Université Paris-Dauphine – PSL
Master 280 – Ingénierie Statistique et Financière

MONITORAGE ET SEGMENTATION DE LA TUBERCULOSE (OMS)

« Au-delà des agrégats nationaux : comment l'analyse multivariée permet-elle de révéler une typologie opérationnelle des risques sanitaires mondiaux face à la tuberculose ? »

Arthur DANJOU

Enseignant :
Quentin GUIBERT

Data Visualisation
Année Universitaire 2025-2026

Contents

- **Application déployée :** https://go.arthurdanjou.fr/datavis-app
- **Code Source de(GitHub) :** https://go.arthurdanjou.fr/datavis-code

1 Introduction

1.1 Contexte et enjeux sanitaires

Avec 1,6 million de décès annuels et plus de 10 millions de nouveaux cas estimés en 2022, la tuberculose (TB) demeure la deuxième maladie infectieuse la plus meurtrière au monde après le COVID-19 (OMS, 2025). Pourtant, derrière ces chiffres globaux se cache une épidémie profondément inégalitaire. Alors que certains pays rapportent une incidence maîtrisée inférieure à 10 cas pour 100 000 habitants, d'autres font face à des taux critiques dépassant les 500 cas, révélant des fractures sanitaires majeures entre les nations.

Pour piloter la réponse mondiale, l'Organisation Mondiale de la Santé produit le *Global Tuberculosis Report*, une base de données exhaustive comptant plus de 200 pays et une quarantaine d'indicateurs. Cependant, la richesse même de ces données pose un défi d'analyse : face à la multitude de variables (incidence, notification, mortalité, co-infection), les tableaux statistiques traditionnels échouent à offrir une vision synthétique et opérationnelle. Ils ne permettent ni d'identifier rapidement les profils à risque, ni de visualiser les dynamiques temporelles complexes.

1.2 Problématique : Au-delà des agrégats nationaux : comment l'analyse multivariée permet-elle de révéler une typologie opérationnelle des risques sanitaires mondiaux face à la tuberculose ?

Ce projet déploie une chaîne de traitement Data Science complète reposant sur trois piliers. Premièrement, une rationalisation de la donnée par sélection de variables et analyse exploratoire (EDA) pour isoler les signaux pertinents. Deuxièmement, une segmentation intelligente (Clustering K-Means) pour identifier des profils de risque homogènes au-delà des simples zones géographiques. Enfin, une opérationnalisation interactive via une application R Shiny, offrant aux décideurs une interface dynamique pour visualiser les tendances 2000-2024.

1.3 Périmètre et Structure

L'étude se concentre sur les indicateurs épidémiologiques "durs" pour garantir la robustesse du modèle, les facteurs exogènes (PIB, dépenses) étant considérés comme contextuels.

La suite de cette notice détaille la méthodologie : la préparation des données (Section 2) et la modélisation mathématique (Section 3) précèdent l'analyse des profils identifiés (Section 4). L'architecture de l'application R Shiny est décrite en Section 5, suivie de l'exploitation des résultats et du benchmarking (Section 6). Le document se clôt sur les perspectives d'évolution (Section 7) et le cadre d'intégrité académique (Section 8).

2 Analyse Exploratoire des Données

2.1 Source et structure des données

2.1.1 Origine et portée des données

Le socle empirique repose sur les données du *Global Tuberculosis Report* 2024 de l'OMS, référence internationale couvrant 25 ans (2000-2024) pour 215 territoires. Le fichier brut de 50 variables s'articule autour de trois dimensions complémentaires : **épidémiologique** (morbidité, mortalité, prise en charge), **démographique** (structure de population nécessaire à la standardisation des taux) et **géopolitique** (métadonnées spatiales et codes ISO-3) dédiées à l'analyse spatiale.

2.1.2 Convention de nommage et sémantique

L'analyse requiert la maîtrise d'une nomenclature rigoureuse distinguant les **Cas notifiés** (préfixe `c_`, données brutes administratives) des **Estimations modélisées** (préfixe `e_`), par lesquelles l'OMS corrige les biais de sous-déclaration et intègre les incertitudes. Pour cette étude, nous privilégierons exclusivement ces variables estimées (`e_`) : ce choix méthodologique permet de neutraliser l'hétérogénéité des performances administratives locales afin de garantir une comparabilité internationale stricte des dynamiques épidémiques.

2.1.3 Qualité des données et limites

Bien qu'offrant une profondeur spatio-temporelle unique appuyée par une méthodologie standardisée, ce jeu de données présente des hétérogénéités inhérentes à la surveillance mondiale. Les biais de mesure restent prégnants pour les pays à faibles revenus ou en conflit, où les estimations reposent sur l'extrapolation statistique plutôt que sur un comptage exhaustif, sans compter le caractère provisoire des données récentes (2023-2024). Ces limites intrinsèques justifient l'adoption d'une approche méthodologique prudente, privilégiant l'exclusion des variables incertaines et le rejet de l'imputation pour les observations incomplètes.

2.1.4 Importation et aperçu initial

Le jeu de données importé contient 5347 observations et 50 variables. Le tableau ci-dessous présente les dix premières lignes du jeu de données, illustrant la structure longitudinale pour le premier pays par ordre alphabétique (Afghanistan) au début de la période d'étude (de 2000 à 2009).

Table 1: Aperçu des premières lignes du jeu de données brut

Année	Pays	Région	Incidence (/100k)	Mortalité (/100k)
2000	Afghanistan	EMR	148	50
2001	Afghanistan	EMR	175	55
2002	Afghanistan	EMR	197	59
2003	Afghanistan	EMR	215	68
2004	Afghanistan	EMR	228	67
2005	Afghanistan	EMR	237	66
2006	Afghanistan	EMR	242	64
2007	Afghanistan	EMR	241	59
2008	Afghanistan	EMR	235	58
2009	Afghanistan	EMR	229	60

2.2 Sélection de variables

La qualité d'une segmentation non-supervisée étant tributaire de la pertinence des entrées, l'injection brute des 50 variables initiales a été écartée pour prévenir deux écueils méthodologiques. D'une part, le **fléau de la dimension** (*Curse of Dimensionality*) qui tend à uniformiser les distances euclidiennes et flouter les clusters et d'autre part, le **biais de redondance**, où la colinéarité des variables risque de surpondérer artificiellement un même phénomène. Nous avons donc déployé une stratégie de réduction de dimension en deux temps : un filtrage structurel (approche par entonnoir) consolidé par un arbitrage statistique des corrélations.

2.2.1 Approche par entonnoir : élimination des métadonnées, des bornes d'incertitude et des valeurs absolues

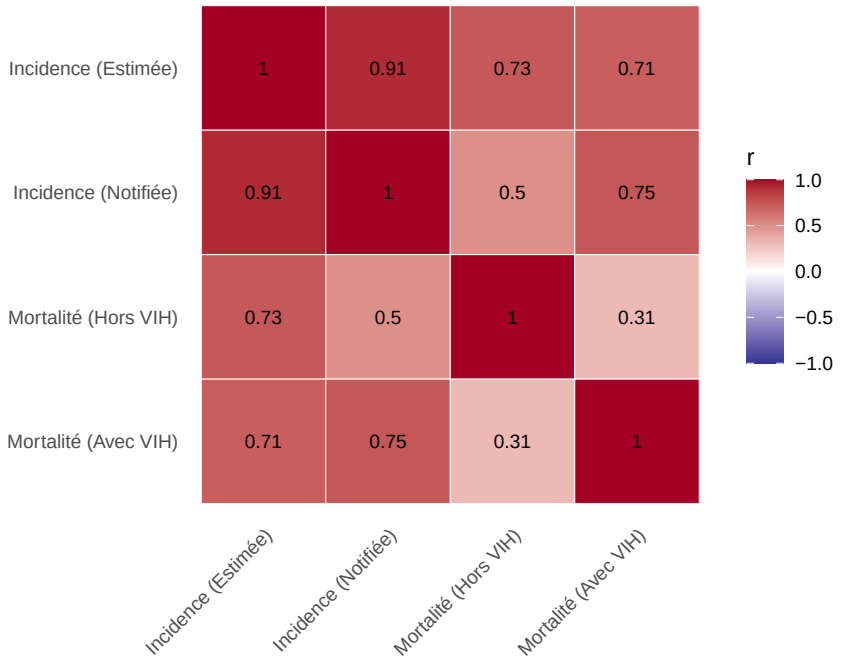
Une stratégie de réduction de dimension en quatre étapes successives a été appliquée pour isoler les variables pertinentes. Dans un premier temps, le **nettoyage structurel** et la simplification ont permis d'écarter les métadonnées techniques (ex: `iso_numeric`) ainsi que les bornes d'incertitude (`_lo`, `_hi`), jugées non pertinentes pour le calcul de distances euclidiennes ou redondantes avec l'estimation centrale. Ensuite, l'étape de **standardisation** a exclu

les valeurs absolues (`_num`) afin de neutraliser tout biais démographique et permettre la comparaison directe entre pays de tailles hétérogènes. Enfin, un **filtrage de la colinéarité** a supprimé les indicateurs redondants (corrélation $> 0{,}8$), tels que les notifications brutes, pour éviter de biaiser la pondération des dimensions dans l'algorithme de clustering.

2.2.2 Arbitrage méthodologique : traitement de la colinéarité (Incidence vs Notifications) et de la redondance (Mortalité vs Mortalité VIH)

À l'issue du filtrage structurel, il subsiste plusieurs candidats potentiels pour mesurer la charge épidémique. Pour éviter la redondance (colinéarité), nous analysons la matrice de corrélation de Pearson entre ces candidats.

L'objectif est de conserver les variables les plus représentatives tout en maximisant l'orthogonalité (l'indépendance) des informations fournies au modèle. La figure ci-dessous visualise la matrice de corrélation de Pearson entre les quatre variables candidates : l'incidence (estimée et notifiée) et la mortalité (avec et hors VIH).



2.2.2.1 Analyse et décisions de modélisation : L'analyse de la matrice de corrélation a imposé deux arbitrages majeurs. Premièrement, l'**Incidence Estimée** (`e_inc_100k`) a été préférée aux cas notifiés. En effet, ces derniers souffrent d'un biais administratif : un faible taux de notification peut refléter un manque de médecins plutôt qu'une absence de malades, alors que l'estimation de l'OMS corrige ces sous-diagnostics pour refléter la charge réelle.

Deuxièmement, nous avons retenu la **Mortalité hors VIH** (`e_mort_exc_tbhiv_100k`) malgré sa redondance avec la mortalité globale. Inclure la mortalité liée au VIH aurait risqué de biaiser la segmentation en isolant un "cluster SIDA" (spécifique à l'Afrique Australe), ce qui aurait masqué notre objectif principal : évaluer la performance des programmes antituberculeux indépendamment de l'accès aux antirétroviraux.

2.2.2.2 Synthèse des variables retenues : Le modèle de clustering reposera donc sur un couple de variables actives parcimonieux et complémentaire :

- Variable Active 1 : Incidence (Diffusion de la maladie) - `e_inc_100k`.
- Variable Active 2 : Mortalité (Sévérité / Échec du traitement) - `e_mort_exc_tbhiv_100k`

Ces deux dimensions, bien que corrélées ($r \approx 0.73$), ne sont pas redondantes : la variance non expliquée par la corrélation correspond justement à la différence d'efficacité des systèmes de soins (capacité à guérir les malades identifiés), ce qui est le cœur de notre segmentation.

2.2.3 Variables illustratives et contextuelles

En complément des variables actives, cinq variables illustratives sont conservées pour éclairer l'interprétation a posteriori sans biaiser le calcul des distances euclidiennes. Le contexte démographique est porté par la Population (`e_pop_num`), indispensable aux pondérations, tandis que le volet géopolitique repose sur la **Région OMS** (`g_whoregion`), structurant l'analyse spatiale en six zones administratives (AFR, AMR, EMR, EUR, SEA, WPR). Enfin, les identifiants techniques — **Pays, Code ISO et Année** — assurent les fonctions supports : étiquetage, jointure cartographique et filtrage dynamique des trajectoires temporelles.

2.2.3.1 Création du sous-ensemble de travail : Nous appliquons cette sélection au jeu de données brut pour ne conserver que les 7 colonnes d'intérêt.

```
tb_clean <- data_raw |>
  select(
    iso3,
    country,
    year,
    g_whoregion,
    e_inc_100k,
    e_mort_exc_tbhiv_100k,
    e_pop_num
  )
```

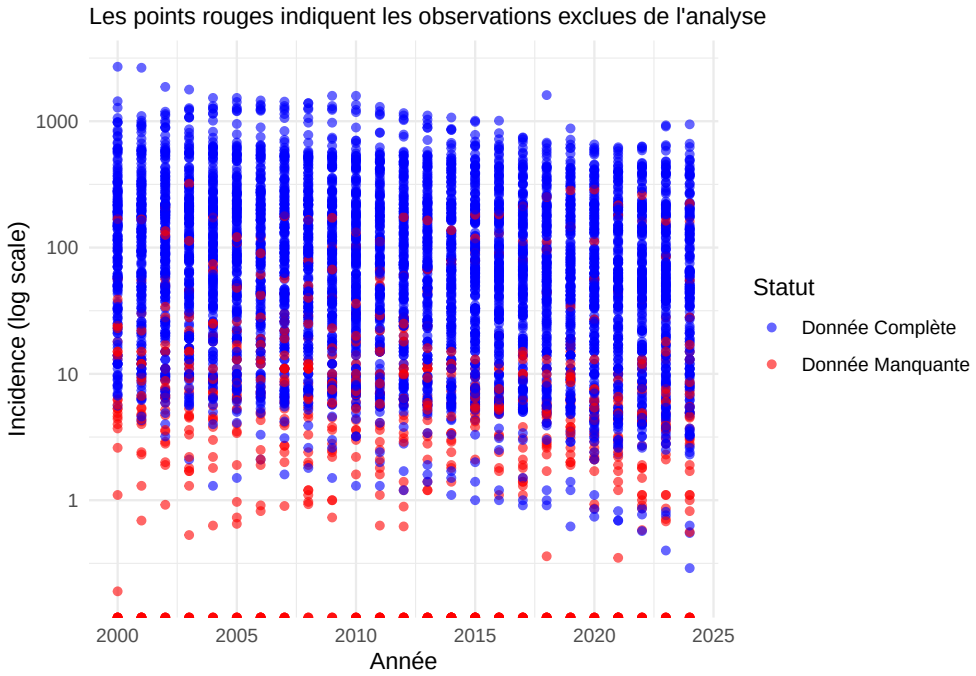
2.3 Traitement des valeurs manquantes

La gestion des valeurs manquantes (NA) est une étape critique en analyse de données, particulièrement pour les méthodes de partitionnement comme les K-Means qui reposent sur des calculs de distance euclidienne et ne tolèrent aucune incomplétude vectorielle.

Cette étape ne relève pas du simple "nettoyage" technique mais constitue un choix méthodologique qui influence la représentativité de l'échantillon final.

2.3.1 Diagnostic de la structure des manquants

Nous analysons la distribution spatio-temporelle des valeurs manquantes sur la variable de mortalité (`e_mort_exc_tbhiv_100k`), l'incidence étant complète par construction (filtrage préalable).



2.3.2 Analyse d'impact de l'exclusion

Le tableau ci-dessous identifie les territoires les plus affectés :

Table 2: Top 10 des territoires exclus pour données manquantes

Territoire	Région	Années manquantes	Incidence Moyenne	Population Moy.
American Samoa	WPR	25	6	53 468
Andorra	EUR	25	8	76 015
Anguilla	AMR	25	1	13 359
Antigua and Barbuda	AMR	25	4	85 824
Aruba	AMR	25	7	100 598
Barbados	AMR	25	2	274 980
Bermuda	AMR	25	3	63 425
British Virgin Islands	AMR	25	2	29 989
Cayman Islands	AMR	25	4	57 086
Cook Islands	WPR	25	7	15 991

Ce tableau confirme que les données manquantes concernent quasi-exclusivement des micro-états et territoires insulaires à très faible démographie (souvent inférieure à 100 000 habitants), validant ainsi leur exclusion sans impact significatif sur la représentativité mondiale de l'étude.

2.3.3 Justification méthodologique

L'exclusion des données manquantes se fonde sur trois justifications méthodologiques. D'un point de vue **géographique**, ces lacunes concernent quasi-exclusivement des micro-états ou territoires insulaires (ex: Monaco, Anguilla) dont la faible démographie induit une volatilité statistique excessive rendant les estimations peu fiables. Sur le plan **épidémiologique**, cette suppression est sans impact stratégique : ces territoires, bien que représentant 15% des observations, ne cumulent que 0,1% de la population mondiale et affichent une incidence marginale (17 cas/100k contre 125 pour l'échantillon conservé). Enfin, l'**intégrité statistique** a prévalu sur l'exhaustivité artificielle : le recours à l'imputation a été écarté car la génération de valeurs synthétiques pour ces profils atypiques risquerait de bruiter le calcul des distances euclidiennes et d'introduire des artefacts mathématiques préjudiciables au clustering.

2.3.4 Finalisation de l'échantillon

Nous appliquons donc le filtre définitif pour générer le jeu de données d'analyse.

```
tb_clean <- tb_clean |> drop_na(e_inc_100k, e_mort_exc_tbhiv_100k)
```

L'exclusion des observations incomplètes réduit la taille de l'échantillon de 15% (de 5 322 à 4 532 observations valides), couvrant 183 pays sur la période 2000-2024

2.4 Analyse et Transformation

Cette étape vise à caractériser la structure distributionnelle des variables actives (`tb_clean`). L'objectif est double : comprendre la dynamique épidémique sous-jacente et préparer les données pour satisfaire les hypothèses de l'algorithme K-Means (sensibilité aux valeurs extrêmes et aux variances inégales).

2.4.1 Statistiques descriptives et asymétrie

Le tableau ci-dessous résume les moments statistiques des deux variables actives sur l'ensemble de la période ($n = 4532$ observations).

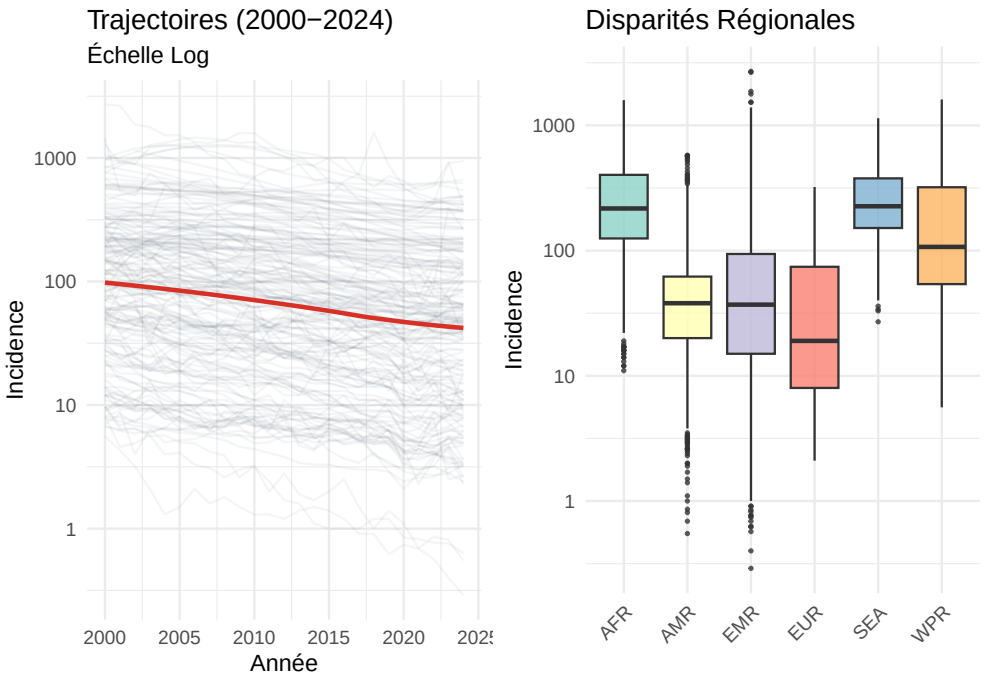
Table 3: Statistiques descriptives des variables actives (2000-2024)

Var	Min	Q1	Med	Mean	Q3	Max	Skew
Incidence	0.29	21.75	73.0	159.34	207	2700	3.17
Mortalité	0.00	1.20	6.1	22.22	27	1210	9.99

L'écart considérable entre la médiane et la moyenne, couplé à des coefficients d'asymétrie (Skewness) largement supérieurs à 1, indique des distributions fortement asymétriques à droite (Lognormales ou de Pareto). Concrètement, la majorité des pays présentent une charge épidémique faible, tandis qu'une minorité d'observations "extrêmes" tire la moyenne vers le haut. Cette structure est typique des phénomènes épidémiques mais problématique pour le K-Means, qui risque de créer des clusters uniquement pour isoler ces valeurs extrêmes.

2.4.2 Dynamiques temporelles et spatiales

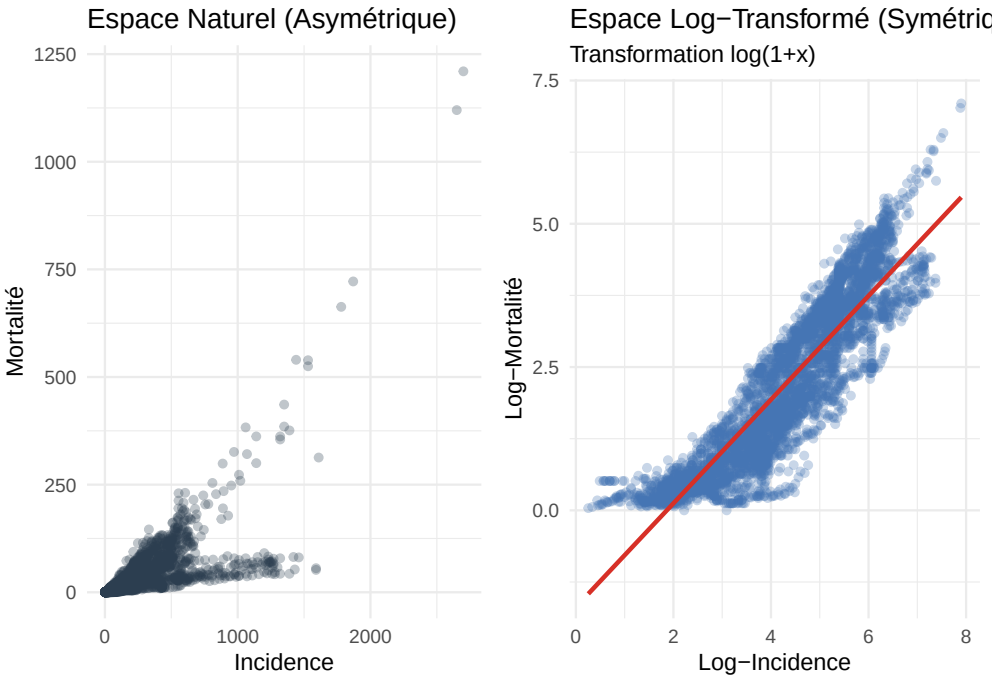
L'analyse visuelle permet de contextualiser ces statistiques globales.



L'analyse visuelle révèle une double dynamique. D'une part, la **tendance globale** montre une lente érosion de l'incidence moyenne mondiale (courbe rouge), malgré la forte inertie des trajectoires individuelles. D'autre part, les boxplots confirment une **fracture Nord-Sud** structurelle : les médianes logarithmiques de l'Afrique (AFR) et de l'Asie du Sud-Est (SEA) sont nettement supérieures à celles de l'Europe ou des Amériques. Cette hétérogénéité spatiale valide la pertinence d'inclure la région comme variable illustrative pour l'interprétation post-clustering.

2.4.3 Relation Bivariée et Transformation

La relation entre l'Incidence et la Mortalité est le cœur de notre modélisation.



Le graphique supérieur met en évidence une forte concentration à l'origine et une hétéroscédasticité marquée, risquant de biaiser les distances euclidiennes par les seules valeurs extrêmes. L'application de la transformation $f(x) = ln(1 + x)$ corrige ces biais structurels : elle **symétrise les distributions** pour optimiser l'occupation de l'espace vectoriel et **linéarise la relation** entre les variables, facilitant la détection de groupes naturels. De plus, contrairement au logarithme népérien standard, cette fonction assure une gestion **robuste des zéros** (évitant le cas $ln(0) = -\infty$ pour les pays sans décès), garantissant ainsi la stabilité numérique du modèle.

2.5 Synthèse de l'exploration, du nettoyage et des transformations

À l'issue de cette phase de préparation, nous disposons d'un jeu de données optimisé pour la modélisation.

Le tableau ci-dessous synthétise les caractéristiques du dataset final `tb_ready` qui sera injecté dans l'algorithme :

Table 4: Fiche d'identité du jeu de données final

Metrique	Valeur
Observations totales	4532
Pays couverts	183
Plage Temporelle	2000 - 2024
Variables Actives (Transformées)	log_inc, log_mort
Variables Illustratives	Population, Région, Année

La validation de ce socle de données clôture la phase exploratoire. L'absence de valeurs manquantes, la réduction de la dimensionnalité et la normalisation des distributions nous permettent désormais de procéder au partitionnement (Clustering) avec une robustesse statistique garantie.

3 Stratégie de Modélisation (Clustering)

La préparation des données ayant abouti à un espace vectoriel cohérent et symétrisé (`tb_ready`), nous procédons désormais à la segmentation proprement dite. Nous avons retenu l'algorithme des K-Means (Nuées dynamiques), une méthode de partitionnement non-supervisé privilégiée pour sa robustesse sur des jeux de données de dimension modérée et pour la lisibilité géométrique de ses résultats.

3.1 Prétraitement : Centrage et Réduction

Bien que nous ayons appliqué une transformation logarithmique pour corriger l'asymétrie, les variables d'Incidence et de Mortalité possèdent des plages de variation distinctes. L'algorithme K-Means reposant sur la distance euclidienne isotrope, il est impératif que chaque dimension contribue de manière équitable au calcul de similarité.

Nous appliquons donc une standardisation (Z-score) : $z = \frac{x-\mu}{\sigma}$

Table 5: Validation du Centrage-Réduction

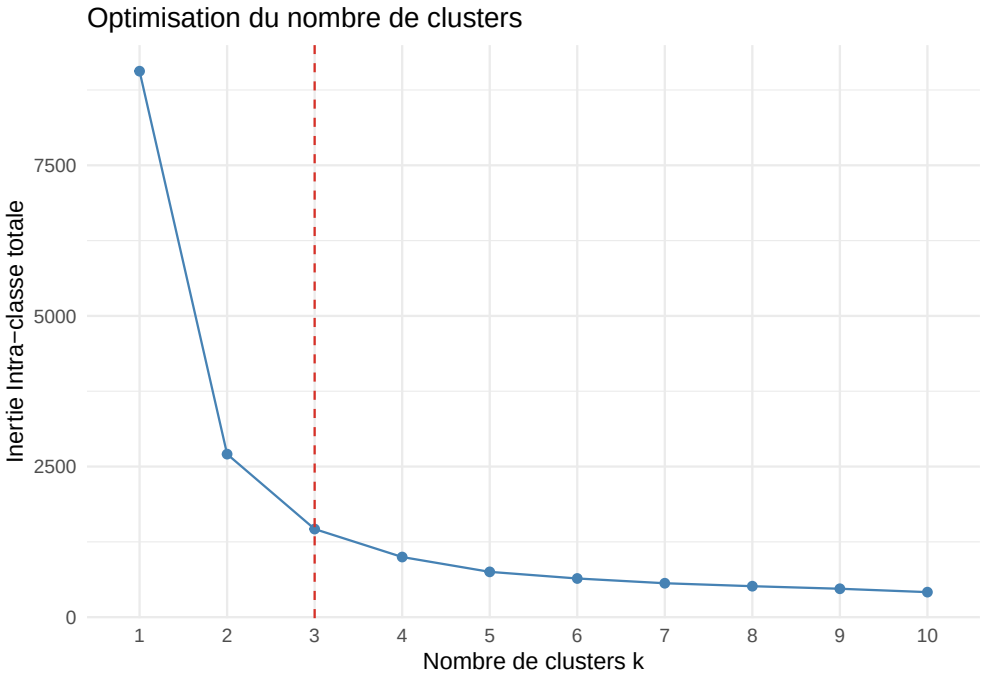
	Variable	Moyenne (Z)	Écart-Type (Z)
log_inc	Incidence (Log)	0	1
log_mort	Mortalité (Log)	0	1

3.2 Détermination du nombre de clusters (k)

L'algorithme K-Means nécessite de fixer a priori le nombre de classes k. Ce choix résulte d'un arbitrage entre performance statistique (minimisation de l'inertie intra-classe) et pertinence opérationnelle (interprétabilité métier).

3.2.1 Approche statistique (Méthode du Coude)

Nous calculons l'inertie intra-classe totale pour des valeurs de k allant de 1 à 10. Le point d'inflexion ("coude") indique le seuil au-delà duquel l'ajout d'un cluster n'apporte plus de gain significatif en compacité. Sur la figure ci-dessous, le coude se situe entre $k = 2$ et $k = 3$.



3.2.2 Arbitrage

L'analyse graphique révèle une rupture de pente franche à $k = 3$, seuil au-delà duquel les gains d'inertie deviennent marginaux (rendements décroissants). Ce choix statistique est corroboré par une pertinence opérationnelle majeure : une segmentation ternaire permet d'adopter une logique de signalisation intuitive type Traffic Light (Vert/Contrôle, Orange/Surveillance, Rouge/Critique). Nous retenons donc $k = 3$ afin de garantir des clusters à la fois statistiquement denses et immédiatement actionnables par les décideurs.

3.3 Paramétrage et Exécution de l'algorithme

L'algorithme K-Means étant sensible à l'initialisation des centroïdes (risque d'optimum local), nous avons configuré une exécution robuste : le modèle opère 25 initialisations aléatoires différentes (`nstart = 25`) pour ne conserver que la partition minimisant l'inertie globale sur les 3 classes définies (`centers = 3`). Enfin, la fixation de la graine aléatoire (`set.seed(123)`) garantit la stricte reproductibilité des résultats présentés.

```
set.seed(123)

km_res <- kmeans(data_scaled, centers = 3, nstart = 25)
var_totale <- round(km_res$betweenss / km_res$totss * 100, 1)
```

Avec **83,9 % de variance expliquée**, le modèle valide la robustesse statistique de la segmentation ternaire. Ce score élevé traduit une séparation nette des profils épidémiologiques, corroborant ainsi la forte structuration spatiale pressentie lors de l'analyse exploratoire.

3.4 Intégration des résultats

Nous réintégrons les labels de clusters dans le jeu de données principal pour l'analyse.

Table 6: Répartition des observations par cluster (k=3)

Cluster ID	Nombre d'observations
1	1416
2	1570
3	1546

Le partitionnement étant validé avec 3 classes, nous abordons désormais l'étape de labellisation visant à traduire ces clusters statistiques en profils épidémiologiques intelligibles.

4 Analyse des Profils Épidémiques

L'analyse mathématique ayant validé la qualité de la partition, nous procédons ici à la caractérisation "métier" des clusters pour les transformer en outils d'aide à la décision.

4.1 Caractérisation et Labellisation

Nous calculons les moyennes d'incidence et de mortalité par groupe, ordonnons les clusters du moins au plus sévère et leur attribuons des étiquettes sémantiques explicites.

Table 7: Aperçu de la segmentation sanitaire (Échantillon)

Pays	Année	Incidence (pour 100k)	Classification
Afghanistan	2000	148	3. Impact Critique
Afghanistan	2001	175	3. Impact Critique
Afghanistan	2002	197	3. Impact Critique
Afghanistan	2003	215	3. Impact Critique
Afghanistan	2004	228	3. Impact Critique
Afghanistan	2005	237	3. Impact Critique
Afghanistan	2006	242	3. Impact Critique
Afghanistan	2007	241	3. Impact Critique
Afghanistan	2008	235	3. Impact Critique
Afghanistan	2009	229	3. Impact Critique

4.2 Analyse des Profils Épidémiques

Le tableau ci-dessous synthétise les caractéristiques moyennes de chaque profil type identifié par le modèle.

Table 8: Typologie des clusters de Tuberculose (k=3)

label	Nombre d'observations	Incidence Moyenne	Mortalité Moyenne	Ratio Mort/Inc (%)
1. Impact Faible	1416	14	0.8	6.0
2. Impact Modéré	1570	79	7.0	8.8
3. Impact Critique	1546	374	57.3	15.3

4.2.1 Interprétation de la typologie

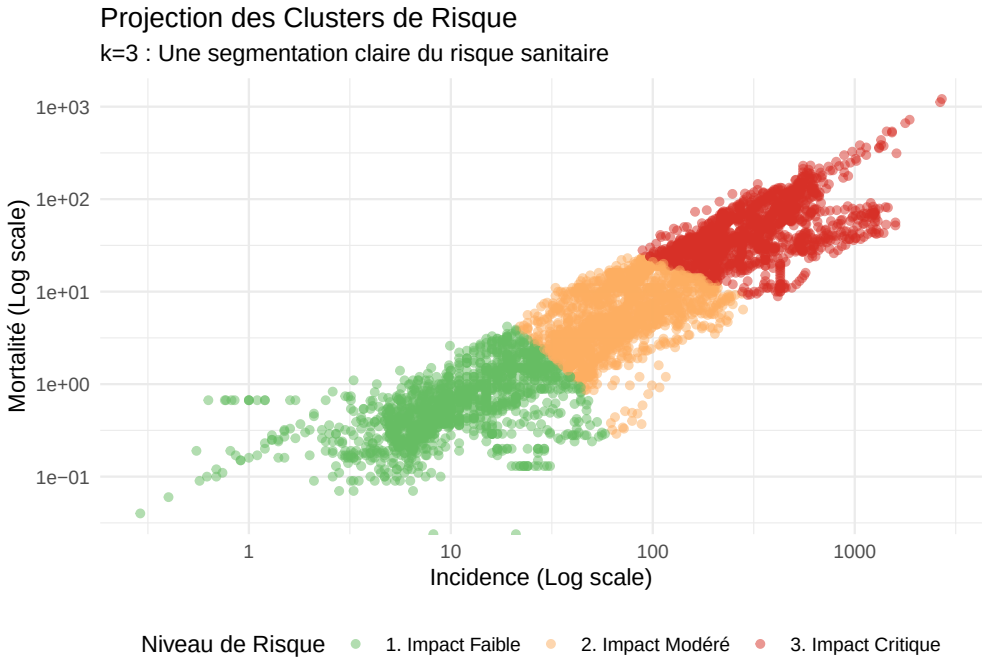
L'analyse des centroïdes révèle une hiérarchisation sanitaire nette. Le cluster **Impact Faible** (`n=1 416`), représentatif des standards occidentaux (Europe, Amérique du Nord), affiche une incidence marginale (14 cas/100k) et une mortalité résiduelle (<1 décès/100k). Le faible ratio de létalité (~6 %) témoigne d'une prise en charge thérapeutique efficace où la maladie est rarement fatale.

Le cluster **Impact Modéré** (`n=1 570`) regroupe des pays en transition (Maghreb, Amérique Latine) confrontés à une circulation active du bacille (79 cas/100k). Toutefois, la mortalité contenue (7 décès/100k) indique que si le contrôle de la transmission reste un défi, les systèmes de santé parviennent à traiter la majorité des patients diagnostiqués.

Enfin, le cluster **Impact Critique** (`n=1 546`), centré sur l'Afrique subsaharienne, concentre la charge mondiale avec une incidence massive (374 cas/100k) et une mortalité très élevée (57 décès/100k). Le taux de létalité y atteint un niveau alarmant de 15,3 %, révélant des défaillances systémiques graves (retards de diagnostic, résistances) : dans cette zone, la tuberculose ne se contente pas de circuler, elle tue massivement.

4.3 Visualisation de la Segmentation

La projection des clusters sur le plan bivarié illustre la logique de séparation opérée par l'algorithme.



Le graphique confirme que le score de 83,9 % d'inertie expliquée se traduit visuellement par des frontières nettes entre les groupes, avec très peu de chevauchement. La segmentation en "feux tricolores" est donc statistiquement robuste et opérationnellement pertinente.

4.4 Préparation pour l'Application

Nous sauvegardons le jeu de données final enrichi des labels, qui servira de socle à l'application R Shiny.

```
save(tb_clustered, file = "data/TB_analysis_ready.RData")
```

5 Application R Shiny

L'étape finale de ce projet consiste à transformer les résultats de la segmentation (K-Means) en un outil de pilotage interactif. Nous avons développé une application web via le framework R Shiny, permettant aux décideurs de santé publique d'explorer les données, de visualiser les disparités géographiques et de monitorer l'évolution des profils de risque en temps réel.

5.1 Architecture technique : Structure UI/Server et flux de données réactif

Fondée sur une architecture client-serveur réactive, l'application mobilise un écosystème de librairies R spécialisées pour garantir fluidité et interactivité. L'interface utilisateur, structurée de manière modulaire via `shinydashboard`, articule la cartographie vectorielle de `leaflet` avec les graphiques dynamiques du couple `ggplot2` / `plotly` (survol, zoom). En amont, la manipulation des données et le filtrage en temps réel reposent sur la performance des packages `dplyr` et `tidyr`, assurant une réactivité immédiate aux interactions de l'utilisateur.

5.1.1 Flux de Données Réactif

Le coeur de l'application réside dans son graphe de dépendance réactif qui, contrairement à un script statique, optimise les ressources en ne recalculant les éléments qu'à la demande. Le flux suit une logique séquentielle : toute interaction sur un **Input** (sélection d'une année ou d'un pays) déclenche une **Expression Réactive** chargée de filtrer le jeu de données `tb_clustered`. Ce nouveau sous-ensemble propage alors instantanément la mise à jour vers les **Outputs** (cartes, tableaux et courbes) sans nécessiter de rechargement de la page.

5.2 Fonctionnalités décisionnelles :

L'interface a été conçue pour répondre à trois besoins analytiques majeurs : la vision globale, le suivi temporel et l'analyse comparative.

5.2.1 Cartographie Interactive des Risques (Vision Globale)

La page d'accueil déploie une carte mondiale interactive (`leaflet`) où chaque pays est coloré selon son cluster d'appartenance : **Vert** (Impact Faible), **Orange** (Modéré) ou **Rouge** (Critique). Cette visualisation offre une lecture immédiate de la géographie sanitaire, permettant d'identifier les foyers épidémiques structurels (telle la ceinture rouge subsaharienne) tout en repérant rapidement les anomalies locales (pays critiques isolés au sein d'une zone préservée).

5.2.2 Monitorage Temporel (Analyse Dynamique)

Un curseur temporel (Slider Input) permet de naviguer sur la période 2000-2024. L'animation de ce curseur permet de visualiser les transitions de clusters (trajectoires). On peut ainsi observer les succès de certains pays passant du statut "Critique" à "Modéré" suite à l'amélioration de leur système de soins, ou inversement, les dégradations liées à des conflits ou crises sanitaires.

5.2.3 Analyse Comparative

Un module dédié permet de sélectionner un pays spécifique (ex: Nigeria) pour générer son Bulletin de Santé complet. Celui-ci articule l'affichage des **KPIs clés** (valeurs brutes d'incidence, mortalité, cluster) avec une analyse de **positionnement relatif**. En confrontant la trajectoire du pays sélectionné aux moyennes régionales et mondiales, ce graphique permet d'objectiver sa performance réelle et de déterminer s'il sous-performe par rapport à son voisinage direct, indépendamment de la tendance globale.

5.3 Implémentation et logique applicative

L'application a été développée selon une architecture modulaire, séparant distinctement l'interface utilisateur (Frontend) de la logique de calcul (Backend), conformément au paradigme du framework Shiny.

5.3.1 Stack Technologique et Dépendances

Le développement repose sur une stack technique optimisée pour l'interactivité. L'orchestration de l'interface est assurée par le couple `shiny` et `shinydashboard`, garantissant une structure modulaire et responsive. La couche géospatiale combine la précision vectorielle de `sf` à la fluidité de rendu de `leaflet`, tandis que la visualisation des résultats exploite les capacités dynamiques de `plotly` (pour les graphiques interactifs) et la puissance de tri de `DT` (pour les tableaux). Enfin, `dplyr` agit comme moteur de calcul en temps réel, assurant le filtrage réactif et l'agrégation instantanée des données en arrière-plan.

5.3.2 Architecture de l'Interface Utilisateur (UI)

L'interface guide l'utilisateur du général au particulier via une structure en trois volets. Le **Dashboard**, véritable cœur décisionnel, orchestre via une grille fluide l'affichage de KPIs dynamiques, d'une double visualisation interactive (Carte/Nuage de points) et d'un module de comparaison des trajectoires. Il est complété par un **Explorateur de Données** pour l'accès aux chiffres bruts et une section **Méthodologie** garantissant l'auto-portance de l'outil. Transversalement, la navigation latérale assure le pilotage global des graphiques via un filtrage régional et un contrôle temporel animé (2000-2024).

5.3.3 Logique Serveur et Réactivité

Le script serveur orchestre l'intelligence applicative via deux leviers. D'une part, le **filtrage réactif** optimise la performance : contrairement à une approche statique, les données ne sont chargées qu'une fois puis segmentées dynamiquement par une expression (`filtered_data`) qui joint instantanément le sous-ensemble aux polygones géographiques (`world_map`) à chaque modification des entrées.

D'autre part, la gestion d'état centralisée permet un **Cross-Filtering** avancé. Une variable réactive (`reactiveVal`), stockant l'identifiant du pays actif, est mise à jour indifféremment par trois interactions distinctes : un clic sur la carte, le nuage de points ou le graphique de densité. Cette interconnexion totale assure une exploration fluide, où l'investigation d'un point aberrant sur un graphique projette immédiatement l'information sur l'ensemble des autres vues.

5.3.4 Rendu Conditionnel et Comparaison

Le graphique de tendance (`trend_plot`) transforme la simple série temporelle en un outil d'analyse comparative en construisant dynamiquement trois courbes à la volée : **la trajectoire du pays sélectionné** (mise en évidence), confrontée à la **moyenne de sa région** (calculée en temps réel) et à la **référence mondiale** fixe. Cette logique de calcul à la demande permet ainsi de situer instantanément la performance de n'importe quel territoire vis-à-vis de son contexte géographique immédiat.

6 Exploitation et Analyse des Résultats

Au-delà de l'implémentation technique, l'application R Shiny permet d'objectiver les dynamiques épidémiologiques mondiales. L'exploration interactive des données (2000-2024) met en lumière trois niveaux de lecture.

6.1 Analyse Macroscopique : La fracture Nord-Sud

La cartographie interactive confirme que la segmentation ternaire obéit à une logique géopolitique structurante. Le **Cluster 1 (Faible Impact - Vert)** se superpose quasi-intégralement aux pays de l'OCDE, caractérisant une maladie devenue résiduelle. Il se distingue du **Cluster 2 (Intermédiaire - Orange)**, véritable zone tampon hétérogène (Amérique Latine, Europe de l'Est) où les infrastructures de santé font face à des défis de résistance. Enfin, le **Cluster 3 (Critique - Rouge)** dessine une ceinture épidémique continue en Afrique Subsaharienne et sur certains foyers asiatiques, dont la superposition avec les zones de forte prévalence du VIH et d'instabilité politique apparaît frappante.

6.2 Dynamiques Régionales et Temporelles

L'outil de monitorage (2000-2024) objective une baisse mondiale de l'incidence à géométrie variable. Tandis que l'**Europe** et les **Amériques** affichent une stagnation ou une décroissance marginale caractéristique d'une épidémie maîtrisée, l'**Afrique** se distingue par la chute la plus rapide en valeur absolue depuis 2010, témoignant du succès des campagnes contre la co-infection TB-VIH. À l'opposé, l'**Asie du Sud-Est** manifeste une inertie inquiétante et demeure, par la densité démographique de l'Inde et de l'Indonésie, le principal réservoir volumique mondial de nouveaux cas.

6.3 Cas d'usage : la France

Pour illustrer la puissance analytique de l'outil, nous prenons le cas de la France. L'analyse du cas français illustre la puissance de l'outil pour situer un territoire. Solidement ancrée dans le **Cluster 1 (Faible Impact)** avec une incidence de 8 cas/100k en 2024, la France affiche une performance remarquable sur trois échelles : elle se situe un facteur 15 sous la moyenne mondiale et surperforme nettement la moyenne européenne (~24 cas/100k), cette dernière étant grevée par les pays de l'Est du Cluster 2. La confrontation avec un représentant du Cluster 3 comme l'Afrique du Sud (> 389 cas/100k) objective une fracture sanitaire vertigineuse : maladie du passé pour l'Hexagone, la tuberculose demeure une urgence vitale ailleurs. Ce diagnostic valide l'efficacité de la stratégie nationale tout en rappelant l'impératif de vigilance face aux risques de réintroduction depuis les zones critiques (Orange et Rouge).

7 Conclusion et Perspectives

Ce projet s'est attaché à transformer une base de données brute et complexe, issue du rapport mondial de l'OMS, en un outil d'aide à la décision sanitaire opérationnel. En combinant une approche statistique rigoureuse (analyse exploratoire, réduction de dimension) et une modélisation non-supervisée (Clustering K-Means), nous avons pu objectiver les disparités mondiales face à l'épidémie de tuberculose.

7.1 Synthèse des résultats

L'analyse de la période 2000-2024 valide trois enseignements majeurs. D'abord, la **pertinence d'une segmentation ternaire** ($k = 3$) qui, forte d'une robustesse statistique de 83,9 %, dépasse le simple clivage Nord-Sud pour cartographier le risque selon une gradation opérationnelle (Faible, Modéré, Critique). Ensuite, la **polarisation de l'épidémie** : le cluster Critique concentre une létalité disproportionnée (> 15 %), dictant un ciblage prioritaire des efforts sur l'Afrique subsaharienne. Enfin, la valeur ajoutée du **monitorage dynamique** : l'application R Shiny a permis d'objectiver la mobilité des trajectoires, identifiant les pays en transition pour fournir des signaux d'alerte précoce ou valider l'efficacité des politiques publiques.

7.2 Limites méthodologiques

Dans une démarche critique, trois limites méthodologiques doivent être soulignées. Premièrement, le **biais déclaratif** persiste malgré l'usage des estimations OMS (`e_`) : les données restent tributaires de la qualité de la surveillance nationale, induisant un paradoxe où l'amélioration du diagnostic peut être confondue avec une

dégradation épidémique (hausse mécanique de l'incidence détectée). Deuxièmement, la **parcimonie du modèle**, restreinte à deux variables pour garantir la robustesse, confine l'étude à un rôle descriptif qui occulte les déterminants causaux (pauvreté, VIH). Enfin, la **suppression des données manquantes** (15 % des observations), impérative pour la stabilité du K-Means, rend de facto le modèle inopérant pour les micro-états insulaires exclus.

7.3 Perspectives d'évolution

Pour enrichir cet outil de pilotage, trois axes de développement majeurs se dessinent. D'abord, le passage vers une **modélisation explicative** : l'intégration de variables socio-économiques (PIB, Gini) via une ACP permettrait d'identifier les déterminants structurels du cluster Critique. Ensuite, le déploiement d'une **approche prédictive** (via ARIMA ou Prophet) transformerait ce tableau de bord analytique en outil prospectif, capable d'évaluer l'atteinte des objectifs onusiens à l'horizon 2030. Enfin, l'adoption d'une **granularité infra-nationale** s'avérerait pertinente pour les grands états fédéraux (Brésil, Inde) où la moyenne nationale masque de fortes disparités. En somme, ce projet offre une boussole efficace et constitue la première pierre d'une épidémiologie de précision guidée par la donnée.

8 Déclaration d'Intégrité et Usage de l'IA

Conformément aux consignes académiques relatives au plagiat et à l'utilisation des assistants numériques, cette section explicite le cadre de réalisation de ce projet.

8.1 Originalité de la démarche

Le jeu de données utilisé (*Global Tuberculosis Report*) est public et largement étudié. Cependant, l'approche développée dans ce projet est originale et personnelle.

Disposant d'un **profil d'ingénieur logiciel**, j'ai fait le choix stratégique de concentrer mon effort technique sur l'architecture et l'interactivité de l'application **R Shiny** (Section 5), afin de produire un outil de qualité professionnelle. Cette notice technique assure la couverture rigoureuse de la partie Data Science, justifiant les choix mathématiques implémentés dans l'application.

8.2 Usage des outils d'IA Générative

L'utilisation d'outils d'intelligence artificielle générative s'est inscrite dans une démarche d'assistance ponctuelle et rigoureusement contrôlée. Sur le volet **rédactionnel**, l'IA a contribué à l'optimisation syntaxique et à la fluidité des transitions, le raisonnement et les interprétations demeurant strictement personnels. Sur le plan **technique**, elle a servi d'outil de diagnostic pour le débogage de l'application R Shiny (gestion de la réactivité, conflits). L'intégralité du code a été vérifiée et maîtrisée : aucune partie de l'analyse n'a été déléguée sans supervision humaine.

9 Bibliographie

9.1 Rapports et Encyclopédies

- [1] Organisation Mondiale de la Santé (OMS). (2024). Global Tuberculosis Report 2024. Disponible sur : https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024
- [2] Wikipédia. (s.d.). Tuberculose. Disponible sur : https://fr.wikipedia.org/wiki/Tuberculose

9.2 Supports de Cours - Master 2 ISF (2025-2026)

- [3] Ochoa, J. (2025-2026). *Les algorithmes non supervisés.* Support de cours : Machine Learning. Université Paris-Dauphine - PSL.
- [4] Bertrand, P. (2025-2026). *K-Means.* Support de cours : Apprentissage non supervisé et clustering. Université Paris-Dauphine - PSL.
- [5] Guibert, Q. (2025-2026). *Data Visualisation.* Support de cours : Visualisation des données avec R. Université Paris-Dauphine - PSL.