Rapport de stage

M2 Mathématiques de l'Aléatoire
Université Paris-Sud
cursus « Probabilités et statistiques »

Alexandre Lecestre

Avril-juillet 2018

Ce mémoire présente le travail effectué en stage durant quatre mois, d'avril à juillet 2018, au sein de l'UMR MISTEA (INRA - Montpellier SupAgro) sous la co-direction de Nicolas Verzelen (statisticien à l'UMR MISTEA) et Mathieu Thomas (généticien des populations au CIRAD). Ce rapport, comme le stage, se divise en deux parties indépendantes. La première partie, plus appliquée, présente l'analyse statistique de données agronomiques. La seconde partie, plus théorique, expose un résultat, accompagné de sa preuve, sur le problème de détection de ruptures à noyaux.

La première partie a été l'objet princial du stage et s'inscrit dans le projet COEX dont l'un des objectifs est d'étudier l'organisation de la diversité cultivée en Afrique de l'Ouest. Dans le cadre de ce stage, l'objectif était de réaliser une analyse statistique sur la base de données qui était à ma disposition, afin d'observer l'organisation de la diversité cultivée à l'échelle des données à ma disposition (Cameroun, Niger, Tchad). L'analyse utilise des méthodes basées sur le modèle LBM (*Latent Block Model*) pour faire du clustering sur les villages échantillonés. Après m'être documenté sur les différentes méthodes et une période de prise en main des données, j'ai pu appréhender leur utilisation en les applicant directement aux données à ma disposition. J'ai implémenté les méthodes statistiques utilisées sous R de façon générique pour faciliter les futures analyses de différents jeux de données dans le cadre du projet COEX. On pourra également utiliser ces méthodes sur des données d'approvisionnement et d'usages ce que je n'ai pas pu faire pendant le stage. À terme, ces méthodes pourraient participer, dans le cadre du projet, à la construction d'une typographie des agriculteurs et à la réalisation d'une cartographie de la diversité cultivée en Afrique de l'Ouest.

La deuxième partie a débuté plus tardivement et reste minoritaire par rapport à la première partie au niveau du temps que j'y ai consacré. Cette partie énonce un résultat en détection de ruptures par noyaux dans la lignée de ce qui est fait par Arlot, Célisse et Harchaoui ou encore Garreau et Arlot. Les preuves sont adaptées d'une analyse du problème de détection de ruptures dans le cadre sous-gaussien pour une série temporelle dans un espace de Hilbert à noyau reproduisant.

1 Analyse statistique de données d'inventaire

1.1 Introduction

Cette partie présente un travail d'analyse statistique de données qui s'inscrit dans le cadre du projet COEX[1]. Un des objectifs du projet est de documenter, à l'échelle de l'Afrique de l'Ouest, la diversité de sources d'approvisionnement en semence, la diversité des usages des plantes cultivées et la diversité des types de variétés utilisées par les agriculteurs. Une des activités du projet consiste à réaliser une méta-analyse des jeux de données de nature similaire collectée à une échelle plus large (Afrique) dans le cadre de précédents projets (FFEM, ARCAD, ATP, Plantadiv, ...). Lors de ce stage, l'objectif a été de réaliser une analyse statistique de la biodiversité cultivée en Afrique subsaharienne (Cameroun, Tchad et Niger) à partir de données d'inventaire issues du projet de recherche Plantadiv[2]. Les outils statistiques utilisés pour mener cette analyse pourront facilement être réutilisés pour analyser d'autres données d'inventaire de culture. Ces données consistent essentiellement en un recensement des espèces cultivées dans différents villages échantillonnés. Au vu des données, l'objectif a été de comprendre comment la biodiversité cultivée se distribue dans l'espace. Pour cela, on a été amené à se poser différentes questions. La première a été de voir si la biodiversité cultivée est distribuée uniformément dans l'espace. Si cette distribution n'est pas uniformément aléatoire, existe-t-il des regroupements de villages basés sur la liste des espèces cultivées dans chacun d'eux ? De la même manière, existe-t-il des (groupes d') espèces spécifiques à certains (groupes de) villages ? Les groupes de villages construits sont-ils cohérents d'un point de vue géographique, culturel, ethnologique ? Les groupes d'espèces co-cultivées se différencient-ils par leur rôle joué dans l'alimentation ?
Le stage a débuté par une période de prise en main des données. Je ne disposais que des données pour une requête sur les 13 espèces les plus communes. Aucun résultat obtenu avec ces données n'a finalement été utilisé. Cependant, cela a permis de mettre en pratique les différents outils d'analyse en parallèle d'une familiarisation avec les objets statistiques utilisés qui correspondent essentiellement à ceux utilisés dans [1]. Cette période a aussi permis de réaliser beaucoup d'analyses exploratoires et ainsi mieux cerner la base de données, de comprendre les différentes variables, quelles informations sont à notre disposition, les quelles sont réellement exploitables d'un point de vue statistique. Tout au long du stage, les données utilisées ont changées plusieurs fois. On ne disposait que de 13 espèces au début, ensuite de toutes les espèces, puis on ne s'est intéressé qu'aux espèces appartenant à l'herbier utilisé pour récolter les données. Finalement, on a aussi retiré les plantes sauvages pour ne conserver que les plantes semées ou cultivées. À chaque fois que les données changent, il faut recommencer l'analyse. Même si les méthodes ne changent pas, les résultats peuvent fortement varier. Cela n'a pas laissé beaucoup de temps pour approfondir l'analyse des résultats,

1. `https://umr-agap.cirad.fr/projets-de-recherche/coex`
2. `http://www.agence-nationale-recherche.fr/Projet-ANR-07-BDIV-0005`

d'autant plus que je n'ai pas forcément les connaissances spécifiques pour pouvoir le faire. Les méthodes utilisées sont facilement transposables à tout type de données d'inventaire. J'ai donc développé un script *R* permettant d'appliquer simplement ces méthodes, réutilisable par des ethnologues et accompagné d'une aide à base d'exemples dans un fichier *html*.

1.2 Présentation des données

Le travail a été effectué exclusivement sur la base de données issue du projet Plantadiv [3]. Cette base de données contient des informations concernant la biodiversité observée dans plusieurs villages de trois pays d'Afrique subsaharienne (Cameroun, Tchad et Niger). Ces données sont issues de prospections et de collectes de semences réalisées entre 2009 et 2012. La base est très riche, on dispose de nombreuses variables assez diverses telles que la nomenclature locale d'une variété, l'ancienneté et l'origine des plantes, l'outillage utilisé, les pratiques, les différents usages, la proximité d'un village au marché, etc. Les données ont été récoltées et renseignées à l'échelle de l'agrosystème. On précisera un peu plus tard ce qu'on entend exactement par « agrosystème » qui est un abus de langage ici mais qu'on utilisera tout le long du rapport. En pratique, les enquêteurs visitent un agrosystème avec un herbier (une mallette) et réalisent un entretien collectif avec des agriculteurs volontaires. L'information principale de la base de données et qui ne contient pas de données manquantes est l'inventaire des espèces cultivées dans un agrosystème. C'est cette information qui est à la base de nos analyses.
Afin de travailler plus efficacement, on a d'abord réalisé un prétraitement sur les données brutes pour obtenir un jeu de données plus adapté aux analyses multivariées. Par soucis de clarté, ce travail fastidieux et spécifique aux données du projet Plantadiv est présenté en annexe. On peut tout de même noter que certains agrosystèmes ont été retirés pour diverses raisons (précisées dans la partie en annexe). Pour clarifier le vocabulaire utilisé, on entendra par « observation » une ligne dans le fichier *csv* obtenu par une requête sur la base de données. On parlera d'« agrosystème » pour désigner un groupe de personnes vivant dans un même village et ayant des pratiques culturelles et agronomiques similaires. À quelques exceptions près, un agrosystème représentera un village. Néanmoins, trois villages différents contiennent deux agrosystèmes. On considère qu'il y a deux agrosystèmes lorsqu'on trouve deux groupes qui habitent le même village mais qui parlent deux langues différentes, qui appartiennent à deux ethnies différentes, qui utilisent des techniques différentes, qui ont des cultures globalement différentes.
De la même manière, on fera un abus de langage en utilisant « variété » non pas au sens de la nomenclature biologique mais pour parler de « type nommé ». Autrement dit, si deux plantes sont nommées de la même manière dans un agrosystème parce que les paysans ne les distinguent pas, alors on ne renseigne qu'une seule variété dans la base de données puisqu'il n'y a qu'un seul type nommé. Une même variété (au sens biologique) n'est pas forcément nommée de la même manière à différents endroits. On distingue parfois deux variétés (au sens biologique) dans un même agrosystème mais on ne peut pas « suivre une variété dans l'espace ». C'est la raison pour laquelle on ne peut pas réaliser les mêmes analyses statistiques au niveau spécifique et variétale. On pourra tout de même utiliser ce qu'on appellera la diversité variétale (nombre de variétés différentes d'une même espèce cultivées dans un agrosystème). Afin de mieux comprendre l'information contenue dans les données et les limites de cette information, il faut comprendre la variable *varCode* expliquée en détail ci-après.

1.2.1 Variable *varCode*

La variable *varCode* est au centre de la base de données et contient l'essentiel de l'information. On explique comment elle est construite avec en exemple l'observation « T-AM-2009-002-01-02-49 ».

- T-AM-2009-002-01-02-49 : initiales de l'informateur.rice
- T-AM-2009-002-01-02-49 : année de l'inventaire
- T-AM-2009-002-01-02-49 : identifiant **unique** d'un transect
- T-AM-2009-002-01-02-49 : identifiant du village sur le transect
- T-AM-2009-002-01-02-49 : identifiant **unique** d'un village

3. `http://www.agence-nationale-recherche.fr/Projet-ANR-07-BDIV-0005`

- T-AM-2009-002-01-02-49 : identifiant de l'agrosystème dans le village
- T-AM-2009-002-01-02-49 : identifiant **unique** d'un agrosystème
- T-AM-2009-002-01-02-49 : identifiant **unique** de l'espèce
- T-AM-2009-002-01-02-49 : identifiant de la variété pour une espèce donnée, dans un agrosystème donné
- T-AM-2009-002-01-02-49 : identifiant **unique** d'une variété dans un agrosystème

Il faut garder en tête ce que contient la variable *varCode* et comment elle est constituée. Cela permet de déterminer quelles analyses on peut faire et celles qu'on ne peut pas faire. L'exemple le plus clair et qu'on a évoqué avant : l'identifiant de la variété permet uniquement de distinguer deux types nommés dans un même agrosystème. On ne peut pas comparer les variétés entre agrosystèmes comme on peut le faire avec les espèces. En revanche, on peut compter le nombre de variété par espèce et par agrosystème. Cela semble assez limité pour faire des analyses au niveau infra-spécifique mais on peut tout de même utiliser cette indice de diversité variétale dans notre analyse.

1.2.2 Matrices d'adjacence

Expliquons comment ont été transformé les données brutes. On peut résumer l'information de biodiversité cultivée au niveau spécifique avec une matrice d'adjacence dont les lignes et les colonnes sont respectivement les agrosystèmes et les espèces, et dont chaque cellule vaut 1 si l'espèce est cultivée dans l'agrosystème et 0 sinon. Cette matrice permet de savoir si un agrosystème cultive ou ne cultive pas une espèce sans avoir à parcourir toutes les observations de notre jeu de données. La figure 1.1 montre la matrice obtenue avec les données.

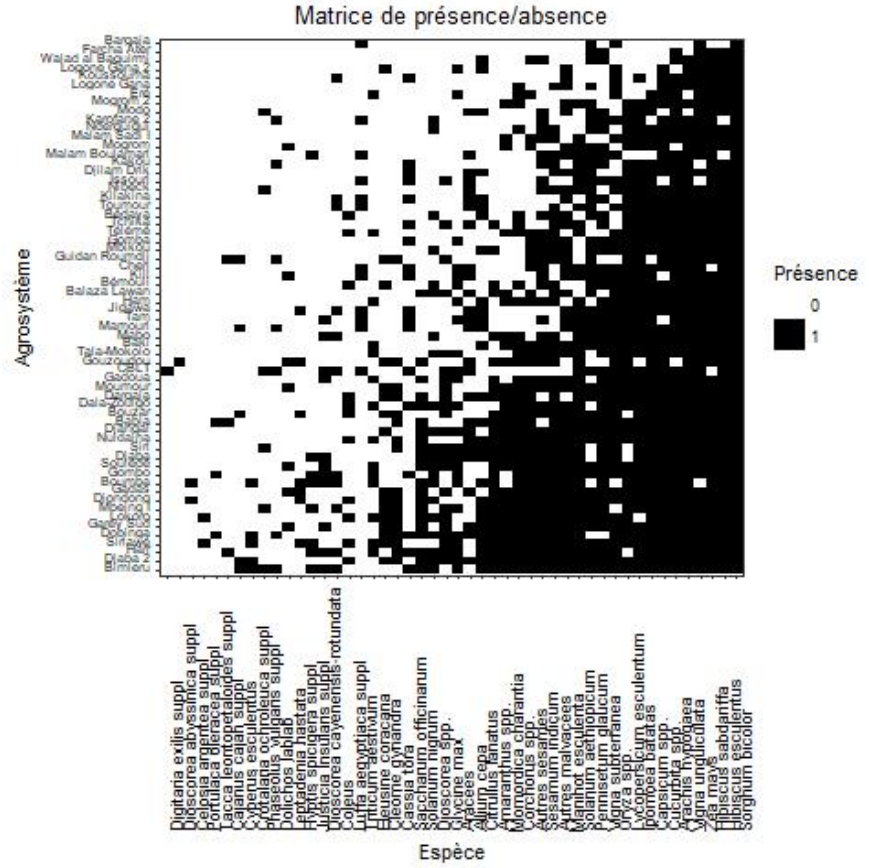


FIGURE 1.1

Sur la matrice de la figure 1.1, les agrosystèmes et les espèces on été ordonnés par leur degré. On parle de « degré » pour le nombre d'espèces cultivées (degré d'un agrosystème) et pour le nombre d'agrosystèmes cultivant une même espèce (degré d'une espèce). Avec des notations mathématiques, la matrice d'adjacence B, pour matrice binaire, est telle que $B_{ij} = 1$ si l'agrosystème i cultive l'espèce j et $B_{ij} = 0$ sinon. Dans la suite, on parlera de matrice binaire ou de matrice de présence/absence.
On peut également résumer l'information de diversité variétale (nombre de variétés cultivées dans un agrosystème) naturellement sous la forme d'une matrice qu'on nomme C, pour matrice de comptage. Dans ce cas, chaque coefficient C_{ij} de la matrice correspond au nombre de variétés de l'espèce j cultivées dans l'agrosystème i. On peut voir ce qu'on obtient avec les données sur la figure 1.2.

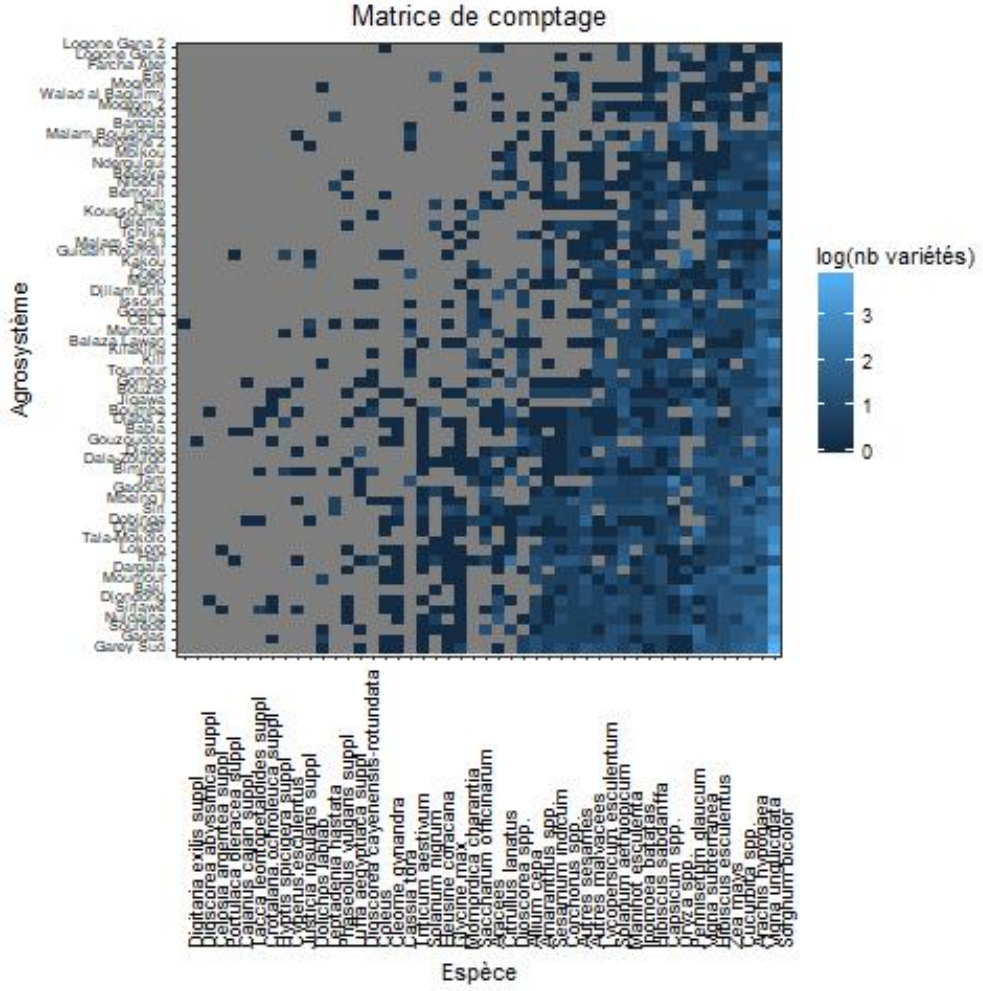


FIGURE 1.2

Le sorgho présente un nombre de variétés très important comparé aux autres espèces. On a donc utilisé une échelle logarithmique pour mieux représenter la diversité variétale. De plus, pour bien les distinguer, on a représenté en gris les coefficients nulles de la matrice, c'est-à-dire lorsqu'une espèce n'est pas cultivée dans un agrosystème (0 variété).
Remarque : Ici, une cellule contient un entier qui est le nombre de variétés cultivées dans le village donné pour l'espèce donnée. On peut noter que cette nouvelle matrice contient plus d'information que la matrice de présence/absence. En effet, on retrouve l'ancienne matrice en mettant à 1 les cellules contenant un entier strictement positif. Plus formellent, on retrouve B à partir de C avec $B_{ij} = 1_{C_{ij}>0}$.

1.2.3 Espèces cultivées : protocoles et catégorie

Un autre aspect très important dans nos données est le fait que les données d'inventaire ont été récoltées sur la base d'une liste fermée de 60 espèces, qu'on peut trouver en annexe. En effet, les entretiens collectifs sont réalisés avec une mallette qui contient un herbier avec ces 60 espèces. Cet herbier est montré aux agriculteurs afin de déterminer ce qu'ils cultivent et ne cultivent pas parmi cette liste d'espèces. Le fait de présenter un herbier avec une liste fermée d'espèces permet normalement de dire avec certitude que si une espèce n'apparaît pas dans les observations c'est qu'effectivement elle n'est pas cultivée et cette information est très importante. En plus de l'identifiant espèce, on dispose d'un niveau agrégé de 48 « espèces » (variable *espCatListes*). On a également des catégories d'espèces (variable *espCat*) qu'on utilise avec le code couleur suivant : oléagineux, brèdes et condiments, légumineuses, tubercules et céréales. L'agrégation et la catégorisation a été faite en concertation par les différents acteurs du projet Plantadiv. Dans l'intégralité du rapport, on utilisera uniquement le niveau agrégé mais on continuera de parler d'« espèce ».

1.2.4 Villages et agrosystèmes

Sachant qu'on s'intéresse à la répartition spatiale de la biodiversité cultivée, il est important de visualiser où se trouvent les différents villages de nos données. Celles-ci concernent 62 agrosystèmes pour 59 villages répartis sur 3 pays. On trouve 17 villages au Cameroun, 15 au Niger et 17 au Tchad. Les coordonnées GPS des villages sont déjà renseignées dans les données. Les 59 villages sont représentés sur la carte ci-dessous avec des couleurs qui correspondent à différents ensembles géographiques.

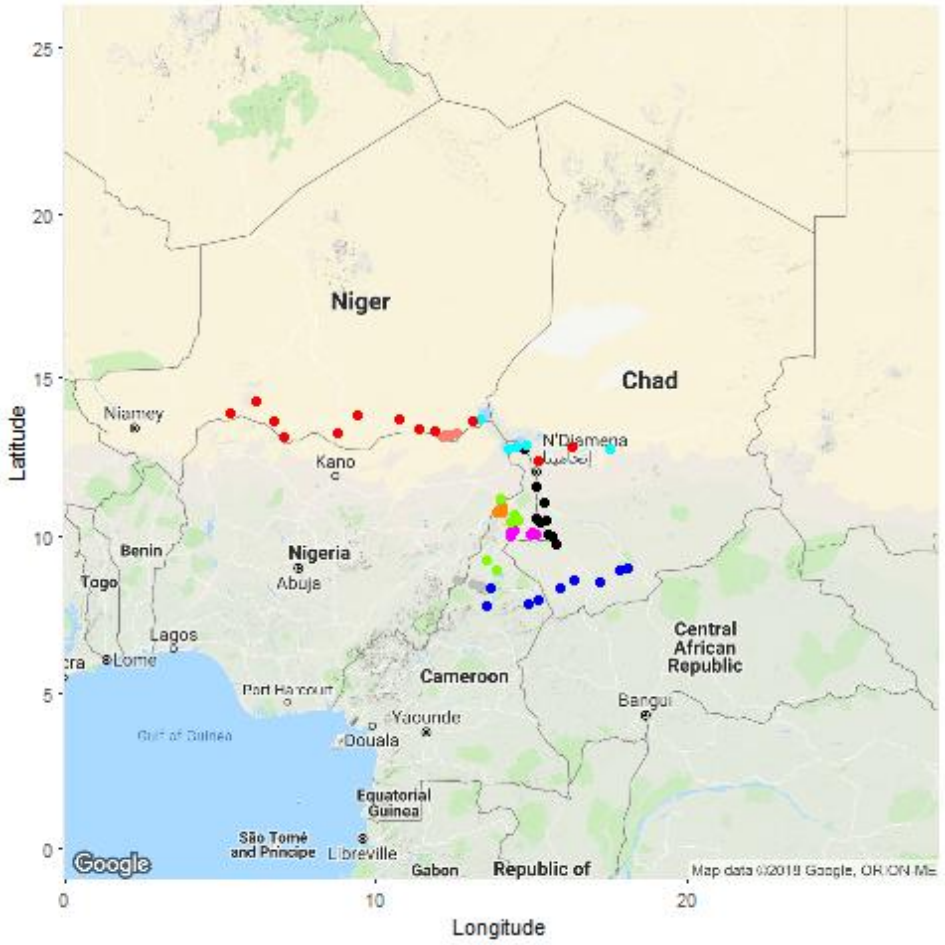


FIGURE 1.3

Dans le cadre du projet Plantadiv ont été distingué les ensemble géographiques suivants :

- le bec de canard (7 villages en magenta),
- le massif d'Adamaoua (3 villages en gris),
- la plaine soudanienne (9 villages en bleu),
- la plaine camerounaise (6 villages en vert),
- la zone inondable du fleuve Logone (9 villages en noir),
- la zone sahélienne (12 villages en rouge),
- la région des lacs Tchad et Fitri (5 villages en cyan),
- les monts Mandara (4 villages en orange),
- et la vallée du fleuve Komadougou Yobé (4 villages en saumon).

Ces ensembles seront éventuellement utilisés lors de l'interprétation des résultats.

1.2.5 Les variables agrosystèmes

Pour la collecte de données, l'unité d'observation a été l'agrosystème. Les analyses ont donc été faites à la même échelle. Le jeu de données contient beaucoup de variables, certaines font référence à un niveau plus fin que l'agrosystème (semence, champ, exploitation, etc.) mais on dispose bien de plusieurs variables qui sont renseignées au niveau de l'agrosystème. Seules celles-ci peuvent être utilisées dans l'analyse et l'interprétation des résultats obtenus mais toutes ne sont pas utilisables. Certaines utilisent une échelle trop précise (canton ou langue par exemple), certaines ne sont que des commentaires (phrases entières inutiles d'un point de vue statistique) et d'autres demanderaient une harmonisation qui n'a pas été faite. Les variables utilisables concernent le type d'agriculture pratiquée (*asTypologie*), la durée ou l'absence de jachères (*asTechJachères*), l'outillage des agriculteurs (*asTechOutillage*), les techniques d'amendement utilisées (*asTechAmendements*), la conduite de l'élevage (*asElevageConduite*), l'importance de l'élevage dans l'agrosystème (*asImportanceElevage*) et l'importance de la pêche dans l'agrosystème (*asPêche*). Ces variables n'ont pas été utilisées dans l'analyse des résultats par manque de temps, mais les figures obtenues sont disponibles en annexe.
La base de données contient beaucoup plus d'information mais on a présenté ici uniquement ce qui est utilisé et important pour comprendre l'analyse qui a été faite. Tout cela est un peu plus détaillé dans la partie de prétraitement des données en annexe. Après avoir présenté les données, la section suivante introduit les méthodes statistiques utilisées pour l'analyse.

1.3 Méthodes statistiques utilisées

Maintenant qu'on a brièvement décrit le jeu de données, on va présenter, dans notre contexte, les méthodes utilisées pour l'analyse des données. Cette analyse portera exclusivement sur les deux matrices d'adjacence de la section 1.2.2, la matrice binaire dans un premier temps puis la matrice de comptage par la suite. Cette analyse s'inspire fortement de [1] et utilise des méthodes statistiques qui permettent de faire de la bi-classification. C'est-à-dire, qu'avec ces méthodes on va pouvoir faire simultanément des groupes d'agrosystèmes et des groupes d'espèces. La méthode principalement utilisée sera l'inférence d'un modèle probabiliste sur nos données appelé "Latent Block Model". On va introduire ce modèle probabiliste dans notre contexte d'inventaire de diversité cultivée au niveau agrosystème et spécifique. C'est un modèle de graphe aléatoire (biparti) que l'on peut également voir comme un modèle de matrice aléatoire en passant par la notion de matrice d'adjacence d'un graphe. On ne va pas du tout utiliser le point de vu « graphe » mais uniquement celui des matrices. Pour l'interprétation en terme de graphes, on peut se référer à l'annexe.
Dans toute la suite de cette section, on reprendra les notations B et C pour la matrice binaire et la matrice de comptage. Lorsqu'on veut rester général et considérer l'une ou l'autre des deux matrices, on parlera d'une matrice M. On parlera de village plutôt que d'agrosystème mais les concepts restent exactement les mêmes. D'un point de vue probabiliste, on peut considérer que chaque coefficient M_{ij} est la réalisation d'une variable aléatoire. Dans ce cas-là, il est intéressant de voir si ces variables sont indépendantes, et si on a un modèle paramétrique, comment varie le(s) paramètre(s) sur une ligne ou sur une colonne par exemple. Il existe plusieurs modèles probabilistes qu'on peut tester sur nos données. L'objectif de l'analyse est de retrouver des

motifs dans la matrice M. On travaille d'abord sur la matrice de présence/absence. Notre démarche est de tester un modèle nul sur cette matrice et de vérifier qu'il est bien rejeté avant d'essayer d'autres modèles.

1.3.1 Modèle nul

Dans cette section, on se situe dans la lignée de l'analyse de modèles nuls tel qu'elles sont faites en écologie. On pourra trouver des exemples et un historique des modèles nuls en écologie dans [5]. Ici, celui qu'on appelle modèle nul suppose que toutes les cellules de la matrice suivent la même loi et sont indépendantes. Autrement dit, le fait que le village i_1 cultive l'espèce j_1 est autant probable que le village i_2 cultive l'espèce j_2 et cela quelque soient les espèces j_1 et j_2 et les villages i_1 et i_2. C'est un modèle uniforme, à la fois au niveau des agrosystèmes et aussi des espèces. Mathématiquement, ce modèle s'écrit

$$B_{ij} \overset{iid}{\sim} \mathcal{B}(p), \forall i, j,$$

où $p \in [0, 1]$. On va tester ce modèle en considérant une statistique qui porte sur la variance empirique des degrés (des espèces et des villages). Le test et le modèle sont détaillés en annexe.

1.3.2 Latent Block Model (LBM)

Lorsque le test précédent rejette l'hypothèse nulle, on veut rechercher une autre structure, plus « complexe » dans la matrice B. On va utiliser dans notre analyse un modèle à variables cachées appelé « Latent Block Model » qui permet de réaliser une bi-classification, c'est-à-dire deux partitions simultanées, une sur les villages et une sur les espèces. Ces partitions sont censées présenter des groupes de villages qui cultivent de la même façon et des groupes d'espèces qui sont globalement cultivées dans les mêmes villages. Pour rester général, on va présenter le modèle à partir d'une matrice M qui peut être une matrice binaire ou une matrice de comptage. En effet, le modèle LBM est assez général et il en existe une version pour chaque matrice.
Un modèle LBM à $K_1 \times K_2$ groupes est décrit par un triplet (α, β, P). On a deux lois de probabilités $\alpha = (\alpha_1, \ldots, \alpha_{K_1})$ et $\beta = (\beta_1, \ldots, \beta_{K_2})$, respectivement sur $\{1, \ldots, K_1\}$ et $\{1, \ldots, K_2\}$. On a aussi un ensemble de paramètres $P = (p_{st})_{1 \leq t \leq K_2}^{1 \leq s \leq K_1}$. Pour $s \in \{1, \ldots, K_1\}$, α_s est la probabilité qu'un village appartienne au groupe de village s. De même, β_t est la probabilité qu'une espèce soit dans le groupe d'espèces t. Le contenu de la matrice P dépend des lois considérées pour le modèle mais l'idée est la même pour toutes les lois paramétriques. Le coefficient p_{st} décrit le comportement des espèces du groupe d'espèces s au sein des villages du groupe de villages t. Pour le village i, on note A_i le groupe de villages auquel il appartient et pour l'espèce j, on note E_j le groupe d'espèces auquel elle appartient. Alors le modèle LBM avec des lois $\mathcal{F}$ s'écrit

$$\cdot A_i \overset{iid}{\sim} \alpha, \forall i,$$

$$\cdot E_j \overset{iid}{\sim} \beta, \forall j,$$

$$\cdot M_{ij} | A_i, E_j \overset{ind.}{\sim} \mathcal{F}(p_{A_i E_j}), \forall i, \forall j.$$

On peut imaginer toute sorte de loi paramétrique pour $\mathcal{F}$. Dans notre cas, on a une matrice binaire et une matrice de comptage donc il nous faut une loi sur $\{0, 1\}$ et une loi sur $\mathbb{N}$. On va d'abord considérer un modèle avec des lois de Bernoulli pour les matrices binaires.

Modèle à lois de Bernoulli

Dans cette partie, on considère $M = B$ puisqu'on utilise des lois de Bernoulli. C'est l'unique modèle possible pour les matrices binaires. Dans ce modèle, la probabilité qu'un village cultive une espèce ne dépend que du groupe auquel appartient l'espèce et du groupe auquel appartient le village. Pour chaque couple de groupes $(s, t) \in \{1, \ldots, K_1\} \times \{1, \ldots, K_2\}$, on a un paramètre $p_{st} \in [0, 1]$ qui donne la probabilité pour un village du groupe s de cultivé une espèce du groupe t. Le modèle s'écrit

$$B_{ij} | A_i, E_j \overset{ind.}{\sim} \mathcal{B}(p_{A_i E_j}), \forall i, \forall j.$$

La figure 1.4 représente un modèle LBM à lois de Bernoulli simulé pour 60 villages, 90 espèces, 3 groupes de villages et 3 groupes d'espèces. Pour bien visualiser la structure d'un tel modèle, les lignes et les colonnes sont rassemblées par groupe et les groupes sont séparés par des lignes bleus.

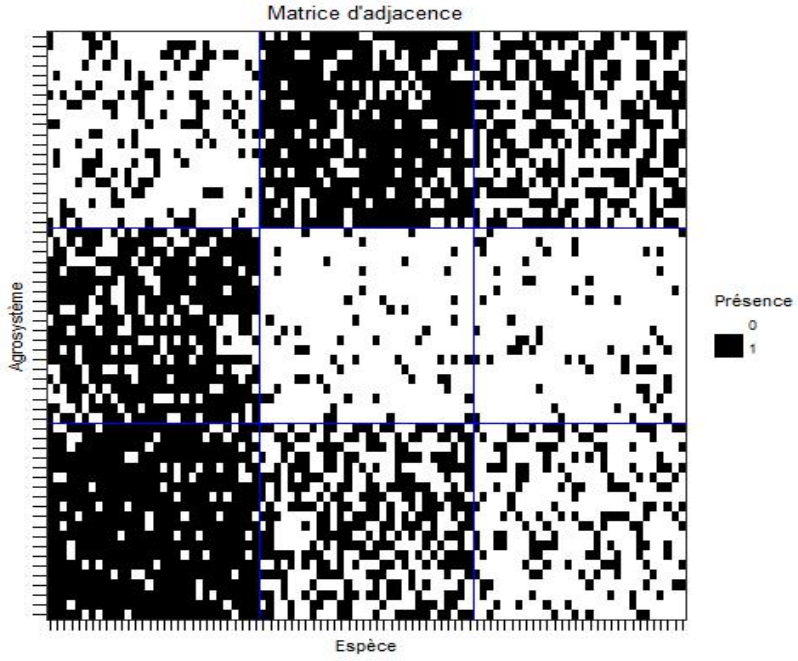


FIGURE 1.4

Ici, on a simulé un modèle dont on a choisi nous-même les paramètres. En pratique, on veut faire le « chemin inverse », c'est-à-dire retrouver le modèle à partir d'une réalisation de ce modèle. Une des difficultés pour réaliser cette opération est qu'on ignore le nombre de groupes que ce soit pour les espèces ou pour les agrosystèmes. Si on veut un modèle le plus « proche » possible de nos données, on va avoir tendance à surestimer fortement le nombre de groupes. Le modèle avec exactement un groupe pour chaque espèce/village, et tel que p_{ij} vaut 1 si le village i cultive l'espèce j et 0 sinon, risque de réaliser l'optimum voulu selon le critère choisi. Pour éviter cette surestimation du nombre de groupes, il faut mettre en place une procédure de sélection de modèles. Dans notre cas, on utilise le package *blockmodels* qui infère un modèle LBM par maximum de vraisemblance pénalisé. En pratique, l'inférence est faite par maximisation du critère ICL (pour *Integrated Complete Likelihood*) avec un algorithme EM (espérance-maximisation) variationnel. Il existe d'autres moyens de faire de la classification (non supervisée) comme la classification hiérarchique ascendante (construction d'un dendrogramme) ou encore le clustering spectral. La consistance du clustering spectral a été prouvé dans [6] pour un modèle SBM (*Stochastic Block Model*) et dont la preuve s'adapterait probablement assez facilement au cas LBM. Bien que ces méthodes ne soient pas inintéressantes et pourraient être utiliser pour analyser les données, elles ne permettent pas de réaliser une bi-classification. Le modèle LBM lie la classification sur les espèces et la classification sur les agrosystèmes. Chaque classification permet d'expliquer l'autre. Réaliser deux classification séparément sur les espèces puis sur les agrosystèmes ne permet pas d'analyser aussi facilement les groupes qu'on obtient.

Remarque : un point important mais qu'on peut vite oublier en regardant les résultats tels qu'ils sont présentés ici est que l'inférence d'un modèle LBM ne donne pas les groupes d'espèces ni les groupes de villages. En effet, on obtient juste une probabilité a posteriori d'appartenance aux différents groupes pour les espèces et les villages. Ces probabilités a posteriori sont données en annexe et permettent de dire avec quelle confiance on a assigné un groupe à un agrosystème. La probabilité a posteriori maximale est bien souvent très proche de 1. On présente donc les résultats en assignant à chaque espèce/village le groupe avec la plus grande probabilité d'appartenance a posteriori. Cela permet bien d'obtenir une bi-classification sur les espèces et les villages.

Modèle à degré corrigé

Une question qu'on peut être amené à se poser est de savoir si les groupes se distinguent vraiment par les espèces qu'ils cultivent plus que par le nombre d'espèces qu'ils cultivent. On peut essayer d'y répondre avec ce qu'on appellera ici un modèle LBM à degré corrigé. On s'inspire du modèle DCBM (*Degree Corrected Stochastic Block Model*) qui est notamment évoqué dans [6]. Pour caractériser le modèle LBM à degré corrigé, on a besoin de deux vecteurs ϕ et ψ qui représentent la correction des degrés sur les villages et sur les espèces. Alors le modèle est défini tel que

$$B_{ij}|A_i, E_j \overset{ind.}{\sim} \mathcal{B}(\text{logit}^{-1}(p_{A_i E_j} + \phi_i + \psi_j)),$$

où $\text{logit}(p) = \log(p/(1-p))$. Les vecteurs ϕ et ψ vont permettre de faire une distinction entre différents villages d'un même groupe et différentes espèces d'un même groupe ce qui n'était pas le cas pour le modèle LBM à lois de Bernoulli. La figure 1.5 présente une simulation d'un modèle LBM à degré corrigé.

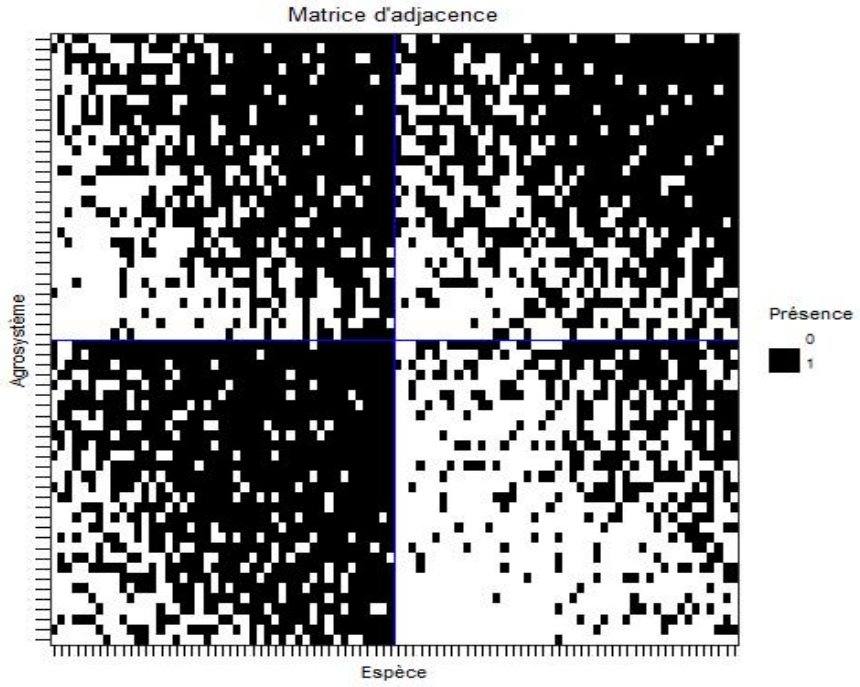


FIGURE 1.5

On voit bien sur la figure 1.5 qu'on n'a pas l'homogénéité au sein d'un groupe qu'on avait sur la simulation précédente, notamment en ce qui concerne les degrés. En pratique, cela nous permettra d'obtenir une bi-classification avec des agrosystèmes et des espèces dont le degré peut fortement varier au sein d'un même groupe. On repère facilement les agrosystèmes « riches » (degré élevé) et les agrosystèmes « pauvres » (degré faible) sur une matrice de présence/absence, de même pour les espèces rares et les espèces communes. Le modèle à degré corrigé permet d'aller plus loin et de détecter des choses qu'on ne perçoit pas immédiatement à la vue de la matrice binaire.

Modèle à lois de Poisson

Après avoir donné deux modèles pour la matrice binaire, concentrons nous sur la matrice de comptage C. On a alors besoin d'une loi paramétrique sur $\mathbb{N}$ et la loi de Poisson est assez commune sur l'ensemble des entiers positifs. On ne peut pas dire que nos données s'accordent bien avec des lois de Poisson mais il existe déjà une fonction dans le package *blockmodels* qui permet d'inférer un modèle LBM avec des lois de Poisson. Dans ce modèle, le nombre de variétés d'une même espèce cultivées dans un village ne dépend que du groupe auquel appartient l'espèce et du groupe auquel appartient le village. Pour chaque couple de groupes

$(s,t) \in \{1,\dots,K_1\} \times \{1,\dots,K_2\}$, on a un paramètre $p_{st} \in \mathbb{R}^+$ qui caractérise le nombre moyen de variétés d'une espèce du groupe t cultivées dans un village du groupe s. Le modèle s'écrit

$$C_{ij}|A_i, E_j \overset{ind}{\sim} \mathcal{P}(p_{A_i E_j}).$$

La figure 1.6 représente un modèle LBM à lois de Poisson simulé pour 60 villages, 90 espèces, 3 groupes de villages et 3 groupes d'espèces. À nouveau, les lignes et les colonnes sont rassemblées par groupe et les groupes sont séparés par des lignes.

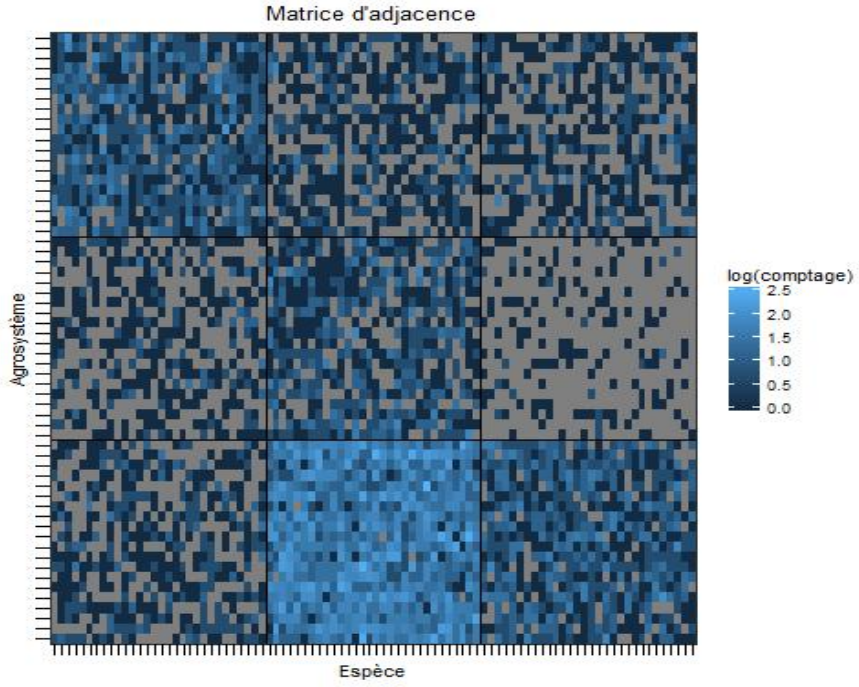


FIGURE 1.6

On utilise une échelle logarithmique pour représenter l'information de comptage et les coefficients nuls sont représentés en gris pour qu'on puisse bien les distinguer. On dispose maintenant d'outils statistiques pour analyser les données à notre disposition. La section suivante présente l'analyse faite sur nos données, à savoir les deux matrices de la section 1.2.2.

1.4 Résultats sur les données du projet Plantadiv

Dans cette section, on présente l'analyse des données du projet Plantadiv et les différents résultats obtenus. Celle-ci va suivre le même enchaînement que la section précédente. Pour commencer, on va d'abord s'intéresser à la matrice de présence/absence. Après avoir vérifié que le modèle nul est bien rejeté par le test, on va réaliser une bi-classification avec un modèle LBM à lois de Bernoulli. On discutera alors des deux classifications obtenues. On pourra notamment se poser la question de savoir si les groupes s'expliquent exclusivement par les degrés. Pour répondre à cette question, on va réaliser une nouvelle bi-classification avec le modèle à degré corrigé que l'on pourra comparer à la première bi-classification. Puis, on utilisera la matrice de comptage pour obtenir une troisième bi-classification avec un modèle LBM à lois de Poisson. Finalement, on résumera l'information délivrée par chacune des bi-classification et on discutera des intérêts des différents modèles, dans notre cas-ci.

1.4.1 Test du modèle nul

En premier lieu, on vérifie que le modèle nul est bien rejeté par le test sur les variances empiriques des degrés. On représente la matrice de présence/absence déjà vue en section 1.2.2.

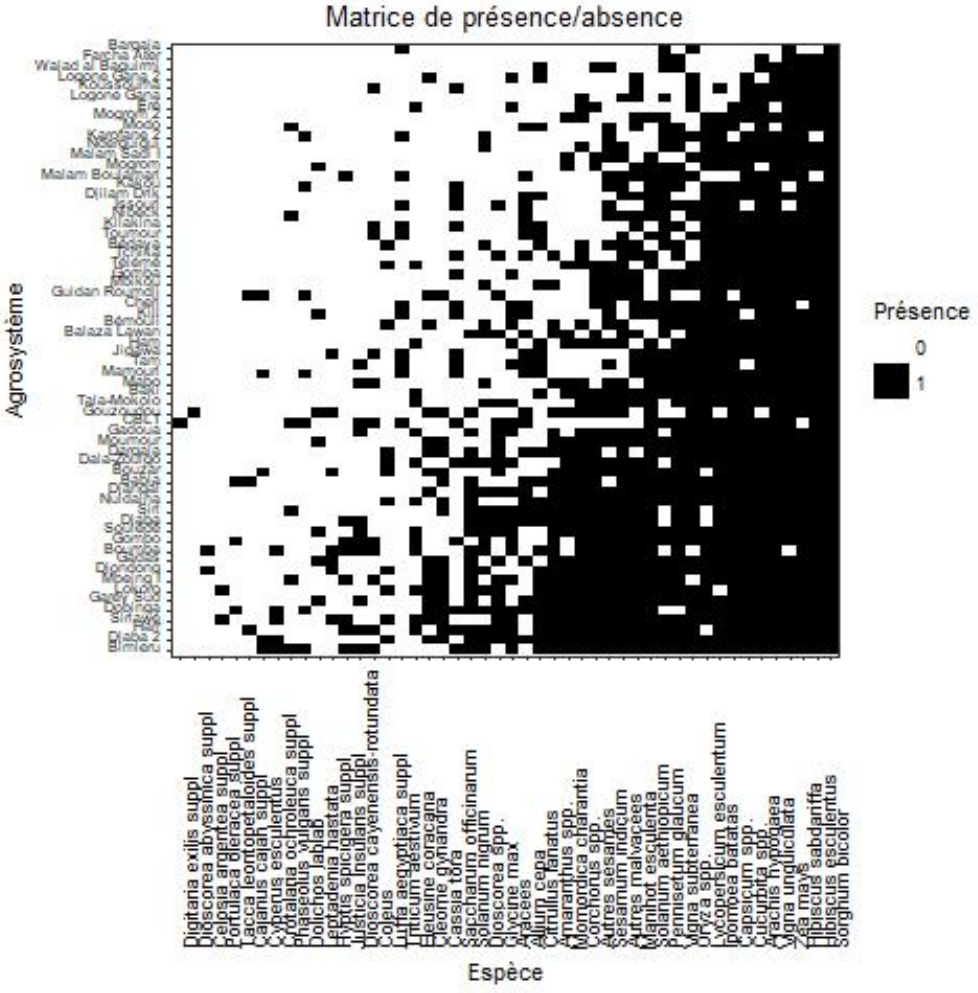


FIGURE 1.7

On réalise 2 000 simulations pour estimer les p-valeurs du test. On obtient les quatre p-valeurs estimées suivantes (dans l'ordre $p\text{val}_{L,1}, p\text{val}_{R,1}, p\text{val}_{L,2}, p\text{val}_{R,2}$ comme c'est présenté en annexe) : 0.9995, 0.0005, 1.0000 et 0.0000. On obtient deux p-valeurs estimées inférieures à 0.05 donc on rejette l'hypothèse du modèle nul. Pour plus d'explications sur ce test, elles se trouvent en annexe ou on peut directement se référer à [1]. L'hypothèse du modèle nul étant rejetée, cela met en évidence que le degré n'est pas distribué de façon uniforme au sein des espèces ainsi qu'au sein des agrosystèmes. Par conséquent, il est pertinent de prolonger la recherche d'une structure dans la matrice. On va appliquer les différentes méthodes de bi-classification présentées en section 3, en commençant avec l'inférence d'un modèle LBM à lois de Bernoulli.

1.4.2 Bi-classification pour la matrice de présence/absence

Après avoir estimé un modèle LBM à lois de Bernoulli sur notre matrice, on a pu assigner un groupe à chaque agrosystème et à chaque espèce. En regroupant les espèces et les agrosystèmes par groupes on obtient la matrice de la figure 1.8. Les groupes sont séparés par des lignes et on a ajouté des couleurs pour distinguer les groupes d'agrosystèmes. On peut trouver en annexe les probabilités a posteriori qui ont permis d'attribuer un groupe à chaque agrosystème.

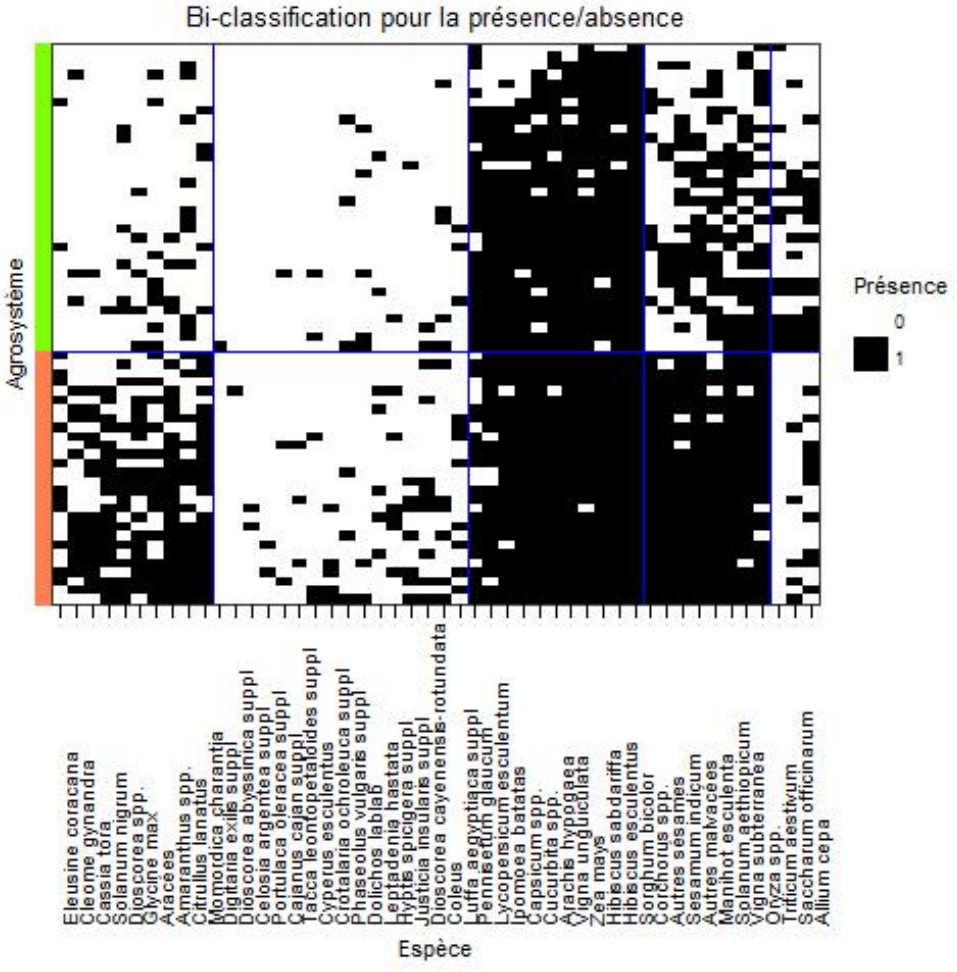


FIGURE 1.8

Le modèle estimé a permis de faire une bi-classification avec 2 groupes d'agrosystèmes et 5 groupes d'espèces. On numérote les groupes d'espèces de 1 à 5 en allant de gauche à droite. On utilisera la même convention pour les autres bi-classification par la suite. On parlera de « densité » pour décrire nos résultats. Par exemple, la densité globale de la matrice est de 0.45. Cela veut dire que 45% des cellules de la matrice contiennent un 1.
Ici, les groupes d'espèces semblent principalement représenter des différences de fréquence avec des espèces couramment utilisées (comme le groupe 3 avec une densité de 0.90) et d'autres moins couramment utilisées (comme le groupe 2, avec une densité de 0.09). Trois autres groupes sont détectés avec des densités intermédiaires (0.36 pour le groupe 1, 0.70 pour le groupe 4 et 0.30 pour le groupe 5). Les groupes 1 et 4 semblent correspondre à des espèces plutôt cultivées dans le groupe d'agrosystèmes orange. Le groupe 5 correspond lui à des espèces majoritairement cultivées dans le groupe d'agrosystème vert. Concernant les groupes d'agrosystèmes, on observe une fréquence plus importante pour le groupe orange qui se voit surtout pour le quatrième

groupe d'espèces. On a une densité de 0.56 pour le groupe orange et une densité de 0.35 pour le groupe vert.

Groupes d'agrosystèmes

On va maintenant s'intéresser plus en détail aux groupes d'agrosystèmes. On a obtenu deux groupes d'agrosystèmes dont la répartition est représentée sur la carte de la figure 1.9.

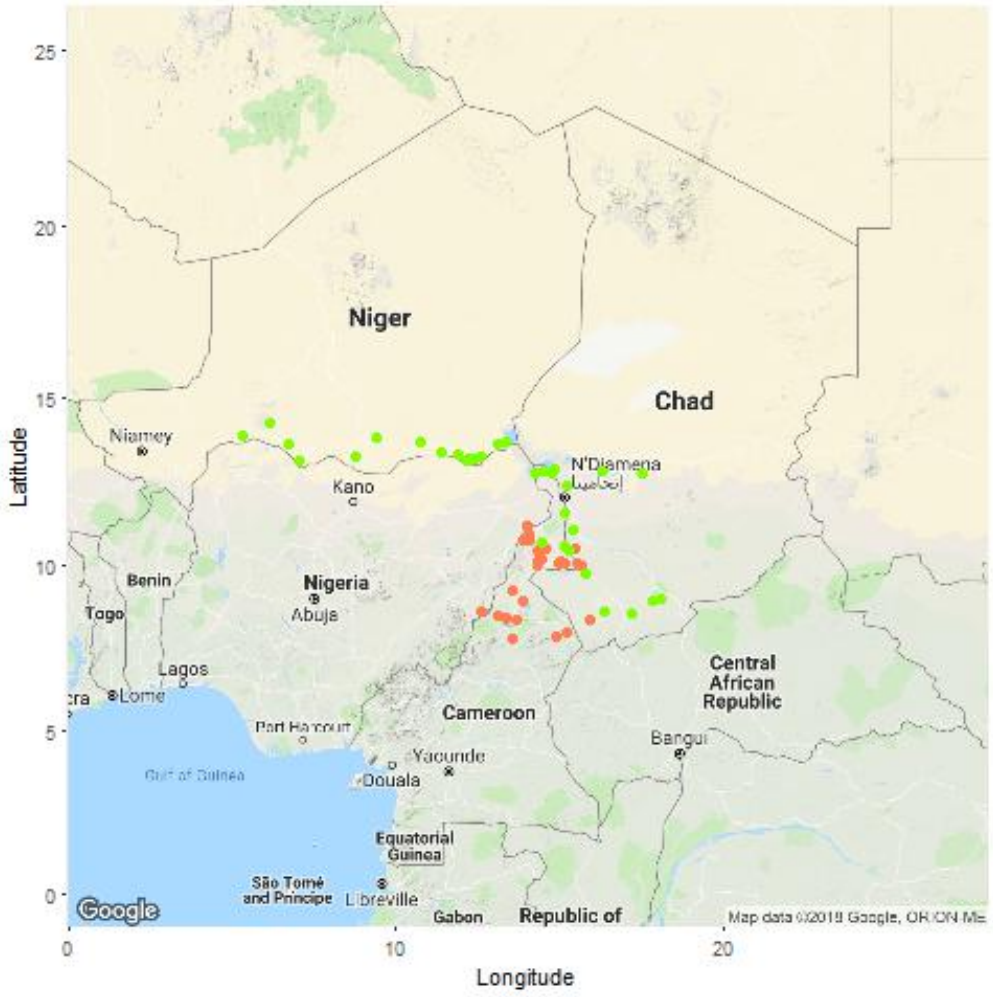


FIGURE 1.9

On peut tout d'abord remarquer qu'il y a bien une réalité géographique derrière les groupes obtenus à partir de la biodiversité cultivée ce qui est rassurant et peut motiver un peu plus l'utilisation de la méthode appliquée ici. Comme on peut le voir, on a une séparation qui est assez proche de la frontière camerounaise. Cette séparation s'explique d'abord par la distinction Nord/Sud entre zone soudanaise et zone sahélienne. On peut voir que même au niveau de la frontière entre le Cameroun et le Tchad, les deux groupes séparent assez bien le bec de canard des zones inondables le long du fleuve Logone.
Remarque : pour toutes les classifications obtenues sur les agrosystèmes, deux agrosystèmes d'un même village appartiennent toujours au même groupe (voir annexe).

Groupes d'espèces

On va discuter un peu plus de la classification sur les espèces. La figure 1.10 montre toujours la bi-classification mais cette fois on a ajouté les couleurs associées aux catégories d'espèces.

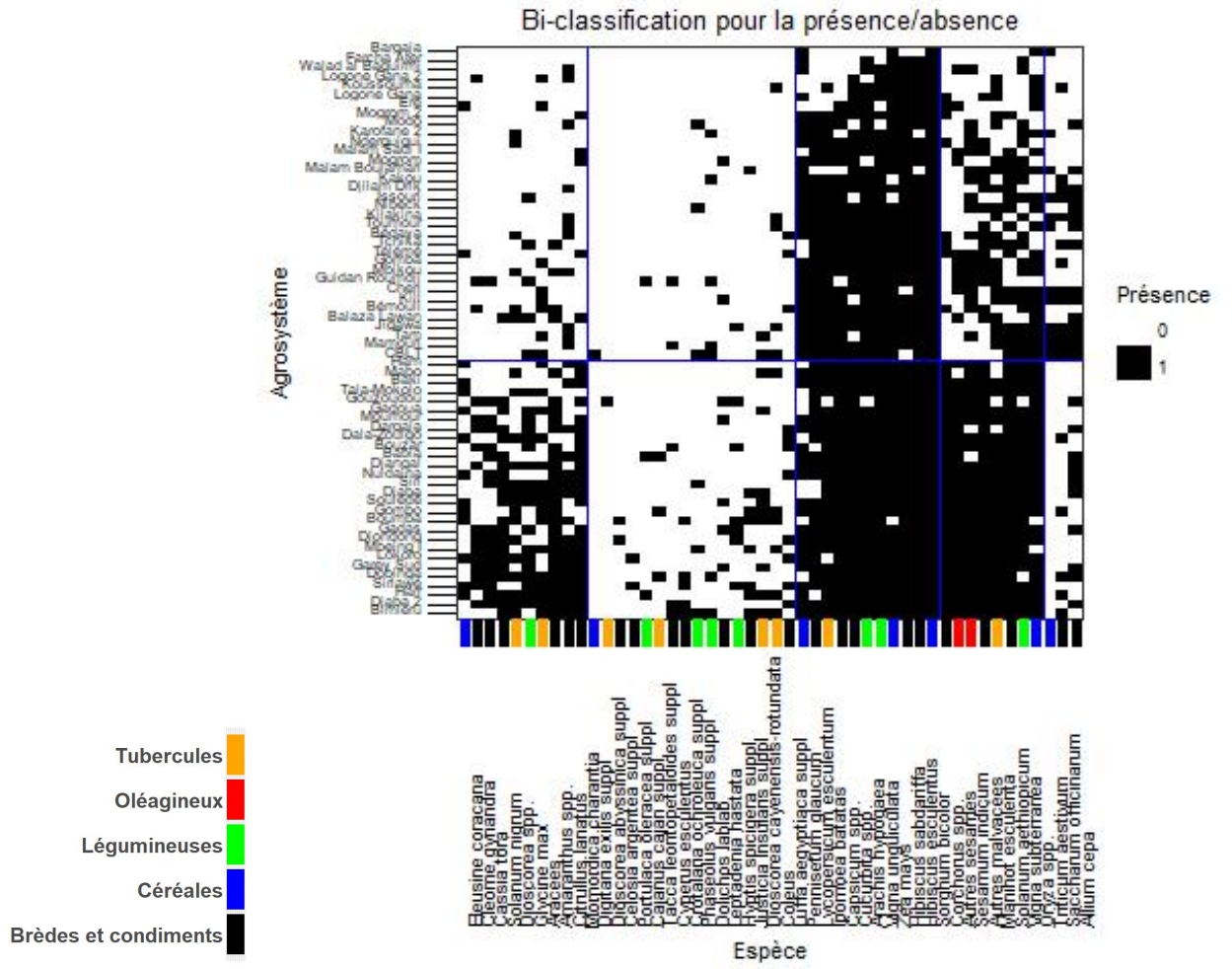


FIGURE 1.10

Après avoir vu la répartition géographique des deux groupes d'agrosystèmes, on peut mieux interpréter les groupes d'espèces. Tout d'abord, le groupe 3 ne distingue pas les groupes d'agrosystèmes, qui cultivent tous deux ces espèces en abondance. On peut raisonnablement estimer que ce groupe d'espèces correspond aux plantes les plus communes de la région. Le groupe 2 ne les distingue pas plus puisque ces espèces sont très peu cultivées dans les deux groupes d'agrosystèmes. À l'inverse, ce sont les plantes les plus rares de la région. Le groupe 1 rassemble des espèces assez peu cultivées qu'on peut rencontrer dans la zone soudanaise mais très rarement au Sahel. Le groupe 4 distingue assez nettement les deux groupes d'agrosystèmes. Les espèces font partie de la culture de base dans la zone soudanaise mais se font beaucoup plus rare au Nord. On distingue nettement le groupe 5 des autres. Il est assez spécifique puisque c'est le seul qui présente une densité plus important au Nord qu'au Sud. Il regroupe trois espèces qui sont le blé, la canne à sucre et l'ail. Pour chaque groupe d'espèces, on peut se demander si le fait que ses espèces sont plus cultivées dans l'un des groupes d'agrosystèmes à cause des contraintes géographiques et climatiques, ou si des agrosystèmes se sont spécialisés dans la culture de ces espèces, dans un objectif commercial par exemple. Naturellement, on a voulu tester la spécialisation des agrosystèmes par rapport aux catégories d'espèces. On a réalisé une bi-classification, c'est-à-dire que les groupes d'agrosystèmes et les groupes d'espèces ont été construits simultanément les uns par rapport aux autres. Pour vérifier s'il y a cette spécialisation des agrosystèmes, il faut regarder comment sont liés catégories et groupes d'espèces. Pour visualiser cela, on utilise la figure 1.11. La largeur des bandes représente les différents effectifs pour chaque couple de catégorie et d'espèce. À première vue, on peut émettre l'hypothèse que les catégories et les groupes obtenus sont indépendants. Autrement dit, il n'y a pas de spécialisation des cultures au sein d'un groupe d'agrosystèmes par rapport aux catégories

d'espèces.

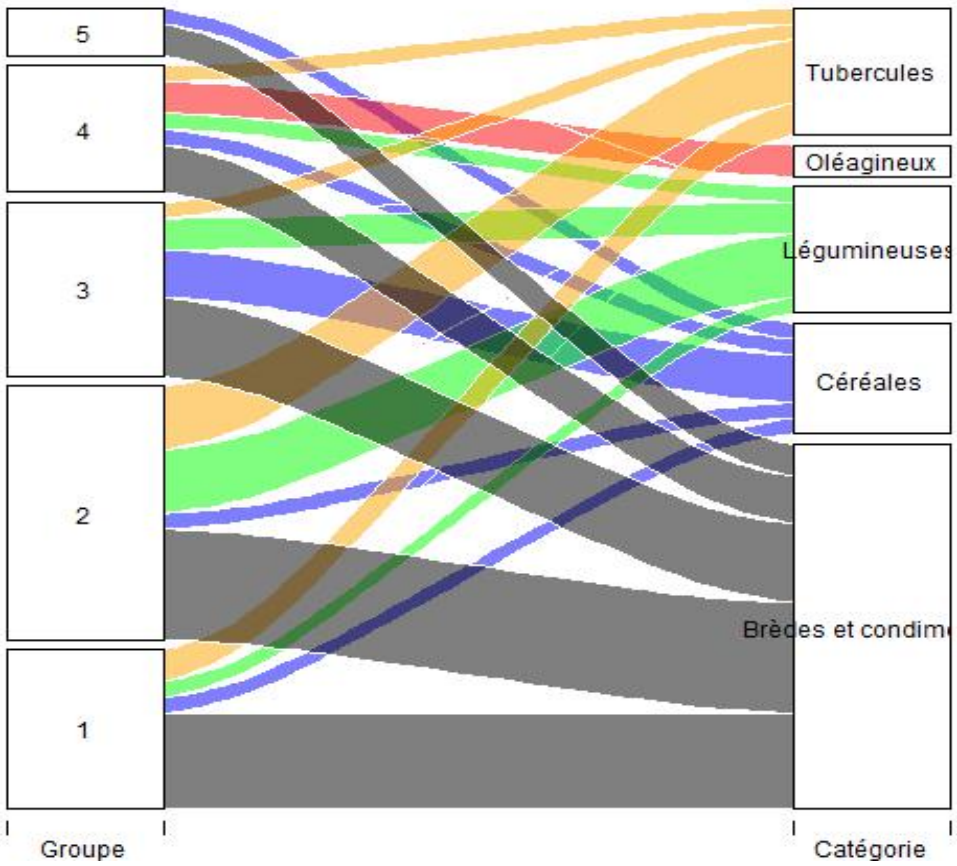


FIGURE 1.11

Pour confirmer l'intuition qu'il n'y pas de spécialisation, on effectue un test d'indépendance du χ^2 sur ces deux partitions des espèces. On obtient une p-valeur de 0.3939 (avec une statistique de test de $\chi^2 = 16.873$ pour 16 degrés de liberté) donc l'hypothèse d'indépendance n'est pas rejetée. On considère donc qu'il n'y a pas de spécialisation par rapport aux catégories.
Globalement, on a l'impression que les différents groupes d'espèces se distinguent par l'abondance de ces espèces dans les différents agrosystèmes. Les espèces sont différenciées par leur degré, par le fait d'être peu ou beaucoup cultivées plus que par le fait d'être cultivées dans certains agrosystèmes plutôt que dans d'autres. Cette remarque est aussi valable pour les groupes d'agrosystèmes. Le groupe orange cultive un nombre important d'espèces alors que le groupe vert en cultivent beaucoup moins. Pour voir si les espèces et les agrosystèmes se distinguent au-delà de leur degré, on va utiliser un modèle LBM à degré corrigé pour réaliser une nouvelle bi-classification.

1.4.3 Bi-classification pour la présence/absence avec degré corrigé

Pour essayer de gommer l'importance (supposée) des degrés dans le modèle précédent, on va appliquer un modèle LBM à degré corrigé. Si les groupes du premier modèle s'expliquent uniquement par des différences de degré, alors le nouveau modèle ne devrait plus détecter différents groupes et on aurait juste 1 groupe d'espèces et 1 groupe d'agrosystèmes. Le modèle LBM à degré corrigé donne la bi-classification de la figure 1.12.

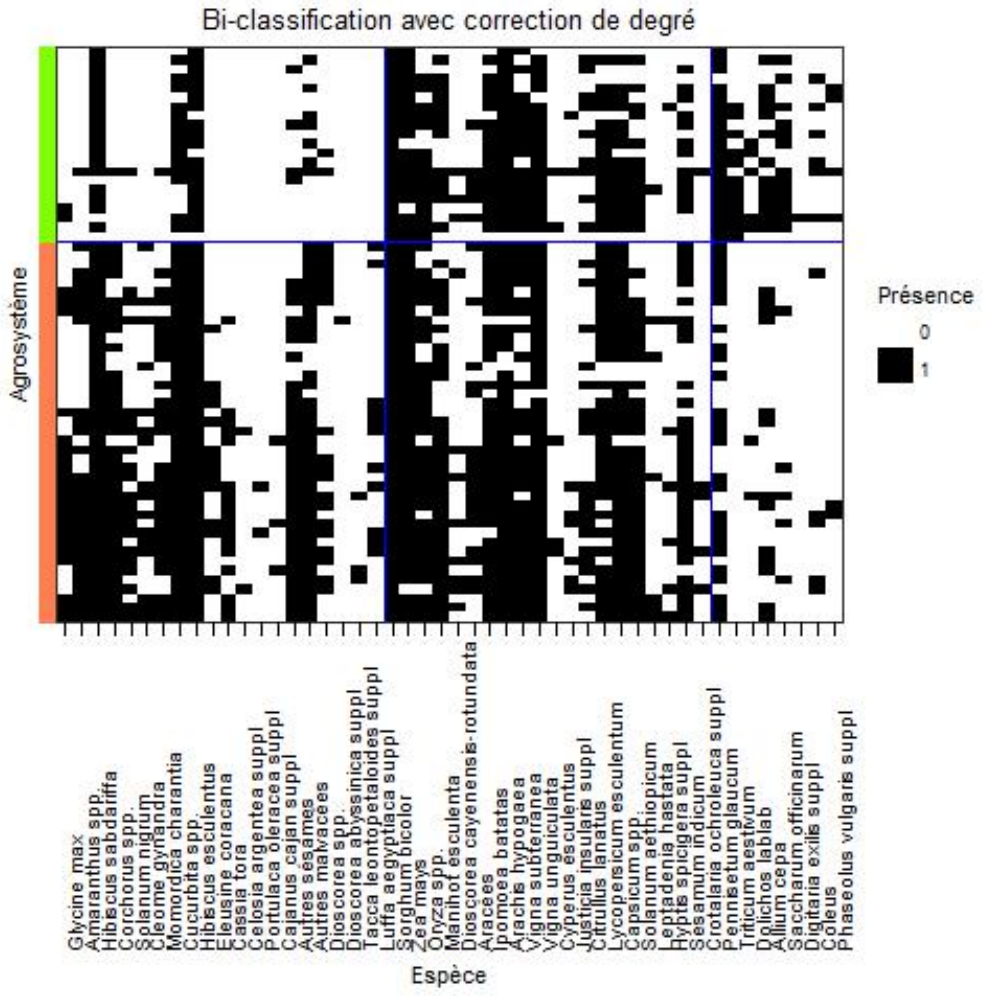


FIGURE 1.12

On obtient toujours deux groupes d'agrosystèmes mais seulement trois groupes d'espèces. On peut d'abord remarquer que les degrés ne semblent pas avoir d'importance dans cette nouvelle bi-classification, ou en tout cas beaucoup moins qu'avant. En effet, les densités semblent beaucoup plus proches que dans le modèle précédent. Le groupe d'agrosystème orange a une densité de 0.49 alors que le groupe vert a une densité de 0.37. Pour les espèces, les groupes 1, 2 et 3 ont des densités respectives de 0.39, 0.58 et 0.26.
On peut remarquer que les densité varient très peu autour de la densité globale (0.45). Cela montre bien que l'importance des degrés dans la constitution des groupes a été gommée par rapport au modèle précédent. On ne peut plus interpréter les groupes juste en observant les degrés. Il vaut mieux regarder des proportions par rapport au degré. Par exemple, au sein du groupe d'agrosystème vert, le groupe d'espèce 1 fait partie des cultures secondaires dans ces agrosystèmes alors que le groupe 3 fait partie des cultures principales. Il se distingue du groupe orange puisque la situation est inversée dans ces agrosystèmes. En revanche on retrouve quand même des agrosystèmes avec des degrés importants et des degrés faibles dans les deux groupes.

Groupes d'agrosystèmes

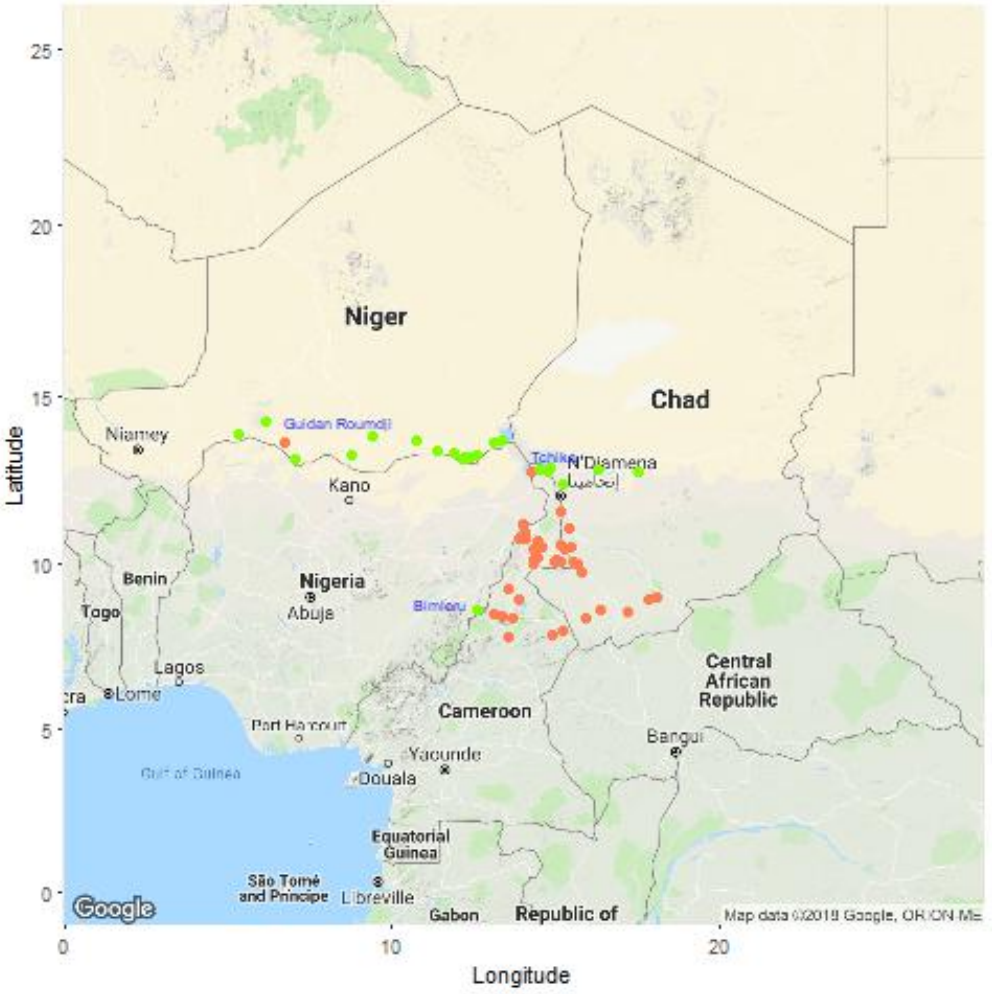


FIGURE 1.13

La carte de la figure 1.13 donne la répartition spatiale des deux groupes d'agrosystèmes. On peut voir que les groupes d'agrosystèmes sont assez similaires à ceux du modèle précédent, ce qui permet de faire plus simplement la comparaison. Là encore, le groupe orange est principalement constitué d'agrosystèmes situés aux Cameroun. On peut remarquer que les agrosystèmes qui sont passés du groupe vert au groupe orange se trouvent surtout dans la zone inondable de la vallée du Logoné et dans la plaine soudanienne du Tchad. Cela s'expliquerait par le fait qu'ils cultivent peu d'espèces différentes mais que la composition des cultures ressemble plus à ce qu'on trouve au Cameroun que dans la zone sahélienne. Dans cette classification, le groupe vert représente essentiellement le Sahel. Trois agrosystèmes se distinguent (comme des points aberrants) puisqu'ils se retrouvent dans le « mauvais » groupe d'un point de vue géographique. On a ajouté le nom de ces trois agrosystèmes (Guidan Roumdji, Tchika et Bimleru) sur la carte pour pouvoir les identifier.

Groupes d'espèces

De nouveau, on a représenté la bi-classification avec des couleurs pour les catégories d'espèces.

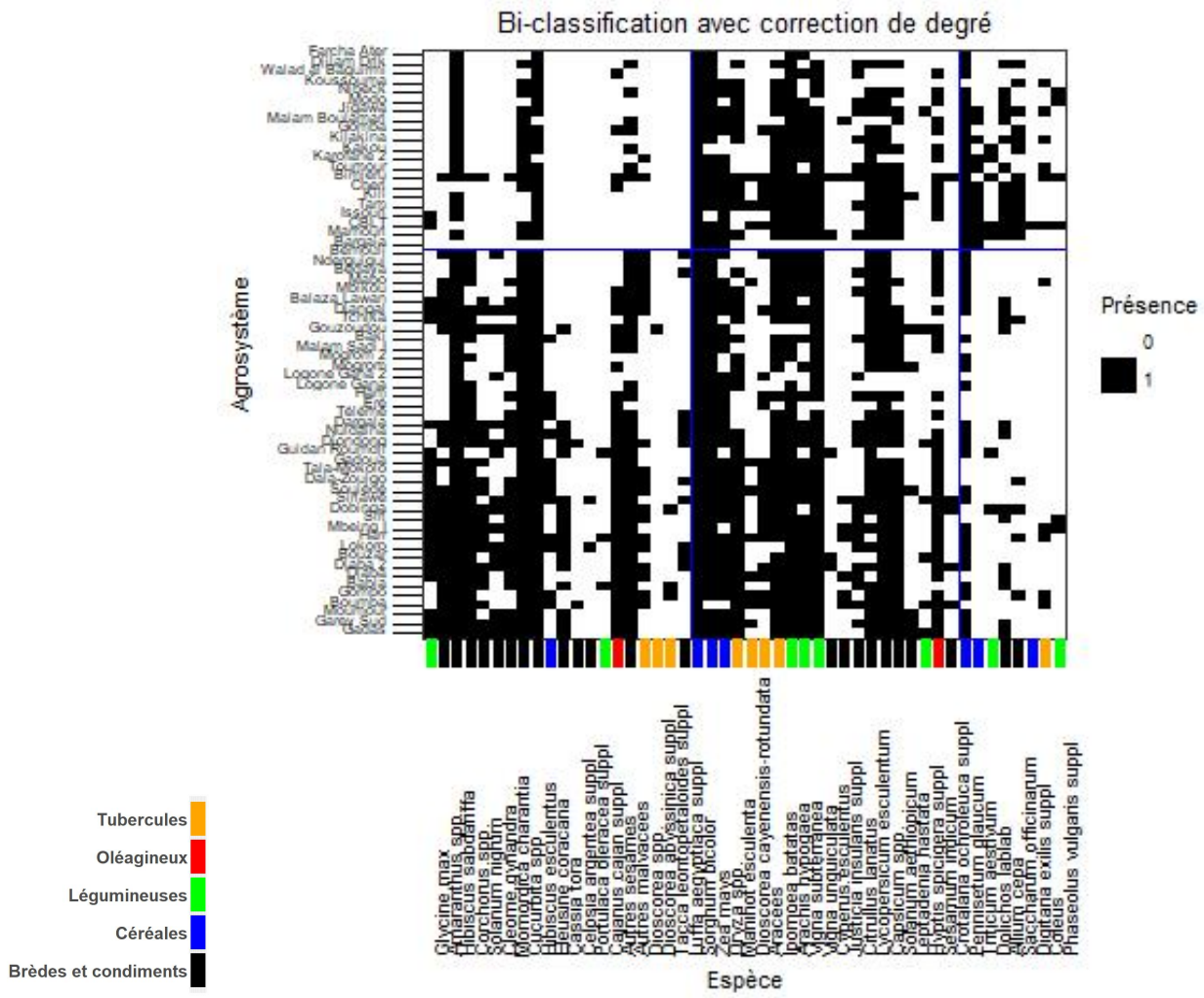


FIGURE 1.14

Les groupes pour le modèle à degré corrigé ne s'expliquent pas comme pour le premier modèle, surtout pour les espèces. En effet, juste en regardant la figure 1.14 on voit qu'il y a des espèces peu cultivées et des espèces fortement cultivées dans chacun des trois groupes. On peut tout même interpréter assez simplement les trois groupes d'espèces. Le groupe 1 rassemble des espèces qui sont plus communes au Sud (groupe orange), tout de même cultivées au Nord mais où elles sont plus marginales. Le groupe 2 correspond aux espèces qui sont cultivées au Nord et au Sud dans les mêmes proportions. Les espèces du groupe 3 sont celles que l'on retrouve le plus souvent au Nord et qui sont plutôt secondaires au Sud. À nouveau, il ne semble pas y avoir de spécialisation par rapport aux catégories à la vue de la figure 1.14. On peut vérifier cette intuition avec la figure 1.15.

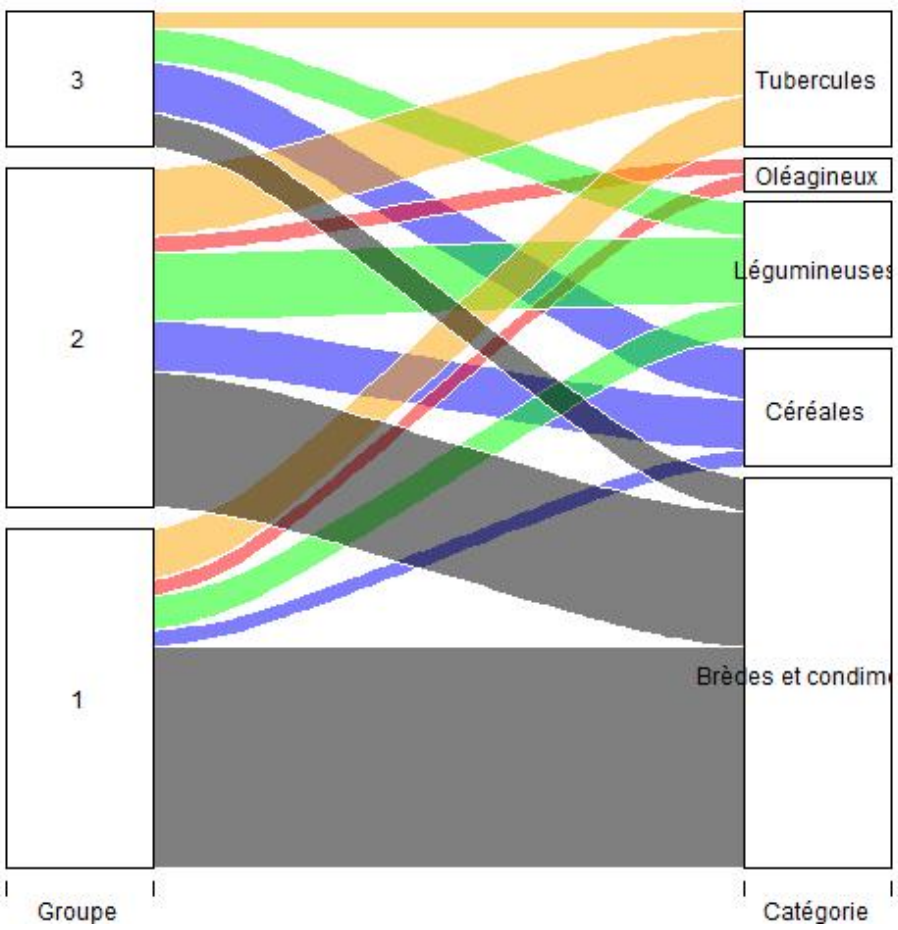


FIGURE 1.15

Excepté pour les oléagineux qui ne contiennent que deux espèces, chaque catégorie est représentée dans les trois groupes. Cela va dans le sens de notre première intuition que l'on peut confirmer avec un test d'indépendance du χ^2. On obtient une p-valeur de 0.419 (avec une statistique de test de $\chi^2 = 8.1494$ pour 8 degrés de liberté). De nouveau, le test du χ^2 ne permet pas de rejeter l'hypothèse d'indépendance. On ne peut pas dire que les groupes d'agrosystèmes ainsi définis soient spécialisés (par rapport aux catégories d'espèces).

On a vu avec le modèle LBM à degré corrigé qu'on pouvait observer certaines choses que ne détectait pas le modèle à lois de Bernoulli. La correction de degré permet de donner de l'importance à la composition des cultures pour créer les groupes d'agrosystèmes, plutôt qu'a leur richesse spécifique (nombre d'espèces cultivées). Dans notre cas, cela a permis d'obtenir une bi-classification simple à interpréter surtout avec le nombre de groupes assez faible. Cette bi-classification fait autre chose que séparer les agrosystèmes riches des agrosystèmes pauvres, et les espèces communes des espèces rares, ce qui semble être le cas lorsqu'on caricature les résultats du premier modèle. Cependant, c'est assez frustrant de ne pas obtenir une bi-classification plus fine. Notamment pour les groupes d'agrosystèmes qui sont encore au nombre de 2 et qui s'expliquent assez intuitivement par les différences climatiques entre zone sahélienne et zone soudanaise. Voyons si on obtient les mêmes résultats lorsque nous utilisons une information plus riche. À la place des données de présence/absence d'une espèce dans un agrosystème donné, nous allons maintenant considérer le nombre de variétés cultivées par espèce et par agrosystème avec la matrice de comptage C. On va donc utiliser un modèle LBM à lois de Poisson.

1.4.4 Bi-classification pour la diversité variétale

On considère la matrice de comptage représentée par la figure 1.2 en section 1.2.2. On va établir une bi-classification à partir de cette matrice toujours par un modèle LBM mais cette fois avec des lois de Poisson. De nouveau, les espèces et les agrosystèmes sont réunis par groupes, qui sont séparés par des lignes sur la matrice ci-dessous.

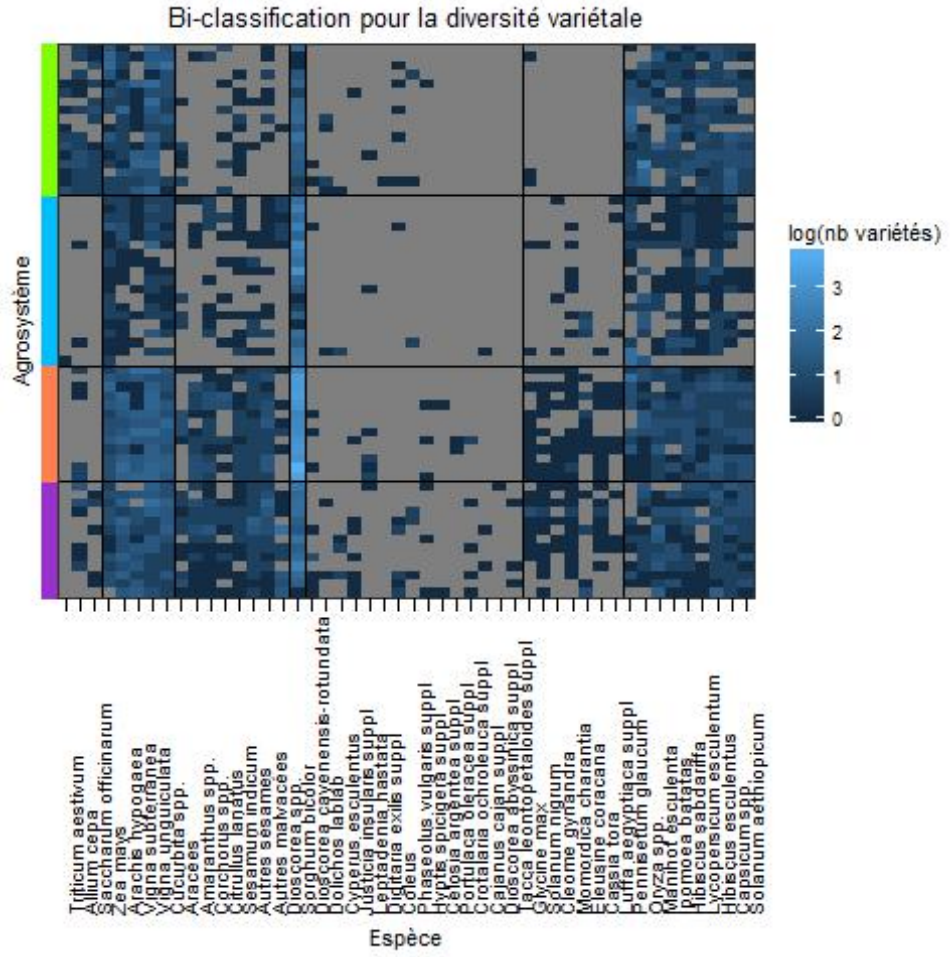


FIGURE 1.16

Pour cette nouvelle bi-classification, on a 4 groupes d'agrosystèmes et 7 groupes d'espèces. On parlera ici aussi de densité mais cela signifiera le nombre moyen de variétés cultivées par espèce et par agrosystème. On va d'abord décrire brièvement les groupes d'espèces. Le groupe 1 composé du blé, de l'ail et de la canne à sucre est le plus singulier. Sa densité est de 0.53. Il est caractéristique du groupe d'agrosystèmes vert où il est quasi-exclusivement cultivé. Le groupe 2, comme le groupe 7, rassemblent des espèces communes qui sont fortement cultivées un peu partout avec des densités respectives de 2.62 et de 1.76. Le groupe 4 est composé d'une seule espèce : le sorgho. Cela s'explique par le fait que le sorgho est cultivé partout avec un nombre de variétés très important (jusqu'à 48) même par rapport à d'autres espèces communes. Il a une densité très importante qui est de 10.2. Le groupe 5 rassemble les espèces rares qui sont très peu cultivées (avec une densité de 0.1) quel que soit le groupe d'agrosystèmes. Les groupes 3 et 6 sont semblables avec des espèces qui sont cultivées principalement dans les groupes d'agrosystèmes orange et violet. Mais les espèces du groupe 6 restent globalement plus rares que celles du groupe 3 (densité de 0.91 pour le groupe 3 et et 0.34 pour le groupe 6).
On décrit maintenant les groupes d'agrosystèmes. Les groupes orange et violet semblent cultiver les mêmes groupes d'espèces et ont des densités semblables (1.28 pour le groupe violet et 1.74 pour le groupe orange).

Il semblerait que ce qui les différencie principalement soit la diversité variétale pour le sorgho, qui est plus importante pour le groupe orange avec en moyenne 24.1 variétés de sorgho cultivées contre 6.85 variétés en moyenne pour le groupe violet. Le groupe bleu semble être le groupe le plus « pauvre », qui cultive le moins d'espèces avec la plus faible diversité variétale (densité de 0.64). Le groupe vert cultive un peu plus d'espèces que le groupe bleu, sa densité est de 0.93. Mais il est surtout caractérisé par le premier groupe d'espèces puisqu'il le cultive très fortement (avec une densité de 1.37) et que ces espèces sont très peu cultivées ailleurs (densités de 0.05, 0.21 et 0.46 respectivement pour les groupe bleu,orange et violet). On peut remarquer que ce premier groupe d'espèces est composé de trois espèces (blé, canne à sucre et ail) qui constituaient déjà un groupe dans la bi-classification du modèle à lois de Bernoulli.

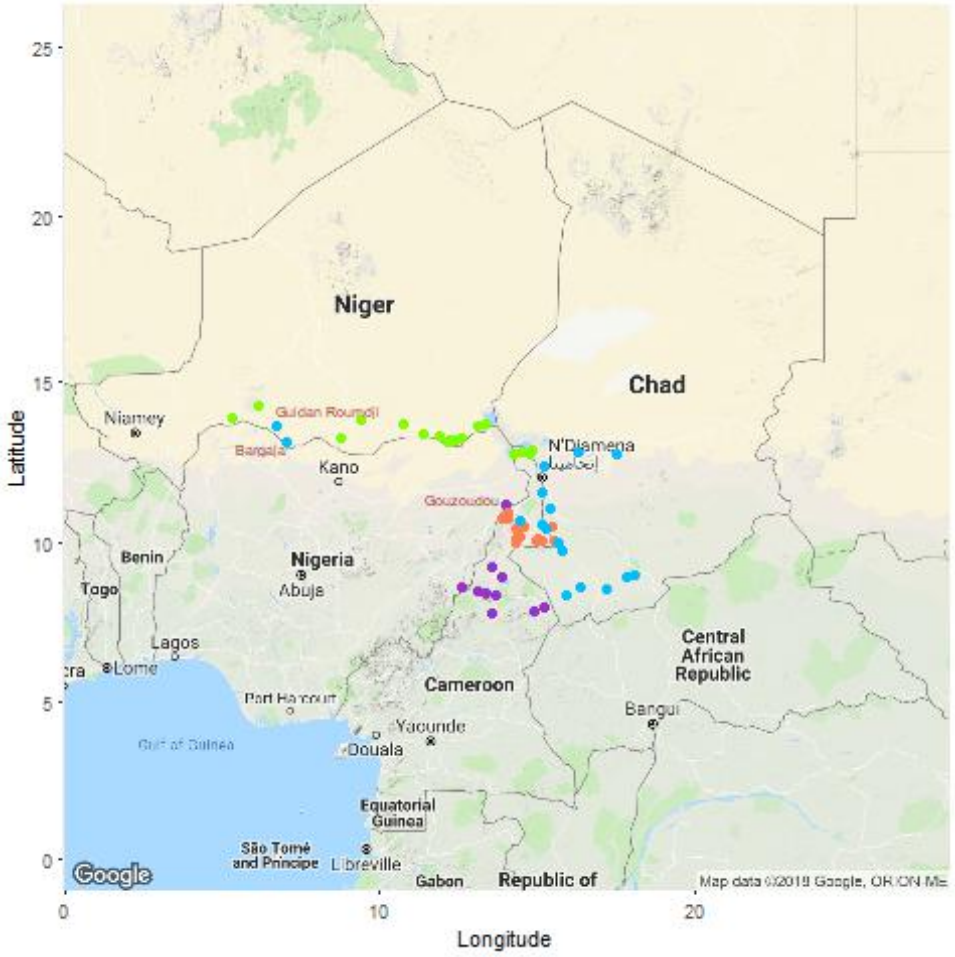


FIGURE 1.17

À nouveau, on peut s'apercevoir que les groupes sont cohérents spatialement. On va décrire ces groupes avec les ensembles géographiques donnés en section 1.2.4. Le groupe orange est essentiellement constitué des agrosystèmes du bec de canard et des monts Mandara. Grossièrement, il regroupe les agrosystèmes se trouvant dans le Nord du Cameroun alors que le groupe violet contient le reste du Cameroun, c'est-à-dire la plaine camerounaise et le massif d'Adamaoua. Le groupe vert représente principalement le Sahel, et contient également les agrosystèmes de la vallée du fleuve Komadougou Yobé. Finalement, le groupe bleu comprend la plaine soudanienne du Tchad et la zone inondable du fleuve Logone. En caricaturant, le groupe vert représente la zone sahélienne du Niger et le groupe bleu les agrosystèmes du Tchad. On peut à nouveau distinguer trois « agrosystèmes aberrants » d'un point de vue géographique (Bargaja, Guldan Roumdji et Gouzoudou), indiqués sur la figure 1.17. On a indiqué le nom de ces trois agrosystèmes sur la carte pour pouvoir les identifier.

Groupes d'espèces

Cette fois-ci, on ne va pas discuter des groupes d'espèces notamment en les comparant aux catégories d'espèces comme avant. En effet, il ne serait pas judicieux d'adopter la même démarche qu'avant au vu du nombre de groupes et des effectifs (le sorgho constitue un groupe à lui tout seul par exemple). On peut tout de même commenter brièvement la classification sur les espèces donnée par le modèle. Il semblerait qu'on ait deux informations qui ressortent principalement de cette classification. La première est la place centrale qu'occupe le sorgho dans les cultures des agrosystèmes échantillonés. La deuxième est qu'il existe un groupe de trois espèces, déjà présent dans la première classification, qui est spécifique du groupe d'agrosystèmes vert et qui témoigne de la spécialisation de ces agrosystèmes puisque ces espèces ne sont que très peu cultivées ailleurs.

1.4.5 Comparaison des différents modèles LBM

Dans cette partie, on va brièvement comparer les classifications d'agrosystèmes obtenues avec les différents modèles. On ne peut pas vraiment comparer les méthodes puisqu'on a analysé un seul jeu de données. On compare uniquement ces méthodes à travers les différents résultats obtenus précédemment. La figure 1.18 montre comment les agrosystèmes se situent dans chaque classification.

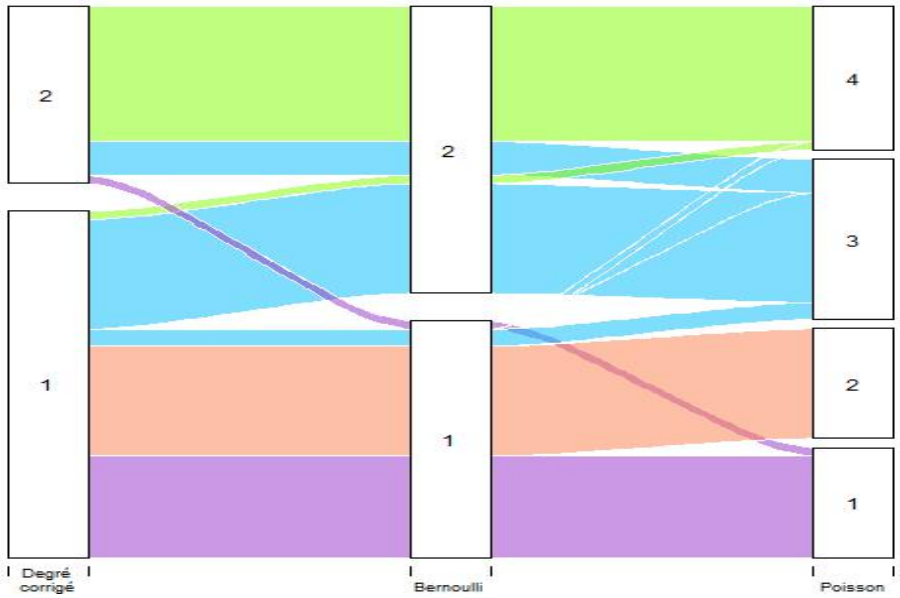


FIGURE 1.18

On peut voir que les classifications sur les agrosystèmes sont loin d'être indépendantes. La diversité variétale permet ici d'avoir une classification plus fine, on a plus de discrimination qu'avec la présence/absence (que ce soit avec ou sans correction des degrés). On peut voir que les groupes violet et orange du modèle LBM à lois de Poisson se retrouvent dans le groupe 1 (orange) des deux autres modèles à l'exception d'un agrosystème (Gouzoudou) qui se retrouve dans le groupe 2 (vert) pour le modèle LBM à degré corrigé. De la même manière, le groupe vert du modèle LBM à lois de Poisson se retrouve toujours dans le groupe 2 (vert) des deux autres modèles à l'exception d'un agrosystème (Tchika). En revanche, il est plus compliqué d'interpréter aussi simplement le groupe bleu sur la figure 1.18. On peut tout de même dire qu'il est principalement constitué des agrosystèmes qui sont passés du groupe 2 (vert) au groupe 1 (orange) entre le modèle LBM à lois de Bernoulli et le modèle à degré corrigé. À la vue de la figure 1.18, il semblerait que les modèles à lois de Bernoulli et à degré corrigé contiennent moins d'information que le modèle à lois de Poisson. Ces deux premiers modèles donnent des résultats plus grossiers mais probablement plus simples à interpréter.

1.5 Conclusion

L'objectif initial de ce stage de quatre mois était de réaliser une analyse statistique de la biodiversité cultivée dans une certaine région d'Afrique subsaharienne à partir de données d'inventaire de culture, notamment afin de comprendre sa répartition géographique. L'analyse présentée dans ce rapport permet globalement de répondre aux différentes questions que l'on s'est posé. On peut regretter que l'analyse ne soit pas plus poussée. Néanmoins les résultats présentés sont intéressants et donnent quelques réponses.
Tout d'abord, on a vu que les regroupements des agrosystèmes à partir de leur composition en espèces respecte une structure spatiale qui correspond aux zones sahélienne et sub-sahélienne. D'autre part, on a pu se rendre compte de la faible spécialisation des agrosystèmes (par rapport aux catégories qu'on trouve dans la base de données). Finalement, l'utilisation de l'information de diversité variétale a permis d'obtenir une classification plus fine des agrosystèmes. Les résultats obtenus mettent en évidence certaines spécificités géographiques qu'on ne voyait pas avant, comme les différences importantes de diversité variétale pour le sorgho (qui est cultivé partout).
Au delà des résultats obtenus avec les données, ce rapport présente également les méthodes utilisées. Ces méthodes peuvent être appliquées pour analyser d'autres données d'inventaire de culture. Pour faciliter cela, j'ai fait en sorte que la plus grande partie possible du code que j'ai produit soit directement réutilisable. J'ai rassemblées les différentes fonctions créées dans un fichier *R* qui est accompagné d'une aide (fichier *html*) présentant des exemples d'utilisation. Cela devrait permettre, avec le rapport, de faciliter l'utilisation des méthodes statistiques utilisées ici et notamment pour des ethnologues.

Bibliographie

[1] Mathieu Thomas, Nicolas Verzelen, Pierre Barbillon, Oliver T. Coomes, Sophie Caillon, Doyle McKey, Marianne Elias, Eric Garine, Christine Raimond, Edmond Dounias, Devra Jarvis, Jean Wencélius, Christian Leclerc, Vanesse Labeyrie, Pham Hung Cuong, Nguyen Thi Ngoc Hue, Bhuwon Sthapit, Ram Bahadur Rana, Adeline Barnaud, Chloé Violon, Luis Manuel Arias Reyes, Luis Latournerie Moreno, Paola De Santis, François Massol. *A Network-Based Method to Detect Patterns of Local Crop Biodiversity : Validation at the Species and Infra-Species Levels.* Advances in Ecological Research, 2015, Volume 53, pages 259-320.

[2] Notes de cours de Stéphane Robin. *Models with Hidden Structure.* `https://www6.inra.fr/mia-paris/content/download/4587/42934/version/1/file/ModelsHiddenStruct-Biology.pdf`

[3] Jean-Benoist Léger. *Blockmodels : A R-package for estimating in LBM and SBM, with many pdf, with or without covariates.* ArXiV :1602.07587v1.

[4] Mahendra Mariadassou, Stéphane Robin, Corinne Vacher. *Uncovering Latent Structure in Valued Graphs : a Variational Approach.* The Annals of Applied Statistics, 2010, Vol. 4, No. 2, 715-742.

[5] Nicholas Gotelli, Gary Graves. *Null Models in Ecology.* Smithsonian Institution Press, Washington, D.C. 1996.

[6] Jing Lei and Alessandro Rinaldo. *Consistency of Spectral Clsutering in Stochastic Block Models.* The Annals of Statistics, 2015, Vol. 43, No. 1, 215-237.

[7] Arthur Dempster, Nan Laird, Donald Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, 1977, Series B. 39 (1) : 1-38.

2 Détection de ruptures dans un espace de Hilbert à noyaux reproduisants

2.1 Introduction

La détection de ruptures est la recherche de fortes discontinuités, de « ruptures », dans la distribution d'une série temporelle. Les procédures de détections de ruptures ont des applications dans différents domaines comme en traitement de signal audio [6], en neurosciences [5] ou encore en sécurité informatique [7]. Le cadre classique de la détection de ruptures considère une série temporelle dont la loi de distribution est constante par morceaux. L'objectif est généralement d'estimer le nombre de ruptures d'une part, et d'autre part de localiser les instants de ces ruptures. La grande majorité des méthodes utilisées fait l'hypothèse que les ruptures se trouvent dans la moyenne ou la variance de la série temporelle considéré. Récemment, Arlot, Célisse et Harchaoui [2] ont généralisé ces méthodes à des variations arbitraires de distribution en s'appuyant sur l'astuce du noyau. Plus précisément, Arlot *et al.* proposent de plonger la série temporelle dans un espace de Hilbert à noyau reproduisant (RKHS) [3]. Si le noyau est bien choisi, des changements de distribution de la série temporelle dans l'espace de départ conduisent à des changements de moyennes dans le RKHS.
Dans [1], Garreau et Arlot démontre un résultat sur la méthode développée dans [2]. Ils utilisent un critère des moindres carrés pénalisé pour obtenir un résultat non-asymptotique sur l'estimation du nombre de ruptures et de leur localisation, qui reste valable en grande dimension. Le théorème 3.1 de Garreau et Arlot garantit l'existence d'un évènement de grande probabilité sur lequel, pour un paramètre de pénalisation bien choisi, le nombre de ruptures et leur localisation sont correctement estimés. Cependant, on n'a pas de garantie d'existence de constante de pénalisation adaptée et le résultat dépend de paramètres dépendants de τ^* et μ^* et qui sont donc inconnus. Notre résultat s'adapte mieux dans le sens où il ne dépend pas de paramètres inconnus. Plutôt que de vouloir estimer le nombre exact de sauts, on montre qu'on estime avec une certaine précision les sauts de haute énergie et qu'on ne surestime pas le nombre de sauts. On s'autorise donc à ne pas estimer les sauts de faible énergie d'une part. Et d'autre part, les sauts de hautes énergies sont estimés avec une meilleure précision que dans [1]. De plus, on n'a pas de condition dépendant de K (paramètre inconnu) dans notre cas, pour la calibration de la constante de pénalisation.

2.2 Modèle et estimateur

La plupart des notations que l'on trouve dans la suite sont analogues à celles de Garreau et Arlot [1], avec toutefois des différences dans les conventions utilisées.

2.2.1 Définitions et Modèle

L'approche développée dans [2] et [1] considère une série temporelle $X_1, \dots, X_n$ de n variables aléatoires indépendantes à valeurs dans un espace mesurable $\mathcal{X}$, et on suppose qu'on dispose d'un noyau semi-défini positif k sur $\mathcal{X}$. On utilise alors le noyau k pour déplacer notre problème de l'espace $\mathcal{X}$ vers un espace de Hilbert $\mathcal{H}$. On effectue la transformation suivante des observations

$$Y_i := k(X_i, \cdot) \in \mathcal{H}, \forall i \in \{1, \dots, n\},$$

où $\mathcal{H}$ est l'espace de Hilbert à noyau reproduisant associé à k. Ici, on ne va pas détailler cette étape largement décrite dans [1] et [2]. On va directement considérer une série temporelle dans un espace de Hilbert $\mathcal{H}$.
Soit $n \in \mathbb{N}$. Soient n variables aléatoires indépendantes $Y_1, \dots, Y_n$ à valeurs dans un espace de Hilbert $\mathcal{H}$. On notera $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ et $\|\cdot\|_{\mathcal{H}}$ le produit scalaire et la norme associés. Le vecteur $L = (\mathcal{L}(Y_1), \dots, \mathcal{L}(Y_n))$ est constant par morceaux ($\mathcal{L}(Y_i)$ désigne la loi de Y_i). On suppose que les Y_i sont bornés, donc intégrables. Formellement, on définit $\varepsilon_i := Y_i - \mathbb{E}[Y_i]$, $\theta_i^* := \mathbb{E}[Y_i]$ et on fait les hypothèses suivantes.

Hypothèses.

- *Les ε_i sont indépendants,*
- *et pour tout $i \in \{1, \dots, n\}$, on a $\|\varepsilon_i\|_{\mathcal{H}} \leq M$ presque sûrement.*

La première hypothèse traduit le fait que les variables Y_i sont indépendantes. Notons que les ε_i sont également centrés, par définition. La deuxième hypothèse nous permet d'obtenir un contrôle sur les déviations de combinaisons linéaires des ε_i, à l'image d'une hypothèse de sous-gaussianité dans un cadre plus « classique ».

2.2.2 Problème de détection de ruptures

On a supposé que $L = (\mathcal{L}(Y_1), \dots, \mathcal{L}(Y_n))$ est constant par morceaux donc $\theta^* = (\theta_1^*, \dots, \theta_n^*)$ est aussi constant par morceaux. Autrement dit, il existe $(K, \mu^*, \tau^*) \in \mathbb{N} \times \mathbb{R}^{K+1} \times [\![0, n+1]\!]$ tel que

$$\forall i \in \{1, \dots, n\}, \theta_i^* = \sum_{k=1}^{K+1} \mu_k^* 1_{i \in [\tau_{k-1}^*, \tau_k^*[} \tag{2.1}$$

et $1 = \tau_0^* < \tau_1^* < \cdots < \tau_K^* < \tau_{K+1}^* = n+1$. Dans un problème de détection de rupture, on ne dispose que d'une réalisation de $(Y_1, \dots, Y_n)$. L'objectif consiste alors à estimer le nombre et le lieux des discontinuité dans L, c'est-à-dire retrouver les paramètres K et τ^* qui sont inconnus. Parfois, on peut chercher à estimer μ^* mais ce n'est pas le cas ici. On peut par contre estimer μ^* très simplement a posteriori, à partir de l'estimation $\hat{\tau}$ de τ^*. Il s'agit alors d'estimer une moyenne sur les différents intervalles donnés par τ^*.

2.2.3 Notations

On fixe les notations K, τ^* et μ^* utilisées juste avant et on définit les notations suivantes.

- $\Delta_k^* := \mu_k^* - \mu_{k-1}^*$ est la hauteur du k-ième saut.
- $\mathcal{T}$ désigne l'ensemble des segmentations τ telles que $1 = \tau_0 < \tau_1 < \cdots < \tau_m < \tau_m = n+1$. On note alors $D_\tau := m$ le nombre de ruptures donné par la segmentation τ.
- $\overline{Y}_{j_1,j_2} := \frac{1}{j_2 - j_1} \sum_{i=j_1}^{j_2-1} Y_i$ est la moyenne de Y entre j_1 (inclus) et j_2 (exclu).
- $\overline{\theta^*}_{j_1,j_2} := \frac{1}{j_2 - j_1} \sum_{i=j_1}^{j_2-1} \theta_i^*$ est la moyenne de θ^* entre j_1 (inclus) et j_2 (exclu).
- $\overline{\varepsilon}_{j_1,j_2} := \frac{1}{j_2 - j_1} \sum_{i=j_1}^{j_2-1} \varepsilon_i$ est la moyenne de ε entre j_1 (inclus) et j_2 (exclu).
- $\tau^{-i} := (\tau_1, \dots, \tau_{i-1}, \tau_{i+1}, \dots, \tau_{D_\tau+1})$ est la segmentation avec les mêmes ruptures que τ exceptée la i-ème.

- De la même manière $\tau^{(k)}$ désigne la segmentation contenant les ruptures de τ et la rupture τ_k^*.

On note $\mathcal{T}_3 = \{t = (t_1, t_2, t_3) : 1 \leq t_1 < t_2 < t_3 \leq n+1\}$ l'ensemble des triplets distincts et ordonnés. Par abus de notation, on appellera parfois t un saut pour faire référence à t_2. Pour $1 \leq k \leq D_{\tau^*}$, on note

$$\mathcal{E}_k^* := \sqrt{\frac{(\tau_{k+1}^* - \tau_k^*)(\tau_k^* - \tau_{k-1}^*)}{(\tau_{k+1}^* - \tau_{k-1}^*)}} \|\Delta_k^*\|_{\mathcal{H}}. \tag{2.2}$$

Cette quantité est appelé énergie du saut $t = (\tau_{k-1}^*, \tau_k^*, \tau_{k+1}^*)$. On prend en compte la hauteur du saut $\|\Delta_k^*\|_{\mathcal{H}}$ mais également la longueur des segments avant et après le saut. Cela permet de mieux détecter un saut de hauteur faible, lorsqu'il est fortement distant de la rupture suivante ou de la rupture précédente. Il vaut donc mieux utiliser l'énergie d'un saut plutôt que sa hauteur pour le caractériser. On verra par la suite que cette quantité apparaît naturellement dans notre analyse, notamment lorsqu'on compare une segmentation τ avec une segementatoin pour laquelle on a inséré ($\tau^{(k)}$) ou retiré (τ^{-i}) une rupture.

2.2.4 Sélection de modèle et estimateur

On introduit le critère de Arlot *et al.* qui est une adaptation du critère des moindres carrés pénalisés. On utilisera ce critère avec une pénalité légèrement différente de [1] pour estimer τ^*. Pour une segmentation $\tau \in \mathcal{T}$, on note F_τ l'ensemble des applications $\{1, \dots, n\} \to \mathcal{H}$ qui sont constantes sur les segments de τ. Pour f dans F_τ, on note $\Pi_\tau f$ la projection orthogonale de f sur F_τ. Parmi les segmentations de même dimension, on choisira celle qui minimise le critère des moindres carrées, c'est-à-dire $\tau \in \mathcal{T}$ qui minimise la quantité suivante

$$\widehat{R}_n(\tau) = \|Y - \Pi_\tau Y\|_{\mathcal{H}^n}^2 = \|Y\|_{\mathcal{H}^n}^2 - \|\Pi_\tau Y\|_{\mathcal{H}^n}^2$$

avec $\|Y\|_{\mathcal{H}^n}^2 = \sum_{i=1}^n \|Y_i\|_{\mathcal{H}}^2$. On considère un critère des moindres carrées pénalisées

$$\widehat{\tau} \in \underset{\tau \in \mathcal{T}_n}{\operatorname{argmin}} \{\operatorname{crit}(\tau) = \widehat{R}_n(\tau) + \operatorname{pen}(\tau)\}.$$

On utilise la pénalité suivante

$$\operatorname{pen}(\tau) = A D_\tau \ln(n)$$

où A est une constante strictement positive. À travers les différentes preuves, on pourra comprendre d'où vient ce choix de pénalité et quelles conditions sur la constante A sont requises pour obtenir les différents résultats avec ce critère

$$\operatorname{crit}(\tau) = \|Y - \Pi_\tau Y\|_{\mathcal{H}^n}^2 + A D_\tau \ln(n). \tag{2.3}$$

Remarque. *Pour $\tau \in \mathcal{T}$, on dispose d'une formule explicite pour $\Pi_\tau(\cdot)$. Soit $f \in \mathcal{H}^n$ et $\tau_{l-1} \leq i < \tau_l$. On a* $(\Pi_\tau f)_i = \frac{1}{\tau_l - \tau_{l-1}} \sum_{j=\tau_{l-1}}^{\tau_l - 1} f_j = \overline{f}_{\tau_{l-1}, \tau_l}$.

2.3 Résultat principal

Avant d'énoncer le principal résultat, on définit la notion d'énergie pour tout $t \in \mathcal{T}_3$. L'énergie du saut t est définie par

$$\mathcal{E}(t) := \sqrt{\frac{(t_3 - t_2)(t_2 - t_1)}{(t_3 - t_1)}} \left\| \overline{\theta^*}_{t_2, t_3} - \overline{\theta^*}_{t_1, t_2} \right\|_{\mathcal{H}}. \tag{2.4}$$

Théorème 1. *Pour $x > 0$, pour $L \geq \frac{M^2}{\ln(n)} \left[1 + \sqrt{2\left(\ln\left(\frac{n^3}{6}\right) + x\right)}\right]^2$ et pour $\sqrt{A} > 22\sqrt{L}$, il existe un évènement Ω de probabilité supérieure à $1 - e^{-x}$ tel que :*

- *on ne détecte pas trop de ruptures*

$$\forall k, |\hat{\tau} \cap \left[(\tau_{k-1}^* + \tau_k^*)/2, (\tau_k^* + \tau_{k+1}^*)/2\right]| \leq 1,$$

- *les sauts de haute énergie sont « bien détectés »*

$$\forall k, \mathcal{E}_k^* > 6(\sqrt{A}+\sqrt{L})\sqrt{\ln(n)} \Rightarrow \min_l |\hat{\tau}_l - \tau_k^*| \leq \left(\frac{6(\sqrt{A}+\sqrt{L})\sqrt{\ln(n)}}{\mathcal{E}_k^*}\right)^2 \min\left\{\frac{\tau_k^* - \tau_{k-1}^*}{2}, \frac{\tau_{k+1}^* - \tau_k^*}{2}\right\},$$

- *et les sauts « non détectés » sont de basses énergies*

$$\tau_k^* \notin \hat{\tau} \Rightarrow \mathcal{E}(\hat{\tau}_l, \tau_k^*, \hat{\tau}_{l+1}) \leq (\sqrt{A}+\sqrt{L})\sqrt{\ln(n)}.$$

À la différence de Garreau et Arlot, notre résultat ne donne rien sur l'estimation exacte du nombre de ruptures tel qu'on l'entend dans [1], c'est-à-dire pour τ^* tel que $\mu_i^* \neq \mu_{i+1}^*$ pour tout $i \in \{1, \dots, D_{\tau^*}\}$. Le théorème 1 garantit tout de même qu'on ne surestime pas le nombre de ruptures dans le sens où $D_{\hat{\tau}} \leq D_{\tau^*}$. Dans notre cas, on s'autorise à ne pas détecter les sauts de basse énergie. En revanche, on garantit bien que les sauts de haute énergie sont détecter avec une précision d'autant plus importante que l'énerge du saut est grande. Le fait d'« ignorer » les sauts de basse énergie permet d'avoir un résultat plus clair, qui ne fait pas d'hypothèse sur le couple (τ^*, μ^*), qui ne fait pas intervenir de constantes dépendantes de ce couple. C'est un avantage par rapport au théorème 3.1 de [1] qui fait intervenir certaines de ces constantes à la fois dans les hypothèses sur la constante de pénalité et dans le résultat. De ce fait, leur résultat ne garantit pas l'existence de constante de pénalisation adaptée (d'un point de vue non-asymptotique). D'un point de vue asymptotique, on a également une précision de l'ordre de $\ln(n)/n$ pour la distance $d_\infty^{(1)}(\tau^*, \hat{\tau})/n$ utilisée dans [1]. L'énergie d'un saut $\mathcal{E}_k^*$ est proportionelle à $\sqrt{n}$ donc tous les sauts sont d'énergie assez haute pour être détecté, pour n assez grand et à condition de choisir les constantes A et L correctement. Pour améliorer le résultat du théorème 1, on pourrait retravailler le contrôle de la probabilité de l'évènement Ω. En utilisant la méthode de chaînage et une inégalité de concentration/déviation plus fine, cela devrait permettre d'améliorer en particulier le lemme 2.

2.4 Ébauche de preuve

On définit d'abord l'évènement Ω par une condition sur nos erreurs $(\varepsilon_i)_i$. On contrôle alors la probabilité de cet évènement via une inégalité de concentration de type Hoeffding pour des variables aléatoires à valeurs dans un espace de Hilbert. Ensuite on démontre différents lemmes (sur Ω) qui reposent uniquement sur des raisonnements déterministes.

2.4.1 Définition de l'évènement Ω

On veut prouver que la procédure de sélection de modèle évoquée précédemment est efficace, c'est-à-dire qu'on estime suffisament bien le vrai nombre de ruptures et le lieu des ruptures. Pour commencer, on peut attendre d'une « bonne » segmentation que le nombre de ruptures détectées soit assez proche du vrai nombre de ruptures. La pénalité doit permettre d'éviter que le nombre de ruptures soit surestimé de manière exagérée. Par exemple, on peut exiger de notre estimateur la condition suivante : qu'il soit impossible d'estimer 3 ruptures sur un même « vrai » segment, i.e. qu'on ne puisse pas avoir $\tau_{k-1}^* \leq \hat{\tau}_{i-1} < \hat{\tau}_i < \hat{\tau}_{i+1} \leq \tau_k^*$. En particulier, cet événement est réalisé si pour tout $k, \tau_{i-1}, \tau_i, \tau_{i+1}$ on a

$$\tau_{k-1}^* \leq \tau_{i-1} < \tau_i < \tau_{i+1} \leq \tau_k^* \Rightarrow \text{crit}(\tau) > \text{crit}(\tau^{-i}). \tag{2.5}$$

Lorsqu'on a $\text{crit}(\tau) > \text{crit}(\tau^{-i})$, cela veut dire que le critère est minimisé en retirant τ_i de la segmentation τ. Or par définition de $\hat{\tau}$, on a $\text{crit}(\hat{\tau}) \leq \text{crit}(\hat{\tau}^{-i})$ pour tout $i \in \{1, \dots, D_{\hat{\tau}}\}$. Par la contraposée de (2.5), on a alors

$$\forall k \in \{1, \dots, D_{\tau^*}+1\}, |\hat{\tau} \cap [\tau_{k-1}^*, \tau_k^*]| \leq 2.$$

Pour $t = (t_1, t_2, t_3) \in \mathcal{T}_3$, on pose

$$Z_t := \frac{(t_3 - t_2)(t_2 - t_1)}{t_3 - t_1} \|\overline{\varepsilon}_{t_2,t_3} - \overline{\varepsilon}_{t_1,t_2}\|_{\mathcal{H}}^2. \tag{2.6}$$

Cette quantité apparaît naturellement puisque dans le cas où $\tau^*_{k-1} \leq \tau_{i-1} < \tau_i < \tau_{i+1} \leq \tau^*_k$ on a en effet $\text{crit}(\tau) - \text{crit}(\tau^{-i}) = A\ln(n) - Z_{(\tau_{i-1},\tau_i,\tau_{i+1})}$. Si $\tau^*_{k-1} \leq \tau_{i-1} < \hat{\tau}_i < \tau_{i+1} \leq \tau^*_k$, on a donc bien

$$\text{crit}(\tau) \leq \text{crit}(\tau^{-i}) \Leftrightarrow Z_{(\tau_{i-1},\tau_i,\tau_{i+1})} \geq A\ln(n).$$

On introduit l'évènement suivant

$$\Omega := \left\{ \sup_{t \in \mathcal{T}_3} Z_t \leq L\ln(n) \right\} \tag{2.7}$$

avec $L < A$. Si l'on se restreint à cet évènement, la condition (2.5) qu'on demandait est toujours satisfaite. Le lemme 1 est donc une conséquence directe de (2.5) et (2.7).

Lemme 1. *Sur Ω, on a*

$$\forall k \in \{1, \dots, D_{\tau^*} + 1\}, |\hat{\tau} \cap [\tau^*_{k-1}, \tau^*_k]| \leq 2. \tag{2.8}$$

On a donc définit un évènement Ω qui donne une garantie par rapport à la surestimation du nombre de sauts. Avant de prouver des résultats plus fins sur la procédure de sélection de modèle sous Ω, on souhaite contrôler la probabilité de cet évènement. En utilisant une inégalité de concentration de type Hoeffding et une borne d'union, on obtient le résultat suivant.

Lemme 2. *Pour $L > M^2/\ln(n)$, on a*

$$\mathbb{P}(\Omega) \geq 1 - \frac{1}{6}\exp\left(3\ln(n) - \frac{1}{2}\left(\frac{\sqrt{L\ln(n)}}{M} - 1\right)^2\right). \tag{2.9}$$

On a donc un évènement Ω dont la probabilité est contrôlée. On a déjà prouvé la propriété (2.8) sur Ω. La suite de la preuve du théorème est purement déterministe. On montrera que sous Ω on a de meilleurs résultats que le lemme 1, qui permettent d'obtenir le théorème 1 sous certaines conditions sur les constantes A et L. On va d'abord utiliser des lemmes techniques pour simplifier les preuves dont les raisonnements peuvent être redondants.

2.4.2 Lemmes techniques

On va utiliser plusieurs fois les deux lemmes suivants dont les preuves se trouvent en section 2.6.

Lemme 3. *Pour toute segmentation $\tau \in \mathcal{T}$, on a :*

$$\forall i \in \{1, \dots, D_\tau\}, crit(\tau^{-i}) - crit(\tau) = \frac{(\tau_{i+1} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}} \|\overline{Y}_{\tau_i,\tau_{i+1}} - \overline{Y}_{\tau_{i-1},\tau_i}\|^2_{\mathcal{H}} - A\ln(n)$$

et de la même manière

$$crit(\tau) - crit(\tau^{(k)}) = \frac{(\tau_{l+1} - \tau^*_k)(\tau^*_k - \tau_l)}{\tau_{l+1} - \tau_l} \|\overline{Y}_{\tau^*_k,\tau_{l+1}} - \overline{Y}_{\tau_l,\tau^*_k}\|^2_{\mathcal{H}} - A\ln(n)$$

pour $k \in \{1, \dots, D_{\tau^*}\}$ et l tels que $\tau_l < \tau^*_k < \tau_{l+1}$.

On va exclusivement travailler avec des transformations locales de segmentations donc on va être constamment confronté à des quantités de la forme

$$\frac{(t_3 - t_2)(t_2 - t_1)}{t_3 - t_1} \|\overline{Y}_{t_1,t_2} - \overline{Y}_{t_2,t_3}\|^2_{\mathcal{H}}$$

et qu'on va vouloir contrôler (minorer ou majorer). Pour cela, on utilisera régulièrement le lemme suivant dans les différentes preuves.

Lemme 4. *Pour $t \in \mathcal{T}_3$, on a*

$$\mathcal{E}(t) + \sqrt{Z_t} \geq \sqrt{\frac{(\tau_{t_3} - \tau_{t_2})(\tau_{t_2} - \tau_{t_1})}{\tau_{t_3} - \tau_{t_1}}} \|\overline{Y}_{t_1,t_2} - \overline{Y}_{t_2,t_3}\|_{\mathcal{H}} \geq |\mathcal{E}(t) - \sqrt{Z_t}|. \tag{2.10}$$

2.5 Raisonnements déterministes sur Ω

Cette partie regroupe les lemmes qui vont composer le résultat principal. Leurs preuves sont toutes des raisonnements par l'absurde et déterministe sur Ω. Elles sont adaptées d'une analyse du problème de détection de ruptures dans le cadre sous-gaussien [8] et se trouvent avec les autres preuves en section 2.6.

Lemme 5. *Sur Ω, pour $k \in \{1, \dots, D_{\tau^*}\}$ et l tels que $\tau_l < \tau_k^* < \tau_{l+1}$, on a*

$$\tau_k^* \notin \hat{\tau} \Rightarrow \mathcal{E}(\tau_l, \tau_k^*, \tau_{l+1}) \leq (\sqrt{A} + \sqrt{L})\sqrt{\ln(n)}.$$

On peut interpréter ce résultat en disant qu'une rupture de haute énergie sera toujours « bien estimée ». Si $\tau_k^* \notin \hat{\tau}$ et que τ_k^* correspond à un saut de haute énergie dans τ^*, alors c'est que τ_l ou τ_{l+1} estime assez bien ce saut. Le lemme suivant vient « confirmer » cette interprétation.

Lemme 6. *Sur Ω et pour $k \in \{1, \dots, D_{\tau^*}\}$, si $\mathcal{E}_k^* > \kappa\sqrt{\ln(n)}$ alors*

$$\min_l |\hat{\tau}_l - \tau_k^*| \leq \min\left[\frac{\tau_{k+1}^* - \tau_k^*}{2}, \frac{\tau_k^* - \tau_{k-1}^*}{2}\right],$$

avec

$$\kappa = \max\left\{6(\sqrt{A} + \sqrt{L}), \frac{3\sqrt{3}}{2}(\sqrt{2A} + \sqrt{L}), (\sqrt{3A} + \sqrt{L})\right\} = 6(\sqrt{A} + \sqrt{L}).$$

On peut encore améliorer ce résultat en liant la précision de l'estimation de τ_k^* à l'énergie du saut $\mathcal{E}_k^*$, ce que fait le corollaire suivant.

Corollaire 1. *Pour $k \in \{1, \dots, D_{\tau^*}\}$ tel que $\mathcal{E}_k^* > \kappa\sqrt{\ln(n)}$, on a*

$$\min_l |\hat{\tau}_l - \tau_k^*| \leq \left(\frac{\kappa\sqrt{\ln(n)}}{\mathcal{E}_k^*}\right)^2 \min\left[\frac{\tau_{k+1}^* - \tau_k^*}{2}, \frac{\tau_k^* - \tau_{k-1}^*}{2}\right].$$

On sait maintenant que les sauts de haute énergie sont bien estimés. On aimerait être sûr qu'on n'estime pas trop de ruptures, la condition (2.5) utilisée pour définir Ω n'est pas entièrement satisfaisante car elle implique uniquement $D_{\hat{\tau}} \leq 2D_{\tau^*}$. Les deux lemmes suivants donnent une condition sur la constante A de la pénalité pour garantir un nombre de ruptures estimées correct, c'est-à-dire $D_{\hat{\tau}} \leq D_{\tau^*}$.

Lemme 7. *Pour $\sqrt{A} > 6\sqrt{L}$, on a*

$$\forall k \in \{0, \dots, D_{\tau^*}\}, |\hat{\tau} \cup [\tau_k^*, (\tau_k^* + \tau_{k+1}^*)/2]| \leq 1$$

et

$$\forall k \in \{1, \dots, D_{\tau^*} + 1\}, |\hat{\tau} \cup [(\tau_{k-1}^* + \tau_k^*)/2, \tau_k^*]| \leq 1.$$

Le lemme 7 ne semble pas vraiment améliorer significativement le résultat du lemme 1 mais c'est une étape préliminaire au résultat suivant qui lui est pleinement satisfaisant.

Lemme 8. *Pour $\sqrt{A} \geq 22\sqrt{L}$, on a*

$$\forall k \in \{1, \dots, D_{\tau^*}\}, |\hat{\tau} \cup [(\tau_{k-1}^* + \tau_k^*)/2, (\tau_k^* + \tau_{k+1}^*)/2]| \leq 1.$$

Ces deux lemmes garantissent que le nombre de sauts estimé est inférieur ou égal au nombre réel de ruptures, i.e. $D_{\hat{\tau}} \leq D_{\tau^*}$.

2.6 Preuves des lemmes

2.6.1 Preuve du lemme 2

On utilise l'inégalité de McDiarmid [4] dont le résultat suivant est un cas particulier.

Propriété 1.
Soit $X_1, \dots, X_n$ des variables aléatoires centrées à valeurs dans un espace de Hilbert et $c_1, \dots, c_n$ tel que $\forall i, \|X_i\| \leq c_i/2$ presque sûrement. Alors, pour tout $t \geq \sqrt{v}$, on a

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n} X_i\right\|_{\mathcal{H}} > t\right) \leq \exp\left(-\frac{(t-\phi)^2}{2v}\right),$$

avec $v := \frac{1}{4}\sum_{i=1}^{n} c_i^2$ *et* $\phi = \mathbb{E}\left[\left\|\sum_{i=1}^{n} X_i\right\|_{\mathcal{H}}\right]$.

Par hypothèse, les variables $(\varepsilon_i)_i$ sont bornées et par définition elles sont centrées. On peut donc appliquer l'inégalité ci-dessus à Z_t pour un certain $t \in \mathcal{T}_3$ en posant

$$X_i = \begin{cases} \frac{1}{t_2-t_1}\varepsilon_i \text{ pour } t_1 \leq i < t_2 \\ -\frac{1}{t_3-t_2}\varepsilon_i \text{ pour } t_2 \leq i < t_3 \end{cases}$$

On a bien $\left\|\sum_i X_i\right\|_{\mathcal{H}} = Z_t$. De plus, l'hypothèse d'indépendance permet d'avoir ici $\phi \leq \sqrt{v} = M\sqrt{\frac{t_3-t_1}{(t_3-t_2)(t_2-t_1)}}$.
On en déduit alors le lemme suivant.

Lemme 9. *Pour $t \in \mathcal{T}_3$ et $L \geq M^2/\ln(n)$, on a*

$$\mathbb{P}\left(Z_t > L\ln(n)\right) \leq \exp\left(-\frac{1}{2}\left(\frac{\sqrt{L\ln(n)}}{M} - 1\right)^2\right).$$

Preuve du lemme 9.
En appliquant l'inégalité de McDiarmid comme décrit, on a $v = \psi M^2$ avec $\psi = \frac{(t_3-t_1)}{(t_3-t_2)(t_2-t_1)}$. On obtient alors,

$$\mathbb{P}\left(\|\overline{\varepsilon}_{t_2,t_3} - \overline{\varepsilon}_{t_1,t_2}\|_{\mathcal{H}} > t\right) \leq \exp\left(-\frac{(t-M\sqrt{\psi})^2}{2M^2\psi}\right).$$

En posant $t = uM\sqrt{\psi}$ avec $u > 1$, on a

$$\mathbb{P}\left(\|\overline{\varepsilon}_{t_2,t_3} - \overline{\varepsilon}_{t_1,t_2}\|_{\mathcal{H}} > uM\sqrt{\psi}\right) \leq \exp\left(-\frac{(u-1)^2}{2}\right).$$

Donc

$$\mathbb{P}\left(Z_{\tau_{i-1},\tau_i,\tau_{i+1}} > u^2M^2\right) \leq \exp\left(-\frac{(u-1)^2}{2}\right).$$

Si $u = \sqrt{L\ln(n)}/M > 1$, alors on a bien le résultat voulu. □

Le lemme précédent contrôle la probabilité de l'évènement $\{L\ln(n) < Z_t\}$ pour un $t \in \mathcal{T}_3$ fixé. Or on s'intéresse à l'évènement $\Omega = \{Z_t \leq L\ln(n), \forall t \in \mathcal{T}_3\}$.

Corollaire 2. *Pour L tel que $\sqrt{L\ln(n)} > M$, on a*

$$\mathbb{P}(\Omega) \geq 1 - \frac{1}{6}\exp\left(3\ln(n) - \frac{1}{2}\left(\frac{\sqrt{L\ln(n)}}{M} - 1\right)^2\right). \tag{2.9}$$

Preuve du corollaire 2.
On utilise une borne d'union pour étendre le résultat du lemme précédent à Ω.

$$\begin{aligned}
\mathbb{P}(\Omega) &= 1 - \mathbb{P}\left(\bigcup_{t\in\mathcal{T}_3} \{Z_t > L\ln(n)\}\right) \\
&\geq 1 - \sum_{t\in\mathcal{T}_3} \mathbb{P}(Z_t > L\ln(n)) \\
&\geq 1 - \binom{n}{3} \exp\left(-\frac{1}{2}\left(\frac{\sqrt{L\ln(n)}}{M} - 1\right)^2\right) \\
&\geq 1 - \frac{n^3}{6} \exp\left(-\frac{1}{2}\left(\frac{\sqrt{L\ln(n)}}{M} - 1\right)^2\right)
\end{aligned}$$

On a bien le résultat souhaité qui correspond également au lemme 2. □

2.6.2 Lemmes techniques

Preuve du lemme 3.
Pour $\tau \in \mathcal{T}$, on a $\mathrm{crit}(\tau) - \mathrm{crit}(\tau^{-i}) = A\ln(n) + \|\Pi_{\tau^{-i}}Y\|^2_{\mathcal{H}^n} - \|\Pi_\tau Y\|^2_{\mathcal{H}^n}$. Il suffit donc de prouver que

$$\|\Pi_\tau Y\|^2_{\mathcal{H}^n} - \|\Pi_{(\tau^{-i})}Y\|^2_{\mathcal{H}^n} = \frac{(\tau_{i+} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}} \|\overline{Y}_{\tau_{i-1},\tau_i} - \overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}}.$$

On a

$$\begin{aligned}
\|\Pi_\tau Y\|^2_{\mathcal{H}^n} - \|\Pi_{\tau^{-i}}Y\|^2_{\mathcal{H}^n} &= (\tau_{i+1} - \tau_i)\|\overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}} + (\tau_i - \tau_{i-1})\|\overline{Y}_{\tau_{i-1},\tau_i}\|^2_{\mathcal{H}} \\
&\quad - (\tau_{i+1} - \tau_{i-1})\|\overline{Y}_{\tau_{i-1},\tau_{i+1}}\|^2_{\mathcal{H}}.
\end{aligned}$$

En remarquant que $\overline{Y}_{\tau_{i-1},\tau_{i+1}} = \frac{\tau_{i+1}-\tau_i}{\tau_{i+1}-\tau_{i-1}}\overline{Y}_{\tau_i,\tau_{i+1}} + \frac{\tau_i-\tau_{i-1}}{\tau_{i+1}-\tau_{i-1}}\overline{Y}_{\tau_{i-1},\tau_i}$, on obtient

$$\begin{aligned}
\|\Pi_\tau Y\|^2_{\mathcal{H}^n} - \|\Pi_{(\tau^{-i})}Y\|^2_{\mathcal{H}^n} &= (\tau_{i+1} - \tau_i)\|\overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}} + (\tau_i - \tau_{i-1})\|\overline{Y}_{\tau_{i-1},\tau_i}\|^2_{\mathcal{H}} \\
&\quad - \frac{(\tau_{i+1} - \tau_i)^2}{\tau_{i+1} - \tau_{i-1}}\|\overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}} - \frac{(\tau_i - \tau_{i-1})^2}{\tau_{i+1} - \tau_{i-1}}\|\overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}} \\
&\quad - 2\frac{(\tau_{i+1} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}}\langle\overline{Y}\tau_i, \tau_{i+1}, \overline{Y}_{\tau_{i-1},\tau_i}\rangle_{\mathcal{H}} \\
&= \frac{(\tau_{i+1} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}}\left[\|\overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}} + \|\overline{Y}_{\tau_{i-1},\tau_i}\|^2_{\mathcal{H}}\right] \\
&\quad - 2\frac{(\tau_{i+1} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}}\langle\overline{Y}\tau_i, \tau_{i+1}, \overline{Y}_{\tau_{i-1},\tau_i}\rangle_{\mathcal{H}} \\
&= \frac{(\tau_{i+} - \tau_i)(\tau_i - \tau_{i-1})}{\tau_{i+1} - \tau_{i-1}}\|\overline{Y}_{\tau_{i-1},\tau_i} - \overline{Y}_{\tau_i,\tau_{i+1}}\|^2_{\mathcal{H}}
\end{aligned}$$

On obtient bien le résultat souhaité. Les calculs sont exactement les mêmes pour $\mathrm{crit}(\tau) - \mathrm{crit}(\tau^{(k)})$. □

Preuve du lemme 4.
Il suffit de remarquer que $\overline{Y}_{j_1,j_2} = \overline{\theta^*}_{j_1,j_2} + \overline{\varepsilon}_{j_1,j_2}$ et d'appliquer l'inégalité triangulaire pour obtenir

$$\mathcal{E}(t) + \sqrt{Z_t} \geq \sqrt{\frac{(\tau_{t_3} - \tau_{t_2})(\tau_{t_2} - \tau_{t_1})}{\tau_{t_3} - \tau_{t_1}}}\|\overline{Y}_{t_1,t_2} - \overline{Y}_{t_2,t_3}\|_{\mathcal{H}} \geq |\mathcal{E}(t) - \sqrt{Z_t}|.$$

□

2.6.3 Raisonnements déterministes sur Ω

Preuve du lemme 1.
Par l'absurde, supposons que $\hat{\tau} = \tau$ avec $k \in \{1, \dots, D_{\tau^*} + 1\}$ et l tels que $\tau^*_{k-1} \leq \tau_{l-1} < \tau_l < \tau_{l+1} < \tau^*_k$. On note alors $t = (\tau_{l-1}, \tau_l, \tau_{l+1})$. Par le lemme 3, on a

$$\mathrm{crit}(\tau^{-l}) - \mathrm{crit}(\tau) = \frac{(\tau_{l+1} - \tau_l)(\tau_l - \tau_{l-1})}{\tau_{l+1} - \tau_{l-1}} \|\overline{\theta}_{\tau_l, \tau_{l+1}} - \overline{\theta}_{\tau_{l-1}, \tau_l}\|^2_{\mathcal{H}} - A \ln(n).$$

Par hypothèse $\mathcal{E}(t) = 0$ donc le lemme 4 donne $\mathrm{crit}(\tau^{-l}) - \mathrm{crit}(\tau) \leq Z_t - A\ln(n)$. Donc sur Ω, on a bien $\mathrm{crit}(\tau^{-l}) - \mathrm{crit}(\tau) \leq (L - A)\ln(n) < 0$ ce qui prouve $\tau \neq \hat{\tau}$. □

Preuve du lemme 5.
Par l'absurde, on suppose que $\hat{\tau} = \tau$ avec $k \in \{1, \dots, D_{\tau^*}\}$ et l tel que $\tau_l < \tau^*_k < \tau_{l+1}$. Pour $t = (\tau_l, \tau^*_k, \tau_{l+1})$, on suppose que $\mathcal{E}(t) > (\sqrt{A} + \sqrt{L})\ln(n)$. Par les lemmes 3 et 4, on a

$$\begin{aligned}
\mathrm{crit}(\tau) - \mathrm{crit}(\tau^{(k)}) &\geq \left(\mathcal{E}(t) - \sqrt{Z_t}\right)^2 - A\ln(n) \\
&\geq \left(\mathcal{E}(t) - \sqrt{L\ln(n)}\right)^2 - A\ln(n).
\end{aligned}$$

L'hypothèse sur $\mathcal{E}(t)$ donne donc $\mathrm{crit}(\tau) - \mathrm{crit}(\tau^{(k)}) > 0$ ce qui prouve bien que $\tau \neq \hat{\tau}$. □

Preuve du lemme 6.
Par l'absurde, on suppose que $\hat{\tau} = \tau$ avec $k \in \{1, \dots, D_{\tau^*}\}$ et l tel que $\tau_l < \tau^*_k - r < \tau^*_k + r < \tau_{l+1}$, où $r = \min\left\{\frac{\tau^*_k - \tau^*_{k-1}}{2}, \frac{\tau^*_{k+1} - \tau^*_k}{2}\right\}$. De plus, on suppose $\mathcal{E}^*_k > \kappa\sqrt{\ln(n)}$. On note $t = [\tau_l, \tau^*_k, \tau_{l+1}]$. Les lemmes 3 et 4 donnent

$$\mathrm{crit}(\tau) - \mathrm{crit}(\tau^{(k)}) \geq \left(\mathcal{E}(t) - \sqrt{Z_t}\right)^2 - A\ln(n). \tag{2.11}$$

Si $\tau_l \geq \tau^*_{k-1}$, alors $\overline{\theta^*}_{\tau_l, \tau^*_k} = \mu^*_k$. Sinon $\tau^*_{k-1} > \tau_l$ et

$$\overline{\theta^*}_{\tau_l, \tau^*_k} = \frac{\tau^*_k - \tau^*_{k-1}}{\tau^*_k - \tau_l}\mu^*_k + \frac{\tau^*_{k-1} - \tau_l}{\tau^*_k - \tau_l}\overline{\theta^*}_{\tau_l, \tau^*_{k-1}} = \mu^*_k + \frac{\tau^*_{k-1} - \tau_l}{\tau^*_k - \tau_l}(\overline{\theta^*}_{\tau_l, \tau^*_{k-1}} - \mu^*_k).$$

Dans tous les cas, on a

$$\overline{\theta^*}_{\tau_l, \tau^*_k} = \mu^*_k + 1_{\tau^*_{k-1} > \tau_l}\frac{\tau^*_{k-1} - \tau_l}{\tau^*_k - \tau_l}(\overline{\theta^*}_{\tau_l, \tau^*_{k-1}} - \mu^*_k).$$

De la même manière

$$\overline{\theta^*}_{\tau^*_k, \tau_{l+1}} = \mu^*_{k+1} + 1_{\tau^*_{k+1} < \tau_{l+1}}\frac{\tau_{l+1} - \tau^*_{k+1}}{\tau_{l+1} - \tau^*_k}(\overline{\theta^*}_{\tau^*_{k+1}, \tau_{l+1}} - \mu^*_{k+1}).$$

En utilisant l'inégalité triangulaire, on obtient alors

$$\mathcal{E}(t) \geq \sqrt{\frac{(\tau_{l+1} - \tau^*_k)(\tau_{l+1} - \tau_l)}{\tau^*_k - \tau_l}}\left(\|\Delta^*_k\|_{\mathcal{H}} - A_\tau - B_\tau\right), \tag{2.12}$$

avec $A_\tau := 1_{\tau_l < \tau^*_{k-1}}\|\mu^*_{k-1} - \overline{\theta^*}_{\tau_l, \tau^*_{k-1}}\|_{\mathcal{H}}$ et $B_\tau := 1_{\tau_{l+1} > \tau^*_{k+1}}\|\mu^*_k - \overline{\theta^*}_{\tau^*_{k+1}, \tau_{l+1}}\|_{\mathcal{H}}$. On fait alors une disjonction de cas pour montrer que $\hat{\tau} \neq \tau$ en utilisant l'hypothèse sur $\mathcal{E}^*_k$.

• **cas 1** $\max(A_\tau, B_\tau) \leq \|\Delta^*_k\|_{\mathcal{H}}/3$

Dans ce cas, on a $\mathcal{E}(t) \geq \sqrt{\frac{(\tau_{l+1}-\tau_k^*)(\tau_k^*-\tau_l)}{\tau_{l+1}-\tau_l}}\frac{\|\Delta_k^*\|_{\mathcal{H}}}{3}$. La fonction définie par $f(x,y) = \frac{(x-\tau_k^*)(\tau_k^*-y)}{x-y}$ pour $x \geq \tau_k^*+r$ et $y \leq \tau_k^*-r$ est croissante en x et décroissante en y donc $f(\tau_{l+1},\tau_l) \geq f(\tau_k^*+r,\tau_k^*-r) = r/2$. On a alors $\mathcal{E}(t) \geq \sqrt{r/2} \times (\|\Delta_k^*\|_{\mathcal{H}}/3)$. Et

$$(\mathcal{E}_k^*/\|\Delta_k^*\|_{\mathcal{H}})^2 = \frac{(\tau_{k+1}^*-\tau_k^*)(\tau_k^*-\tau_{k-1}^*)}{\tau_{k+1}^*-\tau_{k-1}^*} = 2r\left(1-\frac{2r}{\tau_{k+1}^*-\tau_{k-1}^*}\right) \leq 2r.$$

D'où $\mathcal{E}(t) \geq \sqrt{r/2} \times \mathcal{E}_k^*/(3\sqrt{2r}) = \mathcal{E}_k^*/6 \geq \frac{\kappa\sqrt{\ln(n)}}{6} > (\sqrt{A}+\sqrt{L})\sqrt{\ln(n)}$. Avec les inégalités (2.11) et (2.12) on a bien $\text{crit}(\tau^{(k)}) < \text{crit}(\tau)$ donc $\hat{\tau} \neq \tau$.

- **cas 2** $B_\tau \leq \|\Delta_k^*\|_{\mathcal{H}}/3$ et $A_\tau \geq \|\Delta\|_{\mathcal{H}}/3$

On a nécessairement $\tau_l < \tau_{k-1}^*$. On pose alors $u = (\tau_{k-1}^*, \tau_k^*, \tau_{l+1})$. On a

$$\begin{aligned}
\text{crit}(\tau^{(k-1,k)}) - \text{crit}(\tau) &= \text{crit}(\tau^{(k-1,k)}) - \text{crit}(\tau^{(k-1)}) + \text{crit}(\tau^{(k-1)}) - \text{crit}(\tau) \\
&= 2A\ln(n) - \frac{(\tau_{l+1}-\tau_k^*)(\tau_k^*-\tau_{k-1}^*)}{\tau_{l+1}-\tau_{k-1}^*}\left\|\overline{Y}_{\tau_{k-1}^*,\tau_k^*} - \overline{Y}_{\tau_k^*,\tau_{l+1}}\right\|_{\mathcal{H}}^2 \\
&\quad - \frac{(\tau_{l+1}-\tau_{k-1}^*)(\tau_{k-1}^*-\tau_l)}{\tau_{l+1}-\tau_l}\left\|\overline{Y}_{\tau_l,\tau_{k-1}^*} - \overline{Y}_{\tau_{k-1}^*,\tau_{l+1}}\right\|_{\mathcal{H}}^2 \\
&\leq 2A\ln(n) - \frac{(\tau_{l+1}-\tau_k^*)(\tau_k^*-\tau_{k-1}^*)}{\tau_{l+1}-\tau_{k-1}^*}\left\|\overline{Y}_{\tau_{k-1}^*,\tau_k^*} - \overline{Y}_{\tau_k^*,\tau_{l+1}}\right\|_{\mathcal{H}}^2 \\
&\leq 2A\ln(n) - \left(\mathcal{E}(u) - \sqrt{Z_u}\right)_+^2 \\
&\leq 2A\ln(n) - \left(\mathcal{E}(u) - \sqrt{L\ln(n)}\right)_+^2 \text{ par (2.7).}
\end{aligned}$$

De la même manière que pour $\mathcal{E}(t)$, on a

$$\begin{aligned}
\mathcal{E}(u) &\geq \left[\|\Delta_k^*\|_{\mathcal{H}} - 1_{\tau_{l+1}>\tau_{k+1}^*}\left\|\mu_k^* - \overline{\theta^*}_{\tau_k^*,\tau_{l+1}}\right\|_{\mathcal{H}}\right]\sqrt{\frac{(\tau_k^*-\tau_{k-1}^*)(\tau_{l+1}-\tau_k^*)}{\tau_{l+1}-\tau_{k-1}^*}} \\
&= [\|\Delta_k^*\|_{\mathcal{H}} - B_\tau]\sqrt{\frac{(\tau_k^*-\tau_{k-1}^*)(\tau_{l+1}-\tau_k^*)}{\tau_{l+1}-\tau_{k-1}^*}}.
\end{aligned}$$

Comme $B_\tau \leq \|\Delta_k^*\|_{\mathcal{H}}/3$, on a $\mathcal{E}(u) \geq \frac{2\|\Delta_k^*\|_{\mathcal{H}}}{3}\sqrt{\frac{(\tau_k^*-\tau_{k-1}^*)(\tau_{l+1}-\tau_k^*)}{\tau_{l+1}-\tau_{k-1}^*}}$.

D'où $\mathcal{E}(u) \geq \mathcal{E}_k^*\frac{2}{3}\sqrt{\frac{(\tau_{l+1}-\tau_k^*)(\tau_{k+1}^*-\tau_{k-1}^*)}{(\tau_{l+1}-\tau_{k-1}^*)(\tau_{k+1}^*-\tau_k^*)}}$.

Or $\tau_{l+1} \mapsto \frac{\tau_{l+1}-\tau_k^*}{\tau_{l+1}-\tau_{k-1}^*}$ est croissante donc $\frac{(\tau_{l+1}-\tau_k^*)(\tau_{k+1}^*-\tau_{k-1}^*)}{(\tau_{l+1}-\tau_{k-1}^*)(\tau_{k+1}^*-\tau_k^*)} \geq \frac{r(\tau_{k+1}^*-\tau_{k-1}^*)}{(r+\tau_k^*-\tau_{k-1}^*)(\tau_{k+1}^*-\tau_k^*)}$.
Si $2r = \tau_k^* - \tau_{k-1}^* \leq \tau_{k+1}^* - \tau_k^*$, alors

$$\frac{r(\tau_{k+1}^*-\tau_{k-1}^*)}{(r+\tau_k^*-\tau_{k-1}^*)(\tau_{k+1}^*-\tau_k^*)} \geq \frac{r}{r+\tau_k^*-\tau_{k-1}^*} = \frac{1}{3}.$$

Sinon $2r = \tau_{k+1}^* - \tau_k^* < \tau_k^* - \tau_{k-1}^*$, et

$$\begin{aligned}
\frac{r(\tau_{k+1}^*-\tau_{k-1}^*)}{(r+\tau_k^*-\tau_{k-1}^*)(\tau_{k+1}^*-\tau_k^*)} &= \frac{1}{2}\frac{2r+\tau_k^*-\tau_{k-1}^*}{r+\tau_k^*-\tau_{k-1}^*} \\
&= \frac{2r+\tau_k^*-\tau_{k-1}^*}{2r+2(\tau_k^*-\tau_{k-1}^*)} \geq 2/3.
\end{aligned}$$

Dans tous les cas, on a $\mathcal{E}(u) \geq \frac{2}{3\sqrt{3}}\mathcal{E}_k^* > \frac{\kappa 2\sqrt{\ln(n}}{3\sqrt{3}} \geq (\sqrt{L}+\sqrt{2A})\sqrt{\ln(n)}$ donc $\text{crit}(\tau^{(k-1,k)}) < \text{crit}(\tau)$ et $\hat{\tau} \neq \tau$.

- **cas 3** $A_\tau \leq \|\Delta_k^*\|_{\mathcal{H}}/3$ et $B_\tau \geq \|\Delta\|_{\mathcal{H}}/3$ Ce cas est analogue au cas 2.
- **cas 4** $\min(A_\tau, B_\tau) \geq \|\Delta_k^*\|_{\mathcal{H}}/3$

  On a nécessairement $\tau_l < \tau_{k-1}^* < \tau_k^* < \tau_{k+1}^* < \tau_{l+1}$. On note $t_k^* = (\tau_{k-1}^*, \tau_k^*, \tau_{k+1}^*)$ et on a

$$\begin{aligned}
&\mathrm{crit}(\tau^{(k-1,k,k+1)}) - \mathrm{crit}(\tau) \\
&= \mathrm{crit}(\tau^{(k-1,k,k+1)}) - \mathrm{crit}(\tau^{(k-1,k+1)}) + \mathrm{crit}(\tau^{(k-1,k+1)}) - \mathrm{crit}(\tau^{(k-1)}) \\
&\quad + \mathrm{crit}(\tau^{(k-1)}) - \mathrm{crit}(\tau) \\
&\leq A\ln(n) - \left(\mathcal{E}_k^* - \sqrt{Z_{t_k^*}}\right)_+^2 + 2A\ln(n) \quad \text{(par les lemmes 3 et 4)} \\
&\leq 3A\ln(n) - \left(\mathcal{E}_k^* - \sqrt{L\ln(n)}\right)_+^2 \quad \text{(par définition de } \Omega).
\end{aligned}$$

  Par hypothèse $\mathcal{E}_k^* > \kappa\sqrt{\ln(n)} \geq (\sqrt{3A} + \sqrt{L})\ln(n)$, donc $\mathrm{crit}(\tau^{(k-1,k,k+1)}) < \mathrm{crit}(\tau)$ et $\tau \neq \hat{\tau}$.

  Dans tous les cas, $\mathcal{E}_k^* > \kappa\sqrt{\ln(n)}$ garantit bien $\tau \neq \hat{\tau}$.

□

Preuve du corollaire 1.
On suppose que $\mathcal{E}_k^* > \kappa\sqrt{\ln(n)}$. Soit $C > \kappa$ tel que $\mathcal{E}_k^* > C\sqrt{\ln(n)}$. Le résultat du lemme 6 est valable pour τ^* qui vérifie (2.1). Si on ajoute une rupture à τ^*, la nouvelle ségmentation vérifie toujours la propriété (2.1) et on peut alors appliquer le lemme 6 avec cette nouvelle « vraie » segmentation.
On pose donc $\tau^{**} = \tau^* \cup \{\tau_k^* - a, \tau^* + b\} = (\tau_0^*, \ldots, \tau_{k-1}^*, \tau_k^* - a, \tau_k^*, \tau_k^* + b, \tau_{k+1}^*, \ldots, \tau_{D_{\tau^*}+1}^*)$, où $a = \alpha(\tau_k^* - \tau_{k-1}^*)$ et $b = \alpha(\tau_{k+1}^* - \tau_k^*)$ avec $\alpha \in]0,1]$. En notant $\mathcal{E}_k^{**} = \mathcal{E}(\tau_k^* - a, \tau_k^*, \tau_k^* + b)$ l'énergie du saut τ_k^* pour la segmentation τ^{**}, on obtient par un calcul direct $\mathcal{E}_k^{**} = \mathcal{E}_k^*\sqrt{1-\alpha}$. Si on pose $\alpha = 1 - (\kappa/C)^2$, on a bien $\mathcal{E}_k^{**} > \kappa\sqrt{\ln(n)}$ et on peut donc appliquer le lemme 6 avec τ^{**}. Cela donne

$$\min_l |\hat{\tau}_l - \tau_k^*| \leq \min\{b/2, a/2\} = \left(\frac{\kappa}{C}\right)^2 \min\left[\frac{\tau_{k+1}^* - \tau_k^*}{2}, \frac{\tau_k^* - \tau_{k-1}^*}{2}\right].$$

Cette inégalité est vraie pour tout $C > \kappa$ tel que $\mathcal{E}_k^* > C\sqrt{\ln(n)}$ donc en prenant l'infimum sur toutes les constantes C possibles, on retrouve bien

$$\min_l |\hat{\tau}_l - \tau_k^*| \leq \left(\frac{\kappa\sqrt{\ln(n)}}{\mathcal{E}_k^*}\right)^2 \min\left[\frac{\tau_{k+1}^* - \tau_k^*}{2}, \frac{\tau_k^* - \tau_{k-1}^*}{2}\right].$$

□

Preuve du lemme 7.
On va juste montrer $|\hat{\tau} \cap [\tau_k^*, (\tau_k^* + \tau_{k+1}^*)/2]| \leq 1$ pour $k \in \{0, \ldots, D_{\tau^*}\}$ puisque le deuxième résultat se prouve de manière analogue. On suppose par l'absurde que $\hat{\tau} = \tau$ avec $\tau_k^* \leq \tau_{l-1} < \tau_l \leq \frac{\tau_k^* + \tau_{k+1}^*}{2}$. Par le lemme 1, on sait alors que $\tau_{l+1} > \tau_{k+1}^*$. Puisque $\tau = \hat{\tau}$, on a $\mathrm{crit}(\tau) \leq \mathrm{crit}(\tau^{-l})$ qui implique

$$\mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1}) \geq \sqrt{\ln(n)}(\sqrt{A} - \sqrt{L}) \tag{2.13}$$

par les lemmes 3 et 4. De même $\mathrm{crit}(\tau) \leq \mathrm{crit}(\tau^{-l,(k+1)})$ donne

$$\mathcal{E}(\tau_{l-1}, \tau_{k+1}^*, \tau_{l+1}) \leq \mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1}) + 2\sqrt{L\ln(n)}. \tag{2.14}$$

D'autre part, on a par le calcul

$$\mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1}) = \sqrt{\frac{(\tau_{l+1} - \tau_{k+1}^*)(\tau_l - \tau_{l-1})}{(\tau_{l+1} - \tau_l)(\tau_{k+1}^* - \tau_{l-1})}}\mathcal{E}(\tau_{l-1}, \tau_{k+1}^*, \tau_{l+1}).$$

Et on a

$$\begin{aligned}\frac{(\tau_{l+1}-\tau^*_{k+1})(\tau_l-\tau_{l-1})}{(\tau_{l+1}-\tau_l)(\tau^*_{k+1}-\tau_{l-1})} &\leq \frac{\tau_l-\tau_{l-1}}{\tau^*_{k+1}-\tau_{l-1}}\\ &\leq \frac{(\tau^*_{k+1}-\tau^*_k)/2}{\tau^*_{k+1}-\tau^*_k} = 1/2.\end{aligned}$$

D'où $\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) \leq \mathcal{E}(\tau_{l-1},\tau^*_{k+1},\tau_{l+1})/\sqrt{2}$. On note que l'inégalité (2.13) implique $\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) > 0$. L'inégalité précédente et (2.14) donnent

$$\begin{aligned}\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1})\sqrt{2} &\leq \mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) + 2\sqrt{L\ln(n)}\\ \sqrt{2} &\leq 1 + \frac{2\sqrt{L\ln(n)}}{\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1})}\\ \sqrt{2} &\leq 1 + \frac{2\sqrt{L\ln(n)}}{(\sqrt{A}-\sqrt{L})\sqrt{\ln(n)}} \text{ par (2.13)}\\ \sqrt{2} &\leq 1 + \frac{2\sqrt{L}}{\sqrt{A}-\sqrt{L}}.\end{aligned}$$

Or $\sqrt{A} > 6\sqrt{L}$ ce qui implique $\sqrt{2} > 1 + \frac{2\sqrt{L}}{\sqrt{A}-\sqrt{L}}$ et contredit l'inégalité qu'on vient d'établir ci-dessus donc $\tau \neq \hat{\tau}$. □

Preuve du lemme 8.
On suppose par l'absurde que $\tau = \hat{\tau}$ avec $k \in \{1,\dots,D_{\tau^*}\}$ et l tels que $\frac{\tau^*_{k-1}+\tau^*_k}{2} \leq \tau_{l-1} < \tau_l \leq \frac{\tau^*_k+\tau^*_{k+1}}{2}$. Par le lemme 7, on sait qu'on a nécessairement $\tau_{l-1} \leq \tau^*_k \leq \tau_l$. Sans perte de généralité, on peut supposer $\tau_l - \tau^*_k \geq \tau^*_k - \tau_{l-1}$. Les raisonnements sont exactement les mêmes si $\tau_l - \tau^*_k < \tau^*_k - \tau_{l-1}$. On va distinguer deux cas pour prouver qu'on a toujours $\hat{\tau} \neq \tau$. On va également utiliser un lemme intermédiaire pour clarifier la preuve.

Lemme 10.
Si $\tau_{l+1} \leq \tau^*_{k+1}$, on a

$$\mathcal{E}(\tau_{l-1},\tau^*_k,\tau_l) \geq \mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}). \tag{2.15}$$

Si $\tau_{l+1} > \tau^*_{k+1}$, on a les inégalités suivantes.

$$\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) \leq \sqrt{2}\mathcal{E}(\tau_l,\tau^*_{k+1},\tau_{l+1}) + \mathcal{E}(\tau_{l-1},\tau^*_k,\tau_l), \tag{2.16}$$

$$\mathcal{E}(\tau_{l-1},\tau^*_k,\tau_l) \geq \mathcal{E}(\tau_{l-1},\tau_l,\tau^*_{k+1}), \tag{2.17}$$

$$\mathcal{E}(\tau_l,\tau^*_{k+1},\tau_{l+1}) \geq \sqrt{3}\mathcal{E}(\tau^*_k,\tau_l,\tau_{l+1}). \tag{2.18}$$

On fait maintenant la disjonction de cas suivant la position de τ_{l+1} par rapport à τ^*_{k+1}.

- **cas 1** $\tau_{l+1} \leq \tau^*_{k+1}$
  Montrons qu'on ne peut pas avoir $\text{crit}(\tau) \leq \text{crit}(\tau^{-l})$ et $\text{crit}(\tau) \leq \text{crit}(\tau^{(-l,k)})$. En utilisant les lemmes 2.7, 3 et 4 on obtient

$$\begin{aligned}\text{crit}(\tau) \leq \text{crit}(\tau^{-l}) &\Rightarrow \mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) \geq (\sqrt{A}-\sqrt{L})\sqrt{\ln(n)}.\\ \text{crit}(\tau) \leq \text{crit}(\tau^{-l,(k)}) &\Rightarrow \text{crit}(\tau) - \text{crit}(\tau^{(k)}) \leq \text{crit}(\tau^{-l,(k)}) - \text{crit}(\tau^{(k)})\\ &\Rightarrow \mathcal{E}(\tau_{l-1},\tau^*_k,\tau_l) \leq \mathcal{E}(\tau^*_k,\tau_l,\tau_{l+1}) + 2\sqrt{L\ln(n)} = 2\sqrt{L\ln(n)}.\end{aligned}$$

  D'autre part, l'inégalité (2.15) donne $(\sqrt{A}-\sqrt{L}) \leq 2\sqrt{L}$. Or c'est impossible puisque $\sqrt{A} \geq 22\sqrt{L}$ donc $\hat{\tau} \neq \tau$.

• **cas 2** $\tau^*_{k+1} < \tau_{l+1}$

Montrons qu'on ne peut pas avoir $\text{crit}(\tau) \leq \min\{\text{crit}(\tau^{-l}), \text{crit}(\tau^{-l,(k)}), \text{crit}(\tau^{-l,(k+1)})\}$. On suppose par l'absurde que l'inégalité est vérifiée. À nouveau, la définition de l'évènement Ω 2.7 et les lemmes 3 et 4 donnent

$$\begin{aligned}
\text{crit}(\tau) \leq \text{crit}(\tau^{-l,(k)}) &\Rightarrow \text{crit}(\tau) - \text{crit}(\tau^{(k)}) \leq \text{crit}(\tau^{-l,(k)}) - \text{crit}(\tau^{(k)}) \\
&\Rightarrow \mathcal{E}(\tau_{l-1}, \tau^*_k, \tau_l) \leq \mathcal{E}(\tau^*_k, \tau_l, \tau_{l+1}) + 2\sqrt{L\ln(n)} \qquad (2.19a) \\
\text{crit}(\tau) \leq \text{crit}(\tau^{-l,(k+1)}) &\Rightarrow \text{crit}(\tau) - \text{crit}(\tau^{(k+1)}) \leq \text{crit}(\tau^{-l,(k)}) - \text{crit}(\tau^{(k+1)}) \\
&\Rightarrow \mathcal{E}(\tau_l, \tau^*_{k+1}, \tau_{l+1}) \leq \mathcal{E}(\tau_{l-1}, \tau_l, \tau^*_{k+1}) + 2\sqrt{L\ln(n)}. \qquad (2.19b)
\end{aligned}$$

On part de (2.19a) pour montrer

$$\begin{aligned}
\mathcal{E}(\tau_{l-1}, \tau^*_k, \tau_l) &\leq \mathcal{E}(\tau^*_k, \tau_l, \tau_{l+1}) + 2\sqrt{L\ln(n)} \\
&\leq 2\sqrt{\ln(n)} + \frac{1}{\sqrt{3}}\mathcal{E}(\tau_l, \tau^*_{k+1}, \tau_{l+1}) \text{ par (2.18)} \\
&\leq 2\sqrt{L\ln(n)} + \frac{1}{\sqrt{3}}\left(\mathcal{E}(\tau_{l-1}, \tau_l, \tau^*_{k+1}) + 2\sqrt{L\ln(n)}\right) \text{ par (2.19b)} \\
&\leq 2\sqrt{L\ln(n)}\left(1 + \frac{1}{\sqrt{3}}\right) + \frac{1}{\sqrt{3}}\mathcal{E}(\tau_{l-1}, \tau^*_k, \tau_l) \text{ par (2.17)}
\end{aligned}$$

Finalement, on en déduit que $\mathcal{E}(\tau_{l-1}, \tau^*_k, \tau_l) \leq 2(2+\sqrt{3})\sqrt{L\ln(n)}$. De la même manière, (2.19b) donne

$$\begin{aligned}
\mathcal{E}(\tau_l, \tau^*_{k+1}, \tau_{l+1}) &\leq \mathcal{E}(\tau_{l-1}, \tau_l, \tau^*_{k+1}) + 2\sqrt{L\ln(n)} \\
&\leq 2\sqrt{L\ln(n)} + \mathcal{E}(\tau_{l-1}, \tau^*_k, \tau_l) \text{ par (2.17)} \\
&\leq 2\sqrt{\ln(n)} + \mathcal{E}(\tau^*_k, \tau_l, \tau_{l+1}) + 2\sqrt{L\ln(n)} \text{ par (2.19a)} \\
&\leq 4\sqrt{\ln(n)} + \frac{1}{\sqrt{3}}\mathcal{E}(\tau_l, \tau^*_{k+1}, \tau_{l+1}) \text{ par (2.18).}
\end{aligned}$$

Et on a donc $\mathcal{E}(\tau_l, \tau^*_{k+1}, \tau_{l+1}) \leq 2(3+\sqrt{3})\sqrt{L\ln(n)}$. L'inégalité (2.16) donne alors

$$\mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1}) \leq 2\sqrt{L\ln(n)}\left[(1+\sqrt{2})(3+\sqrt{3}) - 1\right].$$

On fait l'hypothèse $\text{crit}(\tau) \leq \text{crit}(\tau^{-l})$ qui implique $\mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1}) \geq (\sqrt{A} - \sqrt{L})\sqrt{\ln(n)}$. On a donc $(\sqrt{A} - \sqrt{L}) \leq 2\left[(1+\sqrt{2})(3+\sqrt{3}) - 1\right]\sqrt{L}$ avec les deux dernières inégalités sur $\mathcal{E}(\tau_{l-1}, \tau_l, \tau_{l+1})$. Or cela contredit $\sqrt{A} \geq 22\sqrt{L}$ et prouve $\text{crit}(\tau) > \min\{\text{crit}(\tau^{-l}), \text{crit}(\tau^{-l,(k)}), \text{crit}(\tau^{-l,(k+1)})\}$, et donc $\hat{\tau} \neq \tau$.

Ceci conclut la preuve du lemme 8. □

Preuve du lemme 10.
On rappelle qu'on a $\frac{\tau^*_{k-1}+\tau^*_k}{2} \leq \tau_{l-1} \leq \tau^*_k < \tau_l \leq \frac{\tau^*_k+\tau^*_{k+1}}{2} < \tau_{l+1}$ avec $\tau_l - \tau^*_k \geq \tau^*_k - \tau_{l-1}$. On va d'abord prouver l'inégalité (2.15). On suppose donc que $\tau_{l+1} \leq \tau^*_{k+1}$. Cela permet d'écrire $\overline{\theta^*}_{\tau_l,\tau_{l+1}} = \mu^*_{k+1}$. On a aussi

$$\overline{\theta^*}_{\tau_{l-1},\tau_l} = \frac{\tau^*_k - \tau_{l-1}}{\tau_l - \tau_{l-1}}\mu^*_k + \frac{\tau_l - \tau^*_k}{\tau_l - \tau_{l-1}}\mu^*_{k+1}$$

d'où

$$\begin{aligned}\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) &= \sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau_l-\tau_{l-1})}{\tau_{l+1}-\tau_{l-1}}}\left\|\overline{\theta^*}_{\tau_{l-1},\tau_l}-\overline{\theta^*}_{\tau_l,\tau_{l+1}}\right\|_{\mathcal{H}}\\ &= \frac{\tau_k^*-\tau_{l-1}}{\tau_l-\tau_{l-1}}\sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau_l-\tau_{l-1})}{\tau_{l+1}-\tau_{l-1}}}\|\Delta_k^*\|_{\mathcal{H}}\\ &= \frac{\tau_k^*-\tau_{l-1}}{\tau_l-\tau_{l-1}}\sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau_l-\tau_{l-1})(\tau_l-\tau_{l-1})}{(\tau_{l+1}-\tau_{l-1})(\tau_l-\tau_k^*)(\tau_k^*-\tau_{l-1})}}\mathcal{E}(\tau_{l-1},\tau_k^*,\tau_l)\\ &= \sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau_k^*-\tau_{l-1})}{(\tau_{l+1}-\tau_{l-1})(\tau_l-\tau_k^*)}}\mathcal{E}(\tau_{l-1},\tau_k^*,\tau_l).\end{aligned}$$

Or $\frac{(\tau_{l+1}-\tau_l)(\tau_k^*-\tau_{l-1})}{(\tau_{l+1}-\tau_{l-1})(\tau_l-\tau_k^*)} \leq \frac{\tau_k^*-\tau_{l-1}}{\tau_l-\tau_{l-1}} \leq 1$ par hypothèse. Donc on a bien prouvé l'inégalité $\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) \leq \mathcal{E}(\tau_{l-1},\tau_k^*,\tau_l)$.
On va maintenant prouver les autres inégalités donc on se place dans le cas où $\tau_{k+1}^* < \tau_{l+1}$. En calculant les énergies intervenant dans les différentes inégalités, on a

$$\begin{aligned}\mathcal{E}^2(\tau_{l-1},\tau_k^*,\tau_l) &= \frac{(\tau_l-\tau_k^*)(\tau_k^*-\tau_{l-1})}{\tau_l-\tau_{l-1}}\|\Delta_k^*\|_{\mathcal{H}}^2,\\ \mathcal{E}^2(\tau_l,\tau_{k+1}^*,\tau_{l+1}) &= \frac{(\tau_{l+1}-\tau_{k+1}^*)(\tau_{k+1}^*-\tau_l)}{\tau_{l+1}-\tau_l}\left\|\mu_{k+1}^*-\overline{\theta^*}_{\tau_{k+1}^*,\tau_{l+1}}\right\|_{\mathcal{H}}^2,\\ \mathcal{E}^2(\tau_{l-1},\tau_l,\tau_{k+1}^*) &= \frac{(\tau_{k+1}^*-\tau_l)(\tau_k^*-\tau_{l-1})^2}{(\tau_{k+1}^*-\tau_{l-1})(\tau_l-\tau_{l-1})}\|\Delta_k^*\|_{\mathcal{H}}^2,\\ \mathcal{E}^2(\tau_k^*,\tau_l,\tau_{l+1}) &= \frac{(\tau_l-\tau_k^*)(\tau_{l+1}-\tau_{k+1}^*)^2}{(\tau_{l+1}-\tau_k^*)(\tau_{l+1}-\tau_l)}\left\|\mu_{k+1}^*-\overline{\theta^*}_{\tau_{k+1}^*,\tau_{l+1}}\right\|_{\mathcal{H}}^2,\\ \mathcal{E}^2(\tau_{l-1},\tau_l,\tau_{l+1}) &= \frac{(\tau_{l+1}-\tau_l)(\tau_l-\tau_{l-1})^2}{\tau_{l+1}-\tau_{l-1})}\left\|\frac{\tau_k^*-\tau_{l-1}}{\tau_l-\tau_{l-1}}\Delta_k^*-\frac{\tau_{l+1}-\tau_{k+1}^*}{\tau_{l+1}-\tau_l}\left(\mu_{k+1}^*-\overline{\theta^*}_{\tau_{k+1}^*,\tau_{l+1}}\right)\right\|_{\mathcal{H}}^2.\end{aligned}$$

On peut alors aisément prouver les inégalités (2.17) et (2.18).On a

$$\mathcal{E}^2(\tau_{l-1},\tau_l,\tau_{k+1}^*) = \frac{(\tau_l-\tau_k^*)(\tau_{k+1}^*-\tau_{l-1})}{(\tau_{k+1}^*-\tau_l)(\tau_k^*-\tau_{l-1})}\mathcal{E}^2(\tau_{l-1},\tau_k^*,\tau_l)$$

et

$$\frac{(\tau_l-\tau_k^*)(\tau_{k+1}^*-\tau_{l-1})}{(\tau_{k+1}^*-\tau_l)(\tau_k^*-\tau_{l-1})} \leq \frac{\tau_l-\tau_k^*}{\tau_{k+1}^*-\tau_l} \leq 1$$

car $\tau_l \leq (\tau_k^*+\tau_{k+1}^*)/2$, ce qui prouve (2.17). D'autre part, on a

$$\mathcal{E}^2(\tau_k^*,\tau_l,\tau_{l+1}) = \frac{(\tau_l-\tau_k^*)(\tau_{l+1}-\tau_{k+1}^*)}{(\tau_{l+1}-\tau_k^*)(\tau_{k+1}^*-\tau_l)}\mathcal{E}^2(\tau_l,\tau_{k+1}^*,\tau_{l+1})$$

et

$$\begin{aligned}\frac{(\tau_{l+1}-\tau^*_{k+1})(\tau^*_k-\tau_{l-1})}{(\tau_{l+1}-\tau^*_k)(\tau^*_{k+1}-\tau_{l-1})} &\leq \frac{\tau^*_k-\tau_{l-1}}{\tau^*_{k+1}-\tau_{l-1}}\\ &= \frac{\tau^*_k-\tau_{l-1}}{\tau^*_{k+1}-\tau^*_k+\tau^*_k-\tau_{l-1}}\\ &\leq \frac{\tau_l-\tau^*_k}{\tau^*_{k+1}-\tau^*_k+\tau_l-\tau^*_k} \text{ par hypothèse } \tau_l-\tau^*_k \geq \tau^*_k-\tau_{l-1}\\ &\leq \frac{(\tau^*_{k+1}+\tau^*_k)/2-\tau^*_k}{\tau^*_{k+1}-\tau^*_k+(\tau^*_{k+1}+\tau^*_k)/2-\tau^*_k}\\ &= \frac{1}{3}\end{aligned}$$

ce qui prouve l'inégalité (2.18). On prouve maintenant l'inégalité (2.16). Une simple inégalité triangulaire donne

$$\begin{aligned}\mathcal{E}(\tau_{l-1},\tau_l,\tau_{l+1}) &\leq \sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau_l-\tau_{l-1})}{\tau_{l+1}-\tau_{l-1}}}\left[\frac{\tau^*_k-\tau_{l-1}}{\tau_l-\tau_{l-1}}\|\Delta^*_k\|_{\mathcal{H}}+\frac{\tau_{l+1}-\tau^*_{k+1}}{\tau_{l+1}-\tau_l}\left\|\mu^*_{k+1}-\overline{\theta^*}_{\tau^*_{k+1},\tau_{l+1}}\right\|_{\mathcal{H}}\right]\\ &= \sqrt{\frac{(\tau_{l+1}-\tau_l)(\tau^*_k-\tau_{l-1})}{(\tau_{l+1}-\tau_{l-1})(\tau_l-\tau^*_k)}}\mathcal{E}(\tau_{l-1},\tau^*_k,\tau_l)\\ &+\sqrt{\frac{(\tau_l-\tau_{l-1})(\tau_{l+1}-\tau^*_{k+1})}{(\tau_{l+1}-\tau_{l-1})(\tau^*_{k+1}-\tau_l)}}\mathcal{E}(\tau_l,\tau^*_{k+1},\tau_{l+1}).\end{aligned}$$

Il suffit alors de montrer

$$\frac{(\tau_{l+1}-\tau_l)(\tau^*_k-\tau_{l-1})}{(\tau_{l+1}-\tau_{l-1})(\tau_l-\tau^*_k)} \leq \frac{\tau^*_k-\tau_{l-1}}{\tau_l-\tau^*_k} \leq 1$$

et

$$\begin{aligned}\frac{(\tau_l-\tau_{l-1})(\tau_{l+1}-\tau^*_{k+1})}{(\tau_{l+1}-\tau_{l-1})(\tau^*_{k+1}-\tau_l)} &\leq \frac{\tau_l-\tau_{l-1}}{\tau^*_{k+1}-\tau_l} = \frac{\tau_l-\tau^*_k+\tau^*_k-\tau_{l-1}}{\tau^*_{k+1}-\tau_l}\\ &\leq \frac{2(\tau_l-\tau^*_k)}{\tau^*_{k+1}-\tau_l} \leq 2.\end{aligned}$$

On a donc bien prouvé l'inégalité (2.16) et donc le lemme 10. □

Bibliographie

[1] Damien Garreau et Sylvain Arlot. *Consistent change-point detection with kernels.* 2017. Arxiv e-prints. Available at `https://arxiv.org/abs/1612.04740v3`.

[2] Sylvain Arlot, Alain Célisse et Zaid Harchaoui. *A kernel multiple change-point algorithm via model selection.* 2012. Arxiv e-prints. Available at `https://arxiv.org/abs/1202.3878v2`.

[3] Nachman Aronszajn. *Theory of Reproducing Kernels.* 1950. Transactions of the American Mathematical Society, vol. 68, no 3, p. 337-404.

[4] C. McDiarmid. *On the method of bounded differences.* 1989. Surveys in Combinatorics, p. 148-188. Cambridge University Press.

[5] Youngser Park, Heng Wang, Tobias Nöbauer, Alipasha Vaziri, et Carey E Priebe. *Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics.* 2015. Neuron, 2(3,000) :4-000.

[6] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, O. Cappé. *A regularized kernel-based approach to unsupervised audio segmentation.* 2009. IEEE Int. Conf. Acoust., Speech, Signal Processing, pages 1665-1668, Taiwan.

[7] A. Tartakovsky, B. Rozovsky, R. Blazek, H. Kim. *A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods.* 2006. IEEE Transactions on Signal Processing, 54(9).

[8] M. Fromont, M. Lerasle, P. Reynaud-Bouret, N. Verzelen. Communication Personnelle.

Annexes

A Prétraitement des données

La base de données dont on dispose est un fichier *.mdb* (base de données réalisée avec le logiciel Microsoft Access). Pour travailler sur la base de données, il faut d'abord faire une requête qu'on peut ensuite utiliser avec *R*. La requête qu'on a utilisé est un fichier *.csv* qui contient 6351 lignes, qu'on appelle « observation » et chacune d'elle est décrite par 42 variables. Celles-ci peuvent être caractéristiques de l'espèce (esp), des cultures (cult), du territoire (ter), de la variété (var) ou encore de l'agrosystème (as) par exemple. On en donne la liste dans le tableau ci-dessous.

cultCodeAS	varAssociation	asTechAmendements
espNomCommun	varFrequenceAssociation	asTechOutilsManuels
espCatListes	varCommerce	asDetailAmendement
espCat	varAncienneté	asPaysageParcArboré
cultAbondance	varAncienneténFormulation	asPaysageRelief
cultAbondanceSource	malNum	asPaysageSols
terNomVillage	asEthnieAgrosystème	asPlantesPhares
terPays	asLanguePrincipale	asElevageComposition
terGPSLat	asCanton	asElevageConduite
terGPSLong	asTypologieChamp	asElevageImportance
terProxMarché	asTypologie	asElevageCommentaires
varCode	asTechJachères	asAutresActivités
varTailleChamp	asTechRotations	asPêche
varTypeChamp	asTechOutillage	varObservation

A.1 La variable *varCode*

Description

On décrit *varCode* avec un exemple d'observation : 'T-AM-2009-002-01-02-49'.

- T-AM-2009-002-01-02-49 : initiales de l'informateur.rice
- T-AM-2009-002-01-02-49 : année de l'inventaire
- T-AM-2009-002-01-02-49 : identifiant *unique* d'un transect
- T-AM-2009-002-01-02-49 : identifiant du village sur le transect
- T-AM-2009-002-01-02-49 : identifiant *unique* d'un village
- T-AM-2009-002-01-02-49 : identifiant de l'agrosystème dans le village
- T-AM-2009-002-01-02-49 : identifiant *unique* d'un agrosystème
- T-AM-2009-002-01-02-49 : identifiant unique de l'espèce
- T-AM-2009-002-01-02-49 : identifiant de la variété pour une espèce donnée, dans un agrosystème donné
- T-AM-2009-002-01-02-49 : identifiant *unique* d'une variété dans un agrosystème

Transformation

En pratique, *varCode* est très riche en information mais est trop compliquée à utiliser de manière directe. On a donc transformé la variable pour accéder plus simplement à l'information. Pour cela, on a d'abord créer de nouvelles variables en segmentant *varCode* pour chaque observation.

$$varCode \rightarrow \text{"T"-init-year-vilNum-agroNum-spNum-varNum}$$

Ces nouvelles variables présentent une information parfois incomplète mais beaucoup plus simple d'utilisation. On a ensuite créé une variable pour l'identifiant d'un agrosystème (*agroID*) en agrégeant plusieurs variables. On a également fait une variable *agroLabel* qui correspond au nom du village (variable *terNomVillage*) auquel on rajoute "2" lorsqu'il s'agit du deuxième agrosystème dans le village.

$$agroID \leftarrow \text{init-year-vilNum-agroNum}$$

Pour bien comprendre ce que l'on a fait ici, on peut regarder ce qu'il se passe pour l'exemple précédent.

varCode	init	year	vilNum	agroNum	spNum	varNum	agroID
T-AA-2010-002-01-01-01	AA	2010	2	1	1	1	AA-2010-2-1
T-AH-2012-001-01-02-03	AH	2012	1	1	2	3	AH-2012-1-1
T-AH-2012-002-01-10-01	AH	2012	2	1	10	1	AH-2012-2-1
T-AM-2009-002-01-02-18	AM	2009	2	1	2	18	AM-2009-2-1
T-BS-2010-003-01-29-01	BS	2010	3	1	29	1	BS-2010-3-1

A.2 Observations retirées

Comme dans la plupart des cas lorsqu'on manipule des données réelles, certaines posent problème et doivent être retirées. Ici, on a retiré les observations qui concernent :

- le village T-AM-2009-003 (arachide non renseigné),
- l'agrosystème T-IB-2009-007-02 (une seule observation),
- et les observations telles que $spNum \neq malNum$ (35 observations).

A.3 La variable *varObservation*

On dispose de la variable *varObservation* qui indique pour chaque plante si elle est « semée ou plantée », « entretenue », « disparue », « non cultivée » ou si on la trouve à l'état « sauvage ». Les effectifs, pour notre jeu de données, sont les suivants :

- · 268 observations pour « Disparue »,
- · 292 observations pour « Non cultivée »,
- · 673 observations pour « Sauvage »,
- · 184 observations pour « Entretenue »,
- · 3036 observations pour « Semée ou plantée »,
- · et 53 observations pour lesquelles *varObservation* n'est pas renseignée.

On s'intéresse à la biodiversité cultivée et non à la biodiversité en général. On ne considère donc pas les plantes sauvages, disparues ou non cultivées et on conserve uniquement les observations pour les plantes renseignées par « Semée ou plantée » ou « Entretenue ». De plus, au vu des effectifs, on ne peut pas travailler sans le niveau « Semée ou plantée ». Le jeu de données final sur lequel on va travailler contient 3220 observations pour 50 variables.

A.4 Autres variables

Parmi les variables à notre disposition, il faut distinguer celles que l'on peut utiliser pour réaliser des analyses statistiques et celles qui ne sont pas adaptées. En général, une variable bien adapté est une variable qui s'apparente au résultat d'un QCM avec une seule réponse possible. Au contraire, les variables qui sont des commentaires, du texte saisi sans contrainte particulière, ne permettent pas l'utilisation d'outils statistiques. On peut en effet constater que :

- certaines variables ne sont pas exploitables comme la variable *varAssociation* (933 niveaux différents, notation non harmonisée),
- ou demandent d'être transformées avant analyse comme la variable *asElevageComposition*.

On évitera tout de même ce genre de variables qui peuvent mener à de mauvaises interprétations. Par exemple, si on a l'observation « bovins, ovins, caprins », on ne connaît pas la proportion de bovins par rapport aux ovins ni même si l'élevage est conséquent dans l'agrosystème. On a bien une information sur la présence de ces élevages mais elle se limite à cela.

B Variables

B.1 Liste fermée d'espèces

spNum	malNum	espCatListes	espCat
1	1	Pennisetum glaucum	Céréales
2	2	Sorghum bicolor	Céréales
3	3	Zea mays	Céréales
6	6	Oryza spp.	Céréales
7	7	Oryza spp.	Céréales
8	8	Triticum aestivum	Céréales
9	9	Manihot esculenta	Tubercules
17	17	Ipomoea batatas	Tubercules
21	21	Arachis hypogaea	Légumineuses
22	22	Vigna subterranea	Légumineuses
23	23	Vigna unguiculata	Légumineuses
30	30	Dolichos lablab	Légumineuses
34	34	Cyperus esculentus	Brèdes et condiments
35	35	Hibiscus sabdariffa	Brèdes et condiments
50	50	Cucurbita spp.	Brèdes et condiments
51	51	Allium cepa	Brèdes et condiments
52	52	Citrullus lanatus	Brèdes et condiments
53	53	Lycopersicum esculentum	Brèdes et condiments
54	54	Hibiscus esculentus	Brèdes et condiments
55	55	Capsicum spp.	Brèdes et condiments
56	56	Capsicum spp.	Brèdes et condiments
57	57	Solanum aethiopicum	Brèdes et condiments
60	60	Saccharum officinarum	Brèdes et condiments
5	5	Digitaria exilis suppl	Céréales
10	10	Dioscorea cayenensis-rotundata	Tubercules
15	15	Aracées	Tubercules
16	16	Aracées	Tubercules
20	20	Coleus	Tubercules
24	24	Phaseolus vulgaris suppl	Légumineuses
31	31	Glycine max	Légumineuses
27	27	Sesamum indicum	Oléagineux
58	58	Leptadenia hastata	Brèdes et condiments
28	28	Autres sésames	Oléagineux
26	26	Hyptis spicigera suppl	Légumineuses
29	29	Autres sésames	Brèdes et condiments
32	32	Amaranthus spp.	Brèdes et condiments
33	33	Amaranthus spp.	Brèdes et condiments
36	36	Autres malvacées	Brèdes et condiments
38	38	Corchorus spp.	Brèdes et condiments
39	39	Corchorus spp.	Brèdes et condiments
40	40	Solanum nigrum	Brèdes et condiments
41	41	Cleome gynandra	Brèdes et condiments
43	43	Cassia tora	Brèdes et condiments
47	47	Momordica charantia	Brèdes et condiments
49	49	Cucurbita spp.	Brèdes et condiments
37	37	Autres malvacées	Brèdes et condiments
4	4	Eleusine coracana	Céréales
11	11	Dioscorea spp.	Tubercules
12	12	Dioscorea spp.	Tubercules
13	13	Dioscorea spp.	Tubercules
19	19	Coleus	Tubercules
42	42	Justicia insularis suppl	Brèdes et condiments
44	44	Celosia argentea suppl	Brèdes et condiments
46	46	Vigna unguiculata	Brèdes et condiments
48	48	Crotalaria ochroleuca suppl	Brèdes et condiments
18	18	Tacca leontopetaloides suppl	Tubercules
25	25	Cajanus cajan suppl	Légumineuses
59	59	Luffa aegyptiaca suppl	Brèdes et condiments
45	45	Portulaca oleracea suppl	Brèdes et condiments
14	14	Dioscorea abyssinica suppl	Tubercules

Remarque : Dans deux cas, un niveau agrégé contient des espèces de différentes catégories. Dans *espCatListes*, le niveau *Vigna unguiculata* regroupe une espèce de *brèdes et condiments* et une espèce légumineuse. *Autres sésames* regroupe une espèce de *brèdes et condiments* et une espèce des *oléagineux*. Dans toute la suite, on utilisera le niveau agrégé de *espCatListes*. Et on placera *Vigna unguiculata* dans la catégorie des légumineuses et *Autres sésames* dans la catégorie des oléagineux.

B.2 Liste des villages

terNomVillage	terPays	terGPSLat	terGPSLong
Bargaja	Niger	13.17	7.05
Mamouri	Niger	13.72	13.35
CBLT	Niger	13.28	12.6
Issouri	Niger	13.23	12.41
Tam	Niger	13.13	12.13
Kill	Niger	13.31	11.92
Cheri	Niger	13.42	11.38
Gadas	Cameroun	10.19	14.43
Garey Sud	Cameroun	10.03	14.33
Moumour	Cameroun	10.12	14.32
Boumba	Cameroun	8.47	13.37
Gombo	Cameroun	8.5	13.12
Bimleru	Cameroun	8.63	12.63
Babla	Cameroun	9.26	13.55
Djaba	Cameroun	8.37	13.7
Bouzar	Cameroun	10.1	15.01
Lokoro	Cameroun	10.16	15.03
Harr	Cameroun	7.78	13.54
Mbeing I	Cameroun	7.87	14.88
Siri	Cameroun	8.02	15.24
Dobinga	Cameroun	8.96	13.9
Sirlawé	Cameroun	10.07	14.95
Soulédé	Cameroun	10.75	13.91
Dala-Zoulgo	Cameroun	10.88	14.06
Tala-Mokolo	Cameroun	10.94	14.07
Gadoua	Cameroun	10.8	14.08
Toumour	Niger	13.67	13.12
Guidan Roumdji	Niger	13.66	6.69
Karofane	Niger	14.3	6.15
Kakou	Niger	13.92	5.32

terNomVillage	terPays	terGPSLat	terGPSLong
Kilakina	Niger	13.72	10.74
Gomba	Niger	13.29	8.75
Malam Boulamari	Niger	13.19	12.23
Jigawa	Niger	13.81	9.43
Djondong	Cameroun	10.09	15.18
Nuldaina	Cameroun	10.06	15.52
Dargala	Cameroun	10.53	14.59
Modo	Tchad	12.76	17.52
Nibeck	Tchad	12.76	14.76
Télémé	Tchad	10.43	15.31
Eré	Tchad	9.76	15.8
Ham	Tchad	10	15.68
Logone Gana	Tchad	11.56	15.13
Mogrom	Tchad	11.1	15.41
Malam Sadi I	Tchad	10.56	15.18
Baki	Tchad	10.5	15.45
Gouzoudou	Cameroun	11.21	14.03
Koussouma	Cameroun	12.82	14.51
Tchika	Cameroun	12.78	14.27
Walad al Baguirmi	Tchad	12.83	16.32
Djilam Drik	Tchad	12.93	14.84
Farcha Ater	Tchad	12.43	15.21
Djangal	Cameroun	10.44	14.3
Balaza Lawan	Cameroun	10.69	14.44
Mbikou	Tchad	8.6	16.39
Mabo	Tchad	8.35	15.91
Bédaya	Tchad	8.92	17.85
Nderguigui	Tchad	8.59	17.21
Bémouli	Tchad	9.03	18.11

B.3 Variables agrosystèmes

La base de données contient beaucoup de variables et certaines contiennent des informations renseignées à l'échelle d'un agrosystème. On pet trouver ces variables dans la liste donnée en section ref. Parmi ces variables, certaines peuvent être utilisées d'un point de vue statistique, notamment pour analyser les groupes d'agrosystèmes obtenus comme *asTypologie*, *asTechJachères*, *asTechOutillage*, *asTechAmendements*, *asElevageConduite*, *asElevageImportance* ou *asPêche*.

C Outils mathématiques et statistiques

Les différents outils ne sont pas présentés dans le contexte de données d'inventaire comme dans le coeur du rapport. Cette partie est purement mathématique.

C.1 Quelques notions de théorie des graphes

Un graphe est un couple $G = (V, E)$ avec V l'ensemble des sommets, *vertices* en anglais, et $E \subset V \times V$ l'ensemble des arêtes, *edges* en anglais. On ne va considérer ici uniquement des graphes non orientés, i.e. $(v_1, v_2) \in E \Rightarrow (v_2, v_1) \in E$. On prend l'exemple $G = (\{1, 2, 3, 4, 5\}, \{(1, 2), (1, 4), (1, 5), (2, 3), (2, 5), (3, 4)\})$.

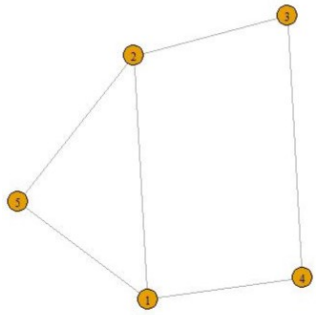
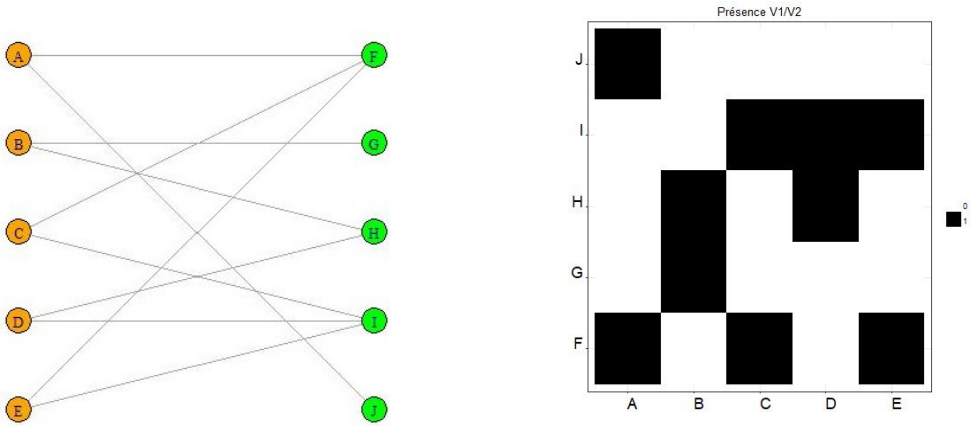


FIGURE C.1

C.1.1 Graphes bipartis

On dit qu'un graphe $G = (V, E)$ est biparti s'il existe une partition de l'ensemble de sommets V en deux sous-ensembles V_1 et V_2 telle que chaque arête ait une extrémité dans V_1 et l'autre dans V_2. On rencontre des graphe bipartis lorsqu'on a réellement une partition de nos individus ou lorsqu'on considère des sommets de nature différente. Dans l'exemple ci-dessous, on a $V = \{A, B, C, D, E, F, G, H, I, J\}$ et le graphe est biparti avec $V_1 = \{A, B, C, D, E\}$ et $V_2 = \{F, G, H, I, J\}$. Un graphe peut toujours être représenté par sa matrice d'adjacence $M = (m_{ij})_{i,j \in V} = (1_{(i,j) \in E})_{i,j \in V}$.



(a) Graphe biparti (b) Matrice d'adjacence

FIGURE C.2

Dans le cas d'un graphe biparti non orienté, on considère la matrice d'adjacence suivante $M = (1_{(i,j) \in E})_{i \in V_1, j \in V_2}$ qui contient bien la même information. Elle est représentée par la figure C.2a pour notre exemple. Ici, pour chaque couple de sommets on a une information binaire qui est la présence (1) ou l'absence (0) d'une arête entre ces deux sommets. On peut avoir des graphes plus compliqués ou on assigne une étiquette à chaque

arête. On peut toujours représenter le graphe sous la forme d'une matrice mais les coefficients ne seront plus forcément des 1 ou des 0.

C.2 Test du modèle nul

Un modèle classique de graphe aléatoire est le modèle d'Erdös-Rényi qui s'adapte parfaitement aux graphes bipartis aléatoires. On l'écrit alors,

$$m_{ij} \overset{iid}{\sim} \mathcal{B}(p), \forall i \in V_1, \forall j \in V_2.$$

Dans la suite, on parlera de modèle nul. Dans un tel modèle, le graphe ne possède aucune structure particulière (en dehors d'être biparti). On peut tester si un graphe biparti est issu du modèle nul. Si tel est le cas, il est inutile de chercher à inférer un modèle LBM. Et au contraire, si ce n'est pas le cas, cela motive la recherche d'une structure plus compliquée comme celaa peut être le cas pour un modèle LBM.
Définissons les statistiques utilisées pour tester le modèle nul. L'hypothèse nulle est la suivante

$$H_0 : \exists p \in [0,1], \forall i \in V_1, j \in V_2, X_{ij} \overset{iid}{\sim} \mathcal{B}(p).$$

On note

- $d_{1,i} = \sum_{j \in V_2} m_{ij}$ le degré du sommet $i \in V_1$,
- $d_{2,j} = \sum_{i \in V_1} m_{ij}$ le degré du sommet $j \in V_2$,
- $N = \sum_{j \in V_2} d_{2,j} = \sum_{i \in V_1} d_{1,i} = \sum_{i \in V_1, j \in V_2} m_{ij}$.

Pour tester le modèle nul, on vérifie si la répartition des degrés est uniforme pour les sommets de V_1 et pour les sommets de V_2. Sous l'hypothèse nulle, on a

- $\forall i \in V_1, d_{1,i} = \sum_{j \in V_2} m_{ij} \sim^{iid} \mathcal{B}(|V_2|, p)$,
- $\forall j \in V_2, d_{2,j} = \sum_{i \in V_1} m_{ij} \overset{iid}{\sim} \mathcal{B}(|V_1|, p)$.

On a donc $\forall i \in V_1, \mathrm{Var}(d_{1,i}) = |V_2|p(1-p)$ et $\forall j \in V_2, \mathrm{Var}(d_{2,j}) = |V_1|p(1-p)$. On utilisera les estimateurs empiriques de différentes quantités comme

- $\hat{p} = N/nm$,
- $\widehat{\mathrm{Var}}(d_1) := \frac{1}{|V_1|-1} \sum_{i \in V_1} (d_{1,i} - |V_2|\hat{p})^2$,
- $\widehat{\mathrm{Var}}(d_2) := \frac{1}{|V_2|-1} \sum_{j \in V_2} (d_{2,j} - |V_1|\hat{p})^2$.

Les statistiques de tests sont $T_1 := \frac{\widehat{\mathrm{Var}}(d_1)}{V_2\hat{p}(1-\hat{p})}$ et $T_2 := \frac{\widehat{\mathrm{Var}}(d_2)}{V_1\hat{p}(1-\hat{p})}$. Pour le test, on va devoir faire des simulations pour calculer les p-valeurs puisqu'on ne connaît pas la loi des variables T_1 et T_2. Pour $k = 1, \ldots, n_{sim}$, on simule un graphe issu du modèle nul $M^{(k)}$ de paramétre $p = \hat{p}$ et on peut alors calculer les quantités $\hat{p}^{(k)}$, $\widehat{\mathrm{Var}}^{(k)}(d_1)$, $\widehat{\mathrm{Var}}^{(k)}(d_2)$, $T_1^{(k)}$ et $T_2^{(k)}$. On a alors une estimation des p-valeurs avec $p\mathrm{val}_{L,1} := \frac{\#\{k : T_1^{(k)} < T_1\}}{n_{sim}}$, $p\mathrm{val}_{R,1} := \frac{\#\{k : T_1^{(k)} > T_1\}}{n_{sim}}$, $p\mathrm{val}_{L,2} := \frac{\#\{k : T_2^{(k)} < T_2\}}{n_{sim}}$ et $p\mathrm{val}_{R,2} := \frac{\#\{k : T_2^{(k)} > T_2\}}{n_{sim}}$. Si l'une de ces quatre p-valeurs est faible (par exemple inférieure à 0.05), alors on rejette l'hypothèse nulle.
Remarque : on teste le modèle en réalisant des simulations avec $p = \hat{p}$. On parle de bootstrap paramétrique.

C.3 Latent Block Model (LBM)

C.3.1 Définitions

Le LBM est un modèle probabiliste de graphe biparti aléatoire inspiré du modèle SBM (*Stochastic Block Model*). On peut aussi le voir comme un modèle de « matrice aléatoire par blocs ». C'est un modèle à variables

cachées (latentes), il est généralement utilisé en statistique pour faire de la (bi-)classification. Un modèle de loi paramétrique $\mathcal{F}$ avec K_1 groupes sur V_1 (lignes de la matrice) et K_2 groupes sur V_2 (colonnes de la matrice) est défini comme suit. Soit des probas multinomiales $\alpha = (\alpha_1, \ldots, \alpha_{K_1})$, $\beta = (\beta_1, \ldots, \beta_{K_2})$ et un semble de paramètres $P = (p_{st})_{\substack{1 \leq s \leq K^1 \\ 1 \leq t \leq K^2}}$. On a

$$A_i \overset{iid}{\sim} \mathcal{M}(\alpha), \forall i,$$

$$B_j \overset{iid}{\sim} \mathcal{M}(\beta), \forall j,$$

$$m_{ij} | A_i, B_j \sim \mathcal{F}(p_{A_i B_j}), \forall i, j.$$

Dans un modèle LBM, la loi du coefficient m_{ij} ne dépend que des variables cachées A_i et B_j, c'est-à-dire du groupe du sommet $i \in V_1$ et du groupe du sommet $j \in V_2$. Dans le rapport, on utilise des modèles à lois de Bernoulli et à lois de Poisson.

C.3.2 Algorithmes EM et package *blockmodels*

Pour l'estimation de modèles à variables latentes, on fait souvent appel à un algorithme EM (voir [7]). Dans le cas d'un modèle LBM, on ne peut pas appliquer directement un algorithme EM pour des motifs computationnels. On utilise des algorithmes dits « EM variationnels ». Dans tous les cas, ces algorithmes ont pour but de maximiser la vraisemblance. Le package *blockmodels* utilise un EM variationnel pour maximiser le critère ICL (*Integrated Complete Likelihood*). Les fonctions qu'on utilise renvoie différents modèles LBM. On conserve uniquement le modèle pour lequel l'ICL est maximisé. On a alors accès à la matrice P aux probas α et β. Le fonction calcule également pour chaque agrosystème et chaque espèce les probas a posteriori d'appartenance aux différents groupes. Dans toute notre analyse, on a assigné à chaque espèce le groupe qui correspond à la plus grande probabilité d'appartenance a posteriori, et on a fait de même pour les agrosystèmes. Ces probabilités a posteriori se trouvent en annexe. Cette partie est très brève mais on peut trouver toute les informations nécessaires dans les références [2] et[3].

C.4 Test d'indépendance du χ^2

Le test d'indépendance du chi^2 permet de tester si deux variables multinomiales sont indépendantes. Soit C, G deux variables multinomiales telles que $C \sim \mathcal{M}(\tau^1)$ et $G \sim \mathcal{M}(\tau^2)$. K_1 et K_2 sont connus mais pas τ^1 ni τ^2. Le test utilise les hypothèses

$$H_0 : C \perp G \text{ vs. } H_1 : {H_0}^C.$$

Le test d'indépendance du χ^2 fait parti des tests statistiques classiques et communément utilisés. On peut trouver des explications détaillées et des références à propos de ce test sur la page Wikipédia [1] associée.

1. https://fr.wikipedia.org/wiki/Test_du_%CF%87%C2%B2#Test_du_%CF%87%C2%B2_d'ind%C3%A9pendance

D Résultats

D.1 LBM à lois de Bernoulli

Les groupes 1 et 2 correspondent respectivement aux groupes orange et vert sur les différentes figures.

Agrosystème	Groupe	Probabilité a posteriori
Bimleru	1	0.998387096774194
Djaba 2	1	0.998387096774194
Harr	1	0.998387096774194
Sirlawé	1	0.998387096774194
Dobinga	1	0.998387096774194
Garey Sud	1	0.998387096774194
Lokoro	1	0.998387096774194
Mbeing I	1	0.998387096774194
Djondong	1	0.998387096774194
Gadas	1	0.998387096774194
Boumba	1	0.998387096774194
Gombo	1	0.998387096774194
Soulédé	1	0.998387096774194
Djaba	1	0.998387096774194
Siri	1	0.998387096774194
Nuldaina	1	0.998387096774194
Djangal	1	0.998387096774194
Babla	1	0.998387096774194
Bouzar	1	0.998387096774194
Dala-Zoulgo	1	0.998387096774194
Dargala	1	0.998387096774194
Moumour	1	0.998387096774194
Gadoua	1	0.998387096774194
Gouzoudou	1	0.897435654930927
Tala-Mokolo	1	0.992658531171575
Baki	1	0.972939222032925
Mabo	1	0.517203026681176
Ham	1	0.882502446211626

Agrosystème	Groupe	Probabilité a posteriori
CBLT	2	0.998387096774194
Mamouri	2	0.998387096774194
Tam	2	0.998387096774194
Jigawa	2	0.998387096774194
Balaza Lawan	2	0.984228341285321
Bémouli	2	0.824569857080803
Kill	2	0.998387096774194
Cheri	2	0.998387096774194
Guidan Roumdji	2	0.995329972114457
Mbikou	2	0.895666372220078
Gomba	2	0.998387096774194
Télémé	2	0.987145833414795
Tchika	2	0.998387096774194
Bédaya	2	0.997603617608335
Toumour	2	0.998387096774194
Kilakina	2	0.998387096774194
Nibeck	2	0.998387096774194
Issouri	2	0.998387096774194
Djilam Drik	2	0.998387096774194
Kakou	2	0.998387096774194
Malam Boulamari	2	0.998387096774194
Mogrom	2	0.998387096774194
Malam Sadi I	2	0.998387096774194
Nderguigui	2	0.998387096774194
Karofane 2	2	0.998387096774194
Modo	2	0.998387096774194
Mogrom 2	2	0.998387096774194
Eré	2	0.998387096774194
Logone Gana	2	0.998387096774194
Koussouma	2	0.998387096774194
Logone Gana 2	2	0.998387096774194
Walad al Baguirmi	2	0.998387096774194
Farcha Ater	2	0.998387096774194
Bargaja	2	0.998387096774194

D.1.1 Figures

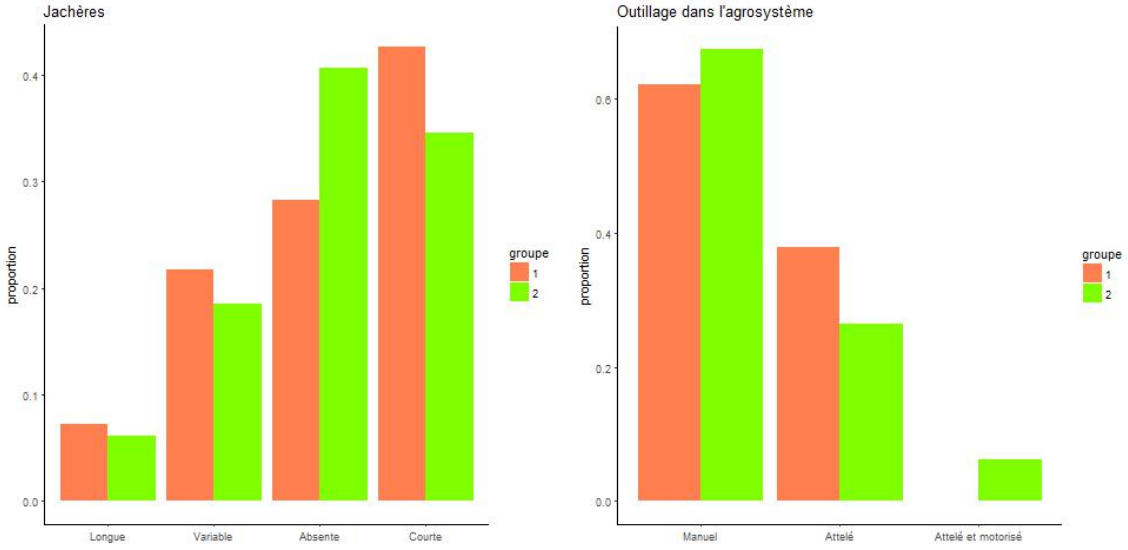


FIGURE D.1

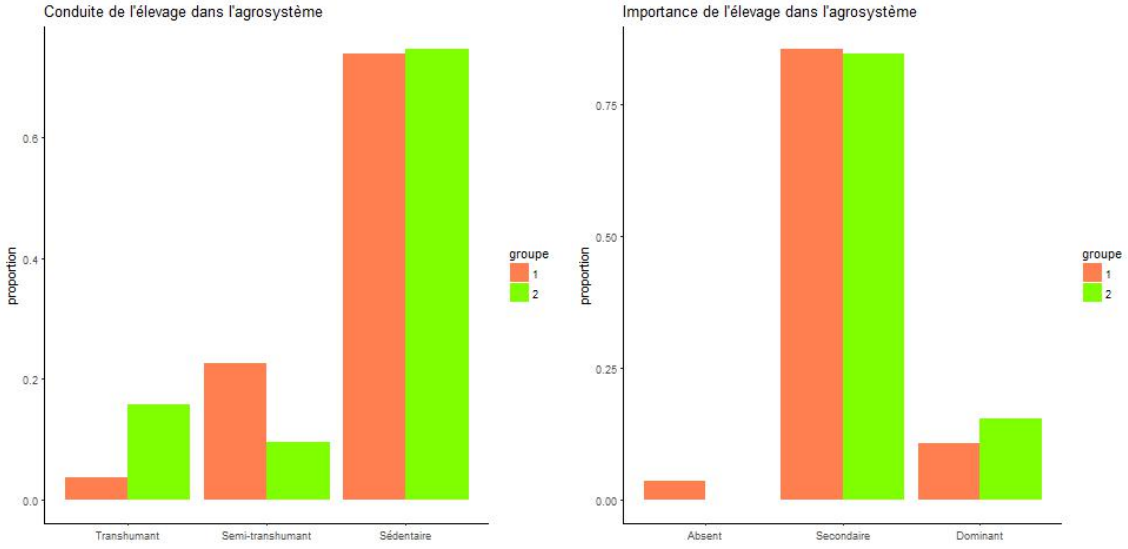


FIGURE D.2

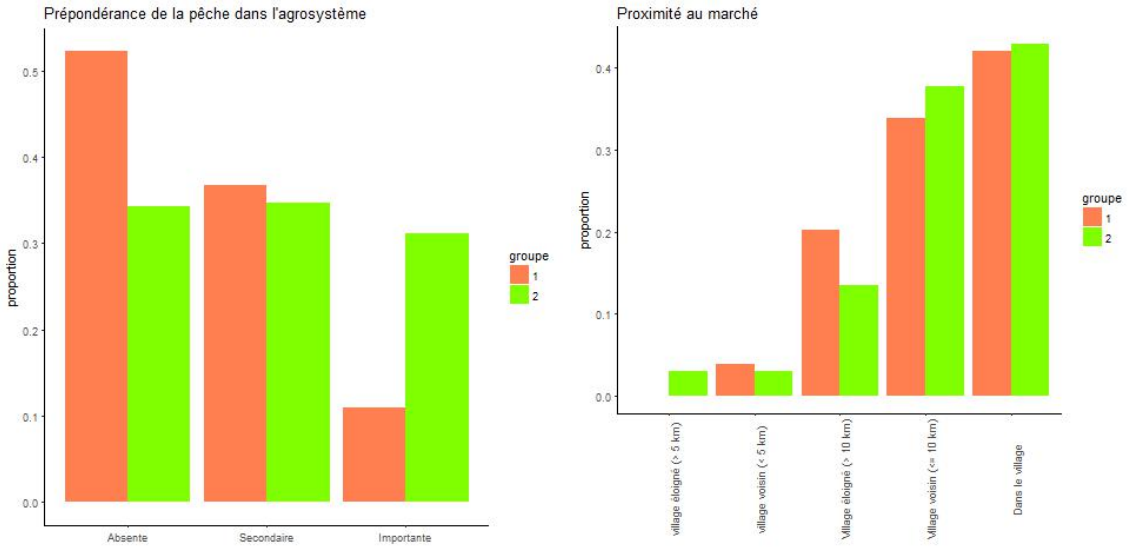


FIGURE D.3

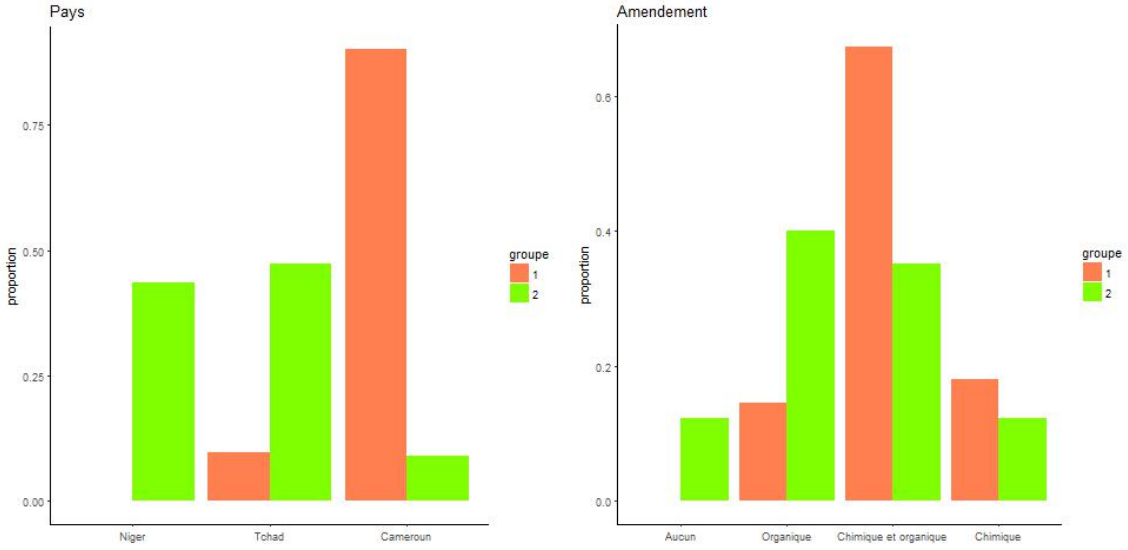


FIGURE D.4

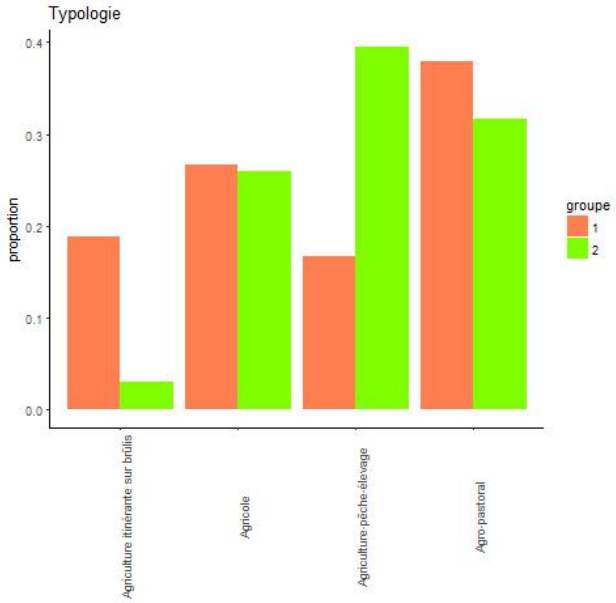

Typologie
proportion
0.4
0.3
0.2
0.1
0.0
groupe
1
2
Agriculture itinérante sur brûlis
Agricole
Agriculture-pêche-élevage
Agro-pastoral

FIGURE D.5

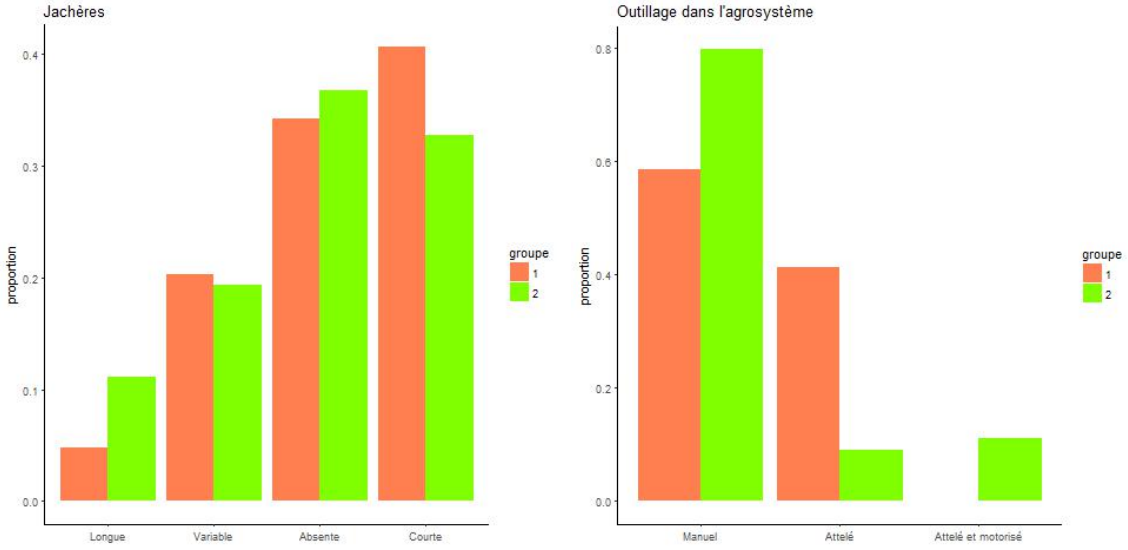
D.2 LBM à degré corrigé

Les groupes 1 et 2 correspondent respectivement aux groupes orange et vert sur les différentes figures.

Agrosystème	Groupe	Probabilité a posteriori
Gadas	1	0.998387096774194
Garey Sud	1	0.998387096774194
Moumour	1	0.998387096774194
Boumba	1	0.998387096774194
Gombo	1	0.998265226183644
Babla	1	0.998387096774194
Djaba	1	0.998387096774194
Djaba 2	1	0.998387096774194
Bouzar	1	0.998387096774194
Lokoro	1	0.998387096774194
Harr	1	0.998387096774194
Mbeing I	1	0.998387096774194
Siri	1	0.998387096774194
Dobinga	1	0.998387096774194
Sirlawé	1	0.998387096774194
Soulédé	1	0.998387096774194
Dala-Zoulgo	1	0.998387096774194
Tala-Mokolo	1	0.998387096774194
Gadoua	1	0.998387096774194
Guidan Roumdji	1	0.998387096774194
Djondong	1	0.998387096774194
Nuldaina	1	0.998387096774194
Dargala	1	0.998387096774194
Télémé	1	0.998387096774194
Eré	1	0.998387096774194
Ham	1	0.998387096774194
Logone Gana	1	0.996548947770472
Logone Gana 2	1	0.998322977211581
Mogrom	1	0.994015654609352
Mogrom 2	1	0.989078944992034
Malam Sadi I	1	0.998387096774194
Baki	1	0.998387096774194
Gouzoudou	1	0.998387096774194
Tchika	1	0.996919900088964
Djangal	1	0.998387096774194
Balaza Lawan	1	0.998387096774194
Mbikou	1	0.998387096774194
Mabo	1	0.998387096774194
Bédaya	1	0.998387096774194
Nderguigui	1	0.998387096774194
Bémouli	1	0.998387096774194

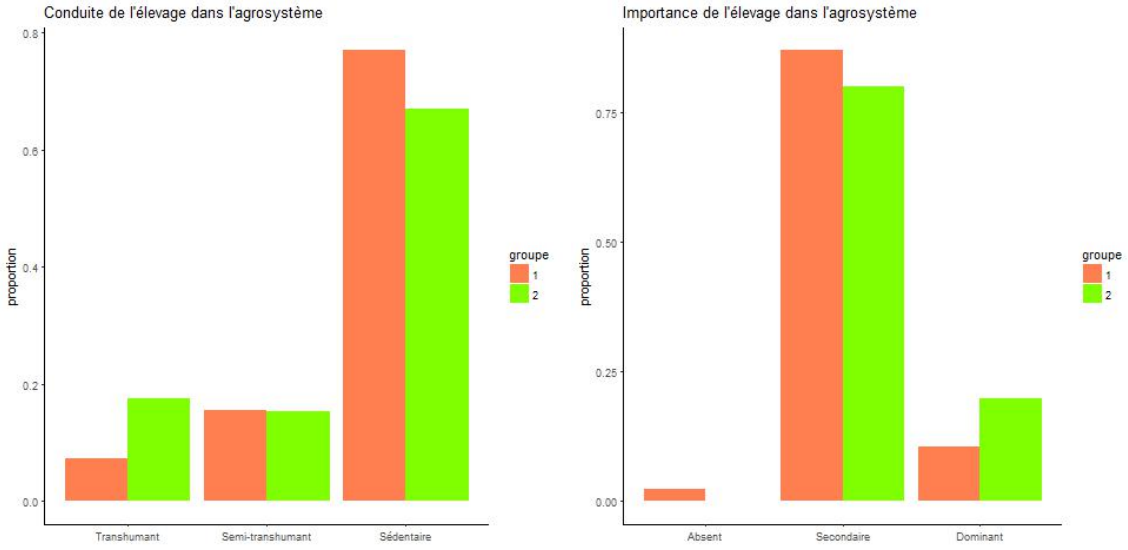
Agrosystème	Groupe	Probabilité a posteriori
Bargaja	2	0.998387096774194
Mamouri	2	0.998387096774194
CBLT	2	0.998387096774194
Issouri	2	0.998387096774194
Tam	2	0.998387096774194
Kill	2	0.998387096774194
Cheri	2	0.998387096774194
Bimleru	2	0.872621081492084
Toumour	2	0.998387096774194
Karofane 2	2	0.998387096774194
Kakou	2	0.995127021630993
Kilakina	2	0.998387096774194
Gomba	2	0.583316157901613
Malam Boulamari	2	0.998387096774194
Jigawa	2	0.998387096774194
Modo	2	0.998387096774194
Nibeck	2	0.996948837551833
Koussouma	2	0.966675252682455
Walad al Baguirmi	2	0.613015598013147
Djilam Drik	2	0.997031789467553
Farcha Ater	2	0.871970558035243

D.2.1 Figures



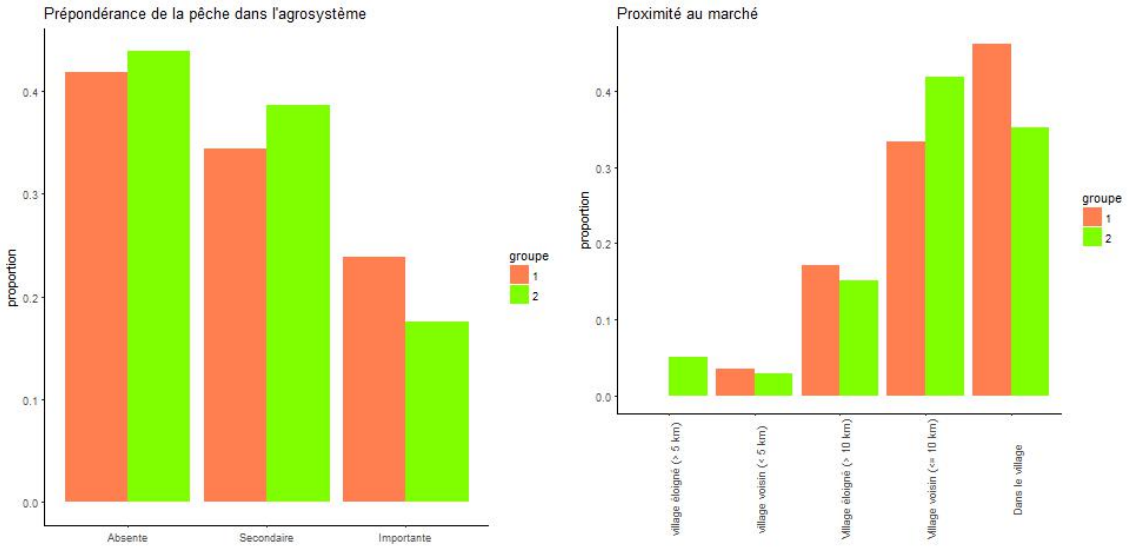
(a) (b)

FIGURE D.6



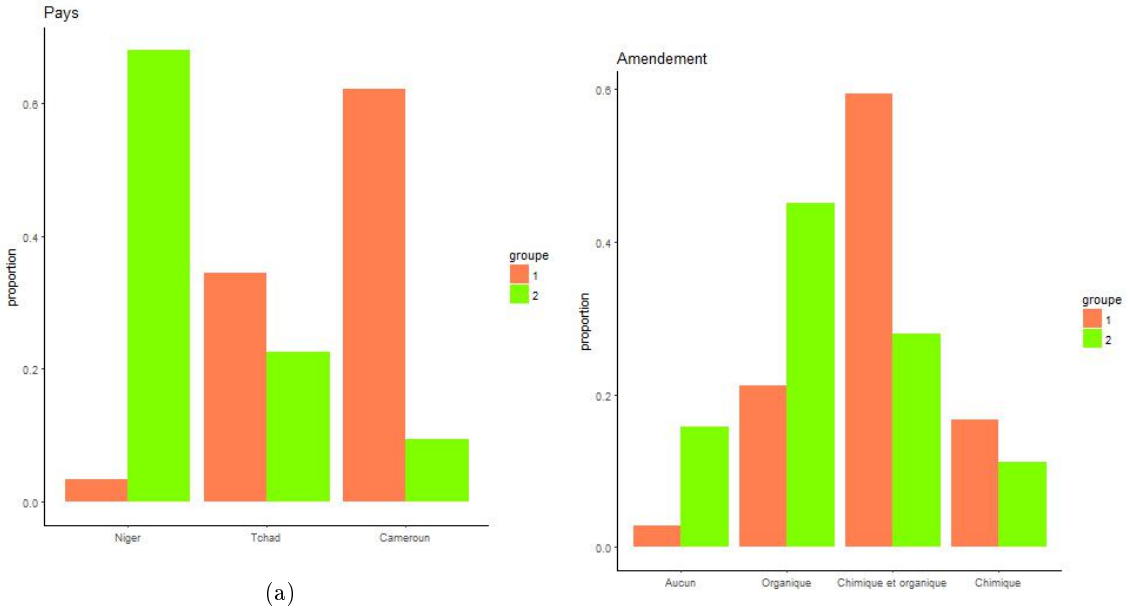
(a) (b)

FIGURE D.7



(a) (b)

FIGURE D.8



(a)

FIGURE D.9

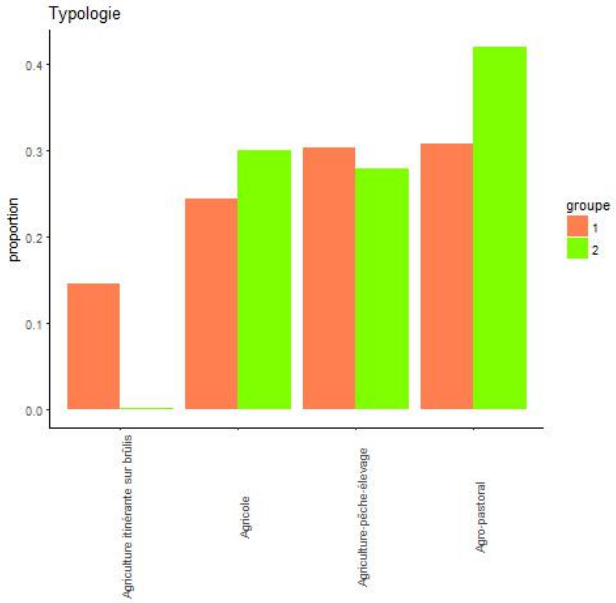
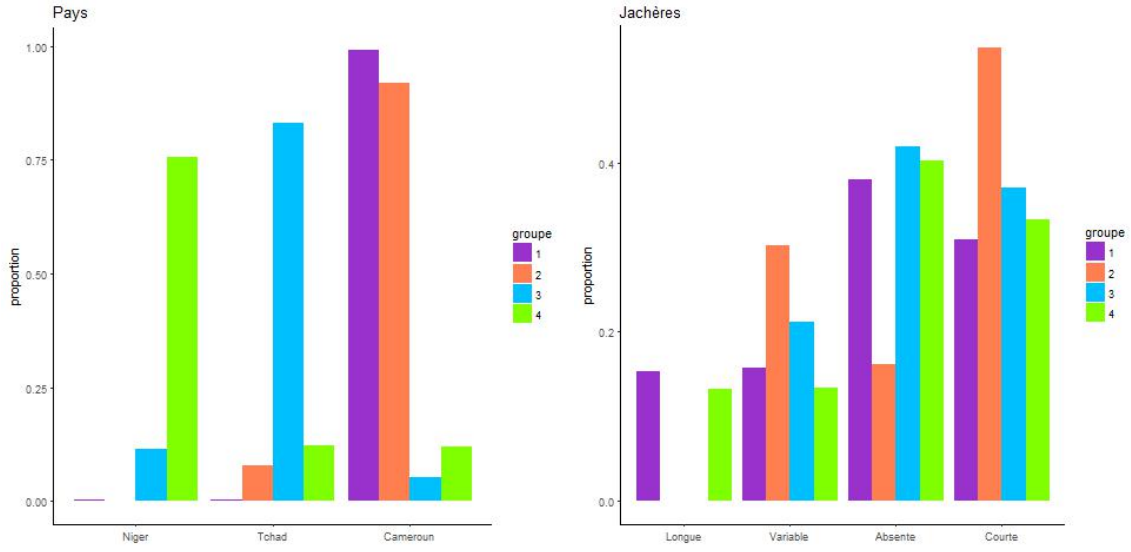

Typologie
proportion
0.4
0.3
0.2
0.1
0.0
groupe
1
2
Agriculture itinérante sur brûlis
Agricole
Agriculture-pêche-élevage
Agro-pastoral

FIGURE D.10

D.3 LBM poisson

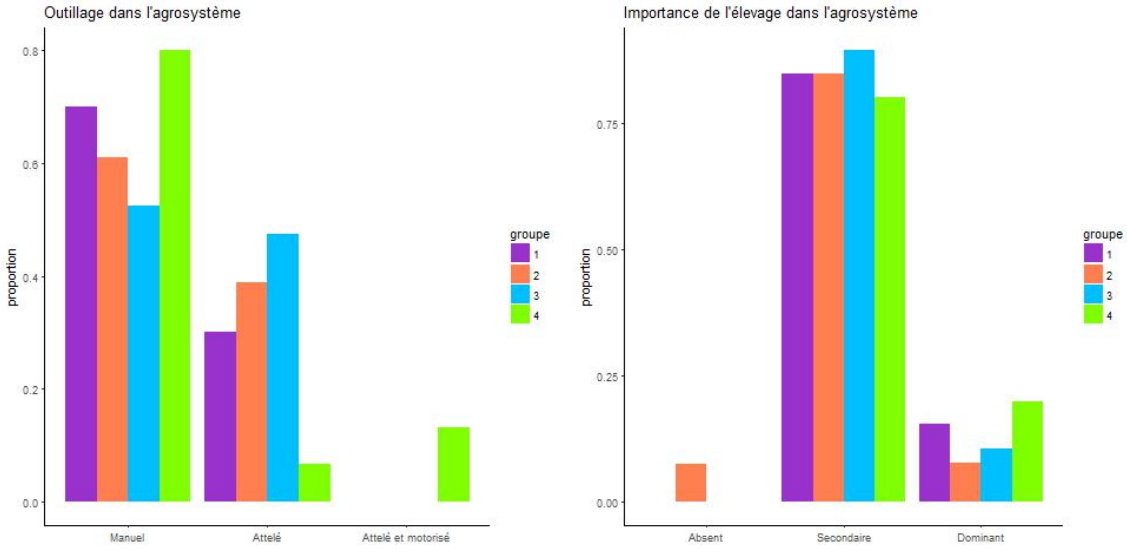
Les groupes 1, 2, 3 et 4 correspondent respectivement aux groupes violet, orange, bleu et vert.

Agrosystème	Groupe	Probabilité a posteriori
Boumba	1	0.995176848874598
Gombo	1	0.995176848874598
Bimleru	1	0.995176848874598
Babla	1	0.9029761039214
Djaba	1	0.995176848874598
Djaba 2	1	0.995176848874598
Bouzar	1	0.995176848874598
Harr	1	0.995176848874598
Mbeing I	1	0.995176848874598
Siri	1	0.995176848874598
Dobinga	1	0.995176848874598
Dala-Zoulgo	1	0.994348882487291
Gouzoudou	1	0.995176848874598
Gadas	2	0.995176848874598
Garey Sud	2	0.995176848874598
Moumour	2	0.995176848874598
Lokoro	2	0.995176848874598
Sirlawé	2	0.995176848874598
Soulédé	2	0.995176848874598
Tala-Mokolo	2	0.995176848874598
Gadoua	2	0.947836241387243
Djondong	2	0.995176848874598
Nuldaina	2	0.995176848874598
Dargala	2	0.995176848874598
Baki	2	0.995176848874598
Djangal	2	0.995176848874598
Bargaja	3	0.995176848874598
Guidan Roumdji	3	0.963673751344303
Modo	3	0.952753746864792
Télémé	3	0.992952961024328
Eré	3	0.995176848874598
Ham	3	0.992531523239874
Logone Gana	3	0.995176848874598
Logone Gana 2	3	0.995176848874598
Mogrom	3	0.995176848874598
Mogrom 2	3	0.995176848874598
Malam Sadi I	3	0.995176848874598
Walad al Baguirmi	3	0.995176848874598
Farcha Ater	3	0.995176848874598
Balaza Lawan	3	0.986210564081437
Mbikou	3	0.995176848874598
Mabo	3	0.986787303747952
Bédaya	3	0.995176848874598
Nderguigui	3	0.995176848874598
Bémouli	3	0.995176848874598
Mamouri	4	0.995176848874598
CBLT	4	0.995176848874598
Issouri	4	0.995176848874598
Tam	4	0.995176848874598
Kill	4	0.995176848874598
Cheri	4	0.995176848874598
Toumour	4	0.995176848874598
Karofane 2	4	0.803019797515509
Kakou	4	0.993742813213701
Kilakina	4	0.995176848874598
Gomba	4	0.995176848874598
Malam Boulamari	4	0.995176848874598
Jigawa	4	0.995176848874598
Nibeck	4	0.992141876462999
Koussouma	4	0.995176848874598
Tchika	4	0.995176848874598
Djilam Drik	4	0.995176848874598



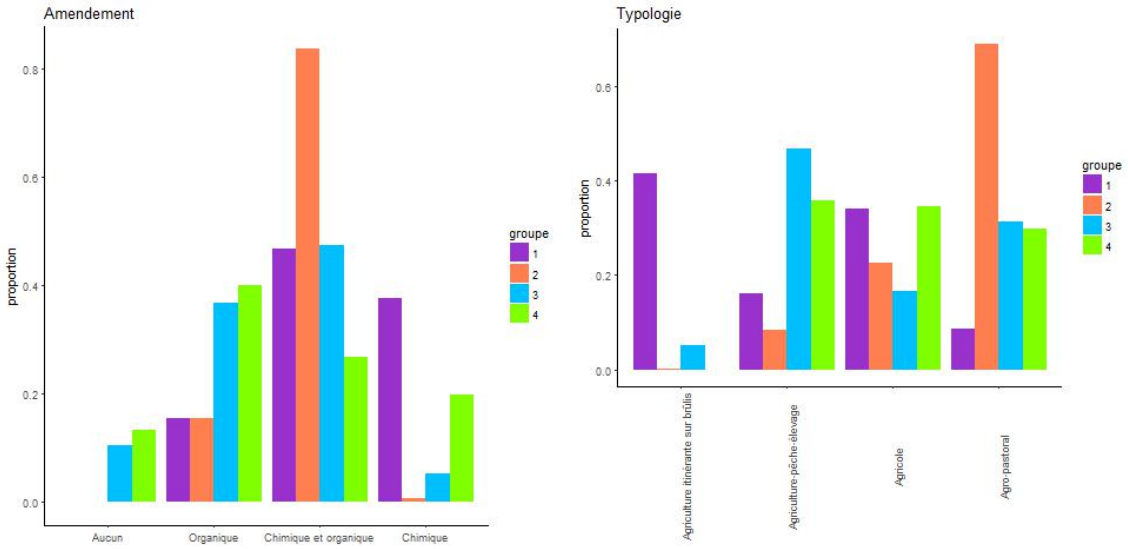
(a) (b)

FIGURE D.11



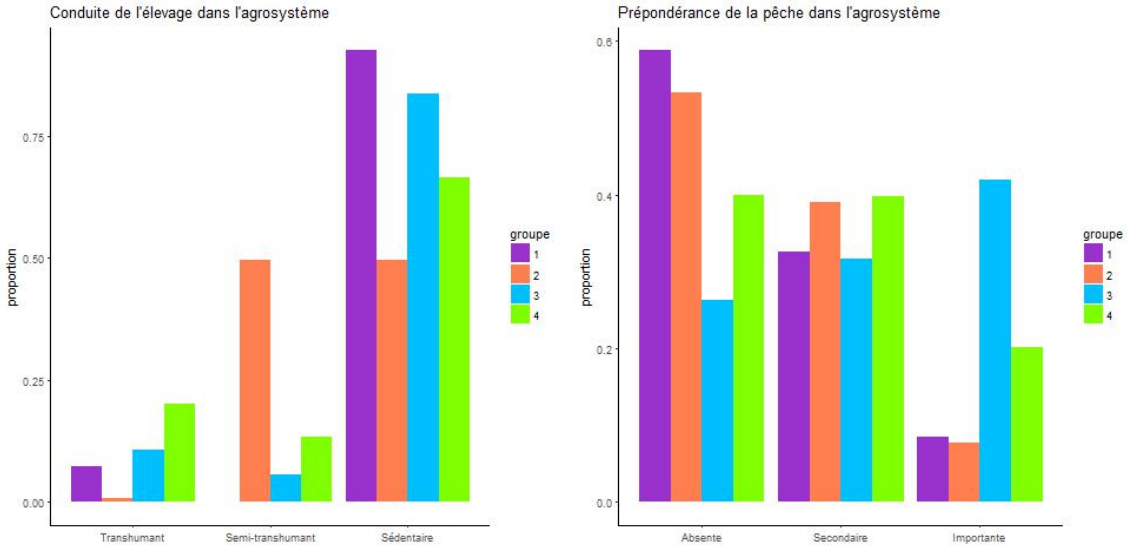
(a) (b)

FIGURE D.12



(a) (b)

FIGURE D.13



(a) (b)

FIGURE D.14

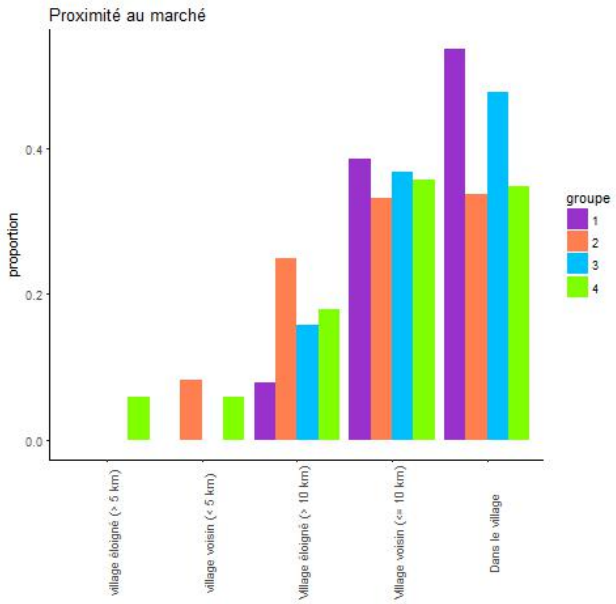

Proximité au marché
proportion
0.0
0.2
0.4
groupe
1
2
3
4
village éloigné (> 5 km)
village voisin (< 5 km)
Village éloigné (> 10 km)
Village voisin (<= 10 km)
Dans le village

FIGURE D.15