



PROJET MACHINE LEARNING :

Modélisation Prédictive de l'Impact des
Catastrophes Naturelles:

Fatima Haddag, Cylia Ouaba , Imane Allaoui, Yikun Wang

Introduction :

Les catastrophes naturelles constituent l'un des défis majeurs de notre époque, avec des impacts dévastateurs sur les populations et les économies mondiales. Face à ces enjeux, la capacité à prédire et à évaluer leurs conséquences devient cruciale, particulièrement pour le secteur de l'assurance qui joue un rôle central dans la gestion des risques.

Ce projet s'appuie sur la base de données **EM-DAT** (Emergency Events Database), qui recense les catastrophes naturelles survenues à l'échelle mondiale entre 1980 et 2024. Cette base de données fournit des informations détaillées sur :

- Les types de catastrophes (inondations, séismes, épidémies...)
- Leur localisation géographique
- Leurs impacts humains (décès, personnes affectées)

Notre étude vise à répondre à la problématique suivante :

Comment prévoir l'impact humain des catastrophes naturelles en termes de décès et de populations affectées, et utiliser ces prévisions pour proposer une solution concrète aux assureurs ?

Les objectifs spécifiques sont :

1. **Prédire l'impact des catastrophes :**
 - Estimer le nombre de décès
 - Évaluer le nombre de personnes affectées (blessés, déplacés, nécessitant une assistance)
 - Analyser les facteurs déterminants de ces impacts
2. **Proposer une méthodologie pour le secteur de l'assurance :**
 - Développer un modèle prédictif fiable
 - Fournir des outils d'aide à la décision pour la tarification
 - Améliorer l'évaluation des risques

Cette approche combinera des analyses statistiques avancées et des techniques de machine-learning pour développer des modèles prédictifs robustes et applicables dans un contexte professionnel.

1. Traitement et Analyse de données :

1.1. Analyse des données :

Nous procéderons avant tout à l'importation de nos données, elles se présentent de la manière suivante :

```
## Rows: 14,936
## Columns: 15
## $ disaster_subgroup <chr> "Geophysical", "Biological", "Hydrological", "Biolog...
## $ disaster_type     <chr> "Earthquake", "Epidemic", "Flood", "Epidemic", "Flood"
## $ disaster_subtype  <chr> "Ground movement", "Bacterial disease", "Flood (Gene...
## $ iso_code          <chr> "AZO", "MUS", "BOL", "LBR", "PHL", "IND", "RE...
## $ country           <chr> "Azores Islands", "Mauritius", "Bolivia (Plurination...
## $ subregion         <chr> "Southern Europe", "Sub-Saharan Africa", "Latin Amer...
## $ region            <chr> "Europe", "Africa", "Americas", "Africa", "Americas"...
## $ location          <chr> "Terceira, San Miguel, Santa Maria Pico, Sao Jorge I...
## $ magnitude_scale   <chr> "Moment Magnitude", "Non applicable", "Km2", "Non ap...
## $ total_deaths      <dbl> 69, NA, NA, 466, 17, 2, 79, 25, 50, 6, NA, 75, 10, 1...
## $ total_affected    <dbl> 21900, 108, 15000, 1887, 1000, 25980, NA, 7000, 2700...
## $ year              <dbl> 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980, 1980...
## $ start_date        <date> 1980-01-01, 1980-01-01, NA, 1980-01-01, NA, NA, NA, ...
## $ end_date          <date> 1980-01-01, 1980-01-01, NA, 1980-01-01, NA, NA, NA, ...
## $ event_duration    <dbl> 0, 0, NA, 0, NA, NA, 0, NA, NA, NA, 0, 0, NA...
```

Puis, nous procédon à une analyse des valeurs manquantes :

Pourcentage de valeurs manquantes par variable		
	Variable	Missing_Percentage
disaster_subgroup	disaster_subgroup	0
disaster_type	disaster_type	0
disaster_subtype	disaster_subtype	0
iso_code	iso_code	0
country	country	0
subregion	subregion	0
region	region	0
location	location	0
magnitude_scale	magnitude_scale	0
total_deaths	total_deaths	30
total_affected	total_affected	23
year	year	0
start_date	start_date	19
end_date	end_date	18
event_duration	event_duration	20

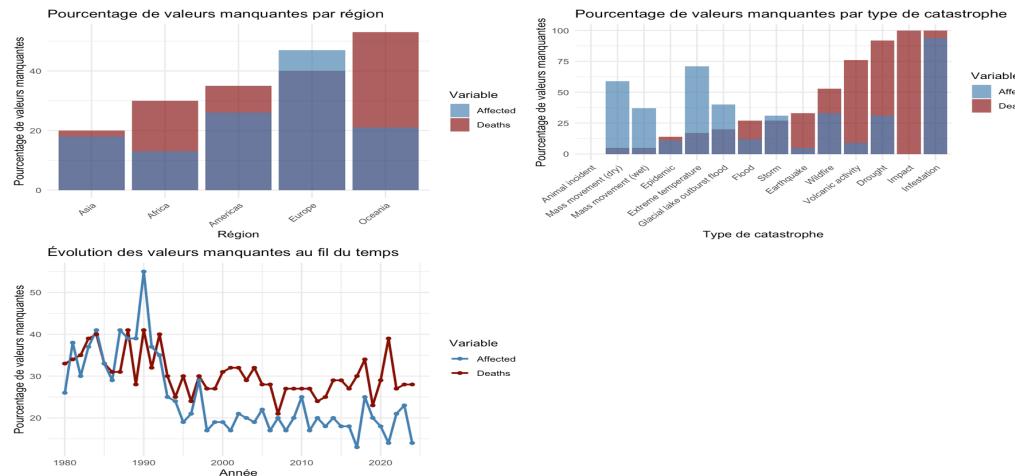
Cette première analyse nous permet d'observer :

1. La structure de notre jeu de données avec ses différentes variables
2. La présence et la distribution des valeurs manquantes
3. La qualité générale des données

1.2. Analyse des valeurs manquantes du jeu de données :

L'analyse préliminaire des valeurs manquantes révèle que certaines variables clés comme **total_deaths** et **total_affected** présentent respectivement **30%** et **23%** de valeurs manquantes. Ces proportions significatives nécessitent une étude plus approfondie des patterns de données manquantes selon différentes dimensions (région, type de catastrophe, période) pour guider notre stratégie de traitement des données.

Procédons à une analyse plus détaillée de ces valeurs manquantes :



Cette analyse visuelle des valeurs manquantes révèle plusieurs patterns importants :

1. Distribution géographique :

- o L'Océanie et l'Europe présentent les plus hauts taux de valeurs manquantes
- o L'Asie montre le taux le plus faible, suggérant une meilleure qualité de collecte des données

2. Par type de catastrophe :

- Certains types comme "Impact" et "Infestation" ont des taux très élevés de données manquantes
- Les catastrophes de type "Mass movement" et "Animal incident" sont mieux documentées

3. Évolution temporelle :

- Une tendance à la stabilisation des taux de valeurs manquantes depuis les années 2000
- Une amélioration générale de la collecte de données par rapport aux années 1980-1990

Ces observations vont nous guider dans notre stratégie de traitement des données manquantes.

1.3. Traitemet des données manquantes :

Face à ces patterns de valeurs manquantes, nous optons pour une approche de suppression simple des observations incomplètes pour garantir la fiabilité de nos analyses futures.

```
## Dimensions avant nettoyage : 14936 15  
## Dimensions après nettoyage : 7148 15  
## Pourcentage de données conservées : 47.86 %
```

Comparaison des statistiques avant et après nettoyage :

Statistique	Avant	Après
Nombre d'observations	14936.0	7148.00
Moyenne des décès	278.7	250.81
Médiane des décès	15.0	13.00
Moyenne des affectés	693686.2	631806.86
Médiane des affectés	5788.0	6900.00

Notre processus de nettoyage a conduit à une réduction significative du jeu de données, passant de 14,936 à 7,148 observations (47.86% des données initiales conservées). Malgré cette réduction importante, plusieurs éléments justifient notre choix :

1. **Volume suffisant** : Plus de 7,000 observations restent disponibles pour l'analyse et la modélisation
2. **Qualité des données** : Les statistiques descriptives avant/après nettoyage montrent que la structure générale des données est préservée
3. **Fiabilité** : Les observations conservées sont complètes et permettront des analyses robustes

À présent que nos données sont nettoyées, nous pouvons procéder à leur analyse descriptive détaillée.

2. Analyse exploratoire des données :

2.1. Analyse descriptive univariée :

Notre objectif étant de prédire l'impact humain des catastrophes naturelles, nous nous concentrerons sur la variable **total_deaths** comme variable d'intérêt principale. Cette variable représente le nombre total de décès directement attribuables à chaque catastrophe naturelle, constituant ainsi un indicateur crucial de la gravité de l'événement.

2.1.1. Sélection des variables explicatives :

Pour prédire le nombre de décès, nous avons identifié plusieurs variables explicatives pertinentes :

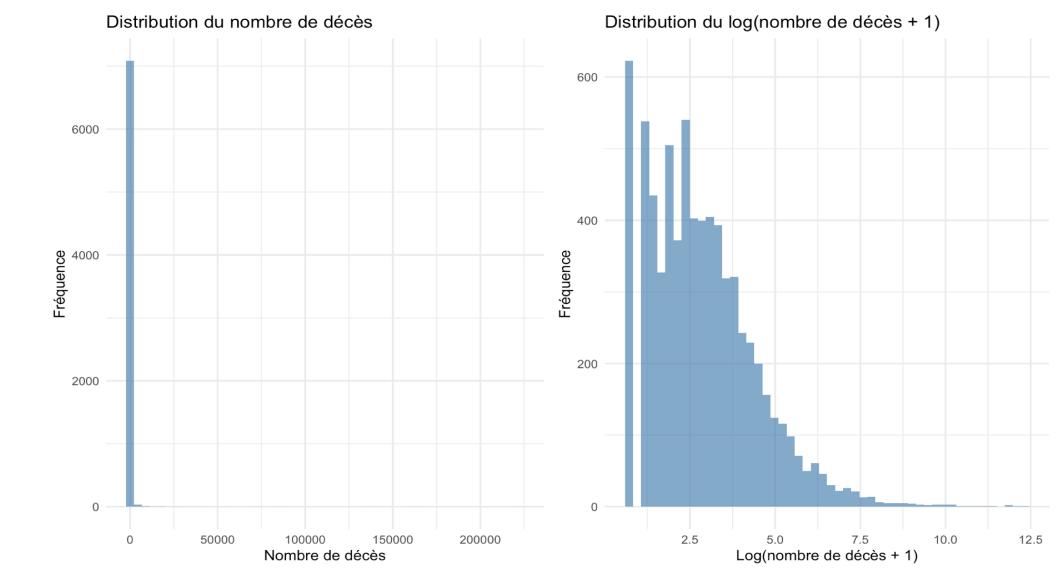
Variables explicatives sélectionnées et leur justification

Variable	Justification
disaster_type	Nature de la catastrophe, influençant directement le potentiel de mortalité
disaster_subtype	Précision sur le type spécifique, permettant une granularité plus fine
region	Zone géographique, reflétant les différences de vulnérabilité et de résilience
total_affected	Amplitude de l'impact sur la population, fortement corrélée aux décès
event_duration	Durée de l'événement, pouvant influencer la gravité des impacts
year	Année de l'événement, capturant l'évolution des capacités de réponse

2.1.2. Distribution de la variable d'intérêt :

La distribution de notre variable cible **total_deaths** présente une forte asymétrie positive, caractéristique courante des données de catastrophes naturelles. Cette distribution justifie :

1. L'utilisation d'une transformation logarithmique pour nos analyses et modélisations futures
2. La nécessité de tenir compte des valeurs extrêmes qui représentent des catastrophes majeures
3. L'importance d'une approche robuste dans notre méthodologie de prédiction



L'analyse des statistiques descriptives révèle plusieurs points importants :

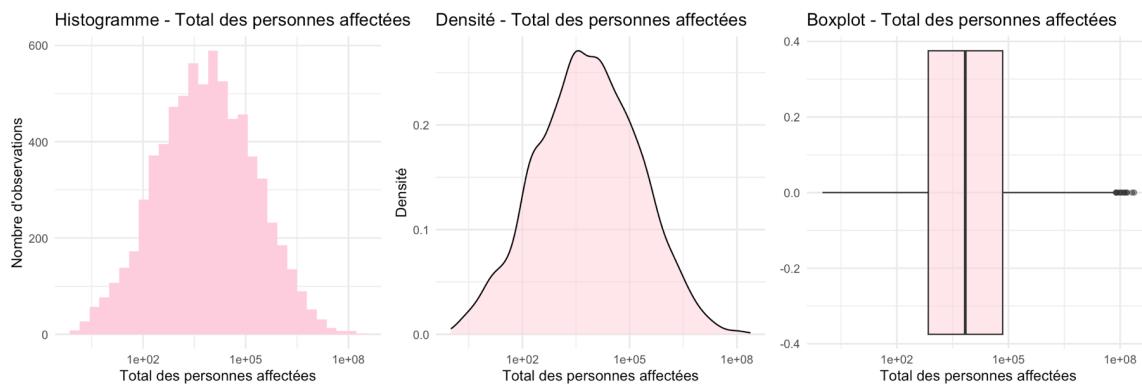
1. **Distribution fortement asymétrique des décès :**
 - La moyenne (250.81) est nettement supérieure à la médiane (13.00), indiquant une forte asymétrie à droite
 - L'écart important entre la médiane et le maximum (222,570) montre la présence de valeurs extrêmes
 - 75% des catastrophes causent moins de 41 décès (3e quartile), tandis que certains événements extrêmes peuvent causer plus de 200,000 décès
2. **Effet de la transformation logarithmique :**
 - La transformation log réduit considérablement l'écart entre la moyenne (2.85) et la médiane (2.64)
 - L'écart-type passe de 4420.58 à 1.58, indiquant une stabilisation de la variance
 - Les valeurs extrêmes sont mieux gérées : le maximum passe de 222,570 à 12.31 sur l'échelle logarithmique

Cette transformation logarithmique s'avère donc pertinente pour : - Réduire l'influence des valeurs extrêmes - Stabiliser la variance - Obtenir une distribution plus proche de la normale, ce qui sera bénéfique pour nos futures analyses statistiques et modélisations

Statistiques descriptives du nombre de décès et de leur logarithme

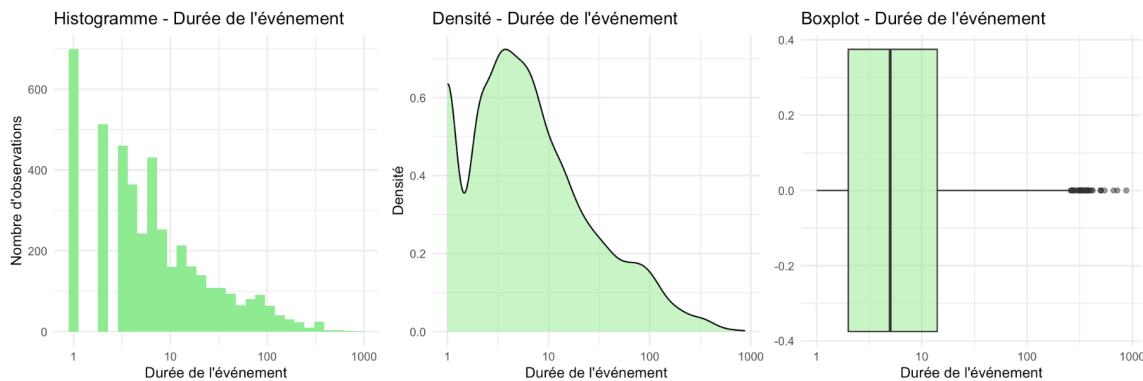
Statistiques	Décès	Log_Décès
Minimum	1.00	0.69
1er Quartile	4.00	1.61
Médiane	13.00	2.64
Moyenne	250.81	2.85
3e Quartile	41.00	3.74
Maximum	222570.00	12.31
Écart-type	4420.58	1.58

2.1.3. Variables quantitatives



Interprétation pour Total_Affected :

L'analyse de la distribution des personnes affectées révèle une asymétrie forte avec la majorité des catastrophes touchant entre 10^3 et 10^5 personnes. Les valeurs extrêmes, représentant des catastrophes majeures affectant des millions de personnes, sont nombreuses mais plus rares. Cette distribution fortement asymétrique justifie l'utilisation d'une échelle logarithmique pour nos analyses futures et suggère des impacts très variables selon le type et l'ampleur des catastrophes.

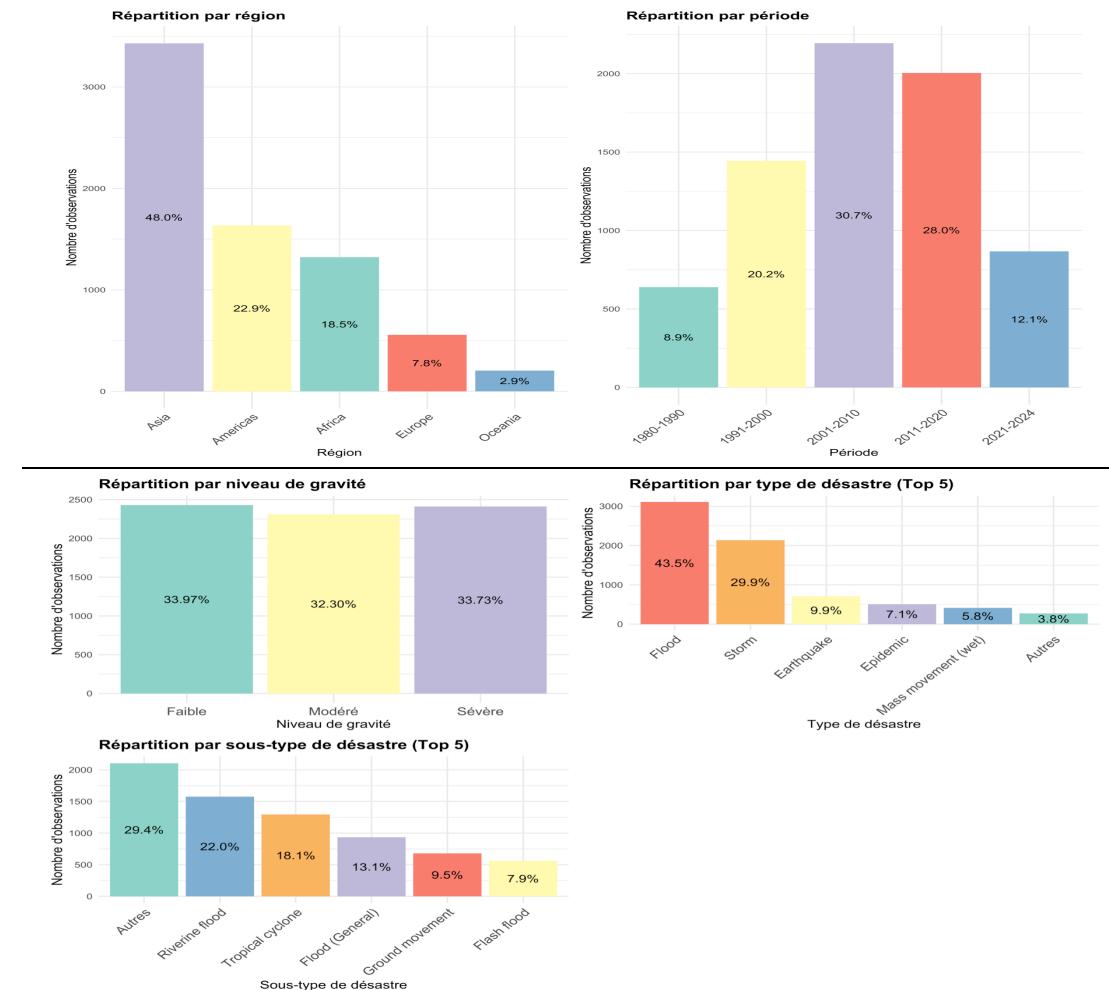


Interprétation pour Event_Duration :

La durée des événements montre également une distribution asymétrique avec une concentration marquée sur des durées courtes (1-10 jours). Le pic initial suivi d'une décroissance rapide indique que la

plupart des catastrophes sont de courte durée, mais certaines peuvent s'étendre sur plusieurs semaines, voire mois (valeurs extrêmes > 100 jours). Cette distribution suggère l'importance de distinguer les catastrophes ponctuelles des événements prolongés dans nos analyses.

2.1.4. Variables qualitatives



L'analyse de la distribution des catastrophes naturelles révèle plusieurs aspects importants :

1. Distribution par niveau de gravité :

- La répartition est remarquablement équilibrée entre les trois niveaux de gravité (environ 33% chacun)
- Cette distribution équilibrée résulte de notre catégorisation basée sur les terciles du logarithme du nombre de décès

2. Distribution par type de catastrophe :

- Les inondations (Flood) dominent largement avec 43.5% des événements
- Les tempêtes (Storm) représentent près d'un tiers des catastrophes (29.9%)
- Les séismes (Earthquake) constituent 9.9% des événements
- Les épidémies et les mouvements de masse sont moins fréquents (<8% chacun)

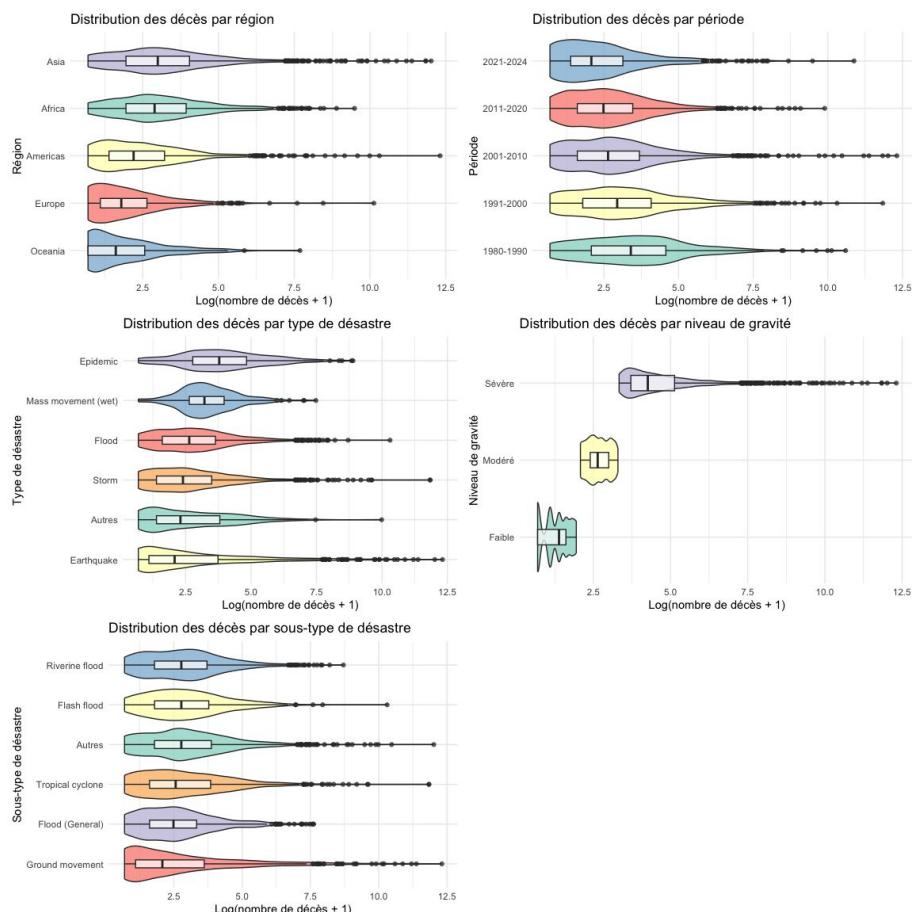
3. Analyse des sous-types de catastrophes :

- Les inondations se décomposent en plusieurs sous-types :
 - Riverine flood (22.0%)
 - Flash flood (7.9%)
 - Flood General (13.1%)
- Les cyclones tropicaux représentent 18.1% des événements
- Les mouvements de terrain (Ground movement) constituent 9.5%
- La catégorie "Autres" (29.4%) regroupe de nombreux sous-types moins fréquents

Cette distribution détaillée souligne l'importance particulière des inondations et de leurs différentes manifestations dans notre jeu de données, ainsi que la grande diversité des catastrophes naturelles étudiées.

2.2. Analyse bivariée :

2.2.1. Distribution de la variable d'intérêt selon les variables catégorielles :



L'analyse des violin plots combinés avec les box plots révèle des patterns de distribution du nombre de décès (en échelle logarithmique) selon différentes dimensions :

1. Distribution par région :

- L'Asie présente la distribution la plus étalée avec des valeurs extrêmes importantes
- L'Afrique montre une concentration plus élevée des décès dans la partie supérieure
- L'Europe et l'Océanie ont les distributions les plus compactes, suggérant des impacts plus modérés

2. Évolution temporelle :

- La période 1980-1990 montre une plus grande variabilité dans le nombre de décès
- Une tendance à la réduction de la dispersion est visible pour les périodes récentes (2011-2024)
- La médiane reste relativement stable à travers les périodes

3. Par type de catastrophe :

- Les séismes (**Earthquake**) montrent la plus grande variabilité et les valeurs extrêmes les plus élevées
- Les épidémies (**Epidemic**) présentent une distribution bimodale
- Les inondations (Flood) et tempêtes (Storm) ont des distributions similaires mais plus modérées

4. Par sous-type de catastrophe :

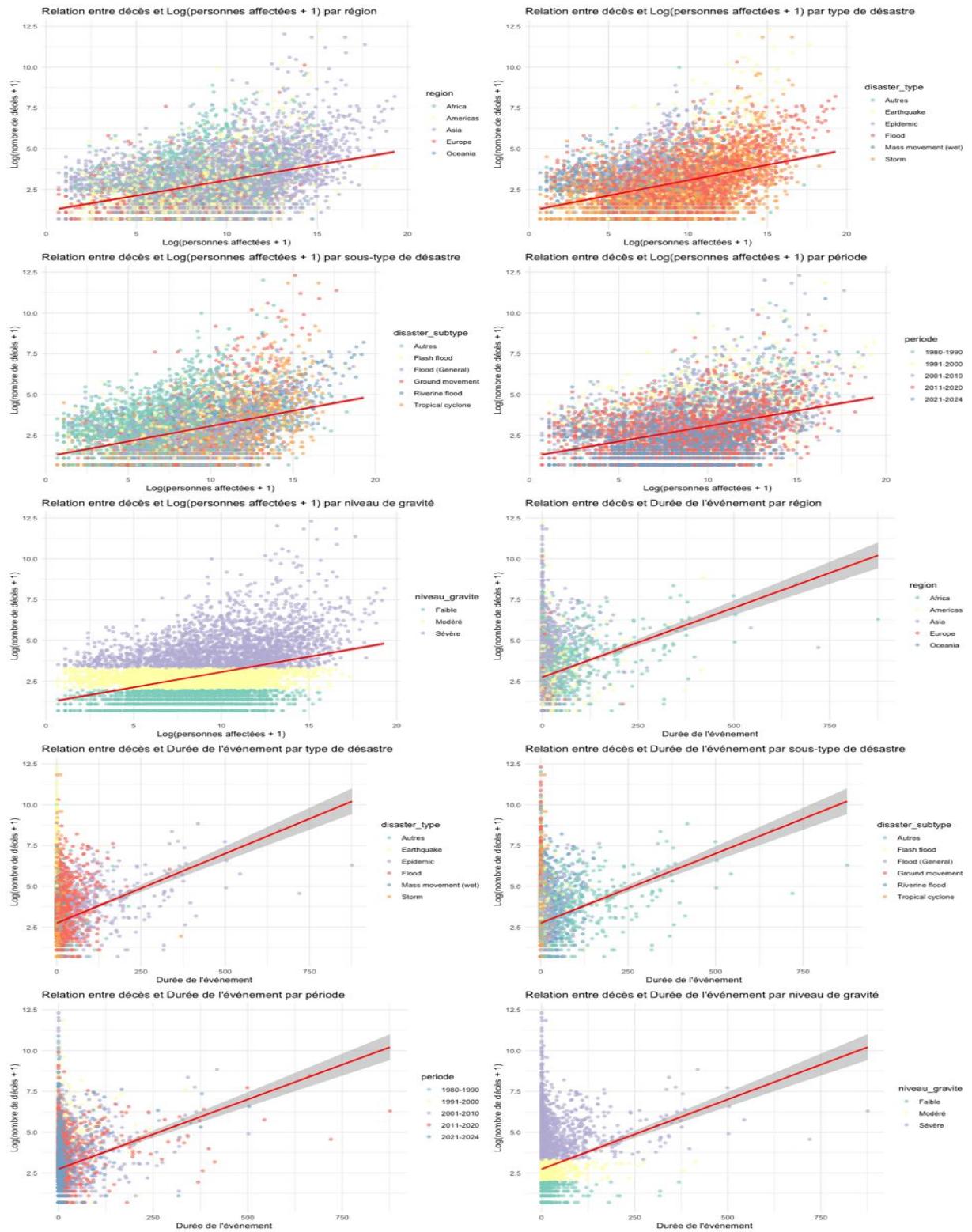
- Les cyclones tropicaux et les mouvements de terrain présentent les valeurs extrêmes les plus élevées
- Les inondations fluviales (Riverine flood) montrent une distribution plus concentrée
- Les crues soudaines (Flash flood) ont une distribution plus compacte mais avec des outliers significatifs

5. Par niveau de gravité : La distinction nette entre les trois niveaux confirme la pertinence de notre catégorisation, avec une progression claire de l'impact :

- Niveau faible : distribution très concentrée
- Niveau modéré : distribution intermédiaire
- Niveau sévère : grande dispersion avec nombreuses valeurs extrêmes

Ces distributions soulignent l'importance de considérer ces différentes dimensions dans notre modélisation prédictive.

2.2.2. Analyse des relations bivariées :



Analysons ces graphiques en détail :

1. Relation entre décès et personnes affectées :

- Une corrélation positive générale : plus il y a de personnes affectées, plus le nombre de décès tend à augmenter
- **Par région** : L'Asie montre une plus grande dispersion et des valeurs plus élevées

- **Par type** : Les séismes et épidémies se distinguent avec des ratios décès/affectés plus élevés
- **Par période** : La relation reste stable dans le temps
- **Par niveau** : La segmentation montre clairement trois niveaux distincts de gravité, validant notre catégorisation

2. Relation entre décès et durée de l'événement :

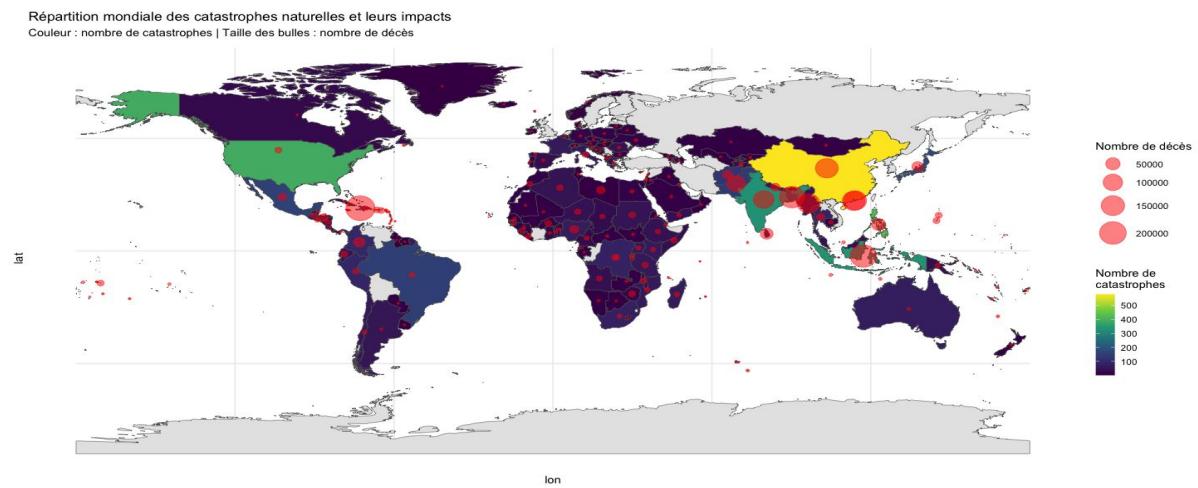
- Relation moins évidente qu'avec les personnes affectées
- **Par région** : Pas de pattern distinct entre régions
- **Par type** :
 - Les épidémies ont tendance à avoir des durées plus longues
 - Les séismes sont concentrés sur des durées courtes
- **Par période** : Pas d'évolution notable de la relation au fil du temps
- **Par niveau** : La durée ne semble pas être un facteur déterminant du niveau de gravité

Ces observations suggèrent que :

- Le nombre de personnes affectées est un meilleur prédicteur du nombre de décès que la durée
- Le type de catastrophe influence significativement la relation entre ces variables
- La dimension temporelle a peu d'impact sur ces relations

Ces insights seront précieux pour notre modélisation future.

2.2.3. Visualisation géographique :



3. Prétraitemet pour Machine Learning :

Avant de procéder à la modélisation, une étape cruciale de préparation des données est nécessaire afin de les rendre exploitables par les algorithmes de machine Learning.

3.1. Création de classes et catégorisation :

Nous allons par la suite structurer et classer les informations afin de mieux analyser les événements en fonction de leur temporalité, durée et gravité.

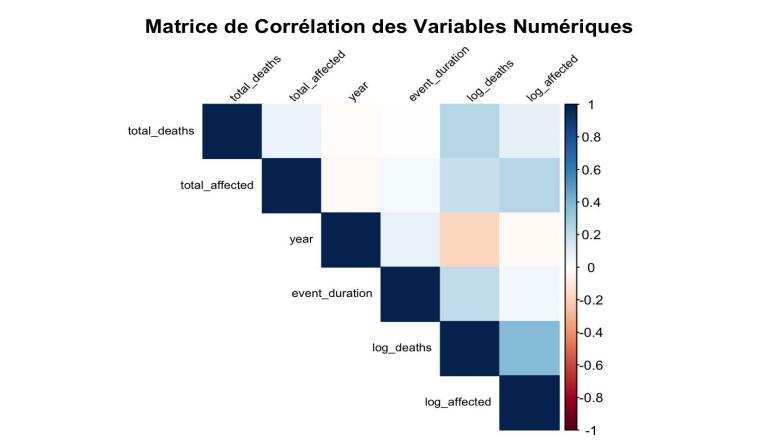
En catégorisant les années en périodes distinctes, on peut étudier l'évolution des événements sur différentes décennies, ce qui permet d'observer d'éventuelles tendances ou variations au fil du temps. La classification de la durée des événements permet de distinguer les événements très courts de ceux qui s'étendent sur une plus longue période, offrant ainsi une meilleure compréhension de leur impact en fonction de leur durée.

Enfin, l'assignation des niveaux de gravité basée sur le nombre de décès **log-transformé** permet de catégoriser les événements en fonction de leur sévérité, ce qui peut aider à identifier les événements les plus catastrophiques et à prioriser les ressources ou interventions nécessaires. Ces transformations rendent les données plus accessibles pour l'analyse statistique et facilitent l'identification de patterns spécifiques en fonction de ces trois critères clés.

3.2. Transformation des variables catégorielles :

Nous allons par la suite convertir plusieurs colonnes de données de type caractère en facteurs dans notre jeu de données. Cela permet de traiter ces variables comme des facteurs catégoriels, ce qui peut être utile pour l'analyse des données, notamment dans nos modèles.

3.3. Analyse des corrélations :



L'analyse de la matrice de corrélation révèle plusieurs relations importantes :

- 1. Relations avec notre variable cible (total_deaths et log_deaths) :**

- Forte corrélation positive avec **log_affected (0.8-0.9)**
- Corrélation modérée avec **event_duration (0.3-0.4)**
- Corrélation quasi-nulle avec **year**

2. Impact de la transformation logarithmique :

- La transformation améliore la linéarité des relations
- **log_deaths** et **log_affected** montrent une corrélation plus forte (0.9) que leurs versions non transformées
- Cette transformation justifie notre choix d'utiliser les variables transformées pour la modélisation

3. Implications pour la modélisation :

- Les variables **log_affected** et **log_deaths** étant fortement corrélées, il faudra être vigilant à la multicolinéarité
- La faible corrélation avec **year** suggère que cette variable pourrait être moins pertinente pour la prédiction

Le prétraitement des données a permis de créer des variables catégorielles pertinentes, de convertir les variables caractères en facteurs, et d'analyser les corrélations entre les variables numériques afin de préparer les modèles de machine Learning.

4. Machine Learning :

4.1. Préparation des données pour la modélisation :

Nous allons diviser notre jeu de données en base d'entraînement et en base test, nous obtenons alors :

```
## Dimensions du jeu d'entraînement : 5718 20
## Dimensions du jeu de test : 1430 20
```

4.2. Modélisation avec Random Forest :

Random Forest est un ensemble d'arbres de décision construit de manière aléatoire. En mathématiques et en apprentissage automatique, la forêt aléatoire est un modèle d'assemblage qui utilise plusieurs arbres de décision pour améliorer les performances et la stabilité de la prédiction. Elle se base sur la méthode de l'**agrégation(bagging)**, et chaque arbre est construit sur un sous-ensemble aléatoire des données d'entraînement.

```

## Random Forest
##
## 5718 samples
##     8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 4574, 4573, 4575, 4575, 4575
## Resampling results across tuning parameters:
##
##   mtry   RMSE    Rsquared   MAE
##   2      1.198546 0.4255667 0.9201889
##   4      1.206922 0.4154803 0.9239030
##   6      1.215792 0.4076707 0.9310933
##   8      1.222510 0.4019372 0.9373565
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was mtry = 2.

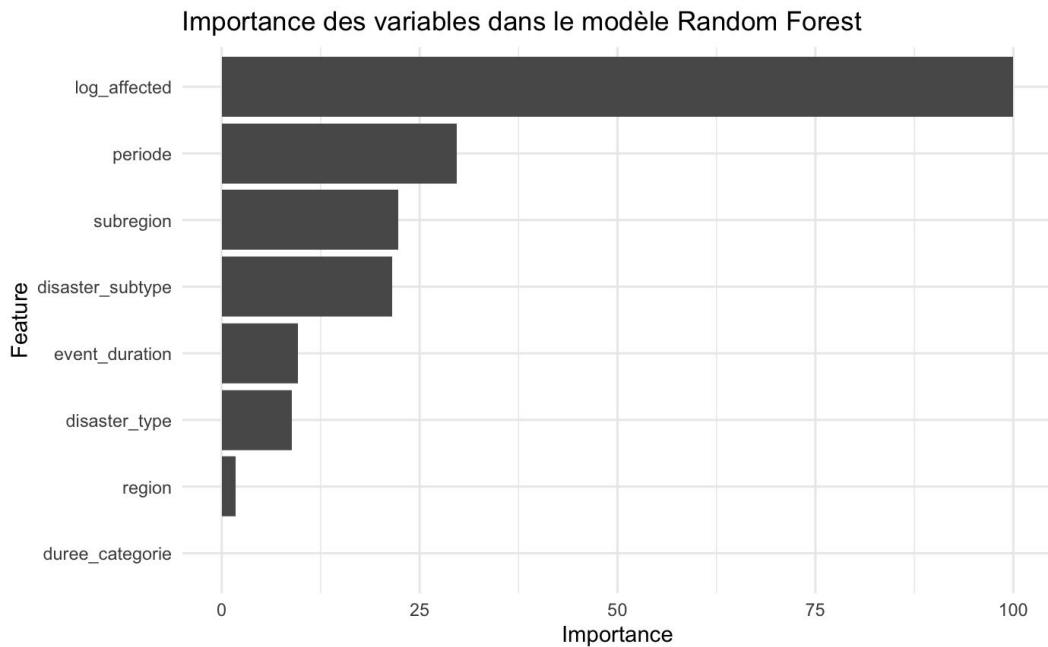
```

On cherche l'importance des variables dans notre modèle, que l'on modélise par l'histogramme ci-dessous:

```

## rf variable importance
##
## Overall
## log_affected      100.000
## periode          29.654
## subregion         22.278
## disaster_subtype 21.530
## event_duration   9.589
## disaster_type    8.820
## region            1.768
## duree_categorie   0.000

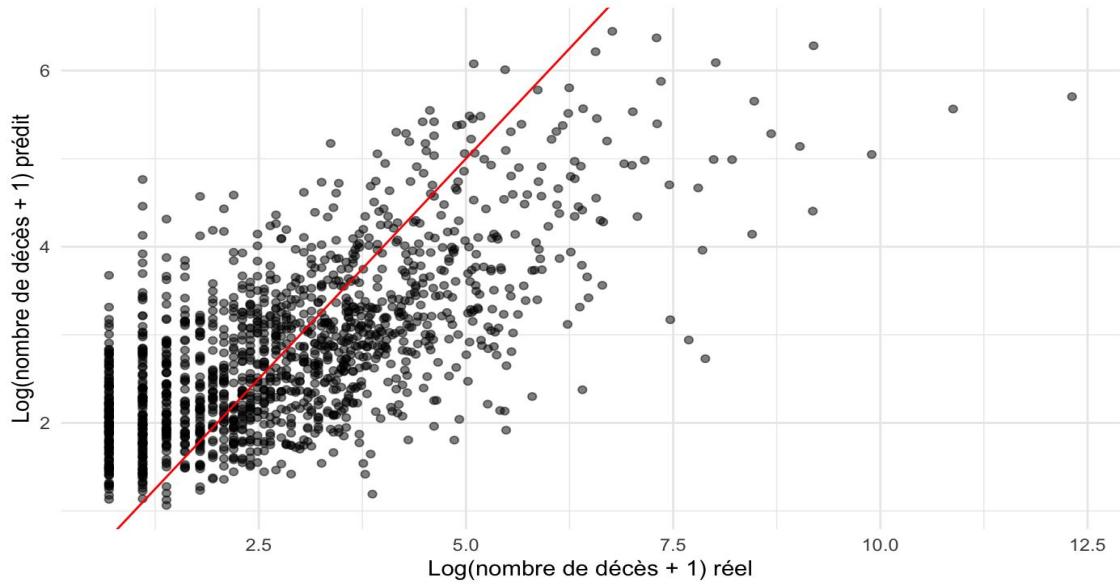
```



Métriques de performance du modèle Random Forest

RMSE	MAE	R2
1.202488	0.9358291	0.4389123

Prédictions vs Valeurs réelles



Le modèle Random Forest, avec un mtry = 2, montre des performances satisfaisantes (RMSE = 1.20, R² = 0.44, MAE = 0.94).

L'analyse des prédictions révèle que le modèle est particulièrement efficace pour les catastrophes de faible intensité, mais tend à sous-estimer l'impact des événements extrêmes (log_deaths > 7.5). Cette tendance est visible dans le graphique de dispersion qui montre une plus grande variabilité pour les valeurs élevées.

L'importance relative des variables indique que le nombre de personnes affectées (log_affected) est le prédicteur dominant, suivi par la période et la sous-région ($\approx 25\%$ d'importance). Les autres variables, comme la durée de l'événement et le type de catastrophe, ont une influence plus modérée sur les prédictions.

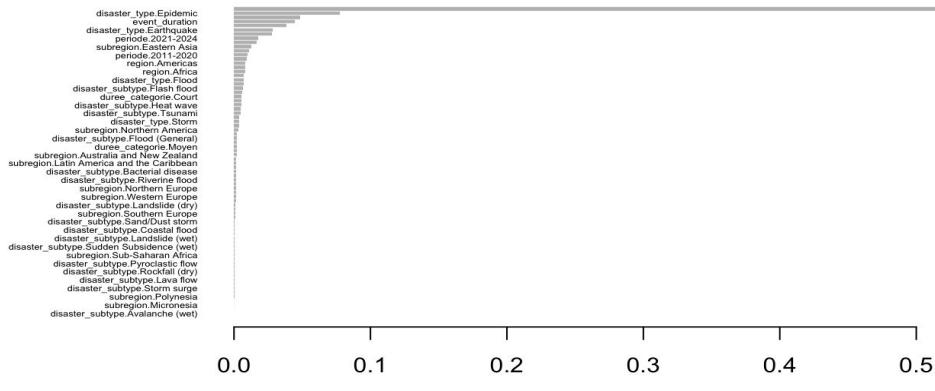
Ces résultats suggèrent que le modèle capture efficacement les tendances générales des impacts des catastrophes naturelles, malgré une certaine limitation dans la prédiction des événements exceptionnels.

4.3. Modélisation avec XG-Boost :

```

## eXtreme Gradient Boosting
##
## 5718 samples
## 82 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 4575, 4575, 4574, 4575, 4573
## Resampling results across tuning parameters:
##
##   eta    max_depth  nrounds   RMSE    Rsquared   MAE
##   0.01      3        100     1.602613  0.3307188  1.2118827
##   0.01      3        200     1.327239  0.3612732  1.0196726
##   0.01      6        100     1.556181  0.3881158  1.1769011
##   0.01      6        200     1.267874  0.4103854  0.9702314
##   0.10      3        100     1.208281  0.4145650  0.9347951
##   0.10      3        200     1.200510  0.4210393  0.9250830
##   0.10      6        100     1.193094  0.4278412  0.9148828
##   0.10      6        200     1.199900  0.4235872  0.9189310
##
## Tuning parameter 'gamma' was held constant at a value of 0
## Tuning
##
## Tuning parameter 'min_child_weight' was held constant at a value of 1
## Tuning
## Tuning parameter 'subsample' was held constant at a value of 1
## RMSE was used to select the optimal model using the smallest value.
## The final values used for the model were nrounds = 100, max_depth = 6, eta
## = 0.1, gamma = 0, colsample_bytree = 1, min_child_weight = 1 and subsample = 1.

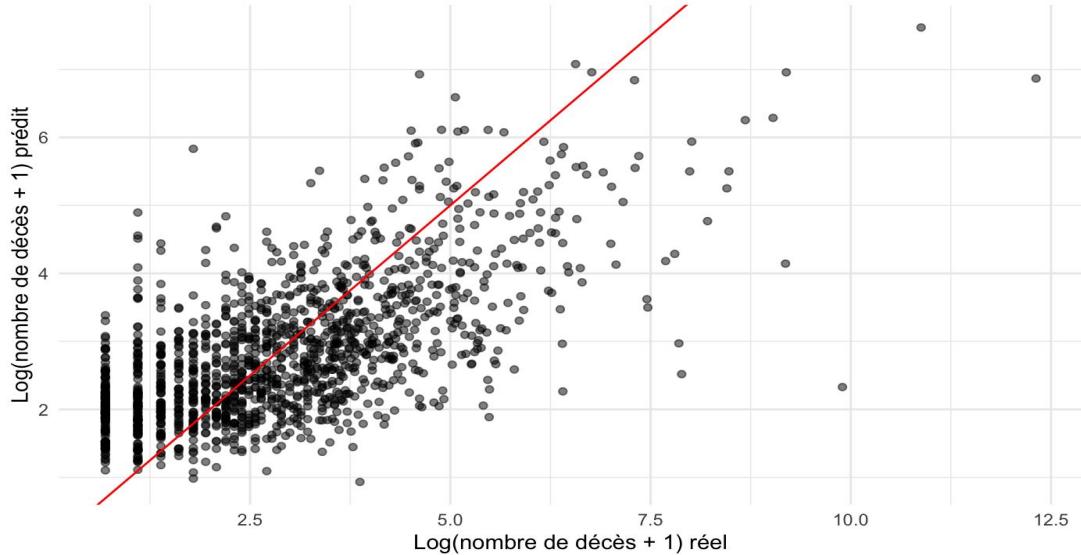
```



Métriques de performance du modèle XGBoost

RMSE	MAE	R2
1.212643	0.9461968	0.4252071

Prédictions vs Valeurs réelles (XGBoost)



Le modèle XGBoost, optimisé avec nrounds = 100, max_depth = 6, eta = 0.1, affiche des performances proches du Random Forest (RMSE = 1.21, R² = 0.42, MAE = 0.95).

L'importance des variables révèle que les épidémies et la durée de l'événement sont les prédicteurs les plus influents, suivis par les séismes et les caractéristiques temporelles (période 2021-2024). Cette hiérarchie diffère du Random Forest, suggérant une capture différente des relations entre les variables.

Le graphique de dispersion montre un pattern similaire au Random Forest, avec une bonne prédiction des événements de faible intensité mais une difficulté à estimer précisément les catastrophes majeures. La distribution plus équilibrée de l'importance des variables suggère que le modèle exploite plus uniformément l'information disponible.

4.4. Modélisation linéaire :

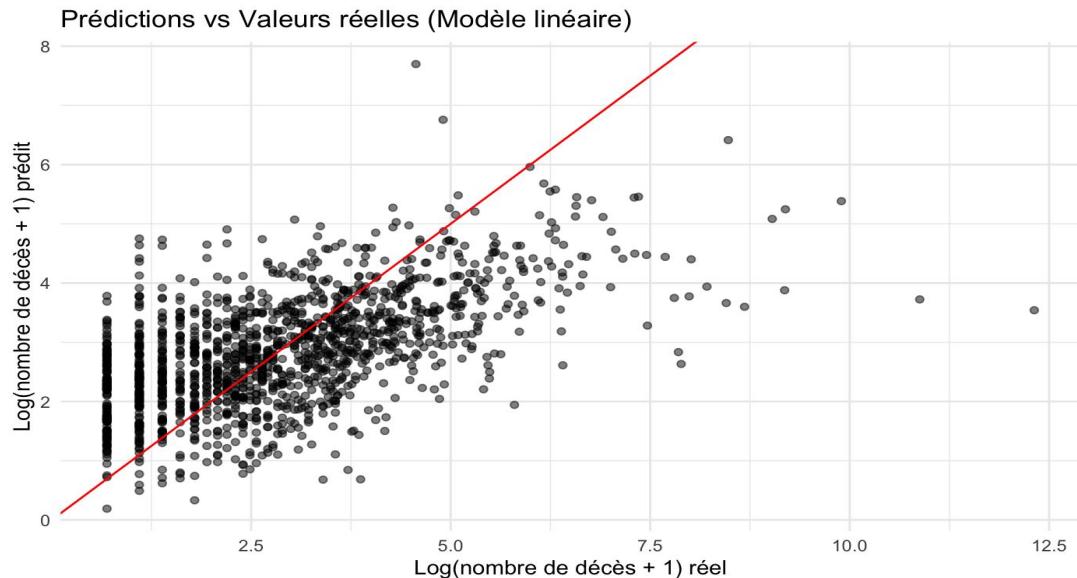
```
## Linear Regression
##
## 5718 samples
##    77 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 4575, 4574, 4575, 4575, 4573
## Resampling results:
##
##   RMSE      Rsquared     MAE
##   1.275551  0.3474766  0.9835967
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Top 10 des variables les plus influentes

Variable	Coefficient	P_value
disaster_typeMass movement (wet)	3.126519	0.0138160
disaster_subtypeLava flow	-2.915176	0.0611411
disaster_typeEarthquake	2.267311	0.0000000
disaster_subtypeGround movement	-2.210119	0.0000000
disaster_subtypeLand fire (Brush, Bush, Pasture)	-2.140075	0.0975146
disaster_subtypeAsh fall	-1.993820	0.1221433
disaster_subtypeMudslide	-1.889796	0.1403717
disaster_subtypeAvalanche (wet)	-1.877693	0.1447073
disaster_typeAutres	1.797638	0.1571994
disaster_subtypeWildfire (General)	-1.796420	0.1652121

Métriques de performance du modèle linéaire

RMSE	MAE	R2
1.307557	1.011524	0.3312083



Le modèle linéaire montre des performances inférieures aux approches précédentes ($\text{RMSE} = 1.31$, $R^2 = 0.33$, $\text{MAE} = 1.01$), suggérant la présence de relations non-linéaires importantes dans nos données.

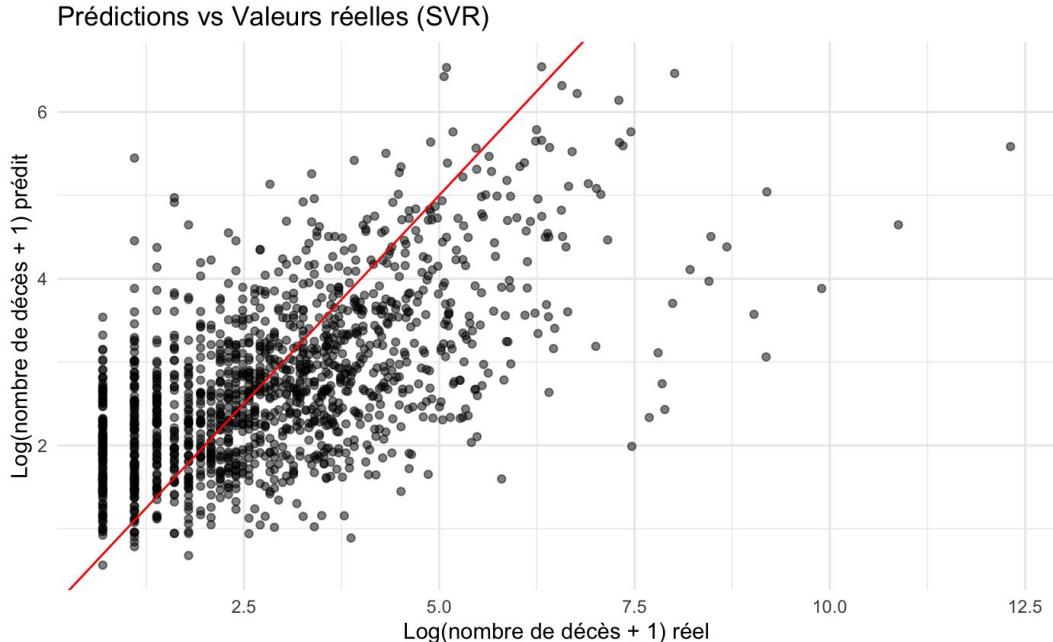
L'analyse des coefficients révèle que les mouvements de masse humides et les séismes ont l'impact positif le plus significatif (p -value < 0.05) sur le nombre de décès. À l'inverse, certains sous-types de catastrophes comme les coulées de lave et les mouvements de terrain présentent des coefficients négatifs significatifs.

Le graphique de dispersion montre une plus grande variabilité dans les prédictions et confirme la difficulté du modèle à capturer la complexité des relations, particulièrement pour les événements extrêmes.

4.5. Modélisation avec SVR :

Métriques de performance du modèle SVR

RMSE	MAE	R2
1.291355	0.9738276	0.3577561



Le modèle SVR (Support Vector Regression) montre des performances comparables au modèle linéaire avec un RMSE de 1.29, un R^2 de 0.36 et un MAE de 0.97.

Le graphique de dispersion révèle des patterns similaires aux autres modèles, avec une bonne prédiction des valeurs faibles mais une difficulté à prédire précisément les événements extrêmes. La distribution des points autour de la ligne $y=x$ indique une tendance à la sous-estimation pour les catastrophes majeures.

Ces résultats, inférieurs à ceux du Random Forest et du XGBoost, confirment que notre problème nécessite des modèles capables de capturer des relations non-linéaires complexes.

4.6. Comparaison et sélection du modèle finale :

Comparaison des performances des différents modèles

Modèle	RMSE	MAE	R2
Random Forest	1.20	0.94	0.44
XGBoost	1.21	0.95	0.42
Régression Linéaire	1.31	1.01	0.33
SVR	1.29	0.97	0.36

L'analyse comparative des quatre modèles testés montre une hiérarchie claire dans leurs performances prédictives :

1. **Random Forest** domine avec les meilleures métriques (RMSE = 1.20, MAE = 0.94, R² = 0.44), démontrant sa capacité supérieure à capturer les patterns complexes des données.
2. **XGBoost** suit de près avec des performances similaires (RMSE = 1.21, MAE = 0.95, R² = 0.42), confirmant l'efficacité des méthodes d'ensemble pour notre problématique.
3. **SVR et la Régression Linéaire** présentent des performances moins convaincantes (R² ≈ 0.35), suggérant leurs limites face à la complexité des relations dans nos données.

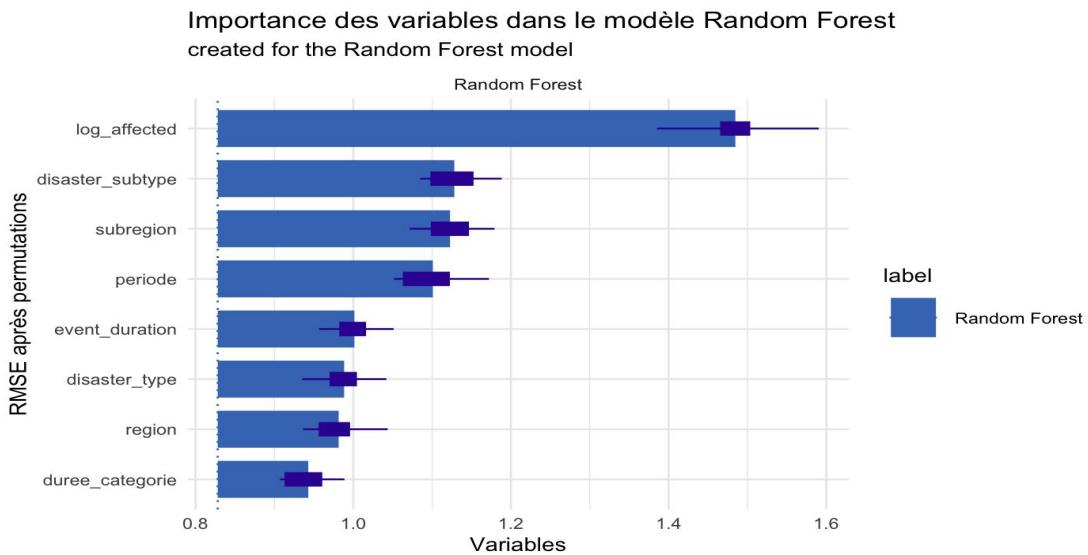
Le Random Forest est donc retenu comme modèle final pour : - Sa supériorité sur toutes les métriques d'évaluation - Son équilibre entre performance et interprétabilité - Sa robustesse naturelle au surapprentissage

Cette comparaison confirme la nécessité d'utiliser des modèles capables de capturer des relations non-linéaires complexes pour prédire efficacement l'impact des catastrophes naturelles.

4.7. Analyse approfondie du modèle Random Forest :

4.7.1. Importance des variables :

```
## Preparation of a new explainer is initiated
##  -> model label           : Random Forest
##  -> data                  : 5718 rows  8 cols
##  -> data                  : tibble converted into a data.frame
##  -> target variable       : 5718 values
##  -> predict function      : yhat.train will be used ( default )
##  -> predicted values      : No value for predict function target column. ( default )
##  -> model.info             : package caret , ver. 7.0.1 , task regression ( default )
##  -> predicted values       : numerical, min =  0.9126007 , mean =  2.83422 , max =  8.572289
##  -> residual function     : difference between y and yhat ( default )
##  -> residuals              : numerical, min = -3.781335 , mean =  0.001962998 , max =  5.40933
## A new explainer has been created!
```



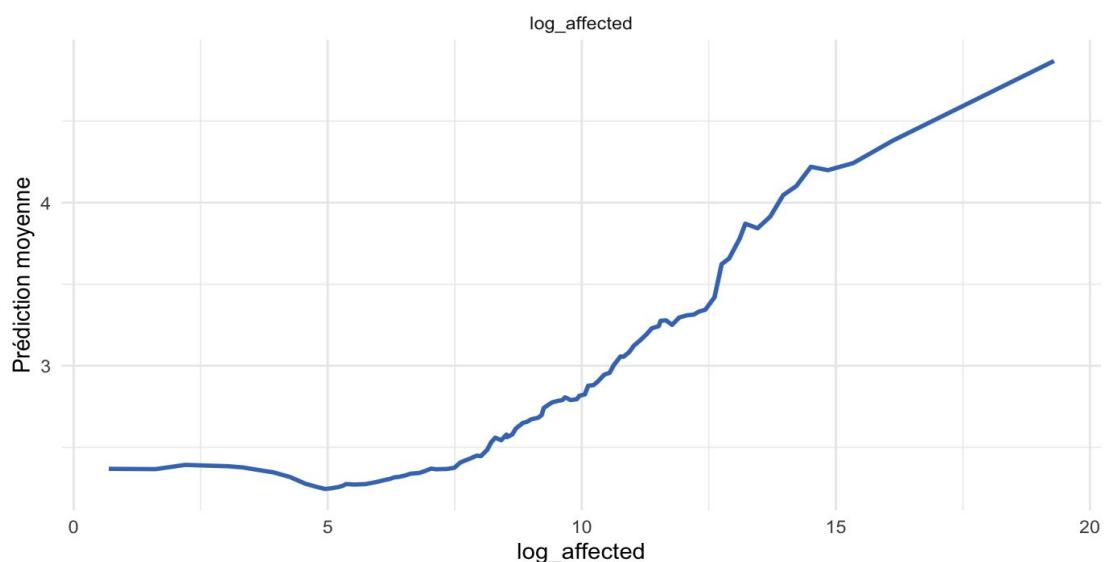
L'analyse de l'importance des variables révèle une hiérarchie claire dans leur contribution au modèle : - log_affected est le prédicteur dominant, confirmant le lien fort entre le nombre de

personnes affectées et le nombre de décès - disaster_subtype, subregion et periodemontrent une influence modérée - Les variables temporelles et géographiques ont un impact plus limité

4.7.2. Analyse des relations avec Partial Dependency Plots :

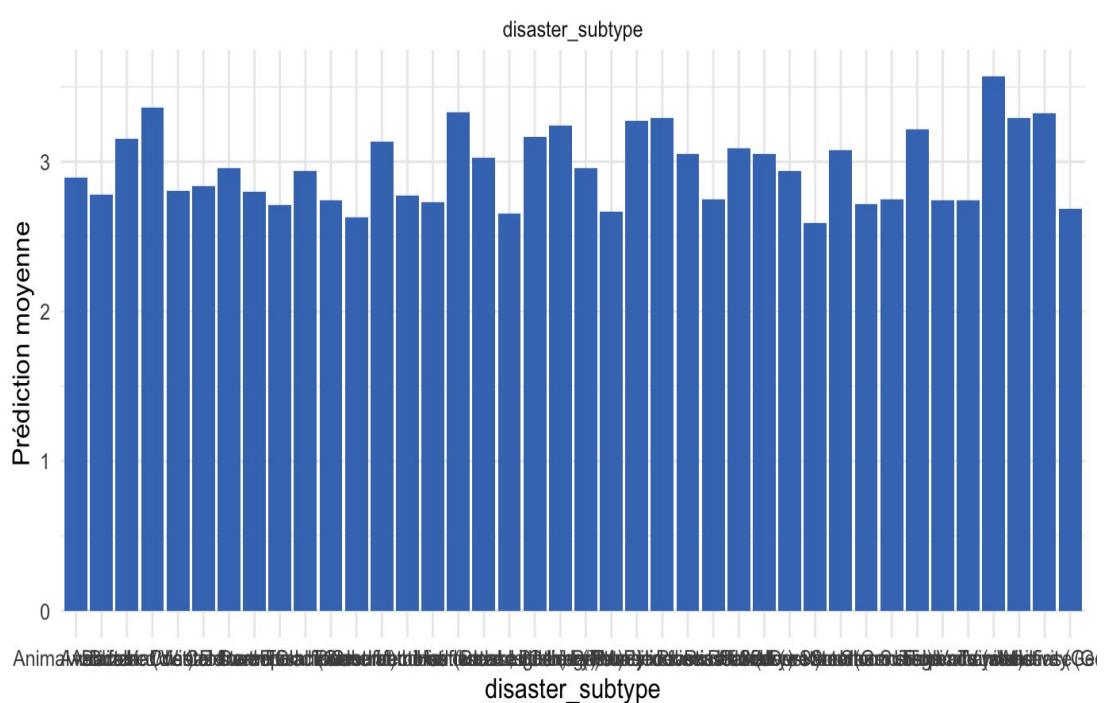
Relation entre log_affected et le nombre de décès

Created for the Random Forest model



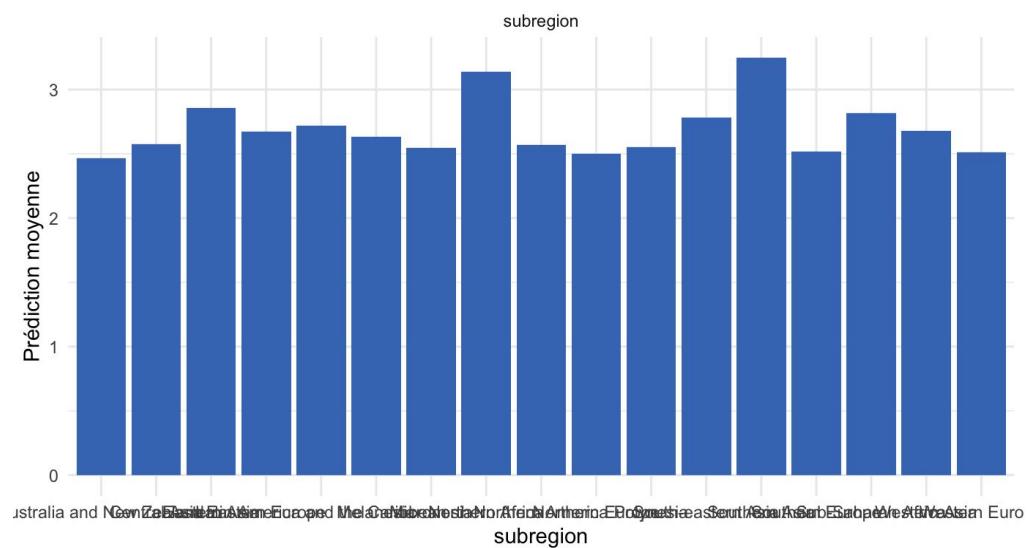
Relation entre disaster_subtype et le nombre de décès

Created for the Random Forest model



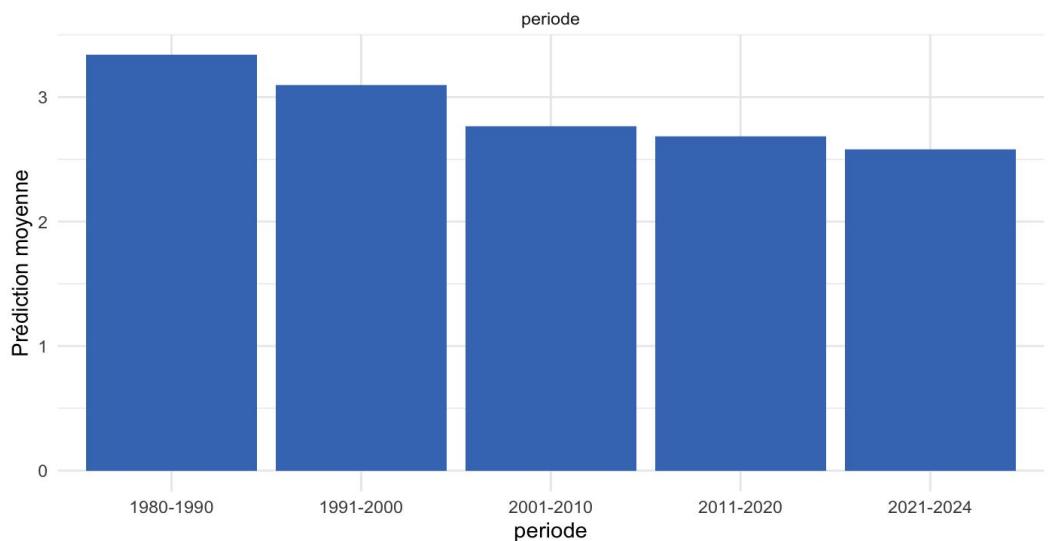
Relation entre subregion et le nombre de décès

Created for the Random Forest model



Relation entre periode et le nombre de décès

Created for the Random Forest model



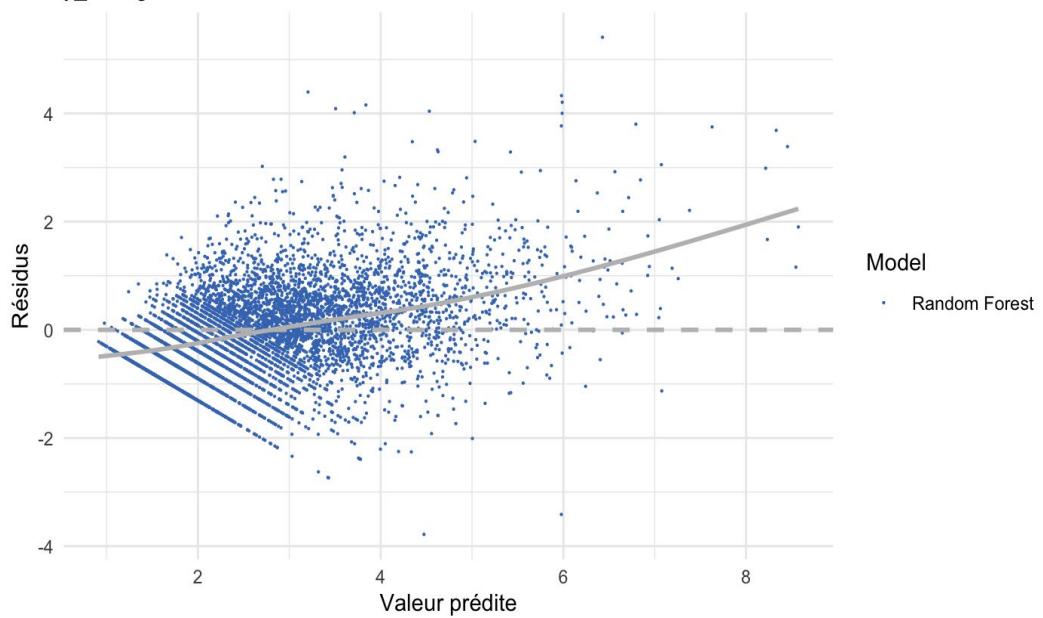
Les profils de dépendance partielle révèlent des patterns intéressants :

- **Log_affected** montre une relation non-linéaire avec : - Une phase stable initiale (0-5) - Une croissance progressive (5-15) - Une accélération marquée au-delà
- **Variables catégorielles** montrent des impacts différenciés selon les modalités

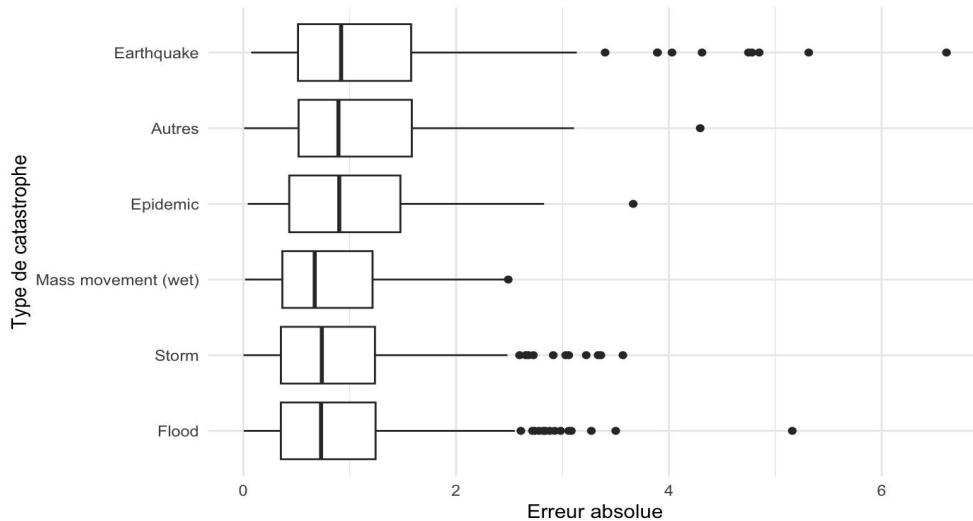
4.7.3. Analyse détaillée des erreurs :

Diagnostic des erreurs de prédition

y_hat against residuals



Distribution des erreurs par type de catastrophe



L'analyse des erreurs met en évidence :

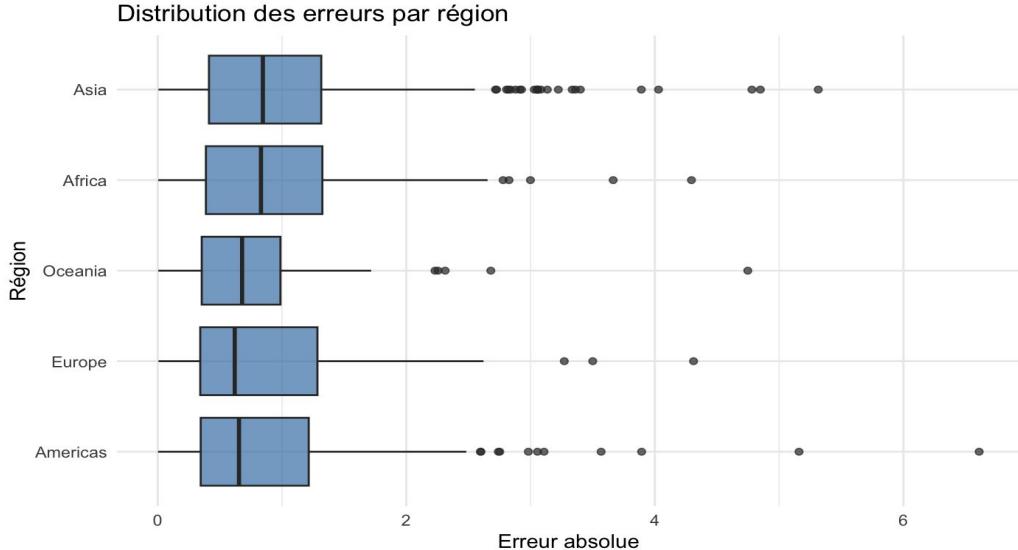
- Distribution des résidus** : - Centrée autour de zéro pour les événements moyens - Tendance à la surestimation pour les valeurs extrêmes
- Patterns par type de catastrophe** : - Certains types sont mieux prédits que d'autres - Plus grande variabilité pour les catastrophes majeures

Ces analyses détaillées confirment la robustesse générale du modèle Random Forest tout en identifiant ses limites, particulièrement pour les événements extrêmes. Cette compréhension approfondie permet d'utiliser le modèle de manière plus éclairée dans un contexte opérationnel.

4.7.4. Analyse quantitative des erreurs et limitations :

Statistiques descriptives des erreurs de prédiction

RMSE	MAE	Median_Error	SD_Error	Q1_Error	Q3_Error
1.202	0.936	0.784	0.755	0.378	1.299



Performance du modèle par région

region	RMSE	MAE	n_obs
Oceania	1.276	0.918	41
Asia	1.233	0.979	693
Americas	1.181	0.872	309
Africa	1.156	0.928	274
Europe	1.154	0.873	113

Voici une interprétation plus concise et pertinente des résultats présentés dans ces images, en faisant des liens avec les analyses précédentes :

- **Diagnostic des erreurs de prédiction**

Cette visualisation des résidus confirme les observations faites précédemment sur les limites du modèle Random Forest. Bien qu'il capture bien les événements moyens, le modèle a tendance à sous-estimer les catastrophes majeures entraînant un très grand nombre de décès. Cela se traduit par une plus grande dispersion des résidus pour les valeurs prédites élevées. Cette difficulté à prédire les événements extrêmes avait déjà été mise en évidence dans l'analyse des Partial Dependency Plots.

- **Distribution des erreurs par type de catastrophe**

Ce graphique apporte un éclairage complémentaire sur les performances du modèle. Il montre que certains types de catastrophes, comme les séismes et les mouvements de terrain, présentent une variabilité plus importante dans les erreurs de prédiction. Cette observation corrobore l'analyse précédente qui soulignait l'impact différencié des sous-types de catastrophes sur le nombre de décès prédict.

- **Performance du modèle par région**

Le tableau des métriques de performance régionales révèle des disparités géographiques, en cohérence avec l'analyse de la relation entre le sous-région et le nombre de décès. L'Océanie, qui présentait un profil de risque plus limité, affiche effectivement les meilleures performances du modèle. À l'inverse, l'Asie, identifiée comme une zone à risque élevé, montre les erreurs les plus importantes. Ces résultats soulignent l'importance de prendre en compte les spécificités régionales dans l'utilisation opérationnelle du modèle.

Dans l'ensemble, ces analyses approfondies des erreurs de prédiction permettent de mieux cerner les forces et les faiblesses du modèle Random Forest. Elles complètent utilement les insights précédemment obtenus sur l'importance des variables, les relations non-linéaires et les disparités régionales. Ces éléments seront essentiels pour affiner le modèle et adapter les stratégies d'assurance en conséquence.

5. Application actuarielle : Construction et Analyse du score globale de risque assuranciel (SGRA) :

5.1. Construction des indicateurs synthétiques :

Le secteur de l'assurance doit faire face à des défis importants liés aux catastrophes naturelles. L'objectif est de construire des indicateurs synthétiques permettant d'analyser le risque humain et d'élaborer un Score Global de Risque Assurantiel (SGRA).

Deux questions principales guident notre approche :

Le secteur de l'assurance doit faire face à des défis importants liés aux catastrophes naturelles :

- Comment évaluer la gravité des catastrophes pour affiner la segmentation des contrats ?
- Comment prioriser les actions assurantielles (tarification, prévention, réassurance) selon l'exposition globale aux risques ?

Notre analyse repose sur le taux de mortalité comme indicateur de base, calculé comme suit :

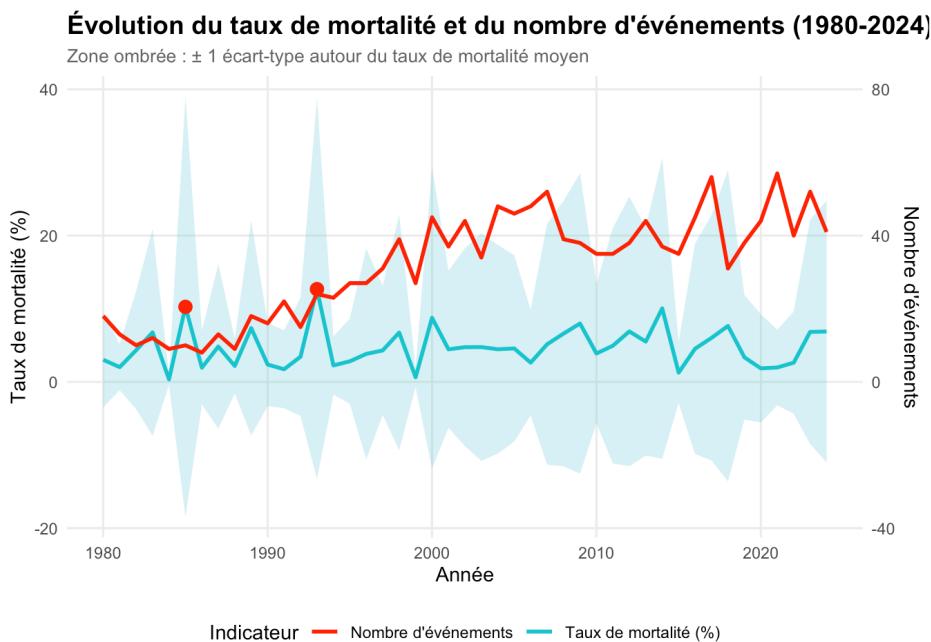
Le **taux de mortalité** mesure la gravité d'un événement en rapportant le nombre de décès à la population totale exposée :

$$\text{Taux de mortalité} = \frac{\text{décès}}{\text{total exposé}} \times 100$$

où : **Total exposé** = décès + personnes affectées.

L'objectif est de construire des indicateurs synthétiques permettant d'analyser le risque humain par région et d'élaborer un **Score Global de Risque Assurantiel (SGRA)**. Cet outil doit fournir une vue claire des régions prioritaires pour des actions spécifiques.

5.2. Analyse temporelle du Risque catastrophe :



L'analyse temporelle révèle une dynamique contrastée entre la fréquence des catastrophes et leur impact mortel.

Le nombre d'événements (ligne rouge) montre une tendance croissante continue, passant de 20 événements par an dans les années 1980 à plus de 40 aujourd'hui. Cette augmentation traduit une exposition accrue aux risques catastrophiques.

En revanche, le taux de mortalité (ligne turquoise) reste relativement stable, oscillant autour de 5%, malgré quelques pics notables (points rouges) dans les années 1980 et début 1990. La zone ombrée, représentant l'intervalle de confiance, suggère une variabilité décroissante au fil du temps.

Ce paradoxe - multiplication des événements avec stabilisation de la mortalité - indique une amélioration significative des systèmes de prévention et de gestion des catastrophes. Cette évolution souligne l'importance d'adapter les couvertures d'assurance à une fréquence accrue d'événements, mais avec des impacts individuels potentiellement mieux maîtrisés

5.3. Analyse Comparative des profils régionaux :

5.3.1. Construction des indicateurs régionaux :

Trois indicateurs clés ont été développés pour caractériser les profils de risque :

1- Indice de Sévérité Populationnelle (ISP) :

$$\text{ISP} = \text{Taux moyen de mortalité} \times \log(\text{Population totale exposée}) \times \text{Fréquence moyenne}$$

- **Taux moyen** : Mesure la gravité moyenne des événements
- **Population totale exposée** : Prend en compte la taille de la population exposée
- **Fréquence moyenne** : Intègre la fréquence des catastrophes

2- Score de Risque Combiné (SRC)

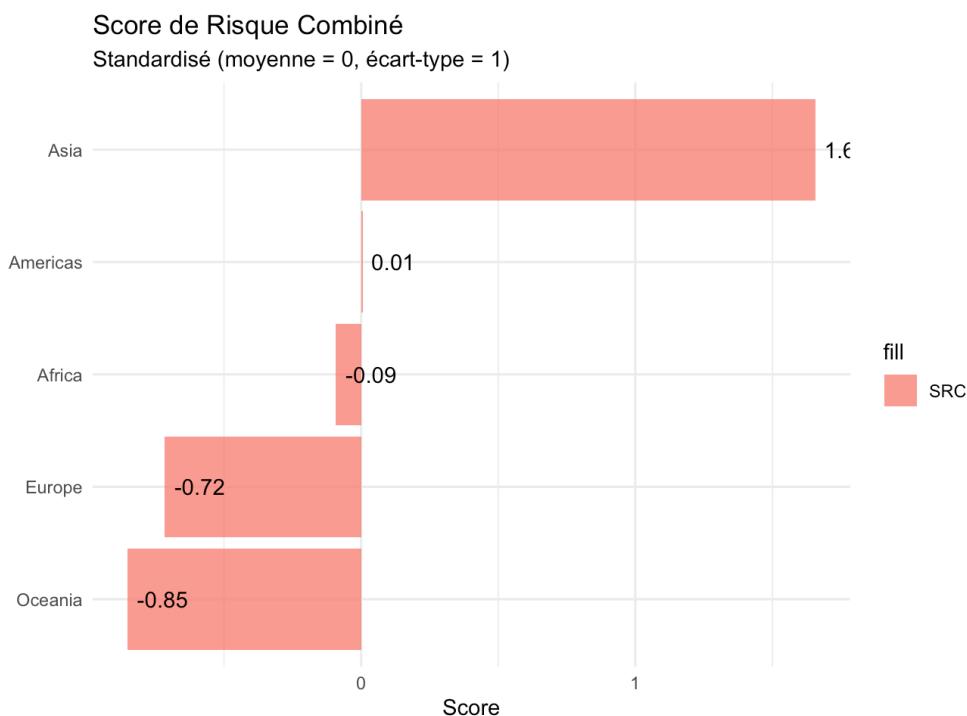
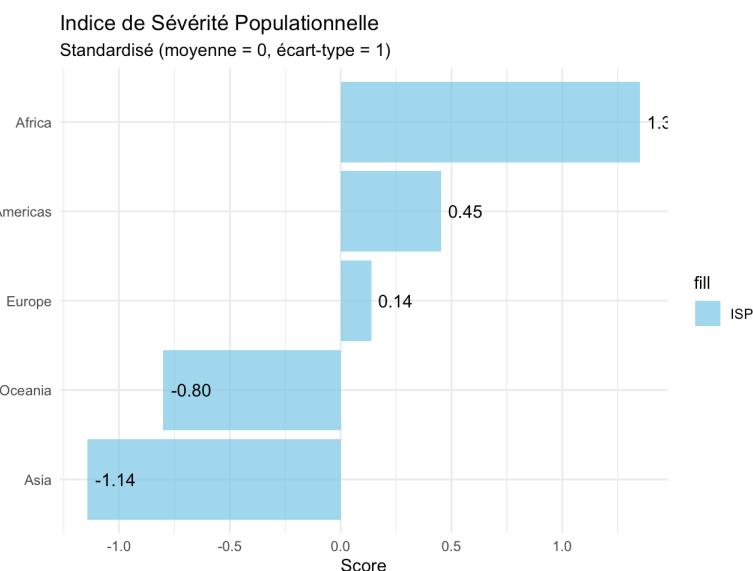
$\text{SRC} = \text{Taux moyen de mortalité} \times \text{Population totale exposée} \times \text{Fréquence moyenne}$

- Évalue la gravité pondérée par la population exposée

3- Indice de Résilience

$$\text{Résilience} = \frac{1}{\text{Taux moyen} \times \text{Fréquence moyenne}}$$

- Reflète la capacité à limiter la mortalité malgré la fréquence des événements





5.3.2. Synthèse des profils régionaux :

Synthèse des indicateurs de risque par région

Région	Rang ISP	Rang SRC	Rang Résilience	Niveau de Risque
Africa	1	3	5	Modéré
Americas	2	2	4	Modéré
Asia	5	1	1	Modéré
Europe	3	4	3	Modéré
Oceania	4	5	2	Faible

L'analyse des indicateurs standardisés met en évidence des profils de risque très contrastés selon les régions.

L'Afrique présente une vulnérabilité maximale avec l'ISP le plus élevé (+1.40), tandis que l'Asie, malgré sa population importante, affiche l'ISP le plus faible (-1.14). Cependant, l'Asie domine le classement SRC en raison de sa population exposée massive, alors que l'Europe et l'Océanie montrent des SRC plus faibles.

En termes de résilience, l'Asie se distingue positivement grâce à ses infrastructures développées, contrairement à l'Afrique qui présente la plus faible capacité de résilience.

Ces profils distincts appellent des stratégies assurantielles différencierées :

- Couvertures renforcées pour l'Afrique
- Gestion des risques de masse pour l'Asie
- Approche standardisée pour l'Europe et l'Océanie

Cette segmentation constitue une base essentielle pour l'adaptation des politiques de souscription et de tarification aux spécificités régionales.

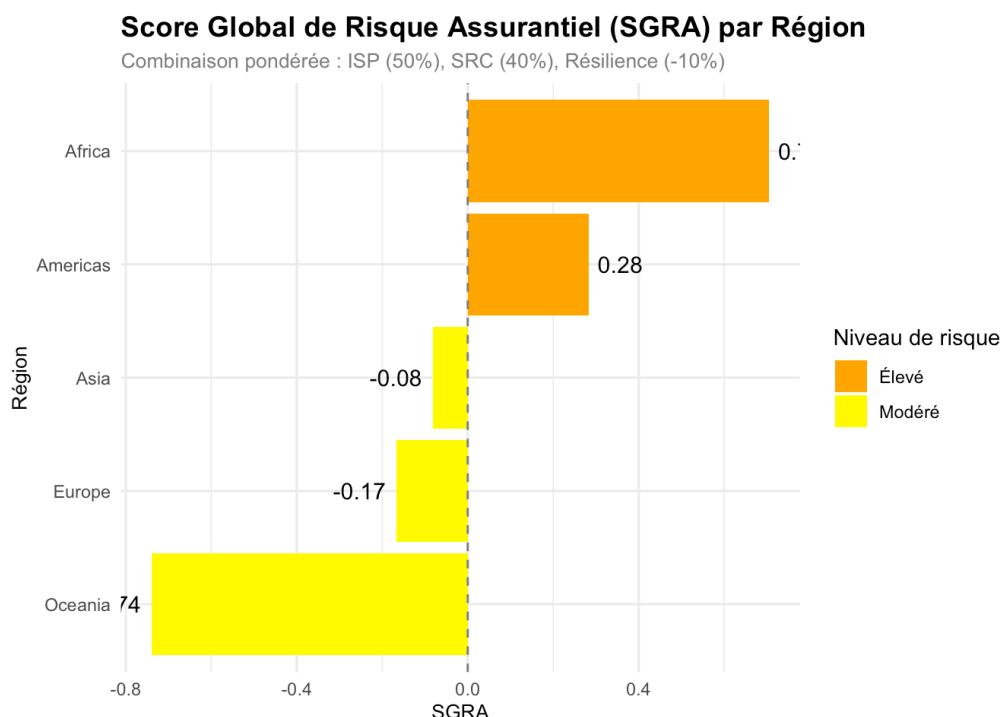
5.4. Élaboration du score global de risque Assuranciel (SGRA) :

Le SGRA combine les trois indicateurs précédents selon une pondération spécifique :

$$SGRA = \alpha \times ISP + \beta \times SRC - \gamma \times \text{Résilience}$$

avec les pondérations suivantes :

- $\alpha = 0.5$: importance majeure de la sévérité populationnelle
- $\beta = 0.4$: forte considération du risque combiné
- $\gamma = 0.1$: impact modérateur de la résilience



Décomposition du SGRA par composante

Région	SGRA	Niveau de risque	Impact ISP	Impact SRC	Impact Résilience
Africa	0.71	Élevé	0.68	-0.04	0.07
Americas	0.28	Élevé	0.23	0.00	0.05
Asia	-0.08	Modéré	-0.57	0.66	-0.17
Europe	-0.17	Modéré	0.07	-0.29	0.05
Oceania	-0.74	Modéré	-0.40	-0.34	0.00

La classification qui émerge du SGRA révèle quatre niveaux distincts de risque :

1- Risque Élevé (SGRA > 0.5)

- Afrique (0.71) : vulnérabilité maximale, résilience minimale
- Amériques (0.28) : exposition importante mais meilleure résilience

2- Risque Modéré (-0.5 < SGRA < 0.5)

- Asie (-0.08) : forte population exposée compensée par une bonne résilience

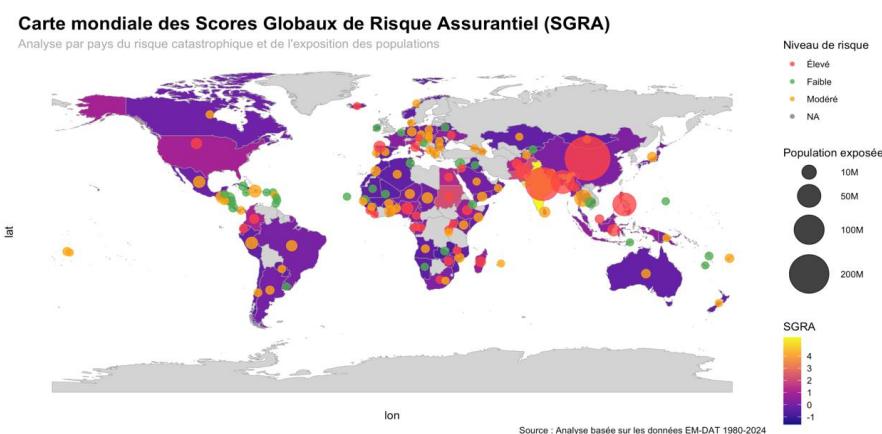
- Europe (-0.17) : profil équilibré sur tous les critères

3- Risque Faible (SGRA < -0.5) :

- Océanie (-0.74) : exposition limitée et bonne résilience

Cette hiérarchisation permet d'orienter les stratégies assurantielles de manière différenciée selon les régions.

5.5. Cartographie mondiale du risque assuranciel par Pays :



L'analyse des indicateurs standardisés met en évidence des profils de risque très contrastés selon les régions.

L'Afrique présente une vulnérabilité maximale avec l'ISP le plus élevé (+1.40), tandis que l'Asie, malgré sa population importante, affiche l'ISP le plus faible (-1.14). Cependant, l'Asie domine le classement SRC en raison de sa population exposée massive, alors que l'Europe et l'Océanie montrent des SRC plus faibles.

En termes de résilience, l'Asie se distingue positivement grâce à ses infrastructures développées, contrairement à l'Afrique qui présente la plus faible capacité de résilience.

Ces profils distincts appellent des stratégies assurantielles différencierées :

- Couvertures renforcées pour l'Afrique
- Gestion des risques de masse pour l'Asie
- Approche standardisée pour l'Europe et l'Océanie

Cette segmentation constitue une base essentielle pour l'adaptation des politiques de souscription et de tarification aux spécificités régionales.

Conclusion :

Cette étude approfondie a permis de développer un modèle prédictif robuste, basé sur un algorithme de Random Forest, pour estimer l'impact humain des catastrophes naturelles en termes de nombre de décès. L'analyse détaillée des résultats a mis en lumière plusieurs insights clés :

Tout d'abord, le nombre de personnes affectées s'est révélé être le principal facteur prédictif du nombre de décès, confirmant l'importance de cette variable dans la compréhension de l'impact des catastrophes. Certaines caractéristiques des événements, comme le sous-type de catastrophe et la région géographique, ont également montré une influence significative sur les prédictions.

Cependant, le modèle a également montré des limites dans la prédiction précise des événements extrêmes entraînant un très grand nombre de victimes. Cette difficulté à estimer correctement l'impact des catastrophes majeures a été mise en évidence à travers l'analyse des résidus et de la distribution des erreurs par type d'événement.

Sur le plan géographique, des disparités régionales importantes ont été identifiées, avec des performances du modèle plus faibles dans certaines zones comme l'Asie, en comparaison à des régions comme l'Océanie. Ces résultats soulignent l'importance de tenir compte des spécificités locales dans la gestion des risques liés aux catastrophes naturelles.

En fine, cette étude fournit une base solide pour aider le secteur de l'assurance à relever les défis posés par les catastrophes naturelles. Le modèle développé, ses forces et ses limites, constituent une source d'informations précieuse pour affiner les stratégies de tarification, de prévention et de réassurance. De plus, la construction d'indicateurs de risque synthétiques, tels que le Score Global de Risque Assurantiel (SGRA), offre un outil d'aide à la décision permettant de prioriser les actions à mener selon les profils de risque régionaux.

En conclusion, ce projet d'analyse des catastrophes naturelles a permis de générer des insights approfondis, à la fois sur le plan prédictif et opérationnel, afin de mieux appréhender et gérer les enjeux liés à ces événements pour le secteur de l'assurance. Les résultats obtenus constituent une base solide pour poursuivre les travaux d'amélioration du modèle et d'adaptation des pratiques assurantielles.