Table des matières

[**Introduction 2**](#_heading=h.30j0zll)

[**I-**](#_heading=h.1fob9te) **Analyse de Données Géologiques 3**

[1.](#_heading=h.3znysh7) Les phénomènes sismiques 3

[2.](#_heading=h.2et92p0) Les méthodes de mesure d’amplitudes 4

[3.](#_heading=h.tyjcwt) Pré-traitement des données 6

[**II-**](#_heading=h.3dy6vkm) **Analyse descriptive 8**

[1.](#_heading=h.1t3h5sf) Magnitude moyenne et médiane sur différentes périodes 8

[2.](#_heading=h.4d34og8) Cartographie des séismes 9

[3.](#_heading=h.2s8eyo1) Répartition des séismes et leur intensités 10

[**III-**](#_heading=h.17dp8vu) **Choix du sujet 11**

[1.](#_heading=h.3rdcrjn) Les essais 11

[2.](#_heading=h.26in1rg) Choix retenu 11

[**IV-**](#_heading=h.lnxbz9) **Modélisation 12**

[1.](#_heading=h.35nkun2) ARIMA 12

[2.](#_heading=h.1ksv4uv) GARCH 15

[3.](#_heading=h.44sinio) Light GBM 16

[4.](#_heading=h.2jxsxqh) LSTM 21

[**Conclusion 25**](#_heading=h.z337ya)

[**Bibliographie 26**](#_heading=h.3j2qqm3)

[**Annexes 27**](#_heading=h.1y810tw)

# Introduction

L'analyse géologique, englobant l'activité sismique et les éruptions volcaniques, revêt une importance cruciale pour la compréhension des processus géologiques de la Terre et la prédiction d'événements futurs. À une époque caractérisée par des avancées technologiques majeures et une sensibilisation mondiale accrue, l'exploration approfondie de ces données émerge comme une pierre angulaire dans la préparation et l'anticipation des catastrophes. Cette démarche se plonge dans les méthodologies utilisées pour scruter les données géologiques en séries temporelles, visant à comprendre les tendances historiques tout en anticipant d'éventuels événements sismiques. La véritable importance de cette entreprise réside dans son potentiel à sauver des vies et à protéger les biens en permettant des actions rapides et informées.

Les données géologiques utilisées dans cette analyse proviennent exclusivement du Service Géologique des États-Unis (USGS), garantissant une base complète et transparente pour notre étude. Le périmètre de notre investigation se concentre sur les tremblements de terre mondiaux survenus depuis 1990.

Les sections suivantes de ce rapport détaillent notre approche méthodologique, s'étendant du prétraitement et de l'analyse exploratoire des données à la sélection du modèle et à la prédiction d'événements. Cette démarche vise à capitaliser sur les avancées de l'analyse des séries temporelles géologiques, renforçant ainsi notre capacité à anticiper et atténuer les impacts des phénomènes géologiques majeurs.

Une composante essentielle de notre démarche réside dans la diversité de choix pour la sélection des variables à modéliser. Ce rapport décrira de manière exhaustive nos différentes tentatives de modélisation, mettant en lumière notre sujet principal : la modélisation du nombre de séismes par mois en Alaska. Cette étape, cruciale dans notre analyse, représente un effort conséquent pour comprendre et prévoir les tendances sismiques dans cette région spécifique.

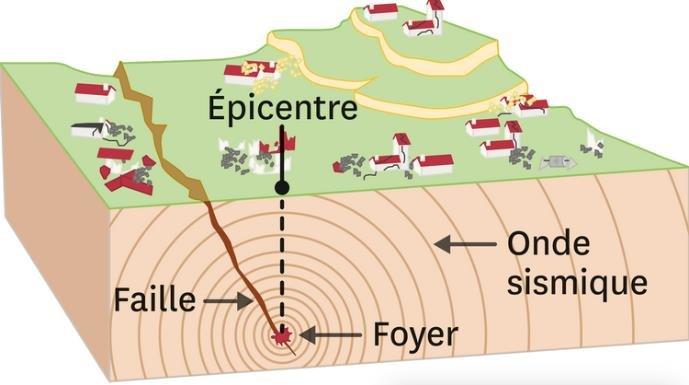
L'objectif ultime de notre démarche est de contribuer au développement de systèmes d'alerte précoce robustes, capables d'avertir les autorités dès la détection de potentiels événements géologiques fréquents. En mettant en avant une approche méthodique et des modèles de prédiction solides, nous aspirons à jouer un rôle significatif dans la sauvegarde de vies et la protection des biens grâce à des interventions promptes et éclairées.

# Analyse de Données Géologiques

## Les phénomènes sismiques

Un séisme, défini comme une secousse du sol, résulte de la libération soudaine d'énergie accumulée par les contraintes exercées sur les roches, se produisant généralement le long d'une faille préexistante. Moins fréquemment, des séismes peuvent être dus à l'activité volcanique ou à des causes artificielles telles que des explosions. Le lieu de la rupture des roches en profondeur est appelé le foyer, et son point de projection à la surface est l'épicentre du séisme. Les mouvements des roches près du foyer génèrent des vibrations élastiques se propageant sous forme de paquets d'ondes sismiques autour du globe terrestre, accompagnées parfois d'un dégagement de chaleur suffisant pour fondre les roches le long de la faille (pseudotachylites).

La prédominance des séismes est observée principalement aux confins des plaques tectoniques, donnant naissance aux séismes inter-plaques. Cependant, des événements sismiques peuvent également se produire à l'intérieur des plaques, caractérisés comme des séismes intra-plaques. Cette répartition sismique à la surface de la Terre trouve son explication dans le concept de tectonique des plaques. Les principales ceintures sismiques mondiales, déterminées par la concentration géographique des activités sismiques, incluent la ceinture de feu du Pacifique (libérant 80 % de l'énergie sismique annuelle), la ceinture alpine (contribuant à hauteur de 15 %) et les dorsales océaniques (responsables de 5 %).



*Figure 1 : Schéma d’un séisme inter-plaques*

Le schéma ci-dessus met en évidence la dynamique des séismes inter-plaques, clarifiant des notions clés telles que le foyer, l'épicentre et les ondes sismiques, fournissant ainsi une compréhension visuelle des mécanismes sismiques fondamentaux. Les séismes intra-plaques, moins fréquents mais potentiellement dévastateurs, se produisent à l'intérieur d'une plaque tectonique, présentant une dynamique distincte par rapport aux séismes inter-plaques résultant de l'interaction de deux plaques.

Situés principalement dans la partie supérieure de la croûte terrestre, les séismes intra-plaques peuvent être classés en deux catégories : le champ de contraintes régional et le champ de contrainte local, chacun capable de déclencher un séisme, avec des mécanismes explicatifs encore en débat. Il est essentiel de noter que les séismes intra-plaques ne se produisent pas de manière isolée ; des interactions avec des séismes inter-plaques sont possibles, comme illustré par le séisme de 2009 aux Samoa, où un séisme intra-plaque a déclenché deux séismes de subduction. Cette interconnexion souligne la complexité des phénomènes sismiques.

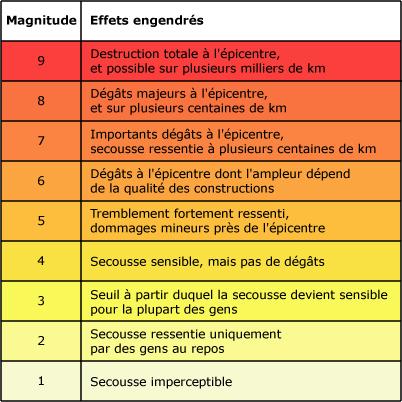
## Les méthodes de mesure d’amplitudes

Les méthodes de mesure des séismes, examinées dans la suite, sont issues des références [[1]](#_heading=h.32hioqz) et [[2]](#_heading=h.32hioqz). Chacune de ces approches a pour objectif de quantifier la puissance du phénomène sismique dans les trois cas présentés.

### 2.1 Magnitude locale

La première évaluation de la magnitude a été introduite en 1935 par Charles Francis Richter pour classer les sismogrammes locaux enregistrés en Californie. Initialement conçue pour mesurer l'amplitude en micromètres sur un sismographe de type Wood-Anderson pour un séisme situé à 100 km, cette mesure est maintenant appelée magnitude locale.

La magnitude de Richter est déterminée en mesurant l'amplitude maximale des ondes sismiques enregistrées par des sismographes. L'amplitude est ensuite ajustée en fonction de la distance entre la source du séisme (l'épicentre) et la station sismique. L'idée fondamentale est que plus un séisme est énergétique, plus les ondes sismiques enregistrées seront importantes.



*Figure 2 : Échelle de Richter*

La formule originale de Richter pour la magnitude locale (ML) est une échelle logarithmique simple :

* A représente l'amplitude maximale
* A0 est une amplitude de référence pour un séisme de magnitude 0 à 100 km.
* est la distance épicentrale
* est une constante d’étalonnage

Les constantes d'étalonnage rendent cette définition valide localement, soulignant son caractère empirique. Par exemple, dans la définition originale pour des séismes modérés en Californie du Sud, enregistrés avec un sismographe de type Wood-Anderson, c 2.76 et log() 2.48.

La magnitude de Richter est déterminée en mesurant l'amplitude maximale des ondes sismiques enregistrées par des sismographes. L'amplitude est ensuite ajustée en fonction de la distance entre la source du séisme (l'épicentre) et la station sismique. L'idée fondamentale est que plus un séisme est énergétique, plus les ondes sismiques enregistrées seront importantes.

### 2.2 Magnitudes dites d’ondes et

L'échelle de Richter, une mesure locale introduite en 1936, a conduit à l'émergence d'une nouvelle magnitude appelée (magnitude des ondes de surface). Proposée par Beno Gutenberg et Charles Richter, cette magnitude se base sur l'amplitude des ondes de surface, en particulier l'onde de Rayleigh sur la composante verticale du sismogramme, pour des distances télésismiques (au-delà de 30°) et une période de 20 secondes (période naturelle des sismographes). La formulation de cette magnitude est similaire à celle de la magnitude locale ().

* est l’amplitude mesurée
* est la distance épicentrale en degrés
* et sont des constantes d’étalonnage

Malgré son caractère empirique et les problèmes de saturation, la mesure (magnitude des ondes de surface) est toujours utilisée aujourd'hui. Cependant, elle présente des limitations pour les séismes profonds (profondeur supérieure à 100 km) et pour l'estimation rapide de la magnitude dans le cadre d'un réseau d'alerte.

Une alternative, la magnitude des ondes de volume notée ( pour body waves), a été introduite en 1956. Elle se base sur le premier train d'onde P, offrant une estimation rapide de l'importance du séisme. Sa formulation dépend de la période dominante T du signal :

* est l’amplitude maximal mesurée
* est la distance épicentrale en degrés
* est la profondeur hypocentrale
* Q est une fonction d’étalonnage dépendant de ces paramètres

La magnitude des ondes de volume (), introduite en 1956, souffre de problèmes de saturation rapide, limitant sa fiabilité, notamment pour les séismes profonds. Parallèlement, d'autres mesures, telles que la magnitude de durée, qui repose sur la mesure du temps d'enregistrement du signal sur le sismogramme, sont utilisées à l'échelle locale ou régionale. La variabilité de ces mesures, influencée par divers facteurs comme le type d'onde, le capteur, la distance et le type de magnitude, complique l'établissement de relations précises entre elles.

### 2.3 Magnitudes des moments

En 1979, les chercheurs Thomas Hanks et Hiroo Kanamori du Caltech en Californie ont introduit une méthode novatrice pour calculer la magnitude des séismes, désignée sous le nom de ou magnitude de moment. Cette approche repose sur un modèle physique de la rupture sismique, prenant en considération la déformation élastique associée à un double-couple de forces de directions opposées et perpendiculaires. Le moment sismique (), exprimé en Newton.mètres (N.m), quantifie l'énergie sismique liée au déplacement sur la faille. La magnitude est calculée à partir du moment sismique selon une formule spécifique suivante :

* est exprimé en N.m. Par exemple, un séisme de magnitude correspond à un moment sismique N.m.

Le moment sismique est défini comme :

* est le nombre de cisaillement, varie entre 30 GPa et 300 GPa dans la Terre
* S est la surface de la faille, calculée à partir de la longueur de la faille L, du pendage de la faille p et de la profondeur de la faille z.
* est le déplacement moyen sur la faille

Cette approche offre une meilleure évaluation de la magnitude des séismes importants en prenant en compte la répartition temporelle de l'énergie libérée. Cependant, sa contrainte inhérente réside dans l'impératif de la présence d'une faille. Par conséquent, son efficacité maximale est atteinte uniquement dans le contexte de phénomènes sismiques d'une magnitude significative.

## Pré-traitement des données

### 3.1 Collecte des données

Comme évoqué précédemment, les données utilisées dans cette analyse ont été extraites du site de l'USGS, l'organisme chargé de la surveillance mondiale de la Terre, qui propose gratuitement une base de données exhaustive. Cette base compile de manière complète les informations provenant des laboratoires de surveillance sismique du monde entier, les centralisant pour faciliter leur accès et leur utilisation. Pour traiter le volume considérable de ces données, le jeu de données examiné dans la suite du rapport résulte de requêtes multiples et combinées.

Il est crucial de souligner que l'ensemble des données n'a fait l'objet d'aucun filtrage préalable. Tous les séismes survenus à l'échelle mondiale entre 1990 et aujourd'hui sont inclus, sans exclusion préalable. La section suivante propose une vue détaillée des variables clés de ces données sismiques, fournissant une description complète de chaque colonne, de son format, et de sa signification. Le Tableau 1 résume de manière structurée ces variables, offrant un aperçu indispensable pour l'analyse ultérieure.

| **Champ** | **Format** | **Description** |
| --- | --- | --- |
| time | Long Integer | Temps de l'événement en millisecondes depuis l'époque (1970-01-01T00:00:00.000Z), sans inclure les secondes intercalaires. Dans certains formats de sortie, la date est formatée pour la lisibilité. |
| place | String | Description textuelle de la région géographique nommée près de l'événement. Il peut s'agir du nom d'une ville ou d'une région de la classification de Flinn-Engdahl. |
| status | String | Indique si l'événement a été examiné par un être humain. |
| tsunami | Integer | Il s'agit d'une série de grandes vagues océaniques généralement causées par une perturbation sous-marine, souvent associée à des tremblements de terre. |
| significance | Integer | Indique l'importance ou le niveau d'impact de l'événement, qui peut être utilisé pour évaluer les conséquences potentielles. |
| data type | String | Type d'événement sismique. |
| magnitudo | Decimal | Magnitude de l'événement. |
| state | String | Représente la division administrative ou l'État où l'événement s'est produit, souvent applicable à des pays spécifiques. |
| latitude / longitude | Decimal | Degrés décimaux de latitude. Valeurs négatives pour les latitudes sud, et degrés décimaux de longitude. Valeurs négatives pour les longitudes ouest. |
| depth | Decimal | Profondeur de l'événement en kilomètres. |
| date | String | Date et heure de l'événement. |

*Tableau 1 : Description des champs des données sur les séismes*

### 3.2 Nettoyage et manipulation des données

Dans la phase initiale de notre analyse, nous avons entrepris une vérification minutieuse de l'intégrité des données liées aux variables spécifiques que nous avons sélectionnées pour notre étude. Heureusement, cette première étape a confirmé que l'ensemble des données était complet, ne présentant aucune lacune. Cette constatation renforce la fiabilité et la qualité intrinsèque des informations que nous avons recueillies du Service Géologique des États-Unis (USGS).

Dans un souci d'assurance qualité supplémentaire, une étape subséquente de notre processus a été consacrée à l'examen de la présence potentielle de doublons au sein de notre ensemble de données. En raison de la complexité des requêtes impliquées dans notre collecte d'informations géologiques, nous avons identifié un total de 16 869 lignes en double parmi les 3 445 751 entrées. Pour garantir la fiabilité et la précision de nos analyses ultérieures, nous avons pris des mesures proactives en effectuant un processus de dédoublonnage pendant la phase de prétraitement des données. Cette démarche vise à garantir la cohérence et l'intégrité des informations que nous utilisons pour notre analyse géologique approfondie.

Pour renforcer la qualité de notre processus, nous avons effectué des vérifications supplémentaires, incluant l'examen des valeurs de profondeur négatives, qui sont incompatibles avec les mesures en kilomètres. De plus, des anomalies telles que des valeurs de magnitude de -5 et -9,99 ont été identifiées et éliminées, car elles ne sont pas physiquement possibles dans le contexte sismique. Ces filtres supplémentaires ont été appliqués avec rigueur lors de la phase de prétraitement des données, contribuant ainsi à garantir l'intégrité et la validité des informations utilisées dans notre analyse géologique approfondie.

Initialement, la variable « time » était exprimée en timestamp, et la décision initiale était de laisser les données telles quelles, le format en millisecondes étant couramment utilisé pour les séries temporelles. Cependant, après des tentatives répétées de modélisation, une conversion a été réalisée. Cela a impliqué le passage du format entier (timestamp en millisecondes) à un format de date. Cette modification a été motivée par la décision ultérieure d'agrégation des données par mois, entraînant une adaptation des objectifs et de la cible du projet pour refléter cette nouvelle approche.

# Analyse descriptive

## Magnitude moyenne et médiane sur différentes périodes

En employant le package ggplot2, la création d'une représentation graphique claire des magnitudes annuelles, intégrant à la fois la moyenne et la médiane, est réalisable. Cette visualisation cherche à déceler une tendance générale, évaluant ainsi une éventuelle corrélation entre ce phénomène naturel et les changements climatiques. Néanmoins, une analyse préliminaire suggère qu'aucune relation significative n'est apparente.

L'analyse temporelle des magnitudes des tremblements de terre, présentée dans le rapport en annexe (**Magnitude moyenne et médiane sur différentes périodes**), couvre la période depuis 1990, offrant ainsi une perspective étendue sur l'évolution de ces phénomènes naturels. Un aspect notable réside dans l'absence apparente de saisonnalité, mais plutôt dans une volatilité significative au sein de la série chronologique. De manière particulière, on observe que la magnitude moyenne annuelle des tremblements de terre était plus importante avant 2007, tandis qu'elle tend à être moins élevée en moyenne et en médiane après cette année. Par conséquent, à première vue, aucune conclusion définitive ne peut être tirée quant à un éventuel lien avec le réchauffement climatique.

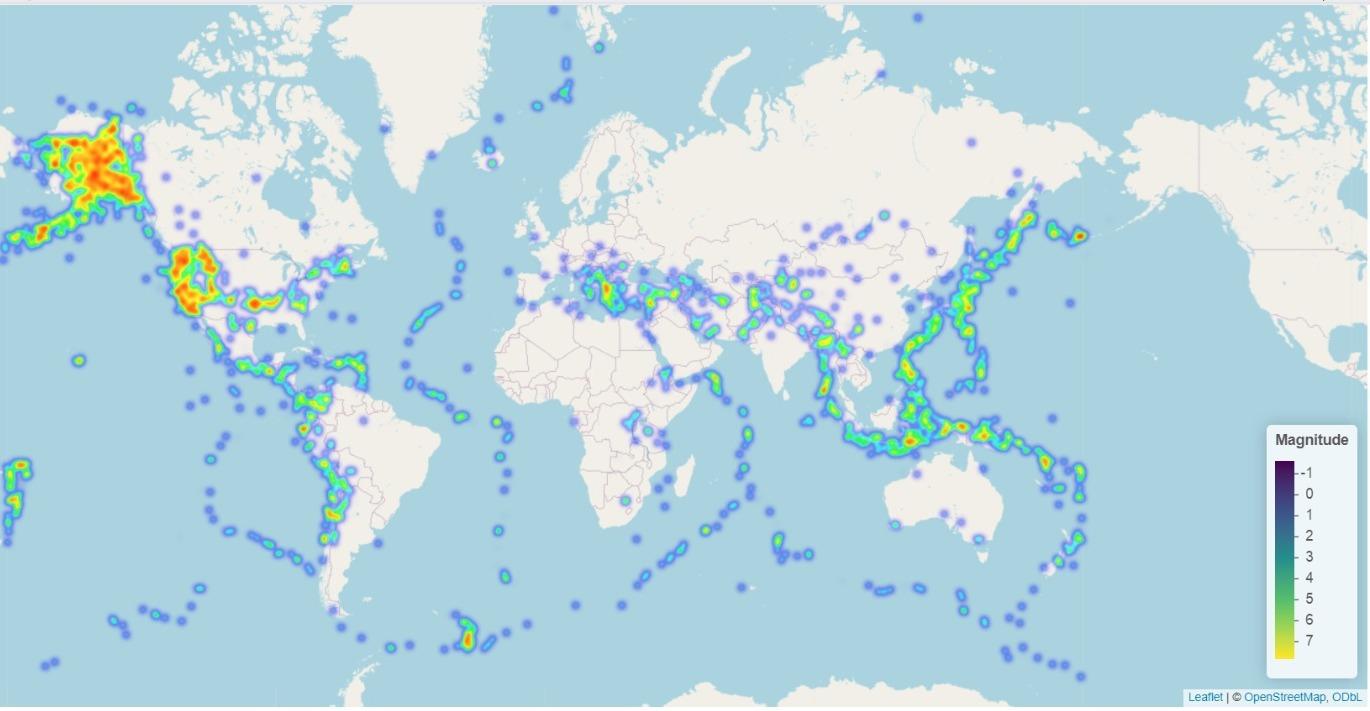
Les boîtes à moustaches (**Boxplot des magnitudes**) fournissent des informations cruciales sur la magnitude moyenne annuelle tout en mettant en évidence la présence de nombreux tremblements de terre très puissants, largement ressentis par les êtres humains. Cependant, dégager une évolution temporelle significative demeure complexe.

Dans cette analyse, l'absence de saisonnalité est confirmée. À la place, une tendance relativement stable se manifeste, conformément aux attentes. Il est important de noter que l'année 2023 ne doit pas être prise en considération, car le dataframe s'arrête en juillet.

Contrairement à la magnitude, une tendance claire émerge en termes de fréquence des tremblements de terre, montrant une augmentation au fil des années. En résumé, bien qu'il n'y ait pas d'augmentation moyenne de la magnitude, on observe une fréquence croissante d'apparition de séismes, indiquant une augmentation du nombre de séismes, principalement de faible intensité.

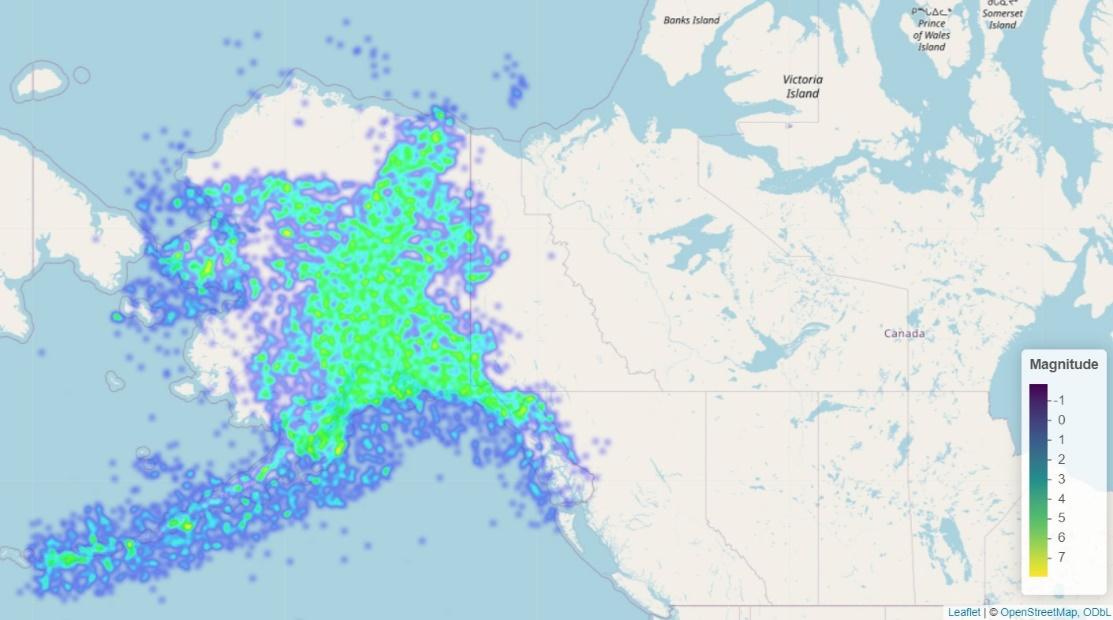
## Cartographie des séismes

Cette section propose une brève analyse des régions fortement touchées par les séismes, mettant en évidence deux pays qui se démarquent : la Californie et l'Alaska. Notre attention se portera particulièrement sur l'Alaska dans le cadre de notre modélisation. Ces emplacements présentent des caractéristiques spécifiques, avec des variables représentées de manière précise. Afin d'obtenir une vision plus générale, il est judicieux de regrouper ces données par pays. La Californie et l'Alaska occupent clairement les premières positions, tandis que le Nevada et Hawaï se classent respectivement en 3ème et 4ème place, bien que leur nombre de séismes soit significativement inférieur en comparaison.



*Figure 3 : Carte des séismes dans le monde*

Le planisphère des séismes dans le monde depuis 1990 constitue une représentation visuelle puissante, offrant des indications claires sur la dynamique des plaques tectoniques à l'échelle mondiale. Grâce à cette visualisation, il est possible de discerner la répartition géographique des failles, fournissant ainsi un aperçu précis des zones sismiques majeures. La concentration notable de tremblements de terre dans les zones montagneuses des États-Unis et, plus particulièrement, en Alaska, souligne l'importance de ces régions sur le plan géologique.



*Figure 4 : Carte des séismes centré sur l’Alaska*

Dans le cadre spécifique de notre examen centré sur l'Alaska, l'inclusion d'une carte détaillée des séismes dans cette région serait une ressource essentielle. Cette carte permettrait de visualiser plus spécifiquement les zones affectées par les séismes en Alaska, mettant en évidence la répartition spatiale de ces événements géologiques. Une observation significative serait la corrélation entre les trajectoires des terres du sud-ouest et l'activité sismique, une relation conforme aux principes de la géologie classique, qui explique que la Micronésie a émergé par soulèvement du terrain. En intégrant une telle carte à notre analyse, nous pourrions mieux comprendre les schémas sismiques locaux et leurs implications géologiques spécifiques.

## Répartition des séismes et leur intensités

Cette section se consacre à la présentation de la distribution de l'intensité des séismes, un aspect crucial étant donné les impacts potentiels considérables de ces phénomènes. Les séismes peuvent entraîner des pertes de vie importantes et générer des déplacements de terrain spectaculaires, atteignant parfois plusieurs mètres. Un exemple emblématique est le séisme de magnitude 9 qui a frappé Tohoku en 2011, provoquant d'importants déplacements horizontaux jusqu'à 50 mètres et un déplacement vertical d'environ 5 mètres, entraînant des conséquences dévastatrices.

À la suite de cet épisode inquiétant, nous cherchons à apporter un certain réconfort en analysant la répartition de ces phénomènes. Une question essentielle est de savoir s'ils se manifestent fréquemment avec une telle ampleur. Pour répondre à cette interrogation, nous avons choisi de représenter graphiquement la distribution des séismes en fonction de leur importance, en utilisant un histogramme basé sur la variable « significance », qui tient compte des conséquences associées. Cette approche graphique nous permettra d'appréhender plus clairement la fréquence des séismes de grande ampleur et d'évaluer leur impact potentiel sur une échelle significative.

Dans cette observation (**Histogramme significativité**), il est à noter que les phénomènes sismiques extrêmes demeurent relativement rares. Il est essentiel de comprendre que ce ne sont pas nécessairement les séismes les plus puissants qui sont qualifiés de plus dévastateurs, même s'il existe une corrélation significative entre puissance et impact (Pearson : 0.9473). Lorsque nous utilisons le terme « dévastateur », nous faisons référence à l'impact sur le bilan humain et les infrastructures. À titre d'exemple, le séisme le plus dévastateur enregistré a eu lieu au Mexique en 2017, près des côtes de Chiapas, avec une magnitude de 8.2.

Une dernière donnée notable concerne la variable « tsunami » (**Tsunami**). Malgré son caractère surprenant, il est important de noter que les tsunamis ne représentent qu'une infime fraction, soit 0.04%, de notre ensemble de données. Cette rareté des tsunamis dans notre jeu de données souligne la spécificité de ces événements par rapport à l'ensemble des séismes enregistrés, soulignant leur caractère exceptionnel au sein de la base de données géologiques considérée.

# Choix du sujet

## Les essais

Nous avons d'abord envisagé de modéliser la magnitude et la localisation potentielle des événements sismiques avec des séries temporelles. Bien que cette approche se révèle impraticable, elle méritait d'être examinée. Notre expérimentation s'est limitée aux modèles de type ARIMA, mais des perspectives intéressantes pourraient émerger avec les modèles VARMA. Ces derniers prennent en compte les interactions entre différentes séries temporelles, offrant ainsi une approche plus complète pour comprendre et prédire les phénomènes géologiques complexes, notamment la magnitude, la profondeur et la significativité des séismes.

À la suite de ces essais, nous avons entrepris du feature engineering. Initialement, dans le but de prédire la magnitude à partir de nos données, nous avons créé plusieurs variables supplémentaires. Celles-ci comprennent des aspects temporels tels que l'année, le mois, le jour et l'heure, ainsi que la variable « time » représentant la différence entre le 1970-01-01 et la date du timestamp en millisecondes. De plus, certains pays et états ont été encodés en one-hot (Alaska et Californie), les autres ont été regroupés sous la catégorie « Other », et les points cardinaux ont été extraits de la variable « place » (si aucun point cardinal n'était présent, la catégorie « Other » était attribuée). Cependant, nous avons choisi de nous concentrer sur un sujet plus accessible, mais tout aussi crucial que la prédiction de la magnitude.

## Choix retenu

Nous avons choisi d'agréger les données par mois, se concentrant ainsi sur la prédiction du nombre de séismes mensuels. Bien que cette approche simplifie la modélisation, elle réduit évidemment la richesse des résultats en termes d'analyse et de système d'alerte. Il est crucial de noter que le problème reste actuellement insoluble. Éventuellement, après la conclusion de ce projet, des analyses sur des périodes plus courtes pourraient être envisagées. Cependant, il est impératif de trouver un compromis entre la disponibilité de données suffisantes pour assurer la signification des résultats et le maintien d'un intérêt commun.

Notre concentration spécifique sur l'Alaska repose sur deux arguments majeurs :

* Intérêt géologique : Un focus sur une zone spécifique s'avère plus pratique pour le développement d'un système d'alerte, en particulier en considérant les caractéristiques géologiques distinctives de cette région.
* Contrainte statistique : La sélection d'une zone avec une quantité de données suffisante est cruciale. Nous avons donc dû choisir entre l’Alaska et la Californie, et notre choix s’est porté sur l’Alaska de manière quelque peu arbitraire.

Enfin, nous avons appliqué un filtre sur les séismes perceptibles par l’homme (magnitude ≥ 3). Ce choix est sujet à débat, et nous laissons aux géologues et statisticiens le soin d’apporter leurs arguments. L’idée sous-jacente est de se concentrer sur les séismes considérés comme importants, bien que cette approche puisse être considérée comme naïve. Nous ne disposons pas de l’expertise métier nécessaire pour déterminer si un séisme, même de faible magnitude, peut être considéré comme un signal précurseur d’un séisme plus important.

# Modélisation

## ARIMA

### Rappels théoriques

La méthodologie de modélisation des séries temporelles ARIMA, inspirée en grande partie de la référence [[3]](#_heading=h.32hioqz), combine les composantes auto-régressives (AR), intégrées (I) et à moyenne mobile (MA). Notée généralement (p, d, q), cette méthode se caractérise par les ordres suivants :

* : Ordre de la composante auto-régressive (AR)
* : Ordre de la différenciation intégrée
* : Ordre de la composante à moyenne mobile (MA)

La formule générale du processus ARIMA est exprimée par l'équation :

où :

* est la valeur observé à l’instant *t*
* : Coefficient auto-régressifs
* : Coefficient de la composante à moyenne mobile
* : Terme d’erreur blanc à l’instant *t*

Pour la composante auto-régressive (AR), le concept est similaire à celui des séries numériques, visant à capturer les dépendances linéaires entre les observations successives. La partie à moyenne mobile (MA) concerne les termes d'erreur, reconnaissant que le terme d'erreur à l'instant *t* dépend des termes d'erreur aux instants *t*-1, *t*-2 et *t*-3 par exemple.

La composante intégrée (I pour « Integrated ») dans le modèle ARIMA est employée pour rendre stationnaire une série temporelle, condition cruciale pour l'application des modèles statistiques. La stationnarité permet de supposer que les propriétés statistiques de la série restent constantes au fil du temps.

Le processus d'intégration (I) consiste à prendre des différences successives entre les observations pour éliminer les tendances ou les structures non stationnaires. Habituellement, une première différence (*d* = 1) est suffisante, bien que dans certains cas, une différence supplémentaire puisse être nécessaire pour atteindre la stationnarité.

La saisonnalité ou la causalité ne seront pas exposées ici. Par définition, nos modèles seront causaux dans le sens où nous cherchons à prédire avec les informations passées, sans chercher à violer les lois de la physique. Par chance, notre série ne présente pas de saisonnalité marquée. En agrégeant par mois, aucune tendance saisonnière significative n'émerge. Introduire une saisonnalité aurait considérablement compliqué le modèle, notamment en impliquant l'utilisation de séries de Fourier.

### Partie pratique

D'abord, quelques remarques sur le traitement des données. La première étape consistait à agréger les données, et il était essentiel de choisir la bonne périodicité pour atteindre le juste équilibre entre disposer de suffisamment de données et de variables pertinentes pour la compréhension des phénomènes liés à la vie humaine. L'agrégation par mois semblait être une option judicieuse pour ce projet, bien que nous laissions aux experts le soin de juger et d'explorer d'autres périodicités qui pourraient être plus adaptées.

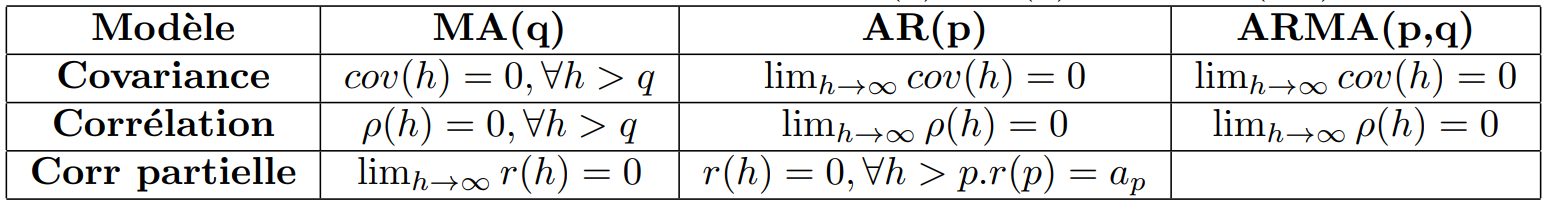
La subtilité inhérente aux séries temporelles, bien que cela puisse sembler évident pour certains, réside dans la séparation entre l'ensemble d'entraînement (train) et l'ensemble de test, ainsi que dans la distinction entre les variables explicatives (X) et la variable cible (Y). Dans ce contexte, la variable à prédire est également la variable prédite, et la division en ensembles de train et de test s'effectue en fonction de la temporalité : l'ensemble d'entraînement comprend les valeurs passées, jusqu'à juillet 2022, tandis que l'ensemble de test englobe les valeurs allant de juillet 2022 à juillet 2023 que l'on cherche à prédire.

À l’issue du traitement des données, nous présentons la première série à notre disposition (**Série brute**). La non-stationnarité du processus est manifeste, comme l'a confirmé le test de Dickey-Fuller, avec une p-value non significative. Ainsi, une différenciation de la série s'avère nécessaire pour obtenir le graphique présent en annexe. (**Série différenciée**). Cependant, il est à noter une volatilité significative au sein de notre série, ce qui pourrait poser des défis importants lors de la prise en compte de ces variations au sein des modèles.

À présent, nous pouvons examiner les résidus, l'autocorrélation (**ACF**) et les autocorrélations partielles (**PACF**). Selon la définition fournie par [[4]](#_heading=h.32hioqz):

* La fonction d'autocorrélation (ACF) au décalage k représente la corrélation entre les valeurs de séries séparées par k intervalles.
* La fonction d'autocorrélation partielle (PACF) au décalage k représente la corrélation entre les valeurs de séries séparées par k intervalles, en tenant compte des valeurs des intervalles intermédiaires.

À titre de rappel, les choix de p et q sont déterminés selon les critères suivants:



*Tableau 2 : Propriétés des processus MA(q), AR(p) et ARMA(p,q) (cours)*

Ici, nous avons deux choix. Soit nous optons pour le modèle le plus simple, une excellente solution en général, en sélectionnant q = 1. Soit, étant donné que nous disposons de données massives, nous tentons q = 8, sachant que le pic est isolé. Les deux stratégies seront essayées.

L'analyse des pics à t-28 est limitée ici, car cela entraînerait l'estimation de nombreux paramètres. En augmentant le nombre de retards jusqu'à t-10, nous accroissons le risque d'obtenir des valeurs inexactes. Cependant, cette limite supérieure demeure un choix nécessaire. Notre approche implique l'exploration de modèles avec des valeurs de p dans l'intervalle [0, 10] et de q dans l'intervalle [0, 10]. L'évaluation sera basée sur les critères AIC et BIC, avec une attention particulière accordée à ce dernier dans notre contexte, étant donné qu'il pénalise davantage les modèles complexes.

Ensuite, on vérifie que les résidus sont bien des bruits blancs à l’aide du test de Box-Ljung. Nous avons testé les modèles ARIMA suivants :

* ARIMA(6,1,10) minimisant le AIC
* ARIMA(1,1,1) minimisant le BIC
* ARIMA(0,1,3) de la fonction auto.arima() à titre informatif
* ARIMA(8,1,1) selon notre hypothèse de départ

Parmi les quatre modèles considérés, seuls deux satisfont au test, à savoir les modèles ARIMA(6,1,10) et ARIMA(8,1,1) présentant une p-value non significative. Par la suite, une prévision a été réalisée pour les 12 mois suivants, correspondant à notre jeu de test, afin de calculer le RMSE des deux modèles. Les résultats ont montré un RMSE de 38.17 pour l'ARIMA(6,1,10) et de 34.36 pour l'ARIMA(8,1,1). On peut noter que le modèle ARIMA(6,1,10) génère des prédictions supérieures aux valeurs réelles, suggérant une tendance à surestimer les données observées. Cette observation soulève des questions quant à la précision et à la fiabilité de ce modèle spécifique dans la prévision du nombre de séisme en Alaska (**Prévision ARIMA**).

On peut noter l’équation de notre modèle ARIMA(8,1,1) qui est la suivante :

En raison de la volatilité constatée dans la série temporelle et de notre insatisfaction quant à l'ajustement du modèle actuel, nous avons entrepris d'explorer une alternative légèrement plus précise. Cette démarche vise à améliorer la qualité des prévisions et à mieux capturer les tendances observées dans les données sismiques en Alaska.

## GARCH

### 2.1 Partie théorique

En présence de volatilité, notamment dans le contexte financier, il est nécessaire de recourir à des modèles robustes à l'hétéroscédasticité. Le modèle GARCH repose sur l'idée principale que la variance conditionnelle de la série temporelle dépend des valeurs passées de la série elle-même et des résidus passés. Cette approche vise à mieux prendre en compte les variations de la volatilité au fil du temps et à fournir des prévisions plus précises dans un contexte caractérisé par des fluctuations importantes.

La formule générale du modèle GARCH(p,q) est définie par :

Où :

* est la variance conditionnelle à l’instant t
* est un terme constant
* et sont des coefficients du modèle
* est le résidus au temps t-i
* est la variance conditionnelle au temps t-j

En résumé, l'approche d'ARIMA se concentre sur la modélisation de la série temporelle elle-même, tandis que GARCH s'intéresse spécifiquement à la modélisation de la volatilité conditionnelle de cette série temporelle. Dans certains cas, il peut être judicieux d'utiliser ces deux types de modèles de manière conjointe afin de capturer différentes composantes d'une série temporelle complexe. C'est précisément ce que nous allons explorer dans la section suivante.

### 2.2 Partie pratique

## Light GBM

### 3.1 Partie théorique

### 3.2 Paramètre par défaut

Dans cette section dédiée à l'utilisation de Light GBM pour la modélisation, nous explorons l'application de cet algorithme afin de prédire le nombre de séismes par mois. Nous débuterons en utilisant les valeurs par défaut de Light GBM, avant d'explorer et de comparer les performances obtenues après l'application de trois méthodes d'optimisation hyperparamétrique. Les résultats de ces différentes approches seront présentés de manière détaillée, offrant ainsi un aperçu complet de l'impact de l'ajustement des hyperparamètres sur la qualité des prédictions dans notre modèle.

Nous avons initié le processus d'entraînement de notre modèle LightGBM en utilisant les paramètres par défaut, dans le but de créer une référence de performance de base. Cette approche nous permet d'évaluer les capacités intrinsèques de l'algorithme sans aucune modification préalable de ses hyperparamètres.

**Entrainement :**

L'entraînement du modèle LightGBM avec les paramètres par défaut s'articule autour de plusieurs étapes clés. Tout d'abord, nous amorçons le processus en définissant le dictionnaire params sans spécifier de paramètres particuliers, laissant ainsi LightGBM utiliser ses valeurs par défaut, généralement adaptées à diverses tâches de modélisation. Ensuite, nous préparons nos ensembles de données d'entraînement (X\_train, y\_train) et de validation (X\_test, y\_test) en les convertissant en formats compatibles avec LightGBM à l'aide de la classe lgb.Dataset. Cette étape essentielle facilite la manipulation des données pour l'entraînement du modèle.

Il est important de souligner que, pour évaluer les performances du modèle, nous utilisons le jeu de test correspondant aux 12 derniers mois de notre série temporelle, avec la dernière valeur test datée de juillet 2023. Cette approche garantit une évaluation réaliste du modèle sur des données temporelles récentes et représente une étape cruciale dans la validation de sa capacité à généraliser à de nouvelles observations.

Par la suite, nous définissons le nombre maximal d'itérations de boosting, noté num\_round, indiquant le nombre total d'arbres à entraîner. Dans l'exemple fourni, nous l'avons fixé à 1000 de manière arbitraire. Enfin, nous appelons la fonction lgb.train() en fournissant les paramètres, les ensembles de données d'entraînement et de validation, ainsi que le nombre maximal d'itérations. Le modèle évalue périodiquement les performances sur l'ensemble de validation à chaque itération, déclenchant un arrêt anticipé après 20 itérations sans amélioration significative (early\_stopping\_rounds=20). À la fin du processus, le modèle entraîné est stocké dans l'objet bst, prêt à être évalué et utilisé pour des prédictions ultérieures.

Une fois que le modèle a été formé avec succès, nous sommes prêts à effectuer des prédictions sur l'ensemble de données de test. Le processus de transformation des prédictions est ensuite initié afin d'obtenir un format de données correspondant à la structure d'origine, composée de nombres entiers, étant donné que le nombre de séismes est une valeur entière. Pour cette raison, les valeurs prédites sont arrondies à l'entier le plus proche. Une fois ces prédictions transformées, nous procédons au calcul du Root Mean Squared Error (RMSE), une métrique de performance couramment utilisée pour évaluer la précision des prédictions par rapport aux valeurs réelles. Il est important de noter que, si nécessaire, la formule mathématique du RMSE peut être consultée en annexe (**Métrique Régression**).

Le RMSE obtenu est de 23.1048 sur l'ensemble de test signifie que, en moyenne, les prédictions du modèle diffèrent d'environ 23.1048 unités du nombre réel de séismes par mois. Un RMSE plus bas indique une meilleure adéquation entre les prédictions du modèle et les données réelles, soulignant ainsi la précision relative du modèle dans la prédiction du nombre de séismes. Comparé à des méthodes alternatives telles que les modèles autorégressifs et GARCH, notre modèle LightGBM démontre une performance notablement supérieure. Cette différence suggère que le modèle LightGBM est plus efficace pour capturer les variations mensuelles du nombre de séismes.

### 3.3 Recherche aléatoire

La recherche aléatoire (RandomSearch CV) par K-Fold est une technique d'optimisation des hyperparamètres d'un modèle d'apprentissage automatique qui combine les avantages de la méthode RandomSearch et de la validation croisée (CV) en k-fold. Elle consiste à définir une grille de recherche d'hyperparamètres, puis à effectuer des itérations aléatoires en sélectionnant des valeurs pour les hyperparamètres à chaque itération. Cependant, au lieu d'évaluer chaque combinaison d'hyperparamètres sur l'ensemble de données d'entraînement, la méthode RandomSearch CV par K-Fold évalue chaque combinaison d'hyperparamètres sur plusieurs échantillons de validation différents.

Le processus de validation croisée en k-fold consiste à diviser l'ensemble de données en k échantillons de taille égale, puis à entraîner le modèle sur k-1 échantillons et à le valider sur le reste. Ce processus est répété k fois pour que chaque échantillon serve une fois de validation. L'erreur moyenne de validation est ensuite calculée pour chaque combinaison d'hyperparamètres. Une explication complète du processus est disponible en annexe, au cas où cela serait nécessaire (**Recherche aléatoire**).

Nous avons décidé de régler certains aspects fondamentaux de notre modèle, à savoir les paramètres num\_leaves, n\_estimators, et le taux d'apprentissage (lr), afin d'optimiser ses performances.

Le paramètre num\_leaves correspond au nombre maximal de feuilles dans chaque arbre de notre modèle. Nous avons exploré une plage de valeurs allant de 20 à 50 avec un pas de 5, cherchant ainsi à comprendre comment la complexité de l'arbre impacte la performance du modèle.

Le paramètre n\_estimators indique le nombre total d'arbres dans notre modèle. Nous avons varié ce paramètre de 50 à 300 avec un pas de 50 pour évaluer comment la taille de l'ensemble d'arbres influence la capacité prédictive globale du modèle.

Quant au taux d'apprentissage (lr), il détermine la taille des pas lors de la mise à jour des poids du modèle. Nous avons examiné des valeurs allant de 0.01 à 0.2 avec un pas de 0.05, cherchant à trouver un équilibre optimal entre une convergence rapide et une convergence stable.

Afin de trouver la combinaison optimale de ces paramètres, nous avons utilisé une méthode de validation croisée temporelle (TimeSeriesSplit) avec 5 splits. Cette approche est cruciale pour préserver l'ordre temporel des données d'entraînement et de test, une considération essentielle dans le contexte des séries temporelles. Elle permet également d'évaluer la performance du modèle de manière robuste sur des données qui reflètent la progression du temps.

Après avoir effectué des itérations sur différentes combinaisons de paramètres, nous avons identifié les valeurs optimales qui maximisent les performances de notre modèle. Ainsi, les paramètres optimaux que nous avons déterminés sont num\_leaves=35, n\_estimators=50 et lr=0.11. Grâce à ces paramètres spécifiques, nous avons obtenu un RMSE de 22.9547, une amélioration significative par rapport aux performances du modèle utilisant les paramètres par défaut.

### 3.3 Méthode bayésienne

Cette optimisation [[5]](#_heading=h.32hioqz) prend en compte les évaluations passées lors du choix de l’ensemble des hyperparamètres. En choisissant ces combinaisons de paramètres qui apporteront les scores de validation les plus prometteurs. Cette approche nécessite généralement moins d’itérations pour obtenir l’ensemble optimal de valeurs d’hyperparamètres. Plus particulièrement parce qu’il ne tient pas compte des zones de l’espace des paramètres qui apporte peu d’information. Il y a quelques étapes principales à cette technique :

* Choisir l’espace de recherche des hyperparamètres
* Définir la fonction objective (dans notre cas l’accuracy a été utilisé)
* Maintenir un modèle probabiliste pour la relation entre les valeurs d’hyperparamètres et les performances du modèle (fonction de substitution et de sélection).
* Alterner entre :

1. Entrainement avec des valeurs d’hyperparamètres qui maximisent l’amélioration attendue.
2. Utiliser les résultats pour mettre à jour le modèle probabiliste initial et ses attentes.

* Sélectionnez les hyperparamètres les plus performants parmi tous les essais.

Dans le contexte de l'optimisation bayésienne, l'objectif est de maximiser la fonction objective. Cependant, l'évaluation de chaque ensemble d'hyperparamètres peut être coûteuse, surtout lorsque de grandes grilles d'hyperparamètres sont utilisées, car cela nécessite l'entraînement du modèle et son évaluation sur l'ensemble de validation. Par conséquent, la fonction objective n'est déclenchée que lorsque nous sommes certains qu'un ensemble d'hyperparamètres conduit à une amélioration du score de validation.

De ce fait, la fonction de substitution et de sélection travaille ensemble pour proposer des paramètres qui, selon elles, apporteront la plus grande précision à la fonction objective. En utilisant le package hyperopt en Python, l'algorithme TPE (Tree Parzen Estimator) a été utilisé comme modèle de substitution de la fonction objective. La fonction de sélection utilisée est l'amélioration attendue (Expected Improvement). Une explication détaillée de cette méthode d'optimisation des hyperparamètres est disponible en annexe, si nécessaire (**Méthode bayésienne**).

En utilisant la recherche bayésienne sur la même plage de valeurs et effectuant 100 itérations, nous avons identifié des paramètres optimaux, à savoir learning\_rate=0.02057, n\_estimators=50, num\_leaves=27. Cette configuration a abouti à un RMSE de 19.83, démontrant une performance notablement supérieure par rapport à la recherche aléatoire.

La recherche bayésienne a identifié un taux d'apprentissage relativement bas (0.02057), indiquant que des pas plus petits lors de la mise à jour des poids ont contribué à une convergence plus stable et efficace du modèle. Cette approche suggère une préférence pour des ajustements plus délicats des paramètres du modèle, permettant ainsi une meilleure adaptation aux caractéristiques subtiles des données.

### 3.4 Algorithme génétique

La méthode de recherche génétique [[6]](#_heading=h.32hioqz) a été employée pour construire le modèle optimal de la garantie dommage collision. La méthode a été implémenté sur Python à l’aide du package DEAP (Distributed Evolutionary Algorithms in Python). L’algorithme de la génétique est un algorithme de recherche heuristique et d’optimisation qui s’inspire de la théorie de l’évolution naturelle de Darwin Charles. Le processus de l’algorithme génétique est disponible en annexe au besoin (**Méthode génétique)**

Dans notre approche, nous avons défini les mêmes plages de valeurs pour les hyperparamètres que celles utilisées par d'autres méthodes. Notre population était constituée de 50 individus, évoluant sur 3 générations. De manière arbitraire, la probabilité de croisement a été fixée à 70%, laissant ainsi 30% pour la mutation. Les hyperparamètres retenus à l'issue de cette exploration génétique sont num\_leaves=24, n\_estimators=64 et lr=0.015. Il est intéressant de noter que la méthode génétique a convergé vers une valeur de lr relativement proche de celle obtenue par la méthode bayésienne. Le RMSE obtenu à la suite de cette optimisation est de 20.13. Cette performance souligne la capacité de la méthode génétique à découvrir des configurations de paramètres efficaces, avec un résultat comparable à la méthode bayésienne dans ce contexte spécifique d'optimisation d'hyperparamètres.

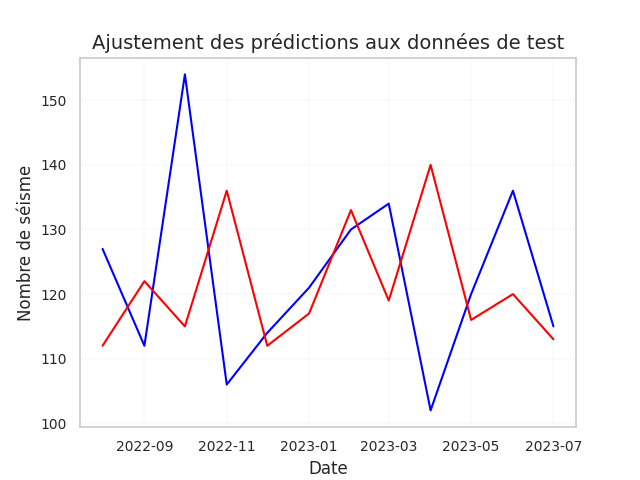
### 3.5 Modèle retenue

Le modèle sélectionné est celui qui minimise le RMSE, à savoir le Light GBM, obtenu grâce à la méthode d'optimisation des hyperparamètres bayésienne. Les métriques associées à la régression sont les suivantes :

| Light GBM Bayésien | |
| --- | --- |
| Métrique | **Valeur** |
| MSE | 393,33 |
| RMSE | 19,83 |
| MAE | 14,83 |
| MAPE | 12,25% |

*Tableau 3 : Métrique Light GBM*

Les résultats du modèle Light GBM obtenus grâce à la méthode d'optimisation bayésienne révèlent une performance notable. Le MSE affiche une valeur de 393,33, indiquant une adéquation générale entre les prédictions du modèle et les données réelles. Le RMSE avec une valeur de 19,83, souligne la précision relative élevée des prédictions par rapport aux valeurs réelles. Le MAE affiche une moyenne de 14,83, indiquant une précision moyenne des prédictions. Le MAPE de 12,25% suggère une précision relative satisfaisante en pourcentage. Ces résultats globaux reflètent la capacité du modèle Light GBM, optimisé par la méthode bayésienne, à fournir des prédictions précises et adaptées aux caractéristiques spécifiques des données de régression.



*Figure 5 : Ajustement des prédictions aux données test*

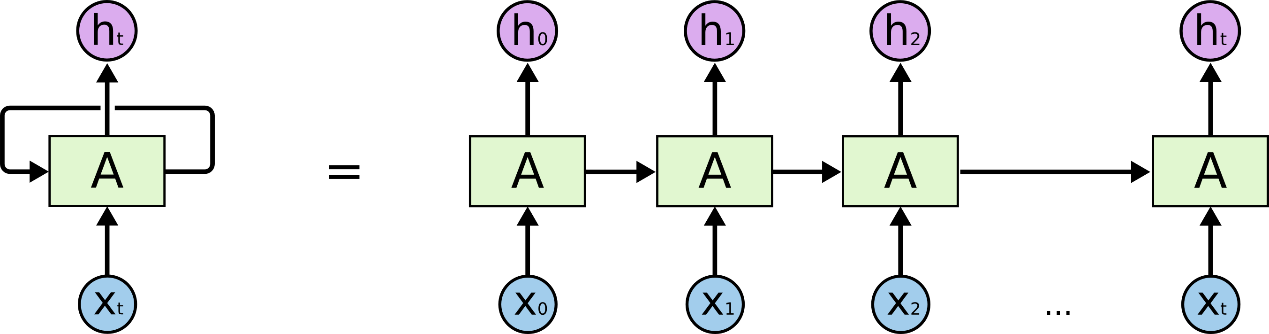
À partir de cette représentation graphique, il est apparent que les prédictions du nombre de séismes sur le jeu de données test présentent, dans l'ensemble, une proximité remarquable entre elles. Cependant, il est intéressant de noter des mois spécifiques, tels qu'août 2022, novembre 2022 et avril 2023, où les prédictions s'écartent sensiblement des valeurs réelles. Ces écarts peuvent être attribués à des facteurs aléatoires qui échappent à la capacité du modèle à les capturer. Ces observations soulignent l'importance de reconnaître les limites d'un modèle, en particulier face à des événements imprévus ou des variations atypiques qui peuvent influencer les prédictions de manière significative.

## LSTM

### 4.1 Partie théorique

**Réseaux de neurones récurrents :**

Les Réseaux Neuronaux Récurrents (ou *Recurrent Neural Network*, RNN) sont une catégorie puissante de modèles de réseaux neuronaux conçus pour traiter des données séquentielles, comme des séquences temporelles, des phrases ou des séquences d'événements.

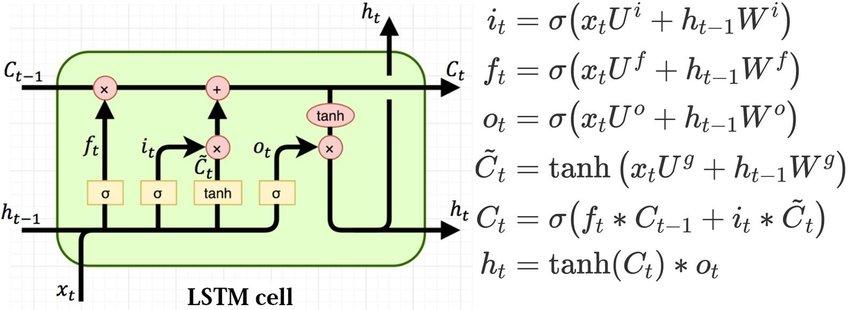


*Figure 6 : Réseaux de Neurones Récurrents*

Un RNN est représenté par un diagramme où un segment du réseau () prend en compte des données d'entrée , produisant une sortie . La particularité réside dans une boucle qui permet la transmission d'informations d'une étape à l'autre, créant ainsi une récurrence dans le processus. Cette structure autorise au réseau de capturer les dépendances séquentielles dans les données, rendant le RNN particulièrement adapté pour traiter des séquences temporelles.

Les RNN rencontrent des limitations liées à l'atténuation des signaux provenant du passé lointain, compliquant l'apprentissage de dépendances complexes sur des séquences temporelles. Cette difficulté est adressée par les Mémoires à Long Terme à Court Terme (ou *Long Short Term Memory*, LSTM), une variante des RNN.

**LSTM :**



*Figure 7 : Structure des LSTM*

Les éléments de base d’un LSTM et leurs fonction sont les suivantes :

* Le nouvel élément des LSTM est l’introduction de l’état de cellule qui permet de stocker et transporter les informations, qui est régulé par 3 portes. Les portes sont composées de fonctions sigmoïdes afin qu’elles génèrent des valeurs comprise entre 0 et 1. La première porte est appelé « *porte de l’oubli*» qui examine et et attribue un nombre compris entre 0 et 1 pour chaque élément du vecteur d’état de cellule . La sortie de la porte d’oubli est calculée comme suit.

Avec :

* est le vecteur de port d’oubli à l’instant
* est la fonction d’activation sigmoïde
* est la matrice de poids porte de l’oubli ()
* est la matrice de poids porte de l’oubli ()
* Ensuite, la *porte d’entrée* décide quelles unités cellulaire doivent être mise à jour avec de nouvelle informations. Pour cela, comme pour la sortie de la porte de l’oubli, une valeur comprise entre 0 et 1 est calculée pour chaque composante d’état de cellule via la logique suivante :

Puis, un ensemble candidat de nouvelles valeurs est créé pour l’état de la cellule à l’aide de et comme entrée. L’état de la cellule candidate est calculée comme suit :

Le nouvel état de la cellule est mis à jour comme suit :

* La tâche suivante consiste à déterminer les états de cellule à afficher car l’état de cellule contient beaucoup d’informations. Pour cela, et passent par une porte de sortie qui génère une valeur comprise entre 0 et 1 pour chaque composantedu vecteur d’état de cellule . La sortie de cette porte est calculée comme suit :
* L’état caché est calculé à partir de l’état de la cellule en faisant passer chacun de ses éléments via une fonction tanh puis en effectuant un produitpar élément avec les valeurs de la sortie :

À titre informatif, la flèche vers le haut à partir de indique que la sortie de l'état caché à cet instant précis est utilisée comme la sortie principale du LSTM à cet instant, contribuant ainsi à la prédiction ou à la représentation finale.

### 4.2 Partie pratique

Dans cette section, nous adopterons une approche de Deep Learning pour anticiper le nombre de séismes en Alaska. Dans un premier temps, nous détaillerons l'architecture de notre modèle. Le tableau synthétique suivant offre un récapitulatif de cette dernière :

| **Couche** | **Type** | **Neurones** | **Forme de sortie** | **Paramètres** |
| --- | --- | --- | --- | --- |
| lstm\_363 | LSTM | 50 | (None, 12, 50) | 10 400 |
| lstm\_364 | LSTM | 200 | (None, 12, 200) | 200 800 |
| dropout\_259 | Dropout | 0 |  | 0 |
| lstm\_365 | LSTM | 150 | (None, 12, 150) | 210 600 |
| dropout\_260 | Dropout | 0 |  | 0 |
| lstm\_366 | LSTM | 100 | (None, 100) | 100 400 |
| dropout\_261 | Dropout | 0 |  | 0 |
| dense\_166 | Dense | 1 | (None, 1) | 101 |

*Tableau 4 : Tableau Architecture*

Notre modèle repose sur quatre couches LSTM, avec une approche consistant à débuter à 50 neurones, augmenter à 200 pour la deuxième couche, et ensuite diminuer progressivement. Chaque couche LSTM traite une séquence de 12 valeurs temporelles. La dernière couche LSTM, orientée vers la prédiction d'une seule valeur, est suivie de couches Dense pour une sortie unique.

Pour éviter le surapprentissage, des couches de régularisation, avec un Dropout de 20%, sont intégrées, visant à équilibrer la régularisation sans compromettre significativement la capacité d'apprentissage. Les essais avec des couches de BatchNormalization et GaussianNoise (0.005, 0.01 et 0.015) n'ont pas amélioré les résultats.

Les couches LSTM (lstm\_363, lstm\_364, lstm\_365) sont configurées avec return\_sequences=True, générant une séquence de valeurs à chaque pas de temps. Cette séquence reflète l'évolution des informations temporelles à différents niveaux d'abstraction, contribuant à une représentation hiérarchique des motifs temporels. Chaque couche LSTM forme une séquence d'informations transmise à la suivante, capturant des dépendances temporelles complexes et améliorant la compréhension des schémas temporels. Enfin, la dernière couche Dense, configurée pour une sortie unique, utilise ces représentations pour la prédiction finale.

Dans le cadre de la compilation de notre modèle, des choix stratégiques ont été opérés pour garantir une optimisation efficace lors de l'entraînement. Tout d'abord, l'optimiseur Adam a été sélectionné pour guider la minimisation de la fonction de perte. Adam combine les avantages de la descente de gradient stochastique (SGD) avec momentum et la méthode de moyenne mobile exponentielle (RMSprop), ce qui en fait un choix courant et performant pour les tâches de séries temporelles. Cette méthode d'optimisation est particulièrement adaptée aux problèmes de régression comme le nôtre.

En ce qui concerne la fonction de perte, nous avons opté pour la « mean\_squared\_error », une mesure appropriée pour les tâches de régression où l'objectif est de minimiser la différence entre les prédictions du modèle et les valeurs réelles. Cette fonction de perte est particulièrement pertinente pour notre tâche spécifique de prédiction dans le domaine des séries temporelles.

Le processus d'entraînement a été soigneusement orchestré sur l'ensemble d'entraînement. Nous avons choisi de réaliser 50 époques, indiquant le nombre de fois où le modèle parcourt l'intégralité de l'ensemble d'entraînement. Ce nombre a été déterminé après une expérimentation empirique visant à trouver un équilibre entre l'efficacité de l'entraînement et la prévention du surapprentissage. En parallèle, la taille de lot (batch\_size) a été fixée à 32, déterminée également de manière empirique pour optimiser l'apprentissage du modèle tout en évitant les risques de surajustement. Ces ajustements judicieux des paramètres ont pour objectif d'assurer une convergence robuste du modèle tout en préservant sa capacité à généraliser efficacement aux données inconnues.

Les résultats du modèle, avec 522 301 paramètres, montrent des performances prometteuses dans la prédiction des séismes mensuels en Alaska. Le RMSE atteint 15.22, le MAE est de 12.09, et le MAPE, exprimé en pourcentage, s'établit à 9.62%, évaluant la précision relative des prédictions par rapport aux observations réelles. Ces métriques attestent de la capacité de ce premier modèle à capturer efficacement l'information liée au nombre de séismes, démontrant ainsi sa pertinence dans la modélisation de cette série temporelle spécifique.

Dans la dernière phase de notre étude, nous avons procédé à l'ajustement de notre modèle LSTM sur les ensembles d'entraînement et de test, détaillé dans l'annexe « **Ajustement LSTM** ». Lors de l'ajustement sur les données d'entraînement, notre modèle a démontré une capacité remarquable à suivre la tendance générale de la série temporelle, tout en capturant de manière précise les variations volatiles inhérentes aux données. Cette adaptabilité suggère une profonde compréhension des structures temporelles sous-jacentes et témoigne de l'efficacité de notre approche.

En revanche, lors de l'ajustement sur les données de test, bien que le modèle conserve une adaptation satisfaisante, son ajustement présente une tendance quasi linéaire, révélant une certaine prudence dans son comportement. Cette approche plus conservatrice peut être interprétée comme une manifestation de l'aversion aux risques du modèle, privilégiant la stabilité plutôt qu'une réponse plus dynamique face aux variations du jeu de test. Malgré cette prudence, le modèle excelle en surpassant les autres modèles examinés au cours de notre étude, soulignant ainsi sa solidité et son efficacité dans la prédiction des séismes mensuels en Alaska. Ces résultats consolidés renforcent la confiance dans la pertinence de notre modèle LSTM pour la modélisation de séries temporelles complexes.

# Conclusion

En conclusion de ce projet, nous a plongés dans l'exploration minutieuse de diverses méthodes de prévision du nombre mensuel de séismes en Alaska. Alors que les approches traditionnelles de séries temporelles comme ARIMA et GARCH ont laissé à désirer en termes de performances, l'adoption de Light GBM a généré des résultats très acceptables. Cependant, c'est l'approche basée sur les réseaux neuronaux LSTM qui a véritablement surpassé nos attentes, offrant des performances exceptionnelles et démontrant une robustesse, rapidité et efficacité remarquables, constituant ainsi des atouts inestimables dans notre analyse.

Il est important de reconnaître les limites de notre approche, notamment le fait que notre focus était spécifiquement l'Alaska, et par conséquent, toute extrapolation à l'échelle mondiale doit être entreprise avec prudence. Une approche plus rigoureuse impliquerait la reproduction de cette analyse pour l'ensemble des pays/états, peut-être en incluant des variables exogènes spécifiques à chaque région. Une étude individuelle de chaque pays pourrait permettre de mieux comprendre les facteurs influençant le nombre de séismes.

L'implémentation d'un système de monitoring du modèle à travers une application Streamlit, dont le lien GitHub est disponible [ici](https://github.com/Alioth108/APP-EARTHQUAKE), a rencontré des obstacles liés à des problèmes de compatibilité de version entre streamlit et tensorflow, qui nécessitent encore une résolution. Ce système aurait permis aux utilisateurs d'importer leurs propres données pour obtenir des prévisions sur les 12 prochains mois. Cependant, il est important de noter que le modèle peut présenter des limites face à des changements sismiques brusques, soulignant le besoin d'intégrer des variables explicatives plus détaillées.

L'idée d'un chatbot pré-entraîné pour fournir des commentaires en temps réel sur les prévisions aurait ajouté une dimension interactive et interprétative, mais sa mise en œuvre n'a pas été achevée dans le temps imparti. De même, la création d'un dashboard interactif, potentiellement incorporant divers KPI du jeu de données, demeure une piste à explorer dans le futur pour une visualisation plus riche et compréhensible des résultats obtenus.

# Bibliographie

[1] Wikipédia. (Consulté le 07/01/2024). « Magnitude (Sismologie) ». Disponible à : https://fr.wikipedia.org/wiki/Magnitude (sismologie).

[2] Musée de sismologie. (Consulté le 07/01/2024). « Magnitude de moment ». Disponible à : https://musee-sismologie.unistra.fr/comprendre-les-seismes/notions-pour-petits-et-grands/la-sismicite/la-magnitude-mw-et-le-moment-sismique/: :text=On%20peut%20la%20calculer%20%C3%A0,ne%20touche%20pas%20la%20surface.

[3] Wikipédia. (Consulté le 10/01/2024). « ARMA ». Disponible à : https://fr.wikipedia.org/wiki/ARMA.

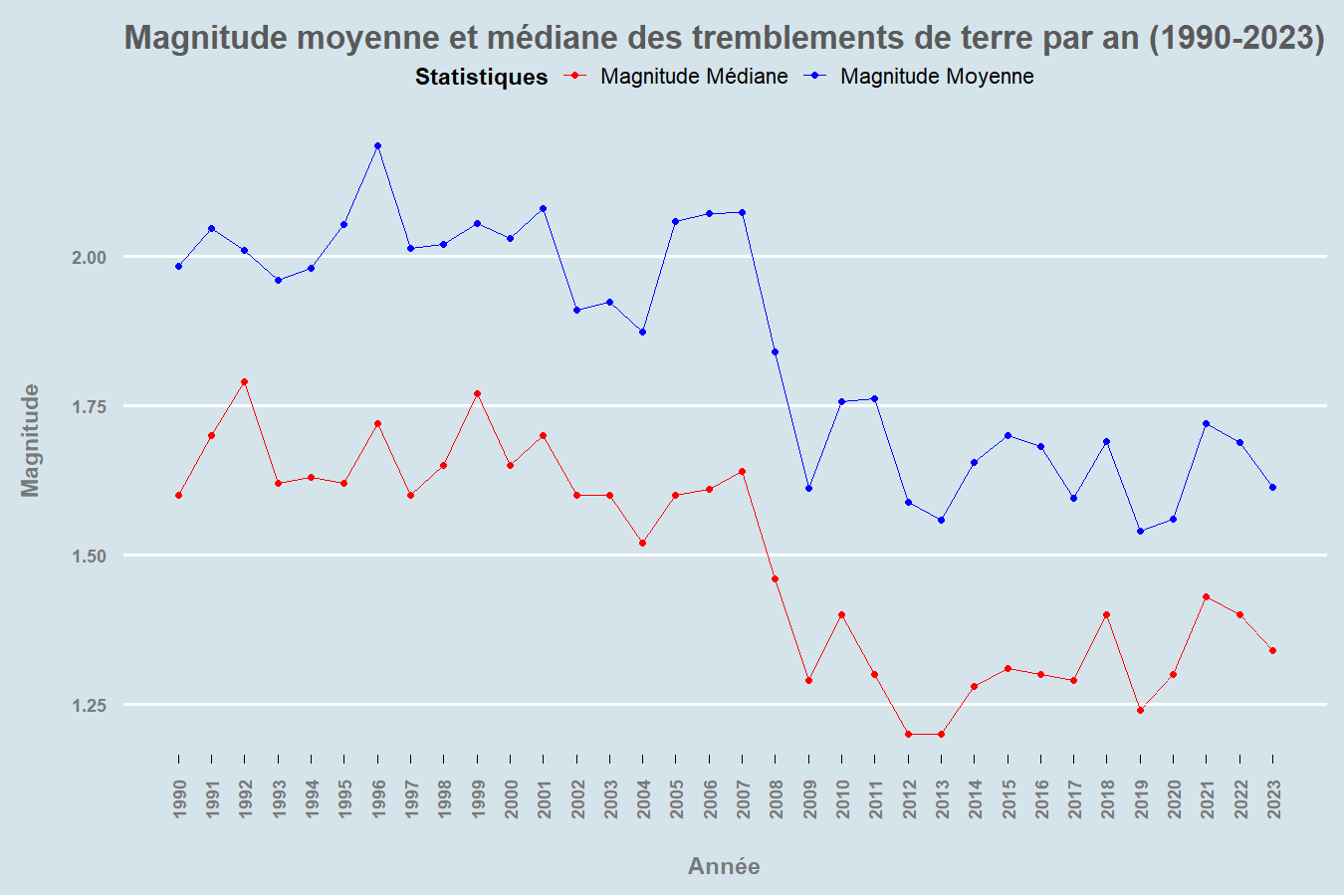
[4] Fonctions d'autocorrélation et d'autocorrélation partielle. IBM. Accessible via : https://www.ibm.com/docs/fr/spss-modeler/saas?topic=data-autocorrelation-partial-autocorrelation-functions.

[5] J. Bergstra ; R. Bardenet ; Y. Bengio et B. Kégl – “*Algorithmes for Hyper-Parameter Optimization”* – Décembre 2011

[6] International Research Journal of Engineering and Technology – “Genetic algorithm for solving simple mathematical equality problem” – Mai 2020

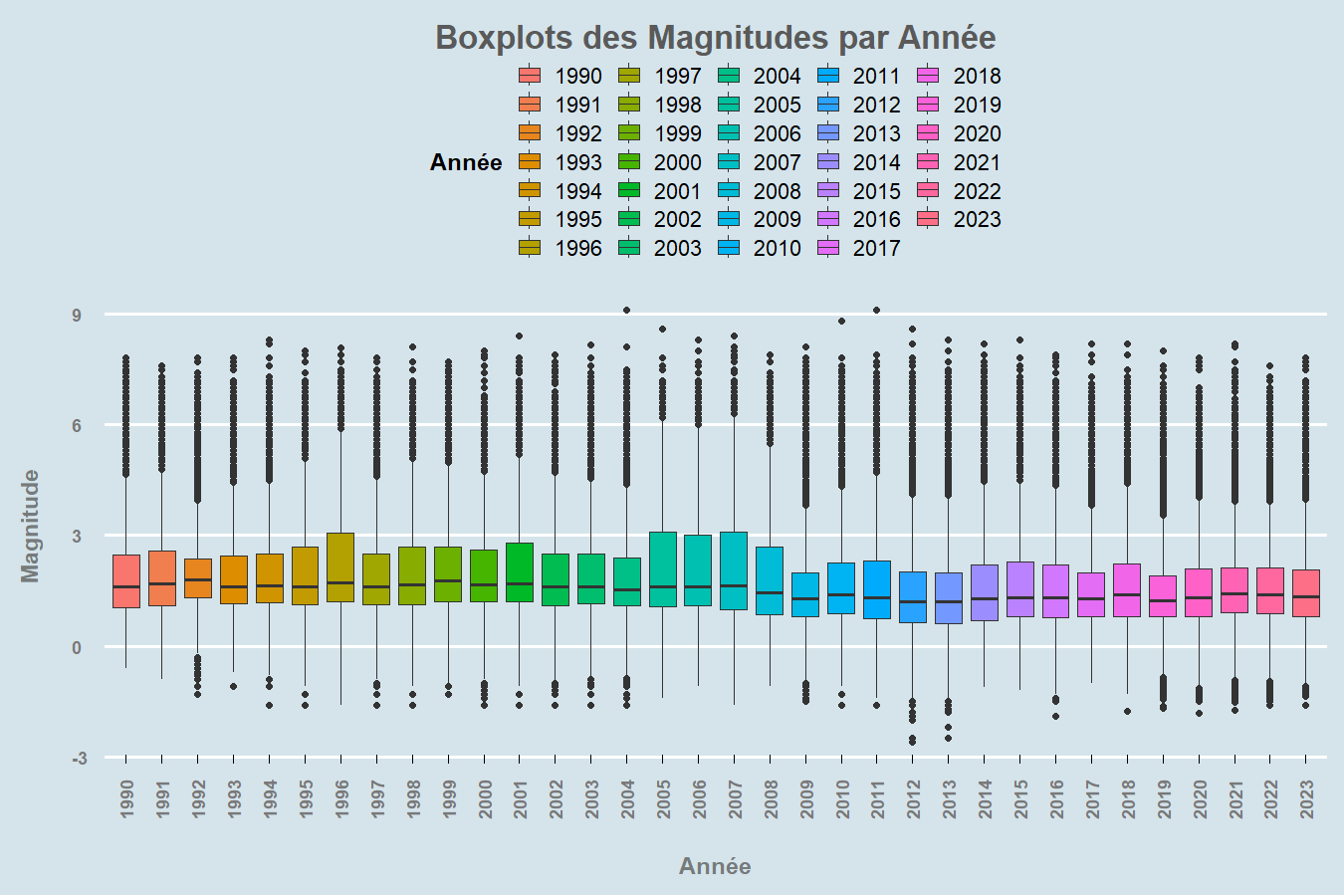
# Annexes

### Magnitude moyenne et médiane sur différentes périodes



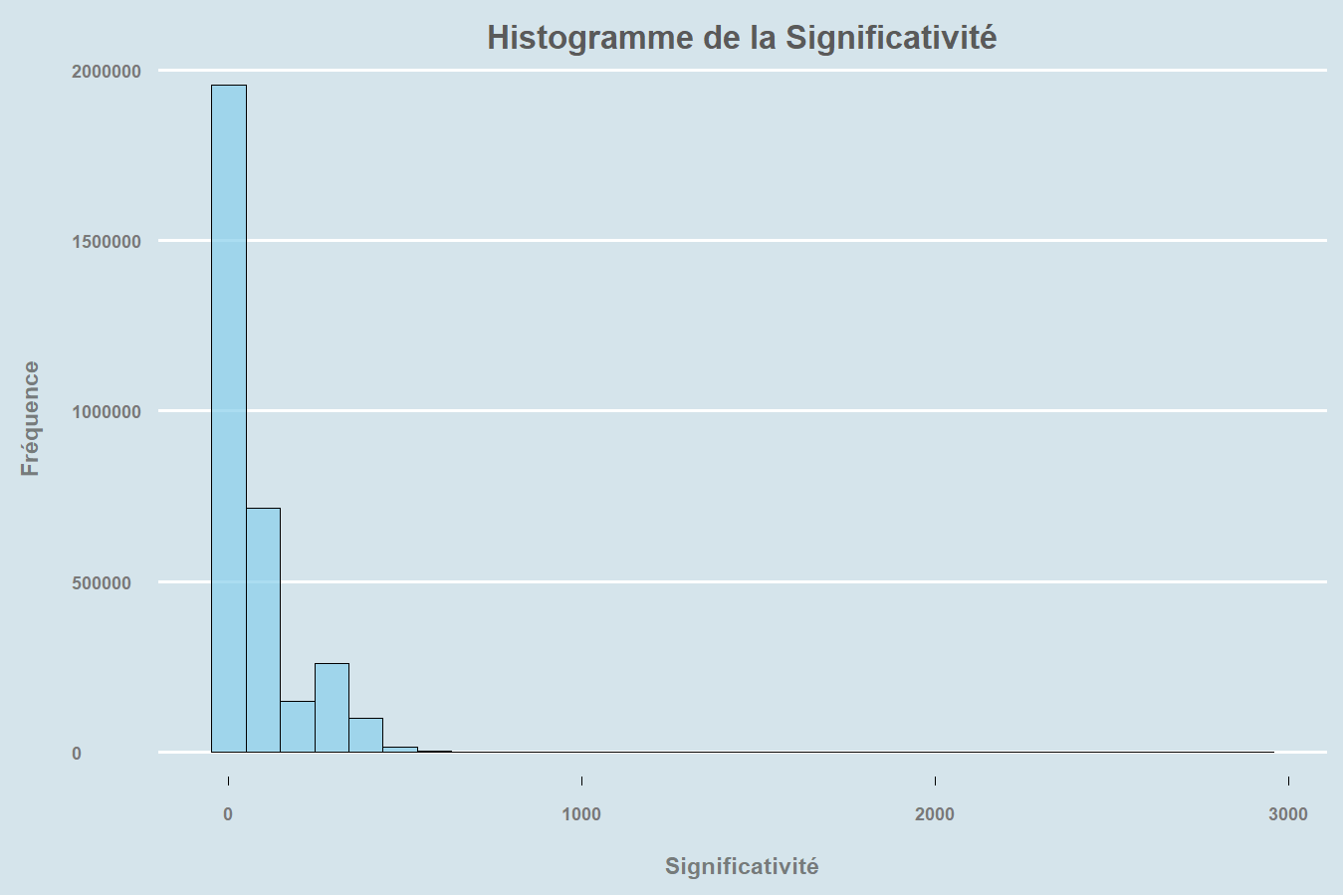
*Figure 8 : Magnitude moyenne et médiane des séismes par année*

### Boxplot des magnitudes



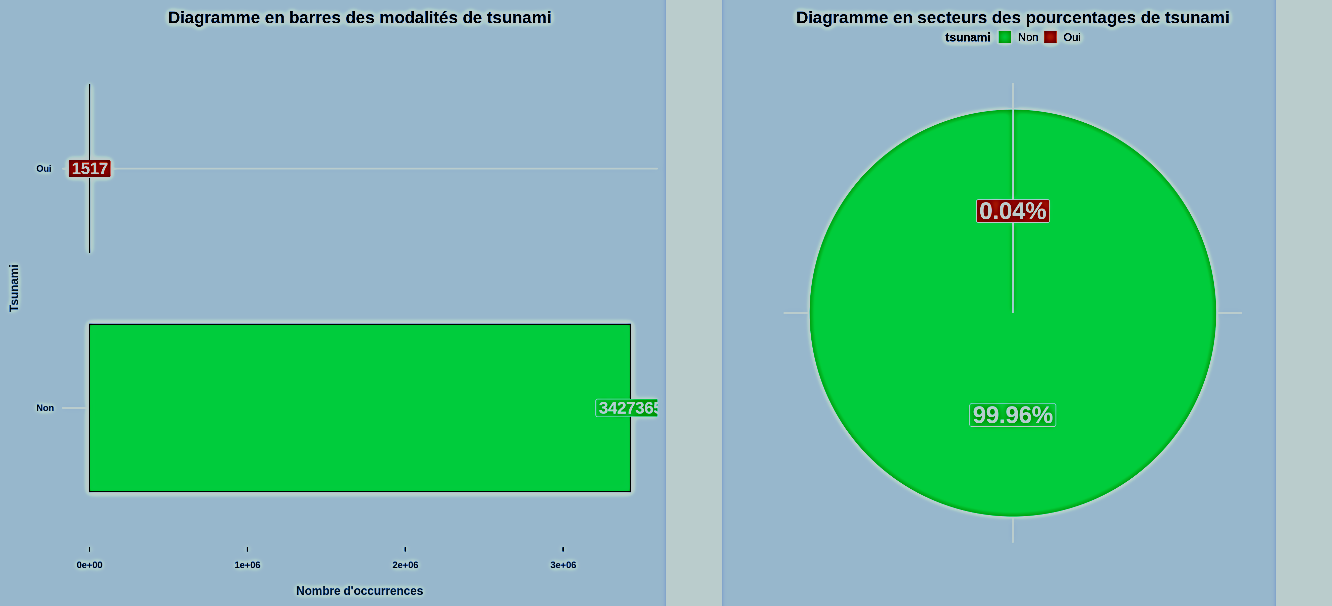
*Figure 9 : Boxplot des magnitudes*

### Histogramme significativité



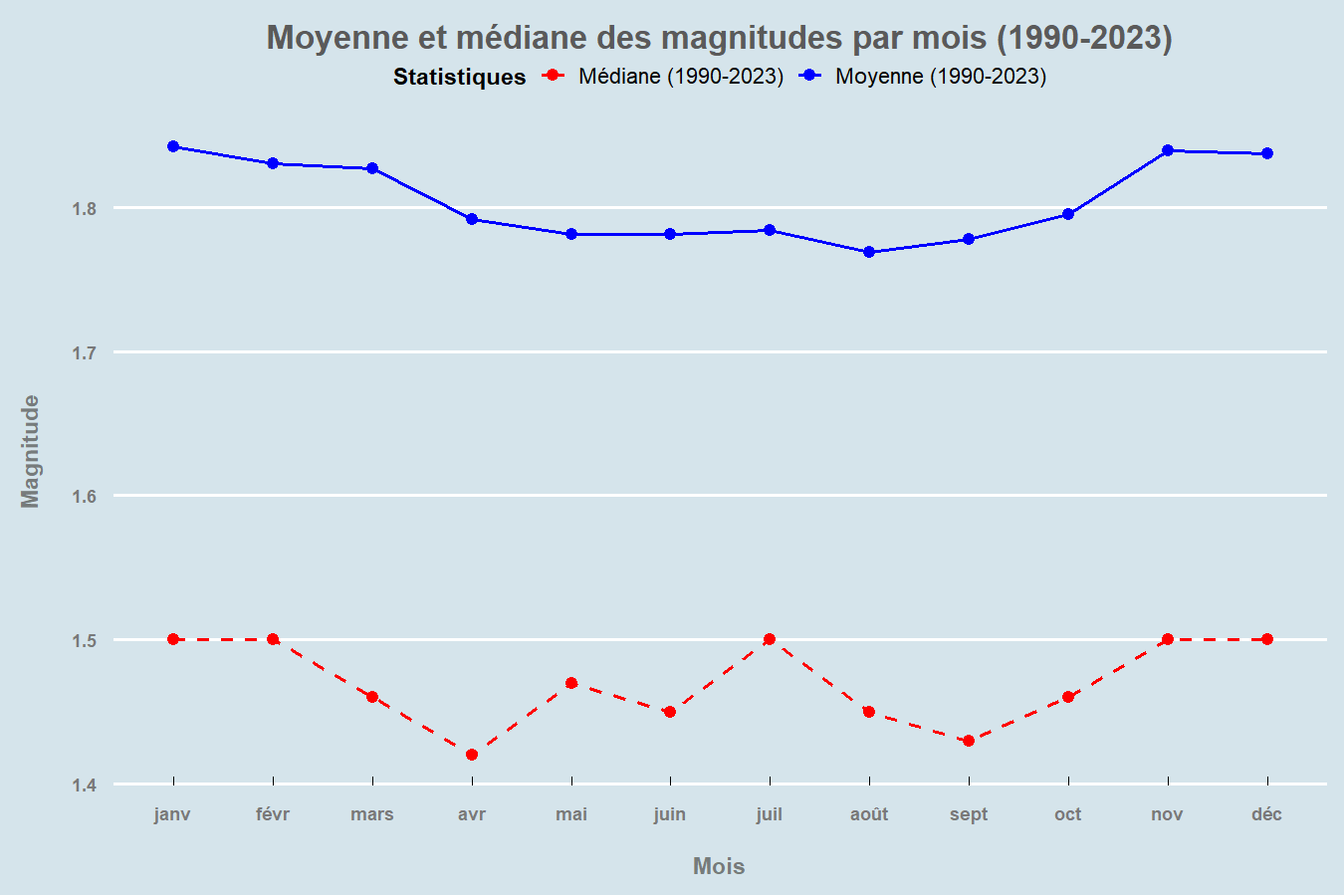
*Figure 10 : Histogramme significativité des séismes*

### Tsunami

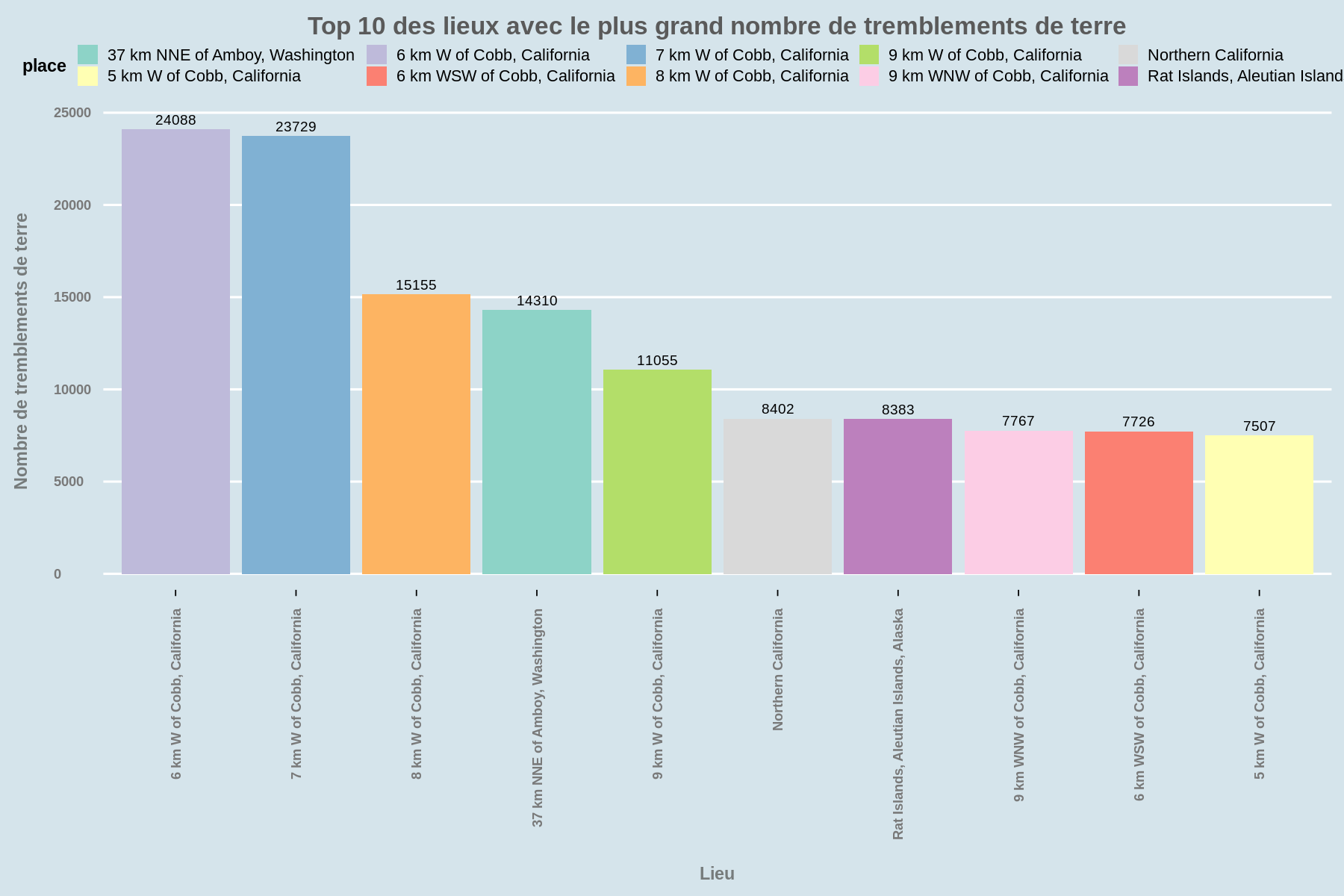


*Figure 11 : Variable Tsunami*

### Moyenne et médiane des magnitudes par mois

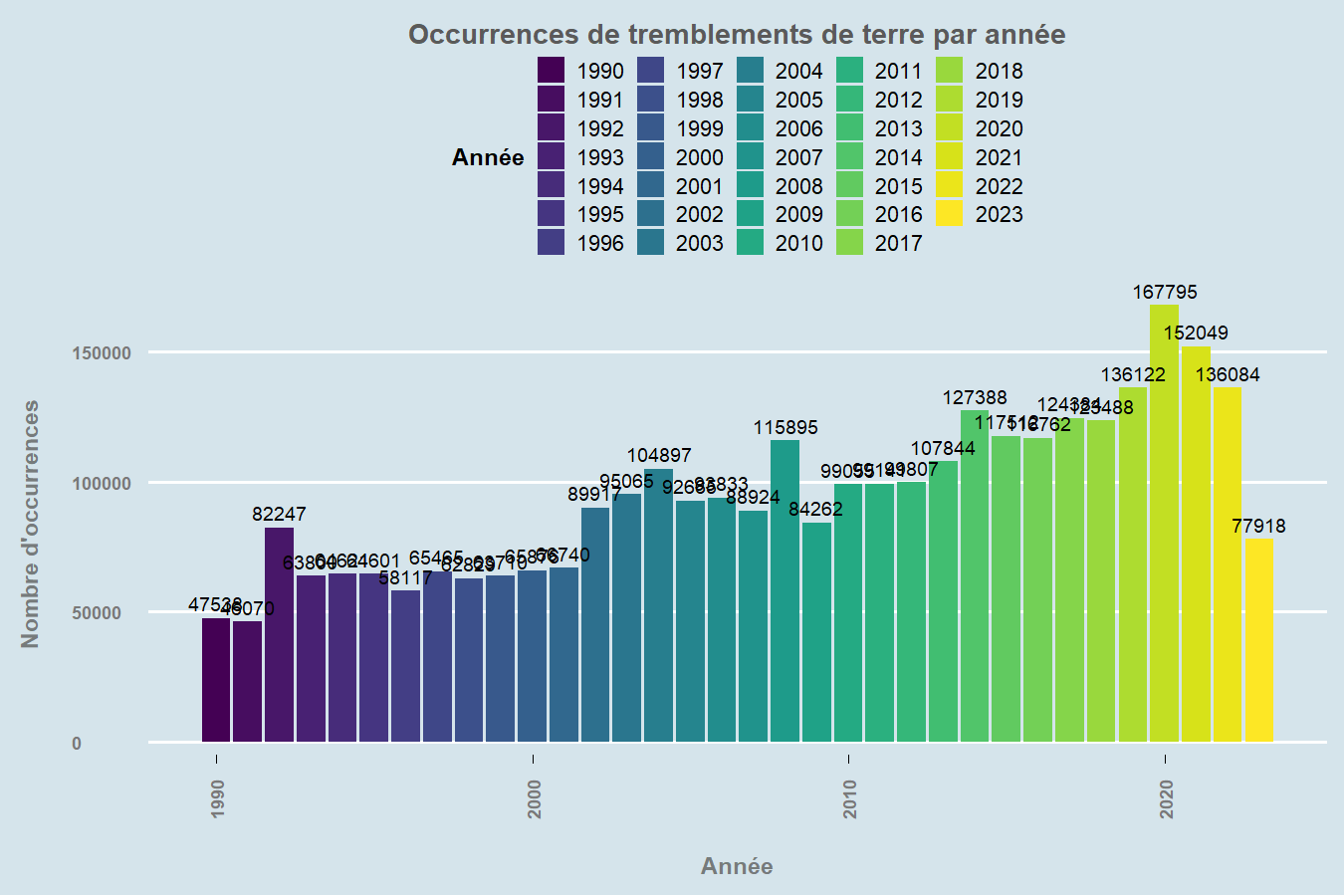
****

### Top 10 des lieux les plus touchés

****

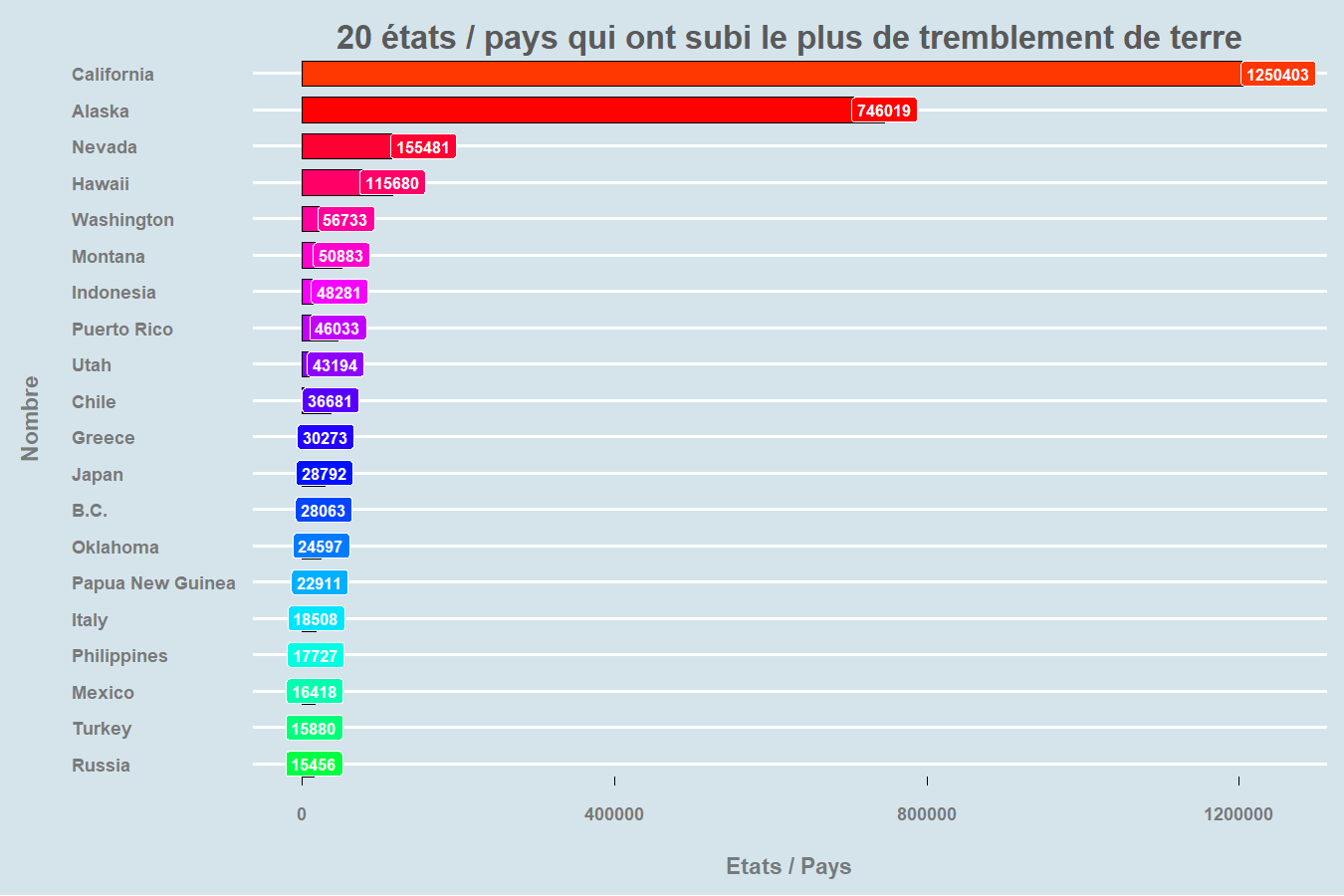
*Figure 12 : Top 10 des lieux les plus touchés*

### Occurrences des séismes par année

****

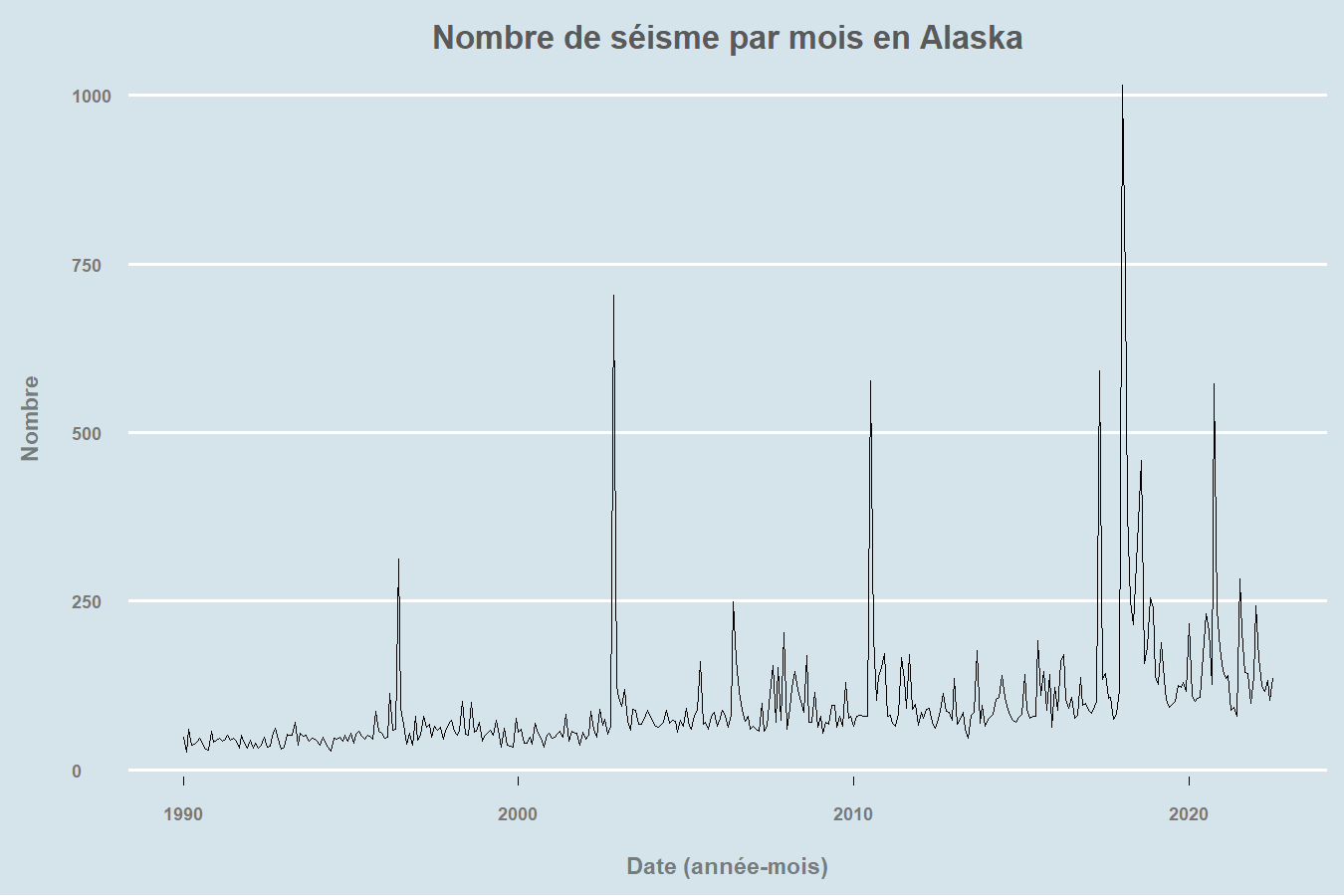
*Figure 13 : Occurrences des séismes par année*

### 20 états/pays les plus touchés



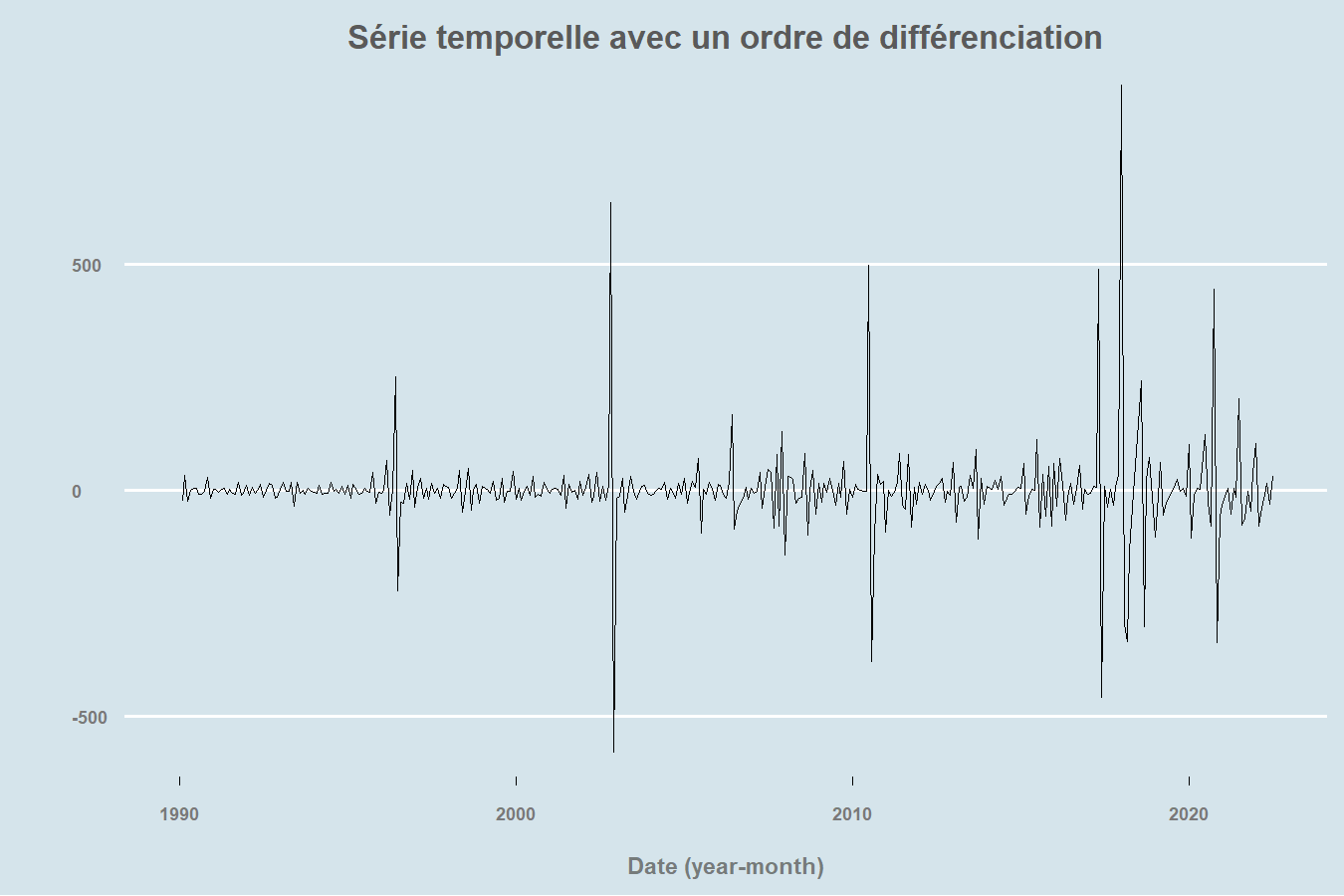
*Figure 14 : 20 états/pays les plus touchés*

### Série brute



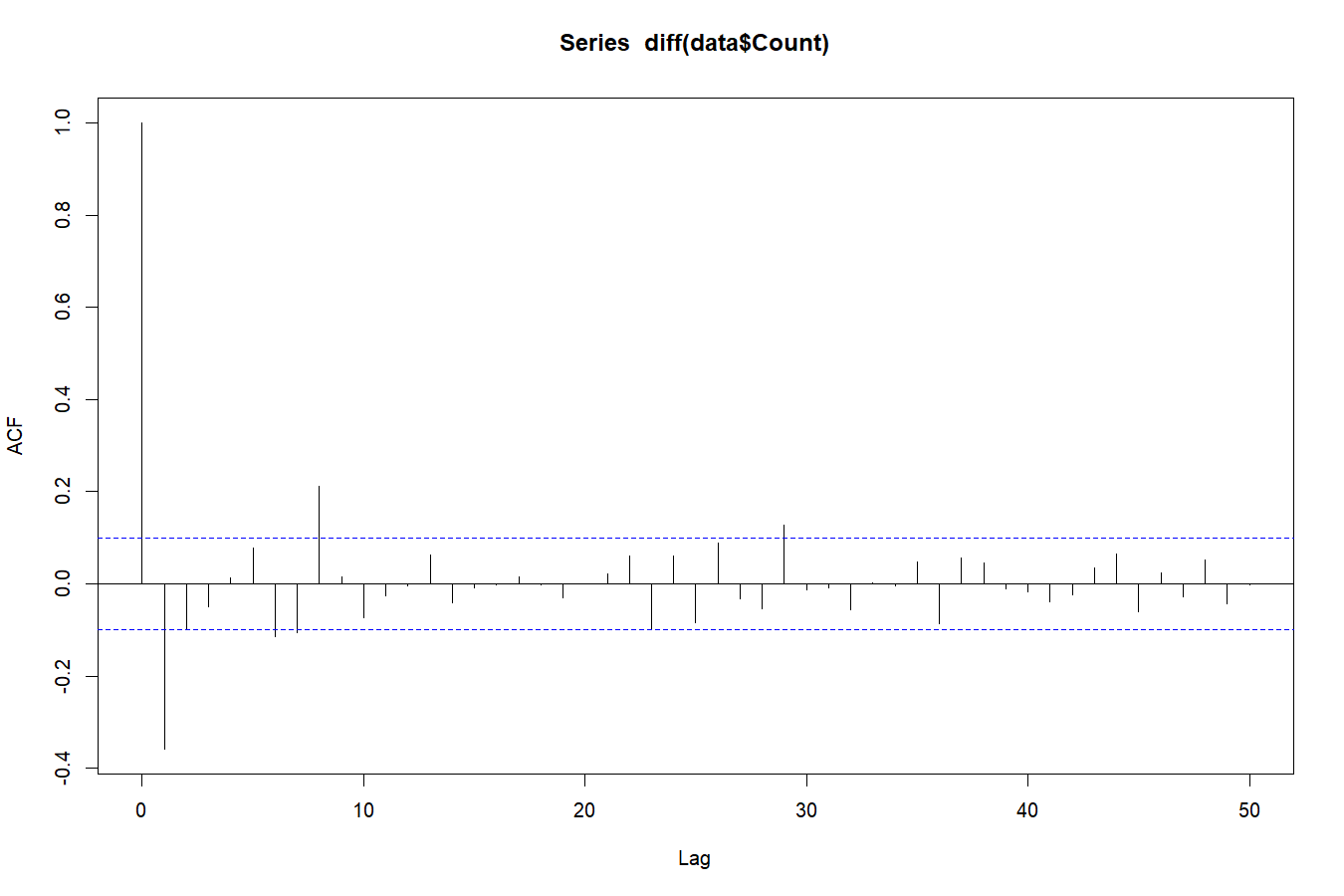
*Figure 15 : Série temporelle brute*

### Série différenciée



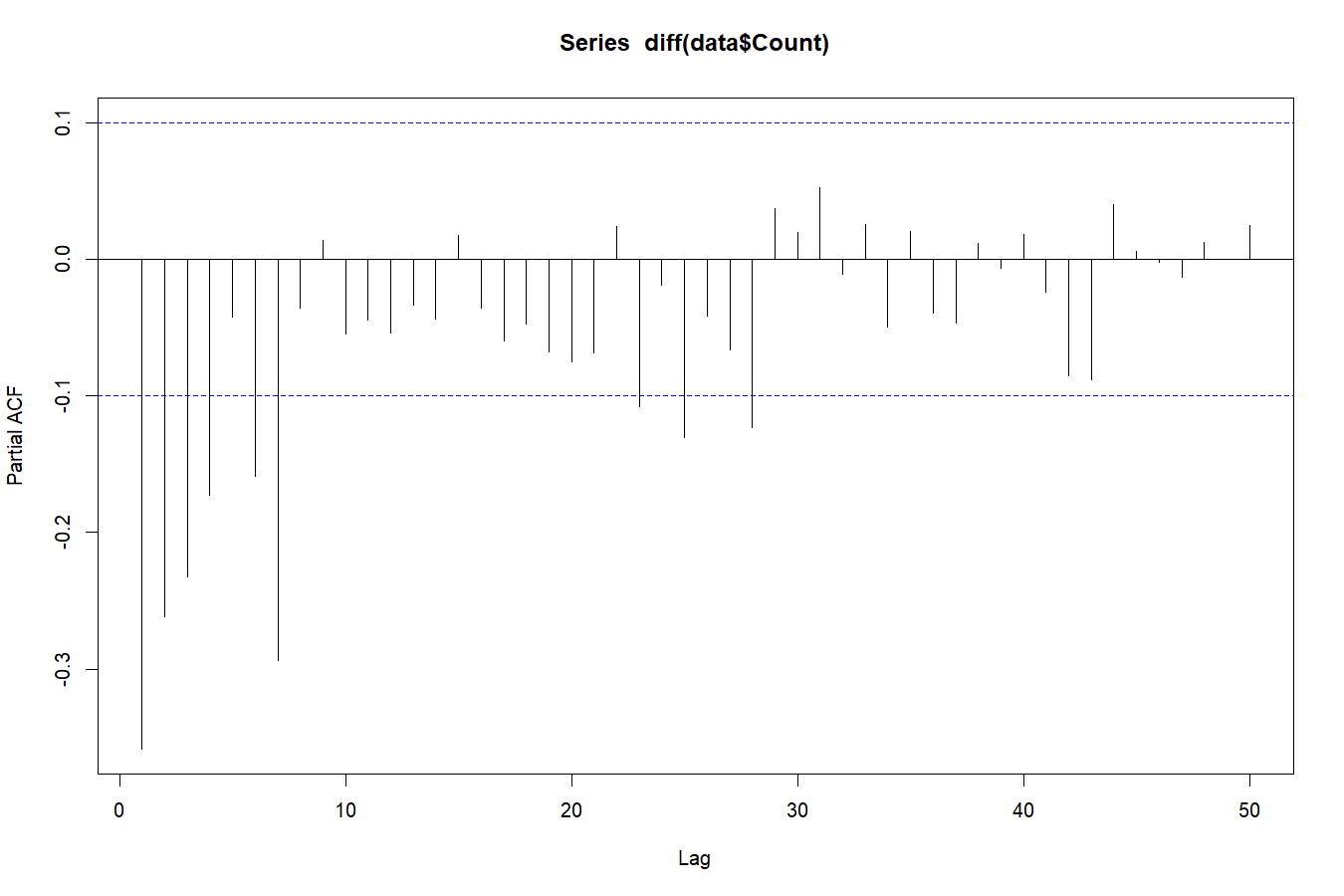
*Figure 16 : Série différenciée*

### ACF



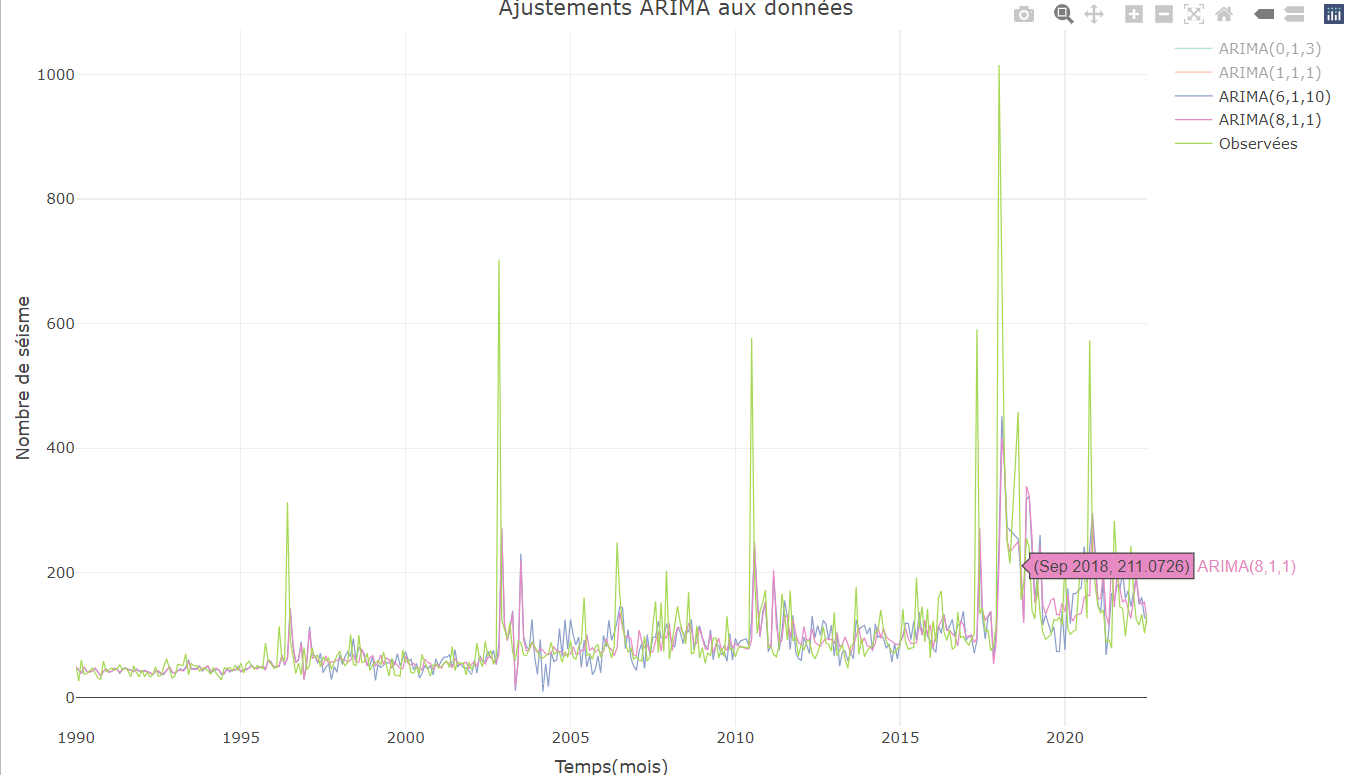
*Figure 17 : ACF*

### PACF



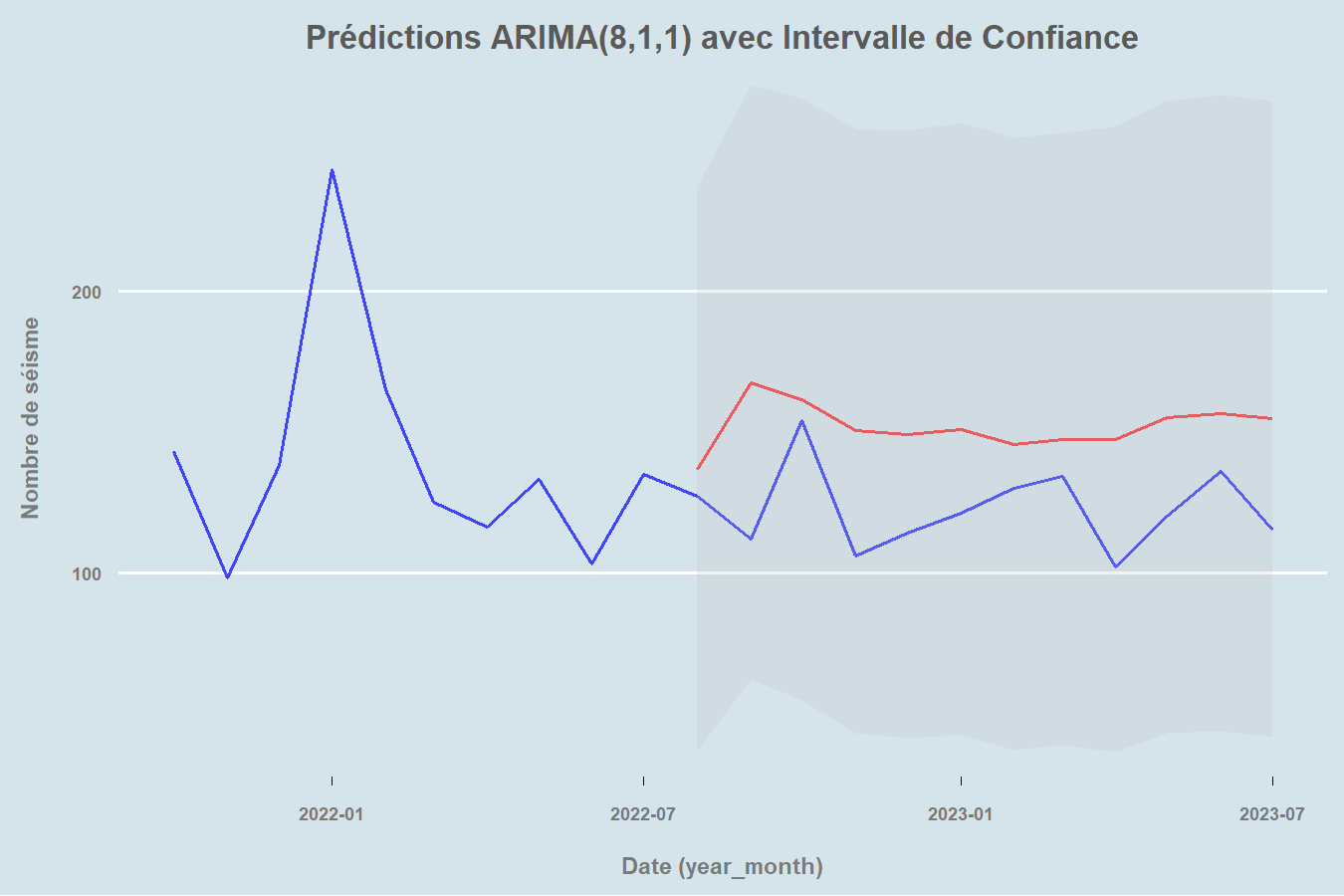
*Figure 18 : PACF*

### Ajustement ARIMA



*Figure 19 : Ajustement ARIMA*

### Prévision ARIMA

**

*Figure 20 : Prévision du modèle sur le jeu de test*

### Métrique Régression

**MSE :**

Le MSE (Mean Square Error) mesure la moyenne des carrés des écarts entre les prédictions du modèle et les valeurs réelles. Le MSE attribue une pondération plus importante aux erreurs importantes en élevant chaque écart au carré, ce qui signifie que les erreurs plus importantes auront un impact disproportionné sur le résultat final. Cette caractéristique en fait une métrique sensible aux valeurs aberrantes, et elle est souvent utilisée dans des contextes où des erreurs importantes doivent être pénalisées de manière significative.

La formule du MSE pour un ensemble de données de taille est la suivante :

Où :

* est la taille de l’ensemble de données
* représentant la valeur réelle de l’observation
* est la valeur prédite par le modèle pour l’observation

Plus le MSE est bas, plus les prédictions du modèle sont considérées comme précises. Cependant, comme le MSE est exprimé en unités au carré de la variable cible, il peut être difficile à interpréter directement. Par conséquent, il est souvent combiné avec d'autres métriques ou transformé en sa racine carrée (Root Mean Squared Error - RMSE) pour obtenir une mesure dans la même unité que la variable cible, facilitant ainsi l'interprétation.

**RMSE :**

La principale différence réside dans le fait que le RMSE (Root Mean Squared Error) prend la racine carrée du MSE, ce qui a pour effet de revenir à l'échelle d'origine de la variable cible, rendant ainsi la métrique plus interprétable. La formule du RMSE pour un ensemble de données de taille n est définie comme suit :

Où :

* est la taille de l’ensemble de données
* représentant la valeur réelle de l’observation
* est la valeur prédite par le modèle pour l’observation

Comme pour le MSE, un RMSE plus bas indique des prédictions plus précises. Cependant, le RMSE a l'avantage d'être dans la même unité que la variable cible, facilitant ainsi l'interprétation directe. Il permet de mesurer l'écart-type des résidus, fournissant ainsi une indication de la dispersion des erreurs du modèle. Une valeur de RMSE plus proche de zéro indique généralement une meilleure adéquation du modèle aux données observées.

**MAE :**

Le MAE (Mean Absolute Error) représente la moyenne des distances absolues entre chaque prédiction du modèle et la valeur réelle correspondante. Il est particulièrement résistant aux valeurs aberrantes, car il ne pondère pas les erreurs en fonction de leur magnitude. La formule du MAE pour un ensemble de données de taille est définie comme suit :

Où :

* est la taille de l’ensemble de données
* représentant la valeur réelle de l’observation
* est la valeur prédite par le modèle pour l’observation

Une valeur de MAE plus basse indique une meilleure adéquation entre les prédictions du modèle et les valeurs réelles. Contrairement au MSE, le MAE est dans la même unité que la variable cible, facilitant ainsi l'interprétation directe de l'erreur moyenne absolue.

**MAPE :**

Le Mape (Mean Absolute Percentage Error) représente la moyenne des pourcentages absolus d'erreur entre chaque prédiction du modèle et la valeur réelle correspondante. Il exprime l'erreur en termes de pourcentage par rapport à la valeur réelle. Le Mape est particulièrement utile pour évaluer la précision d'un modèle lorsque les échelles des variables cibles peuvent varier. La formule du MAE pour un ensemble de données de taille est définie comme suit :

Où :

* est la taille de l’ensemble de données
* représentant la valeur réelle de l’observation
* est la valeur prédite par le modèle pour l’observation

Un Mape plus bas indique une meilleure adéquation entre les prédictions du modèle et les valeurs réelles, exprimée en pourcentage. Une valeur de Mape de 0 signifierait que les prédictions du modèle sont parfaites, tandis qu'une valeur plus élevée indique une erreur plus importante en pourcentage par rapport aux valeurs réelles. Le Mape est utile pour évaluer la précision relative d'un modèle, en particulier dans des domaines où l'échelle des valeurs cibles peut varier significativement.

### Recherche aléatoire

Le processus de la méthode aléatoire en termes d’étapes algorithmiques est le suivant :

1. Définir une grille de recherche d’hyperparamètres à explorer.
2. Sélectionner aléatoirement une combinaison d’hyperparamètres dans la grille de recherche.
3. Diviser l’ensemble de données en k échantillons de taille égale pour la validation croisée en k-fold.
4. Entraîner le modèle sur k-1 échantillons et le valider sur le reste.
5. Répéter les étapes 3 et 4 k fois en utilisant des échantillons de validation différents à chaque fois.
6. Calculer l’erreur moyenne de validation pour la combinaison d’hyperparamètres.
7. Répéter les étapes 2 à 6 un nombre prédéfini de fois en sélectionnant à chaque fois une nouvelle combinaison d’hyperparamètres aléatoire.
8. Sélectionner la combinaison d’hyperparamètres qui donne la plus petite erreur moyenne de validation.

### Méthode bayésienne

Afin de comprendre ce processus, il est essentiel d'appréhender les éléments spécifiques de la fonction de sélection.

**Fonction de sélection :** Elle est le critère par lequel le prochain ensemble d’hyperparamètres est choisi à partir de la fonction de substitution. Le choix de critère le plus courant est l’amélioration attendue (expected improvement).

Où :

* est la valeur réelle de la fonction objectif utilisant les hyperparamètres
* est une valeur seuil définit par la variable qui représente le quantile des scores de précision négatifs (observés jusqu’à présent). Le seuil est fixé à 15% par défaut.
* est l’ensemble d’hyperparamètres proposé
* est le modèle de probabilité de substitution exprimant la probabilité de donné .

Si l’intégrale est positive, cela signifie que les hyperparamètres devraient donner un meilleur résultat que la valeur seuil.

Au lieu de représenter directement , on utilise à la place la formule suivante :

est la probabilité des hyperparamètres compte tenu du score sur la fonction objective et s’exprime par :

Où :

* est la densité formée en utilisant les observations des évaluations passées, de sorte que la perte correspondante (métrique : MAPE) est inférieur à un certain seuil
* est la densité formée en utilisant les observations restantes.

De ce fait sous l’algorithme TPE, on peut montrer que maximiser EI revient à choisir des valeurs de qui minimisent le rapport i.e. que nous voudrions des points avec des valeurs élevées sous et des valeurs faibles sous . A chaque itération, l’algorithme tire plusieurs échantillons de , les évalue en termes de et renvoie le candidat avec le plus haut .

Ci-dessous une démonstration de la formule en fonction de et .

En posant :



On a alors :

On pose :

Ainsi, on a :

Le terme à l’extrême droite est la partie la plus importante. C’est là que l’on voit que est proportionnelle au rapport et qu’il faut donc minimiser ce rapport. Et donc nous obtiendrons des valeurs de paramètres plus susceptibles sous que sous .

### Méthode génétique

1. Initialisation : Une population initiale de solutions est créée. Chaque solution représente une combinaison d'hyperparamètres pour le modèle d'apprentissage automatique. La population 0 échantillonnée au hasard est générée.
2. Evaluation : on évalue la valeur MAPE de chaque individu de la population selon le modèle d’apprentissage automatique utilisé et obtenir les scores de validation croisée.
3. Sélection : L’idée de la sélection est de choisir les individus les plus aptes de la population et de les approuver pour la prochaine génération suivante. Les individus ayant un MAPE élevé ont plus de chances d’être sélectionnés. Cela favorise la propagation des bonnes solutions dans la population.
4. Reproduction : Les solutions sélectionnées sont combinées pour créer de nouvelles solutions, appelées enfants. Cela peut se faire en utilisant des opérations telles que la recombinaison (crossover) et la mutation. La recombinaison mélange les hyperparamètres des solutions parentales, tandis que la mutation modifie aléatoirement certains hyperparamètres.
5. Remplacement : Les nouveaux enfants remplacent les solutions les moins performantes de la population d'origine. Cela permet de maintenir la diversité génétique tout en favorisant les solutions prometteuses.
6. Répétition : Les étapes 2 à 5 sont répétées pendant un certain nombre d'itérations ou jusqu'à ce qu'un critère d'arrêt prédéfini soit atteint. Les critères d'arrêt courants sont le nombre maximal d'itérations.
7. Solution : L’algorithme se termine si la population a convergé. À la fin, la meilleure solution trouvée dans la population finale est renvoyée comme combinaison d'hyperparamètres optimisée.

### Optimiseur Adam

L'optimiseur Adam est utilisé pour ajuster les poids et les biais lors de l'entraînement d'un modèle, y compris pour la couche LSTM. Plus précisément, Adam s'occupe de la mise à jour des paramètres du modèle (poids et biais) de manière itérative afin de minimiser la fonction de perte définie lors de la compilation du modèle. Ci-dessous quelques étapes sont présentées sur le fonctionnement de l’optimiseur Adam :

* **Calcul du gradient :** Adam calcule le gradient de la fonction de perte par rapport à chaque paramètre du modèle. Ce gradient représente la direction dans laquelle chaque paramètre doit être ajusté pour réduire la perte.
* **Calcul des Moments du Premier Ordre (m) et du Deuxième Ordre (v) :** Adam maintient deux moments pour chaque paramètre. Le premier moment (m) est une moyenne mobile exponentielle des gradients, et le deuxième moment (v) est une moyenne mobile exponentielle des carrés des gradients. Ces moments sont utilisés pour adapter le taux d'apprentissage de chaque paramètre. D'un point de vue mathématique, la formulation serait la suivante :

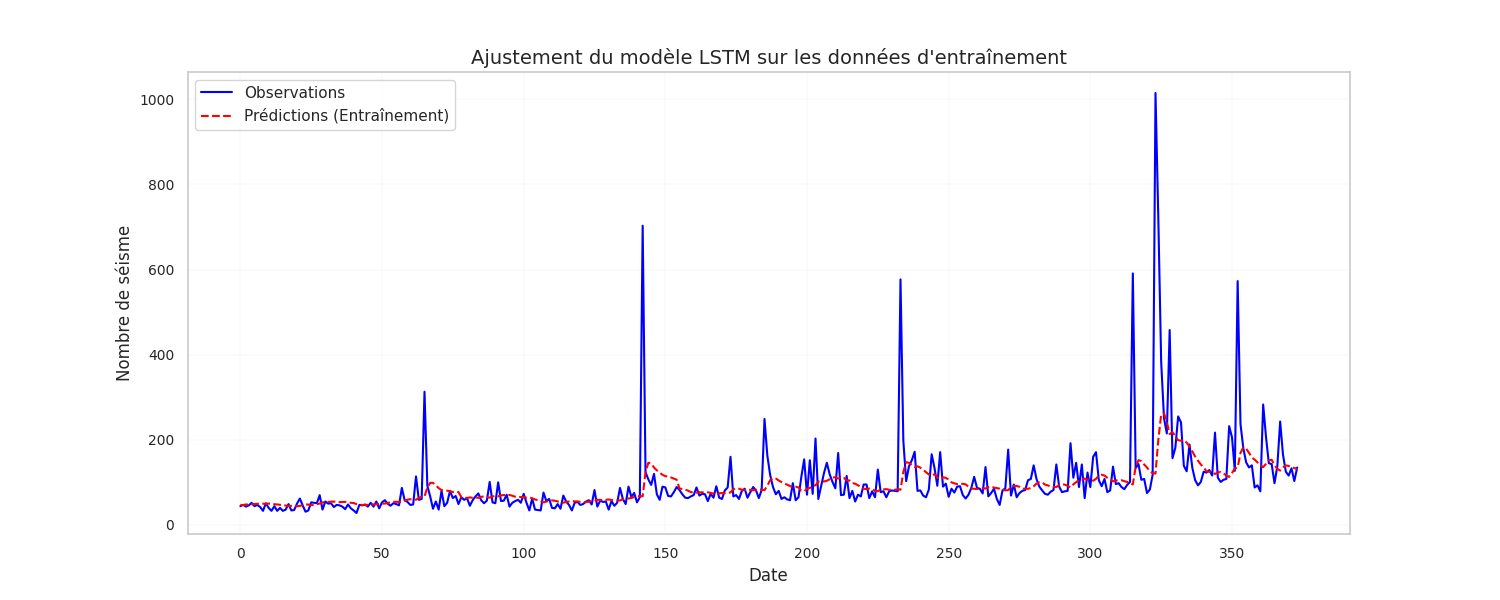
où est le moment du premier ordre au temps ; est le coefficient d'oubli pour le premier ordre (habituellement fixé à 0.9) ; est le moment du premier ordre au temps précédent, et est le gradient au temps .

où est le moment du deuxième ordre au temps ; est le coefficient d'oubli pour le deuxième ordre (habituellement fixé à 0.999) ; est le moment du deuxième ordre au temps précédent, et est le carré du gradient au temps .

* **Correction des Biais (biais-correctif) :** Adam effectue une correction des biais pour compenser les biais introduits par la moyenne mobile exponentielle. Cela permet d'ajuster les moments pour les premières itérations de l'entraînement.
* Mise à Jour des Paramètres : Les paramètres du modèle (poids et biais) sont mis à jour en utilisant les moments calculés et le taux d'apprentissage. Cette mise à jour vise à minimiser la fonction de perte.

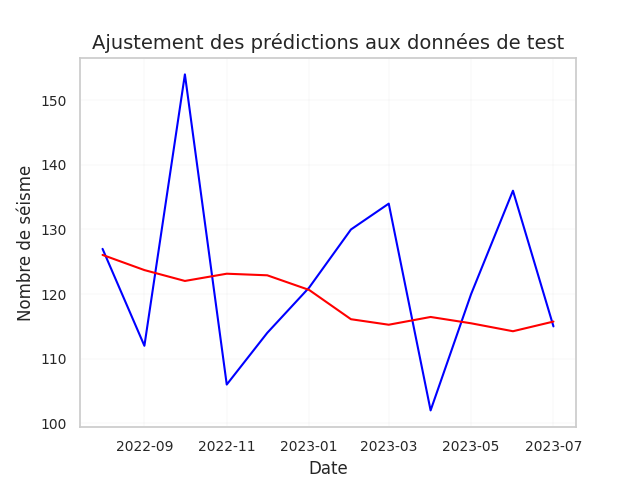
Où est le taux d’apprentissage (ici fixé à 0.001), et sont les moments corrigés, et est une petite constante pour éviter la division par zéro (fixé à 1e-07)

### Ajustement LSTM



*Figure 21 : Ajustement sur les données d’entrainements*

L'analyse visuelle du graphique révèle de manière évidente l'adaptation de notre architecture LSTM aux données d'entraînement. Ce dernier semble étroitement suivre les tendances générales des données, ce qui constitue un indicateur très positif de la performance du modèle. La capacité du modèle à s'ajuster de manière précise aux schémas et variations présents dans l'ensemble d'entraînement renforce la confiance dans sa capacité à saisir les structures sous-jacentes des données temporelles. Cette observation visuelle offre un aperçu encourageant quant à la qualité de l'apprentissage réalisé par notre modèle, suggérant qu'il capture de manière effective les motifs et les comportements inhérents aux séquences temporelles étudiées.



*Figure 22 : Ajustement sur les données de test*

De manière similaire, nous avons entrepris l'ajustement de notre modèle sur les données de test, afin d'évaluer sa capacité à généraliser aux séquences temporelles inconnues. Les résultats de cette étape révèlent une excellente adéquation entre notre architecture LSTM et la tendance globale de la série temporelle de test. Cependant, une observation minutieuse révèle une caractéristique notable : le modèle semble adopter une approche plus conservatrice, présentant un ajustement presque linéaire avec une volatilité relativement faible.