L'INSTITUT agro Rennes Angers

L'Institut Agro Rennes-Angers
☐ Site d'Angers ☒ Site de Rennes

BORDEAUX SCIENCES AGRO

UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

Année universitaire : 2022 - 2023 Spécialité : Ingénieur en agronomie Spécialisation (et option éventuelle) : Sciences halieutiques et aquacoles, préparée à l'Institut Agro Rennes-Angers (Ressources et Ecosystèmes Aquatiques)	**Mémoire de fin d'études** ☒ d'ingénieur de Bordeaux Sciences Agro ☐ de master de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de l'Institut Agro Montpellier (étudiant arrivé en M2) ☐ d'un autre établissement (étudiant arrivé en M2)

Impact des conditions environnementales et socio-économiques sur les sorties de port de flottilles de pêche du quartier maritime de Bayonne et projections futures dans le cadre du changement climatique

Par : Grégoire BOUILLON

Soutenu à Rennes le 14/09/2023

Devant le jury composé de :

Président : Didier GASCUEL

Maître de stage : Benoît SAUTOUR

Enseignant référent : Etienne RIVOT

Jury extérieur : Verena TRENKEL

Fiche de confidentialité et de diffusion du mémoire

Confidentialité

☒ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible (1).

Date et signature du maître de stage (2) : 14/09/23
(ou de l'étudiant-entrepreneur)

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur(3) **Nom Prénom** BOUILLON Grégoire

autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☒ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☐ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'auteur : 14/09/2023

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☒ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☒ la diffusion papier du mémoire uniquement(4)

☒ la diffusion papier du mémoire et la diffusion électronique du résumé

☒ la diffusion papier et électronique du mémoire

Date et signature de l'enseignant :

Pr Didier GASCUEL
Directeur Pôle halieutique,
mer et littoral
L'Institut Agro Rennes-Angers

(1) L'administration, les enseignants et les différents services de documentation de l'Institut Agro Rennes-Angers s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3) Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Remerciements

Je voudrais remercier ici tout d'abord Benoît SAUTOUR et le groupe FUTUR-ACTS que ce soit pour l'opportunité qu'a représenté ce stage et pour les différents évènements où je fus convié et auxquels je participai avec plaisir.

Je voudrais également remercier l'Ifremer d'Anglet l'aide précieuse apportée en soutien de mon stage sans oublier les bons moments lors des pauses. Un merci particulier à Nathalie

Je voudrais également remercier le site de l'UPPA de Montaury pour toutes les personnes qui m'ont accueilli et en particulier Mamadou, Houssam et Mohamad qui m'ont accompagné au quotidien.

Je voudrais remercier Claire qui tout en vivant un heureux événement a tout fait pour m'apporter de l'aide et m'aiguiller sur la bonne piste en tant qu'apprenti chercheur.

Enfin, un merci particulier et très sincère à Noëlle qui m'a encadré tout au long de mon stage et soutenu du mieux qu'elle pouvait dans une période qui fut très compliquée les derniers mois sur le plan personnel. Sa bienveillance, ses conseils m'ont ainsi permis de finaliser mon stage.

Table des matières

Introduction 4
I – Données de départ et préprocessing 5
I.A Présentation des données 5
I-A-1 Données de pêche 5
I-A-2 Les données environnementales 7
I-B Préprocessing : traitement des données 8
I-B-1 Traitement des données de pêche 8
I-B-2 Traitement des données environnementales 15
I-B-3 Création du jeu de données final 16
II – Sélection des variables cibles et création de seuils 17
II-A Présentation de la méthode 17
II-B Résultats 17
II-B-1 Présentation des résultats 17
II-B-2 Objectif de la méthode 20
II-B-3 Limites 20
III – Création des réseaux Bayésiens 21
III-A Présentation de la méthode 21
III-B Apprentissage du réseau bayésien 22
III-B-1 Présentation 22
III-B-2 Les algorithmes d'apprentissage de la structure 22
III-B-3 Choix de l'algorithme et de l'outil de création de réseaux bayésiens 23
III-C Création des réseaux Bayésiens 25
III-D Utilisation de notre réseau bayésien 27
III-D-1 Comparaison des réseaux bayésiens 28
III-D-2 Evaluation à la journée 29
III-D-4 Evolution en fonction des tendances 29
III-D-5 Ajout de connaissances à dire d'experts 31
III-D-6 Limites 32
Conclusion 33
Bibliographie 34
Annexes 36

Introduction

Les flottilles de pêche sont sensibles aux facteurs environnementaux, réglementaires, biologiques et socio-économiques. Les ressources qu'elles exploitent sont fluctuantes, avec une abondance et une localisation variant selon les saisons et les années. L'exploitation de ces ressources est soumise à la réglementation, notamment à travers les quotas ou les arrêtés restreignant les périodes temporelles ou les zones d'exploitation, et son accès est conditionné par plusieurs facteurs. Parmi ces facteurs, certains peuvent être intrinsèques aux navires (longueur, puissance et tonnage jauge brute), liés au contexte socio-économique (prix de vente, quantité de quota restant, rendements de la zone de pêche…) ou encore liés aux conditions océaniques et météorologiques (vagues, direction et force du vent, qualité de l'eau). Ces facteurs peuvent tant restreindre l'accès de manière directe (accès impossible en raison de la capturabilité) qu'indirect (impact sur le rendement) impliquant un choix pour le professionnel. Le choix d'un pêcheur de sortir en mer peut donc être lié à l'état de la mer qui pourrait porter atteinte à la sécurité des pêcheurs mais aussi impacter le comportement des engins, des poissons et in fine entraîner des conséquences sur les rendements.

L'océan n'échappe pas aux conséquences du changement climatique qui impliquent des modifications des paramètres biologiques et physico-chimiques des masses d'eau, avec des modifications déjà observées. Le sixième rapport du GIEC (Intergovernmental Panel on Climate Change (IPCC), 2023) s'accorde à estimer une augmentation de la température de l'air, mais aussi de la surface de l'océan, de son acidification, ou encore d'une augmentation du nombre de jours de vagues de chaleur à l'échelle mondiale. D'autres travaux comme ceux de Bricheno et Wolf (2018) observent une diminution de la moyenne, mais aussi une augmentation des extrêmes pour la hauteur des vagues sur la côte Atlantique européenne. Ces changements sont susceptibles d'impacter le milieu d'évolution des navires et potentiellement de conditionner leur activité. Des changements ont également été décrits sur la ressource, avec des modifications de l'abondance ou des distributions latitudinales ou en profondeur (Brander, 2010; Perry et al., 2005).

C'est dans ce contexte global que la région Nouvelle-Aquitaine a fait preuve de proactivité en mobilisant un groupe d'experts scientifiques pluridisciplinaires pour rassembler les connaissances nécessaires aux acteurs du territoire afin de construire leur stratégie d'adaptation au changement climatique. Deux ouvrages furent produits, le premier ayant été centré sur les impacts en région Aquitaine tandis que le second a aussi intégré une composante anticipation (AcclimaTerra and Le Treut, 2018; Le Treut, 2013). Ces deux ouvrages contiennent chacun un chapitre sur le volet pêche. Dans la continuité de ces travaux, le Réseau Régional de Recherche Futurs-ACT fut créé en 2020, toujours sous l'impulsion de la région Nouvelle-Aquitaine. Son objectif principal est d'anticiper le changement climatique en mobilisant la communauté scientifique pour accompagner les territoires dans leur transition.

Un premier projet indépendant de cette dynamique sur le changement climatique fut piloté par le comité interdépartemental des pêches maritimes et des élevages marins 64-40 (CIDPMEM64-40) associant les scientifiques et les pêcheurs. Focalisé sur la flottille du quartier maritime de Bayonne, celui-ci a permis d'estimer le poids socio-économique de la filière pêche en établissant une typologie (EPOSE) de navires et en mobilisant les données provenant de différentes sources complémentaires (Caill-Milly et al., 2019; Gallet et al., 2019). Un deuxième projet Vents&Marées financé par DLAL-FEAMP a eu lieu entre 2021 et 2022 et fut conduit par une équipe pluridisciplinaire issue de différents organismes. En s'appuyant sur le premier projet, il permit l'identification de certains facteurs clés conditionnant les sorties en mer des navires de pêche en lien avec le changement climatique. Des facteurs environnementaux liés à la météorologie marine et socio-économiques ont pu être identifiés. Des seuils de conditions environnementales et socio-économiques notamment sur le prix furent mis en évidence et confrontés aux professionnels du milieu (Bru et al., 2022).

Partant de ces résultats, et en lien avec les objectifs de Futurs-ACT qui finance ce stage, l'étape suivante est de proposer des scénarios d'évolution de l'activité de flottilles immatriculée au quartier maritime de Bayonne reposant sur les probabilités de sorties qui dépendent de la variation des facteurs clés. Ces facteurs clés sont représentés à l'aide de données historiques ou en intégrant des scénarios à dire d'experts en s'appuyant sur la documentation. La méthodologie ici utilisée est la modélisation par réseaux bayésiens. Le réseau bayésien est un modèle graphique probabiliste représenté sous la forme d'un graphe acyclique utilisant des boîtes représentant les variables aléatoires, reliées entre elles par des arcs orientés qui symbolisent les relations de dépendance probabiliste. Cette méthode a pu être utilisée dans différents travaux qui concernent le milieu halieutique, notamment pour analyser les risques pour les stratégies politiques européennes en synthétisant les connaissances (Bastardie and Brown, 2021) ou

pour prédire les réponses de l'écosystème au changement de température, de production primaire ou des captures (Trifonova et al., 2017). Elle permet de pallier certaines limitations rencontrées lors de la construction de nombreux modèles écologiques comme celle des données manquantes ou celle des covariables fortement corrélées entre elles, qui peuvent entraver l'estimation de paramètres. Il apparaît que cette nouvelle méthodologie n'est pas toujours utilisée en écologie à son plein potentiel (Ramazi et al., 2021).
L'objectif de ce projet est d'introduire une méthodologie utilisant les réseaux bayésiens pour étudier l'impact du changement climatique et des variables socio-économiques sur la probabilité de sortie en mer de la flottille du quartier maritime de Bayonne. Pour cela, une première étape nécessite d'agréger les données, de les nettoyer et de créer la variable cible (***Sortie***) et les différentes variables ayant potentiellement un impact. Parmi ces variables seront identifiés des variables clés lors d'une seconde phase de datamining utilisant des analyses par arbres de décisions conditionnelles. Cette seconde phase permettra aussi de mettre en évidence des seuils pour les variables clés qui permettra de discrétiser nos variables pour la troisième phase. Cette dernière phase du projet consiste à créer nos réseaux bayésiens à partir uniquement de nos données et d'étudier les différentes possibilités d'utilisations de ces réseaux en ayant pour objectif de pouvoir prédire une probabilité de sortie en fonction d'états précis des différentes variables ou de prédire l'évolution de cette probabilité en fonction de scénarios climatiques. Ce projet a pour ambition future d'intégrer plus fortement un volet social, ce qui implique une réelle volonté d'utiliser des méthodes dont la présentation est simple pour faciliter la concertation. La possibilité d'apporter une connaissance experte à posteriori a ainsi était explorée.

I – Données de départ et préprocessing

Avant de créer des réseaux bayésiens, il est nécessaire d'effectuer toute une phase préparatoire pour transformer les données sous une forme exploitable et en adéquation avec le sujet. Pour commencer, le préprocessing consiste à récupérer les données, les nettoyer et les traiter en vue des phases suivantes.

Un préprocessing des données disponibles a déjà été mis en place lors du projet Vents&Marées (2000-2019). Dans le cadre de ce stage de master 2, deux nouvelles années ont pu être prises en compte dans l'étude (2020-2021). La période d'étude s'étend donc de l'année 2000 jusqu'à 2021.

I.A Présentation des données

Les données utilisées pour ce projet sont divisées en 2 catégories : les données de pêche (incluant les données socio-économiques) et les données environnementales.

I-A-1 Données de pêche

Les données de pêche sont issues de la base de données de l'Ifremer HARMONIE. Cette base de données utilise SACROIS, un algorithme qui croise les données en provenance de différentes sources :

- Les déclarations de pêche dans les logbooks ou par les fiches de pêche ;
- Les ventes des pêcheurs issues des criées (données RIC) ;
- Les données qui peuvent être extrapolées à partir des données satellitaires issues par exemple de la VMS ;
- D'autres données issues d'enquêtes (par exemple enquêtes d'activité de pêche).

Cette base de données a pour objectif d'estimer au mieux qui pêche, comment, où, quoi, combien, l'effort investi et ce qui en est retiré (tonnages et valeurs).

Dans le cadre du projet Vents&Marées, les données utilisées couvraient la période 2000-2019. Elles ont donc été reprises et ont été complétées par une extraction de toutes les séquences de pêche sur la période 2000-2021 des navires de pêche ayant pour quartier maritime Bayonne, tous engins confondus. Une séquence de pêche correspond à une action de pêche qui est définie par le navire, l'engin, la date et le rectangle statistique CIEM où l'action de pêche est effectuée. Une séquence de pêche peut donc avoir plusieurs lignes pour chaque espèce prélevée et pour chaque rectangle statistique traversé lors de l'action de pêche. Ces données ont été anonymisées pour respecter les règles par souci de confidentialité.

Tableau 1. Résumé du contenu de la base « activité flottilles » de départ.

Variable	Classe	Premieres_valeurs
year	double	2000, 2000, 2000, 2000, 2000, 2000
month	double	9, 9, 9, 4, 9, 9
MAREE_ID	integer	13293082, 13293080, 13293091, 13510385, 13293087, 13293079
MAREE_DATE_DEP	double	2000-09-09, 2000-09-11, 2000-09-22, 2000-04-03, 2000-09-14, 2000-09-23
MAREE_DATE_RET	double	2000-09-09 23:59:59, 2000-09-11 23:59:59, 2000-09-22 23:59:59, 2000-04-03 00:00:00, 2000-09-14 23:59:59, 2000-09-23 23:59:59
LIEU_COD_DEP_SACROIS	integer	XBA, XBA, XBA, CBA, XBA, XBA
LIEU_COD_RET_SACROIS	integer	XBA, XBA, XBA, CBA, XBA, XBA
SEQ_ID	integer	26578506, 26637185, 26630683, 26898608, 26630060, 26614701
DATE_SEQ	double	2000-09-09, 2000-09-11, 2000-09-22, 2000-04-03, 2000-09-14, 2000-09-23
SECT_COD_SACROIS_NIV5	character	R16E8, R16E8, R16E8, R16E8, R16E8, R15E8
ENGIN_COD	integer	GEN, GEN, GEN, GEN, GEN, GEN
ENGIN_COD_SACROIS	integer	GTR, GEN, GEN, GNS, GTR, GEN
METIER_DCF_6_COD	integer	GTR_DEF_0_0_0, GNS_DEF_0_0_0, GNS_DEF_0_0_0, GNS_CEP_0_0_0, GTR_DEF_0_0_0, GNS_DEF_0_0_0
DIMENSION	integer	12500, 12500, 12500, 12500, 12500, 12500
MAILLAGE	double	NA, NA, NA, NA, NA, NA
TP_NAVIRE_SACROIS	double	14, 13, 11, NA, 9, 12
TP_NAVIRE_SIPA	double	14, 13, 11, NA, 9, 12
TPS_MER	double	8, 8, 8, NA, 8, 8
ESP_COD_FAO	integer	HOM, HOM, CTL, CTL, SOL, HOM
QUANT_POIDS_VIF_SACROIS	double	30, 25, 12, 89.1, 20, 4
MONTANT_EUROS_SACROIS	double	18.56, 7.87, 21.31, 192.26, 208.26, 2.71
Qam	integer	BA, BA, BA, BA, BA, BA
annee_construction	integer	1993, 1993, 1993, 1993, 1993, 1993
Loa	double	10.65, 10.65, 10.65, 10.65, 10.65, 10.65
Power_Main	double	131, 131, 131, 131, 131, 131
Ton_Ref	double	4, 4, 4, 4, 4, 4
Vms_Code	integer	N, N, N, N, N, N
typo_epose	integer	Fileyeur_10-12_m, Fileyeur_10-12_m, Fileyeur_10-12_m, Fileyeur_10-12_m, Fileyeur_10-12_m, Fileyeur_10-12_m
Cod_Nav	integer	Nav_7717, Nav_7717, Nav_7717, Nav_7717, Nav_7717, Nav_7717

Rows: 4,681,405
Columns: 29

Tableau 1 : Tableau présentant les données de pêche extraites de la base HARMONIER de l'IFREMER. Extrait du rapport final de Vents&Marées.

Le quartier maritime de Bayonne est composé de 4 ports : Capbreton (**ABA**), Ciboure / Saint-Jean-de-Luz (**CBA**), Bayonne / Boucau (**XBA**) et ceux qui opèrent sur l'Adour (**GBA**).
Nous allons nous concentrer sur les navires immatriculés dans le quartier maritime de Bayonne et qui partent principalement d'un de ces 4 ports.

A noter qu'une correction a été apportée sur les fileyeurs partant de Bayonne / Boucau et utilisant l'engin « **GND** » (filets maillants dérivants) : ils ont été identifiés comme opérant sur l'Adour, ce qui correspond à la réalité selon les pêcheurs.

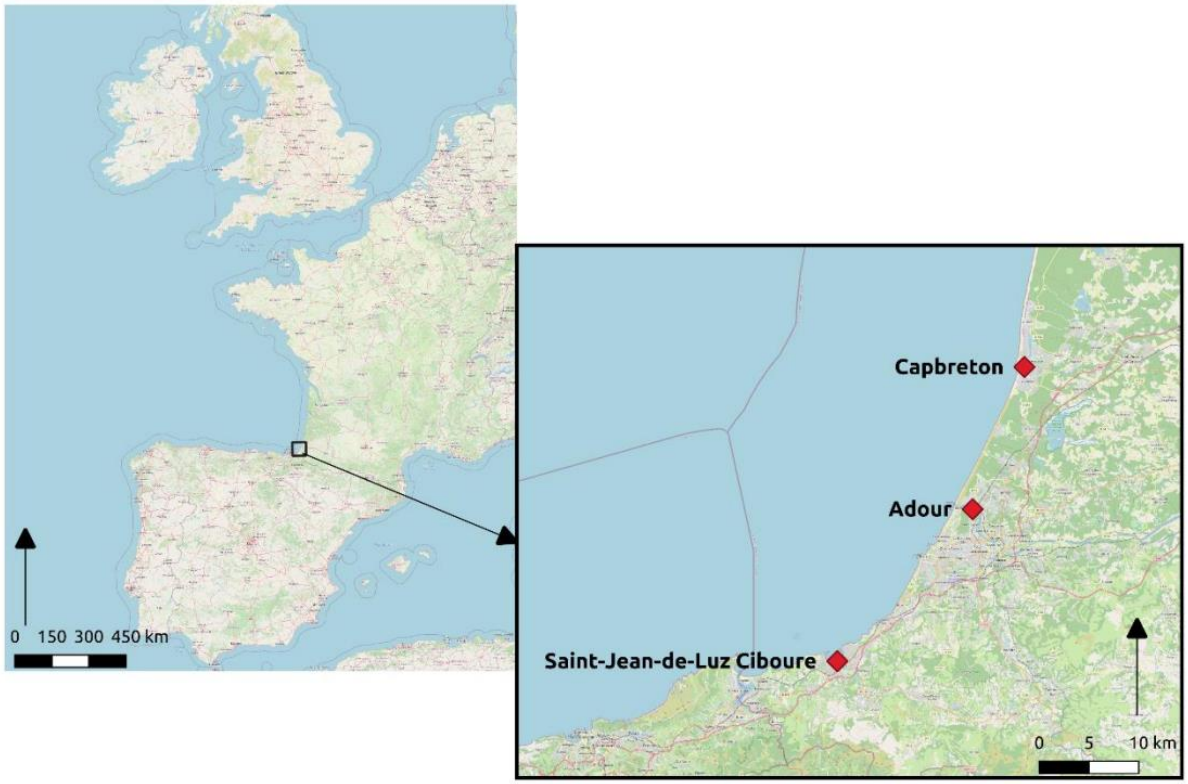


Figure 1 : Cartes présentant la région d'étude et les ports d'intérêts. Extrait du rapport final de Vents&Marées.

I-A-2 Les données environnementales

Les données environnementales englobent plusieurs données qui concernent le milieu dans lequel évoluent les navires :

- des données sur les vagues ;
- des données diverses comme la température, des indices (Tempête, hydrologique et Liga) ou encore le débit de l'Adour.
- des données de vent.

Les données sur les vagues :

Ces données correspondent à la direction (DIR), la hauteur en mètres (Hs) et la période en secondes (Tp). Ces données sont disponibles de 1993 à 2021, donc sur toute notre période d'étude. Elles sont à l'échelle horaire, mesurées dix minutes avant l'heure et ne sont pas spatialisées car obtenues par la bouée au large Anglet, en un point précis. Les données sont ensuite retravaillées pour être corrigées à l'aide des travaux d'Aurélien Callens (Callens et al., 2020) pour correspondre à la localisation de la côte au lieu de la bouée.

Les données de vent

La force du vent (FF) et sa direction (DD) sont les deux composantes prises en compte. Les données proviennent de Météo France où elles sont mesurées dix minutes avant chaque heure en un point précis qui correspond à la station météo de Biarritz-Anglet. La période extraite est de 2000 à 2021 pour compléter la base de données existante et correspondre à notre période d'étude.

Les données diverses

Ces données regroupent différentes variables journalières permettant de caractériser des situations hydroclimatiques. Nous pouvons retrouver :

- ***Débits_Adour*** et ***Débits_Nivelle***, deux variables indiquant le débit de la rivière qu'ils désignent ;
- Les variables de températures (***Tmin_air***, ***Tmax_air***, ***Tmoy_air***) ;
- Des indices journaliers synthéthiques : ***Indice_Liga***, ***Indice_Tempete***, ***Indice_Hydrologique*** (tous sans unité).
  Le liga est un mucilage fruit du stress d'un phytoplancton qui s'agglomère dans les filets de pêche et pouvant entraîner une baisse du rendement selon les pêcheurs. Son indice est calculé à partir du rayonnement global (source de lumière pour la photosynthèse), de la température maximum de l'air représentant un proxy du réchauffement des eaux de surface, du débit de l'Adour qui est une source de sels nutritifs pour la photosynthèse et de la hauteur des vagues (Susperrégui et al., 2015).
  Le second indice est lui calculé à partir de la vitesse maximale du vent et de la hauteur des vagues.

Nous voulions aussi considérer des données traitant de la qualité de l'eau, cette composante étant une inquiétude et un élément que les pêcheurs évoquent et prennent en considération. Malheureusement, nous n'avons pas pu obtenir ces données, une piste ayant été les données de rejets des stations épurations qui peuvent être des données sensibles. Malgré tout, ***Indice_Liga*** est la variable que l'on peut considérer d'une certaine façon comme un proxy de la qualité de l'eau (Susperregui et al., 2012, 2010).

I-B Préprocessing : traitement des données

Lors de cette phase de nettoyage et traitement, une méthodologie proche de celle utilisée dans Vents&Marées fut appliquée mais dans une version améliorée, en corrigeant des oublis ou en optimisant le script R par exemple.

I-B-1 Traitement des données de pêche

La première étape était de sélectionner les navires en fonction de critères pour pouvoir répondre au mieux à la question. Notre objectif était donc de sélectionner les navires représentatifs de la flottille locale. Il était nécessaire de faire un tri en fonction du métier (engin de pêche) et de l'intensité de l'activité.

Sélection de la flottille d'intérêt en fonction de ses caractéristiques techniques

Un travail préliminaire a pu être effectué (Caill-Milly et al., 2019; Gallet et al., 2019) pour établir une typologie intitulée « EPOSE » qui fragmente la flottille de Bayonne en plusieurs groupes définis en fonction du métier principalement pratiqué et des caractéristiques du navire. Elle présente un intérêt important dans notre étude car elle permet de regrouper les navires dans des groupes ayant des pratiques similaires, ce qui permet d'effectuer des opérations en fonction de ces derniers quand cela est pertinent.

Suivant les objectifs du travail préliminaire EPOSE, la typologie a été définie et appliquée pour caractériser les navires présents en 2016. Cette information est, de ce fait, absente pour les navires présents avant ou après cette date, notamment pour ceux qui sont sortis de la flottille, entrés ou ont subi une reconversion. Il a donc fallu dans un premier temps l'étendre pour correspondre à notre période de prise en compte de l'activité.

Pour pouvoir faire cela, nous avons utilisé les informations disponibles dans la base et procédé en plusieurs étapes :

1. Une première variable a été calculée par codage : ***TYPE_ENGIN***. Pour chaque ligne, nous récupérons la valeur de la variable « ***ENGIN_COD_SACROIS*** » qui est le code FAO de l'engin utilisé par le navire qui a été déterminé par l'algorithme SACROIS, puis nous lui attribuons un type d'engin : Filet, Ligne, Bolinche, Chalutier ou Autres.

2. Certaines lignes de la variable « ***ENGIN_COD_SACROIS*** » étant vides, nous avons imputé l'information en lui affectant la valeur pour la même ligne de la variable « ***ENGIN_COD*** » si cette dernière est non nulle. La première variable correspond à l'engin que l'algorithme SACROIS détecte et attribue à l'opération en croisant les données. La seconde correspond à ce qui est déclaré par le pêcheur. Même si l'information est généralement identique, il peut subsister des différences, que ce soit dû à une erreur humaine (oubli par exemple) ou de la machine qui attribue aucune valeur. Même si nous préférons utiliser la variable issue de SACROIS considérée plus fiable, la seconde est appréciable si la première est absente.

3. Puis une seconde variable « ***TYPE_ENGIN_FAV*** » a été créée : celle-ci prend les mêmes valeurs que « ***TYPE_ENGIN*** ». Elle correspond au type d'engin le plus utilisé sur une année. Ceci est déterminé par année en prenant l'engin qui a été utilisé le plus de jours dans l'année. Un jour est compté quand une séquence utilisant l'engin apparaît à un jour donné (***DATE_SEQ***).

4. Nous pouvons maintenant attribuer les typologies EPOSE pour chaque année dans une variable nommée « ***new_typo_epose*** » à l'aide de la variable précédemment créée et de « ***Loa*** » qui correspond à la longueur du navire. Nous avons donc les typologies suivantes :

- Les fileyeurs < à 10 m
- Les fileyeurs de 10 à 12 m
- Les fileyeurs de 12 à 20 m
- Les fileyeurs > 20 m
- Les ligneurs < 15 m
- Les ligneurs > 15 m
- Les bolincheurs
- Les chalutiers < 15 m
- Les chalutiers de 15 à 25 m
- Les chalutiers > 25 m
- Autres navires

Un navire avait une unique valeur typologie EPOSE lors du projet originel, mais elle peut désormais évoluer en fonction des années si le navire de pêche a changé de métier principal.

=> Dans ce projet, nous nous concentrons uniquement sur les ligneurs, les fileyeurs et les bolincheurs sélectionnés par filtrage sur ***TYPE_ENGIN_FAV***

Nous voulions aussi nous concentrer sur les navires de pêche qui sont potentiellement les plus impactés par les facteurs climatiques. Nous avons considéré que cela correspondait aux navires effectuant des courts séjours et qui sortent de manière quotidienne, à savoir : les ligneurs inférieurs à 15 mètres, les fileyeurs inférieurs à 20 mètres et les bolincheurs.

=> Nous avons donc filtré les lignes qui ont à la fois la valeur « ***TYPE_ENGIN_FAV*** » et celle de « ***Loa*** » qui correspondent à ces conditions. La combinaison des deux correspond à une typologie EPOSE : le filtre aurait pu être appliqué sur la variable ***new_typo_epose*** pour un résultat identique.

Enfin, il est nécessaire de sélectionner le navire en fonction de sa réelle activité en partance des ports d'exploitation du quartier maritime de Bayonne. Nous avons donc procédé également en plusieurs étapes :

1. Tout d'abord, nous créons une variable « ***ports*** » qui est une liste contenant les quatre codes attribués aux quatre ports du quartier maritime de Bayonne : « XBA, CBA, ABA, GBA » ;
2. Nous avons ensuite créé la variable « ***Port_Fav*** » qui de manière analogue à ***TYPE_ENGIN_FAV*** détermine chaque année le port d'où le navire est parti le plus souvent en utilisant la variable « ***LIEU_COD_DEP*** ». Pour cela, la valeur n'est comptée qu'une seule fois par marée, représentée par la variable « ***ID_MAREE*** » ;
3. Puis une dernière variable fut créée, « ***part_dep_BA*** ». Elle représente la part des marées qui sont en partance d'un des ports du quartier maritime de Bayonne.

$$partdepBA_{année} = \frac{nombre\ de\ marées\ avec\ LieuCodDep\ \in ports}{nombre\ total\ de\ marées}\ pour\ chaque\ année$$

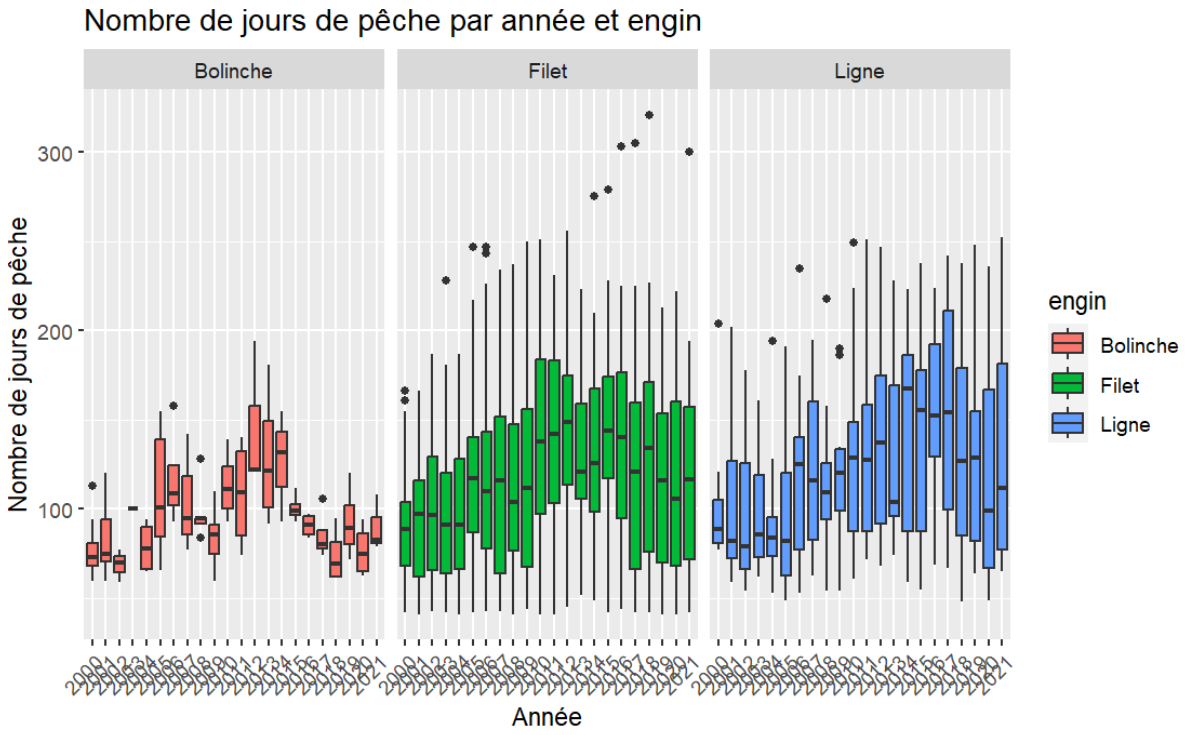
=> L'objectif étant de cibler la flottille opérant réellement sur le quartier maritime de Bayonne, nous avons conservé les navires de pêche ayant pour valeur de « ***Port_Fav*** » un des quatre ports du quartier maritime et dont la valeur de « ***part_dep_BA*** » est de minimum 50 %.

Sélection en fonction de l'activité du navire

Il est ensuite nécessaire de sélectionner des navires de pêche ayant une activité représentative de leur flottille, donc considérés comme actifs pour pouvoir espérer des résultats robustes. Nous avons donc créé des indicateurs globaux d'activité sur toute la période et annuels pour calculer le nombre de jours de pêche effective. Cette étude sera menée par engin en raison de la spécificité de chacun en matière d'activité.

Engin	Moyenne	Minimum	Q1	Médiane	Q3	Max	Ecart-type
Bolinche	74,69	12	58,5	82,09	92,15	153	31,54
Filet	78,67	4.67	40,13	72,92	110,52	223	47,77
Ligne	79,15	3	45,5	73	100,67	210,09	52,42
Non distingué	79.26	3	40,53	74,39	109,57	223	48,46

Tableau 2 : Indicateurs globaux du nombre de séquences par année en fonction du type d'engin sur toute la période 2000-2021



Nous pouvons voir sur la figure 3 des tendances similaires, mais une grande variabilité entre les engins si l'on compare les premiers et troisièmes quartiles. L'objectif étant de détecter les navires peu actifs, il semble judicieux de s'intéresser au premier quartile par flottille séparément vu la disparité des niveaux des premiers quartiles en fonction du type d'engin comme on peut le voir dans le tableau sur toute la période ci-dessus.

=> le seuil utilisé pour identifier des navires actifs est donc au sein de chaque engin, le premier quartile global (voir tableau 1)

Changer d'unité statistique : ne garder qu'une date de sortie par jour de pêche

Lorsque les différents tris ont pu être effectués pour sélectionner les navires représentatifs sur la période considérée, nous avons décidé de ne conserver qu'une information à la journée. Ce qui nous intéresse, c'est de savoir si le navire est sorti ou non une journée donnée et mettre en relation cette donnée avec une information de contexte souvent disponible quotidiennement.

L'objectif est donc d'agréger toutes les lignes d'une date (***DATE_SEQ***) fixée et un navire donné en une unique ligne, où sont conservées différentes variables d'intérêts comme l'espèce principalement ciblée, son tonnage et son prix. On perdra alors l'information du carré statistique s'il y en avait plusieurs.

Pour faire cela, nous avons procédé en plusieurs étapes :

1. Création d'une nouvelle variable « ***Total_kg*** » qui correspond au tonnage total de la journée pour le navire ;
2. Nous pouvons à partir de cette variable calculer la part sur la journée de chaque espèce et créer une variable « ***ESP_cible*** » qui correspond à l'espèce ayant le tonnage le plus important sur la journée, la valeur étant donc égale pour toutes les lignes d'un même navire (***Cod_Nav***) pour une même date (***DATE_SEQ***). Lorsque deux espèces ont le même tonnage, celle sélectionnée comme espèce cible est l'espèce ayant le plus de valeurs. Il était possible de faire la sélection dans l'autre sens. S'il y a égalité parfaite, alors une valeur est sélectionnée de manière arbitraire ;
3. Pour ne conserver qu'une ligne unique par navire et date, il suffit alors de sélectionner les lignes où l'espèce cible correspond à l'espèce pêchée (***ESP_COD_FAO***).

Ces lignes traduisent implicitement le fait que le bateau est bien sorti le jour correspondant. Nous pouvons alors ajouter la variable « ***Sortie*** » qui a pour toutes les lignes la valeur 1. La variable « ***Sortie*** » est notre variable cible.

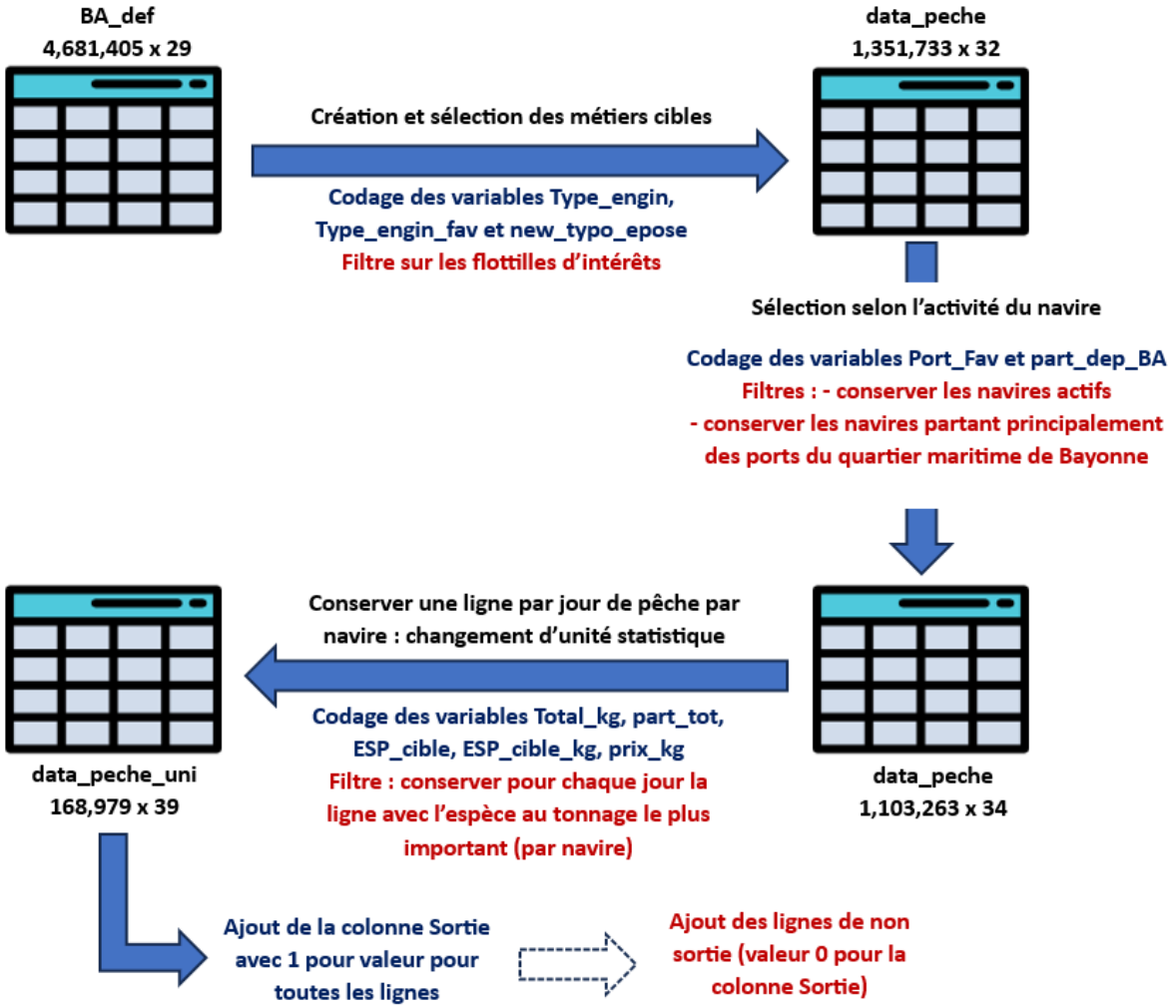


Figure 3 : Récapitulatif des manipulations effectuées sur les données
En bleu : opérations sur les colonnes
En rouge : opérations sur les lignes

Reconstruire les jours de non sortie : créer les zéros

Maintenant, notre jeu de données contient bien une ligne pour chaque jour où le navire a déclaré avoir pêché. Il nous faut rajouter les lignes où le navire n'est pas sorti (***Sortie*** = 0).

Pour cela, nous avons créé une fonction qui prend pour paramètre un jeu de données similaire au précédent mais seulement pour un navire pour une année et qui procède de la façon suivante :

1. Nous regardons le premier départ et le dernier retour de marée du navire sur l'année. Nous considérons ainsi que la fin de l'année et le début de l'année sont des sorties qui ne sont pas liées aux conditions, mais des zéros que l'on qualifie de structurels (période de non-travail par choix) ;
2. Une liste contenant toutes les dates entre le premier départ et le dernier retour est créée. Avec notre jeu de données en paramètre, nous avons ainsi les jours où le bateau est effectivement sorti. Nous pouvons donc retirer ces jours de la liste précédemment créée, et il reste les jours où le bateau est censé être en activité mais a fait le choix de ne pas sortir ;
3. Nous pouvons ensuite compléter chaque ligne qui correspond à chaque date de non-sortie avec les variables de notre jeu de données en assignant 0 à la variable ***Sortie*** et en complétant ce qui est possible, notamment les valeurs « redondantes » comme les caractéristiques du navire, la typologie EPOSE ou le port favori qui sont constantes.

Il suffit ensuite de lancer cette fonction pour tous les navires pour chaque année. Nous obtenons à la fin un jeu de données avec toutes les non-sorties que nous assemblons alors avec le jeu de données précédent.

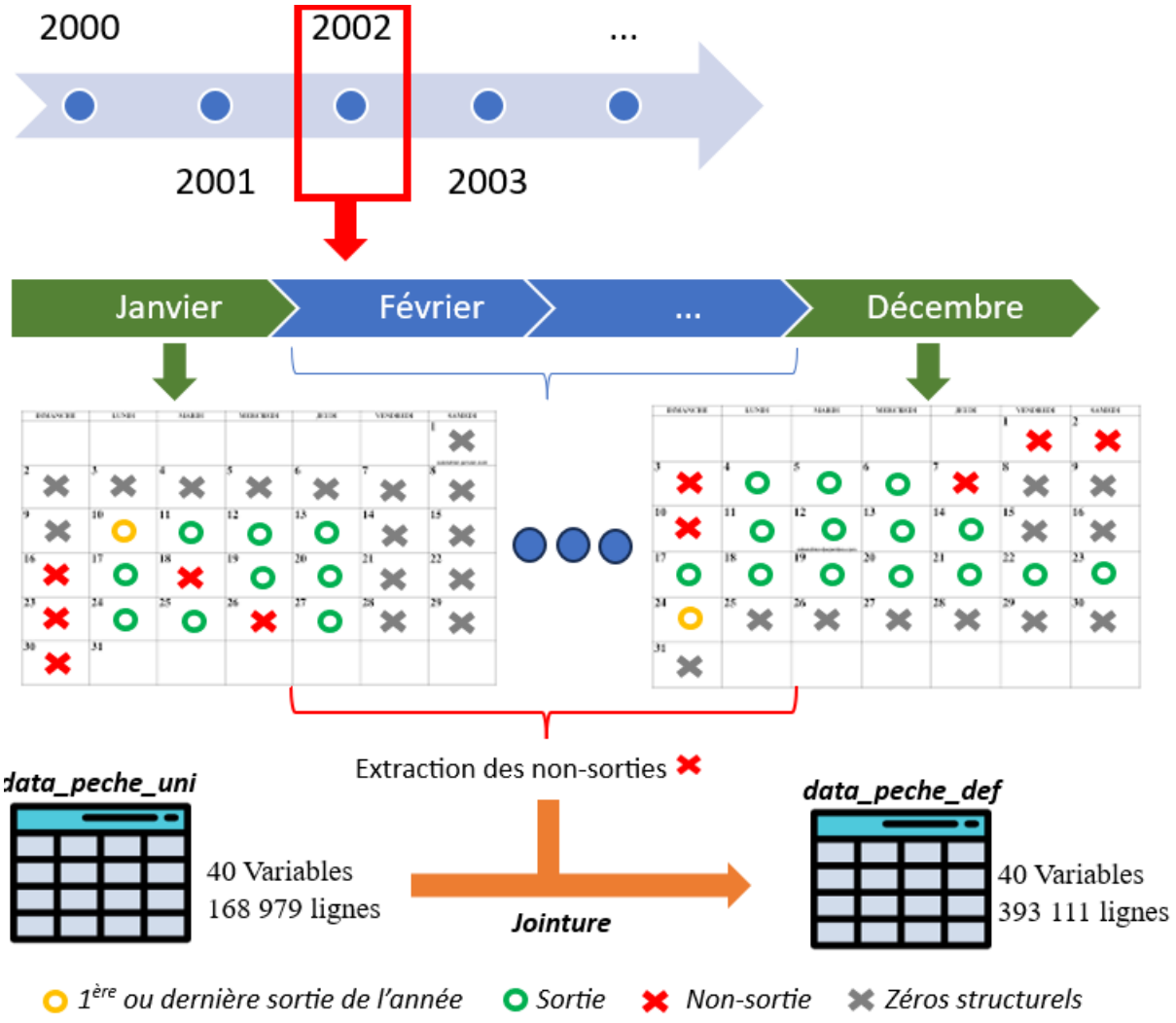


Figure 4 : Graphique présentant le fonctionnement de la fonction de création des zéros (non-sorties). Exemple de l'année 2022.

Lorsque nous créons les zéros, certaines variables importantes n'ont alors aucune valeur, en particulier ***ESP_cible*** et ***prix_kg*** :

- Pour l'espèce ciblée, nous avons choisi d'affecter à chaque jour où la valeur est manquante la valeur la plus représentée dans les autres navires ayant une valeur pour ce jour et qui font partie de la même typologie EPOSE. Si aucun navire de la même typologie n'est sorti, alors c'est la dernière espèce ciblée connue du navire qui est affectée.
- Pour le prix, le prix assigné est le prix au kilogramme moyen des navires de la même typologie étant sortis le jour où la valeur est manquante. Si aucun navire de la même typologie n'est sorti, alors c'est la dernière valeur du prix au kilogramme connue du navire qui est affectée.

Traitement des zéros structurels

Nous avons évoqué cette notion lorsque nous avons traité la création des non-sorties dans notre jeu de données. Un zéro structurel est une non-sortie qui ne peut pas être attribuée aux conditions climatiques ou socio-économiques. Conserver ces zéros pourrait biaiser le résultat en corrélant une non-sortie avec des facteurs climatiques, alors que l'on est sûr que cela n'est pas lié.

Tout d'abord, nous avions évoqué la possibilité qu'un navire appartienne au quartier maritime de Bayonne sans opérer sur la zone. Malgré notre sélection précédemment exposée, certains navires peuvent tout de même effectuer une partie de la saison ou exceptionnellement des actions de pêche dans d'autres lieux. Ces lignes ne nous intéressent pas, car nous nous focalisons sur la sortie / non-sortie depuis les ports du quartier maritime de Bayonne. Mais ces jours-là ne correspondent pas à des non-sorties, puisque le navire de pêche est parti pêcher, mais depuis un autre port. Nous avons alors décidé de supprimer ces lignes, ce qui laisse des dates où il n'y a ni sortie, ni non-sortie. Nous pouvons tout de même soulever un premier biais pour ces navires-là : si le bateau n'est pas sorti durant une période où il opère dans un autre port, cette non-sortie est prise en compte.

D'autres zéros structurels peuvent exister pour des raisons logistiques. Lorsque l'on s'intéresse aux nombres de sorties par jour de pêche, nous remarquons que, pour les ligneurs et les fileyeurs qui sont partis depuis Ciboure / Saint-Jean-de-Luz (CBA), il y a beaucoup moins de sorties les vendredis et les samedis. Cette observation est liée au fait que la criée est fermée le samedi et le dimanche. Nous avons donc supprimé tous les jours correspondant à des vendredis et des samedis, que ce soit des 0 pour les fileyeurs et les ligneurs ayant pour port favori Ciboure ou Saint-Jean-de-Luz ou des 1 s'ils sont partis depuis ce port. Nous avons effectué les mêmes actions pour les bolincheurs, ce qui s'explique par le fait que les bolincheurs partent tous de ce port et qu'ils ciblent des espèces nécessitant un circuit court, la différence du nombre de sorties entre ces deux jours et les autres jours de la semaine est d'autant plus marquée.

Enfin, il existe différents arrêtés et fermetures temporaires limitant l'accès à l'Adour pour les pêcheurs le samedi et le dimanche. De manière analogue, nous avons donc supprimé ces deux jours pour les ligneurs et les fileyeurs liés à ce port (GBA).

Que faire pour les années Covid ?

Rapidement, la question s'est posée de la prise en compte des années 2020 et 2021 qui sont spécifiques en raison de l'épidémie de COVID qui a probablement altéré les comportements. C'est pour cela que nous nous sommes intéressés à la tendance du nombre de séquences sur la période

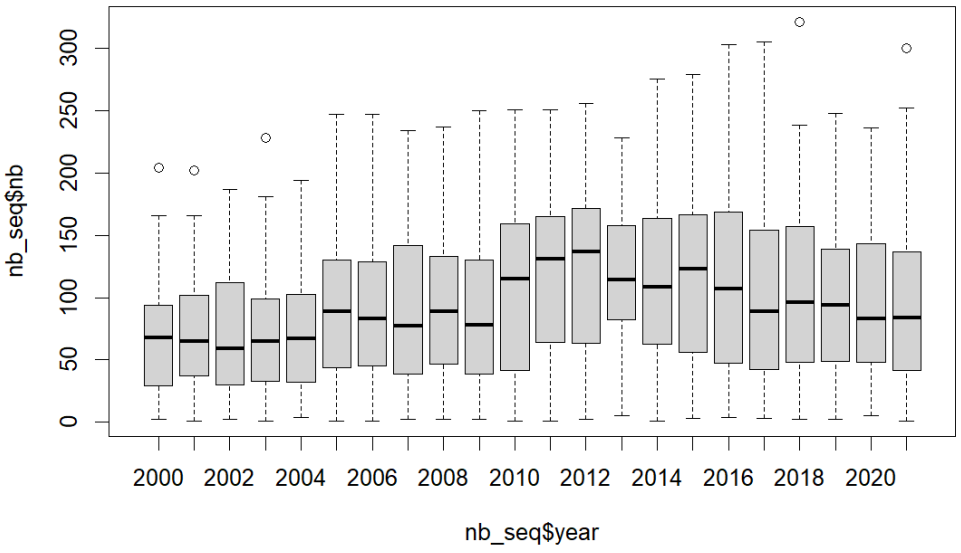


Figure 5 : Boxplots du nombre de séquences en fonction des années

Graphiquement, il ne semble pas y avoir de différences marquées entre ces années-là et les années précédentes. Nous conserverons ces années pour notre étude.

I-B-2 Traitement des données environnementales

Obtention des données sur les vagues et sur le vent

Notre jeu de données sur les vagues est récupéré à l'aide d'un script R issu du travail de A. Callens (Callens et al., 2020) qui interroge diverses données libres accessibles en ligne. Son objectif est de corriger les données de vagues de la bouée d'Anglet pour qu'elles correspondent aux vagues au niveau du port. Lors du projet précédent Vents&Marées, A. Callens avait alors fourni un jeu de données directement corrigé. Dans ce projet, nous avons étendu la période en ajoutant les deux années 2020 et 2021. Nous avons donc dû reproduire ce travail pour ajouter les valeurs corrigées de la période manquante.

Pour cela, nous avons tout d'abord récupéré les données de la bouée d'Anglet sur le site de CANDHIS pour l'année 2020 et 2021 que nous avons ensuite joint aux données disponibles précédemment. Nous avons ensuite récupéré sur le site COPERNICUS le modèle « Atlantic -Iberian Biscay Irish- Ocean Wave Reanalysis » qui fournit un champ instantané sur le golfe de Gascogne les divers paramètres des vagues (ici, la période, la hauteur et la direction sont les paramètres qui nous intéressent) dans un fichier .nc dont nous avons réduit la zone sur la côte basque sur ces deux années.

Le script issu du papier est divisé en quatre parties :

- la première a pour objectif d'extraire les données du fichier .nc ;
- La deuxième rassemble les données issues du fichier COPERNICUS aux données de la bouée d'Anglet, dans un unique fichier csv « *data.assembled.csv* ». Nous avons créé ce fichier pour les deux nouvelles années, que nous avons ensuite joint avec le même fichier issu du projet précédent Vents&Marées qui contient les années manquantes de la période, pour ainsi obtenir un fichier final « *data_assembled_final.csv* » qui concerne toute la période de notre projet ;
- Une fois la nomenclature corrigée sur les scripts, nous avons lancé la phase d'entraînement (le troisième script) ;
- et pu voir les résultats du modèle sur toute la période pour enfin obtenir le jeu de données final à l'aide du quatrième script.

Pour le jeu de données sur le vent, une convention entre Météo France et l'Ifremer nous permet d'utiliser ces données.

Agrégation journalière et traitement des données vagues et météo

Les données de pêche sont journalières, il est donc nécessaire que les données environnementales aient la même échelle temporelle.

Le pas de temps de mesure des données de vague étant de dix minutes avant chaque heure, nous avons donc plusieurs valeurs pour chaque journée. Avant d'agréger ces valeurs, nous avons traité les données, en particulier les données de vent qui présentaient des zéros (Force nulle et direction nulle). Nous avons considéré que ces valeurs étaient des erreurs de mesures lorsque notamment elles étaient précédées et suivies de valeurs fortes.

Nous avons ensuite agrégé les valeurs de chaque jour en une ligne avec de nombreuses caractéristiques d'agrégation. Ces indicateurs agrégés journaliers sont présentés dans la figure suivante.

Figure 6 : Présentation des indicateurs journaliers agrégés

Une fois ces indicateurs journaliers créés, nous nous sommes intéressés à l'effet mémoire. Les pêcheurs, lorsqu'ils prennent leur décision de sortir ou non ne vont pas forcément regarder la météo du jour uniquement mais la météo dans sa continuité et considérer les jours précédents, par exemple comment le vent a soufflé les trois derniers jours. Nous avons alors ajouté des variables décalées dans le temps (de 1, 2 ou 3 jours précédents) pour chaque indicateur journalier agrégé : la variable prend donc la valeur de son indicateur journalier d'il y a X jours (avec X = 1, 2 ou 3).

Nous avons aussi pu considérer que lorsqu'un pêcheur met son engin à l'eau, il doit pouvoir le récupérer le lendemain (en particulier les fileyeurs). Ils peuvent donc potentiellement projeter leurs prévisions sur plusieurs jours, et nous avons donc créé de manière analogue des variables qui prennent la valeur de son indicateur journalier 1, 2 ou 3 jours après.

Chaque indicateur journalier agrégé a donc 7 informations : celle pour le jour même, 6 pour les 3 jours précédents et les 3 jours suivants.

I-B-3 Création du jeu de données final

La création du jeu de données final se fait par simple adjonction. Nous rassemblons donc les 3 jeux de données environnementaux en un seul en faisant la jonction sur la date, puis nous faisons une seconde jonction une nouvelle fois sur la date pour que chaque sortie ou non sortie ait les variables environnementales correspondant à la date.

Nous avons finalement un jeu de données de 393 111 lignes pour 540 variables. Nous avons conservé un large panel de variables en ayant pour but d'avoir toutes les variables qui ont potentiellement un lien avec la sortie ou non d'un navire. Ces variables présentent des liens entre elles, en particulier celles traitant un même facteur environnemental, nous voulons extraire les variables qui ont le plus fort impact sur notre variable cible de ***Sortie*** et que nous qualifierons de variables clés. L'obtention de ces variables clés font l'office d'une seconde phase de datamining.

II – Sélection des variables cibles et création de seuils

À la suite de la phase de préprocessing, nous obtenons un jeu de données final correspondant mieux à notre objectif. La taille de ce jeu de données est importante puisqu'il contient plus de 500 variables.
Nous avons dans un premier temps chercher à identifier les variables les plus pertinentes pour expliquer la sortie afin d'aboutir à des modèles plus faciles à manipuler en temps de calcul, plus parcimonieux et plus interprétables.
Les phénomènes étant souvent décrits par des effets seuils, nous avons donc cherché à évaluer des valeurs seuils expliquant la non-sortie et donc regarder en suivant la méthode déjà mise en place dans le projet Vents&Marées, des arbres de décisions conditionnels.

A partir de ce moment, le choix a été fait, comme dans le projet précédent, de séparer les métiers qui ont des comportements différents mais aussi en fonction des ports, qui peuvent avoir des comportements différents pour des raisons de logistiques (accessibilité du port / criée), réglementaires ou socio-économiques. Nous allons donc construire des arbres de décisions pour les bolincheurs, les fileyeurs « CBA », les fileyeurs « XBA », les fileyeurs « ABA », les ligneurs « ABA » et les ligneurs « CBA ». Les pêcheurs opérant sur l'Adour (« GBA ») seront traités à part en tant que cas particulier.

II-A Présentation de la méthode

Un arbre de décision est un arbre binaire aidant à la décision d'une valeur précise d'une variable cible pour un individu tout en connaissant l'état de covariables.

Le choix a été fait d'utiliser la librairie Rattle de R pour effectuer ces arbres. Rattle possède une interface facilitant grandement la phase de datamining. Dans cette interface, deux arbres de décisions sont proposés : ceux dits conditionnels avec l'algorithme "ctree" et les traditionnels avec l'algorithme "rpart".

La méthode "rpart" permet d'obtenir un arbre de régression de type CART. Cet algorithme effectue une recherche exhaustive de toutes les coupures possibles en maximisant une information mesurant l'impureté des nœuds en sélectionnant la covariable qui effectue la meilleure coupure. Ce type d'algorithme est sensible aux problèmes de surajustement et possède un biais de sélection envers les covariables ayant plusieurs coupures possibles (Hothorn et al., 2015).

Nous avons choisi de mettre en œuvre des arbres de décisions conditionnels qui sont implémentés avec l'algorithme "ctree" (Hothorn et al., 2015). Cet algorithme implémente le partitionnement binaire récursif introduit par Strasser and Weber (1999). Il se base sur la distribution conditionnelle de statistiques qui mesurent l'association entre les réponses et les covariables. Ceci permet une sélection impartiale parmi les covariables mesurées à différentes échelles. De plus, plusieurs procédures de test sont appliquées pour déterminer s'il est possible d'affirmer qu'aucune association significative entre l'une des covariables et la réponse n'existe, ce qui indique quand la récursion doit s'arrêter.

Les avantages de ces arbres de décision sont multiples :
- Ils sont simples à comprendre et à interpréter ;
- Ils permettent d'obtenir des valeurs en lien avec un scénario ;
- Ils demandent peu de préparation des données, s'accommodant notamment des valeurs manquantes et des données continues ou discrètes ;
- Ils sont peu sensibles aux valeurs extrêmes ou aux relations non linéaires ;
- Ils peuvent être combinées avec d'autres méthodes.

II-B Résultats

II-B-1 Présentation des résultats

Nous allons ici décrire les résultats pour un arbre. Nous avons choisi les ligneurs de Capbreton (*ABA*) pour présenter le procédé de la lecture d'un arbre. Les arbres des différentes typologies sont disponibles en annexe. Deux arbres ont été construits par typologie. Le premier arbre contient les caractéristiques de saisonnalité (année, mois, espèces). Le second arbre ne possède pas ces caractéristiques. Pour la création de ces arbres, nous avons limité à une profondeur de cinq pour conserver une certaine lisibilité, à l'exception des fileyeurs de Capbreton où une profondeur de quatre est choisie, toujours pour une question de lisibilité. Les autres paramètres possibles dans rattle sont la Division min (minsplit) et Compartiment min (minbucket) que nous avons laissé à la valeur défaut, respectivement 20 et 7. La

variable cible est "***Sortie***".

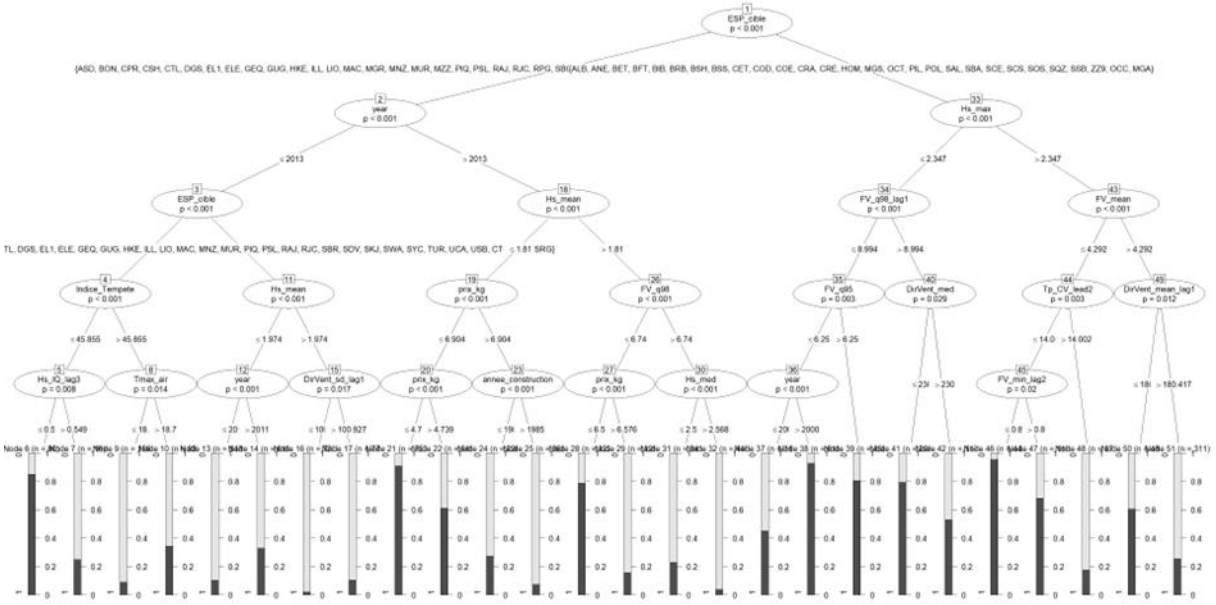


Figure 7 : Arbre de décision pour les ligneurs de Capbreton prenant en compte les variables de saisonnalités

On peut voir que la variable la plus discriminante est l'espèce cible, puis à gauche l'année et à droite, la première variable non temporelle, la hauteur moyenne des vagues. Dans notre projet, nous ne sommes pas particulièrement intéressés par l'effet saison, mais plutôt par l'importance des facteurs environnementaux ou sociaux-économiques dans le choix de sortir ou non. Ces variables de saisonnalité étant présentes dans tous les arbres si nous les conservons, nous avons choisi pour la suite de les exclure dans nos analyses.

Intéressons-nous donc à l'analyse de l'arbre de décision conditionnelle des ligneurs de Capbreton (**ABA**) :

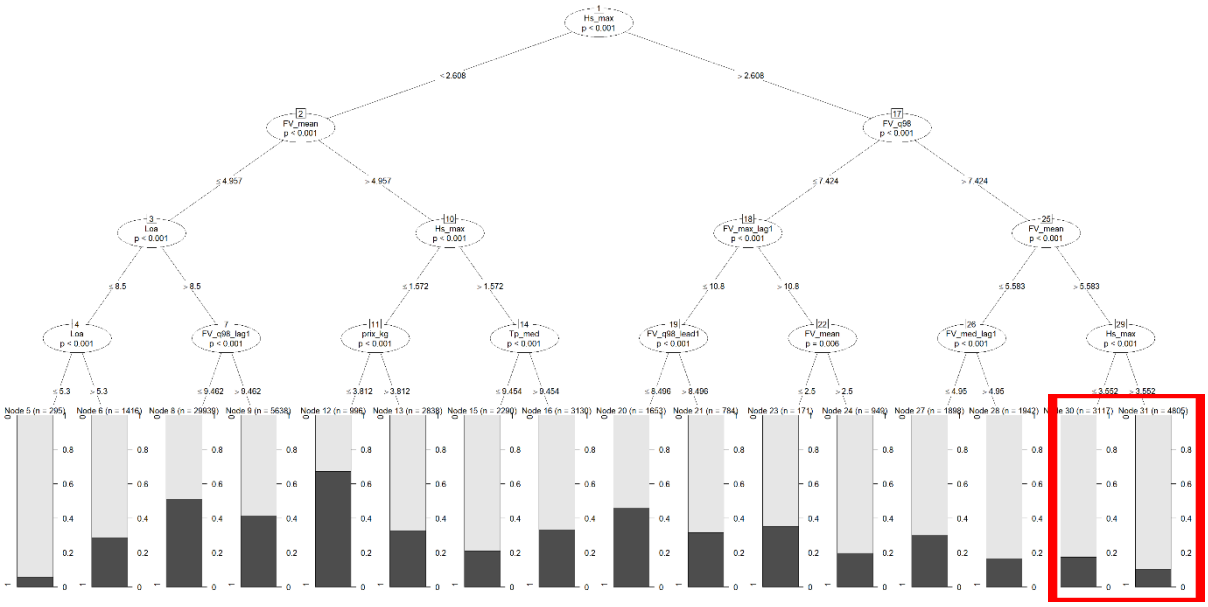


Figure 8 : Arbre de décision pour les ligneurs de Capbreton ne prenant pas en compte les variables de saisonnalité

Pour cette typologie, la variable la plus discriminante est la hauteur maximale journalière des vagues.

Voici comment nous pouvons analyser cet arbre :

- A gauche, si la hauteur maximale des vagues est modeste (<2,6 m), alors l'arbre classifie les données en fonction de la force moyenne journalière des vents, puis de la taille du navire ou encore une fois de la hauteur maximale des vagues avec un autre seuil (1,7 m) ;
- A droite, si la hauteur maximale des vagues est importante (>2,6 m), alors l'arbre classifie les données en fonction du quartile à 98% de la force du vent journalière.

Nous pouvons aussi voir des issues où les proportions de zéros sont exceptionnelles. Par exemple, la zone bleue met en exergue une probabilité de sortie élevée (autour de 70%) quand la hauteur des vagues maximale est inférieure ou égale à 1.6 mètres, que la force moyenne journalière du vent est inférieure à 5 et que le prix au kilogramme de l'espèce cible est inférieure à 3 euros et 82 centimes. A l'inverse, la zone rouge montre une issue où les probabilités de sorties sont inférieures à 20% quand la hauteur maximale des vagues est inférieure à 2,6 mètres, que le quartile à 98% est supérieur à 7,4 et que la force moyenne journalière du vent est supérieure à 5,6. L'erreur de classement moyenne en lien avec le modèle est de 31%.

Nous avons pu voir précédemment comment lire un arbre de décision et il est possible de reproduire des analyses similaires avec les autres arbres disponibles en annexe. Seuls les arbres excluant la saisonnalité sont présents dans l'annexe, étant les arbres utilisés par la suite.
Pour chaque arbre, un certain nombre de variables ressortent parmi nos plus de 540 variables de départ. Nous avons résumé l'occurrence de ces variables dans les arbres en fonction du métier et du port de départ. Nous ne prenons pas en compte la position de la variable sur l'arbre, et nous comptons chaque occurrence même si elle est multiple sur un même arbre.

La variable qui sort le plus souvent est ***prix_kg*** avec 28 occurrences. Nous avons aussi de notable ***Power_main*** (18) en deuxième et ***Hs_max***, ***Fv_mean*** et ***annee_construction*** en troisième (14). Les deux variables environnementales qui semblent être les plus déterminantes pour notre variable cible ***Sortie*** sont donc la hauteur maximale des vagues et la force moyenne des vents.

Le tableau suivant résume le pourcentage de variables apparaissant dans les arbres en fonction du nombre total de variables qui sont sorties dans l'arbre (colonne total) en lien avec la catégorie concernée. Certaines variables ont été regroupées, notamment celles d'un même phénomène.

Total	**FV**	**DirVent**	**Hs**	**DirWave**	**Tp**	**Caractéristiques**	**Autres**	**Total**
Fileyeurs	0,24	0,11	0,10	0,02	0,13	0,25	0,16	63
Ligneurs	0,25	0,06	0,17	0,04	0,02	0,38	0,08	48
Bolincheurs	0,47	0,26	0,05	0,05	0,05	0,05	0,05	19
CBA	0,27	0,10	0,11	0,03	0,07	0,29	0,13	70
ABA	0,34	0,06	0,17	0,03	0,06	0,34	0	35
XBA	0,20	0,24	0,04	0,04	0,12	0,12	0,24	25
Total	36	15	15	4	10	35	15	130

Tableau 3 : Fréquences des apparitions des variables dans les différents arbres. La colonne « Total » représente le nombre total de variables apparues dans les arbres, les répétitions étant incluses.

En regroupant par catégorie, avec “Caractéristiques” représentant les variables décrivant les caractéristiques du bateau et “Autres” les indices ou la température, on peut voir que les variables apparaissant le plus sont largement la force du vent et les caractéristiques des navires. La direction des vagues est une variable presque pas représentée, et ne semble pas très importante. En revanche, bien que la hauteur des vagues semble peu représentée, elle est au final très importante, avec la hauteur maximale des vagues étant la première variable sortant dans les arbres pour 4 sur les 6 arbres ! Cette variable ***Hs_max*** semble donc être une variable clé, avec la force moyenne journalière des vents sortant en premier sur 1 arbre et apparaissant sur la plupart des arbres.

Si on s’intéresse par catégorie, on peut voir que les bolincheurs sont très sensibles à la force du vent, représentant la moitié des variables avec la moyenne étant la première variable sortie. Si on additionne l’autre variable traitant des vents, la direction, on monte à plus de 70% des variables. Les autres variables apparaissent de manière équivalente.

Pour les fileyeurs et les ligneurs, on semble avoir des tendances similaires avec les variables apparaissant le plus souvent dans les arbres étant les caractéristiques du navire et la force du vent. Il est tout de même important que hormis les fileyeurs de Boucau (**XBA**) qui ont la puissance du navire (***Power_main***) qui sort en premier dans l’arbre, pour toutes les autres typologies, c’est la hauteur maximale des vagues qui sort en premier.

Si nous nous focalisons sur les ports de départ, on peut noter que l’arbre des pêcheurs de Capbreton n’a pas sorti de règles de classification impliquant un indice ou la température, à l’inverse des pêcheurs de Ciboure / St-Jean-de-Luz (13%) ou en particulier Boucau (24%). L’arbre des pêcheurs de Boucau ont aussi presque 50% des variables ressortant dans l’arbre qui sont des variables en lien avec le vent.

II-B-2 Objectif de la méthode

Cette méthode avait pour objectif principal de sélectionner un pool de variables et les seuils auxquels elles ont un impact sur la décision de sortie ou non du port pour la suite de notre étude. A partir de notre jeu de données (cf. Partie I) avec plus de 540 variables, nous le subdivisons en plusieurs jeux de données. Chaque jeu de données représente un métier en lien avec un port de départ favori. Pour chaque sous-ensemble, nous ne conservons pas les 540 variables mais nous réduisons le nombre de variables. La sélection des variables correspond à celles qui ressortent dans les arbres correspondant à la typologie en excluant la saisonnalité.

Nous avons aussi la volonté de discrétiser les variables pour la suite du projet. La discrétisation correspond à la transformation de nos variables quantitatives continues à des variables quantitatives discrètes ou qualitatives. Nous avons pris les seuils issus des règles de décision de l’arbre pour transformer nos variables. Si une variable apparaît plusieurs fois dans un arbre, chaque seuil est considéré : si la variable apparaît N fois avec N seuils, la variable aura alors dans le jeu de donnée correspondant à sa typologie N+1 valeurs.

II-B-3 Limites

Plusieurs limites existent avec les arbres de décisions. En particulier, ceux-ci sont sensibles aux variables discrètes ayant de nombreuses modalités. Ces nombreuses modalités peuvent créer un biais de l’arbre de décision envers cette variable.
Ce biais peut particulièrement être présent pour les arbres prenant en compte la saisonnalité, où des variables comme “***year***”, “***month***” ou “***ESP_cible***” sont présents et ont pu apparaître dans les résultats alors qu’il y a de nombreuses modalités.

Une seconde limite est que les résultats peuvent être de pauvres prédicteurs lorsque l’arbre est surajusté avec les données. Comme présenté lors de la section précédente, l’objectif dans ce travail quand on utilise cette méthode n’est pas dans un but prédictif mais sélectif pour obtenir des variables clés. Nous sommes donc moins sensibles à cette limite.

III – Création des réseaux Bayésiens

III-A Présentation de la méthode

Les réseaux bayésiens font partie de la classe des modèles graphiques représentant les relations de dépendance conditionnelle entre les variables d'un système. Ils prennent la forme d'un Graphique Acyclique Orienté que l'on peut qualifier par l'acronyme **DAG** (**D**irected **A**cyclic **G**raph). Ce graphique comporte un ensemble de variables aléatoires $\mathbf{X = \{X_1, X_2, \ldots, X_n\}}$ avec $\mathbf{n \in \mathbb{N}}$, chacune représentée sous la forme d'une boîte appelée nœud, noté $\mathbf{n_i \in N}$. Ces nœuds peuvent être reliés entre eux par un arc dirigé $\mathbf{a_i \in A}$ qui représentent une dépendance probabiliste directe. Le réseau bayésien peut donc être représenté sous la forme d'un graphique avec un ensemble de nœuds **N** reliés entre eux par les arcs orientés de l'ensemble **A** sans que ceux-ci ne forment un cycle et donc on utilise la notation **G = (N, A)** pour désigner un réseau ou graphe :

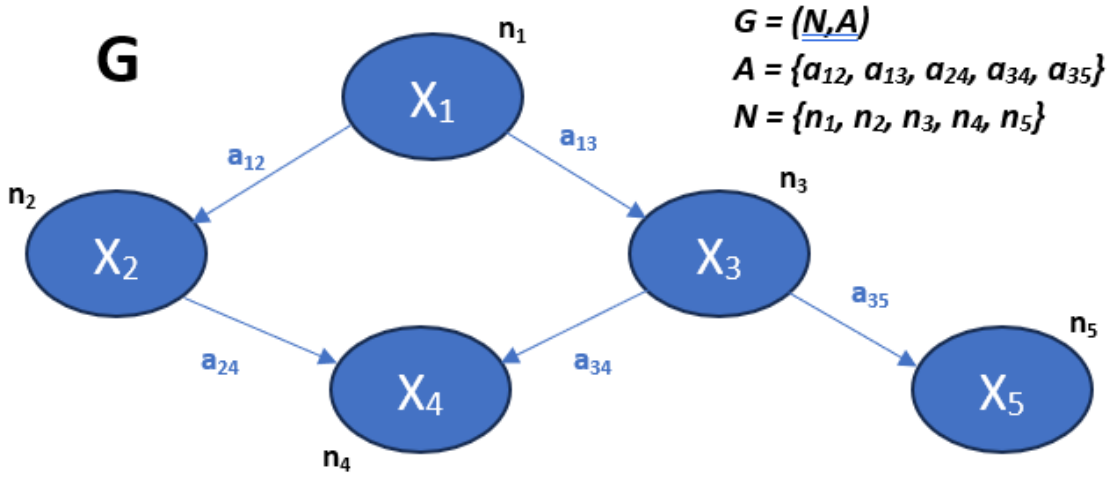


Figure 9 : Exemple d'un DAG d'un réseau bayésien G

La factorisation de la probabilité globale de notre réseau bayésien prend donc la forme suivante :

$$P(X) = \prod_{i=1}^{n} P(X_i | \Pi_{X_i})$$

Nous pouvons définir comme les parents d'une variable aléatoire X_i du nœud n_i toute variable aléatoire X_j d'un nœud n_j ayant un arc dirigé n_j orienté vers le nœud n_i. On a alors la probabilité marginale suivante :

$$P(X_i) = \prod_{j=1}^{n} P(X_i \mid X_j)$$

Nous pouvons ainsi définir trois types de réseau bayésien :

- Les réseaux bayésiens multinomiaux qui ne présentent que des variables aléatoires X discrètes avec des distributions multinomiales numériquement spécifiées.
- Les réseaux bayésiens gaussiens présentent des variables aléatoires X continues suivant une distribution gaussienne dont seul l'espérance dépend des parents de manière affine.
- Les réseaux bayésiens hybrides, mélangeant variables aléatoires discrètes et continues et permettant d'introduire n'importe quelle distribution de probabilités. Malheureusement, cette liberté impose des réseaux bayésiens bien plus complexes et difficiles à manipuler, ce qui implique une littérature et des outils très réduits pour manipuler ces objets à ce jour.

Notre variable cible étant la variable ***Sortie*** qui ne présente que des 1 ou des 0, utiliser un réseau bayésien nous avons préféré nous concentrer principalement sur les réseaux bayésiens les plus répandus dans la littérature : les réseaux bayésiens multinomiaux.

L'avantage des réseaux bayésiens est la possibilité de visualiser à travers les arcs les relations de dépendance probabilistes qui relient les variables entre elles, et ainsi de visualiser les relations de dépendances entre les variables.

Pour créer notre réseau bayésien, nous devons tout d'abord découvrir son DAG (sa structure), ce que l'on appelle l'apprentissage de la structure. Il est ensuite nécessaire d'apprendre les paramètres de notre réseau bayésien, ce qui consiste à résoudre les tables de probabilités conditionnelles de chaque nœud de notre DAG, notamment à travers la formule de Bayes qui donne son nom à la méthode. Enfin, une fois notre réseau bayésien créé, nous pourrons explorer les différentes voies d'utilisation et d'application dans la réalité.

III-B Apprentissage du réseau bayésien

III-B-1 Présentation

Pour créer notre réseau bayésien, la première phase est la phase d'apprentissage. Cette phase se divise en 2 parties :

- L'apprentissage de la structure correspond à la découverte de notre graphe : les nœuds et les arcs dirigés qui les relient entre eux
- La seconde partie est l'apprentissage des paramètres qui pour le DAG découvert à l'étape précédente, estime les tables de probabilités conditionnelles de chaque nœud.

Pour chaque partie, il y a deux voies. La première option est ce que l'on nomme le « unsupervised learning », où l'apprentissage est effectué à l'aide des données disponibles tandis que la seconde option, « supervised learning », se base sur les connaissances à dires d'expert pour décider la structure ou les paramètres. Il est possible de combiner ces deux options pour chaque étape.

III-B-2 Les algorithmes d'apprentissage de la structure

Nous avons de nombreuses données disponibles, et nous avons donc décidé de commencer par obtenir une première structure à partir de celle-ci. Mais il existe une multitude d'algorithme pour apprendre cette structure que nous pouvons regrouper en trois catégories : les algorithmes par contraintes, par maximisation de score ou hybride.

Algorithmes par contrainte

Ces algorithmes sont basés sur les travaux fondateurs de Pearl (Verma and Pearl, 1991) qui a fourni un cadre pour l'apprentissage de la structure des réseaux bayésiens avec l'algorithme IC qui utilise des tests d'indépendance conditionnelle. Sa première implantation fut dans l'algorithme PC (Peter-Clark), aujourd'hui toujours utilisé dans une version améliorée PC-stable.

Algorithmes par maximisation de score

Ce sont des algorithmes qui appliquent l'idée d'optimisation à l'apprentissage de la structure d'un réseau bayésien. Ainsi, l'algorithme produit de nombreux réseaux bayésiens candidats auxquels il assigne un score de réseau traduisant sa qualité en lien avec les données. Son objectif étant d'optimiser ce score, il renvoie le réseau bayésien ayant le meilleur score de réseau. Plusieurs critères de score sont disponibles, tels que le score AIC ou BIC. L'objectif étant de sélectionner les arcs pertinents qui seront présents dans

notre réseau bayésien (Confirmation / Falsification), nous avons choisi le score BIC (Aho et al., 2014).

Algorithmes hybrides

Ces algorithmes mélangent les deux précédents dans l'objectif de compenser les faiblesses respectives. Il y a deux étapes principales : une première étape de restriction qui est une phase de contrainte dans laquelle. La seconde phase de maximisation consiste à utiliser un algorithme de maximisation du score.

III-B-3 Choix de l'algorithme et de l'outil de création de réseaux bayésiens

Face à cette multitude de choix, il est difficile de sélectionner l'algorithme adéquat. Une première manière de discriminer est de sélectionner des outils numériques pour créer des réseaux bayésiens, ce qui réduira les algorithmes à ceux disponibles.

Nous avons considéré 3 choix d'outils :

- GENIE, un logiciel propriétaire de BayesFusion qui présente une interface permettant une utilisation simplifiée pour créer, apprendre et analyser des réseaux Bayésiens. Malheureusement, étant propriétaire, les sources ne sont pas disponibles.
- Bnlearn (Scutari, 2010) qui est un package de R complet.
- Bnstruct (Franzin et al., 2017) qui est aussi un package de R permettant d'apprendre la structure et les paramètres d'un réseau bayésien dans différentes situations, et en particulier dans les cas où il y a des données manquantes.

Outil	**GENIE**	**BNlearn**	**BNstruct**
Algorithmes disponibles	Bayesian Search, PC, Greedy Thick Thining, (Tree) Augmented Naive Bayes et Naive Bayes	PC, Grow Shrink, Incremental Association, Fast / Interleaved Incremental Association Hill-climbing, Tabu Search, Max-Min Hill-Climbing, Restricted Maximization, Hybrid HPC, Max-Min / Hiton parents and children.	Silander-Myllymaki (sm) Max-Min Parent-and-Children Hill Climbing Max-Min Hill-Climbing heuristic Structural Expectation-Maximization (sem)
Prise en compte des NA	Non	Non sauf exception	Oui
Avantage	Facile d'utilisation	Complet	Spécialisé (prise en compte NA)
Présence dans la Bibliographie explorée	Très faible	Importante	Faible

Tableau 4 : Comparaison des différents outils considérés. Sont surlignés les algorithmes que nous voulons utiliser.

Notre objectif principal était d'apprendre entièrement notre réseau bayésien à partir de nos données. Les réseaux bayésiens sont communément construits manuellement à partir de la connaissance des experts, mais dans le papier (Ramazi et al., 2021), les auteurs proposent une approche d'apprentissage de la structure et les paramètres entièrement à partir des données. Étant donné qu'ils concluent que cette démarche permet des prédictions plus précises, nous allons l'utiliser.

Pour l'apprentissage de la structure, ils utilisent l'algorithme de Silander & Myllymaki (2012) qui est implémenté dans le package bnstruct sous le nom « sm ». Ils notent que, dans les cas trop complexes (plus de 25 nœuds), la structure *a priori* était apprise avec bnlearn.

L'auteur de bnlearn a aussi produit des travaux pour déterminer quel algorithme modélise le mieux la structure des réseaux en termes de précision et de vitesse (Scutari et al., 2019). La vitesse n'est pas un aspect qui nous intéresse. Avec les premiers tests, nous avons observé que nos réseaux se construisent avec une vitesse de l'ordre de la minute, plus ou moins rapide selon l'algorithme. La différence n'étant pas un problème en termes de temps, le choix de l'algorithme ne se fera pas selon ce critère. En revanche, en termes de précision, ces travaux soulignent que les algorithmes de contraintes sont globalement moins précis que l'algorithme "tabu search" mais pas que les algorithmes "recuits simulés". En revanche, ils sont plus précis dans la plupart des situations que les autres algorithmes de score et il n'y a pas de différences significatives avec les algorithmes hybrides. L'algorithme "tabu search" (Russell and Norvig, 2009) serait ainsi globalement le plus précis. Ce classement ne varierait pas en fonction de la taille du réseau.
Pour compléter, il est précisé que pour les données complexes, seuls les algorithmes de score produisent de larges réseaux dans lesquels des dépendances d'ordres supérieurs (dépendances spatiales de longue distance) sont profondément représentées, ce qui est clé dans les données climatiques.

Bien que GENIE soit simple d'utilisation avec un logiciel qui permet de créer son réseau bayésien en « click-button », plusieurs aspects ont fait que nous l'avons écarté :

- C'est un logiciel propriétaire aux sources non disponibles
- Son utilisation est peu représentée dans la littérature explorée
- Les algorithmes d'apprentissage utilisables ne sont pas ceux qui nous intéressent

Nous avons considéré 2 options : l'utilisation de l'algorithme sm avec bnstruct ou celui du tabu search avec bnlearn. Pour effectuer ce choix final, nous nous sommes posé 2 questions :

- Avons-nous besoin de considérer nos données manquantes ? Si c'est le cas, cela nous dirigerait vers l'algorithme sm qui les prend en compte. Mais quand on s'intéresse aux données manquantes, le travail préliminaire présenté en partie I fait qu'il n'en reste plus beaucoup, de l'ordre de 1%. Il existe une exception : les typologies où la variable « Indice_Tempete » est sortie. Cette variable n'a été mesurée qu'à partir de 2009, ce qui induit de nombreux NA. Nous avions alors 4 options : les conserver en tant que NA, les conserver en tant que valeur discrète « non mesuré », les supprimer ou ne pas considérer la variable dans la création de notre réseau bayésien. Pour les 2 premières options, nous craignions d'introduire un biais, tandis que pour la troisième, cela engendrerait une forte perte d'information. Nous avons donc opté pour la dernière option. On peut ainsi présenter l'exemple des ligneurs de Ciboure / St-Jean-de-Luz qui a la variable Indice_Tempete qui est sorti dans l'arbre de décision conditionnelle. En prenant en compte cette variable, on a 17 066 NA présents dans nos données sur 52 398, ce qui descend à 336 zéros si on ne la considère pas, soit moins de 1%.

- Quelle est la facilité de produire le réseau bayésien et de poursuivre les différentes étapes ? Sur ce point, bnstruct est plus complexe d'utilisation, avec des classes propres. Pour ce qui est d'inférence, d'utilisation des réseaux bayésiens produits, bnstruct n'a pas de continuité et propose des fonctions qui permet de changer les classes des objets pour permettre de basculer sur d'autres packages, et en particulier bnlearn. Il semble donc plus simple de tout faire sur bnlearn, qui est plus complet.

Le choix final est donc de produire nos réseaux bayésiens à l'aide de l'algorithme tabu search disponible dans le package bnlearn.

III-C Création des réseaux Bayésiens

Nous commençons par créer un premier réseau bayésien à partir des données. Pour cela, nous avons choisi arbitrairement de commencer par les ligneurs de Capbreton (ABA), et les prochains exemples présentant notre démarche ont pour support ces données. Il n'y a donc ici aucune connaissance experte, le réseau bayésien est entièrement appris à partir des données.

Sur le graphique, l'épaisseur des arcs représentent leur score BIC : Si le retrait de l'arc entraîne une forte baisse du score, l'arc est épais (et inversement). Nous avons aussi mis en rouge la variable cible, ***Sortie*** par souci de lisibilité.

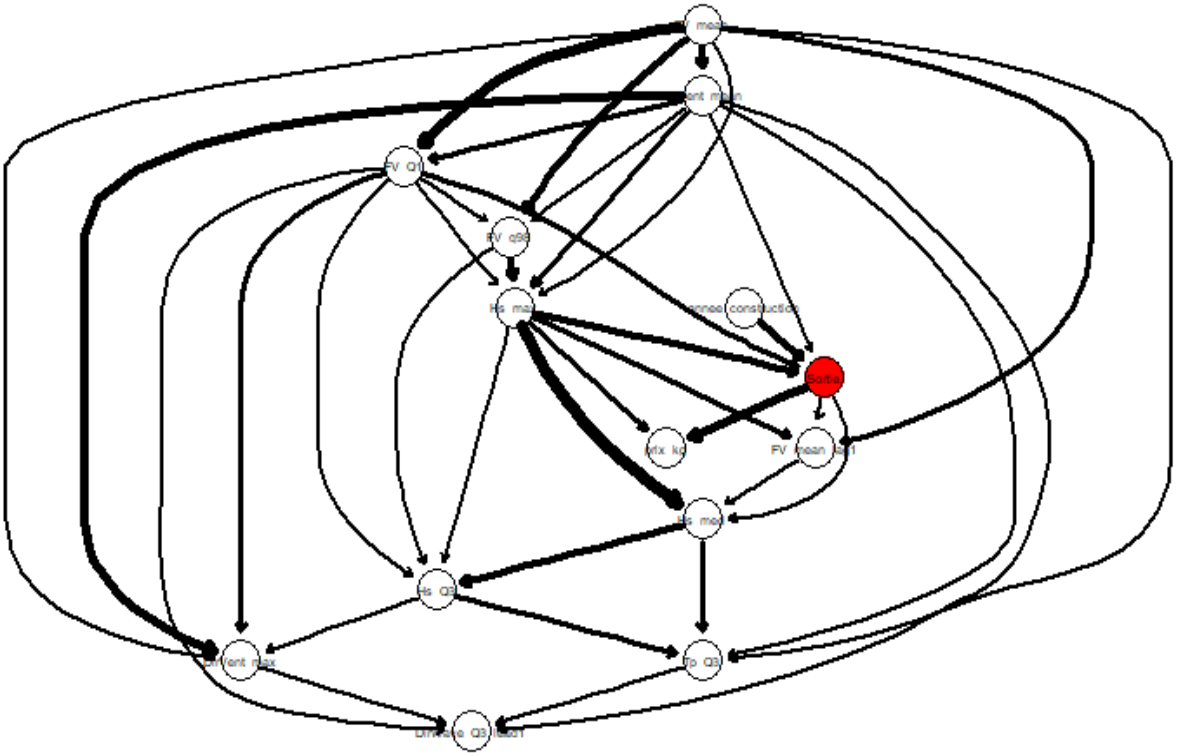


Figure 10 : Réseau bayésien des ligneurs de Capbreton, algorithme tabu

Le DAG du réseau bayésien est plutôt chargé et difficilement lisible. Nous pouvons le simplifier. En effet, nous pouvons remarquer de nombreux liens que l'algorithme a jugé pertinents car augmentant le score du réseau bayésien, mais si on s'intéresse à son sens logique dans la réalité, le lien est impossible :

- Tout d'abord, hormis la variable prix_kg, la variable cible ***Sortie*** ne doit pas avoir d'arcs dirigés vers une autre variable. De manière générale, les variables non environnementales ne peuvent pas avoir un arc dirigé vers les variables environnementales ;
- Les variables présentent une temporalité. Nous avons des variables représentant des phénomènes qui ont eu lieu les jours précédents (lag) et d'autres qui représentent des phénomènes qui n'ont pas encore eu lieu (lead). On peut donc considérer qu'une variable X_1 ne peut pas influencer une variable X_2 si elle a lieu avant. Dans l'exemple ci-dessus, on peut donc considérer que la variable **FV_mean_lag1** ne peut pas être influencée par les autres variables dans ce système (aucune autre variable lag) alors qu'à l'inverse, la seule variable lead **DirWave_Q3_lead1** ne peut pas influencer sur une variable autre que ***Sortie***. Nous pouvons en effet considérer que leur possible état évalué par les pêcheurs (avec la météo) influe sur leur choix de sortir ou non ;
- Les variables traitant des vagues ne peuvent pas influencer le vent, c'est plutôt l'inverse.

Nous avons donc créé une "blacklist" interdisant tous ces différents arcs. On obtient alors un nouveau réseau bayésien.

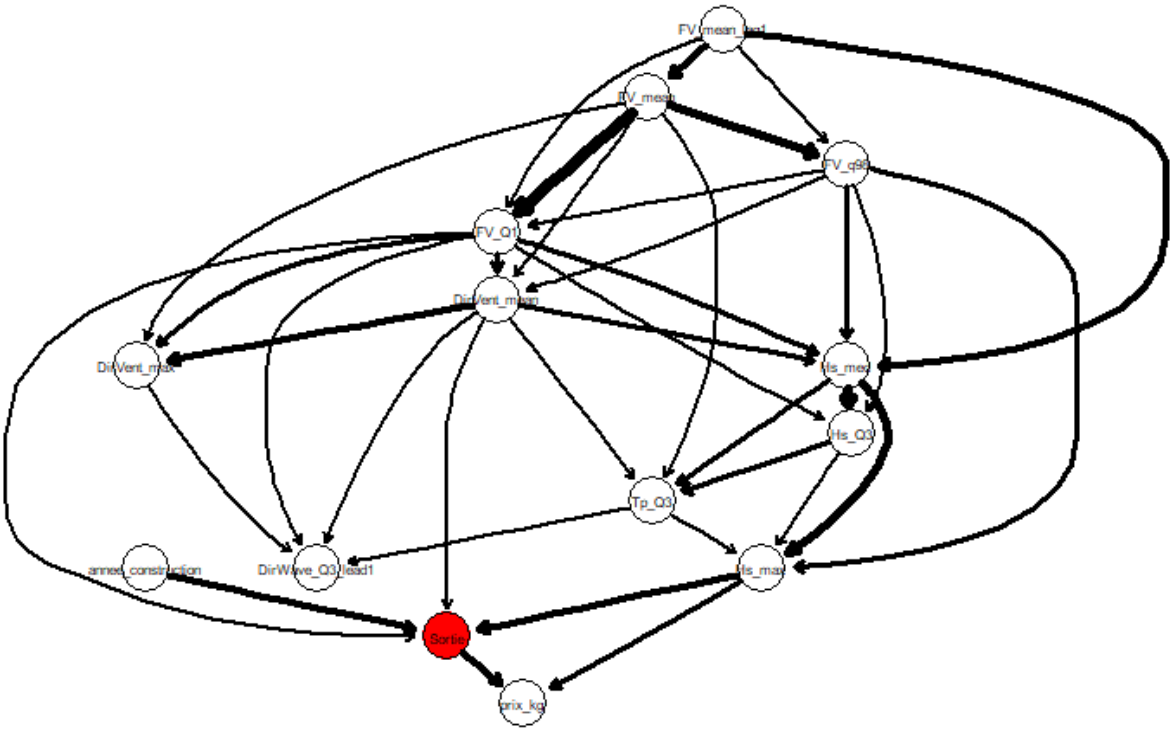


Figure 11 : Réseau bayésien des ligneurs de Capbreton, algorithme tabu avec une blacklist

Nous voulons ensuite simplifier ces DAG. On peut ainsi voir que certaines variables d'un même phénomène sont présentes plusieurs fois. Par exemple, il y a sur le DAG ci-dessus, 4 variables sur la force du vent (**FV**) ou encore 3 sur la hauteur des vagues (**Hs**). Nous avons alors choisi de ne conserver qu'une seule variable par phénomène environnemental. Pour faire notre choix, nous avons choisi la première variable qui est sortie dans les arbres de décision conditionnels, en conservant plusieurs seuils si la même variable est sortie plusieurs fois.

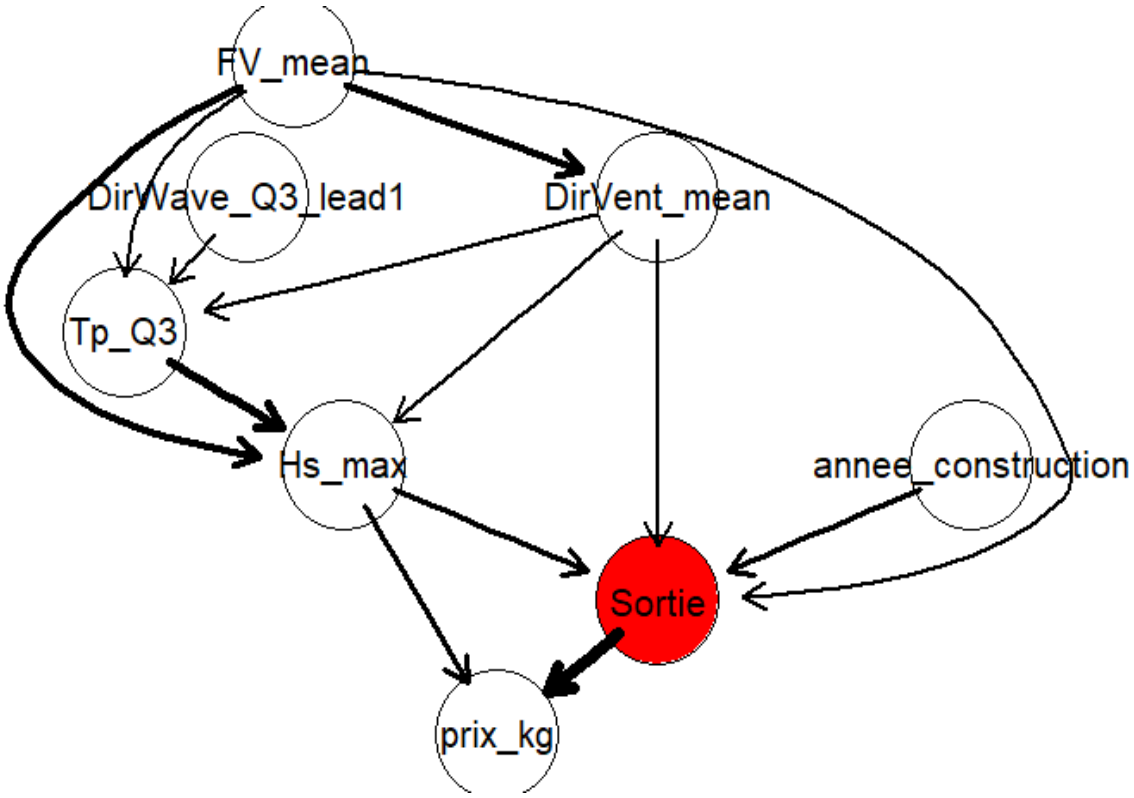


Figure 12 : Réseau bayésien des ligneurs de Capbreton en version simplifiée, algorithme tabu avec une blacklist

Une fois la structure de nos réseaux créée, nous avons ensuite appris leurs paramètres, c'est-à-dire obtenu les tables de probabilités conditionnelles de chaque nœud (variable) à partir des données correspondantes : encore une fois, nous avons choisi l'option sans l'introduction de connaissances expertes.

Comparons les différents réseaux obtenus précédemment pour voir s'il existe une réelle différence. Nous allons pour cela faire de l'inférence exacte, programmée dans le package gRain (gRaphical model inference). Le réseau bayésien est alors transformé en un arbre de jonction, ce qui permet d'accélérer les calculs de probabilités conditionnelles. Un arbre de jonction est une structure graphique utilisée dans le contexte des réseaux bayésiens pour représenter et calculer des probabilités conjointes, conditionnelles et marginales. En effet, l'arbre de jonction est créé en identifiant des clusters (groupes) regroupant des variables fortement liées entre elles. Ces clusters facilitent la propagation des probabilités et permettent de calculer rapidement et efficacement des probabilités conditionnelles.

Une fois l'arbre de jonction créé pour nos différents réseaux bayésiens, nous pouvons alors comparer les différentes probabilités marginales obtenues.

Variables	**BN Complet**	**BN Blacklist**	**BN simplifié**
P(Sortie = 0)	0.603*	0.603*	0.602
P(FV_mean < 4.312)	0.59	0.59	0.59
P(Hs_max < 2.189)	0.650*	0.650*	0.651
P(DirVent_mean < 222.857)	0.551	0.550	0.551
P(Tp_Q3 > 12.9)	0.186	0.184	0.182
P(DirWave_Q3_lead1 > 302)	0.704*	0.704	0.704

Tableau 5 : Comparaison des paramètres appris à partir des données des différents réseaux bayésiens
** : valeurs différentes si plus de décimales sont affichées*

Nous pouvons voir que les différences sont minimes, avec parfois une valeur égale. Cela semble logique puisque les paramètres sont appris à partir d'un grand jeu de données. Nous préférons par la suite utiliser le réseau bayésien simplifié.

III-D Utilisation de notre réseau bayésien

Nos réseaux bayésiens créés et prêts à l'emploi, 3 options d'utilisation ont été envisagées :

- Utiliser le réseau bayésien pour évaluer lors d'une journée X avec des conditions connues la probabilité qu'un navire sorte ;
- Avec des tendances connues de nos variables environnementales en lien avec des scénarios climatiques, on observe alors l'évolution de nos probabilités de sorties ;
- A partir de notre réseau bayésien, nous pouvons ajouter une ou plusieurs variables aléatoires et des arcs évoquées par les pêcheurs ou la littérature mais pour laquelle nous n'avions pas de données et donc auxquels nous assignons des probabilités manuellement pour étudier l'impact sur la probabilité de sortie : à partir de notre réseau bayésien créé à partir des données, nous créons un réseau bayésien dérivé incluant des dires d'experts.

Pour ces exemples, nous nous sommes concentrés sur 4 typologies. Nous avons tout d'abord les ligneurs de Capbreton (ABA) précédemment utilisés ainsi que les ligneurs et les fileyeurs de Ciboure / St-Jean de Luz (CBA). Nous avons aussi analysé les bolincheurs car ils partent tous de Ciboure / St-Jean de Luz, ce qui nous permet de comparer les trois métiers pour un même port ainsi que la différence au sein d'un

même métier. Pour la création de ces réseaux bayésiens, nous avons préféré choisir la version simplifiée (la première variable qui sort dans les arbres pour les variables environnementales). Pour le cas des fileyeurs CBA, nous n'avons pas pris les variables de la période car 4 variables sont sorties à profondeur 5, ce qui aurait grandement complexifié le réseau tandis que le prix au kilogramme et la longueur ont été conservés alors que la règle de décision apparaît à la même profondeur. De manière similaire, 2 variables sur la force du vent sont présentes à la profondeur 2, nous avons ici conservé les deux.

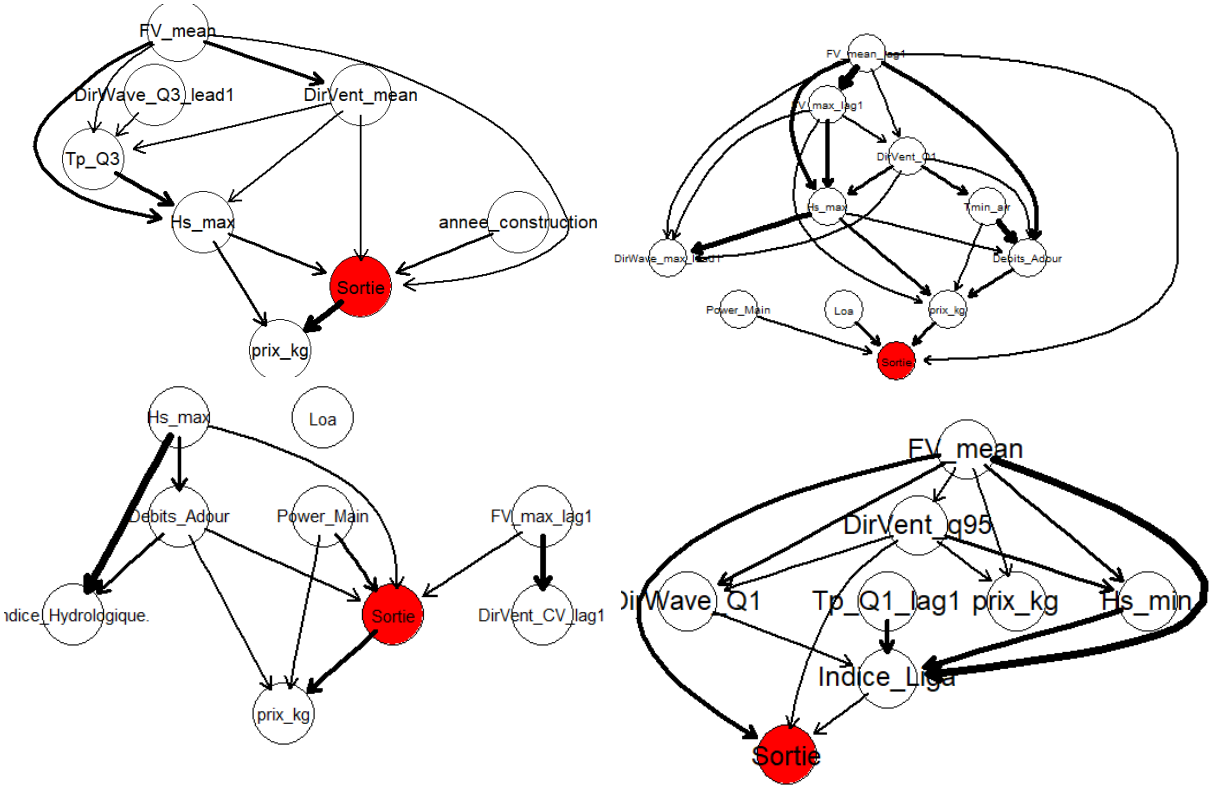


Figure 13 : Diagrammes des réseaux bayésiens pour différentes combinaisons de typologies et ports
En haut : Ligneurs ABA à gauche, Ligneurs CBA à droite
En bas : Fileyeurs CBA à gauche, Bolincheurs à droite

III-D-1 Comparaison des réseaux bayésiens

Tout d'abord, nous pouvons noter qu'après sélection des variables à travers l'analyse des arbres, il n'y a ici qu'un seul cas où une variable sélectionnée n'a pas d'arcs : la longueur du navire pour les fileyeurs **CBA**. Deux raisons peuvent l'expliquer : la blacklist interdit la plupart des liens, seules un arc partant de ***Loa*** vers ***Sortie*** peut exister, et dans l'arbre de décision cette variable n'a qu'une seule règle de décision à la profondeur 5.

En effet, la position dans les arbres de décisions de variables semble avoir un lien avec la force des arcs. On peut remarquer que les arcs ayant la force la plus importante (le plus épais) sont souvent liées à au moins un nœud correspondant à une variable sortie en première position dans les arbres (***Hs_max*** pour les ligneurs et les fileyeurs et ***FV_mean*** pour les bolincheurs).

Certaines variables peuvent être reliées différemment selon l'arbre, en particulier ***prix_kg***. Pour les ligneurs **ABA** et les fileyeurs **CBA**, ***prix_kg*** a un arc dirigé vers la variable ***Sortie*** alors que c'est l'inverse pour les ligneurs **CBA** et qu'ils ne sont pas reliés pour les bolincheurs. Le prix dépend peut dépendre de la quantité et de la demande (loi de l'offre et de la demande). Il aurait été intéressant de considérer le prix au kilogramme le jour précédent (***prix_kg_lag1***, variable non créé) qui peut probablement influencer le choix de sortir du pêcheur (appât potentiel du gain malgré des conditions défavorables).

III-D-2 Evaluation à la journée

Dans ce scénario, nous imaginons qu'un autre modèle ou une personne possède en sa possession un scénario d'une journée précise avec la taille des vagues, la force du vent… Il souhaite alors savoir dans son scénario quelle est la probabilité que le navire sorte. Ce résultat pourra potentiellement être intégré dans un modèle global ou pour appuyer des décisions.
Pour cela, on peut introduire une évidence **E**. Cette évidence contient des valeurs fixées pour différentes variables aléatoires. Nous allons considérer les variables ***FV_mean*** et ***Hs_max***, respectivement la force moyenne du vent est la hauteur maximale des vagues qui sont les deux variables environnementales qui apparaissent le plus souvent dans les arbres.

Soit un jour **N**, nous avons l'information de l'état de ces variables, que nous rentrons dans l'évidence **E**. En propageant cette information dans l'arbre, nous pouvons observer l'évolution de la probabilité de sortie.

Pour les bolincheurs, seul ***FV_mean*** est présent sur le graphique, seul ***Hs_max*** est présent sur les ligneurs **CBA** et les fileyeurs **CBA** tandis que les deux variables sont présentes pour les ligneurs **ABA**.

Scénario	**Ligneurs ABA**	**Ligneurs CBA**	**Fileyeurs CBA**	**Bolincheurs**
Sans évidence	0.60	0.57	0.62	0.68
Hs_max fort	0.80	0.54	0.72	NA
Hs_max faible	0.50	0.63	0.60	NA
FV_mean fort	0.73	NA	NA	0.79
FV_mean faible	0.51	NA	NA	0.63
Les deux fort	0.87	NA	NA	NA
Les deux faibles	0.47	NA	NA	NA

Tableau 6 : Probabilité de non-sortie en fonction de différentes évidences et des différentes typologies. Fort signifie que l'on est au-dessus du seuil de la variable, et faible en dessous. Nous prenons toujours le seuil le plus extrême s'il y a plus de 2 seuils

III-D-4 Evolution en fonction des tendances

Des scénarios climatiques existent indiquant l'évolution des conditions environnementales globalement, notamment dans les rapports du GIEC. Dans son résumé pour les décideurs sur la partie "L'océan et la cryosphère dans le contexte du changement climatique" (GIEC, 2019), des changements et des risques sont projetés. Une augmentation de la température de l'air avec des impacts sur les débits des fleuves et des aléas locaux, une acidification continue en surface de l'océan ainsi qu'une perte d'oxygène, une élévation du niveau marin, ou encore diminution de la biomasse totale des populations d'animaux marins, de leur production et de leur potentiel de capture des pêcheries sont des exemples de projections (liste non exhaustive).

Régionalement, deux synthèses pluridisciplinaires ont été produites par le comité scientifique régional Acclimaterra visant à regrouper les connaissances scientifiques sur le contexte et au niveau régional, le comité scientifique régional AcclimaTerra qui "traite du contexte et des enjeux du climat pour l'Aquitaine, des défis pour ses ressources, ses activités et sa qualité de vie" comme il se présente sur son site internet a produit deux synthèses pluridisciplinaires. Le premier traite de l'impact du changement climatique tandis que le second se focalise sur l'anticipation et comment agir à l'échelle du territoire (Le

Treut, 2013 ; AcclimaTerra, Le Treut, 2018). Dans ces deux ouvrages, des chapitres sont dédiés aux ressources marines et à son exploitation ainsi qu'aux modifications physiques du littoral. Ces chapitres mettent en évidence que le changement climatique entraînera principalement :

- un changement de la climatologie des vagues, de l'intensité et de la fréquence des événements extrêmes de vagues ;
- un changement des débits des fleuves (cela concerne les périodes, les quantités) ;
- une modification de la circulation océanique liée à l'intensité du Gulf Stream ;
- une élévation du niveau moyen des mers ;
- une augmentation de la température (non seulement en surface mais aussi au fond) ;
- une baisse possible de la productivité marine ;
- une acidification des eaux …

Dans le projet précédent Vent&Marées, la tendance d'évolution des conditions environnementales a pu être étudiée localement, à la même échelle que ce projet. Deux variables ont en particulier été étudiées : la force moyenne du vent et la hauteur max des vagues, ici nommées ***FV_mean*** et ***Hs_max***. Une tendance plutôt décroissante a été observée pour la vitesse moyenne journalière du vent tandis qu'il n'y a pas de tendance observée pour la hauteur maximale des vagues, bien que la variabilité semble accrue depuis 2010.

Dans cette partie, nous avons eu pour objectif de modifier la probabilité de dépasser ou non le seuil des variables environnementales. Pour cela, il y a toujours 2 manières de procéder :

- On récupère des données de prédictions des variables, et on peut ainsi apprendre de nouveaux paramètres à partir de ces données et du réseau bayésien. On peut alors observer l'évolution des paramètres de notre variable cible.
  Nous n'avons pas de données à notre disposition pour cela. Nous avons considéré les données disponibles sur Copernicus, mais nous n'avons pas approfondi par manque de temps ;
- A partir de la structure de nos réseaux bayésiens appris à partir de nos données, nous n'apprenons plus les paramètres mais nous les fixons manuellement grâce à la connaissance experte. Nous pouvons tout de même extraire pour base les paramètres précédemment appris.

Pour introduire cette méthodologie, nous avons essayé avec la deuxième option. Nous avons choisi de voir comment évolue la probabilité de sortie quand nous modifions manuellement les tables de probabilités conditionnelles.
Nous allons appliquer cette méthode aux variables ***FV_mean*** et ***Hs_max***. Si nous voulons par exemple simuler une tendance à la baisse de ***Hs_max***, nous allons diviser par 2 les probabilités conditionnelles dans le cas où ***Hs_max*** est supérieur au seuil, et inversement pour une tendance à la hausse en divisant par 2 les probabilités conditionnelles où la variable est inférieure au seuil.

Nous allons ici nous concentrer sur les résultats pour les ligneurs de Capbreton (ABA) pour présenter la méthode. En effet, pour les résultats, les valeurs seront similaires à la méthode précédente, avec une hausse ou une baisse moins marquée. Nous modifierons en même temps les tables de probabilités de ***FV_mean*** et ***Hs_max*** à la hausse ou à la baisse.

Paramètres appris	**FV_mean < 4.312**				**FV_mean >= 4.312**			
	Tp_Q3 <= 12.9		**Tp_Q3 > 12.9**		**Tp_Q3 <= 12.9**		**Tp_Q3 > 12.9**	
DirVent_mean	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**
Hs_max < 2.189	0.84	0.91	0.42	0.34	0.63	0.47	0.23	0.09
Hs_max >= 2.189	0.16	0.09	0.58	0.66	0.37	0.53	0.77	0.91

Valeurs Tendance à la baisse	**FV_mean < 4.312**				**FV_mean >= 4.312**			
	Tp_Q3 <= 12.9		**Tp_Q3 > 12.9**		**Tp_Q3 <= 12.9**		**Tp_Q3 > 12.9**	
DirVent_mean	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**	**<222.857**	**>=222.857**
Hs_max < 2.189	0.92	0.955	0.71	0.67	0.815	0.735	0.615	0.545
Hs_max >= 2.189	0.08	0.045	0.29	0.33	0.185	0.265	0.385	0.455

Valeurs Tendance à la hausse	**FV_mean < 4.312**				**FV_mean >= 4.312**			
	Tp_Q3 <= 12.9		**Tp_Q3 > 12.9**		**Tp_Q3 <= 12.9**		**Tp_Q3 > 12.9**	
DirVent_mean	<222.857	>=222.857	<222.857	>=222.857	<222.857	>=222.857	<222.857	>=222.857
Hs_max < 2.189	0.42	0.455	0.21	0.17	0.315	0.235	0.115	0.045
Hs_max >= 2.189	0.58	0.545	0.79	0.83	0.685	0.765	0.885	0.955

*Tableau 7 : Tableaux présentant l'évolution des probabilités de non sorties en fonction des tendances d'évolution des variables **Hs_max** et **FV_mean.***

III-D-5 Ajout de connaissances à dire d'experts

Nos réseaux bayésiens ont pu être créés à partir des données disponibles. Mais certaines données ne sont pas disponibles, et la variable en lien ne peut donc pas être présente dans nos réseaux bayésiens.
Lors d'entretien, certains pêcheurs ont pu évoquer leur intérêt dans leur prise de décision de la qualité de l'eau. Une eau de mauvaise qualité serait pour eux synonyme de non sortie. Ils ont notamment pu pointer les rejets des stations d'épuration, mais nous n'avons pas réussi à obtenir les données correspondantes.
Dans cette situation, nous avons décidé de rajouter une variable représentant la qualité de l'eau appelée « ***Qualeau*** ». Cette variable a un seul arc, dirigé vers la variable cible « ***Sortie*** ». Elle a deux états, « **propre** » et « **impropre** ». Dans le cas où la variable prend la valeur « **propre** », nous pouvons considérer que cela ne change pas le comportement du pêcheur et que les tables de probabilités restent inchangées (cas précédent). En revanche, quand l'eau a pour valeur « **impropre** », cela influence le comportement des pêcheurs qui préféreraient ne pas sortir. Nous avons choisi de réduire par 2 les probabilités conditionnelles de non-sortie pour l'exemple.

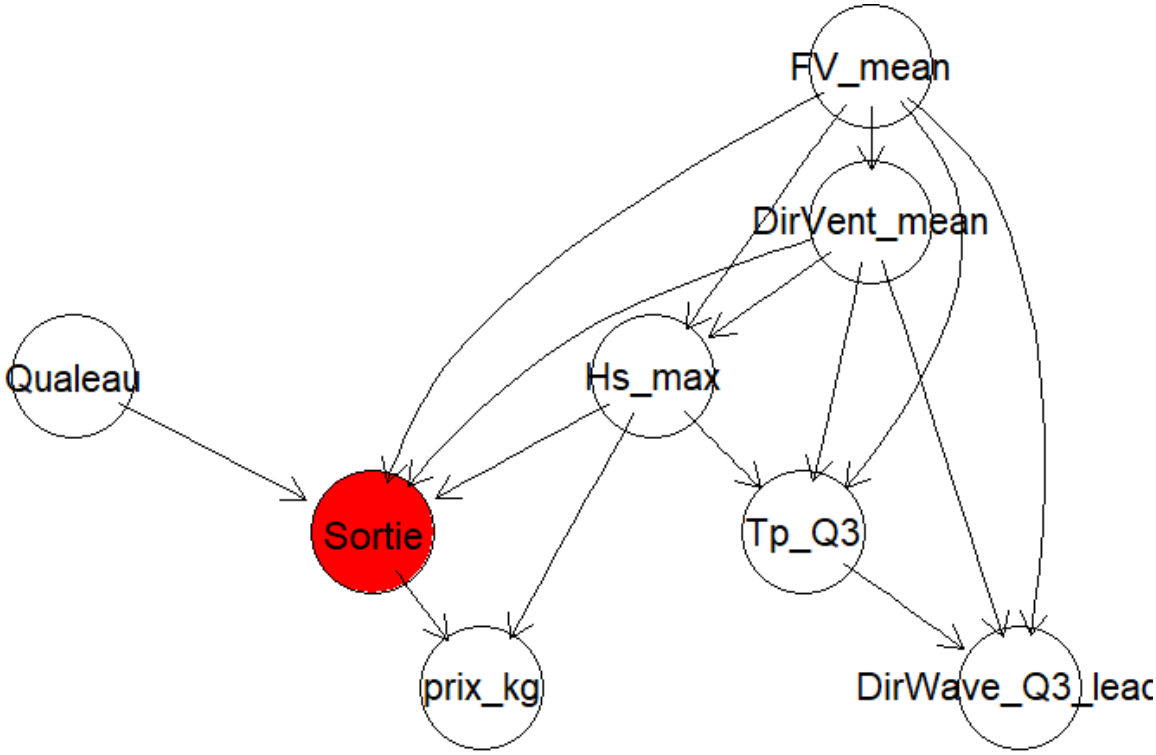


Figure 14 : Probabilité de non-sortie en fonction de différentes évidences et des différentes typologies. Fort signifie que l'on est au-dessus du seuil de la variable, et faible en dessous. Nous prenons toujours le seuil le plus extrême s'il y a plus de 2 seuils

Logiquement, on a une probabilité marginale de ***Sortie*** = 0 qui passe de 0.60 à 0.62. Ce cas-là fait office d'exemple inspiré de la réalité, mais il est voué à être mis en pratique dans les travaux qui suivront ce stage.

III-D-6 Limites

Au fil de la construction de nos réseaux bayésiens et lors de leur utilisation, plusieurs interrogations sont apparues, mettant en exergue les différentes limites :

- Les variables testées sont-elles suffisantes ou pertinentes ? Bien sûr, on peut toujours faire plus. C'est notamment l'intérêt de pouvoir compléter avec de la connaissance experte lorsque les données ne sont pas disponibles, ce qui a été présenté juste précédemment. Mais certaines variables peuvent poser des questions, notamment ***prix_kg***. Lors de la construction du premier réseau bayésien (ligneurs **ABA**), ***Sortie*** avait un arc dirigé vers ***prix_kg***, alors que l'on attendait à priori l'inverse, ce qui est le cas pour les ligneurs **CBA**. Nous avions alors considéré d'ajouter le ***prix_kg*** de la veille pour compléter.
- D'autres variables qui apparaissent souvent comme ***annee_construction*** peuvent poser soucis. Pour certaines typologies, une année peut désigner qu'un seul navire, et elle ne prend pas en compte les possibles rénovations. Est-elle vraiment informative ?
- Est-il pertinent de sélectionner autant de variables environnementales ? N'est-il pas mieux de se concentrer sur quelques variables d'intérêt ? Bien que les données puissent soulever des variables clés, est-il nécessaire de conserver plusieurs variables pour un même phénomène ? Cette question pourrait potentiellement être résolue lors d'entretiens avec les pêcheurs.
- Après la phase de datamining, nous avons discrétisé nos variables à l'aide des seuils obtenus avec la méthode des arbres de décision conditionnels, pouvant donner un certain côté binaire (au-dessus ou en dessous du seuil). Il est peut-être préférable de conserver l'aspect continue des variables pour pouvoir mieux capter la finesse des situations. Il est nécessaire pour cela d'essayer de faire des réseaux bayésiens gaussiens ou hybrides, mais la méthodologie ici

présentée ne s'applique pas et il sera nécessaire d'en développer une nouvelle tout en prenant en compte que la bibliographie et les outils disponibles sont encore plus limités.

-

Conclusion

A travers ce stage, nous avons pu développer une méthodologie permettant, à partir des données disponibles, de créer des réseaux bayésiens intégrant une variable cible représentant la sortie ou non en mer pour différentes typologies du quartier maritime de Bayonne.
Une première phase a permis de nettoyer et de préparer les données tout en créant les différentes variables potentiellement intégrées dans notre réseau bayésien.
La seconde phase a permis de sélectionner des variables clés avec des seuils pour discrétiser ces variables à l'aide d'arbres de décision conditionnels. Avec l'aide de ces variables clés discrétisées et de notre variable cible, nous avons pu créer des réseaux bayésiens continus. Nous avons pu montrer que nous pouvions utiliser ces réseaux bayésiens dans plusieurs situations :

- pour évaluer dans des conditions environnementales précises la probabilité de sortie d'un navire;
- pour évaluer cette même probabilité dans des scénarios climatiques indiquant une tendance d'évolution des variables clés ;
- pour intégrer de la connaissance experte sans avoir les données et observer l'évolution de notre variable cible.

Pour la première phase, j'ai pu me reposer sur les acquis obtenus durant mon cursus à l'Institut Agro tandis que pour la seconde phase, la méthodologie était similaire à celle du projet précédent Vents&Marées. En revanche, la troisième phase était une méthodologie peu explorée par les encadrants et par moi-même, et j'ai pu entreprendre une vraie démarche de prise en main d'une méthode pour créer et adapter une méthodologie à une question.

L'avantage clair de cette méthodologie créée est son aspect graphique et simple qui permet de comprendre facilement les enjeux de la méthode et une interprétation facilitée. Cet avantage permet de présenter les travaux à des non-initiés tout en assurant une certaine compréhension des enjeux, ce qui permettra à l'avenir d'échanger avec les personnes concernées, ici les pêcheurs, pour améliorer notre modèle.
En effet, le volet social n'a pas pu être abordé durant ce stage. Ce travail a permis d'introduire une base qui ne demande qu'à être améliorée à travers des échanges avec les pêcheurs concernés pour confirmer, ajouter, supprimer des variables ou des liens expliquant les éléments qui peuvent influencer leur choix de sortir ou non.

D'autres limites peuvent être soulevées car plusieurs choix forts ont pu être faits dans cette méthodologie. Tout d'abord, il y a le choix de travailler dans un cadre discret, il peut être intéressant d'explorer la piste des réseaux bayésiens gaussiens ou hybrides malgré leur faible présence dans la littérature scientifique. Et enfin plusieurs limites ont pu être mises en avant durant la seconde et la troisième phase, avec la pertinence de certaines variables pouvant être remise en cause (***annee_construction)*** ou encore l'amélioration du processus de sélection des variables clés pour la création des réseaux bayésiens.

Bibliographie

AcclimaTerra, Le Treut, H., 2018. (dir). Anticiper les changements climatiques en Nouvelle-Aquitaine. Pour agir dans les territoires. Éditions Région Nouvelle-Aquitaine, 2018, 488 p.

Aho, K., Derryberry, D., Peterson, T., 2014. Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95, 631–636. https://doi.org/10.1890/13-1452.1

Bastardie, F., Brown, E.J., 2021. Reverse the declining course: A risk assessment for marine and fisheries policy strategies in Europe from current knowledge synthesis. Mar. Policy 126, 104409. https://doi.org/10.1016/j.marpol.2021.104409

Brander, K., 2010. Impacts of climate change on fisheries. J. Mar. Syst., Impact of climate variability on marine ecosystems: A comparative approach 79, 389–402. https://doi.org/10.1016/j.jmarsys.2008.12.015

Bricheno, L., Wolf, J., 2018. Future Wave Conditions of Europe, in Response to High-End Climate Change Scenarios. J. Geophys. Res. Oceans 123. https://doi.org/10.1029/2018JC013866

Bru, N., Kermorvant, C., Caill-Milly, N., Lissardy, M., 2022. Rapport final : Projet Vents Et Marées.

Caill-Milly, N., Lissardy, M., Bru, N., Dutertre, M.-A., Saguet, C., 2019. A methodology based on data filtering to identify reference fleets to account for the abundance of fish species: Application to the Striped red mullet (Mullus surmulletus) in the Bay of Biscay. Cont. Shelf Res. 183, 51–72. https://doi.org/10.1016/j.csr.2019.06.004

Callens, A., Morichon, D., Abadie, S., Delpey, M., Liquet, B., 2020. Using Random forest and Gradient boosting trees to improve wave forecast at a specific location. Appl. Ocean Res. 104, 102339. https://doi.org/10.1016/j.apor.2020.102339

Franzin, A., Sambo, F., Di Camillo, B., 2017. bnstruct: an R package for Bayesian Network structure learning in the presence of missing data. Bioinforma. Oxf. Engl. 33, 1250–1252. https://doi.org/10.1093/bioinformatics/btw807

Gallet, F., Ducommun-Rigole, L., Caill-Milly, N., Lesueur, M., Gueguen, A., Lissardy, M., Morandeau, G., Le Grand, C., 2019. Etude du poids socio-économique de la filière pêche dans le quartier maritime de Bayonne.

Hothorn, T., Hornik, K., Zeileis, A., 2015. ctree: Conditional Inference Trees.

Intergovernmental Panel on Climate Change (IPCC), 2023. Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1st ed. Cambridge University Press. https://doi.org/10.1017/9781009325844

Le Treut, H., 2013. Les impacts du changement climatique en Aquitaine : un état des lieux scientifique. Pessac : Presses Universitaires de Bordeaux : LGPA-Editions, 2013, 365 p. (Dynamiques environnementales, HS 2013).

Perry, A.L., Low, P.J., Ellis, J.R., Reynolds, J.D., 2005. Climate change and distribution shifts in marine fishes. Science 308, 1912–1915. https://doi.org/10.1126/science.1111322

Ramazi, P., Kunegel-Lion, M., Greiner, R., Lewis, M.A., 2021. Exploiting the full potential of Bayesian networks in predictive ecology. Methods Ecol. Evol. 12, 135–149. https://doi.org/10.1111/2041-210X.13509

Russell, S., Norvig, P., 2009. Artificial Intelligence: A Modern Approach, 3rd edition. ed. Pearson, Upper Saddle River.

Scutari, M., 2010. Learning Bayesian Networks with the bnlearn R Package. https://doi.org/10.48550/arXiv.0908.3817

Scutari, M., Graafland, C.E., Gutiérrez, J.M., 2019. Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. Int. J. Approx. Reason. 115, 235–253. https://doi.org/10.1016/j.ijar.2019.10.003

Silander, T., Myllymaki, P., 2012. A simple approach for finding the globally optimal Bayesian network structure. https://doi.org/10.48550/arXiv.1206.6875

Strasser, H., Weber, C., 1999. On the Asymptotic Theory of Permutation Statistics. Asymptot. Theory Permut. Stat., Report Series SFB "Adaptive Information Systems and Modelling in Economics and Management Science."

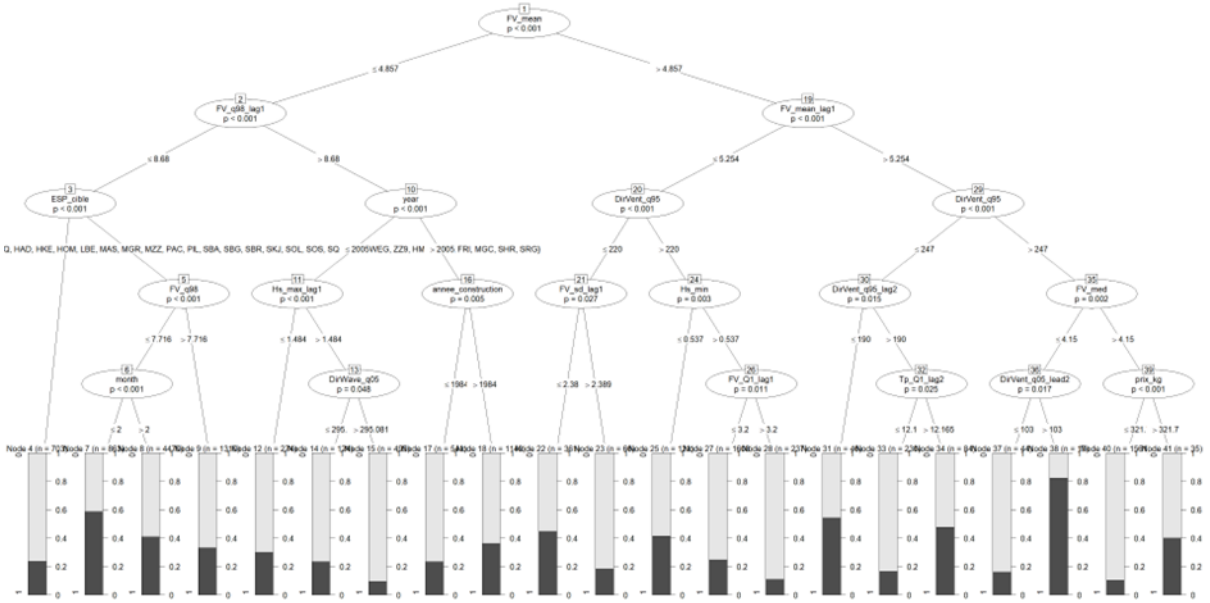
Susperrégui, N., Dela Amo, Y., Bru, N., d'Amico, F., Pigot, T., Gaudin, P., 2015. Analyse historique des tendances des facteurs abiotiques contrôlant l'apparition de liga (=mucilages marins) sur le littoral basco landais (Rapport Technique FFP).

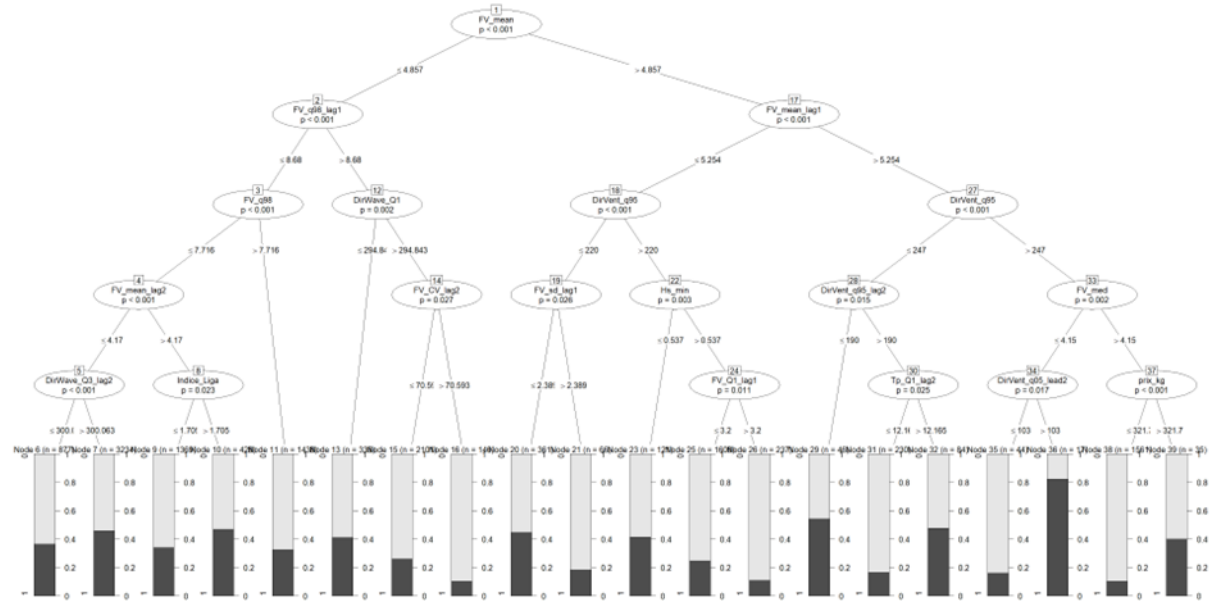
Susperregui, N., D'Elbée, J., Maton, V., Rihouey, D., Maneux, E., Etcheber, H., Sautour, B., Othéguy,

P., Monperrus, M., Maron, P., Soulier, L., Gallet, F., Dubois, L., 2010. Etude d'une substance appelée « liga » sur le littoral basque : identification, origine et facteurs influençant son apparition. (Rapport technique).
Susperregui, N., Gallet, F., Gaudin, P., Fossecave, P., 2012. Etude du phénomène « LIGA » sur le littoral basco-landais : Janvier 2011 - Juin 2012. (Rapport technique).
Trifonova, N., Maxwell, D., Pinnegar, J., Kenny, A., Tucker, A., 2017. Predicting ecosystem responses to changes in fisheries catch, temperature, and primary productivity with a dynamic Bayesian network model. ICES J. Mar. Sci. 74, 1334–1343. https://doi.org/10.1093/icesjms/fsw231
Verma, T., Pearl, J., 1991. Equivalence and Synthesis of Causal Models 221–236. https://doi.org/10.1145/3501714.3501732

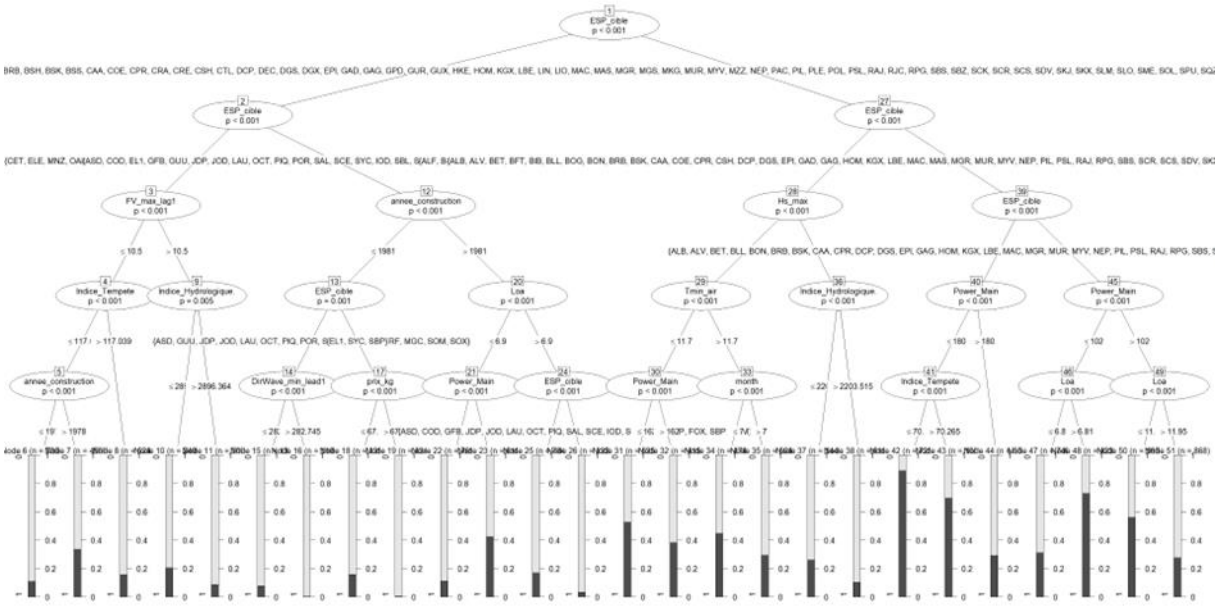
Annexes

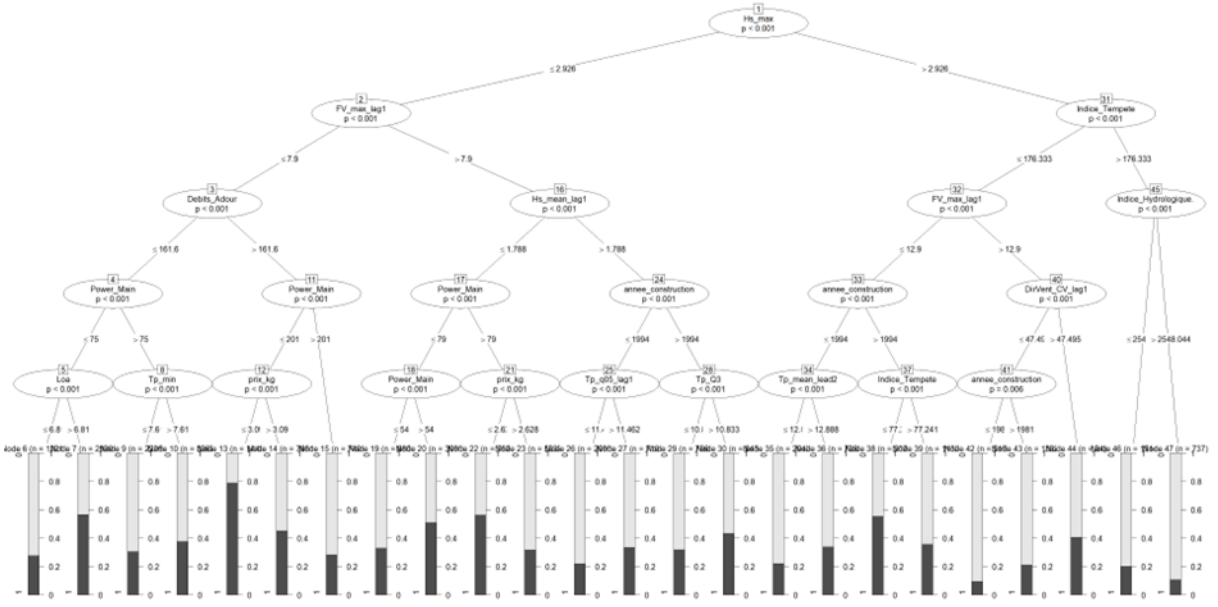
ANNEXE I : ARBRE DE DÉCISION CONDITIONNEL POUR LES BOLINCHEURS AVEC PUIS SANS LES VARIABLES DE SAISONNALITÉS



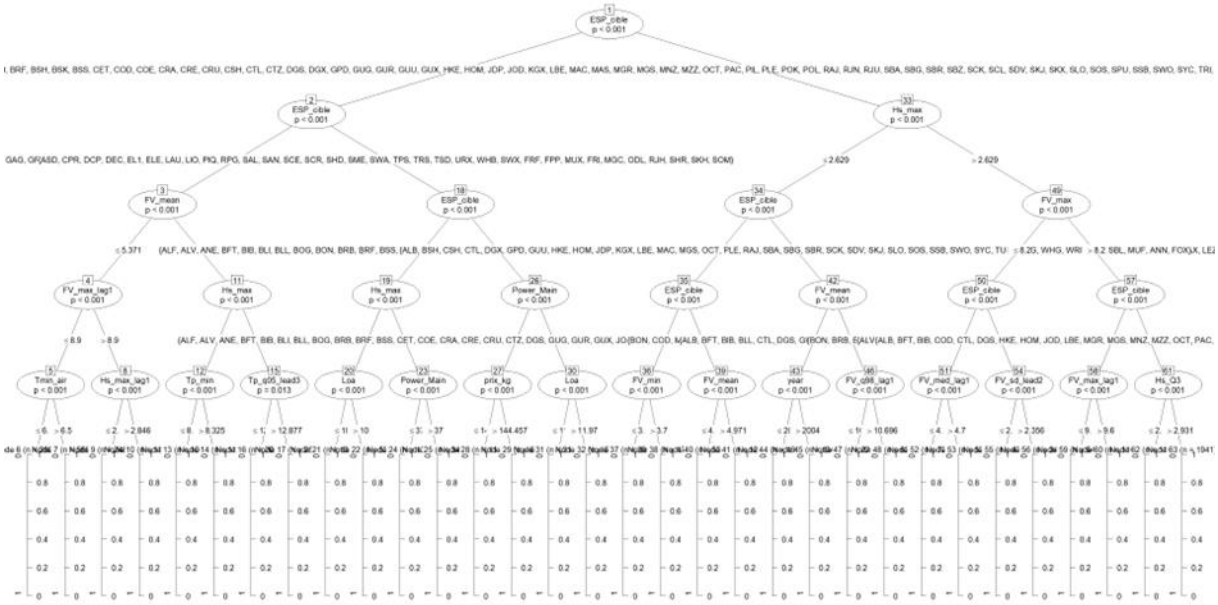


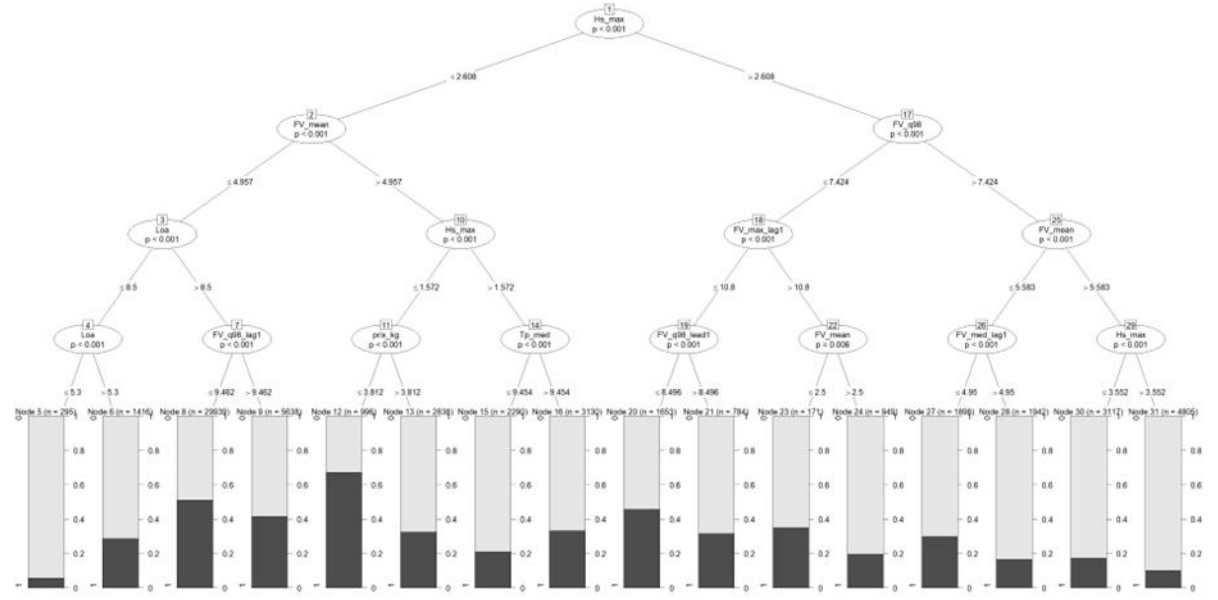
ANNEXE 2 : ARBRES DE DÉCISION CONDITIONNEL POUR LES FILEYEURS DE CIBOURE / ST-JEAN-DE-LUZ AVEC PUIS SANS LES VARIABLES DE SAISONNALITÉS



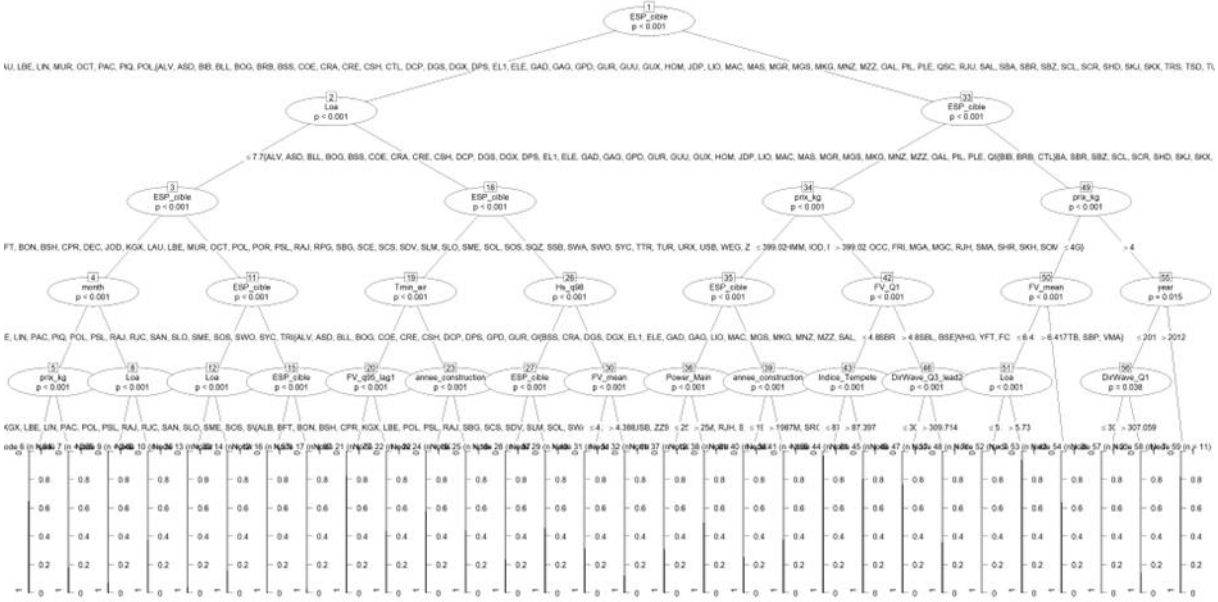


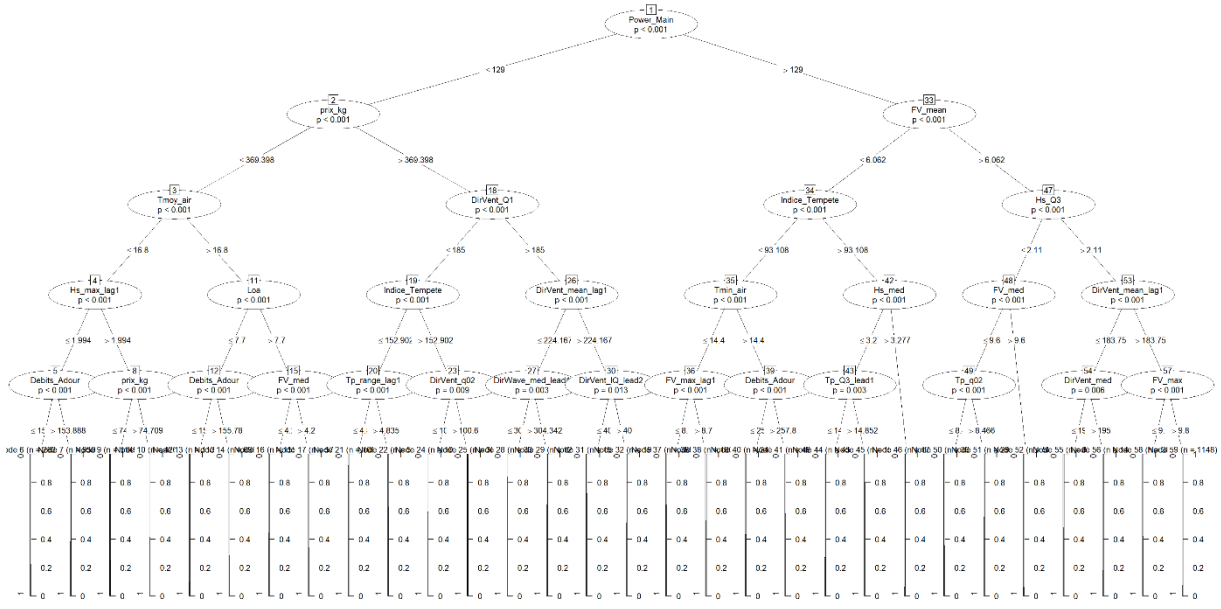
ANNEXE 3 : ARBRES DE DÉCISION CONDITIONNEL POUR LES FILEYEURS DE CAPBRETON AVEC PUIS SANS LES VARIABLES DE SAISONNALITÉS





ANNEXE 4 : ARBRES DE DÉCISION CONDITIONNEL POUR LES FILEYEURS DE BAYONNE / BOUCAU AVEC PUIS SANS LES VARIABLES DE SAISONNALITÉS





ANNEXE 5 : ARBRES DE DÉCISION CONDITIONNEL POUR LES LIGNEURS DE CIBOURE / ST-JEAN-DE-LUZ AVEC PUIS SANS LES VARIABLES DE SAISONNALITÉS