ETUDE DE CAS, STATISTIQUE APPLIQUEE LES CYCLONES

Rakim FOFANA*, Israël MOIELWAY NGARTI**,
Ayindé TOUKOUROU***

adama-rakim.fofana@etu.u-bordeaux.fr*
israel.moielway.ngarti@etu.u-bordeaux.fr**
ayinde.toukourou@etu.u-bordeaux.fr***

Résumé : Face à l'impact croissant des phénomènes climatiques extrêmes, cette étude vise à approfondir notre compréhension des ouragans dans les régions du monde. Nous examinerons leurs modèles d'évolution, leurs répercussions socio-économiques, afin d'élaborer une analyse statistique bien précise.

1. Introduction

Dans un monde en constante évolution, les phénomènes climatiques extrêmes occupent une place prépondérante, affectant des régions du globe de manière souvent dévastatrice. Parmi ces manifestations, les ouragans se distinguent par leur puissance destructrice et leur fréquence croissante. Cette étude s'inscrit dans le contexte alarmant des changements climatiques, cherchant à approfondir notre compréhension des ouragans à l'échelle mondiale. Nous explorons leurs modèles d'évolution, les variations selon les régions, ainsi que leurs répercussions socio-économiques.

Face à cette menace grandissante, notre objectif est d'analyser à l'aide de méthodes statistiques et de formuler une problématique sur le déroulement des ouragans en fonction de ses différentes composantes à savoir la vitesse du vent, la durée, etc.

2. Préliminaires

- Les **ouragans** et les **cyclones** sont tous des termes pour désigner **le même phénomène météorologique** : une violente tempête tropicale caractérisée par des vents cycloniques circulaires. La principale différence réside **dans la région du globe** où se produit le phénomène : on parle d'ouragan dans **l'Atlantique Nord** et de cyclone dans **l'Océan Indien** et **le Pacifique Sud**. On peut ajouter à cette liste le **typhon**, qui lui se produit au **Pacifique Nord-Ouest.**
- Concernant l'Atlantique Nord, il y a eu **2 ouragans** de catégorie 5 entre les années 1992 et 1998. Ce chiffre est à la hausse entre les années 2010 et 2019 !

En effet, **12 ouragans** ont été répertoriés ; tous repartis entre 2016 (2), 2017 (4), 2018 (2) et 2019 (4).

- L'ouragan **Agnes** de 2012 n'est pas le seul événement du nom. Un ouragan Agnes a déjà été enregistré en 1989 au niveau de la côte Est des Etats-Unis. En fait, les noms de cyclones tropicaux sont réutilisés tous les six ans, à moins qu'un cyclone particulièrement dévastateur ait son nom retiré de la liste par l'Organisation Météorologique Mondiale.
- En 2021, il y a eu **112 cyclones** dans le monde selon la base de données *CatNat*. Le bilan humain s'est élevé à **4 201** morts et le coût financier total était d'environ **133 milliards de dollars**. Comparativement, en 2021, il y a eu **16 044 séismes** recensés avec une magnitude supérieure à 4.0 à l'échelle de Richter (faible et modéré) et plus de **6 000 inondations**. Les bilans humains et financiers varient en fonction de plusieurs facteurs.

3. Données

Nous possédons pour notre étude une base de données téléchargée sur EM-DAT - The international disaster database (emdat.be). Il s'agit d'une base de données de **15825 évènements** climatiques sur l'étendue planétaire depuis 1950 jusqu'à 2022. Et ce, pour tout type de catastrophe naturelle. Pour notre étude, nous nous focaliserons uniquement sur les catastrophes de type ouragan.

NOM DE VARIABLE	DESCRIPTION	FORMAT
Year	Année	Entier
Disaster Group	Nature des désastres	Caractère
Disaster Type	Type de désastre	Caractère
Continent	Continent	Caractère
Dis Mag Value	Intensité des vents	Entier
Dis Mag Scale	Echelle de l'intensité	Caractère
Start Day	Jour de début du désastre	Entier
End Day	Jour de fin du désastre	Entier
Total Deaths	Total de décès enregistré	Entier
TotalDamages, Adjusted ('000 USD$)	Coûts induits ajustés (en KiloDollars)	Entier

TAB. 1 – *Tableau récapitulatif de la description de certaines variables.*

3.1 Préparation des données

Les logiciels utilisés pour notre étude statistique sont le tableur *Excel* et *le langage R* (via l'environnement *RStudio*). La préparation consiste à faire le « nettoyage » des données. On transforme d'abord les variables quantitatives en facteur, ensuite nous vérifions si nos données contiennent des valeurs manquantes et/ou aberrantes qu'on remplacera par des valeurs appropriées, ou supprimera tout simplement. Une colonne a été ajoutée à notre tableau pour mesurer la durée de chaque événement (en jours) ; il s'agit de la différence entre le jour de fin et le début d'une catastrophe.

3.2 Valeurs manquantes et aberrantes

- ### Valeurs manquantes

Nous distinguons deux cas de « valeurs manquantes » dans notre jeu de données : celles manquantes et celles qui ne figurent pas du tout pour certaines variables.

Les valeurs manquantes dans notre base de données indiquent des champs qui étaient remplis mais dont les données ont été perdues ou ne sont plus disponibles, tandis que les valeurs qui ne figurent pas du tout n'ont jamais été collectées ou enregistrées.

C'est le cas par exemple de la variable *Continent*. En effet, l'Europe n'enregistre aucune catastrophe de type « Storm » en 1950.

- ### Valeurs aberrantes

Les valeurs aberrantes sont des observations généralement éloignées des autres observations et qui peuvent fausser l'analyse et les résultats si elles ne sont pas correctement traitées.

Pour les valeurs aberrantes, nous avons d'abord calculer leur pourcentage et simultanément évaluer leur impact sur la moyenne. Nous déciderons en fonction des cas.

4. Analyse

- **Fréquence par grande catégories de région**

Après filtration du tableau de données sur la colonne *Disaster Type* sur Excel, on obtient une base de données restreinte aux évènements de type Storm, donc ouragans. Toujours avec Excel, on s'aide d'un tableau croisé dynamique qui nous affiche le nombre de désastre, en fonction des régions où elles se sont déroulées.

Régions	Nombre de Disaster Type
Australia and New Zealand	100
Caribbean	341
Central America	244
Central Asia	5
Eastern Africa	183
Eastern Asia	726
Eastern Europe	100
Melanesia	111
Micronesia	27
Middle Africa	22
Northern Africa	19
Northern America	716
Northern Europe	105
Polynesia	65
Russian Federation	4
South America	92
South-Eastern Asia	597
Southern Africa	44
Southern Asia	437
Southern Europe	91
Western Africa	36
Western Asia	66
Western Europe	258
Total général	**4389**

TAB. 2 – *Tableau croisé dynamique des régions par ouragans enregistrés*

Pour la fréquence en pourcentage, on décide d'ajouter une colonne où nous effectuons le rapport de l'effectif sur le total général. Par ce tableau, on voit de prime abord que l'Asie recense le plus de 40% de catastrophe de type ouragan. Les régions les plus touchées sont l'Asie de l'Est (environ 17%), l'Asie du Sud et du Sud-Est (environ 10% et 14%). Ensuite vient l'Amérique avec 30% dont les principales régions concernées sont l'Amérique du Nord (16%), le Moyen Caraïbe (8%) l'Amérique Centrale (6%). Comparativement, les chiffes en Afrique, Europe et Océanie pèsent moins. En effet, 7% sont répertoriés en Afrique et en Océanie puis 13% en Europe.

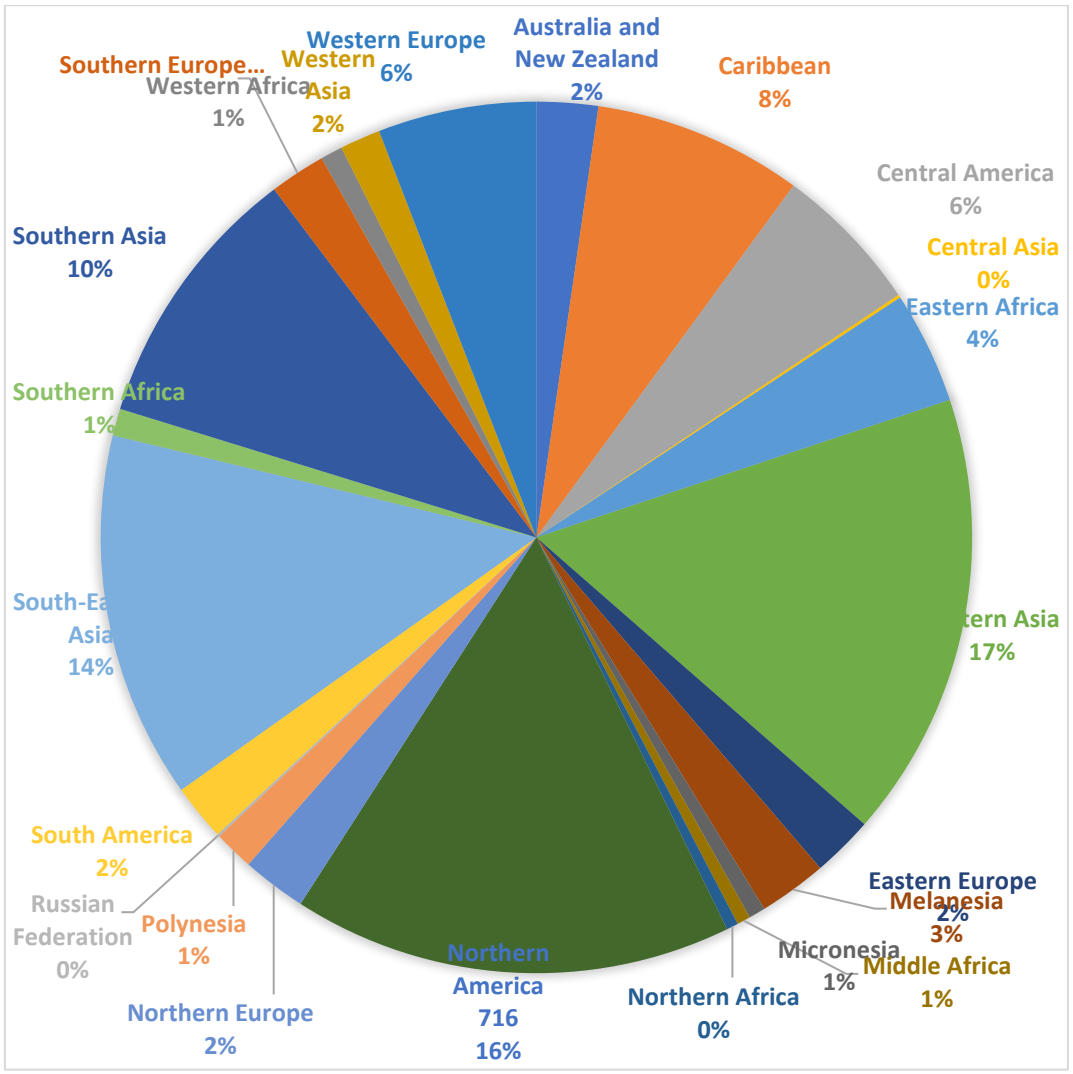

FIG.1– *Quantité d'ouragans par Régions.*

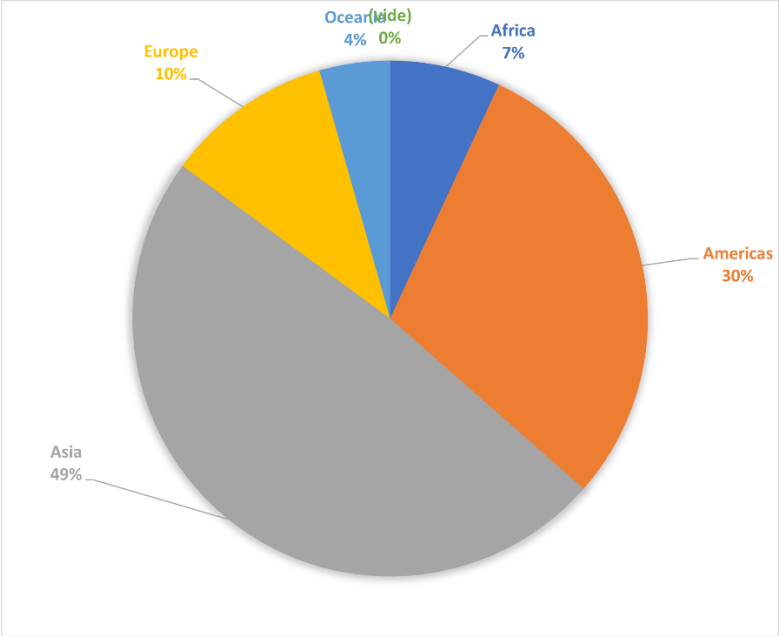

FIG.2 – *Quantités d'ouragans par continent.*

- **Evénement Agnes**

Par filtration de notre tableau, il y a 6 évènements nommés Agnes. Il n'y en aura plus en Atlantique Nord car, comme indiqué dans les préliminaires, le nom est souvent retiré en raison de catastrophes dévastatrices. L'ouragan Agnes a été retiré après la saison 1972 par son intensité et des dommages occasionnés en Atlantique Nord.

- **Analyse temporelle du nombre d'ouragan**

Identiquement à l'analyse des fréquences, nous avons effectué, après filtration de notre tableau, un tableau croisé dynamique affichant le nombre d'ouragans au fil des années. On obtient le tableau TAB.3 suivant :

Années	Nombre de Disaster Type
1950	7
1951	11
1952	9
1953	11
1954	10
1955	19
1956	14
1957	9
1958	13
1959	16
1960	20
1961	14
1962	13
1963	21
1964	22
1965	19
1966	27
1967	25
1968	25
1969	25

TAB. 3 –*Extrait du tableau croisé dynamique des années par ouragans enregistrés.*

A l'aide du tableau précédent, on obtient graphiquement une courbe d'évolution des ouragans par années (figures 3). On observe une croissance irrégulière du nombre d'ouragan au fil des années sans jamais atteindre les 160 ouragans enregistrés. Cette croissance irrégulière commence à partir de l'année 1982 puis se poursuit jusqu'en 2022, date de fin d'observations pour la base de données. L'année 1990 a quant à elle connue le maximum d'ouragans (135) comparativement à l'année de début d'observation 1950, qui elle n'a enregistrée que 7 ouragans. Toutefois, la présence des valeurs manquantes, que le tableau croisé dynamique n'a pas pris en compte, ne nous permet pas d'affiner nos observations. Cependant, la tendance générale n'est pas tant impactée.

La figure 3b est l'identique de la 2a à l'exception qu'elle a été faite sous R, à l'aide du package *ggplot2*.

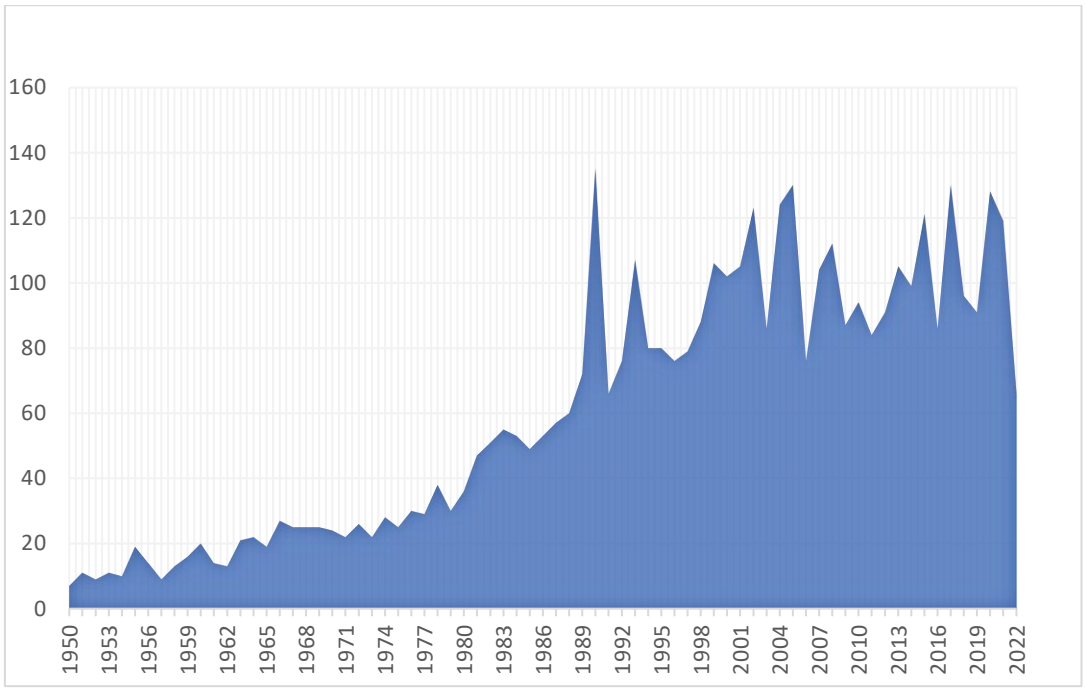

FIG. 3a –*Evolution du nombre ouragans au fil des années (avec Excel)*

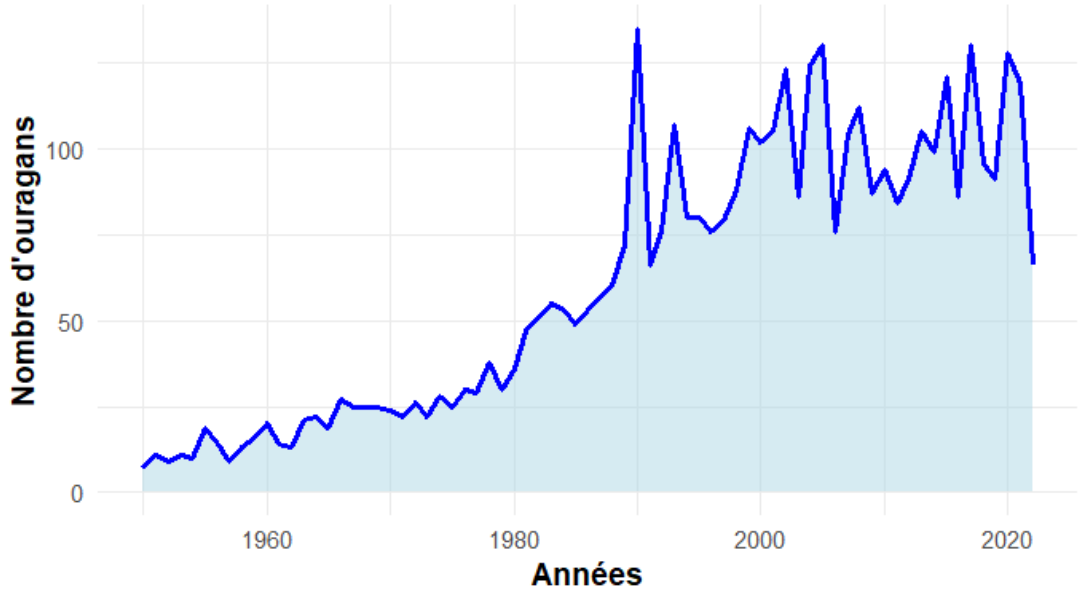

FIG. 3b –*Evolution du nombre d'ouragans au fil des années (avec RStudio).*

- **Analyse temporelle du nombre d'ouragans, Amérique vs Europe**

Cette fois, un tableau croisé dynamique sur les années, le continent et le nombre de désastres enregistrés a été fait. Une fois le tableau obtenu, nous avons appliqué un filtre sur la colonne *Continent* pour afficher uniquement les valeurs *Europe* et *Americas* (voir tableau 4).

Années	Americas	Europe	Total général
1950	5		5
1951	2		2
1952	2		2
1953	6	1	7
1954	6		6
1955	11		11
1956	4		4

TAB.4 – *Extrait du tableau croisé dynamique du nombre d'ouragans en Amérique et Europe au fil des années*.

A partir du tableau croisé dynamique, il est observé que le nombre d'ouragans enregistrés en Amérique est significativement élevé que celui en Europe. En effet, 1 393 ouragans répertoriés en Europe contre 558 en Amérique. Cette observation souligne une disparité notable dans la fréquence des ouragans entre les deux régions. Pour en apprendre d'avantage, nous avons tracé une courbe de fréquence aussi bien sur Excel, que sur RStudio.

Par observation de la courbe de tendance, on note une croissance irrégulière des nombres d'ouragans en Amérique, tandis que l'Europe présente une tendance plus erratique. Cette tendance se justifie d'une part par des « valeurs manquantes », qui ne figurent pas du tout. Par exemple, le vide observé en 1950 sur le tableau 4 n'est pas réellement manquante, mais plutôt que l'Europe n'avait aucun évènement de type Storm à cette année dans la base de données. D'où les « vides » sur les figures 4a et 4b.

Particulièrement en 1988, l'Amérique a connu un nombre substantiel d'ouragans par rapport à l'Europe. L'Europe comptabilise elle, cette année une augmentation significative du nombre de catastrophes, atteignant plus de 65 ouragans enregistrés en 1990 contre 26 en Amérique cette même année.

Par la suite, le nombre d'ouragans en Europe a montré une diminution après cette année de pointe, suivie d'une tendance irrégulière. En revanche, l'Amérique maintient une croissance assez modérée.

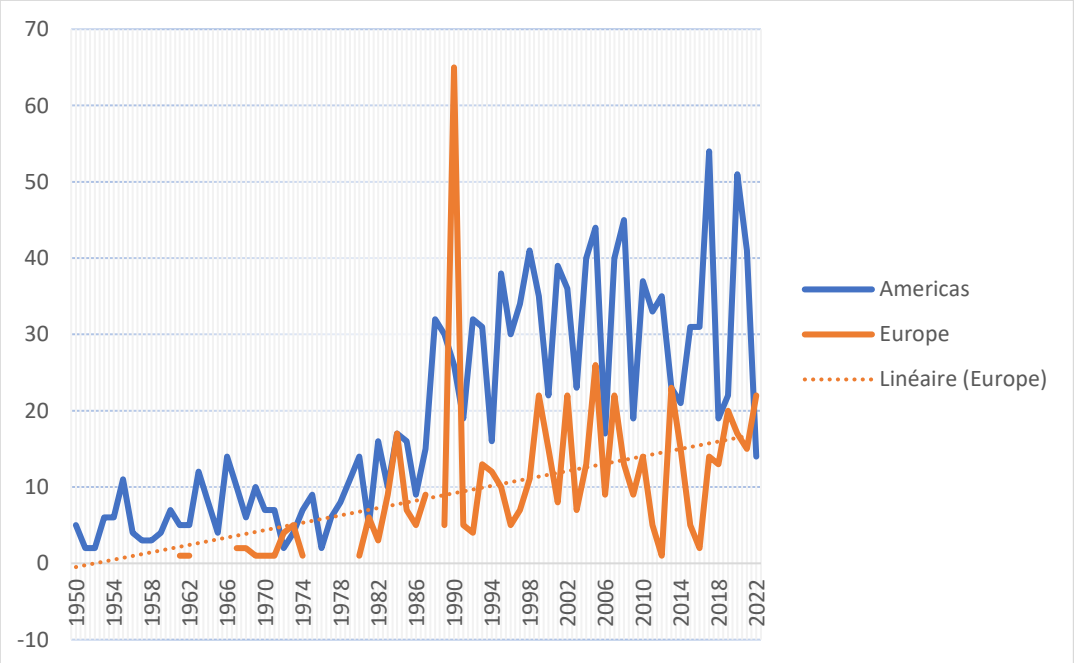

FIG.4a – *Evolution du nombre d'ouragan en Amérique et Europe au fil des années (avec Excel)*

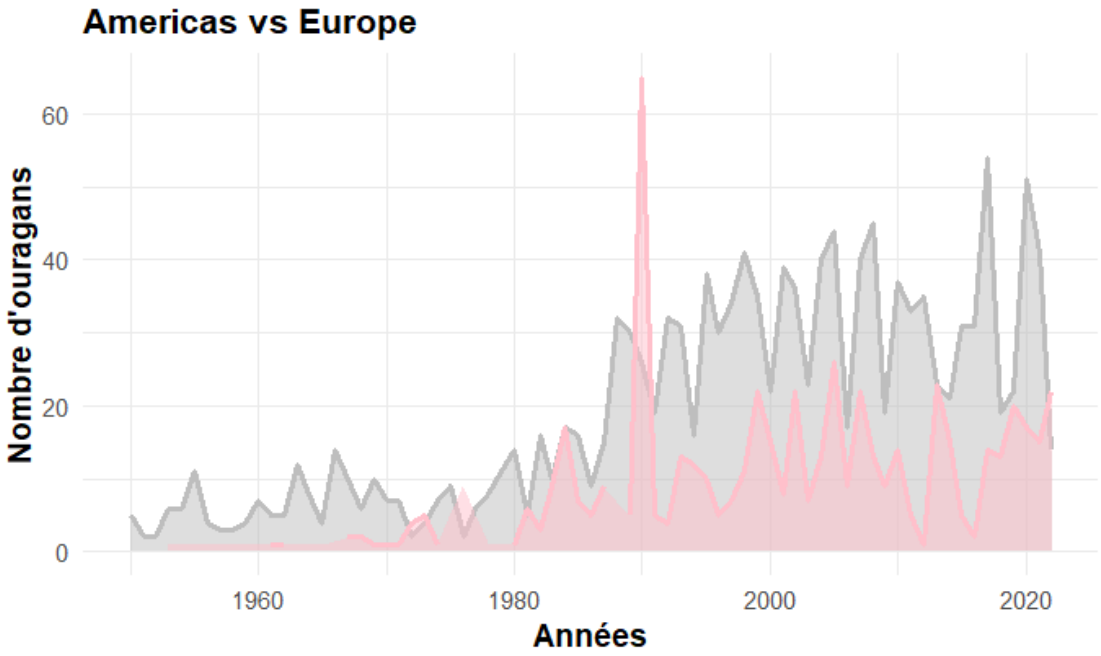

FIG.4b – *Evolution du nombre d'ouragans en Amérique et en Europe avec (Rstudio). Rose : Europe ; Gris : Amérique.*

- **Evénement années 2000, Europe vs Amérique, intensités des vents**

En considérant les colonnes Dis Mag Value des continent Europe et Amérique, on constate des valeurs manquantes et aberrantes.

Nous avons observé le pourcentage en valeurs manquantes grâce au package *naniar* (FIG.5). La figure 4 nous montre qu'il manque environs 87.5% d'observations pour les intensités en Europe et environs 75% pour les intensités en Amérique. Les observations des intensités en Amérique sont plus nombreuses que celles en Europe en raison des observations qui n'ont pas été faites à certaines années en Europe. Le tableau croisé dynamique ne comptabilise pas les valeurs manquantes. De plus, dans nos recherches, les chiffres ne concordent pas souvent avec les sources ; la moyenne des observations ne paraissant pas bon indicateur face à ces grandes quantités, nous les garderons donc vides, sans aucune modification.

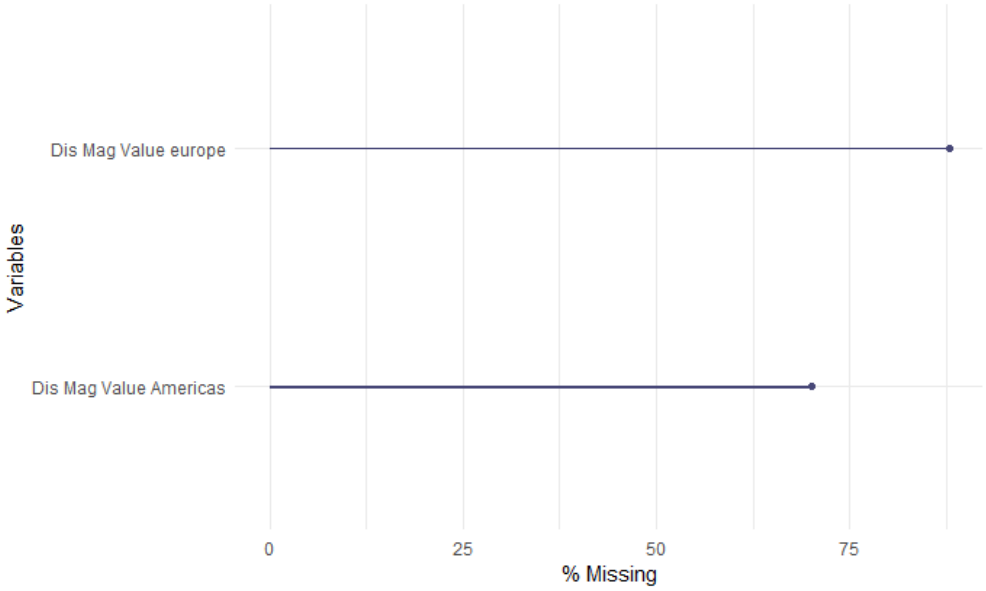

FIG.5 – *Pourcentage des valeurs manquantes des intensités manquantes en Amérique et en Europe.*

A l'aide de la fonction *SI.CONDITIONS* et *LIGNES*, on trouve le pourcentage de valeur aberrantes pour les deux variables intensité. 20% des observations américaines contre 10% des observations européennes. Les cyclones sont de base des phénomènes très variables, leur imputé une intensité, qu'elle soit moyenne ou autres fausseront toute étude. De plus, pour des études plus poussées, notamment la prédiction des intensités, supprimer ces valeurs aberrantes ou les recoder reviendrait à sous-estimer l'impact des événements futurs…Nous décidons de poursuivre notre étude avec ces valeurs.

En faisant un tableau croisé dynamique, on obtient la moyenne des intensités européenne et américaines par années (FIG.6).

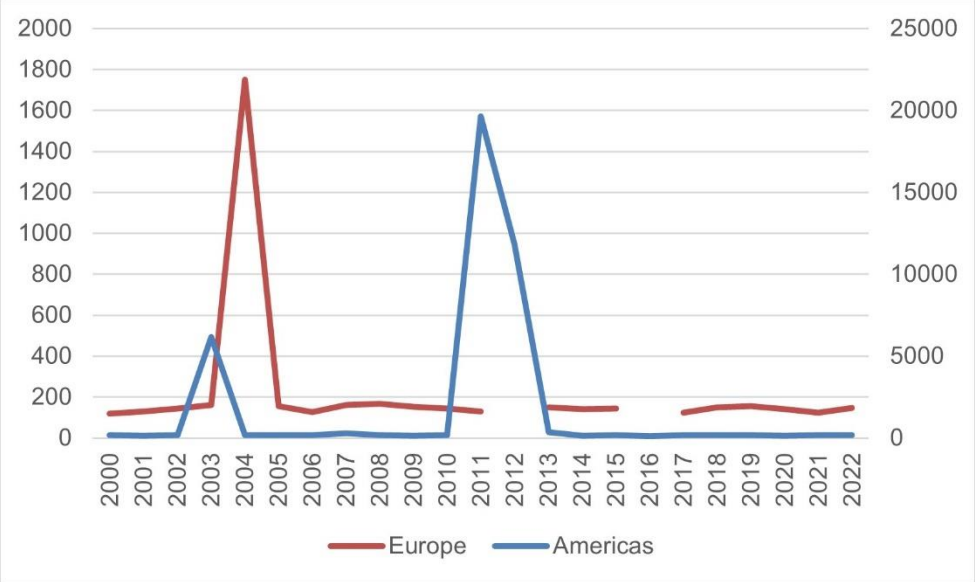

FIG.6- *Intesité moyenne des ouragans entre l'Europe et l'Amerique de 2000 à 2022.*

Bien qu'on ait vu dans les points précédents que les ouragans étaient plus fréquents en Amérique qu'en Europe, on remarque que l'Europe subit des ouragans à forte intensités, et ce, entre les années 2003 et 2005. L'Amérique elle, présente une croissance périodique ; en effet, les années 2002 et 2004 avaient des vitesses à 400 km/h tandis que 5 ans après, c'est le « boom ». On observe des vitesses dépassants les 1600 km/h ! Bien entendu, ces intensités irrégulières ont pour cause les valeurs aberrantes non traitées. Cela nous montre une fois de plus l'irrégularité de cette catastrophe climatique. Ainsi donc, l'Europe risque de subir dans l'avenir (aussi bien que l'Amérique) des catastrophes à hyper intensités…

- **Lien entre intensités, décès et coûts induits**

Nous avons croisé notre tableau en fonction de la moyenne des vents, et du total des décès et des couts induits enregistrés par année. Rappelons que les valeurs aberrantes n'ont pas été supprimées. Nous exportons notre tableau et décidons d'étudier une possible relation entre les intensités, les décès et les couts induits. La figure 7 ci-dessous représente un nuage de point sous R, à l'aide de la fonction *plot()*,en fonction des trois variables à étudier. On remarque que les points sont presque tous confondus en raison des valeurs aberrantes qui augmentent l'échelle de notre figure. Nous décidons de faire une régression linéaire multiple pour observer la nature de la relation entre les variables.

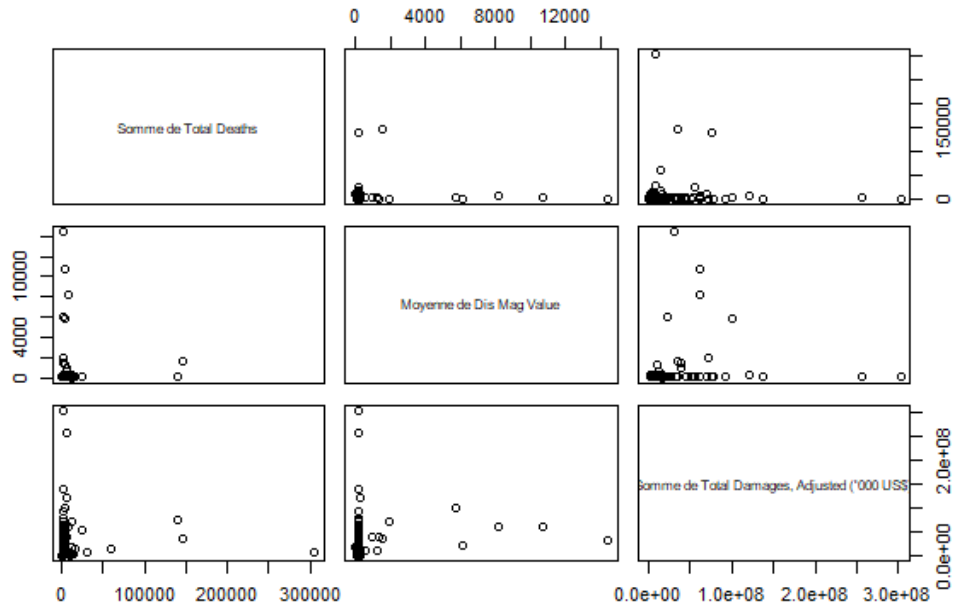

FIG.7 – *Nuage de points entre les intensités, les décès enregistrés et les coûts induits.*

Avec la fonction *lm()* sur RStudio, on effectue une régression multiple, qui consiste à voir l'impact de l'intensité des vents sur le nombre de mort et les coûts induits. La figure 7 montre le résultat de la régression faite. On observe que l'intercept, $\beta_0 = 40$ 896 235 qui est la somme des totaux des décès et couts induits lorsque la moyenne des intensités est nulle n'est pas pertinent ici, puisque l'intensité des vents n'a jamais été nulle. β1=1001 est l'estimation du coefficient de la moyenne de *Dis Mag Value*, avec une p-value (p-value = 0.721). La moyenne de l'intensité des catastrophes n'a donc pas d'impact significatif sur le total de la somme des décès et des couts induits. Le coefficient de détermination $R^2 \approx 0.24$ indique que le modèle explique de 24% la variance du total des décès et des couts induits. Enfin, le F-statistic, qui est le test de Fisher, qui en fait teste si au moins un des prédicteurs est significativement différent de zéro. F-statistic < p-value, le modèle dans son ensemble n'est pas significatif.

La régression linéaire ne nous permet donc pas de conclure sur le lien entre l'intensité, les couts induits et le total de décès. Nous calculons néanmoins la corrélation entre ces variables pour confirmer l'hypothèse de non-linéarité. La figure 9 nous montre les différentes corrélations entre les variables, corrélations calculées avec RStudio.

```
Call:
lm(formula = (couts_induits$`Somme de Total Deaths` + couts_induits$`Somme de Total Damages, Adjusted ('000 US$)`) ~
    couts_induits$`Moyenne de Dis Mag Value`, data = donnees_sans_na)

Residuals:
      Min        1Q    Median        3Q       Max
-39697443 -28957584 -21553558  11029125 261292039

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       40896235    8101305   5.048  5.4e-06 ***
couts_induits$`Moyenne de Dis Mag Value`     1001       2793   0.358    0.721
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56210000 on 54 degrees of freedom
  (17 observations effacées parce que manquantes)
Multiple R-squared:  0.002372,  Adjusted R-squared:  -0.0161
F-statistic: 0.1284 on 1 and 54 DF,  p-value: 0.7215
```

FIG.8- Résultat *régression multiple entre l'intensité, les couts induits et les décès.*

```
                                          Somme de Total Deaths
Somme de Total Deaths                                1.00000000
Moyenne de Dis Mag Value                            -0.03396416
Somme de Total Damages, Adjusted ('000 US$)          0.03385626
                                          Moyenne de Dis Mag Value
Somme de Total Deaths                                -0.03396416
Moyenne de Dis Mag Value                              1.00000000
Somme de Total Damages, Adjusted ('000 US$)           0.04872266
                                          Somme de Total Damages, Adjusted ('000 US$)
Somme de Total Deaths                                                       0.03385626
Moyenne de Dis Mag Value                                                    0.04872266
Somme de Total Damages, Adjusted ('000 US$)                                 1.00000000
```

FIG.9 *Corrélations entre les différentes variables*.

Dans l'ensemble, les corrélations de la figure 9 suggèrent qu'il y a peu d'association linéaire entre ces variables.

Effectuons le test du Chi-deux pour évaluer l'indépendance des variables. Toujours avec RStudio, on utilise la fonction *chisq.test()* de notre tableau sans prendre en compte les valeurs manquantes. La statistique du Chi-deux (X-squared) est calculée à environ 4 787 386, les degrés de libertés pour ces tests sont 110 et la p-value très petite (2.2exp(-16)) suggèrent que la distribution des données dans le tableau n'est pas aléatoire. Cela indique qu'il y a une relation significative entre les variables dans notre tableau.

Il y a en effet un lien statistique entre les coûts induits, les décès et la vitesse des vents. Ce lien n'est pas linéaire.

```
>
> # Afficher les résultats du test
> print(resultats_chi2)

        Pearson's Chi-squared test

data:  couts_induits_sans_na
X-squared = 4787386, df = 110, p-value < 2.2e-16
```

FIG.10 – *Résultat du test du Chi-deux sur le tableau croisé dynamique contenant les coûts induits, les décès et la vitesse des vents (avec RStudio).*

- **Liens entre plusieurs variables**

Pour calculer la durée de chaque événement, nous avons d'abord ajouté deux colonnes *Date début* et *Date fin*. A l'aide de la fonction *DATE* sur Excel, nous avons regroupé les jours de début/fin, les mois de début/fin et l'année de début/fin de chaque catastrophe. Ensuite, par l'ajout d'une colonne supplémentaire *Durée,* on effectue la différence entre *Date fin* et *Date début*.

Start Year	Start Month	Start Day	End Year	End Month	End Day	Date Début	Date fin	Durée
2005	8	29	2005	9	19	29/08/2005	19/09/2005	22
2021	8	28	2021	9	2	28/08/2021	02/09/2021	6
2017	9	19	2017	9	21	19/09/2017	21/09/2017	3
2017	8	25	2017	8	29	25/08/2017	29/08/2017	5
2017	9	10	2017	9	28	10/09/2017	28/09/2017	19
2021	2	10	2021	2	20	10/02/2021	20/02/2021	11
2004	9	15	2004	9	16	15/09/2004	16/09/2004	2
2005	10	24	2005	10	24	24/10/2005	24/10/2005	1
2018	10	10	2018	10	11	10/10/2018	11/10/2018	2
2020	8	27	2020	8	28	27/08/2020	28/08/2020	2
2011	4	22	2011	4	29	22/04/2011	29/04/2011	8
2004	8	13	2004	8	13	13/08/2004	13/08/2004	1
2011	5	20	2011	5	25	20/05/2011	25/05/2011	6
2004	9	5	2004	9	5	05/09/2004	05/09/2004	1
2016	10	8	2016	10	12	08/10/2016	12/10/2016	5
2018	9	12	2018	9	18	12/09/2018	18/09/2018	7
2020	8	8	2020	8	12	08/08/2020	12/08/2020	5
2016	4	10	2016	4	15	10/04/2016	15/04/2016	6
2014	5	18	2014	5	23	18/05/2014	23/05/2014	6

TAB.5 – *Extrait du nouveau tableau contenant les nouvelles variables Durée, Date début, Date fin.*

Pour connaitre s'il existe un lien entre durée (en jours) et le maximum des intensité/couts induits/ total de décès, on effectue des histogrammes.

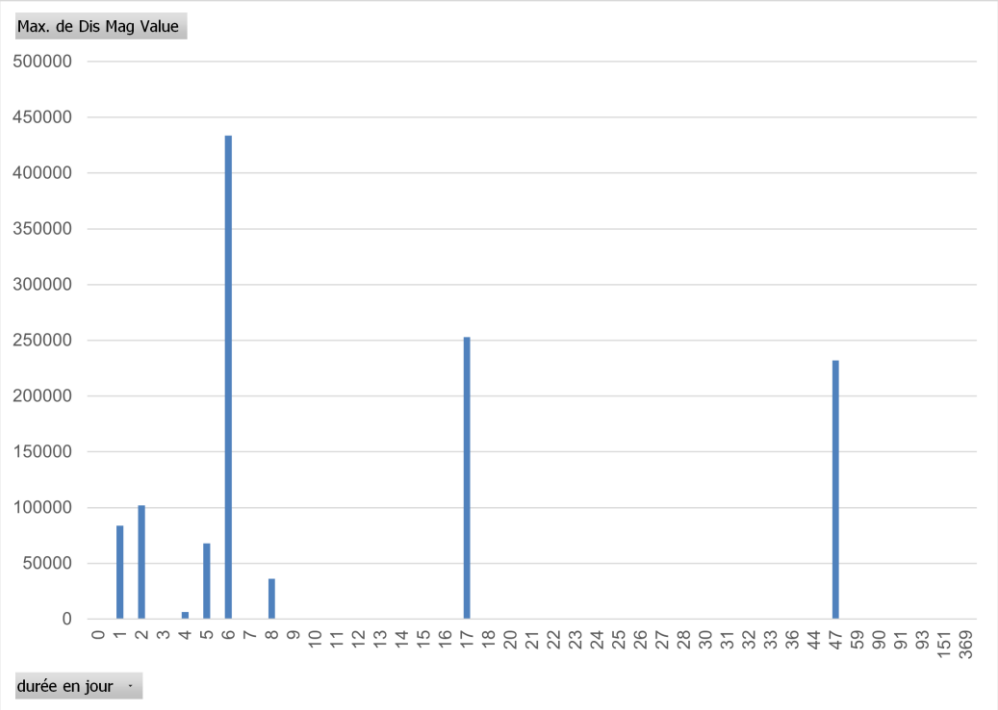

FIG.11 – *Histogramme du maximum des intensités par durée.*

On voit qu'en moyenne, les ouragans gagnent en intensité pendant les 20 premiers après leur apparition. Après le 20ème jour, l'intensité maximale régresse. Ces intensités maximales sont élevées par la présence des valeurs aberrantes. Face aux changements climatiques constants, il ne serait pas impossible qu'après le 20ème jour, l'intensité augmente davantage. Ces valeurs aberrantes nous mettent en garde d'une hausse éventuelle des intensités. Notre étude est aussi limitée par les observations manquantes.

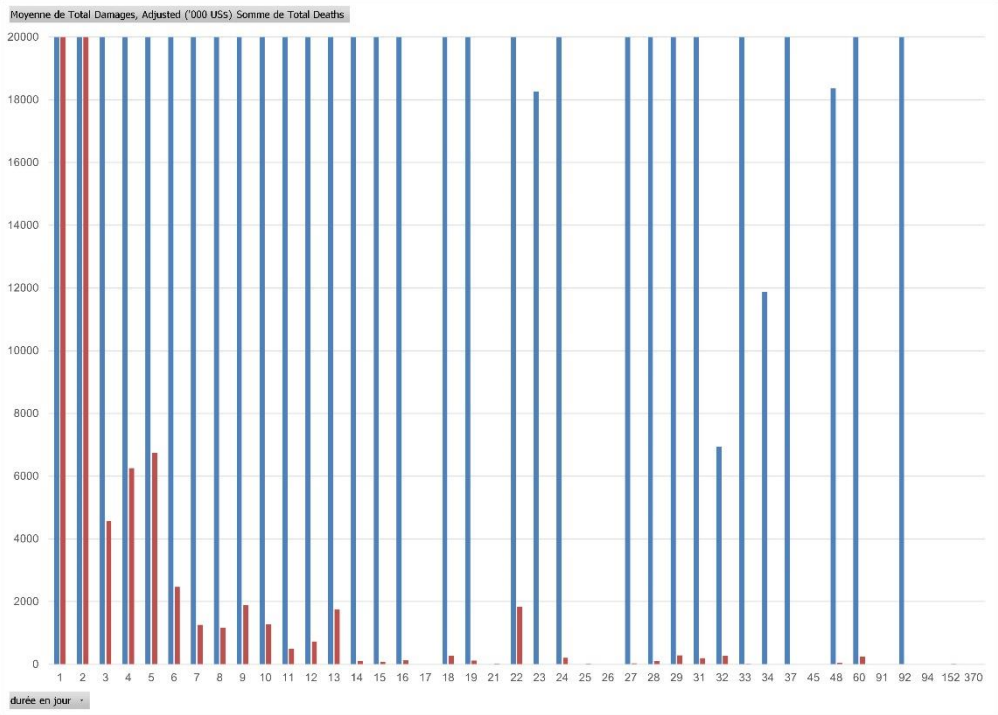

FIG.12 – *Histogramme du nombre de décès et la moyenne des coûts induits par durée (en jours). En rouge les décès, en bleu la moyenne des coûts induits.*

De façon générale, la figure 11 nous montre que les coûts induits sont élevés toute la durée de vie des ouragans. De plus, on remarque que les évènements dont la durée de vie est inférieure à 5 jours est en supériorité numérique. Ces évènements justifient emmagasinent à eux seuls la majorité des décès. On s'aperçoit, grâce aux figures 11 et 12 que l'intensité des vents et les coûts induits sont liés.

5. Conclusion

Dans le dernier point abordé, nous avons vu que les ouragans éphémères, c'est-à-dire ceux dont la durée ne dépasse pas une journée sont les plus nombreux à être observés. De plus, on observe que la vitesse des vents augmente dans le temps ainsi que les coûts induit. Quel est l'impact de cette vitesse sur les coûts induits par les ouragans, en mettant particulièrement en lumière l'influence des ouragans éphémères (de moins de 5 jours), et comment ces informations peuvent contribuer à une meilleure anticipation et gestion des conséquences économiques des phénomènes cycloniques ?

6. Références

Begin'R (u-bordeaux.fr)

Pierre-André, CORNILLON. Al. (2018). R pour la statistique et la science des données. Collection « Pratique de la statistique ». Editions Pur.

Liste des noms retirés d'ouragans — Wikipédia (wikipedia.org)

Bilan statistique des catastrophes naturelles en France et dans le monde en 2021 (catnat.net)

Historique des cyclones en Atlantique depuis 1950 - Météo Tropicale (meteo-tropicale.fr)