Revue Internationale du Chercheur
ISSN: 2726-5889
Volume 6 : Numéro 1

Analyse de données par les méthodes factorielles : Application aux maladies sexuellement transmissibles à la Division Provinciale de la Santé (DPS) du Kasai Central.

Data analysis by factorial methods: application to sexually transmitted diseases for the provincial division of Health (DPS) of Central Kasai.

Jean MUTOMBO NTUMBA
Doctorant
Université Pédagogique de Kananga
République Démocratique du Congo

Date de soumission : 07/02/2025
Date d'acceptation : 18/03/2025

Pour citer cet article :

MUTOMBO NTUMBA. J (2025) « Analyse de données par les méthodes factorielles : Application aux maladies sexuellement transmissibles à la Division Provinciale de la Santé (DPS) du Kasai Central.», Revue Internationale du chercheur « Volume 6 : Numéro 1 » pp : 1061 - 1082

Résumé

Cet article a analysé, par les méthodes factorielles, les maladies sexuellement transmissibles (MST) qui constituent un problème de santé publique dans la province du Kasaï Central, en République Démocratique du Congo.

Les MST sont des maladies infectieuses, contagieuses, d'étiologies très diverses par le biais de contacts sexuels non protégés, qu'ils soient vaginaux, anaux ou oraux. Elles peuvent être causées par des bactéries, des virus ou des parasites ; et peuvent avoir des symptômes variés ou être asymptomatiques, ce qui rend leur détection précoce importante pour éviter des complications graves. Après l'analyse en composantes principales (ACP), les Zones de Santé les plus touchées sont : Masuika, Tshibala, Kananga, Katende, Mikalayi, Bena tshiadi, Bena Leka et Luambo, parmi les 26 Zones de santé de la Division Provinciale de Santé du Kasaï Central.

La recherche suggère de mener une sensibilisation à travers les organisations à assise communautaire et autres canaux de communication pour lutter efficacement contre ce phénomène dans la province du Kasaï Central.

Mots clés : Analyse; Méthode factorielle; Maladies ; Sexuellement ; Transmissibles.

Abstract

This article analyzed, by factorial methods, sexually transmitted diseases (MST) which constitute a public health problem in the province of Central Kasai, in the Democratic Republic of Congo.

MST are infectious, contagious diseases, very diverse etiologies through unprotected sexual contacts, whether vaginal, anal or oral. They can be caused by bacteria, viruses or parasites; and can various symptoms or be asymptomatic which makes their early detection significant to avoid serious complications. After the main component analysis (ACP), the most affected health areas are: Masuika, Tshibala, Kananga, Katende, Mikalayi, Bena tshiadi, Bena Leka and Luambo, among the 26 health areas of the provincial health division of Central Kasai.

Research suggests carrying out awareness through organizations with community Assisi and other communication channels to effectively fight this phenomenon in the province of Central Kasai.

Keywords: analysis; Factorial method; Diseases; Sexually transmitted

Introduction

Les maladies sexuellement transmissibles (MST) constituent un problème de santé publique du fait de leur fréquence, de leur recrudescence (Santi, 2018) et des complications cliniques qui en résultent (Janier, 2009).

Elles ne sont pas seulement une cause importante de morbidité chez les adultes, mais elles peuvent provoquer des complications avec des séquelles, telles que la stérilité chez les hommes et les femmes, une grossesse ectopique, le cancer du col de l'utérus, un faible poids de naissance, ainsi que la prématurité ou la conjonctivite du nouveau-né. En effet, l'infection génitale basse non traitée peut évoluer vers une infection génitale haute, elle-même à l'origine par la suite de grossesse ectopique, de trouble de la fertilité tubaire, dont elles représentent la première cause chez la jeune femme (Goma, 2018).

Dans de nombreux pays en développement, en RDC et particulièrement dans la province du Kasaï Central, les infections sexuellement transmissibles (IST) figurent depuis des décennies parmi les cinq principales maladies qui amènent les adultes à s'adresser aux services de santé. Il y a rarement de contrôle fiable, et l'ampleur du problème est souvent inconnue. Là où l'on dispose de données, elles font état d'une proportion de malades beaucoup plus élevée dans la tranche d'âge des 15-44 ans.

Par ailleurs, l'analyse de données regroupe deux familles de méthodes suivant les deux objectifs :

Une partie des méthodes cherche à représenter de grands ensembles de données par peu de variables, c'est-à-dire, recherche les dimensions pertinentes de ces données. Les variables ainsi déterminées permettent une représentation synthétique recherchée. Parmi ces méthodes, de nombreuses analyses sont issues de l'analyse factorielle, telles que l'analyse en composantes principales, l'analyse factorielle de correspondances, l'analyse de correspondances multiples, ou encore l'analyse canonique.

L'analyse factorielle des correspondances a été conçue pour l'étude des tableaux de contingence obtenus par croisement de variables qualitatives. Cette analyse permet de traiter des variables qualitatives, et est surtout adaptée à ce type de variables. L'analyse factorielle des correspondances multiples est une extension de l'analyse factorielle des correspondances qui ne permet que le croisement de deux variables qualitatives (Kostov, et al., 2023).

Une autre partie des méthodes cherche à classer les données de manière automatique. Ces méthodes sont complémentaires avec les précédentes pour synthétiser et analyser les données afin de répondre plus particulièrement à l'objectif fixé de caractériser les proximités entre

individus et celles entre variables. C'est donc dans cette optique qu'il faut situer notre étude basée sur l'analyse en composantes principales des données de maladies sexuellement transmissibles

Depuis les premières publications annonçant les principes de l'Analyse Discriminante des Données symboliques, beaucoup de travaux ont été réalisés. La question centrale de recherche en Analyse des Données Symboliques (ADS) est celle d'étendre de façon inéluctable les outils de l'Analyse des Données classiques et du Data Mining aux cas des données symboliques (Afonso, et al., 2018).

Kasiama a mené une étude dont l'objectif était d'étendre les techniques de l'Analyse Factorielle Classique (en l'occurrence, l'ACP et l'AFD) aux cas de données symboliques afin d'en extraire des connaissances. L'auteur a cherché à améliorer la méthode des centres, une extension de la méthode standard d'ACP à des données symboliques du type intervalle, afin d'obtenir une nouvelle méthode dite méthode des moyennes géométriques (Kasiama, 2020).

Dans son mémoire de master en santé publique, option épidémiologie, intitulé « Prévalence des IST et VIH/SIDA à la clinique de santé sexuelle des Halles de Bamako », Coulibaly a abordé ce thème en se focalisant sur l'analyse univariée et multivariée. Il a fait une régression logistique qui est utilisée lorsqu'on cherche à étudier la relation entre une variable dépendante binaire Y et des variables explicatives qui peuvent être qualitatives ou quantitatives, avec comme objectif d'expliquer la variation de Y en fonction des variables explicatives (Coulibaly, 2018).

Dans leur article intitulé « Usage de TIC dans la prévention des IST et du VIH chez les adolescents et les jeunes en Côte d'Ivoire : à l'assaut du sida », pour assurer une prévention efficace du VIH-sida, l'étude avait pour objectif d'examiner l'impact de l'utilisation de réseaux sociaux numériques sur la motivation autodéterminée les comportements de prévention chez les jeunes en Côte d'Ivoire. Il visait aussi à comprendre comment les interactions et les informations diffusées via ces plateformes numériques influencent les motivations intrinsèques et extrinsèques des jeunes, ainsi que leur engagement dans les programmes de dépistage du VIH et l'adoption de comportements sexuels responsables (Silué, et al., 2024).

Pour notre part, l'étude s'articulera autour des questions ci-après :

Comment représenter graphiquement les relations entre les zones de santé par l'évaluation de leur ressemblance?

Comment représenter graphiquement les relations entre les variables par l'évaluation de leurs liaisons ?

L'analyse en composantes principales pourrait représenter graphiquement les relations entre individus (Zones de Santé) par l'évaluation de leurs ressemblances, ainsi que les relations entre variables par l'évaluation de leurs liaisons.

Nous avons opté pour les méthodes d'analyse factorielle notamment l'analyse en composantes principales aux données de maladies sexuellement transmissibles, pour en extraire des connaissances utiles, en nous appuyant sur la technique documentaire, l'outil d'analyse ayant servi pour ce travail étant le logiciel SPAD 5.0.

Pour cerner la réalité de ce phénomène de maladies sexuellement transmissibles, nous avons subdivisé notre réflexion en trois grands thèmes notamment aperçu général de maladies sexuellement transmissibles, les méthodes factorielles, présentation et traitement de données.

Cadre de référence théorique

Le choix de la théorie dans cet article a pour objectif de préciser les facteurs essentiels de la question de départ de cette étude. Dans ce cadre, la théorie de facteurs latents (Spearman, 1904) a été mobilisée en analyse factorielle. En effet, un facteur latent est une variable non observée qui influence des variables observées. Dans le cadre des MST, un facteur latent pourrait être le comportement à risque qui influence plusieurs variables mesurées, comme le nombre de partenaires sexuels, l'usage de préservatifs, ou la fréquence de test de dépistage. Ainsi, avons-nous examiné la corrélation qui existe entre les cas contacts parmi les nouveaux cas de MST et les nouveaux cas de MST.

1. Aperçu général de maladies sexuellement transmissibles

1.1. Les infections sexuellement transmissibles(IST)

Les IST sont des maladies infectieuses, contagieuses, d'étiologies très diverses et d'expressions cliniques variées qui se transmettent entre les personnes par le biais de contact sexuel (y compris lors d'un rapport vaginal, anal ou oral).

Elles sont à distinguer des autres types d'infection du tractus génital notamment les infections endogènes qui sont la croissance excessive de micro-organismes normalement présents dans le vagin. Exemple: Vaginose bactérienne. Les infections iatrogènes sont la conséquence de l'introduction dans le tractus génital de micro-organismes lors des procédures Médico-chirurgicales sans respect des règles d'asepsie. Exemple: avortement provoqué, curetage, pose de stérilet. Les maladies sexuellement transmissibles sont les infections dont le mode de transmission prédominant est l'activité sexuelle par contact avec la muqueuse génitale, anale ou orale (Nouchi, 2018).

1.2. Classification des IST

Deux classifications sont utilisées : une classification basée sur les signes cliniques (syndromique).

Les IST avec écoulements : Urétrite chez l'homme, Cervico-vaginite chez la femme.

Les IST avec ulcérations : Syphilis, Chancre mou, Herpès génital, Lymphogranulomatose vénérienne ou Maladie de Nicolas Favre, Donovalose ou granulome inguinal.

Les IST avec végétations : Condylomes, Verrues.

Les IST avec douleurs pelviennes chez la femme : Salpingite, endométrite.

Une classification basée sur l'étiologie : Selon l'agent infectieux en cause, on distingue : les IST bactérienne, virale, parasitaire, fongique ou ectoparasite.

1.3. Modes de transmission

La voie sexuelle : c'est la principale voie de contamination du VIH et des germes responsables d'IST ;

La voie maternelle (mère enfant) : deux voies de transmission sont possibles :

La voie verticale transplacentaire (syphilis congénitale VIH), la voie filière génitale (conjonctivite du nouveau-né).

La voie sanguine : hépatite B et C, et VIH.

2. Les méthodes factorielles

Dans ce point nous allons montrer les différentes méthodes de l'analyse de données et plus particulièrement les méthodes factorielles. Ces méthodes d'analyse des données ont largement démontré leur efficacité dans l'étude de grandes masses complexes d'informations. Ce sont des méthodes dites multidimensionnelles en opposition aux méthodes de la statistique descriptive qui ne traitent qu'une ou deux variables à la fois. (Escofier & Pages, 2023).

L'analyse multivariée réunit un grand nombre de méthodes, souvent complexes, qui tentent de donner une image simplifiée des multiples relations entre les variables d'une enquête ou d'une base de données (Stafford & Bodson, 2006).

Les méthodes de la statistique exploratoire sont la visualisation et les méthodes de classification. Les méthodes de la statistique décisionnelle sont les méthodes de discrimination et les méthodes de régression. (Kasiama, 2020).

Les méthodes factorielles ont l'ambition de représenter un grand nombre de variables dans un espace de faible dimension. On y représente également les unités statistiques, soit individuellement, soit selon des groupes ou catégories en utilisant un principe barycentrique.
La possibilité de réduire la dimension provient de l'existence de corrélations entre les variables. Cette réduction s'effectue non pas par sélection d'un sous- ensemble de variables mais par la construction de variables synthétiques, combinaisons linéaires des variables initiales.
On distingue les variables selon leur nature : qualitatives ou quantitatives et selon leur fonction dans l'analyse : actives ou illustratives (on dit aussi supplémentaires). Seules les variables actives participent à la détermination de l'espace de représentation appelé espace factoriel.
Les variables actives doivent être toutes de même nature ce qui conditionne la méthode d'analyse : composantes principales pour les variables quantitatives, correspondances pour les variables qualitatives (Saporta, 2021).
Bien évidemment, avant d'effectuer une analyse multidimensionnelle sophistiquée, il est recommandé de prendre contact avec les données au moyen des outils classiques de la statistique descriptive ou de ceux plus récents de la statistique exploratoire :
Tris à plat avec histogrammes, box-plot (appelés parfois boites à pattes ou boites à moustache), courbes de densité pour les variables quantitatives ; camemberts et autres diagrammes pour les variables qualitatives.
Tris croisés qui consistent à ventiler les observations selon deux variables afin d'étudier leur liaison.

2.1. Les méthodes de visualisation

Ces méthodes, basées sur l'algèbre linéaire et la géométrie euclidienne, ont pour but la visualisation du nuage des observations.
Cette visualisation se fait sur un espace de dimension réduite choisie de sorte à minimiser la déformation du nuage de points sur cet espace. Cela revient à chercher un petit nombre de variables synthétiques qui résume au mieux l'ensemble des variables et qui engendre un espace de projection conservant aux données le maximum de variation. Ces méthodes dites descriptives consistent à résumer, visualiser et synthétiser les informations. Il s'agit par exemple de l'Analyse Factorielle des Correspondances, l'Analyse en Composantes Principales et de Classification Ascendante Hiérarchique.
Les méthodes sont différentes selon le type des tableaux utilisés :

L'analyse en composantes principales (ACP), traite les tableaux de données quantitatives, et dont le but est de résumer l'information contenue dans un tableau composé d'un nombre élevé de lignes et de colonnes ;

L'analyse factorielle de correspondances (AFC), traite les tableaux de contingence et des fréquences ; son objectif général est de mettre en évidence les relations dominantes entre modalités des variables nominales initiales (Crucianu, 2021). Elle vise à rassembler en un nombre réduit de dimensions la plus grande partie de l'information initiale en s'attachant non pas aux valeurs absolues mais aux correspondances entre les variables, c'est-à-dire aux valeurs relatives. C'est une méthode factorielle de réduction de dimension pour l'exploration statistique d'une table de contingence définie par deux variables qualitatives. L'analyse factorielle des correspondances s'applique aux tableaux rectangulaires. La méthode proposée pour étudier des tableaux multiples est de choisir le tableau binaire optimal, c'est- à-dire ayant la plus grande variance et d'en faire l'AFC (Choulakian, 1988).

L'analyse factorielle de correspondances multiples (AFCM), traite les tableaux de données qualitatives ;

L'analyse factorielle d'un tableau de distances (AFTD), dont le but est de représenter les points à partir de leurs distances.

Notons que l'analyse en composantes principales est une des techniques les plus utilisées en analyse de données multidimensionnelles. Elle réduit un vecteur par projection orthogonale sur les sous-espaces de dimension fixée a priori qui maximise la variance des projetés. Dans le cas où la variabilité traitée est la variabilité totale, sa solution exacte est le sous-espace engendré par les premiers vecteurs propres de la matrice de covariance, dans l'ordre décroissant des valeurs propres. Cette méthode garantie une erreur minimale (erreur de reconstruction) entre vecteurs initiaux et vecteurs projetés au sens euclidien du terme (principe de moindre inertie). L'un des inconvénients majeurs de l'ACP est l'absence d'un modèle génératif des données et d'une densité de probabilité. En fait, l'ACP suppose implicitement que la distribution des données est un hyper-ellipsoïde caractérisé par sa moyenne et sa matrice de covariance globale (Bouallegue, 2013).

2.2. Les méthodes de classifications

Elles visent à mettre en évidence une typologie des individus autrement dit une structuration des individus en classes homogènes. En intelligence artificielle, la classification automatique est considérée comme un procédé permettant à l'ordinateur de découvrir une information

d'ordre sémantique qui n'était pas dans le tableau initial sous forme claire : on parle alors d'apprentissage sans professeur ou non supervisé. Plusieurs types de méthodes existent selon la structure classificatoire recherchée : partition, hiérarchie ou recouvrement. Toutes ces méthodes produisent en sortie des regroupements homogènes qu'on appelle classes. La caractérisation d'une classe peut être constituée par l'énumération de ses éléments ou par une description représentant l'ensemble de ses éléments. En analyse des données, cette description s'appuie sur des indicateurs statistiques tels que les indicateurs centraux, de dispersion, de distribution, etc.

Les deux types de techniques de classification sont :

La classification non hiérarchique ou partitionnement : la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence mais le nombre m de classes est fixé à l'avance.

La classification hiérarchique : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents.

Les méthodes de classification automatique ne nécessitant pas d'apprentissage offrent un intérêt important lorsque les données sont complètement inconnues. Elles permettent ainsi de dégager des classes qui ne sont pas évidentes a priori. Les deux principales méthodes développées sont la méthode des centres mobiles (apparentée à la méthode des k-means) ou des nuées dynamiques (comme un cas particulier) et la classification hiérarchique ascendante ou descendante. Nous pouvons aussi citer les approches fondées sur les graphes et hypergraphes. La méthode des centres mobiles consiste à associer les individus à des centres de classes choisis aléatoirement, puis à recalculer ces centres jusqu'à obtenir une convergence. La difficulté consiste dans le choix d'une distance appropriée. La classification hiérarchique ascendante (respectivement descendante) consiste à regrouper les individus selon leur ressemblance (respectivement dissemblance). Toute la difficulté est dans la définition d'une mesure de ressemblance et de la distance associée.

2.3. Les méthodes de discrimination et les méthodes de régression

Sous le nom d'analyse discriminante, on désigne toute une série des méthodes explicatives, descriptives et surtout prédictives destinées à étudier une population comportant k classes d'individus (Christophe & Mounir, 2016). Chaque individu est caractérisé par un ensemble de *p* variables quantitatives et une variable qualitative identifiant la classe à laquelle appartient cet

individu. Il s'agit d'un apprentissage supervisé utilisant un ensemble d'exemples où les classes d'appartenance sont connues au préalable. A partir de cet ensemble, les normes (ou règles) d'affectation sont définies.

Quand la variable à expliquer est quantitative, on utilise la méthode de régression. En effet, ces méthodes cherchent à expliquer les valeurs prises sur d'autres variables, dites variables explicatives.

Parmi les méthodes issues de l'analyse discriminante et directement rattachées à l'analyse de données, il y a l'analyse linéaire discriminante, la régression logistique, les k plus proches voisins ou encore les arbres de décision. D'autres méthodes issues de l'intelligence artificielle et du monde de la reconnaissance des formes peuvent être rattachées à l'analyse discriminante telles que le perceptron multicouche (et les autres réseaux de neurones) et les chaines de Markov (Laurent, 2024) ou encore issues de la théorie de l'apprentissage statistique telle que les machines à vecteurs de supports (Myriam, et al., 2018).

Si ces dernières ne sont pas toujours considérées comme faisant partie de l'analyse de données, elles sont parfaitement intégrées dans le data mining. Le terme « data mining » est apparu aux Etats-Unis au milieu des années 1990. Il s'appuie sur la métaphore suivante : « les bases de données des entreprises constituent une mine à partir de laquelle on peut extraire des pépites sous la forme de connaissances synthétiques permettant une aide à la prise de décision ». La traduction en français la plus proche est le terme « fouille de données ».

La science de données (en anglais Data science) est l'extraction de connaissances d'ensemble de données. C'est une discipline qui s'appuie sur des outils mathématiques, des statistiques, d'informatiques et de visualisation de données (Alain, 2022). Elle est en plein développement dans le monde universitaire ainsi que dans le secteur privé et le secteur public.

La Science de Données est un domaine inter-disciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées de nombreuses données structurées et non structurées. Elle est souvent associée aux données massives et l'analyse de données.

Le décideur dans l'entreprise étant celui qui engage la pérennité de l'entreprise, il doit s'entourer de différents moyens lui permettant une prise de décision la plus pertinente. Parmi ces moyens, les entrepôts des données, en anglais : Data Warehouse ont une place primordiale. Le problème fondamental réside dans l'exploitation de ces informations. Pour cela, il est nécessaire d'utiliser Data Mining.

Data Mining est un nouveau champ d'applications à l'interface de la statistique et de nouvelles technologies de l'information (bases de données, intelligence artificielle, apprentissage,…). La métaphore data mining signifie qu'il y a des trésors ou pépites cachés sous des montagnes de données que l'on peut découvrir avec des outils spécialisés (classification, visualisation avec des méthodes telles que ACP, AFC et ACM, etc.)

L'analyse linéaire discriminante est aussi appelée analyse factorielle discriminante car elle est en fait une analyse en composantes principales supervisée. Elle décrit les individus en classes (celles-ci sont données par une variable issue de l'apprentissage) et ensuite affecte de nouveaux individus dans ces classes .C'est donc une méthode à la fois descriptive et prédictive. Elle permet de traiter aussi bien des variables quantitatives que qualitatives.

La régression logistique consiste à exprimer les probabilités a posteriori d'appartenance à une classe *p(C/x)* comme une fonction de l'observation (David et al., 2013).

Bien souvent c'est la régression linéaire qui est employée. L'approche des k plus proches voisins repose sur l'idée simple d'attribuer un nouvel individu à la classe majoritaire parmi ses k plus proches voisins (individus de la base d'apprentissage les plus proches au sens d'une certaine distance).

Les arbres de décision nécessitent souvent une construction délicate et difficilement généralisable si les données d'apprentissage sont peu représentatives de la réalité. La méthode CART (Classification And Regression Tree) possède une construction d'arbre aux propriétés intéressantes pour la segmentation.

3. Présentation et traitement de données

Les données que nous avons recueillies proviennent de la base de données présentée par la division provinciale de la santé du Kasaï Central (DPS) de 2022 à 2024 pour les 26 Zones de santé.

Tableau N°1 : **Données d'IST de 2022 à 2024**

Zone de santé	CC 2022	CC 2023	CC 2024	NC 2022	NC 2023	NC 2024	T.a.S 2022	T.a.S 2023	T.a.S 2024	T.a.E 2022	T.a.E 2023	T.a.E 2024
Bena Leka	1598	2 025	2 532	3780	4 437	6 216	3139	4 896	3 978	106	439	695

Bena Tshadi	1183	1 059	1 065	2891	2 074	1 849	1679	1 677	1 792	95	206	68
Bilomba	587	556	836	2440	3 094	3 661	2576	3 328	2 838	95	80	193
Bobozo	584	636	630	4422	2 931	3 230	2774	3 038	2 756	18	169	129
Bunkonde	50	64	439	3781	5 969	5 219	5378	4 833	5 867	50	107	303
Demba	1366	1 536	1 061	4772	6 940	7 311	5438	5 957	5 185	1593	1 654	1 285
Dibaya	2656	4 332	5 367	4110	9 106	11 452	5980	11 182	8 704	154	235	122
Kalomba	2108	2 697	3 425	2921	3 248	4 029	2483	3 254	2 933	8	241	488
Kananga	1444	2 051	1 932	13496	12 075	11 591	9674	10 253	11 118	218	971	804
Katende	629	963	653	1416	2 308	2 087	1507	1 777	1 553	433	490	255
Katoka	357	399	376	3729	4 210	4 371	3554	4 343	4 204	0	0	0
Luambo	11978	14 095	12 742	8627	14 460	13 143	11843	12 877	14 026	133	138	197
Lubondaie	1884	2 100	2 407	3409	5 527	5 613	4926	5 284	5 140	85	152	193
Lubunga	1101	1 223	1 524	2297	4 405	3 798	3070	3 619	4 329	47	127	134
Luiza	1240	1 273	1 323	4924	7 592	7 849	5435	6 685	6 756	258	640	948
Lukonga	1026	1 314	1 136	6745	5 076	6 341	6530	6 196	4 885	10	19	6
Masuika	1664	1 956	2 183	9480	11 762	12 782	9980	12 203	11 276	118	141	245
Mikalayi	523	560	1 193	2419	3 126	4 004	2721	3 635	2 951	97	134	222
Muetshi	1332	1 625	1 653	2443	2 576	2 678	1992	2 585	2 399	167	116	77
Mutoto	2454	2 129	1 964	2904	3 270	3 097	3420	2 438	2 429	825	580	445
Ndekesha	1402	1 611	1 168	1315	3 670	4 270	2790	3 978	3 010	445	440	158
Ndesha	1557	1 349	1 696	5250	4 962	5 870	4896	5 486	4 645	299	366	271
Tshibala	2645	2 639	2 817	12003	12 064	11 282	12040	10 295	11 068	127	0	73
Tshikaji	1006	662	757	3500	3 874	3 599	4063	3 464	3 669	12	17	8
Tshikula	1364	1 613	1 815	2259	4 303	4 289	2646	4 098	3 941	252	141	111
Yangala	563	671	563	1329	1 768	2 172	1034	1 979	1 550	137	138	112
Total	44301	51 138	53 257	116662	144 827	151 803	121 568	139 360	133 002	5 782	7 741	7 542
Valeur en %	29,8	34,4	35,8	28,2	35,0	36,7	30,9	35,4	33,8	27,4	36,7	35,8

Source : Division Provinciale de la Santé Kasaï Central

Légende :

CC : Cas Contacts parmi les nouveaux cas d'IST

NC : Nouveaux Cas d'IST

T.a.S : Traités selon l'approche Syndromique

T.a.E : Traités selon l'approche Etiologique

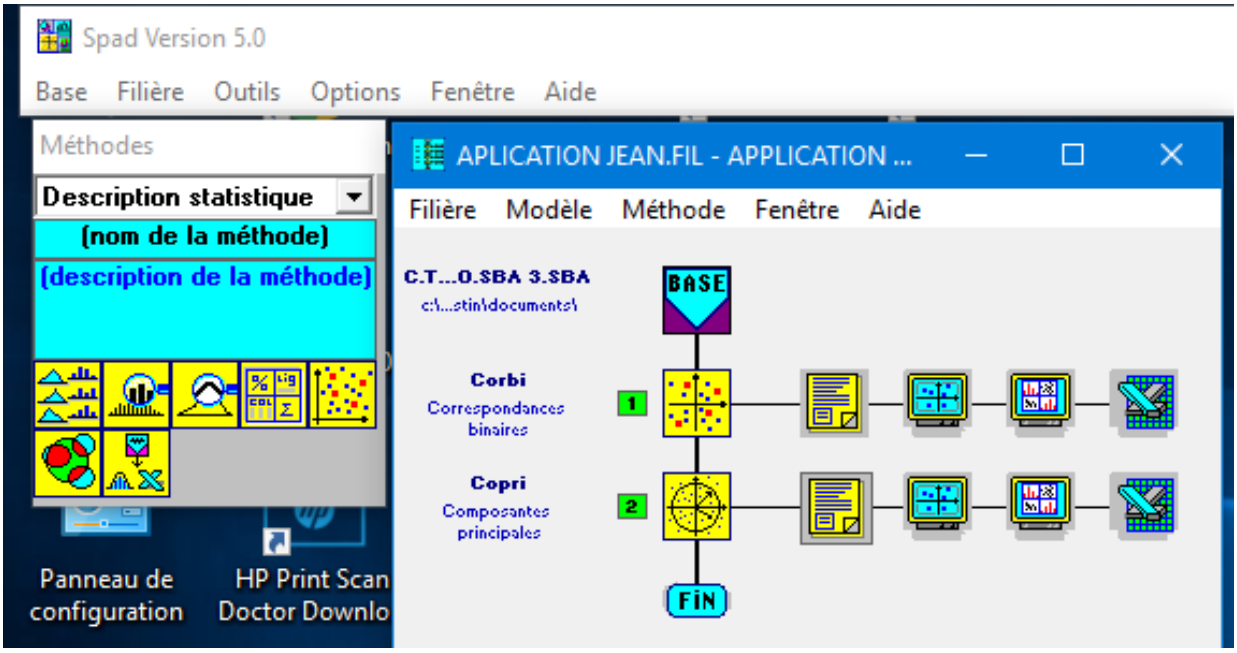
L'approche syndromique du traitement des IST est une méthode de diagnostic et de traitement basée sur les symptômes présentés par le patient. Par contre, l'approche étiologique du traitement des IST consiste à diagnostiquer et traiter spécifiquement la cause de l'infection, plutôt que de se baser uniquement sur les symptômes. C'est l'approche la plus précise et efficace à long terme, mais elle nécessite souvent des ressources de laboratoire et du temps pour obtenir les résultats. Elle est utilisée dans les environnements où les tests de laboratoire sont facilement accessibles.

D'après ce tableau, nous avons observé la croissance de cas contacts et de nouveaux cas d'IST au cours de trois années d'étude. En effet, pour l'ensemble de 26 Zones de Santé, les cas contacts représentaient respectivement 29,8% en 2022, 34,4% en 2023 et 35,8% en 2024. De même pour les nouveau cas d'IST, soit 28,2% en 2022, 35,0% en 2023 et 36,7% en 2024. Notons que les cas contacts sont de personnes qui ont été en contact étroit avec quelqu'un qui a été testé positif à une maladie sexuellement transmissible. Ainsi, l'idée est de surveiller de près les individus pour voir s'ils développent de symptômes et éventuellement les tester pour éviter la propagation de maladies. Quant aux approches de traitement, le traitement selon l'approche syndromique représente 30,9% de cas en 2022, 35,4 en 2023 et 33,8% en 2024. Le traitement selon l'approche étiologique représente 27,5% de cas en 2022, 36,8% en 2023 et 35,8% en 2024.

3.1. Traitement de données

L'analyse des données s'est faite par l'ordinateur sur la feuille Excel avec le logiciel SPAD (version 5.0) où toutes les productions statistiques (effectifs, poids, écart-type, moyenne, minimum et maximum, corrélation) ont été effectuées. En voici l'interface :

Figure1 : interface du logiciel SPAD 5.0



Source : Auteur

Cette figure représente l'interface du logiciel d'analyse SPAD 5.0 qui nous a servi à traiter les données.

Tableau N°2 : Statistiques sommaires des variables continues

Libellé de la variable	Effectif	Poids	Moyenne	Ecart-type	Minimum	Maximum
CC	24	24,00	1730,000	2246,010	50,000	11978,000
2022	26	26,00	281,885	433,846	1,000	1598,000
CC	26	26,00	415,423	276,486	1,000	956,000
2023	26	26,00	805,000	1426,710	1,000	4422,000
CC	26	26,00	409,808	519,614	1,000	2419,000
2024	26	26,00	3527,270	3691,220	3,000	13496,000
NC	26	26,00	266,269	894,852	2,000	3780,000
2022	26	26,00	360,962	284,937	2,000	962,000
NC	26	26,00	827,269	1495,330	2,000	5378,000
2023	26	26,00	390,077	556,171	1,000	2721,000

NC	26	26,00	3768,190	3639,940	3,000	12040,000

Source : Auteur

Ce tableau donne l'effectif de Zones de Santé considérées, leurs poids, leur moyenne, l'Ecart-type, le minimum et le maximum de valeurs initiales.

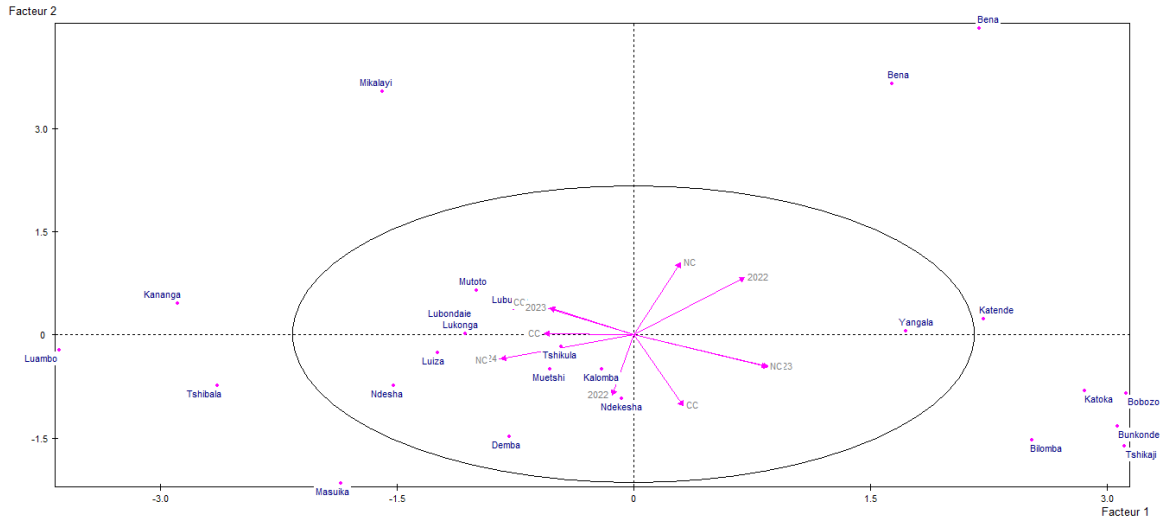
Tableau N°3 : Matrice des corrélations

	CC	**2022**	**CC**	**2023**	**CC**	**2024**	**NC**	**2022**	**NC**	**2023**	**NC**
CC	1,00										
2022	-0,20	1,00									
CC	-0,27	-0,15	1,00								
2023	-0,31	0,30	0,34	1,00							
CC	0,17	-0,28	-0,40	-0,42	1,00						
2024	0,41	-0,54	-0,05	-0,44	0,22	1,00					
NC	0,00	0,75	-0,44	-0,16	-0,20	-0,25	1,00				
2022	0,05	-0,42	0,40	-0,02	-0,19	0,06	-0,37	1,00			
NC	-0,31	0,27	0,31	0,95	-0,43	-0,43	-0,11	-0,01	1,00		
2023	-0,10	-0,21	-0,31	-0,37	0,81	0,10	-0,18	0,03	-0,38	1,00	
NC	0,58	-0,58	-0,06	-0,51	0,20	0,95	-0,26	0,13	-0,49	0,09	1,00

Source : Auteur

La matrice de corrélation joue un rôle crucial dans la compréhension de relations entre les différentes variables et dans la réduction de la dimensionnalité de données. Ainsi, si les variables sont fortement corrélées, cela signifie qu'elles portent des informations similaires. Ce qui peut suggérer qu'une réduction de dimensionnalité pourrait être efficace, car certaines variables redondantes peuvent être combinées sans perdre trop d'informations.

Figure 2 : Cercle de corrélation



Source : Auteur

Cette figure nous donne une vue d'ensemble de relations entre les variables originales et les composantes principales, elle nous permet de visualiser quelles variables sont les plus importantes dans la structure de données, et d'interpréter la position des individus dans l'espace de composantes principales. En effet, les variables les plus importantes dans la définition de deux premières composantes principales sont : nouveaux cas d'IST 2022, nouveaux cas d'IST 2023 et nouveaux cas d'IST 2024.

Tableau N°4 : Matrice des valeurs-tests

	CC	**2022**	**CC**	**2023**	**CC**	**2024,00**	**NC**	**2022,00**	**NC**	**2023,00**	**NC**
CC	99,99										
2022	-0,99	99,99									
CC	-1,33	-0,79	99,99								
2023	-1,59	1,56	1,83	99,99							
CC	0,86	-1,44	-2,14	-2,26	99,99						
2024	2,11	-3,08	-0,23	-2,41	1,17	99,99					
NC	0,00	4,98	-2,39	-0,81	-1,05	-1,30	99,99				
2022	0,27	-2,28	2,18	-0,12	-0,98	0,32	-1,96	99,99			
NC	-1,55	1,40	1,64	9,22	-2,34	-2,33	-0,56	-0,06	99,99		
2023	-0,49	-1,07	-1,62	-1,96	5,70	0,54	-0,91	0,16	-2,05	99,99	

NC	3,28	-3,39	-0,31	-2,88	1,03	9,27	-1,36	0,69	-2,74	0,45	99,99

Source : Auteur

Cette matrice fait référence aux nouvelles coordonnées des données projetées dans l'espace de composantes principales. Chaque colonne de cette matrice représente une composante principale, et chaque ligne correspond à une observation dans l'espace de nouvelles dimensions.

Tableau N°5 : Valeurs propres

Trace de la matrice:	**11**		
Numéro	**Valeur propre**	**Pourcentage**	**Pourcentage cumulé**
1	3,9811	36,19	36,19
2	2,4717	22,47	58,66
3	1,6798	15,27	73,93
4	1,0081	9,16	83,10
5	0,7195	6,54	89,64
6	0,4471	4,06	93,70
7	0,4228	3,84	97,55
8	0,1254	1,14	98,69
9	0,0922	0,84	99,53
10	0,0381	0,35	99,87
11	0,0142	0,13	100,00

Source : Auteur

Ce tableau permet de comprendre l'importance de chaque composante principale dans la représentation de données. Chaque valeur propre représente la variance expliquée par la composante principale. Plus la valeur propre est grande, plus la composante principale capture de variance dans les données. Ainsi les composantes principales avec les valeurs propres les plus élevées sont les plus importantes pour décrire la structure des données.

3.2. Discussions

Le tableau N°2 à la page 11: affiche des informations essentielles pour comprendre la distribution et les caractéristiques de données initiales. En effet, la moyenne trouvée représente

la tendance centrale de données pour chaque variable par année. L'écart-type trouvé dans ce tableau représente la dispersion au tour de la moyenne. Leurs valeurs indiquent que les données sont largement dispersées au tour de la moyenne ;

Le tableau N°3 à la page 11:présente les diagonales qui sont égales à 1 par le fait que chaque variable est parfaitement corrélée avec elle-même. Une corrélation proche de 1 ou de -1 indique que les deux variables sont fortement corrélées positivement ou négativement, la corrélation proche de zéro, montre qu'il n'y a pas de relation linéaire entre les variables. Dans ce tableau plus le chiffre est élevé, plus la corrélation entre les variables est forte ; nous constatons qu'il existe une corrélation positive entre les nouveaux cas et les cas contacts;

La *figure* 2 à la page 12 : illustre bien le cercle de corrélation qui est donc une aide à l'interprétation de la représentation des observations dans l'espace des facteurs. En effet, la lecture directe du cercle de corrélation montre les groupes des zones de santé ayant les caractéristiques similaires. Il s'agit de :

Groupe 1 : zones de santé de Katoka, Bobozo, Bunkonde, Tshikaji, et Bilomba.

Groupe 2 : zones de sante de Tshikula, Muetshi, Kalomba, Ndekesha.

Groupe 3 : zones de santé de Mutoto, Lubondayi, Lukonga et Lubunga.

En suite les zones de santé (individus) qui sont éloignés du centre du cercle constituent les cas extrêmes qui indiquent une forte hétérogénéité dans l'ensemble de données. C'est le cas de : zones de santé de Masuika, Tshibala, Kananga, Katende, Mikalayi, Bena tshiadi, Bena Leka et Luambo.

Le tableau N°4 à la page 12 : affiche les valeurs-test qui sont des scores indiquant où chaque observation se situe par rapport à chaque composante principale;

Le tableau N°5 à la page 13: affiche les valeurs propres d'une matrice de dimension 11 dont la trace est égale à 11, la première valeur propre (3,9811) représente 36,19% de la variance totale. Ceci indique qu'un seul axe factoriel explique déjà une grande partie de l'information. Les trois premières valeurs propres cumulent 73,93% de la variance totale. Cela suggère qu'une réduction de dimension en gardant uniquement ces trois axes pourrait conserver l'essentiel de l'information.

Eu égard à tout ce qui précède, nous recommandons ce qui suit :

Que le PNLS intensifie les activités de lutte contre le VIH/SIDA et les maladies sexuellement transmissibles dans les zones de santé à forte concentration des activités minières et intenses

échanges commerciaux comme Demba, Luiza, Luambo, Masuika, Tshibala, Lubunga, Ndekesha, Ndesha.

Qu'il soit mené d'intenses activités de sensibilisation et d'information de la population sur les modes de transmission des maladies sexuellement transmissibles, les moyens de prévention contre ces maladies dans les zones les plus touchées ci-haut citées. Ainsi nous recommandons que le PNLS organise les campagnes de prévention dans les milieux les plus touchés en impliquant les jeunes et les adultes à promouvoir des comportements de prévention efficace contre les MST comme préconisé dans l'étude de Silué et al.(2024);

Qu'il soit maintenu un système de contrôle permanant dans la population pour réduire les cas contacts, qui sont d'ailleurs en corrélation positive avec les nouveaux cas.

Conclusion

Cette recherche a été réalisée dans le but d'examiner le phénomène des maladies sexuellement transmissibles dans les 26 Zones de santé de la Division Provinciale du Kasaï-Central. Elle a été axée sur la technique documentaire, et l'analyse en composantes principales.

Après l'analyse, nous avons abouti aux constats suivants : les zones de santé représentant des caractéristiques similaires en matière de comportements sexuels des individus se regroupent en :

Groupe 1 : zones de santé de Katoka, Bobozo, Bunkonde, Tshikaji, et Bilomba.

Groupe 2 : zones de santé de Tshikula, Muetshi, Kalomba, Ndekesha.

Groupe 3 : zones de santé de Mutoto, Lubondaie, Lukonga et Lubunga.

En outre, les zones de santé à forte concentration d'activités minières et d'échanges commerciaux dans lesquelles les gens affichent des comportements de santé à risque sont les plus touchées. Il s'agit de : Masuika, Tshibala, Kananga, Katende, Mikalayi, Bena tshiadi, Bena Leka et Luambo. Il y a nécessité d'améliorer les actions de sensibilisation et de prévention dans ces zones, à travers les organisations à assise communautaire et autres canaux de communication pour une lutte efficace contre les MST. La théorie développée dans cet article implique bien à appliquer l'adoption des comportements sexuels favorables (le nombre de partenaires sexuels, l'usage de préservatifs, ou la fréquence de test de dépistage) des jeunes et des adultes pour la bonne gestion des cas contacts, ce qui entraînera une lutte efficace dans la réduction des risques de nouveaux cas. Dès lors, les perspectives futures devraient se concentrer sur la connaissance des principaux facteurs associés à une incidence élevée de MST dans les

différentes zones de santé afin d'améliorer les campagnes de sensibilisation à d'autres problèmes de santé publique au Kasaï Central.

BIBLIOGRAPHIE

- Afonso F., Diday E. et Toque C. (2018). Data Science par Analyse des Données Symboliques, Ed. TECHNIP, Paris.
- Alain C. (2022). La Data Science pour modéliser les systèmes complexes : optimiser la prédiction, l'estimation et l'interprétation, Ed. Dunod, Paris.
- Bouallegue M. (2013). L'Analyse Factorielle pour la Modélisation Acoustique des Systèmes de reconnaissance de la parole, Thèse de Doctorat, Université d' Avignon et des Pays de Vaucluse.
- Choulakian V. (1998). Analyse factorielle des correspondances de tableaux multiples, Revue de statistique appliquée, tome36, n°4.
- Christophe L. & Mounir M. (2016). Biostatistique et analyse informatique de données de santé avec R. Vol 1, Ed. ISTE, Paris.
- Coulibaly H. (2018). Prévalence des IST et VIH/SIDA à la clinique de santé sexuelle de Halles de Bamako, mémoire de Master2 Santé Publique, Université des Sciences, des Techniques et des Technologies de Bamako.
- Crucianu M. (2021). Apprentissage Statistique : Modélisation Descriptive et Introduction aux Réseaux de Neurones (RCP 208). Méthode d'analyse factorielle, Conservatoire Nation des Arts et Métiers, Paris.
- David W-H, Stanley Lemeshow Jr. et Rodney X-S. (2013). Applied Logistic Regression, 3ème Ed. Wiley, Hoboken.
- Escofier B. & Pages J. (2023). Analyses Factorielles simples et multiples, 5ème Edition, Dunod, Paris.
- Goma A. (2018). Prévention des infections sexuellement transmissibles à Chlamydia Trachomatis: Analyse des facteurs associés au non usage du préservatif chez les étudiants âgés de 18-24 ans, participant à la cohorte I-Share, Mémoire de mater2, Université de Versailles-Saint-Quentin-En-Yvelines.
- Janier M. (2009). Les maladies sexuellement transmissibles, Elsevier Masson, Paris.
- Kasiama J. (2020). Contribution à l'analyse factorielle discriminante des données symbolique ; Application aux données médicales liées aux maladies cardiovasculaires à l'INRB. Thèse de Doctorat, UPN, Kinshasa.
- Kostov, B., Alvarez-Esteban, R., Bécue-Bertaut, M., & Husson, F. (2023). « Multilingual textual data: an approach through multiple factor analysis. » in Statistica Applicata - Italian Journal of Applied Statistics, 35(3), pp : 339-357

- Laurent J. (2024). L'analyse de séquences, Ed. Dunod, Paris.
- Myriam M-B, Saporta G. et Christine T-A. (2018). Apprentissage statistique et données massives, Ed. TECHNIP, Paris.
- Nouchi A. (2018). Maladies sexuellement transmissibles acquises en voyage, étude rétrospective mono centrique de 140 cas, Thèse de doctorat, Université Paris Descartes, Paris.
- Santi P. (2018). « Très forte hausse des infections sexuellement transmissibles en France », Le Monde, n.r. pp : 1-6
- Saporta G. (2021). Notions sur les méthodes factorielles, Conservatoire National des Arts et Métiers, Paris.
- Spearman C. (1904). « General Intelligence », objectively Determined and Measured. American Journal of Psychology, 15, 201-292.
- SILUE N., KPANGBA B., & BONGBA E. (2024). «Usage des TIC dans la prévention des IST et du VIH chez les adolescents et les jeunes en Côte d'Ivoire : à l'assaut du sida », in Revue Francophone vol : 2 Numéro : 3, pp : 138-155.
- Stafford J., & Bodson P. (2006). L'analyse multivariée avec SPSS, Presses de l'Université du Québec, Québec.