Université de Haute-Alsace
Faculté des Sciences et Techniques
Département de Mathématiques

Rapport de Projet de Master 1

Analyse de l'évolution du travail des enfants à partir de séries temporelles

Rédigé par :
Hocine Rayane ARHAB
Zine Elabidine AIDOUD

Encadrante universitaire :
Suzy MADDAH

Année universitaire 2024–2025

Résumé

Cette étude analyse l'évolution du travail des enfants dans le monde entre 2010 et 2023 à partir de données harmonisées de l'Organisation Internationale du Travail. En examinant plus de deux millions d'enregistrements par pays, tranche d'âge, sexe et statut scolaire, nous mettons en évidence les grandes tendances et ruptures dans la dynamique du travail des enfants. Les résultats révèlent d'importants pics en 2015 et 2022, touchant toutes les catégories, avec une prédominance chez les garçons et les adolescents de 15 à 17 ans. L'analyse montre que la majorité des enfants travailleurs restent scolarisés, mais que la proportion de ceux qui quittent l'école augmente fortement lors des années de crise. Ces constats servent de base à des recommandations ciblées et soulignent la nécessité d'un suivi continu pour mieux protéger les enfants vulnérables.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction générale

Le travail des enfants demeure un enjeu mondial majeur, compromettant les droits et le développement des plus jeunes. Malgré les efforts internationaux et l'adoption d'objectifs de développement durable, des millions d'enfants continuent d'exercer une activité qui nuit à leur éducation et à leur bien-être. Comprendre l'ampleur et l'évolution de ce phénomène est essentiel pour concevoir des politiques efficaces.

Ce projet s'appuie sur un jeu de données exhaustif de l'Organisation Internationale du Travail couvrant la période 2010-2023 afin d'explorer les tendances de fond et les changements soudains du travail des enfants. L'analyse croisée par sexe, tranche d'âge et statut scolaire permet d'identifier non seulement les évolutions à long terme, mais aussi les ruptures liées aux crises ou aux changements de politique. Les visualisations proposées offrent une lecture fine des interactions entre travail des enfants, scolarisation et facteurs démographiques, mettant en lumière à la fois les progrès réalisés et les vulnérabilités persistantes. Ce travail apporte ainsi des éléments concrets pour l'élaboration de réponses adaptées et souligne les efforts encore nécessaires pour éradiquer le travail des enfants.

Chapitre 2

Contexte du travail des enfants

2.1 Définition

Le travail des enfants, tel que défini par l'Organisation internationale du travail (OIT), à savoir : « L'ensemble des activités qui privent les enfants de leur enfance, de leur potentiel et de leur dignité, et nuisent à leur scolarité, santé, développement physique et mental.»
Il s'agit d'un travail qui :

- est mentalement, physiquement, socialement ou moralement dangereux et nocif pour les enfants ;
- interfère avec leur scolarité, en les privant d'aller à l'école ou en les obligeant à quitter l'école prématurément, ou à combiner école et travail excessif.

Toutes les tâches réalisées par des enfants ne sont pas considérées comme du travail des enfants à éliminer : Seul le travail qui compromet la santé, l'éducation ou le développement de l'enfant est visé par les conventions internationales. [1].

2.2 Statistiques mondiales (basées sur les données récentes)

- En 2020, **160 millions d'enfants** (soit 1 sur 10 dans le monde) étaient engagés dans le travail des enfants, selon l'OIT et l'UNICEF.
- Le phénomène est plus répandu chez les garçons que chez les filles à tous les âges.
- Plus d'un tiers des enfants astreints au travail ne sont pas scolarisés.

— L'Afrique subsaharienne concentre la plus forte proportion et le plus grand nombre absolu d'enfants travailleurs, suivie par l'Asie et le Pacifique.
— La tendance mondiale, après avoir diminué entre 2000 et 2016, est repartie à la hausse depuis 2020, notamment sous l'effet de la pandémie de Covid-19 et de l'aggravation de la pauvreté.

2.3 Facteurs influençant le travail des enfants

On peut regrouper les cause en trois grandes catégories : socio-économiques, éducatives et culturelles/structurelles.

2.3.1 Facteurs socio-économiques

— **Pauvreté** : principal moteur du travail des enfants, exacerbée par les crises économiques, les conflits ou encore la pandémie.
— **Chômage ou précarité des parents** : les familles en difficulté sont souvent contraintes de faire travailler leurs enfants pour subvenir aux besoins essentiels.
— **Absence ou faiblesse de la protection sociale** : manque d'allocations familiales ou de filets sociaux pour compenser la perte de revenus.

2.3.2 Facteurs éducatifs

— **Difficulté d'accès à l'école** : éloignement géographique, frais de scolarité, infrastructures insuffisantes ou manque d'enseignants.
— **Qualité de l'éducation** : écoles peu attractives, violences scolaires, manque de perspectives professionnelles.
— **Combinaison travail/école** : certains enfants tentent de concilier les deux, souvent au détriment de leur réussite scolaire.

2.3.3 Facteurs culturels et structurels

— **Normes sociales** : dans certaines cultures, le travail des enfants est perçu comme normal, voire valorisé.
— **Genre** : les garçons sont souvent plus visibles dans le travail formel, tandis que les filles sont surreprésentées dans les tâches domestiques invisibles.

— **Secteurs d'activité** : l'agriculture familiale et le secteur informel concentrent la majorité des enfants travailleurs.

Chapitre 3

Fondements des séries temporelles

3.1 Définition

Une série temporelle est une suite d'observations répétées d'un même phénomène à des dates différentes. Les dates sont souvent équidistantes (séries journalières, mensuelles, trimestrielles ou annuelles).

Définition mathématique : Une série temporelle est une suite de variables aléatoires $(X_t)_{t\in T}$, où T est un ensemble d'indices temporels (souvent $\mathbb{N}$ ou $\mathbb{Z}$). En pratique, on observe une réalisation $(x_t)_{1\leq t\leq n}$.

On représente une série temporelle $(x_t)_{1\leq t\leq T}$ (où t désigne le numéro de l'observation) à l'aide d'un graphique avec en abscisse les dates et en ordonnée les valeurs observées. Ou bien : l'évolution au cours du temps d'un phénomène,

dans le but de décrire, expliquer puis prévoir ce phénomène dans le futur. On dispose ainsi d'observations à des dates différentes, c'est-à-dire d'une suite de valeurs numériques indexées par le temps. [2].

3.2 Les principales composantes d'une série temporelle

— **Tendance** : Elle représente l'évolution globale de la variable étudiée sur le long terme, montrant une croissance, une décroissance ou une stabilité. Elle reflète l'évolution globale sans tenir compte des variations de court terme.

— **Saisonnalité** : C'est un cycle qui se répète régulièrement dans le temps, souvent lié à des périodes fixes comme les saisons, les mois, les jours de la semaine, etc. La saisonnalité correspond aux fluctuations périodiques qui se reproduisent à intervalles réguliers et qui sont prévisibles.
— **Résidus** : Ce sont les variations aléatoires ou imprévisibles qui ne peuvent pas être expliquées par la tendance ou la saisonnalité. Elles correspondent aux phénomènes accidentels ou au bruit de la série.
— **Cycle** : Ce cycle correspond à des phénomènes répétitifs mais dont la période est plus longue, variable ou inconnue. Cette composante est moins systématique que la saisonnalité et peut être incluse dans certaines décompositions.

Afin d'analyser ces différentes composantes, plusieurs méthodes de décomposition peuvent être utilisées. L'une des méthodes les plus utilisées aujourd'hui est la décomposition STL, que nous présentons ci-dessous. [3].

3.2.1 Décomposition STL (Seasonal-Trend decomposition using Loess)

Cette méthode sépare la série en tendance, saisonnalité et résidus, ce qui permet d'analyser et de prévoir chaque composante séparément. Elle repose sur l'utilisation de modèles de régression ajustés localement (LOESS) pour effectuer un lissage. L'algorithme STL fonctionne en deux boucles : une boucle interne qui alterne le calcul et le retrait des composantes saisonnières et de tendance, et une boucle externe qui réduit l'influence des valeurs aberrantes. [4].

$$y_i = s_i + t_i + r_i$$

où :
— y_i est la valeur de la série temporelle au point i,
— s_i est la composante saisonnière au point i,
— t_i est la composante de tendance au point i,
— r_i est la composante résiduelle (ou bruit) au point i.

Décomposition mathématique :

$$X_t = T_t + S_t + R_t$$

où T_t est la tendance, S_t la saisonnalité, R_t le résidu.

3.3 Propriétés mathématiques importantes

3.3.1 Stationnarité

Une série (X_t) est dite stationnaire si :
— $\mathbb{E}[X_t] = \mu$ (espérance constante)
— $\text{Var}(X_t) = \sigma^2$ (variance constante)
— $\text{Cov}(X_t, X_{t+h})$ ne dépend que de h (l'autocovariance ne dépend que du décalage)

3.3.2 Autocorrélation

La fonction d'autocorrélation (ACF) mesure la dépendance entre les valeurs de la série à différents décalages :

$$\rho(h) = \frac{\text{Cov}(X_t, X_{t+h})}{\text{Var}(X_t)}$$

La fonction d'autocorrélation partielle (PACF) permet d'isoler l'effet direct d'un décalage donné.

3.4 Importance de l'analyse temporelle pour comprendre l'évolution du travail des enfants

L'analyse temporelle est importante pour assimiler la dynamique du travail des enfants à l'échelle mondiale ou nationale. Grâce à l'étude des séries temporelles, il est possible de :

— Détecter les tendances de fond : On peut identifier si le phénomène du travail des enfants est en augmentation, en diminution ou stable sur plusieurs années.
— Repérer les ruptures ou les anomalies : Les analyses temporelles révèlent des années atypiques (comme une forte baisse ou un pic soudain), souvent liées à des changements de méthode de collecte, à des crises, ou à des événements socio-économiques majeurs.
— Comprendre l'absence ou la présence de saisonnalité : Si les données sont suffisamment fines, on peut observer des variations saisonnières, ce qui permet d'adapter les stratégies de prévention.

— Appuyer la prévision : La modélisation des séries temporelles permet d'anticiper l'évolution future du travail des enfants, d'identifier les périodes à risque et d'orienter les ressources vers les moments et les lieux les plus critiques.

3.5 Modèles statistiques de prévision des séries temporelles

Les modèles statistiques permettent de prédire les valeurs futures en exploitant les dépendances temporelles de la série. Parmi les approches traditionnelles, [5]. on retrouve :

3.5.1 ARIMA (AutoRegressive Integrated Moving Average)

— Ce modèle combine des composantes autorégressives, de moyenne mobile et de différenciation pour modéliser la tendance et l'autocorrélation dans les données. Il est particulièrement adapté aux séries sans saisonnalité marquée et permet de produire des prévisions fiables à court terme.

Le modèle ARIMA est défini par trois paramètres :
— p : ordre autorégressif (AR)
— d : ordre de différenciation
— q : ordre de la moyenne mobile (MA)

Formulation générale

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{3.1}$$

où :
— Y_t : valeur de la série à l'instant t
— ϕ_i : coefficients autorégressifs
— θ_j : coefficients de moyenne mobile
— $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$: bruit blanc

Exemple ARIMA(1,1,1)

Après différenciation d'ordre 1 ($Y'_t = Y_t - Y_{t-1}$) :

$$Y'_t = c + \phi_1 Y'_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t \tag{3.2}$$

Identification des paramètres

— Analyse ACF/PACF pour p et q
— Test de Dickey-Fuller augmenté (ADF) pour d
— Critère d'information AIC/BIC pour sélection modèle

3.5.2 Lissage exponentiel

— Cette méthode attribue plus de poids aux observations récentes pour estimer la tendance. Elle est simple à mettre en œuvre et efficace pour les séries présentant peu de variations structurelles.

Formule de base

$$S_t = \alpha Y_t + (1-\alpha) S_{t-1}, \quad 0 < \alpha < 1 \tag{3.3}$$

où :
— S_t : valeur lissée
— α : facteur de lissage
— Interprétation :
  — $\alpha \to 1$: poids fort aux dernières observations
  — $\alpha \to 0$: lissage important

Méthode de Holt (tendance)

Système d'équations pour données avec tendance :

$$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1}) \tag{3.4}$$
$$T_t = \beta(S_t - S_{t-1}) + (1-\beta) T_{t-1} \tag{3.5}$$

avec :
— T_t : estimation de la tendance
— $\beta \in [0,1]$: paramètre de lissage de tendance

3.5.3 Avantages comparatifs

Caractéristique	ARIMA	Lissage exponentiel
Données stationnaires	Requis	Non requis
Capture tendance	Oui (via différenciation)	Oui
Capture saisonnalité	SARIMA	Holt-Winters
Complexité mathématique	Élevée	Modérée

TABLE 3.1 – Comparaison des approches de prévision

Chapitre 4

Introduction à l'IA dans la prévision des séries temporelles

L'intelligence artificielle a profondément changé notre manière de prévoir l'évolution des données dans le temps, surtout dans les cas où les méthodes statistiques classiques ne suffisent plus à en saisir toute la complexité. Grâce à l'apprentissage profond, les modèles d'IA peuvent traiter d'importantes quantités de données, repérer des liens difficiles à percevoir, faire émerger des tendances invisibles à l'œil nu, ou encore modéliser des phénomènes saisonniers complexes. Cela en fait des outils particulièrement utiles pour aborder des enjeux sociaux sensibles, comme celui de l'évolution du travail des enfants.

4.1 Modèles d'IA utilisés en prévision temporelle

4.1.1 Réseaux neuronaux récurrents (RNN) et LSTM

Les réseaux neuronaux récurrents (RNN) et leurs variantes, notamment les réseaux LSTM (Long Short-Term Memory), sont spécialement conçus pour traiter des données séquentielles. Leur architecture leur permet de mémoriser des informations sur de longues périodes, ce qui est important pour la prévision de phénomènes temporels qui dépendent de leur propre passé. Les LSTM, en particulier, sont reconnus pour leur efficacité à modéliser les dépendances à long terme dans les séries temporelles, ce qui les distingue des modèles traditionnels comme ARIMA

4.1.2 Facebook Prophet

Facebook Prophet est un modèle additif développé par Meta, conçu pour être facile à utiliser tout en restant puissant ; il est performant pour les séries avec rupture ou saisonnalité. Il décompose la série temporelle en trois composantes principales : la tendance, la saisonnalité et les effets des jours fériés. Prophet s'appuie sur un modèle additif généralisé (GAM) pour intégrer ces composantes, ce qui le rend particulièrement adapté aux séries présentant des ruptures de tendance ou des saisonnalités multiples.

4.2 Préparation et entraînement des modèles IA

Le bon fonctionnement d'un modèle d'intelligence artificielle repose en grande partie sur la qualité des données qu'on lui fournit. Avant d'entraîner un modèle, il est essentiel de passer par plusieurs étapes clés : collecter les données, les nettoyer, les mettre à l'échelle si nécessaire, et les organiser sous forme de séquences temporelles. Par exemple, pour entraîner un modèle LSTM, on transforme souvent les données en suites d'entrées et de sorties. En revanche, Prophet fonctionne avec une structure beaucoup plus simple : une colonne pour les dates et une autre pour les valeurs à prévoir.

Une fois les données prêtes, l'entraînement du modèle consiste à ajuster ses paramètres (les hyperparamètres), valider ses performances à travers une validation croisée adaptée au temps, et mesurer sa précision à l'aide d'indicateurs comme le RMSE ou le MAPE. Bien que les modèles d'IA soient très puissants, ils ont généralement besoin d'un volume important de données pour éviter de trop s'adapter à l'historique (ce qu'on appelle le surajustement).

4.3 Clustering

4.3.1 Définition

Le clustering (ou regroupement) est la tâche consistant à diviser une population ou un ensemble de points de données en plusieurs sous-groupes(Clusters), de sorte que les points appartenant à un même sous-groupe soient similaires entre eux, et différents des points des autres sous-groupes. Il s'agit essentiellement d'un regroupement d'objets basé sur leur similarité ou dissimilarité.

Le clustering aide à comprendre les groupements naturels présents dans un jeu de données. Son objectif est de diviser les données en ensembles co-

hérents et significatifs. La qualité du regroupement dépend des méthodes utilisées et de la capacité à identifier des motifs cachés. Le principal avantage du clustering par rapport à la classification est qu'il peut s'adapter aux changements et permet d'identifier les caractéristiques utiles qui différencient les groupes entre eux. [7]

4.3.2 Types de Clustering

On retrouve plusieurs types de clustering qui différent selon les données, les méthodes utilisées et les objectifs a atteindre [8] :

Type de clustering	**Description**
Clustering basé sur les partitions	Divise les données en groupes distincts (clusters) non chevauchants selon un critère d'optimisation (souvent la minimisation de la distance intra-cluster). L'algorithme le plus connu est K-means, qui attribue chaque point à un cluster en fonction de sa proximité avec un centroïde. Le nombre de clusters doit être défini à l'avance.
Clustering hiérarchique	Construit une arborescence (dendrogramme) de clusters, soit en fusionnant progressivement les points (agglomératif), soit en divisant successivement les groupes (divisif). Les clusters sont imbriqués à différents niveaux et leur nombre n'a pas besoin d'être fixé à l'avance. On visualise ainsi la structure multi-niveaux des données.
Clustering basé sur la densité	Les clusters sont formés autour des zones de forte densité de points, séparées par des régions de faible densité. Cette méthode permet de détecter des clusters de forme arbitraire et d'identifier les points isolés comme du bruit. Elle est particulièrement adaptée aux jeux de données contenant des outliers ou des clusters non sphériques (ex : DBSCAN).
Clustering basé sur les séries temporelles	Spécialisé pour les données chronologiques, ce clustering regroupe les séries présentant des motifs, tendances ou comportements temporels similaires. Il utilise des mesures adaptées (comme la distance DTW) pour comparer la forme des séries, indépendamment de leur phase ou de leur amplitude. Il permet de révéler des dynamiques cachées dans les données temporelles.

TABLE 4.1 – Principaux types de clustering et leurs descriptions

4.3.3 Clustering basé sur les séries temporelles

Le **clustering de séries temporelles** est une méthode non supervisée qui consiste à regrouper des séries temporelles présentant des comportements ou motifs similaires au fil du temps. Contrairement au clustering classique, il prend en compte la dimension temporelle des données, ce qui est essentiel pour analyser des phénomènes évolutifs comme le travail des enfants dans le monde. [9]

Importance du clustering temporel

Cette technique permet de :

- Identifier des profils d'évolution communs entre différents pays ou régions, par exemple des tendances similaires dans la réduction ou l'augmentation du travail des enfants.
- Simplifier l'analyse en regroupant des milliers de séries en quelques clusters représentatifs.
- Détecter des ruptures, des cycles ou des motifs saisonniers dans les données temporelles.
- Aider à orienter les politiques publiques en ciblant les groupes de pays présentant des dynamiques similaires.

Formulation mathématique

Soit un ensemble de séries temporelles $X = \{X_1, X_2, \ldots, X_n\}$, où chaque série $X_i = (x_{i1}, x_{i2}, \ldots, x_{iT})$ est une séquence temporelle de longueur T. Le clustering vise à partitionner X en k clusters $C_1, C_2, \ldots, C_k$ en minimisant la somme des distances entre chaque série et le centroïde de son cluster :

$$\min_{C_1,\ldots,C_k} \sum_{j=1}^{k} \sum_{X_i \in C_j} D(X_i, \mu_j)$$

où D est une mesure de distance adaptée aux séries temporelles (par exemple, la distance Dynamic Time Warping) et μ_j est le centroïde du cluster C_j. [10]

Exemples d'algorithmes

- **K-means avec DTW** : adaptation de l'algorithme K-means utilisant la distance DTW, qui permet d'aligner les séries temporelles malgré des décalages ou variations de vitesse. Très utilisé pour regrouper des séries présentant des formes similaires.

— **Clustering hiérarchique** : construit une hiérarchie de clusters (dendrogramme) avec des distances adaptées (DTW ou autres). Utile pour explorer la structure multi-niveaux des données.
— **DBSCAN temporel** : méthode basée sur la densité, capable de détecter des clusters de forme arbitraire et d'ignorer le bruit, adaptée aux séries avec des motifs locaux.

Défis et bonnes pratiques

— **Choix de la distance** : la mesure DTW est privilégiée pour sa robustesse aux décalages temporels, mais son coût computationnel est élevé.
— **Normalisation** : les séries doivent être normalisées (moyenne/variance ou min-max) pour éviter que l'amplitude domine la similarité.
— **Gestion des données manquantes** : interpolation ou imputation nécessaire pour garantir la qualité du clustering.
— **Évaluation** : utilisation d'indices comme le score de silhouette pour valider la qualité des clusters.

4.4 Avantages et limites des modèles IA

4.4.1 Avantages

— **Capacité à modéliser des relations complexes** : Les modèles d'IA, en particulier les LSTM, sont capables de détecter des tendances, des saisonnalités et des anomalies qui échappent aux modèles traditionnels.
— **Adaptabilité** : Les modèles IA peuvent être mis à jour régulièrement pour intégrer de nouvelles données et s'adapter à des environnements changeants.
— **Flexibilité** : Ils permettent d'intégrer facilement des variables externes et de traiter des séries multivariées.

4.4.2 Limites

— **Besoins en données** : Les LSTM nécessitent de grandes quantités de données pour éviter le surajustement, ce qui peut être problématique dans certains contextes.

— **Interprétabilité réduite** : Les modèles d'IA, en particulier les réseaux de neurones, sont souvent considérés comme des «boîtes noires», ce qui rend difficile l'interprétation des résultats.
— **Coût computationnel** : L'entraînement des modèles d'IA est généralement plus coûteux en temps et en ressources que celui des modèles statistiques classiques.

4.5 Comparaison avec les modèles statistiques classiques

Des études comparatives montrent que les LSTM obtiennent souvent de meilleures performances en termes de précision sur des séries temporelles complexes, tandis que Prophet offre un bon compromis entre facilité d'utilisation et interprétabilité. Les modèles statistiques classiques restent pertinents pour des séries simples ou en cas de données limitées.

Chapitre 5

Base de données temporelles et InfluxDB

5.1 Définition d'une base de données de séries temporelles

Une base de données de séries temporelles (TSDB) est un système spécialisé conçu pour les ensembles de données axés sur une période donnée. Elle offre un stockage optimisé pour l'étude des métriques à travers le temps. [6]

5.2 Caractéristiques et avantages d'InfluxDB

InfluxDB est un moteur de base de données orienté séries temporelles, conçu pour optimiser l'ingestion, le stockage et l'interrogation de données horodatées. Initialement doté d'un langage de requête similaire au SQL, il a récemment introduit Flux, un langage de script fonctionnel plus expressif, destiné à des analyses temporelles avancées. [11]

Ses atouts principaux sont :

- **Performances**
  - Débit d'écriture élevé : Jusqu'à des millions de points de données par seconde, idéal pour des données annuelles massives.
  - Requêtes rapides : Optimisé pour les analyses temporelles (ex : `SELECT ... WHERE time > '2010-01-01'`).
- **Scalabilité**
  - Stockage évolutif : Gère des téraoctets de données grâce à une compression jusqu'à 4,5x.

  - Architecture modulaire : Séparation entre l'ingestion des données et les requêtes pour une scalabilité horizontale.

- **Langage de requête**
  - Flux : Langage dédié aux séries temporelles, permettant des opérations complexes (ex : calcul de moyennes mobiles).
  - Compatibilité SQL : Depuis la version 3.0, pour une intégration aisée avec des outils comme Grafana.

5.3 Structure des données dans InfluxDB

InfluxDB organise les données autour de quatre concepts clés :

Composant	**Rôle**	**Exemple**
Measurement	Conteneur thématique	travail_enfants
Tags	Métadonnées indexées (filtrage rapide)	pays="RDC",sexe="Total", age_group="5-17"
Fields	Valeurs numériques ou booléennes	value=6373050 (nombre d'enfants travailleurs)
Timestamp	Horodatage (obligatoire)	2020-01-01T00 :00 :00Z

TABLE 5.1 – Exemple de structure générale dans InfluxDB

Chapitre 6

Mise en œuvre

6.1 Présentation des données

Le jeu de données utilisé dans ce projet porte sur l'évolution du travail des enfants dans le monde, en s'appuyant sur les principales sources internationales : l'Organisation internationale du Travail (OIT) et l'UNICEF, co-responsables du suivi de la cible 8.7 des Objectifs de développement durable. Les données sont issues de plus de 100 enquêtes nationales auprès des ménages, couvrant environ deux tiers de la population mondiale des enfants âgés de 5 à 17 ans. Elles sont régulièrement mises à jour et permettent de suivre les tendances mondiales et régionales depuis l'an 2000.

Le fichier comprend, pour chaque pays et chaque année, des informations sur :

— Le nombre d'enfants travailleurs (effectifs et taux), ventilé par tranche d'âge (5-11 ans, 12-14 ans, 15-17 ans), par sexe et par statut scolaire.

— Des indicateurs de fiabilité des données et des notes méthodologiques.

Les estimations sont principalement fondées sur l'extrapolation des résultats d'enquêtes nationales, avec une harmonisation méthodologique pour permettre la comparaison internationale. En 2020, selon ces données, 160 millions d'enfants étaient engagés dans le travail des enfants, soit près d'un sur dix dans le monde, avec une prévalence plus forte en Afrique subsaharienne et en Asie. Le jeu de données permet ainsi d'analyser l'évolution du phénomène, d'identifier les groupes les plus vulnérables et de suivre l'impact des politiques publiques.

6.2 Définition des variables

6.2.1 Structure des variables

Variable	Description	Exemple
ref_area.label	Pays/zone géographique	Afghanistan
source.label	Source primaire des données	HIES - Households Living Conditions Survey
indicator.label	Type d'indicateur	Children in employment
sex.label	Répartition par genre	Male/Female/Total
classif1.label	Tranche d'âge	5-11 ans / 12-14 ans / 15-17 ans
classif2.label	Statut scolaire	En scolarité/Hors scolarité/Total
time	Année de mesure	2018
obs_value	Valeur numérique	12300
obs_status.label	Fiabilité	Unreliable/Break in Series

TABLE 6.1 – Description des variables du jeu de données

6.2.2 Méthodologie de collecte

Les données sont collectées via :

— Enquêtes ménages standardisées (MICS, DHS)
— Recensements nationaux
— Systèmes administratifs nationaux
— Modélisation statistique pour les pays sans données directes

L'OIT applique un processus d'harmonisation incluant :

— Pondération des échantillons
— Ajustement des définitions nationales
— Imputation des valeurs manquantes
— Validation croisée inter-pays

6.2.3 Portée analytique

Ce jeu de données permet d'analyser :

— L'évolution temporelle du travail des enfants
— Les disparités genre/âge/régions
— Le lien travail des enfants-scolarisation

6.3 Chargement et exploration des données

6.3.1 Chargement

La première étape pour commencer notre prétraitement de données est le chargement du dataset avec les bibliothèques adéquates.

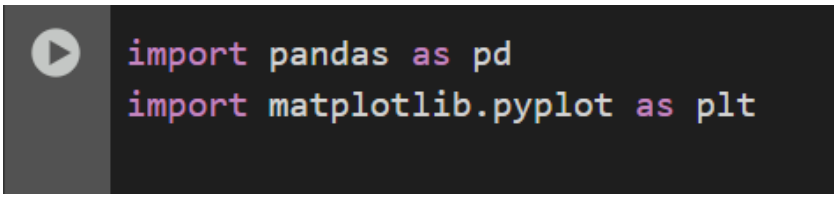


FIGURE 6.1 – Chargement des bibliothèques

Présentation de Pandas : Pandas est une bibliothèque Python qui permet de manipuler facilement des données à analyser [12]. :

- Manipulation de tableaux de données avec étiquettes de colonnes et lignes ;
- Ces tableaux sont appelés *dataframes* (similaires à R) ;
- Lecture/écriture de fichiers tabulés ;
- Tracé de graphiques avec `matplotlib` à partir de dataframes.

Présentation de matplotlib : Matplotlib est une bibliothèque de visualisation de données pour Python, largement utilisée dans les domaines scientifiques et de l'analyse de données. Elle permet de créer une grande variété de graphiques statiques, animés et interactifs avec un haut degré de personnalisation. [13].

FIGURE 6.2 – Import du dataset

6.3.2 Exploration du dataset

Après le chargement, on s'intéresse à la structure du dataset. On affiche les premières lignes et la taille de ce dernier :

```
print(data.head())
print(data.shape)
```

```
  ref_area.label                                source.label  \
0    Afghanistan  HIES - Households Living Conditions Survey
1    Afghanistan  HIES - Households Living Conditions Survey
2    Afghanistan  HIES - Households Living Conditions Survey
3    Afghanistan  HIES - Households Living Conditions Survey
4    Afghanistan  HIES - Households Living Conditions Survey

                                      indicator.label sex.label  \
0  Children in employment by sex, age and school ...     Total
1  Children in employment by sex, age and school ...     Total
2  Children in employment by sex, age and school ...     Total
3  Children in employment by sex, age and school ...     Total
4  Children in employment by sex, age and school ...     Total

                    classif1.label  \
0  Age (Child labour bands): '5-17
1  Age (Child labour bands): '5-17
2  Age (Child labour bands): '5-17
3  Age (Child labour bands): '5-17
4  Age (Child labour bands): '5-11

                                    classif2.label  time  obs_value  \
0                   Educational attendance: Total  2014   3261.161
1                Educational attendance: Attending  2014   1556.560
2             Educational attendance: Not attending  2014    185.107
3  Educational attendance: Not elsewhere classified  2014   1519.494
4                   Educational attendance: Total  2014   1254.495

  obs_status.label note_indicator.label  \
0              NaN                   NaN
1              NaN                   NaN
2              NaN                   NaN
3              NaN                   NaN
4              NaN                   NaN

                                  note_source.label
0  Repository: ILO-STATISTICS - Micro data proces...
1  Repository: ILO-STATISTICS - Micro data proces...
2  Repository: ILO-STATISTICS - Micro data proces...
3  Repository: ILO-STATISTICS - Micro data proces...
4  Repository: ILO-STATISTICS - Micro data proces...
(8745, 11)
```

FIGURE 6.3 – Aperçu initial du dataset

Informations sur les colonnes : On trouve les variables/colonnes mentionnées en haut.

6.4 Nettoyage des données

6.4.1 Filtrage des données

Dans cette étape, nous essayons d'optimiser la structure de notre jeu de données en :

— renommant les colonnes les plus pertinentes avec des noms plus significatifs et moins complexes ;
— supprimant les lignes sans valeur et les doublons ;
— transformant la colonne `année` au format `datetime`.

```
[5] print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8745 entries, 0 to 8744
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   ref_area.label        8745 non-null   object
 1   source.label          8745 non-null   object
 2   indicator.label       8745 non-null   object
 3   sex.label             8745 non-null   object
 4   classif1.label        8745 non-null   object
 5   classif2.label        8745 non-null   object
 6   time                  8745 non-null   int64
 7   obs_value             8153 non-null   float64
 8   obs_status.label      1894 non-null   object
 9   note_indicator.label  749 non-null    object
 10  note_source.label     7665 non-null   object
dtypes: float64(1), int64(1), object(9)
memory usage: 751.7+ KB
None
```

FIGURE 6.4 – Structure et types des colonnes

```
df = data[['ref_area.label', 'sex.label', 'classif1.label','classif2.label', 'time', 'obs_value','obs_status.label']]

# Renommer pour plus de clarté
df.columns = ['Country', 'Sex', 'Age_group','Education_Status', 'Year', 'Value','Fiability_data']

# Supprimer les lignes avec valeur manquante
df_clean = df.dropna(subset=['Value']).copy()

# Transformer l'année en format datetime
df_clean['Year'] = pd.to_datetime(df_clean['Year'], format='%Y')

# Supprimer les doublons si existants
df_clean = df_clean.drop_duplicates()


df_clean['Age_group'] = df_clean['Age_group'].str.replace(
    "Age \(Child labour bands\): '?",  # '? signifie "0 ou 1 apostrophe"
    "",
    regex=True
)

df_clean['Education_Status'] = df_clean['Education_Status'].str.replace(
    "Educational attendance: ",  # Texte à supprimer
    "",                          # Remplacer par une chaîne vide
    regex=True                # Désactiver les regex pour un remplacement littéral
)


# Afficher un aperçu
print("\nDonnées nettoyées :")
print(df_clean.head())

# Sauvegarder le résultat
df_clean.to_csv('cleaned_data.csv', index=False)

print("\nFichier nettoyé enregistré sous 'cleaned_child_labour_data.csv'.")
```

FIGURE 6.5 – Application du filtrage

```
Données nettoyées :
       Country    Sex Age_group          Education_Status       Year  \
0  Afghanistan  Total      5-17                     Total 2014-01-01
1  Afghanistan  Total      5-17                 Attending 2014-01-01
2  Afghanistan  Total      5-17             Not attending 2014-01-01
3  Afghanistan  Total      5-17  Not elsewhere classified 2014-01-01
4  Afghanistan  Total      5-11                     Total 2014-01-01

      Value Fiability_data
0  3261.161            NaN
1  1556.560            NaN
2   185.107            NaN
3  1519.494            NaN
4  1254.495            NaN
```

FIGURE 6.6 – Résultat du filtrage

6.4.2 Filtrage des régions

Dans cette étape, on affiche les entités avec le plus d'enfants travailleurs.

```
[ ] # Filtrer sur Age_group, Sex et Education_Status = "Total"
    df_filtered = child[
        (child['Age_group'] == '5-17') &
        (child['Sex'] == 'Total') &
        (child['Education_Status'] == 'Total')
    ]

    # Trier par Value décroissant et afficher les 10 premiers résultats
    top_n = 20
    top = df_filtered[['Country', 'Year', 'Age_group', 'Sex', 'Education_Status', 'Value']].sort_values(by='Value', ascending=False).head(top_n)

    print("Top valeurs pour Age_group=Total, Sex=Total, Education_Status=Total :")
    print(top)
```

```
Top valeurs pour Age_group=Total, Sex=Total, Education_Status=Total :
                                Country       Year Age_group    Sex  \
6729                              World  2020-01-01      5-17  Total
6765                             Africa  2020-01-01      5-17  Total
6837                 Sub-Saharan Africa  2020-01-01      5-17  Total
7269                Asia and the Pacific  2020-01-01      5-17  Total
6909                     Eastern Africa  2020-01-01      5-17  Total
6981                     Western Africa  2020-01-01      5-17  Total
7449                     Southern Asia   2020-01-01      5-17  Total
7341  South-Eastern Asia and the Pacific  2020-01-01      5-17  Total
7377                 South-Eastern Asia  2020-01-01      5-17  Total
2440                           Ethiopia  2015-01-01      5-17  Total
6873                    Central Africa   2020-01-01      5-17  Total
7017                          Americas   2020-01-01      5-17  Total
7385                       Eastern Asia  2020-01-01      5-17  Total
7053    Latin America and the Caribbean  2020-01-01      5-17  Total
7485          Europe and Central Asia    2020-01-01      5-17  Total
2404                           Ethiopia  2021-01-01      5-17  Total
4694                           Nigeria   2022-01-01      5-17  Total
6801                   Northern Africa   2020-01-01      5-17  Total
```

FIGURE 6.7 – Les entités avec le plus d'enfants travailleurs

On a remarqué que certaines lignes du fichier représentent des régions entières et non des pays. Cela pose problème pour suivre l'évolution des données dans le temps, car les valeurs des régions sont souvent très élevées et peuvent écraser celles des pays, ce qui fausse l'analyse. Pour éviter cela, on a décidé de supprimer les régions et de garder uniquement les pays.

```python
import pandas as pd

# Liste des régions à exclure
regions = [
    'World', 'Africa', 'Northern Africa', 'Sub-Saharan Africa', 'Central Africa', 'Eastern Africa',
    'Southern Africa', 'Western Africa', 'Americas', 'Latin America and the Caribbean', 'Caribbean',
    'Central America', 'South America', 'Northern America', 'Arab States', 'Asia and the Pacific',
    'Eastern Asia', 'South-Eastern Asia and the Pacific', 'South-Eastern Asia', 'Pacific Islands',
    'Southern Asia', 'Europe and Central Asia', 'Northern Southern and Western Europe', 'Northern Europe',
    'Southern Europe', 'Western Europe', 'Eastern Europe', 'Central Asia', 'Central and Western Asia',
    'Western Asia'
]

# Charger le fichier
df = pd.read_csv("cleaned_data_multiplied.csv")

# Filtrer pour ne garder que les pays (pas les régions) et les totaux pertinents
df_pays = df[
    (df['Sex'] == 'Total') &
    (df['Age_group'] == '5-17') &
    (df['Education_Status'] == 'Total') &
    (~df['Country'].isin(regions))
]

# Trier par Value décroissant
top_n = 10
top_countries = df_pays[['Country', 'Year', 'Value']].sort_values(by='Value', ascending=False).head(top_n)

print("Pays et années avec les plus grandes valeurs de travail des enfants :")
print(top_countries)
```

```
Pays et années avec les plus grandes valeurs de travail des enfants :
                                Country        Year       Value
2440                           Ethiopia  2015-01-01  19044691.0
2404                           Ethiopia  2021-01-01   9603262.0
4694                            Nigeria  2022-01-01   9015391.0
3278                              India  2023-01-01   6904617.0
1082  Democratic Republic of the congo  2020-01-01   6373050.0
6280                             Uganda  2021-01-01   5491148.0
6196       United Republic of Tanzania  2020-01-01   5128527.0
3309                              India  2022-01-01   5123563.0
6244       United Republic of Tanzania  2014-01-01   5066889.0
3340                              India  2021-01-01   4642897.0
```

FIGURE 6.8 – Les pays avec le plus d'enfants travailleurs

6.4.3 Filtrage des données peu fiables

Maintenant, nous excluons les données peu fiables présentes dans le jeu de données. Nous avons identifié deux types principaux de données jugées peu fiables :

— Unreliable : Ces données sont considérées comme peu fiables par la source statistique, en raison de facteurs tels qu'un échantillon trop faible, une méthode de collecte non conforme ou des réponses jugées douteuses.

— Break in series : Cela indique une rupture dans la continuité de la série statistique, généralement due à des changements dans le questionnaire, la définition du travail des enfants, ou encore dans la population cible (par exemple, en termes d'âge ou de zones géographiques).

Il est donc préférable de les exclure.

```python
# Exclure les lignes peu fiables
df = df[~df['Fiability_data'].str.contains("Unreliable|Break in series", na=False)]
```

FIGURE 6.9 – Exclusion des données jugées peu fiables

Nous sauvegardons maintenant le résultat du nettoyage dans un nouveau fichier CSV, qui sera utilisé pour les étapes suivantes.

```
Nombre total de lignes dans le dataset original : 8153
Nombre de lignes après nettoyage : 5771
Nombre de lignes supprimées : 2382
```

FIGURE 6.10 – Dataset après nettoyage

6.5 Intégration avec InfluxDB

Dans cette partie, on va intégrer nos données dans InfluxDB, ce dernier étant adapté aux séries temporelles, on pourra visualiser les différents tags facilement .

6.5.1 Chargement de la bibliothèque Influx sur python

Pour utiliser InfluxDB on aura besoin de la bibliothèque suivante :

```
import influxdb_client
```

FIGURE 6.11 – Bibliothèque influx

influxdb-client est une bibliothèque Python qui permet de se connecter à une instance InfluxDB (versions 2.x et 3.x), d'y écrire des données (write), d'en interroger (query) à l'aide du langage Flux ou SQL, et de gérer les buckets, organisations et autres ressources InfluxDB directement depuis Python.

6.5.2 Connexion à InfluxDB

Nous devons maintenant nous connecter à InfluxDB et définir les variables Sex, Age-group, et Education-status comme tags, tandis que value sera utilisé comme valeur.

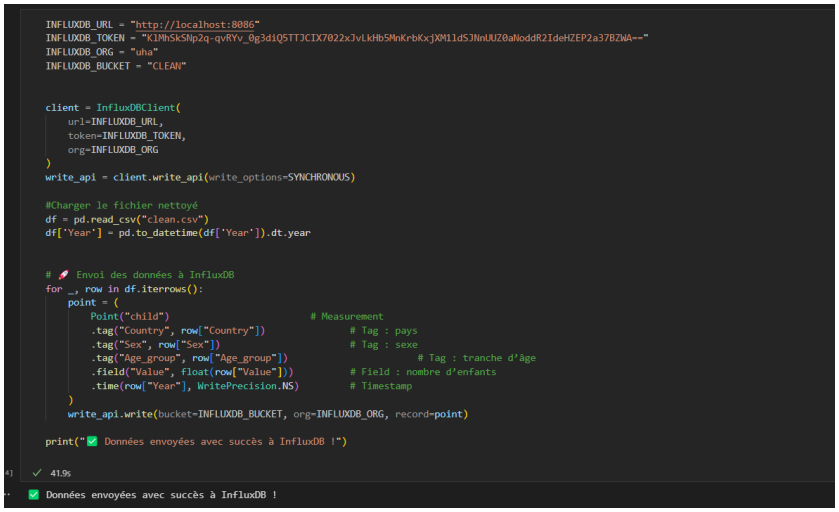


FIGURE 6.12 – Connexion à influx

La connexion a été établie et le bucket a été rempli avec nos données. Nous pouvons maintenant vérifier si les données ont été correctement insérées en accédant à InfluxDB et en affichant directement notre dataset.

table	_measurement	_field	_value	_start	_stop	_time	Age_group	Country	Education_Status	Sex
0	child_labor	Value	138533	1999-12-31T23:00:00.000Z	2023-12-31T23:00:00.000Z	2014-01-25T08:24:00.000Z	12-14	Afghanistan	Attending	Female
1	child_labor	Value	439381	1999-12-31T23:00:00.000Z	2023-12-31T23:00:00.000Z	2014-01-25T08:24:00.000Z	12-14	Afghanistan	Attending	Male
2	child_labor	Value	577914	1999-12-31T23:00:00.000Z	2023-12-31T23:00:00.000Z	2014-01-25T08:24:00.000Z	12-14	Afghanistan	Attending	Total
3	child_labor	Value	25735	1999-12-31T23:00:00.000Z	2023-12-31T23:00:00.000Z	2014-01-25T08:24:00.000Z	12-14	Afghanistan	Not attending	Female
4	child_labor	Value	22035	1999-12-31T23:00:00.000Z	2023-12-31T23:00:00.000Z	2014-01-25T08:24:00.000Z	12-14	Afghanistan	Not attending	Male

‹ 1 2 3 4 5 … 1535 ›

FIGURE 6.13 – Affichage des données avec InfluxDB

6.6 Analyse des évolutions

Dans cette partie, on pourra facilement visualiser et analyser l'évolution de nos données :

6.6.1 Évolution globale :

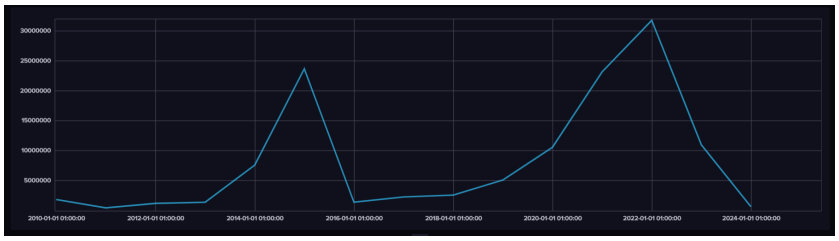


FIGURE 6.14 – Évolution globale au fur des années

— Observation : La courbe montre deux pics majeurs : un en 2015 (23,6 millions) et un autre, encore plus élevé, en 2022 (31,7 millions), suivis d'une baisse rapide en 2023-2024.

— Analyse : Ces pics indiquent des ruptures ou des événements exceptionnels dans la collecte ou la réalité du travail des enfants. La tendance générale est à la hausse entre 2017 et 2022, puis une chute brutale.

— Interprétation : Les pics peuvent être liés à des changements méthodologiques, à des crises (par exemple : pandémie, conflits), ou à l'intégration de nouveaux pays ou groupes dans les statistiques.

6.6.2 Comparaison de l'évolution des 2 sexes :

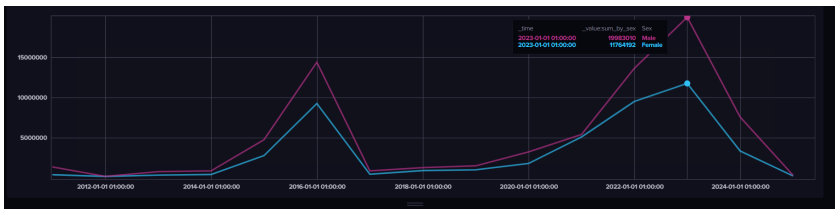


FIGURE 6.15 – Évolution des deux sexes au fur des années

— Observation : Les garçons (courbe bleue) sont systématiquement plus nombreux que les filles (courbe rouge) parmi les enfants travailleurs. Les deux suivent des tendances similaires, avec des pics en 2015 et 2022.

— Analyse : L'écart entre les sexes reste stable, ce qui suggère que les facteurs qui influencent le travail des enfants touchent davantage les garçons, mais affectent aussi les filles de façon parallèle.

— Interprétation : Les politiques de lutte contre le travail des enfants doivent cibler les deux sexes, avec une attention particulière aux garçons.

6.6.3 Comparaison de l'Évolution selon les groupes d'age :

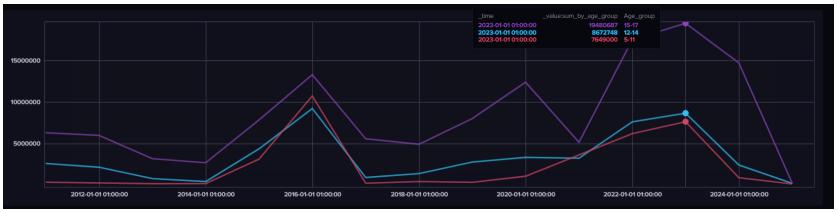


FIGURE 6.16 – Évolution selon les groupes d'âge au fur des années

— Observation : Les adolescents de 15-17 ans sont de loin les plus nombreux parmi les enfants travailleurs, suivis des 12-14 ans et des 5-11 ans . Les pics de 2015 et 2022 sont portés principalement par les 15-17 ans.

— Analyse : Le travail des enfants concerne surtout les plus âgés, mais touche aussi les plus jeunes lors des pics.

— Interprétation : Les interventions doivent être adaptées à chaque tranche d'âge, avec un accent particulier sur les adolescents.

6.6.4 Comparaison de l'évolution selon statut de scolarité :

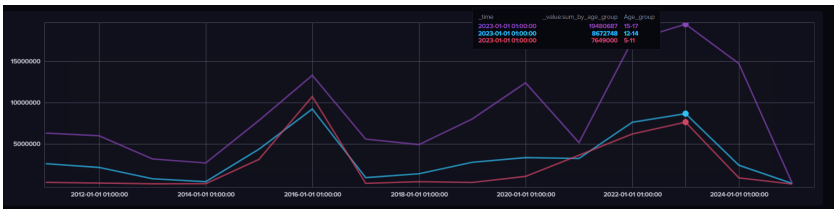


FIGURE 6.17 – Évolution selon le statut de scolarité au fur des années

— Observation : La majorité des enfants travailleurs sont scolarisés ("Attending", mais la part des non-scolarisés ("Not attending") augmente fortement lors des pics.

— Analyse : Le travail des enfants ne signifie pas toujours abandon scolaire : beaucoup cumulent école et travail. Cependant, lors des crises, le nombre d'enfants non scolarisés explose.

— Interprétation : Les politiques doivent viser à maintenir les enfants à l'école et à réduire le travail des enfants, en particulier lors des périodes de crise.

6.6.5 Analyse et comparaison générale

— **Crises économiques et augmentation de la pauvreté** : La pauvreté demeure la principale raison structurelle du travail des enfants. Pendant les crises économiques, les familles qui perdent des revenus sont plus enclines à envoyer leurs enfants travailler pour répondre aux besoins essentiels de la maison.

— **Pandémie de Covid-19** : La crise sanitaire mondiale a largement affecté les populations, surtout en Afrique subsaharienne et dans d'autres régions vulnérables. Les confinements, la fermeture des écoles, la perte d'emplois des parents et la diminution des aides sociales ont amené beaucoup d'enfants à travailler.

— **Problèmes dans les systèmes éducatifs** : L'accès difficile à l'école, les frais de scolarité, la distance et la fermeture temporaire des établissements favorisent l'abandon scolaire et la nécessité de faire travailler les enfants.

— **Facteurs démographiques et structurels** : Dans certaines parties du monde, la croissance rapide de la population, couplée à un manque d'emplois convenables pour les adultes, augmente la pression sur les enfants pour qu'ils aident financièrement la famille.

— En 2015, le rapport mondial de l'OIT met déjà en avant l'impact des difficultés d'insertion des jeunes sur le marché du travail, ce qui peut rendre le travail des enfants plus intéressant pour les familles en difficulté.

— En 2022, la pandémie de Covid-19 est vue comme la principale raison ayant causé le retour à la hausse du travail des enfants, avec une détérioration des conditions économiques et sociales.

— **Conclusion** : Les hausses de 2015 et 2022 montrent l'effet combiné des crises économiques, de la pandémie, de la pauvreté persistante, de l'insuffisance des protections sociales et des difficultés d'accès à l'éducation.

6.7 Implémentation des modèles statistiques classiques

6.7.1 Décomposition et analyse

On commence notre analyse par décomposer notre série temporelle et afficher ses composantes pour les analyser.

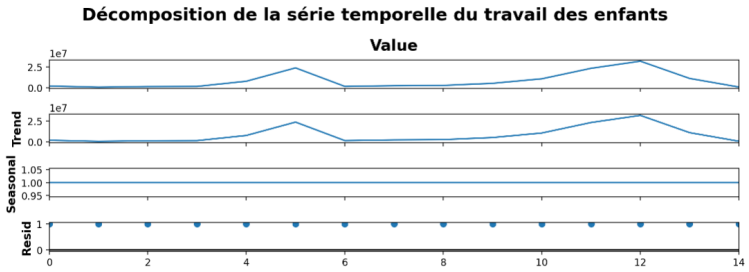


FIGURE 6.18 – Décomposition de la série temporelle

Analyse de la décomposition de la série temporelle du travail des enfants

L'étude de la série temporelle du travail des enfants nécessite l'examen séparé de chacune de ses composantes, telles qu'affichées dans le graphique :

1. **Série originale (« Value »)**
   La première courbe représente l'évolution globale du nombre d'enfants au travail sur la période considérée. On observe des fluctuations notables, caractérisées par des pics marqués autour des cinquième et douzième unités de temps, suivis d'une diminution après ces sommets.
2. **Tendance (« Trend »)**
   La seconde courbe isole la tendance de fond de la série. Celle-ci suit la forme générale de la série originale, mais en atténuant les variations rapides. La tendance indique une croissance progressive jusqu'à un maximum (vers les cinquième et douzième unités de temps), puis une baisse, ce qui suggère que le phénomène du travail des enfants traverse des phases d'augmentation et de diminution au cours de la période étudiée.
3. **Saisonnalité (« Seasonal »)**
   La troisième courbe illustre la composante saisonnière. Celle-ci est quasiment constante et proche de 1, traduisant l'absence de variations saisonnières significatives : le travail des enfants ne présente pas

de fluctuations régulières associées à des cycles saisonniers ou des périodes fixes.

4. **Résidus (« Resid »)**
   Enfin, la dernière courbe présente les résidus. Les valeurs sont proches de zéro, ce qui indique que la dynamique de la série est majoritairement captée par la tendance, avec peu de fluctuations aléatoires ou d'anomalies restantes.

6.7.2 Ajustement des modèles

On a remarque que des pics existent en 2015 et 2022 , afin d'éviter ces forts pics on peut enlever les points et faire une interpolation. L'interpolation permet ici de combler les années de pics qui sont 2015 et 2022 pour garantir une série temporelle exploitable, continue et cohérente, facilitant ainsi l'analyse, la modélisation et la prévision du phénomène.

On segmente notre série en 3 segments : avant 2015 , 2015-2022 , après 2022 Découper la série temporelle en sous-périodes homogènes séparées par des points de rupture permet d'ajuster des modèles spécifiques à chaque segment, ce qui améliore la capture des tendances et variations propres à chaque période, évite que les ruptures ne biaisent la modélisation globale, et rend les prévisions plus fiables et réalistes en tenant compte des dynamiques particulières de chaque phase.

FIGURE 6.19 – Interpolation et segmentation de la série temporelle

Après cette transformation on implémente nos 3 modèles en entraînant chacun sur chaque segment , en divisant les données en 80% de données d'entraînement et 20% pour le test.

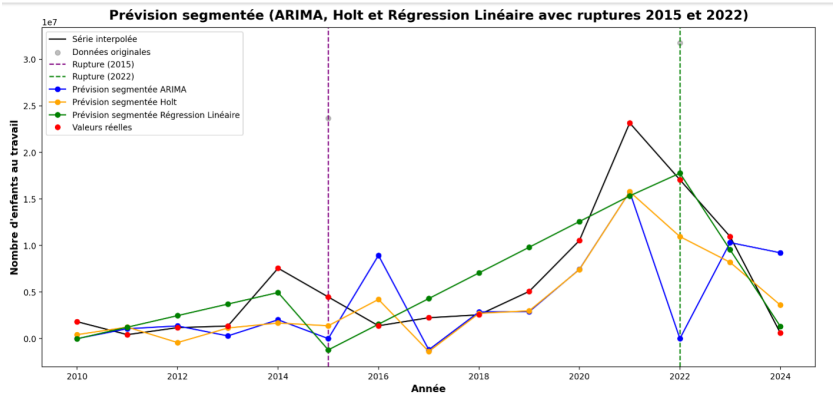


FIGURE 6.20 – Test des différents modèles

Analyse des résultats

— Jusqu'en 2015, les trois modèles respectent généralement la tendance modérée de la série, avec des différences assez minimes comparées aux valeurs effectives.
— Entre 2015 et 2022, la série enregistre une augmentation significative accompagnée de fluctuations importantes. Les modèles segmentés essaient de s'ajuster à cette nouvelle dynamique :
— La régression linéaire (en vert) suit la tendance à la hausse,
— Holt (orange) et ARIMA (bleu) manifestent des réactions distinctes face aux fluctuations récentes, en fonction de leur sensibilité respective.
— Après l'année 2022, la tendance se modifie à nouveau et la série connaît une chute abrupte. Les modèles, dont l'entraînement se limite à ce segment, anticipent des variations très diverses :
— ARIMA (en bleu) prévoit une stabilisation ou une légère augmentation,
— Holt (en orange) et la régression linéaire (en vert) continuent de suivre la tendance à la baisse.

L'avantage de la segmentation est clair :
Chaque modèle ici est ajusté à la dynamique spécifique de son segment, permettant ainsi des prévisions plus réalistes pour chaque sous-période.
Les dynamiques sont ensuite pondérées par une moyenne pour essayer d'avoir des résultats cohérents.

6.7.3 Test des prévisions

On peut tester les modèles pour une prévision de 5 années, on aura les résultats suivants :

Année	Prévision Régression Linéaire	Prévision Holt	Prévision ARIMA
2025	667109	611367	899127
2026	722851	611367	1186887
2027	778593	611367	1474647
2028	834334	611367	1762407
2029	890076	611367	2050167
2030	945818	611367	2337927

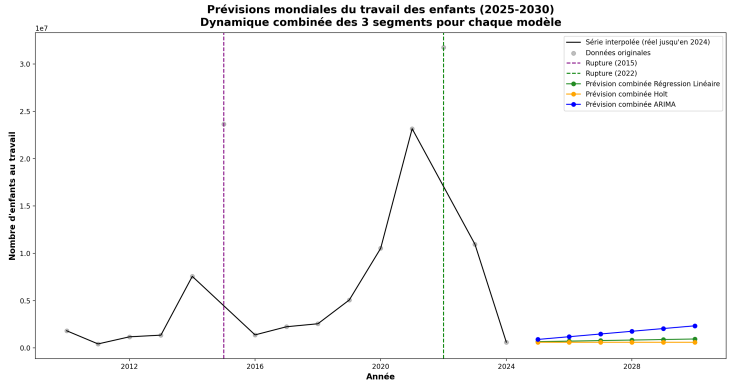


FIGURE 6.21 – Résultats de la prévision

Les prévisions 2025-2030 sont issues de la combinaison pondérée des tendances de chaque segment :

— Régression linéaire (vert) et Holt (orange) prévoient une stabilisation à un niveau bas, prolongeant la tendance récente.
— ARIMA (bleu) anticipe une légère hausse, reflétant une dynamique différente issue de la combinaison des trois segments.

On constate que :

— Les trois modèles, bien qu'utilisant la même logique de combinaison des dynamiques passées, donnent des trajectoires assez proches mais pas identiques : cela traduit l'incertitude sur la reprise ou la stabilisation du phénomène après la forte chute post-2022.
— Les ruptures de 2015 et 2022 (traits verticaux) structurent nettement la série : elles correspondent à des changements brutaux de tendance qui rendent la prévision difficile.
— Les valeurs prévues restent très inférieures au pic de 2022, indiquant que, selon tous les modèles, un retour à un niveau aussi élevé est peu probable à court terme.

6.7.4 Test de fiabilité

Dans cette section on va tester la fiabilité des modèles en calculant la MAPE (*Mean Absolute Percentage Error*, ou Erreur Absolue Moyenne en Pourcentage) est un indicateur qui mesure la précision d'un modèle de prévision. Elle se calcule ainsi :

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100$$

où y_i est la valeur réelle, $\hat{y}_i$ la valeur prédite et n le nombre d'observations. [14].

Plus la MAPE est faible, plus le modèle est fiable. Par exemple, une MAPE de 10 % signifie que la prévision s'écarte en moyenne de 10 % de la valeur réelle.

```
Fiabilité (MAPE sur 2023-2024) :
  Régression linéaire pondérée : 47.47%
  Holt pondéré                 : 47.21%
  ARIMA pondéré                : 48.52%
```

FIGURE 6.22 – Résultats MAPE

Ces valeurs sont très élevées : elles signifient que, sur la période 2023-2024, les prévisions issues de chaque modèle s'écartent en moyenne de près de la moitié de la valeur réelle.

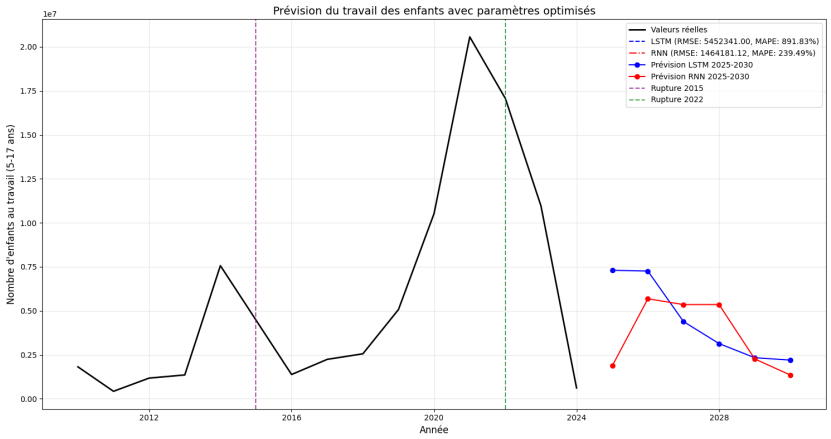
Aucun des trois modèles n'est fiable : en général, une MAPE inférieure à 20 % est considérée comme acceptable pour des prévisions de séries temporelles ; ici, on est largement au-dessus.

Les trois méthodes testées (régression linéaire pondérée, Holt pondéré, ARIMA pondéré) donnent des résultats très proches, ce qui montre qu'aucune ne parvient à bien capter la dynamique réelle récente de la série. Cela peut s'expliquer par des ruptures très marquées, des dynamiques hétérogènes ou une forte instabilité des données causée par les pics présents et le manque de saisonnalité.

6.8 Implémentation de modèles d'IA

6.8.1 Réseaux neuronaux récurrents (RNN) et LSTM

Dans cette partie, on s'intéresse aux modèles de deep learning et leurs fonctionnements :



Prévisions 2025-2030:

Année	LSTM	RNN
2025	7,301,007	1,871,151
2026	7,259,502	5,683,525
2027	4,389,568	5,353,235
2028	3,136,140	5,355,211
2029	2,334,133	2,263,053
2030	2,198,522	1,350,041

FIGURE 6.23 – Résultats modèles IA

Interprétation et explications

Volatilité et ruptures :
Les valeurs réelles montrent de grandes fluctuations, ce qui rend la tâche difficile pour les modèles séquentiels classiques (LSTM/RNN). Les ruptures (2015, 2022) correspondent probablement à des changements de méthodologie, de politique, ou à des crises majeures.

Différences LSTM/RNN :
Le LSTM semble « surestimer » la persistance du pic observé autour de 2022, puis anticipe une forte correction.
Le RNN, moins sensible à la mémoire longue, propose des prévisions plus basses, mais reste instable sur la période 2026-2028.

Scores d'erreur élevés :
Les MAPE très élevés (surtout LSTM : près de 900 %) indiquent que les modèles ne parviennent pas à bien capturer la dynamique réelle, probablement à cause des ruptures et de la non-stationnarité de la série.

Tendance générale :
Malgré les différences, les deux modèles prévoient une baisse du travail des

enfants à l'horizon 2030, mais avec des incertitudes majeures sur le niveau réel attendu.

6.8.2 Prophet

Prophet ici est puissant car il gère les coupures dans les séries.

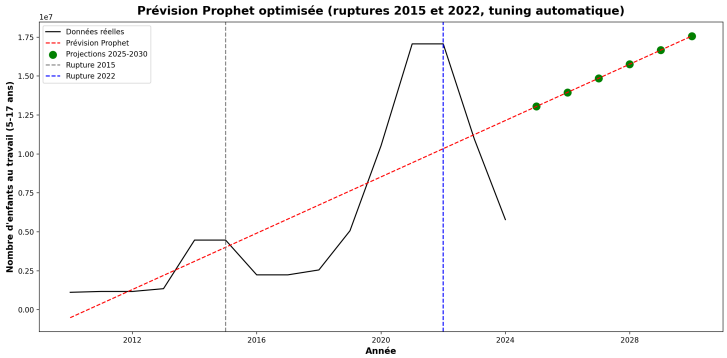


FIGURE 6.24 – Visualisation avec prophet

```
Prévisions Prophet 2025-2030 :
          ds          yhat
15 2024-12-31  1.304865e+07
16 2025-12-31  1.395195e+07
17 2026-12-31  1.485525e+07
18 2027-12-31  1.575856e+07
19 2028-12-31  1.666433e+07
20 2029-12-31  1.756763e+07
```

FIGURE 6.25 – Résultats prophet

Analyse de la segmentation et des prévisions

Le modèle Prophet (ligne rouge pointillée) prévoit :
— Un rebond significatif à partir de 2024
— Une augmentation constante et quasi-linéaire sur toute la période 2025-2030
— Des valeurs passant progressivement de 13,848 millions en 2025 à 17,527 millions en 2030
— Un retour progressif vers les niveaux maximaux observés en 2022

Cette prévision suggère que la baisse observée après 2022 n'était que temporaire, et que la tendance de fond reste à la hausse sur le long terme.

Fiabilité du modèle

La fiabilité du modèle, mesurée par le MAPE (Mean Absolute Percentage Error) de 43,07%, indique une précision modérée. Cette métrique représente l'écart moyen en pourcentage entre les valeurs réelles et prédites. Un MAPE de 43% signifie que les prévisions du modèle s'écartent en moyenne de 43% des valeurs réelles - un niveau d'erreur relativement élevé qui s'explique par :

- La forte volatilité de la série temporelle
- Les ruptures structurelles prononcées en 2015 et 2022
- La difficulté inhérente à modéliser des phénomènes socio-économiques complexes

Implications et limites

Cette analyse met en évidence la nature cyclique et instable du travail des enfants à l'échelle mondiale. L'optimisation du modèle Prophet avec points de rupture explicites permet de mieux capturer ces dynamiques, mais la fiabilité modérée (MAPE de 43%) nous rappelle les limites de toute projection dans un contexte aussi volatile.

6.8.3 Clustering

Dans cette partie, on va essayer de créer des clusters pour regrouper les pays avec des données ou des métrique similaires, On commence par importer la bibliothèque Tslearn :

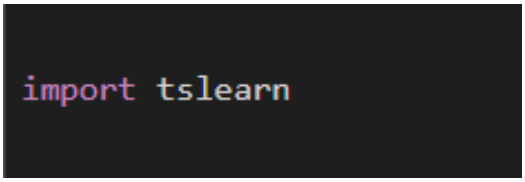
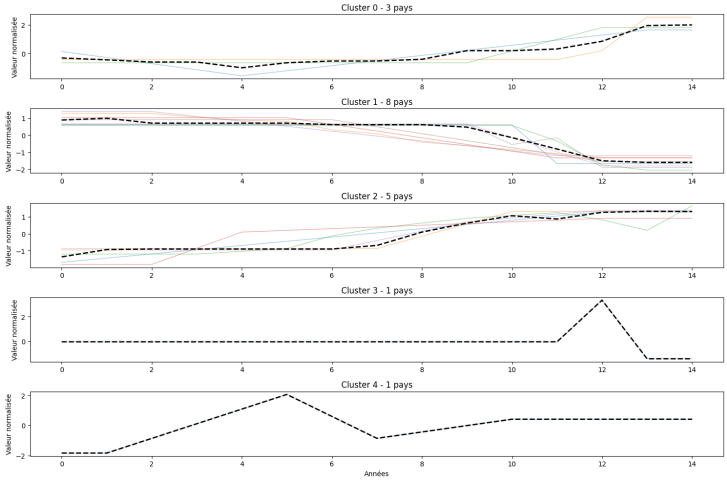


FIGURE 6.26 – Import de Tslearn

Tslearn est une bibliothèque Python qui fournit des outils d'apprentissage automatique pour l'analyse des séries temporelles. Cette bibliothèque est construite sur les bases de Scikit-learn, Numpy et Scipy, dont elle dépend. [15].

Résultats obtenus




```
Pays par cluster :

Cluster 0 (3 pays):
Egypt, India, Mexico

Cluster 1 (8 pays):
Colombia, Ethiopia, Gabon, El Salvador, Cambodia, Timor-Leste, Togo, Zambia

Cluster 2 (5 pays):
Ecuador, Senegal, Niger, Burundi, Burkina Faso

Cluster 3 (1 pays):
Dominican Republic

Cluster 4 (1 pays):
Malawi
```

FIGURE 6.27 – Résultat du clustering

On obtient 5 clusters en tout.

Analyse des résultats obtenus :

Cluster	**Pays**	**Similitudes**
0	Égypte, Inde, Mexique	Grandes économies émergentes avec tendance à l'augmentation en fin de période ; fortes populations ; économies diversifiées
1	Colombie, Éthiopie, Gabon, El Salvador, Cambodge, Timor-Leste, Togo, Zambie	Tendance descendante ; pays en transition et développement ; amélioration des indicateurs du travail des enfants
2	Équateur, Sénégal, Niger, Burundi, Burkina Faso	Principalement des pays africains ; tendance ascendante ; prédominance du travail agricole ; vulnérabilité aux changements climatiques
3	République Dominicaine	Profil atypique avec pic soudain puis chute brutale ; dynamique singulière
4	Malawi	Profil volatil avec pic marqué suivi d'une stabilisation ; cas isolé avec schéma distinct

TABLE 6.2 – Répartition des pays par cluster et caractéristiques principales

Cluster 0 - Grandes économies émergentes (3 pays) Ce cluster regroupe trois pays aux économies importantes qui présentent une évolution similaire :

- **Profil temporel** : Tendance relativement stable initialement, suivie d'une augmentation significative vers la fin de la période observée
- **Caractéristiques communes** :
  - Pays à revenus intermédiaires avec de grandes populations
  - Coexistence de secteurs économiques formels et informels
  - Marchés du travail diversifiés mais avec persistance du travail des enfants

Cluster 1 - Pays en amélioration (8 pays) Ce groupe hétérogène géographiquement partage une tendance positive :

- **Profil temporel** : Courbe descendante avec valeurs positives au début et négatives à la fin
- **Caractéristiques communes** :
  - Mise en place de politiques efficaces de protection de l'enfance
  - Transition économique et sociale progressive
  - Renforcement des systèmes éducatifs
  - Programmes internationaux d'aide actifs

Cluster 2 - Pays à défis persistants (5 pays) Majoritairement africains, ces pays font face à des difficultés structurelles :

- **Profil temporel** : Tendance globalement ascendante partant de valeurs négatives vers des valeurs positives
- **Caractéristiques communes** :
  - Forte prévalence du travail agricole saisonnier
  - Importance de l'économie informelle
  - Vulnérabilité aux changements climatiques
  - Pression démographique significative
  - Instabilité politique dans certains cas

Cluster 3 - Cas atypique (République Dominicaine) Ce pays présente un profil unique qui justifie son isolement dans un cluster séparé :

- **Profil temporel** : Pic extrêmement prononcé à l'année 12 suivi d'une chute brutale
- **Caractéristiques** : l'exclusion scolaire et sociale de certains enfants, notamment ceux d'origine haïtienne, à la traite transfrontalière, et à

l'augmentation des contrôles sur les lieux de travail. Ces éléments ont révélé une exacerbation des vulnérabilités sociales, aggravées par la crise du Covid et des changements législatifs discriminatoires.

Cluster 4 - Cas spécifique (Malawi) Ce cluster composé d'un seul pays présente un profil distinct :

- **Profil temporel** : Valeurs très négatives au début, pic prononcé au milieu, puis stabilisation
- **Caractéristiques** : Ce comportement atypique s'explique par une crise agricole causée par des inondations et une sécheresse, entraînant une insécurité alimentaire massive. Cette situation, combinée à la pauvreté rurale persistante, a accru la dépendance des familles au travail des enfants.

Chapitre 7

Conclusion générale

Avec ce projet, on a examiné l'évolution du travail des enfants au niveau global entre 2010 et 2023, en intégrant des approches statistiques classiques, des techniques d'IA ainsi qu'une base de données temporelle comme InfluxDB. En utilisant des données fiables de l'OIT et de l'UNICEF, l'étude a révélé deux pics significatifs en 2015 et 2022, liés aux crises socio-économiques, aux méthodes de rupture et aux impacts du COVID-19.

Les résultats indiquent que les enfants travailleurs continuent d'être profondément touchés par l'éducation absente, les conflits et les lacunes structurelles, avec une présence notable chez les garçons et les jeunes âgés de 15 à 17 ans. Il est à noter qu'une large proportion des enfants reçoit encore une éducation, ce qui complique la compréhension de ce phénomène.

Les résultats montrent que la réalisation de prévisions à moyen et long terme est particulièrement ardue en raison de la nature annuelle des données, ce qui limite la profondeur de l'analyse. En outre, le faible volume des données disponibles (Small-Data) représente un défi considérable pour les modèles d'intelligence artificielle comme le LSTM, qui requièrent des ensembles de données étendus et riches pour apprendre efficacement sans sur-ajuster.

Ainsi, bien que des modèles tels que les prévisions ou les approches de segmentation aient contribué à formuler des hypothèses concernant les tendances pour la période 2025-2030, les erreurs de prévision significatives rappellent que chaque prédiction doit être interprétée avec rigueur.

Enfin, un cluster temporel facilite l'identification des dossiers nationaux en fonction de leur degré de développement, ce qui met en lumière la nécessité de politiques adaptées à chaque contexte national.

Ce projet souligne l'importance de la surveillance continue, d'une collecte de données plus fréquente et harmonisée, ainsi que de l'utilisation de solutions technologiques et de décisions politiques pour prévenir et éradiquer le travail des enfants sur le long terme.

Bibliographie

[1] Organisation Internationale du Travail, *IPEC - Programme international pour l'abolition du travail des enfants*, http://ilo.org/ipec/facts/lang--fr/index.htm.

[2] ENSAl, *Séries temporelles*, https://ensai.fr/wp-content/uploads/2019/06/Polyseriestemp.pdf.

[3] Université Mouloud Mammeri de Tizi-Ouzou, *Chapitre 3*, https://www.ummto.dz/fsecsg/wp-content/uploads/2021/09/Chapitre-3-1.pdf.

[4] Yannig Goude, *Cours 2 : Tendance et composante saisonnière*, https://www.imo.universite-paris-saclay.fr/~yannig.goude/Materials/time_series/cours2_tendance_composante_saisonniere.pdf.

[5] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting : Principles and Practice* (3rd ed.). OTexts. Retrieved from https://otexts.com/fpp3/

[6] Pure Storage, *What is a time series database ?*,https://www.purestorage.com/fr/knowledge/what-is-a-time-series-database.html.

[7] ENSA Khouribga, *Data Mining Texte Mining , chapitre :Clustering* https://v-assets.cdnsw.com/fs/Root/e4dhq-Chap_4_Clustering.pdf

[8] Introduction to clustering in data science, *SUBEX a telecom ai company* https://www.subex.com/blog/introduction-to-clustering-in-data-science/

[9] Clustering de séries temporelles, *Heka AI* https://www.heka.ai/fr/nos-publications/clustering-de-series-temporelles

[10] Bridging the Gap : A Decade Review of Time-Series Clustering Methods, *Arxiv* https://arxiv.org/html/2412.20582v1

[11] Raphaël Bruchez, *Les bases de données NoSQL, comprendre et mettre en œuvre*, 3 édition, 2022, pp. 131 et 236.

[12] Documentation officielle de pandas, https://pandas.pydata.org/docs/.

[13] Documentation officielle de matplotlib, https://matplotlib.org/stable/index.html.

[14] Oracle Help Center, https://docs.oracle.com/cloud/help/fr/pbcs_common/PFUSU/insights_metrics_MAPE.htm#PFUSU-GUID-C33B0F01-83E9-468B-B96C-413A12882334.

[15] Documentation officielle de tslearn, https://tslearn.readthedocs.io/en/stable/index.html.