

Analyse factorielle des correspondances des causes de décès par pays

MKRTCHYAN Yester, TOURE Fatoumata Salihou

Table of contents

1	Introduction	1
2	Analyse factorielle des correspondances des causes de décès par pays	2
2.1	Préparation des données	2
2.2	Analyse exploratoire	3
2.2.1	Chargement des données	3
2.2.2	Visualisation des données par barplots	3
2.2.3	Comparaison des pays	5
2.2.4	Profils ligne et colonne	5
2.2.5	Comparaison des pays, graphiques de profils ligne et colonnes avec <code>ggplot2</code>	6
2.3	Analyses avancées (AFC et Tests)	8
2.3.1	Test du khi-deux et V de Cramer	8
2.4	Lancement de l'AFC	9
2.4.1	Critère du bâton brisé pour sélectionner les axes	13
2.4.2	Analyse des contributions des lignes et colonnes	14
2.4.3	Distances au centre de gravité	26
3	Conclusion	28
4	Annexe	29
4.1	Source du jeu de données initiale	29
4.2	Tableau de l'encodage des différentes causes de décès	29

1 Introduction

Les données sur la santé ont toujours constitué un élément central de l'analyse statistique. Le souhait de mesurer et d'évaluer la santé nationale a été le moteur de la mise en place

des premiers systèmes d'analyse statistique. La répartition des causes de décès constituent notamment un indicateur central de l'état de santé des populations. Cette répartition présente toutefois de fortes disparités selon les pays et les régions du monde, reflétant la diversité des contextes sanitaires, démographiques et sociaux. À partir des données mondiales de 2019 relatives aux causes de décès, ce projet propose une analyse comparative des pays afin d'étudier la structure de leurs profils de mortalité. Il vise, d'une part, à examiner l'existence éventuelle de relations entre les causes de décès et certaines caractéristiques globales des pays, telles que leur niveau de développement, et, d'autre part, à identifier des groupes de pays présentant des profils de mortalité similaires. Cette approche permet ainsi de mettre en évidence des proximités ou des oppositions entre pays au regard de la répartition des causes de décès. L'analyse repose sur la mise en œuvre d'une Analyse Factorielle des Correspondances appliquée aux données considérées.

2 Analyse factorielle des correspondances des causes de décès par pays

2.1 Préparation des données

Pour cette analyse, nous avons utilisé un jeu de données disponible sur Kaggle, portant sur les causes de mortalité dans le monde dont le lien se trouve en annexe. Le jeu de données original couvre la période 1990 à 2019 et recense les effectifs de décès pour 204 pays et territoires, organisés sous la forme d'un tableau pays \times causes de décès.

Afin de concentrer l'étude sur l'année la plus récente, nous avons extrait uniquement les observations de 2019. Une vérification préalable a confirmé l'absence de valeurs manquantes dans les variables retenues. Les données extraites ont ensuite été enregistrées dans un nouveau fichier CSV, distinct du fichier initial.

La variable correspondant à l'année a été supprimée du fait que toutes les lignes concernent 2019. De même, parmi les deux colonnes descriptives des pays (nom complet et abréviation), seule la colonne des abréviations a été conservée, les correspondances avec les noms complets étant stockées dans un fichier séparé.

Pour améliorer la lisibilité et faciliter l'interprétation graphique dans l'Analyse Factorielle des Correspondances (AFC), nous avons procédé à un encodage des différentes causes de décès, chaque cause étant associée à un identifiant plus court. Cet encodage a été enregistré dans un document séparé, qui sera également disponible en annexe.

Le jeu de données final comporte ainsi 204 lignes correspondant aux abréviations des pays et territoires, et des colonnes représentant les différentes causes de décès encodées, avec pour valeurs les effectifs observés. En étant deux variables qualitatives il convient de procéder à une AFC.

2.2 Analyse exploratoire

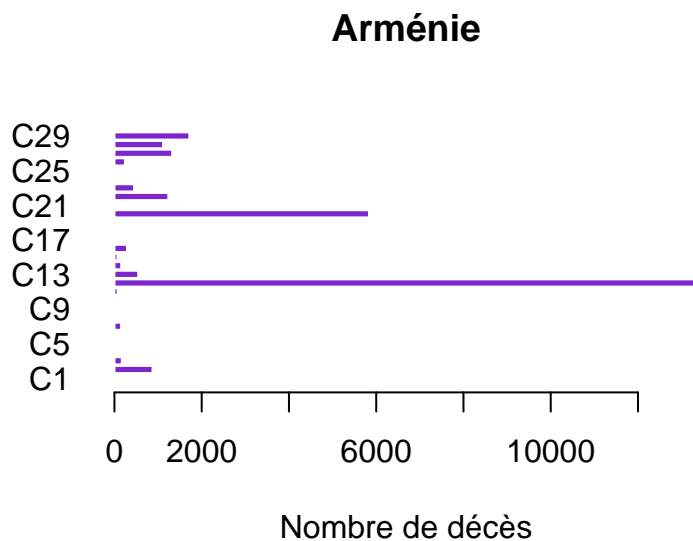
2.2.1 Chargement des données

La première étape de l'analyse consiste à charger les librairies nécessaires ainsi que le jeu de données à analyser dans le script R qui est le fichier CSV nettoyé. Après l'importation, une première visualisation globale du tableau est réalisée afin de vérifier la structure des données et la cohérence des variables.

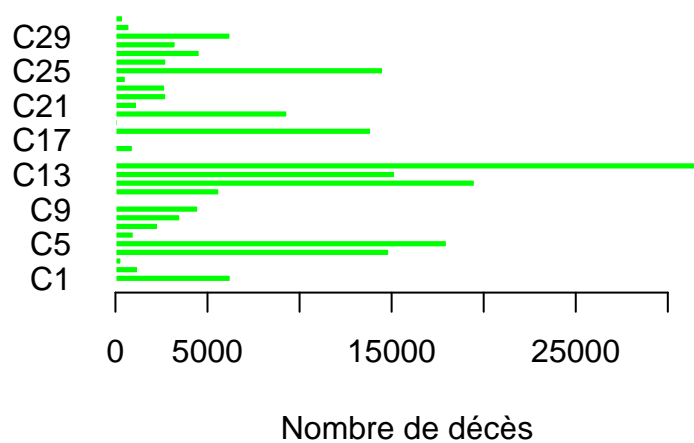
Au moment du chargement de données, la première colonne du jeu de données est désigné comme étant la colonne d'identification des lignes et elle est ensuite supprimée, afin de ne conserver que les causes de décès sous forme numérique. Cette étape permet de faciliter les traitements ultérieurs et l'accès direct aux informations par pays.

Le jeu de données obtenu est stocké dans la variable "morts" pour le reste de l'analyse. Une vérification du jeu de données à l'aide de l'affichage des premières lignes confirme la bonne manipulation des données et l'absence d'anomalies apparentes.

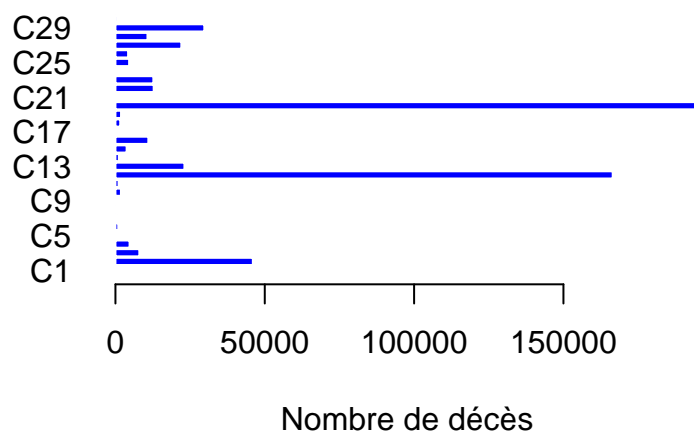
2.2.2 Visualisation des données par barplots



Mali



France



2.2.3 Comparaison des pays

Les Figures 1 à 3 présentent des diagrammes en barres horizontales illustrant le nombre de décès par cause respectivement en Arménie, au Mali et en France. Ces représentations graphiques ont pour objectif de comparer la structure de la mortalité selon les principales causes de décès dans chacun des pays étudiés.

Les causes de décès sont représentées sur l'axe des ordonnées, tandis que l'axe des abscisses indique le nombre de morts. Ce choix de visualisation permet une lecture claire et facilite la comparaison entre catégories.

L'analyse met en évidence des profils de mortalité contrastés. En Arménie, en 2019, les maladies cardiovasculaires (C12) sont responsables de plus de 12 000 décès, tandis que le néoplasme (C20) en représente environ 6 000. En France, ces mêmes causes figurent également parmi les principales causes de mortalité, avec plus de 15 000 décès pour le néoplasme et environ 20 000 pour la cause cardiovasculaires. À l'inverse, au Mali, les maladies transmissibles ainsi que certaines maladies infectieuses occupent une place plus importante dans la mortalité totale. Ils ont entraîné environ 32 000 décès, les maladies cardiovasculaires (C12) près de 20 000 décès, et l'automutilation et la malnutrition (C18 et C25) environ 14 000 décès chacune.

Cette analyse descriptive constitue une première étape exploratoire. Elle permet de dégager des tendances générales et justifie la poursuite de l'étude par des analyses comparatives plus approfondies entre pays et groupes de causes de décès.

2.2.4 Profils ligne et colonne

Afin d'approfondir l'analyse descriptive, des profils lignes et des profils colonnes ont été construits à partir du tableau de contingence des décès par cause et par pays. Le profil ligne d'un pays correspond à la répartition relative des causes de décès à l'intérieur de ce pays : chaque valeur représente la proportion d'une cause donnée parmi l'ensemble des décès du pays considéré. Par exemple, en Afghanistan la meningite représente 0.7% des causes des décès en 2019 (

$$0,007 = \frac{1563}{1563 + 1775 + \dots + 485 + 1940}$$

). Ainsi, la somme des proportions sur une ligne est égale à 1, ce qui permet de comparer les structures de mortalité indépendamment du niveau absolu de mortalité.

Table 1: Profils lignes pour quelques pays et causes sélectionnés

	C1	C2	C3	C4	C5	C6	C7	C8
AFG	0.0071816	0.0081556	0.0025730	0.0057158	0.0024352	0.0077513	0.0230425	0.0185535
ARM	0.0001788	0.0312187	0.0060435	0.0001430	0.0000000	0.0015735	0.0054713	0.0002861
MLI	0.0332092	0.0065039	0.0017241	0.0788586	0.0954897	0.0052678	0.0122810	0.0186735

	C1	C2	C3	C4	C5	C6	C7	C8
FRA	0.0003979	0.0810118	0.0140813	0.0083357	0.0000000	0.0016622	0.0009473	0.0000810
MOZ	0.0109682	0.0061453	0.0013705	0.0141875	0.0825047	0.0028777	0.0072786	0.0083276
AUS	0.0002859	0.0698592	0.0149352	0.0010193	0.0000000	0.0012182	0.0020013	0.0000808
ARG	0.0013440	0.0357363	0.0090283	0.0039460	0.0000000	0.0016793	0.0085476	0.0011185
USA	0.0004042	0.0507657	0.0113620	0.0021482	0.0000000	0.0012751	0.0062466	0.0003492

Table 2: Profils colonnes pour quelques pays et causes sélectionnés

	C1	C2	C3	C4	C5	C6	C7	C8
AFG	0.0066205	0.0010940	0.0015440	0.0049481	0.0008240	0.0071161	0.0121089	0.0205699
ARM	0.0000212	0.0005381	0.0004659	0.0000159	0.0000000	0.0001856	0.0003694	0.0000408
MLI	0.0265160	0.0007557	0.0008961	0.0591263	0.0279850	0.0041887	0.0055897	0.0179312
FRA	0.0009573	0.0283575	0.0220484	0.0188297	0.0000000	0.0039820	0.0012990	0.0002343
MOZ	0.0115594	0.0009424	0.0009402	0.0140408	0.0319154	0.0030202	0.0043727	0.0105549
AUS	0.0001948	0.0069279	0.0066253	0.0006523	0.0000000	0.0008268	0.0007775	0.0000662
ARG	0.0019188	0.0074241	0.0083898	0.0052901	0.0000000	0.0023875	0.0069563	0.0019205
USA	0.0048542	0.0887060	0.0888084	0.0242233	0.0000000	0.0152487	0.0427591	0.0050431

De manière complémentaire, les profils colonnes décrivent, pour chaque cause de décès, la répartition relative des pays. Chaque valeur indique la contribution d'un pays donné au total des décès observés pour une cause spécifique . En 2019, l'Afghanistan représentait 0,66% (

$$0,0066 = \frac{1563}{1563 + 13 + \dots + 2065 + 1450}$$

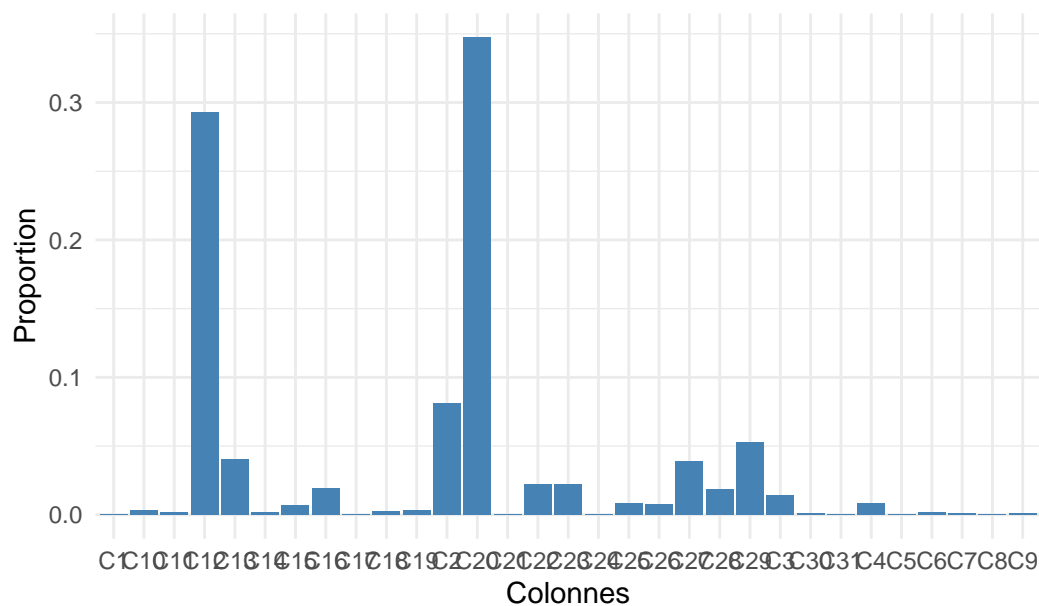
)de la mortalité mondiale due à la méningite. Là encore, la normalisation par colonne permet de comparer les pays entre eux pour une cause donnée, sans être influencé par les différences globales de population ou de mortalité.

2.2.5 Comparaison des pays, graphiques de profils ligne et colonnes avec ggplot2

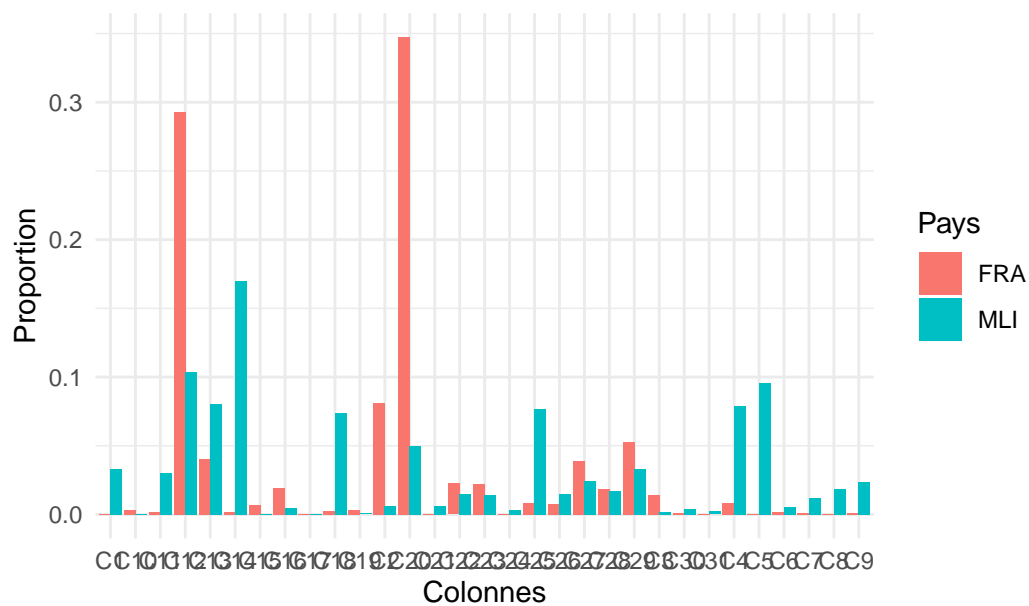
Les profils lignes sont ensuite représentés graphiquement afin de faciliter l'interprétation. Le profil ligne de la France met en évidence la prédominance des maladies non infectueuses dans la structure de la mortalité, confirmant les observations issues de l'analyse par diagrammes en barres. Une comparaison des profils lignes de la France et du Mali souligne des contrastes marqués : la France présente une concentration plus forte des décès sur les maladies chroniques, tandis que le Mali affiche une part relativement plus importante de maladies infectueuses et de causes évitables. Ces différences traduisent des stades distincts de transition épidémiologique et

fournissent une base pertinente pour des analyses multivariées ultérieures, telles que l'analyse des correspondances.

Profil ligne de la France (FRA)



Comparaison des profils ligne : France vs Mali



2.3 Analyses avancées (AFC et Tests)

2.3.1 Test du khi-deux et V de Cramer

Afin d'évaluer statistiquement l'existence d'un lien entre les pays et les causes de décès, un test du khi-deux d'indépendance est réalisé à partir du tableau de contingence. L'hypothèse nulle H_0 suppose l'indépendance entre le pays et la cause de décès, tandis que l'hypothèse alternative H_1 postule l'existence d'une dépendance entre ces deux variables.

Le test met en évidence une statistique du khi-deux très élevée (36 396 969) associée à une p-value nulle, conduisant au rejet de l'hypothèse d'indépendance au seuil usuel. Cela suggère que la répartition des causes de décès diffère significativement selon les pays considérés. Toutefois, l'interprétation de ce résultat doit être nuancée, car la structure du tableau, notamment la présence de nombreux effectifs élevés et de fortes disparités entre pays limite la pertinence stricte du test du khi-deux.

Afin de quantifier l'intensité de la liaison entre les deux variables, le coefficient de V de Cramér est calculé. La valeur obtenue étant 0.149 indique une liaison faible entre le pays et la cause de décès, malgré le rejet de l'hypothèse d'indépendance. Ce résultat souligne que la dépendance statistique observée est réelle mais modérée.

Sous l'hypothèse d'indépendance, les profils attendus pour différents pays sont égaux. Nous le confirmons avec les exemples de la France et du Mali.

	C1	C2	C3
AFG	945.158167	6495.35413	1452.071117
ALB	96.560813	663.58912	148.348892
DZA	835.052350	5738.68049	1282.912682
ASM	1.563386	10.74397	2.401871
AND	2.549188	17.51863	3.916384

Table 3: Résidus attendus (extrait)

	C1	C2	C3	C4	C5	C6	C7	C8
AFG	945.158167	6495.35413	1452.071117	1006.519544	2575.043961	1949.101598	1658.070310	785.907640
ALB	96.560813	663.58912	148.348892	102.829715	263.075902	96.963688	169.394523	80.291197
DZA	835.052350	5738.68049	1282.912682	1889.265458	2275.065259	1838.536392	1464.914083	394.353648
ASM	1.563386	10.74397	2.401871	1.664884	4.259380	1.569909	2.742614	1.299970
AND	2.549188	17.51863	3.916384	2.714686	6.945157	2.559824	4.471985	2.119673
AGO	752.770465	5173.21959	1156.500877	801.641684	2050.891699	755.911207	1320.568770	625.935510
ATG	2.510103	17.25003	3.856337	2.673064	6.838672	2.520576	4.403420	2.087174
ARG	1463.729034	10059.097782	248.765050	1558.756968	3987.868633	1469.836068	2567.787844	1217.104047

L'examen des effectifs attendus sous l'hypothèse d'indépendance ainsi que des résidus standardisés permet d'identifier les principales contributions au khi-deux. Ces écarts traduisent des phénomènes d'attraction et de répulsion : un résidu standardisé positif indique qu'un pays et une cause se rencontrent plus fréquemment que prévu (attraction), tandis qu'un résidu standardisé négatif montre une occurrence moindre que prévu (répulsion). Un résidu supérieur à 2 ou inférieur à -2 est généralement considéré comme significatif.

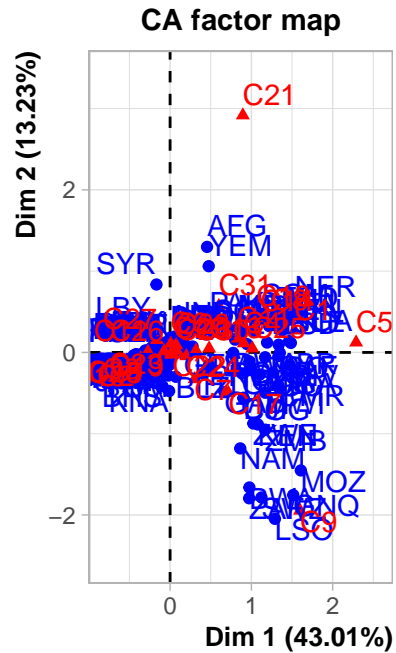
Table 4: Résidus standardisés (extrait)

	C1	C2	C3	C4	C5	C6	C7	C8
AFG 20.1809321	-	-	7.5178633	-	24.0525057	82.9225228	116.4484601	
		59.5831619	23.5358700		40.622433			
ALB	-	9.9895002	8.2107427	-	-	-	-	-
	8.5238559			9.6717903	16.319765	6.2058964	8.6705160	8.6431180
DZA	-	-	0.0024504	-	-	-	-	-
	18.8668478	7.1114257		26.1042200	48.067428	10.8357278	26.4292529	2.1462813
ASM	-	-	0.3872354	0.2603221	-	1.9437291	0.1560142	-
	0.4515643	1.1596590			2.076152			0.2635712
AND	-	4.4829677	1.5634126	-	-	-	-	-
	1.6001045			1.6514625	2.651111	1.6034536	2.1228199	1.4585534
AGO 64.6547205	-	-	78.4804784	194.3012081	3540972	-	57.8761968	
		56.9798157	26.2857592			9.5887733		
ATG	-	-	0.5843410	0.2004320	-	0.9338863	0.2853888	-
	0.9552307	0.3055684			2.630708			1.4473287
ARG	-	20.1654607	16.8570314	-	-	-	6.2240087	-
	26.5582819			5.8256207	63.724356	23.7003494		24.1994379

À titre d'illustration, le Bangladesh (BGD) présente un résidu standardisé de $-99,09$ pour la malaria (C5), ce qui traduit une répulsion extrêmement marquée : cette cause est très fortement sous-représentée dans ce pays par rapport à ce qui serait attendu sous l'hypothèse d'indépendance. À l'inverse, l'Afghanistan (AFG) affiche un résidu standardisé très élevé (116,45) pour les troubles maternels (C8), indiquant une attraction très forte entre ce pays et cette cause de décès, largement surreprésentée par rapport au modèle d'indépendance.

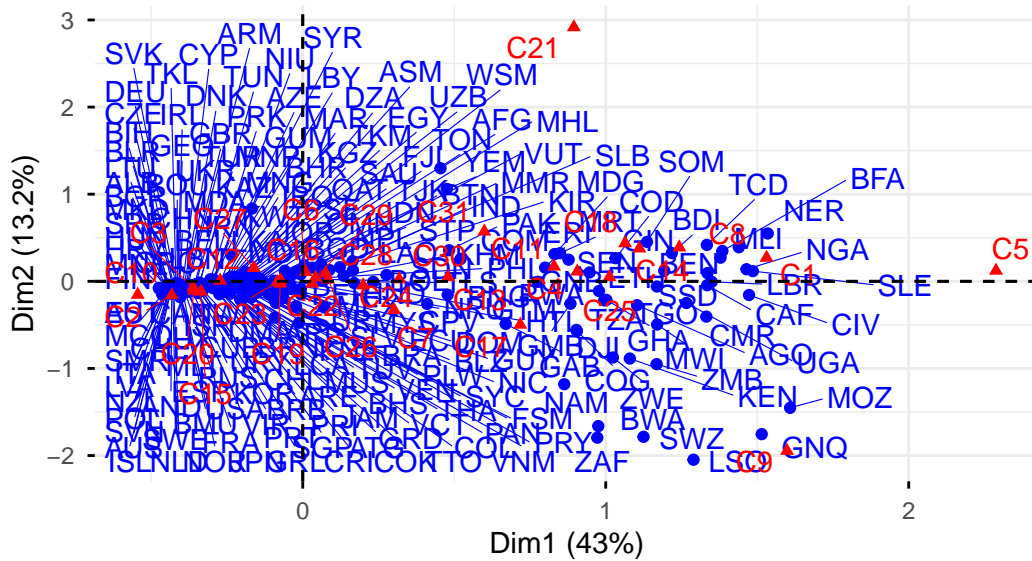
2.4 Lancement de l'AFC

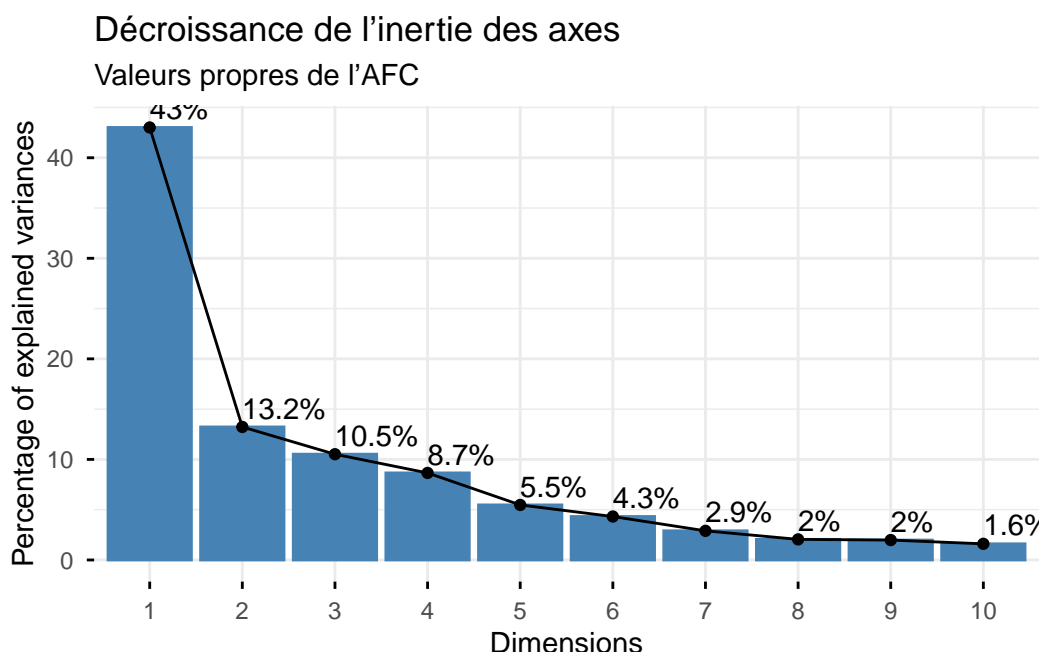
Pour explorer les relations entre les pays et les causes de décès, une analyse factorielle des correspondances (AFC) est réalisée. Cette méthode multivariée est particulièrement adaptée aux tableaux de contingence et permet de visualiser les dépendances entre lignes (pays) et colonnes (causes) tout en réduisant la dimensionnalité des données.



Analyse factorielle des correspondances (AFC)

Représentation conjointe des pays et des causes de décès (axes 1 et 2)

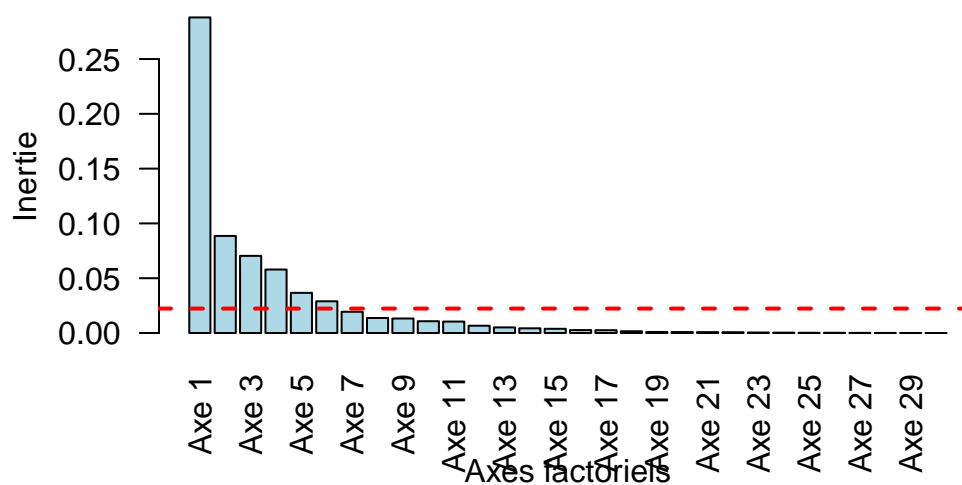




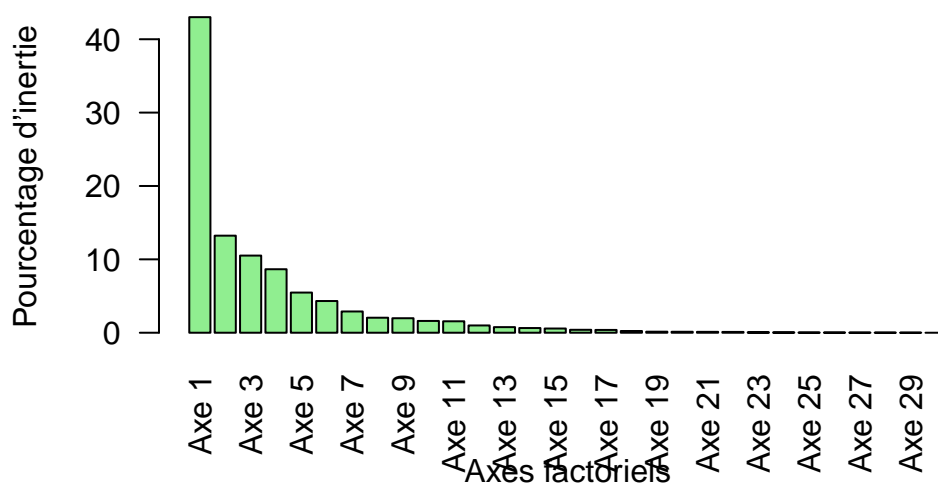
L'AFC est appliquée aux 204 pays du tableau, considérés comme lignes actives. Les valeurs propres associées aux axes factoriels sont examinées afin de déterminer la part d'inertie expliquée par chaque axe et d'identifier le nombre d'axes significatifs à retenir. La décroissance des valeurs propres est représentée graphiquement pour faciliter l'interprétation. La somme des valeurs propres correspond à l'inertie totale du tableau, qui peut également être calculée à partir de la statistique du khi-deux normalisée par l'effectif total (Φ^2). Dans les deux cas la valeur trouvée est 0.6695183.

Le critère de Kaiser est utilisé pour identifier les axes factoriels dont l'inertie dépasse la moyenne qui est 0.02231728, indiquant qu'ils expliquent une proportion d'information supérieure à celle attendue par hasard. Dans cette analyse, six axes présentent une inertie supérieure à la moyenne et sont donc retenus pour l'interprétation.

Décroissance de l'inertie des axes



Décroissance de l'inertie (en %)



Les diagrammes de l'inertie totale et du pourcentage d'inertie permettent de visualiser la contribution relative de chaque axe à la structure globale des données, préparant ainsi l'interprétation des positions des pays et des causes sur les plans factoriels. Cette étape constitue un prérequis essentiel avant de produire les représentations graphiques des individus et des variables dans

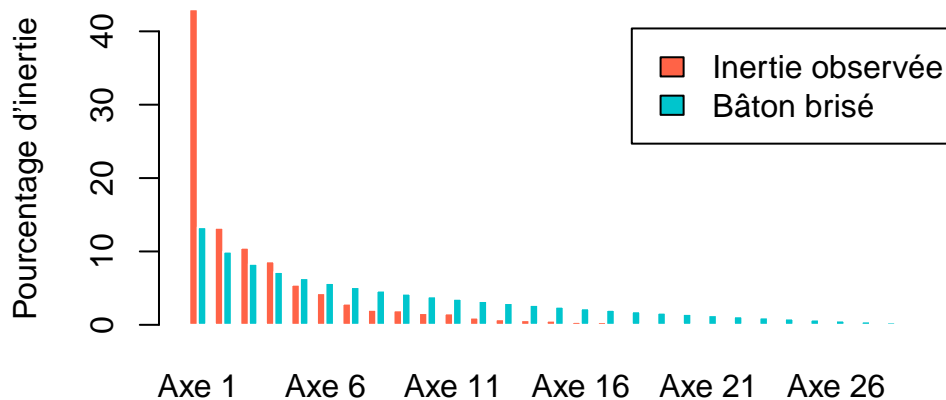
l'espace factoriel.

2.4.1 Critère du bâton brisé pour sélectionner les axes

Pour compléter l'identification des axes significatifs, le critère du bâton brisé est utilisé. Les valeurs propres observées sont ainsi mises en relation avec les proportions théoriques du bâton brisé, permettant de distinguer les axes apportant une information réelle de ceux qui ne contiennent qu'un bruit statistique.

dim 1	dim 2	dim 3	dim 4	dim 5	dim 6	dim 7	dim 8	dim 9	dim 10	dim 11
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dim 12	dim 13	dim 14	dim 15	dim 16	dim 17	dim 18	dim 19	dim 20	dim 21	dim 22
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
dim 23	dim 24	dim 25	dim 26	dim 27	dim 28	dim 29	dim 30			
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE			

Sélection des axes par le critère du bâton brisé



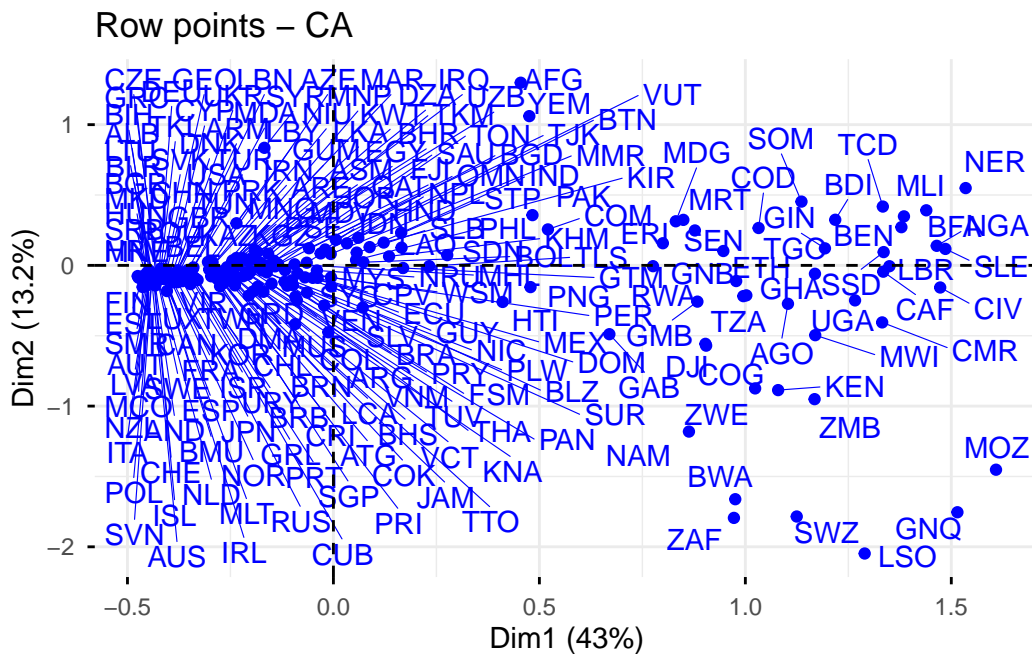
La comparaison numérique indique que les axes dont l'inertie observée dépasse celle du bâton brisé sont considérés comme pertinents. Dans cette analyse, quatre axes présentent une inertie supérieure à la valeur attendue par le bâton brisé et sont donc retenus pour l'interprétation factorielle.

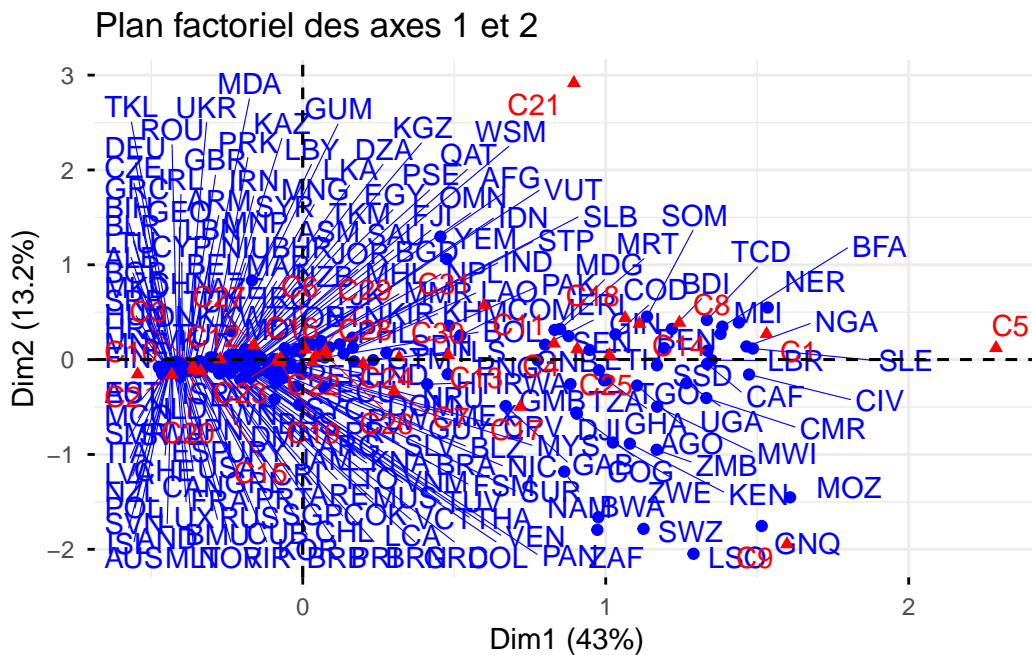
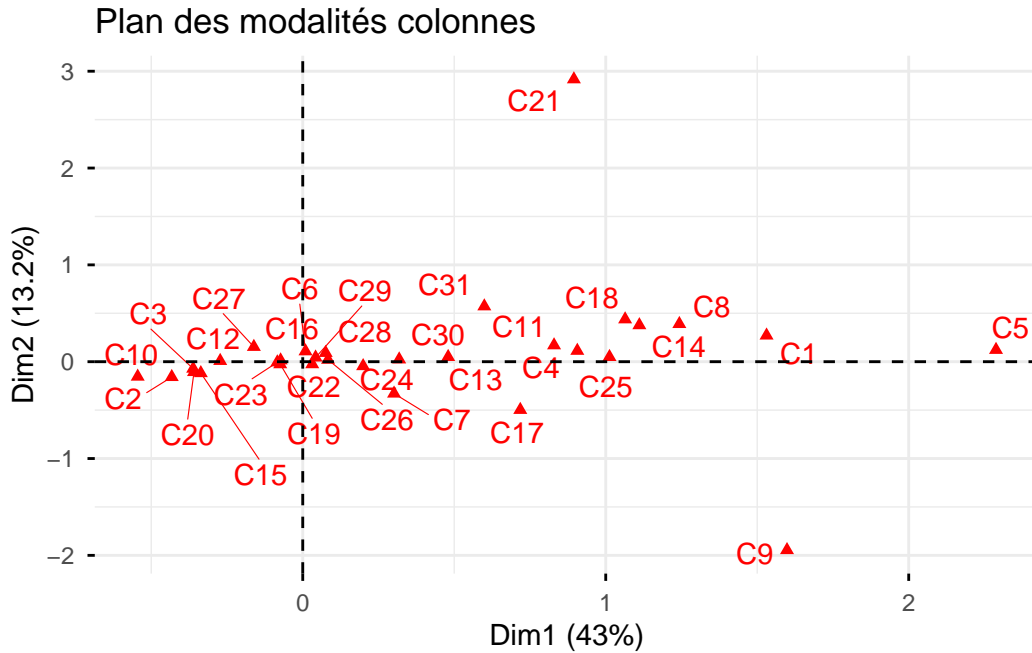
Un diagramme comparatif illustre visuellement la sélection des axes : les barres rouges représentant l'inertie observée sont mises côte à côte avec celles du bâton brisé, représentées

en bleu. Cette représentation permet de visualiser aisément quels axes dépassent le seuil et confirment leur pertinence pour l'étude des relations entre les pays et les causes de décès. Ce critère identifiant moins d'axes pertinents que le critère de Kaiser, son résultat sera privilégié pour la suite de l'analyse.

2.4.2 Analyse des contributions des lignes et colonnes

Après sélection des axes significatifs, l'AFC permet de représenter les modalités des lignes (pays) et des colonnes (causes) dans l'espace factoriel. Les diagrammes factoriels des lignes et des colonnes mettent en évidence les similitudes et les différences structurelles : les pays proches sur le plan factoriel présentent des profils de mortalité similaires, comme Djibouti et le Congo, tandis que les causes de décès rapprochées indiquent une co-occurrence relative dans différents pays. Par exemple, les troubles néonataux (C14), troubles maternels (C8) et maladies diarrhéiques (C18) apparaissent proches sur le graphique, suggérant qu'elles sont présentes dans des proportions comparables dans les mêmes pays. À l'inverse, la malaria (C5), très éloignée de l'origine et des autres modalités, présente un profil fortement spécifique, caractéristique d'un nombre limité de pays.

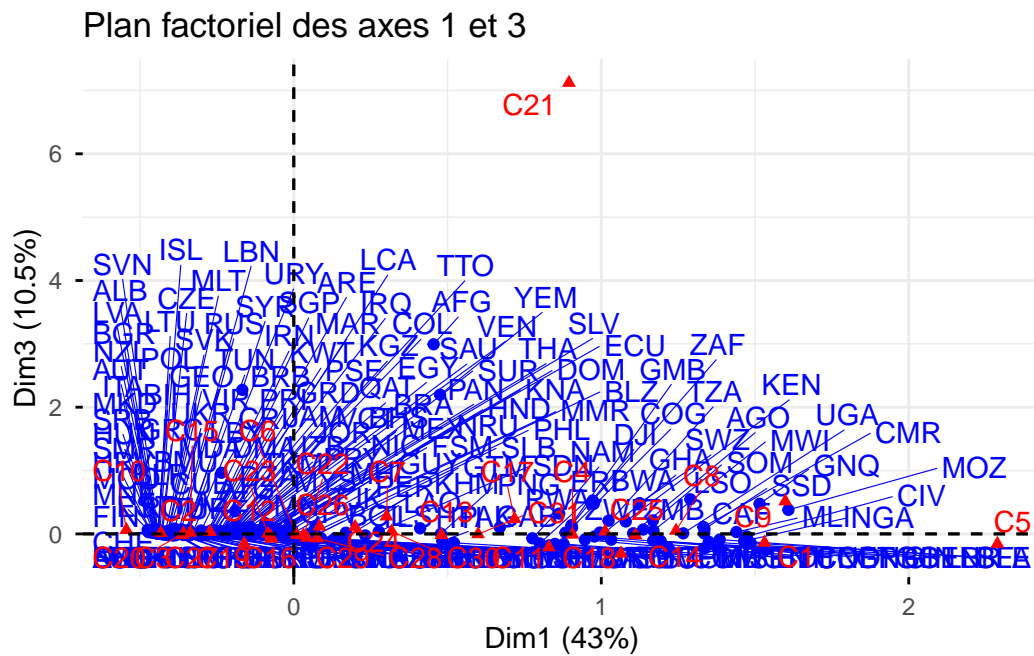




Les biplots combinant simultanément les modalités des lignes (pays) et des colonnes (causes) permettent de visualiser les relations entre pays et causes de décès sur différents plans factoriels. Le premier plan factoriel, défini par les axes 1 et 2, est le plus informatif, concentrant une part importante de l'inertie totale avec 43 % pour l'axe 1 et 13,2 % pour l'axe 2. Il permet

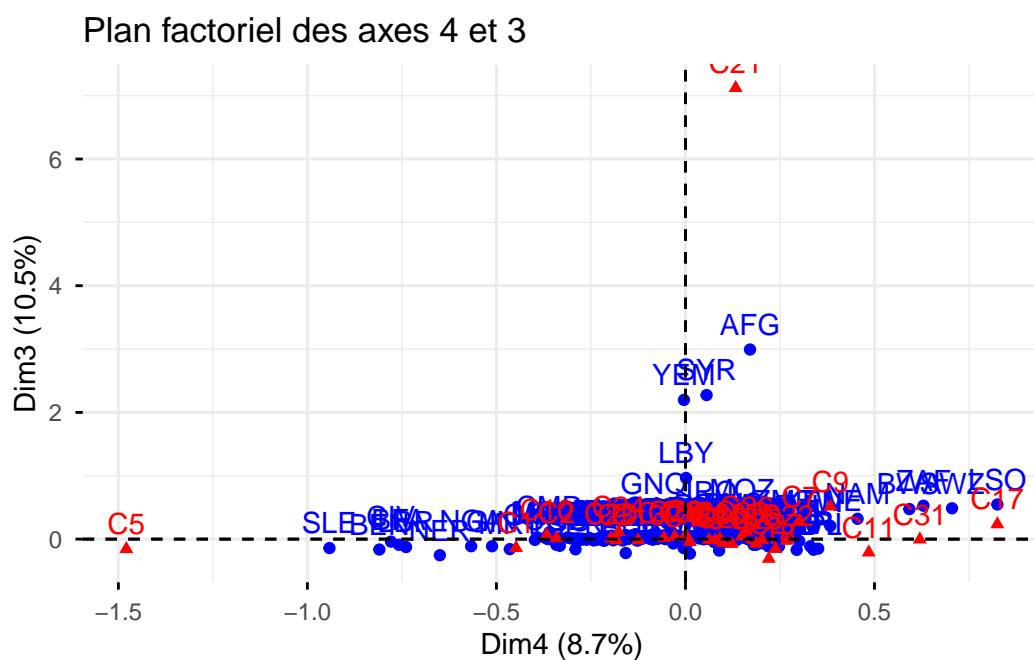
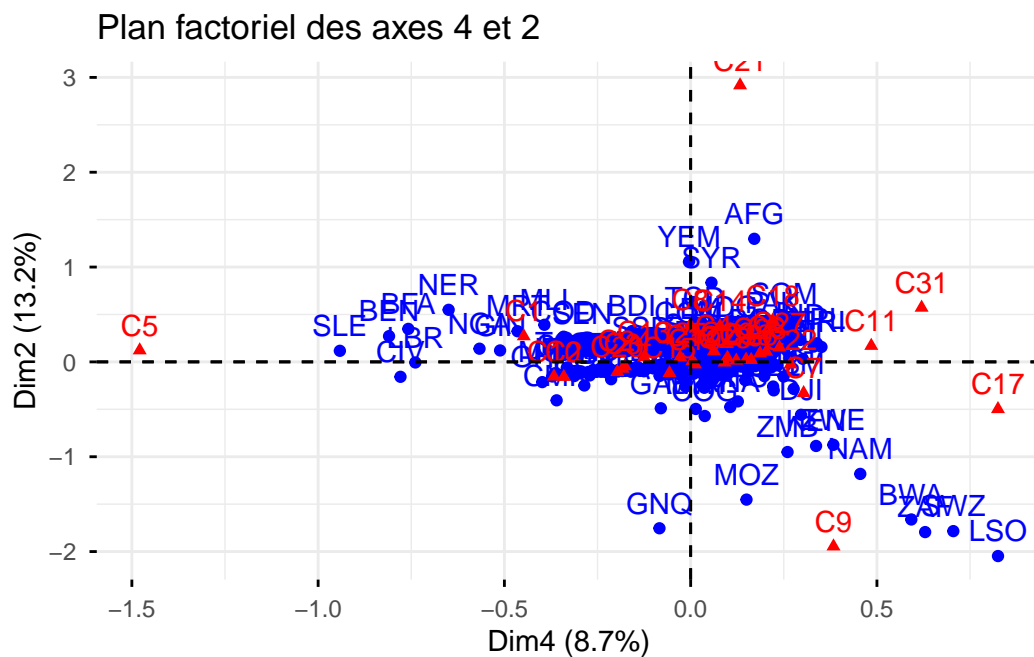
d'identifier les groupes de pays aux profils de mortalité similaires ainsi que les causes de décès les plus structurantes.

Les plans factoriels secondaires, explorés grâce aux combinaisons d'axes 1-3, 1-4, 3-2, 4-2 et 4-3, apportent des informations complémentaires sur les dimensions moins dominantes. Par exemple, le plan 1-3 explique 43 % et 10,5 % de l'inertie, le plan 1-4 contribue à 43 % et 8,7 %, et les autres plans permettent de mettre en évidence des variations supplémentaires entre pays et causes de décès. L'option `repel = T` utilisée dans les biplots améliore la lisibilité en évitant le chevauchement des étiquettes. L'analyse combinée de ces différents plans factoriels permet ainsi de mettre en évidence à la fois les structures principales et les relations secondaires dans les données.



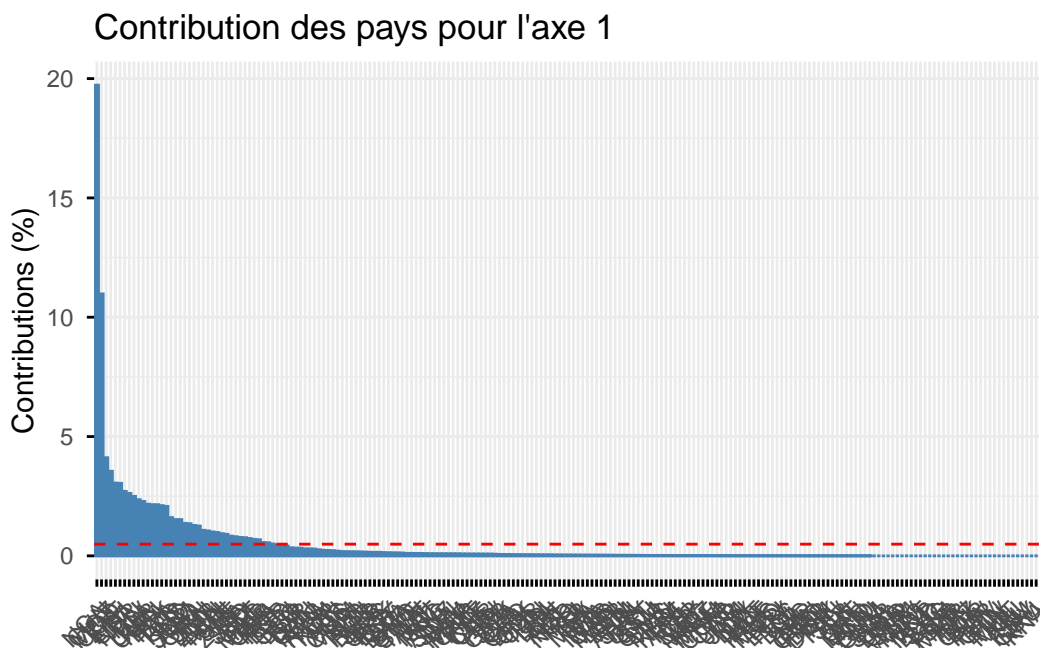
PCA plot showing the first two principal components (Dim1 and Dim4) for 20 countries. The x-axis represents Dim1 (43%) and the y-axis represents Dim4 (8.7%). Countries are colored blue (Group 1) or red (Group 2) and shaped as circles (Group 1) or triangles (Group 2). A dashed vertical line is at Dim1 = 0.

Country	Group	Dim1 (43%)	Dim4 (8.7%)
C1	2	1.5	-0.3
C2	1	-0.5	0.2
C3	1	-0.2	0.1
C4	1	-0.1	0.0
C5	2	2.2	-1.5
C6	1	-0.3	0.1
C7	1	0.0	0.2
C8	1	0.1	0.1
C9	2	1.5	0.4
C10	1	-0.2	0.0
C11	2	0.5	0.6
C12	1	-0.4	0.0
C13	1	-0.3	-0.1
C14	1	-0.2	0.0
C15	1	0.0	0.0
C16	1	0.1	-0.1
C17	2	0.5	0.5
C18	1	0.2	-0.1
C19	1	0.0	0.0
C20	1	-0.1	0.0
C21	1	-0.2	0.0
C22	1	-0.3	0.0
C23	1	-0.4	0.0
C24	1	0.1	-0.2
C25	1	-0.2	-0.1
C26	1	-0.3	-0.2
C27	1	-0.4	-0.3
C28	1	-0.5	-0.4
C29	1	-0.6	-0.5
C30	1	-0.7	-0.6
C31	2	0.5	0.6
C32	1	-0.1	0.1
C33	1	0.0	0.1
C34	1	0.1	0.2
C35	1	0.2	0.3
C36	1	0.3	0.4
C37	1	0.4	0.5
C38	1	0.5	0.6
C39	1	0.6	0.7
C40	1	0.7	0.8
C41	1	0.8	0.9
C42	1	0.9	1.0
C43	1	1.0	1.1
C44	1	1.1	1.2
C45	1	1.2	1.3
C46	1	1.3	1.4
C47	1	1.4	1.5
C48	1	1.5	1.6
C49	1	1.6	1.7
C50	1	1.7	1.8
C51	1	1.8	1.9
C52	1	1.9	2.0
C53	1	2.0	2.1
C54	1	2.1	2.2
C55	1	2.2	2.3
C56	1	2.3	2.4
C57	1	2.4	2.5
C58	1	2.5	2.6
C59	1	2.6	2.7
C60	1	2.7	2.8
C61	1	2.8	2.9
C62	1	2.9	3.0
C63	1	3.0	3.1
C64	1	3.1	3.2
C65	1	3.2	3.3
C66	1	3.3	3.4
C67	1	3.4	3.5
C68	1	3.5	3.6
C69	1	3.6	3.7
C70	1	3.7	3.8
C71	1	3.8	3.9
C72	1	3.9	4.0
C73	1	4.0	4.1
C74	1	4.1	4.2
C75	1	4.2	4.3
C76	1	4.3	4.4
C77	1	4.4	4.5
C78	1	4.5	4.6
C79	1	4.6	4.7
C80	1	4.7	4.8
C81	1	4.8	4.9
C82	1	4.9	5.0
C83	1	5.0	5.1
C84	1	5.1	5.2
C85	1	5.2	5.3
C86	1	5.3	5.4
C87	1	5.4	5.5
C88	1	5.5	5.6
C89	1	5.6	5.7
C90	1	5.7	5.8
C91	1	5.8	5.9
C92	1	5.9	6.0
C93	1	6.0	6.1
C94	1	6.1	6.2
C95	1	6.2	6.3
C96	1	6.3	6.4
C97	1	6.4	6.5
C98	1	6.5	6.6
C99	1	6.6	6.7
C100	1	6.7	6.8
C101	1	6.8	6.9
C102	1	6.9	7.0
C103	1	7.0	7.1
C104	1	7.1	7.2
C105	1	7.2	7.3
C106	1	7.3	7.4
C107	1	7.4	7.5
C108	1	7.5	7.6
C109	1	7.6	7.7
C110	1	7.7	7.8
C111	1	7.8	7.9
C112	1	7.9	8.0
C113	1	8.0	8.

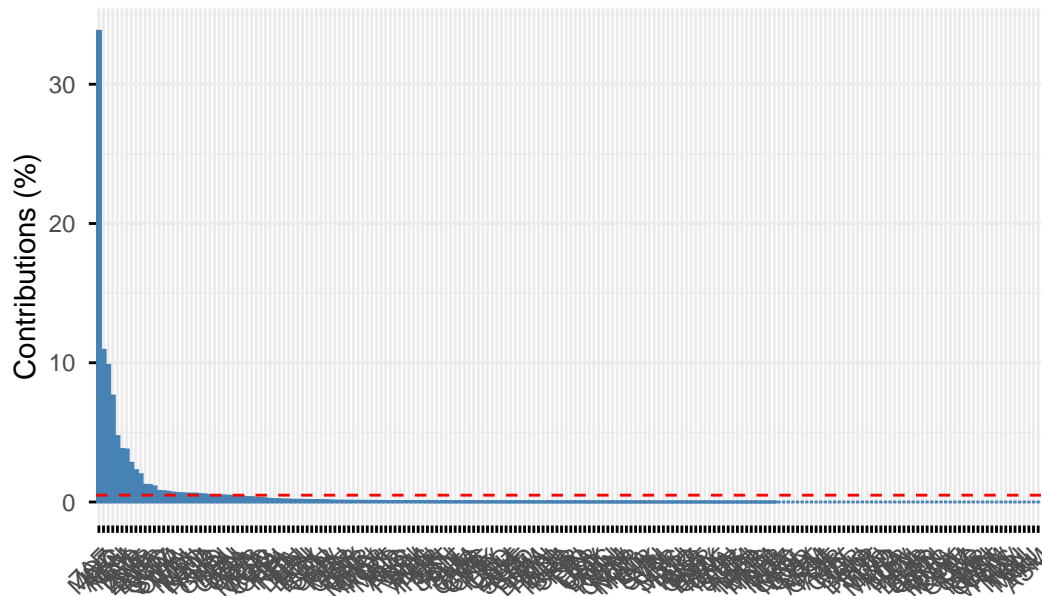


Pour approfondir l'interprétation, les contributions des lignes et des colonnes aux axes factoriels sont examinées. Les modalités les plus contributives sont identifiées, ce qui permet de se concentrer sur les pays et causes qui expliquent le mieux la variance sur chaque axe. Étant donné le grand nombre de pays (204), une sélection des dix premières contributions pour chaque

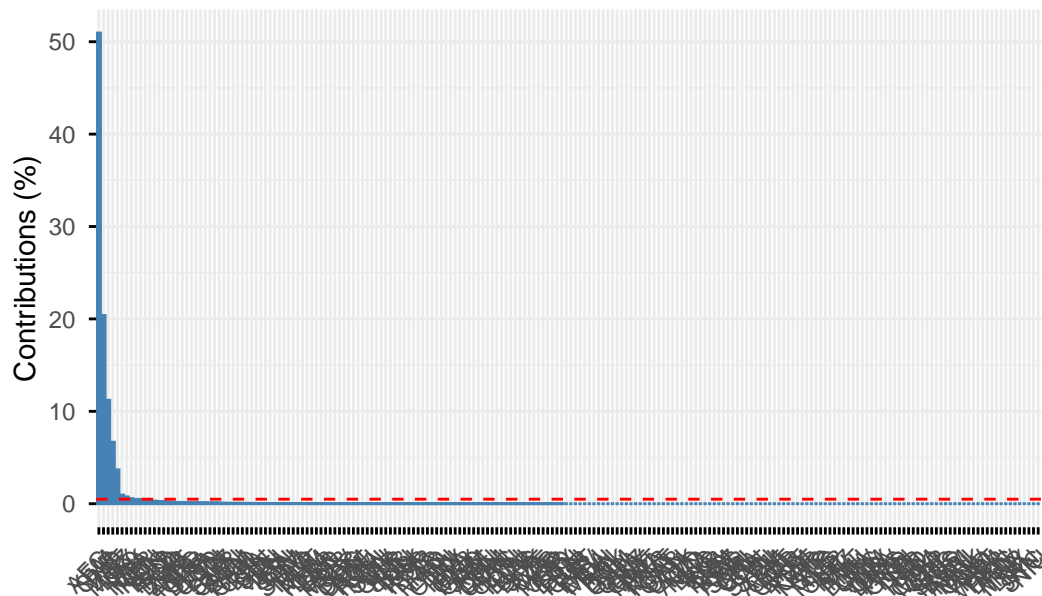
axe facilite la lecture et l'interprétation des résultats. Sur l'axe 1, le Nigeria (NGA, 20%), la Chine et le Mozambique sont les principaux contributeurs, tandis que la malaria (C5) domine parmi les causes de décès (21%). L'axe 2 est fortement influencé par l'Afrique du Sud (ZAF, 33,8%) et le VIH (C9) (68%). L'axe 3 met en avant l'Afghanistan (AFG, 50,9%) et les conflits et terrorisme C21 (plus de 80%), et l'axe 4 est porté par l'Inde (IND, 24,2%) et la malaria (plus de 40%).

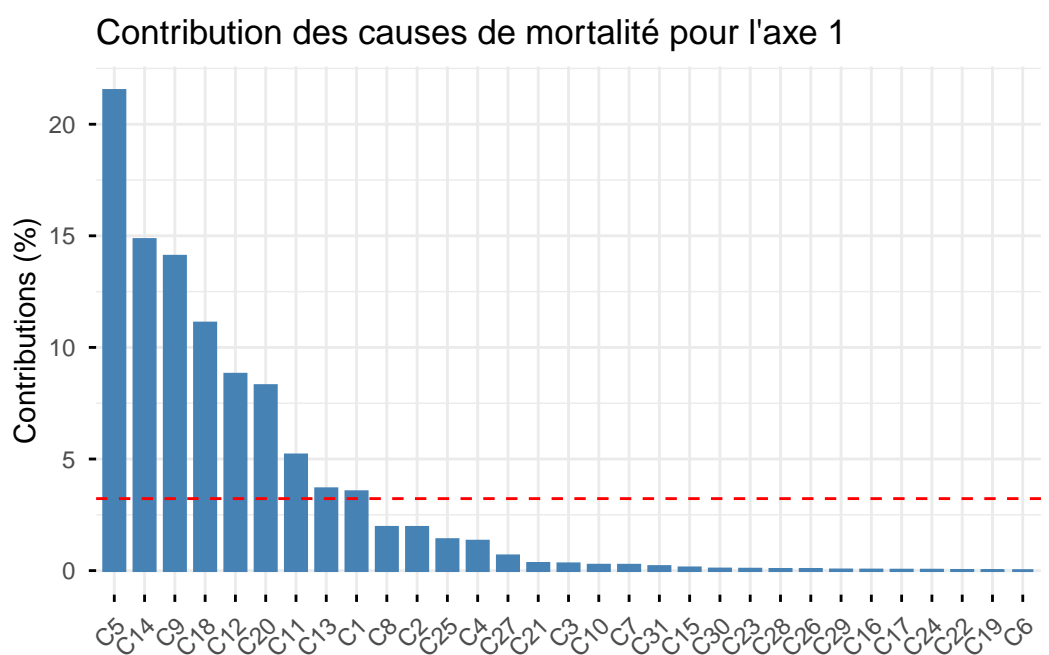
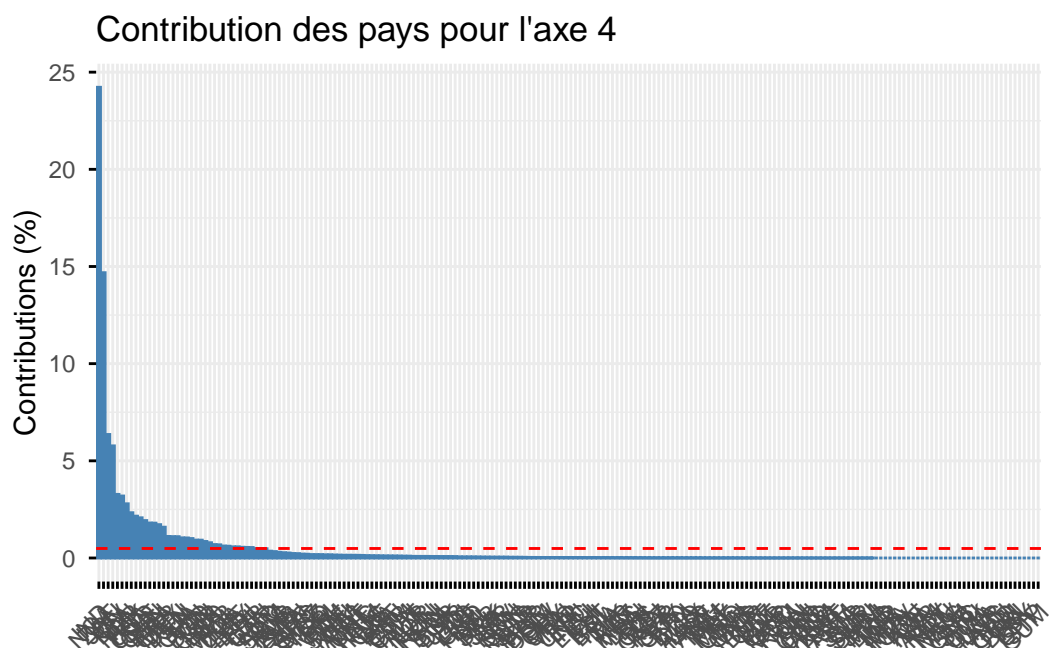


Contribution des pays pour l'axe 2

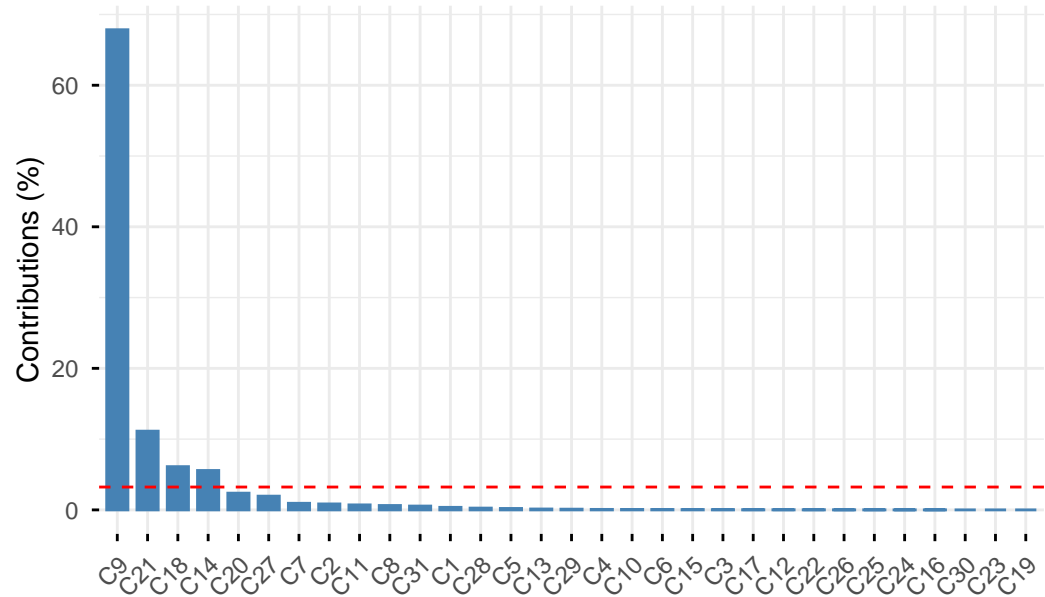


Contribution des pays pour l'axe

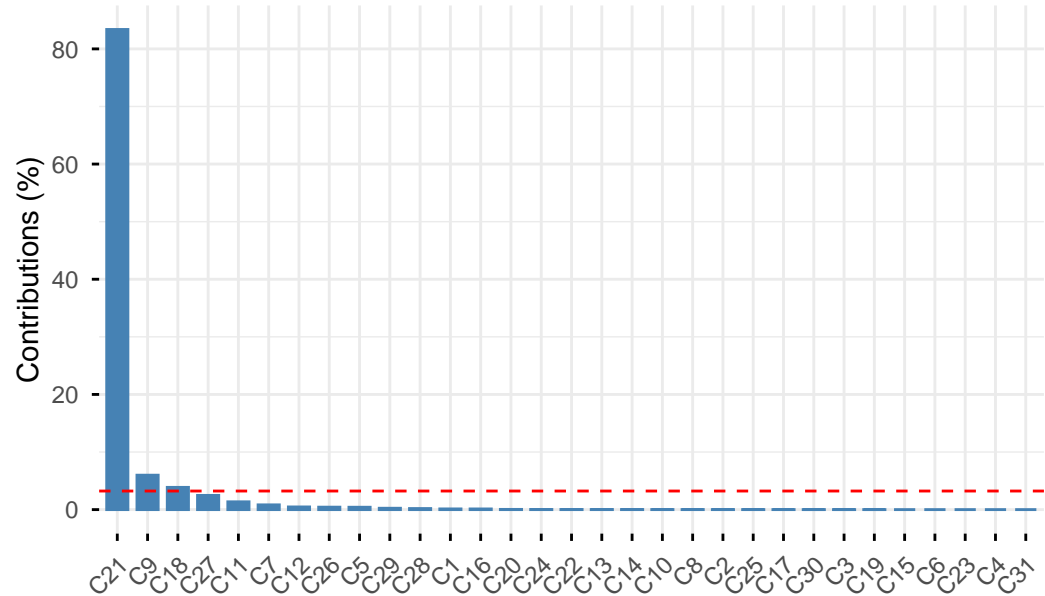


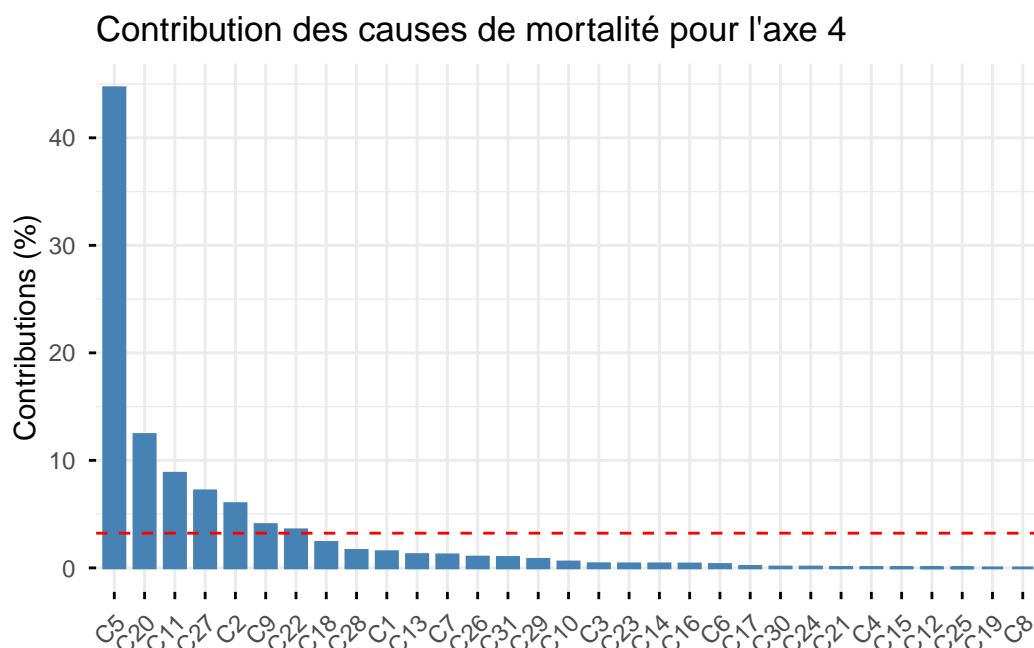


Contribution des causes de mortalité pour l'axe 2



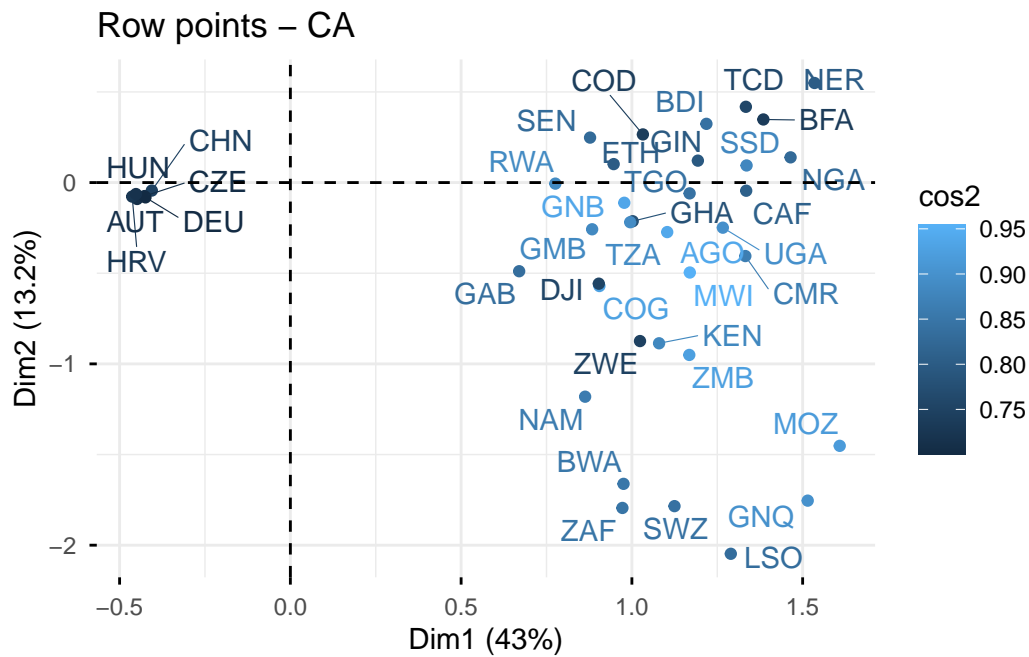
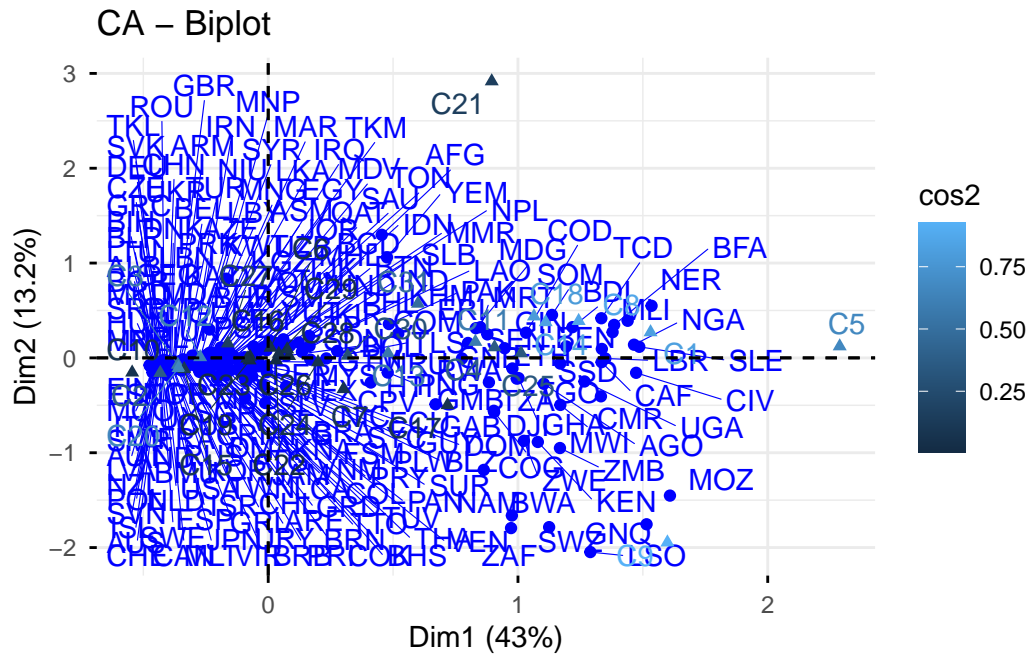
Contribution des causes de mortalité pour l'axe 3

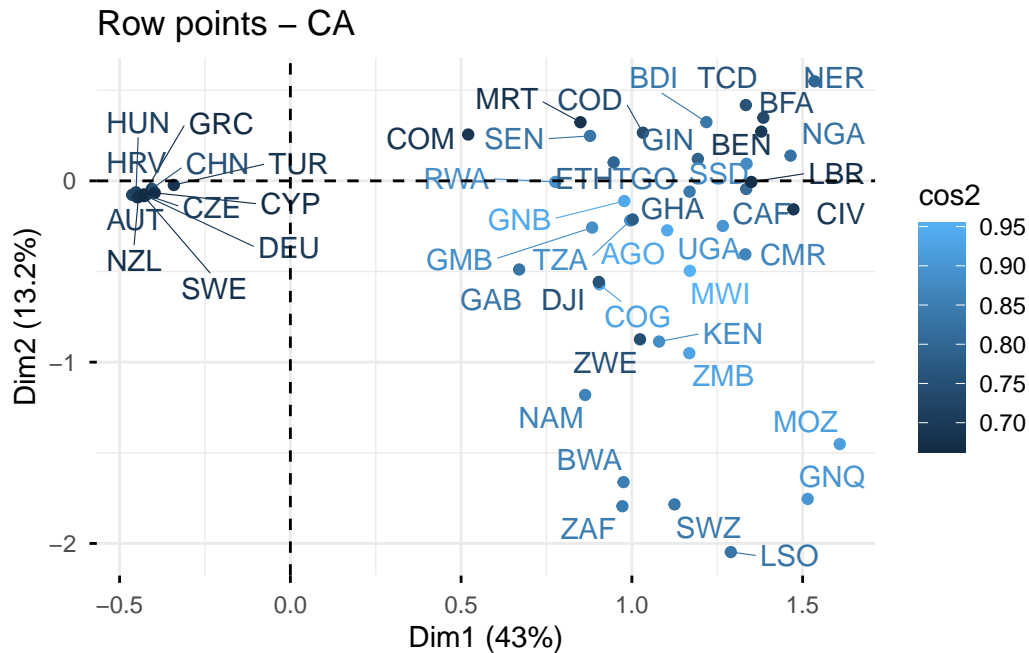




La qualité de représentation de chaque modalité est évaluée à l'aide du \cos^2 , qui mesure la proportion de la variance de la modalité expliquée par les axes retenus. Les modalités présentant un \cos^2 élevé sont mieux représentées sur le plan factoriel et constituent des points de référence fiables pour l'interprétation. Pour rendre les graphiques plus lisibles, nous avons d'abord filtré les modalités avec un \cos^2 supérieur à 0,7, garantissant ainsi que seules celles bien expliquées par les axes sont affichées. Ensuite un second filtrage a permis de sélectionner les 50 mieux représentées pour produire un autre graphique. Ces derniers utilisent un code couleur : les modalités les mieux représentées, avec un \cos^2 proche de 1, apparaissent en couleur claire, tandis que celles avec un \cos^2 proche de 0,7 apparaissent en couleur plus foncée.

PCA plot showing the first two dimensions (Dim1: 43%, Dim2: 13.2%) of the data. The plot includes a color scale for \cos^2 values ranging from 0.25 to 0.75. The samples are labeled with their IDs (C1 to C27) and arrows pointing to their positions. A vertical dashed line is at Dim1 = 0 and a horizontal dashed line is at Dim2 = 0.





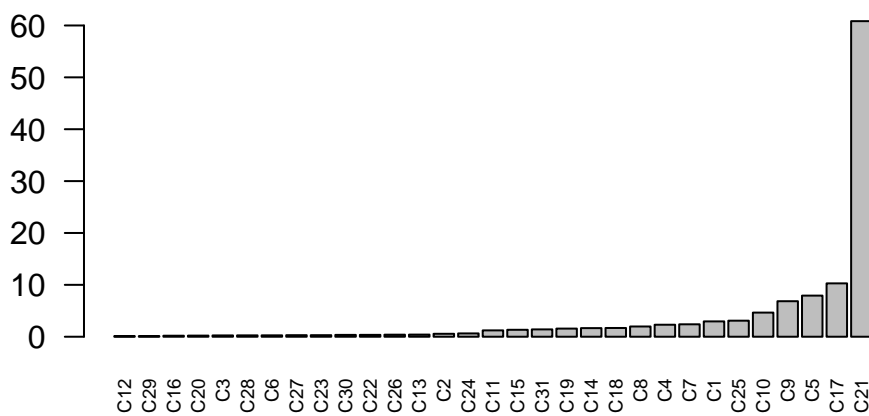
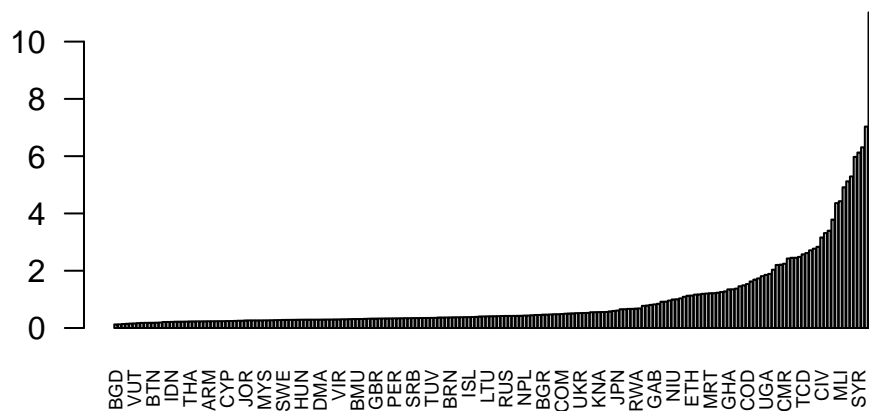
Cette analyse graphique et quantitative des contributions et de la qualité de représentation fournit une base solide pour tirer des conclusions sur les profils de mortalité et les relations entre pays et causes dans l'ensemble du jeu de données.

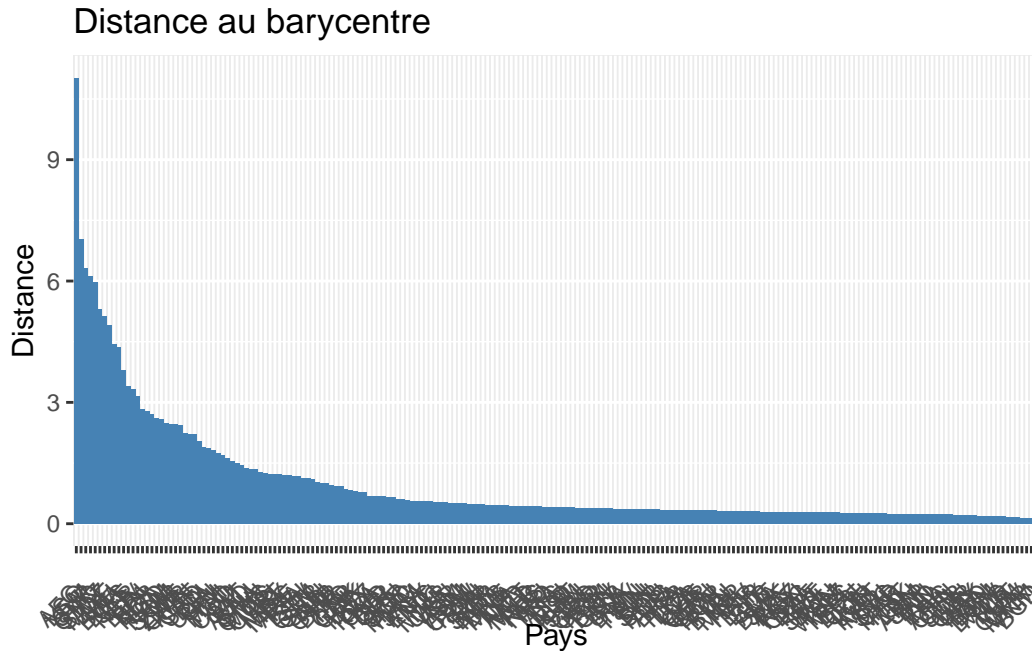
2.4.3 Distances au centre de gravité

L'AFC permet de mesurer l'écart de chaque pays ou cause par rapport au profil moyen à travers la distance au centre de gravité (barycentre). Cette distance est calculée en rapportant l'inertie de chaque modalité à sa masse. Par exemple, dans notre analyse, le Bangladesh (BGD) présente la distance la plus faible (0,12), ce qui signifie que son profil est très proche du profil moyen, tandis que l'Afghanistan (AFG) a la distance la plus élevée (11,02), indiquant un profil fortement atypique.

Pour les causes, certaines présentent également des écarts importants : les conflits et terrorismes (C21) se distinguent par une distance très élevée (60,83), ce qui suggère un profil extrêmement spécifique par rapport aux autres causes. Ces mesures permettent ainsi de détecter les pays et causes qui se rapprochent le plus du profil moyen et ceux qui s'en écartent fortement.

La visualisation de ces distances, sous forme de diagrammes en barres verticales ou horizontales, facilite l'interprétation. Les barres les plus courtes correspondent aux profils proches de la moyenne, tandis que les barres plus longues signalent des profils atypiques.





3 Conclusion

L'objectif de ce travail était d'étudier l'existence d'un lien entre les causes de décès et les pays, et d'identifier d'éventuelles similarités de profils de mortalité entre ces derniers, en mobilisant l'analyse factorielle des correspondances. Les résultats obtenus permettent d'apporter des réponses claires aux questions formulées en introduction.

Tout d'abord, le test du khi-deux, complété par l'indice de V de Cramer, met en évidence l'existence d'une dépendance entre les pays et les causes de décès. Bien que cette dépendance demeure d'intensité modérée, elle est néanmoins significative et confirmée par la structure révélée par l'AFC. Cette analyse montre que les causes de décès ne se répartissent pas aléatoirement selon les pays, mais qu'elles sont associées à des profils nationaux spécifiques.

L'examen du plan factoriel met en lumière des regroupements cohérents de pays présentant des profils de mortalité similaires. Les pays dits développés, majoritairement situés en Europe, en Amérique du Nord ou dans certaines régions d'Asie, apparaissent proches du profil moyen. Leur mortalité est principalement dominée par des maladies chroniques et dégénératives, telles que les maladies cardiovasculaires, les maladies neurologiques ou celles liées au vieillissement. À l'inverse, de nombreux pays africains se situent à distance du centre de gravité et présentent des profils proches les uns des autres, caractérisés par une forte prévalence de maladies infectieuses, de la malnutrition, ainsi que de décès liés aux conflits ou à l'instabilité politique. Ces causes, fortement contributrices aux axes factoriels, expliquent une part importante de la dépendance

observée. De manière similaire, certains pays en situation de crise hors du continent africain, tels que l’Afghanistan ou le Yémen, se distinguent également par des profils très éloignés du profil moyen, en lien avec des causes de mortalité spécifiques comme les conflits armés, le terrorisme, le paludisme ou le VIH.

Ainsi, les résultats obtenus apparaissent globalement cohérents avec le contexte géographique, géopolitique et socio-économique mondial. Ils confirment l’existence d’un lien entre le niveau de développement des pays et la structure de leur mortalité, ainsi que la présence de profils de décès similaires parmi des pays partageant des conditions de développement et de stabilité comparables. Ces observations doivent toutefois être interprétées avec prudence, au regard des limites liées à l’année étudiée et à l’agrégation des causes de décès.

4 Annexe

4.1 Source du jeu de données initiale

URL : <https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world>

4.2 Tableau de l’encodage des différentes causes de décès

Table 5: Encodage des causes de décès

Meningitis	C1
Alzheimer’s Disease and Other Dementias	C2
Parkinson’s Disease	C3
Nutritional Deficiencies	C4
Malaria	C5
Drowning	C6
Interpersonal Violence	C7
Maternal Disorders	C8
HIV/AIDS	C9
Drug Use Disorders	C10
Tuberculosis	C11
Cardiovascular Diseases	C12
Lower Respiratory Infections	C13
Neonatal Disorders	C14
Alcohol Use Disorders	C15
Self-harm	C16
Exposure to Forces of Nature	C17

Meningitis	C1
Diarrheal Diseases	C18
Environmental Heat and Cold Exposure	C19
Neoplasms	C20
Conflict and Terrorism	C21
Diabetes Mellitus	C22
Chronic Kidney Disease	C23
Poisonings	C24
Protein-Energy Malnutrition	C25
Road Injuries	C26
Chronic Respiratory Diseases	C27
Cirrhosis and Other Chronic Liver Diseases	C28
Digestive Diseases	C29
Fire, Heat, and Hot Substances	C30
Acute Hepatitis	C31
