UNIVERSITÉ DU BURUNDI
GRENIER DU SAVOIR DU BURUNDI
Dépôt institutionnel officiel
https://repository.ub.edu.bi

Analyse exploratoire des piqures de moustiques, cas de l'Anopheles gambiae sensu lato et l'Anopheles funestus sensu lato au Burundi

Hategekimana, William; Sous la direction de : Dr Jean de Dieu Nkurunziza

2024
UB, ENS
https://repository.ub.edu.bi/handle/123456789/1937

Document téléchargé depuis le **Grenier du Savoir du Burundi** — https://repository.ub.edu.bi

UNIVERSITE DU BURUNDI

ECOLE NORMALE SUPERIEURE

MASTER EN DIDACTIQUE DES SCIENCES

OPTION : MATHEMATIQUES

ANALYSE EXPLORATOIRE DES PIQÛRES DE MOUSTIQUES, Cas de *l'Anopheles gambiae* sensu lato et *l'Anopheles funestus* sensu lato au Burundi.

Par

William HATEGEKIMANA

Sous la direction de :

Dr Jean de Dieu NKURUNZIZA

Mémoire présenté et défendu publiquement en vue de l' obtention du diplôme de Master en Didactique des Sciences, Option Mathématiques

Bujumbura, Juin 2024

Composition des membres du jury

President : Dr. MBAZUMUTIMA Vianney

Secretaire : Dr. SINZINKAYO Denis

Membre : Dr. NKURUNZIZA Jean de Dieu

Dédicaces

A

Ma chère épouse

MANIRAMBONA Belyse

Mon fils :

HATEGEKIMANA Yam Cherubin

Mes chers parents

NDABIHAYE Salathiel

Et

NTACOBUBAHA Edine

Mes frères et sœurs,

Toute ma famille

Et

Toutes mes connaissances

Tous les étudiants en didactique des sciences mathématiques

HATEGEKIMANA William

Remerciements

Avant toutes choses, nous voulons signaler que ce travail n'est pas le fruit de nos seuls efforts, raison pour laquelle nous présentons nos sentiments de gratitude à toutes les personnes qui ont contribué à sa réalisation.

Nos sincères remerciements sont adressés au Directeur de memoire, Professeur **Dr Jean de Dieu NKURUNZIZA** de l'Ecole Normale Supérieure. Ses précieux conseils, sa disponibilité, sa rigueur scientifique et ses suggestions nous ont été d'une grande utilité. Qu'il trouve ici l'expression de notre profonde et respectueuse gratitude.
Nos remerciements vont aussi au **Programme National Integree de Lutte contre le Paludisme (PNILP)** pour nous avoir partagee les donnees que nous avons utilisee dans notre etude, votre contribution a été inestimable pour le succes de cette etude.

Nous remercions également tout le staff de l'Institut de Pedagogie Appliquee de l'Universite du Burundi et de l'Ecole Normale Supérieure pour la formation ornée qu'il nous a donné sans passer en revers de la main. Les membres du jury pour le temps qu'ils ont bien voulu consacrer pour lire ce travail ; leurs précieux commentaires pour l'amélioration de ce travail nous seront d'une importance capitale.

Nos sentiments de reconnaissance sont également dirigés à l'endroit de tous ceux qui ont contribué à notre formation depuis le cycle primaire jusqu' à l'Ecole Normale Supérieure et à l'Université du Burundi plus particulièrement tous les professeurs du département de mathématiques de l'Ecole Normale Supérieure et de l'Institut de Pédagogie Appliquée. Qu'ils trouvent ici notre profonde gratitude pour nous avoir dotés d'une formation tant humaine que scientifique.

RÉSUMÉ

Ce mémoire explore les variations des piqûres de moustiques en se concentrant sur les espèces *An. gambiae* s.l. et *An. funestus* s.l. qui sont les espèces les plus fréquents au Burundi, les piqures qui ont été prises en considération sont de piqures qui ont été collectées entre 18h du soir et 6h du matin. L'objectif principal est d'effectuer une analyse exploratoire de ces piqures à l'aide de l'Analyse en Composantes Principales (ACP) et la Classification Ascendante Hiérarchique (CAH). Les données recueillies par le Programme National Integree de Lutte contre le Paludisme (PNILP) lors des enquêtes entomologiques menées sur la période de 2019 à 2021 ont été soumises à l'ACP pour réduire leur dimensionnalité et ensuite, la CAH a été utilisée pour regrouper les individus en clusters suivant leurs caractéristiques. Pour notre cas, nous avons considéré deux types d'individus et les résultats ont révélé que les vecteurs *An. funestus* s.l. ont été agressifs à l'intérieur en 2019 et en 2020, les vecteurs *An. gambiae* s.l. ont été agressifs à l'extérieur en 2019 et en 2020 pendant les tranches d'heure de 6-7pm, 7-8pm, 8-9pm et 9-10pm. A l'extérieur, les vecteurs *An. funestus* s.l. ont été agressifs en 2019, 2020 et 2021 pendant les tranches d'heures de 12-1 am, 1-2 am et 1-3am. Dans les mêmes tranches d'heures, les vecteurs *An. gambiae* s.l. quant à eux ont été agressifs à l'intérieur et à l'extérieur en 2019, 2020 et 2021. A l'intérieur et à l'extérieur les vecteurs *An. funestus* s.l. ont été agressifs en 2019 et 2020 à MPANDA, KIREMBA, MATANA, GIHOFI et MUTAHO et les vecteurs *An. gambiae* s.l. ont été agressifs à l'intérieur et à l'extérieur en 2019,2020 et 2021 à MABAYI, CANKUZO, VUMBI et NYANZA LAC. Ces informations sont cruciales pour la mise en place de stratégies de prévention et de contrôle des maladies transmises par les moustiques au Burundi.

Mots clés : Piqures de moustiques, tranches d'heures, Analyse en composantes principales, Classification ascendante hiérarchique, *An. gambiae* s.l. et *An. funestus* s.l.

SUMMARY

This dissertation explores variations in mosquito bites, focusing on the species *An. gambiae* s.l. and *An. funestus* s.l. which are the most frequent species in Burundi. We analysed bites occured between 18h in the evening and 6h in the morning. The main objective is to carry out an exploratory analysis of these bites using Principal Component Analysis (PCA) and Hierarchical Ascending Classification (HAC). The data collected by PNILP during entomological surveys conducted over the period 2019 to 2021 were subjected to PCA to reduce their dimensionality and then HAC was used to group individuals into clusters according to their characteristics. For our case, we considered two types of individuals and the results revealed that *An. funestus* s.l. vectors were aggressive indoors in 2019 and 2020, *An. gambiae* s.l. vectors were aggressive outdoors in 2019 and 2020 during the 6-7pm, 7-8pm, 8-9pm and 9-10pm time slots. Outdoor *An. funestus* s.l. vectors were aggressive in 2019, 2020 and 2021 during the 12-1am, 1-2am and 1-3am time slots. *An. gambiae* s.l. vectors were aggressive both indoors and outdoors during the same time periods in 2019, 2020 and 2021. *An. funestus* s.l. vectors were aggressive both indoors and outdoors in 2019 and 2020 at MPANDA, KIREMBA, MATANA, GIHOFI and MUTAHO and the vectors *An. gambiae* s.l. were aggressive indoors and outdoors in 2019, 2020 and 2021 in MABAYI, CANKUZO, VUMBI and NYANZA LAC. This information is crucial for the implementation of prevention and control strategies for mosquito-borne diseases in Burundi.

Key words: Mosquito bites, time slices, Principal component analysis, Hierarchical ascending classification, *An. gambiae* s.l. and *An. funestus* s.l.

TABLE DES MATIERES

Composition des membres du jury i
Dédicaces ii
Remerciements iii
RÉSUMÉ iv
SUMMARY v
TABLE DES MATIERES vi
SIGLE ET ABRÉVIATIONS viii
LISTE DES FIGURES ix
LISTE DES TABLEAUX x
AVANT PROPOS xi
INTRODUCTION GÉNÉRALE 1
CHAPITRE 1 : METHODOLOGIE ET DESCRIPTION DES VARIABLES 4
1.1. Zone de l'etude 4
1.2. Variables de l'étude 4
1.2.1. Les variables quantitatives 5
I.2.2. Présentation des individus 6
1.3. Méthodes d'analyse des données 8
I.4. Conclusion 8
CHAPITRE 2 : THEORIES SUR L'ANALYSE EN COMPOSANTES PRINCIPALES ET LA CLASSIFICATION ASCENDANTE HIERARCHIQUE. 9
2.1. Analyse en composantes principales (ACP) 9
2.1.1. Introduction 9
2.1.2. Principe de l'ACP 10
2.1.3. Tableaux de données 10
2.1.4. Centre de gravité 10
2.1.5. Matrice de corrélations 11
2.1.6. Le nuage des individus 12
2.1.7. Le nuage des variables. 14
2.1.8. Représentation des variables 14
2.1.9. Nombre d'axe à retenir 15
2.1.10. Interprétation et qualité de représentation d'une ACP 15
2.2 Classification hiérarchique ascendante (CAH) 17
2.2.1. 1Introduction 17
2.2.2. Modèles de classification 17

2.3. Conclusion 20

CHAPITRE 3 : ANALYSE EXPLORATOIRE DES PIQURES DE MOUSTIQUES, CAS DE *L'ANOPHELES GAMBIAE* S.L. ET L'*ANOPHELES FUNESTUS* S.L. AU BURUNDI 21

3.1. Analyses en composantes principales 21

3.1.1. Analyse en composantes principales pour les tranches d'heures 21

3.1.2. Analyse en composantes principales pour les localités (sites sentinelles) 30

3.2. Application de la CAH aux données de piqûres de moustiques 40

3.2.1. Classification ascendante hiérarchiques pour les tranches d'heures 41

3.2.2. Classification ascendante hiérarchique pour les sites sentinelles 46

3.2.3. Discussions des résultats 50

3.2.4. Les limites de l'etude 51

CONCLUSION GÉNÉRALE 53

RÉFÉRENCES BIBLIOGRAPHIQUES 55

ANNEXE A: LOGICIEL R 58

ANNEXE B : APPLICATION 58

SIGLE ET ABRÉVIATIONS

1er Q : Premier quartile

3e Q : Troisieme quartile

ACP : Analyse en composantes principales

CAH : Classification ascendante hiérarchique

CTR : Contribution

Det: Déterminant

IAF: Indoor *An. funestus* s.l.

IAG: Idoor *An. gambiae* s.l.

IC: Intervalle de confiance

IRS: Indoor Residual Spraying

ITN: Insecticide-treated Nets

Kurt: Kurtosis

Max : Maximum

Media : Médiane

MII : Moustiquaire Imprégnée insecticide

Min: Minimum

OAF: Outdoor *An. funestus* s.l.

OAG : Outdoor *An. gambiae* s.l.

OMS : Organisation Mondial de la Sante

PCA : Principal Component analysis

PID: Pulverisation Intra domiciliare

PNILP : Programme National Integre de Lutte contre le Paludisme

QLT : Qualité

Sd : Standard deviation

Skew: Skewness

USAID: United States Agency for International Development

WHO: World Health Organization

LISTE DES FIGURES

Figure 3. 1. Matrice des corrélations entre les variables 23
Figure 3. 2. projections des tranches d'heures 24
Figure 3. 3. Projections des variables 25
Figure 3. 4. Le biplot 26
Figure 3. 5. Eboulie des valeurs propres 28
Figure 3. 6. Cercle des corrélations 30
Figure 3. 7. Matrice des corrélations 32
Figure 3. 8. Projections des sites sentinelles 34
Figure 3. 9. Projections des variables 35
Figure 3. 10. Le biplot pour les sites sentinelles 36
Figure 3. 11. Cercle des corrélations 40
Figure 3. 12. Dendrogramme pour les tranches d'heures 41
Figure 3. 13. Les clusters (groupes) et des tranches d'heures en 3D 44
Figure 3. 14. Clusters (groupes) et des tranches d'heures sur la carte des facteurs 45
Figure 3. 15. Dendrogramme pour les sites sentinelles 46

LISTE DES TABLEAUX

Tableau 1. 1. Présentation des variables quantitatives 5
Tableau 1. 2. Individus selon les tranches d'heures de piqures 6
Tableau 1. 3. Individus selon les sites sentinelles 6
Tableau 1. 4. Visualisation des données sur les tranches d'heures 7
Tableau 1. 5. Visualisation des données sur les sites sentinelles 7
Tableau 3. 1. Statistiques sommaires des variables continues 21
Tableau 3. 2. Tableau des valeurs propres 27
Tableau 3. 3. Contribution pour les valeurs propres 29
Tableau 3. 4. Statistiques sommaires pour les sites sentinelles 30
Tableau 3. 5. Les valeurs propres pour les sites sentinelles 37
Tableau 3. 6. Eboulie des valeurs propres 38
Tableau 3. 7. Contribution pour les composantes principales 39
Tableau 3. 8.Caractérisation du cluster1 42
Tableau 3. 9. Caractérisation du cluster 3 43
Tableau 3. 10. Caractérisation du cluster 6 43
Tableau 3. 11. Caractérisation du cluster 1 47
Tableau 3. 12. Caractérisation du cluster 2 47
Tableau 3. 13. Caractérisation du cluster 3 48
Tableau 3. 14. Factor map des clusters (groupes) et des sites sentinelles en 3D 49
Tableau 3. 15. Factor map des clusters (groupes) et des sites sentinelles sur la carte des facteurs 50

AVANT PROPOS

Ce memoire intitulé : **« Analyse exploiratoire des piqures de moustiques , cas de l'Anopheles gambiae sensu lato et l'Anopheles funestus sensu lato au Burundi »** s'nscrit dans le cadre de la lutte contre le palludisme, un enjeu de sante publique majeur au Burundi. En tant que maladie parasitaire transmise par les moustiques du genre Anopheles, le paludisme touche des millions de personne chaque annee , et les efforts de prevention , de contrôle et d'erradication doivent etre soutenus par des donnees precises et une comprehension approfondie des vecteurs de transmission impliques. En revanche, soutenue par l'USAID, au Burundi, la surveillance entomologique est mise en œuvre à travers 9 sites sentinelles comme épine dorsale de la prise de décision en matière de lutte antivectorielle**.**

C'est ainsi que notre objectif principal est de mener une analyse exploratoire des piqures de moustiques en se concentrant sur deux espèces fréquents au Burundi dont l'*An. gambiae* s.l. et l*'An. funestus* s.l. et fournir des résultats de cette recherche afin que les décideurs puissent en tenir en considération pour éliminer du paludisme au Burundi, specifiquement il s'agit de de regrouper les tranches d'heures de piqures en utilisant l'ACP et la CAH et de regrouper les localités où sont implantées les sites sentinelles en utilisant l'ACP et CAH.

Au cours de ce travail , des difficultes n'ont pas manque. Par exemple, notre étude se limite à l'analyse des données entomologiques recueillies entre 2019 et 2021, ce qui peut ne pas refléter les dynamiques actuelles des populations de moustiques. De plus, les conditions environnementales et les pratiques humaines qui influencent les populations de moustiques peuvent avoir changé depuis la période de collecte des données, ce qui limite la généralisation de nos résultats à l'ensemble du Burundi ou à d'autres périodes.L'etude se concentre sur certaines regions du Burundi notammenet les regions où sont instalées les sites sentinelles, ceci peut limiter la representativite des resultats a l'ensemble du pays. Les données de notre étude se limites a la période allant de Janvier 2019 à Décembre 2021, ceci ne nous a pas premis de répresenter pleinement la diversité des populations de moustiques et des conditions climatiques au Burundi.

INTRODUCTION GÉNÉRALE

De nos jours, les informations nous parviennent de plusieurs canaux et différents moyens tels que la presse, les médias lourds, les réseaux sociaux, ...etc. Souvent nous sommes noyés dans ces données. L'excès des informations sous forme de données multidimensionnelles générées par différents domaines tels que le domaine de la santé, l'industrie, marketing, médical, banques, les instutitions d'enseignement sont stockées dans des bases de données, ce qui constitue un problème concret et embarrassant **(Rachid et Sabrina, 2019).**

Parmi ces données figurent les données sur les piqûres de moustiques qui constituent une exposition principale du paludisme, plus particulièrement celles des espèces qui sont responsables de la transmission du paludisme, constituant ainsi un enjeu majeur de la santé publique dans de nombreux pays du monde, y compris le Burundi **(PNILP**, **2018)**.Ce problème nécessite une exploration détaillée basée sur des normes scientifiques pour pouvoir les comprendre et les explorer.

En effet, le paludisme reste la principale cause de morbidité et de mortalité principalement en Afrique subsaharienne où plus de 95 % des cas estimés dans le monde ont été signalés par l'Organisation Mondiale de Santé **(WHO,2022).** Plus de 400 espèces d'anophèles femelles existent mais *An. gambiae* s.l. et *An. funestus* s.l. sont les principaux vecteurs de transmission de cette maladie en Afrique **(WHO, 2017).**

Au Burundi, comme dans d'autres pays d'Afrique subsaharienne, le paludisme est la principale cause de consultations ambulatoires et de mortalité hospitalière avec *An. gambiae* s.l. et *An. funestus* s.l. comme vecteurs primaires **(PNILP,2021).**

Pour lutter contre cette maladie, l'Organisation mondiale de la santé (OMS) recommande fortement la distribution à grande échelle de moustiquaires imprégnées d'insecticide et la pulvérisation intra domiciliaire à effet rémanent comme deux interventions a haut impact permettant de controler les vecteurs du paludisme **(Diawara, Konaté et al, 2021).**

Mais l'émergence de la résistance des vecteurs aux insecticides est une grave préoccupation pour les pays car l'un des éléments clés de la lutte contre le paludisme est la suppression des moustiques vecteurs notamment les anophèles femelles **(Ahan-Moss Y et al, 2020).** La surveillance entomologique est l'une des activités essentielles proposées par l'OMS pour suivre l'efficacité des interventions de lutte contre les vecteurs du paludisme**.**

La lutte antivectorielle au Burundi est principalement menée avec la distribution de moustiquaires depuis 2003 à travers le marketing social ou les campagnes de masse et les services de soins prénatals, ainsi que par la PID qui est menée régulièrement dans quatre districts sanitaires (Kiremba, Buye, Gashoho et Muyinga) depuis fin 2016 sous le soutien du Fonds mondial **(Sinzinkayo D. et al, 2021).**

Soutenue par l'USAID, la surveillance entomologique est mise en œuvre à travers 9 sites sentinelles comme épine dorsale de la prise de décision en matière de lutte antivectorielle**.**

Dans ce contexte, cette étude vise à explorer les variations des piqûres de moustiques entre **18h et 6h** du matin au Burundi, en se concentrant spécifiquement sur *An. gambiae* s.l. et *An. funestus* s.l. .En utilisant des techniques d'analyse en composantes principales **(ACP)** et de classification ascendante hiérarchique **(CAH),** cette étude cherche à identifier les structures et les tendances dans les profils de piqûres de moustiques, offrant ainsi des perspectives importantes pour la lutte antivectorielle et la gestion du paludisme dans la région.

Plusieurs facteurs sont à l'origine de la recrudescence du paludisme dans le monde. Les sites de reproduction des vecteurs, particulièrement *An. gambiae* s.l., connaissent une extension sous l'effet de la déforestation **(Tajougang et al, 1994)** et des aménagements hydro-agricoles, notamment avec l'introduction de la riziculture irriguée et la pisciculture, du fait de la pression démographique **(Philippon B. et al, 1976).** En effet, l'accroissement démographique en Afrique sub-saharienne et les irrégularités de la production agricole sous pluies incitent depuis plusieurs décennies le développement d'aménagements hydrauliques, notamment la riziculture irriguée.

De façon générale, sur le plan écologique, ces aménagements hydrauliques entraînent, de manière quasi-constante, des modifications hydrologiques géodynamiques et écosystémiques qui peuvent favoriser la prolifération de certains vecteurs de maladies infectieuses transmissibles liées à l'eau, notamment le paludisme et les schistosomoses. Les rizières constituent un modèle de modifications de l ' environnement dont l'impact entomologique et particulièrement épidémiologique est différent selon la situation locale et le niveau de stabilité du paludisme. Elles occasionnent une pullulation de moustiques, notamment les anophèles vecteurs du paludisme **(J.M.C., Doamnio et al,2002).**

L'objectif principal de cette étude est de mener une analyse exploratoire des piqures de moustiques en se concentrant sur deux espèces fréquents au Burundi dont l'*An. gambiae* s.l. et l*'An. funestus* s.l. et fournir des résultats de cette recherche afin que les décideurs puissent en tenir en considération pour éliminer du paludisme au Burundi.

Quant aux objectifs spécifiques, il s'agira :

- De regrouper les tranches d'heures de piqures en utilisant l'ACP et la CAH
- Regrouper les localités où sont implantées les sites sentinelles en utilisant l'ACP et CAH

Notre travail est subdivisé en trois chapitres. Dans le premier chapitre nous allons passer en revue la méthodologie utilisée pour récolter les données. Dans le deuxième chapitre, nous mettons en évidence la revue la théorie de l'analyse en composante principale et de la classification ascendante hiérarchique. Enfin dans le troisième chapitre il **est** question d'appliquer la théorie développée dans le second chapitre pour effectuer une analyse exploratoire des piqures de moustiques en se concentrant sur les piqures de l'*An. gambiae* s.l. et de l'*An. funestus* s.l. au Burundi.

CHAPITRE 1 : METHODOLOGIE ET DESCRIPTION DES VARIABLES

Ce présent chapitre met en évidence l'introduction, les données, les variables de l'étude, la présentation des individus et les méthodes d'analyse des données. Dans ce chapitre nous explorons la source des données qui vont être analysées dans le cadre de notre étude et nous clarifions les méthodes d'analyse des résultats et les outils à utiliser pour y parvenir.

1.1.Zone de l'etude

Le Burundi est un pays d'Afrique de l'Est d'une superficie de 27 834 km2. Le pays est subdivisé en 18 provinces sanitaires comprenant chacune 2 à 3 districts sanitaires. L'USAID PMI soutient le PNILP pour la surveillance entomologique dans plus de la moitié du pays, à travers neuf provinces sanitaires réparties de manière représentative. Les 9 sites sentinelles entomologiques sont : (1) Cankuzo, (2) Gihofi, (3) Kiremba, (4) Mabayi, (5) Matana, (6) Mpanda, (7) Mutaho, (8) Nyanza-Lac, et (9) Vumbi (**The VectorLink Burundi Project, December 2019).**

1.2.Variables de l'étude

Cette étude porte sur une analyse exploratoire des piqures de moustiques en se concentrant sur les *An. gambiae* s.l. et *An. funestus* s.l. . Les variables sont quantitatives et sont constituées des piqures de moustiques de deux vecteurs d'anopheles notamment l'*An. gambiae* s.l. et l*'An. funestus* s.l. à l'intérieur et à l'extérieur sur une période de 3 ans, cela entraine la démultiplication des variables jusqu'à en avoir douze variables.

Les données de notre étude ont été collectées par le PNILP sous financement de l'USAID dans la période allant de janvier 2019 à Décembre 2021 sur les différentes sites sentinelles entomologiques ci haut citées.

1.2.1. Les variables quantitatives

Ce tableau décrit la présentation des variables quantitatives

Tableau 1. 1. Présentation des variables quantitatives

N^{O}	Description	Variables
1	Indoor *An. gambiae* s.l. en 20219 (Piqures à l'intérieur du vecteur *An. gambiae* s.l.)	IAG2019
2	Outdoor *An. gambiae* s.l. en 2019 (Piqures à l'extérieur du vecteur *An. gambiae* s.l.)	OAG2019
3	Indoor *An. funestus* s.l. en 2019 (Piqures à l'intérieur du vecteur *An. funestus* s.l.)	IAF2019
4	Outdoor *An. funestus* s.l. en 2019 (Piqures à l'extérieur du vecteur *An. funestus* s.l.)	OAF2019
5	Indoor *An. gambiae* s.l. en 2020 (Piqures à l'intérieur du vecteur *An. gambiae* s.l.)	IAG2020
6	Outdoor *An. gambiae* s.l. en 2020 (Piqures à l'extérieur du vecteur *An. gambiae* s.l.)	OAG2020
7	Indoor *An. funestus* s.l. en 2020 (Piqures à l'intérieur du vecteur *An. funestus* s.l.)	IAF2020
8	Outdoor *An. funestus* s.l. en 2020 (Piqures à l'extérieur du vecteur *An. funestus* s.l.)	OAF2020
9	Indoor *An. gambiae* s.l. en 2021 (Piqures à l'intérieur du vecteur *An. gambiae* s.l.)	IAG2021
10	Outdoor *An. gambiae* s.l. en 2021 (Piqures à l'extérieur du vecteur *An. gambiae* s.l.)	OAG2021
11	Indoor *An. funestus* s.l. en 2021 (Piqures à l'intérieur du vecteur *An. funestus* s.l.)	IAF2021
12	Outdoor *An. funestus* s.l. en 2021 (Piqures à l'extérieur du vecteur *An. funestus* s.l.)	OAF2021

I.2.2. Présentation des individus

Les individus de l'étude sont constitués des tranches d'heure de 18h du soir à 6h du matin d'une part et d'autre part des localités **où** sont implantées les sites sentinelles.

Tableau 1. 2. Individus selon les tranches d'heures de piqures

Nº	**Description**	**Individus**
1	18h-19h	6-7pm
2	19h-20h	7-8pm
3	20h-21h	8-9pm
4	21h-22h	9-10pm
5	22h-23h	10-11pm
6	23h-Minuit	11-12pm
7	Minuit-1h	12-1am
8	1h-2h	1-2am
9	2h-3h	2-3am
10	3h-4h	3-4am
11	4h-5h	4-5am
12	5h-6h	5-6am

Tableau 1. 3. Individus selon les sites sentinelles

Nº	**Individus**
1	CANKUZO
2	GIHOFI
3	KIREMBA
4	MABAYI
5	MATANA
6	MPANDA
7	MUTAHO
8	NYANZA LAC
9	VUMBI

Le tableau 4 explore les données sur les tranches d'heures

Tableau 1. 4. Visualisation des données sur les tranches d'heures

Time	IAG 2019	OAG 2019	IAF 2019	OAF 2019	IAG 2020	OAG 2020	IAF 2020	OAF 2020	IAG 2021	OAG 2021	IAF 2021	OAF 2021
10-11 pm	584	653	143	68	446	537	100	59	385	496	108	86
11-12 pm	565	666	156	68	463	567	119	82	372	481	111	89
1-2 am	828	889	164	118	592	722	124	139	463	606	106	120
12-1 am	766	878	140	117	543	758	113	110	481	567	115	105
2-3 am	777	864	175	115	606	690	135	150	449	549	123	124
3-4 am	678	805	164	132	531	658	139	132	391	474	110	112
4-5 am	596	702	155	113	478	527	123	115	375	378	94	83
5-6 am	461	612	128	79	321	469	83	88	223	325	50	52
6-7pm	329	334	45	31	301	304	55	28	254	271	47	34
7-8pm	505	503	60	48	407	462	85	41	311	383	69	57
8-9 pm	502	554	86	52	404	465	87	62	338	411	94	65
9-10 pm	561	610	104	63	465	483	84	79	391	434	78	74

Le tableau 5 met en exergue les données sur les sites sentinelles

Tableau 1. 5. Visualisation des données sur les sites sentinelles

Localites	IAG 2019	OAG 2019	IAF 2019	OAF 2019	IAG 2020	OAG 2020	IAF 2020	OAF 2020	IAG 2021	OAG 2021	IAF 2021	OAF 2021
CANKUZO	1043	887	389	206	531	599	296	162	235	220	117	62
GIHOFI	397	434	22	26	222	196	12	8	222	211	4	13
KIREMBA	217	163	28	15	171	227	25	31	94	174	39	38
MABAYI	42	58	161	138	32	44	133	118	21	22	186	152
MATANA	2	33	0	0	1	26	2	3	0	26	0	0
MPANDA	3019	4080	149	55	2744	3430	150	151	2339	2699	190	91
MUTAHO	651	487	1	2	639	577	16	8	469	420	1	1
NYANZA LAC	1180	1066	595	370	927	884	504	349	675	720	405	245
VUMBI	601	862	175	192	289	659	108	254	378	883	163	399

1.3. Méthodes d'analyse des données

L'analyse des données a été effectué à l'aide de l'ACP qui nous a permis de réduire les dimensionnalités de la base de données en procédant par la visualisation du nuage des individus projetés, du cercle des corrélations, le biplot combinant les deux, la détermination des valeurs propres à partir desquelles nous avons déterminé les axes factoriels et enfin sélectionner les composantes principales. La CAH a été aussi utilisé pour regrouper les observations en clusters homogènes selon leur similarité par la construction de l'arbre de classification : Dendrogramme.

I.4. Conclusion

L'objectif de ce chapitre est de montrer les méthodes mises en avant pour pouvoir effectuer une analyse exploratoire des piqures de moustiques selon les tranches d'heures de piqures et selon les localités où sont implantées les sites sentinelles.

CHAPITRE 2 : THEORIES SUR L'ANALYSE EN COMPOSANTES PRINCIPALES ET LA CLASSIFICATION ASCENDANTE HIERARCHIQUE.

Ce chapitre nous retrace les théories sur lesquelles notre travail de recherche est fondée, ces theories sont entre autre l'ACP et la CAH.

2.1. Analyse en composantes principales (ACP)

2.1.1. Introduction

L'Analyse en Composantes Principales (ACP) est une méthode statistique utilisée pour explorer et résumer les variations dans un ensemble de données multivariées. L'ACP prend sa source dans un article qui a été publiee par Karl Pearson en 1901 **(Pearson, K., 1901).** Ce Karl Pearson qui est aussi le père du test du χ^2 prolonge ses travaux dans le domaine de la régression et des corrélations entre plusieurs variables. Il utilise ces corrélations non plus pour expliquer une variable à partir des autres (comme en régression), mais pour décrire et résumer l'information contenue dans ces variables. C'est dans les années 1930 que l'ACP est formalisée par l'économiste et statisticien américain Harold Hotelling **(Harold Hotelling, 1933).** Il développe par la suite l'analyse canonique des corrélations, généralisation des analyses factorielles dont fait partie l'ACP.

Son principe repose sur la transformation des variables originales en un nouvel ensemble de variables non corrélées appelées composantes principales, qui capturent le maximum de variance dans les données d'origine. L'ACP fait partie du groupe des méthodes descriptives multidimensionnelles appelées méthodes factorielles. Ces méthodes qui sont apparues au début des années 30 et ont été surtout développées en France dans les années 60, en particulier par Jean-Paul Benzecri qui a beaucoup exploité les aspects géométriques et les représentations graphiques **(Jean-Paul Benzécri, 1973)**.

En principe, il existe deux types d'analyse en composantes principales (ACP) (Principal Component Analysis, PCA en anglais). Il existe l'ACP non normée dans laquelle les variables quantitatives du tableau sont uniquement centrées (moyenne = 0) et l'ACP normée dans laquelle les variables quantitatives du tableau sont préalablement centrées réduites (moyenne = 0 et variance = 1) **(Rachid et Sabrina, 2019).**

Du fait que les variables d'un tableau sont souvent exprimées dans des unités de mesure différentes ou avec des ordres de grandeur différents (intervalles et écarts-types bien différents),

l'utilisation de l'ACP normée est bien plus courante. Elle est d'ailleurs l'option par défaut dans les fonctions R permettant de calculer une ACP. Par conséquent, nous détaillons dans cette section uniquement l'ACP normée.

Autrement dit, le recours à une ACP non normée est plus rare et s'applique uniquement à la situation suivante : toutes les variables du tableau sont mesurées dans la même unité; il pourrait être ainsi judicieux de conserver leurs variances respectives**.**

2.1.2. Principe de l'ACP

D'un point de vue géométrique, il n'est pas facile de représenter graphiquement plusieurs variables par un nuage de points **(Pr Zineb SAYL,2020).** Pour cette raison l'Analyse en Composantes Principales (ACP) est venue comme une réponse car elle permet de résumer les informations contenant dans un tableau a n individus et p variables et de remplacer les p variables avec q nouvelles variables (q < p), on cherche à définir q nouvelles variables dites combinaison linéaire des p variables initiales contenant le plus d'informations possible.

Ces nouvelles variables portent le nom de "composantes principales ", les axes qu'elles déterminent sont " les axes principaux " et les formes linaires associées sont " les facteurs principaux".

2.1.3. Tableaux de données

Considérons pour notre cas, pour les tranches d'heures, nous avons n=12 individus (les tranches d'heures) caractérisés par p=12 variables (les piqures de moustiques), ces données sont présentées dans un tableau rectangulaire appelé la matrice des données de dimension 12*12.

Les données sont les mesures effectuées sur n individus ou unités $\{e_1, e_2, e_3 \dots \dots, e_{12}\}$.
Les p variables quantitatives qui représentent ces mesures sont $\{x_1, x_2, x_3 \dots \dots, x_{12}\}$.
On possède ainsi un tableau rectangulaire de mesure noté X dont les colonnes sont les variables et les lignes représentent les individus statistiques **(Rachid et Sabrina, 2019)**.

2.1.4. Centre de gravité

Afin de faciliter les représentations graphiques du nuage de points, on choisit en général comme origine le centre de gravité du nuage. Le nuage se retrouve alors centré autour de cette nouvelle origine.

Le centre de gravité est le vecteur des moyennes arithmétiques de chaque variable, on le note souvent par g appelé aussi individu moyen ou point moyen. Pour cela, on définit le centre de gravité du nuage des individus par :

$$\mathbf{g} = (\overline{\mathbf{x_1}}, \overline{\mathbf{x_2}}, \dots\dots\dots \overline{\mathbf{x_{12}}})^{\mathbf{t}} \ \textbf{avec} \ \overline{\mathbf{x}}_{\mathbf{j}} = \frac{\mathbf{1}}{\mathbf{12}} \sum_{\mathbf{i=1}}^{\mathbf{12}} \mathbf{x_{ij}}. \qquad \textbf{(2.1)}$$

Le centrage des données nous permet de ramener toutes les colonnes de X a la même origine zéro dans une matrice appelée matrice centrée qui est notée par X_c de terme général :

$$\mathbf{X_c} = \mathbf{x_{ij}} - \overline{\mathbf{x}}_{\mathbf{j}} \qquad \textbf{(2.2)}$$

Le centrage et la réduction des données nous permet de ramener toutes les variables a un même écart-type 1, d'après le tableau X_c on construit un tableau standard noté par X_{cr} de terme général :

$$\mathbf{X_{cr}} = \frac{\mathbf{x_{ij}}}{\mathbf{s_j}} = \frac{\mathbf{x_{ij}} - \overline{\mathbf{x}}_{\mathbf{j}}}{\mathbf{s_j}} \qquad \textbf{(2.3)}$$

avec s_j l'écart-type or l'écart-type provient de la variance qui est donnée par la formule :

$$\mathbf{s_j^2} = \frac{\mathbf{1}}{\mathbf{12}} \sum_{\mathbf{i=1}}^{\mathbf{12}} (\mathbf{x_{ij}} - \overline{\mathbf{x}}_{\mathbf{j}})^{\mathbf{2}}. \qquad \textbf{(2.4)}$$

Nous avons devant nous trois matrices X qui est la matrice des données brutes, X_c la matrice des données centrées et X_{cr} la matrice des données centrées réduites pour décrire les mêmes individus ce qui implique que nous aurons trois nuages de points différents.

2.1.5. Matrice de corrélations

La matrice de corrélation est l'ensemble des coefficients de corrélation regroupé dans un même tableau noté par R dont les termes diagonaux valent 1.

$$r_{x^i x^j} = \frac{\mathrm{cov}(x^i, x^j)}{S_{x^i} S_{x^j}} = \frac{\sum_{i=1}^{12} (x_k^i - \bar{x}^i)(x_k^j - \bar{x}^j)}{\sqrt{\sum_{i=1}^{12} (x_k^i - \bar{x}^i)^2} \sqrt{\sum_{i=1}^{12} (x_k^j - \bar{x}^j)^2}} \qquad \textbf{(2.5)}$$

L'interprétation du coefficient de corrélation se fait comme suit :

Le coefficient de corrélation varie entre -1 et 1 et son interprétation est la suivante :

- ✓ Si r est proche de 0, il n'y a pas de relation linéaire entre x et y ;
- ✓ Si r est proche de -1, il existe une forte relation linéaire négative entre x et y ;
- ✓ Si r est proche de 1, il y a une forte relation linéaire entre x et y.

La matrice de variances-covariances ou la matrice des corrélations se décompose en valeurs propres et en vecteurs propres. Les valeurs propres d'une matrice A sont les scalaires λ tels que :

$$Ax = \lambda X$$

$$\Leftrightarrow (A - \lambda I) = 0 \qquad \textbf{(2.6)}$$

Où I est une matrice unité d'ordre 12. En développant le déterminant suivant, on obtient le polynôme caractéristique de A dont les racines sont ses valeurs propres.

$$\det(A - \lambda I) = 0 \qquad \textbf{(2.7)}$$

$$\det\begin{pmatrix} a_{11} - \lambda & a_{12} & \dots. & a_{1.12} \\ a_{21} & a_{22} - \lambda & \dots.. & . \\ . & . & & . \\ a_{12.1} & a_{12.2} & \dots... & a_{12.12} - \lambda \end{pmatrix} = 0$$

Pour obtenir les vecteurs propres de la matrice A, on diagonalise cette matrice. Une matrice carrée A est dite diagonalisable s'il existe une matrice inversible P et une matrice diagonale D satisfaisant la relation :

$$A = PDP^{-1}$$

Où les λ_i sont les valeurs propres de la matrice A et aussi la diagonale de D et les colonnes de P sont les vecteurs propres associés.

2.1.6. Le nuage des individus

Chaque individu étant un point défini par 12 coordonnées est considéré comme un élément d'un espace vectoriel $\mathbb{R}^{12}$ appelé l'espace des individus. L'ensemble des 12 individus est un " nuage " de points dans $\mathbb{R}^{12}$ et g en est le centre de gravité. L'espace $\mathbb{R}^{12}$ est muni d'une structure euclidienne afin de pouvoir définir des distances entre individus**.**

Pour faire une représentation géométrique des individus, il faut choisir une distance entre deux points de l'espace.

La distance en analyse de données est définie comme une fonction (d) permettant de déterminer à quel point deux observations sont semblables ou différentes l'une de l'autre. Elle doit respecter certaines conditions comme la positivité, le principe d'identité des indiscernables, la symétrie et l'inégalité triangulaire.

La distance la plus couramment utilisée par l'analyse des composantes principales dans un espace où sont représentées les unités (individus) est **la distance euclidienne classique** .La distance entre deux unités (individus) e_i et e_j est défini par :

$$d^2(e_i, e_j) = \sum_{j=1}^{12}(e_i - e_j)^2 \qquad \textbf{(2.8)}$$

L'inertie est une quantité permettant de décrire la dispersion des observations d'un jeu de données. Cette mesure dépend à la fois des données (nombres d'observations et de variables, échelle des variables) et de la mesure de distance retenue entre deux observations. Plus spécifiquement, l'inertie correspond à la somme des distances entre chaque observation et le centre du jeu de données.

$$I_{tot} = \frac{1}{12}\sum_{i=1}^{12} d^2(x_i, g)$$

$$I_{tot} = \frac{1}{12}\sum_{i=1}^{12}\sum_{j=1}^{12}(x_{ij} - \overline{x_j})^2 \quad \textbf{(2.9)}$$

L'inertie mesure la dispersion des points du nuage par rapport à son centre de gravité : plus l'inertie est grande, plus le nuage est dispersé, et à l'inverse plus l'inertie est petite, plus le nuage est concentré autour de son centre de gravité.

$$I_{tot} = \sum_{j=1}^{12} s_j^2 \qquad \textbf{(2.10)}$$

L'inertie totale du nuage de points est donc égale à la somme des variances des variables. Dans le cas où les variables sont centrées réduites, la variance de chaque variable vaut 1. L'inertie totale est alors égale à 12 (nombre de variables).

La recherche d'axes portant le maximum d'inertie équivaut à la construction de nouvelles variables de variance maximale.

En d'autres termes on procède au changement de repère dans $\mathbb{R}^{12}$ de façon **à** se placer dans un nouveau système de représentation ou le premier axe apporte le plus possible de l'inertie totale du nuage, le deuxième axe le plus possible de l'inertie non prise en compte par le premier, et ainsi de suite. Cette réorganisation s'appuie sur la diagonalisation de la matrice de variances covariances.

Les axes principaux de l'inertie sont des axes de direction des vecteurs propres de V normés à 1. Ces axes sont donc au nombre de p. Le premier axe U^1 est celui associe à la plus grande valeur propre, le deuxième axe U^2 est celui associe à la deuxième valeur propre, …

A chaque axe principal il est associé une variable **appelée composante principale**.

La composante C^1 est le vecteur renfermant les coordonnées des projections des individus sur l'axe 1.

La composante C^2 est le vecteur renfermant les coordonnées des projections des individus sur l'axe 2.

Pour obtenir ces coordonnées, il suffit d'écrire chaque composante principale comme une combinaison linéaire des variables initiales.

$$C^1 = U_1^1 X^1 + U_2^1 X^2 + U_3^1 X^3 + \cdots + U_p^1 X^p \qquad \textbf{(2.11)}$$

La j^{eme} composante principale fournit les coordonnées des n individus sur le j^{eme} axe principal. Dans ce cas, si on désire **une représentation plane des individus**, la meilleure sera celle réalisée grâce aux deux premières composantes principales.

Dans le cas où, on travaille avec un tableau centré réduit X_{cr} associé à X, on utilisera la métrique M = 1, ce qui implique que la matrice de variance-covariance et les facteurs principaux sont tout simplement les vecteurs propres de la matrice de corrélation C rangé selon l'ordre décroissant des valeurs propres **(Ouattara, 2012).**

2.1.7. Le nuage des variables.

Chaque variable x_i est en fait une liste de 12 valeurs numériques. On la considère comme un vecteur x_i d'un espace $\mathbb{R}^{12}$, à n dimensions appelées espace des variables.

Dans une ACP normée, les liaisons entre les variables deux à deux sont évaluées avec le coefficient de corrélation qui est donnée par l'expression (2.5), soit la moyenne du produit des deux variables centrées réduite.

2.1.8. Représentation des variables

Dans la représentation des variables on s'intéresse beaucoup au cosinus de l'angle θij entre deux variables i et j.

$$\cos\theta_{ij} = \frac{<x^i, x^j>}{\|x^i\|\|x^j\|} = \frac{S_{ij}}{S_i S_j} \qquad \textbf{(2.12)}$$

Bref, le cosinus de l'angle entre deux variables s'interprète comme une corrélation.

Si pour le cas du nuage des individus on s'interesse aux distances entres points, pour le cas du nuage des variables on s'intéressera plutôt aux angles **(Rachid B. et Sabrina T., 2019).**

Les « proximités » entre **les composantes principales et les variables initiales** sont mesurées par les covariances, et surtout les coefficients de corrélation. La corrélation entre une composante principale C_j et une variable initiale x_i est égale à :

$$r(x_i, c_j) = \frac{\mathrm{cov}\,(x_i, c_j)}{S_i S_j} \qquad \textbf{(2.13)}$$

Et comme les variables sont centrées :

$$\cos\theta_{ij} = r(x_i, C_j) \qquad \textbf{(2.14)}$$

2.1.9. Nombre d'axe à retenir

Dans la plupart des cas la qualité des estimations auxquelles conduit l'ACP dépend, de façon évidente, du choix de nouvelles variables, c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentation **(Rachid B. et Sabrina T., 2019).** Différentes procédures sont complémentaires notamment la règle de Kaiser, la part de l'inertie, la règle du coude, …

2.1.10. Interprétation et qualité de représentation d'une ACP

L'ACP construit de nouvelles variables dites artificielles, et des représentations graphiques permettant de visualiser les relations entre variables, ainsi que l'existence d'éventuels groupes d'individus et de groupes de variables.

L'objectif est d'obtenir la représentation des individus dans un espace de dimension plus faible que 12.

- **Interprétation des axes**

Pour chaque axe retenu, il est nécessaire de vérifier les variables et les individus qui participent le mieux à la formation de l'axe. Pour ces raisons, il faut chercher la contribution des individus,

la contribution des variables à la formation de la composante principale, il faut aussi déterminer la qualité de la représentation des individus et la qualité de la représentation des variables.

- **La contribution de l'individu i à la composante C_k est défini par** :

$$\mathrm{CTR}_k(e_i) = \frac{p_i c_{ki}^2}{\lambda_k} \qquad \textbf{(2.15)}$$

En pratique, nous retenons pour l'interprétation les individus dont la contribution est supérieure à la contribution moyenne. La séléction des individus suit la formule suivante : $|c_{ik}| > \sqrt{\lambda_k}$.

- **La contribution des variables j à la composante ck est défini par** :

$$\mathrm{CTR}_k(x_j) = \frac{r^2(c^k, x^j)}{\lambda_k} = (u_j^k)^2 \textbf{ (2.16)}$$

Nous retenons les variables dont la contribution est supérieure à la contribution moyenne,

$$|u_{jk}| > \frac{1}{\sqrt{p}}$$

Une fois les axes interpréter, nous regardons les graphiques et analyser les proximités entre les points.

- **La qualité de représentation de l'individu i sur l'axe k :**

$$\mathrm{QLT}_k(e_i) = \cos^2(\theta_{ik}) = \frac{c_{ki}^2}{\|e_i\|^2} \quad \textbf{(2.17)}$$

Lorsque l'angle est proche de 0, le cosinus est proche de 1, donc ce cas, l'individu est bien représenté. Dans le cas inverse, l'individu est mal représenté.

- **La qualité de représentation de la variable j sur l'axe k :**

$$\mathrm{QLT}_k(x_j) = \cos^2(\theta_{ik}) = \frac{c_{ki}^2}{\|x_j\|^2} \quad \textbf{(2.18)}$$

Une variable est d'autant mieux représentée sur un axe qu'elle est proche du bord du cercle des corrélations et de l'axe, d'autant plus mal représentée qu'elle est proche de l'origine.

Il faut chaque fois choisir le nombre de composantes à retenir, c'est- à-dire la dimension des espaces de représentation. Les axes factoriels sont interprétés par rapport aux variables initiales bien représentées. Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales **(Ouattara, 2012).**

La classification ascendante hiérarchique (CAH) étant une méthode d'analyse statistique qui permet de regrouper des individus ou des variables en fonction de leurs similarités. Lorsqu'elle est combinée avec l'analyse en composante principale (ACP), elle peut être utilisée pour explorer et regrouper les données sur les piqûres de moustiques causées par l'Anophèle Gambiae et l'Anophèle Funestus au Burundi. Pour cette raison, parlons un peu de la classification ascendante hiérarchique.

2.2 Classification hiérarchique ascendante (CAH)

2.2.1. 1Introduction

La classification consiste à rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque individu pris deux à deux. Elle est dite classification ascendante hiérarchique si elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes **(Gabor J. Székely et al., 2005)**.

Les modèles usuellement utilisés en classification sont les partitions et les hiérarchies. Dans les deux cas, les objets qui se ressemblent sont regroupés en classes. Pour les partitions, les classes sont deux à deux disjointes, pour les hiérarchies, elles peuvent être emboîtées.

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque individu pris deux à deux.

L'objectif de cette méthode pour notre cas est d'effectuer une classification des tranches d'heures de piqures de moustiques au Burundi cas de l'anophèle gambiae et de l'anophèle funestus.

La classification est alors la recherche d'une typologie ou segmentation c'est-à-dire d'une partition ou répartition des individus en classes ou catégories. Ceci est fait en optimisant un critère visant à regrouper les individus dans des classes les plus homogènes et les plus distinctes entre elles. Cet objectif est à distinguer des procédures de discrimination ou encore de classement (en anglais classification) pour lesquelles une typologie est a priori connue, au moins pour un échantillon d'apprentissage. Nous sommes dans une situation d'apprentissage non-supervise ou en anglais de clustering (**Alain B. et al., 2004).**

2.2.2. Modèles de classification

En classification, pour notre cas, toute l'information utile est contenue dans un tableau 12×12 donnant les dissemblances entre les n individus à classer. La notion de dissemblance est plus large que celle de distance.

2.3.1.1.Dissimilarité (Indice de dissemblance) et distance

La notion de dissimilarité est essentielle en classification **(Ouattara, 2012).** Une dissimilarité sur un ensemble d'objets X est une fonction d de $X \times X$ dans l'ensemble des réels positifs $\mathbb{R}+$ telle que les propriétés ci-dessous soient satisfaites :

- $d(x, y) \geq 0$ pour tous x, y appartenant à X (positivité)
- $d(x, x) = 0$ pour tout x appartenant a X
- $d(x, y = d(y, x)$ pour tous x, y appartenant à X (Symétrie)
- $d(x, y) = 0$ ce qui implique que x=y pour tous x,y appartenant à X
  Si cette 4ème propriété est vérifiée nous avons une dissimilarité propre.

Une dissimilarité propre d est dite distance si elle satisfait l'inégalité triangulaire donnée par l'expression suivante :

- $d(x, y) \leq d(x, z) + d(z, y)\ ,\ \forall\, x, y \in X$

Un sous ensemble K de 2^X (ensemble des parties de X) sera appelé symétrie de classes sur X si et seulement si, il vérifie les propriétés suivantes :

- $X \in K$ et $\emptyset \in K$,
- pour tout $x \in X$, $\{x\} \in K$
- pour tous $A, B \in K, A \cap B \neq \emptyset$ entraine $A \cap B \in K$

Par définition, une hiérarchie est un système de classe pour lequel, deux classes sont toujours incluses l'une dans l'autre, soit d'intersection vide. De ce fait, une hiérarchie est représentée sous la forme d'un arbre appelé plus précisément dendrogramme où chaque classe est représentée par un segment.

Un sous ensemble $R = \{p_1, p_2, \ldots \ldots, p_k\}$ de 2^X sera appelé recouvrement de X si et seulement si, il vérifie les propriétés suivantes :

- pour tout $1 \leq i \leq k, p_k \neq \emptyset$
- pour tous $1 \leq i \neq j \leq k, p_i \not\subset p_j$ et $p_j \not\subset p_i$
- $p_1 \cup p_2 \cup \ldots \ldots \cup p_k = K$

Une partition P est un recouvrement tel que les classes sont donc deux à deux disjointes. Formellement, on a : pour toute classes $A\ et\ B\ de\ P : A \cap B = \emptyset\ si\ A \neq B$.

2.3.1.2.Méthodes de partitionnement

Une partition $P = \{C1, C2, \cdots, Ck\}$ sur X est de meilleure qualité si l'inertie intraclasse est minimisée. La partition à n éléments possède une inertie intraclasse nulle. L'objectif est donc de trouver une partition $k < n$ qui minimise l'inertie intraclasse. Il est irréaliste de vouloir essayer toutes les partitions possibles car le nombre de partitions possibles explose exponentiellement avec n. Les algorithmes de partionnement utilisés sont tous des heuristiques. Ils trouvent la plupart du temps une partition acceptable, mais sans garantie d'optimalité **(Ouattara, 2012).**

2.3.1.3. Etapes de l'algorithme des k-means

Les paragraphes suivants présentent les principales étapes de l'algorithme des $k\ means$ ou des centres mobiles :

1. A l'étape initiale, on choisit au mieux k centres de classe provisoires formés des points individus, appelés aussi noyaux. Les k noyaux $\{x(1) , \cdots , x(k)\}$ permettent de construire une classification initiale $\{c_1^0,\ c_2^0 \ldots\ldots, c_k^0\}$ des n individus en affectant chaque individu j à la classe dont il est le plus proche ($d\ (xj\ ,\ x\ (i))\ minimum$). On calcule ensuite les centres de gravité : $\{g_1^0, g_2^0, \ldots\ldots, g_k^0\}$ des classes $\{c_1^0,\ c_2^0 \ldots\ldots, c_k^0\}$. Ces centres de gravité $\{g_1^0, g_2^0, \ldots\ldots, g_k^0\}$ deviennent les nouveaux centres de classe.
2. A l'étape 2, les nouveaux centres de classe $\{g_1^0, g_2^0, \ldots\ldots, g_k^0\}$ permettent d'obtenir une nouvelle partition $\{c_1^1,\ c_2^1 \ldots\ldots, c_k^1\}$ en affectant chaque individu j à la classe dont il est le plus proche $(d(x_j, g_i^0)\text{minimum})$. On calcule ensuite les centres de gravité $\{g_1^1, g_2^1, \ldots\ldots, g_k^1\}$ des classes $\{c_1^1,\ c_2^1 \ldots\ldots, c_k^1\}$. Ces centres de gravité $\{g_1^1, g_2^1, \ldots\ldots, g_k^1\}$ deviennent les nouveaux centres de classe.
3. A l'étape m, les centres de classe $\{g_1^{m-1}, g_2^{m-1}, \ldots\ldots, g_k^{m-1}\}$ obtenus à l'étape précédente permettent d'obtenir une nouvelle partition $\{c_1^m,\ c_2^m \ldots\ldots, c_k^m\}$ en affectant chaque individu j à la classe dont il est le plus proche $(d(x_j, g_i^{m-1})\text{minimum})$. On calcule ensuite les centres de gravité $\{g_1^m, g_2^m, \ldots\ldots, g_k^m\}$ des classes $\{c_1^m,\ c_2^m \ldots\ldots, c_k^m\}$. Ces centres de gravité $\{g_1^m, g_2^m, \ldots\ldots, g_k^m\}$ deviennent les nouveaux centres de classe.

L'algorithme est convergent. Il s'arrête lorsque deux itérations successives conduisent à une décroissance du critère de la somme des carrés interclasses ou à sa stabilisation **(Ouattara, 2012).**

2.3.1.4. Etapes de l'algorithme de classification ascendante hiérarchique (CAH)

Cet algorithme consiste à regrouper itérativement les individus en commençant par le bas : les deux plus proches, en construisant progressivement un arbre ou un dendrogramme regroupant finalement tous les individus en une seule classe à la racine. Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ou la distance entre deux groupes **(Ouattara, 2012).**

Il peut être présenté comme suit :

1. A l'étape initiale, chaque individu forme une classe. Il y a donc n classes.
2. A la première étape, on utilise la matrice de dissimilarité entre les n individus et on regroupe les deux individus les plus proches. Ces deux individus forment une nouvelle classe et on obtient une nouvelle partition de $n - 1$ classes.
3. A la deuxième étape, une nouvelle matrice de dissimilarité entre les $n - 1$ classes est construite en utilisant le critère d'agrégation. On regroupe les deux classes les plus proches et on obtient une nouvelle partition de $n - 2$ classes.
4. On répète le processus jusqu'à l'étape $n - 1$ où le regroupement des classes de la partition de l'étape précédente conduit à une seule classe

En bref, l'algorithme de CAH est une méthode générale de construction d'une hiérarchie à partir de données décrites par une dissimilarité.

2.3. Conclusion

L'objectif de ce chapitre était de définir les outils nécessaires à l'analyse des données comme l'ACP qui est utilis**é** pour effectuer l'exploration et résumer les variations dans un ensemble de données multivariées. Son principe repose sur la transformation des variables originales en un nouvel ensemble de variables non corrélées appelées composantes principales, qui capturent le maximum de variance dans les données d'origine. Et la CAH consiste à rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle.

CHAPITRE 3 : ANALYSE EXPLORATOIRE DES PIQURES DE MOUSTIQUES, CAS DE *L'ANOPHELES GAMBIAE* S.L. ET L'*ANOPHELES FUNESTUS* S.L. AU BURUNDI

3.1. Analyses en composantes principales

3.1.1. Analyse en composantes principales pour les tranches d'heures

L'analyse en composante principale (ACP) est une méthode statistique utilisée pour explorer et visualiser la structure des données multidimensionnelles. Dans le cas des piqûres de moustiques, en particulier l'*An. gambiae* s.l. et l'*An. funestus* s.l. au Burundi, une ACP pourrait être réalisée pour comprendre la variabilité des données et identifier des tendances significatives.

3.1.1.1. Statistiques sommaires des variables continues

Le tableau qui suit donne les statistiques descriptives sur les variables continues à savoir l'effectif le poids, la moyenne, l'écart-type, le minimum et le maximum.

Tableau 3. 1. Statistiques sommaires des variables continues

	2019				2020				2021			
	IAG	**OAG**	**IAF**	**OAF**	**IAG**	**OAG**	**IAF**	**OAF**	**IAG**	**OAG**	**IAF**	**OAF**
Min	329	334	45	31	301	304	55	28	223	271	47	34
1er Q	504.2	596	99.5	60.25	406.2	468	84.75	61.25	331.2	381.8	75.75	63
Média	574.5	659.5	1415	73.50	464	532	106.5	85	380	454	100	84.5
Mean	596	672.5	126.7	83.67	463.1	553.5	103.9	90.24	369.4	447.9	92.08	83.42
3e Q	700	819.8	158	115.5	534	666	123.2	119.2	405.5	509.2	110.2	106.75
Max	828	889	175	132	606	758	139	150	481	606	123	124
Sd	144.9	167.7	43.19	33.65	96.13	131.9	25.51	39.27	78.57	100.2	25.49	28.38
Skew	0.02	-0.31	-0.67	0.01	-0.15	-0.03	-0.27	0.00	-0.37	-0.10	-0.61	-0.10
Kurt	-1.03	-0.9	-1.13	-1.67	-1.17	-1.05	-1.21	-1.44	-0.98	-1.19	-1.20	-1.34
IC	[513.9; 60.9]	[577.5; 767.4]	[102.2; 151.1]	[64.6 ; 102.7]	[408.4 ; 517.4]	[478.8 ; 628.1]	[89.4 ; 118.3]	[68.2 ; 112.6]	[324.9 ;413.8]	[391.2; 504.5]	[77.6 ; 106.5]	[67.3 ; 99.4]

Globalement, nous constatons une grande amplitude de données pour une même variable. Les variables ont presque un minimum plus faible que le maximum. Cette amplitude laisse à penser qu'une forte variabilité des données est représentée dans le jeu de données.

1. Les variables OAG2019, OAG20 et OAG2021 montrent respectivement la variation des piqures enregistrées à l'extérieur pour le vecteur *An. gambiae* s.l. sur chaque tranche d'heures en 2019, 2020 et 2021.
2. Les variables OAF2019, 2020 et OAF2021 ont en moyenne moins de piqures de moustiques enregistrées à l'extérieur pour le vecteur *An. funestus* s.l. sur chaque tranche d'heure en 2019, 2020 et 2021.

3. Les variables IAG2019, IAG20 et IAG2021 montrent respectivement la variation des piqures enregistrées à l'intérieur pour le vecteur *An. gambiae* s.l. sur chaque tranche d'heures en 2019, 2020 et 2021.
4. Les variables IAF2019, IAF2020 et IAF2021 ont en moyenne moins de piqures de moustiques enregistrées à l'extérieur pour le vecteur *An. funestus* s.l. sur chaque tranche d'heure en 2019, 2020 et 2021.

Pour le vecteur *An. gambiae* s.l. il y a eu une grande variance de piqures à l'intérieur en 2019 (mean=596 ; IC [513.9 ;690.9] ; sd=144.9), en 2020 (mean = 463.1 ; IC [408.4 ,517.4] ; sd = 96.13) il y a eu une légère diminution en 2021 (369.4 ; IC [324.9,413.8] ; sd = 78.57). A l'extérieur, la grande variance a été observée en 2019 (mean=672.5 ; IC [577.5 ,767.4] ; sd=167.7), une légère diminution a été observée en 2020 (mean=553.5 ; IC [478.8,628.1] ; sd=131.9) et en 2021 (mean=447.9 ; IC [391.2,505.5] ; sd=100.2).

Pour le vecteur *An. funestus* s.l., il y a également eu une grande variance de piqures à l'intérieur en 2019 (mean=126.7 ; IC [102.2 ;151.1] ; sd= 43.19), en 2020 (mean = 103.92 ; IC [89.4 ; 118.3] ; sd = 25.51) il y a eu une diminution voire en 2021 (mean = 92.08 ; IC [76.6,106.5,413.8] ; sd = 25.49). A l'extérieur, la grande variance a été observée en 2019 (mean=83.7 ; IC [64.6,102.7] ; sd=33.65), en 2020 (mean=90.24 ; IC [68.2,112.6] ; sd=39.27) et en 2021 (mean=83.42 ; IC [67.3 ;99.4] ; sd=28.38).

Les variables IAG2019, OAG2020, IAF2019 ont une grande variance, nous allons faire une ACP normée.

3.2.1.2. Matrices de corrélation

Le tableau ci-après représente la matrice des corrélations entre les variables quantitatives.

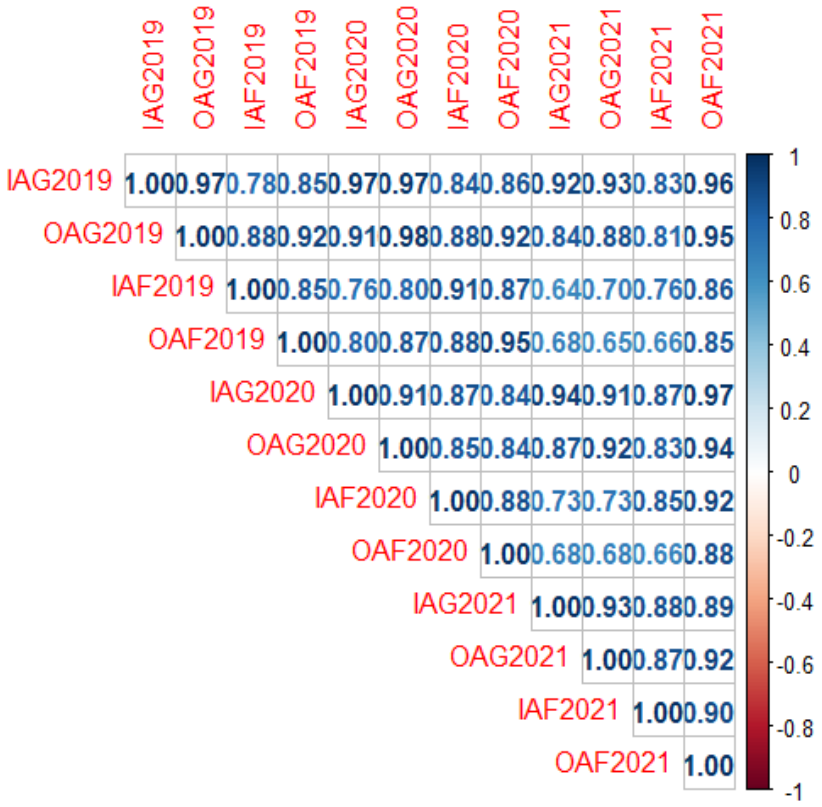


Figure 3. 1. Matrice des corrélations entre les variables

Toutes les corrélations linéaires sont positives ce qui signifie que toutes les variables varient, en moyenne, dans le même sens. Certaines corrélations sont très forte, d'autres moyennes. Par exemple, il y a une très forte corrélation positive (0,96) entre les caractéristiques IAG2019 – OAG2019 d'une part, IAF2019-IAG2019 d'autre part (0,78).

- La variation des piqures observée pour les vecteurs *An. gambiae* s.l. et *An. funestus* s.l. à l'intérieur et à l'extérieur ont significativement tendance à avoir une forte corrélation entre eux. Du fait de cette matrice, on pourra constater qu'il y a une forte corrélation pour les piqures qui sont dans la même année.

- Les variables IAG2019 et IAG2020 sont fortement liées positivement avec le coefficient de 0.97 cela signifie que lorsqu'une personne a été piqué par une moustique du vecteur *An. gambiae* s.l. en 2019 a l'intérieur il l'a été aussi en 2020 et lorsqu'elle n'a pas été piqué par ce vecteur en 2019 elle ne l'a pas été aussi en 2020.

- Les variables IAF2019 et IAF2020 sont fortement liées positivement avec le coefficient de 0.88 cela signifie que lorsqu'une personne a été piqué par une moustique du vecteur *An. funestus* s.l. en 2019 à l'intérieur il l'a été aussi en 2020 et lorsqu'elle n'a pas été piqué par ce vecteur en 2019 elle ne l'a pas été aussi en 2020.

3.2.1.3. Visualisation du nuage des individus projetés et du cercle des corrélations

3.2.1.3.1. Analyse en Composantes Principales normée (ACP normée)

Etant donné que les données sont toutes dans la même unité, la question se pose de réaliser une ACP normée ou non. En faisant une ACP normée : nous nous assurons que chaque variable à la même importance quelle que soit son amplitude. Une ACP non normée nous assure au contraire qu'une variation d'une piqure à la même importance. Nous optons pour l'ACP normée. Dans ce contexte précis où les variables ont toutes à peu près la même amplitude, il n'y a pas grande différence. Nous allons réaliser une ACP sur les piqures.

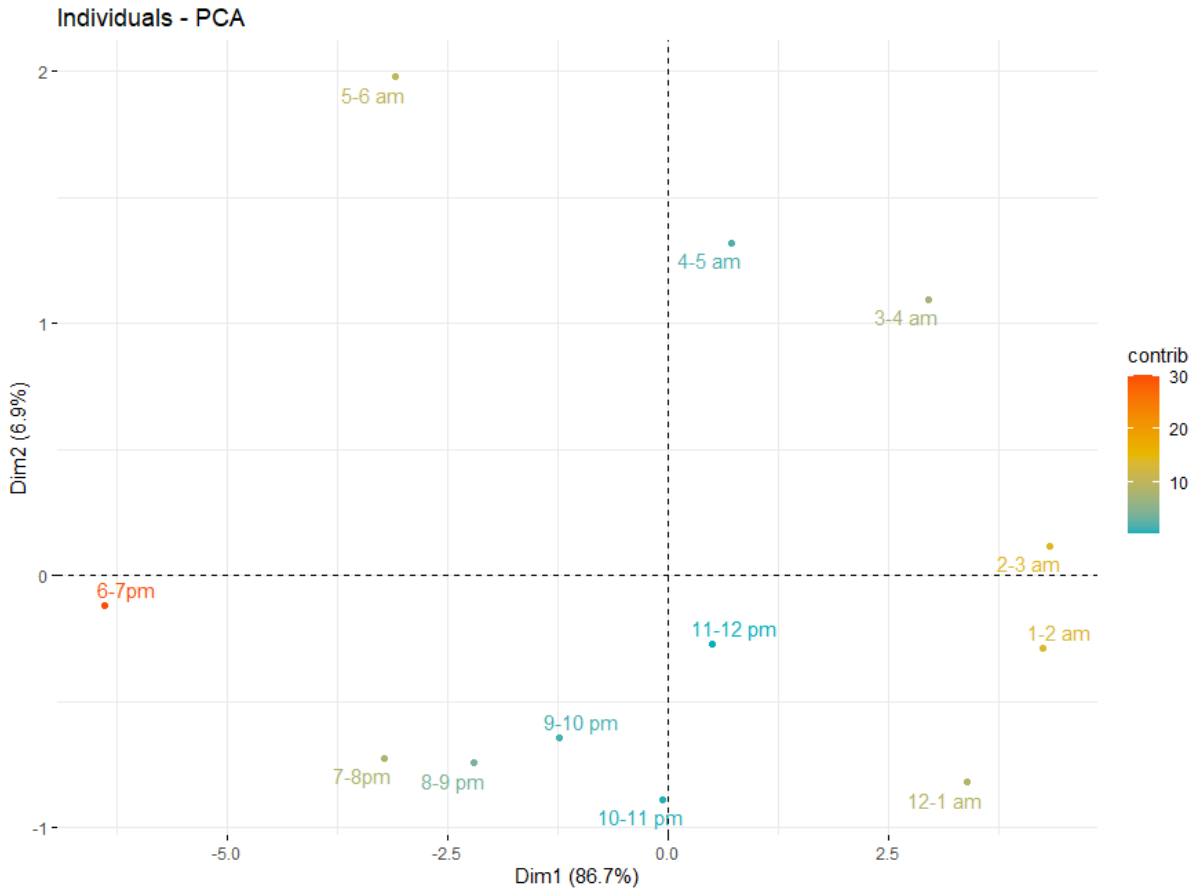


Figure 3. 2. projections des tranches d'heures

- ✓ Ce graphe montre que les tranches d'heures 1-2 am et 2-3 am sont proches ce qui signifie que les piqures moyennes a 1-2 am et 2-3 am sont à peu près les mêmes et ce que ce soit à l'intérieur ou à l'extérieur.
- ✓ Par contre les tranches d'heures 2-3am et 6-7pm ont des comportements différents c'est à dire que ses deux tranches d'heures sont complètement opposées sur le premier axe.

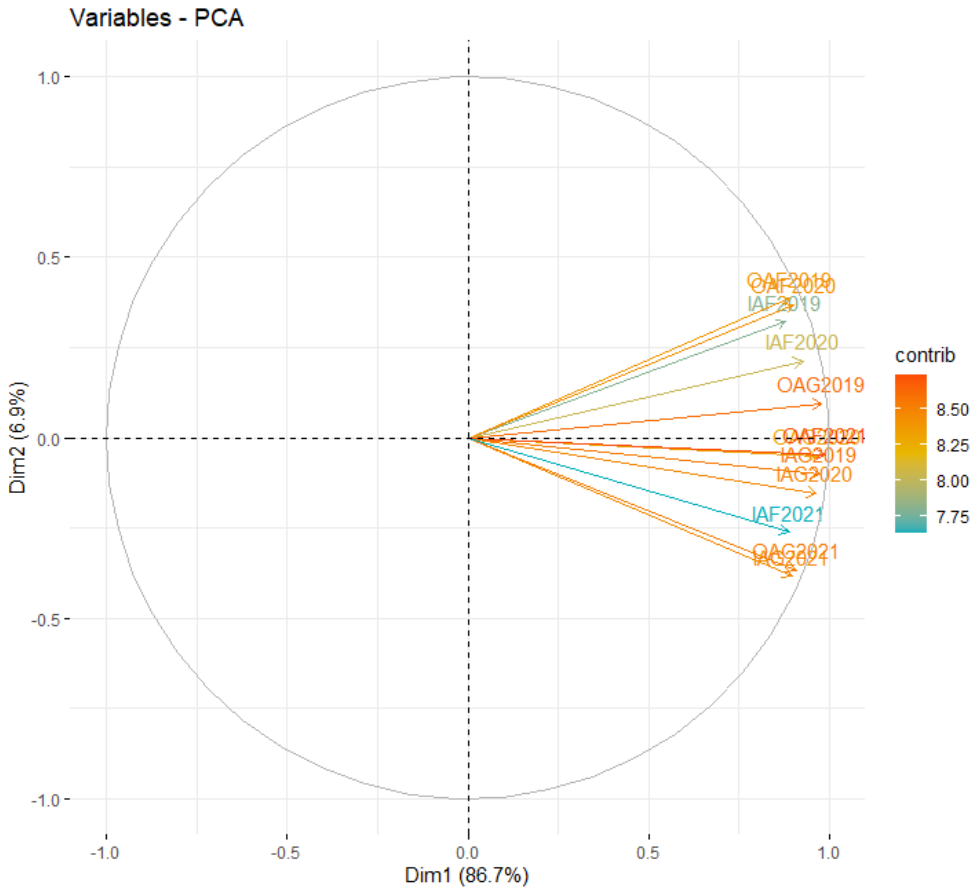


Figure 3. 3. Projections des variables

- ✓ Toutes les variables sauf les variables IAF2021 et IAF2019, sont corrélées positivement à l'axe 1
- ✓ Les tranches d'heures qui sont à droite ont connues des fréquences de piqures qui sont élevées sur la période d'étude.

3.2.1.3.2. Le biplot

Le biplot est une représentation graphique qui permet de visualiser à la fois les individus (lignes) et les variables (colonnes) d'un tableau de données dans un même espace.

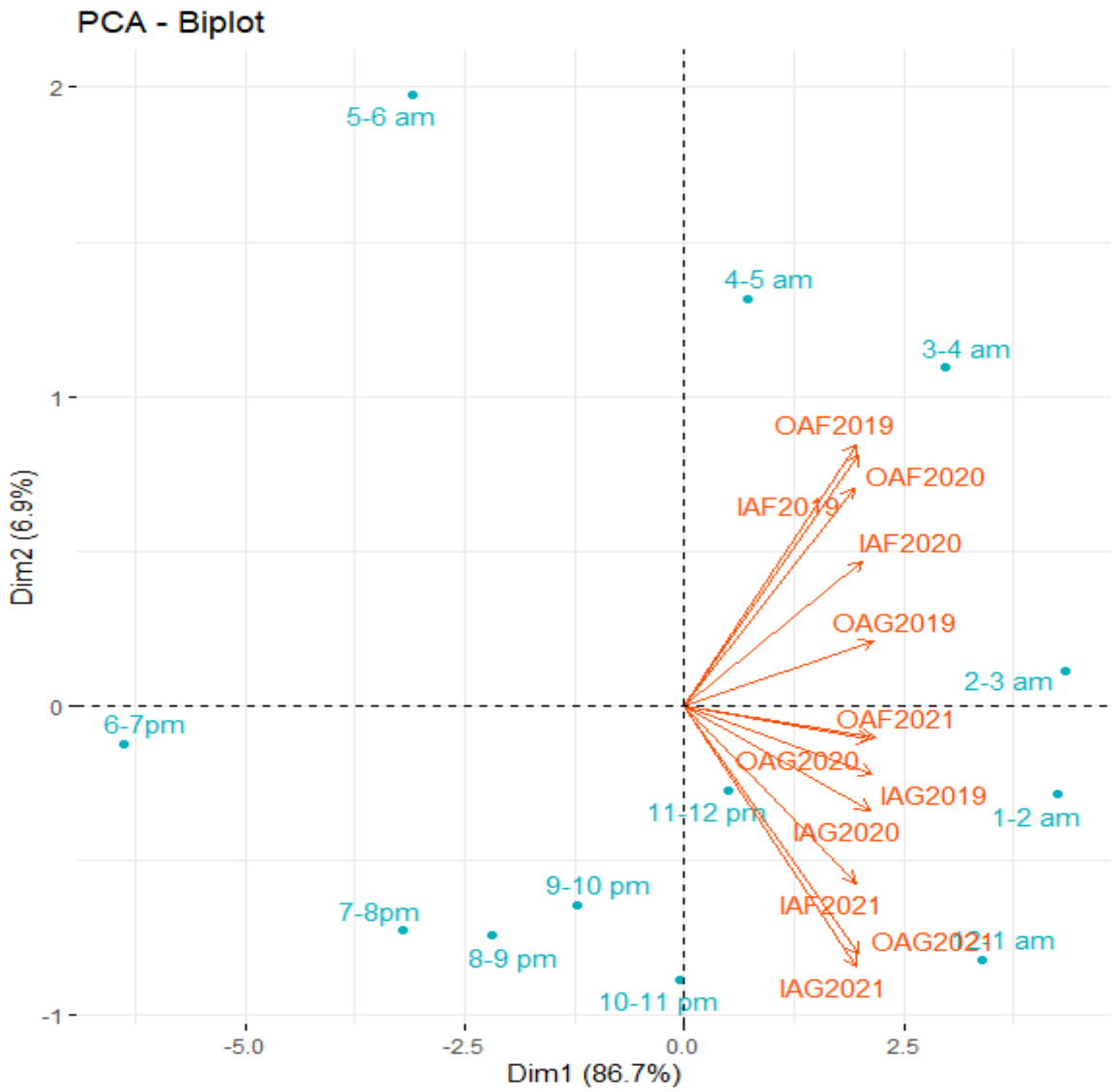


Figure 3. 4. Le biplot

- ✓ Nous observons un premier axe fortement et positivement corrélé à toutes les piqures : les tranches d'heures ayant des coordonnées positives sur cet axe ont tendance à avoir des piqures plus élevées que la moyenne sur tous les variables de l'étude (tranches d'heures remarquables : 2-3 am, 1-2 am), celles qui ont des coordonnées négatives ont tendance à avoir des piqures moins- élevées que la moyenne sur toutes les variables de l'étude (tranches d'heure remarquables 6-7pm).

- ✓ Le second axe est corrélé positivement aux piqures du vecteur *An. funestus* s.l. en 2019, 2020 a l'interieur et a l'exterieur et et le vecteur *An.gambiae* s.l. en 2019 . Il est corrélé négativement aux piqures du vecteur *An. gambiae* s.l. en 2019, 2020, 2021 a l'interieur eten 2020, 2021 a l'exterieur et le vecteur *An. funestus* s.l. en 2021 a l'interieur et a l'exterieur : les tranches d'heures ayant des coordonnées positives sur cet axe ont tendance à avoir des piqures du vecteur *An. funestus* s.l. plus élevées que la moyenne et des piqures du vecteur *An. gambiae*, les tranches d'heures ayant des coordonnées négatives sur cet axe ont tendance à avoir plus des piqures du vecteur *An. gambiae* et moins des piqures du vecteur *An. funestus* s.l.

3.2.1.3.3. Eboulie des valeurs propres

L'eboulie des valeurs propres est un graphique qui montre les valeurs propres d'une matrice de covariance ou de covariance, généralement utilisées dans l'analyse en composantes principales (ACP). Ces valeurs propres représentent la variance expliquée par chaque axe principal. L'eboulie des valeurs propres est un outil important pour déterminer le nombre optimal de composantes à analyser dans l'ACP, car elle permet de visualiser la répartition de la variance expliquée par chaque axe.

Nous regardons maintenant les valeurs propres.

Tableau 3. 2. Tableau des valeurs propres

	Valeurs propres	Pourcentage	Pourcentage cum
Dim1	10.40	86.70	86.70
Dim2	0.83	6.93	93.63
Dim3	0.35	2.95	96.57
Dim4	0.17	1.48	98.06
Dim5	0.10	0.84	98.90
Dim6	0.06	0.58	99.48
Dim7	0.03	0.27	99.74
Dim8	0.01	0.14	99.88
Dim9	0.01	0.07	99.96
Dim10	0.01	0.03	99.99
Dim11	0.00	0.01	100.00

On retrouve dans cette colonne pourcentage les valeurs de 86.70% et 6.93% qui étaient indiquées dans les étiquettes d'axes de graphique ci-dessus. Ainsi, avec une représentation 2D (correspondant aux deux premiers axes principaux) on arrive à représenter

86.70%+6.9%=93.63% de l'information contenue dans les données les piqures de moustiques (sur 3 ans) au Burundi. Ici, on a 93.63% de l'info sur 2 axes, alors qu'on avait 12 variables à la base. Avant de réaliser l'ACP, deux dimensions correspondaient donc à 8 fois un dixième de l'inertie du jeu de données, soit 80%.

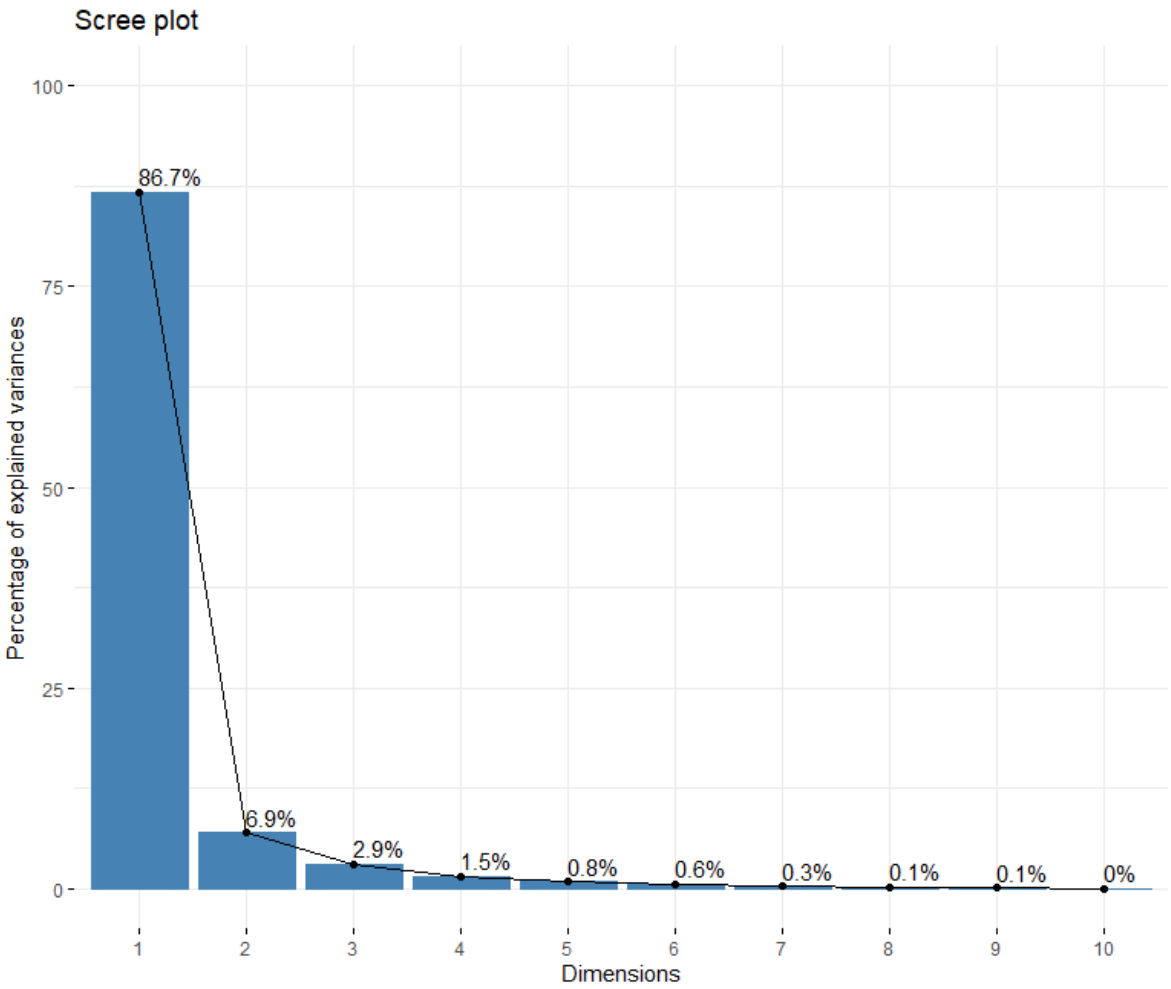


Figure 3. 5. Eboulie des valeurs propres

- ✓ Le premier axe concentre visiblement une très grande partie de la variance. Le deuxième axe représentant à lui seul plus de 6.9% de la variabilité, il peut être intéressant de le garder pour bien interpréter les résultats. Pourcentage d'inertie :
- ✓ Sur le premier plan factoriel (Axe1 et Axe 2) on est parvenu à conserver 93.63% d'information cela signifie que si on résume les 12 variables par deux dimensions on récupère 93.63% d'information contenue dans le jeu de tableau. Nous avons un parfait résumé qui synthétise toutes les 12 variables.

3.2.1.3.4. Contribution pour les composantes principales

Les contributions pour les composantes principales dans le cadre de l'analyse en composantes principales (ACP) sont des mesures qui indiquent la part de variance expliquée par chaque variable dans la construction des composantes principales. Ces contributions permettent de comprendre l'importance de chaque variable dans la formation des axes principaux de l'ACP.

Tableau 3. 3. Contribution pour les valeurs propres

	Dim1	Dim2	Dim3	Dim4	Dim5
IAG2019	9.14	1.21	9.10	0.16	0.36
OAG2019	9.22	1.09	3.44	8.40	1.19
IAF2019	7.42	12.41	17.63	20.46	14.06
OAF2019	7.64	17.90	5.45	2.77	21.21
IAG2020	8.92	2.84	0.96	18.01	8.48
OAG2020	8.99	0.30	4.82	10.36	20.28
IAF2020	8.24	5.46	16.94	8.41	4.46
OAF2020	7.79	16.42	4.75	3.86	16.56
IAG2021	7.72	17.77	1.18	8.05	0.10
OAG2021	7.91	16.14	0.47	18.05	2.68
IAF2021	7.59	8.20	34.75	0.81	7.40
OAF2021	9.42	0.26	0.53	0.67	3.223

Presque toutes les variables ont assez contribué pour la construction de la première dimension mais celles qui ont contribué le plus sont les variables OAF2021, OAG2019 et IAG2019. Les variables qui ont contribué d'une manière significative pour la construction de la deuxième dimension sont OAF2019 et IAG2021.

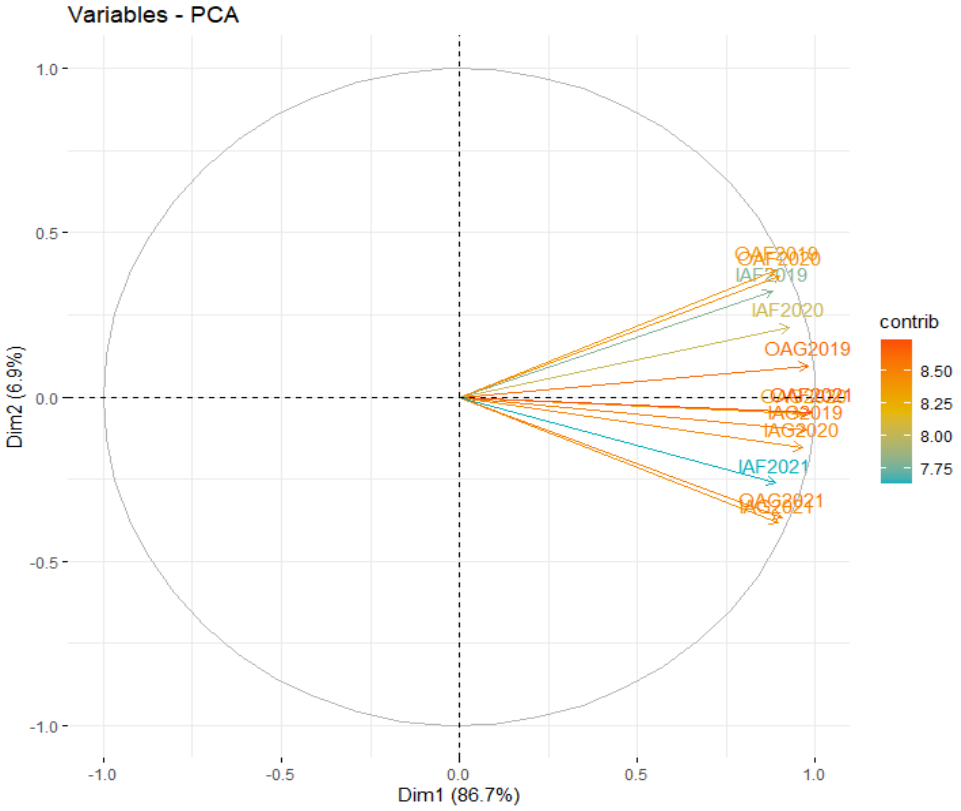


Figure 3. 6. Cercle des corrélations

3.1.2. Analyse en composantes principales pour les localités (sites sentinelles)

3.1.2.1. Statistiques sommaires des variables

Le tableau qui suit donne les statistiques descriptives sur les variables continues à savoir l'effectif le poids, la moyenne, l'écart-type, le minimum et le maximum.

Tableau 3. 4. Statistiques sommaires pour les sites sentinelles

	2019				2020				2021			
	IAG	**OAG**	**IAF**	**OAF**	**IAG**	**OAG**	**IAF**	**OAF**	**IAG**	**OAG**	**IAF**	**OAF**
Min	2.0	33.0	0.0	0.0	1.0	26	2.0	3.0	0.0	22.0	0.0	0.0
1er Q	217.0	163.0	22.0	15.0	171.0	196	16.0	8.0	94.0	174.0	4.0	13.0
Médi	601.0	487.0	149.0	55.0	289.0	577	108.0	118.0	235.0	220.0	117.0	62.0
Mean	794.7	896.7	168.9	111.6	617.3	738	138.4	120.4	492.6	597.2	122.8	111.2
3e Q	1043.0	887.0	175.0	192.0	639.0	659	150.0	162.0	469.0	720.0	186.0	152.0
Max	3019.0	4080.0	595.0	370.0	2744	3430	504.0	349.0	2339	2699	405.0	399.0
Sd	928.89	1251.65	202.6	126.08	852.3	1052.56	166.59	122.3	726.2	842.1	132.6	134.5
Skew	1.37	1.73	0.97	0.78	1.62	1.76	1.08	0.56	1.73	1.63	0.82	1.01
Kurt	0.78	1.17	-0.51	-0.79	1.38	1.81	-0.19	-1.20	1.70	1.38	-0.46	-0.42
IC	[269.1 ; 1320.3]	[188.4 ; 1604.8]	[54.2 ; 283.6]	[40.2; 182.8]	[135.1; 1099]	[142.4 ;1 333.5]	[42.4; 234.3]	[51.1 ; 189.6]	[81.6 ; 903.4]	[120;1 073.3]	[47.7 ; 197.8]	[35.1 ; 187.3]

Globalement, nous constatons une grande amplitude de données pour une même variable. Les variables ont presque un minimum plus faible que le maximum. Cette amplitude laisse à penser qu'une forte variabilité des données est représentée dans le jeu de données.

1. Les variables OAG2019, OAG20 et OAG2021 montrent respectivement la variation des piqures enregistrées à l'extérieur pour le vecteur *An. gambiae* s.l. sur chaque localité (site sentinelle) en 2019, 2020 et 2021.
2. Les variables OAF2019, OAF2020 et OAF2021 ont en moyenne moins de piqures de moustiques enregistrées à l'extérieur pour le vecteur *An. funestus* s.l. sur chaque localité (site sentinelle) en 2019, 2020 et 2021.
3. Les variables IAG2019, IAG20 et IAG2021 montrent respectivement la variation des piqures enregistrées à l'intérieur pour le vecteur *An. gambiae s.l.* sur chaque localité (site sentinelle) en 2019, 2020 et 2021.
4. Les variables IAF2019, I020 et IAF2021 ont en moyenne moins de piqures de moustiques enregistrées à l'extérieur pour le vecteur *An. funestus* s.l.sur chaque localité (site sentinelle) en 2019, 2020 et 2021.

Pour le vecteur *An. gambiae* s.l. il y a eu une grande variance de piqures à l'intérieur en 2019 (mean=794.7 ; IC [269.1 ;1320.3] ; sd=928.89), en 2020 (mean = 617.3 ; IC [135.1 ,1099] ; sd = 852.3) il y a eu une légère diminution en 2021 (mean=492.6 ; IC [81.6 ;903.4] ; sd = 726.2). A l'extérieur, la grande variance a été observée en 2019 (mean=896.7 ; IC [188.4 ,1604.8] ; sd=1251.65), une légère diminution a été observée en 2020 (mean=738 ; IC [142.4,1333.5] ; sd=1052.5) et en 2021 (mean=597.2 ; IC [120.7,1073.3] ; sd=842.1).

Pour le vecteur *An. funestus* s.l., il y a également eu une grande variance de piqures à l'intérieur en 2019 (mean=168.9 ; IC [54.2 ;283.6] ; sd= 202.6), en 2020 (mean = 138.4 ; IC [42.4 ; 234.1] ; sd = 166.59) il y a eu une diminution voire en 2021 (mean = 122.8 ; IC [47.7,197.8] ; sd = 132.6). A l'extérieur, la grande variance a été observée en 2019 (mean=111.6 ; IC [40.2,182.8] ; sd=126.1), en 2020 (mean=120.4 ; IC [51.1,189.6] ; sd=122.3) et en 2021 (mean=111.2 ; IC [35.1,187.3] ; sd=134.5).

Les variables OAG2019, OAG2020 et IAF2019 ont une grande variance, nous allons faire une ACP normée.

3.1.2.2. Matrices de corrélation

Le tableau ci-après représente la matrice des corrélations entre les variables quantitatives. Toutes les corrélations linéaires sont positives ce qui signifie que toutes les variables varient, en moyenne, dans le même sens. Certaines corrélations sont très forte, d'autres moyennes. Par exemple, il y a une très forte corrélation positive (0,96) entre les caractéristiques IAG2019 – OAG2019 d'une part, IAF2019-IAG2019 d'autre part (0,78).

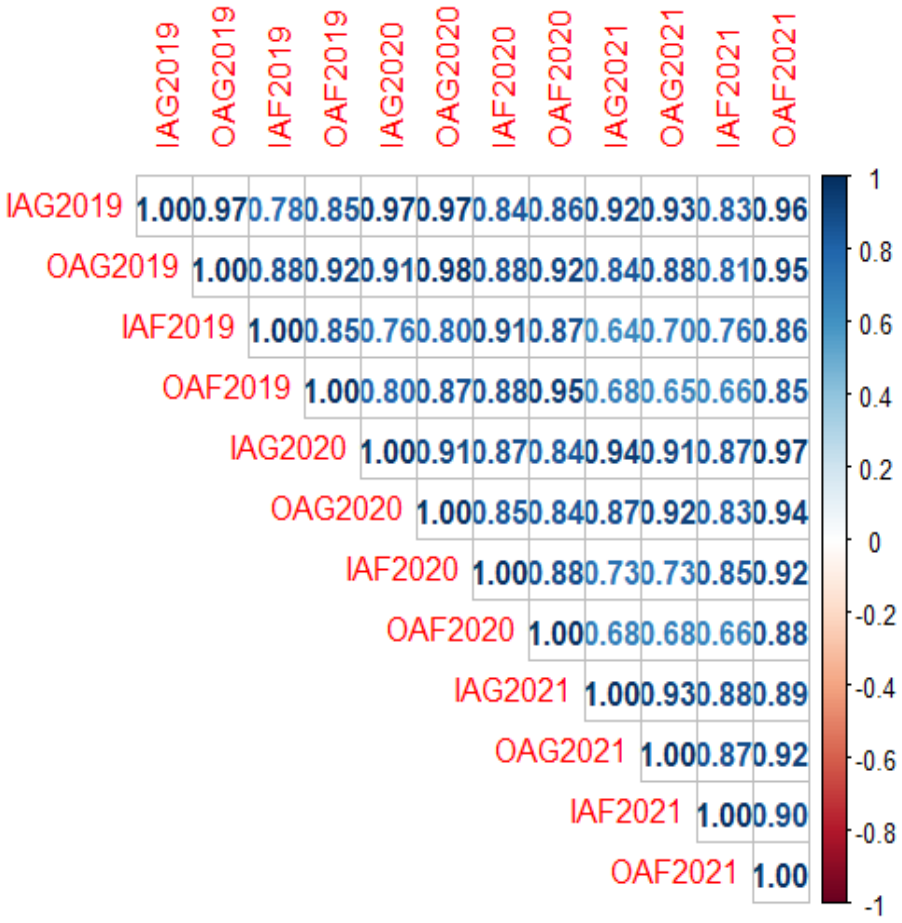


Figure 3. 7. Matrice des corrélations

- La variation des piqures observée pour les vecteurs *An. gambiae* s.l. et *An. funestus* s.l. à l'intérieur et à l'extérieur ont significativement tendance à avoir une forte corrélation entre eux. Du fait de cette matrice, on pourra constater qu'il y a une forte corrélation pour les piqures qui sont dans la même année.

- Les variables IAG2019 et IAG2020 sont fortement liées positivement avec le coefficient de 0.97 cela signifie que lorsqu'une personne a été piqué par une moustique du vecteur

An. gambiae s.l. en 2019 à l'intérieur il l'a été aussi en 2020 et lorsqu'elle n'a pas été piqué par ce vecteur en 2019 elle ne l'a pas été aussi en 2020.

- Les variables IAF2019 et IAF2020 sont fortement liées positivement avec le coefficient de 0.88 cela signifie que lorsqu'une personne a été piqué par une moustique du vecteur *An. funestus* s.l. en 2019 à l'intérieur il l'a été aussi en 2020 et lorsqu'elle n'a pas été piqué par ce vecteur en 2019 elle ne l'a pas été aussi en 2020.

3.1.2.3. Visualisation du nuage des individus projetés et du cercle des corrélations

3.1.2.3.1Analyse en Composantes Principales normée (ACP normée)

Etant donné que les données sont toutes dans la même unité, la question se pose de réaliser une ACP normée ou non. En faisant une ACP normée : nous nous assurons que chaque variable à la même importance quelle que soit son amplitude. Une ACP non normée nous assure au contraire qu'une variation d'une piqure à la même importance. Nous optons pour l'ACP normée. Dans ce contexte précis où les variables ont toutes à peu près la même amplitude, il n'y a pas grande différence. Nous allons réaliser une ACP sur les piqures.

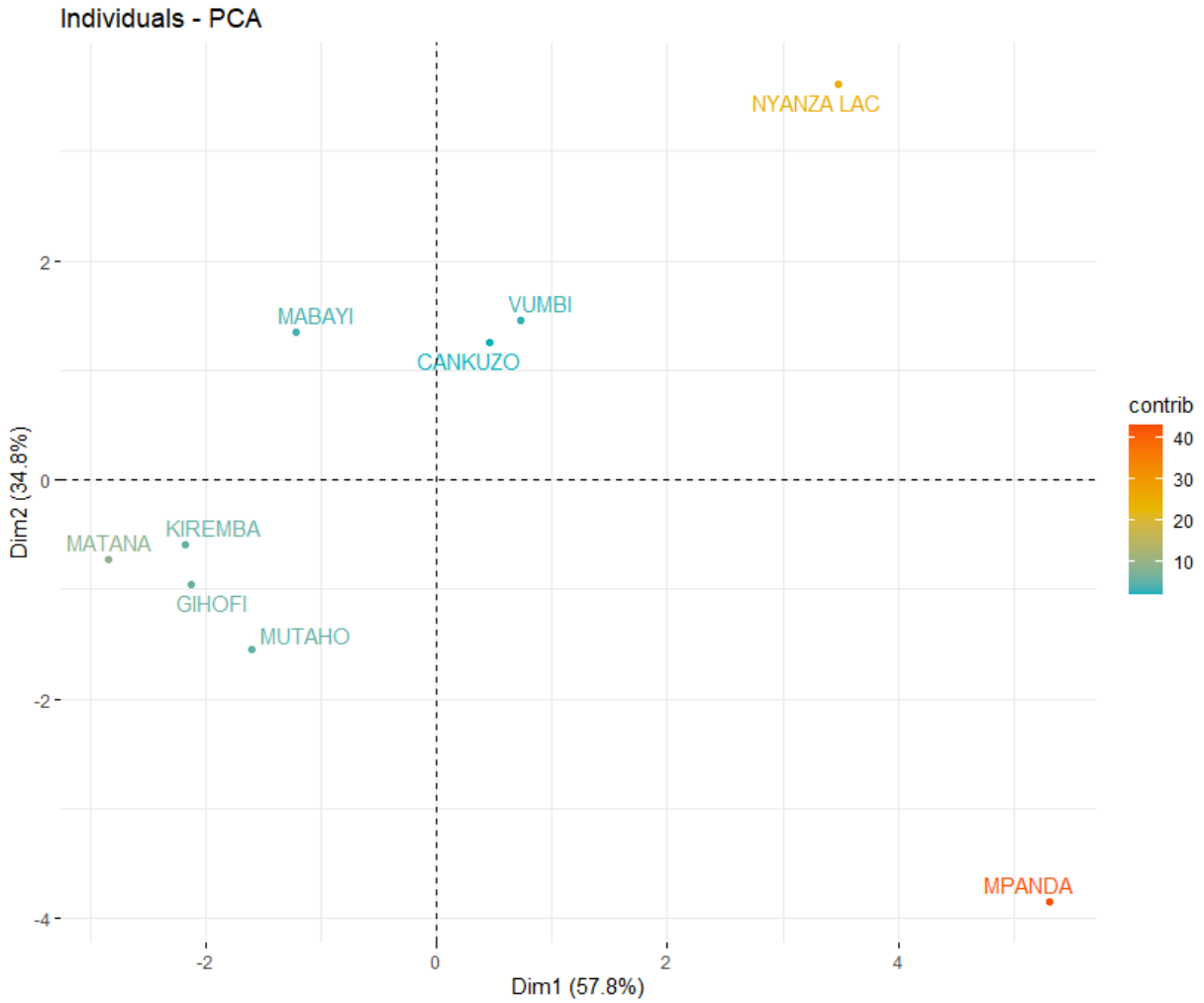


Figure 3. 8. Projections des sites sentinelles

- ✓ Ce graphe montre que les sites sentinelles NYANZA LAC et MPANDA sont opposés sur le premier axe. Les comportements des moustiques lors des piqures sur les deux sont différents.
- ✓ Par contre les sites MUTAHO, KIREMBA, GIHOFI et MATANA sont très proches ce qui signifie que les piqures moyennes sont à peu près les mêmes que ce soit à l'intérieur et à l'extérieur.

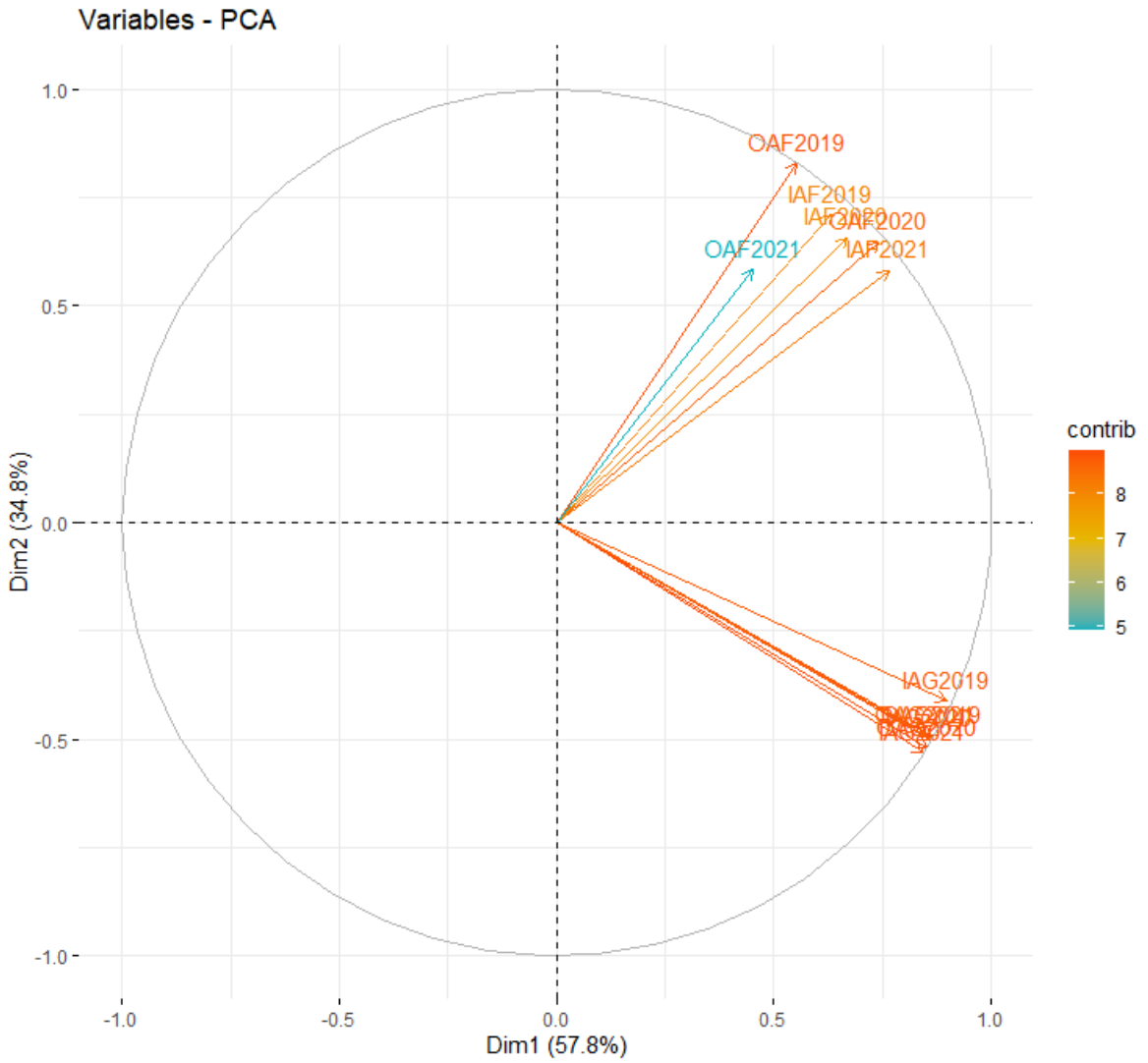


Figure 3. 9. Projections des variables

- ✓ Toutes les variables sauf la variable OAF2021, sont corrélées positivement à l'axe 1
- ✓ Les sites sentinelles qui sont à droite ont connues des fréquences de piqures qui sont élevées sur la période d'étude.

3.1.2.3.2. Le biplot

Le biplot est une représentation graphique qui permet de visualiser à la fois les individus (lignes) et les variables (colonnes) d'un tableau de données dans un même espace.

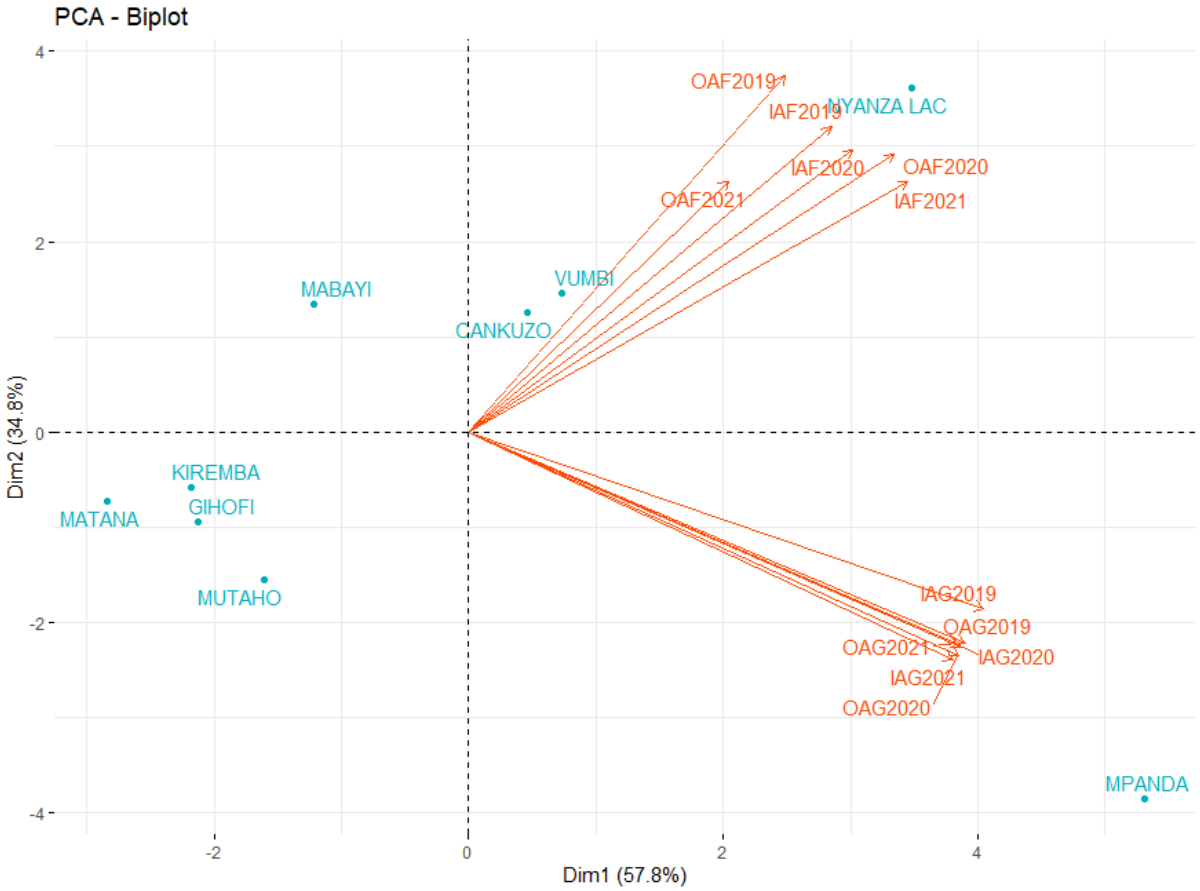


Figure 3. 10. Le biplot pour les sites sentinelles

- ✓ Nous observons un premier axe fortement et positivement corrélé à toutes les piqures : les sites sentinelles ayant des coordonnées positives sur cet axe ont tendance à avoir des piqures plus élevées que la moyenne sur tous les variables de l'étude (sites sentinelles remarquables : NYANZA LAC, VUMBI), celles qui ont des coordonnées négatives ont tendance à avoir des piqures moins- élevées que la moyenne sur toutes les variables de l'étude (sites sentinelles KIREMBA, MATANA, GIHOFI, MUTAHO).

- ✓ Le second axe est corrélé positivement aux piqures du vecteur *An. funestus* s.l. en 2019, 2020 a l'interieur et a l'exterieur . Il est corrélé négativement aux piqures du vecteur *An. gambiae* s.l. en 2019, 2020, 2021 a l'interieur et a l'exterieur: les tranches d'heures ayant des coordonnées positives sur cet axe ont tendance à avoir des piqures du vecteur *An. funestus* s.l. plus élevées que la moyenne et des piqures du vecteur *An. gambiae*, les tranches d'heures ayant des coordonnées négatives sur cet axe ont tendance à avoir plus des piqures du vecteur *An. gambiae* et moins des piqures du vecteur *An. funestus* s.l.

3.1.2.3.3. Eboulie des valeurs propres

L'eboulie des valeurs propres est un graphique qui montre les valeurs propres d'une matrice de covariance ou de covariance, généralement utilisées dans l'analyse en composantes principales (ACP). Ces valeurs propres représentent la variance expliquée par chaque axe principal. L'eboulie des valeurs propres est un outil important pour déterminer le nombre optimal de composantes à analyser dans l'ACP, car elle permet de visualiser la répartition de la variance expliquée par chaque axe.

Tableau 3. 5. Les valeurs propres pour les sites sentinelles

	Valeurs propres	Pourcentage	Pourcentage cumulees
Dim1	6.93	57.8	57.7
Dim2	4.18	34.81	92.59
Dim3	0.75	6.28	98.8
Dim4	0.11	0.93	99.87
Dim5	0.01	0.10	99.91
Dim6	0.01	0.05	99.96
Dim7	0.003	0.03	99.99
Dim8	0.000	0.00	1000

On retrouve dans cette colonne pourcentage les valeurs de 57.8% et 34.81% qui étaient indiquées dans les étiquettes d'axes de graphique ci-dessus. Ainsi, avec une représentation 2D (correspondant aux deux premiers axes principaux) on arrive à représenter 57.8%+34.81%=92.59 % de l'information contenue dans les données les piqures de moustiques (sur 3 ans) au Burundi. Ici, on a 92.59% de l'info sur 2 axes, alors qu'on avait 12 variables à la base.

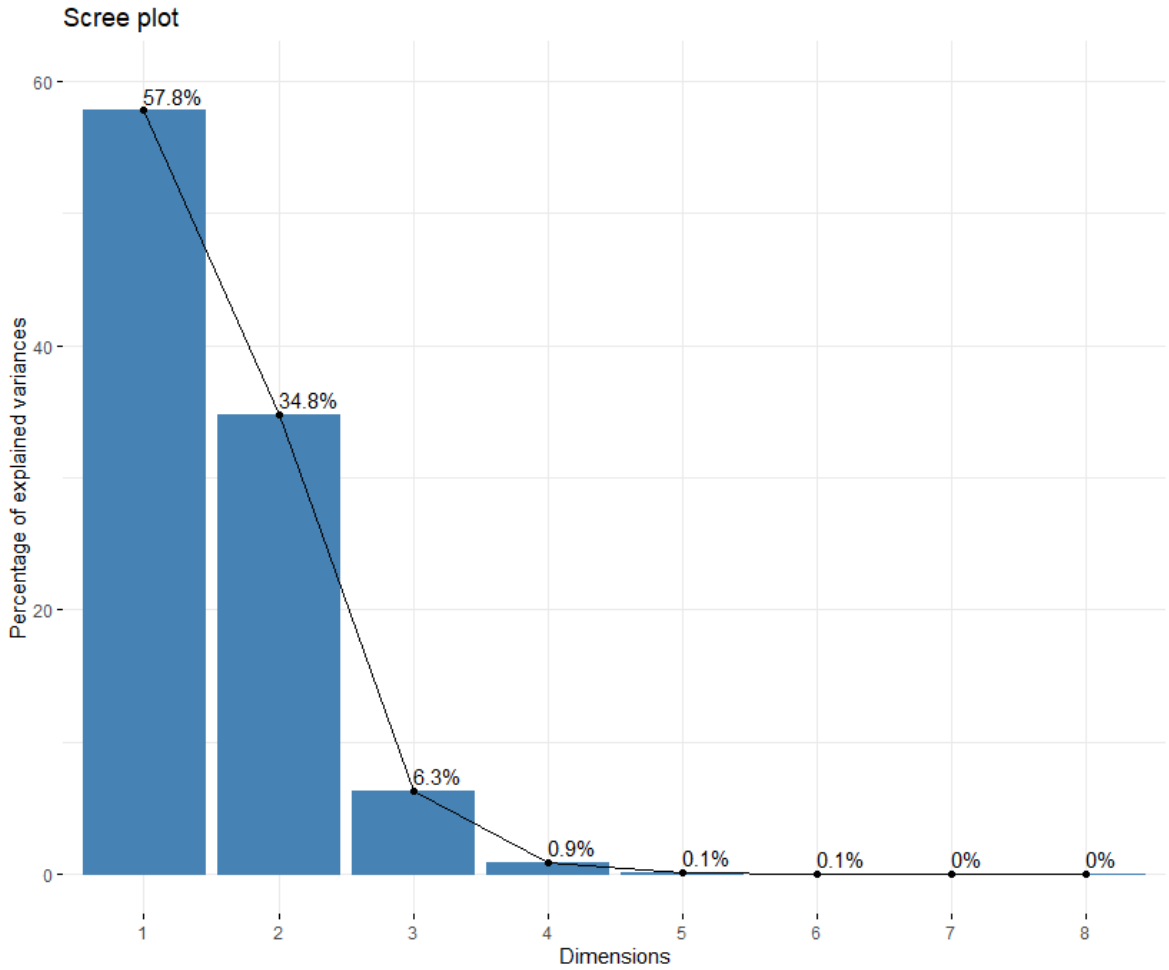


Figure 3. 11. Eboulie des valeurs propres

- ✓ Le premier axe concentre visiblement une très grande partie de la variance. Le deuxième axe représentant à lui seul plus de 34.81% de la variabilité, il peut être intéressant de le garder pour bien interpréter les résultats.
- ✓ Sur le premier plan factoriel (Axe1 et Axe 2) on est parvenu à conserver 92.59% d'information cela signifie que si on résume les 12 variables par deux dimensions on récupère 92.59% d'information contenue dans le jeu de tableau. Nous avons un parfait résumé qui synthétise toutes les 12 variables.

3.1.2.3.4. Contribution pour les composantes principales

Les contributions pour les composantes principales dans le cadre de l'analyse en composantes principales (ACP) sont des mesures qui indiquent la part de variance expliquée par chaque variable dans la construction des composantes principales. Ces contributions permettent de comprendre l'importance de chaque variable dans la formation des axes principaux de l'ACP.

Tableau 3. 6. Contribution pour les composantes principales

	Dim1	**Dim2**	**Dim3**	**Dim4**	**Dim5**
IAG2019	0.98	-0.10	-0.18	-0.02	0.02
OAG2019	0.98	0.10	-0.11	-0.12	-0.03
IAF2019	0.88	0.32	0.25	-0.19	0.12
OAF2019	0.89	0.39	-0.14	0.07	-0.15
IAG2020	0.96	-0.15	-0.06	0.18	0.09
OAG2020	0.97	-0.05	-0.13	-0.14	-0.14
IAF2020	0.93	0.21	0.24	0.12	-0.07
OAF2020	0.90	0.37	-0.13	0.08	0.13
IAG2021	0.90	-0.38	-0.06	0.12	0.01
OAG2021	0.91	-0.37	-0.04	-0.18	0.05
IAF2021	0.89	-0.26	0.35	0.04	-0.09
OAF2021	0.99	0.05	-0.04	0.03	0.06

Presque toutes les variables ont assez contribué pour la construction de la première dimension mais celles qui ont contribué le plus sont les variables IAG2019, OAG20219, IAG2020,OAG2020 et OAF2021. Les variables qui ont contribué d'une manière significative pour la construction de la deuxième dimension sont OAF2019 et OAF2020.

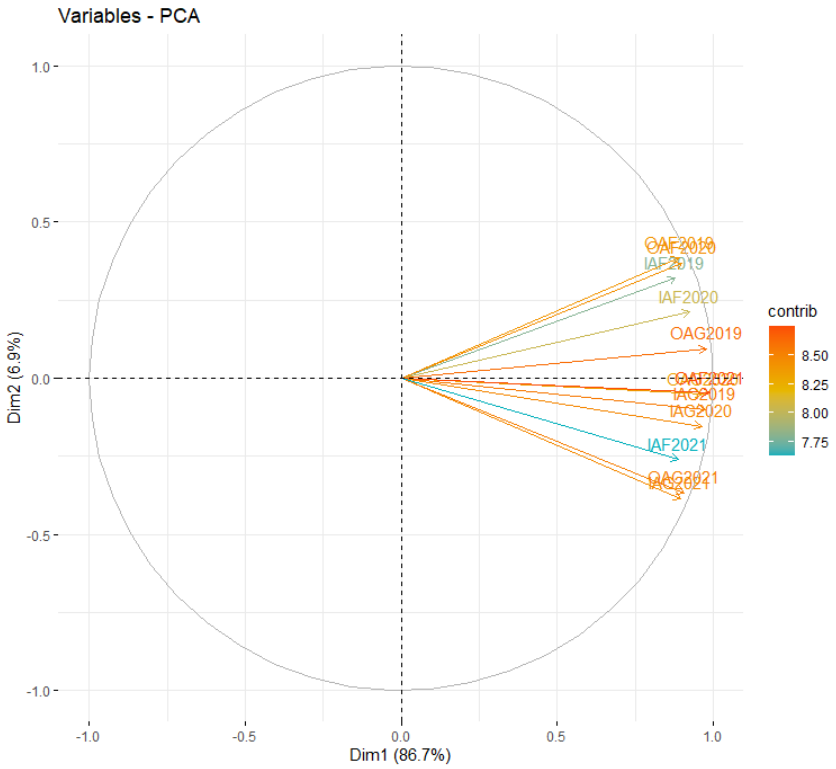


Figure 3. 12. Cercle des corrélations

3.2. Application de la CAH aux données de piqûres de moustiques

L'application de la Classification Ascendante Hiérarchique (CAH) aux données de piqûres de moustiques nous a permis de regrouper les observations en clusters homogènes en fonction de leur similarité.

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante.

3.2.1. Classification ascendante hiérarchiques pour les tranches d'heures

3.2.1.1. Construction de l'arbre de classification

Dendrogramme

Pour rassembler les individus en construisant un arbre, on les assemble 2 à 2 en commençant par les individus les plus proches. Ainsi, la construction de l'arbre de classification repose sur les distances entre individus.

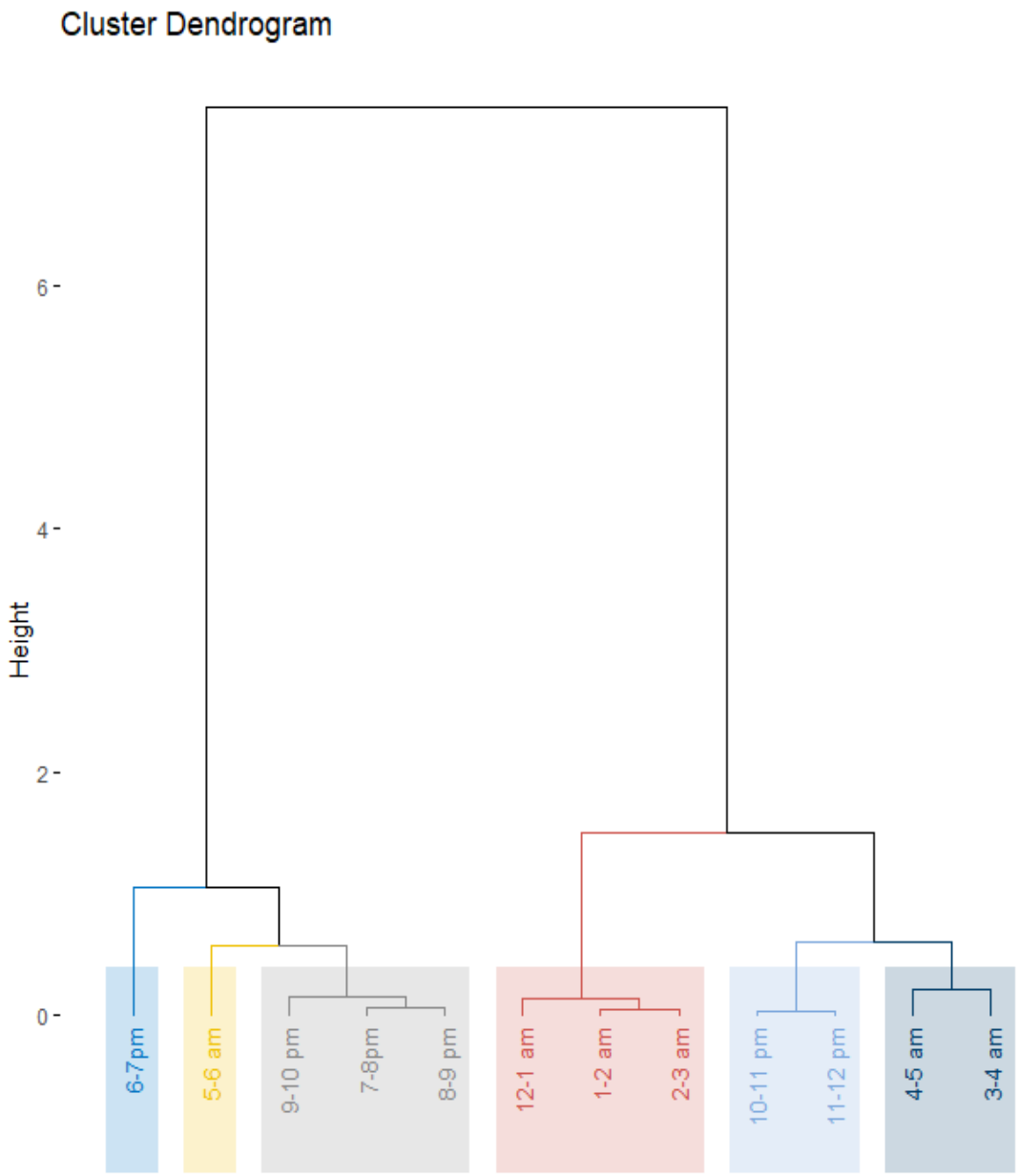


Figure 3. 13. Dendrogramme pour les tranches d'heures

Ce dendrogramme a été créer avec la subdivision finale de 6 groupes selon leur similarité de niveau de variation des piqures.

Le premier groupe comporte 1 seule tranche d'heure, le second comporte 1 seule tranche, le troisième comporte 3 tranches d'heure, le 4eme comporte 3 tranches d'heure, le 5eme comporte 2 tranches d'heures et le dernier groupe comporte aussi 2 tranches d'heure. Cette classification est basée sur les 12 variables utilisés pour l'ACP.

Caractérisation du cluster1

La tranche d'heure 6-7 pm est celui qui représente le cluster1 car elle est la seule dans ce cluster.

Tableau 3. 7.Caractérisation du cluster1

Cluster 1	Mean in category	Overall mean	p.value
IAF 2019	45	126.6	0.04
OAG2020	303.5	553.2	0.04
IAF 2020	54.5	103.6	0.04
OAG2019	334	672.5	0.03

Les variables énumérées dans le tableau sont les plus significativement associées avec le cluster 1. Leurs valeurs moyennes dans le cluster 1 sont inférieures à la moyenne globale correspondantes de chaque variable dans tous les groupes. Par conséquent, le groupe 1 se caracterise par un taux faible :

- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2019 ;
- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2020 ;
- ✓ Des piqures du vecterur *An. gambiae* a l'exterieur en 2019 ;
- ✓ Des piqures du vecterur *An. gambiae* a l'exterieur en 2020

Caractérisation du cluster 2

La tranche d'heure 5-6 am est celui qui représente le cluster2 car elle est la seule dans ce cluster. Il n'y a pas de variables qui caractérisent cette tranche d'heure donc il y a pas eu de piqures dans cette tranche d'heure. Il en est de même que pour les tranches d'heures 10-11pm, 11-12 pm et 4-5, 3-4 am. Représentant respectivement les clusters 4 et 5.

Caractérisation du cluster 3

Les tranches d'heure 7-8 pm, 8-9 pm et 9-10 pm représentent le cluster 3 car elles sont les trois dans ce cluster.

Tableau 3. 8. Caractérisation du cluster 3

Cluster 6	Mean in category	Overall mean	p.value
IAF2019	83.3	126.6	0.04

La variable énumérée dans le tableau est la plus significativement associées avec le cluster3. Sa valeur moyenne dans le cluster 3 est inferieure à valeur moyenne globale dans tous les clusters. Par conséquent, le groupe 3 se caracterise par un taux faible :

- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2019 ;

Caractérisation du cluster 6

Les tranches d'heure 12-1 am, 1-2 am et 2-3 am représentent le cluster 6 car elles sont les trois dans ce cluster.

Tableau 3. 9. Caractérisation du cluster 6

Cluster 6	Mean in category	Overall mean	p.value
IAG2019	790.3	596.00	0.007
OAG2020	723.3	553.2	0.01
OAG2021	574	447.9	0.01
IAG2020	580.3	463.1	0.01
OAG2019	877	672.5	0.01
IAG2021	464.3	369.4	0.01
OAF2021	116.3	83.4	0.02
OAF2020	133	90.4	0.03
OAF2019	116.6	83.6	0.04

Les variables énumérées dans le tableau sont les plus significativement associées avec le cluster 6. Leurs valeurs moyennes dans le cluster 6 sont supérieures aux valeurs moyennes globales dans tous les clusters. Par conséquent, on peut conclure que le cluster se caractérise par un nombre élevé :

- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2019
- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2020
- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2021
- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2020
- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2019
- ✓ De piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2021
- ✓ De piqures des vecteurs *An. funestus* s.l. a l'exterieur en 2019
- ✓ De piqures des vecteurs *An. funestus* s.l. a l'exterieur en 2020
- ✓ De piqures des vecteurs *An. funestus* s.l. a l'exterieur en 2021

3.2.1.2. Représentation des clusters (groupes) et des tranches d'heures en 3D

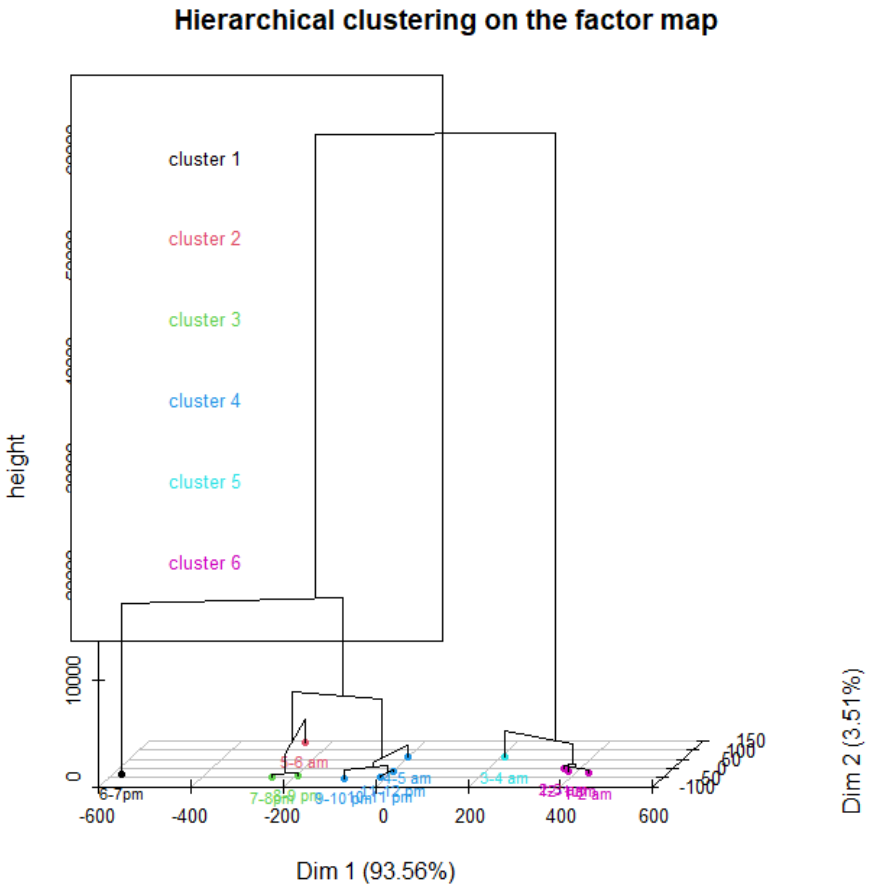


Figure 3. 14. Les clusters (groupes) et les tranches d'heures en 3D

La représentation graphique de la projection des tranches d'heures sur les axes principaux et du dendrogramme ainsi que les groupes d'appartenance des tranches d'heures selon leur ressemblance est alors donnée par cette figure en trois dimensions.

3.2.1.3. Représentation des clusters (groupes) et des tranches d'heures sur la carte des facteurs

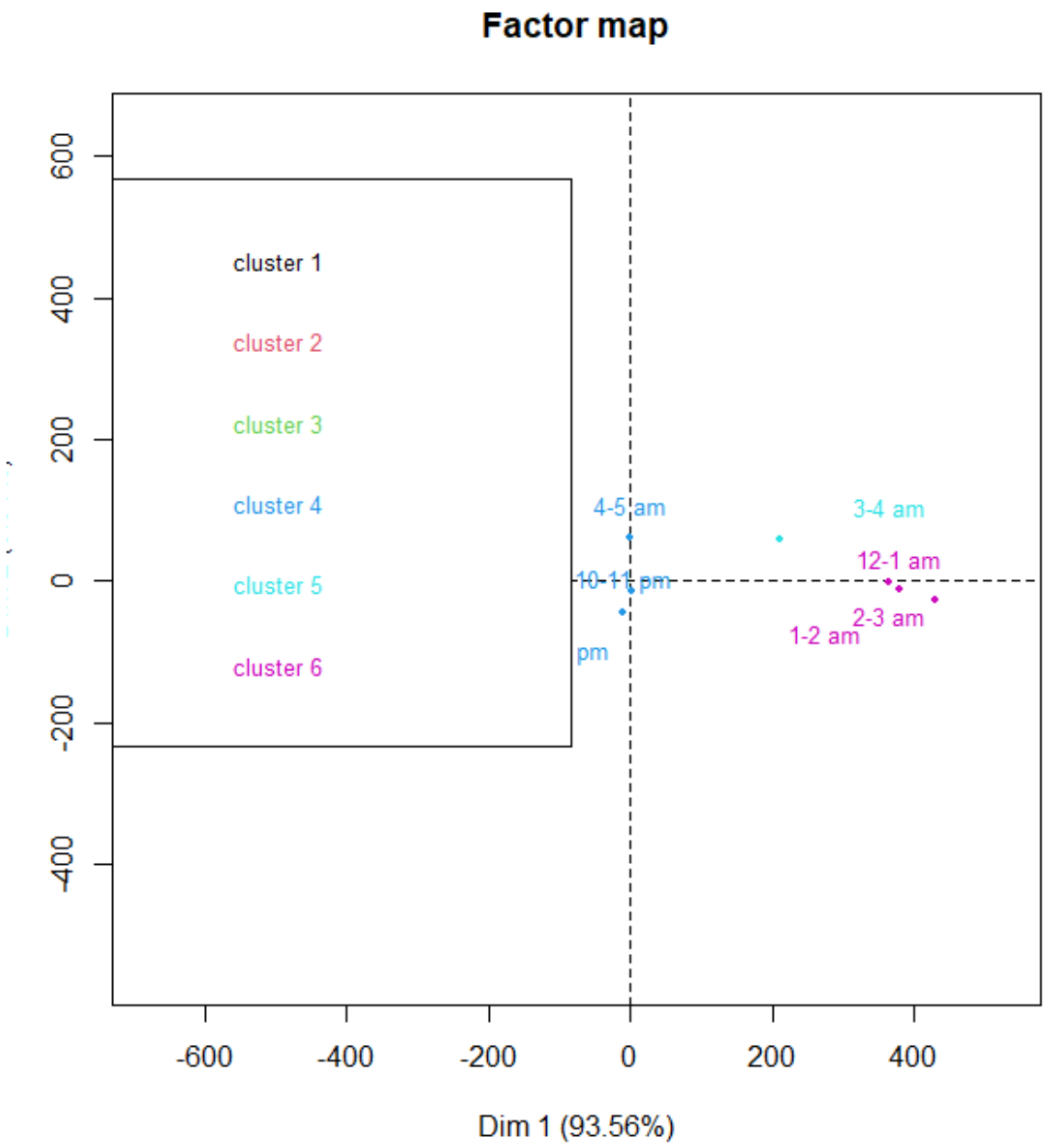


Figure 3. 15. Clusters (groupes) et tranches d'heures sur la carte des facteurs

3.2.2. Classification ascendante hiérarchique pour les sites sentinelles

3.2.2.1. Construction de l'arbre de classification

Dendrogramme

Pour rassembler les individus en construisant un arbre, on les assemble 2 à 2 en commençant par les individus les plus proches. Ainsi, la construction de l'arbre de classification repose sur les distances entre individus.

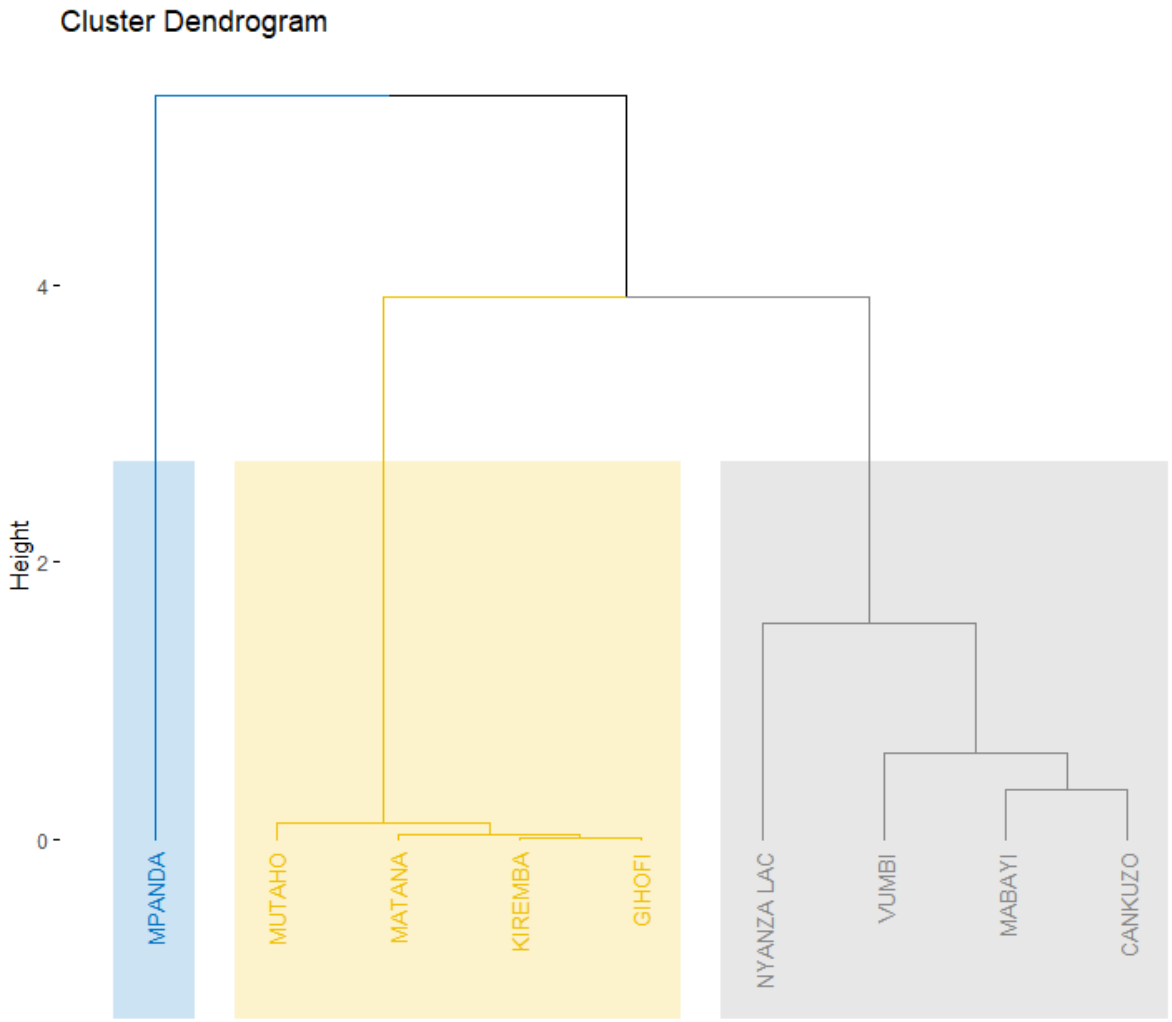


Figure 3. 16. Dendrogramme pour les sites sentinelles

Ce dendrogramme a été créé avec la subdivision finale de 3 groupes selon leur similarité de niveau de variation des piqures.

Le premier groupe comporte un seule site sentinelle, le second comporte 4 sites sentinelles, le troisième comporte 4 sites sentinelles. Cette classification est basée sur les 12 variables utilisés pour l'ACP.

Caractérisation du cluster1

Les sites sentinelles de KIREMBA, MATANA, GIHOFI, MABAYI et MUTAHO est celui qui représente le cluster1 car elles sont les cinq dans ce cluster.

Tableau 3. 10. Caractérisation du cluster 1

Cluster 1	Mean in category	Overall mean	p.value
OAF 2019	36.2	111.5	0.04
IAF2020	37.6	138.4	0.04
IAF 2019	42.6	168.8	0.03
OAF2020	33.6	120.4	0.01

Les variables énumérées dans le tableau sont les plus significativement associées avec le cluster 1. Leurs valeurs moyennes dans le cluster 1 sont inférieures à la moyenne globale correspondantes de chaque variable dans tous les groupes. Par conséquent, le groupe 1 se caracterise par un taux faible :

- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2019 ;
- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2020 ;
- ✓ Des piqures du vecterur *An. funestus* s.i. a l'exterieur en 2019 ;
- ✓ Des piqures du vecterur *An. funestus* s.l. a l'exterieur en 2020

Caractérisation du cluster2

Les sites sentinelles de CANKUZO, VUMBI et NYANZA LAC représentent le cluster 2 car elles sont les trois dans ce cluster.

Tableau 3. 11. Caractérisation du cluster 2

Cluster 2	Mean in category	Overall mean	p.value
OAF2019	256	1111.5	0.01
OAF2020	255	120.4	0.01
IAF2019	386.3	168.8	0.02
IAF2020	302.6	138.4	0.03

Les variables énumérées dans le tableau sont les plus significativement associées avec le cluster 2. Leurs valeurs moyennes dans le cluster 2 sont inférieures à la moyenne globale correspondantes de chaque variable dans tous les groupes.

Par conséquent, le groupe 2 se caracterise par un taux faible :

- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2019 ;
- ✓ Des piqures du vecteur *An. funestus* s.l. a l'interieur en 2020 ;
- ✓ Des piqures du vecterur *An. funestus* s.i. a l'exterieur en 2019 ;
- ✓ Des piqures du vecterur *An. funestus* s.l. a l'exterieur en 2020

Caractérisation du cluster3

Les sites sentinelles de MPANDA représentent le cluster 3 car elle est la seule dans ce cluster.

Tableau 3. 12. Caractérisation du cluster 3

Cluster3	Mean in category	Overall mean	p.value
OAG2020	3430	738	0.006
OAG2019	4080	896.6	0.006
IAG2021	2339	492.5	0.007
OAG2021	2699	597.2	0.008
IAG2020	2744	617.3	0.008
IAG2019	3019	794.6	0.01

Les variables énumérées dans le tableau sont les plus significativement associées avec le cluster 3. Leurs valeurs moyennes dans le cluster 3 sont supérieures aux valeurs moyennes globales dans tous les clusters. Par conséquent, on peut conclure que le cluster se caractérise par un nombre élevé :

- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2019 ;
- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2020 ;
- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'exterieur en 2021 ;
- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2019 ;
- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2020 ;
- ✓ Des piqures des vecteurs *An. gambiae* s.l. a l'interieur en 2021

3.2.2.2. Représentation des clusters (groupes) et des tranches d'heures en 3D

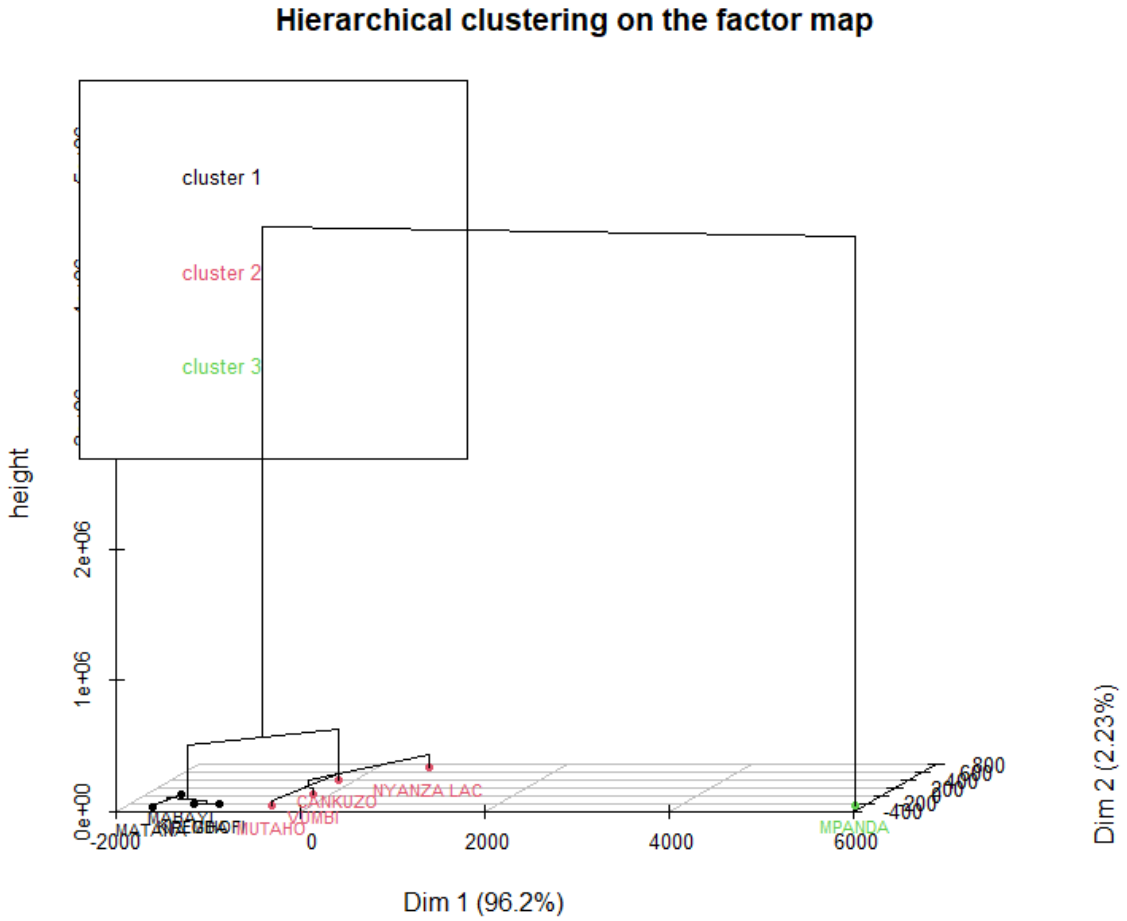


Tableau 3. 17: Les clusters (groupes) et les sites sentinelles en 3D

La représentation graphique de la projection des tranches d'heures sur les axes principaux et du dendrogramme ainsi que les groupes d'appartenance des tranches sites sentinelles selon leur ressemblance est alors donnée par cette figure en trois dimensions.

3.2.2.3. Représentation des clusters (groupes) et des sites sentinelles sur la carte des facteurs

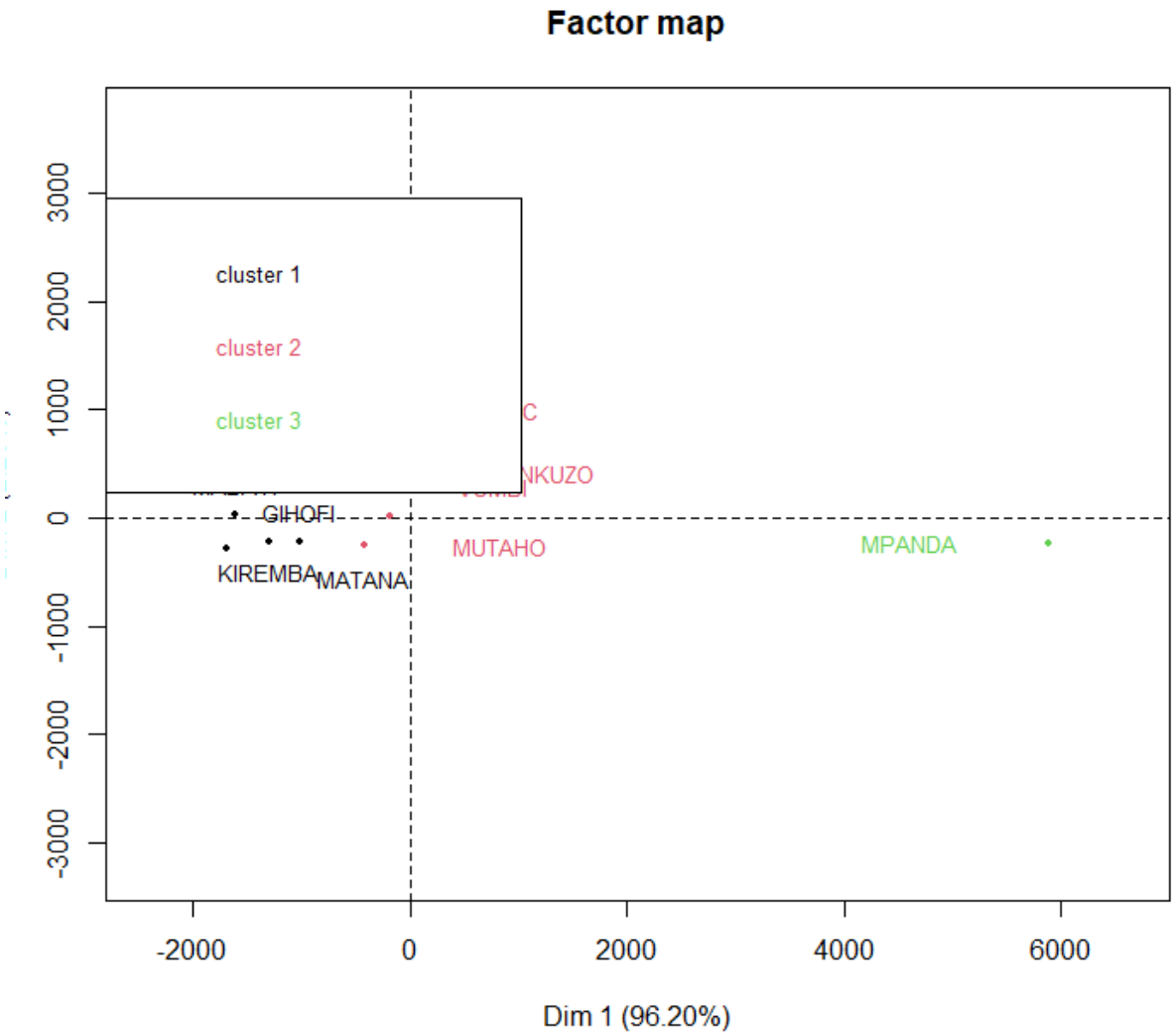


Tableau 3. 18: Clusters (groupes) et sites sentinelles sur la carte des facteurs

3.2.3. Discussions des résultats

Les résultats de notre étude montrent que les piqures de moustiques varient selon les tranches d'heures. Le cycle d'agressivité horaire *d'An. gambiae* s.l. est marqué par un pic au milieu de la nuit entre minuit et 4 heures aussi à l'intérieur qu'à l'extérieur sur toute la période de l'étude. Les vecteurs *An. gambiae* s.l. ont été agressifs à l'extérieur en 2019 et en 2020 pendant les tranches d'heure de 6-7pm, 7-8pm, 8-9pm et 9-10pm. Dans les tranches d'heures de 12-1 am, 1-2 am et 1-3 am, les vecteurs *An. gambiae* s.l. ont été agressifs à l'intérieur et à l'extérieur en 2019, 2020 et 2021.

Ces résultats sont similaires aux résultats trouvés dans l'étude menée par J.M.C. Doannio et ses coequipiers en 2002 sur la dynamique de la transmission du paludisme à Kafiné, un village rizicole en zone de savane humide de Côte d'Ivoire ou ils ont trouvé que « Le cycle général d'agressivité d'*An. gambiae* s.l. à l'intérieur des maisons peut être divisé en trois périodes. Entre 18heures et 21 heures entre 22 heures et 24 heures, cette activité s'accroît et devient intense entre 0 heure et 4 heures, avec un pic entre 1 heure et 2 heures. L'agressivité majeure des femelles *d'An*. *gambiae* s.l. , à l'intérieur comme à l'extérieur des habitations de Kafiné, se situe donc entre 21 heures et 5 heures ».

Le cycle d'agressivité horaire *d'An*. *funestus* s.l. est marqué par un pic au milieu de la nuit entre 2 heures et 5 heures aussi à l'intérieur qu'à l'extérieur sur toute la période de l'étude. Les vecteurs *An*. *funestus* s.l. ont été agressifs à l'intérieur en 2019 et en 2020, A l'extérieur les vecteurs *An. funestus* s.l. ont été agressifs en 2019, 2020 et 2021 pendant les tranches d'heures de 12-1 am, 1-2 am et 1-3am. Ces résultats sont aussi semblables aux résultats de l'étude menée par The PMI VectorLink Project Burundi Annual report : December 30, 2020.

Pour les sites sentinelles les résultats de notre étude montrent que les piqures de moustiques varient selon les sites sentinelles au Burundi. A l'intérieur et à l'extérieur les vecteurs *An. gambiae* s.l. ont été agressifs en 2019 et 2020 à MPANDA, KIREMBA, MATANA, GIHOFI et MUTAHO et les vecteurs *An. funsestus* s.l. ont été agressifs à l'intérieur et à l'extérieur en 2019,2020 et 2021 à MABAYI, CANKUZO, VUMBI et NYANZA LAC. Ces informations sont cruciales pour la mise en place de stratégies de prévention et de contrôle des maladies transmises par les moustiques au Burundi. Ces résultats sont aussi semblables aux résultats de l'étude menée par The PMI VectorLink Project Burundi Annual report : December 30, 2020. Dans cette etude, l'activité de piqûre *d'An. gambiae* s.l. et d'*An. funestus* s.l. a varié dans le temps et dans les sites sentinelles. Les taux de piqûre les plus élevés ont été observés à MPANDA, à NYANZA-LAC, à MUTAHO, à MATANA, a GIHOFI et KIREMBA. Dans la plupart des sites, les taux de piqûres de vecteurs du paludisme ont diminué après les interventions de lutte antivectorielle, sauf à MPANDA et à KIREMBA où les taux de piqûre sont restés élevés après la campagne de distribution de masse de MII.

3.2.4. Les limites de l'etude

Notre étude se limite à l'analyse des données entomologiques recueillies entre 2019 et 2021, ce qui peut ne pas refléter les dynamiques actuelles des populations de moustiques. De plus, les conditions environnementales et les pratiques humaines qui influencent les populations de

moustiques peuvent avoir changé depuis la période de collecte des données, ce qui limite la généralisation de nos résultats à l'ensemble du Burundi ou à d'autres périodes.

L'etude se concentre sur certaines regions du Burundi notammenet les regions où sont instalées les sites sentinelles, ceci peut limiter la representativite des resultats a l'ensemble du pays.

Les données de notre étude se limites a la période allant de Janvier 2019 à Décembre 2021, ceci ne nous a pas premis de répresenter pleinement la diversité des populations de moustiques et des conditions climatiques au Burundi.

Les variations climatiques d'une année à une autre peuvent aussi influencer les résultats, rendant difficile la généralistions des conclusions.

CONCLUSION GÉNÉRALE

Dans ce travail, qui portait sur l'analyse éxploratoire des piqûres de moustiques, cas de l'*An. gambiae* s.l. et l*'An. funsetus* s.l. , nous avons mené une analyse en composante principale et une classification ascendante hiérarchique qui nous ont permis de réduire la dimensionnalité de la base de données et ensuite de regrouper les individus en clusters suivant leurs caractéristiques. L'utilisation de ces techniques nous a permis d'identifier les tranches d'heures où le risque de piqûres de moustiques est plus élevé mais aussi les sites sentinelles où les piqures de moustiques sont plus élevés.

Ces résultats fournissent des informations précieuses pour les autorités de santé publique et les organisations impliquées dans la lutte contre les maladies vectorielles au Burundi. Ils soulignent également l'importance de la surveillance continue et de l'adaptation des stratégies de contrôle en fonction des variations temporelles des populations de moustiques. Enfin, cette étude ouvre la voie à de futures recherches comme :

- ✓ L'utilisation des modeles mathematiques et informatiques pour predire les epidemies de paludisme en fonctions des donnees environnementales, climatiques et entomologiques, ceci pourait ameliorer les reponses rapides et les mesures de prevention.
- ✓ Etudier l'impact des changements climatiques sur la distribution et le comportements des vecteurs moustiques vecteurs du paludisme permettrait de prevoir et de s'adapter aux nouvelles menaces emeregentes ;….

Ces recherches aideront à affiner les méthodes d'analyse des données pour une meilleure compréhension des dynamiques des maladies transmises par les moustiques, C'est pour cette raison que nous invitons tout autre chercheur désirant étendre cette recherche sur les autres variables, qu'il soit le bienvenu. Pour ameliorer les conditions de lutte contre le paludisme, nous formulons les recommandations suivantes :

Au niveau du Programme de lutte contre le paludisme :

- ✓ Renforcer la surveillance entomologique par l'amelioration de la collecte des donnees en utilisant des technologies avencees et en foramant davantage de personnels qualifies ;
- ✓ Developper des approches integrees de lutte antivectorielle par la combinaison des methodes de lutte chimique, biologiques et environementales pour une gestion plus efficace des populations de moustiques ;
- ✓ Sensibiliser et eduquer la communaute par l'augmentation des efforts de sensibilisation pour encourager l'adoption des mesures preventives par les communautes locales ;

- ✓ Encourager les leaders communautaires a joure un role actif dans la sensibilisation et l'education de la population sur les mesures de prevention du palusme ;
- ✓ Appuiyer la recherche et innovation par l'encouragement de la recherche sur de nouvelles interventions et technologies pour la prevention ;
- ✓ Developper des partenariats et collaboration par la promotion de la collaboration entre les agences de sante, les gouvrnements locaux et les organisations internationales pour une reponse coordonnee au paludisme.

Au niveau des populations :

- ✓ Utuliser regulierement les moustiquaires impregnees d'insecticide non seulement la nuit mais aussi lors des siestes pendant la journee ;
- ✓ Reduire les habitats de reproduction des moustiques par la bonne gestion de l'envieronnement ;
- ✓ Promouvoir l'utilisation de vetements couvrants (ayant de manches longues et pentalons) surtout au crepiscule et a l'aube lorsque les moustiques sont plus actifs ;
- ✓ Promouvoir l'utilisation repulsifs anti moustiques sur la peau et les vetement surtout pour les personnes qui restent a l'exterieur pendant les heures de picd'activite des moustiques ;

RÉFÉRENCES BIBLIOGRAPHIQUES

1. **Alain BACCINI et Philippe BESSE**, *« Data mining, publication du laboratoire de statistique et probabilités »*, Université PAUL SABATIER, septembre 2004 ;
2. **Burke, A., Dahan-Moses, Y.,Duncan,F. & Coetzee,M.,***Anopheles funestus mosquitoes from rural southern Mozambique are highly succeptible to pyrtroid insecticides*. Parasites & Vectors.2019; 12(1),175
3. **Chanyalew T, Natea G, Amenu D, Yewhalaw D,** Simma EA. *Composition of mosquito fauna and insecticide resistance status of Anopheles gambiae* sensu lato *in Itang special district, Gambella, Southwestern Ethiopia*. Malaria Journal. 2022;21(1):1–10.
4. **Dahan-Moss Y, Hendershot A, Dhoogra M, Julius H, Zawada J, Kaiser M, et al**. *Member species of the Anopheles gambiae complex can be misidentified as Anopheles leesoni*. Malaria Journal. 2020;19(1):1–9;
5. **Gabor J. Székely et Maria L. Rizzo,** *Hierarchical clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method*., Journal of Classification, vol. 22, nº 2, septembre 2005, p. 151-183;
6. **Ward, J.** *Hierarchical Grouping to Optimize an Objective Function.* Journal of the American Statistical Association, 1936; 58(301), 236-244.
7. **Harold Hotelling,** *Analysis of a Complex of Statistical Variables with Principal Components »*, 1933, Journal of Educational Psychology.
8. **Jean-Paul Benzécri** ; *Analyse des données*. T2 (leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du Laboratoire de statistique de l'Université de Paris 6. T. 2 : l'analyse des correspondances), Dunod Paris Bruxelles Montréal, 1973.
9. **Jean-Paul Benzécri et Al.,** *Pratique de l'analyse des données*. T2 (abrégé théorique. Études de cas modèle), Dunod Paris, 1984
10. **J.M.C. Doamnio et al**, *La dynamique de la transmission du paludisme à Kafiné, un village rizicole en zone de savane humide de Côte d'Ivoire*, Bull Soc Pathol Exot,2002,1,11-16.

11. **Keïta M, Doumbia S, Sissoko I, Touré M, Diawara SI, Konaté D, et al.** I*ndoor and outdoor malaria transmission in two ecological settings in rural Mali*: implications for vector control. Malar J. 2021 Mar 4;20(1):127.

12. **Muturi, E. J., et al.** *(2008). Ecology and distribution of Anopheles mosquito species in relation to malaria transmission in East African highlands*. *Journal of Vector Ecology, 33(1), 1-8.*

13. **Ndayiragije, A., et al.** (2017). Malaria resurgence in Burundi: community perception and strategies for prevention and control. *Malaria Journal, 16(1), 1-9.*

14. **Organization WHO, World malaria report 2022**. *World Health Organization*; 2022.

15. **Organization WHO, Global vector control response 2017-2030**. *Global vector control response* 2017-2030. 2017;

16. **Ouattara Fangman Alain,** *Traitement et analyse statistiques de l'information comptable et financière produite par les compagnies d'assurance de la conférence interafricaine des marches d'assurances*, CIMA BP LIBREVILLE GABON, Novembre 2012 ;

17. **Pearson, K.,** *On Lines and Planes of Closest Fit to Systems of Points in Space* », Philosophical Magazine, vol. 2, n° 6, 1901, p. 559–572;

18. **PNILP B.** *Plan Stratégique National de Lutte contre le paludisme 2021-2027*. Burundi ; 2021 Nov p. 12–83.

19. **Philippon B. & Mouchet J** - *Répercussion des aménagements hydrauliques à usage agricole sur l'épidémiologie des maladies à vecteurs en Afrique Intertropicale. Colloque International : l'eau et les activités agricoles*, Paris, 3-5 mars. C a h i e r C.E.N.E.C.A, 1976, 3212, 13pp.

20. **Philippe Apparicio et Jeremy Gelb**, *Méthodes quantitatives en Sciences Sociales*, 24 février 2023, page 661.

21. **Pr Zineb SAYL**, Cours *: Analyse des données,* A/A 2019_2020

22. **Rachid Bounsiar et Sabrina TEBANI**, *Etude des indicateurs du développement et classification des pays africains : Approche statistique, Mémoire de Master en Mathématiques,* Université MOULOOUD MAMMERI, 2019 ;

23. **Sinzinkayo D, Baza D, Gnanguenon V, Koepfli C.** *The lead-up to epidemic transmission: malaria trends and control interventions in Burundi 2000 to 2019*. Malar J. 2021 Jul 2;20(1):298.

24. **TAJOUGANG I, SAME EKOBO A, EBO'O EYENGA V & ENYONG P**, *Etude de la faune vectorielle du site du futur barrage de Memve'ele(Cameroun),* Bull Soc PatholExot' 1994, 87,261-266

25. **The PMI VectorLink Project. Burundi Annual report**: January 31, 2019. Rockville, MD. The PMI VectorLink Project, Abt Associates Inc.

26. **The PMI VectorLink Project. Burundi Annual report**: December 30, 2020. Rockville, MD. The PMI VectorLink Project, Abt Associates Inc.

27. **The VectorLink Burundi Project. December 2019**. *Entomological Monitoring Annual Report. October 2018-September 2019.* Rockville, MD. The VectorLink Project, Abt Associates Inc.

ANNEXE A: LOGICIEL R

- Le langage R est un langage de programmation et un logiciel libre destiné aux statistiques environnement mathématique utilisés pour le traitement de données. Il permet de faire des analyses statistiques aussi bien simples que complexes comme des modèles linéaires ou non-linéaires, des tests d'hypothèse, de la modélisation de séries chronologiques, de la classification, etc. Il dispose également de nombreuses fonctions graphiques très utiles et de qualité professionnelle.

- R a été créé par Ross Ihaka et Robert Gentleman en 1996 du département de statistique de l'Université d'Auckland, en Nouvelle Zélande, et est maintenant développé par la R développement Core Team. Il est conçu pour pouvoir être utilisé avec les système d'exploitation Unix, Linux, Windows et MacOS. Le R est une application n'offrant qu'une invite de commande il est basé sur la notion de vecteur, ce qui simplifie les calculs mathématique et réduit considérablement le recours aux structures itératives (boucles for, ...etc.). Programmes courts, en général quelques lignes de code seulement. Temps de développement très court.

ANNEXE B : APPLICATION

##Les packages ##

ade4, FactorMineR, reader, NCmisc, ggplot, factoextra, psych, cluster, knitr, corrplot

##Programmation##

```
library(reader)
library(NCmisc)
library(FactoMineR)
library(ggplot2)
library(factoextra)
library(psych)
library(cluster)
library(knitr)
library(corrplot)
library(ade4)
```

#########1. Analyse en composantes principales ##################################

##1.1. Les statistiques de base pour les tranches d'heures ##########################

##1.1.1. Importation des données ##

HBTA2019_2021 <- read.csv ("HBTA2019_2021.csv", sep = " ;", header =T, row. names = "Time")
HBTA2019_2021
dim (HBTA2019_2021)
str (HBTA2019_2021)
summary (HBTA2019_2021)
describe (HBTA2019_2021)

##1.1.2. Les intervalles de confiance ##

ICAN_IAG201911 <-596-1.96*144.97/sqrt(12)
ICAN_IAG201912 <-596+1.96*167.75/sqrt(12)
ICAN_IAG201911
ICAN_IAG201912

ICAN_OAG201911 <-672.5-1.96*167.75/sqrt(12)
ICAN_OAG201912 <-672.5+1.96*167.75/sqrt(12)
ICAN_OAG201911
ICAN_OAG201912

ICAN_IAF201911 <-126.67-1.96*43.19/sqrt(12)
ICAN_IAF201912 <-126.67+1.96*43.19/sqrt(12)
ICAN_IAF201911
ICAN_IAF201912

ICAN_OAF201911 <-83.67-1.96*33.65/sqrt(12)
ICAN_OAF201912 <-83.67+1.96*33.65/sqrt(12)
ICAN_OAF201911
ICAN_OAF201912

ICAN_IAG202011 <-463.08-1.96*96.13/sqrt(12)
ICAN_IAG202012 <-463.08+1.96*96.13/sqrt(12)
ICAN_IAG202011
ICAN_IAG202012

ICAN_OAG202011 <-553.5-1.96*131.9/sqrt(12)
ICAN_OAG202012 <-553.5+1.96*131.9/sqrt(12)
ICAN_OAG202011
ICAN_OAG202012

ICAN_IAF202011 <-103.92 -1.96*25.51/sqrt(12)
ICAN_IAF202012 <-103.92 +1.96*25.51/sqrt(12)
ICAN_IAF202011
ICAN_IAF202012

```
ICAN_OAF202011 <-90.42-1.96*39.27/sqrt(12)
ICAN_OAF202012 <-90.42+1.96*39.27/sqrt(12)
ICAN_OAF202011
ICAN_OAF202012

ICAN_IAG202111 <-369.4-1.96*78.57/sqrt(12)
ICAN_IAG202112 <-369.4+1.96*78.57/sqrt(12)
ICAN_IAG202111
ICAN_IAG202112

ICAN_OAG201911 <-447.9-1.96*100.2/sqrt(12)
ICAN_OAG201912 <-447.9+1.96*100.2/sqrt(12)
ICAN_OAG201911
ICAN_OAG201912

ICAN_IAF201911 <-92.08-1.96*25.49/sqrt(12)
ICAN_IAF201912 <-92.08+1.96*25.49/sqrt(12)
ICAN_IAF201911
ICAN_IAF201912

ICAN_OAF201911 <-83.42-1.96*28.38/sqrt(12)
ICAN_OAF201912 <-83.42+1.96*28.38/sqrt(12)
ICAN_OAF201911
ICAN_OAF201912
```

##1.1.3. Construction de la matrice de corrélation ##################################

```
corre<-  round(cor(HBTA2019_2021),2)
corrplot (corre, method="number", type = "upper")
```

##1.1.4. Analyse en Composantes Principales normée pour les tranches d'heures ######

```
res. HBTA2019_2021 <- PCA (HBTA2019_2021[,1 :12])

par(mfrow=c(1,2))

plot(res,choix="ind", cex=0.8, title="", graph.type="classic")

plot(res,choix="var", cex=0.8, title="", graph.type="classic")
```

##1.1.5. Biplot pour les tranches d'heures ##

```
Combinaison des graphiques

fviz_pca_biplot(res, repel = TRUE,
                col.var = "#FC4E07", # Variables color
                col.ind = "#00AFBB" # Individuals color)
```

1.1.6. Eboulie des valeurs propres pour les tranches d'heures

```
valeur.propre <- get_eigenvalue(res.pca)
print (valeur.propre, hist="valeur.propre")

kable(res$eig, caption = "Tableau des valeurs propres")
fviz_eig(res, addlabels = TRUE, ylim = c(0, 100))
```

1.1.7. Contribution pour les composantes principales

```
kable(round(res$var$contrib,2), caption = "Tableau de contribution des variables")

fviz_pca_var(res, col.var = "contrib",

        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")

)
```

##1.2. Les statistiques de base pour les sites sentinelles ##############################

##1.2.1. Importation des données ##

```
HBTL2019_2021 <- read.csv ("HBTL2019_2021.csv", sep = ";", header =T, row. names =
"Locality")
HBTL2019_2021
dim(HBTL2019_2021)
str(HBTL2019_2021)
summary(HBTL2019_2021)
describe(HBTL2019_2021)
```

##1.**2.2. Les intervalles de confiance ##**

```
ICAN_IAG201911 <-794.7-1.96*928.89/sqrt(12)
ICAN_IAG201912 <-794.7+1.96*928.89/sqrt(12)
ICAN_IAG201911
ICAN_IAG201912

ICAN_OAG201911 <-896.67-1.96*1251.65/sqrt(12)
ICAN_OAG201912 <-896.67+1.96*1251.65/sqrt(12)
ICAN_OAG201911
ICAN_OAG201912

ICAN_IAF201911 <-168.9-1.96*202.69/sqrt(12)
ICAN_IAF201912 <-126.67+1.96*206.69/sqrt(12)
ICAN_IAF201911
ICAN_IAF201912
```

```
ICAN_OAF201911 <-111.56-1.96*126.08/sqrt(12)
ICAN_OAF201912 <-111.56+1.96*126.08/sqrt(12)
ICAN_OAF201911
ICAN_OAF201912

ICAN_IAG202011 <-617.33-1.96*852.37/sqrt(12)
ICAN_IAG202012 <-617.33+1.96*852.37/sqrt(12)
ICAN_IAG202011
ICAN_IAG202012

ICAN_OAG202011 <-738-1.96*1052.56/sqrt(12)
ICAN_OAG202012 <-738+1.96*1052.56/sqrt(12)
ICAN_OAG202011
ICAN_OAG202012

ICAN_IAF202011 <-138.44 -1.96*169.59/sqrt(12)
ICAN_IAF202012 <-138.44 +1.96*169.59/sqrt(12)
ICAN_IAF202011
ICAN_IAF202012

ICAN_OAF202011 <-120.44-1.96*122.38/sqrt(12)
ICAN_OAF202012 <-120.44+1.96*122.38/sqrt(12)
ICAN_OAF202011
ICAN_OAF202012

ICAN_IAG202111 <-492.56-1.96*726.26/sqrt(12)
ICAN_IAG202112 <-492.56+1.96*726.26/sqrt(12)
ICAN_IAG202111
ICAN_IAG202112

ICAN_OAG202111 <-597.22-1.96*842.12/sqrt(12)
ICAN_OAG202112 <-597.22+1.96*842.12/sqrt(12)
ICAN_OAG202111
ICAN_OAG202112

ICAN_IAF202111 <-122.78-1.96*132.69/sqrt (12)
ICAN_IAF202112 <-122.78+1.96*132.69/sqrt (12)
ICAN_IAF202111
ICAN_IAF202112

ICAN_OAF202111 <-111.22-1.96*134.51 /sqrt (12)
ICAN_OAF202112 <-111.22+1.96*134.51/sqrt (12)
ICAN_OAF202111
ICAN_OAF202112
```

1.2.3. Construction de la matrice de corrélation##################################

```
corre<-  round(cor(HBTL2019_2021),2)
corrplot (corre, method="number", type = "upper")
```

##1.2.4. Analyse en Composantes Principales normée pour les sites sentinelles #########

```
res.HBTL2019_2021 <- PCA(HBTL2019_2021[,1:12])
par(mfrow=c(1,2))
plot(res.HBTL2019_2021,choix="ind", cex=0.8, title="", graph.type="classic")
plot(res.HBTL2019_2021,choix="var", cex=0.8, title="", graph.type="classic")
```

##1.2.4.1. Cercle des corrélations ##

```
fviz_pca_var(res.HBTL2019_2021, col.var = "contrib",
        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
)
```

##1. 2.4.2. Projection des individus ##

```
plot(res.HBTL2019_2021, choix = "ind", habillage = 2)
fviz_pca_ind(res.HBTL2019_2021, col.ind = "contrib",
        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
        repel = TRUE # Avoid text overlapping (slow if many points)
)
```

1.**2.5. Biplot**

```
#Combinaison des graphiques
fviz_pca_biplot(res.HBTL2019_2021, repel = TRUE,
          col.var = "#FC4E07", # Variables color
          col.ind = "#00AFBB" # Individuals color
)
```

##1. 2.6. Eboulie des valeurs propres ##

```
valeur.propre <- get_eigenvalue(res.HBTL2019_2021)
print(valeur.propre, hist="valeur.propre")
kable(res$eig.HBTL2019_2021, caption = "Tableau des valeurs propres")
fviz_eig(res, addlabels = TRUE, ylim = c(0, 100))
```

1.**2.7. Contribution pour les composantes principales ############################**

```
kable(round(res$var$contrib,2), caption = "Tableau de contribution des variables")

fviz_pca_var(res, col.var = "contrib",

        gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")

)
```

##2. Classification hiérarchique ascendante#####################################

##2.1. Classification hiérarchique ascendante (CAH) pour les tranches d'heures ######

```
library(ade4)
library(factoextra)
library(NbClust)
```

###2.1.1. Construction de l'arbre de classification#################################

```
res. hcpc <- HCPC(res.pca,graph = FALSE)

fviz_dend(res.hcpc,cex = 0.7,palette = "jco",rect = TRUE, rect_fill = TRUE,rect_border =
"jco",labels_track_height = 0.8 )

fviz_cluster(res.hcpc,repel = TRUE,show.clust.cent = TRUE,palette = "jco",ggtheme =
theme_minimal(),main = "Factor map")
```

##2.1.2. Représentation des clusters (groupes) et des villes en 3D##################

```
plot (BD2.val.prop1, choice ="3D.map",draw.tree = F,

    repel = TRUE)
```

##2.1.3. Représentation des clusters (groupes) et des tranches d'heures sur la carte des facteurs ##

```
plot (BD2.val.prop1, choice ="map", draw. tree = F,

    repel = TRUE)
```

##2.2. Classification hiérarchique ascendante (CAH) pour les sites sentinelles##########

```
library(ade4)

library(cluster)
```

##2.2.1. Construction de l'arbre de classification#################################

```
res.hcpc <- HCPC(res.HBTL2019_2021,graph = FALSE)

fviz_dend(res.hcpc,cex = 0.7,palette = "jco",rect = TRUE, rect_fill = TRUE,rect_border =
"jco",labels_track_height = 0.8 )
```

```
fviz_cluster(res.hcpc,repel = TRUE,show.clust.cent = TRUE,palette = "jco",ggtheme =
theme_minimal(),main = "Factor map")
```

##2.2.2. Représentation des clusters (groupes) et des villes en 3D ####################

```
plot (BD2.val.prop1, choice ="3D.map", draw. tree = F,
    repel = TRUE)
```

##2.2.3. Représentation des clusters (groupes) et des villes sur la carte des facteurs######

```
plot (BD2.val.prop1, choice ="map", draw.tree = F,
    repel = TRUE)
```