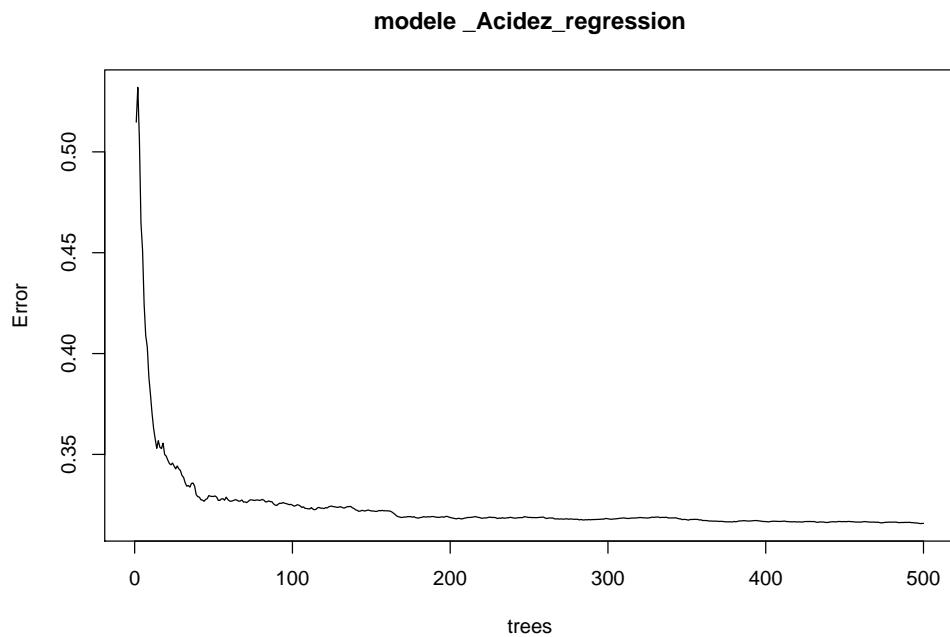


FIGURE 4.21 – Performances de la régression pour la variable *Acidez*

Call :
randomForest(x = x, y = y, mtry = param\$mtry,
importance = TRUE,
proximity = .3, tunegrid = .1)
Type of random forest: regression
Number of trees: 500
No. of variables tried at each **split**: 2

Mean of squared **residuals**: 0.3157974
% Var explained: -1.63

Listing 4.4 – Test du modèle de classification

TABLE 4.9 – Les 20 variables les plus importantes du modèle.

	Overall
dtr10	100.00
tmin9	98.16
tmean3	96.84
tmax8	96.40
tmax7	94.54
tmax6	93.25
tmin3	92.78
TmeanTotalAvg	90.41
DtrTotalAvg	89.54
tmax3	89.17
TmaxTotalAvg	88.40
tmax5	87.07
tmax10	85.38
TminTotalAvg	84.12
dtr8	83.90
tmean10	83.73
tmin5	83.27
tmax1	82.81
tmean4	82.49
tmean9	81.66

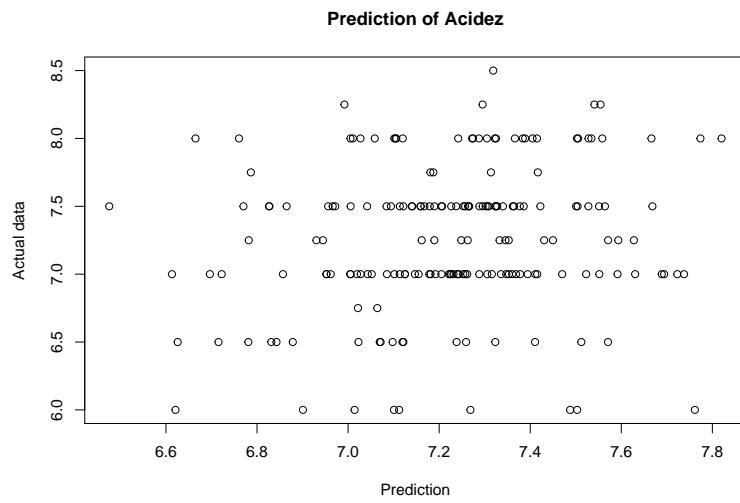


FIGURE 4.22 – Comparaison des prédictions et des valeurs actuelles pour la variable *Acidez*

4.2.1.5 Régression sur la variable dépendante *Dulzor*

Notre modèle contient :

- 439 échantillons
- 72 variables indépendantes

TABLE 4.10 – Réglage du modèle

mtry	RMSE	Rsquared
2	0.7996682	0.02423378
9	0.8030381	0.03690104
17	0.8082138	0.04453890
25	0.8112174	0.04482055
33	0.8141780	0.04452465
40	0.8163036	0.04853452
48	0.8187376	0.04858482
56	0.8211659	0.04628629
64	0.8209675	0.04722086
72	0.8226232	0.04976240

On observe qu'avec un *mtry* de 2, la RMSE est la plus petite. Cependant, une fois entraîné le modèle donne l'erreur suivante : RMSE = 0.8747 et Rsquared = 0.0009.

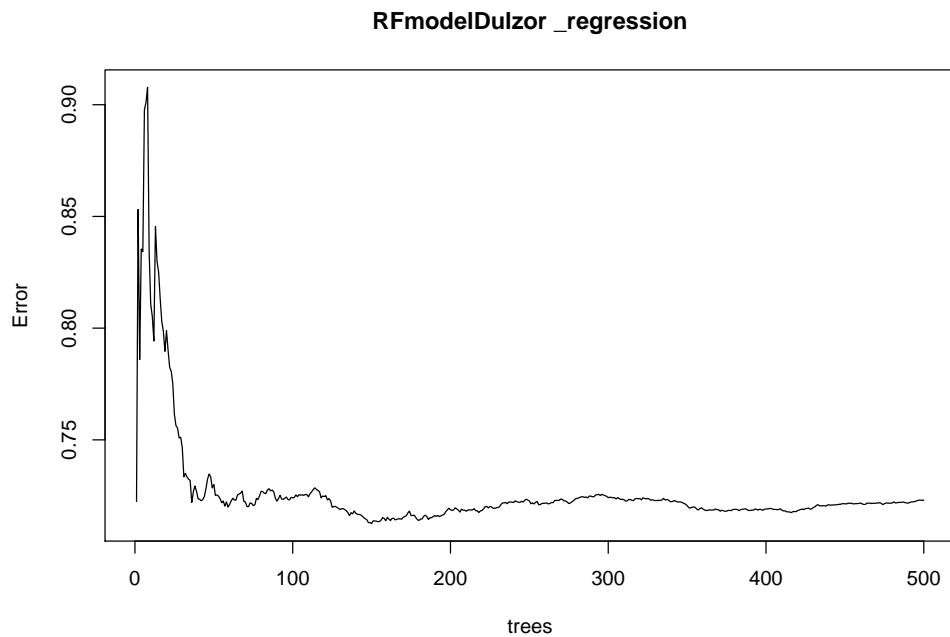


FIGURE 4.23 – Performances de la régression pour la variable *Dulzor*

Call:
 randomForest(x = x, y = y, mtry = param\$mtry,
 importance = TRUE,
 proximity = .3, tunegrid = .1)
 Type of random forest: regression
 Number of trees: 500
 No. of variables tried at each split: 2

Mean of squared residuals: 0.7230219
 % Var explained: -9.25

Listing 4.5 – Test du modèle de classification

TABLE 4.11 – Les 20 variables les plus importantes du modèle.

	Overall
dtr4	100.00
tmax4	99.29
tmax9	96.55
tmin4	93.13
tmean4	92.17
TmaxTotal	91.49
tmin5	90.85
TmeanTotal	90.64
prec6	89.90
tmin2	89.84
dtr10	89.69
tmax6	89.17
tmax5	87.47
tmax7	87.16
dtr8	87.15
tmax2	86.37
dtr2	86.23
tmean1	85.78
TmeanTotalAvg	85.08
DtrTotal	84.98

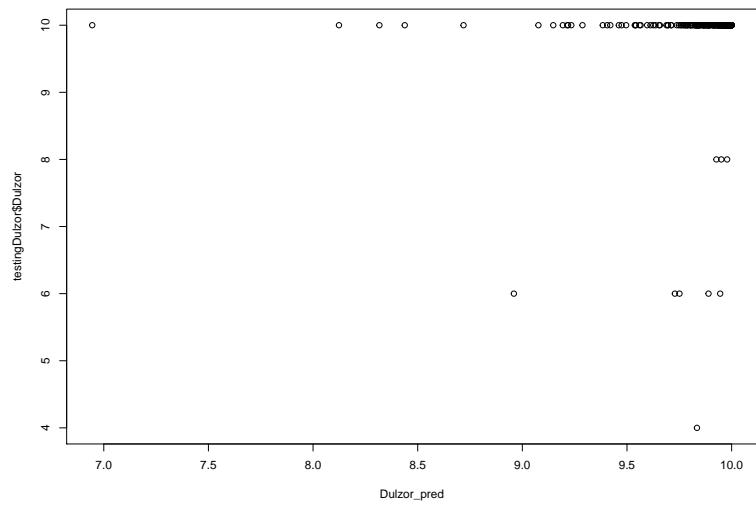


FIGURE 4.24 – Comparaison des prédictions et des valeurs actuelles pour la variable *Dulzor*

Discussion des résultats pour Random Forest Avec une *accuracy* de seulement 0.47 (47% d'échantillons classés correctement) et un test de Kappa d'uniquelement 0.0416⁵, la classification à l'aide de Random Forest ne semble pas fiable avec les données en notre possession.

Pour mesurer la qualité de la régression, nous utiliserons comme mesure le coefficient r^2 fournit par les résultats de Random Forest. Plus le coefficient s'approche de 1 plus la variation totale des sorties prédites par le modèle est faible et donc plus le modèle est bon. Les résultats obtenus avec les 3 variables de sortie différentes sont les suivants :

- Puntaje Total : $r^2 = 0.00431$
- Acidez : $r^2 = 0.0261$
- Dulzor : $r^2 = 0.0009$

4.2.2 Random Forest avec variables à haute variabilité

Afin de tester le modèle sans données inutilement redondantes et fortement corrélées, un second modèle Random Forest a été entrainé avec les variables indépendantes qui avaient un taux de variabilité supérieur à un certain niveau. Le premier niveau a été placé à 10, ce qui a pour effet de ne garder que 25 variables. Le second a été placé à 20, ce qui a pour effet de ne garder que 21 variables. Les résultats se sont avérés tout aussi imprécis, voir par exemple les résultats de prédiction sur la figure 4.25 avec le seuil à 10 pour la variable *Puntaje Total* et sur la figure 4.26 avec le seuil à 20. Les cafés ayant un total de zéro point ont ici été gardés afin d'observer une éventuelle prédiction correspondante. Sur la figure 4.25 on peut observer qu'un des café ayant zéro points a été prédit à moins de 40 points, ce qui signifie que certaines variables indépendantes peuvent tirer la prédiction dans le bon sens. Les variables montrée dans le tableau 4.12 sont celles utilisées pour créer le modèle avec un seuil de 10. On peut voir que les précipitations ont une grande importance et cela rejoint l'idée observée dans la section 4.1.2.4 que la quantité de précipitations est importante pour la qualité du grain Cependant cela ne permet que d'avoir une approximation de la quantité de défauts physiques du grain et non de la qualité générale.

5. Le test kappa de Cohen fournit un résultat proche de 1 si les prédictions sont parfaites, et proche de 0 s'il n'y a aucune correspondance.

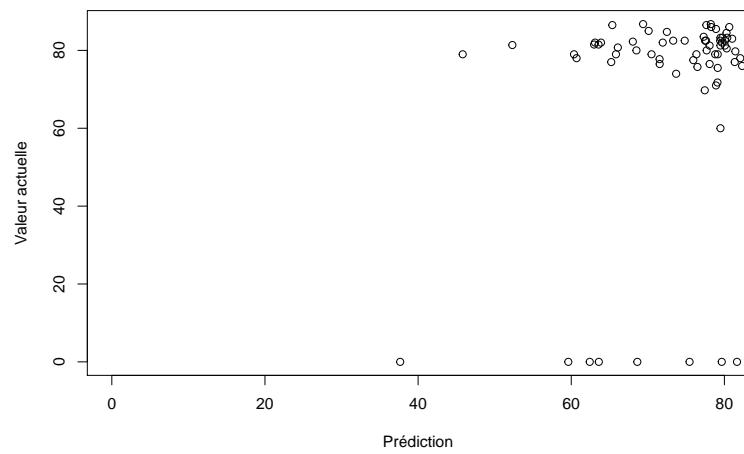


FIGURE 4.25 – Prédiction du total de points avec 25 variables indépendantes

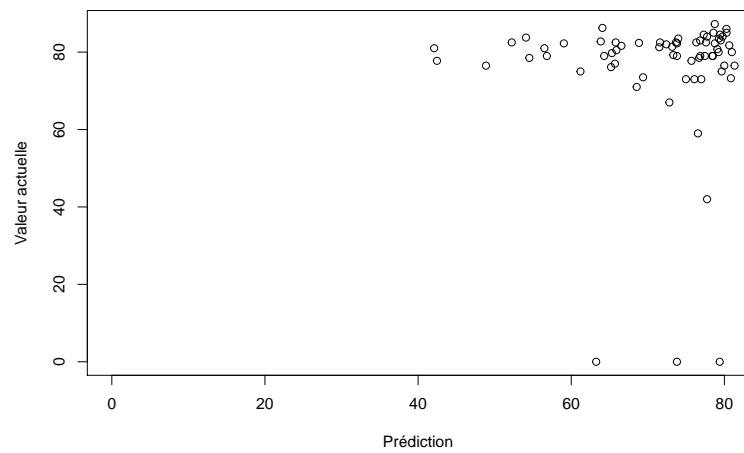


FIGURE 4.26 – Prédiction du total de points avec 21 variables indépendantes

TABLE 4.12 – Variables à haute variabilité avec un coefficient de variabilité supérieur à 10

	%IncMSE	IncNodePurity
ASNM	5.0408727	15044.3045
Luminosidad	5.4787179	3642.4806
prec1	7.6033403	11395.8391
prec2	8.4894841	12991.3588
prec3	9.6477012	14018.6199
prec4	7.6742913	16432.9873
prec5	8.5170477	13554.8590
prec6	9.3037427	14994.6836
prec7	7.3229410	14325.6941
prec8	9.0394682	13509.3978
prec9	9.6867243	13287.9725
prec10	12.0962566	17472.3424
dtr5	10.2926692	16738.1273
dtr10	9.1992584	15261.9880
PrecTotal	12.0991208	15483.9130
PrecTotalAvg	11.6456487	13780.0364
OrientationNum	2.4731816	10925.6272
Slope	0.1353807	15492.9382
org_avg	2.7056077	6520.6197
Franco	1.5936970	1931.0307
Arcilloso	-2.9621275	1746.0451
Limoso	2.7839655	1631.1038
Arenoso	-1.5199547	831.4292
Cascajoso	-0.5423810	2330.3915

4.2.3 Partial Least Squares

Cette méthode utilise à la fois le principe de régression et d'analyse en composante principale (PCA ou ACP). La méthode *plsreg1* de R fournit en retour plusieurs objets dont les corrélations entre variables. Leur représentation graphique permet de visualiser aisément où se placent les variables de sorties choisie par rapport aux autres.

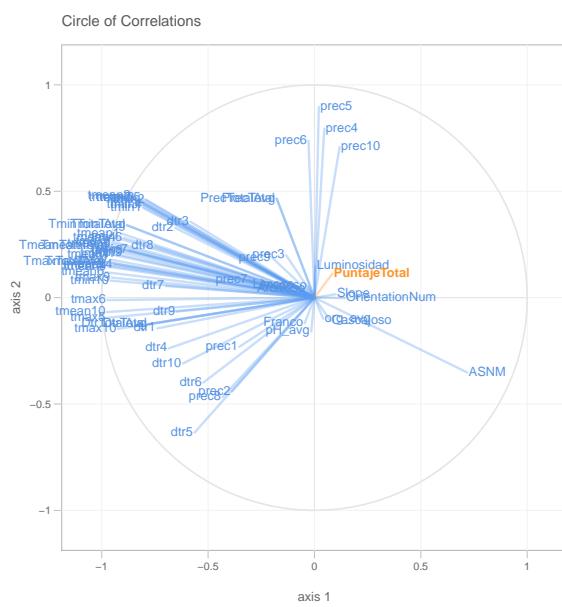


FIGURE 4.27 – Représentation des corrélations entre les variables indépendantes et la variable dépendante *Puntaje Total*

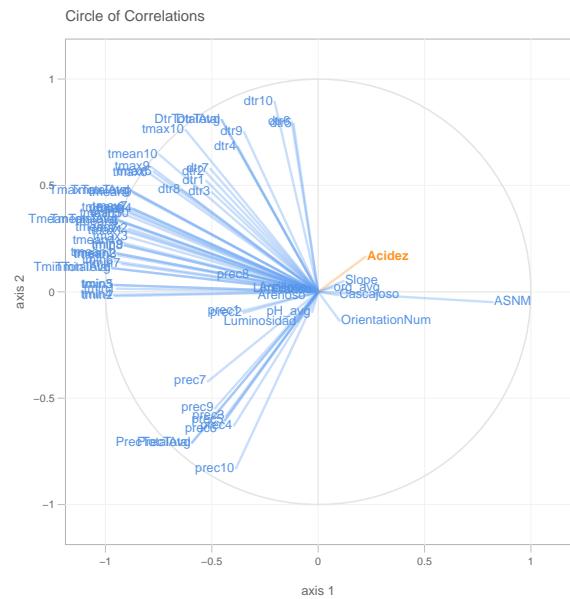


FIGURE 4.28 – Représentation des corrélations entre les variables indépendantes et la variable dépendante *Acidez*

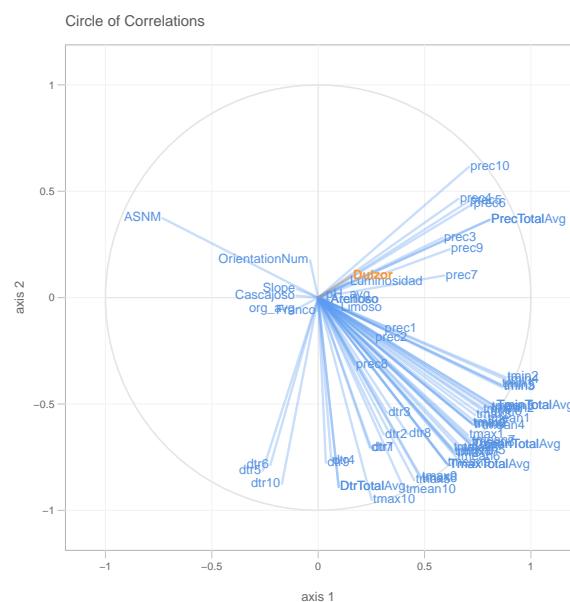


FIGURE 4.29 – Représentation des corrélations entre les variables indépendantes et la variable dépendante *Dulzor*

4.2.3.1 Analyse des résultats

La prédiction avec Partial Least Squares ne donne pas de résultats plus convaincants que Random Forest. Cependant et afin de mitiger ces résultats, la méthode n'a pas été exploitée avec son plein potentiel car un seul block de données à été utilisé (la totalité des variables indépendantes) alors que PLS et Multiblock PLS peuvent éventuellement être amélioré en réglant les paramètres plus finement.

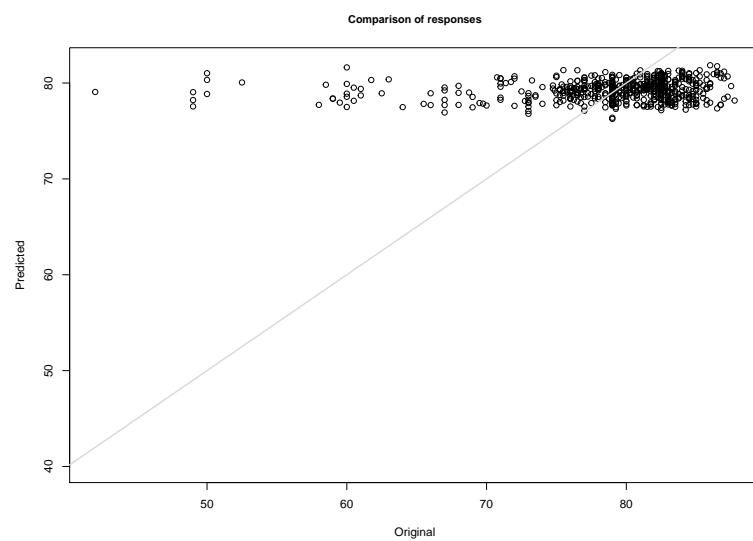


FIGURE 4.30 – Comparaison des prédictions et des valeurs actuelles pour la variable Puntaje Total

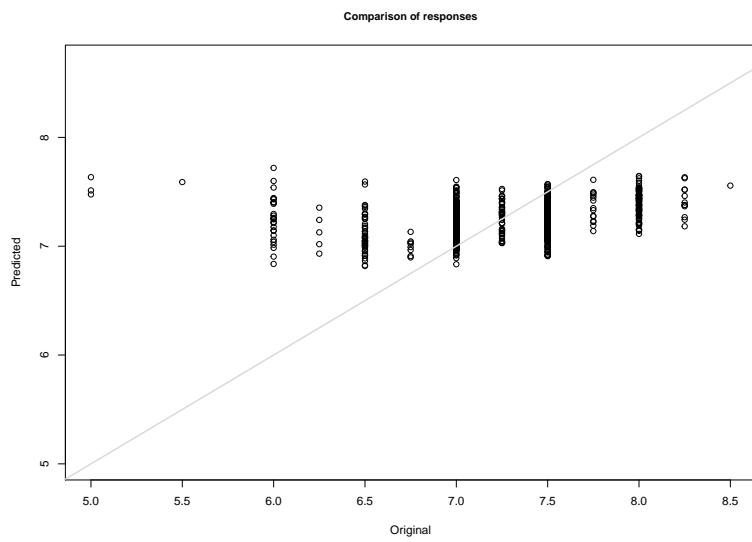


FIGURE 4.31 – Comparaison des prédictions et des valeurs actuelles pour la variable Acidez

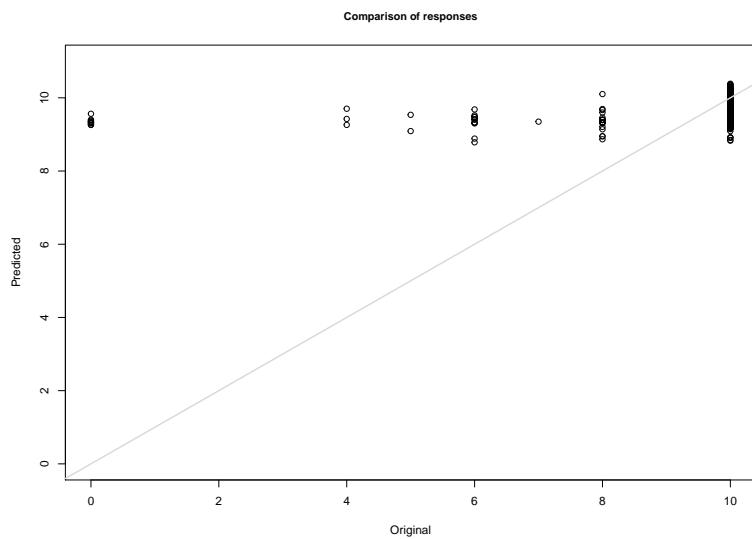


FIGURE 4.32 – Comparaison des prédictions et des valeurs actuelles pour la variable Dulzor

5 Discussion des résultats

Les différentes analyses réalisées, tant au niveau de l'apprentissage supervisé que non-supervisé n'ont pas permis de créer un modèle fiable de description ou de prédiction de la qualité du café. Cependant, certains résultats peuvent être mis en avant et certaines critiques peuvent être faites sur les données et les méthodes utilisées.

Résultats exploitables Concernant les résultats exploitables, une relation a été observée entre la quantité de pluie durant l'année et le nombre de défauts physiques du grain. Plus il y a de pluie, plus il y a de grains ayant des défauts et plus il y a de cafés ayant la note zéro (voir figure 4.17). Rappelons que les grains défectueux sont éliminés par un processus industriel permettant l'élimination des grains ayant une couleur ou une densité anormale. Cependant, certains défauts, comme la présence de trous dûs aux parasites (Broca, ou Broca de punto), sont plus difficilement détectables et peuvent amener un goût désagréable au café. Un seul grain peut modifier le goût d'une tasse ! On remarque aussi que l'année 2011 a beaucoup moins de cafés dont la note dépasse les 80 (voir figure 2.9), score minimal pour avoir ma mention *Specialty Coffee*. Il est donc raisonnable de penser que les fortes quantités de pluie, nuisent à la qualité du café et que plus il pleut, moins il y a de chance d'avoir de très bon cafés.

Données disponibles Certain points concernant la qualité des données reçues sont à mettre en avant. Premièrement, les données de qualités de sols ne contenant que le pH, le taux de matières organiques et la texture, toutes les relations chimiques éventuelles entre les minéraux du sol et les arômes n'ont pas pu être explorées. La composition chimique du sol n'est pas la seule à avoir une influence possible sur la qualité et les saveurs du café. Les différents traitement du café une fois récolté, comme la fermentation ou le séchage, peuvent avoir une grande influence mais malheureusement aucune donnée n'a été fournie à ce sujet non plus.

Deuxièmement, les données de dégustation contenaient parfois, de manière difforme, des notes sur les arômes ou saveurs du café. Ces données peuvent être décrites comme sur la figure 5.1 et pourraient permettre de décrire d'une manière plus sensorielle le goût du café et ainsi faire un éventuel lien avec les composants chimiques du sol et les pratiques culturelles ou même le climat. Malheureusement, le manque de données en nombre d'une part et d'uniformité d'autre part n'ont pas permis d'intégrer ces informations au set de données et donc d'en analyser les éventuelles corrélations.



FIGURE 5.1 – Roue des parfums du café

Il serait intéressant de prélever d'autres données, directement au niveau des fermes. Beaucoup de facteurs influencent la croissance d'une plante et la maturation de ses fruits en particulier les facteurs influençant la photosynthèse. Si une plante est particulièrement exposée à l'est, et donc au soleil du matin, la photosynthèse sera plus efficace car l'humidité ambiante est plus élevée, ce qui améliore les capacités d'assimilation de la lumière par la plante. Par contre, si la plante grandit sur un terrain ombragé, donc entourée d'autres plantes, l'humidité sera aussi plus élevée mais la plante devra fabriquer plus de feuilles pour capter plus de lumière et aura donc moins d'énergie à mettre dans les fruits. L'altitude peut aussi avoir une influence importante. Sur la luminosité, en particulier par rapport à la couverture nuageuse qui peut être parfois plus importante en montagne, mais aussi sur la température. Dans les lieux plus froids, la plante fonctionne plus lentement et les fruits ont besoin de plus de temps pour murir, ce qui laisse plus de temps aux arômes pour se développer et produit de manière générale des cafés de qualités.

Durant le projet, une réunion a été organisée avec Felipe Rincón, coordonnateur de gestion au comité départemental des caféculteurs de Risaralda, qui est à l'origine des données sur le café que nous avons reçues. La possibilité m'a été donnée de poser des

questions entre autre sur le processus de récolte et de centralisation des données, processus jusqu’alors inexistant. M. Rincón m’a alors confié que les différentes demandes effectuées pour ce travail ainsi que le résultat final de l’agglomération des données disponibles, l’ont poussé à commencer un processus de normalisation de la récolte des données. Il est donc envisageable de pouvoir retenter une étude de ce type lorsque les données existantes auront éventuellement été complétées et lorsque de nouvelles données, plus complètes et possédant moins de défauts, auront été générées. Parmi les défauts importants signalés à M. Rincón, on notera qu’aucune date de récolte n’a été fournie. Seul les dates de dégustation étaient disponibles et les périodes de croissance de la plante ont dû être estimées.

Méthodes utilisées Il est aussi important de mentionner les méthodes utilisées. Random Forest, PLS, ou le clustering, peuvent être paramétrées de beaucoup de manières différentes et leur efficacité peut être testée et améliorée. Dans le cadre de ce projet, avec le peu d’heures à disposition, ces méthodes ont été utilisées de manière exploratoire afin d’avoir une première impression sur les possibilités de modélisation. D’un même point de vue, d’autres méthodes auraient pu être explorées en jouant sur les particularités de chacune d’entre-elles par rapport aux types de données en notre possession. Pour améliorer cette partie, un bagage théorique supplémentaire en statistique serait nécessaire afin de mieux cerner les subtilités des méthodes de modélisation.

6 Conclusion

Ce projet de recherche a permis de mettre au jour des relations entre le café et le climat sans pour autant parvenir à des résultats de prédiction ou de caractérisation satisfaisants. Des données supplémentaires et du temps de recherche seraient nécessaires afin de pouvoir améliorer les résultats obtenus.

Les recherches effectuées ont eu un impact positif sur la FNC qui, en la personne de M Felipe Rincón, viennent d'engager un processus de normalisation de la récolte des données afin d'améliorer les possibilités d'analyses dans un future proche.

La relation entre la mauvaise qualité des grains et les fortes précipitations pourrait permettre de mettre en place des procédures particulières en cas de fortes pluies afin d'éviter des pertes inutiles. Une étude devrait cependant être faite afin d'identifier les éventuels facteurs reliant la pluie aux défauts et ainsi pouvoir lutter contre, ces facteurs pouvant être naturels, humains ou logistiques.

Les objectifs du cahier des charges ont été atteint dans la mesure où les données ont pu être analysées et les résultats interprétés. Cette expérience a été très enrichissante tant au niveau professionnel que personnel. Il s'agissait d'un cas réel d'étude avec des données et des analyses du terrain, pour le terrain.

A Description des données

A.1 Dataset

Le dataset présenté ci-dessous a été réalisé en regroupant les informations de plusieurs classeurs Excel ainsi que de documents au format PDF. Il s'agit d'une version contenant la totalité des données nécessaires au pré-traitement comme les coordonnées géographiques, les noms des fermes etc. Des sous-sets ont été extraits selon les besoins afin de réaliser les calculs avec, par exemple, uniquement les valeurs numériques et une variable dépendante spécifique.

TABLE A.1 – Dataset partie 1

SICA	Numéro d'identification unique par parcelle
Cedula	Numéro de document d'identité du caféculteur
Municipio	Municipio
Vereda	Vereda
Finca	Ferme
EPSG :3116_X	Coordonnées
EPSG :3116_Y	Coordonnées
EPSG :4326_Y	Coordonnées
EPSG :4326_X	Coordonnées
Variedad	Variété
FechaAnalysis	Date d'analyse
year	Année d'analyse
Occurence	Nombre d'occurrence du café
UV	Pas d'information
Olor	Pas d'information
Humedad	Humidité
Merma	Ullage
Pergamino	Produit après lavage
Almendra	Uniformité des grains
AlmendraTotal	Total de grain
AlmendraSana	Total de grain sains
NegrosYVinagres	Défaut physique
Broca	Défaut physique

TABLE A.2 – Dataset partie 2

BrocaDePunto	Défaut physique
Veteado	Défaut physique
Mordido	Défaut physique
Inmaduro	Défaut physique
Flojo	Défaut physique
Sobresecado	Défaut physique
Arrugado	Défaut physique
Aplastado	Défaut physique
Cristalizado	Défaut physique
Reposado	Défaut physique
Granizo	Défaut physique
Conchas	Défaut physique
Partido	Défaut physique
Ambar	Défaut physique
DefectosTotales	Total des défauts
ASNM	Altitude [mètres]
Luminosidad	Luminosité (3 catégories)
prec1-10	Précipitations sur 10 mois
tmin1-10	Températures min sur 10 mois
tmax1-10	Températures max sur 10 mois
tmean1-10	Températures moyennes sur 10 mois
dtr1-10	Diurnal Temperature Range sur 10 mois
PrecTotal	Total des précipitations
TminTotal	Total des températures minimales
TmaxTotal	Total des températures maximales
TmeanTotal	Total des températures moyennes
DtrTotal	Total des DTR
PrecTotalAvg	Moyenne des prec sur 10 mois
TminTotalAvg	Moyenne des tmin sur 10 mois
TmaxTotalAvg	Moyenne des tmax sur 10 mois
TmeanTotalAvg	Moyenne des tmean sur 10 mois
DtrTotalAvg	Moyenne des DTR sur 10 mois
OrientationNum	Orientation N-S-E-W [8 catégories]
Slope	Pente [pourcentage]
Soil Profile	Profil du sol
Unidad_c_1	Sous-profil
pH_avg	pH moyen sur 1m

TABLE A.3 – Dataset partie 3

org_avg	Matière organique moyenne sur 1m [pourcentage]
Franco	Sol Franco [0,1,2,3]
Arcilloso	Sol Argilleux [0,1,2,3]
Limoso	Sol Limoneux [0,1,2,3]
Arenoso	Sol Sabloneux [0,1,2,3]
Cascajoso	Sol Gravilloneux [0,1,2,3]
Taza1	Tasse une (limpia ou non)
Taza2	Tasse deux (limpia ou non)
Taza3	Tasse trois (limpia ou non)
Taza4	Tasse quatre (limpia ou non)
Taza5	Tasse cinq (limpia ou non)
Aroma-Fragancia	Parfum-Arome [0-10 pts]
Acidez	Acidité [0-10 pts]
Cuerpo	Corps [0-10 pts]
Sabor	Saveur [0-10 pts]
SaborResidual	Saveur résiduelle [0-10 pts]
Dulzor	Douceur [0-10 pts]
Uniformidad	Uniformité [0-10 pts]
Balance	Equilibre [0-10 pts]
TazaLimpia	Tasse "propre" [0-10 pts]
PuntajeCatador	Points dégustateur [0-10 pts]
PuntajeTotal	Points totaux [0-100 pts]
Category	Catégorie [1,2,3,4]

A.2 Données de dégustation brutes

Les nombreux fichiers sont disponibles dans les sources du projet.

A.3 Données climatiques et géographiques brutes

Les données climatiques consistent en l'extrapolation de multiples points (stations météo) sur une partie du territoire. Ces données sont décrites dans le tableau A.4. L'accès aux données brutes n'est pas possible via Github compte tenu du poids des dites données.

TABLE A.4 – Format des données climatiques brutes

Nom de la colonne	Description (valeur par défaut)
NCOLS	Nombre de colonnes (720)
NROWS	Nombre de lignes (720)
XLLCORNER	Longitude, coin inférieur gauche [Degré decimal] (-77.50042)
YLLCORNER	Latitude, coin inférieur gauche [Degré decimal] (3.500417)
CELLSIZE	Taille des cellules sur la carte [Degrés decimal] (-0.004166667)
NODATA_value	Valeur « pas de données » (-9999)
Datas	Données climatique concernée, tableau de 720 par 720

FIGURE A.1 – Shapefile des différentes orientations (Nord - Sud - Est - Ouest) des points dans le département de Risaralda

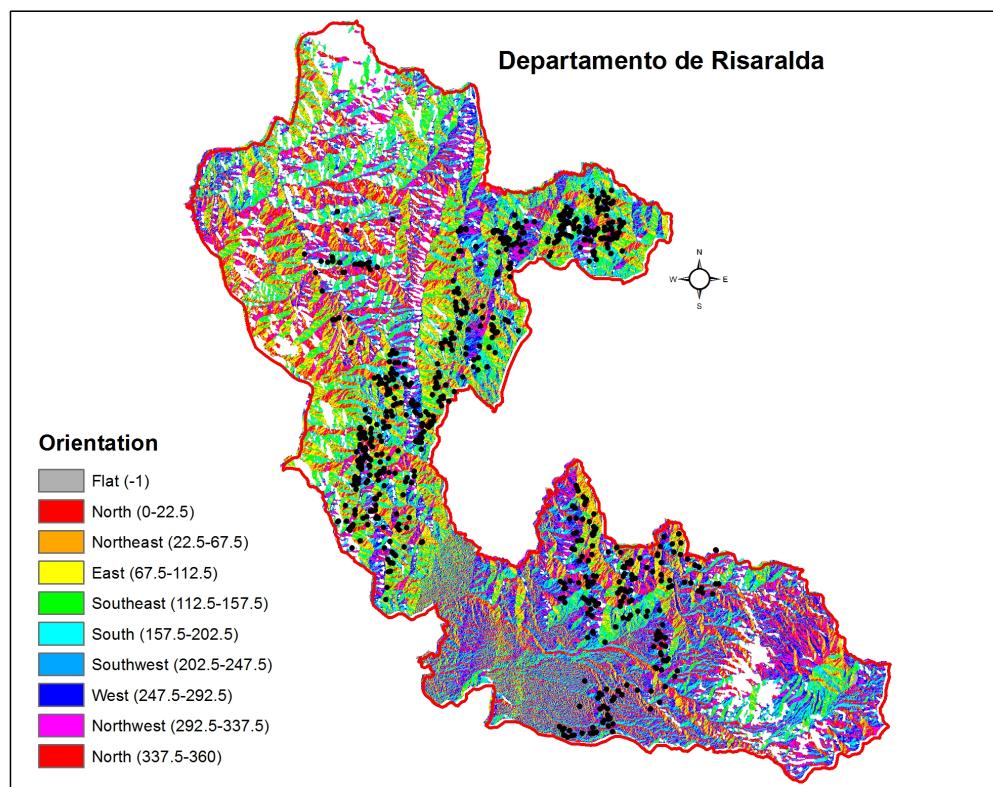
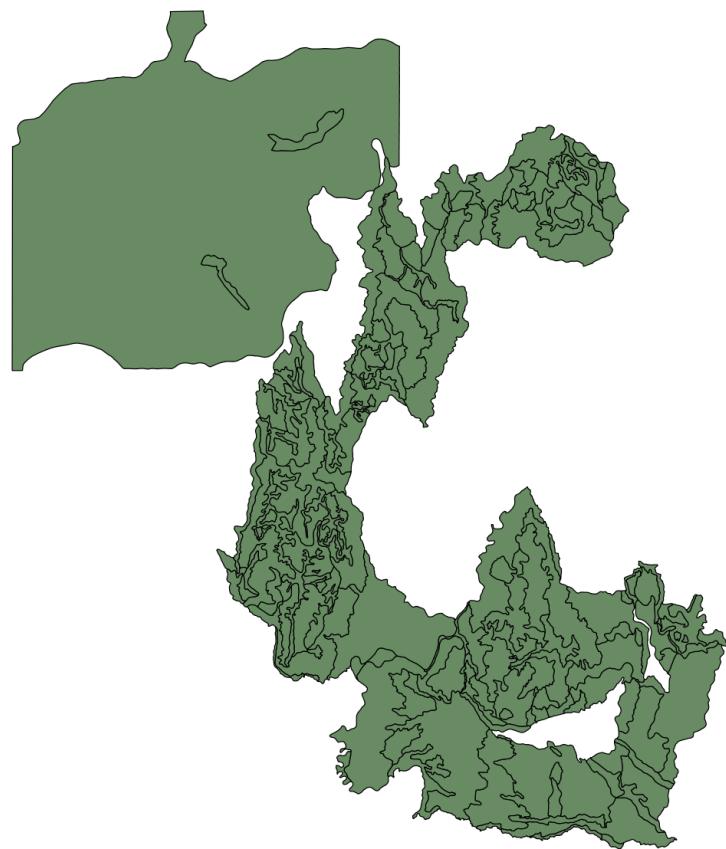


FIGURE A.2 – Shapefile des différents profils de sol dans le département de Risaralda



B Importances des variables par cluster

TABLE B.1 – HCPC avec trois clusters comparés à la sortie Acidez

	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 B.2 – Tableau des clusters pour la sortie Puntaje Total

	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 B.3 – Importance des variables lors de la réalisation des clusters

	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

C Autres informations

Les sources du projet se trouvent sur un dépôt *Github* à l'adresse https://github.com/ThibaultSchowing/Bachelor_Thesis_2017_Sources.

Le présent document et les fichiers associés se trouvent sur un dépôt *Github* à l'adresse https://github.com/ThibaultSchowing/Bachelor_Thesis_2017.

Les partial plots sont disponibles sur le dépôt *Github* du projet à l'adresse https://github.com/ThibaultSchowing/Bachelor_Thesis_2017_Sources/tree/master/Notebook/R/Projects/BachelorThesis/PartialPlots.

Les analyses par variétés sont disponible sur le dépôt *Github* du projet à l'adresse https://github.com/ThibaultSchowing/Bachelor_Thesis_2017_Sources/tree/master/Notebook/R/Projects/BachelorThesis/VARIETY_ANALYSIS. Les répertoires créés ne sont pas forcément représentatifs des analyses effectuées et peuvent être vides.

Bibliographie

- [1] Cafe import , http://www.cafeimports.com/origin_colombia.
- [2] Centre du commerce international, <http://www.intracen.org>.
- [3] Condensé d'informations sur la caféculture en colombie, https://fr.wikipedia.org/wiki/Caf%C3%A9iculture_en_Colombie.
- [4] Ideam - instituto de hidrología, meteorología y estudios ambientales, <http://www.ideam.gov.co/>.
- [5] Informations générales sur la colombie, www.colombia.com.
- [6] Organisation internationnale du café, <http://www.ico.org/>.
- [7] Scaa cuping protocol, <http://www.scaa.org/pdf/resources/cupping-protocols.pdf>.
- [8] Sistema information cafetera, www.federaciondecafeteros.org.
- [9] Toutes les relations avec le café colombien, <http://www.cafedecolombia.com>.
- [10] Wikipédia - climat colombien, https://fr.wikipedia.org/wiki/Climat_de_la_Colombie.
- [11] *Guide de l'exportateur de café.* - 3éme 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.
- [12] Colombie : El niña provoque chaleur et sécheresse, juan manuel santos appelle à la vigilance. *Actu Latino*, Jan 2016.
- [13] Issam El Alaoui. Les méthodes ensemblistes pour algorithmes de machine learning. *Octo blog*, 2014.
- [14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction*. Springer, 2 edition, 2009.
- [15] Teuvo Kohonen. Neurocomputing : Foundations of research. chapter Self-organized Formation of Topologically Correct Feature Maps, pages 509–521. MIT Press, Cambridge, MA, USA, 1988.
- [16] OMMpublic.wmo.int. D'une intensité exceptionnelle, l'épisode el niña a amorcé son déclin, mais ses effets perdurent. 2016.
- [17] Leo Breiman Statistics and Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.
- [18] 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 températures 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-ecologica/el-suelo-como-fuente-de-vida-propiedades-ii-	7
2.4	Emplacement des fermes avec coloration selon le nombre de points attribués au café. Nord.	8
2.5	Emplacement des fermes avec coloration selon le nombre de points attribués au café. Sud.	9
2.6	Répartition des variétés de café dans le département.	10
2.7	Nombre de points totaux vs altitude	12
2.8	Répartition des cafés par rapport au nombre de points totaux	12
2.9	Répartition des points totaux par année	14
3.1	Schéma simple du fonctionnement de Random Forest. Source : https://www.youtube.com/watch?v=ajTc5y3OqSQ	16
3.2	Méthode PLS. X est représentée par son score t et Y par u. Une première estimation de U est multipliée à travers X pour obtenir une approximation du poids ω_t . Le poids est normalisé pour être de longueur 1 et re-multiplié à travers X pour produire t. A partir de t et de Y, le poids q^T est obtenu ce qui donne un nouveau vecteur u. Cette opération est répétée jusqu'à la convergence de t. [18]	17
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 . [18]	18
3.4	Pays organisés en SOM d'après des indicateurs de pauvreté. Source : http://www.cis.hut.fi/research/som-research/worldmap.html	20
3.5	Pays correspondants à la carte SOM de la figure 3.4 Source : http://www.cis.hut.fi/research/som-research/worldmap.html	20
4.1	Étapes de construction du set de données et pertes d'observations.	25
4.2	Matrice de corrélation entre les différentes sorties.	27

4.3	Matrice de corrélation entre toutes les variables.	28
4.4	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.	29
4.5	Résultats de la PCA sous forme graphique. Réalisé avec la totalité des variables	30
4.6	PCA avec Prcomp : Coloration par année	31
4.7	PCA avec Prcomp : Coloration par catégorie	32
4.8	HCPC et proposition de nombre de cluster	33
4.9	HCPC arbre 3D	34
4.10	Saut d'inertie du dendrogramme	35
4.11	U-Matrix de la carte SOM	36
4.12	U-Matrix avec les points. En jaune les cafés de catégorie 2,en bleu catégorie 3 et en rouge catégorie 4. Pour rappel, les catégories sont expliquées au point 2.2	37
4.13	Composants de qualité - Variables de sortie	38
4.14	Précipitations	39
4.15	SOM - Autres données	40
4.16	Altitude, points totaux et défauts lors d'une seconde exécution du réseau de neurones	41
4.17	Répartition par année et mise en avant des composants intéressants de la SOM	42
4.18	Matrice de confusion de la classification avec Random Forest	46
4.19	Performances de la régression pour la variable <i>Puntaje Total</i>	48
4.20	Comparaison des prédictions et des valeurs actuelles pour la variable <i>Puntaje Total</i>	50
4.21	Performances de la régression pour la variable <i>Acidez</i>	51
4.22	Comparaison des prédictions et des valeurs actuelles pour la variable <i>Acidez</i>	53
4.23	Performances de la régression pour la variable <i>Dulzor</i>	54
4.24	Comparaison des prédictions et des valeurs actuelles pour la variable <i>Dulzor</i>	56
4.25	Prédiction du total de points avec 25 variables indépendantes	58
4.26	Prédiction du total de points avec 21 variables indépendantes	58
4.27	Représentation des corrélations entre les variables indépendantes et la variable dépendante <i>Puntaje Total</i>	60
4.28	Représentation des corrélations entre les variables indépendantes et la variable dépendante <i>Acidez</i>	61
4.29	Représentation des corrélations entre les variables indépendantes et la variable dépendante <i>Dulzor</i>	61
4.30	Comparaison des prédition et des valeurs actuelle pour la variable Puntaje Total	62

4.31	Comparaison des prédition et des valeurs actuelle pour la variable Acidez	63
4.32	Comparaison des prédition et des valeurs actuelle pour la variable Dulzor	63
5.1	Roue des parfums du café	66
A.1	Shapefile des différentes orientations (Nord - Sud - Est - Ouest) des points dans le département de Risaralda	74
A.2	Shapefile des différents profils de sol dans le département de Risaralda . .	75

Liste des tableaux

2.1	Nombre de points aux dégustations par catégories	10
2.2	Altitude par catégorie	11
2.3	Précipitations par année	13
2.4	Températures moyennes par année	13
2.5	Points totaux par année	14
4.1	Catégories de cafés d'après le nombre de points Source : http://www.scaa.org/?page=resources&d=cupping-protocols .	26
4.2	Tableau des rotations des six premiers composants de la PCA avec mise en évidence des variables les plus importantes par composante.	30
4.3	HCPC avec trois clusters comparés à la sortie catégories	35
4.4	Réglage du modèle	45
4.5	Les vingt variables les plus importantes par catégorie (sur 72) pour la classification	46
4.6	Réglage du modèle	48
4.7	Les 20 variables les plus importantes du modèle.	49
4.8	Réglage du modèle	51
4.9	Les 20 variables les plus importantes du modèle.	52
4.10	Réglage du modèle	54
4.11	Les 20 variables les plus importantes du modèle.	55
4.12	Variables à haute variabilité avec un coefficient de variabilité supérieur à 10	59
A.1	Dataset partie 1	71
A.2	Dataset partie 2	72
A.3	Dataset partie 3	73
A.4	Format des données climatiques brutes	74
B.1	HCPC avec trois clusters comparés à la sortie Acidez	77
B.2	Tableau des clusters pour la sortie Puntaje Total	78
B.3	Importance des variables lors de la réalisation des clusters	79

Liste des abréviations

CIAT Centre Internationnal de recherche pour l'Agriculture Tropicale

DTR Diurnal Temperature Range

EPSG European Petroleum Survey Group

FNC Fédération Nationale des Caféculteurs

GIS Geographical Information System

HCPC Hierarchical Clustering on Principal Components

SCAA Specialty Coffee Association of America

SICA Sistema de Información Cafetera

SOM Self Organizing Map

SRID Spatial Reference System Identifier

tmax Température maximale

tmean Température moyenne

tmin Température minimale