Mémoire présenté devant l'ENSAE ParisTech
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 07/02/2018

Par : **Sadi Aoun**

Titre : **Solution dynamique d'assurance paramétrique agricole**

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

M. Nicolas BARADEL

Entreprise : Liberty Specialty Markets
Nom : M. Eric SUGIER
Signature :

Membres présents du jury de l'Institut des Actuaires

M. Pierre LACOSTE
Mme. Florence PICARD
Mme. Maissetou TOURE-COULIBALY

Directeur du mémoire en entreprise :

Nom : M. Jean-Christophe GARAIX
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*
Signature du responsable entreprise

Secrétariat :

Signature du candidat

Bibliothèque :

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Remerciements

Tout d'abord, je souhaite adresser mes remerciements les plus sincères à toute l'équipe de Liberty Specialty Markets de Paris qui m'a offert un accueil des plus chaleureux. Tous ont contribué, sans exception, à faire de ce stage un moment cordial et très enrichissant sur de nombreux aspects tant professionnels que personnels.

En particulier, je remercie Jean-Christophe Garaix pour sa pédagogie, son soutien et la motivation qu'il m'a apportés et avec qui ce fût un réel plaisir de travailler. Je tiens aussi à remercier chaleureusement Abdessamad El Angoudi pour son accompagnement, son professionnalisme mais aussi sa bonne humeur qui ont grandement participés à la réussite de ce stage. Je souhaite également à remercier Salah Dhouib, Hervé Grenier et Françoise Bollotte pour toutes les discussions pertinentes que nous avons eues et les précieux conseils qu'ils m'ont apportés.

Je remercie tout particulièrement Nicolas Baradel, encadrant académique de mon mémoire qui m'a suivi et instruit sur divers projets et matières qu'il enseigne mais aussi pour ses conseils judicieux au sujet de mes choix académiques. J'ai également une pensée particulière pour Julien Randon-Furling qui a éveillé ma passion des mathématiques appliquées à l'Université Paris 1 et qui m'a d'autant plus donné l'envie d'intégrer l'ENSAE ParisTech.

Pour finir, je remercie également ma famille qui m'a toujours soutenu et encouragé dans mon travail et plus particulièrement mon frère, Moujib Aoun, qui m'a fait découvrir ce passionnant univers qu'est l'actuariat et qui a sans aucun doute inspiré ma carrière professionnelle.

Résumé

Ce mémoire a pour vocation de proposer une solution innovante afin de résorber les difficultés financières auxquelles est confronté le marché de l'assurance agricole. En effet, comme nous pourrons le constater, la pluralité des risques climatiques auxquels est exposé l'agriculture incite ses agents économiques à se couvrir contre une multitude de périls dont le prix peut s'avérer élevé relativement au profit que l'activité engendre.

Au travers une étude de marché, nous constaterons ainsi que l'Etat est bien souvent amené à inciter les agriculteurs à se couvrir en finançant une partie de la prime versée à l'assureur. Néanmoins, malgré ces incitations, seule une part minoritaire des agriculteurs souscrit à des couvertures agricoles. De ce constat, nous comprenons qu'il existe un réel enjeu économique à innover ce marché par la commercialisation d'un produit d'assurance offrant une couverture similaire à celles présentes tout en garantissant un prix inférieur.

La majorité des agriculteurs, en particulier en France, souscrivent à des couvertures agricoles dites multi-périls dont l'indemnisation est effectuée sur les pertes réelles. Cette couverture indemnitaire nécessite dès lors une expertise coûteuse, qui peut, dans le cas de contestation d'une des parties, engendrer des coûts supplémentaires et un retard de paiement qui pourrait s'avérer contraignant pour l'agriculteur au vu de son cycle de production annuel. Ces problématiques nous amèneront donc à nous intéresser au marché particulier de l'assurance paramétrique dont l'avantage premier est de réduire fortement les coûts relatifs à la gestion de sinistre tout en garantissant une transparence pour l'assuré comme pour l'assureur, ce qui conduira à une indemnisation immédiate en cas de sinistre.

De par les récentes avancées mathématiques, en particulier en apprentissage statistique, et grâce à de longs processus de collecte et d'analyse de données, l'innovation est aujourd'hui au cœur de l'assurance. Au vu de la difficulté de commercialisation relative à la complexité des indices de références, nous avons été ainsi amenés à modéliser un indice de rendement à l'aide de méthodes statistiques et de données climatiques, agricoles et satellitaires. D'une part, l'objectif de ce mémoire sera de traiter les données convenablement afin de construire par la suite un modèle prédictif de l'indice de rendement et d'autre part développer une méthode de tarification adaptée à la solution proposée. Pour finir, nous ajouterons que cette couverture paramétrique sera d'autant plus innovante du fait qu'une souscription dynamique sera envisagée où la prime variera au fil des stades phénologiques de la culture considérée.

Abstract

The purpose of this thesis is to offer an innovative solution in order to reduce the financial difficulties faced by the agricultural insurance market. Indeed, as we can notice, the plurality of climatic risks to which farmers are exposed encourages economic agents to hedge themselves against a multitude of perils which, could be expensive relatively to the profit that the activity generates.

Through a market study, we will observe that often the state has to encourage farmers to hedge themselves by financing part of the premium paid to the insurer. Nevertheless, despite these incentives, only a minority of farmers subscribe to agricultural coverage. From this observation, we understand that there is a real economic stake to innovate this market by producing a new insurance product offering similar coverage while guaranteeing a lower price.

Most farmers, particularly in France, subscribe to multiple-peril agricultural insurances, the compensation of which are based on actual losses. This indemnity coverage therefore requires costly expertise which, in the event of dispute by one of the parties involved, may also result in additional costs and a delay in payment which could be binding for the farmer regarding his annual production cycle. These issues led us to focus on the parametric insurance market, the main advantage of which is to greatly reduce the costs related to the management of claims while ensuring transparency for both the insured and the insurer which will consequently enable the immediate compensation in case of damages.

Thanks to recent advances in mathematics, particularly in machine learning, and through a long process of data collection and data analysis, innovation is now is at the very heart of the insurance business. Given the inherent basis risk in parametric hedging and the difficulty of marketing these products due to the complexity of their benchmarks, we consequently modelled a yield index using climatic and agricultural data as well as satellite images employing statistical methods. On the one hand, the purpose of this thesis will be to process the data appropriately in order to subsequently build a predictive model of the yield index and on the other hand to develop a pricing method adapted to the proposed solution. Finally, this parametric coverage will be all the more innovative since a dynamic underwriting will be envisioned where the premium will vary over the phenological stages of the underlying crop.

Table des matières

Première partie

Partie I : Gestion du risque agricole et climatique

Cette première partie vise à présenter ce qu'est le risque climatique, thème central du mémoire. Les risques climatiques auxquels nous sommes tous confrontés ont des impacts économiques et environnementaux considérables. Nous nous intéresserons aux nombreux secteurs d'activités de l'économie, sensibles à ces impacts, et dont une couverture en assurance s'avère essentielle, et en particulier au secteur de l'agriculture, qui en est le plus sensible.

Malgré cette forte exposition au risque, le marché de l'assurance agricole est très peu développé. En effet, ce sont généralement les gouvernements qui supportent cette charge via des subventions. L'assurance indemnitaire étant la couverture dominante en assurance agricole, nous présenterons alors son principe de fonctionnement et ses limites. Par ailleurs, nous comparerons les avantages et inconvénients de cette dernière avec l'assurance paramétrique.

Après avoir constaté les limites des couvertures actuellement proposées, nous pourrons alors introduire la solution innovante de couverture agricole qui fait l'objet de ce mémoire. D'une part, nous présenterons l'intérêt de la solution mise en œuvre et d'autre part le fonctionnement de cette dernière. Cette solution étant fondée sur un indice de rendement modélisé, nous serons amenés à présenter les multiples données utilisées dans le cadre de cette étude.

1 Gestion du risque climatique et agricole

1.1 L'exposition au risque climatique

1.1.1 Météo-sensibilité de l'économie

Selon la zone géographique étudiée, les conditions météorologiques environnantes exposent de nombreux secteurs d'activités de l'économie à des risques climatiques. Selon une récente étude [1] de l'ONU [2], environ 30% de l'activité économique mondiale serait exposée au climat et 70% des entreprises seraient météo-sensibles. Les conséquences directes de ces conditions climatiques se répercutent en général par de lourdes pertes financières. Cette dépendance au climat menace de nombreux secteurs de l'économie au point que le gouvernement soit parfois amené à servir des subventions afin d'inciter la souscription de couvertures auprès d'assureurs.

Le secteur agricole figure sans aucun doute parmi les secteurs économiques les plus exposés du fait de sa proximité avec la nature. Il n'en reste pas moins qu'il existe encore bien d'autres activités fortement sensibles tel que le secteur énergétique ou encore touristique. Du fait d'une hausse de température pendant la période hivernale, les stations de ski et les fournisseurs d'électricité souffriront d'une baisse de fréquentation ou de consommation d'énergie. Voici ci-dessous une liste non-exhaustive des différents risques climatiques auxquels sont exposés certains secteurs d'activités :

Secteur d'activité	Risques Climatiques
Energie	Température
Agriculture	Gel, précipitation et température
Tourisme et loisirs	Ensoleillement, enneigement et température
Santé	Hiver très froid et été très chaud
Transports	Vent, pluie, neige, verglas, gel
Bâtiment et travaux publics	Vent, pluie, neige, verglas, gel

FIGURE 1 – Expositions climatiques selon le secteur d'activité

L'agriculture reste tout de même l'un des secteurs d'activités les plus sensible de par la multiplicité des facteurs climatiques. Le risque de gel est un des principaux risques auquel l'agriculture est confrontée. Les précipitations ou températures défavorables peuvent quant à elles provoquer l'apparition de bioagresseurs dévastateurs pour les cultures. La sécheresse, pouvant causée par ailleurs des incendies, affecterait bien entendu le rendement de l'agriculteur. Un agriculteur a donc tout intérêt, s'il le peut financièrement, à se couvrir contre ces risques climatiques pour s'assurer un revenu minimum en cas de sinistre. De plus, il est important de préciser que l'agriculteur est aussi soumis au risque climatique pouvant survenir à l'étranger qui détermine en partie les prix sur les marchés financiers. Il est ainsi commun que les agriculteurs souscrivent à la fois à des contrats d'assurance couvrant une perte de rendement et des contrats à terme couvrant une baisse du prix de marché.

1. Quelles protections face aux aléas climatiques ?
2. Organisation des Nations Unies

1.1.2 Les catastrophes naturelles

Les catastrophes naturelles engendrent de lourdes pertes et peuvent freiner considérablement le développement économique d'un pays, ce qui peut être d'autant plus tragique pour un pays en voie de développement. Leur apparition peut autant amener à la destruction à grande échelle de champs, maisons, immeubles et villages que de nombreuses pertes humaines. Selon un récent rapport [3] de Munich Re, les catastrophes naturelles survenues en 2017, 3ème année la plus dévastatrice, ont engendré 330 milliards de dollars de pertes économiques dont 135 milliards à la charge des assureurs. L'ouragan Irma a été l'événement le plus coûteux pour les assureurs avec 32 milliards de dollars de pertes assurées, l'ouragan Harvey et Maria se plaçant juste après avec une perte estimée à 30 milliards. Les événements extrêmes ont d'autant plus de répercussions que les personnes et les biens sont concentrés dans des zones à risque. En effet, certains phénomènes conjugués entrainent une augmentation de ces pertes en partie liée à la concentration croissante des populations et des richesses dans des zones exposées à ces événements extrêmes où l'activité humaine détériore toujours d'avantage l'environnement.

En outre, ces catastrophes naturelles peuvent prendre des formes diverses et variées selon leur origine, qu'elles soient climatiques (gel, tempête, sécheresse, ouragan, etc) ou géologiques (tremblement de terre, éruptions volcaniques, etc). Selon l'origine considérée, ce risque de catastrophe naturelle complexifie grandement le processus de modélisation statistique contrairement aux risques communs. En effet, sa faible fréquence et sa forte sévérité en font un risque particulier qui incitera l'assureur à prendre d'autant plus de précautions lors de la modélisation de ce risque, la survenance d'un tel évènement pouvant amener à la faillite de certains acteurs du marché. Pour mieux illustrer l'impact financier des risques extrêmes et la participation des assureurs, nous présenterons par la suite les indemnités versées par ces derniers lors de la survenance des évènements les plus importants de ces trente dernières années.

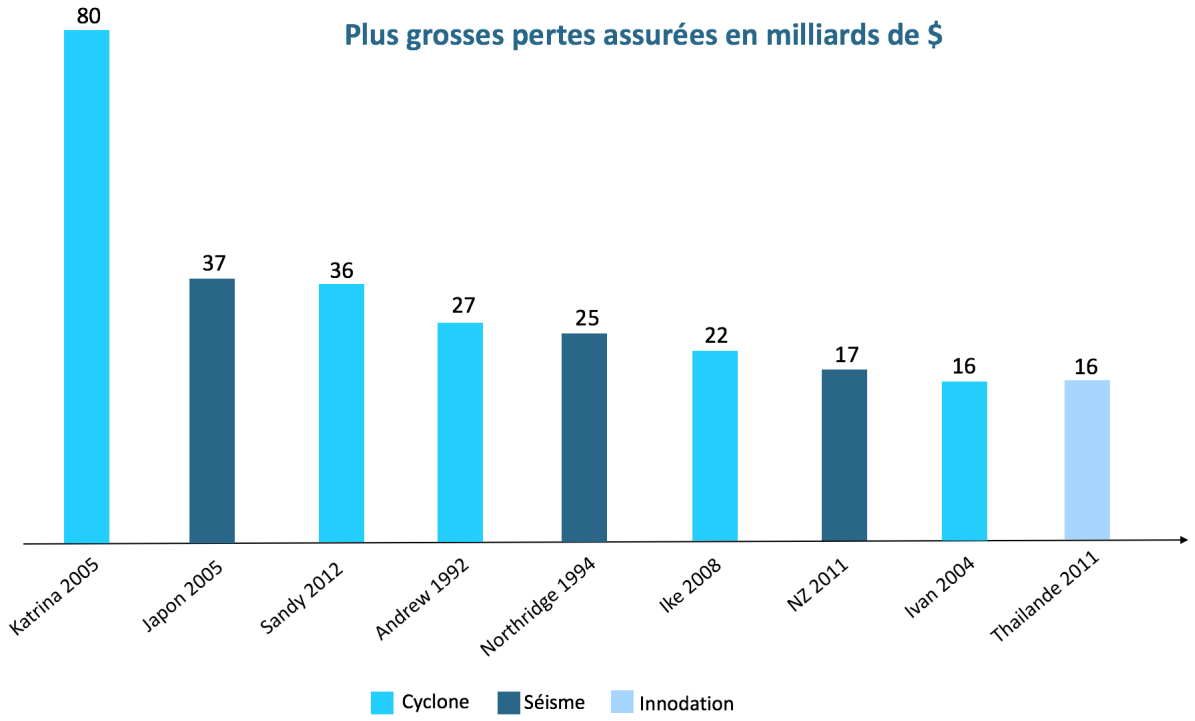

FIGURE 2 – Indemnisations versées par le secteur de l'assurance depuis 1992

3. Munich Re puts 2017 incured losses at $135bn

Comme nous pourrons le constater par la liste [4] des principaux événements extrêmes recensés par la FFA [5] depuis 1988, la France est régulièrement touchée par des catastrophes naturelles, et en particulier par des inondations, tempêtes, orages de grêles, particulièrement dévastatrices pour le secteur agricole :

Principaux événements naturels depuis 1988 (Cat-Nat + TGN)

(vu au 30 mai 2017)

Régime d'assurance	Exercice de survenance	Désignation de l'événement	Coût marché national en M€	Coût marché national en M€ constants (1) (indice FFB)
Cat-Nat	2016	Inondations bassins Seine et Loire/mai-juin	1 420	**1 420**
Cat-Nat	2015	Inondations et crue éclair dans le SE/3 oct.	600	**602**
TGN	2014	Orages de grêle France/8-10 juin 2014	850	**857**
TGN	2013	Orages de grêle France/été 2013	850	**866**
Cat-Nat	2011	Subsidence	800	**852**
Cat-Nat	2010	Inondations du Var/juin	615	**686**
Cat-Nat	2010	Inondations tempête Xynthia/février	745	**830**
TGN	2010	Tempête Xynthia/février	735	**819**
TGN	2009	Tempête Klaus/janvier	1 680	**1 942**
Cat-Nat	2003	Subsidence	1 300	**2 071**
Cat-Nat	2003	Inondations du Rhône/décembre	670	**977**
Cat-Nat	2002	Inondations du Gard/septembre	700	**1 059**
TGN	1999	Tempête Lothar et Martin/février	6 860	**11 421**
Cat-Nat	1998	Subsidence	320	**541**
Cat-Nat	1996	Subsidence	360	**629**
Cat-Nat	1995	Inondations du Nord/janvier-février	360	**645**
Cat-Nat	1991	Subsidence	250	**512**
TGN	1990	Tempête Daria/février	1 315	**2 819**
Cat-Nat	1990	Subsidence	355	**761**
Cat-Nat	1989	Subsidence	230	**506**
Cat-Nat	1988	Inondations Nîmes/octobre	290	**665**

(1) Coût en euros constant revalorisés par l'indice FFB à la fin 2016 *Sources : FFA et CCR*

FIGURE 3 – Principales catastrophes naturelles climatiques en France depuis 1988

La définition précise des catastrophes naturelles dépend du cadre juridique considéré et donc du pays concerné. En France [6], sont considérées comme effets des catastrophes naturelles, "les dommages matériels directs non assurables ayant eu pour cause déterminante l'intensité anormale d'un agent naturel lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises". La garantie « catastrophes naturelles » offre une couverture contre les dommages causés par inondation, glissement de terrain, avalanche, tremblement de terre, etc. tout en couvrant les bâtiments, le matériel, les véhicules, le bétail et les récoltes engrangées. La garantie « catastrophe naturelle » se déclenche seulement si un arrêté interministériel paru au Journal officiel constate l'état de catastrophe naturelle.

4. Données clés 2016, annexe (cf. page 47)
5. Fédération Française de l'Assurance
6. Article L125-1 du Code des Assurances

1.2 Le marché de l'assurance agricole

1.2.1 L'attractivité des couvertures

Pour mesurer l'attractivité d'un produit d'assurance pour une ligne d'affaire bien précise, il est commun de se référer à une statistique de marché appelée « taux de pénétration ». Ce taux, qui correspond au rapport entre le nombre d'assurés et le nombre d'agents économiques présents sur un marché considéré, permet d'apprécier l'utilité des couvertures proposées. Nous illustrerons la situation [7] du marché Européen de l'assurance agricole en présentant les taux de pénétration de ces pays pour divers risques sous-jacents à une couverture agricole :

	Allemagne	Autriche	Belgique	Danemark	Espagne	Finlande	France	Grèce	Italie	Norvège	Pays-Bas	Pologne	Portugal	Suède	Suisse	Turquie	Tchéquie	Royaume-Uni
tempête	O	O	C[1]	O	P	O	C	S	N	P	O	O	O	O	C	O	O	O
cyclone	O	O	C[1]	O	P	O	C	S	N	P	O	N	O	O	C	O	N	O
inondation	S	O	C[1]	N	P	O	C	S	O	P	N	O	O	O	C	O	O	O
grêle	O	O	O	O	O	O	O	S	O	S	O	O	C[2]	O	C	S	O	O
glissement de terrain	S	O	C[1]	O	S	O	C	S	O	P	S	O	O	O	C	O	O	O
neige	S	O	O	O	O	O	O	S	O	N	O	O	N	O	C	O	O	O
gel	O	O	O	O	O	N	O	S	O	O	O	O	N	O	O	N	O	O
avalanche	S	O	N	N	O	O	C	N	O	P	N	O	N	O	C	N	O	N
sécheresse	N	O	N	N	S	N	C	N	N	N	N	N	N	O	N	N	S	N
affaissement de terrain	S	O	C	N	S	N	C	S	N	N	N	O	O	O	N	O	O	O
séisme	O	N	C[1]		P		C	O	N	P	N	O	O	O	C	C	O	O
incendie de forêt	O	S	N	O	S	O	S	S	S	N	O	O	O	O	N	S	N	N
éruption volcanique	O	N	N		P		C	O	N	P	N	O	O	O			O	N
foudre	O	O	O		O	O	O	O	O	O	O			O		O	O	O

[1] *seulement pour les risques individuels simples*
[2] *seulement si la grêle résulte d'une tempête*

Type de couverture d'assurance		Taux de pénétration de la couverture
C	*couverture obligatoire par la loi*	> 75%
P	*pool obligatoire*	25-75%
O	*couverture optionnelle*	10-25%
S	*couverture disponible, mais peu souscrite*	< 10 %
N	*couverture non existante*	*non connu*

FIGURE 4 – Taux de pénétration du marché Européen en 2009

Au vu du faible taux de pénétration des couvertures agricoles dans le monde, le gouvernement intervient et aide les agriculteurs financièrement. Chaque pays impose ses lois et peut rendre obligatoire la souscription de certaines couvertures. Dans certains pays, le gouvernement prend à sa charge les sinistres de catastrophes naturelles, dans d'autres, il propose des subventions pour certains risques à couvrir. De nombreux gouvernements prennent ainsi en charge une partie de la prime versée à l'assureur afin d'inciter les agriculteurs à souscrire à des contrats d'assurance, la quote-part [8] pouvant aller jusque 65% en France. Malgré tout, la prime relative à la couverture est encore trop élevée à tel point que ces contrats d'assurance ont du mal à trouver preneur, le rendement agricole étant corrélé à une multitude de risques d'ordre climatique, les prix de ces contrats présentent un prix élevé que les agriculteurs ne peuvent payer en intégralité. Selon un récent rapport [9] émanant du ministère de l'agriculture en 2016, seulement 25.7% de la superficie agricole hors prairie était assurée contre l'aléa climatique.

7. Couverture d'assurance des catastrophes naturelles en Europe, modifié de Mills 2009 (cf. Figure 3)
8. Indemnistation des pertes occasionnées par des aléas climatiques
9. La gestion des risques en agriculture

L'APCA [10] tente d'alerter les agriculteurs de l'importance des contrats d'assurance avec des publications et des notes d'informations [11] dans lesquelles elle explique :" Parmi tous les aléas auxquels sont confrontés les agriculteurs, le climat est l'un des plus difficiles à maîtriser. Chaque année se produisent des accidents climatiques qui, localement voire dans un département ou une région, provoquent des pertes qui peuvent atteindre des montants considérables et menacer l'équilibre économique des exploitations sinistrées". Or trop peu d'agriculteurs en France souscrivent à des contrats d'assurance agricole, ou bien ne se couvrent que partiellement. Il est donc nécessaire de trouver des solutions à ce problème, néanmoins, selon une étude publiée par RFIB, les couvertures agricoles ont plus que quadruplé de volume depuis le début du siècle. Comme nous avons pu le voir, l'évolution de ce marché semble être positivement corrélée à l'accroissement des catastrophes naturelles, des subventions étatiques et de la communication des associations d'agriculteurs :

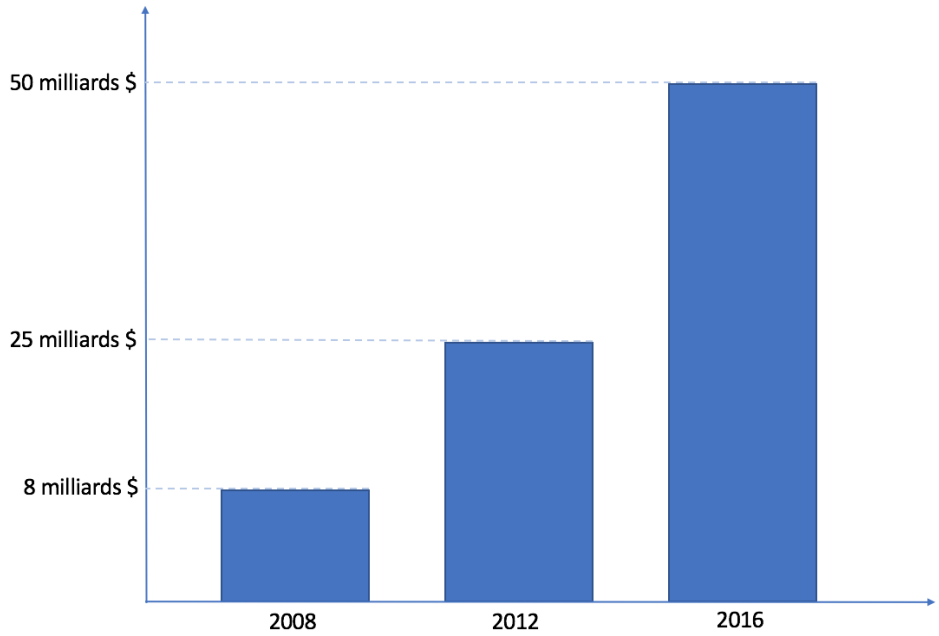

FIGURE 5 – Evolution du volume de primes agricoles mondiales depuis 2008

Nous ajouterons que l'APREF [12] précise par une note [13] que "le développement des garanties socle qui sont en cours de discussion dans le marché, est un facteur positif de développement du marché. Une question importante demeure celle du soutien futur que l'Etat souhaite apporter à l'assurance récoltes dans le cadre de la nouvelle PAC 2014-2020. Une aide mal calibrée, non en ligne avec ce qui se pratique dans les autres pays, constituerait selon nous un frein au développement serein du marché, développement qui bénéficierait pourtant à tous les acteurs, compte tenu de l'ampleur de l'impact potentiel d'un évènement climatique majeur pour l'ensemble de la ferme France".

10. Assemblée Permanente des Chambres d'Agriculture
11. La gestion des crises en Chambre d'agriculture(cf. page 4)
12. Association des professionnels de la réassurance en France
13. Développement du Marché Assurance et Réassurance Récoltes en France,(cf. page 3)

1.2.2 Evolution des rendements agricoles

Du fait des progrès techniques, en général liés aux recherches scientifiques ou industrielles, le marché de la couverture agricole porte une dimension tendancielle qui complexifie la modélisation du rendement agricole. En particulier, de par les récentes innovations tant à l'échelle des machines agricoles qu'au niveau des variétés semées, on observe, quelle que soit la culture considérée, une amélioration des rendements d'une année sur l'autre. Le rendement historique d'un agriculteur sera sans aucun doute bien inférieur au rendement que ce même agriculteur serait capable de produire aujourd'hui. Pour mieux illustrer l'innovation au coeur du secteur agricole, nous présenterons les données [14] de rendement français de multiples cultures de 1862 à 2007 collectées auprès du service public de statistiques ministérielles AGRESTE [15] :

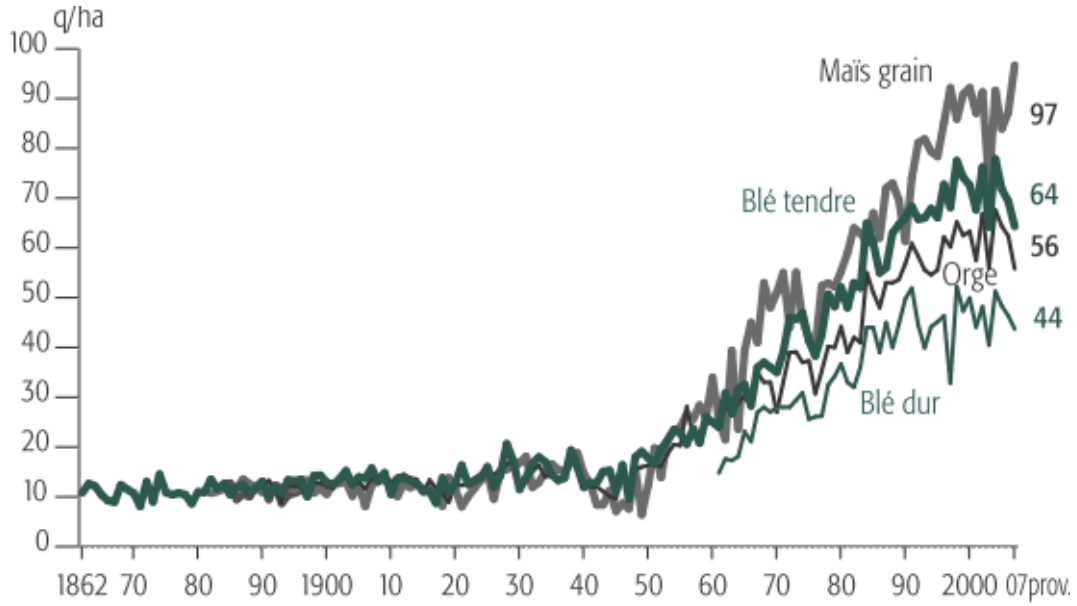

FIGURE 6 – Evolutions des rendements Français de 1862 à 2007

Il est intéressant de souligner que la croissance des rendements s'est fortement accélérée dès la fin de la seconde guerre mondiale suite aux innovations majeures développées lors de cette guerre et des échanges commerciaux qui s'en sont suivis. Encouragé par le plan Marshall, la France a ainsi importé des Etats-Unis les engrais et les machines agricoles performantes qui lui ont permis d'accroître significativement le rendement de ses agriculteurs dès la fin de la guerre.

14. Agreste Primeur(cf. page 2)

15. Service de la statistique et de la prospective du Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt

1.3 Le marché de la couverture

1.3.1 L'assurance indemnitaire

Le principe de l'assurance indemnitaire porte sur les dommages causés sur les biens ou les dommages corporels. Il s'agit d'un fondement de base de l'assurance, qui consiste à ce qu'en cas de dommage matériel ou corporel subi par l'assuré, elle indemnise la victime :

— en fonction du préjudice subi
— dans les limites des garanties souscrites
— sans contribuer à l'enrichissement de l'assuré

L'assuré est indemnisé en fonction du préjudice subi qui nécessitera une expertise bien particulière selon la nature du sous-jacent. Qu'il s'agisse d'un sous-jacent agricole ou automobile, un expert tel qu'un garagiste devra être dépêché afin d'estimer le montant du préjudice. En résumé, la prestation d'indemnisation, par définition de l'assurance indemnitaire, est calculée en fonction du préjudice subi, et non sur la base d'un capital au montant préétabli. La particularité de la souscription d'une assurance indemnitaire, si elle est bien structurée, est que l'assuré observera une indemnisation fortement corrélée au préjudice subi. Dans le cas d'une assurance agricole, l'agriculteur recevra une indemnisation correspondant à la perte de rendement observée uniquement si l'un des risques climatiques couvert par le contrat se réalise. Son indemnisation sera donc en toute logique corrélée aux pertes réelles.

Le risque de grêle fût l'une des premières couvertures indemnitaires proposée aux agriculteurs dès la fin du XIX^{e} siècle. Néanmoins, au vu des différents risques climatiques auxquels est exposé le secteur agricole, il est commun pour les agriculteurs de souscrire à des contrats d'assurance indemnitaire dits multi-périls sur récoltes ou assurance aléa climatique. Commercialisée au début de ce siècle, cette solution de couverture a pour but de couvrir la perte de rendements des agriculteurs causée par des risques climatiques variés dont les principaux aléas sont la grêle, la tempête, la sécheresse, le gel et les inondations. Le récent essor du volume de données disponibles explique en partie l'apparition de ces nouveaux contrats dont la modélisation statistique s'avère particulièrement délicate. En effet, il est primordial pour l'assureur d'étudier la corrélation de ces différents risques afin d'appréhender au mieux le risque dans sa globalité. Cette complexité est d'autant plus accentuée du fait que les données récoltées manquent de fiabilité et de précision, ce qui rend délicate la modélisation statistique qui en découle. L'APREF [16] le fait remarquer à travers ce commentaire [17] :"Les assureurs et réassureurs ont en effet besoin des séries de rendements par culture et par zone géographique homogène pour construire leurs tarifs et évaluer les engagements maximum potentiels. Par « zone homogène », on entend une région géographique dans laquelle les paramètres de sols, de climat et de pratiques culturales sont similaires. Il faudrait donc idéalement avoir accès au niveau le plus fin possible (par code postal ou par défaut par canton ou arrondissement) aux données de superficies cultivées et à la production récoltée.". Cette difficulté d'évaluation des engagements sous-jacents au contrat incite les acteurs du marché à faire preuve de solutions innovantes.

16. Association des professionnels de la réassurance en France
17. Développement du Marché Assurance et Réassurance Récoltes en France (cf. page 10)

Malgré toutes les révolutions technologiques et informatiques de ces dernières années, la gestion des sinistres en assurance peine à être automatisée et reste une charge importante pour le secteur. Lorsqu'un sinistre survient lors d'une couverture indemnitaire, plusieurs étapes successives s'enchaînent avant de procéder au versement d'une indemnisation. Une fois que l'assuré a déclaré son sinistre dans le délais imparti, l'assurance doit évaluer l'ampleur des dommages causés par un état des lieux effectué par un expert. Cet expert, en contact avec le client, détermine le coût du préjudice et le degré de responsabilité du client afin de calculer le montant de la prise en charge à laquelle l'assuré aura le droit. Ce processus, généralement très long et couteux, à la charge de l'assureur, est forcément pris en compte lors de la tarification de la prime d'assurance. De plus, dans un contexte d'assurance agricole, ce type de contrat n'est pas tellement adapté car l'agriculteur ne peut se permettre d'attendre trop longtemps l'indemnisation de son préjudice. Du fait que l'agriculteur a un cycle de production annuel bien souvent autofinancé, il y a une certaine urgence à l'indemniser au plus vite en cas de sinistre.

De plus, la fraude à l'assurance, notion dont il n'existe pas de définition légale en France, est un autre inconvénient qui subsiste toujours dans ce type de gestion de sinistre. L'Alfa [18] apporte toutefois la définition [19] – avec l'aval des praticiens – qu'il y a fraude à l'assurance chaque fois qu'un « acte volontaire permettant de tirer un profit illégitime d'un contrat d'assurance » a été identifié. Frauder suppose donc un acte volontaire et malintentionné. La fraude peut être présente tout au long de la vie d'un contrat, de la souscription à la déclaration du sinistre, et dans toutes les branches de l'assurance (automobile, agriculture...). Selon l'Argus de l'assurance les 46 255 fraudes recensées [20] en 2015 par l'Alfa représentent 265 millions d'euros d'économies réalisées. Par ailleurs, il arrive que l'expert en charge d'estimer le préjudice soit également complice de la fraude.

1.3.2 L'assurance paramétrique

Depuis le début du siècle, nous assistons au développement d'un nouveau marché dans le secteur de l'assurance, communément appelé assurance indicielle ou paramétrique, qui a pour particularité de considérer des indices observables à tout instant et par chacune des parties. Dans un contrat d'assurance paramétrique, l'indemnisation est déclenchée et calculée sur la seule base de l'indice de référence retenu par l'assuré. On y opposera ainsi l'incontournable assurance indemnitaire qui consiste à indemniser la survenance de sinistre sur la base de pertes réelles. Les indices de référence, principalement climatiques, sont généralement retenus en fonction de la fiabilité et de l'historique disponible des données que l'on peut obtenir auprès d'une entité centrale garante des données comme par exemple une station météo. Le développement de la plante étant fortement corrélé aux conditions climatiques environnantes, il est aisé pour un assureur d'observer l'apparition d'un sinistre uniquement à partir des données météorologiques. Si l'on considère par exemple le cas d'une couverture agricole où l'agriculteur souhaite se prémunir d'une éventuelle perte de rendement suite à une sécheresse, il peut être intéressant de construire une couverture où l'indemnisation sera déclenchée si le cumul des pluviométries durant la période à risque ne dépasse pas un certain niveau fixé contractuellement, ce qui est équivalent au prix d'exercice d'une option financière. Avant de souscrire à une éventuelle couverture indicielle, il est alors fondamental de procéder à une analyse approfondie de la corrélation entre le risque sous-jacent considéré et l'indice de référence retenu :

18. Agence pour la lutte contre la fraude à l'assurance
19. Site ALFA (cf. La fraude)
20. Lutte contre la fraude : 265M euros récupérés par les assureurs en 2015

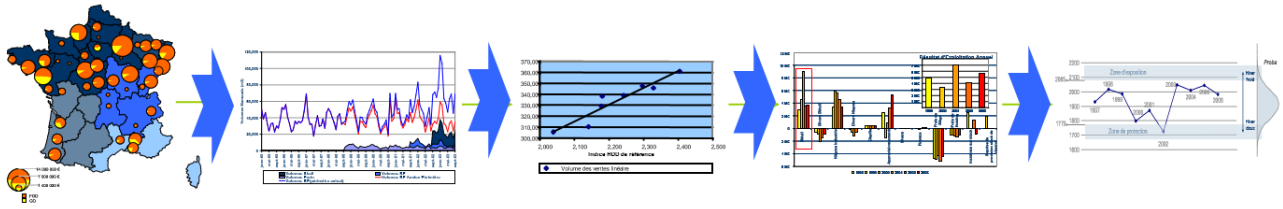

FIGURE 7 – Analyse de la corrélation entre l'indice de référence et le risque sous-jacent

Cet attrait pour les couvertures paramétriques s'explique en partie du fait que l'assuré comme l'assureur peut observer objectivement la survenance ou non d'un sinistre ce qui conduit à une indemnisation plus transparente. En effet, comme nous l'avons vu précédemment lors de la souscription d'une couverture indemnitaire, l'inconvénient est qu'il est nécessaire d'attendre l'évaluation des pertes par un expert avant de procéder à l'indemnisation. Selon le risque sous-jacent au contrat, cette évaluation des pertes implique une gestion de dossier après sinistre qui est un processus coûteux plus ou moins long à la charge de l'assureur qui le répercutera bien entendu sur la prime commerciale dont doit s'acquitter l'assuré. Ce procédé permet à la fois de réduire les frais résultant d'un long processus de gestion et les coûts susceptibles de survenir au vu du risque judiciaire. Ce gain en temps de gestion se répercute alors sur le chargement appliqué au prix du contrat. Cette réduction des coûts dans la construction d'un produit est une étape essentielle afin de rendre sa commercialisation la plus efficiente possible.

L'évaluation même des pertes par l'expert nécessite une certaine transparence afin d'éviter les possibles conflits juridiques résultant d'une contestation d'un signataire du contrat. L'assurance indicielle implique une transparence de l'ensemble des parties, ce qui réduit fortement le risque de contestation, à condition de la bonne tenue de l'entité garante des données. Il est par ailleurs aujourd'hui possible de gérer l'ensemble des transactions d'un produit paramétrique, que ce soit l'achat d'une couverture ou une indemnisation, à l'aide d'un système automatisé, décentralisé et encore plus transparent appelé « block-chain ».

Néanmoins, lors d'une solution d'assurance indicielle, l'exposition au risque est déterminée par la corrélation entre un évènement climatique et son impact sur l'activité d'un agent économique. Or, l'activité d'une entité est en général exposée à tout un ensemble d'évènements climatiques. La modélisation de cet ensemble d'évènements à risque peut parfois nécessiter de multiples indices, ce qui peut rendre d'autant plus complexe le procédé de couverture indicielle. Dans l'exemple précédent, il est possible que l'agriculteur soit confronté à un risque de gel qui nécessitera de considérer un cumul de température minimale en plus de l'indice sécheresse construit. L'assurance paramétrique peut ainsi entraîner un écart entre le montant réel des préjudices subis et l'indemnisation calculée sur la base de l'indice retenu si la modélisation venait à être défaillante, c'est ce que l'on appelle le risque de base.

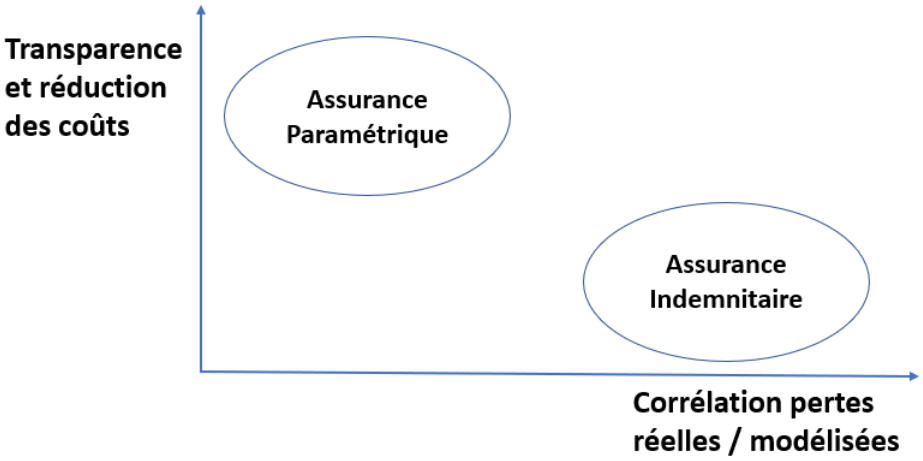

FIGURE 8 – Limites de l'assurance paramétrique et indemnitaire

La commercialisation d'un produit d'assurance paramétrique peut s'avérer parfois compliquée si l'indice de référence n'est pas compris par le client. Il est pourtant possible de considérer des indices climatiques ou satellitaires qui peuvent parfois être significativement corrélés à l'activité. De plus, il peut s'avérer difficile pour un assuré de convenir d'un seuil à partir duquel il souhaite être couvert, en particulier si ce dernier s'avère complexe.

2 Présentation de la solution indicielle

2.1 Fonctionnement du produit d'assurance

2.1.1 Intérêts et caractéristiques de la couverture

Comme nous avons pu le voir, le caractère particulier du marché de l'assurance agricole conduit à des prix élevés qui incitent les assureurs à porter une attention particulière à la tarification des contrats. En outre, il est aussi important pour eux d'innover ce marché par l'apport de nouveaux produits présentant un prix inférieur aux couvertures multi-périls déjà présentes, tout en comportant un spectre de couverture tout aussi large. Dans une telle problématique de réduction des prix, l'assurance indemnitaire étant à ce jour très chère, il semble alors judicieux de se rediriger vers le développement d'un produit alternatif reposant sur de l'assurance paramétrique. Comme évoqué précédemment, cette approche paramétrique, plus économe et transparente, permettra au produit d'assurance d'avoir en premier lieu un prix réduit, ce qui améliorera en conséquence le taux de pénétration.

Néanmoins, la capacité de couverture devant être similaire à celle d'une couverture multi-périls, il apparait évident que nous serons amenés à considérer une multitude d'indices pertinents afin de modéliser convenablement les différents risques couverts par l'assurance multi-périls. Qu'il s'agisse du risque de sécheresse, de gel ou encore d'inondation, la modélisation statistique qui va s'en suivre devra alors dans un premier temps se fonder sur des indices suffisament représentatifs de ces évènements extrêmes. Pour ce faire, nous collecterons des données climatiques, agricoles et satellitaires. Une couverture multi-indicielles présentera potentiellement un prix légèrement inférieur à la couverture multi-périls dont l'écart de prix risque d'être peu significatif. Ainsi, on comprend que la couverture paramétrique multi-indicielle sera encore sans doute trop chère pour capter les parts de marchés restantes.

Enfin, la recherche effectuée sur les phénomènes physiques régissant le sous-jacent se complexifie au fil du temps, ce qui se répercute également sur les produits. Au-delà de la croissance continue des données, les récentes découvertes technologiques ont par exemple permis de développer des indices relativement complexes tel que des indices satellitaires dont on verra plus en détails le fonctionnement par la suite. Ces recherches amènent ainsi naturellement à une complexification des produits qui rendent leur compréhension d'autant plus difficile pour le potentiel assuré. De ce fait, la complexité d'une couverture multi-indicielle peut être un frein à la commercialisation de ce dernier si les potentiels assurés n'en comprennent pas l'intérêt ni même les sous-jacents.

De ce constat, la motivation a été de construire un produit d'assurance paramétrique reposant sur un unique indice, similaire au sous-jacent d'un contrat indemnitaire, comportant l'ensemble des caractéristiques d'une couverture multi-indicielle. Cet exercice de modélisation, essentiellement fondé sur des statistiques, a donc pour objectif de répliquer le rendement d'un agriculteur à partir de multiples données afin de construire ce que l'on appellera l'indice de rendement potentiel. C'est sur la base de cet indice de rendement potentiel, considéré comme réel à la fin du cycle de production, que l'agriculteur souscrira à une couverture de perte de rendement. En effet, l'usage d'un tel indice permet de contourner la complexité relative à la commercialisation d'une couverture multi-indicielle. La souscription de couvertures paramétriques reposant sur un tel indice modélisé permet à quiconque de souscrire une garantie sans connaître les composantes et les phénomènes régissant le sous-jacent du contrat.

FIGURE 9 – Modélisation de l'indice de rendement potentiel

Néanmoins, comme n'importe quelle couverture indicielle, l'indemnisation comporte intrinsèquement un risque de base, du fait que le rendement réel peut différer de l'indice de rendement potentiel. Afin de limiter ce risque, l'assureur est incité à modéliser l'indice ainsi construit par une formule explicite dite fermée afin de garantir à tous les agents une compréhension réelle des déterminants. La modélisation indicielle se doit donc de conserver la transparence d'un produit d'assurance paramétrique classique. La construction de modèles non-paramétriques complexes faisant office de « boîte noire », tel que des réseaux de neurones artificiels, est donc à proscrire dans un cadre de commercialisation de produit d'assurance régulé tel que celui-ci :

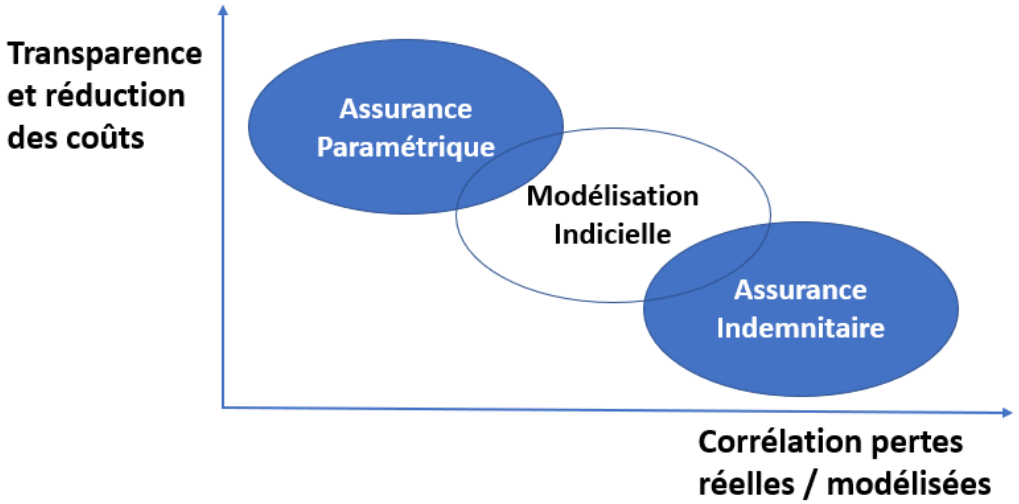

FIGURE 10 – Limites de la modélisation indicielle

2.1.2 Couverture de perte de revenu

Lors de la souscription d'une couverture paramétrique, une valeur unitaire P^Y propre à l'indice de référence retenu est définie. Lors de la souscription d'une couverture de prix, cette dernière est égale à 1 alors que dans le cas d'une couverture agricole ou climatique, il est nécessaire de définir une valeur monétaire à la perte engendrée par la variation. En effet, l'indice pouvant être en degré celsius ou en tonne par hectare, il est nécessaire de définir contractuellement l'indemnisation versée à chaque variation de l'indice. La particularité de ce produit sera donc de considérer la valeur unitaire P^Y fixe durant toute la période de risque, celle-ci étant assimilable au coût de production relatif à la production d'une tonne de la culture considérée. Ce prix incorpore donc par exemple les coûts relatifs à l'eau, aux graines semées, etc. Cette valeur variera ainsi d'un contrat à l'autre uniquement si la culture change, le maïs étant par exemple moins consommateur en eau que le blé. Nous présenterons par la suite le fonctionnement du produit dans le cas d'une culture de blé tendre d'hiver [21] où la valeur unitaire P^Y sera fixée à 120 euros par tonne.

L'intérêt de ce produit est donc de couvrir le revenu minimum d'un agriculteur lié à une perte de rendement au niveau de l'exploitation. Nous serons donc amenés à couvrir son revenu R correspondant au produit de son rendement potentiel à maturité Y_T, exprimé en tonne par hectare, et du prix d'une tonne associé à la culture semée :

$$R = P^Y * Y_T \tag{1}$$

La modélisation indicielle s'effectuant à l'échelle de la parcelle de l'agriculteur et la garantie à l'échelle de l'exploitation, il est important de préciser qu'un rendement au niveau de l'exploitation sera alors calculé à partir d'une pondération de la superficie $(S_i)_{i\in[|1,N|]}$ des N parcelles associées. Nous pourrons alors exprimer le rendement potentiel de l'exploitation en fonction du rendement potentiel $(Y_i)_{i\in[|1,N|]}$ des N parcelles :

$$Y = \sum_{i\in[|1,N|]} Y_i \frac{S_i}{S_N} = \frac{1}{S_N} \sum_{i\in[|1,N|]} Y_i S_i \quad \text{où} \quad S_N = \sum_{i\in[|1,N|]} S_i \tag{2}$$

Lors de la souscription d'une couverture, l'assuré sera amené à définir une garantie à partir de laquelle il souhaitera être indemnisé en cas de sinistre. Selon le risque auquel est exposé l'assuré, cette garantie s'inscrira dans une couverture contre un risque de hausse ou à l'inverse de baisse. Une fois les caractéristiques du produit définies, il est alors possible de construire une fonction de perte $\mathcal{F}^L(.)$ qui s'exprime en fonction de l'indice à maturité et de la garantie souscrite Y^L :

$$\mathcal{F}^L : \left| \begin{array}{rcl} \mathcal{L}(\Omega, \mathcal{Y}) * \mathcal{Y} & \longrightarrow & \mathcal{L}(\Omega, \mathbb{R}_+) \\ (Y, Y^L) & \longmapsto & \mathcal{F}^L(Y, Y^L) = L \end{array} \right. \tag{3}$$

Dans un tel cadre, il apparait naturel de considérer des couvertures à la baisse où pour un niveau de garantie choisi par l'assuré, l'assureur fera face à une potentielle perte L. Il est important de souligner que le niveau de garantie sera limité par une garantie maximum Y^ϵ fonction de l'indice de rendement auquel est retiré une valeur ϵ_Y fixée à une tonne par hectare. Une fois cette valeur de garantie définie, la somme assurée est ainsi définie comme la perte potentielle à venir résultant d'une couverture à la baisse sur l'indice de rendement potentiel à maturité :

$$L = \mathcal{F}^L(Y, Y^L) = P^Y(Y^L - Y)\mathbb{1}_{Y\leq Y^L} \quad \text{où} \quad Y^L \leq Y^\epsilon \quad \text{et} \quad Y^\epsilon = Y - \epsilon_Y \tag{4}$$

21. Blé tendre d'hiver

2.1.3 La couverture dynamique

Le rendement agricole a la particularité d'évoluer en fonction des conditions climatiques présentes sur l'ensemble du cycle de production. Cette remarque importante nous a donc conduit à modéliser l'indice de rendement à tout instant en actualisant les données utilisées. Sur la base de cet indice de rendement agrégé au niveau de l'exploitation et d'un prix fixe à la tonne propre à la culture, l'agriculteur pourra alors souscrire une couverture à n'importe quelle période du cycle. Les périodes sélectionnées afin d'opérer la discrétisation correspondent aux stades phénologiques [22] successifs comme l'épiaison, la montaison ou encore la maturation. Néanmoins, il est important de souligner que les dates de début et de fin de chacun de ces stades dépendent de la culture étudiée. Ainsi, nous considèrerons l'ensemble de ces périodes $\mathcal{T} = [|1, ..., T|]$ où T correspond à la période de récolte, équivalent à la maturité du contrat.

Dans un cadre de souscription dynamique, l'assuré pourra alors décider de souscrire à un complément de couverture P_t^Y à n'importe quel stade phénologique $t \in [|1, T-1|]$ du cycle de production à condition que le cumul des couvertures souscrites n'excèdent pas la somme assurée maximale définie par P^Y. Ainsi, dans un tel cadre, l'assureur fera face à une perte potentielle dont l'indemnisation sera réalisée en fonction du niveau de l'indice de rendement potentiel effectif en fin de cycle de production Y_T. Nous rappelons que le niveau de la garantie Y_t^L sélectionné par l'agriculteur est limité par l'indice de rendement potentiel Y_t. Cette particularité nous amènera donc à définir notre perte potentielle L_t de la manière suivante :

$$L_t = P_t^Y (Y_t^L - Y_T)\mathbb{1}_{Y_T \leq Y_t^L} \quad \text{où} \quad Y_t^L \leq Y_t^\epsilon = Y_t - \epsilon_Y \quad \text{et} \quad \sum_{t \in [|1,T-1|]} P_t^Y \leq P^Y \tag{5}$$

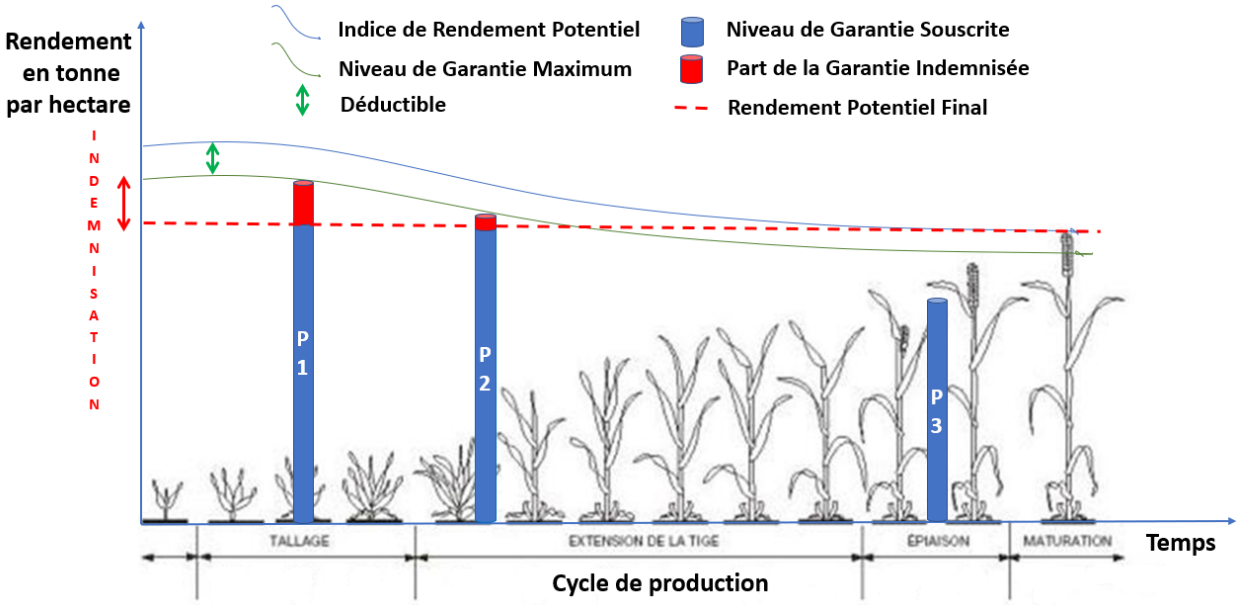

FIGURE 11 – Fonctionnement du produit d'assurance

A maturité, l'indice de rendement potentiel, considéré comme réel, est alors comparé aux multiples couvertures souscrites $(P_t^Y, Y_t^L)_{t \in [|1,T-1|]}$ afin de constituer l'indemnisation éventuelle :

$$L = \sum_{t \in [|1,T-1|]} L_t = \sum_{t \in [|1,T-1|]} P_t^Y (Y_t^L - Y_T)\mathbb{1}_{Y_T \leq Y_t^L} \tag{6}$$

22. BBCH : une échelle universelle pour identifier le stade des cultures

2.2 Présentation de la base de données

2.2.1 Présentation de la région étudiée

L'objectif premier dans l'élaboration d'un tel produit est de construire un modèle statistique afin de prédire le rendement d'un agriculteur pour une culture donnée. Plusieurs méthodes d'apprentissage statistique existent parmi lesquelles l'apprentissage supervisé. Contrairement à d'autres disciplines qui ne nécessiterait par exemple aucune donnée comme l'apprentissage par renforcement, nous serons ici amenés à construire une base de données historiques $\mathcal{D}_n$ représentative de la relation aléatoire à modéliser. Cette base contiendra ainsi la variable à expliquer, le *rendement* à l'échelle de la parcelle, la *variété* utilisée, la *géolocalisation satellitaire* ainsi que le *type de sol* présent. Ces informations nous permettront par la suite de collecter des données complémentaires essentielles à l'étude statistique qui s'en suit. Collectée auprès de la coopérative à couvrir, nous disposerons ainsi initialement des données historiques de multiples agriculteurs situés dans l'Est de la France :

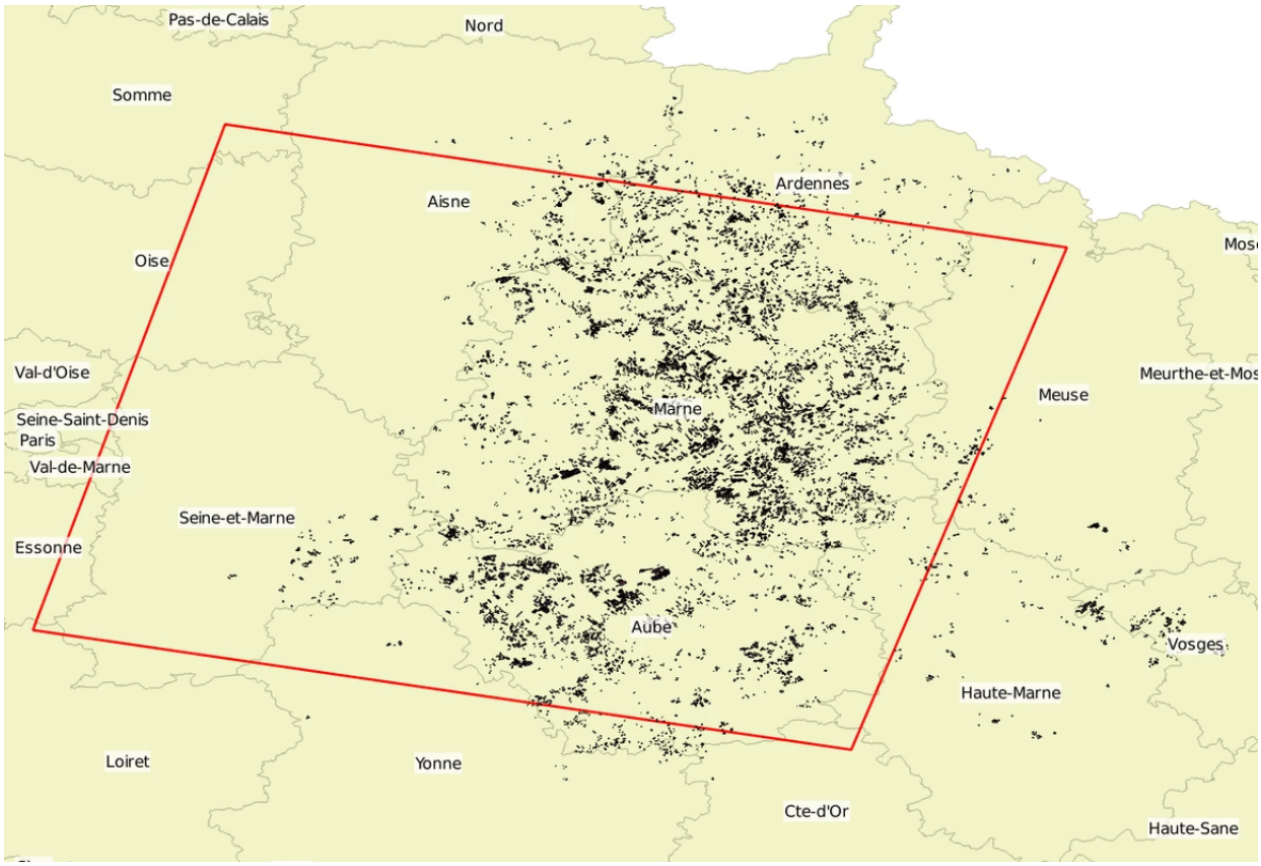

FIGURE 12 – Parcelles agricoles de la zone d'étude

Cette base de données d'environ $n = 50\,000$ observations s'étend sur une plage de temps d'une vingtaine d'années. Au-delà du volume d'observations à disposition, il est aussi crucial dans un tel cadre de disposer d'un nombre important d'années afin de tenir compte des évènements climatiques passés. Comme nous le verrons par la suite, nous devrons prendre d'autant plus de précaution car la région considérée dans l'étude est confrontée à un risque de précipitation qui aura causé l'an dernier l'une des pires récoltes en France depuis la sécheresse de 1976.

Selon l'article [23] des Echos : « Les champs de blé ont souffert des mauvaises conditions météorologiques. Les températures trop froides, les pluies excessives et le manque de lumière ont entraîné des maladies comme la fusariose ou encore la prolifération d'insectes ravageurs... Résultat : des épis peu chargés et des grains de médiocre qualité. », évènement climatique que l'on illustrera à partir des données [24] de rendements moyens dans la région considérée pour la culture de blé tendre d'hiver :

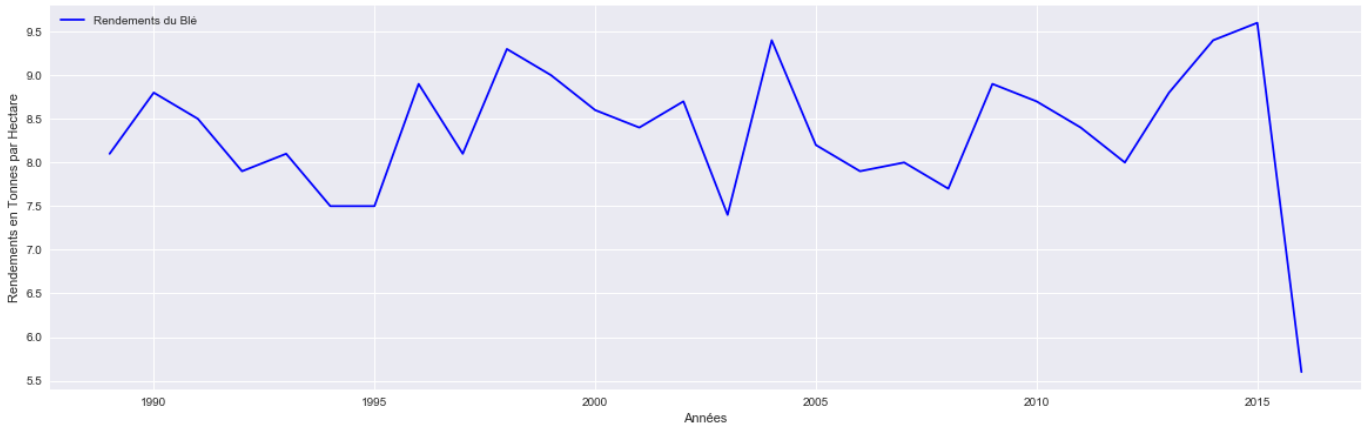

FIGURE 13 – Rendements agricoles moyens de la région considérée

2.2.2 Collecte de données agro-climatiques

Comme nous avons pu le voir, nous serons amenés à modéliser les différents risques susceptibles d'engendrer une perte inhabituelle de rendement au travers de données climatiques. Pour étudier ces risques climatiques tel que le gel, la sécheresse ou les inondations, nous nous intéresserons aux variables hydriques ou de températures suivantes :

Variables Climatiques		
Températures	Hydriques	Diverses
Température minimale	Présence de précipitations	Vitesse du vent
Température maximale	Hauteur des précipitations	Ensoleillement
Température médiane	Niveau d'humidité	...

FIGURE 14 – Variables climatiques collectées

En général, nous collectons l'ensemble des données climatiques issue des stations météorologiques situées à proximité de la région d'étude afin d'établir une cartographie des différentes variables et ainsi obtenir une valeur à n'importe quel endroit de la région étudiée. Cette cartographie du climat, appelée *krigeage*, est une méthode de géostatistique d'interpolation spatiale dont le terme provient du nom de l'ingénieur Danie K. Krige à la base de cette méthode. Néanmoins, suite au désastre climatique survenu dans la région, nous comprenons que la modélisation statistique sous-jacente doit s'étendre au-delà de variables simplement climatiques. Les informations relatives à la variété semée définissent par exemple la résistance au froid, aux bioagresseurs et aux élements naturels dont il est bon de rappeler qu'une partie des récoltes de 2016 fût aussi détruite par l'apparition de bioagresseurs.

23. « 2016, la pire récolte de blé en France depuis 40 ans »
24. Données de production Grandes Cultures (départements)

Néanmoins, les variétés ne sont en général commercialisées que pendant 3 ou 4 ans. Décrites dans la base de données par leur noms, ces innovations successives ont pour effet de créer une discontinuité temporelle de la variable *variété* qui rend donc son utilisation impossible. C'est pourquoi il a été question de collecter des données complémentaires décrivant leurs caractéristiques physiologiques et leurs résistance aux bioagresseur que nous illustrerons par les données [25] mise à disposition par un institut technique agricole française ARVALIS, membre de l'ACTA [26] :

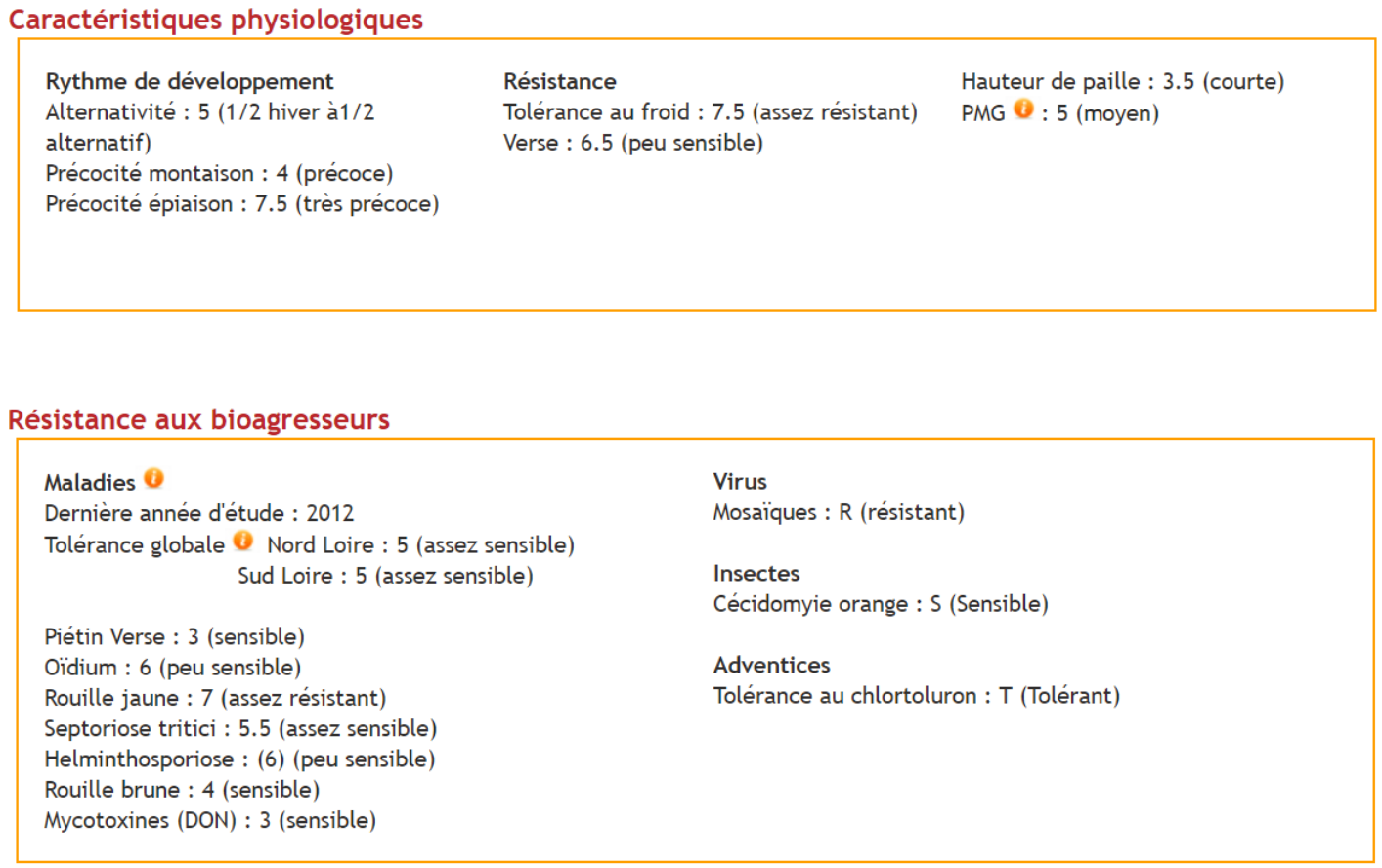

FIGURE 15 – Résistance et caractéristiques physiologiques de la *variété* ACCROC

En outre, la nature du sol présent peut aussi fortement impacter le rendement agricole, en particulier lors d'évènements extrêmes. En effet, lors des intempéries de 2016, les parcelles constituées de sable ont d'avantage été impactées que les autres du fait de la réaction chimique au contact de l'eau. Pour diverses raisons explicitées par la suite, nous avons été amenés à nous intéresser à la composition chimique du *type de sol* par une classification obtenue par un triangle des textures [27] collecté auprès de l'USDA [28], le ministère de l'agriculture des Etats-Unis :

25. Variétés de blé tendre
26. Association de Coordination Technique Agricole
27. Triangle des textures
28. United States Department of Agriculture

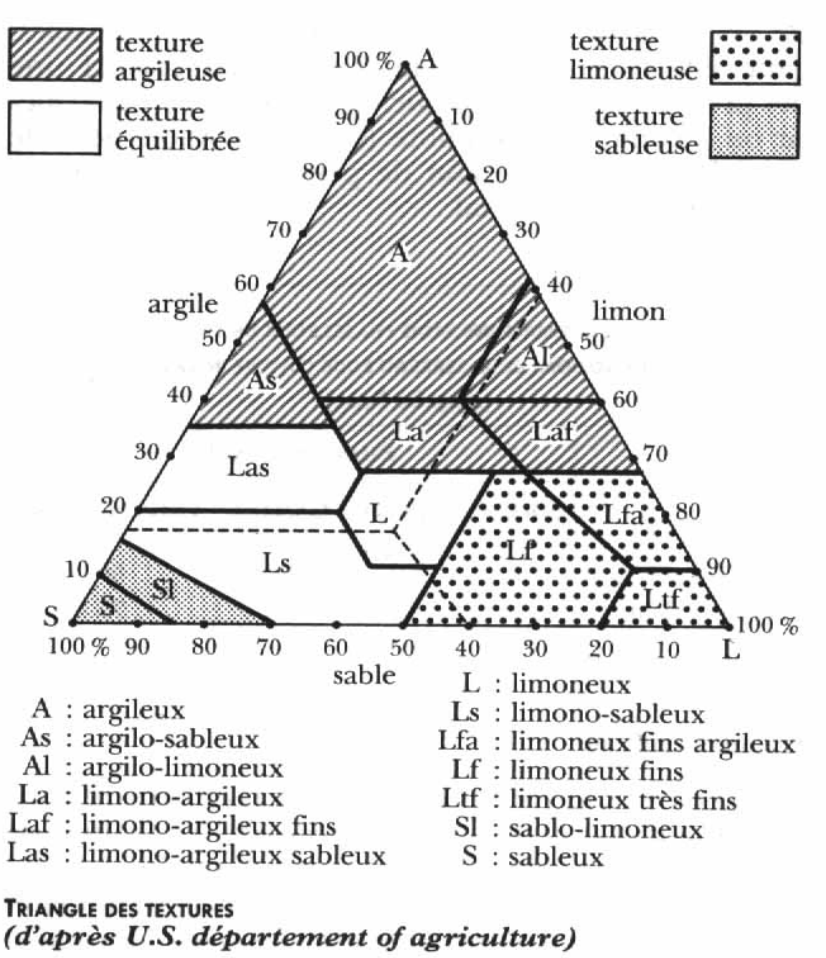

FIGURE 16 – Texture du *type de sol*

2.2.3 Construction d'un indice satellitaire

Dans le monde agricole, il est fréquent que l'agriculteur souhaite observer sa parcelle à l'aide d'un satellite afin d'en tirer des informations relatives à l'état de santé de ses cultures. Si une culture est en bonne santé, elle absorbe la majorité de la lumière visible et reflète une grande proportion de la lumière dans le proche infrarouge. Au contraire, une culture stressée reflète plus de lumière visible et moins de lumière du proche infrarouge. Par une analyse des pixels contenus dans l'image satellitaire, il est possible de construire plusieurs indices satellitaires dont la vocation est bien souvent d'informer sur le niveau de masse végétale présente. Nous précisons que ces images satellitaires ont été collectées auprès d'un partenaire de la collectivité agricole pour toute la durée de l'étude. De plus, ce partenaire garanti une résolution spatiale d'une précision de 5m sur 5m.

Utilisé pour la première fois en 1973, l'*Indice de Végétation par Différence Normalisé*, aussi appelé *Normalized Difference Vegetation Index* (*NDVI*), est utilisé dans le but d'analyser la densité végétale présente au sein d'une zone agricole. Construit à partir des longueurs d'ondes proches du rouge et de l'infrarouge, cet indice de végétation permet de détecter et de quantifier l'activité chlorophyllienne en incluant dans son calcul à la fois la réflectivité [29] dans le proche infrarouge et dans le rouge. Cet indice est donc corrélé à la santé de la plante. Le processus de construction rend cet indice sensible à la vigueur et à la quantité de la végétation.

29. Signatures spectrales des principales surfaces naturelles

Si l'on considère PIR comme la réflectivité dans le proche infrarouge et R la réflectivité dans la région du rouge visible alors nous pouvons écrire la formule de l'indice $NDVI$ de la manière suivante :

$$NDVI = \frac{PIR - R}{PIR + R} \tag{7}$$

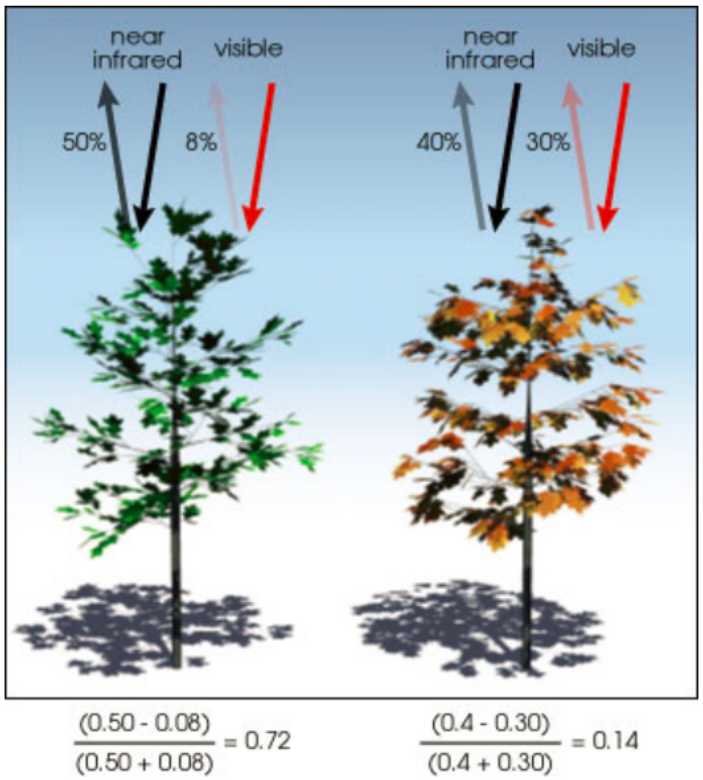

FIGURE 17 – Fonctionnement de l'*Indice de Végétation par Différence Normalisé*

Comme nous pouvons l'observer par l'équation (cf. Equation 7) et le schéma [30] (cf. Figure 17) collecté auprès de la NASA [31], les valeurs de l'indice sont comprises entre -1 et +1, les valeurs négatives correspondant aux surfaces autres que les parties végétales, comme la neige, l'eau ou les nuages pour lesquelles la réflectivité dans le rouge est supérieure à celle du proche infrarouge. Pour les sols nus, la réfléctivité étant à peu près du même ordre de grandeur dans le rouge et le proche infrarouge, le $NDVI$ présente des valeurs proche de 0. Les formations végétales quant à elles, ont des valeurs de $NDVI$ positives, généralement comprises entre 0,1 et 0,7. Les valeurs les plus élevées correspondant aux surfaces les plus denses.

En résumé, cet indice compris entre -1 et 1 indique la présence d'une surface végétale lorsqu'il est positif, plus la valeur est haute, plus la surface présente une forte masse végétale. Il est important de préciser que cet indice ne capte pas l'excès de précipitation ou de chaleur d'où l'intérêt de compléter la base par des variables agro-climatiques.

30. Fonctionnement de l'indice de végétation par différence normalisé
31. National Aeronautics and Space Administration

L'une des variables prédominantes de cette étude fût ainsi construite de manière journalière, tout comme les données climatiques, afin d'obtenir une série temporelle dont nous présenterons ainsi quelques images satellitaires :

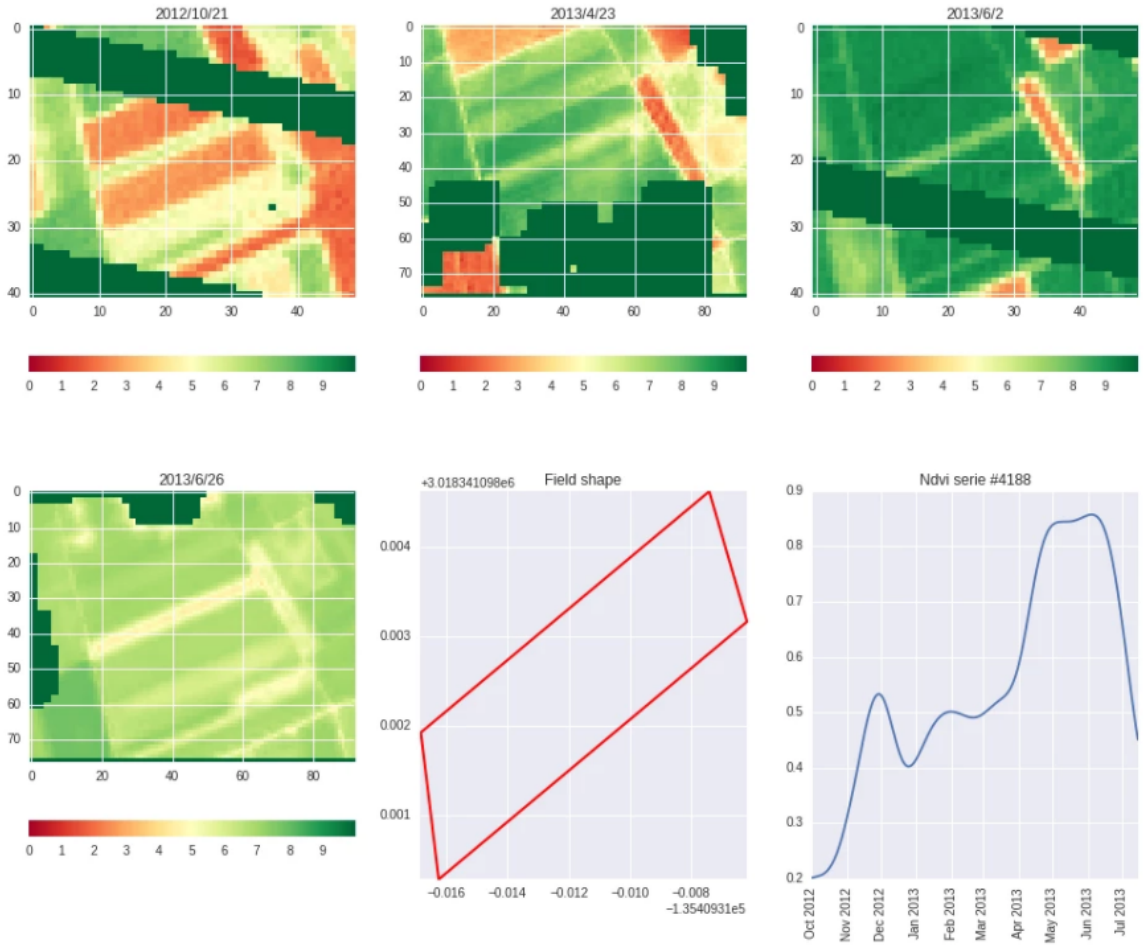

FIGURE 18 – Construction de l'indice NDVI pour l'année 2012-2013

Deuxième partie

Exploration, Analyse et Traitement des données

Dans cette partie, nous nous intéresserons à l'ensemble des méthodes statistiques fondamentales à l'exploration et l'analyse d'une base de données $\mathcal{D}_n$. L'objectif d'une telle analyse est d'une part extraire de l'information et d'autre part traiter ces informations. Cette analyse effectuée en amont de toute modélisation prédictive est primordiale au bon fonctionnement de cette dernière. En général, les méthodes adoptées doivent être réalisées en fonction du phénomène étudié, à savoir le rendement d'un agriculteur.

Tout d'abord, nous nous intéresserons exclusivement à l'analyse du rendement des agriculteurs pour la culture considérée. Au travers d'une estimation de la densité de probabilité et d'une analyse temporelle, l'intérêt de l'exercice sera de modéliser convenablement le sous-jacent. Néanmoins, nous continuerons l'analyse par une brève présentation des méthodes adoptées lors de la détection des anomalies présentes dans la base, très fréquentes dans un contexte agricole. Ces anomalies susceptibles de biaiser la modélisation du rendement doivent alors être retirées du jeu de données. Par la suite, nous développerons l'ingénierie des caractéristiques opéré sur quelques variables climatiques et agricoles. Ces manipulations ont pour objectif principal d'accroître la significativité des variables explicatives présentes au travers de regroupements de modalités ou par la construction de nouvelles variables. Enfin, l'objectif étant de construire un modèle prédictif, nous présenterons quelques méthodes afin de sélectionner les variables explicatives les plus pertinentes pour modéliser l'indice de rendement potentiel. Nous ferons remarquer qu'au vu de la confidentialité de ce mémoire, nous limeterons la présentation des résultats où nous préciserons par ailleurs qu'un coefficient a été appliqué à l'ensemble des variables présentes.

Cette recherche d'information repose en général sur des méthodes statistiques issues de l'apprentissage non-supervisé dont l'objectif est de tirer de l'information d'une suite de données univariée $\mathcal{D}_n = (Y_i)_{i\in[|1,n|]}$ en caractérisant par exemple la loi $\mathcal{L}$ ayant engendrée ces données ou en créant des classes au sein des variables explicatives. Dans le dernier cas, nous considèrerons alors $\mathcal{D}_n = (X_i)_{i\in[|1,n|]}$ avec $X_i = (X_i^1, ..., X_i^p)$ comme l'observation de l'individu i et $X^j = (X_1^j, ..., X_n^j)$ les observations de la variable $j \in [|1, p|]$. Dans un tel contexte, aucune relation aléatoire liant deux processus n'est recherchée, l'apprentissage non-supervisé se concentre exclusivement à l'analyse des relations au sein d'un même ensemble d'apprentissage.

3 Analyse du risque sous-jacent

3.1 Caractérisation de la loi de probabilité

Il est intéressant de débuter une étude statistique par la caractérisation de la loi de probabilité qui a engendrée les données de la variable d'intérêt Y. Cette analyse est primordiale afin de mieux comprendre le processus aléatoire régissant le phénomène étudié. Cette étape est d'autant plus cruciale si l'on souhaite construire un modèle prédictif car selon l'hypothèse $(\mathcal{H}_{\mathcal{L}})$ émise, des modèles statistiques bien précis devront être adoptés. Afin de prendre en compte l'hétérogénéité des risques, nous supposerons les observations $(Y_i)_{i\in[|1,n|]}$ indépendantes et identiquement distribuées selon une certaine loi $\mathcal{L}$ caractérisée par un paramètre θ à identifier :

$$(\mathcal{H}_{\mathcal{L}}) \quad | \quad Y_i \rightsquigarrow \mathcal{L}(\theta_i) \quad i.i.d \quad \text{où} \quad \theta_i \in \Theta \tag{8}$$

Pour appréhender au mieux la loi de probabilité, nous nous intéressons à la *fonction de répartition* $F_Y(.)$ ou encore sa *densité de probabilité* $f_Y = F_Y^{'}$ p.s. L'objectif est alors de valider l'adéquation d'une loi au travers d'une analyse empirique $\hat{F}_Y(.)$:

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i\in[|1,n|]} \mathbb{1}_{y_i\leq y} \underset{n\longrightarrow\infty}{\longrightarrow} \mathbb{E}[\mathbb{1}_{Y\leq y}] = \mathbb{P}[Y\leq y] = F_Y(y) \tag{9}$$

Cette caractérisation de la loi peut s'opérer au travers d'une approche paramétrique ou non-paramétrique, la différence résidant dans l'hypothèse $(\mathcal{H}_{\mathcal{L}})$ émise sur la loi. Comme nous le verrons par la suite, l'approche paramétrique émet un à priori en cherchant la loi satisfaisante parmi un ensemble de lois tandis que l'approche non-paramétrique n'en émet aucun.

3.1.1 Estimation Paramétrique

L'estimation paramétrique d'une densité de probabilité consiste à sélectionner une famille de lois susceptible de contenir celle ayant engendrée le processus. La validité de l'hypothèse émise se doit donc d'être vérifiée par la suite à l'aide d'outils statistiques, ces outils relevant à la fois d'une analyse quantitative que qualititave. Dans le cas d'un rendement agricole, il est en général conseillé de se référer à la famille exponentielle ou celles des valeurs extrêmes.

Néanmoins, afin de déterminer la loi adéquate $\mathcal{L}$ parmi celles présentes dans la famille, nous serons amenés à procéder à des *tests statistiques* dont l'intérêt est de confirmer ou non la véracité de l'hypothèse émise sur la distribution de la variable d'intérêt. Les tests les plus communément réalisés sont le test du maximum de vraisemblance, le test de Kolmogorov-Smirnov ou encore le test d'Anderson-Darling. L'idée principale de ces tests est de construire une procédure statistique $\mathcal{T}_{\mathcal{L}}(.)$ afin de vérifier l'hypothèse *nulle* $(\mathcal{H}_0^{\mathcal{L}})$ et *alternative* $(\mathcal{H}_1^{\mathcal{L}})$ suivante :

$$(\mathcal{H}_{\mathcal{L}}) \quad | \quad \mathcal{T}_{\mathcal{L}}(F_Y) \quad \left\{ \begin{array}{ll} (\mathcal{H}_0^{\mathcal{L}}) & \hat{F}_Y = F_{\mathcal{L}} \\ (\mathcal{H}_1^{\mathcal{L}}) & \hat{F}_Y \neq F_{\mathcal{L}} \end{array} \right\} \tag{10}$$

Cette procédure statistique se succède en général d'une analyse de l'histogramme par un ajustement des densités de probabilité des lois retenues :

FIGURE 19 – Estimation de densité paramétrique

La loi de Gumbel possède l'inconvénient d'avoir une longue queue de distribution à gauche qui aura tendance à accroître le support de la variable d'intérêt $\mathcal{Y}$, au risque donc de dévier de la réalité. D'autant plus que ces valeurs extrêmes peuvent en partie s'expliquer par la présence d'anomalies au sein de la série, ce qui biaisera de ce fait l'estimation de densité opérée. De ce constat, nous privilégierons ainsi la loi de probabilité gaussienne (ou normale) $\mathcal{L} = \mathcal{N}(\mu, \sigma^2)$ dont nous rappellerons la densité de probabilité $f_{\mathcal{L}}(.)$:

$$(\mathcal{H}_{\mathcal{L}}) \quad | \quad f_Y(y) = f_{\mathcal{L}}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2} \tag{11}$$

Une fois l'analyse quantitative terminée, il est en général commun de la compléter par des outils graphiques tel qu'un *diagramme quantitle-quantile* dont l'intérêt est d'apprécier visuellement l'adéquation d'une loi de probabilité au travers des quantiles $F_{\mathcal{L}}^{-1}(\alpha)$ avec $\alpha \in [0, 1]$. La fonction de répartition $F_{\mathcal{L}} = F_{\mathcal{N}(\mu,\sigma^2)}$ sélectionnée dans l'étude d'un diagramme quantile-quantile correspondra à celle obtenu empiriquement $\hat{F}_Y(.)$ si le graphique présente une stricte linéarité entre le rendement observé $\hat{y}$ et le rendement théorique y :

$$\{(y, \hat{y}) \in [0, 1]^2; y = F_{\mathcal{L}}^{-1}(\frac{i}{n}), \hat{y} = \hat{F}_Y^{-1}(\frac{i}{n}); i \in [|1, n|]\} \tag{12}$$

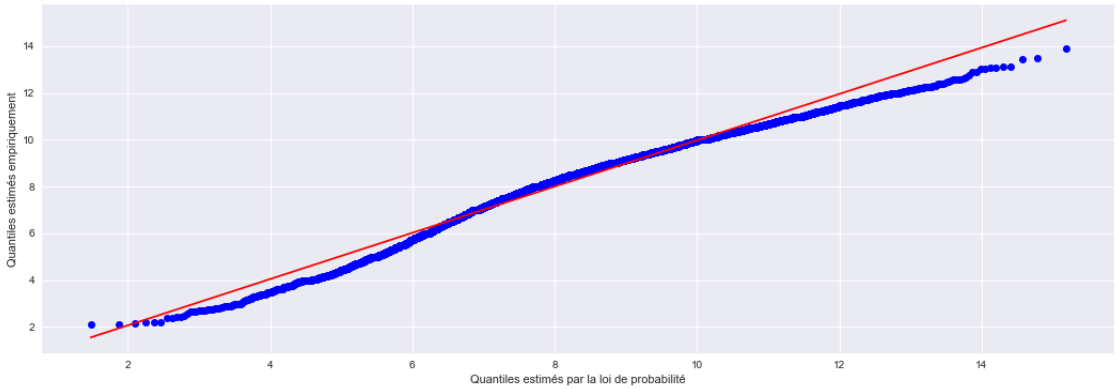

FIGURE 20 – Diagramme quantile-quantile

3.1.2 Estimation non-paramétrique

Une autre approche moins restrictive, dite non-paramétrique, est une étude à ne pas négligée dans le cas où l'adéquation de la loi $\mathcal{L}$ ne paraitrait pas évidente. L'*estimation par noyau* est une méthode non-paramétrique reposant sur la définition d'un *noyau* $K : \mathcal{Y} \longrightarrow \mathbb{R}_+$ dont on rappellera les propriétés similaires à celle d'une densité de probabilité symétrique :

$$\int_{\mathcal{Y}} K(y)d\lambda(y) = 1 \quad \text{ET} \quad \forall y \in \mathcal{Y}, K(y) = K(-y) \tag{13}$$

Cette définition s'accompagne en général d'un *paramètre de lissage* $h \in \mathbb{R}_+^*$, communément appelé *fenêtre* ou encore *bande passante*. La densité de probabilité estimée $\hat{f}_{\mathcal{L}}^h(.)$ à partir des données $(y_i)_{i\in[|1,n|]}$ s'écrira alors en fonction d'un noyau $K_h(.)$ de telle sorte que :

$$(\mathcal{H}_{\mathcal{L}}) \quad | \quad \hat{f}_{\mathcal{L}}^h(y) = \frac{1}{n}\sum_{i=1}^{n} K_h(y - y_i) \quad \text{où} \quad K_h(y - y_i) = \frac{1}{h}K(\frac{y - y_i}{h}) \tag{14}$$

Dans cette étude, nous nous intéresserons au noyau $K(.)$ bien souvent choisi comme la densité d'une fonction normale centrée ($\mu = 0$) et réduite ($\sigma^2 = 1$). Cette approche nous amènera donc à l'estimation suivante :

$$(\mathcal{H}_{\mathcal{L}}) \quad | \quad \hat{f}_{\mathcal{L}}^h(y) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{y - y_i}{h}) \quad \text{où} \quad K(y) = f_{\mathcal{N}(0,1)}(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2} \tag{15}$$

Si le choix du noyau influe en général très peu l'estimateur, il n'en est pas de même pour le paramètre de lissage. Un paramètre trop petit provoque un sur-apprentissage qui fait apparaitre d'énorme perturbation alors qu'une valeur trop grande effacerait au contraire la majorité des caractéristiques. Une manière répandue de choisir ce paramètre est de supposer que l'échantillon est distribué selon une loi paramétrique donnée, par exemple selon la loi normale. L'hypothèse $(\mathcal{H}_{\mathcal{L}})$ considérée plus haut (cf. Equation 11) nous amènera donc à une approximation donnée par $\hat{h} = (\frac{4\sigma^5}{3n})^{\frac{1}{5}} = 1.06\sigma n^{-\frac{1}{5}}$. Pour mieux illustrer la différence entre une approche paramétrique et non-paramétrique, nous présenterons l'histogramme empirique auquel chacune des densités seront ajustées :

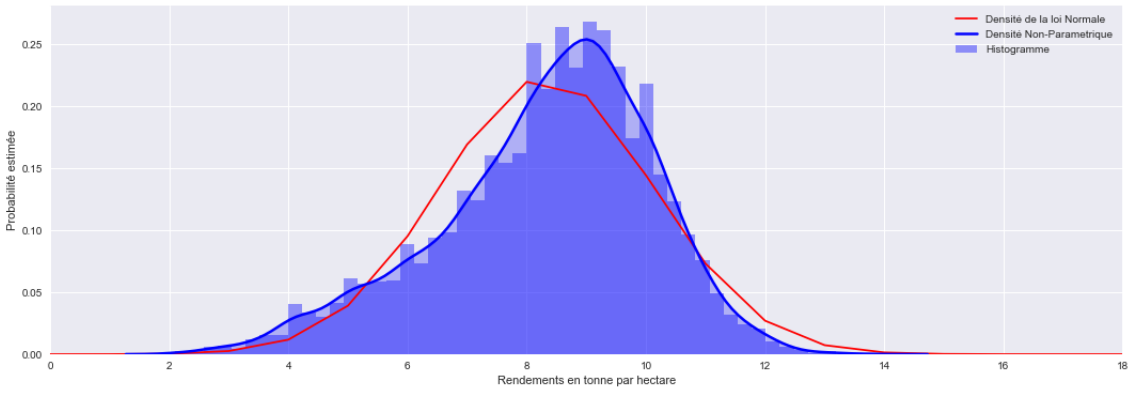

FIGURE 21 – Estimation de densité non-paramétrique

Comme on pouvait s'y attendre de par l'estimation de densité précédente (cf. Figure 20), on observe une sur-évaluation et sous-évaluation de la densité normale aux extrémités du support. La modélisation des points extrêmes étant une part importante, on comprend que le modèle prédictif construit par la suite devra tenir compte de ces inadéquations.

3.2 Analyse d'une série temporelle

La récolte de donnée effectuée initialement omet bien souvent des données essentielles à l'analyse de l'évolution du risque. En effet, il est probable que certaines variables explicatives déterminantes n'aient pas pu être prises en compte du fait de l'absence de celle-ci lors de la récolte ou la difficulté même de récupérer cette donnée. Dans un cadre de tarification agricole, après avoir récolté la *variété* utilisée, le *type de sol* présent, la *gólocalisation satellitaire*, et encore bien d'autres variables, il est dans certain cas difficile de connaître avec certitude l'ensemble des équipements dont bénéficient les parcelles de l'agriculteur. Qu'il s'agisse de l'engrais ou des machines agricoles utilisées, ces données ne sont en général pas disponibles. Nous le verrons plus en détails lors de l'ingénierie des caractéristiques mais il est important de souligner que tout problème d'apprentissage comporte souvent un problème de censure du fait que l'on ne détienne pas l'ensemble des données susceptibles d'expliquer la relation.

Dans cette partie, nous nous intéresserons ainsi aux méthodes statistiques dont le but est de recouvrir de l'information manquante dans un cadre temporel. Dans un contexte d'apprentissage stochastique, il est fréquent de vouloir analyser la présence d'une *tendance* au sein d'une série $(Y_t)_{t\in\mathcal{T}}$ sur une certaine période $\mathcal{T}$, et si présente, la retirer. La présence d'une tendance dans le processus peut engendrer un biais à la modélisation prédictive à venir du fait que les données passées ne sont plus représentatives des données futures. Tout l'intérêt de cette étude est alors de traiter et modifier les données historiques de rendements afin qu'elles puissent être représentatives des données futurs observables.

3.2.1 Décomposition de la tendance

Plus rigoureusement, cette tendance μ se définie comme la variation de rendement d'une année à l'autre que l'on retrouve plus ou moins les années suivantes. Pour commencer une étude, il est en général usuel de considérer des modèles additifs où pour un rendement déterministe $\bar{Y_0}$ indépendant du temps, on peut exprimer le rendement Y_t en fonction d'une composante tendancielle μ_t et d'une composante aléatoire ϵ_t :

$$(\mathcal{H}_{\mathcal{T}}^{\mu}) \quad | \quad \forall t \in \mathcal{T}, Y_t = \bar{Y_0} + \mu_t + \epsilon_t \quad \text{AVEC} \quad \mathcal{F}_{\mathcal{T}}(t) = \bar{Y_0} + \mu_t \tag{16}$$

L'objectif étant de s'intéresser aux variations du rendement d'une année à l'autre, il est important d'introduire l'*opérateur de différence* Δ_T :

$$\forall t \geq T, \Delta_T Y_t = Y_t - Y_{t-T} \tag{17}$$

Si l'on ne s'intéresse par exemple qu'au premier moment Δ_1 et que l'on suppose raisonnablement $\mu_t \approx \mu_{t-1}$, on peut alors montrer que dans le cadre définie précédemment, nous pouvons retirer la tendance par la simple différence :

$$\forall t \geq 1, \Delta_1 Y_t = Y_t - Y_{t-1} = (\bar{Y_0} - \bar{Y_0}) + (\mu_t - \mu_{t-1}) + (\epsilon_t - \epsilon_{t-1}) \approx (\epsilon_t - \epsilon_{t-1}) \tag{18}$$

Une fois le modèle sélectionné $\mathcal{F}_{\mathcal{T}}(.)$, l'analyse de la tendance s'effectue en trois étapes successives où la première consiste à déterminer les valeurs $(\hat{Y_t})_{t\in\mathcal{T}}$ relativement au modèle :

$$\forall t \in \mathcal{T}, \hat{Y_t} = \mathcal{F}_{\mathcal{T}}(t) \tag{19}$$

Ensuite, il sera question de rendre les données historiques représentatives de celles observables aujourd'hui à partir d'une récente date t_p, que l'on appelle communément *pivot*. Cette date est en général définie comme la plus grande présente au sein de $\mathcal{T}$: $t_p = \underset{t\in\mathcal{T}}{max}\ t$. Cependant, dans le cas où celle-ci correspond à un évènement extrême, comme c'est d'ailleurs le cas dans notre étude, nous préconiserons de retirer l'observation de l'étude ou de choisir $t_p^{'} = \underset{t\in\mathcal{T}-\{t_p\}}{max}\ t$.
Une fois cette date définie, l'exercice consiste alors à modifier les valeurs historiques relativement à la donnée récente $\hat{Y}_p = \mathcal{F}_{\mathcal{T}}(t_p)$ et aux valeurs modélisées $(\hat{Y}_t)_{t\in\mathcal{T}}$. Ainsi, nous obtiendrons des valeurs historiques $(\tilde{Y}_t)_{t\in\mathcal{T}}$ ajustées de la manière suivante :

$$\forall t \in \mathcal{T}, \quad \tilde{Y}_t = Y_t - \hat{Y}_t + \hat{Y}_p \tag{20}$$

De par la méthodologie adoptée, on remarquera que la seule et unique valeur historique inchangée correspond à celle du pivot : $\tilde{Y}_p = Y_p - \hat{Y}_p + \hat{Y}_p = Y_p$. De plus, les valeurs récentes ayant les mêmes propriétés tendancielles, on observera que celles-ci devront elles aussi être inchangées.

3.2.2 Correction de la tendance

Afin de déterminer cette tendance μ_t, il est nécessaire de préciser une forme à cette dernière où des approches linéaires peuvent être envisagées :

$$(\mathcal{H}_{\mathcal{T}}^{\mu}) \quad | \quad \forall t \in \mathcal{T}, Y_t = \bar{Y}_0 + t\bar{Y}_1 + \epsilon_t \quad \text{où} \quad \mathcal{F}_{\mathcal{T}}(t) = \bar{Y}_0 + t\bar{Y}_1 \tag{21}$$

Néanmoins, cette approche linéaire présuppose une tendance inchangée d'une période à l'autre, ce qui semble irréaliste au vu des séries de rendements présentées en première partie (cf. Figure 6). En effet, les innovations successives dans le monde agricole se déroulent périodiquement et non en continue, ne serait-ce que par le temps consacré à la mise en place de nouvelles technologies. Cette remarque importante nous amènera alors à s'intéresser à des approches par voisinage dont la *régression locale pondérée* développée par Cleveland et Devlin en 1988.

Le principe général d'une approche locale est d'approximer la fonction par une évaluation au sein d'un voisinage $\mathcal{N}_\nu$ dont on pourra alors se référer pour définir le paramètre de lissage $h = \frac{card(\mathcal{N}_\nu)}{card(\mathcal{T})}$. Si le paramètre est trop faible, l'estimation manque de précision car le voisinage est trop petit, à l'inverse, si le voisinage couvre l'ensemble des observations ($card(\mathcal{N}_\nu) = card(\mathcal{T})$) alors on retrouve la droite d'une régression linéaire ($h = 1$). Il est important de remarquer que cette méthode requiert plus de temps de calcul qu'une simple approche linéaire du fait que pour $card(\mathcal{T})$ observations, on doit faire $card(\mathcal{T})$ régressions. Cette fois, nous ne considèrerons pas un noyau gaussien $K_h(.)$ mais une fonction bien particulière définie par Cleveland. Cette procédure à la particularité d'attribuer des poids à chacune des observations $w_h(t_i, t) = \frac{K_h(t_i - t)}{\sum_{i\in[|1,n|]} K_h(t_i - t)}$ selon une fonction de type tri-cubique en fonction de la distance au centre de la classe :

$$(\mathcal{H}_{\mathcal{T}}^{\mu}) \quad | \quad \mathcal{F}_{\mathcal{T}}(t) = (\bar{Y}_0 + t\bar{Y}_1)w_h(t) \quad \text{où} \quad K_h(t) = (1 - (\frac{|t|}{h})^3)^3 \mathbb{1}_{|t|<h} \tag{22}$$

Il est important de souligner que la correction de la tendance peut être réalisée au niveau de l'historique de la parcelle si jamais celle-ci venait à être disponible ou au niveau de l'exploitation de l'agriculteur. Dans notre cas, l'estimation sera effectuée à l'échelle de la parcelle :

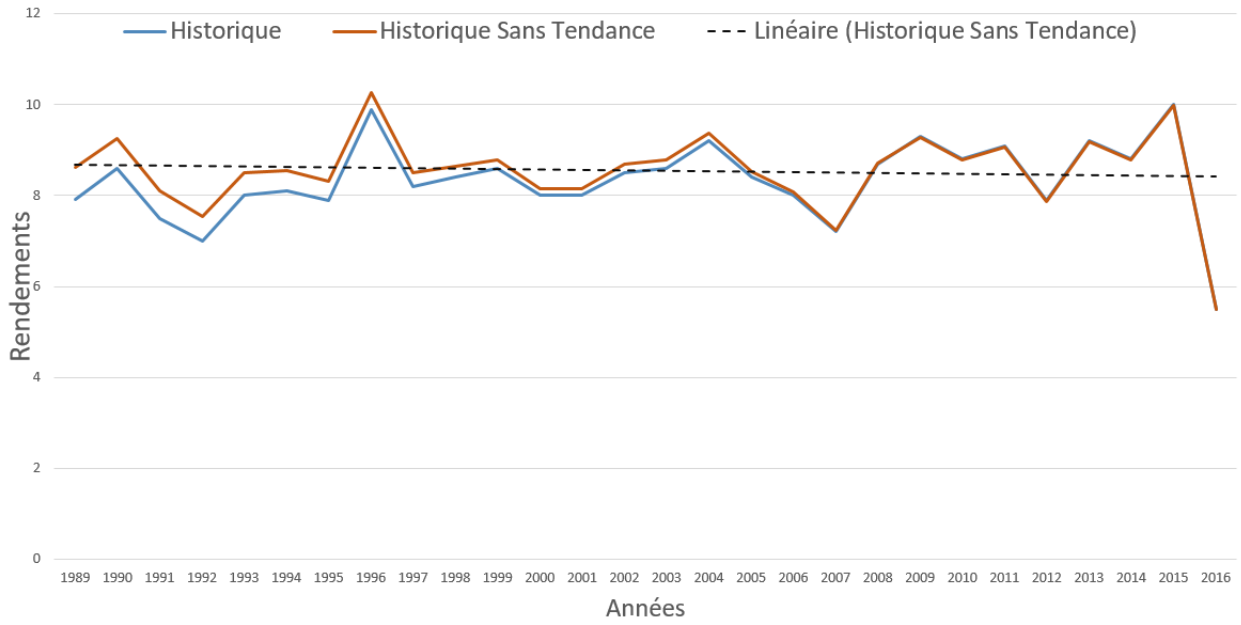

FIGURE 22 – Correction de la tendance d'une série de rendements

4 Détection de données aberantes

La détection des données aberantes est une étape essentielle afin d'améliorer la robustesse des modèles construits par la suite. En effet, l'analyse de ces données est une étape majeure, effectuée en amont de toute modélisation, car leur présence peut biaiser les résultats ainsi que le modèle construit à partir de l'ensemble des données. Ces données si particulières, que l'on appelle communément anomalie, est une observation différente du reste des données présentes dont leur apparition dans la base s'explique en général par le processus de collecte de données à la source de l'étude statistique, qui peut s'avérer parfois maladroit. La recherche de ces anomalies est un problème dont la résolution doit prendre en compte le coût relatif entre celui d'une détection à tort et celui de la non détection, rapport dépendant du contexte de l'étude.

Dans le cas où aucune donnée ou caractéristique définit l'anomalie, on parle aussi bien de classification à une classe que de détection de nouveauté. Ces dernières appellations introduisent néanmoins une nuance dans l'objectif. Il s'agit soit de la détection d'anomalies dans un ensemble de données, soit de déterminer si une nouvelle observation est cohérente ou non avec les données déjà disponibles. Il est important de préciser que nous ne disposons pas ici d'historique identifiant les anomalies. Cette absence de caractérisation nous empêchera d'évaluer l'efficience des méthodes utilisées qui sont par ailleurs les mêmes dans les deux cas.

Chaque méthode projette sa conception de ce qu'est une anomalie. De ce fait, il est probable et rassurant que les différentes méthodes utilisées s'accordent sur la détection des observations très atypiques. La détection d'anomalies est donc un problème complexe sans solution uniformément meilleure car le choix de la méthode à utiliser dépend amplement du contexte. De plus, et heureusement, la classe des anomalies est très généralement sous représentée, engendrant les problèmes classiques de déséquilibres des classes.

L'identification des anomalies présentes dans une base se ramène alors à un problème classique de classification binaire dont la résolution peut être réalisée à l'aide de différentes approches. Cette détection peut en effet s'opérer de manière univariée en ne s'intéressant uniquement à la variable d'intérêt ou de manière multivariée en observant tout un ensemble de variables explicatives. De plus, chacune des approches peut être suivie selon des méthodes issues de l'apprentissage non-supervisé ou supervisé.

4.1 Approche univariée

4.1.1 Approche par quantile

L'analyse univariée de la variable d'intérêt Y repose essentiellement sur une approche probabiliste où l'on s'intéresse aux observations ayant une probabilité d'occurrence anormale relativement aux autres. Etant confronté à une variable stochastique, il peut être intéressant d'effectuer l'analyse pour chacune des années présentes dans l'historique. Une approche répandue est de s'intéresser au diagramme appelé *boîte à moustache* dont la particularité est de faire ressortir les caractéristiques de la distribution à travers quelques statistiques usuelles tel que la médiane $\mathcal{S}_{med}$ et certains quantiles $q_\alpha = F_Y^{-1}(\alpha)$. Tout l'intérêt de ce diagramme est de construire des bornes $[B_{inf}, B_{sup}]$ à partir desquelles nous pourrions considérer les données comme étant des anomalies.

Nous considèrerons dans notre cas la boîte à moustache suivante :

$$\{B_{inf}, q_{25\%}, \mathcal{S}_{med}, q_{75\%}, B_{sup}\} \quad \text{où} \quad \left\{ \begin{array}{l} B_{inf} = q_{25\%} - 1.5 * (q_{75\%} - q_{25\%}) \\ B_{sup} = q_{75\%} + 1.5 * (q_{75\%} - q_{25\%}) \end{array} \right\} \tag{23}$$

Sous l'hypothèse que $Y \rightsquigarrow \mathcal{L} = \mathcal{N}(\mu, \sigma^2)$, nous avons la relation suivante :

$$\mathbb{P}_{\mathcal{L}}[Y \in [B_{inf}, B_{sup}]] = 0.7 \tag{24}$$

Néanmoins, cette approche probabiliste s'avère peu rigoureuse du fait qu'aucun test statistique n'a été réalisé pour mesurer l'incertitude de l'hypothèse. Cette critique nous amènera alors à continuer l'étude des anomalies à l'aide de tests statistiques propre à l'hypothèse $(\mathcal{H}_{\mathcal{L}})$ émise.

4.1.2 Approche paramétrique

Il est possible à l'aide d'une approche paramétrique de construire un test statistique afin de déterminer avec une certaine probabilité les observations aberrantes. L'idée principale de ces tests est de construire une procédure afin de vérifier s'il existe une seule et unique anomalie dans une série de données. Dans une telle configuration, le test se doit d'être répété à plusieurs reprises afin d'identifier la présence de l'ensemble des anomalies, l'hypothèse nulle $(\mathcal{H}_0^{\neq})$ stipulant qu'il n'existe aucune anomalie tandis que l'hypothèse alternative $(\mathcal{H}_1^{\neq})$ stipule qu'il en existe une seule et unique. Sous l'hypothèse que la variable d'intérêt suive une loi normale, il est possible de considérer le test de Grubb's $\mathcal{T}_{\neq}^{G}(.)$ et la statistique de test associée :

$$\mathcal{T}_{\neq}^{G}(Y) = \frac{1}{\sigma_Y^2} \underset{i \in [|1,n|]}{max} |Y_i - \bar{Y}| \tag{25}$$

Nous présenterons les résultats obtenus lors de son application sur la variable d'intérêt pour une année considérée. Afin de faire ressortir l'intérêt de l'étude, nous présentons la densité de probabilité de la variable d'intérêt avant et après que le test n'ait été effectué :

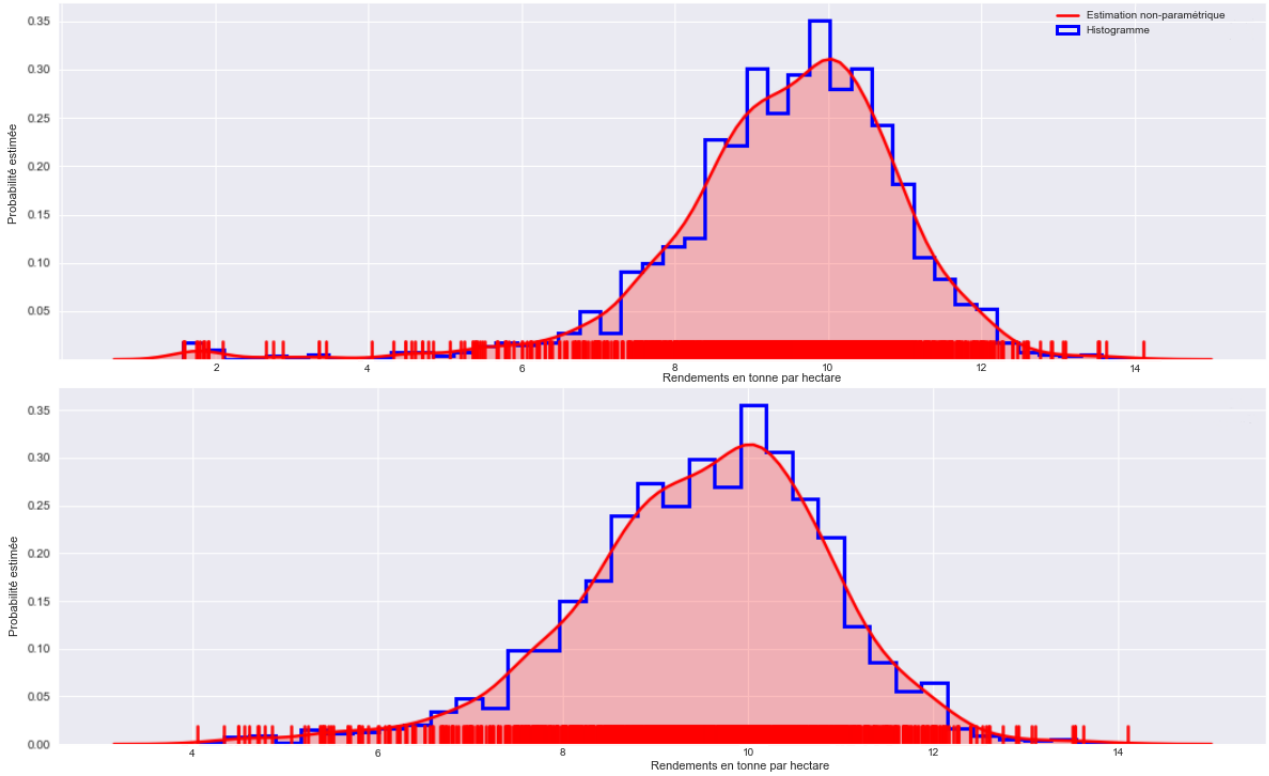

FIGURE 23 – Détection de données aberantes par le test de Grubb's

Au vu des deux distributions, on observe ainsi que les faibles valeurs ont été considérées par le test comme étant des données aberantes. Cette analyse est d'autant plus intéressante que l'année considérée ne présente pas d'évènement extrême ayant engendré de tel rendement. La suppression de ces valeurs nous ramène par ailleurs à une densité de probabilité proche de celle d'une loi normale. Néanmoins, comme on peut s'en douter, rien n'affirme que ces faibles valeurs n'ont pas été réalisées à cause de la particularité de la parcelle liée à la variété semée par exemple. Les limites de l'approche univariée nous conduira alors à s'intéresser aux variables explicatives au travers d'une analyse multivariée.

4.2 Analyse multivariée

4.2.1 Local Outlier Factor

L'approche *Local Outlier Factor* (*LOF*) développée dans les années 2000 par Breunig reste l'une des méthodes de détection d'anomalies les plus utilisée parmi celles fondées sur la densité. Cette méthode consiste à comparer la densité locale d'une observation avec la densité de ses $k-$*plus proches voisins* afin d'y affecter une mesure de l'anomalie. Par souci de clarté, nous noterons ici $\forall i \in [|1,n|], d_i(X) = ||X - x_i||$ et $d(\tilde{X}, X) = ||\tilde{X} - X||$. Pour un entier $k \in \mathbb{N}_*$, cet espace euclidien nous permettra alors de définir les *statistiques de rang* $\{r_1, ..., r_k\}$:

$$r_k(X) = i^*, \quad \text{SI ET SEULEMENT SI} \left\{ \begin{array}{l} d_{i^*}(X) = \underset{1\leq i\leq n; i\neq r_1,...r_{k-1}}{min} d_i(X) \\ d_{i^*}(X) < \underset{1\leq i<i*; i\neq r_1,...,r_{k-1}}{min} d_i(X) \end{array} \right\} \tag{26}$$

Cette méthode nous amènera alors à s'intéresser à l'ensemble des distances $\mathcal{D}_k^{\tilde{X}}$ qui séparent notre donnée $\tilde{X}$ de l'ensemble de ses voisins $\mathcal{N}_k^{\tilde{X}}$. Il est important de préciser que du fait de possibles équidistances entre les voisins de $\tilde{X}$, il est possible d'observer $card(\mathcal{N}_k^{\tilde{X}}) \geq k$.

$$\left\{ \begin{array}{l} \mathcal{N}_k^{\tilde{X}} = \{X_i \in \mathcal{X}; i \in \{r_1(\tilde{X}), ..., r_k(\tilde{X})\}\} \\ \mathcal{D}_k^{\tilde{X}} = \{d(\tilde{X}, X) \in \mathbb{R}_+; X \in \mathcal{N}_k^{\tilde{X}}\} \end{array} \right\} \tag{27}$$

Une fois ces notions précédentes définies, nous nous intéresserons désormais à la distance $d_k^{\tilde{X}}(.)$ au coeur de la méthode *LOF* :

$$d_k^{\tilde{X}}(X) = max\{d(X, \tilde{X}), \mathcal{D}_k^X\} \tag{28}$$

Cette pseudo-distance $d_k^{\tilde{X}}(X)$, appelée *distance d'atteignabilité*, est la distance de X à $\tilde{X}$ si X est assez éloigné de X, mais si X est dans le voisinage de $\tilde{X}$, $d_k^{\tilde{X}}(X)$ est minoré par $\mathcal{D}_k^{\tilde{X}}$. Les observations du k-voisinage de $\tilde{X}$ sont considérées équidistances afin d'apporter une forme de lissage, contrôlé par le paramètre k. Cette approche se fonde sur la densité des k-plus proches voisins d'une donnée $\tilde{X}$ suspectée d'être une anomalie. La densité locale d'atteignabilité $f_k^{\tilde{X}}(.)$ est ensuite définie par l'inverse de la moyenne de la distance d'atteignabilité dans le k-voisinage de X. De là, une valeur locale $LOF_k(\tilde{X})$ est ainsi calculée :

$$LOF_k(\tilde{X}) = \frac{\sum_{X\in\mathcal{N}_k^{\tilde{X}}} \frac{f_k^{\tilde{X}}(X)}{f_k^{\tilde{X}}(\tilde{X})}}{card(\mathcal{N}_k^{\tilde{X}})} = \frac{\sum_{X\in\mathcal{N}_k^{\tilde{X}}} f_k^{\tilde{X}}(X)}{f_k^{\tilde{X}}(\tilde{X}) card(\mathcal{N}_k^{\tilde{X}})} \quad \text{où} \quad f_k^{\tilde{X}}(X) = 1/\frac{\sum_{X\in\mathcal{N}_k^{\tilde{X}}} d_k^{\tilde{X}}(X)}{card(\mathcal{N}_k^{\tilde{X}})} \tag{29}$$

Une valeur de 1 correspond à une observation dans la norme de la distribution, mais une borne au-delà de laquelle une observation est atypique n'est pas explicite, cela dépend du contexte et des dispersions relatives. L'avantage premier de cette méthode est qu'elle n'émet aucune hypothèse sur la distribution des données :

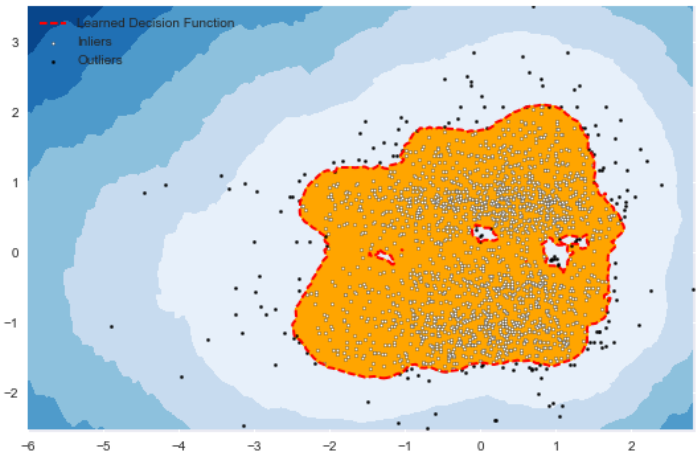

FIGURE 24 – Détection d'anomalies par une approche de densité locale

Cette approche par les voisins atteint rapidement ses limites du fait que le paramètre k sélectionné influence grandement le résultat et qu'aucune méthode efficiente ne permet de déterminer le paramètre optimal. Cette approche locale ne tient d'autant plus pas compte des possibles corrélations entre les variables explicatives. Ces remarques nous conduiront alors à nous intéresser à des approches plus stables fondées sur des arbres de décision.

4.2.2 L'arbre d'isolation

Impulsé par Breiman, l'adaptation des forêts aléatoires à un problème de classification non-supervisée a été développée au début de ce siècle. Néanmoins, nous nous attarderons ici par celle développée par Liu en 2008, *l'arbre d'isolation*. Quelque soit la méthode préconisée, ces versions des forêts aléatoires sont adaptées de façon très spécifique à la détection d'anomalies.

La version non-supervisée des forêts aléatoires est aussi une application de la notion de proximité entre les observations. Le principe de la méthode de Liu est de construire un ensemble d'arbres de décision complètement aléatoires où la division opérée dans chaque nœud est issue du tirage d'une variable et d'un seuil aléatoire. La construction de chacun des arbres est poursuivie jusqu'à l'obtention d'une unique observation par feuille. La quantification de l'isolement ou de l'anomalie d'une observation est ainsi obtenue par la longueur du chemin atteignant cette observation. Plus celui-ci est court, plus l'observation est considérée isolée ou atypique. Il en découle comme précédemment la construction d'une mesure de proximité et donc de distance. La mesure de l'anomalie est ensuite calculée pour chacune des observations comme la moyenne des longueurs des chemins. Ce score indique donc pour chaque observation sa plus ou moins grande proximité avec toutes les autres observations.

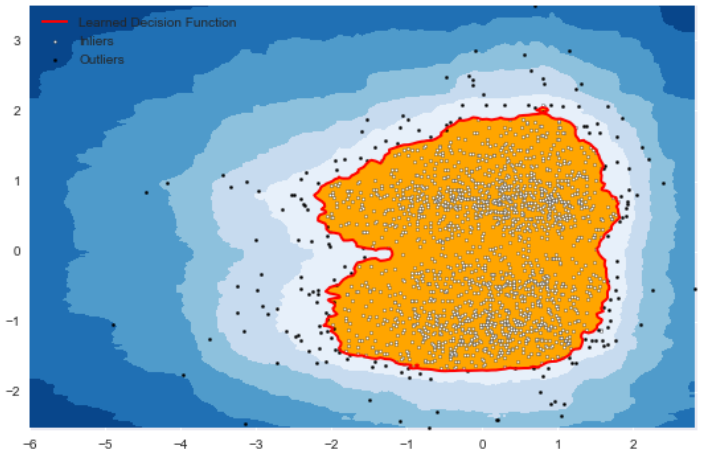

FIGURE 25 – Détection d'anomalies par des arbres d'isolation

Naturellement, la différence est marquée entre les deux types d'anomalies détectées par les forêts aléatoires par rapport à une approche fondée sur la distribution globales des observations. Cette fois-ci, le nuage constitué des observations normales est plein et plus cohérent que celui présenté précédemment. De plus, on remarquera que cette approche a pu tenir compte de l'effet de seuil procuré par la variable explicative en ordonnée.

5 Ingénierie des caractéristiques

5.1 Classification de variables explicatives

Dans cette partie, nous nous intéresserons à un exercice incontournable de l'ingénierie des caractéristiques : la classification de variables explicatives. L'intérêt de l'exercice est de construire des classes d'individus homogènes au sein d'un ensemble de variables explicatives ou même au sein d'une seule et unique variable. Pour ce faire, il est primordiale de considérer l'espace d'entrée $(\mathcal{X}, ||.||)$ muni d'une distance associée $d(.)$ permettant de mesurer la différence entre deux individus $(X_i, X_{i'})$:

$$\forall i, i' \in [|1, n|], d(X_i, X_{i'}) = ||X_i - X_{i'}|| = \sqrt{\sum_{j \in [|1,p|]} (X_i^j - X_{i'}^j)^2} \tag{30}$$

Néanmoins, il est en général nécessaire de se rapporter à une représentation centrée réduite des individus $(\tilde{X}_i)_{i \in [|1,n|]}$ du fait de l'existence de mesures disparates entre les variables explicatives :

$$\tilde{X}_i^j = \frac{X_i^j - \bar{X}^j}{\sigma_j} \quad \text{où} \quad \bar{X}^j = \frac{1}{n} \sum_{i \in [|1,n|]} X_i^j \quad \text{et} \quad \sigma_j^2 = \frac{1}{n} \sum_{i \in [|1,n|]} (X_i^j - \bar{X}^j)^2 \tag{31}$$

Par la suite, nous considérerons alors le *centre de gravité* $g_{\mathcal{N}}$ et le *nuage de points* $\mathcal{N}$ associée aux individus $(X_i)_{i \in [|1,n|]}$ dont les poids $(p_i)_{i \in [|1,n|]}$ peuvent être considérés uniformes $(p_i = \frac{1}{n})$:

$$\mathcal{N} = \{(X_i, p_i); X_i \in \mathcal{X}, p_i \in]0, 1[; \sum_{i \in [|1,n|]} p_i = 1\} \quad \text{où} \quad g_{\mathcal{N}} = \sum_{i \in [|1,n|]} p_i X_i \tag{32}$$

De ce nuage de point, on s'intéressera à sa dispersion autour de son centre de gravité pour laquelle l'inertie est minimale. L'inertie d'un nuage $\mathcal{I}_{\mathcal{N}}$ étant la moyenne pondérée des carrés des distances du centre de gravité :

$$\mathcal{I}_{\mathcal{N}} = \sum_{i \in [|1,n|]} p_i d^2(g_{\mathcal{N}}, X_i) \quad \text{ou} \quad \mathcal{I}_{\mathcal{N}} = \sum_{i \in [|1,n|]} p_i ||X_i||^2 \quad \text{si} \quad g_{\mathcal{N}} = 0 \tag{33}$$

L'objectif étant de construire K classes homogènes d'individus, nous serons amenés à construire une partition $\mathcal{P}^{\mathcal{X}} = (\mathcal{X}_i)_{i \in [|1,K|]}$ de $\mathcal{X}$ où l'on considèrera $(\mathcal{N}_i)_{i \in [|1,K|]}$ les nuages de points associés. Ces K classes seront d'autant plus homogènes que les inerties correspondantes $\mathcal{I}_{\mathcal{N}_i}$ seront faibles. Cette remarque nous amènera alors à s'intéresser à la somme des inerties de chacune de ces classes, l'*inertie intraclasse* $\mathcal{I}_{intra}$:

$$\mathcal{I}_{intra} = \sum_{i \in [|1,K|]} \mathcal{I}_{\mathcal{N}_i} \tag{34}$$

Le problème de classification se réduit alors à chercher la partition de sorte à ce que l'inertie intraclasse soit minimale, de manière à obtenir en moyenne des classes homogènes. A cette notion s'ajoute celle de la dispersion de chacune des classes par rapport à leur centre de gravité $g_{\mathcal{N}_i}$, l'*inertie interclasse* $\mathcal{I}_{inter}$:

$$\mathcal{I}_{inter} = \sum_{i \in [|1,K|]} \bar{p}_i d^2(g_{\mathcal{N}}, g_{\mathcal{N}_i}) \quad \text{où} \quad p_i = \frac{card(\mathcal{N}_i)}{n} \tag{35}$$

Le but de ces méthodes de classification est de construire une partition afin de regrouper les individus en classes homogènes, or l'inertie du nuage, étant la somme de son inertie intraclasse et de son inertie interclasse ($\mathcal{I}_{\mathcal{N}} = \mathcal{I}_{intra} + \mathcal{I}_{inter}$), cela revient au même de chercher à maximiser l'inertie interclasse ou minimiser l'inertie intraclasse. Par la suite, nous présenterons deux des méthodes de classification les plus répandues afin de regrouper les modalités de la variable *type de sol*. Ces méthodes de classifications sont particulièrement utiles pour effectuer des regroupements de modalités au sein de certaines variables explicatives. Ces regroupements de modalités s'effectueront si la variable dénombre un trop grand nombre de modalités ou si l'influence de ses modalités sur la variable d'intérêt est la même. En effet, un grand nombre de modalités peut rendre une variable non discriminante ou engendrer du sur-apprentissage. Nous distinguerons de ces méthodes de classification celles qui permettent d'obtenir une partition optimale en un nombre de classes fixé à priori et celles qui aboutissent à la meilleure partition par améliorations successives d'une partition initiale plus ou moins arbitraire.

5.1.1 Méthode des Centres Mobiles

L'*algorithme de Forgy*, ou *méthode des centresmobiles*, est un algorithme qui consiste à améliorer progressivement une partition arbitraire. On construit alors une nouvelle partition en agglomérant les éléments de l'espace d'entrée autour des centres de gravité de chacune des classes. L'affectation d'un élément à une classe est telle que la distance entre le centre de gravité et l'élément soit minimum. On calcule les centres de gravité des classes de la nouvelle partition, et on itère le processus. Ainsi, d'une étape à l'autre, l'inertie interclasse ne peut pas diminuer, ou, ce qui est équivalent, que l'inertie intraclasse ne peut pas augmenter.

D'autres variantes de l'algorithme de Forgy existent tel que la méthode des $k - means$ de Mac Queen ou encore la méthode $ISODATA$ de Ball et Hall. Ce dernier est considéré comme l'algorithme le plus sophistiqué des méthodes de centroïdes. Les fondements étant les mêmes, ce dernier ajoute des perfectionnements supplémentaires. Il y a division d'une classe lorsque l'écart-type d'une variable quelconque dans la classe est supérieur à un certain nombre de fois son écart-type. De plus, il y a fusion de deux classes uniquement lorsque leurs centres de gravité sont à une distance inférieure à un seuil fixé en amont.

Certes, ces méthodes sont pratiques pour traiter rapidement des ensembles d'effectifs élevé, il n'en reste pas moins que l'inconvénient majeur de ces algorithmes est que la partition finale dépend de la partition de départ, celle-ci étant choisie à l'initialisation de l'algorithme. De plus, il est nécessaire de préciser le nombre de classes souhaité qui est, en absence de réelle expertise, bien difficile à justifier. Ainsi, ces méthodes des centres mobiles ne semblent pas adaptées à notre problème. Nous allons donc nous intéresser par la suite sur une seconde méthode de classification dîte hiearchique.

5.1.2 Classification Ascendante Hiearchique

Les méthodes de *classification hiérarchiques* produisent des suites de partitions emboîtées obtenues par regroupements successifs de parties. Il existe des méthodes descendantes et ascendantes qui partent des singletons ou de l'espace d'entrée, et procèdent par regroupements ou segmentations successifs. Les méthodes descendantes ne sont en général très peu utilisées du fait qu'elles reposent sur des principes de segmentation et non de regroupement. On appelle *hiérarchie totale* un ensemble $\mathcal{P}^{\mathcal{X}} \subset \mathcal{P}(\mathcal{X})$ contenant l'espace d'entrée $\mathcal{X}$ et ses singletons où deux éléments de l'espace d'entrée sont emboîtés ou d'intersection vide :

$$\left\{ \begin{array}{c} \mathcal{X} \in \mathcal{P}^{\mathcal{X}} \quad \text{ET} \quad \forall X \in \mathcal{X}, \{X\} \in \mathcal{P}^{\mathcal{X}} \\ \forall X, X' \in \mathcal{P}^{\mathcal{X}}, X \cap X' \in \{X, X', \varnothing\} \end{array} \right\} \tag{36}$$

L'algorithme de construction de la hiérarchie ascendante peut être décrit de la manière suivante :

— **Initialisation** : Partition initiale $\mathcal{P}_0^H$ composée de chacun des singletons
— **Regroupement** : Tant que la partition n'est pas égale à l'espace d'entrée $\mathcal{P}_{n-1}^{\mathcal{X}} = \mathcal{X}$
— $\mathcal{P}_{k+1}^{\mathcal{X}}$ est construit en regroupant les deux classes les plus proches parmi $\mathcal{P}_k^{\mathcal{X}}$
— **fin**

Une hiérarchie de parties est donc un arbre satisfaisant l'axiome suivant :

$$\forall X, X_1, X_2 \in \mathcal{P}^{\mathcal{X}}; X \subset X_1, X \subset X_2 \Longrightarrow X_1 \subset X_2 \quad \text{OU} \quad X_2 \subset X_1 \tag{37}$$

Ce système représente une hiérarchie de parties indicée que l'on visualise graphiquement sous la forme d'un arbre hiérarchique appelé *dendrogramme*. Chaque élément de la hiérarchie de parties est représenté par un nœud, et les nœuds sont reliés par des branches. Dans le cas par exemple du *type de sol*, nous observerons des similitudes au sein des modalités qui nous amènent à construire l'arbre :

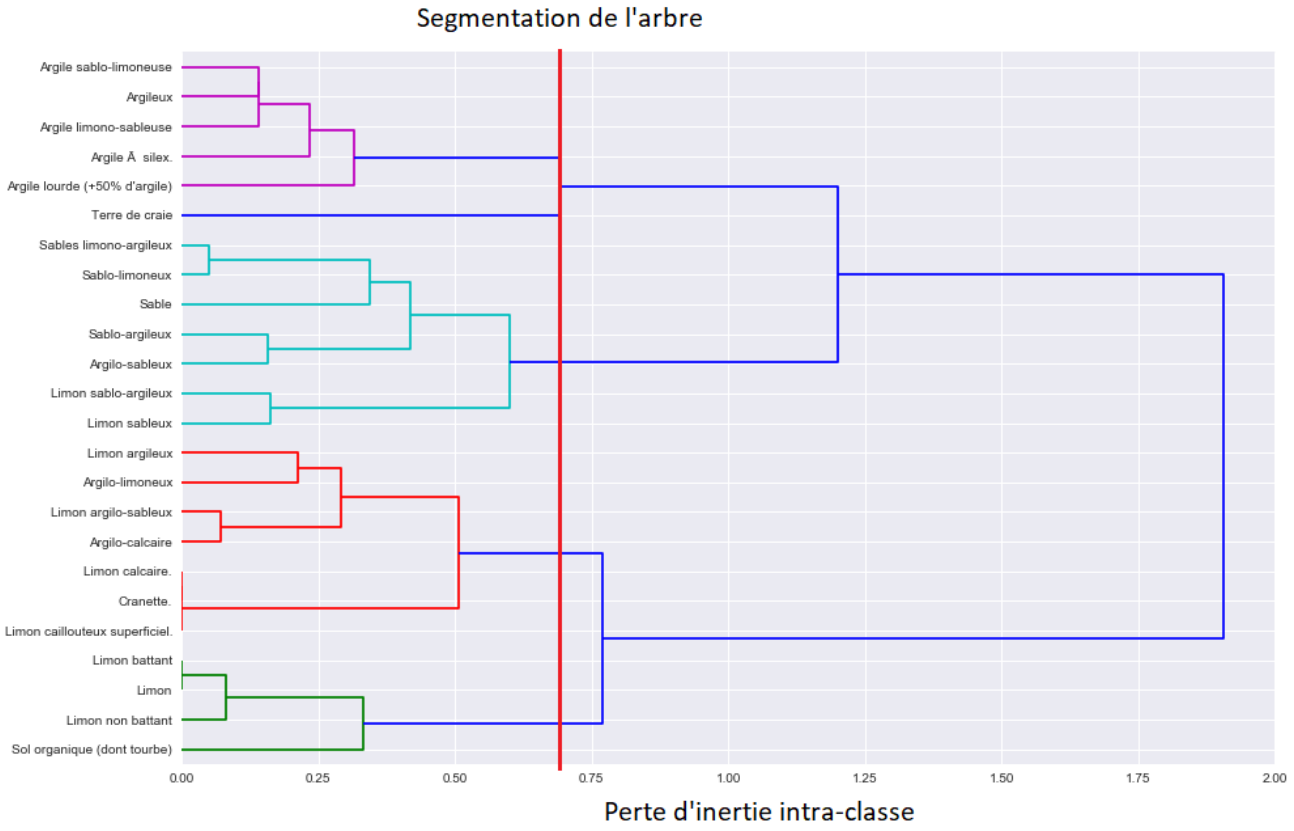

FIGURE 26 – Dendrogramme de la Classification Ascendante Hiérarchique du *type de sol*

Il est intéressant de compléter l'étude par une *Analyse en Composante Principale* (ACP) afin d'avoir une appréciation visuelle de la classification opérée en amont et du nuage de points qui en découle. Cette méthode, fondée sur la projection d'espaces vectoriels, a pour objectif principal de réduire la dimension d'un espace afin d'en avoir une meilleure visualisation :

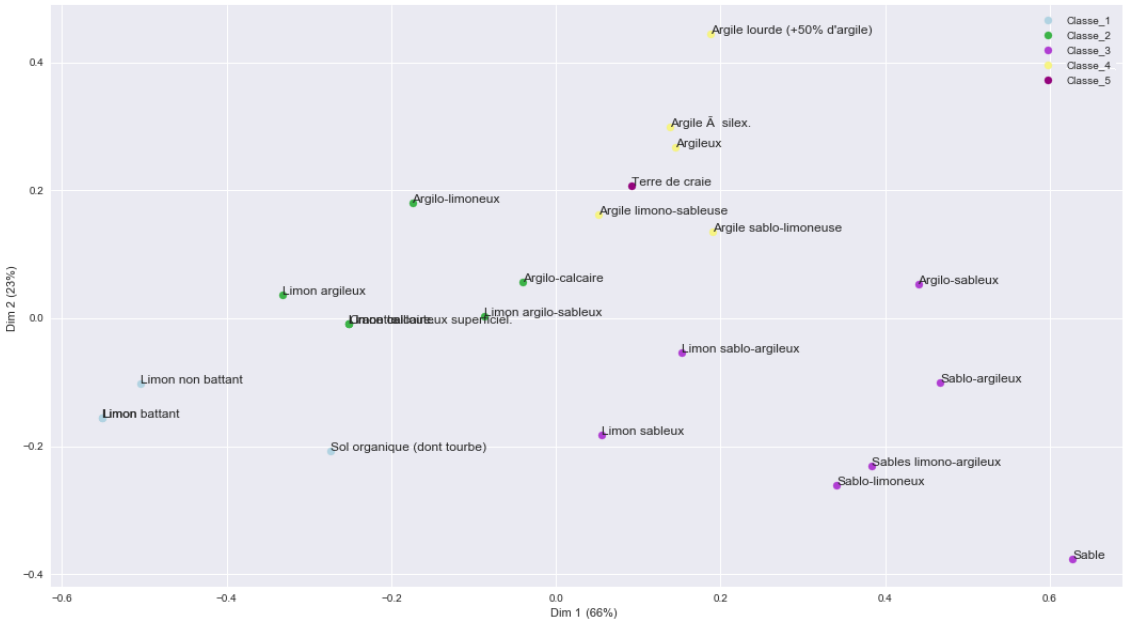

FIGURE 27 – Visualisation de la classification à travers une ACP

5.2 Ingénierie des variables climatiques

5.2.1 Construction de variables significatives

S'il est bon de considérer que l'on peut expliquer un processus à partir d'une récolte de données, il va de soi d'admettre que la base de données manque sans doute d'informations pour expliquer parfaitement le phénomène aléatoire en question. Afin de résorber ce manque, nous nous concentrerons dans la suite à la construction de variables explicatives spécifiques au problème étudié. Cette ingénierie des variables est en général promulgué par un expert du sous-jacent, par exemple un agronome dans le cas de la modélisation d'un rendement agricole. Ainsi, un travail d'analyse de données et d'expertise, propre à chacune des cultures, permet de déterminer les variables explicatives pertinentes à construire ainsi que les différents seuils auxquels opérer des manipulations. Nous rappellons que nous avons été amenés à partitionner le cycle de production en plusieurs périodes successives $\mathcal{T} = [|1, ..., T|]$ correspondant aux stades phénologiques de la culture étudiée.

Selon la culture, il existe ce que l'on appelle une temperature de base (T_B=0 °C pour le blé) en dessous de laquelle la croissance biologique est nulle. De même, il existe une température maximale au delà de laquelle la plante ne croît plus (30 °C). Ainsi, pour chacun des stades phénologiques, il apparaît naturel de construire des premiers indicateurs à partir de ces différents seuils. Si l'on considère la température de base et une suite de température minimale $(T_i^m)_{i\in[|1,n_t|]}$ atteinte sur une période $t \in \mathcal{T}$ composée de n_t jours alors nous nous intéresserons au nombre de jours ou aux valeurs qui dépassent ce seuil T_B :

$$\forall t \in \mathcal{T}, X_t^j = \sum_{i \in [|1,n_t|]} \mathbb{1}_{T_i^m \leq T_B} \quad \text{OU} \quad X_t^j = \sum_{i \in [|1,n_t|]} |T_i^m - T_B| \mathbb{1}_{T_i^m \leq T_B} \tag{38}$$

En choisissant $T_B = 1$ cm, ces variables explicatives peuvent être adaptées dans le cas où l'on souhaiterait déterminer la *présence* ou le *cumul de précipitations* sur une période. Néanmoins, il est commun de construire ces dépassements de seuils à partir d'indices plus complexes que la température. On peut par exemple considérer l'indice appelé *dégre jour de croissance* qui sert à mesurer l'accumulation de chaleur nécessaire au développement biologique de la culture. Si l'on considère $(T_i^M)_{i \in [|1,n_t|]}$ une suite de température maximale atteinte dans chacune des journées, on peut dès lors cumuler la valeur des degrés jours DJ_i :

$$\forall t \in \mathcal{T}, X_t^j = \sum_{i \in [|1,n_t|]} DJ_i \mathbb{1}_{DJ_i \geq T_B} \quad \text{OÙ} \quad DJ_i = \frac{T_i^m + T_i^M}{2} - T_B \tag{39}$$

5.2.2 Construction de l'ensemble stochastique

L'objectif étant de constuire une solution de couverture dynamique, nous avons été amenés à discréditer le cycle de production. Afin d'éviter des variations non-significatives du rendement potentiel à cause de faibles variations des variables explicatives, les périodes sélectionnées ont été celles des stades phénologiques de la culture considérée. La construction de l'ensemble d'apprentissage $\mathcal{X}$ au fil des stades phénologiques est une étape primordiale dans la modélisation du rendement. Nous précisons que les variables indépendantes du temps tel que la *variété* ou encore le *type de sol* seront à chaque fois incorporées dans l'ensemble d'apprentissage.

Dans cette étude, nous nous intéresserons alors à la construction des sous-ensembles $(\mathcal{X}_t)_{t \in \mathcal{T}}$ correspondant aux variables stochastiques, à savoir le $NDVI$ et les variables climatiques. La construction de ces ensembles peut s'opérer de plusieurs manières selon les hypothèses émises par l'analyste. Néanmoins, afin de conserver l'information des périodes précédentes, nous serons forcément amenés à considérer une suite croissante d'ensembles :

$$(\mathcal{H}_{\mathcal{X}}) \quad | \quad \forall t, t' \in \mathcal{T}; t \leq t' \Longrightarrow \mathcal{X}_t \subset \mathcal{X}_{t'} \tag{40}$$

Une première idée serait de considérer une variable stochastique X_t^j indépendante d'elle même d'une période à l'autre. Dans une telle approche, l'idée est de considérer que chacune des variables a un effet sur le sous-jacent qui pourrait s'avérer contraire d'une période à une autre. Comme on a pu le voir à travers l'année 2016, on peut par exemple penser aux précipitations qui sont essentielles pour la croissance de la plante en début de cycle mais qui s'avèrent dévastatrice par la suite :

$$(\mathcal{H}_{\mathcal{X}}^{\perp}) \quad \forall j \in [|1,p|], \forall t, t' \in \mathcal{T}; t \neq t', X_t^j \perp X_{t'}^j \tag{41}$$

Sous une telle hypothèse $(\mathcal{H}_{\mathcal{X}}^{\perp})$, il s'agirait alors de construire chacune des variables stochastiques pour chacun des stades ce qui nous amènerait à démultiplier le nombre de variables initiales par le nombre de périodes considérées. A chacune des périodes, l'ensemble d'apprentissage $\mathcal{X}_t$ ainsi que sa dimension C_t s'écriront alors par récurrence de la manière suivante :

$$\text{SOUS} \quad (\mathcal{H}_{\mathcal{X}}^{\perp}) \quad | \quad \mathcal{X}_t = ((X_1^1, ..., X_1^p), ..., (X_t^1, ..., X_t^p)) \Longrightarrow \forall t \in \mathcal{T}, \quad C_t = t * p \tag{42}$$

On remarque ainsi que sous $(\mathcal{H}_{\mathcal{X}}^{\perp})$, le nombre de variables explicatives augmente considérablement au cours des périodes au risque d'entraîner le même effet quant à la complexité de ce dernier. Il est important de rappeler que quelque soit l'algorithme sélectionné par la suite, il ne sera pas en capacité de reconnaître d'une période à une autre que c'est la même variable si jamais celle-ci venait à être sélectionnée plusieurs fois. De ce constat, l'idée d'agréger les variables au fil des stades phénologiques semble contribuer à une stabilité de la complexité de l'ensemble d'apprentissage tout en levant l'hypothèse d'indépendance.

Par la suite, nous avons alors construit le modèle en agrégeant les variables explicatives uniquement si jamais celles-ci venaient à être considérées pertinentes sur la période considérée. Au vu de l'ingénierie des caractéristiques opérée et de la volonté de conserver l'impact de chacune des variables sur chacune des périodes, le choix de l'opération à appliquer pour les agréger s'est naturellement porté sur la somme. Ainsi, l'ensemble des entrées s'écrira de manière récursive de tel manière que $\forall t \in \mathcal{T}^*$, on cumelera la variable X_t^j d'une période à l'autre uniquement sur des périodes considérées pertinentes $\mathcal{T}_j \subset \mathcal{T}$:

$$\text{SOUS} \quad (\mathcal{H}_{\mathcal{X}}^{\sum}) \quad | \qquad \mathcal{X}_t = ((\sum_{i \in \mathcal{T}_1; i \leq t} X_i^1), ..., (\sum_{i \in \mathcal{T}_p; i \leq t} X_i^p)) \Longrightarrow \forall t \in \mathcal{T}, C_t = p \tag{43}$$

On remarque ainsi que dans cette dernière approche $(\mathcal{H}_{\mathcal{X}}^{\sum})$, peu importe la période considérée, la complexité de ce modèle est constante sur l'intégralité du cycle et bien plus faible que sous l'hypothèse précédente $(\mathcal{H}_{\mathcal{X}}^{\perp})$.

6 Sélection de variables explicatives

6.1 Sélection univariée

Les méthodes de sélection de variables par une analyse univariée s'intéresse à la seule relation entre la variable d'intérêt et une variable explicative considérée. Le coefficient sélectionné dans une étude univariée se doit d'être choisie avec précaution du fait que la relation entre les deux peut être linéaire ou non. Chacun des coefficients porte donc un à priori sur la relation entre ces deux variables. Le calcul de ce coefficient pour l'ensemble des variables explicatives amène à la construction d'une *matrice de corrélation*.

Ces méthodes sont particulièrement utiles pour réduire l'ensemble possible des variables explicatives avant toute modélisation que ce soit. Le faible coût de calcul qui en résulte en font des méthodes prisées tout particulièrement au début de l'étude. L'usage de telles méthodes permet de réduire dans un premier temps la dimension de l'ensemble $\mathcal{X}$ auquel nous serons amenés par la suite à user de méthodes beaucoup plus coûteuses.

6.1.1 Le coefficient de Pearson

Le *coefficient de Pearson* $\rho_{(X,Y)}$ permet d'étudier la relation linéaire entre deux variables en s'intéressant au rapport entre leur covariance $\mathbb{C}\text{ov}[X,Y]$ et leur écart-types respesctifs (σ_X, σ_Y). Ce coefficient, compris entre -1 et 1, informe d'une relation linéaire lorsque ce dernier se voit être proche de 1 en valeur absolu, le signe indiquant si la relation est positivement ou négativement corrélée. Calculée pour chacune des variables explicatives, nous obtenons la matrice symétrique :

$$\rho_{(X,Y)} = \frac{\mathbb{C}\text{ov}[X,Y]}{\sigma_X \sigma_Y} \quad \text{où} \quad \mathbb{C}\text{ov}[X,Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{44}$$

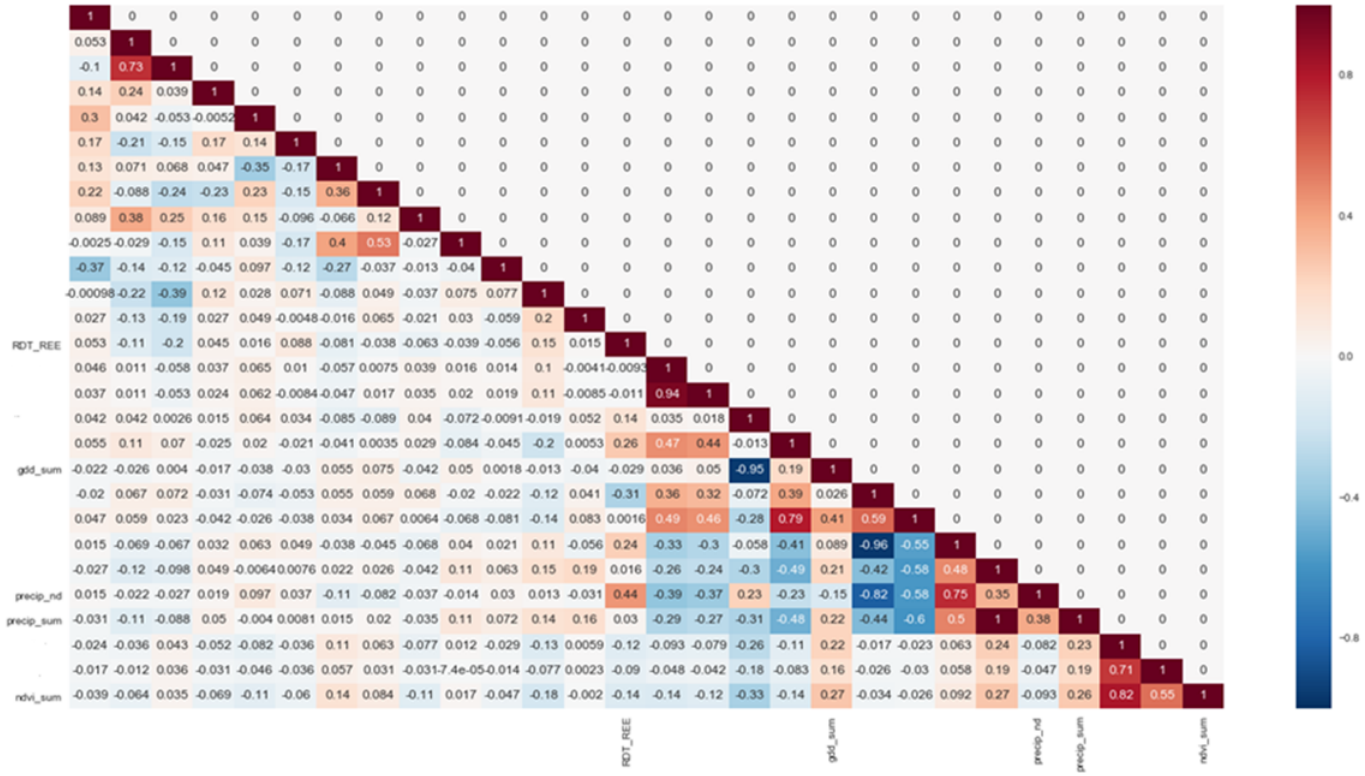

FIGURE 28 – Matrice de Corrélation

De cette matrice de corrélation, nous observerons que notre variable d'intérêt RDT_REE présente par exemple une forte corrélation avec la variable $precip\_nd$ correspondant au nombre de jours où il a pu pleuvoir (cf. Equation 38) tandis que gdd_sum (cf. Equation 39) est quasiment nulle et $ndvi\_sum$ très faible. Néanmoins, nous constatons dans l'ensemble des variables peu de corrélations linéaires. En effet, un nombre important de variables présentent un coefficient de pearson nul. Cette remarque nous conduira alors à s'intéresser à de possibles corrélations non-linéaires au travers de l'*information mutuelle*.

6.1.2 L'information mutuelle

A l'inverse, nous nous intéresserons cette fois aux possibles relations non-linéaires dont une matrice de corrélation similaire à celle présentée sera déterminée. Néanmoins, la forme de cette information mutuelle $\mathcal{I}_{(X,Y)}$ varie selon la nature de la loi jointe (X, Y) considérée. Dans un cas continu, nous nous intéresserons aux densités marginales (f_X, f_Y) et jointe $f_{(X,Y)}$:

$$\mathcal{I}_{(X,Y)} = \int_{\mathcal{Y}} \int_{\mathcal{X}} f_{(X,Y)}(x,y) ln(\frac{f_{(X,Y)}(x,y)}{f_x(x) f_Y(y)}) d\lambda(x) d\lambda(y) \tag{45}$$

Si les variables aléatoires X et Y s'avèrent être indépendantes, il adviendra alors que la loi jointe sera égale au produit des lois marginales. Quelque soit la nature de la variable d'intérêt, il s'en suivra alors que l'information sera nulle. On comprend de ce fait que l'information mutuelle sera d'autant plus grande que la variable X sera corrélée à la variable Y. On remarque que peu importe la nature des variables, l'information mutuelle sera toujours positive.

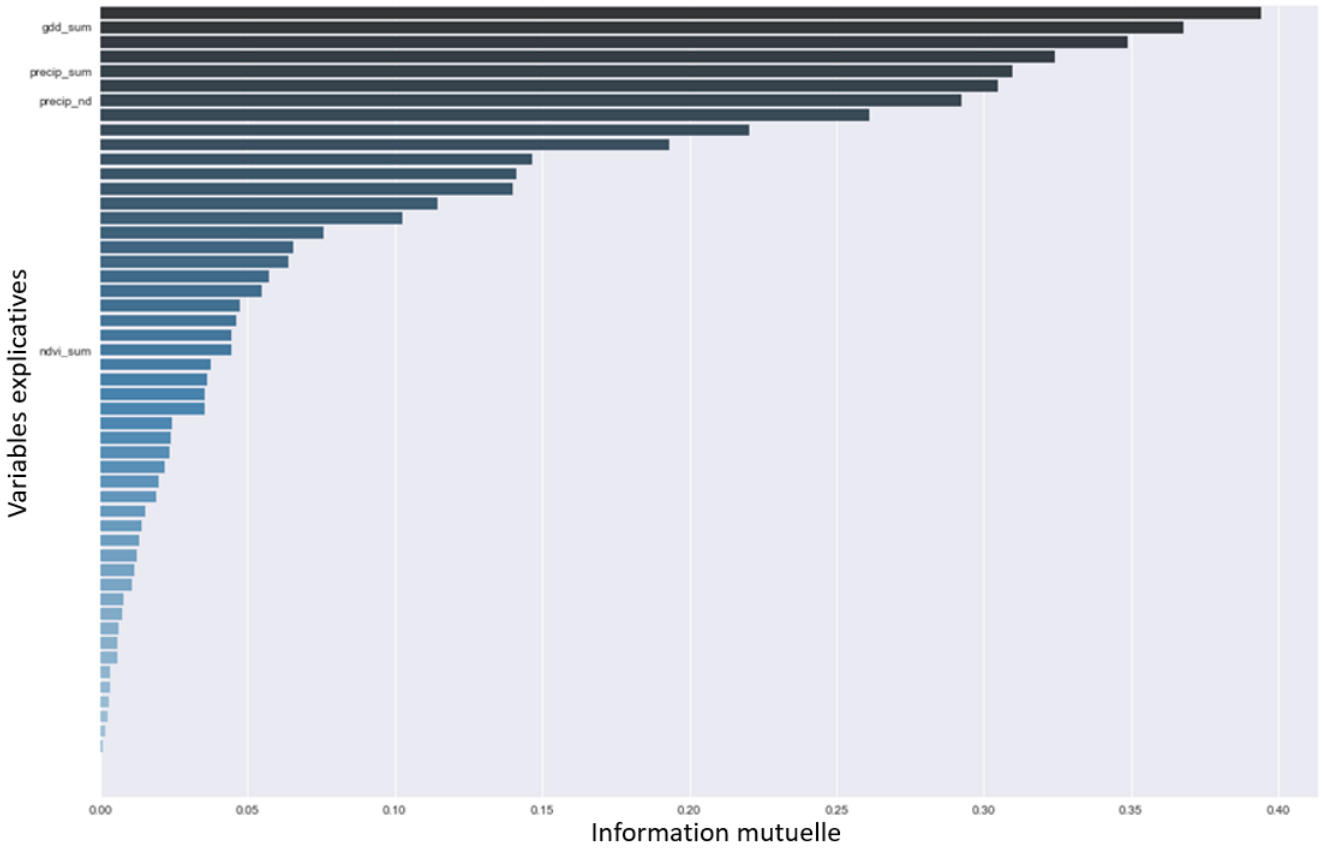

FIGURE 29 – Sélection de variable par l'information mutuelle

Cette fois-ci, nous remarquerons que la significativité de la relation entre la variable d'intérêt et les variables explicatives gdd_sum, $precip\_sum$ et $ndvi\_sum$ s'est considérablement accrue par rapport à l'anayse linéaire précédente.

6.2 Sélection par itération

Proche des méthodes de sélection par modèles abordées dans la prochaine partie, la sélection par itération a la particularité de réduire la dimension de l'espace d'entrée au cours de plusieurs étapes successives de sélection de variables. On distingue de cette sélection par itération les méthodes par *élimination récursive* et les méthodes de *sélection successive*. L'élimination récursive consiste à construire un modèle à partir de l'ensemble d'entrée initiale où l'on retirera successivement les variables explicatives les moins importantes. Afin de mieux illustrer la différence avec une méthode de sélection successive, nous nous rapporterons au schéma suivant :

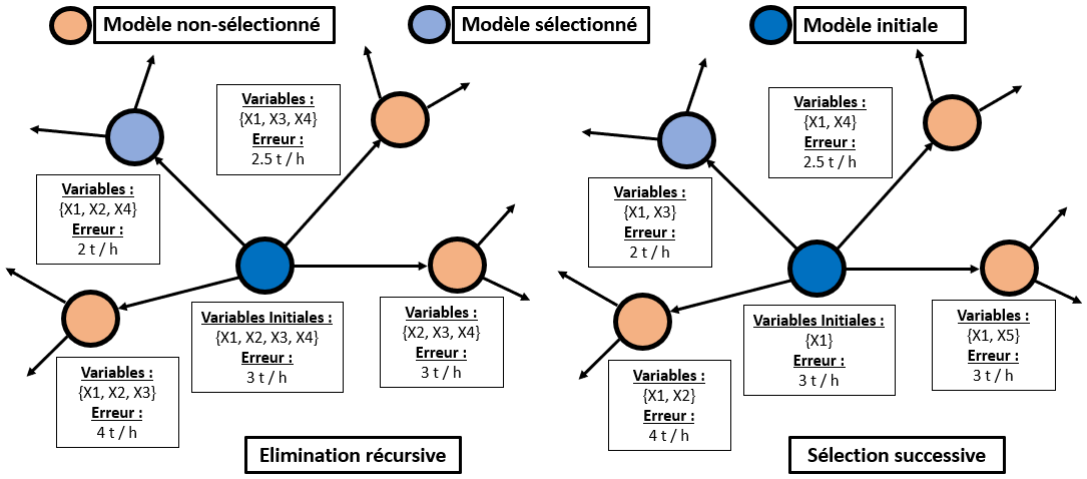

FIGURE 30 – Séléction de variables par itération

Il est important de souligner que la méthode de sélection successive n'est pas adaptée pour les modèles non-paramétriques fondés sur le principe de partition. En effet, si l'on prend par exemple le cas de l'arbre de décision, cette méthode de sélection de variables débutera par une unique variable pour laquelle un nombre important de seuils sera définies. L'initialisation de cette méthode débute alors par du surapprentissage qui risque de biaiser la suite de la procédure. Ainsi, nous avons préconisé l'approche successive lors de la modélisation paramétrique et l'approche éliminatoire lors de la modélisation non-paramétrique.

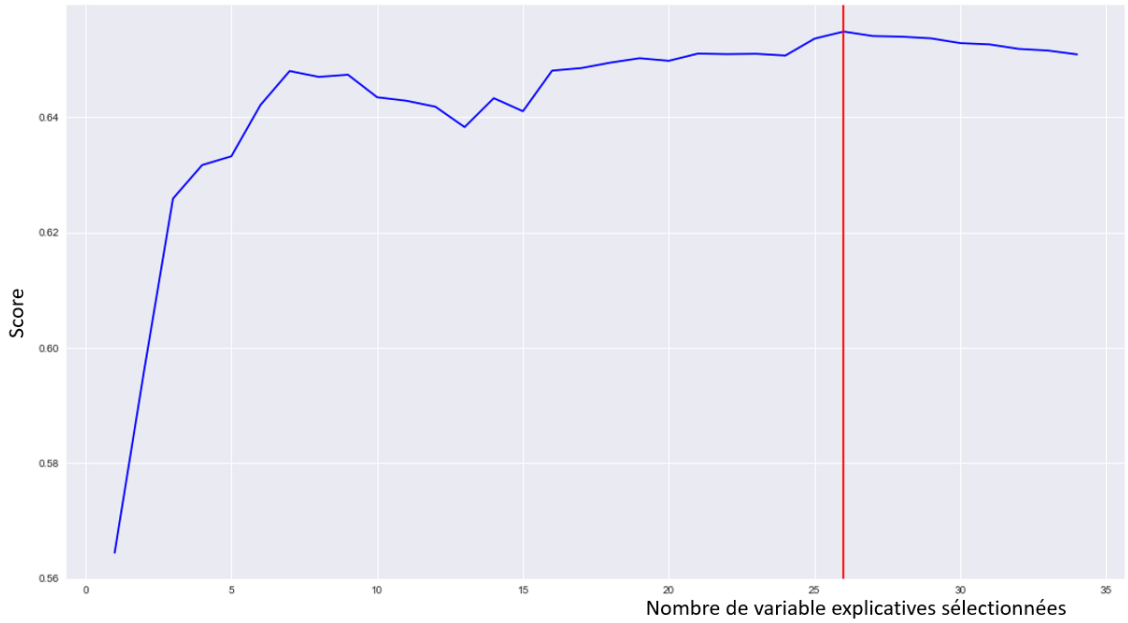

FIGURE 31 – Sélection de variables par élimination récursive

Troisième partie

Modélisation de l'indice et tarification du produit

Dans cette dernière partie, nous nous intéresserons à la modélisation de l'indice de rendement à partir des données climatiques, agricoles et satellitaires. Le but étant d'expliquer une relation aléatoire à partir de données, nous serons amenés à détailler ce que l'on appelle communément la théorie de l'apprentissage supervisé. Dans un cadre d'apprentissage supervisé, nous observons un ensemble d'apprentissage $\mathcal{D}_n = (D_i)_{i\in[|1,n|]}$ composé de n individus dont nous distinguerons de ces individus $D_i = (X_i, Y_i)$, la variable à expliquer $Y \in \mathcal{Y}$ et la variable explicative $X \in \mathcal{X}$ avec $\mathcal{D} = \mathcal{X} * \mathcal{Y}$.

Plusieurs approches peuvent être envisagées selon l'hypothèse émise sur la loi de probabilité ayant engendrée ces données. Ainsi, nous développerons les différentes approches employées lors de la modélisation de l'indice de rendement potentiel dont nous présenterons aussi les résultats obtenus. Nous soulignerons que, l'objectif étant de construire un produit d'assurance paramétrique, la transparence de ce dernier nous limitera à des approches interprétables. Cette remarque nous incitera alors à construire un modèle linéaire généralisé et un arbre de décision.

Pour finir, une fois la modélisation de l'indice de référence achevée, nous serons amenés à mettre au point une méthode de tarification adaptée au problème et au modèle considéré. L'approche linéaire étant limitée, la tarification reposera sur les données présentes dans les feuilles de l'arbre de décision. Néanmoins, la couverture étant dynamique, nous constaterons qu'il est bien difficile d'évaluer la performance de la solution proposée.

7 Modélisation de l'indice de rendement

7.1 Théorie de l'apprentissage supervisé

7.1.1 Le problème d'apprentissage

Afin de caractériser l'aspect aléatoire de la relation liant Y à X, nous considérons que les données ont été engendrées par une certaine loi de probabilité $\mathcal{L}$ caractérisée par un paramètre $\theta \in \Theta$. Par la suite, nous distinguerons l'apprentissage supervisé paramétrique qui consiste à émettre une hypothèse sur la loi $\mathcal{L}$ et l'apprentissage non-paramétrique qui n'en émet aucune. Comme on a pu le voir dans la partie précédente, l'hypothèse $(\mathcal{H}_{\mathcal{L}})$ est en général soigneusement étudiée à l'aide de méthodes statistiques issues de l'apprentissage non-supervisé. Néanmoins, afin de prendre en compte l'hétérogénéité des individus $D_i = (X_i, Y_i)$, nous supposerons ces individus d'être des réalisations indépendantes et identiquements distribuées :

$$(\mathcal{H}_{\mathcal{L}}) \quad Y_i|X_i = x \rightsquigarrow \mathcal{L}(\theta_i) \quad i.i.d \quad \text{où} \quad \theta_i \in \Theta \tag{46}$$

L'objectif de l'apprentissage supervisé est de prévoir Y associée à toute nouvelle entrée X, sous-entendu être une nouvelle réalisation de la loi $\mathcal{L}$, indépendante des réalisations précédemment observées. Dans une telle démarche, nous chercherons alors une fonction $\mathcal{F} \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$ qui a toute entrée X associe une prédiction $\hat{Y} = \mathcal{F}(X)$:

$$(\mathcal{P}_{\mathcal{F}}) \quad \mathcal{F} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathcal{Y} \\ X & \longmapsto & \hat{Y} \end{array} \right. \tag{47}$$

Cette fonction de prédiction engendrera une possible perte qui sera une mesure de la qualité de prédiction. Afin de mesurer cette perte, il est alors nécessaire d'introduire une pseudo-distance $l(.)$ qui nous permettra de mesurer la différence entre une prédiction $\hat{Y} = \mathcal{F}(X)$ et une supposée réelle observation Y. Selon que l'espace de sortie $\mathcal{Y}$ est au plus dénombrable (i.e finie ou dénombrable) ou simplement non-dénombrable, on parlera de problème de classification $(\mathcal{H}_{\mathcal{Y}})$ ou de régression $(\bar{\mathcal{H}}_{\mathcal{Y}})$:

$$(\mathcal{H}_{\mathcal{Y}}) \quad | \quad card(\mathcal{Y}) < \infty \quad \text{ou} \quad \exists f : \mathbb{N} \longrightarrow \mathcal{Y}; \forall y \in \mathcal{Y}, \exists! n \in \mathbb{N}; y = f(n) \tag{48}$$

Cependant, selon l'hypothèse considérée, une fonction de perte bien spécifique au problème d'apprentissage devra être définie. Selon que le problème d'apprentissage soit une régression ou une classification, cette fonction de perte aura une certaine forme :

$$(\bar{\mathcal{H}}_{\mathcal{Y}}) \quad | \quad l_p : \left| \begin{array}{ccc} \mathcal{Y} * \mathcal{Y} & \longrightarrow & \mathbb{R} \\ (Y, \hat{Y}) & \longmapsto & |Y - \hat{Y}|^p \end{array} \right. \tag{49}$$

ou dans un problème de classification :

$$(\mathcal{H}_{\mathcal{Y}}) \quad | \quad l : \left| \begin{array}{ccc} \mathcal{Y} * \mathcal{Y} & \longrightarrow & \{0, 1\} \\ (Y, \hat{Y}) & \longmapsto & \mathbb{1}_{Y \neq \hat{Y}} \end{array} \right. \tag{50}$$

On remarquera que dans le cas d'une régression, plusieurs fonctions de perte $l_p(.)$ $(p \geq 1 fixe)$ peuvent être considérées, il n'en reste pas moins qu'il existe une relation d'équivalence entre ces dernières ce qui confortera le choix porté sur une fonction de perte quadratique (i.e p=2). Dans ce cas précis, on parle en général de régression aux moindres carrés.

7.1.2 Prédicteur de Bayes

La qualité d'une fonction de prédiction est mesurée par son risque $\mathcal{R}_{\mathcal{L}}$ (ou *erreur de généralisation*) qui est l'espérance par rapport à la loi $\mathcal{L}$ de la perte encourue sur la donnée (X, Y) :

$$\mathcal{R}_{\mathcal{L}} : \left| \begin{array}{rcl} \mathcal{F}(\mathcal{X}, \mathcal{Y}) & \longrightarrow & \mathbb{R} \\ \mathcal{F} & \longmapsto & \mathbb{E}_{\mathcal{L}}[l(Y, \mathcal{F}(X)] \end{array} \right. \tag{51}$$

Dans un tel cadre, le problème $(\mathcal{P}_{\mathcal{F}})$ est donc de trouver la fonction $\mathcal{F}_*^{\mathcal{L}} \in \mathcal{F}(\mathcal{X}, \mathcal{Y})$, appelée fonction oracle ou prédicteur de Bayes, minimisant le risque :

$$(\mathcal{P}_{\mathcal{F}}) \quad \mathcal{F}_*^{\mathcal{L}} \in \underset{\mathcal{F} \in \mathcal{F}(\mathcal{X}, \mathcal{Y})}{argmin} \mathcal{R}_{\mathcal{L}}[\mathcal{F}] \tag{52}$$

Dans un cadre d'apprentissage stochastique, l'objectif est alors d'estimer pour chacun des stades phénologiques successifs une fonction de prédiction $\mathcal{F}_t(.) \in \mathcal{F}(\mathcal{X}_t, \mathcal{Y})$ dont la construction de l'ensemble stochastique fût détaillée dans la partie précédente (cf. Equation 43). Dans un tel cadre, l'objectif est de trouver la fonction oracle pour chacune des périodes relatives au cycle de production $\mathcal{T} = [|1, T|]$ de la culture considérée ce qui revient à considérer le problème suivant :

$$(\mathcal{P}_{\mathcal{F}}) \quad \forall t \in \mathcal{T}, \mathcal{F}_t^{\mathcal{L}} \in \underset{\mathcal{F}_t \in \mathcal{F}(\mathcal{X}_t, \mathcal{Y})}{argmin} \mathcal{R}_{\mathcal{L}}[\mathcal{F}_t] \quad ; \quad \forall t' \leq t, \quad \mathcal{X}_{t'} \subset \mathcal{X}_t \tag{53}$$

Nous montrerons (cf. Annexe [32]) que s'il existe une fonction $\mathcal{F}_*^{\mathcal{L}}(.)$ minimisant le risque pour n'importe quel point de l'espace d'entrée $\mathcal{X}$ alors elle minimisera aussi l'erreur de généralisation :

$$\forall x \in \mathcal{X}; \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y \in \mathcal{Y}}{argmin}\ \mathbb{E}_{\mathcal{L}}[l(Y, y)|X = x] \Longrightarrow \mathcal{F}_*^{\mathcal{L}} \in \underset{\mathcal{F} \in \mathcal{F}(\mathcal{X}, \mathcal{Y})}{argmin} \mathcal{R}_{\mathcal{L}}[\mathcal{F}] \tag{54}$$

Cette propriété fondamentale nous permettra alors de déterminer les fonctions oracles dans un cadre de régression aux moindres carrés (cf. Annexe [33]) et dans un cadre de classification binaire (cf. Annexe [34]) :

$$(\mathcal{P}_{\mathcal{F}}) \quad | \quad \left\{ \begin{array}{ll} (\bar{\mathcal{H}}_{\mathcal{Y}}) & y_* = \bar{\mathcal{F}}_*^{\mathcal{L}}(x) = \mathbb{E}_{\mathcal{L}}[Y|X = x] \\ (\mathcal{H}_{\mathcal{Y}}) & y_* = \mathcal{F}_*^{\mathcal{L}}(x) = \mathbb{1}_{\bar{\mathcal{F}}_*^{\mathcal{L}}(x) \geq \frac{1}{2}} \end{array} \right\} \tag{55}$$

Le but de l'apprentissage supervisé est de trouver une fonction de prédiction dont le risque est aussi faible que possible. Autrement dit, ce risque doit être aussi proche que possible du risque des prédicteurs oracles or la loi de probabilité générant les données pouvant être supposée inconnue, les prédicteurs oracles et le risque sont alors eux aussi inconnus. Néanmoins, le risque peut être estimé par son équivalent empirique $\hat{\mathcal{R}}_n[\mathcal{F}]$:

$$\hat{\mathcal{R}}_n[\mathcal{F}] = \frac{1}{n} \sum_{i \in [|1,n|]} l(Y_i, \mathcal{F}_n^{\mathcal{L}}(X_i)) \tag{56}$$

Si nous supposons que le risque de la fonction de prédiction admet un moment d'ordre 2, alors la loi forte des grands nombres et le théorème central limite permettent d'affirmer que :

$$\hat{\mathcal{R}}_n[\mathcal{F}] \underset{n \longrightarrow \infty}{\longrightarrow} \mathcal{R}_{\mathcal{L}}[\mathcal{F}] \quad \text{ET} \quad \sqrt{n} * \hat{\mathcal{R}}_n[\mathcal{F}] \underset{n \longrightarrow \infty}{\longrightarrow} \mathcal{N}(\mathcal{R}_{\mathcal{L}}[\mathcal{F}], \mathbb{V}_{\mathcal{L}}[l(Y, \mathcal{F}(X)]) \tag{57}$$

32. Minimisation de l'erreur de généralisation
33. Problème de régression aux moindres aux carrés
34. Problème de classification binaire

Puisque nous cherchons une fonction qui minimise l'erreur de généralisation approchée par le risque empirique $\hat{\mathcal{R}}_n[.]$, il est naturel de considérer le problème d'apprentissage de *minimisation du risque empirique* :

$$(\mathcal{P}_{\mathcal{F}}) \quad \mathcal{F}_n^{\mathcal{L}} \in \underset{\mathcal{F}\in\mathcal{F}(\mathcal{X},\mathcal{Y})}{argmin}\, \hat{\mathcal{R}}_n[\mathcal{F}] \tag{58}$$

Dans un problème de régression aux moindres carrés ou de classification binaire, nous serons alors amenés à résoudre le problème suivant :

$$(\mathcal{P}_{\mathcal{F}}) \quad | \quad \left\{ \begin{array}{ll} (\bar{\mathcal{H}}_{\mathcal{Y}}) & \mathcal{F}_n^{\mathcal{L}} \in \underset{\mathcal{F}\in\mathcal{F}(\mathcal{X},\mathcal{Y})}{argmin}\ \frac{1}{n}\sum_{i\in[|1,n|]}(Y_i - \mathcal{F}(X_i))^2 \\ (\mathcal{H}_{\mathcal{Y}}) & \mathcal{F}_n^{\mathcal{L}} \in \underset{\mathcal{F}\in\mathcal{F}(\mathcal{X},\mathcal{Y})}{argmin}\ \frac{1}{n}\sum_{i\in[|1,n|]}\mathbb{1}_{Y_i\neq\mathcal{F}(X_i)} \end{array} \right\} \tag{59}$$

7.1.3 Dilemme de biais-variance

Ce n'est pas une bonne idée de prendre $\mathcal{F}(\mathcal{X},\mathcal{Y})$ car cela entraîne un problème de choix puisqu'en général, pour tout ensemble d'apprentissage $\mathcal{D}_n$, il existe une infinité de fonctions de prédiction minimisant le risque empirique. Cela mène en général à un surapprentissage dans la mesure où la fonction résultant a un risque empirique qui peut être très inférieure à son risque réel. En pratique, il faut considérer un sous-ensemble $\hat{\mathcal{F}}(\mathcal{X},\mathcal{Y}) \subset \mathcal{F}(\mathcal{X},\mathcal{Y})$ suffisamment grand pour pouvoir raisonnablement approcher toute fonction tout en ne le prenant pas trop grand pour éviter de surapprendre. Soit $\hat{\mathcal{F}}_*^{\mathcal{L}}(.)$ la fonction oracle minimisant le risque sur $\hat{\mathcal{F}}(\mathcal{X},\mathcal{Y})$, on obtient alors l'inégalité générale suivante :

$$\mathcal{R}_{\mathcal{L}}[\hat{\mathcal{F}}_n^{\mathcal{L}}] \geq \mathcal{R}_{\mathcal{L}}[\hat{\mathcal{F}}_*^{\mathcal{L}}] \geq \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}] \tag{60}$$

Par cette inégalité (cf. Equation 60), nous remarquerons que l'excès de risque de la fonction de prédiction empirique se décompose respectivement en deux termes positifs, appelés *erreur stochastique* (ou variance) et *erreur systématique* (ou biais). Plus $\hat{\mathcal{F}}(\mathcal{X},\mathcal{Y})$ est grand, plus le biais est faible mais plus la variance est grande :

$$\mathcal{R}_{\mathcal{L}}[\hat{\mathcal{F}}_n^{\mathcal{L}}] - \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}] = (R_{\mathcal{L}}[\hat{\mathcal{F}}_n^{\mathcal{L}}] - \mathcal{R}_{\mathcal{L}}[\hat{\mathcal{F}}_*^{\mathcal{L}}]) + (R_{\mathcal{L}}[\hat{\mathcal{F}}_*^{\mathcal{L}}] - \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}]) \tag{61}$$

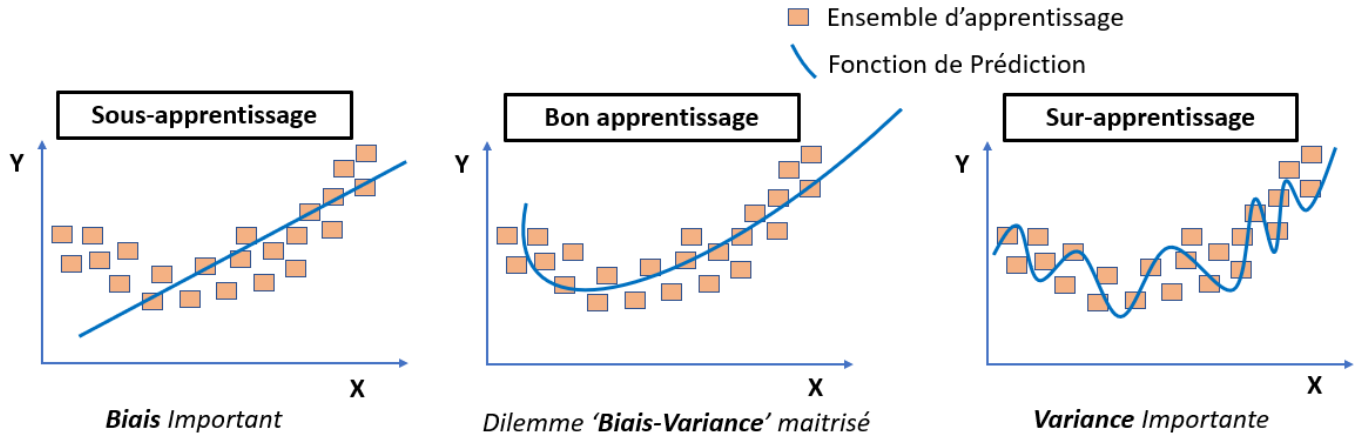

FIGURE 32 – Dilemme de Biais-Variance

Il y a donc un compromis à trouver dans le choix de $\hat{\mathcal{F}}(\mathcal{X},\mathcal{Y})$, problème souvent appelé *dilemme biais−variance*. En choisissant bien l'ensemble de prédicteurs, il est alors possible de construire un prédicteur avec une faible variance tout en minimisant le risque empirique. Le choix de cet ensemble repose sur un a priori de la relation à expliquer.

7.1.4 Robustesse de l'estimation

Quel que soit l'approche préconisée par la suite, qu'elle soit paramétrique ou non, il n'en reste pas moins que pour apprécier la robustesse de la fonction de prédiction estimée, nous serons amenés à diviser notre base de donnée initiale $\mathcal{D}_n$ en une base d'apprentissage $\mathcal{D}_n^{\mathcal{A}} \subset \mathcal{D}_n$ et une base de test $\mathcal{D}_n^{\mathcal{T}} \subset \mathcal{D}_n$. Cette répartition des données entre ces deux bases est un choix qui repose en général sur la grandeur des données disponibles n, cependant il est commun de considérer la population n en base 100 et ainsi considérer $\mathcal{D}_n^{\mathcal{T}} = 10\%\mathcal{D}_n$ lors de la phase de test et $\mathcal{D}_n^{\mathcal{A}} = 90\%\mathcal{D}_n$ lors de la phase d'apprentissage. De plus, dans un problème de régression, il est commun de se référer à certaines mesures spécifiques tel que l'*erreur moyenne quadratique* $\mathcal{RMSE}$. Tout comme la fonction de perte, l'objectif est d'obtenir une erreur moyenne la plus faible possible.

$$\mathcal{RMSE} = \sqrt{\hat{\mathcal{R}}_n[\mathcal{F}^{\mathcal{L}}]} = \sqrt{\frac{1}{n}\sum_{i\in\mathcal{D}_n^{\mathcal{T}}} l(Y_i, \mathcal{F}(X_i))} = \sqrt{\frac{1}{n}\sum_{i\in\mathcal{D}_n^{\mathcal{T}}} (Y_i - \mathcal{F}(X_i))^2} \tag{62}$$

Enfin, nous introduirons une seconde mesure tout aussi utile pour déterminer la qualité d'une fonction de prédiction en s'intéressant cette fois à la dispersion du nuage de points par rapport à une fonction de prédiction. On remarquera que cette mesure $\mathcal{R}_2$, appelée *coefficient de détemination*, converge vers 1 lorsque l'erreur diminue $\mathcal{F}(X_i) \approx Y_i$:

$$\mathcal{R}_2 : \left| \begin{array}{rcl} \mathcal{F}(\mathcal{X}, \mathcal{Y}) & \longrightarrow & \mathbb{R} \\ \mathcal{F} & \longmapsto & \frac{\sum_{i\in\mathcal{D}_n^{\mathcal{T}}}(\bar{Y}_n - \mathcal{F}(X_i))^2}{\sum_{i\in\mathcal{D}_n^{\mathcal{T}}}(\bar{Y}_n - Y_i)^2} = 1 - \frac{\sum_{i\in\mathcal{D}_n^{\mathcal{T}}}(Y_i - \mathcal{F}(X_i))^2}{\sum_{i\in\mathcal{D}_n^{\mathcal{T}}}(\bar{Y}_n - Y_i)^2} \end{array} \right. \tag{63}$$

Etant dans un cadre dynamique, nous présenterons les mesures présentées plus haut à chacun des stades phénologiques pour différentes approches explicitées par la suite :

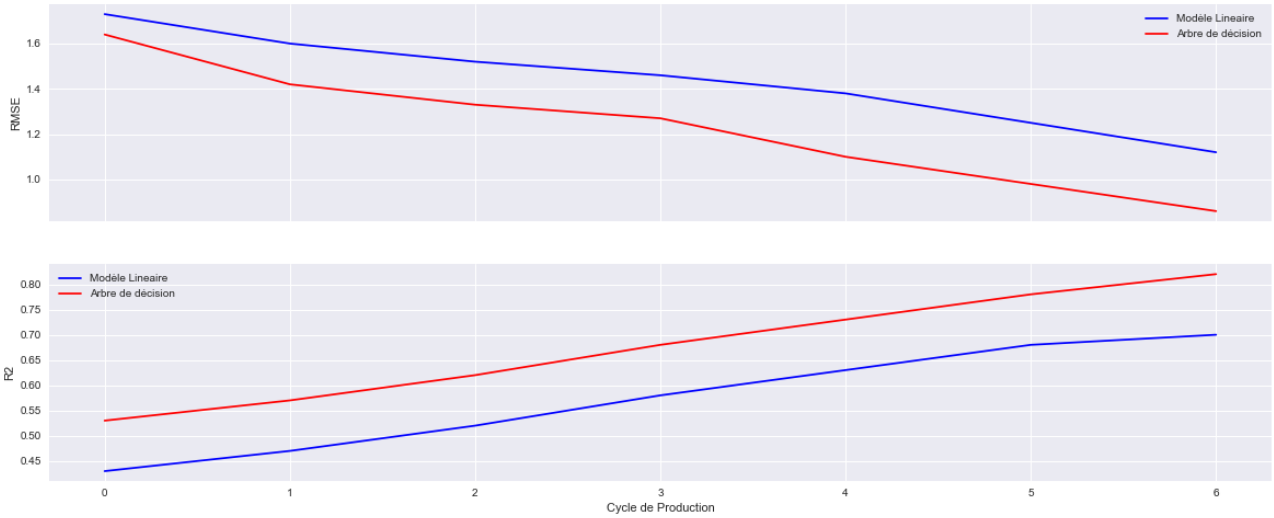

FIGURE 33 – Robustesse de l'approche paramétrique et non-paramétrique

Comme on pouvait s'y attendre, au vu des relations non-linéaires, l'arbre de décision présente de meilleurs résultats qu'une simple approche linéaire. De plus, on remarquera que la vitesse de convergence de l'arbre de décision est bien supérieure à celle de l'approche linéaire au vu des relations qui se complexifient au fil des stades phénologiques. Peu importe l'approche retenue, on remarque qu'au fil des périodes, l'erreur type se réduit tandis que le coefficient de détermination s'améliore. Cette évolution est aussi un indicateur pertinent afin de mesurer la qualité des traitements opérées en amont sur la base de données.

7.2 Apprentissage paramétrique

7.2.1 La fonction de lien

L'approche paramétrique consiste à réduire l'ensemble possible des fonctions de prédiction $\mathcal{F}(\mathcal{X}, \mathcal{Y})$ par l'émission d'une hypothèse concernant la loi $\mathcal{L}$ qui a engendrée les données $(D_i)_{i\in[|1,n|]}$. Par souci d'homoscédasticité, il est primordial de définir une fonction de lien $\mathcal{G}_{\mathcal{L}}(.)$ permettant de relier le paramètre inconnu de la loi θ et la prédiction $\hat{Y}$ effectuée :

$$\mathcal{G}_{\mathcal{L}} : \left| \begin{array}{ccc} \mathcal{Y} & \longrightarrow & \Theta \\ \hat{Y} & \longmapsto & \hat{\theta} \end{array} \right. \tag{64}$$

Ainsi, sous l'hypothèse que la loi est connue et sous l'hypothèse que la fonction de lien soit inversible, nous pouvons alors redéfinir $(\mathcal{H}_{\mathcal{L}})$ de la manière suivante :

$$(\mathcal{H}_{\mathcal{L}}^{\Theta}) \quad Y_i|X_i = x \rightsquigarrow \mathcal{L}(\theta_i) \quad i.i.d \quad \text{où} \quad \theta_i = \mathcal{G}_{\mathcal{L}}(\mathcal{F}_*^{\mathcal{L}}(x)) \Longleftrightarrow \mathcal{F}_*^{\mathcal{L}}(x) = \mathcal{G}_{\mathcal{L}}^{-1}(\theta_i) \tag{65}$$

Cette relation met en lumière l'importance d'approcher convenablement la loi dont on déterminera par la suite, la fonction de lien. De nombreuses méthodes issues de l'apprentissage non-supervisé permettent de déterminer la véracité de l'hypothèse faite sur la loi de probabilité. De l'estimation effectuée en première partie, nous venons de restreindre l'ensemble des fonctions de prédictions par une hypothèse sur $\mathcal{L} = \mathcal{N}(\mu, \sigma^2)$.

Néanmoins, il est nécessaire de spécifier une forme à la fonction afin d'estimer la sortie à partir de l'entrée. Des approches à la fois linéaires et non-linéaires peuvent être envisagées dans une approche de modélisation paramétrique. Ainsi, une fois la loi déterminée, une fonction $\mathcal{F}_*^{\Theta}(.)$ est ensuite approximée afin d'estimer le paramètre :

$$\mathcal{F}_*^{\Theta} : \left| \begin{array}{ccc} X & \longrightarrow & \Theta \\ X & \longmapsto & \hat{\theta} \end{array} \right. \tag{66}$$

Dans un tel cadre d'apprentissage, il en vient alors une nouvelle relation de prédiction :

$$\text{sous} \quad (\mathcal{H}_{\mathcal{L}}^{\Theta}) \Longrightarrow \mathcal{F}_*^{\mathcal{L}}(x) = \mathcal{G}_{\mathcal{L}}^{-1}(\theta_i) = \mathcal{G}_{\mathcal{L}}^{-1}(\mathcal{F}_*^{\Theta}(x)) \tag{67}$$

7.2.2 Le modèle linéaire généralisé

Pour débuter l'étude, nous présenterons l'approche linéaire communément appelée Modèle Linéaire Généralisé dont la particularité est d'approcher le paramètre par une fonction linéaire :

$$\text{sous} \quad (\mathcal{H}_{\mathcal{L}}^{\Theta}) \quad \Longrightarrow \mathcal{F}_*^{\Theta}(x) = \beta_0 + \sum_{j\in[|1,p|]} \beta_j x_j = x'\beta = \theta \tag{68}$$

Afin de se ramener à la fonction de prédiction $\mathcal{F}_*^{\mathcal{L}}(.)$, il est important de composer le paramètre par l'inverse de la fonction de lien $\mathcal{G}_{\mathcal{L}}^{-1}(.)$ spécifique à la loi considérée :

$$(\mathcal{H}_{\mathcal{L}}^{\Theta}) \Longrightarrow \mathcal{F}_*^{\mathcal{L}}(x) = \mathbb{E}_{\mathcal{L}}[Y|X = x] = \mathcal{G}_{\mathcal{L}}^{-1}(\theta) = \mathcal{G}_{\mathcal{L}}^{-1}(\mathcal{F}_*^{\Theta}(x)) = \mathcal{G}_{\mathcal{L}}^{-1}(x'\beta) \tag{69}$$

Comme nous avons pu le voir en première partie, nous nous sommes attardés à la famille exponentielle (cf. Annexe [35]), et plus particulièrement à la loi normale. Dans notre étude, nous nous intéresserons exclusivement à la loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on identifiera la fonction de lien $\mathcal{G}_{\mathcal{L}}(x) = x = \theta$.

35. Famille Exponentielle

7.2.3 Limites de l'approche paramétrique

Cette hypothèse sur la loi de probabilité permet la formulation d'un paramètre à estimer que l'on peut au moyen d'une certaine relation approximer à l'aide d'une formule fermée. Cette particularité d'émettre ainsi des formules fermées, propre à la modélisation paramétrique, lui accorde une place privilégiée dans une étude statistique à des fins commerciales. En effet, il apparaît crucial autant pour l'assureur que l'assuré que le modèle sous-jacent à la tarification soit interprétable. Parfois, et tout particulièrement lors de l'indemnisation, la régulation en place exige en général la formulation d'une formule fermée afin de rendre compte de son impact. Dans un cadre de transparence, l'approche linéaire peut donc être préconisée lors de la modélisation de l'indice de rendement en fin de cycle de production, lorsque ce dernier est considéré comme réel.

Cependant, l'apprentissage paramétrique repose très souvent sur un prédicteur linéaire qui en général ne correspond pas à la réalité et au problème donné. Cette restriction linéaire implique une relation tout aussi linéaire entre la variable d'intérêt et la variable explicative représentée par un coefficient β_j. Or il est probable qu'une même variable ait des effets contraires selon qu'elle soit plus ou moins élevée, comme une température par exemple. D'autres modèles, dît additifs généralisés permettent de rester dans un cadre paramétrique tout en relâchant cette hypothèse linéaire.

Néanmoins, comme on a pu le voir lors de la caractérisation de la loi de probabilité, l'hypothèse paramétrique portée sur cette loi peut engendrer une erreur d'estimation qui entraînera par conséquent un biais à la modélisation qui s'en suit. C'est d'ailleurs cette restriction qui nous incitera par la suite à développer des modèles dit non-paramétrique dont la particularité première est de lever cette hypothèse sur la loi qui sera alors considérée comme inconnue. Néanmoins, étant dans une étude à but commerciale, nous nous limiterons à des méthodes interprétables tel que l'arbre de décision. Ainsi, nous n'aborderons pas dans l'étude les méthodes plus complexes difficilement interprétables.

7.3 Apprentissage non-paramétrique

7.3.1 Procédure de validation croisée

L'apprentissage non-paramétrique a cette spécificité de considérer la loi $\mathcal{L}$ qui a engendrée l'ensemble d'apprentissage $\mathcal{D}_n = (D_i)_{i\in[|1,n|]}$ inconnue. L'ensemble des méthodes d'apprentissage non-paramétrique que nous développerons par la suite se fonde sur le principe de partition obtenu à l'aide d'algorithme dont nous serons amenés à estimerer les paramètres. Pour rappel, nous souhaitons mettre en œuvre un modèle linéaire ainsi qu'un arbre de décision. Si le découpage de la base initiale $\mathcal{D}_n$ en un ensemble d'apprentissage $\mathcal{D}_n^{\mathcal{A}}$ et une base de test $\mathcal{D}_n^{\mathcal{T}}$ est suffisant pour construire le modèle paramétrique, ce n'est pas le cas des algorithmes d'apprentissage non-paramétrique. En effet, ces derniers nécessitent en général une procédure dite de *validation croisée* dont l'objectif est de déterminer le paramètre optimal. Afin d'éviter un possible surapprentissage, cette recherche de paramètre s'effectue sur une tierce base de données dite de validation $\mathcal{D}_n^{\mathcal{V}}$. Dans un tel cadre nous serons alors amené à séparer notre base de donnée initiale en une base :

1. *d'apprentissage* qui permet de construire le modèle linaire et l'arbre de décision
2. de *validation* qui permet d'obtenir le paramètre optimum θ_* de l'arbre
3. de *test* qui permet d'estimer les erreurs commises pour chacune des approches

Néanmoins, construire la base de validation distinctement de la base d'apprentissage engendrerait un manque de données pour l'approche linéaire dans le cas où l'on souhaiterait les comparer de manière équitable. Plusieurs méthodes de validation-croisée plus ou moins complexes existent dont leurs particularités reposent essentiellement sur le principe de partition. L'idée générale étant alors de constuire une multitude de modèles pour un même paramètre sur des ensembles d'apprentissages différents. Cette méthode permet alors d'obtenir une multitude de modèles pour un paramètre donnée auquel une erreur est par la suite estimée :

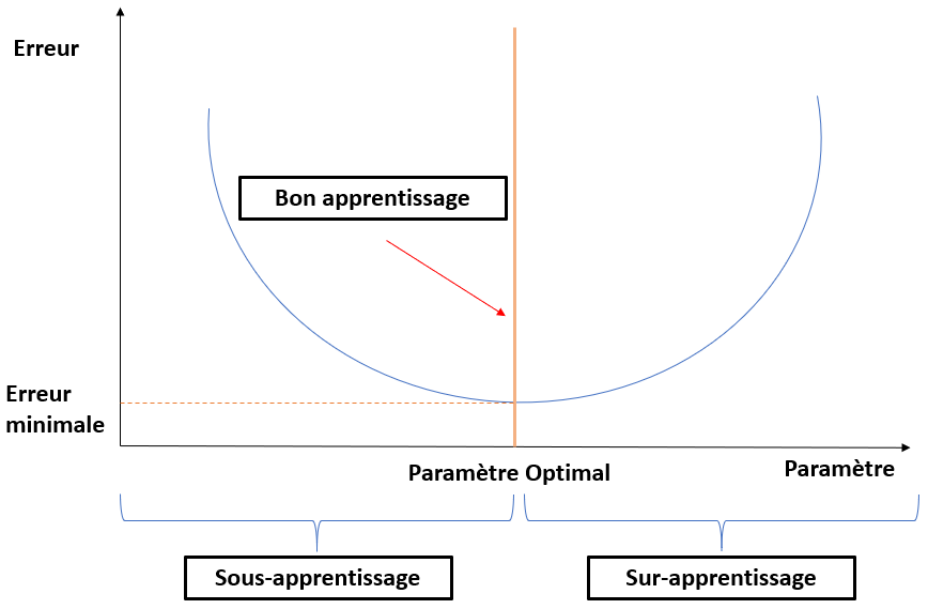

FIGURE 34 – Procédure de Validation Croisée

Dans notre étude, nous nous intéresserons à la procédure de validation croisée (cf. Annexe [36]) appelée $k-Fold$, dont la particularité est de scinder l'ensemble d'apprentissage en une multitude où seul l'un d'entre eux servira de base de validation.

36. Procédure de Validation Croisée : k-Fold

7.3.2 Consistance du prédicteur

Cette approche empirique de l'approximation nous amène à se poser la question de la consistance d'une approche d'apprentissage dont la qualité dépend donc grandement de la taille de l'échantillon n. La qualité d'une prédiction étant définie comme son risque, on s'intéressera alors son espérance par rapport à l'ensemble d'apprentissage $\mathcal{D}_n$ considéré comme aléatoire. On dira alors qu'un algorithme d'apprentissage est *consistant* par rapport à la loi $\mathcal{L}$ si et seulement si :

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_{\mathcal{L}}[\mathcal{F}_n]] \underset{n\longrightarrow\infty}{\longrightarrow} \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}]] \tag{70}$$

La loi étant supposée elle-même inconnue, on sera amené à s'intéresser à des algorithmes d'apprentissage *universellement consistant* où la consistance sera vérifiée pour toute loi de probabilité sur $\mathcal{D}$. Néanmoins, les résultats de consistance universelle ne disent pas le nombre de données nécessaires pour avoir une inégalité garantie par un $\epsilon \in \mathbb{R}_+$ fixé : $\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_{\mathcal{L}}[\mathcal{F}_n]] \leq \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}]] + \epsilon$. La consistance universelle n'affirmant que :

$$\sup_{\mathcal{L}} \lim_{n\longrightarrow\infty} \{\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_{\mathcal{L}}[\mathcal{F}_n]] - \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}]]\} = 0 \tag{71}$$

Or, pour que ce nombre existe, il faudrait avoir un résultat de *consistance universelle uniforme* :

$$\lim_{n\longrightarrow\infty} \sup_{\mathcal{L}} \{\mathbb{E}_{\mathcal{D}_n}[\mathcal{R}_{\mathcal{L}}[\mathcal{F}_n]] - \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}]]\} = 0 \tag{72}$$

Cette absence d'algorithme universellement uniformément consistant nous amène alors à définir un algorithme d'apprentissage efficient comme étant un algorithme universellement consistant et ayant une propriété de convergence uniforme sur une classe de probabilités paraissant pertinente pour le problème à traiter. Cette classe constitue dès lors un a priori implicite sur l'ensemble de fonction de prédiction. Dans cette partie, nous nous intéresserons à des classes bien particulières fondées sur le principe de partition.

Afin d'estimer cette fonction de prédiction, il est important d'émettre l'hypothèse que celle-ci est mesurable. C'est-à-dire que nous pouvons approximer tout élément possible de l'espace de sortie. Dès lors que la fonction de prédiction est mesurable, nous pouvons utiliser un théorème de la théorie de la mesure assurant que toute fonction mesurable peut être approchée par une fonction étagée dont on peut écrire sa forme générale à l'aide d'une partition $(\mathcal{X}_i)_{i\in[|1,K|]}$. Cette discrétisation par une partition nous fait d'autant plus comprendre l'intérêt de disposer d'un large échantillon où $N_i = \sum_{i\in[|1,n|]} \mathbb{1}_{\mathcal{X}_i}(X_i)$ le nombre d'observations présentes dans la partie $\mathcal{X}_i$ s'avère crucial. Cette forme particulière des fonctions étagées nous amènera alors à considérer l'ensemble des prédicteurs suivant :

$$\hat{\mathcal{F}}(\mathcal{X},\mathcal{Y}) = \{\mathcal{F}_*^{\mathcal{L}} : \mathcal{X} \longrightarrow \mathcal{Y}; \mathcal{F}_*^{\mathcal{L}}(X) = \sum_{i\in[|1,K|]} \mathcal{F}_i^{\mathcal{L}}(X)\mathbb{1}_{\mathcal{X}_i}(X)\} \tag{73}$$

Dans un problème de régression aux moindres carrés, il existe un unique minimiseur de risque empirique donnée par :

$$\mathcal{F}_*^{\mathcal{L}}(X) = \sum_{i\in[|1,K|]} \hat{Y}_{\mathcal{X}_i}\mathbb{1}_{\mathcal{X}_i}(X) \quad \text{où} \quad \mathcal{F}_i^{\mathcal{L}}(X) = \hat{Y}_{\mathcal{X}_i} = \frac{1}{N_i}\sum_{i\in[|1,n|]} Y_i\mathbb{1}_{\mathcal{X}_i}(X_i) \tag{74}$$

Dans la suite, nous nous intéresserons aux algorithmes les plus populaires comme par exemple l'algorithme des k-plus proches voisins ou encore les arbres de décision. Tout les deux utilisant des classes qui sont constituées des fonctions constantes par morceaux sur une partition. Ce qui diffère ces deux algorithmes est la façon dont la partition en question est choisie.

7.3.3 Les k-plus proches voisins

La particularité de la partition engendrée par l'algorithme des k-plus proches voisins en fait un algorithme très utile dans certain cas précis. Pour un entier $k \in [|1, n|]$ fixé, les statistiques de rang $\{r_1, ..., r_k\}$ définies précédemment (cf. Equation 26) conduisent à définir la partition suivante $\forall i \in [|1, C_n^k|]$:

$$\mathcal{X}_i = \{x \in \mathcal{X}; (r_1(x), ...r_k(x)) = c_i\} \tag{75}$$

En d'autres termes, pour un k fixé, les ensembles $\mathcal{X}_i$ sont les parties de $\mathcal{X}$ sur lesquelles l'application $x \longrightarrow \{r_1(x), ..., r_k(x)\}$ est constante. Néanmoins, pour que le prédicteur soit consistant, le paramètre doit en général croître avec la taille de l'échantillon. Les petites valeurs induisent une grande volatilité du résultat et, par conséquent, favorisent le surapprentissage. En revanche, si le paramètre est très grand, la partition en sera d'autant plus grande et le prédicteur qui en résulte peu flexible. Il faut donc trouver le paramètre garantissant une flexibilité suffisante tout en évitant le surapprentissage.

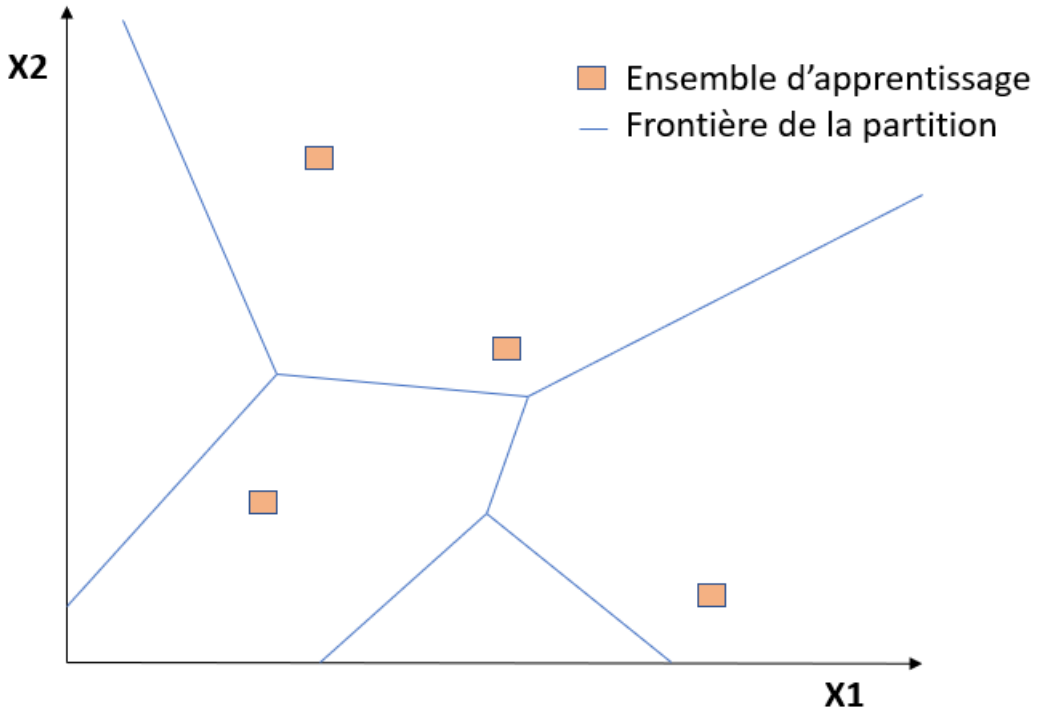

FIGURE 35 – Partition construite par l'algorithme des k-plus proches voisins

Comme on a pu l'observer, l'algorithme construit une partition uniquement à partir de l'ensemble d'entrées. Ainsi, lorsqu'il sera question de retrouver les valeurs manquantes de la variable *type de sol*, de par la spécificité de sa partition, l'algorithme des k-plus proches voisins sera priviligié. L'analyse des corrélations au sein des variables explicatives permet de mettre en lumière des relations de dépendances et contribue en général à sélectionner par la suite les variables pertinentes pour estimer ces valeurs manquantes.

Néanmoins, dans notre cas, il apparaît déjà naturel de souhaiter retrouver ces valeurs à partir de la *golocalisation satellitaire* de la parcelle agricole. Afin de mieux présenter le principe de partition décrit plus haut, nous présenterons ainsi la partition $(\mathcal{X}_i)_{i\in[\![1,C_n^k]\!]}$ obtenue pour un $k = 8$:

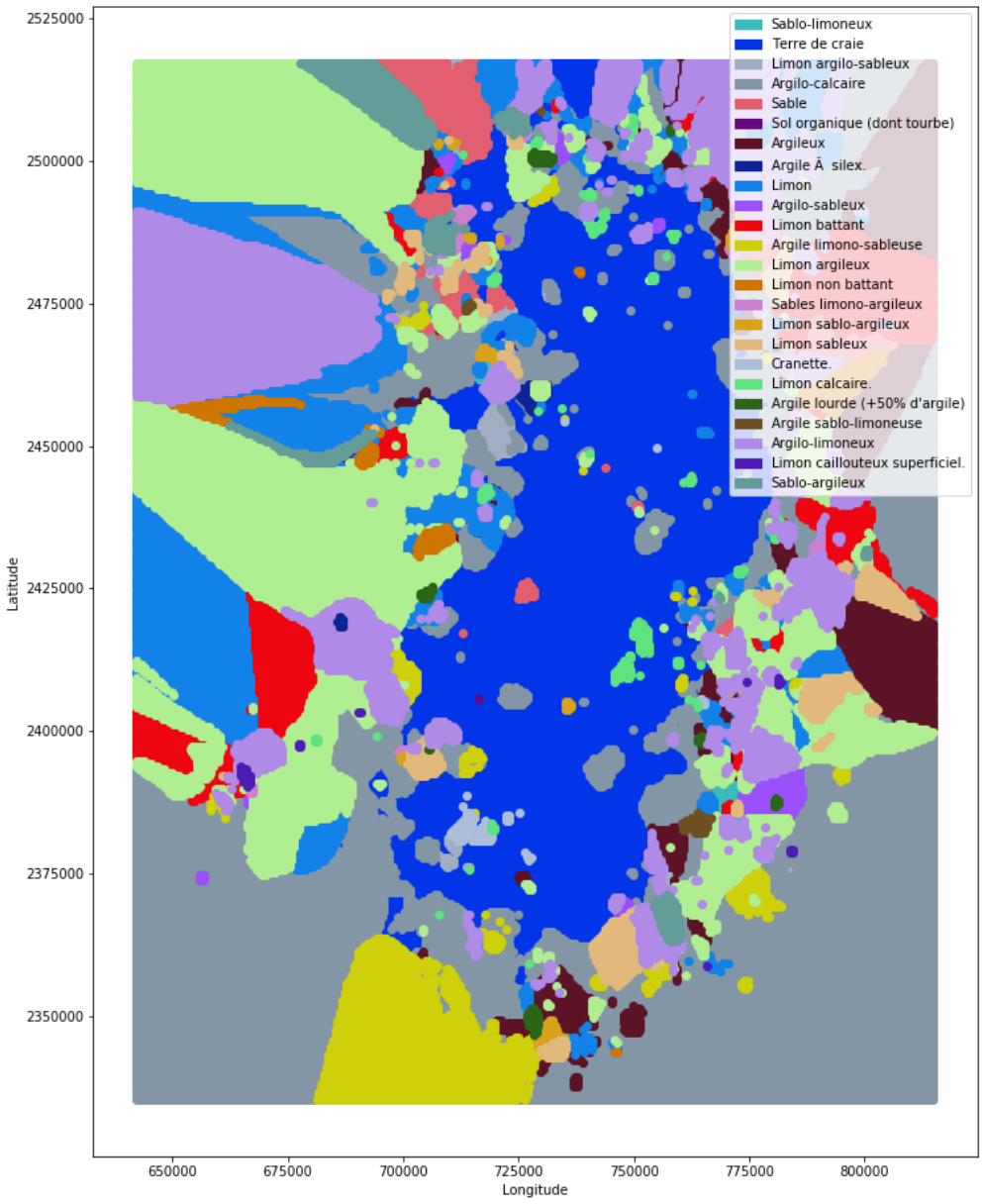

FIGURE 36 – Partition obtenue par l'algorithme des k-plus proches voisins

7.3.4 L'arbre de décision

Les arbres de décision représentent un autre cas particulier des méthodes à partition à la différence très importante que la partition générée est basée non seulement sur les variables explicatives mais sur aussi sur les sorties observées, à l'instar de l'algorithme des k-plus proches voisins. Cette particularité en fait donc un algorithme prédictif très intéressant pour modéliser l'indice de rendement potentiel. Il y a plusieurs façons de construire un arbre de décision dont les plus répandus sont le C4.5 et le CART (Classification And Regression Trees), à préciser que l'algorithme CART est celui utilisé dans l'étude.

Dans un arbre de décision, chaque noeud correspond à un sous-ensemble $\mathcal{N}^{\mathcal{X}} \in \mathcal{X}$ et à un test $\mathcal{T}_{\mathcal{N}}(.)$, appelé *critère de segmentation*, auquel on soumet une variable explicative X^j. Si l'on considère que ce test $\mathcal{T}_{\mathcal{N}}(.)$ peut donner lieu à K résultats différents alors :

$$\mathcal{T}_{\mathcal{N}} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & [|1, K|] \\ \mathcal{N}^{\mathcal{X}} & \longmapsto & \mathcal{T}(\mathcal{N}^{\mathcal{X}}) \end{array} \right. \tag{76}$$

Ainsi, le noeud correspondant à $(\mathcal{N}^{\mathcal{X}}, \mathcal{T}_N)$ donne naissance à K noeuds-fils, tel que l'ensemble $\mathcal{N}_K^{\mathcal{X}}$ associé au $K^{ème}$ fils peut s'écrire sous la forme suivante :

$$\mathcal{N}_k^{\mathcal{X}} = \{X \in \mathcal{N}^{\mathcal{X}}; \mathcal{T}_{\mathcal{N}}(X) = K\} \tag{77}$$

Dans notre étude, nous nous intéresserons uniquement aux arbres de décision dît binaire du fait que le test ne donne lieu qu'à deux résultats (i.e K=2). Selon une telle architecture, il est alors possible d'obtenir facilement une représentation de la modélisation :

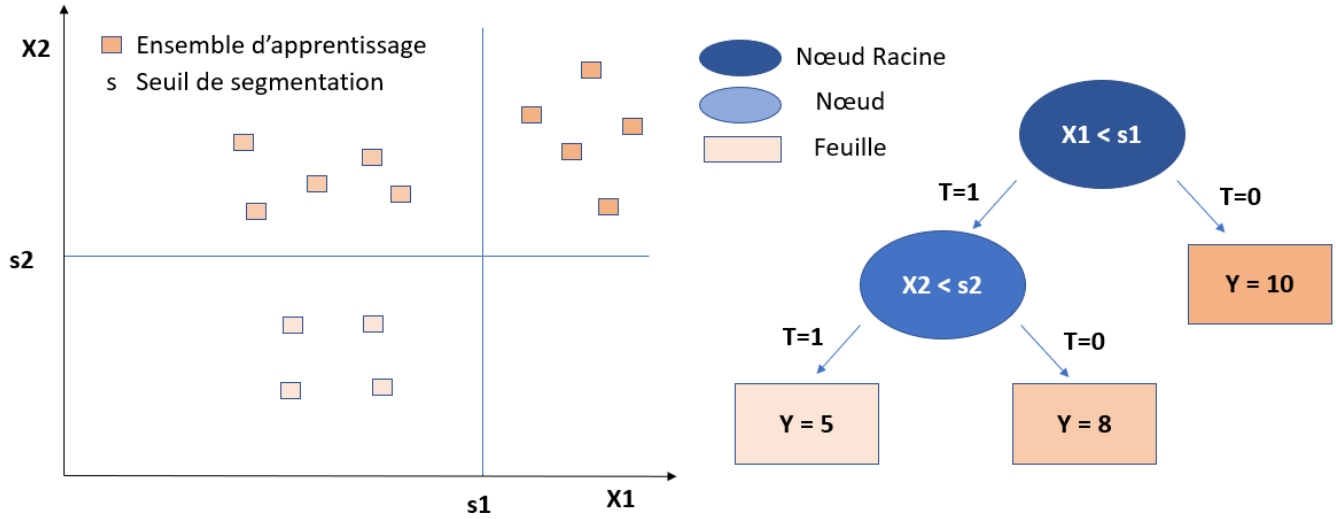

FIGURE 37 – Schéma d'un arbre de décision binaire

La construction d'un tel arbre est donc un processus récursif, initialisée par le noeud racine correspondant à $\mathcal{N}^{\mathcal{X}} = \mathcal{X}$. De façon schématique, le processus se suit d'une série de test afin de choisir un critère de segmentation et étendre l'arbre par la création d'un nœud-fils. On évalue la qualité d'une segmentation par une *fonction d'hétérogénéité* $i(.)$ qui a la particularité de croître en fonction de l'hétérogénéité des valeurs prises par la variable d'intérêt au sein d'un même nœud. Tout comme la fonction de perte, cette fonction d'hétérogénéité possède une forme bien particulière selon l'hypothèse $(\mathcal{H}_{\mathcal{Y}})$ considérée. Le processus continue tant que qu'une condition d'arrêt prescrite à l'initialisation n'est pas respectée. Il est commun d'arrêter la construction de l'arbre par une prescription sur sa profondeur.

Néanmoins, l'expansion d'une branche par la segmentation d'un nœud pouvant être non significative, il est important d'apporter quelques spécificités bien précises à la construction de l'arbre. Par exemple, il est commun de spécifier au nœud la présence d'un effectif minimum avant de segmenter. Ces spécificités peuvent être renseignées avant ou après la construction de l'arbre. Néanmoins, le fait de préciser de tels conditions à l'initialisation peut empêcher la segmentation de nœuds qui pourraient s'avérer significatif par la suite. C'est pourquoi il est en général conseiller de construire d'abord l'arbre de décision où l'on procède par la suite à la suppression des branches qui n'améliore pas significativement le risque estimé. Un tel algorithme peut être décrit de la sorte :

— **Initialisation** : ARBRE = Nœud Racine = $\mathcal{N}_{\mathcal{X}} = \mathcal{X}$
— **Expansion** : Pour chaque nœud $\mathcal{N}^{\mathcal{X}}$ de ARBRE
— si $\mathcal{N}^{\mathcal{X}}$ ne vérifie pas la condition d'arrêt
— choisir le critère de segmentation $\mathcal{T}_{\mathcal{N}}$ qui optimise Δ_K
— créer les nœuds-fils $(\mathcal{N}_k^{\mathcal{X}})_{k\in[|1,K|]}$
— mettre à jour ARBRE = ARBRE + {noeuds-fils}
— **fin**
— **Elagage** : Pour chaque nœud $\mathcal{N}^{\mathcal{X}}$ de ARBRE
— si $\mathcal{N}^{\mathcal{X}}$ vérifie la condition d'élagage
— ARBRE = ARBRE - {le nœud $\mathcal{N}^{\mathcal{X}}$ et ses descendants}
— **fin**

L'algorithme CART utilisé dans l'étude repose sur un arbre de décision binaire utilisant comme critère de segmentation l'indice de Gini dans le cas de la classification (cf. Annexe [37]) et la variance dans le cas de la régression. Pour mesurer la *réduction d'hétérogénéité* ou le *gain d'homogénéité* $\Delta_K(.)$ suite à la segmentation d'un nœud $\mathcal{N}^{\mathcal{X}}$ en K nœuds-fils $\{\mathcal{N}_1^{\mathcal{X}}, ..., \mathcal{N}_K^{\mathcal{X}}\}$, on applique la formule suivante :

$$\Delta_K : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}_+ \\ \mathcal{N}^{\mathcal{X}} & \longmapsto & = i(\mathcal{N}^{\mathcal{X}}) - \sum_{k\in[|1,K|]} \mathbb{P}[\mathcal{N}_k^{\mathcal{X}}] i(\mathcal{N}_k^{\mathcal{X}}) \end{array} \right. \quad \text{où} \quad \forall k \in [|1,K|], \mathbb{P}[\mathcal{N}_k^{\mathcal{X}}] = \frac{card(\mathcal{N}_k^{\mathcal{X}})}{card(\mathcal{N}^{\mathcal{X}})} \tag{78}$$

Dans le cas d'une régression, le gain d'homogénéité d'une segmentation est mesurée par la *variance intra – nud* $\Delta_K^{\bar{\mathcal{H}}_{\mathcal{Y}}}(.)$:

$$\Delta_K^{\bar{\mathcal{H}}_{\mathcal{Y}}}(\mathcal{N}^{\mathcal{X}}) = \mathbb{V}[Y|\mathcal{N}^{\mathcal{X}}] - \sum_{k\in[|1,K|]} \mathbb{P}[\mathcal{N}_k^{\mathcal{X}}] \mathbb{V}[Y|\mathcal{N}_k^{\mathcal{X}}] \tag{79}$$

Après avoir opérée une simplification de calcul, nous estimerons alors la variance intra-noeud $\hat{\Delta}_K^{\bar{\mathcal{H}}_{\mathcal{Y}}}(.)$ de la manière suivante :

$$\hat{\Delta}_K^{\bar{\mathcal{H}}_{\mathcal{Y}}}(\mathcal{N}^{\mathcal{X}}) = \frac{1}{card(\mathcal{N}^{\mathcal{X}})^2} \sum_{i;X_i\in\mathcal{N}^{\mathcal{X}}} (y_i - \bar{y}_{\mathcal{N}^{\mathcal{X}}})^2 - \Big(\sum_{k\in[|1,K|]} \sum_{i;X_i\in\mathcal{N}_k^{\mathcal{X}}} (y_i - \bar{y}_{\mathcal{N}_k^{\mathcal{X}}})^2 \Big) \tag{80}$$

Malheureusement, l'arbre de décision construit est trop large pour pouvoir le présenter clairement dans son ensemble. De ce fait, nous avons privilégié de présenter uniquement les nœuds à partir desquelles le risque extrême correspondant à l'année 2016 a pu être segmenté. Comme nous l'avons présenté dans le schéma (cf. Figure 37), l'opacité de la couleur présente dans l'arbre représente le niveau plus ou moins élevé de l'indice de rendement agricole modélisé. Par une analyse de l'opacité, nous remarquons alors que cette dernière s'atténue de droite à gauche de l'arbre de décision. Nous comprenons alors de ce fait que le rendement prédit en sera d'autant plus faible que la couleur présente est faible. Nous remarquerons tout particulièrement dans l'arbre (cf. Figure 38) l'excès de précipitation survenu en 2016 mis en lumière par la variable *precip_sum* correspondant au cumul des précipitations (cf. Equation 38).

37. Hétérogéneité dans un problème de classification

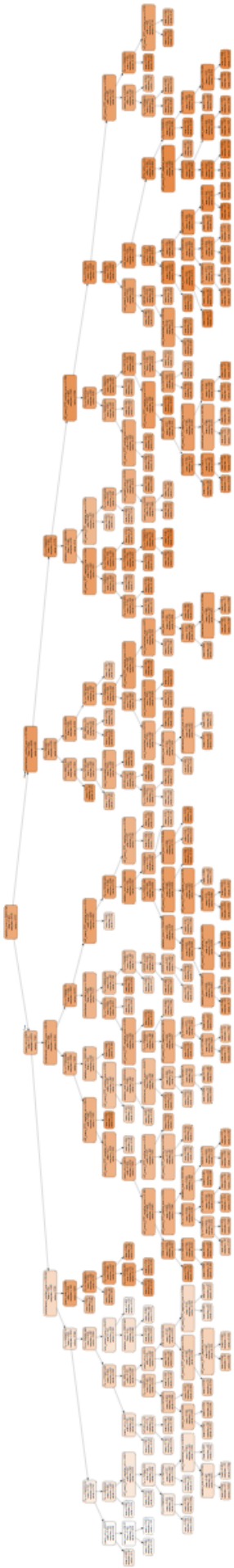

FIGURE 38 – Arbre de décision modélisant l'indice de rendement

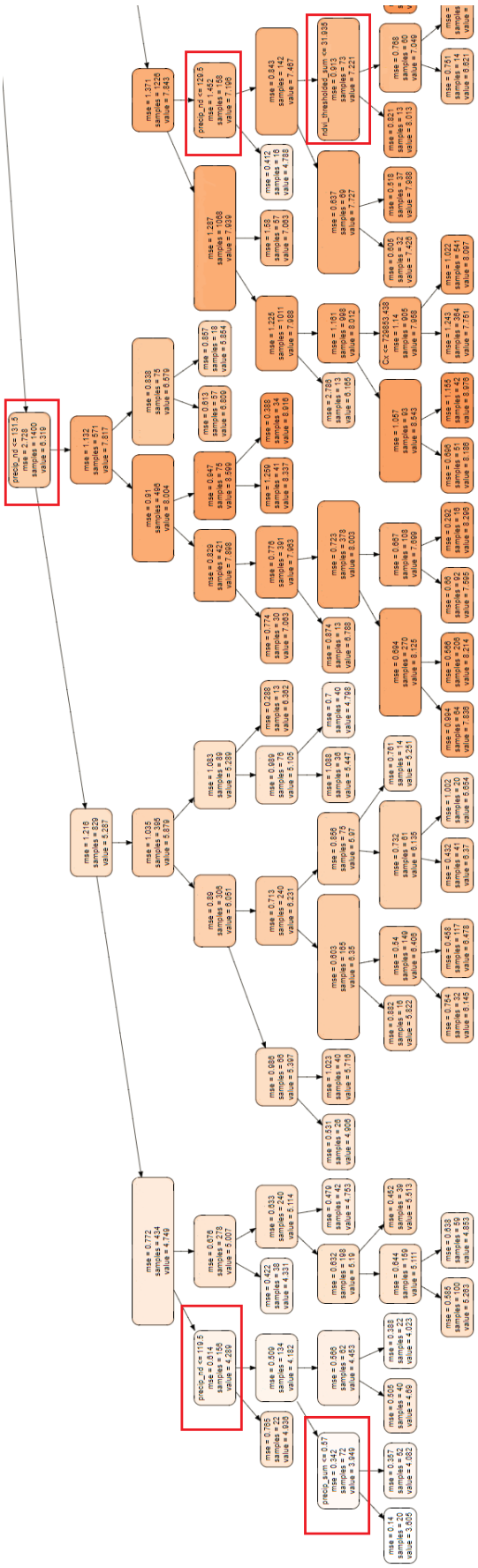

FIGURE 39 – Partie de l'arbre de décision modélisant l'indice de rendement

7.3.5 Les méthodes d'agrégations

Les arbres de décision sont en général peu robustes dans le sens où les prédictions peuvent être significativement différentes de nouvelles observations, si jamais celles-ci s'avéraient être différentes de celles présentes dans les feuilles de l'arbre. En effet, plus l'arbre est étendu et plus les observations présentes dans les feuilles correspondent à des profils bien particuliers. Ainsi, les estimations peuvent s'avérer très bonnes comme très mauvaises dès lors qu'il s'agit de prévoir des données autres que celles qui ont été utilisées pour la construction de l'arbre.

Il existe une méthode d'agrégation, développée par Berk en 2004, dont le but est de surmonter ce problème de robustesse afin d'assurer une stabilité et une fiabilité aux prédictions. Cette méthode repose sur la construction d'un grand nombre d'arbres de décision pour ensuite agréger les résultats afin d'obtenir une unique prédiction de la sortie. Ainsi, pour une nouvelle donnée, on estimera la variable d'intérêt en regroupant les différents résultats obtenus à partir de l'ensemble des arbres construits. La construction de ces arbres est opérée par tirage avec remise où nous serons amenés à tirer aléatoirement n individus un grand nombre de fois pour obtenir N échantillons à partir desquels on va construire N arbres de décision. Cette procédure permet ainsi de construire un estimateur consistant en réduisant considérablement la variance de l'estimation.

L'agrégation des sorties de chacun des arbres diffère selon la nature de la variable d'intérêt. Si celle-ci s'avère discrète, la sortie associée correspondra à la modalité la plus représentée dans les prédictions. A l'inverse, dans le cas d'une variable d'intérêt continue, nous effectuerons la moyenne des prédictions de chacun des arbres. Dans ce dernier cas, considérons une suite $(\mathcal{F}_i^{\mathcal{L}})_{i\in[|1,N|]}$ de N arbres de décisions où l'estimateur agrégé $\mathcal{F}_N^{\mathcal{L}}(.)$ peut être obtenu par la moyenne des N estimateurs :

$$\mathcal{F}_N^{\mathcal{L}} = \frac{1}{N}\sum_{i\in[|1,N|]} \mathcal{F}_i^{\mathcal{L}} \tag{81}$$

Si on émet l'hypothèse que :

$$(\mathcal{H}_\sigma^\rho) \quad \forall i \in [|1,n||, \mathbb{V}[\mathcal{F}_i^{\mathcal{L}}] = \sigma^2 \quad \text{ET} \quad \forall i \neq j, \mathbb{C}\text{o}\mathbb{v}_{\mathcal{L}}(\mathcal{F}_j^{\mathcal{L}}, \mathcal{F}_i^{\mathcal{L}}) = \rho \tag{82}$$

Il en vient que nous pourrons alors exprimer (cf. Annexe [38]) la variance de cet estimateur agrégé $\mathbb{V}[\mathcal{F}_N^{\mathcal{L}}]$ en fonction des constantes définies par l'hypothèse $(\mathcal{H}_\sigma^\rho)$:

$$\text{SOUS} \quad (\mathcal{H}_\sigma^\rho) \Longrightarrow \mathbb{V}[\mathcal{F}_N^{\mathcal{L}}] = \frac{\sigma^2}{N} + \rho(1 - \frac{1}{N^2}) \tag{83}$$

Ainsi, on remarque par cette relation que la variance se réduit lorsque le nombre d'arbres augmente. De plus, a ce stade, nous remarquerons aussi que cette variance sera d'autant plus grande que la covariance entre les estimateurs le sera. Afin de réduire d'avantage la variance de l'estimateur agrégé, il est possible d'introduire un aléa lors de la construction d'un arbre en choisissant aléatoirement les variables explicatives soumises au critère de segmentation. Cet aléa qui est à l'origine des forêts aléatoires, a pour but de renforcer l'indépendance entre les arbres. Si les échantillons s'avèrent être indépendants, alors on peut montrer que la variance de l'estimateur agrégé convergera vers 0 lorsque le nombre d'estimateurs tend vers l'infini (cf. Annexe [39]).

38. Variance de l'estimateur agrégé
39. Convergence de la variance de l'estimateur agrégé

La méthode des forêts aléatoires (cf. Annexe [40]), développée par Breiman en 2001, est similaire dans la manière d'agréger les modèles. Cependant, cette méthode se distingue de la précédente par le fait qu'avant la division d'un nœud, à la place de sélectionner l'ensemble des variables explicatives à soumettre au critère de segmentation, on tirera aléatoirement un nombre de variables. Le fait d'ajouter un tel aléa dans la construction des arbres permet de rendre les arbres plus indépendants les uns des autres ce qui engendrera en conséquence une réduction de la variance de l'estimateur agrégé. Si on note [.] la partie entière d'un réel, le nombre de variables explicatives tirées aléatoirement p' est choisi de la manière suivante :

$$p' = [\sqrt{p}] \quad \text{OU} \quad p' = [\frac{p}{3}] \tag{84}$$

Dans le cas d'une étude statistique en grande dimension où le nombre de variables explicatives est important, la sélection aléatoire de variables explicatives permet de sélectionner des variables qui n'auraient peut-être pas été choisies si elles avaient été prises toutes ensemble. Cette particularité en fait donc un modèle très apprécié lors de la sélection des variables explicatives. Certains modèles, comme l'arbre de décision, se construisent en sélectionnant à chaque étape les variables explicatives les plus pertinentes. Se basant en général sur des mesures de performance, il est alors possible d'estimer l'importance d'une variable au sein de la modélisation. Nous serons ainsi amenés à constuire un modèle, tout comme la sélection par itération, où un score sera affecté à chacune des variables explicatives sélectionnées dans l'étude. Contrairement à une sélection univariée, cette approche a l'avantage de prendre en compte les possibles corrélations entre les variables. Certaines variables qui se sont avérées non-significatives lors d'une sélection univariée peuvent lors de cette approche présentées une certaine importance au vu des corrélations qu'elles peuvent avoir.

Dans une approche non-paramétrique fondée sur l'arbre de décision, l'importance d'une variable explicative dans la modélisation repose sur l'accroissement d'homogénéité qu'elle procure lors de la segmentation des nœuds successifs. Plus concrètement, on définira une fonctionnelle qui a une variable aléatoire explicative X^j associe son *importance* $\mathcal{I}(.)$ définie comme la moyenne pondérée des accroissements des nœuds où se trouve la variable explicative considérée :

$$\mathcal{I} : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}_+ \\ X^j & \longmapsto & \mathbb{E}[\Delta_K | X^j \in \mathcal{N}^{\mathcal{X}}] = \sum_{i; X^j \in \mathcal{N}_i^{\mathcal{X}}} \mathbb{P}[\mathcal{N}_i^{\mathcal{X}}] \Delta_K(\mathcal{N}^{\mathcal{X}}) \quad \text{où} \quad \mathbb{P}[\mathcal{N}_i^{\mathcal{X}}] = \frac{card(\mathcal{N}_i^{\mathcal{X}})}{card(\mathcal{N}^{\mathcal{X}})} \end{array} \right. \tag{85}$$

Il est en général préférable d'user de forêts aléatoires afin d'accroître l'estimation de l'importance d'une variable explicative. Si l'on considère une forêt de N arbres de décision, alors nous obtiendrons une estimation de l'importance d'une variable X^j par la moyenne de l'ensemble des accroissements compris dans les nœuds de chacun des arbres :

$$\mathcal{I}(X^j) = \frac{1}{N} \sum_{N} \sum_{i; X^j \in \mathcal{N}_i^{\mathcal{X}}} \frac{card(\mathcal{N}_i^{\mathcal{X}})}{card(\mathcal{N}^{\mathcal{X}})} \Delta_K(\mathcal{N}^{\mathcal{X}}) \tag{86}$$

De plus, afin de comparer les importances des variables entre elles, il peut être intéressant d'opérer une normalisation afin d'obtenir un score de tel sorte que :

$$w_j = \frac{\mathcal{I}(X^j)}{\sum_{j \in [|1,p|]} \mathcal{I}(X^j)} \quad \text{où} \quad \sum_{j \in [|1,p|]} w_j = 1 \tag{87}$$

.

40. Algorithme des Forêts Aléatoires

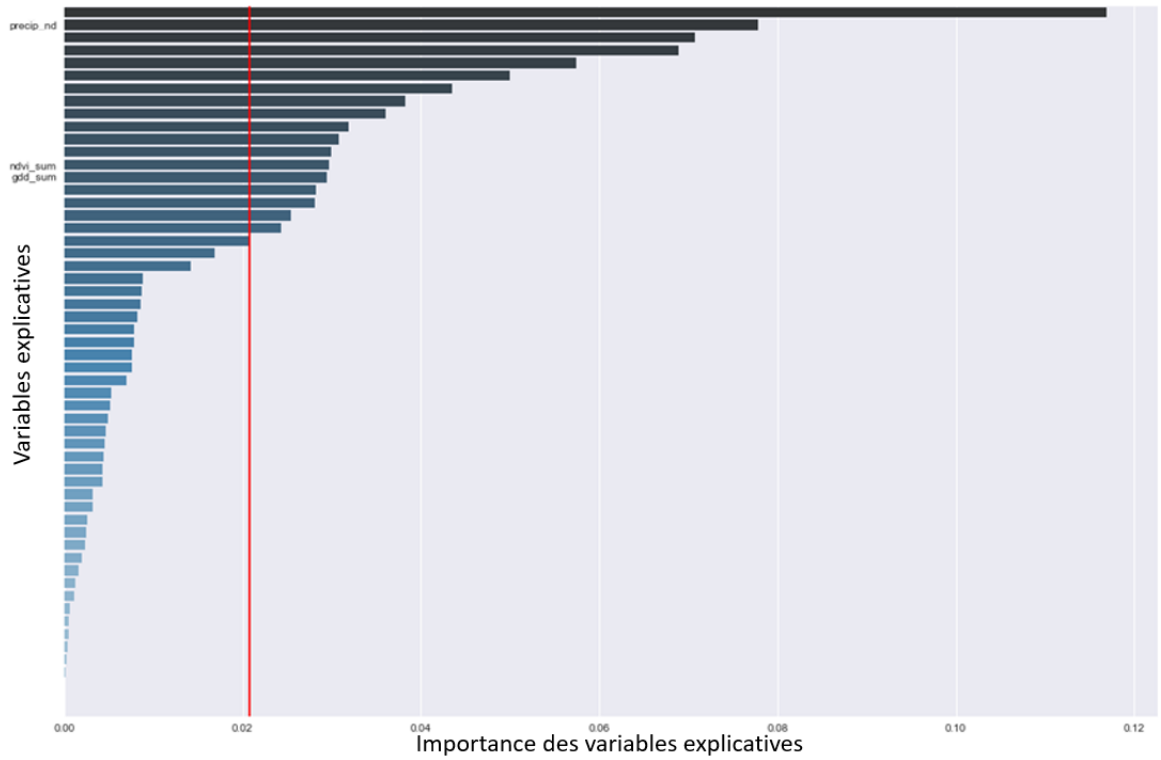

FIGURE 40 – Importance des variables explicatives par une forêt aléatoire

De cette sélection de variables explicative, nous retrouverons sans surprise la variable explicative *precip_nd* précédemment sélectionnée lors de la sélection univariée. Néanmoins, nous remarquerons que les variables *ndvi_sum* et *gdd_sum*, qui semblaient non significatives au vu des corrélations linéaires étudiées en amont, ont cette fois une importance significative.

7.3.6 Limites de l'approche non-paramétrique

Dans un cadre de commercialisation de produit régulé, au vu de la nécessité d'interpréter le modèle construit, l'approche non-paramétrique se voit très limitée. Si celle-ci peut conduire à un prédicteur plus robuste, son interprétation est d'autant plus compliquée. Dans une telle démarche, le besoin de communiquer sur ses avantages et inconvénients peut nécessiter l'explication de quelques-uns des principes du modèle qui peuvent s'avérer bien difficile dans le cas d'une forêt aléatoire ou d'un réseau de neurones artificiels.

Néanmoins, l'approche non-paramétrique, en particulier pour l'arbre de décision, a en général l'avantage de présenter une approche non-linéaire qui peut s'avérer plus proche de la réalité qu'une simple modélisation linéaire. En effet, la relation régissant le phénomène aléatoire est dans bien souvent des cas non-linéaires. Cependant, cette approche non-linéaire se doit d'être interprétable. Cette remarque importante nous amènera alors à se limiter à un arbre de décision.

Il est important de souligner que le coût de calcul des algorithmes d'apprentissage non-paramétrique n'est pas non plus une composante à négliger. Dans le cas d'une forêt aléatoire par exemple, ce coût d'exécution peut-être d'autant plus élevé que le nombre d'arbres de décision construit. En pratique, il est rare que les algorithmes ne soient qu'une fois exécutée et l'exécution de ces derniers peut parfois prendre plusieurs jours voir plusieurs semaines selon la dimension de l'ensemble d'apprentissage et la complexité de l'algorithme.

8 Tarification du produit d'assurance

8.1 Les principes de prime

8.1.1 Principe de prime

La tarification est sans aucun doute une étape cruciale dans la commercialisation d'un produit, qu'il soit financier, actuariel ou même industriel. Cependant, contrairement à l'industrie qui détermine essentiellement son prix par une évaluation des coûts de production, l'industrie du risque se doit d'ajuster son prix en fonction du risque sous-jacent à la couverture. En effet, dès lors que l'on décide de commercialiser un produit dont le profit comporte une dimension aléatoire, il est nécessaire de procéder à une gestion de ce risque afin de tirer une estimation adéquate du prix. Le vendeur d'une couverture, que ce soit une option financière ou une police d'assurance, est confronté à un système de production inversé où le profit résultant de la transaction ne sera connu qu'à maturité du contrat. On comprend alors que l'activité de couverture comporte un risque inhérent qui induit inéluctablement de porter une grande attention aux méthodes adoptées lors de la détermination du prix du contrat.

Ce prix, appelé *prime*, sera donc naturellement une fonction du risque encouru par la couverture. Le processus de tarification devra alors s'aider de méthodes prédictives afin de modéliser la perte aléatoire inhérente à celle-ci. Ainsi, on définit un *principe de prime* $\mathcal{P}(.)$ comme une fonctionnelle qui à une perte aléatoire $L = \mathcal{F}^L(Y_T, Y^L)$ associe un montant de prime π non aléatoire :

$$(\mathcal{H}_{\mathcal{P}}) \qquad \mathcal{P} : \left| \begin{array}{rcl} \mathcal{L}(\Omega, \mathbb{R}_+) & \longrightarrow & \mathbb{R}_+ \\ L & \longmapsto & \pi \end{array} \right. \tag{88}$$

Il est intéressant de présenter brièvement quelques propriétés importantes que l'on pourrait attendre de cette fonction de prime. Selon toute logique, cette fonction se doit d'être positivement corrélée à la perte encourue :

$$\forall L_1, L_2 \in \mathcal{L}(\Omega, \mathbb{R}_+); L_1 \geq L_2 \text{P.S}, \mathcal{P}(L_1) \geq \mathcal{P}(L_2) \tag{89}$$

De plus, cette dernière ne doit pas être non plus trop élevée par rapport aux pertes maximales sinon l'intérêt pour l'assuré sera nul :

$$\forall L \in \mathcal{L}(\Omega, \mathbb{R}_+), \underset{L \in \mathcal{L}(\Omega, \mathbb{R}_+)}{sup} L \geq \mathcal{P}(L) \tag{90}$$

Cependant, afin d'assurer une certaine solvabilité à l'assureur, il est important que la prime soit au moins supérieure aux pertes espérées dont la notion d'espérance fait référence à la loi régissant le phénomène aléatoire :

$$\forall L \in \mathcal{L}(\Omega, \mathbb{R}_+), \mathcal{P}(L) \geq \mathbb{E}_{\mathcal{L}}(L) \tag{91}$$

Enfin, dans un contexte de mutualisation du risque d'un portefeuille, la propriété de sous-additivité insinue que la prime d'un risque mutualisé doit être inférieure aux primes relatives à chacun des risques considérés individuellement :

$$\forall L_1, L_2 \in \mathcal{L}(\Omega, \mathbb{R}_+), \mathcal{P}(L_1 + L_2) \leq \mathcal{P}(L_1) + \mathcal{P}(L_2) \tag{92}$$

8.1.2 Principes de tarification

Toute tarification actuarielle se fonde sur le calcul d'une perte espérée où certains chargements peuvent être appliqués selon le cadre considéré. Une tarification qui se fonde sur cette perte espérée est communément appelée un *principe de prime pure*. Tout comme nous avons pu le voir lors de la recherche de la fonction oracle, ce principe de prime repose sur le fait que l'espérance est solution d'un problème de minimisation (cf. Equation 55) :

$$(\mathcal{H}_{\mathcal{P}}^{\mathbb{E}}) \quad | \quad \forall L \in \mathcal{L}(\Omega, \mathbb{R}_+), \mathcal{P}[L] = \mathbb{E}_{\mathcal{L}}[L] = \pi \tag{93}$$

On remarque cependant que cette notion amène à la considération de quelques autres principes du fait qu'elle ne tient pas compte du risque encouru défini comme la variance de la perte. De ce fait, l'assureur est amené à ajuster la prime pure en considérant une partie de cette variance. Le chargement appliqué à la variance est modélisé de manière à tenir compte des frais internes issus de la gestion du dossier ainsi que des frais externes pouvant résulter de frais de courtage. Cette prise en compte du risque $\mathbb{V}$ et des frais α dans la tarification conduisent à la définition du *principe de variance* suivant :

$$(\mathcal{H}_{\mathcal{P}}^{\mathbb{V}}) \quad | \quad \forall L \in \mathcal{L}(\Omega, \mathbb{R}_+), \forall \alpha \in \mathbb{R}_+, \mathcal{P}[L] = \mathbb{E}_{\mathcal{L}}[L] + \alpha * \mathbb{V}_{\mathcal{L}}[L] \tag{94}$$

8.2 Valorisation de la solution paramétrique

8.2.1 Homogénéisation du risque

Sans distinction de l'assuré, la prime pourrait s'avérer dans certains cas, pas assez élevée face aux agents risqués, et dans d'autres, trop élevée pour attirer les agents peu risqués. L'hétérogénéité des risques au sein d'un portefeuille d'assurés se doit alors d'être homogénéisée à l'aide d'une fine segmentation, sans quoi un phénomène de marché dît d'anti-sélection se produirait. Cette segmentation a pour objectif de distinguer les niveaux de risques présents dans le portefeuille afin d'ajuster la prime en conséquence. Ce phénomène d'anti-sélection amène donc l'actuaire à procéder à un travail de segmentation puis de tarification. A préciser que nous nous intéresserons ici uniquement à la tarification à priori qui ne tient pas compte de l'historique de l'assuré à l'instar de la tarification à posteriori. La prime devant être représentative du risque encouru, nous serons alors amenés à regrouper les agriculteurs en classes homogènes de risques.

Néanmoins, cette classification des agriculteurs ne peut pas être effectuée à l'aide du modèle linéaire généralisé du fait que pour chaque agriculteur, nous obtiendrons une unique prédiction de rendement. Nous rappelons que ce modèle ne peut être utilisé que à maturité lorsque l'indice de rendement est considéré comme réel. Dans un contexte d'assurance agricole paramétrique où seul un indice de référence est retenu comme base d'indemnisation, on observe en général au pire une fois le sinistre. Cette particularité nous amène à s'intéresser à des méthodes statistiques différentes de l'approche fréquence-sévérité communément utilisée en assurance non-vie. Comme nous avons pu l'évoquée lors des principes de primes, la tarification repose sur l'espérance et la variance des pertes espérées. De ce constat, nous avons alors sélectionné l'arbre de décision qui associe à toute entrée une feuille $\mathcal{N}^{\mathcal{X}}$ contenant une distribution de rendements, la prédiction émise par l'arbre de décision n'étant rien d'autre que la moyenne des rendements présente dans la feuille. Ainsi, sur la base de cette distribution, nous serons alors en mesure d'estimer convenablement la potentielle perte et le risque associé à l'estimation.

Selon la méthode de segmentation envisagée, nous considérerons alors chacune des feuilles comme étant des classes homogènes distinctes où les observations présentes comportent exactement les mêmes caractéristiques. En effet, au vu de la construction de l'arbre, les seuils de chacun des nœuds permettent d'obtenir des feuilles homogènes où toutes les observations respectent l'ensemble des conditions prescrites par la branche. Notons qu'un noeud $\mathcal{N}^{\mathcal{X}}$ est correctement segmenté si à quelques exceptions près, les observations présentes dans ce noeud sont similaires. De ce fait, nous disposerons alors d'un ensemble de K feuilles $(\mathcal{N}_i^{\mathcal{X}})_{i\in[|1,K|]}$:

$$(\mathcal{H}_{\mathcal{N}}) \quad | \quad \forall i \in [|1,K|], \forall X_i, X_i^{'} \in \mathcal{N}_i^{\mathcal{X}}, X_i \approx X_i^{'} \quad \text{ET} \quad Y_i \approx Y_i^{'} \tag{95}$$

Selon la taille des effectifs présents dans chacune des feuilles, on peut envisager une tarification fondée sur les données historiques présentes. Cependant, cette approche empirique est relativement limitée au nombre de sinistres effectifs dans la feuille, celui-ci étant lié au nombre d'observations présentes et à la garantie souhaitée. Cette approche peut donc conduire à une sinistralité nulle, si la garantie est improbable ou s'il n'y a pas assez de données, et par conséquent une prime nulle. Limités par le nombre d'observations, nous nous intéresserons uniquement à la valorisation par simulations.

8.2.2 Valorisation par Monte Carlo

Au vu de la méthode de segmentation et de tarification retenue, nous chercherons à estimer les lois de probabilité $(\mathcal{L}_i)_{i\in[|1,K|]}$ ayant engendrées les distributions présentes dans chacune des feuilles de l'arbre. Si l'hypothèse $(\mathcal{H}_{\mathcal{N}})$ s'avère véridique, les données présentes dans la feuille doivent alors être proches d'une moyenne. Ainsi, la segmentation opérée se traduira alors en toute logique en une distribution proche de celle d'une loi normale. Cette hypothèse doit bien entendu être vérifiée à l'aide de méthodes statistiques semblables à celles présentées dans la première partie (cf. Equation 10). Pour mieux illustrer l'hypothèse, nous présenterons l'évolution d'une même observation au sein des feuilles :

$$(\mathcal{H}_{\mathcal{N}}) \quad | \quad \forall i \in [|1,K|], \mathcal{N}_i^{\mathcal{X}} \rightsquigarrow \mathcal{N}(\mu_i, \sigma_i) = \mathcal{L}_i \tag{96}$$

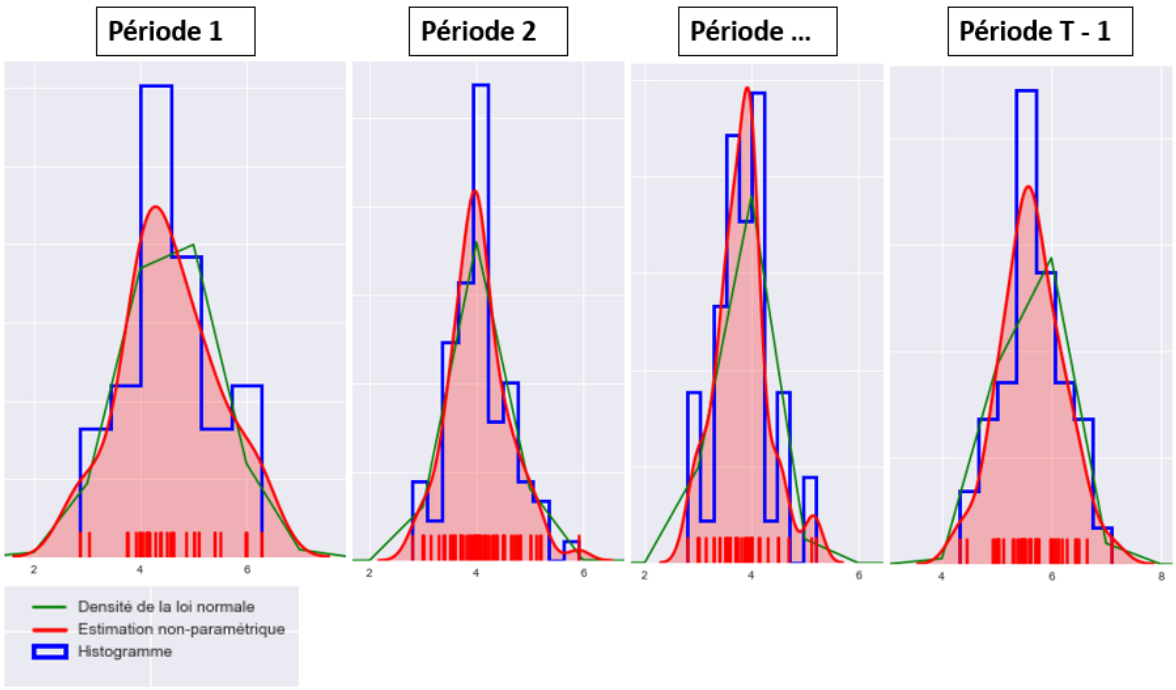

FIGURE 41 – Estimation des densités de probabilité des feuilles de l'arbre

De par l'évolution de la densité de probabilité estimée, on remarquera que l'écart-type se réduit d'un stade phénologique à l'autre. Cette remarque s'explique en partie du fait que le modèle prédictif s'améliore au fil des périodes ce qui améliorera en conséquence la segmentation du portefeuille d'assurés. En effet, en début de période, là où le modèle est le moins performant, nous pourrons constater des faibles rendements associés à des rendements élevés, ce qui aura pour effet d'accroître l'écart-type de la densité. Dans le cas d'une tarification fondée sur le principe de variance, cette segmentation d'autant plus fine devrait alors en toute logique engendrer une diminution de la prime d'une période à l'autre. De même, on observera que la prime pure estimée sera d'autant plus représentative du risque de l'assuré au vu de l'homogénéisation opérée.

L'évaluation de la prime par simulation Monte Carlo présuppose de connaître la loi de probabilité régissant le phénomène aléatoire à simuler supposée être une loi normale. Au vu de l'hypothèse $(\mathcal{H}_{\mathcal{N}})$ émise précédemment, et de la nature dynamique du produit d'assurance, la procédure afin d'évaluer la prime π_t relative à la période $t \in \mathcal{T} = [|1, T-1|]$ est alors la suivante :

— **Initialisation** : Déterminer les feuilles contenant les parcelles de l'exploitation
— **Estimation** : Pour la parcelle et donc la feuille $\mathcal{N}^{\mathcal{X}}$ considérée
— Déterminer les paramètres de la loi $\mathcal{N}(\mu, \sigma^2)$
— **Simulation** : Pour un nombre de fois N_S relativement élevé
— Tirer au hasard une trajectoire du rendement Y_i
— Calculer la perte relative à la souscription (Y_t^L, P_t^Y)
— **Estimation** : Calculer la moyenne des pertes simulées $(\mathcal{F}^{\mathcal{L}}(Y_i))_{i\in[|1,N_S|]}$
— **Estimation** : Agrégation des primes à l'échelle de l'exploitation
— **fin**

Dans un tel cadre, la précision de l'estimation est inversement liée au nombre de trajectoires simulées N_S. Cette remarque conduit généralement à prendre une valeur $N_S = 10\ 000$ élevée tout en garantissant un temps de calcul raisonnable. Si l'on considère $\mathcal{F}^L(.)$ la fonction associée au produit d'assurance (cf. Equation 5), il en vient alors d'après la loi des grands nombres que la prime pure π_t pour une période t considérée s'exprimera de la manière suivante :

$$\hat{\pi}_t = \frac{1}{N_S}\sum_{i=1}^{N_S} \mathcal{F}^L(Y_i) = \frac{1}{N_S}\sum_{i=1}^{N_S} P_t^Y(Y_t^L - Y_i)\mathbb{1}_{Y_i \leq Y_t^L} \underset{N_S \longrightarrow \infty}{\longrightarrow} \mathbb{E}_{\mathcal{L}}[P_t^Y(Y_t^L - Y)\mathbb{1}_{Y \leq Y_t^L}] \qquad (97)$$

Une fois la modélisation statistique et la méthode de tarification actuarielle achevée, l'assureur est souvent amené à se référer à des indicateurs de performance dont l'intérêt est de mettre en exergue l'efficience d'une couverture d'assurance. De ces indicateurs, l'assureur considère bien souvent le rapport entre les sinistres observés et les primes perçues, appelé *ratio de perte*. Ce ratio se doit donc d'être inférieur à 1 pour que l'assureur soit bénéficiaire. D'autres indicateurs tel que le rapport entre les pertes observées et la somme assurée peuvent aussi être envisagés, nous parlerons cette fois de *ratio de coût*.

Néanmoins, les primes et les garanties étant sujettes à une souscription dynamique $(P_t^Y, Y_t^L)_{t\in[|1,T-1|]}$, l'ensemble des indicateurs présentés précédemment reposent sur des pertes, des primes et des sommes assurées qui s'avèrent inconnues avant maturité du contrat T. Sans considérer des hypothèses par rapport au comportement de souscription de l'assuré, il devient alors difficile voire impossible d'évaluer efficacement les performances d'un tel produit. D'autant plus qu'évaluer les performances d'un tel produit sur une année historique présente dans l'ensemble d'apprentissage conduirait à une estimation biaisée.

9 Conclusion

Dans le cadre d'une solution de couverture dynamique, nous ne disposons pas d'une grille de tarification permettant d'évaluer de manière précise et efficace les gains ou pertes relatifs à la commercialisation du produit d'assurance. Le risque de l'agriculteur variant d'un stade phénologique à l'autre, il est impossible d'évaluer son risque lors de l'octroi du contrat. De plus, la souscription étant assujetties à des niveaux de garantis et des niveaux de prix inconnus, il est d'autant plus difficile d'évaluer la performance d'une telle solution. Nous ajouterons à cette remarque, que lors de la commercialisation d'un tel produit, il est en général prescrit à une zone restreinte ce qui dans le cas d'une année extrême risque d'être dévastateur pour l'assureur. Il est donc important pour ce dernier de mutualiser le risque de son portefeuille d'assurés au travers une diversification tant géographique que par type de culture.

Cependant, la complexité d'un tel produit ne doit pas être négligée lors de la commercialisation de ce dernier. Celui-ci peut en effet se trouver limité aux assurés les plus aguerris. La nécessité de construire des partenariats avec des intermédiaires spécifiques au domaine sous-jacent sera une première voie d'accès afin de faciliter la compréhension du produit. Dans un contexte agricole, il sera par exemple préconisé de démarcher des coopératives agricoles capables d'expliquer en retour l'intérêt de ces produits aux agriculteurs. Une piste d'amélioration de ce produit est de proposer des couvertures contre une perte de revenu incorporant, en plus d'une protection contre une perte du rendement, une couverture contre la variation des prix de vente. Néanmoins, au vu de la souscription dynamique propre au produit structuré, la prise en compte d'une telle dimension ajouterait une forte complexité au produit. Dans une démarche aussi innovante, il sera sûrement préférable d'attendre quelques années avant de le proposer.

Grâce aux récents progrès réalisés ces dernières années, en particulier dans le domaine de l'apprentissage statistique, et l'émergence toujours plus accrues des données, on assiste depuis une dizaine d'années au développement de nouveau type de produit d'assurance, spécialement paramétrique. Néanmoins, comme on a pu l'observer tout au long de ce mémoire, la modélisation du rendement de l'assuré se doit d'être pensée avec précaution au vu du risque de base. Ce risque inhérent au processus de couverture est sans doute l'une de ses limites les plus contraignantes à ce jour. Les limites technologiques que l'on a pu soulever seront sans doute levées dans les prochaines années au vu des récentes avancées. Ne serait-ce qu'au niveau spatial, la résolution des images se verra accroître ce qui améliorera de ce fait la qualité des indices satellitaires construits à partir des images satellitaires récoltées. Ces futures avancées engendreront dès lors une diminution du risque de base ce qui aura sans doute pour conséquence d'accroître l'intérêt de ces produits paramétriques innovants. Il n'est pas non plus écarté que d'autres technologies se développent ce qui ouvrira alors la voie vers de nouvelles approches.

Quatrième partie
Annexes

10 La Famille Exponentielle

Les lois de cette famille Exponentielle ont une particularité d'admettre une densité de probabilité $f_{(\theta,\phi)}(.)$ de la forme générale suivante :

$$f_{(\theta,\phi)}(x) = e^{\frac{x\theta - b(\theta)}{a(\phi)} + c(x,\phi)} \tag{98}$$

où $a(.)$, $b(.)$, $c(.)$ sont des fonctions à identifier et ϕ un paramètre de nuisance connu.

On peut montrer, en intégrant la densité et sous certaines hypothèses de dérivabilité, que l'espérance ainsi que la variance peuvent s'exprimer en fonction de θ par le biais de $b(.)$:

$$\mu = \mathbb{E}_{\mathcal{L}}[Y|X=x] = b'(\theta) \quad \text{ET} \quad \mathbb{V}_{\mathcal{L}}[Y|X=x] = \phi b''(\theta) = \phi b''(b'^{-1}(\mu)) \tag{99}$$

Cette relation met en lumière l'importance de la fonction b"(.), appelée *fonction variance*, qui caractérise entièrement la loi de la famille Exponentielle. Chacune des lois de cette famille posséde une fonction de lien $\mathcal{G}_{\mathcal{L}}(.)$ spécifique permettant de relier l'espérance μ au paramètre θ. Le lien est tel que :

$$\mathcal{G}_{\mathcal{L}}(\mu) = \theta \Longrightarrow \mathcal{G}_{\mathcal{L}}(.) = b'(.)^{-1} \Longrightarrow \mathcal{G}_{\mathcal{L}}(.)^{-1} = b'(.) \tag{100}$$

Dans le cas d'une étude de régression, nous nous intéresserons à la loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on peut identifier ces paramètres $\theta = \mu$ et $\phi = \sigma^2$, la fonction de lien $\mathcal{G}_{\mathcal{L}}(\mu) = \mu = \theta$ ainsi que les fonctions associées :

$$a(\phi) = \phi, b(\theta) = \frac{1}{2}\theta^2, c(x,\phi) = -\frac{1}{2}(\frac{x^2}{\phi} + ln(2\pi\phi)), x \in \mathbb{R} \tag{101}$$

11 Minimisation du risque généralisé

Supposons que $\forall x \in \mathcal{X}, \inf_{y\in\mathcal{Y}} \mathbb{E}_{\mathcal{L}}[l(Y,y)|X=x]$ est atteint.

Soit $\mathcal{F}, \mathcal{F}_*^{\mathcal{L}} \in \mathcal{F}(\mathcal{X},\mathcal{Y})$; $\forall x \in \mathcal{X}, \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y\in\mathcal{Y}}{argmin}\ \mathbb{E}_{\mathcal{L}}[l(Y,y)|X=x]$, on a :

$\mathcal{R}_{\mathcal{L}}[\mathcal{F}] := \mathbb{E}_{\mathcal{L}}[l(Y,\mathcal{F}(X))] = \int_{\mathcal{X}} \mathbb{E}_{\mathcal{L}}[l(Y,\mathcal{F}(X))|X=x]\mathrm{d}\mathbb{P}_X(x)$

$$= \int_{\mathcal{X}} \mathbb{E}_{\mathcal{L}}[\underset{y\in\mathcal{Y}}{min}\ l(Y,y)|X=x]\mathrm{d}\mathbb{P}_X(x)$$
$$\geq \int_{\mathcal{X}} \mathbb{E}_{\mathcal{L}}[l(Y,\mathcal{F}_*^{\mathcal{L}}(x))|X=x]\mathrm{d}\mathbb{P}_X(x) = \mathbb{E}_{\mathcal{L}}[l(Y,\mathcal{F}_*^{\mathcal{L}}(X))] = \mathcal{R}_{\mathcal{L}}[\mathcal{F}_*^{\mathcal{L}}].$$

Nous venons ainsi de montrer que si une fonction de prédiction minimise le risque généralisé en tout point de l'ensemble d'entrée $\mathcal{X}$, alors elle minimisera le risque généralisé dans son ensemble.

12 Problème de régression aux moindres carrés

Dans un problème de régression aux moindres carrés (i.e où p=2), nous sommes amenés à résoudre le problème de minimisation suivant :

$$(\mathcal{P}_{\mathcal{F}}^{\bar{\mathcal{H}}_{\mathcal{Y}}}) \quad | \quad \forall x \in \mathcal{X}; \bar{\mathcal{F}}_*^{\mathcal{L}}(x) \in \underset{y \in \mathcal{Y}}{argmin}\ \mathbb{E}_{\mathcal{L}}[(Y-y)^2|X=x] = \phi(y) \tag{102}$$

De par la linéarité de l'espérance, on remarque que $\phi(.)$ peut s'écrire de manière polynomiale :

$$\phi(y) = E_{\mathcal{L}}[Y^2|X=x] - 2yE_{\mathcal{L}}[Y|X=x] + y^2 \tag{103}$$

Cette forme polynomiale permet alors de déterminer facilement une solution $y_* = \bar{\mathcal{F}}_*^{\mathcal{L}}(x)$ à partir de la condition du 1er ordre $\frac{\partial \phi}{\partial y}(y) = 0$:

$$(\mathcal{S}_{\mathcal{F}}^{\bar{\mathcal{H}}_{\mathcal{Y}}}) \quad | \quad y_* = \mathbb{E}_{\mathcal{L}}[Y|X=x] = \bar{\mathcal{F}}_*^{\mathcal{L}}(x) \tag{104}$$

13 Procédure de validation croisée : k-Fold

Dans notre étude, nous nous intéresserons à la méthode communément appelé $k-Fold$. Pour un $k \in \mathbb{N}^*$ fixé, l'ensemble d'apprentissage initial $\mathcal{D}_n^{\mathcal{A}}$ est subdivisé en k sous-échantillons où seul l'un d'entre eux est sélectionné comme ensemble de validation $\mathcal{D}_i^{\mathcal{V}}$. Il est souvent précisé lors de la construction de ces k échantillons de le réaliser de manière à respecter la répartition des modalités des variables explicatives. Les $k-1$ sous-échantillons restant $(\mathcal{D}_i^{\mathcal{A}})_{i\in[|1,k-1|]}$ serviront par la suite à la construction de l'arbre de décision dont l'évaluation de la robustesse sera opérée sur l'ensemble restant $\mathcal{D}_i^{\mathcal{V}}$.

$$\forall i \in [|1,k|]; card(\mathcal{D}_i^{\mathcal{A}}) = \frac{k-1}{k} card(\mathcal{D}_n^{\mathcal{A}}) \quad \text{ET} \quad card(\mathcal{D}_i^{\mathcal{V}}) = \frac{1}{k} card(\mathcal{D}_n^{\mathcal{A}}) \tag{105}$$

L'opération se répète ainsi k fois pour ensuite calculer la moyenne des erreurs commises au cours de chacune des opérations. L'exercice de validation croisée est par la suite réalisé pour tous les paramètres possibles θ afin d'obtenir le paramètre optimal θ^*. Il s'en suivra alors une agrégation des erreurs $(l_i)_{i\in[|1,k|]}$ obtenues de chacun des arbres construits sur la base de ces k sous-échantillons :

$$(\mathcal{P}_\Theta) \quad \theta_n^* \in \underset{\theta\in\Theta}{argmin}\ l = \frac{1}{k}\sum_{i=1}^{k} l_k = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{card(\mathcal{D}_i^{\mathcal{V}})} \sum_{i;X_i\in\mathcal{D}_i^{\mathcal{V}}} (y_i - \mathcal{F}_i(X_i))^2 \tag{106}$$

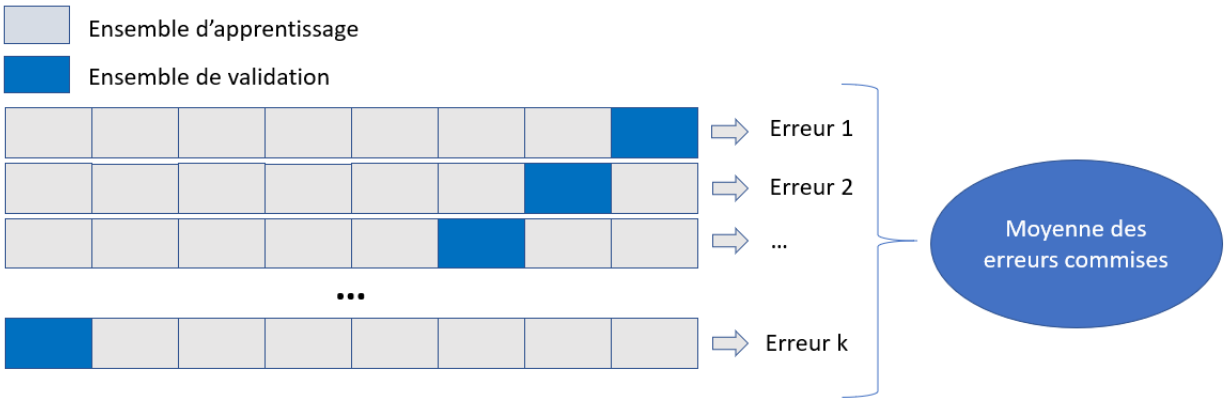

FIGURE 42 – Procédure de Validation Croisée k-Fold

14 Problème de classification binaire

Dans un problème de classification binaire, nous serons alors aménés à résoudre le problème :

$$(\mathcal{P}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad \forall x \in \mathcal{X}; \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y \in \mathcal{Y}}{argmin}\ \mathbb{E}_{\mathcal{L}}[\mathbb{1}_{Y \neq y} | X = x] = \mathbb{P}_{\mathcal{L}}(Y \neq y | X = x) = \phi(y) \tag{107}$$

On remarque cependant que $\phi(y) = 1 - \mathbb{P}_{\mathcal{L}}(Y = y | X = x) = 1 - \bar{\phi}(y)$:

$$(\mathcal{P}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad \underset{y \in \mathcal{Y}}{argmin}\ \phi(y) \Longleftrightarrow \underset{y \in \mathcal{Y}}{argmin}\ (1 - \bar{\phi}(y)) \Longleftrightarrow -\underset{y \in \mathcal{Y}}{argmin}\ \bar{\phi}(y) \Longleftrightarrow \underset{y \in \mathcal{Y}}{argmax}\ \bar{\phi}(y) \tag{108}$$

Grâce à la forme plus générale de $\bar{\phi}(.)$, nous pourrons réécrire le problème en fonction de la masse de dirac $\delta(.)$:

$$(\mathcal{P}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad \forall x \in \mathcal{X}; \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y \in \mathcal{Y}}{argmax}\ \bar{\phi}(y) = \mathbb{P}_{\mathcal{L}}(Y = y | X = x) = \sum_{y \in \mathcal{Y}} \delta_{Y=y} \mathbb{P}[Y = y | X = x] \tag{109}$$

Dans un problème de classification binaire où $\mathcal{Y} = \{0, 1\}$, il adviendra que nous pourrons écrire :

$$(\mathcal{P}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad \forall x \in \mathcal{X}; \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y \in \{0,1\}}{argmax}\ \bar{\phi}(y) = \delta_{y=0} \mathbb{P}[Y = 0 | X = x] + \delta_{y=1} \mathbb{P}[Y = 1 | X = x] \tag{110}$$

Si l'on considère $\bar{\mathcal{F}}_*^{\mathcal{L}}(x) = \mathbb{E}_{\mathcal{L}}[Y | X = x] = \mathbb{P}_{\mathcal{L}}[Y = 1 | X = x]$, alors :

$$(\mathcal{P}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad \forall x \in \mathcal{X}; \mathcal{F}_*^{\mathcal{L}}(x) \in \underset{y \in \{0,1\}}{argmax}\ \bar{\phi}(y) = \delta_{y=0}(1 - \bar{\mathcal{F}}_*^{\mathcal{L}}(x)) + \delta_{y=1} \bar{\mathcal{F}}_*^{\mathcal{L}}(x) \tag{111}$$

Ainsi, la solution à ce problème non différentiable sera alors de la forme :

$$(\mathcal{S}_{\mathcal{F}}^{\mathcal{H}_{\mathcal{Y}}}) \quad | \quad \mathcal{F}_*^{\mathcal{L}}(x) = y_* = \mathbb{1}_{\bar{\mathcal{F}}_*^{\mathcal{L}}(x) \geq \frac{1}{2}} \tag{112}$$

15 Hétérogéneité dans un problème de classification

Dans le cas d'une variable discrète dont le support $\mathcal{Y} = \{y_1, ..., y_m\}$ est constitué de m modalités, nous nous intéresserons à des fonctions d'hétérogéneités $i(.)$ de la forme suivante :

$$i : \left| \begin{array}{ccc} \mathcal{X} & \longrightarrow & \mathbb{R}_+ \\ \mathcal{N}^{\mathcal{X}} & \longmapsto & \sum_{i\in[|1,m|]} g(p_i^{\mathcal{N}^{\mathcal{X}}}) \end{array} \right. \quad \text{AVEC} \quad p_i^{\mathcal{N}^{\mathcal{X}}} = \mathbb{P}[Y = y_i | X \in \mathcal{N}^{\mathcal{X}}] = \frac{card(\mathcal{N}^{\mathcal{X}}(y_i))}{card(\mathcal{N}^{\mathcal{X}})} \tag{113}$$

où $card(\mathcal{N}^{\mathcal{X}}(y_i))$ correspond à l'effectif du noeud $\mathcal{N}^{\mathcal{X}}$ présentant la modalité y_i. La probabilité $p_i^{\mathcal{N}^{\mathcal{X}}}$ correspond donc à la probabilité que la variable d'intérêt soit égale à la modalité y_i tout en sachant la donnée dans le noeud $\mathcal{N}^{\mathcal{X}}$.

Le choix de la fonction $g : [0, 1] \longrightarrow \mathbb{R}_+$ influence peu le résultat où les trois les plus couramment utilisées sont :

1. L'indice de Gini : $g(p) = p(1 - p)$
2. L'erreur de Bayes : $g(p) = min\{p, (1 - p)\}$
3. La fonction d'entropie : $g(p) = -pln(p)$

Ainsi, lors de la construction d'un tel arbre de décision, le gain d'homogéneité Δ est déterminé de la manière suivante :

$$\hat{\Delta}_K^{\mathcal{H}_{\mathcal{Y}}}(\mathcal{N}^{\mathcal{X}}) = \sum_{i\in[|1,m|]} g(p_i^{\mathcal{N}^{\mathcal{X}}}) - \sum_{k\in[|1,K|]} \mathbb{P}[\mathcal{N}_k^{\mathcal{X}}] \sum_{i\in[|1,m|]} g(p_i^{\mathcal{N}_k^{\mathcal{X}}}) \tag{114}$$

L'estimation de ce gain $\hat{\Delta}$ sera donc :

$$\hat{\Delta}_K^{\mathcal{H}_{\mathcal{Y}}}(\mathcal{N}^{\mathcal{X}}) = \sum_{i\in[|1,m|]} g(\frac{card(\mathcal{N}^{\mathcal{X}}(y_i))}{card(\mathcal{N}^{\mathcal{X}})}) - \sum_{k\in[|1,K|]} \frac{card(\mathcal{N}_k^{\mathcal{X}})}{card(\mathcal{N}^{\mathcal{X}})} \sum_{i\in[|1,m|]} g(\frac{card(\mathcal{N}_k^{\mathcal{X}}(y_i))}{card(\mathcal{N}_k^{\mathcal{X}})}) \tag{115}$$

16 Variance de l'estimateur agrégé

Si on émet l'hypothèse que :

$$(\mathcal{H}_\sigma^\rho) \quad \forall i \in [|1,n||, \mathbb{V}[\mathcal{F}_i^\mathcal{L}] = \sigma^2 \quad \text{ET} \quad \forall i \neq j, \mathbb{C}\text{ov}_\mathcal{L}(\mathcal{F}_j^\mathcal{L}, \mathcal{F}_i^\mathcal{L}) = \rho \tag{116}$$

Il en vient que nous pourrons alors exprimer la variance de cet estimateur agrégé $\mathbb{V}[\mathcal{F}_N^\mathcal{L}]$ en fonction des constantes définies par l'hypothèse $(\mathcal{H}_\sigma^\rho)$:

$$\mathbb{V}_\mathcal{L}[\mathcal{F}_N^\mathcal{L}] = \mathbb{V}_\mathcal{L}[\frac{1}{N}\sum_{i\in[|1,N|]}\mathcal{F}_i^\mathcal{L}] = \frac{1}{N^2}(\sum_{i\in[|1,N|]}\mathbb{V}_\mathcal{L}[\mathcal{F}_i^\mathcal{L}] + 2\sum_{i,j\in[|1,N|];i\neq j}\mathbb{C}\text{ov}_\mathcal{L}[\mathcal{F}_i^\mathcal{L}, \mathcal{F}_j^\mathcal{L}]) \tag{117}$$

$$\text{SOUS} \quad (\mathcal{H}_\sigma^\rho) \Longrightarrow \mathbb{V}[\mathcal{F}_N^\mathcal{L}] = \frac{1}{N^2}(N\sigma^2 + 2\rho N(\frac{N}{2} - 1)) = \frac{\sigma^2}{N} + \rho(1 - \frac{1}{N^2}) \tag{118}$$

17 Convergence de la variance de l'estimateur agrégé

Si l'on suppose les fonctions indépendantes, il en vient un covariance nulle :

$$\forall i \neq j, \mathcal{F}_i^\mathcal{L} \perp \mathcal{F}_j^\mathcal{L} \Longrightarrow \forall i \neq j, \mathbb{C}\text{ov}_\mathcal{L}(\mathcal{F}_j^\mathcal{L}, \mathcal{F}_i^\mathcal{L}) = \rho = 0 \tag{119}$$

De ce fait, nous pourrons réecrire la variance de l'estimateur et calculer sa limite :

$$\mathbb{V}[\mathcal{F}_N^\mathcal{L}] = \frac{\sigma^2}{N} \underset{N\longrightarrow\infty}{\longrightarrow} 0 \tag{120}$$

18 Algorithme des Forêts Aléatoires

Pour mieux illustrer la méthode, nous présenterons l'algorithme pour N arbres de décision :

— si $k \leq N$
— **Initialisation** : Création du k^{e} ensemble par tirage avec remise
— **Initialisation** : ARBRE = Noeud Racine = $\mathcal{N}_{\mathcal{X}} = \mathcal{X}$
— **Expansion** : pour chaque noeud $\mathcal{N}^{\mathcal{X}}$ de ARBRE
— si $\mathcal{N}^{\mathcal{X}}$ ne vérifie pas la condition d'arrêt
— Tirage aléatoire de p' variables explicatives parmi l'ensemble des p variables
— choisir un critère de segmentation $\mathcal{T}_{\mathcal{N}}$
— créer les noeuds-fils $(\mathcal{N}_k^{\mathcal{X}})_{k\in[|1,K|]}$
— mettre à jour ARBRE = ARBRE + {noeuds-fils}
— **fin**
— **Elagage** : Pour chaque noeud $\mathcal{N}^{\mathcal{X}}$ de l'arbre
— si $\mathcal{N}^{\mathcal{X}}$ vérifie la condition d'élagage
— ARBRE = ARBRE - {le noeud $\mathcal{N}^{\mathcal{X}}$ et ses descendants}
— **fin**
— **fin**

Afin de compléter l'illustration, nous pourrons nous référer au schéma suivant :

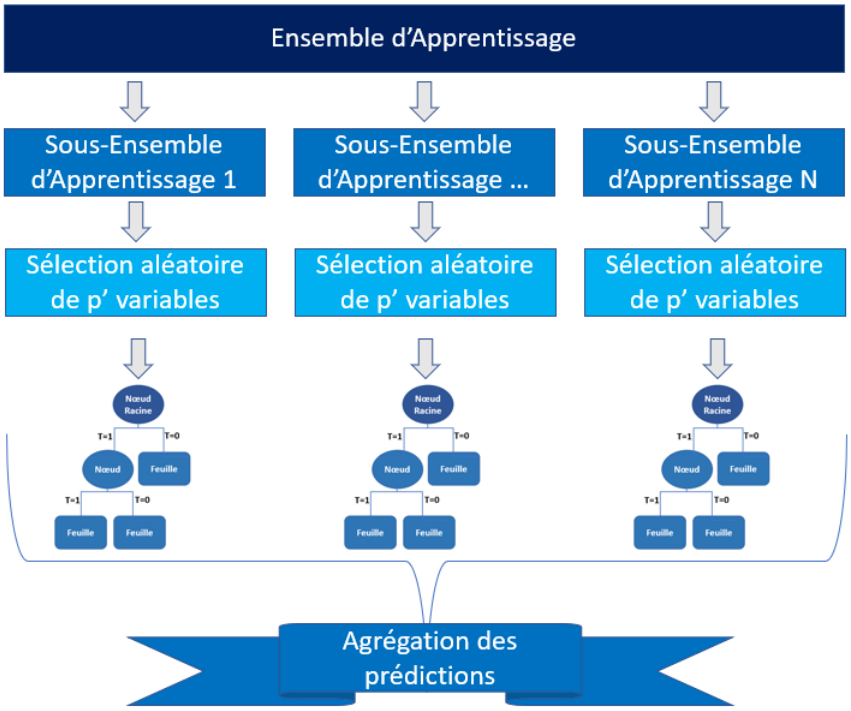

FIGURE 43 – Schéma de l'algorithme des forêts aléatoires

Cinquième partie

Glossaire

— **ONU** : Organisation des Nations Unies

— **FFA** : Fédération Française de l'Assurance

— **APCA** : Assemblée Permanente des Chambres d'Agriculture

— **APREF** : Association des professionnels de la réassurance en France

— **AGRESTE** : Service de la statistique et de la prospective du Ministère de l'Agriculture, de l'Agroalimentaire et de la Forêt

— **ALFA** : Agence pour la lutte contre la fraude à l'assurance

— **NASA** : National Aeronautics and Space Administration

— **ACTA** : de Coordination Technique Agricole

— **USDA** : United States Department of Agriculture

Sixième partie

Bibliographie

— OCDE, **Assurance et risques environnementaux : Une analyse comparative du rôle de l'assurance dans la gestion des risques liés à l'environnement** (2004)

— OCDE, **Gestion des risques dans l'agriculture** (2010)

— Barnett, B.J., O. Mahul, **Weather index insurance for agriculture and rural areas in lower-income countries** (2007)

— OCDE, **Climate Change and Agriculture** (2010)

— AXA Coporate Solutions, **Parametric Insurance : A fitting solution for the weather-sensitive** (2016)

— The World Bank, **Weather Index Insurance For Agriculture : Guidance for Development Practitioners** (2011)

— Hardaker J.B., Gudbrand L., Anderson J.R. , Huirne R.B.M, **Coping with Risk in Agriculture 3rd Edition : Applied Decision Analysis** (2015)

— Barrieu P., **Produits dérivés météorologiques et environnement. PhD thesis, HEC** (2002)

— Hull J., **Options, Futures, and Other Derivatives** (2015)

— Aderson R.W., Danthine J., **The time pattern of hedging and the volatility of future prices** (1983)

— Jewson S., Brix A., **Weather Derivative Valuation : The Meteorological, Statistical, Financial and Mathematical Foundations** (2005)

— Paglia A., Phelippe-Guinvarc'H M.V., **Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique. Bulletin Français d'Actuariat Vol. 11, n22** (2010)

— Matheron G., **Traité de géostatistique appliquée, tome I. In E. Technip (ed.), "Mémoires du Bureau de recherches géologiques et minières"** (1963)

— Pettorelli N., **The Normalized Difference Vegetation Index** (2013)

— Richer de Forges A.C., Feller C., Jamagne M., Arrouays D., **Perdus dans le triangle des textures. Étude et Gestion des sols. Volume 15, numéro 2** (2008)

— Moeys J., **The Soil Texture Wizard** (2009)

— Institut technique de l'agriculture biologique, **Comparaison de variétés de céréales en agriculture biologique** (2016)

— Cleveland, W. S., **Robust Locally Weighted Regression and Smoothing Scatterplots, Journal of the American Statistical Association, Vol. 74** (1979)

— Cleveland, W. S., Devlin, S. J., Locally Weighted Regression : **An Approach to Regression Analysis by Local Fitting, Journal of the American Statistical Association, Vol. 83** (1988)

— Berk R.A. **An introduction to Ensemble Methods for Data Analysis** Department of Statistics UCLA (2004)

— Breiman L. **Random Forests. Statistics Department, Berkeley** (2001)

— M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. **LOF : identifying density-based local outliers** (2000)

— Brownlee J. **Machine Learning Algorithms From Scratch** (2016)

— Brownlee J. **Introduction to Time Series Forecasting with Python** (2016)

— Denuit M., Charpentier A. **Mathématiques de l'assurance non vie. Tome 1 : Principes fondamentaux de théorie du risque** (2004)

— Denuit M., Charpentier A. **Mathématiques de l'assurance non vie. Tome 2 : Tarification et Provisionnement** (2005)

Septième partie

Sources

— **Quelles protections face aux aléas climatiques ?**
https://www.insurancespeaker-wavestone.com/2015/11/5019/

— **Munich Re puts 2017 incured cat losses at $135bn**
http://www.insuranceinsider.com/?page_id=1271728&utm_source=Insider-Publishing&utm_medium=Email&utm_content=Untitled2&utm_campaign=Munich+Re+puts+2017+insured+cat+losses+at+%24135bn&utm_cid=40343

— **Données clés 2016 FFA**
https://www.ffa-assurance.fr/content/assurance-de-personnes-donnees-cles-par-annee

— **Article L125-1 du Code des Assurances**
https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006073984&idArticle=LEGIARTI000006792610&dateTexte=&categorieLien=cid

— **Couverture d'assurance des catastrophes naturelles en Europe, modifié de Mills (2009) cf. Figure 3**
http://freakonometrics.hypotheses.org/42766

— **Indemnistation des pertes occasionnées par des aléas climatiques**
https://www.service-public.fr/professionnels-entreprises/vosdroits/F22259

— **La gestion des crises en Chambre d'agriculture**
http://agriculture.gouv.fr/la-gestion-des-risques-en-agriculture

— **Guide Pratique : La gestion des crises en Chambre d'agriculture**
http://www.chambres-agriculture.fr/fileadmin/user_upload/National/002_inst-site-chambres/pages/exploitation_agri/Guide_pratique_gestion_risques.pdf

— **APREF, Développement du Marché Assurance et Réassurance Récoltes en France**
https://www.apref.org/sites/default/files/espacedocumentaire/4-1_note_recoltes_juin_2014.pdf

— **Agreste Primeur**
http://agreste.agriculture.gouv.fr/IMG/pdf/primeur210.pdf

— **Alfa asso, La fraude**
http://www.alfa.asso.fr/fr/content/la-fraude

— **Argus de l'Assurance, Lutte contre la fraude**
http://www.argusdelassurance.com/institutions/lutte-contre-la-fraude-265-m-recuperes-par-les-assureurs-en-2015.109398

— **Les Echos, « 2016, la pire récolte de blé en France depuis 40 ans »**

https://www.lesechos.fr/22/07/2016/lesechos.fr/0211150682458_2016--la-pire-recolte-de-ble-en-france-depuis-40-ans.htm

— **Fonctionnement de l'indice de végétation par différence normalisé**
https://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php

— **Variétés de blé tendre**
http://www.fiches.arvalis-infos.fr/liste_fiches.php?fiche=var&type=512

— **Triangle des textures**
https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_054167