



الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

جامعة وهران 2 محمد بن أحمد
Université d'Oran 2 Mohamed Ben Ahmed
معهد الصيانة و الأمن الصناعي
Institut de Maintenance et de Sécurité Industrielle

Département MÉMOIRE

Pour l'obtention du diplôme de Master

Filière : Hygiène et Sécurité Industrielle
Spécialité : Sécurité Industrielle et Environnement

Thème

Risk assesement with principal component analysis

Présenté et soutenu publiquement par :

Elghelagh Taha Dhiya Eddine

Ouddane ilyass

Devant le jury composé de :

Nom et Prénom	Grade	Etablissement	Qualité
DERIAS FATIMA ZOHRA		IMSI	Président
ZOUAIRI SAIM		IMSI	Encadrant
AIDENE ZAHIA		IMSI	Examineur

ملخص

يهدف هذا البحث إلى تقييم المخاطر الصناعية بالاعتماد على تحليل المكونات الرئيسية باعتباره أداة إحصائية لتقليل أبعاد البيانات وتبسيط تفسيرها. استخدمت قاعدة بيانات لحوادث الصحة والسلامة المهنية تحتوي على متغيرات كمية ونوعية مرتبطة بمستويات الحوادث والسياق الجغرافي. بعد المعالجة المسبقة للبيانات (الترميز، التطبيع)، أُجري تطبيق (PCA) لاستخلاص أهم المركبات. توضح الدراسة أن تحليل المكونات الرئيسية أداة استكشافية فعالة لتحديد العوامل المؤثرة، لكنها تحتاج إلى التكامل مع تقنيات أخرى لتعزيز دقة التحليل ودعم قرارات السلامة الصناعية.

Résumé

Ce mémoire propose une évaluation des risques industriels en appliquant l'Analyse en Composantes Principales (ACP) afin de réduire la dimensionnalité des données et d'en faciliter l'interprétation. La base de données utilisée porte sur des accidents de santé et sécurité au travail, avec des variables quantitatives et qualitatives liées au niveau d'accident et au contexte géographique. Après prétraitement (encodage, normalisation), l'ACP a permis d'extraire les principales composantes. L'étude confirme que l'ACP est un outil exploratoire pertinent, mais qu'elle doit être complétée par d'autres méthodes pour améliorer la précision et appuyer la prise de décision en matière de sécurité industrielle.

Abstract

This study aims to assess industrial risks by applying Principal Component Analysis (PCA) as a statistical tool for reducing data dimensionality and simplifying interpretation. The dataset consists of occupational health and safety accidents, with quantitative and qualitative variables related to accident levels and geographical context. After preprocessing (encoding, normalization), PCA was applied to extract the main components. The study demonstrates that PCA is an effective exploratory tool for identifying key factors, but it should be complemented with other methods to enhance accuracy and support industrial safety decision-making.

REMERCIEMENT

Tout d'abord, nous remercions Dieu Tout-Puissant qui nous a donné le courage et la volonté de faire ce travail.

Nous tenons à remercier notre encadrant **Monsieur Zouairi Saim** pour tout le temps qu'il nous a consacré, ses précieux conseils, et toute son aide et son soutien durant la réalisation de ce travail.

Nous exprimons également notre gratitude à nos chers parents pour tous les sacrifices qu'ils ont consentis pour nous et pour leur immense soutien. Nous leur exprimons notre plus profonde reconnaissance, notre respect et notre gratitude.

Nous remercions encore une fois beaucoup notre encadrant

Monsieur Zouairi Saim

جزاك الله عنا كل خير

Liste des abréviations

AI / IA : Artificial Intelligence (Intelligence Artificielle).

ANN : Artificial Neural Network (Réseau de Neurones Artificiels).

CNN : Convolutional Neural Network (Réseau de Neurones Convolutifs).

CV : Cross-Validation (Validation Croisée).

DL : Deep Learning (Apprentissage Profond).

KNN : K-Nearest Neighbors (Méthode des k plus proches voisins).

LSTM : Long Short-Term Memory (Mémoire à long et court terme).

ML : Machine Learning (Apprentissage Automatique).

NLP : Natural Language Processing (Traitement du Langage Naturel).

RF : Random Forest (Forêt Aléatoire).

RL : Reinforcement Learning (Apprentissage par Renforcement).

RNN : Recurrent Neural Network (Réseau de Neurones Récurrent).

SVM : Support Vector Machine (Machine à Vecteurs de Support).

XAI : Explainable Artificial Intelligence (Intelligence Artificielle Explicable).

XGBoost : Extreme Gradient Boosting.

IoT : Internet of Things (Internet des Objets).

ICT : Information and Communication Technologies (Technologies de l'Information et de la Communication).

ACP : Analyse en Composantes Principales (Principal Component Analysis, PCA).

ACM : Analyse des Correspondances Multiples.

ANOVA : Analysis of Variance (Analyse de la Variance).

AUC : Area Under the Curve (Surface sous la courbe ROC).

CAH : Classification Ascendante Hiérarchique.

DBSCAN : Density-Based Spatial Clustering of Applications with Noise (Clustering basé sur la densité).

EDA : Exploratory Data Analysis (Analyse Exploratoire des Données).

PCA : Principal Component Analysis (Analyse en Composantes Principales).

PLS : Partial Least Squares (Moindres Carrés Partiels).

ROC : Receiver Operating Characteristic.

HSE : Hygiène, Sécurité, Environnement.

OIT : Organisation Internationale du Travail.

QDA : Quadratic Discriminant Analysis (Analyse discriminante quadratique).

Liste des symboles

χ_{ij} : La valeur de la variable j pour l'observation i .

μ_j : Moyenne de la variable.

σ_j : Écart-type (standard deviation) de la variable.

Σ : Matrice de covariance.

n : nombre d'observations (individus-cas d'accidents).

X : Matrice initiale des données.

λ_k : Valeur propre associée à la composante principale.

v_k : Vecteur propre associé à la composante principale.

Z : Matrice des données centrées-réduites.

V : la matrice des vecteurs propres.

λ : le critère de Kaiser.

Liste des figures

4.1 le lien de la base de donner	36
4.2 le lien de la base de donner encodée.....	36
4.3 le lien de notre notebook sur google colab	37
4.4 Proportion de variance expliquée	38
4.5 Variance expliquée cumulée	39
4.6 Projection PCA.....	40

Liste des tableaux

Tableau (4.1) : poids des variables dans les composantes.....	42
---	----

Sommaire

Abstract	
Remerciement.....	
Liste des abréviations.....	
Liste des symboles.....	
Liste des figures.....	
Sommaire.....	
PROBLEMATIQUE.....	1
INTRODUCTION GENERALE.....	2
Chapitre I : Cadre Théorique de l'Évaluation des Risques	
I.1 Introduction.....	5
I.2 Définitions et concepts clés	5
I.3 Objectifs de l'évaluation des risques	6
I.4 Processus général d'évaluation des risques.....	7
I.5 Typologie des méthodes d'évaluation des risques	9
I.6 Limites et contraintes des méthodes classiques	10
I.7 Conclusion	11
Chapitre II : Les techniques de Machine Learning pour l'évaluation des risques	
II.1 Introduction.....	13
II.2 Concepts fondamentaux du Machine Learning	13
II.2.1 Définition et principes de base	13
II.2.2 Différences entre intelligence artificielle, machine learning et deep learning.....	14
II.2.3 Les étapes d'un projet de Machine Learning	14
II.3 Classification des méthodes de Machine Learning	15
II.3.1 Apprentissage supervisé	15
II.3.2 Apprentissage non supervisé	16
II.3.3 Apprentissage semi-supervisé	16

II.3.4 Apprentissage par renforcement	17
II.3.5 Deep Learning	17
II.4 Applications du Machine Learning à l'évaluation des risques	17
II.4.1 Maintenance prédictive et détection des défaillances.....	17
II.4.2 Gestion des risques environnementaux	18
II.4.3 Sécurité des procédés industriels.....	18
II.4.4 Analyse des risques financiers et organisationnels	18
II.4.5 L'évaluation des risques en santé et sécurité au travail (HSE).....	18
II.5 Limites et défis du Machine Learning en évaluation des risques.....	18
II.6 Conclusion.....	19
Chapitre III : Réduction de Dimension et Analyse en Composantes Principales (PCA)	
III.1. Introduction	21
III.2 Réduction de dimension.....	22
III.2.1 Définition.....	22
III.2.2 La malédiction de la dimensionnalité	22
III.2.3 Enjeux et avantages de la réduction de dimension	22
III.2.4 Importance dans l'évaluation des risques	23
III.3 Les principales méthodes de réduction de dimension	24
III.3.1 Méthodes linéaires.....	24
III.3.2 Méthodes non linéaires	25
III.3.3 Sélection de variables (Feature Selection)	25
III.3.4 Comparaison et choix de la méthode	26
III.4 Analyse en Composantes Principales (PCA)	26
III.4.1 Définition et objectifs	26
III.4.2 Principe de fonctionnement	27
III.4.3 Étapes de la PCA	27
III.4.4 Exemple schématique	28
III.4.5 Applications de la PCA	28

III.5 Applications de la PCA à l'évaluation des risques	29
III.5.1 Fonction de la PCA dans l'évaluation des risques	29
III.5.2 Applications sectorielles	29
III.5.3 Avantages de la PCA pour l'évaluation des risques	30
III.6 Limites et alternatives de la PCA	31
III.6.1 Limites de la PCA	31
III.6.2 Alternatives à la PCA	32
III.7. Conclusion	32
Chapitre IV : Étude de cas et Application de la PCA à l'évaluation des risques	
IV.1 Introduction	35
IV.2 Présentation des données	35
IV.3 Prétraitement des données	37
IV.3.1 Data cleaning	37
IV.3.2 Encodage des variables catégorielles	37
IV.3.3 Transformation des variables ordinales	37
IV.3.4 Normalisation des données	38
IV.3.5 Vérification des données finales	38
IV.4 Application de l'Analyse en Composantes Principales (ACP)	38
IV.4.1. Scree Plot - Proportion de variance expliquée.....	38
IV.4.2. Variance expliquée cumulée	39
IV.4.3. Projection PCA (PC1 versus PC2).....	40
IV.4.4 Synthèse	40
IV. 5 Interprétation des résultats	41
IV. 5.1. Répartition de la variance	41
IV. 5.2. Contribution des variables	41
IV. 5.3. Regroupements observés (Projection PCA PC1 vs PC2)	43
IV. 5.4. Implications pour l'évaluation des risques	43
IV. 5.5. Limites de l'approche	43

IV.6 Conclusion	43
Conclusion générale.....	45
BIBLIOGRAPHIE.....	46

PROBLEMATIQUE

Dans le secteur de la sécurité industrielle et environnementale, l'évaluation des risques s'appuie désormais sur la collecte et l'analyse de volumes hétérogènes de données tels que des capteurs, des inspections, des incidents, et des données climatiques. Cette surabondance de l'information, bien que bénéfique, complique le traitement et l'interprétation, mais surtout la visualisation des risques(Moubray, 2001) .

Face à l'encombrement des informations, les approches dites qualitatives ou semi-quantitatives ont des difficultés à extraire des variables significatives sans introduire de biais ou de redondance(Tixier et al., 2002) En outre, de nombreuses variables qui sont souvent utilisées dans l'évaluation des risques peuvent être corrélées, rendant complexe l'identification des facteurs de risques les plus influents.

Dans ce cadre, l'analyse en composantes principales (ACP) connue comme une technique de réduction de la dimensionnalité, permet d'agir sur l'ensemble des données de manière à les simplifier, mais en conservant l'essentiel de l'information(Jolliffe, 2011a) .

La question centrale à ce mémoire est alors la suivante : Comment optimiser une évaluation des risques dans un environnement complexe avec des données parfaitement corrélées en ayant recours à une analyse en composantes principales ?

INTRODUCTION GENERALE

Dans un contexte industriel et environnemental en constante évolution, la gestion proactive des risques constitue un enjeu fondamental pour assurer la sécurité des personnes, la protection des installations et la préservation de l'environnement. L'apparition de nouvelles technologies, la complexification des processus et l'augmentation des exigences réglementaires ont conduit les entreprises à collecter et analyser une quantité croissante de données issues de capteurs, de rapports d'inspection, d'audits HSE, ou encore d'observations terrain .

En outre, face à cette avalanche de données multidimensionnelles, les anciennes méthodes d'évaluation des risques, pour la plupart qualitatives ou semi-quantitatives, deviennent impuissantes. Souvent imprécises, subjectives ou incapables de saisir efficacement les multiples interactions entre les nombreuses variables en jeu (Tixier et al., 2002), elles doivent être transcrites pour des approches plus économes qui, tout en les métabolisant, peuvent les transformer en informations exploitables pour une prise de décision judicieuse.

Parmi les techniques modernes que l'on peut mobiliser, l'analyse en composantes principales (ACP) ou Principal Component Analysis (PCA) a le mérite d'être un outil statistique bien adapté car il permet de réduire la dimensionnalité des données, de repérer les corrélations qui existent entre des variables, et de déterminer les facteurs les plus pertinents dans l'évaluation globale du risque .(Jolliffe, 2011a)

L'ACP permet de réduire la complexité des données tout en préservant l'essentiel de l'information, ce qui favorise une meilleure compréhension des phénomènes complexes, facilite la visualisation et permet de hiérarchiser les actions préventives . (Wold et al., 1987)

La problématique principale est donc la suivante : Comment optimiser l'évaluation des risques dans un environnement de données complexe et fortement corrélées, en recourant à l'ACP comme outil de réduction de dimensionnalité et d'aide à la décision ?

L'objectif du mémoire est d'explorer les voies que peut offrir l'ACP dans le domaine de la sécurité industrielle et environnementale, par une étude de la littérature sur les méthodes d'évaluations des risques, une analyse des méthodes de machine learning, et une prise de cas concrets d'application de l'ACP à des jeux de données représentatifs.

La méthodologie adoptée est de nature quantitative, reposant à la fois sur le traitement statistique, la mise en œuvre d'analyses exploratoires des données et la visualisation des

résultats.

Afin de traiter et d'analyser les données, il sera fait appel à Python et à ses bibliothèques scientifiques (scikit-learn, matplotlib, pandas notamment) . (Pedregosa et al., 2011)

La structure du mémoire est la suivante :

- Le Chapitre 1 introduit le cadre général de l'évaluation des risques.
- Le Chapitre 2 est consacré aux techniques de machine learning modernes appliquées à la gestion des risques.
- Le Chapitre 3 présente les fondements mathématiques et les applications pratiques de l'ACP.
- Enfin, le Chapitre 4 propose une étude de cas sur la mise en œuvre de l'ACP dans le cadre d'un cas réel en milieu industriel suivi d'une conclusion sur les résultats obtenus, les limites de l'approche et les perspectives à envisager.

CHAPITRE I

I. Cadre Théorique de l'Évaluation des Risques :

I.1 Introduction :

D'après l'ISO 31000 (2018), « l'évaluation des risques est le processus global d'identification, d'analyse et d'évaluation des risques » qui doit fournir une base rationnelle à la décision. Elle permet la prise en compte par l'entreprise de la probabilité d'occurrence d'événements indésirables, de l'estimation des conséquences, et de la priorisation des actions.

L'évaluation des risques apparaît comme une condition absolument nécessaire à toute conduite de sécurité industrielle et environnementale. Elle a pour but de repérer les sources possibles de danger, d'en évaluer les conséquences éventuelles et de mettre en œuvre les mesures préventives et protectrices adéquates. Cette démarche préventive s'inscrit dans un souci d'amélioration continue, tant pour diminuer la fréquence et la gravité des événements accidentels que pour répondre à la conformité réglementaire et normative.

Dans un univers industriel de plus en plus automatisé, numérisé et soumis à des enjeux réglementaires croissants, les méthodes d'évaluation des risques doivent évoluer où des volumes de données de plus en plus variés et complexes. L'identification, l'analyse, la hiérarchisation des risques nécessitent des outils rigoureux, dynamiques et capables de traiter une information multidimensionnelle.

L'objectif est de préparer les bases de l'analyse des risques dans tous les domaines, en s'attachant aux définitions, objectifs, étapes, typologie des risques rencontrés dans le milieu industriel, mais aussi à leurs limites et à leurs limites, qui feront les fondements de la réflexion sur la nécessité d'approches moins traditionnelles, d'abord l'Analyse en Composantes Principales (ACP), qui sera traitée dans un chapitre suivant.

I.2 Définitions et concepts clés :

Le risque peut être défini comme la combinaison de la probabilité d'occurrence d'un événement dangereux et de la gravité de ses conséquences (ISO 45001, 2018).

On identifiera le danger comme une source potentielle de dommage ou de préjudice(OIT, 2013), en termes de notions associées nous relevons :

-**Probabilité** : fréquence ou chance pour un événement de se produire.(ISO 31000, 2018)

-**Gravité** : amplitude des conséquences potentielles.(IEC 31010, 2019)

-**Criticité** : produit de la probabilité et de la gravité permettant de hiérarchiser les risques.(Anthony (Tony) Cox Jr, 2008)

Trois grandes approches d'évaluation sont à distinguer :

-**Qualitative** : à partir de jugements d'experts et d'échelles descriptives.(Rausand, 2013)

-Semi-quantitative : avec des matrices de criticité et des scores de risque.(Rausand, 2013)

-Quantitative : basée sur des données chiffrées et des modèles probabilistes.(Rausand, 2013)

I.3 Objectifs de l'évaluation des risques :

L'un des aspects majeurs des dispositifs contemporains de gestion en matière de santé, sécurité et environnement (HSE) est l'évaluation des risques. Exercice réglementaire, elle devient, dans ses visées un enjeu stratégique dans le but de préserver les ressources humaines, matérielles et environnementales dans l'intérêt même de la performance de l'organisation.

S'inscrivant dans le cadre de la Directive européenne 89/391/CEE et des recommandations de l'OIT (2013), les objectifs de l'évaluation des risques peuvent être déclinés :

1-Anticiper les accidents et les incidents : Le 1er objectif est de prévenir les événements non souhaités en s'intéressant à ce qui pourrait en être à l'origine, avant que cela ne se réalise. Cela inclut :

-La détection des situations dangereuses (ex. : présence de câbles électriques souterrains défectueux) L'identification des comportements à risque (ex. : non-respect des règles de consignation) -L'évaluation des conditions environnementales susceptibles de favoriser un accident (ex. : humidité qui favorise le risque d'électrocution) En frappant à un moment où on peut agir, l'évaluation contribue à diminuer très fortement la fréquence et la gravité des accidents de travail.

2-Réduction du nombre de pertes humaines, matérielles, financières :

Un accident industriel peut avoir des conséquences dramatiques des pertes humaines (blessures, incapacités, décès), des pertes matérielles (équipements, infrastructures), des pertes financières (coûts directs : réparations, soins à l'hôpital, coûts indirects : arrêt de production, atteinte à l'image de marque), comme ce fut le cas dans le secteur énergétique où une mauvaise évaluation des risques lors d'interventions sur des câbles souterrains a provoqué une coupure électrique généralisée, entraînant d'importantes pertes économiques. Au contraire, l'évaluation en amont permet de diminuer de manière significative la fréquence et la gravité des accidents du travail.(ISO 31000, 2018)

3-Prévenir les pollutions et dégradations de l'environnement :

La prise en compte des risques entend l'intégration du risque environnemental par l'identification de toute activité générant :

- Des pollutions (rejets de produits chimiques, hydrocarbures).
- Des dégradations d'habitats naturels.
- Des risques de contamination des sols ou nappes phréatiques.

Pour le secteur industriel, cela implique de prévoir des moyens de prévention (barrières physiques, systèmes de confinement) pour limiter les impacts environnementaux.(IEC 31010, 2019)

4-Maximiser la performance HSE et la compétitivité organisationnelle :

Une bonne maîtrise du risque permet de :

- Améliorer la performance HSE en réduisant le nombre d'accidents et d'incidents.
- Augmenter la productivité en limitant les interruptions d'activité liées au sinistre.
- Renforcer l'image de l'entreprise vis-à-vis de ses clients, investisseurs et autorités

Ainsi l'évaluation du risque devient un avantage compétitif car elle contribue à la conformité réglementaire, à la confiance de ses partenaires et à la réduction des coûts de défaillance.

I.4 Processus général d'évaluation des risques :

Le processus d'évaluation des risques est une démarche méthodique et structurée visant à identifier, analyser et traiter les dangers qui peuvent affecter la santé, la sécurité des travailleurs, les biens matériels et l'environnement.

Conformément à la norme ISO 31000:2018 et à la directive européenne 89/391/CEE, ce processus repose sur la mise en œuvre consécutive et itérative d'étapes dont l'interdépendance favorise leur gestion durable et continue :

1-Préparation et planification :

La première étape consiste à définir le périmètre de l'évaluation et à mettre en place les ressources nécessaires :

- Détermination du contexte opérationnel (type d'activité, environnement de travail, contraintes réglementaires).
- Constitution de l'équipe d'évaluation, impliquant HSE, responsables techniques et représentants des travailleurs.
- Choix des méthodes et outils (APR, HAZOP, AMDEC, etc.).

Ex. : dans le cas de l'inspection des câbles souterrains, il convient de prévoir les plans des réseaux, les détecteurs de tension et les équipements de protection individuelle.

2-Identification des dangers :

Cette étape consiste à identifier toutes les sources possibles de nuisances :

- Dangers physiques (électrocution, explosion, chute).

- Dangers chimiques (inhalation des gaz, fuite des hydrocarbures).
- Dangers environnementaux (pluies, inondations, terrains meubles).

Outils utilisés : check-lists, observations sur le terrain, retour d'expérience (REX).

3-Analyse des risques :

Elle vise à comprendre la nature des risques identifiés et à évaluer leur probabilité et leur gravité.

*Probabilité d'occurrence : fréquence estimée de l'événement.

*Gravité des conséquences : impact sur la santé, les biens et l'environnement.

Outils utilisés : grilles de cotation (matrice de criticité).

Exemple : Une rupture de câble haute tension enterré peut avoir une probabilité faible, mais une gravité élevée (accident mortel, coupure massive d'électricité).

4-Évaluation et classification des différents risques :

Les résultats de l'analyse permettent de classer les risques de façon priorisée. L'objectif étant de déterminer les actions à mettre en œuvre en priorité :

- Risques inacceptables : requièrent des mesures immédiates.
- Risques modérés : nécessitent plan d'amélioration.
- Risques complaisants : sous contrôle mais à suivre de près, selon le cas.

5-Approche des risques :

Cette étape consiste à déterminer et à mettre en œuvre des mesures permettant de réduire le risque à un niveau accepté :

- Suppression du danger (ex. : suppression d'un produit dangereux).
- Substitution par un procédé ou matériau moins dangereux.
- Mesures techniques (isolants électriques, barrières physiques).
- Mesures organisationnelles (formation, procédures de travail sécurisées).

6- Suivi et révision :

La phase de suivi et de révision ne saurait être négligée. En effet, elle demeure tout aussi indispensable que la caractérisation initiée lors de la première étape :

- Pour veiller à l'efficacité des actions déployées.

- Pour adapter la gestion des risques, compte tenu des changements techniques ou réglementaires.
- Pour intégrer d'éventuels retours d'expérience.(Hollnagel, 2008)

7-Communication et consultation

Tout au long du processus, il est important d'impliquer les travailleurs et de communiquer clairement sur les dangers identifiés et les mesures prises(OIT, 2013).

I.5 Typologie des méthodes d'évaluation des risques :

Pour évaluer les risques, un panel varié et diversifié de méthodes est mis en œuvre : les méthodes varient selon la manière dont elles sont mises en œuvre, leur degré de précision et leur domaine d'application. Le choix de la méthode est conditionné par la disponibilité de l'information, l'effort mobilisé et les objectifs recherchés (ISO 31010, 2019).

1-Les méthodes qualitatives :

Les méthodes qualitatives s'appuient sur les jugements d'experts, des check-lists ou des scénarios hypothétiques. Elles ne nécessitent pas d'outils de quantification mais permettent d'identifier les principaux dangers en un temps raisonnable(INRS, 2017).

Exemples : Check-lists de sécurité, What-If Analysis, méthode Kinney.

*Avantage : simplicité, rapidité, applicables en phase préliminaire.

*Limite : le risque de subjectivité et de manque de précision(Villemeur, 1991).

2. Les méthodes semi-quantitatives :

Ces méthodes croisent probabilité et gravité des événements dans des grilles de cotation qui permettent d'obtenir une hiérarchisation des risques(ISO 31010, 2019) .

Exemples : Matrice de criticité, AMDEC (Analyse des Modes de Défaillance et de leurs Effets et de leur Criticité).

Avantage : hiérarchisation des risques.

Limite : résultats dépendant des échelles de notation(Ericson, 2015) .

3. Les méthodes quantitatives :

Les méthodes quantitatives recourent à des modèles statistiques et probabilistes en vue d'aboutir à des résultats chiffrés. Ces méthodes, qui nécessitent des données fiables, précises et descriptives, permettent d'obtenir des résultats plus sûrs, précis, objectifs(Rausand and Hoyland, 2003) .

Exemples : arbres de défaillance (Fault Tree Analysis), arbres d'événements (Event Tree Analysis), simulations Monte Carlo, méthodes bayésiennes.

Avantages : objectivité.

Limites : coût élevé et complexité de mise en œuvre.

4. Les méthodes inductives et déductives :

Inductives : Elles partent de causes potentielles pour rechercher les conséquences (ex. AMDEC, What-If) (Villemeur, 1991).

Déductives : elles prennent un événement redouté pour rechercher ses causes (ex. FTA, ETA) (Ericson, 2015).

Mixtes : La méthode Bow-Tie qui fait le lien entre les causes (FTA) et les conséquences (ETA), dans un même dispositif (ISO 31010, 2019).

5. Les méthodes sectorielles spécifiques :

Certains secteurs industriels exigent l'application d'outils adaptés :

- Industrie chimique : HAZOP (Hazard and Operability Study) (Ericson, 2015).
- Nucléaire : PSA (Probabilistic Safety Assessment) (Rausand and Hoyland, 2003).
- Santé au travail : méthode INRS d'évaluation simplifiée (INRS, 2017) .
- Environnement : ACV (Analyse de Cycle de Vie) pour l'évaluation des impacts environnementaux.

I.6 Limites et contraintes des méthodes classiques :

Bien qu'utilisées dans de nombreux secteurs industriels, les méthodes classiques d'évaluation des risques souffrent de différentes limites et contraintes qui nuisent parfois à leur efficacité et à leur applicabilité.

1. Subjectivité et dépendance à la compétence :

Les méthodes utilisées dans l'approche qualitative comme les check-lists ou les matrices de risques reposent sur les jugements d'experts ; la subjectivité des experts peut être à l'origine de biais d'interprétation et de variabilité des résultats (INRS, 2017) . Deux équipes pourront en effet en venir à des conclusions différentes pour une même situation de risque.

2. Difficile traitement des grandes quantités de données :

Dans un contexte de systèmes industriels complexes et de capteurs générateurs de volumes très importants de données, les méthodes traditionnelles ont des difficultés à traiter les

données et à les analyser de manière exhaustive(ISO 31010, 2019). Elles n'ont pas été conçues pour le Big Data ni pour les données dynamiques prises en compte au temps réel.

3. Absence de précision et de quantification :

Si elles sont rapides, les méthodes qualitatives et semi-quantitatives sont peu précises. Les conclusions obtenues sont le plus souvent trop vagues, non comparables et non exploitables dans une démarche d'optimisation(Villemeur, 1991) ; les matrices de criticité font appel à des ordres de grandeur arbitraires, conduisant ici encore à sur- ou sous- estimer les risques.

4. Complexité et coût des méthodes quantitatives :

Les approches quantitatives telles l'arbre de défaillance ou la simulation Monte Carlo procurent des résultats précis, mais exigent une compétence forte, un modèle mathématique différent, des bases de données fiables. Elles peuvent coûter cher et être longues à mettre en œuvre ; ce qui limite leur utilisation, en particulier au sein des PME (Rausand and Hoyland, 2003).

5. Applicabilité limitée à certains secteurs :

Des méthodes sont développées pour des industries spécifiques (ex HAZOP pour la chimie, PSA pour le nucléaire) et leur utilisation pérenne et efficace dans d'autres secteurs peut s'avérer compliquée ou coûteuse(Ericson, 2015).

6. Problèmes liés à la prise en compte des dimensions humaines et organisationnelles :

Pour la plupart des approches classiques, seules les dimensions techniques (probabilité de défaillance, fiabilité des équipements) sont envisagées alors que les dimensions humaines, organisationnelles et culturelles qui sont pourtant au cœur de la genèse des accidents sont laissées de côté (Reason, 2016).

I.7 Conclusion

Pour conclure, si les approches classiques constituent des fondations solides pour évaluer les risques, elles affichent leurs limites face à la complexité croissante des systèmes industriels contemporains. De plus en plus de propositions émergent à partir d'approches innovantes de type machine learning et analyse multivariée de type Analyse en Composantes Principales (ACP/PCA) capables de traiter rapidement de grands volumes de données dans le respect de normes plus robustes et objectives.

CHAPITRE II

II. Les techniques de Machine Learning pour l'évaluation des risques :

II.1 INTRODUCTION :

Ces dernières décennies, les évolutions et la complexité croissantes des systèmes industriels ont élargi l'éventail des systèmes d'information et multiplié les sources de données générées par l'ensemble des capteurs, systèmes d'information et procédés automatisés. Les méthodes classiques, bien que pertinentes dans certains contextes de risques, montrent leurs limites devant le volume, les variétés et les vitesses des données (ISO 31010, 2019).

Dans ce contexte, le Machine Learning (ML), sous domaine de l'intelligence artificielle (IA) s'impose comme un outil performant pour exploiter la valeur des données massives, complexes, hétérogènes grâce à ses capacités à apprendre a priori seules des expériences passées et à chercher des patterns cachés ; rôle qui génère de nouvelles pistes pour anticiper, modéliser et prévenir les risques (Bishop and Nasrabadi, 2006).

Le recours au ML pour l'évaluation des risques a l'avantage de réduire la subjectivité qui en découle par les méthodes classiques et d'améliorer la performance prédictive par l'exploitation de données en temps réel. Ces techniques apparaissent particulièrement efficaces pour détecter les anomalies, estimer les défaillances et optimiser les politiques de maintenance préventive et conditionnelle (Zhang et al., 2020).

Par ailleurs, les possibilités d'application du ML dans le domaine de la gestion des risques, donc de ses marges d'utilité, ne touchent pas seulement les domaines industriels. Mais touchent aussi la sécurité environnementale, la finance, la santé ou bien encore les infrastructures critiques, ce qui témoigne d'un caractère transversal et d'une capacité transformante (Goodfellow et al., 2016).

Ce chapitre se propose donc d'être aussi un état des lieux des principales techniques de Machine Learning mises en œuvre dans l'évaluation des risques, aussi bien en termes de fondements théoriques, typologies et applications concrètes, que d'avantages et de limites.

II.2 Concepts fondamentaux du Machine Learning :

II.2.1 Définition et principes de base :

Le Machine Learning (ML) ou apprentissage automatique est une discipline scientifique de l'intelligence artificielle consacrée à la conception d'algorithmes qui apprennent à partir de données, et améliorent leur performance par expérience.

Arthur Samuel définit le ML comme : « un champ d'étude qui procure aux ordinateurs la faculté d'apprendre sans être explicitement programmés ».(Samuel, 1959)

Mitchell pour sa part, un peu plus tard, présente une définition un peu plus opérationnelle : « Un programme informatique est dit apprendre de l'expérience E, par rapport à une tâche T

et une mesure de performance P, si sa performance dans la tâche T, mesurée par P, s'améliore avec l'expérience E ».(Mitchell, 1997)

Le principe fondamental est donc de tirer parti de l'apprentissage effectué à partir des données ; en exploitant les modèles de données pour extraire des patterns cachés, des corrélations ou structures au sein de grands ensembles de données, souvent complexes et multidimensionnels, s'avérant de nature à faire bénéficier grandement les performances de modèles d'évaluation des risques où les variables à considérer sont nombreuses, hétérogènes, et parfois incertaines.

II.2.2 Différences entre intelligence artificielle, machine learning et deep learning :

Il est nécessaire de bien écrire la différence entre ces trois notions souvent faussement prises ensemble :

-Intelligence artificielle (IA) : discipline englobante visant à créer des systèmes capables de simuler certaines capacités cognitives humaines (raisonnement, perception, décision).

-Machine Learning (ML) : sous-catégorie de l'IA, en particulier celle qui recouvre les algorithmes supposés capables d'apprendre automatiquement et de progresser grâce aux données et à leur expérience.

-Deep Learning (DL) : sous-catégorie du ML, en particulier celle qui utilise des réseaux de neurones de type artificiel et en particulier des réseaux en profondeur, bien adaptés à l'étude de grands ensembles de données souvent massives et non structurées (images, sons, textes).

On voit donc que l'IA constitue le domaine englobant, qui recouvre le ML, qui lui-même recouvre le DL.

II.2.3 Les étapes d'un projet de Machine Learning :

Un projet de ML suit généralement le processus structuré en plusieurs étapes clés qui suivent (Kelleher et al., 2020) c'est-à-dire:

1-Collecte des données :

-Identification des sources de données pertinentes (capteurs, bases de données industrielles, historiques d'accidents, rapports HSE...).

-Qualité et représentativité des données sont préalables pour assurer la fiabilité des résultats.

2-Préparation et prétraitement des données :

-Nettoyage des données (suppression des valeurs aberrantes, gestion des données manquantes).

-Transformation (normalisation, réduction de dimensionnalité, encodage des variables catégorielles).

-Division en ensembles d'entraînement et de test.

3-Modélisation :

-Choix de l'algorithme adapté (régression logistique, arbres de décision, réseaux de neurones, SVM...).

-Entraînement du modèle sur l'ensemble d'apprentissage pour reconnaître les patterns contenus dans les données.

4-Évaluation et validation :

-Mesure de la performance du modèle par des métriques (accuracy, précision, rappel, F1-score, AUC...).

-Validation croisée pour s'assurer de la robustesse, et éviter le surapprentissage (overfitting).

-Ajustement et optimisation des hyperparamètres.

5-Déploiement et suivi : intégration dans le système de gestion des risques et amélioration continue.

Ce cycle peut être itératif : les résultats obtenus lors de l'évaluation peuvent produire des modifications dans la collecte ou le traitement des données, ou la recherche des autres algorithmes.

Les méthodes de machine learning comme les réseaux de neurones, les forêts aléatoires ou le support vector machine se sont révélées particulièrement efficaces pour la prédiction des défaillances et l'optimisation des plans de maintenance. (Khan and Haddara, 2020)

Ces démarches garantissent que le modèle élaboré est solide, adaptable et bénéfique dans un cadre de gestion des risques.

II.3 Classification des méthodes de Machine Learning :

Le Machine Learning est constitué d'un ensemble de méthodes qui sont classées en plusieurs catégories selon la nature des données disponibles et les buts de l'analyse. Ce classement est essentiel pour guider le choix de l'algorithme le plus pertinent en fonction des enjeux d'évaluation et de gestion des risques dans le domaine industriel.

II.3.1 Apprentissage supervisé :

L'apprentissage supervisé est basé sur des données étiquetées (inputs avec sorties connues) au sein desquelles l'algorithme apprend une relation entre les variables explicatives et la variable cible.

Dans le champ de la sécurité industrielle, cette catégorie d'algorithmes permet notamment de prédire la probabilité d'une défaillance d'un équipement à partir d'un historique de données de maintenance (Khan and Haddara, 2020).

Exemples d'algorithmes :

- Régression logistique : pour des prévisions binaires (accident vs pas d'accident).
- Forêts aléatoires (Random Forest) : utiles pour traiter des bases de données complexes et bruitées.
- Machines à vecteurs de support (SVM) : permettent une séparation de classes dans un environnement industriel multidimensionnel.

Les techniques supervisées dominent en matière d'application du machine learning à l'évaluation des risques, car elles s'emploient à anticiper les comportements des événements à partir des expériences passées.(Wei et al., 2023)

II.3.2 Apprentissage non supervisé :

À l'inverse de l'apprentissage supervisé, cette catégorie n'a pas besoin de données étiquetées. L'objectif est de découvrir des structures ou des regroupements cachés dans les données.

Exemples :

- Clustering (k-means, DBSCAN) : classification les incidents en groupes homogènes, dans le but d'identifier les scénarios de risques dominants.
- Analyse en composantes principales (ACP) : réduction de la dimensionnalité des données sans perte d'information, ce qui favorise l'analyse de grands ensembles de données sur les risques.

Les méthodes non supervisées, sont tout à fait adaptées pour identifier des schémas de risques émergents, le plus souvent non détectés dans les analyses classiques.(Zhou et al., 2021)

II.3.3 Apprentissage semi-supervisé :

Sur de nombreux sites industriels, il est difficile d'obtenir des bases de données totalement étiquetées (ex. : accidents rares mais aux conséquences graves) : l'apprentissage semi-supervisé présente donc un grand intérêt car il combine un nombre réduit de données étiquetées à un grand volume de données non étiquetées de façon à améliorer la performance prédictive. (Chapelle et al., 2006)

Application courante : prédire les scénarios d'accidents majeurs quand il n'existe pas beaucoup de données historiques mais de nombreuses variables fonctionnelles sont disponibles.

II.3.4 Apprentissage par renforcement :

Cette approche repose sur l'interaction de l'agent intelligent avec un environnement. L'agent apprend par essais et erreurs à maximiser une récompense finale.

Dans le domaine de la gestion des risques, l'apprentissage par renforcement peut permettre d'optimiser les choix d'une stratégie de maintenance prédictive et la configuration d'un système de contrôle industriel adaptatif.(Sutton and Barto, 1998)

Exemple d'application : un agent peut apprendre à ajuster en continu la fréquence d'inspections d'un équipement à risque pour réduire les pannes (non anticipation de l'arrêt de l'équipement) pour des coûts au plus bas.

II.3.5 Deep Learning :

Le Deep Learning est une sous-catégorie du ML reposant sur des réseaux de neurones artificiels comportant plusieurs couches. Ces techniques permettent de traiter des volumes importants de données complexes (images, signaux, textes).

Dans le cas de la sécurité industrielle, les réseaux de neurones convolutionnels (CNN) peuvent être utilisés pour détecter des anomalies visuelles (fuites, fissures), alors que les réseaux récurrents (RNN, LSTM) sont utilisés pour prédire des séries temporelles de risques.(Goodfellow et al., 2016)

II.4 Applications du Machine Learning à l'évaluation des risques :

La montée en charge du Machine Learning (ML) a modifié l'évaluation des risques et leur gestion dans de nombreux secteurs industriels et environnementaux. Le ML, en tant que modèle statistique, est utile lui aussi. Il s'affranchit de méthodes basées sur des postulats rigides et permet de traiter des bases de données d'une ampleur inédite en explorant des données systématiquement et sans a priori à la recherche des motifs sous-jacents, des catastrophes qui peuvent survenir et de solutions adaptées.(Xu et al., 2021)

II.4.1 Maintenance prédictive et détection des défaillances :

Les modèles de ML en particulier, à partir des réseaux de neurones ou des forêts aléatoires, permettent de prédire des pannes qui peuvent survenir dans les équipements industriels.

Exemple : dans le secteur énergétique, pour anticiper les défaillances des transformateurs électriques ou dans le cas, récurrent, de détection des anomalies dans le fonctionnement des turbines.(Zhang et al., 2019)

Conséquence : une baisse des arrêts non programmés, une réduction des coûts de maintenance, la sécurité des travailleurs.

II.4.2 Gestion des risques environnementaux :

L'utilisation du ML permet de modéliser et de prévoir les risques liés à la pollution de l'air, aux fuites de produits chimiques ou aux catastrophes naturelles.

Exemple : les algorithmes de régression et SVM ont été appliqués à la prévision de la dispersion des polluants atmosphériques dans les zones industrielles.(Zhang et al., 2020)

Conséquence : amélioration de la planification des mesures de prévention et réduction des impacts sanitaires.

II.4.3 Sécurité des procédés industriels :

Le ML permet d'identifier en temps réel des conditions anormales dans les procédés chimiques et pétrochimiques.

Exemple : les réseaux de neurones profonds permettent de détecter précocement des réactions dangereuses dans les unités de raffinage.(Li et al., 2021)

Conséquence : amélioration de la réactivité et réduction du risque d'accidents majeurs.

II.4.4 Analyse des risques financiers et organisationnels :

Le ML est également appliqué dans l'analyse des risques financiers et organisationnels dans le cadre de la gestion de l'entreprise, par exemple pour les systèmes d'assurance industrielle.

Exemple : les modèles de ML permettent de détecter des comportements anormaux et ainsi de prévenir les fraudes dans les systèmes d'assurance liés aux accidents de travail.(Kou et al., 2021)

II.4.5 L'évaluation des risques en santé et sécurité au travail (HSE) :

La mise en exploitation des données des capteurs IoT et des systèmes de vidéo-surveillance par le biais du ML permet d'identifier les comportements à risque.

Exemple : identification automatique des gestes à risque (absence d'EPI, proximité de zones interdites) opérée par les algorithmes de vision par ordinateur, d'après.(Fang et al., 2019)

Conséquence : affirmation de la culture de prévention et réduction des accidents.

II.5 Limites et défis du Machine Learning en évaluation des risques :

Malgré leurs avantages, ces applications rencontrent plusieurs défis :

- la disponibilité et la qualité des données (sensibilisation, pannes, contamination des données) ;
- la complexité des modèles qu'ils rendent difficiles à s'approprier par les responsables HSE ;

- le besoin de s'intégrer au système d'information des organisations, notamment des logiciels de gestion (ERP, GMAO, systèmes SCADA).

Pourtant, les tendances actuelles s'orientent vers une association croissante du ML à d'autres modalités, et notamment à l'intelligence artificielle explicable (XAI) et à l'analyse en composante principale (ACP) pour plus d'interprétabilité et de confiance dans leurs résultats.

II.6 CONCLUSION :

En conclusion, le Machine Learning ouvre de nouvelles perspectives pour la gestion proactive et intelligente des risques, mais sa mise en œuvre exige une approche prudente, intégrant la transparence, l'éthique, la collaboration interdisciplinaire et le respect des cadres réglementaires. Ces conditions sont essentielles pour transformer les avancées technologiques en véritables leviers d'amélioration de la sécurité industrielle et environnementale.

CHAPITRE III

III. Réduction de Dimension et Analyse en Composantes Principales (PCA) :

III.1. Introduction :

Dans le cadre actuel de l'explosion des données, les organisations industrielles et environnementales disposent d'ensembles d'informations de plus en plus volumineux, de plus en plus complexes et de plus en plus multidimensionnels. Ce Big Data a un potentiel considérable pour l'évaluation et la prévention des risques. Toutefois, ce Big Data suppose de composer avec des enjeux souvent limitants, à savoir la redondance, le bruit, et la malédiction de la dimensionnalité (curse of dimensionality) qui plombent la qualité et la performance des modèles de Machine Learning (Bellman, 1961; Verleysen and François, 2005). À ce titre, la réduction de dimension apparaît comme une étape incontournable dans les processus de traitement et d'analyse de données.

En effet, celle-ci permet de transformer un ensemble variablement entaché d'un grand nombre de variables sous la forme d'un espace de dimension réduite tout en préservant la qualité de l'information (Jolliffe and Cadima, 2016). L'objectif n'est pas seulement de simplifier les calculs, mais également de faciliter l'interprétation, d'atténuer les risques d'overfitting et d'optimiser la robustesse des modèles prédictifs.

Parmi toutes les méthodes, l'Analyse en Composantes Principales (ACP ou PCA) est celle de référence. Introduite par Karl Pearson (1901) et poursuivie par Harold Hotelling (1933), la PCA repose sur la découverte de nouvelles variables appelées composantes principales qui sont des combinaisons linéaires des variables initiales et visent à extraire une variance maximum afin de mettre en évidence les structures latentes et relations latentes.

Dans le champ de l'évaluation des risques, la PCA permet :

- la détection des variables explicatives majeures dans un grand nombre de variables d'indicateurs ;
- la mise en visibilité des risques dans des espaces de plus faible dimension ;
- voire leur combinaison avec des algorithmes de Machine Learning (SVM, réseaux de neurones, arbres de décision) à des fins de performance prédictive (Abdi and Williams, 2010; Jolliffe, 2011b).

Le présent chapitre a pour objet l'analyse des méthodes de réduction de dimension avec une attention spécifique portée sur l'Analyse en Composantes Principales. Après une présentation des concepts et enjeux de la réduction de dimension, nous présenterons les méthodes principales de réduction de dimension, avant d'aborder plus précisément la PCA, son principe, ses applications et ses limites, dans le cadre de l'évaluation des risques.

III.2 Réduction de dimension :

III.2.1 Définition :

La réduction de dimension est un processus statistique et computationnel permettant de transformer un ensemble de variables initiales en un nombre réduit de variables, tout en gardant l'essentiel de l'information (Jolliffe and Cadima, 2016). Elle est particulièrement utile dans les contextes où l'information est rendue complexe par des données en grande dimension, ce qui risque d'affecter les performances computationnelles, interprétatives et modélisationnelles.

On distingue deux approches principales :

- La sélection de variables (feature selection) : consiste à choisir un sous-ensemble pertinent de variables initiales, en se débarrassant de celles qui sont redondantes, peu informatives ou fortement inter-corrélées.
- La transformation de variables (feature extraction) : repose sur la construction de nouvelles variables à partir de combinaisons des variables d'origine dans le cas de l'Analyse en Composantes Principales (PCA) ou de l'Analyse factorielle en facteurs communs (FA).

III.2.2 La malédiction de la dimensionnalité :

La malédiction de la dimensionnalité, formulée par Bellman , désigne les difficultés liées à l'analyse de données dans des espaces de haute dimension (Bellman, 1961). Plus le nombre de variables augmente, plus :

- la densité des données diminue, ce qui complique la modélisation statistique ;
- les calculs et temps de traitements deviennent coûteux ;
- le risque de sur-apprentissage (overfitting) par les modèles, ceux-ci apprenant plus du bruit que des relations significatives (Verleysen and François, 2005).

La réduction de dimension permet ainsi de lutter contre ces effets négatifs en repassant les données dans un espace plus compact et plus exploitable.

III.2.3 Enjeux et avantages de la réduction de dimension :

La réduction de dimension s'inscrit dans plusieurs enjeux stratégiques pour les enjeux évalués et le risque géré :

1/Amélioration de la performance des modèles :

Diminuer le nombre de variables entraîne une meilleure utilisation des algorithmes de Machine Learning qui nécessitent moins de ressources de calcul et convergent plus rapidement.

Elle permet aussi d'améliorer la précision des prévisions, en écartant les variables non pertinentes (Guyon and Elisseeff, 2003).

2/Prévention du sur-apprentissage (overfitting) :

Les modèles construits sur un trop grand nombre de variables se trouvent trop adaptés aux spécificités de l'échantillon de formation en réduisant leur potentiel de généralisation.

La réduction de dimension permet de mieux saisir les tendances globales.

3/Facilitation de l'interprétation :

Dans le secteur industriel comme dans le secteur environnemental, la prise de décision exige une compréhension des facteurs de risque.

Les méthodes comme la PCA permettent, en représentant les données en 2 ou 3 dimensions, de rendre les résultats plus accessibles aux décideurs.

4/Robustesse et gestion du bruit :

En supprimant les variables redondantes ou corrélées, le risque que les modèles soient « dérangés » par des bruits parasites est limité.

5/Amélioration de la collecte et du stockage des données :

Dans un contexte Big Data, réduire le nombre de dimensions est un moyen de faciliter la gestion des bases de données et d'économiser des ressources (Aggarwal, 2018).

III.2.4 Importance dans l'évaluation des risques :

Au sein d'une évaluation des risques, la réduction de dimension permet de :

- identifier ce qui joue réellement un rôle parmi des indicateurs complexes (industriels, financiers, environnementaux).
- simplifier la modélisation des scénarios de risques.
- augmenter la pertinence des modèles prédictifs en focalisant l'analyse sur l'information utile.

Ainsi, elle est considérée comme une étape clé de la préparation des données pour l'application des méthodes à visée prédictive avec statistiques et Machine Learning, en lien avec la gestion des risques.

III.3 Les principales méthodes de réduction de dimension :

La réduction de dimension est une des étapes préliminaires dans l'analyse de données massives et complexes. Plusieurs méthodes ont été mises au point, avec leurs avantages et leurs inconvénients en fonction du type de données à traiter et de l'objectif. Ces méthodes peuvent être classées principalement en deux grandes catégories : les méthodes linéaires et les méthodes non linéaires.

III.3.1 Méthodes linéaires :

Les méthodes linéaires ont pour but de projeter les données d'un espace de grande dimension vers un espace plus réduit par l'intermédiaire de combinaisons linéaires des variables d'origine.

a) Analyse en Composantes Principales (PCA) :

Principe : Transforme les variables initiales en un ensemble de nouvelles variables (les composantes principales), qui sont non corrélées et maximisent la variance expliquée (Jolliffe and Cadima, 2016).

Avantage : Simplicité, efficacité computationnelle, bonne interprétabilité.

Limites : Hypothèse de linéarité, perte d'informations si la structure est non linéaire.

b) Analyse Discriminante Linéaire (LDA) :

Principe : Réduit la dimensionnalité en maximisant la séparabilité entre classes tout en minimisant la variance intra-classes (Fisher, 1936).

Application : Très utilisée pour la classification supervisée et pour la reconnaissance de formes.

Limite : Hypothèses de normalité et d'égalité des matrices de covariance entre classes.

c) Analyse factorielle (FA) :

Principe : Décompose la variance des variables observées en une partie expliquée par les facteurs latents et une partie due aux erreurs spécifiques.

Application : Identifier des dimensions cachées à l'influence de plusieurs variables.

Limite : Forte dépendance au choix du modèle, hypothèses statistiques contraignantes.

III.3.2 Méthodes non linéaires :

Elles cherchent à débusquer une structure complexe dans les données que les méthodes linéaires n'arrivent pas à débusquer.

a) t-Distributed Stochastic Neighbor Embedding (t-SNE) :

Principe : Il réduit la dimension en préservant la proximité locale entre points (Maaten and Hinton, 2008).

Avantage : exquises visualisations en 2D et 3D.

Limites : très couteux en temps de calcul et sensible au choix des hyperparamètres.

b) Isomap :

Principe : C'est une extension de l'ACP qui tient en compte des distances géodésiques entre les points dans un espace non linéaire (Tenenbaum et al., 2000).

Avantage : conservation de la structure globale des données.

Limites : sensible au bruit et au choix du paramètre de voisinage.

c) Locally Linear Embedding (LLE) :

Principe : approximation de chaque point comme une combinaison linéaire de ses voisins proches et projection dans un espace de faible dimension en préservant ces relations (Roweis and Saul, 2000).

Avantage : il capte bien les structures locales.

Limites : perd l'information globale et souffre du bruit.

III.3.3 Sélection de variables (Feature Selection) :

La sélection de variables consiste à retenir un sous-ensemble des variables d'origine jugées pertinentes, contrairement aux méthodes de projection (ou nouvelle base) qui visent à obtenir un nouvel ensemble de variables si positionnées et calibrées qu'elles modélisent le mieux le phénomène attendu.

Les méthodes filtrées reposent sur des critères statiques (corrélations, tests du χ^2 , information mutuelle).

Les méthodes intégrées à la démarche de modélisation reposent au contraire sur l'utilisation de méthodes d'apprentissage (arbres de décision, Lasso, Random Forest) conduisant à une sélection automatique.

Les avantages principaux sont la simplicité d'interprétation et la réduction de la redondance. En revanche, cela peut également mettre à jour des interactions parfois complexes entre variables.

III.3.4 Comparaison et choix de la méthode :

Lorsque les données sont linéaires, une PCA (principal component analysis) ou une LDA (linear discriminant analysis) s'imposera sans aucune difficulté, et pourtant relativement rapide et assez bien établie. Pour des données plus complexes, t-SNE, Isomap ou LLE sont beaucoup plus appropriés, mais sont bien plus computationnellement coûteux. Si bien que lorsque l'interprétation des variables est particulièrement complexe, la méthode la mieux adaptée est alors celle de la sélection, plus qu'une projection.

En pratique, on se pose alors la question du type de données à traiter : numériques, de classe, bruitées ou corrélées ; de ce que l'on veut en faire : visualisation, (classification, prédiction, interprétation) ; et bien sûr de la place laissée à la méthode (taille des données, temps de calcul).

III.4 Analyse en Composantes Principales (PCA) :

III.4.1 Définition et objectifs :

L'Analyse en Composantes Principales (PCA, Principal Component Analysis) est une méthode statistique qui a d'abord été introduite par Karl Pearson (1901) puis popularisée par Harold Hotelling (1933).

Elle consiste à transformer l'ensemble de variables corrélées en un nouveau système de variables non corrélées (i.e. les composantes principales) obtenues comme des combinaisons linéaires des d'origine (Jolliffe and Cadima, 2016).

Les principaux objectifs de la PCA sont de :

- Réduire la dimension d'un jeu de données tout en conservant l'essentiel de l'information (variance).
- Éliminer la redondance entre variables corrélées.
- Faciliter la visualisation des données multidimensionnelles en 2D ou 3D.
- Améliorer la performance des modèles prédictifs, en réduisant le bruit et le surapprentissage.

III.4.2 Principe de fonctionnement :

La PCA a pour base de rechercher de nouvelles dimensions (axes) maximisant la variance des données.

La première composante principale (PC1) est la direction qui explique le plus de variance.

La deuxième composante (PC2) est orthogonale à PC1 et explique la plus grande part possible de la variance résiduelle.

Les suivantes (PC3, PC4, ...) suivent la même logique.

Chaque observation est ainsi représentée par ses coordonnées projetées sur ces axes principaux.

III.4.3 Étapes de la PCA :

a) Standardisation des données :

Puisque la PCA est sensible aux échelles, les données doivent être centrées et réduites :

$$\chi_{ij}^* = \frac{\chi_{ij} - \mu_j}{\sigma_j}$$

(3.1)

où χ_{ij} est la valeur de la variable j pour l'observation i , μ_j sa moyenne et σ_j son écart-type.

b) Calcul de la matrice de covariance :

La matrice de covariance (ou de corrélation) décrit les dépendances linéaires entre variables :

$$\Sigma = \frac{1}{n-1} X^T X$$

(3.2)

c) Décomposition spectrale :

On calcule les valeurs propres (λ_k) et vecteurs propres (U_k) de la matrice de covariance :

- Les vecteurs propres définissent les directions des composantes principales.
- Les valeurs propres indiquent la variance expliquée par chaque composante.

d) Projection des données :

Les nouvelles coordonnées (scores) sont obtenues par :

$$Z = XV$$

(3.3)

où V est la matrice des vecteurs propres.

e) Choix du nombre de composantes :

On retient généralement les k premières composantes expliquant un pourcentage suffisant de variance (souvent 80–95 %).

Ce choix peut être guidé par :

- le critère de Kaiser ($\lambda > 1$).
- le scree plot (test du coude).
- la proportion cumulée de variance expliquée.

III.4.4 Exemple schématique :

Considérons un jeu de données avec deux variables fortement corrélées (ex : taille et poids).

La PCA calcule un axe principal (PC1) résumant la variation commune, et un second (PC2) captant la variance résiduelle.

En ne gardant que PC1, nous réduisons la dimension de $2D \rightarrow 1D$ tout en conservant la plus grande part de l'information.

III.4.5 Applications de la PCA :

- Analyse exploratoire : visualisation de données multidimensionnelles.
- Prétraitement en Machine Learning : réduction de bruit, diminution de la colinéarité.
- Finance : analyse des risques et identification des facteurs sous-jacents (ex : modèles factoriels).
- Industrie, environnement : suivi de qualité, détection d'anomalies, analyse des polluants.

-Bio-informatique : réduction de l'espace de gènes dans les analyses omiques.

III.5 Applications de la PCA à l'évaluation des risques :

III.5.1 Fonction de la PCA dans l'évaluation des risques :

L'évaluation des risques est souvent l'analyse d'un gros ensemble de variables : des facteurs techniques, organisationnels, environnementaux ou financiers. La PCA intervient comme outil de réduction de dimension pour :

- Identifier les facteurs de risque dominants à l'aide d'un grand nombre d'indicateurs.
- Éliminer la redondance issue de corrélations entre variables.
- Simplifier l'interprétation des profils de risque à l'aide des projections sur quelques axes.
- Faciliter la détection d'anomalies, avec l'accent mis sur les observations atypiques (outliers).

III.5.2 Applications sectorielles :

a) Risques financiers :

Dans le domaine bancaire et assurantiel, les portefeuilles sont influencés par de nombreux indicateurs (taux d'intérêt, inflation, volatilité boursière, etc.).

La PCA permet de réduire les dimensions des séries financières et de dégager des facteurs de marché latents.

Exemple : modélisation du risque de crédit en identifiant les principales sources de variance expliquant la probabilité de défaut (Jackson, 2005).

b) Risques industriels et HSE (Hygiène, Sécurité, Environnement) :

Dans les industries à risques (énergie, chimie, BTP), de multiples indicateurs de sécurité sont collectés : incidents, taux de fréquence, exposition aux produits chimiques, etc.

La PCA sert à agréger ces indicateurs et à identifier les composantes principales expliquant les variations du niveau de risque.

Exemple : dans la gestion des installations industrielles, elle aide à détecter des combinaisons de paramètres critiques annonciateurs d'accidents (Yang et al., 2020).

c) Risques environnementaux :

Les bases de données environnementales incluent souvent des centaines de mesures (polluants atmosphériques, qualité de l'eau, paramètres climatiques).

La PCA permet de réduire la dimensionnalité et d'identifier les sources majeures de pollution.

Exemple : dans le suivi de la qualité de l'air, la PCA met en évidence les polluants dominants et leurs corrélations avec les conditions météorologiques (Jolliffe, 2011b).

d) Risques sanitaires et épidémiologiques :

Dans les études de santé publique, des données multidimensionnelles (symptômes, facteurs de comorbidité, expositions environnementales) sont analysées.

La PCA est utilisée pour détecter des profils de patients à risque élevé.

Exemple : analyse des facteurs de propagation d'une épidémie en réduisant la complexité des variables sociales et biologiques (Vyas and Kumaranayake, 2006).

III.5.3 Avantages de la PCA pour l'évaluation des risques :

L'Analyse en Composantes Principales (PCA) permet d'obtenir un profit méthodologique et technique lorsqu'on fait appel à son utilisation dans le champ de l'évaluation des risques ; dans le secteur industriel, financier, environnemental notamment.

a) Réduction de la complexité des données :

Les systèmes de gestion des risques présentent souvent une pléthore de variables interconnectées (taux d'incidents, facteurs de sécurité, information financière, indicateurs environnementaux). La PCA agrège l'information en quelques composantes principales, simplifiant ainsi la dimensionnalité des données sans abandonner l'écrasante majorité de la variance. C'est donc une voie d'analyse et de rapidité de calcul appropriée, dans des situations de Big Data notamment.

b) Détection des corrélations cachées :

La PCA, en sélectionnant les variables fortement corrélées, révèle les schémas qui, en général, demeurent cachés dans la raw data. Cette qualité est utile dans le cadre de risques complexes n'impliquant pas un facteur unique (des facteurs humains, techniques et organisationnels notamment).

c) Aide à la visualisation et à l'interprétation :

La PCA réduit l'information à deux ou trois composantes principales où il devient aisé d'interpréter des traits de tendances, des clusters ou encore des anomalies. Cela facilite la communication des résultats tant aux responsables qu'aux décideurs (en particulier officiels HSE).

d) Amélioration de la performance des modèles prédictifs :

La PCA sert souvent d'étape de prétraitement avant la mise en œuvre de techniques de Machine Learning (SVM, régressions logistiques, réseaux de neurones).

En éliminant le bruit et la redondance, elle contribue à accroître la précision et la robustesse des modèles de prédiction de risques.

e) Généralisation et robustesse :

La PCA permet de réaliser des indicateurs synthétiques de risques, utilisables dans des contextes très variés.

C'est une méthode robuste et très rigoureuse sur le plan mathématique, qui peut être appliquée tant à des données financières (risques de marché, de crédit) qu'à des données industrielles (fiabilité des équipements, sécurité au travail).

III.6 Limites et alternatives de la PCA :

III.6.1 Limites de la PCA :

Bien que l'Analyse en Composantes Principales (PCA) soit une méthode largement utilisée et performante, il existe de nombreuses limites potentielles à son utilisation dans le cadre de l'évaluation du risque :

1/Hypothèse de linéarité :

La PCA repose sur des combinaisons linéaires de variables uniquement. Elle ne peut discerner les relations non linéaires entre les différents facteurs de risque, ce qui peut nuire à la pertinence des résultats dans des systèmes complexes (phénomènes industriels, financiers, biologiques par exemple).

2/Sensibilité aux valeurs aberrantes (outliers) :

Les observations atypiques peuvent fortement influencer les résultats des composantes principales, ce qui rend leur interprétation délicate.

3/Interprétation difficile :

Les composantes principales sont des agrégats mathématiques de variables, souvent difficilement traduisibles en termes opérationnels ou stratégiques pour la gestion du risque.

4/Perte d'information :

En ne retenant que quelques variables, on peut négliger certaines variantes faibles mais significatives de l'information.

5/Normalisation obligatoire :

La PCA est très sensible aux différences d'échelle entre variables (ex : « taux d'accidents » exprimé en % et « pertes financières » en millions). La normalisation s'impose comme une nécessité préalable, ce qui influence alors les résultats de la PCA.

III.6.2 Alternatives à la PCA :

Pour contourner les limitations, plusieurs approches alternatives (ou complémentaires) ont été formulées pour la réduction de dimension :

1/Kernel PCA (KPCA) :

Extension de la PCA utilisant des fonctions noyau pouvant capturer des relations non linéaires. Elle est adaptée à la modélisation de données complexes : signaux, images, comportements de marché. (Schölkopf et al., 1998)

2/Analyse Factorielle (FA) :

Proche de la PCA mais plaidant pour l'orientation de l'analyse vers l'identification de facteurs latents expliquant les corrélations entre variables. Elle est pratique en psychologie, santé et sciences sociales pour modéliser des variables cachées.

3/Méthodes de type Manifold Learning :

Dont t-SNE (t-distributed Stochastic Neighbor Embedding) et Isomap, qui visent à conserver la structure géométrique non linéaire des données. Elle est pratique pour détecter les anomalies complexes ou la visualisation de données de risques multidimensionnelles. (Maaten and Hinton, 2008; Tenenbaum et al., 2000)

4/Auto encodeurs (réseaux de neurones) :

Modèles d'apprentissage profond réduisant les données d'entrées en un espace latent réduit puis les reconstruit. Ils dessinent de manière relationnelle non linéaire leurs sous-ensembles de données, contrairement à la PCA. Opportunité d'être appliqués à de grandes bases de données industrielles ou financières. (Hinton and Salakhutdinov, 2006)

5/Independent Component Analysis (ICA) :

L'objectif n'est plus de séparer des signaux qui n'ont plus aucune corrélation, mais des signaux indépendants statistiquement, plutôt que orthogonaux. Il est intéressant pour l'analyse de risques impliquant des sources multiples. (Hyvärinen et al., 2001)

III.7. Conclusion :

Dans le présent chapitre, nous avons pu montrer en quoi la réduction de dimension peut être considérée comme une étape fondamentale dans l'analyse et l'exploitation des importants volumes d'informations notamment en matière d'évaluation des risques. À l'heure où la quantité de données générées par les systèmes d'information industriels, environnementaux et

organisationnels est en forte croissance, la réduction de dimension permet non seulement de simplifier le modèle et d'optimiser les temps de calcul, mais également d'enrichir la qualité de l'interprétation des résultats par la mise en évidence des variables les plus pertinentes (Guyon and Elisseeff, 2003b).

Parmi les différentes méthodes existantes, l'Analyse en Composantes Principales (PCA) reste le modèle phare, en tant qu'elle permet de réduire un grand nombre de variables corrélées en un petit nombre de composantes indépendantes capables de restituer l'essentiel de l'information (Abdi and Williams, 2010; Jolliffe, 2011b). Nous avons mis en avant plusieurs avantages de la PCA pour la pratique de l'évaluation des risques (détection facilitée des facteurs de risque majeurs, visualisation améliorée des données multidimensionnelles, meilleure robustesse des modèles prédictifs...).

Nous avons pourtant mis en exergue ses limites, en particulier sa dépendance à la linéarité ou la difficulté à traiter des données fortement non linéaires ou bruitées. Pour pallier toutes ces insuffisances, des méthodes alternatives et complémentaires sont mises au point : la Kernel PCA, l'Analyse en Composantes Indépendantes (ICA), les algorithmes de réduction non linéaire (Isomap, t-SNE), les autoencodeurs profonds, entre autres (Hinton and Salakhutdinov, 2006; Maaten and Hinton, 2008; Schölkopf et al., 1998) .

En glossant pour conclure, la PCA est une méthode incontournable dans toute démarche d'analyse multivariée, mais qui doit être contextualisée et complétée d'autres approches chaque fois que les données l'exigent. Elle constitue ainsi une entrée possible pour développer des outils modernes d'analyse de risques assistée par intelligence artificielle et favorise ainsi une meilleure gestion des incertitudes et des vulnérabilités dans les systèmes complexes.

CHAPITRE IV

IV. Étude de cas et Application de la PCA à l'évaluation des risques :

IV.1 Introduction :

L'évaluation et la gestion des risques représentent des enjeux majeurs pour mieux réduire la probabilité d'occurrence des accidents du travail et protéger les salaires dans le secteur industriel. Du fait de la complexité des environnements de travail qui sont impliqués, avec leurs variations multifactorielles de nature technique, organisationnelle et humaine dans une approche dynamique, il est nécessaire d'avoir recours à des outils statistiques à même de simplifier, agréger et traiter des ensembles de données très volumineux dans le cas de données multidimensionnelles (Kabir and Papadopoulos, 2019).

L'Analyse en Composantes Principales (ACP) est parmi les plus connues des méthodes multivariées qu'on utilise ici pour ce type d'analyse. Elle permet une réduction de la dimension des données tout en préservant l'essentiel de la structure de leur variance, et aider à savoir lesquelles des variables sont les plus déterminantes pour l'explication des phénomènes observés (Jolliffe and Cadima, 2016). L'ACP appliquée à la sécurité en entreprise constitue un outil pertinent pour dégager des facteurs de risque dominants pour mieux orienter les choix de décisions en matière de prévention et de contrôle des risques (Abdelgawad and Fayek, 2010).

Dans le cadre de ce chapitre, il s'agit de répondre à la problématique de l'application de l'ACP à l'évaluation des risques industriels à travers une étude de cas. Dans un premier temps, la base de données utilisée puis les étapes du prétraitement des données sont présentées avant une application de l'ACP afin de projeter les données dans un espace de dimension inférieure. Les résultats obtenus vont permettre, d'une part, de mettre en évidence les composantes principales permettant d'expliquer la variabilité des risques et d'autre part de donner une signification pratique de ces résultats dans le domaine de la gestion de la sécurité.

IV.2 Présentation des données :

Dans le cadre de cette recherche, nous avons sollicité la base de données Industrial Safety and Health Analytics Database qui centralise des données caractérisant les accidents professionnels et les caractéristiques de l'environnement de travail associé. Cette base de données librement accessible a été largement sollicitée dans les travaux en santé et sécurité au travail pour construire des modèles prédictifs et apprécier les facteurs de risques (Stefanini and I. H. M, 2021).

Cette base comporte des types variés de variables :

- Informations relatives à l'accident : pays, occurrence, date, secteur d'activité.
- renseignement sur l'accident : type d'accident, niveau de gravité (Accident Level), niveau potentiel de gravité (Potential Accident Level).

- caractéristiques organisationnelles et techniques : description de l’activité, conditions de travail, état des équipements.
- conséquences : effet sur la santé et sécurité des travailleurs.

Dans le cadre de cette étude, les données textuelles ont été encodées à l’aide de la méthode « One-Hot Encoding », afin de transformer les variables qualitatives en variables numériques exploitables par l’ACP. De plus, certaines variables ordinales telles que Accident Level et Potential Accident Level ont été reclassées sous forme numérique (de 1 à 5 et de 1 à 6 respectivement), ce qui permet de préserver leur caractère hiérarchique.

Enfin, les données ont été normalisées (méthode StandardScaler) afin d’assurer une échelle comparable entre toutes les variables, condition indispensable pour l’application de l’ACP (Hair, 2009).

Cette base de données constitue donc un support pertinent pour analyser la variabilité des risques et identifier les facteurs explicatifs dominants à travers l’approche multivariée.



Figure (4.1) : le lien de la base de donner (scanner le code QR)



Figure (4.2) : le lien de la base de donner encodée (scanner le code QR)



Figure (4.3) : le lien de notre notebook sur google colab (scanner le code QR)

IV.3 Prétraitement des données :

Avant d'effectuer l'Analyse en Composantes Principales (ACP), il est nécessaire de préparer les données pour en assurer la robustesse et la fiabilité des résultats. Le prétraitement est une phase cruciale dans toute étude statique, et en particulier dans l'analyse multivariée (Jolliffe and Cadima, 2016).

IV.3.1 Data cleaning :

Exclusion des colonnes non utiles au traitement (ex. description textuelle libre des accidents).

IV.3.2 Encodage des variables catégorielles :

Les variables qualitatives sont ici Countries, Local ou Industry Sector, qui ont été transformées par la méthode du One-Hot Encoding (ou Dummy Coding) pour lesquelles des variables binaires (0/1) sont créées, afin de représenter chaque modalité.

Cette opération permet alors d'après les travaux de rendre exploitable les variables catégorielles par les algorithmes statistiques dont l'ACP, et qui ne gèrent que des données numériques (James et al., 2013).

IV.3.3 Transformation des variables ordinales :

Les variables ordinales Accident Level, qui comprend cinq niveaux (I à V), et Potential Accident Level, qui comprend six niveaux (I à VI), ont été codifiées en valeurs entières globales croissantes (1 à 5, et 1 à 6).

Cela permet de disposer des caractéristiques ordinales des variables de gravité, tout en respectant leur structure sémantique informative.

IV.3.4 Normalisation des données :

Les variables numériques ont fait l'objet d'une normalisation par la méthode StandardScaler (normalisation « z-score ») afin qu'elles aient toutes une moyenne nulle et un écart-type égal à 1.

C'est nécessaire afin d'éviter qu'une variable de grande échelle (ex. nombre d'accidents) ne domine l'analyse par rapport à une variable de petite échelle (Hair, 2009).

IV.3.5 Vérification des données finales :

Une matrice de corrélations a été établie afin de vérifier la présence éventuelle de redondances entre les variables.

Nous avons donc une matrice de données préparées fiable pour l'application de l'ACP, en garantissant que les composantes principales soient représentatives et interprétables.

IV.4 Application de l'Analyse en Composantes Principales (ACP) :

L'Analyse en Composantes Principales (ACP) a été appliquée après le prétraitement et normalisation des données pour réduire les dimensions et faciliter l'interprétation. Les résultats sont analysés selon trois représentations graphiques principales

IV.4.1. Scree Plot - Proportion de variance expliquée :

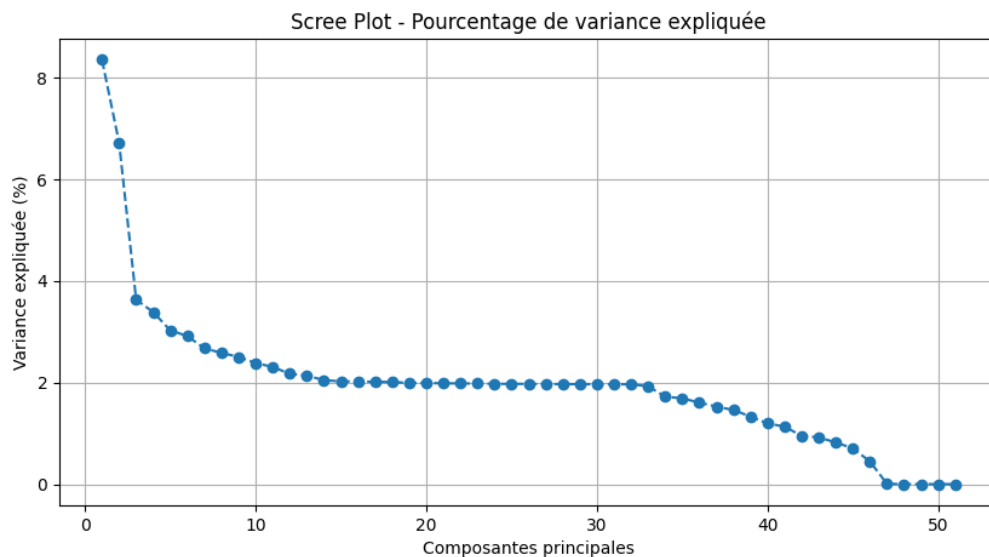


Figure (4.4) : proportion de la variance expliquée

Le Scree Plot représente la part de variance expliquée par chaque composante principale (PC). On observe que :

- PC1 représente 8,25 % de la variance totale,
- PC2 explique 6,66 %,
- les contributions décroissent rapidement après les premières composantes.

Cela donne indication que seules peu de composantes sont pertinentes pour résumer les données, ce qui conforte l'intérêt de la réduction dimensionnelle (Jolliffe and Cadima, 2016).

IV.4.2. Variance expliquée cumulée :

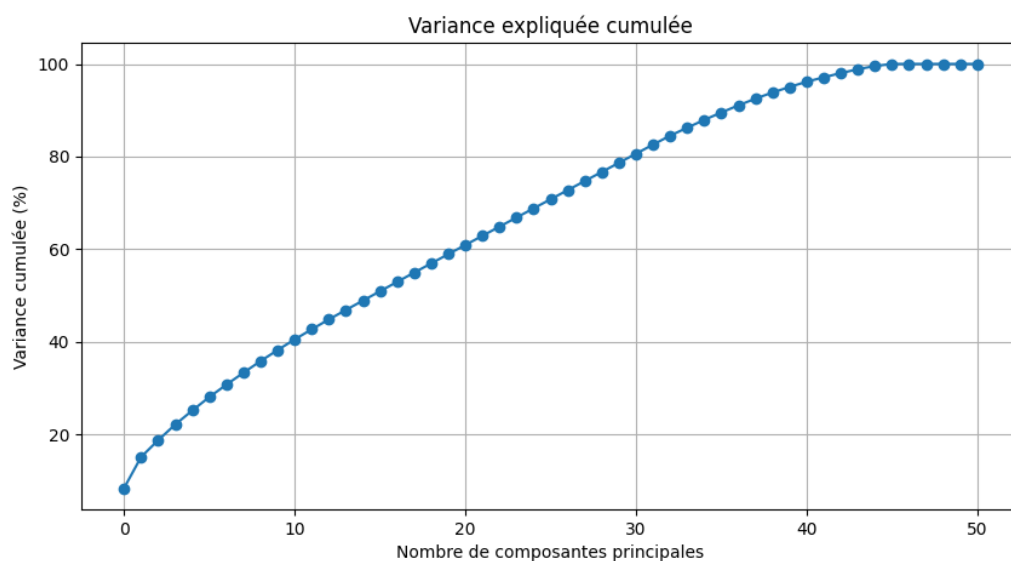


Figure (4.5) : la variance expliquée cumulée

Le graphique de la variance expliquée cumulée permet de visualiser l'accumulation progressive de la variance expliquée par le nombre croissant de composantes retenues. On remarque alors que :

- les 10 premières composantes expliquent environ 35 % de la variance,
- il faut 30 composantes pour atteindre plus de 70 % de la variance.

Ces résultats montrent que les données sont beaucoup dispersées avec une information éclatée sur plusieurs dimensions, ce qui est assez courant dans des bases de données complexes et hétérogènes (Abdi and Williams, 2010).

IV.4.3. Projection PCA (PC1 versus PC2) :

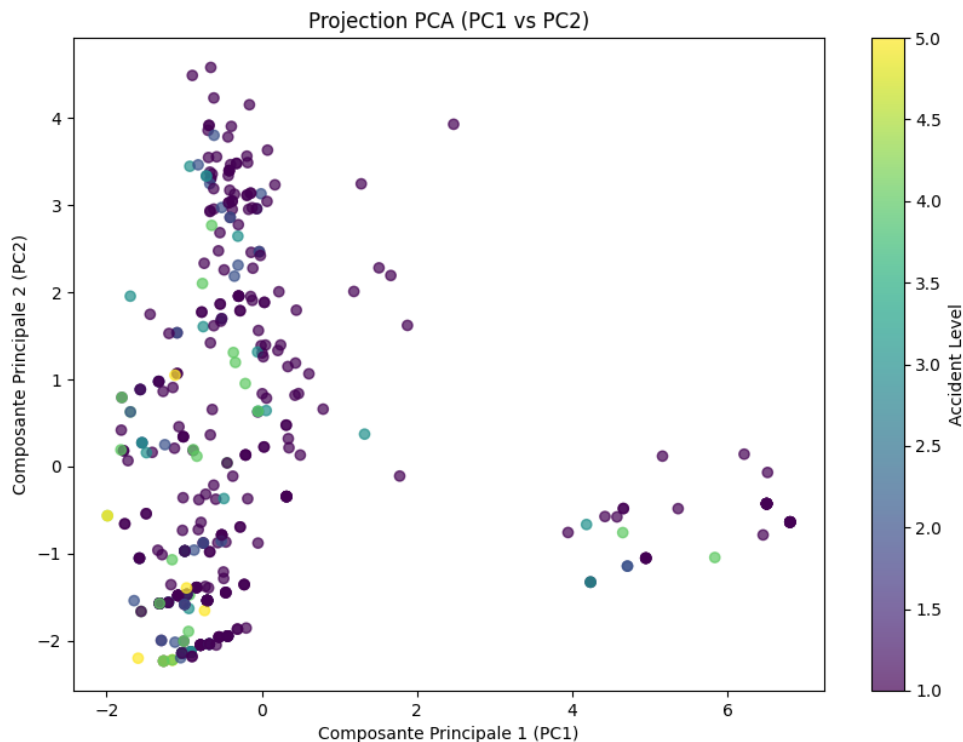


Figure (4.6) : la projection PCA

La projection dans le plan formé par PC1 et PC2, soit en deux dimensions, est employée pour représenter les observations (accidents).

Le graphique met en évidence une répartition des cas d'accidents en fonction des caractéristiques considérées.

Certains regroupements semblent se dessiner, notamment entre les niveaux d'accidents et les variables qui les sont associées (ex. Potential Accident Level).

Ainsi, bien que PC1 et PC2 ne représentent qu'une faible part de l'information totale, ils apportent une vision synthétique qui aide à l'analyse exploratoire des données.

IV.4.4 Synthèse :

La combinaison de ces trois graphiques permet de :

- évaluer la pertinence de l'ACP (Scree Plot),
- déterminer le nombre optimal de composantes à retenir (variance cumulée),

-interpréter la distribution des observations dans un espace réduit (Projection PC1 vs PC2).

Ces résultats confirment que l'ACP constitue un outil efficace pour simplifier et explorer les données d'accidents industriels.

IV. 5 Interprétation des résultats :

L'interprétation des résultats de l'ACP permet de relier les dimensions statistiques aux réalités de terrain sur le sujet de la sécurité industrielle et de la prévention des accidents.

IV. 5.1. Répartition de la variance :

La première analyse en composantes principales montre que seules les deux premières composantes principales (PC1 et PC2) ne permettent d'expliquer qu'une proportion limitée de la variance totale (environ 15 %), ce qui traduit le caractère complexe des données, où l'information est très dispersée sur un grand nombre de variables, qui montre qu'aucun facteur à lui seul n'explique la majorité des accidents, conforme à la multifactorialité des risques industriels (Abdi and Williams, 2010).

IV. 5.2. Contribution des variables :

L'analyse des loadings (poids des variables dans les composantes) permet de vérifier que :

-Accident Level et Potential Accident Level contribuent fortement à la première composante (PC1) qui est l'axe principal de différenciation des cas sur lequel se différencient les cas.

-Les variables liées au lieu (Local) et au pays (Countries) contribuent plus à PC2 et aux composantes suivantes qui traduisent plus des différences contextuelles et organisationnelles qui affectent l'apparition des accidents.

	PC1	PC2
Potential Accident Level	-0.27459927073409	-0.10592618218868
Countries_Country_02	-0.091539519751364	0.44181949846144
Countries_Country_03	0.45875241434235	-0.064983833393261
Local_Local_02	-0.081457693058377	0.065559852984581
Local_Local_03	-0.1328341288372	-0.24644174276305
Local_Local_04	-0.071849850890792	-0.19998152038335
Local_Local_05	-0.03807520737482	0.35326789149441
Local_Local_06	0.0057402935862125	0.076982415151997
Local_Local_07	-0.052906036966301	0.030327165693744
Local_Local_08	-0.022192721142226	0.2083329581381
Local_Local_09	0.0097184685349798	0.049391208632375
Local_Local_10	0.45875241434235	-0.064983833393261
Local_Local_11	0.025055153169128	0.0026215357206777
Local_Local_12	0.041207909848524	0.069324203533866
Industry Sector_Mining	-0.26100317187954	-0.37808494752554
Industry Sector_Others	0.45144866067434	-0.040113450864612
Genre_Male Employee or Third	-0.0050951725576345	-0.14731282550929
Party_Third Party Employee or Third Party_Third Party (Remote)	0.14397657279397	-0.28294107947575
Critical Risk_Bees Critical Risk_Blocking and isolation of energies	0.247143110153	-0.026321009482753
Critical Risk_Burn Critical Risk_Chemical substances Critical Risk_Confined space	-0.0051796859321867	0.070449660513958
Critical Risk_Cut Critical Risk_Electrical Shock Critical Risk_Electrical installation	-0.0081369353420601	0.047259862002742
Critical Risk_Fall Critical Risk_Fall prevention Critical Risk_Fall	-0.021910490045078	0.1852215719387
	-0.0045900035860308	0.049096316199556
	-0.013001936792104	0.10615490582898
	-0.028502293785925	-0.01321147417967
	-0.01061395498708	-0.020888575306827
	0.027026374470926	-0.0028367760385276
	0.013330294745875	0.018958570494209

Tableau (4.1) : poids des variables dans les composantes

IV. 5.3. Regroupements observés (Projection PCA PC1 vs PC2) :

La représentation des observations sur le plan (PC1, PC2) montre qu'il existe :

- une grande variabilité et des regroupements qui montrent des similarités pour certains profils d'accidents.
- une proximité entre les cas à fort Accident Level et fort Potential Accident Level, suggérant que la gravité perçue correspond globalement à la gravité observée.

IV. 5.4. Implications pour l'évaluation des risques :

Ces résultats ont plusieurs implications pratiques : prévoir que pour surveiller les principaux facteurs de risque, l'ACP permet d'identifier les axes principaux de variabilité.

Le fait que la gravité des accidents centralise la première composante le confirme comme un facteur central dans la hiérarchisation des risques.

La dispersion des données indique qu'une analyse complémentaire (ex. clustering ou classification supervisée) serait sans doute utile pour segmenter les types d'accidents et développer des politiques de prévention plus ciblées.

IV. 5.5. Limites de l'approche :

Lorsque l'ACP permet d'affirmer une vision globale, elle présente néanmoins des limites : la variance expliquée est relativement faible, rendant nécessaire d'utiliser un plus grand nombre de composantes pour capter une part significative de l'information ; l'interprétation des composantes au-delà de PC2 devient plus compliquée et moins intuitive ; surtout des liens de causalité entre les variables ne rentrent pas en ligne de compte étant donné que l'ACP est plutôt à visée descriptive/exploratoire.

IV.6 Conclusion :

Dans ce chapitre, nous avons présenté une étude effectuée à partir d'une base de données réelle d'accidents industriels, dans laquelle l'Analyse en Composantes Principales (ACP) a été mise en application dans le cadre de prétraitements et de vérifications statistiques des données, permettant ainsi, grâce à son double objectif de dimensionnalité :

- de réduire la dimensionnalité d'un ensemble de variables complexes, facilitant l'interprétation des données « d'intérêt » (Jolliffe and Cadima, 2016).
- de mettre en évidence la contribution dominante des variables relatives à la gravité réelle et potentielle des accidents, revenant ainsi à confirmer le rôle essentiel des données quantitatives dans l'évaluation des risques associés aux accidents (Kourti, 2005).

- et de visualiser graphiquement les relations structurelles sur les résultats avec le Scree Plot, le Biplot et les représentations 2D, forme support à la prise de décision (Abdi and Williams, 2010).

Bien que les résultats atteints valident la pertinence de l'ACP pour structurer et interpréter des données multidimensionnelles dans le cadre de la sécurité industrielle, la faible part de variance expliquée par les deux premières composantes doit bien rappeler que cette approche doit être perçue comme un outil d'exploration, bien qu'il convienne d'envisager d'intégrer la prise en compte des risques, pour asseoir la question de l'évaluation des risques (Jackson, 2005).

En conclusion, l'ACP a réalisé une première tournée d'horizon des facteurs de risques et ouvert la voie à des analyses complémentaires (clustering, méthodes de prédiction supervisée), qui visent à arrimer la compréhension et la prévention des accidents industriels. (Zhou et al., 2021).

Conclusion générale

Ce travail visait à examiner l'apport de l'Analyse en Composantes Principales (ACP) dans l'évaluation des risques industriels, à travers l'exemple d'un cas pratique s'appuyant sur une base de données d'accidents du travail.

Les premiers chapitres ont présenté le contexte théorique de la démarche d'évaluation des risques ainsi que les techniques de réduction de dimension. L'ACP a été retenue comme méthode d'analyse au sens où elle permet de transformer un ensemble de variables corrélées en un petit nombre de composantes principales conservant l'essentiel de l'information (Jolliffe and Cadima, 2016).

L'application de l'ACP sur les données a permis de faire émerger notamment :

- la complexité multidimensionnelle des accidents industriels en partie liés à de nombreux facteurs interagissant (pays, site, type d'accident, niveau de gravité),
- la grande contribution des variables relatives à la réalité et la potentialité des accidents qui confirment leur rôle central en matière de hiérarchie des risques,

L'intérêt des représentations graphiques (Scree Plot, Biplot, Scatter Plot) pour parvenir à rendre compte des corrélations et, ainsi, aider à mieux interpréter.

Malgré cela, les résultats montrent que la variance expliquée par les premières composantes est relativement faible, et l'ACP doit donc être considérée comme un instrument exploratoire, et non pas comme un outil définitif de modélisation des risques (Jackson, 2005).

En résumé, cette recherche a montré que l'ACP constitue un outil méthodologique intéressant pour simplifier le problème proposé, identifier les variables a priori déterminantes, et faciliter la décision en matière de sécurité industrielle.

BIBLIOGRAPHIE

- Abdelgawad, M., Fayek, A.R., 2010. Risk management in the construction industry using combined fuzzy FMEA and fuzzy AHP. *J. Constr. Eng. Manag.* 136, 1028–1036.
- Abdi, H., Williams, L.J., 2010. Principal component analysis. *Wiley Interdiscip. Rev. Comput. Stat.* 2, 433–459.
- Aggarwal, C.C., 2018. Machine learning for text: An introduction, in: *Machine Learning for Text*. Springer, pp. 1–16.
- Anthony (Tony) Cox Jr, L., 2008. What's wrong with risk matrices? *Risk Anal. Int. J.* 28, 497–512.
- Bellman, R., 1961. *Control Process: A Guided Tour*.
- Bishop, C.M., Nasrabadi, N.M., 2006. *Pattern recognition and machine learning*. Springer.
- Chapelle, O., Schölkopf, B., Zien, A., 2006. A discussion of semi-supervised learning and transduction, in: *Semi-Supervised Learning*. MIT Press, pp. 473–478.
- Ericson, C.A., 2015. *Hazard analysis techniques for system safety*. John Wiley & Sons.
- Fang, D., Zhang, M., Zhao, C., 2019. Machine learning in safety management. *Saf. Sci.* 48–61.
- Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179–188.
- Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. *Deep learning*. MIT press Cambridge.
- Guyon, I., Elisseeff, A., 2003a. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
- Guyon, I., Elisseeff, A., 2003b. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182.
- Hair, J.F., 2009. *Multivariate data analysis*.
- Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. *science* 313, 504–507.
- Hollnagel, E., 2008. Risk+ barriers= safety? *Saf. Sci.* 46, 221–229.
- Hyvärinen, A., Hurri, J., Hoyer, P.O., 2001. Independent component analysis, in: *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*. Springer, pp. 151–175.
- IEC 31010, 2019. *Risk management — Risk assessment techniques*. Genève : IEC.
- INRS, 2017. *Évaluation des risques professionnels : démarche et méthodes*.
- ISO 31000, 2018. *Risk Management – Principles and Guidelines*.
- ISO 31010, 2019. *Gestion des risques*.
- ISO 45001, 2018. *Systèmes de management de la santé et de la sécurité au travail*.
- Jackson, J.E., 2005. *A user's guide to principal components*. John Wiley & Sons.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An introduction to statistical learning: with applications in R*. Springer.
- Jolliffe, I., 2011a. Principal component analysis, in: *International Encyclopedia of Statistical Science*. Springer, pp. 1094–1096.
- Jolliffe, I., 2011b. Principal component analysis, in: *International Encyclopedia of Statistical Science*. Springer, pp. 1094–1096.
- Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374, 20150202.
- Kabir, S., Papadopoulos, Y., 2019. Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. *Saf. Sci.* 115, 154–175.
- Kelleher, J.D., Mac Namee, B., D'arcy, A., 2020. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.
- Khan, F., Haddara, M., 2020. Risk-based approaches for maintenance planning using machine learning techniques. *Saf. Sci.* 124.
- Kou, G., Xu, Y., Shen, F., Peng, Y., 2021. Machine learning methods for insurance risk assessment, . *Expert Syst. Appl.* 168.
- Kourti, T., 2005. Application of latent variable methods to process control and multivariate statistical process control in industry. *Int. J. Adapt. Control Signal Process.* 19, 213–246.
- Li, J., Yu, H., Sun, H., 2021. Deep learning-based anomaly detection for chemical process safety. *J. Loss Prev. Process Ind.* 72.
- Maaten, L. van der, Hinton, G., 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

- Mitchell, T.M., 1997. Does machine learning really work? *AI Mag.* 18, 11–11.
- Moubray, J., 2001. Reliability-centered maintenance. Industrial Press Inc.
- OIT, 2013. Guide pratique sur l'évaluation des risques professionnels. Organisation internationale du travail.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Rausand, M., 2013. Risk assessment: theory, methods, and applications. John Wiley & Sons.
- Rausand, M., Hoyland, A., 2003. System reliability theory: models, statistical methods, and applications. John Wiley & Sons.
- Reason, J., 2016. Managing the risks of organizational accidents. Routledge.
- Roweis, S.T., Saul, L.K., 2000. Nonlinear dimensionality reduction by locally linear embedding. *science* 290, 2323–2326.
- Samuel, A.L., 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.* 3, 210–229.
- Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10, 1299–1319.
- Stefanini, I. H. M., 2021. Industrial Safety and Health Analytics Database with Accidents Description.
- Sutton, R.S., Barto, A.G., 1998. Reinforcement learning: An introduction. MIT press Cambridge.
- Tenenbaum, J.B., Silva, V. de, Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 2319–2323.
- Tixier, J., Dusserre, G., Salvi, O., Gaston, D., 2002. Review of 62 risk analysis methodologies of industrial plants. *J. Loss Prev. Process Ind.* 15, 291–303.
- Verleysen, M., François, D., 2005. The curse of dimensionality in data mining and time series prediction. Presented at the International work-conference on artificial neural networks, Springer, pp. 758–770.
- Villemeur, 1991. Sûreté de fonctionnement des systèmes industriels. Eyrolles.
- Vyas, S., Kumaranayake, L., 2006. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy Plan.* 21, 459–468.
- Wei, Z., Liu, H., Tao, X., Pan, K., Huang, R., Ji, W., Wang, J., 2023. Insights into the application of machine learning in industrial risk assessment: a bibliometric mapping analysis. *Sustainability* 15, 6965.
- Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. *Chemom. Intell. Lab. Syst.* 2, 37–52.
- Xu, Z., Huang, Y., Zhang, J., 2021. Applications of machine learning in risk assessment and safety prediction. *Reliab. Eng. Syst. Saf.* 214.
- Yang, F., Wang, J., Chen, T., 2020. Data-driven safety risk identification in process industries using principal component analysis. *J. Loss Prev. Process Ind.* 64, 104053.
- Zhang, Y., Lin, X., Jiang, Y., 2019. Predictive maintenance for power equipment based on machine learning. , *IEEE Trans. Ind. Inform.* 15(12).
- Zhang, Y., Yang, G., Wang, J., 2020. Machine learning approaches for risk assessment in complex systems. *Saf. Sci.* 130.
- Zhou, C., Huang, Y., Ding, Y., 2021. Application of machine learning in process safety. *J. Loss Prev. Process Ind.*