Modélisation Prédictive de l'Impact des Catastrophes Naturelles

Groupe : WAOH
Noms : Cylia, Fatima, Imane, Yikun

Sommaire

1 | Présentation générale et des données

1.1 | Contexte et problématique

- **Contexte des catastrophes naturelles**
- **Base des données EM-DAT**
- **La problématique de l'étude : 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 ?**
- **Proposer une application pour le secteur de l'assurance**

1.2 | Présentation des données

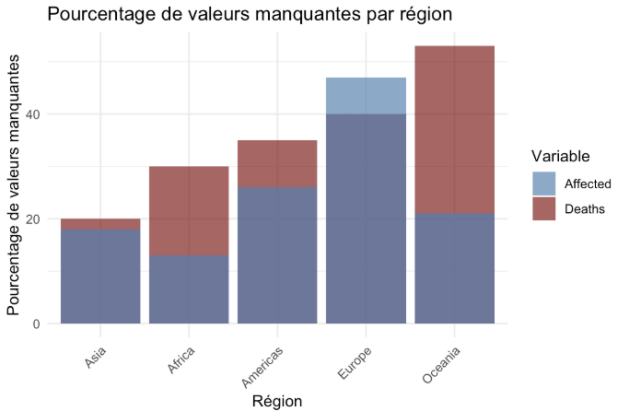
- **La base de données EM-DAT (Emergency Events Database)**
  - Les **types** de catastrophes (inondations, séismes, épidémies, etc.)
    - **Type de catastrophes**
    - **Sous-type de catastrophes**
  - Leur **localisation :**
    - **Continent**
    - **Pays**
  - Leurs impacts sur les populations (nombre de décès, de personnes affectées)
    - **Nombre de décès**
    - **Nombre de personnes affectés**

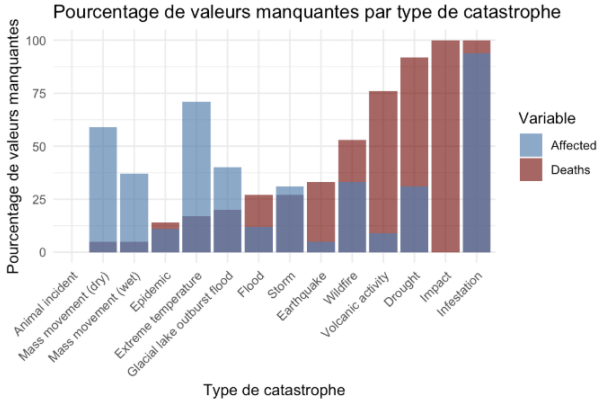
1.3 | Analyse et traitement des valeurs manquantes

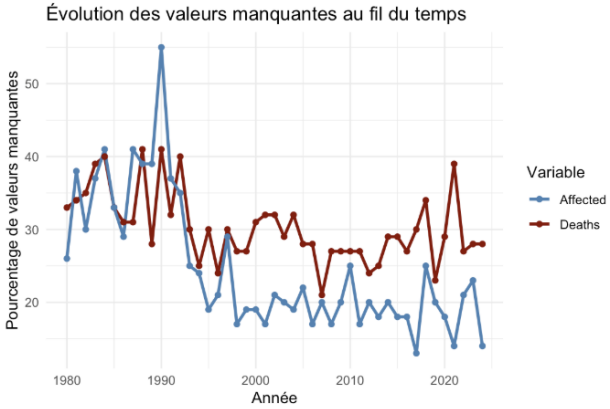
Pourcentage de valeurs manquantes par variable

	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

1.3 | Analyse et traitement des valeurs manquantes







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

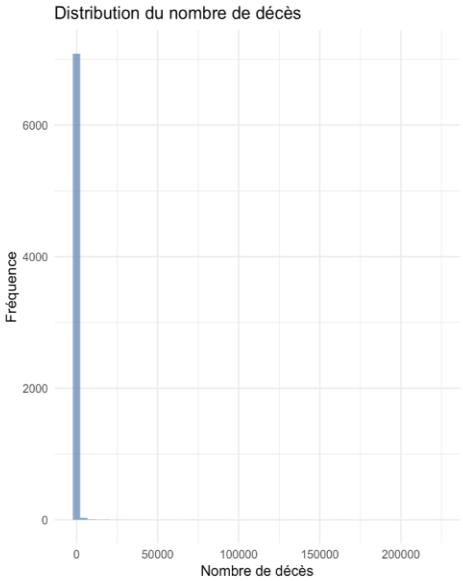
Pourcentage de données conservées : 47.86%

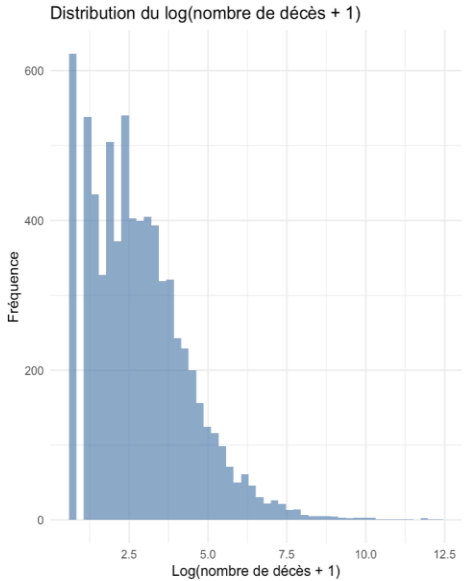
2 | Exploration et Analyse des Données

2.1 | Analyse Descriptive Univariée

- **Variable d'intérêt :**
  - total_deaths (nombre de décès pour chaque catastrophe naturelle)
- **Variables explicatives :**
  - disaster_type : Type de catastrophes
  - disaster_subtype : Sous-type de catastrophes
  - Region : Continents
  - total_affected : Personnes affectés (blessés, déplacés, nécessitant une assistance)
  - event_duration : Durée de l'évènement
  - Year : Année de l'évènement

2.1 | Analyse Descriptive Univariée : Variable Cible

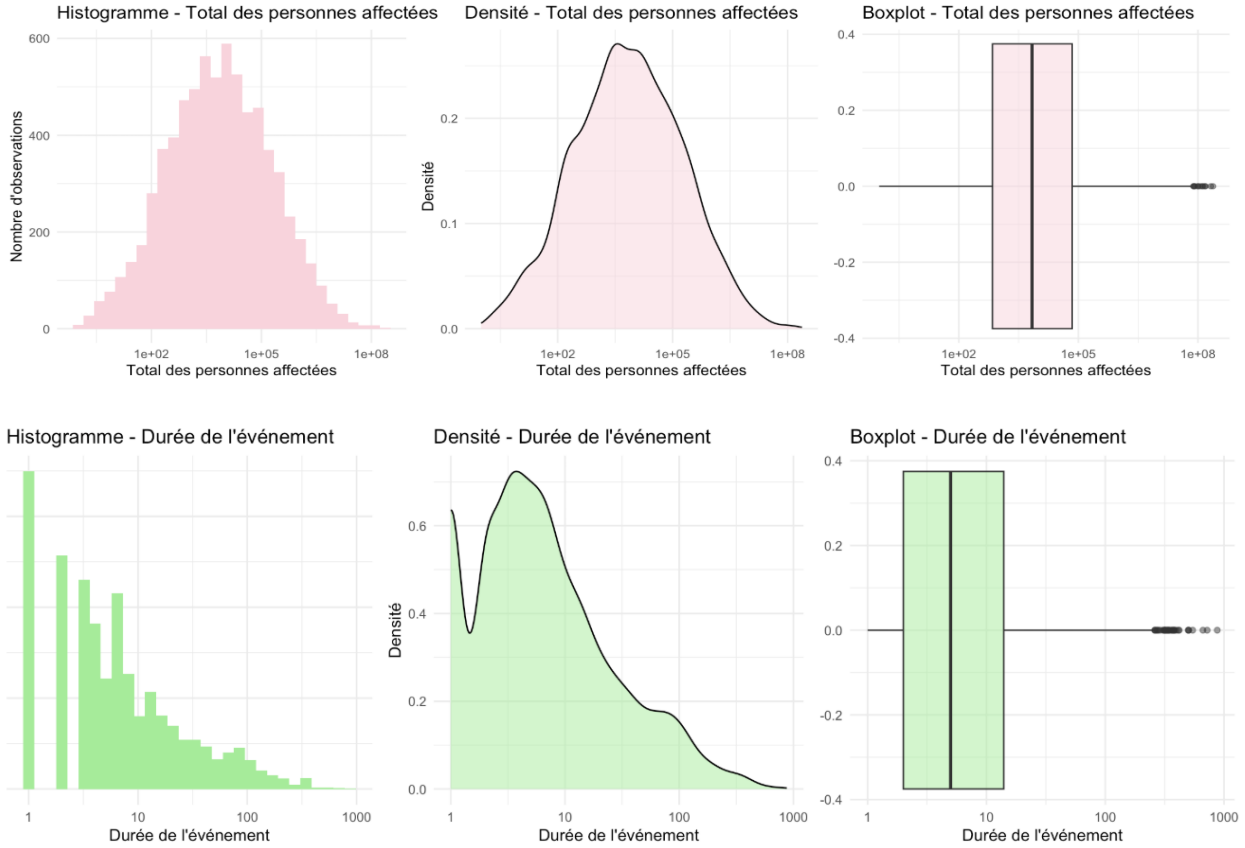




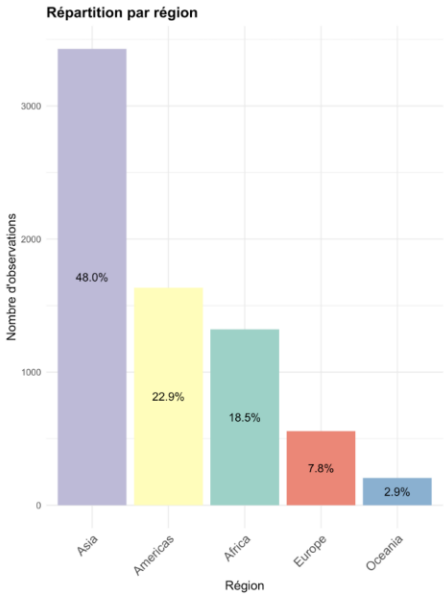
Statistiques descriptives du nombre de décès et de leur logarithme

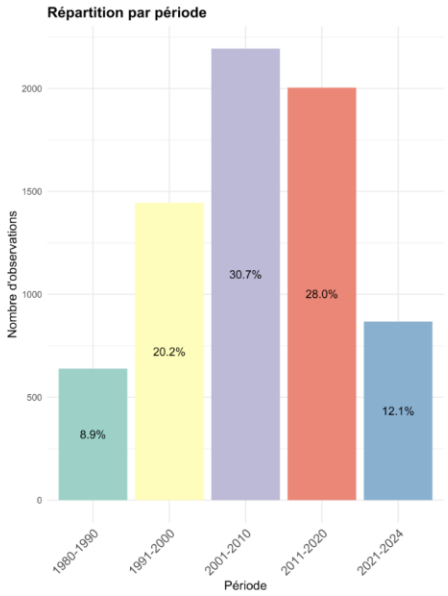
Statistiques	Décès	Log_Décès
Minimum	1.00	0.69
1er Quartile	4.00	1.61
Médiane	13.00	2.64
Moyenne	250.81	2.85
3e Quartile	41.00	3.74
Maximum	222570.00	12.31
Écart-type	4420.58	1.58

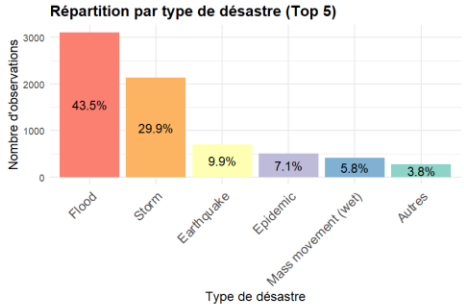
2.1 | Analyse Descriptive Univariée : Variables Quantitatives

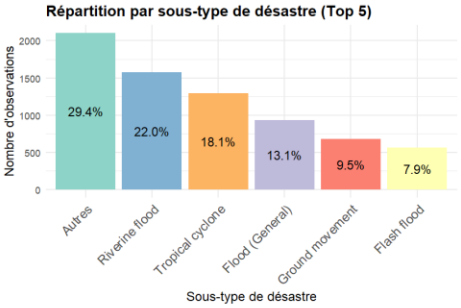


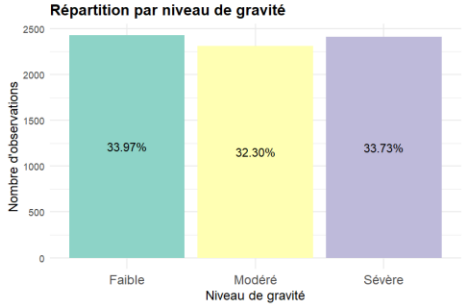
2.1 | Analyse Descriptive Univariée : Variables Qualitatives



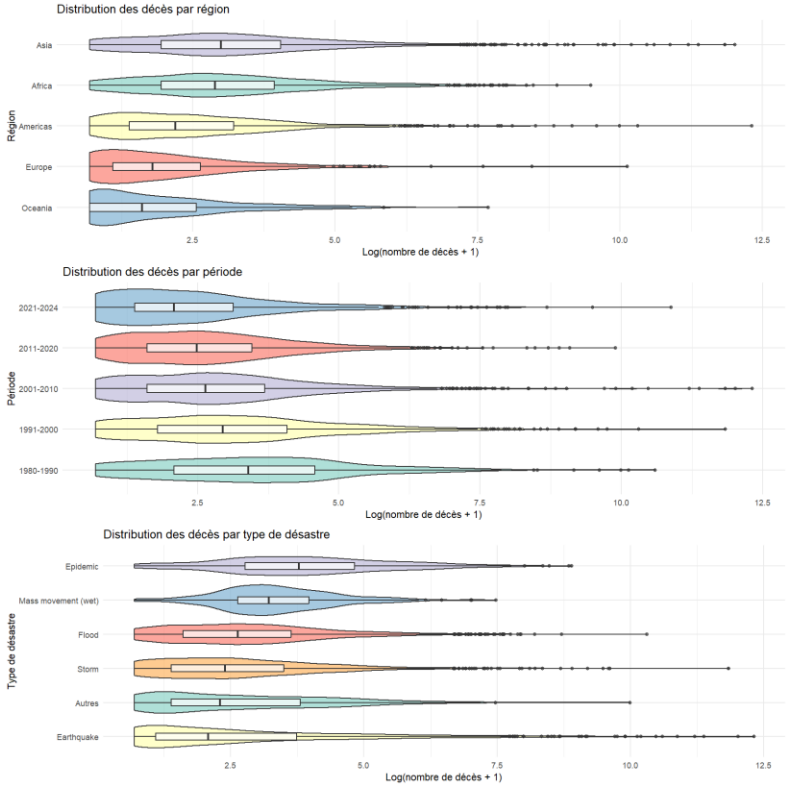


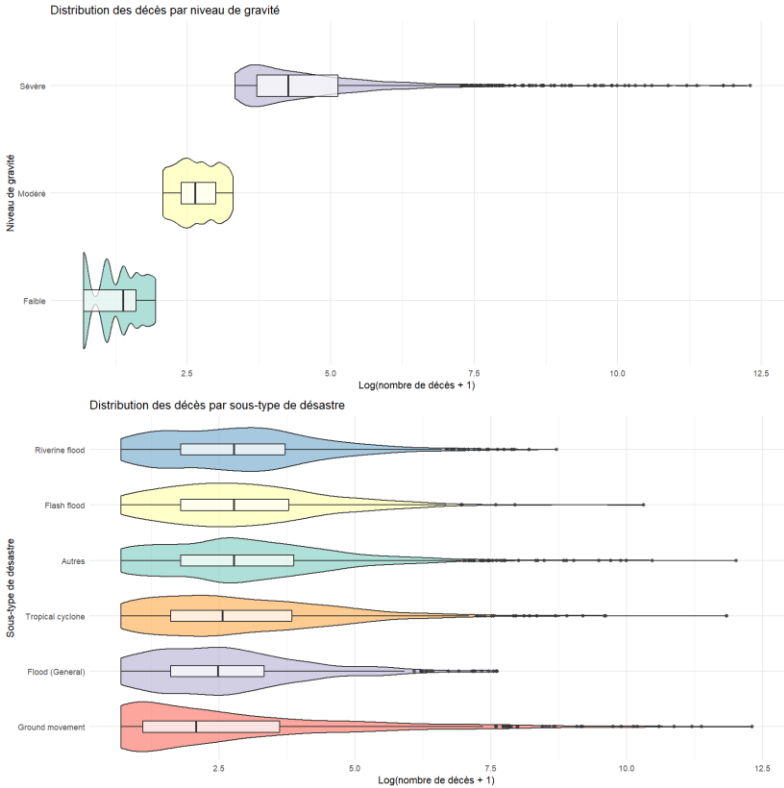




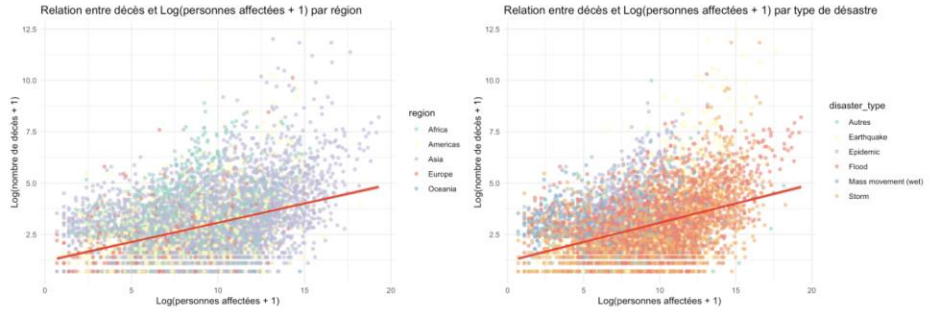


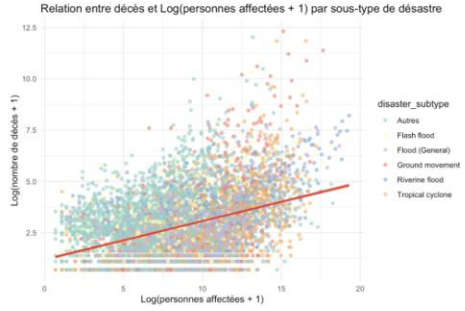
2.2 | Analyse Bivariée

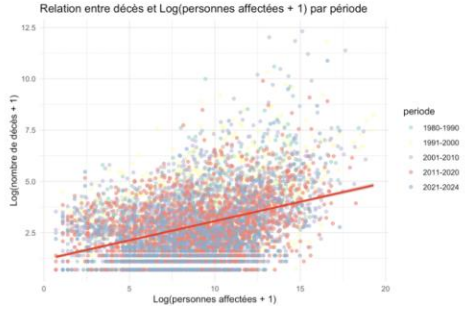


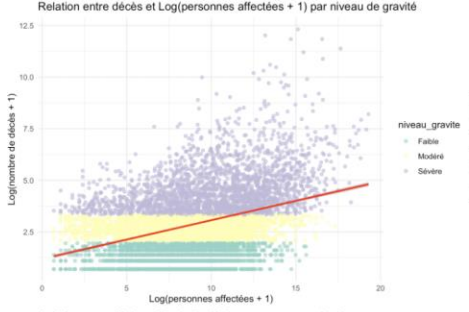


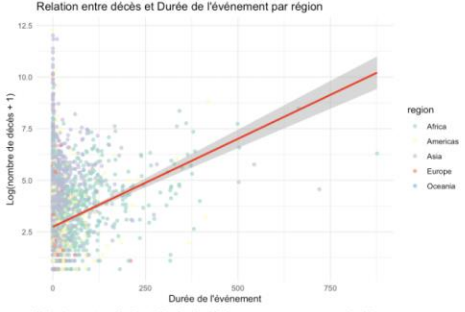
2.2 | Analyse Bivariée

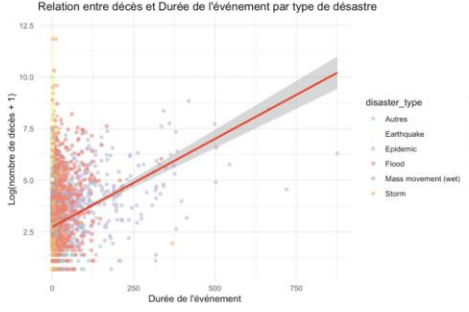


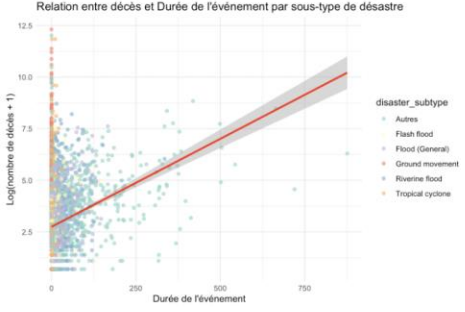




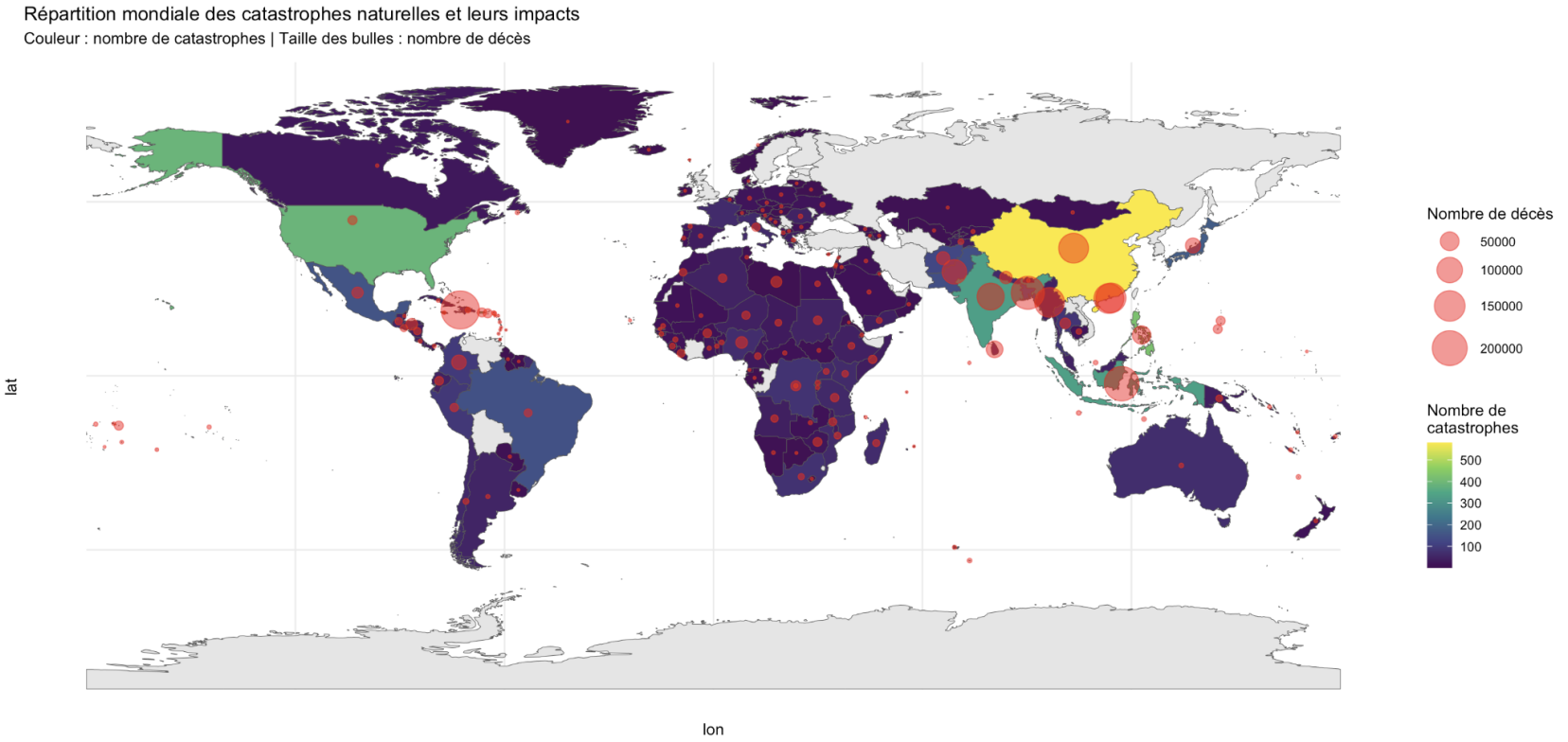








2.3 | Visualisation géographique



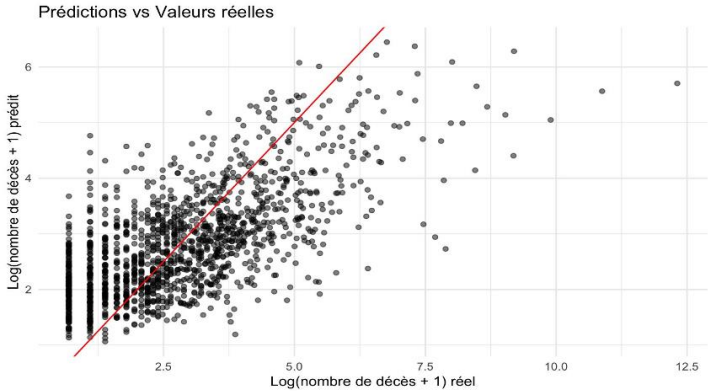
3 Modèles de Machine Learning

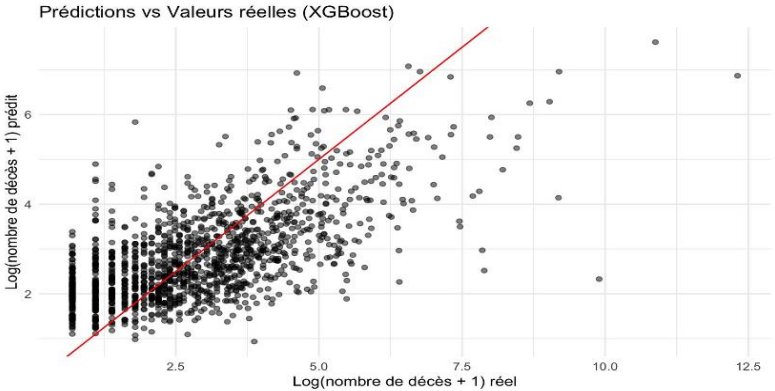
3.1 | Les modèles et les préparations des données

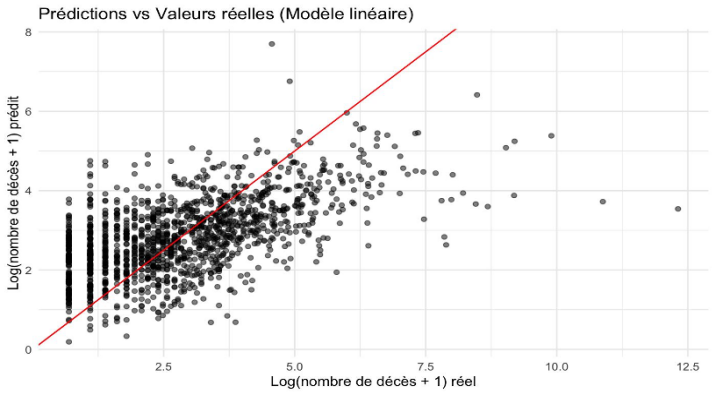
Modèle	Type	Justification
Random Forest	Ensemble Learning	Robuste aux outliers, gère naturellement les variables catégorielles, bonne capacité de généralisation
XGBoost	Gradient Boosting	Performance reconnue sur des données complexes, capture efficacement les relations non-linéaires
Régression Linéaire	Modèle linéaire	Modèle de référence simple, facilement interprétable, base de comparaison
SVR	Support Vector Machine	Efficace pour les relations non-linéaires, robuste avec les données normalisées

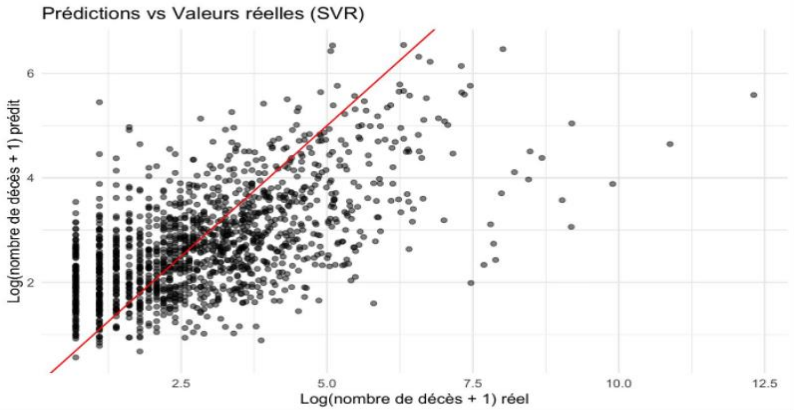
- **Ensemble d'entraînement : 5718 lignes (80% des données).**
- **Ensemble de test : 1430 lignes (20% des données).**

3.2 | Nos modèles :







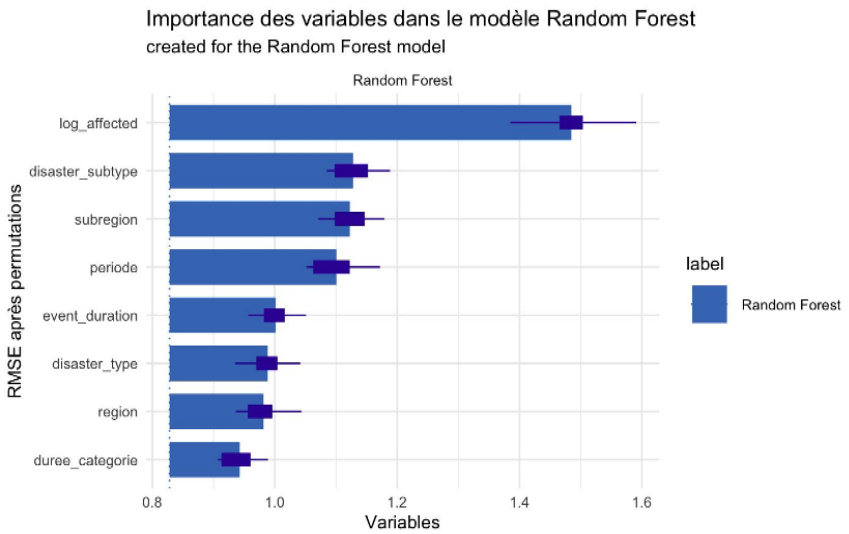


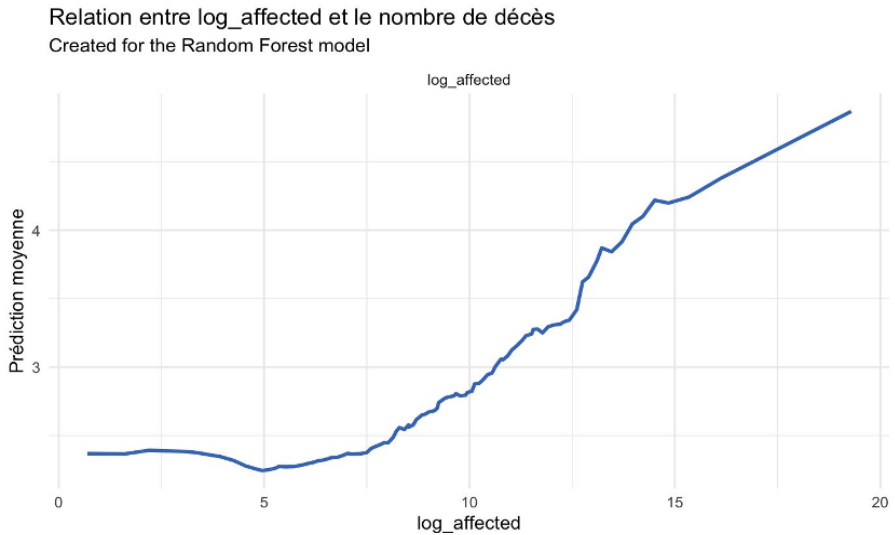
3.3 | Comparaison des modèles :

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

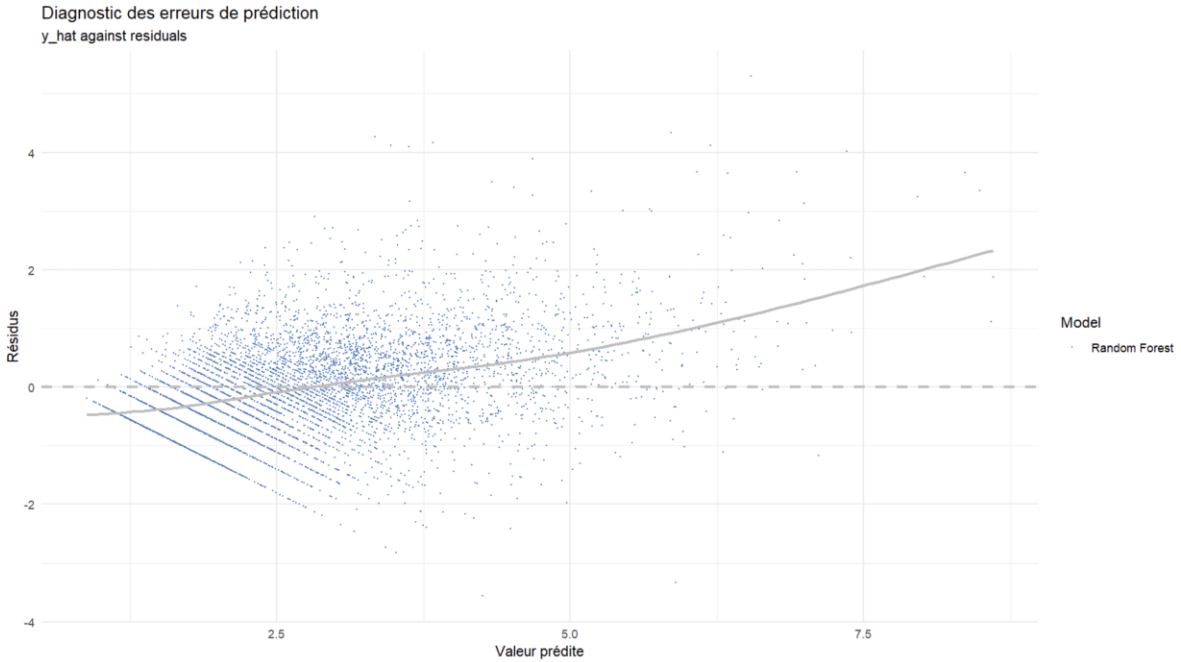
3.4 | Analyse approfondie du modèle Random Forest :





3.5 | Analyse approfondie du modèle Random Forest

- **Diagnostic des erreurs de prédiction**



3.6 | Analyse approfondie du modèle Random Forest

- **Analyse quantitative des erreurs et limitations**
  - Statistiques descriptives des erreurs

Métrique	Valeur
RMSE	1.202
MAE	0.936
Erreur médiane	0.784
Écart-type des erreurs	0.755
Q1	0.378
Q3	1.299

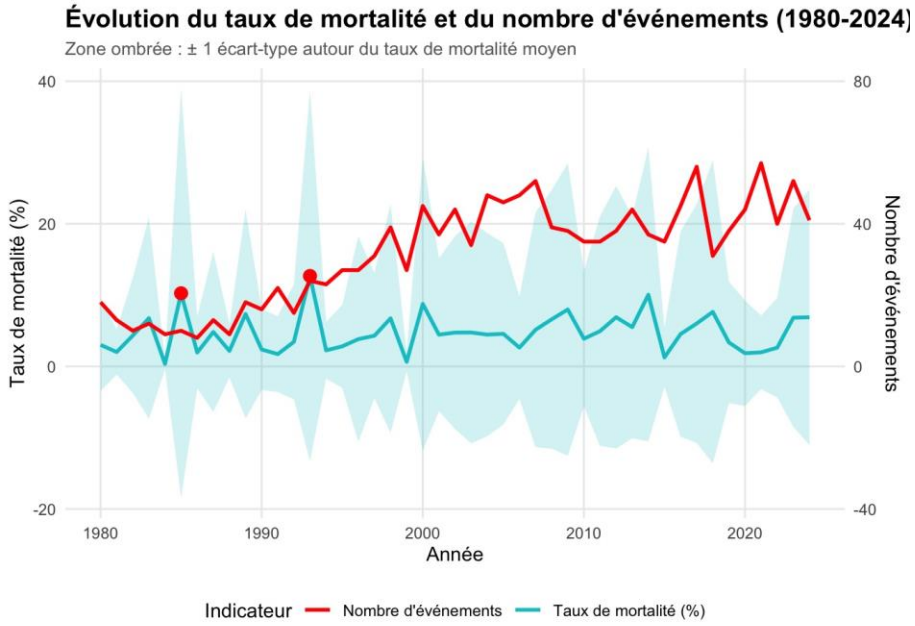
  - Performance des erreurs par région

4 Application actuarielle

4.1 | Approche de notre application

- **Construire d'indicateurs pour établir un Score Global de Risque Assurantiel (SGRA)**

- **Objectif :**
  - **Evaluer la gravité des catastrophes** pour affiner la segmentation des contrats
  - **Prioriser les actions assurantielles** selon l'exposition globale aux risques



- **Taux de mortalité :**

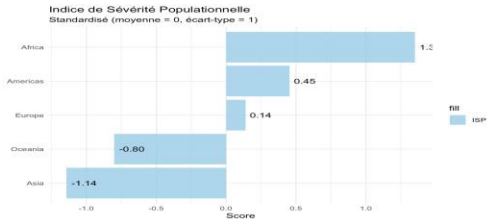
$$\text{Taux de mortalité} = \frac{\text{décès}}{\text{total exposé}} \times 100$$

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

4.2 | Construction d'indicateurs régionaux

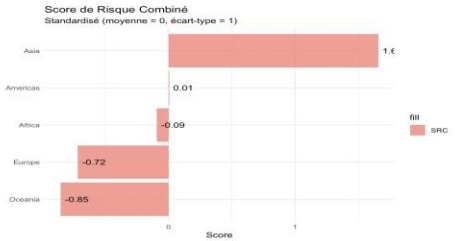
- **Indice de Sévérité Populationnelle (ISP)**

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



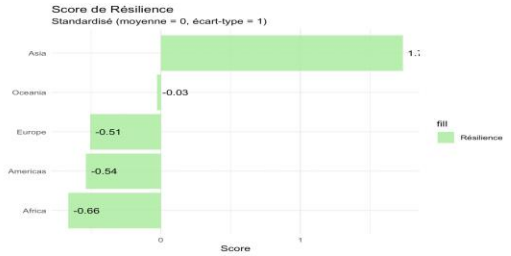
- **Score de Risque Combiné (SRC)**

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



- **Indice de Résilience**

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



4.3 | Synthèse des profils régionaux

Région	Rang ISP	Rang SRC	Rang Résilience
Africa	1	3	5
Americas	2	2	4
Asia	5	1	1
Europe	3	4	3
Oceania	4	5	2

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

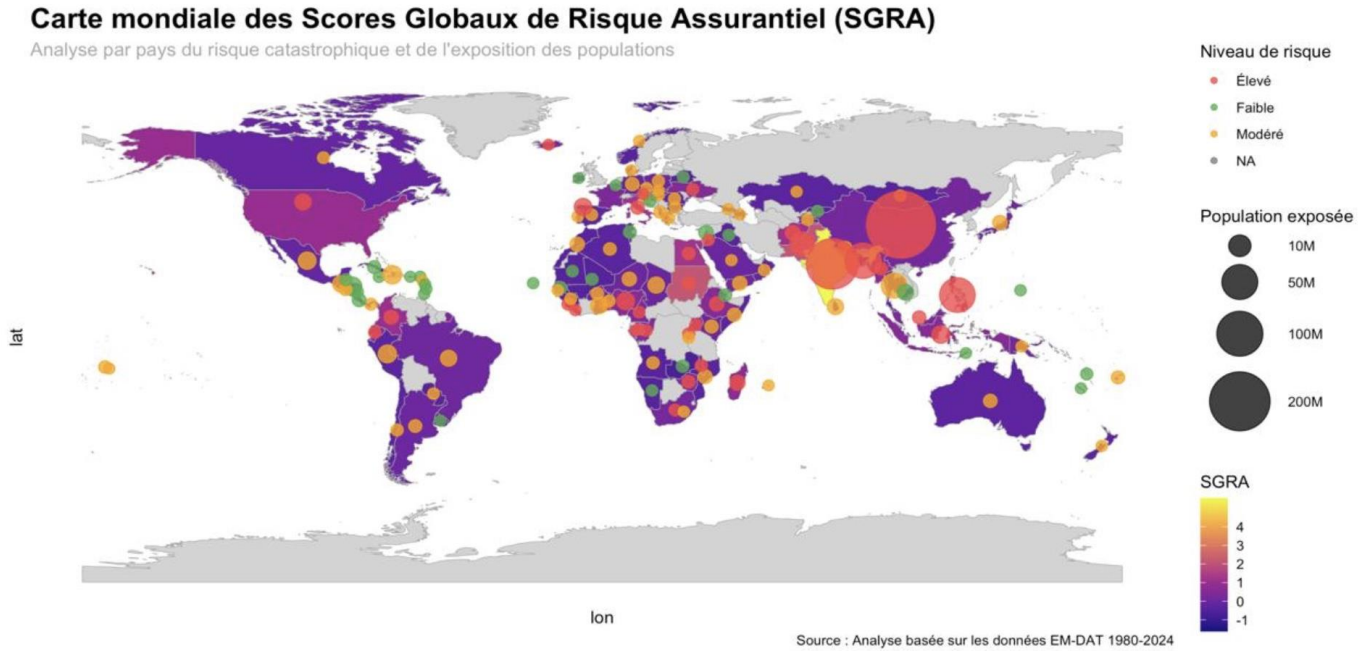
4.4 | Elaboration du Score Global de Risque Assurantiel

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

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

- Afrique : vulnérabilité maximale, résilience minimale
- Amériques : exposition importante mais meilleure résilience
- Asie : forte population exposée compensée par une bonne résilience
- Europe : profil équilibré sur tous les critères
- Océanie : exposition limitée et bonne résilience

4.4 | Analyse régionale du risque et de l'exposition des populations



5 Conclusion

4.4 | Dashboard Data Visualisation