ENSAI

École nationale
de la statistique
et de l'analyse
de l'information

# Projet Série Temporelle
# Prévision du nombre de séisme en Alaska

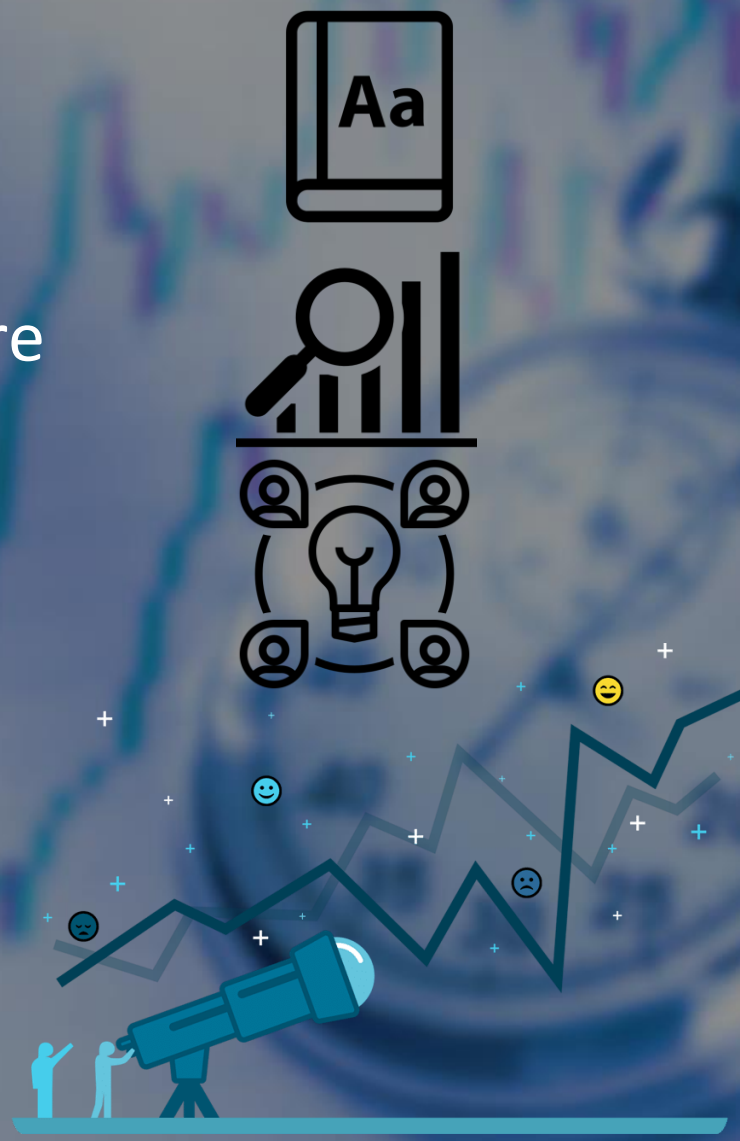
***Davyd BAYARD, Guillaume POIRIER***

***Enseignant : ESSATFA Youssef***

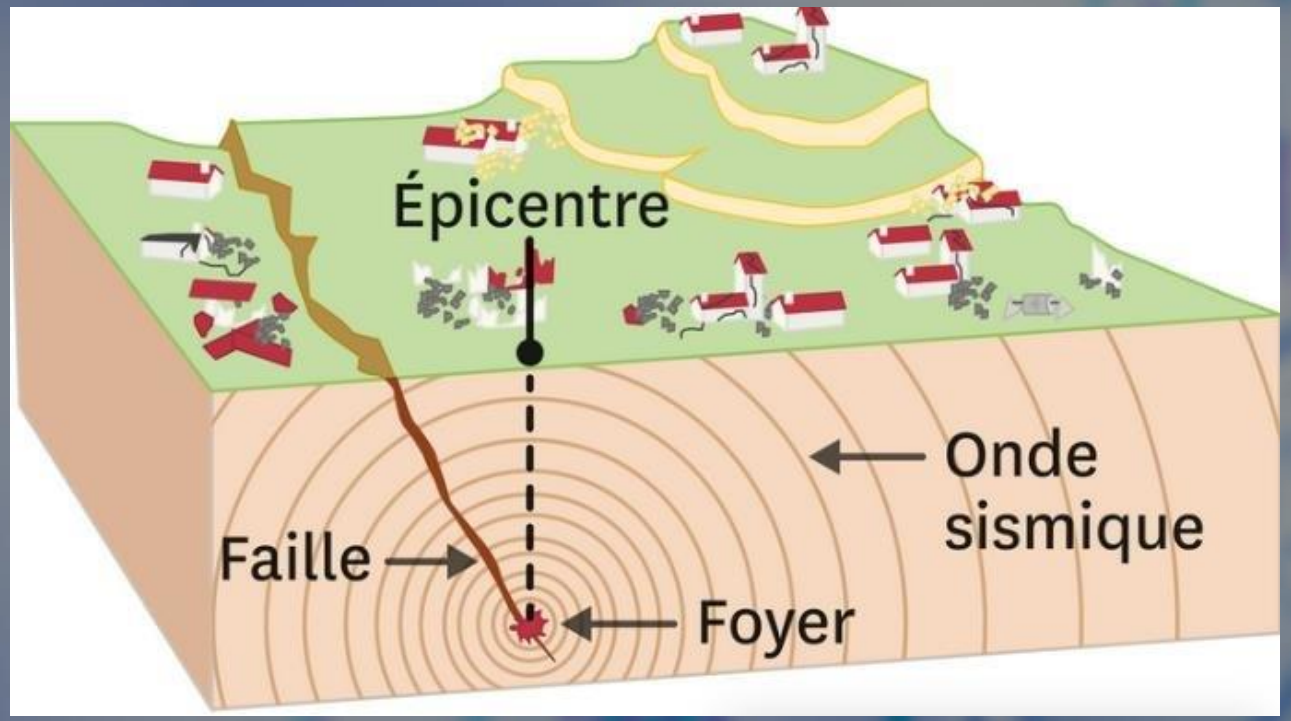
***2023/2024***

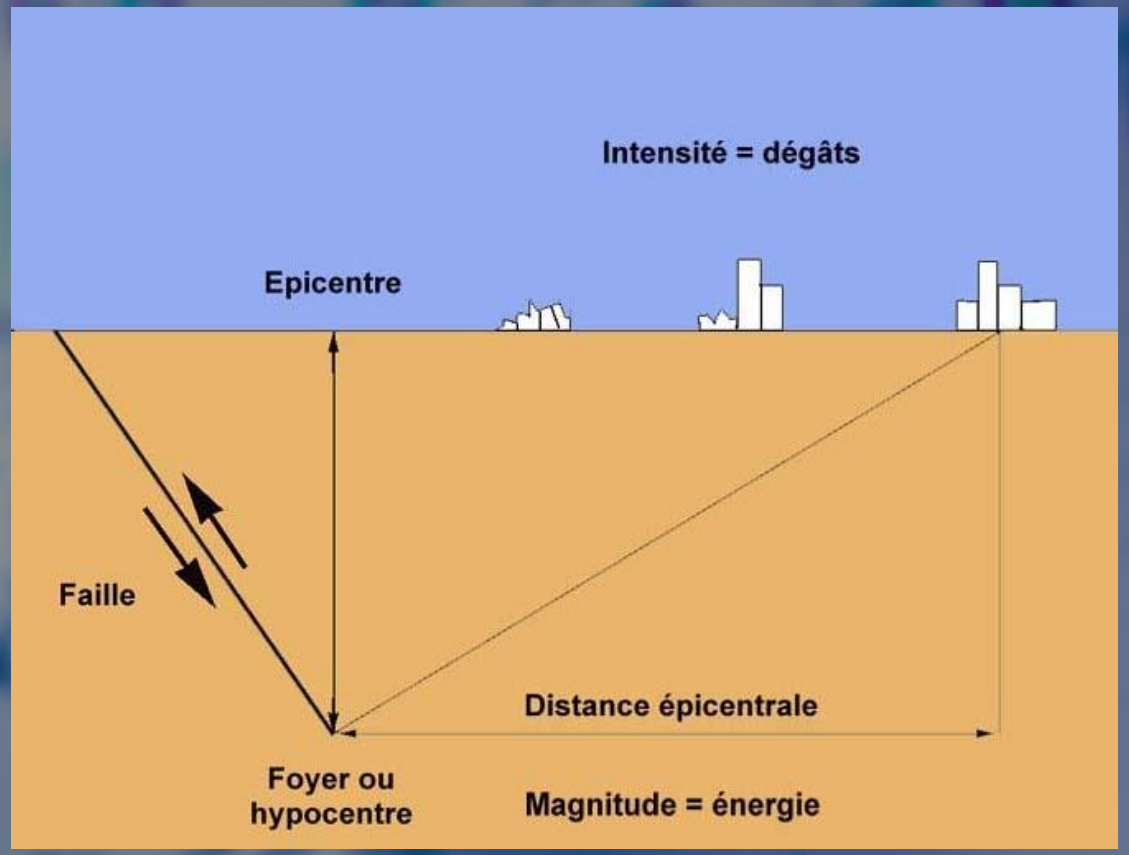
# Sommaire:





# Les séismes


Épicentre
Onde sismique
Faille
Foyer


Intensité = dégâts
Epicentre
Faille
Distance épicentrale
Foyer ou hypocentre
Magnitude = énergie

# I - Data Description :

| 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'évènement |

**Rapide synthèse :**

Nombre d'enregistrements (lignes) : 3445751
Nombre de caractéristiques (colonnes) : 12
Nombre d'entrées dupliquées : 16869 (a supprimer)
Nombre de valeurs manquantes : 0
Nombre de valeurs mag > 10 : 0

Valeurs de profondeur négatives
Anomalie magnitude négatives (-5 et -9,99)

**NB Modalité par variable**

| Variable | Nb modalité |
|---|---|
| place | 531130 |
| status | 6 |
| data_type | 25 |
| state | 858 |

```
table(data$status)

automatic AUTOMATIC    manual    MANUAL  reviewed  REVIEWED
   204715      1100         8         4   3209021     14034
```

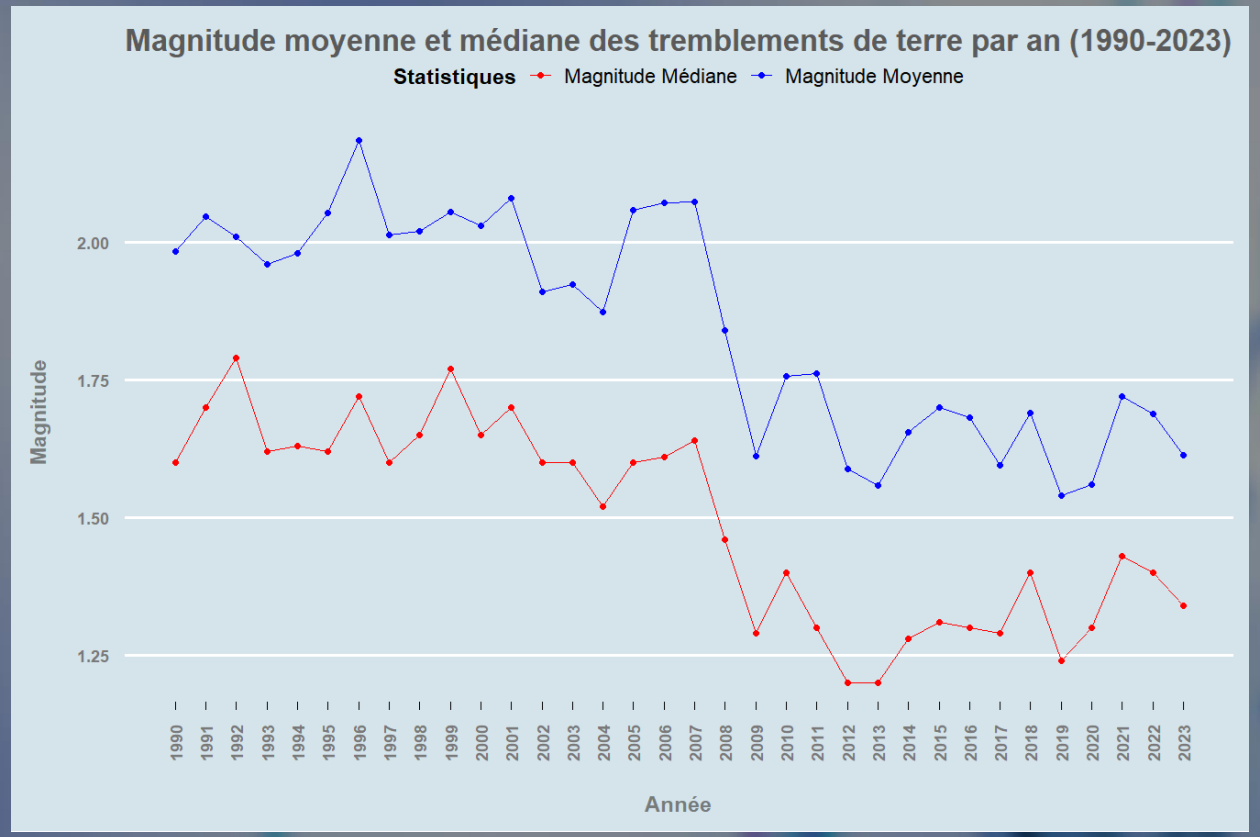
```
[ ]  table(data$status)

    AUTOMATIC    MANUAL  REVIEWED
       205815        12   3223055
```

# II – Analyse Exploratoire

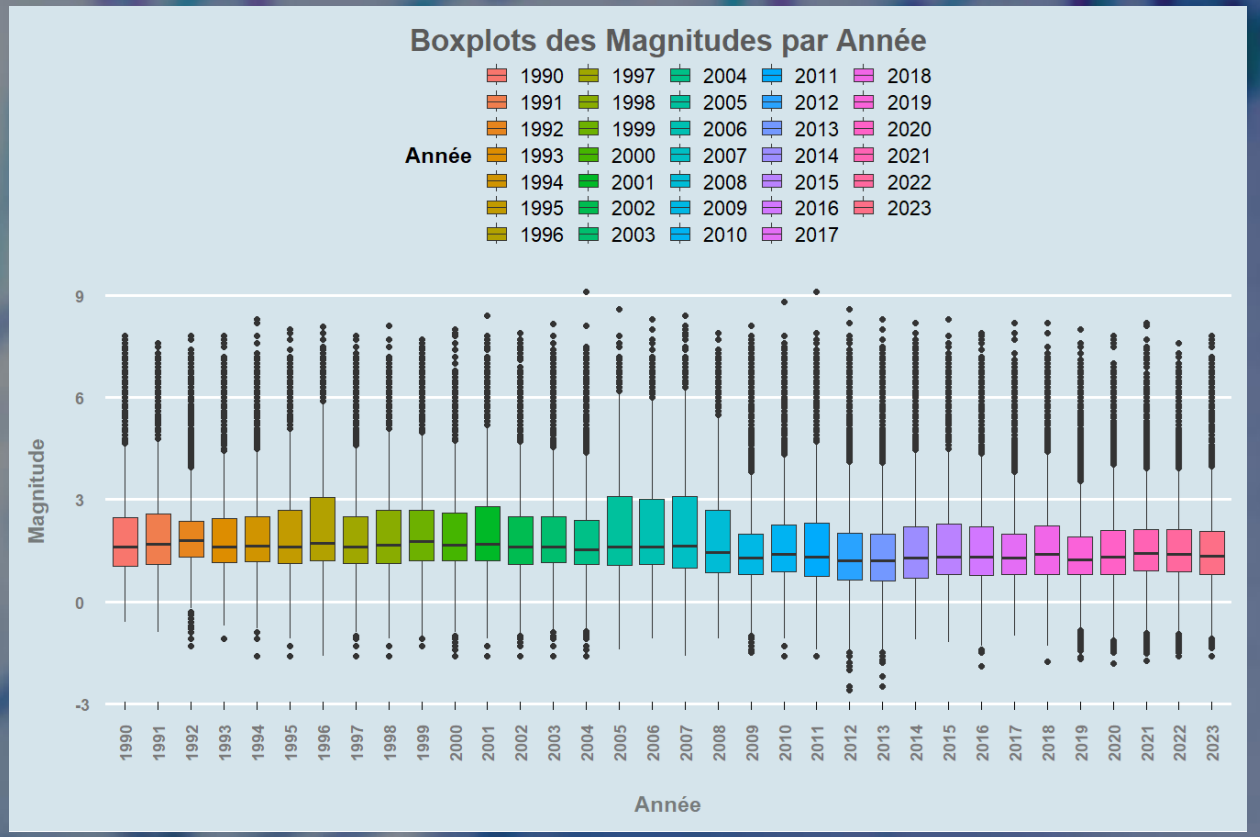
## a) Magnitude

### Moyenne et médiane par année




- 2 groupes : avant 2007 et après 2009

### Boxplot

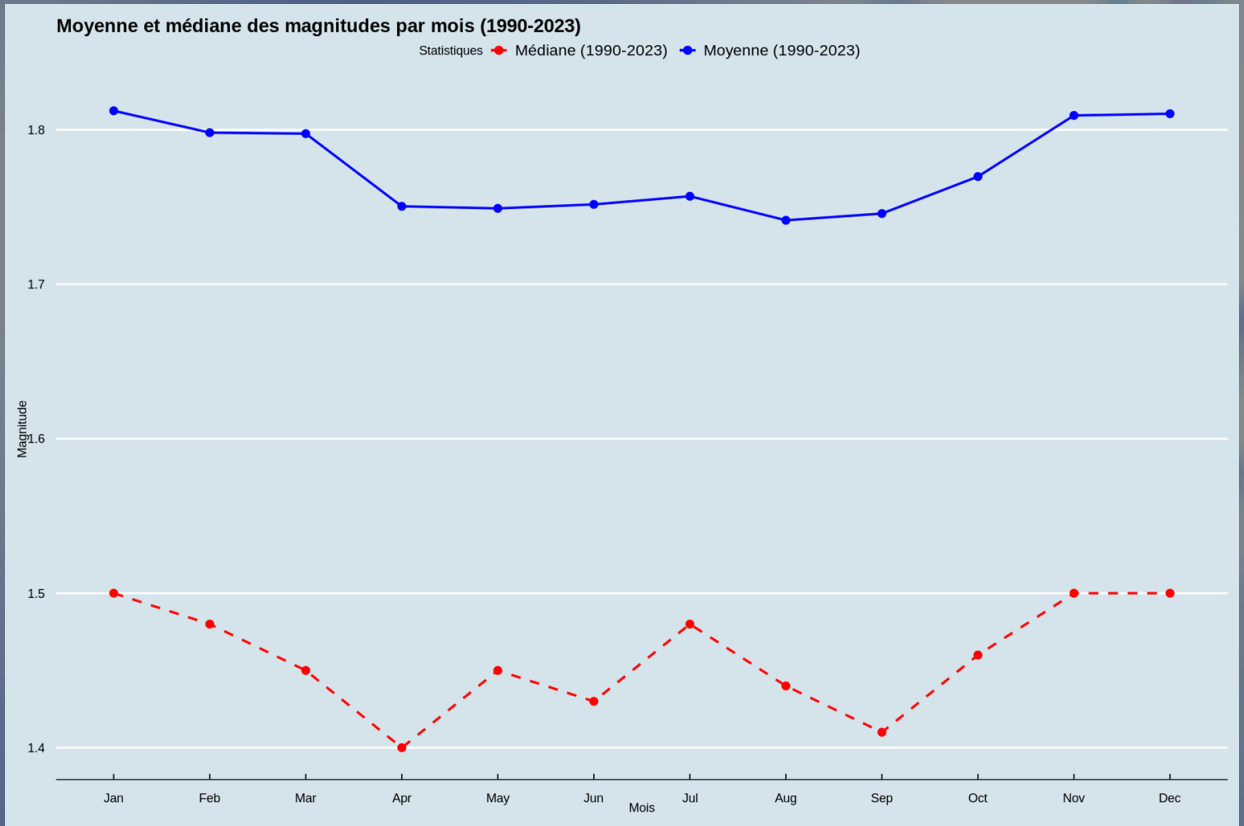



- Présence de séismes puissants
- Tendance stable au fur et à mesure des années

# II – Analyse Exploratoire

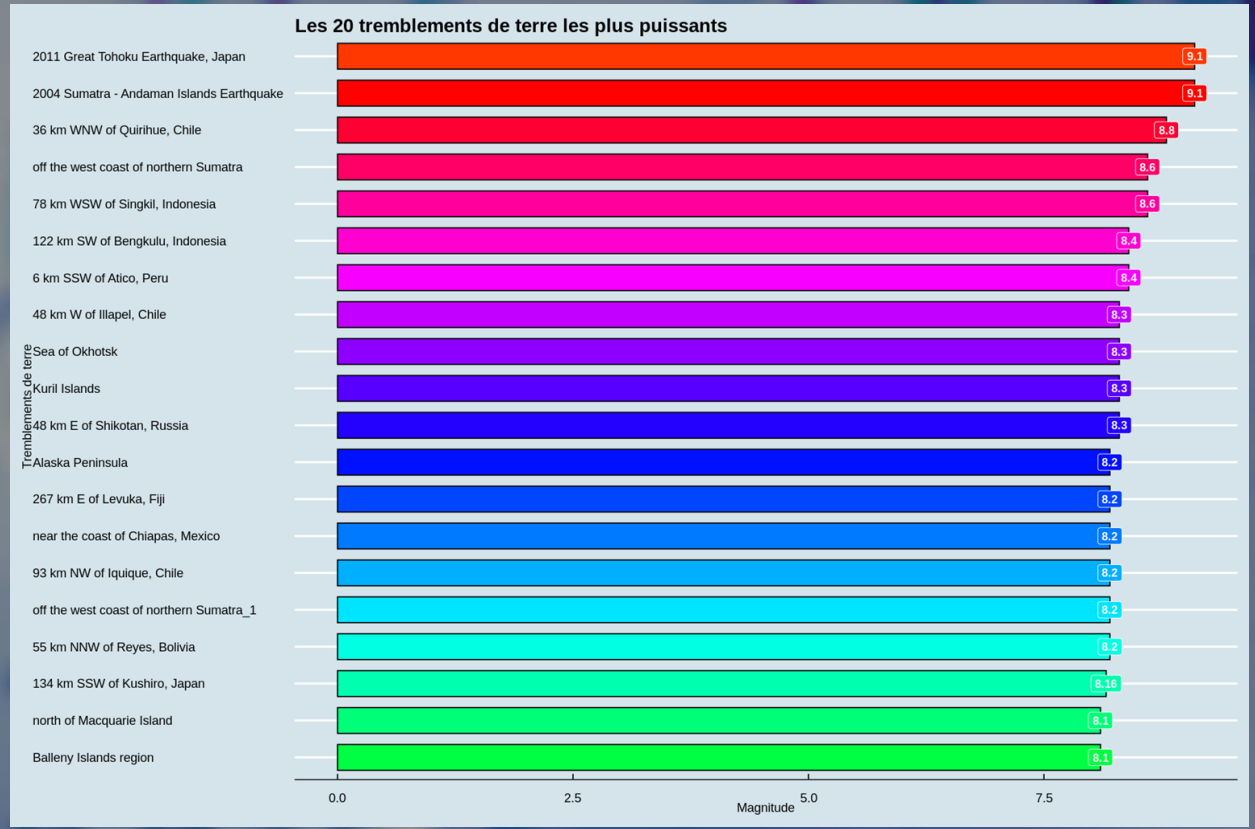
## a) Magnitude

### Magnitude par Mois




- Janvier, Février, Novembre et Décembre : mag puissants
- Avril et Septembre : mag plus faible

### Top 20 des séismes

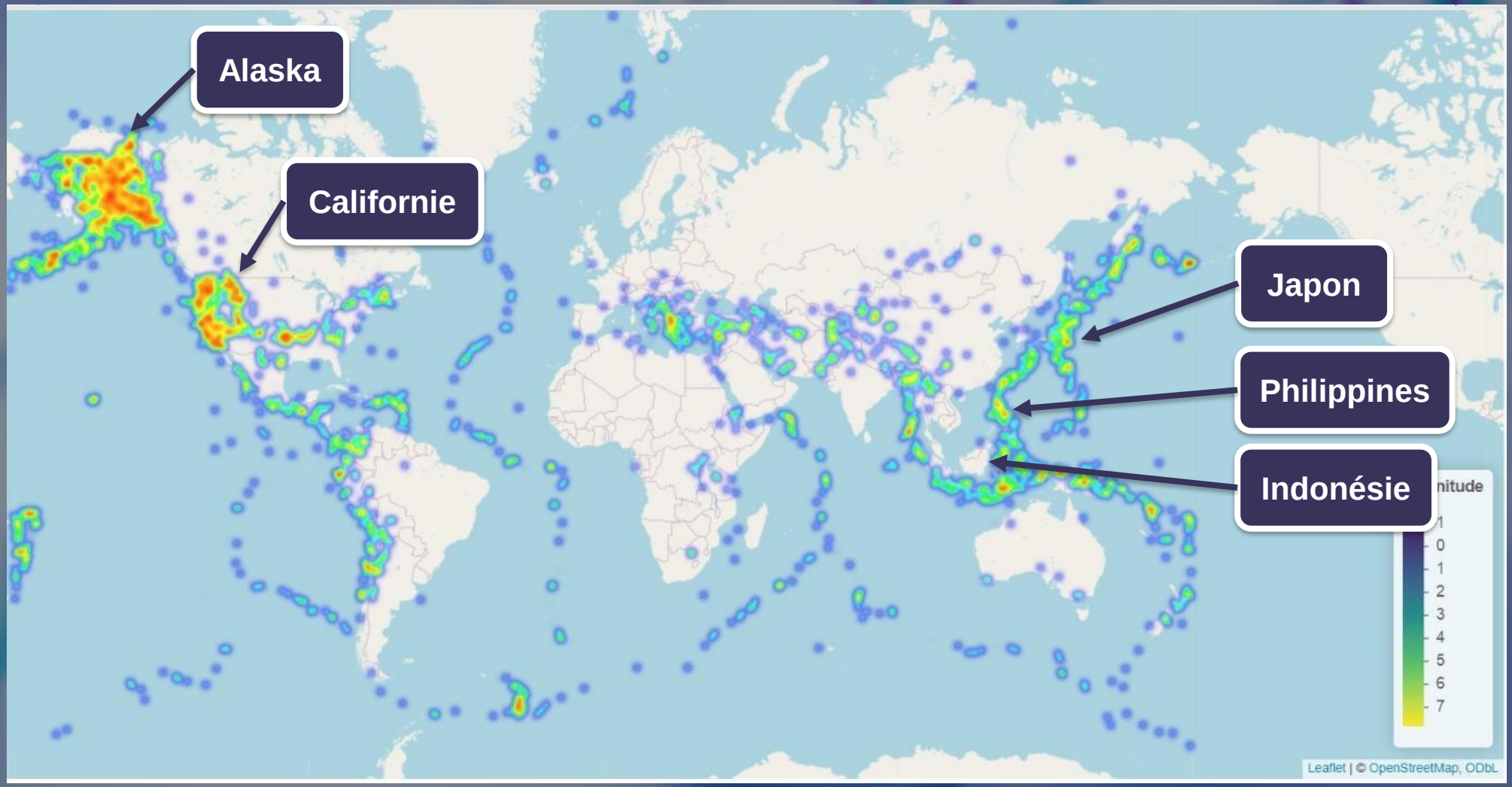



- Pays : Alaska, Japon, Chili, Indonésie, Pérou et Russie.
- Localisation : île et bordure de mer (mer d'Okhotsk)

Mag : magnitude

# II – Analyse Exploratoire

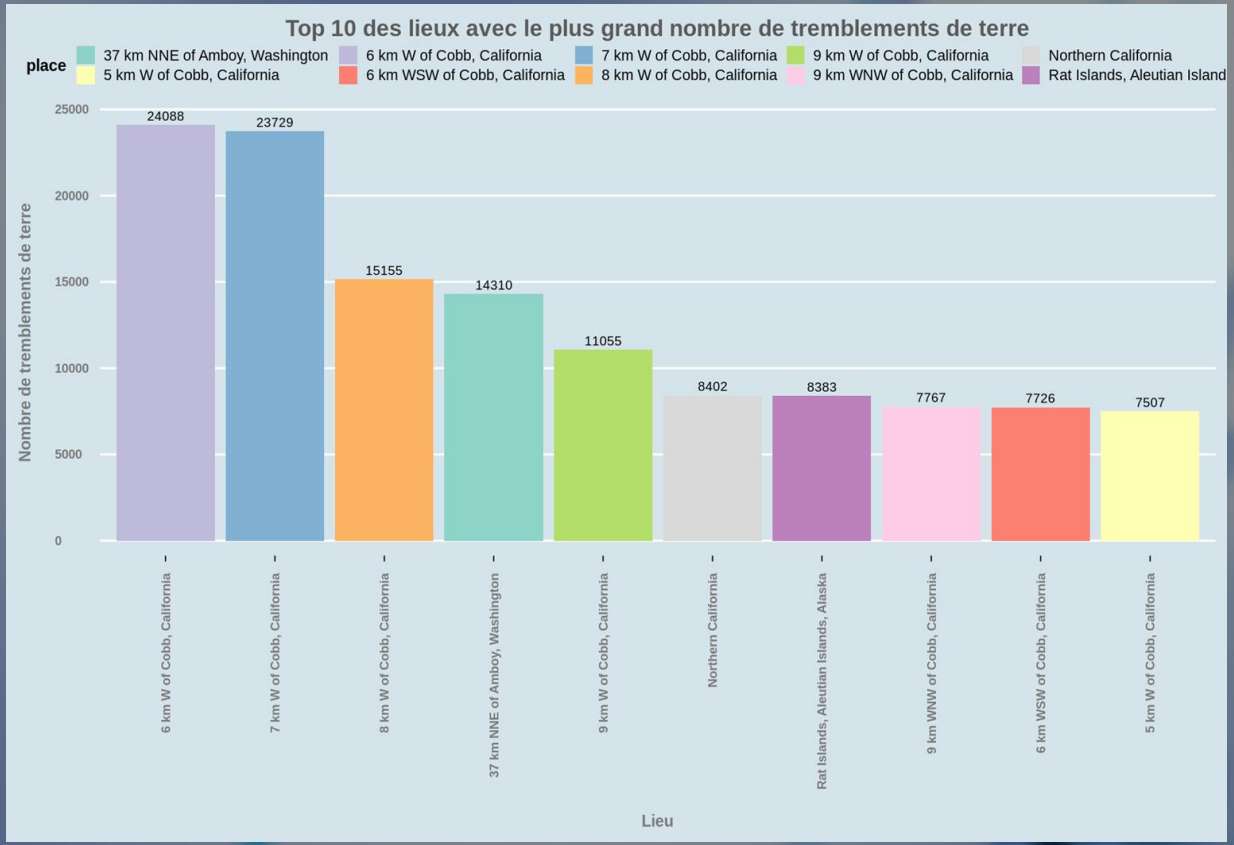
## b) Carte des séismes selon leur magnitude



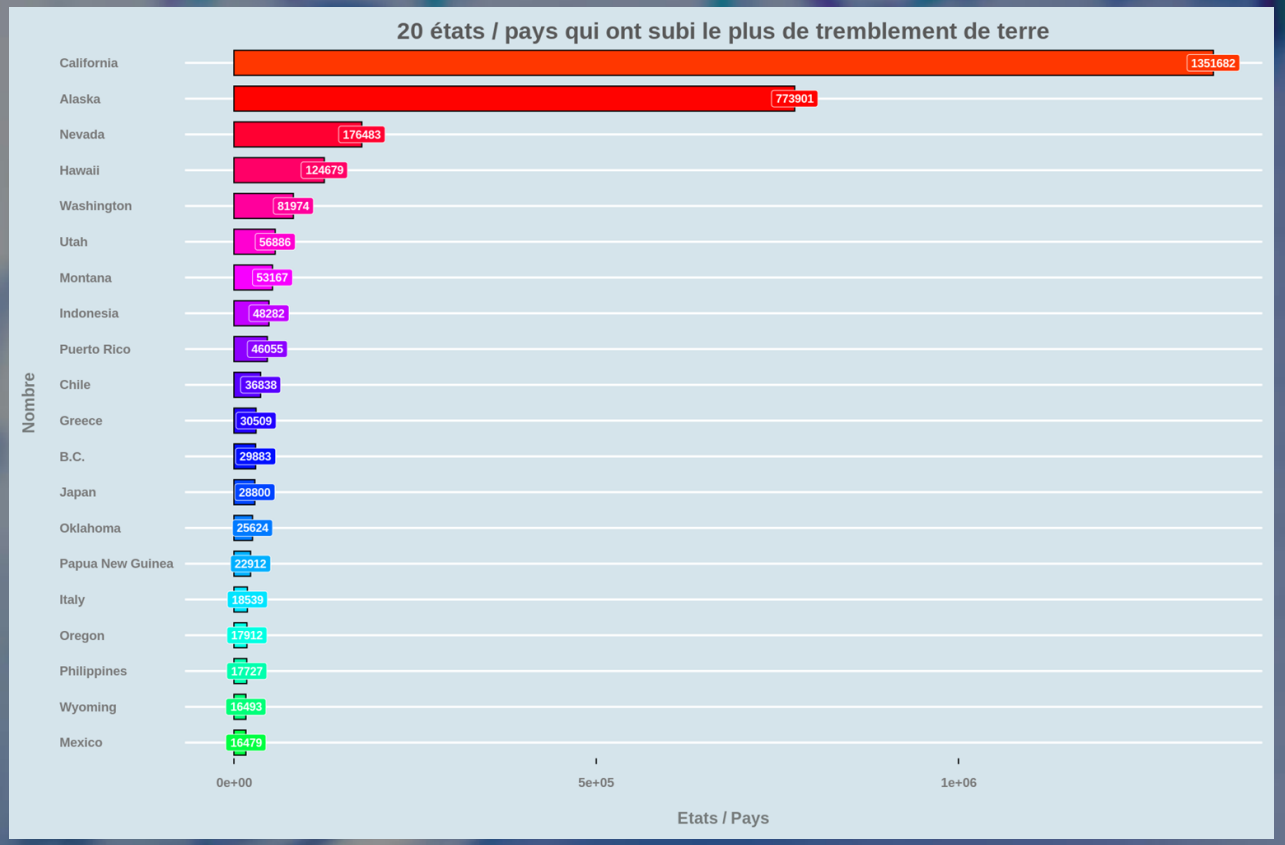
# II – Analyse Exploratoire

## c) Nb de tremblement de terre

### Top 10 des lieux les plus exposés



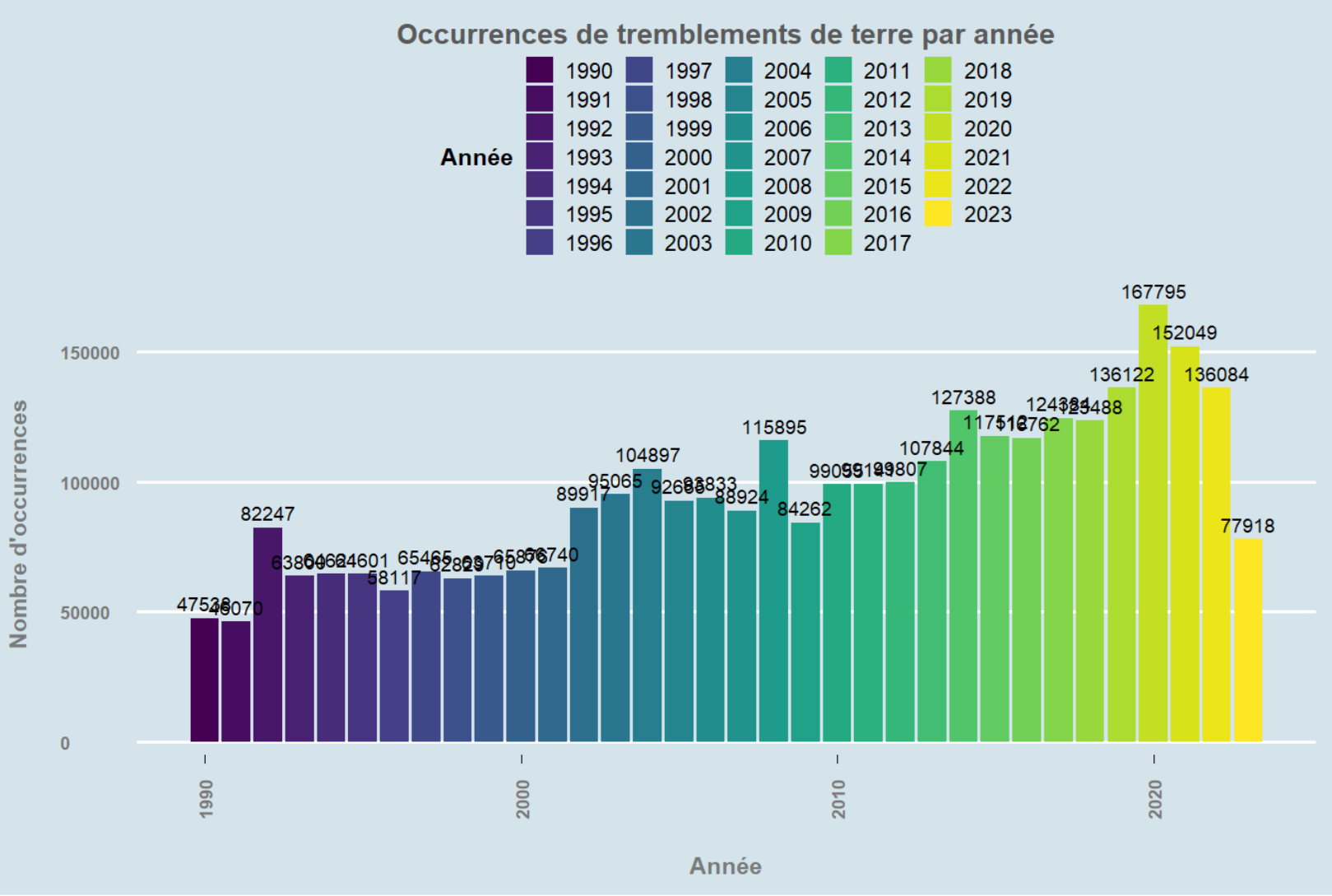

### Top 20 états / pays les plus exposés




- Exposition sur les états américains : Californie, Alaska et Washington
- 2 états américains très affectés : Californie et Alaska

# II – Analyse Exploratoire

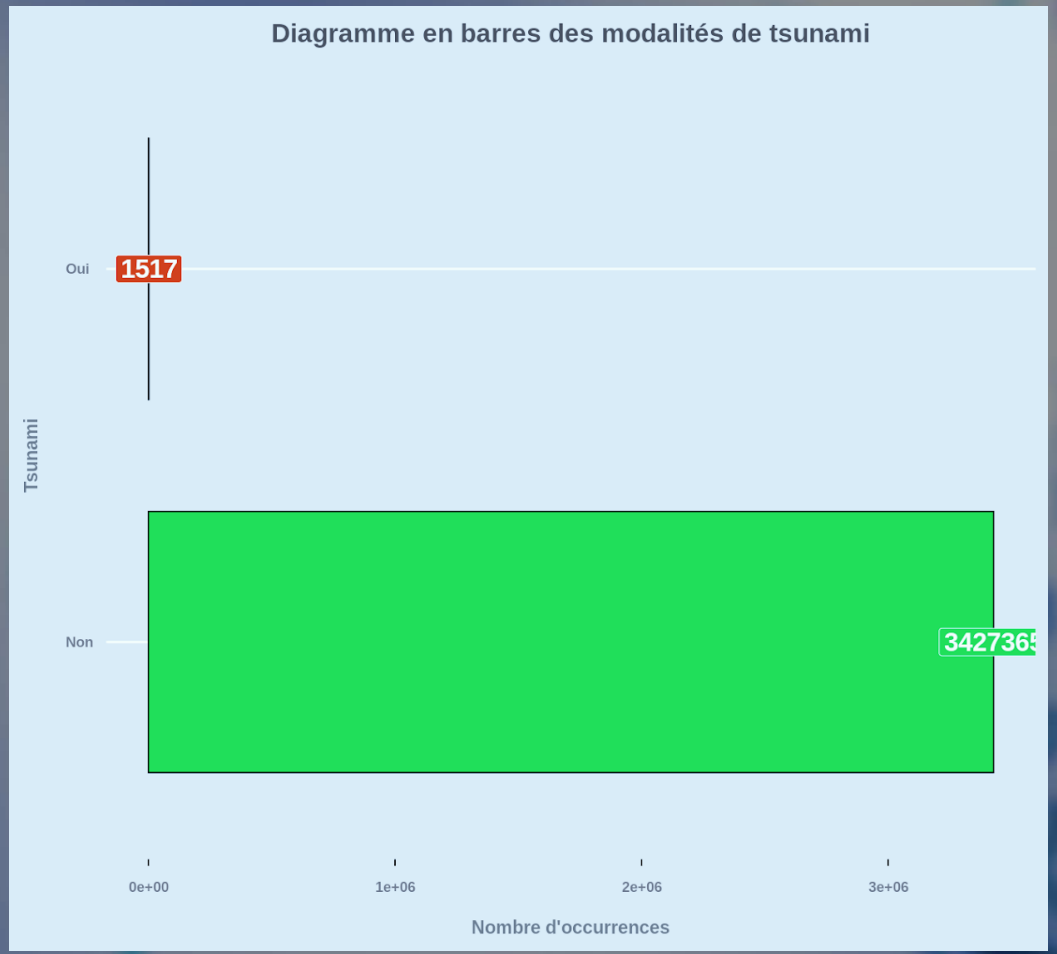
## c) Nb de tremblement de terre

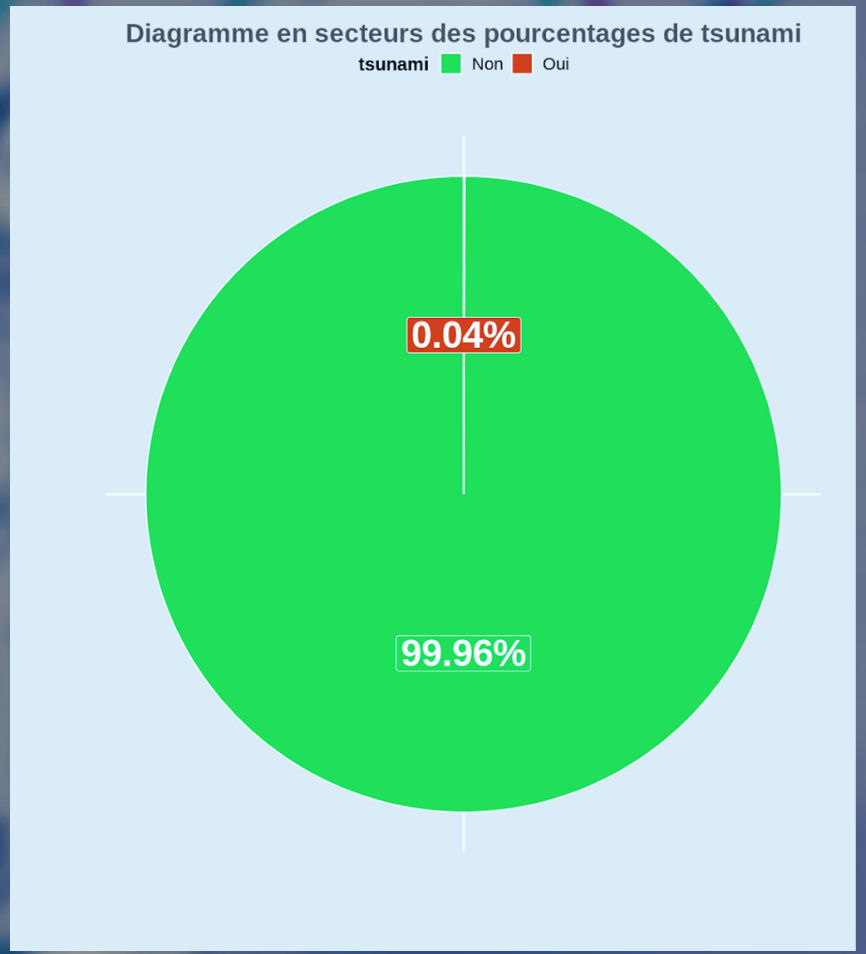



- Augmentation continue
- 2021 : plus de séisme
- 1991 : moins de séisme
- 2023 n’est pas à prendre en compte

# II – Analyse Exploratoire

## d) Tsunami

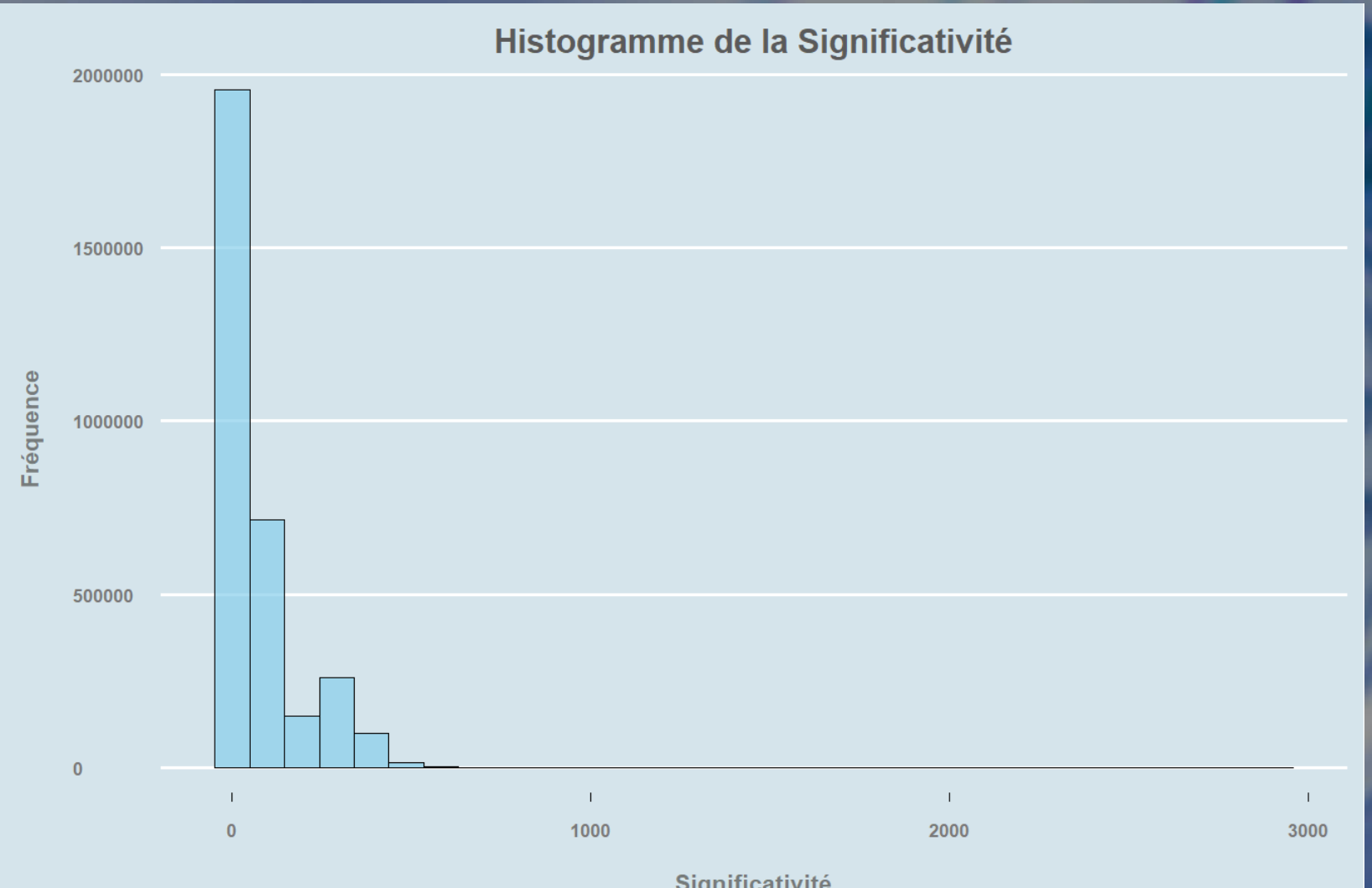



- Tsunamis très peu représentés

# II – Analyse Exploratoire

## e) Significativité




- Séismes dévastateurs sont très rares
- Corrélation mag et significance : 0,9473

# III – Le sujet

## Feature engineering :

- Aspect temporel numérique (année, mois, jours et heure)
- Pays / états encodé en One Hot Encoding (Alaska et Californie)
- Points cardinaux présent dans la variable place

## Sujets de série temporelle possibles

1. Prédiction de la magnitude
2. Prédiction des tsunamis suite à un séisme
3. Prédiction de l'impact des séismes
4. Prédiction du nombre de séisme

Le sujet 1 a été tenté mais aucun bon résultat
=> Sujet 4 avec agrégation par mois et focus sur l'Alaska

# IV – Modélisation

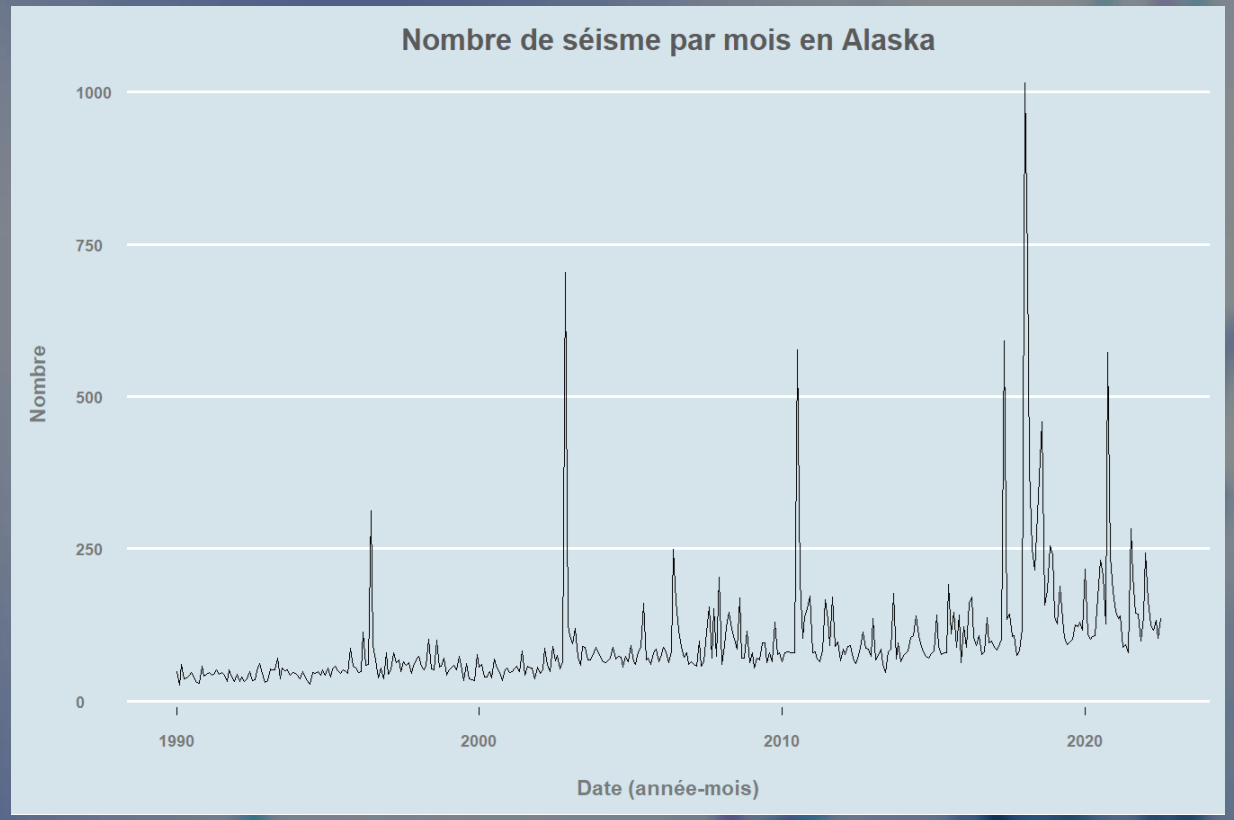
## a) ARIMA

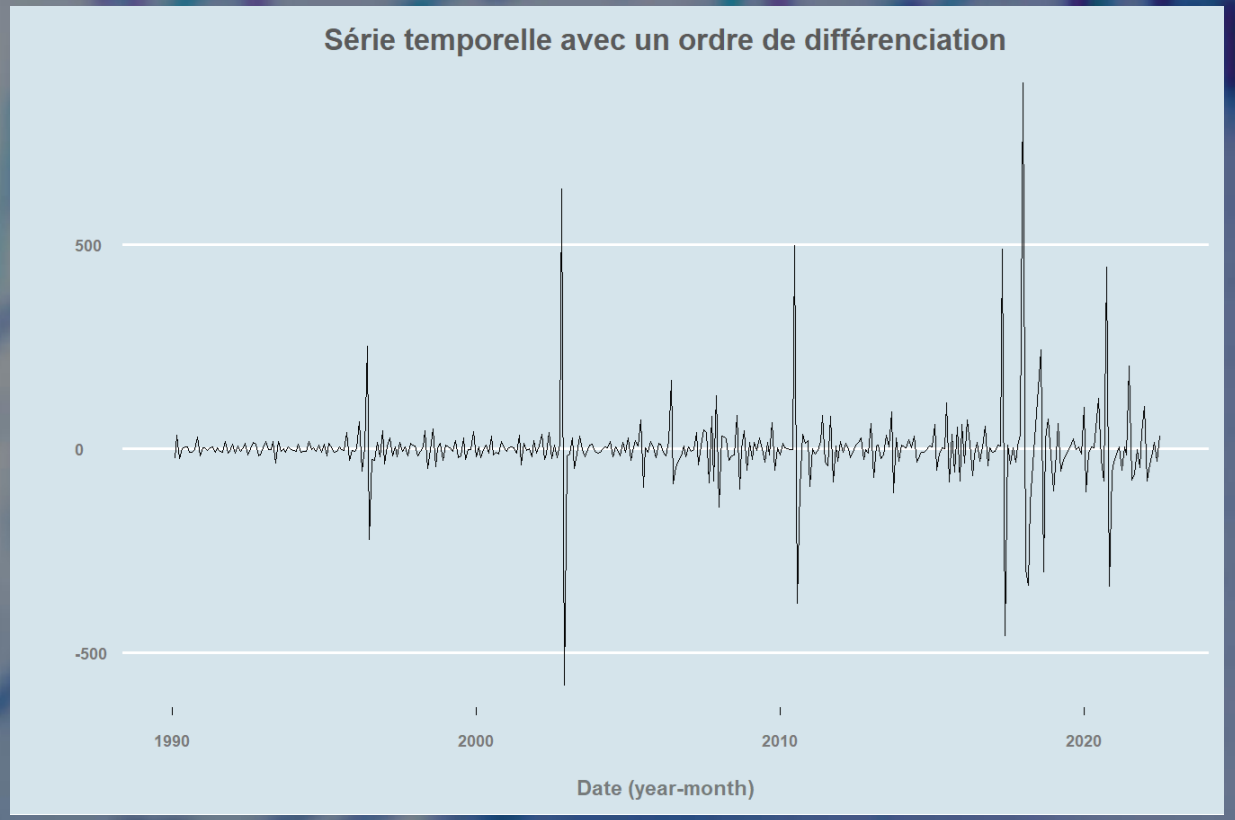
$$X_t = \sum_{k=1}^{p} a_k X_{t-k} - \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t$$

| Modèle | MA(q) | AR(p) | ARMA(p,q) |
|---|---|---|---|
| Covariance | $cov(h) = 0, \forall h > q$ | $\lim_{h\to\infty} cov(h) = 0$ | $\lim_{h\to\infty} cov(h) = 0$ |
| Corrélation | $\rho(h) = 0, \forall h > q$ | $\lim_{h\to\infty} \rho(h) = 0$ | $\lim_{h\to\infty} \rho(h) = 0$ |
| Corr partielle | $\lim_{h\to\infty} r(h) = 0$ | $r(h) = 0, \forall h > p . r(p) = a_p$ | |

# IV – Modélisation

## a) ARIMA



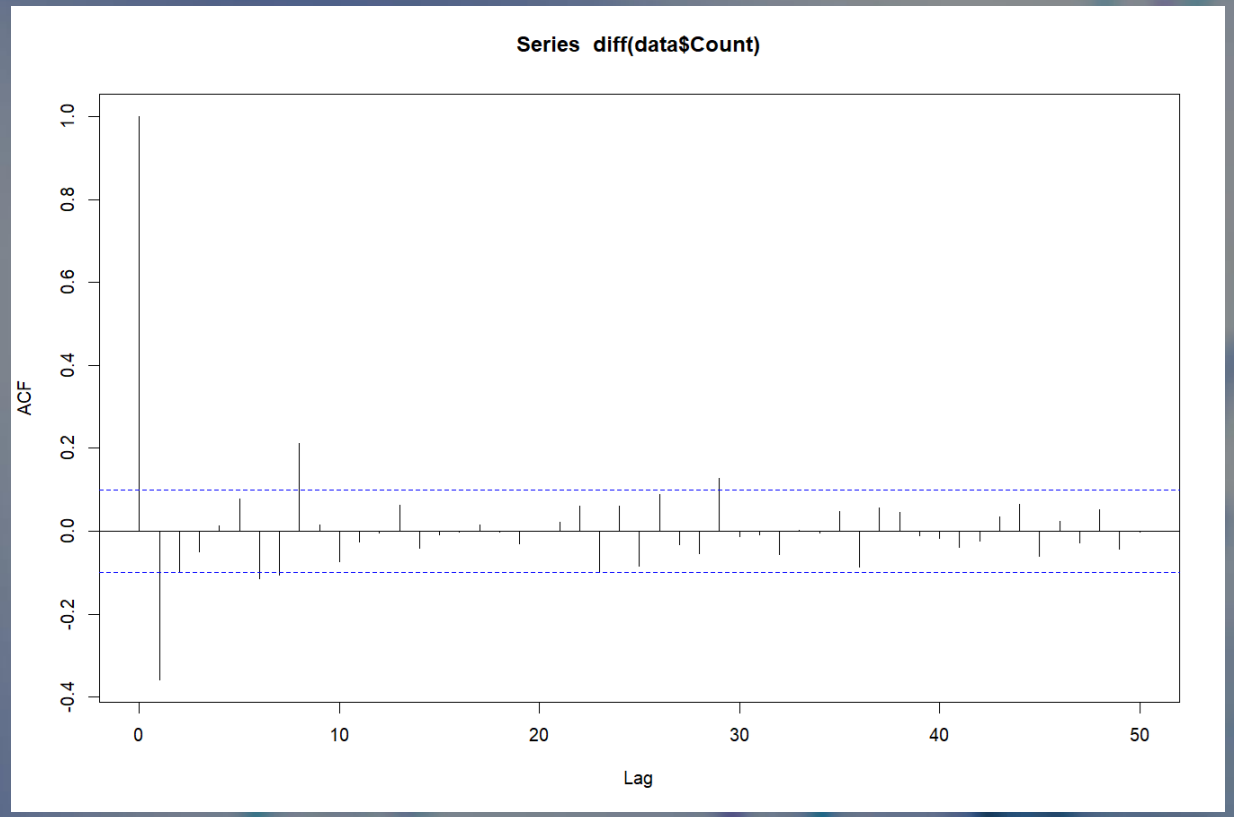

- Volatilité présente

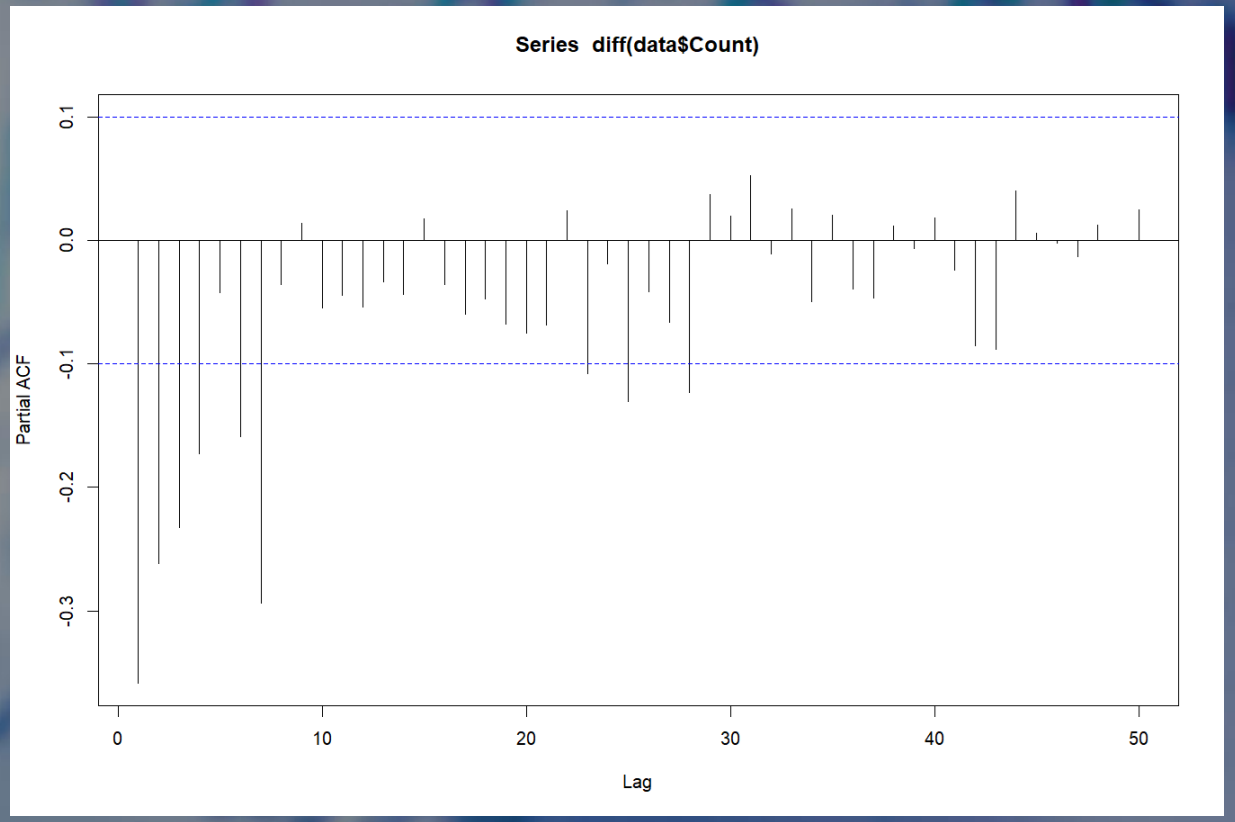
- Test de Dickey-Fuller: p-value = 0,01

⇒ Série stationnaire

# IV – Modélisation

## a) ARIMA




- Q=1
- P = 8

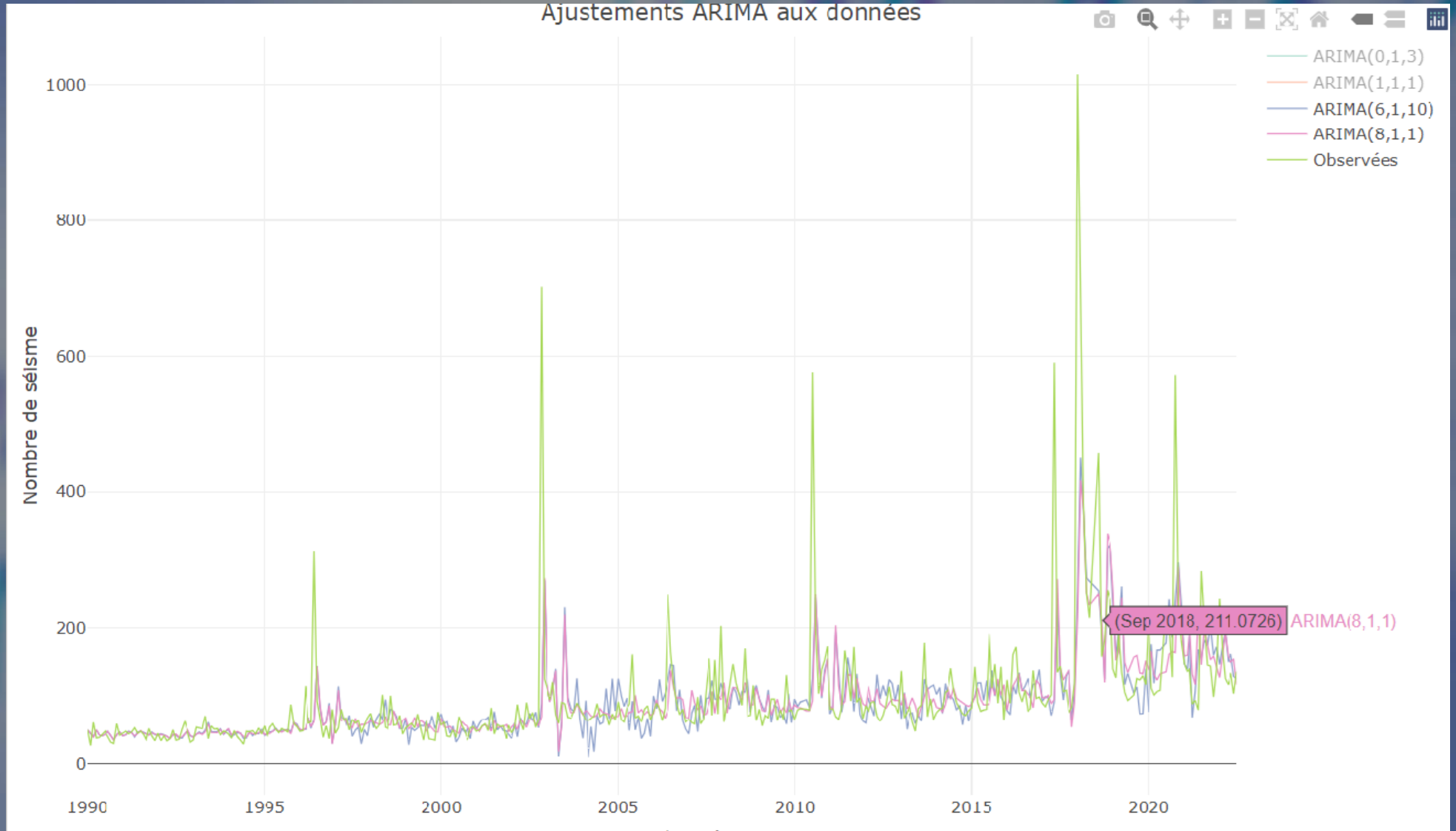
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_8 Y_{t-8} + \epsilon_t - \theta_1 \epsilon_{t-1}$$

- Recherche grille :

- ARIMA(6,1,10) : AIC (L-jung Box : p-value **non significative**)
- ARIMA(1,1,1) : BIC (L-jung Box : p-value **significative**)
- ARIMA(0,1,3) : auto.arima() (L-jung Box : p-value **significative**)

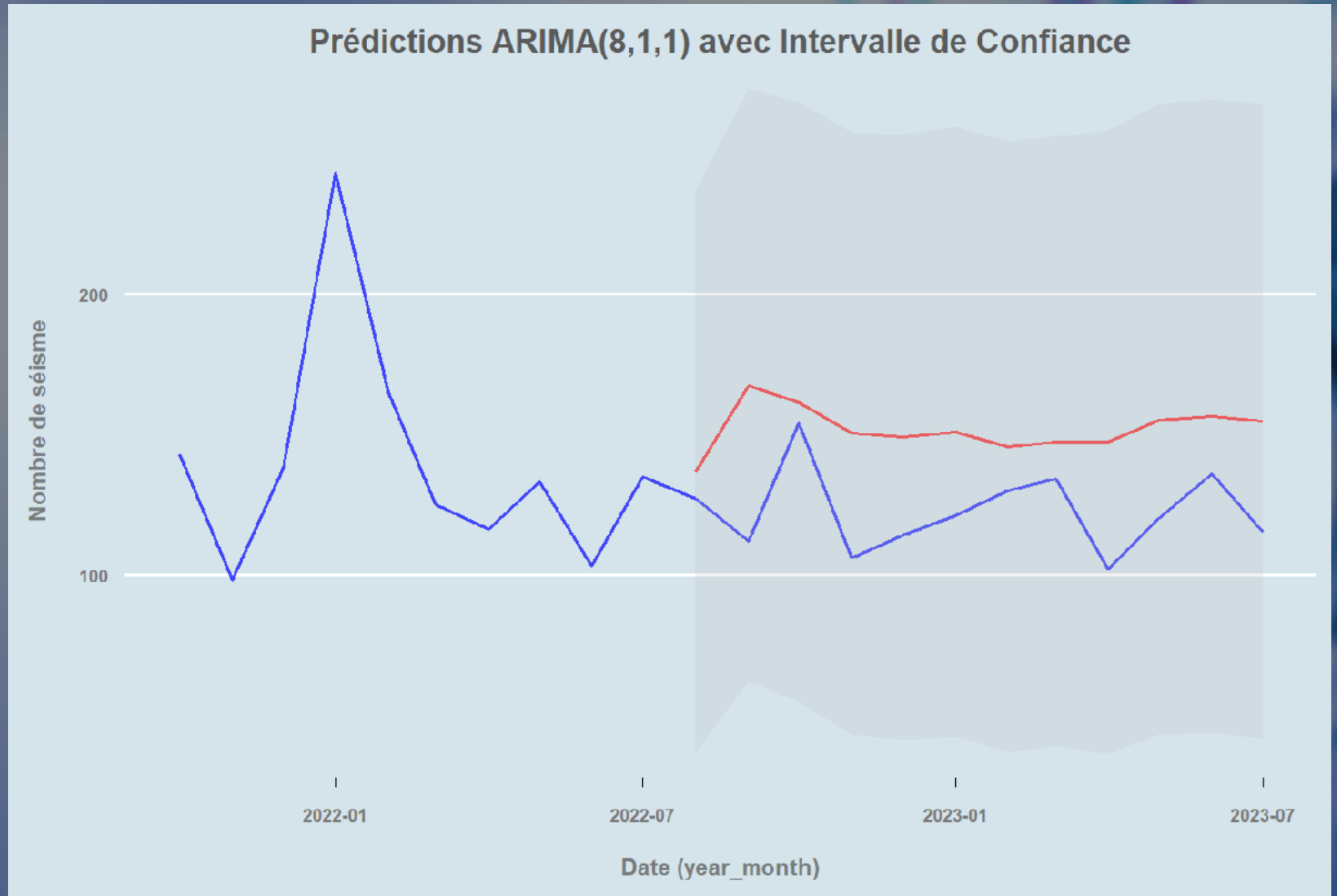
# IV – Modélisation

## a) ARIMA



# IV – Modélisation

## a) ARIMA




- ARIMA(8,1,1) ; RMSE : 34,36
- ARIMA(6,1,10) ; RMSE : 38,17
- Ne capture pas la volatilité

# IV – Modélisation

## b) GARCH

$$X_t = \sum_{k=1}^{p} a_k X_{t-k} - \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t$$

$$\varepsilon_t \sigma_t^2 = \sigma_t \eta_t$$

$$\sigma_t^2 = \omega + \alpha \sum_{i=1}^{p} \varepsilon_{t-i}^2 + \beta \sum_{j=1}^{q} \sigma_{t-j}^2$$

# IV – Modélisation

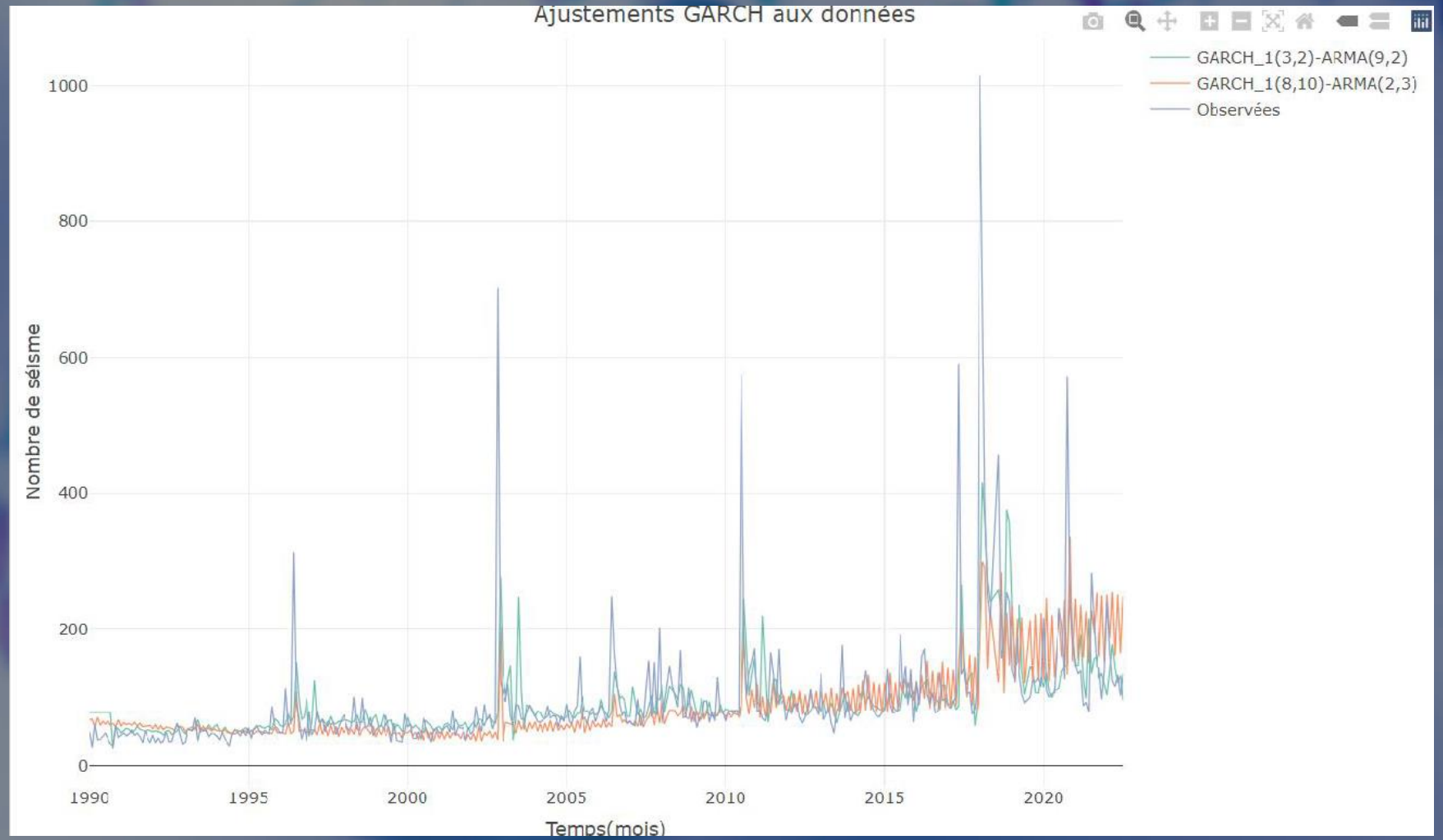
## b) GARCH

- Méthode grille search:

- ordre p et q GARCH de 1 à 10
- ordre p et q ARMA de 1 à 3

- AIC : GARCH(8,10) – ARMA(2,3)
- Expérimental : GARCH(3,2) ARMA(9,2)

- P=9 pacf() sans différenciation
- Q=2 ordre petit comme ARIMA
- Ordre GARCH choisi arbitrairement



# IV – Modélisation

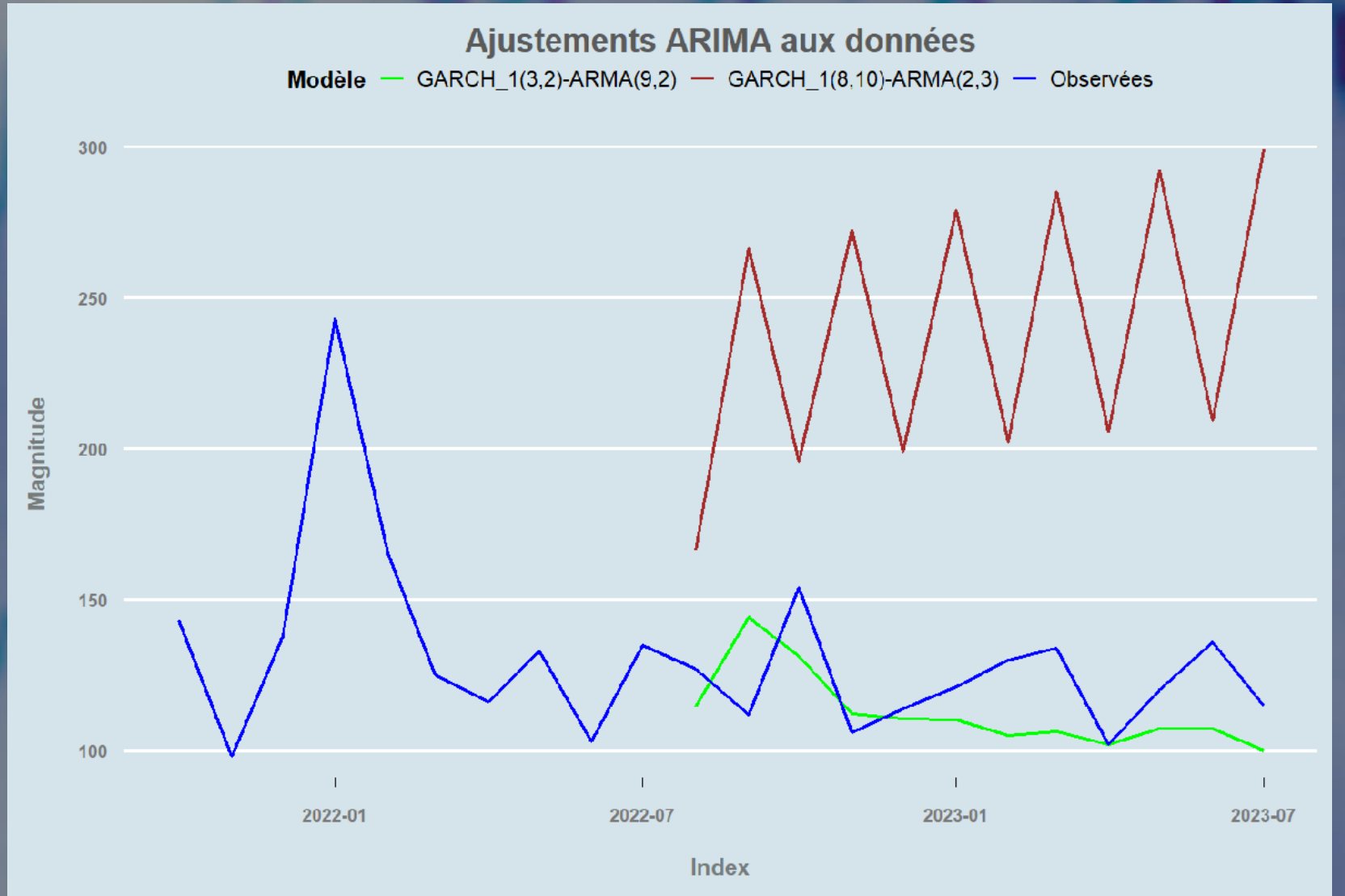
## b) GARCH

RMSE:

- GARCH(3,2) – ARMA(9,2) : 25,9
- GARCH(8,10) – ARMA(2,3) : 64,075

- Ajustement meilleur que ARIMA

- Difficulté à prédire sur le long terme



# IV – Modélisation

c) Light GBM

## Boosting:

- Combinaison de plusieurs modèles faibles
- Correction des erreurs des précédents modèles

## Gradient Boosting:

- Ajusté sens opposé du gradient de la fonction de perte

⇒ Réduire les erreurs de prédictions

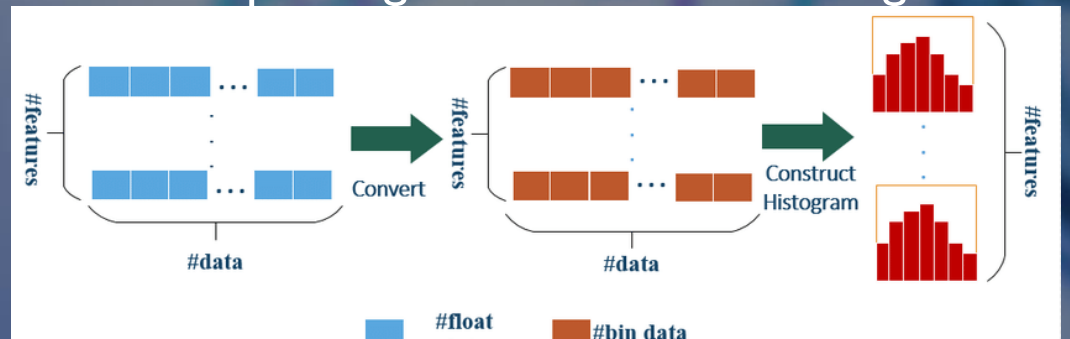
## Optimisation basé sur l'histogramme :

- Light GBM <> Boosting
- Construire les arbres sur les intervalles

⇒ Réduction du nombre de point de coupure

(Arbre : utilise les valeurs uniques)

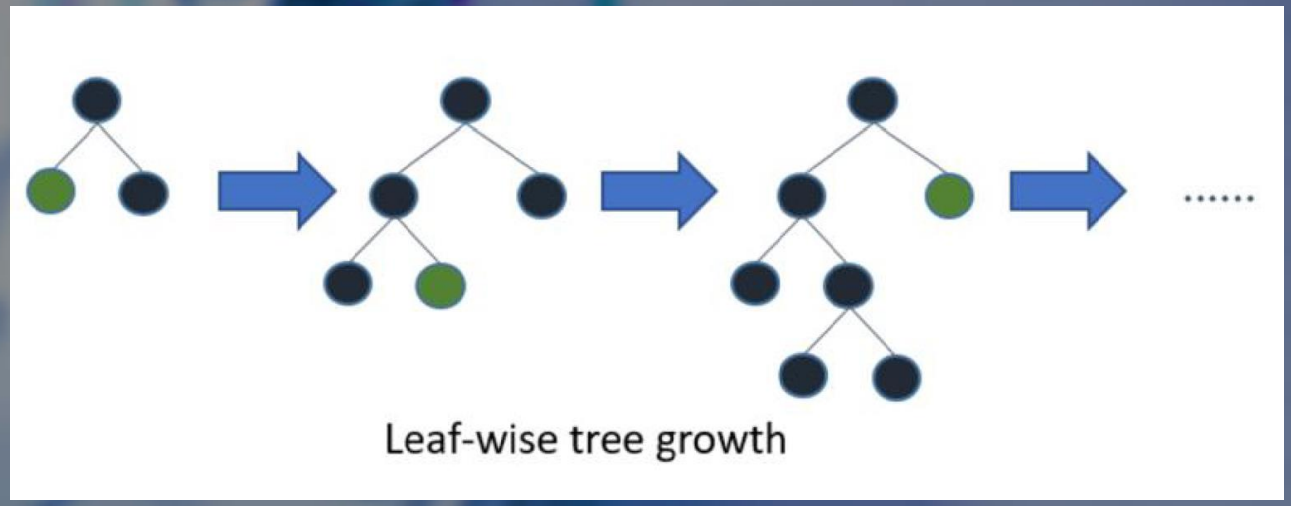
Seuil de coupure light GBM : basé sur les gradients




## Leaf-wise Splitting:

- Choisir le nœud avec la perte la plus élevée

⇒ Arbres plus profond mais plus performant




## Échantillonnage basé sur le gradient d'un côté:

- Construction de l'arbre : donner un poids aux observations qui contribue le plus aux erreurs en fonction du gradient (fort)

⇒ Efficacité de l'entraînement

# IV – Modélisation

## c) Light GBM

### Paramètre par défaut :

**Data Spliting :** train et test
Puis convertir dans un format adapté pour Light GBM : lgb.Dataset()
Data Test : 12 derniers mois de la série temporelle jusqu'à juillet 2023
Nombre d'itération de boosting à fixer : 1000
**RMSE** : 23,1048

### Recherche aléatoire (Random Search):

Grille de paramètre avec recherche aléatoire (nb max de feuille, nb total d'arbre et taux d'apprentissage)

- nb max de feuille : 20 à 50 pas de 5 (optimal : 35)
- nb total d'arbre : 50 à 300 pas de 50 (optimal : 50)
- taux d'apprentissage : 0,01 à 0,2 pas de 0,05 (optimal : 0,11)

**RMSE** : 22,9547

### Méthode bayésienne :

Prend en compte les évaluations passées en fonction des hyperparamètres.
⇒ Moins d'itérations
nb max de feuille : 27
nb total d'arbre : 50
taux d'apprentissage : 0,02057
**RMSE** : 19,83

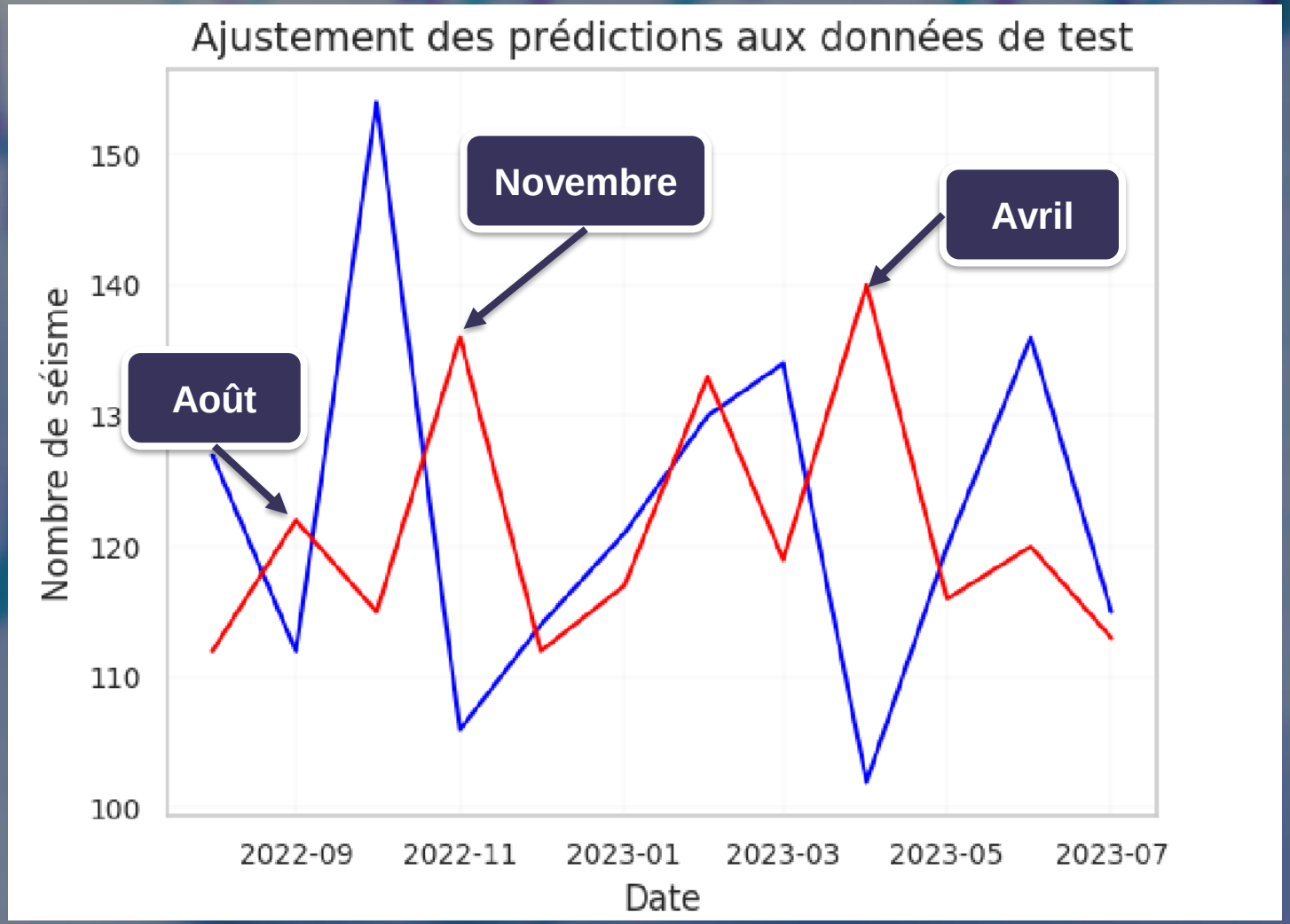
### Algorithme génétique:

Générer une population initiale
Evaluer et sélectionner les meilleurs
Opérations de croisement et de mutation => évoluer vers des configurations optimales au fil des générations.
nb max de feuille : 24
nb total d'arbre : 64
taux d'apprentissage : 0,015
**RMSE** : 20,13

# IV – Modélisation

## c) Light GBM

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



# IV – Modélisation

## d) RNN

**Réseau de neurones récurrent :**

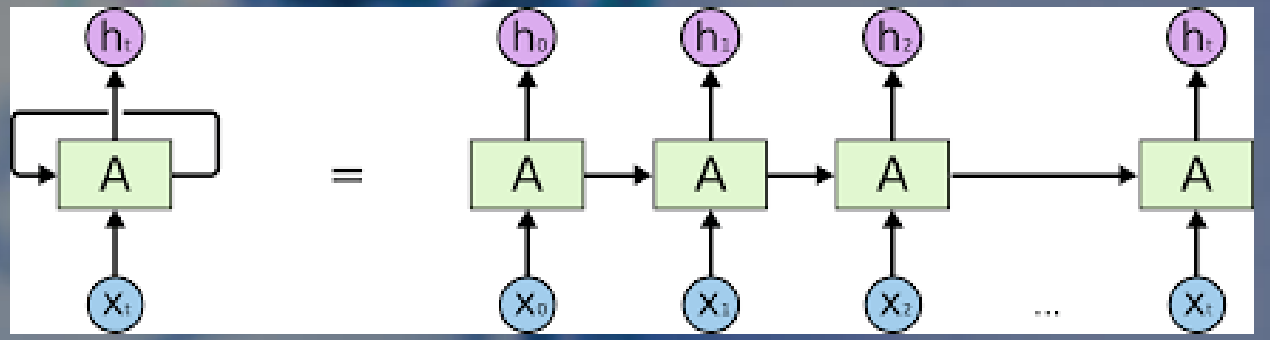
ht : sortie xt: données d'entrées

Données séquentielles (langage, données temporelles)

Capture les informations au fur et à mesure

Limites : atténuations du passé lointain

Mauvais sur des données avec dépendance complexe




**LSTM**

# IV – Modélisation

## d) LSTM

**Cellule Ct et Porte de l'oubli:**

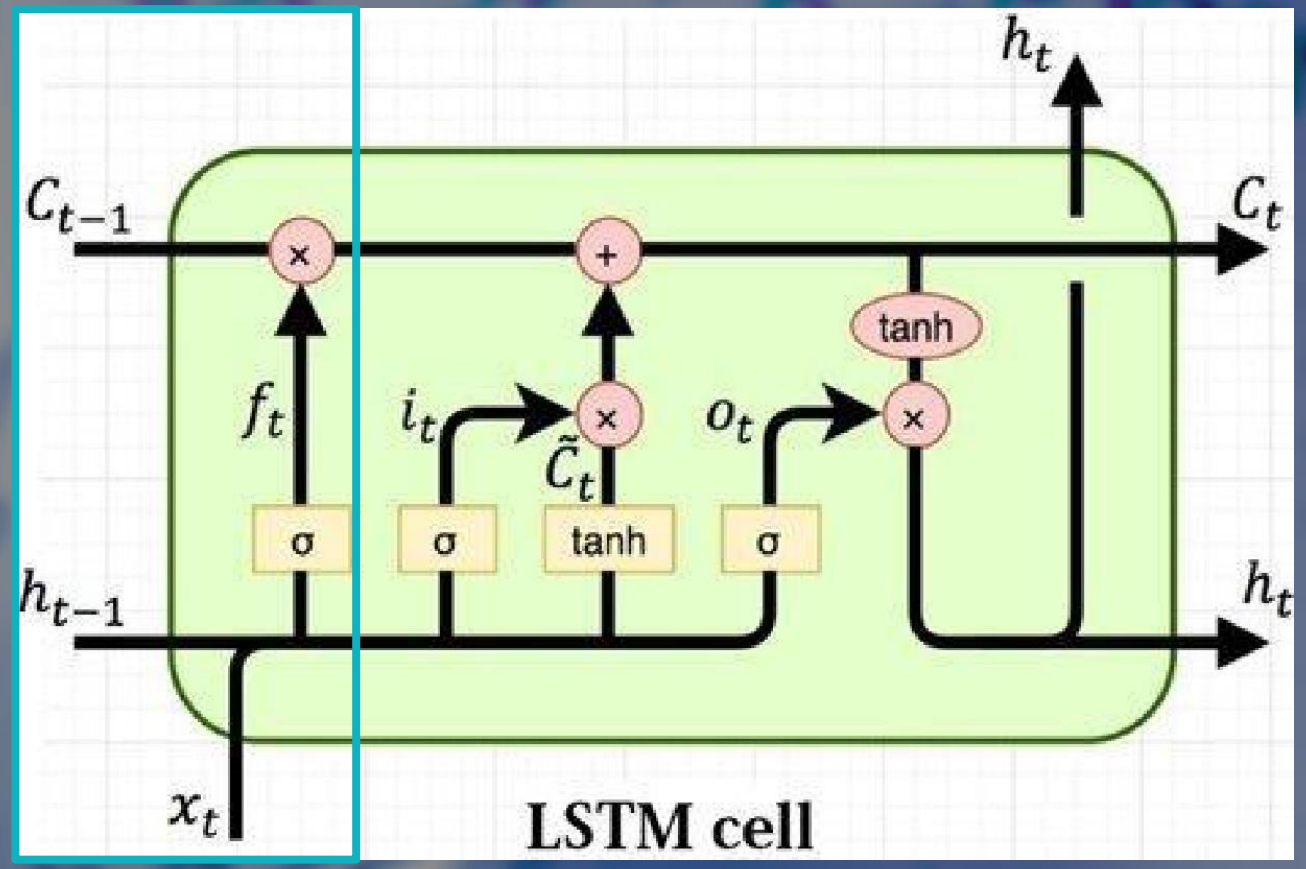
- Cellule Ct (vecteur) : stocker les informations [0,1]

- Porte de l'oubli : examine xt et ht-1 et attribue un nombre entre 0 et 1 pour chaque élément de Ct-1

$$f_t = \sigma(W_f x_t + U_f h_{t-1})$$

Avec :

- $f_t$ est le vecteur de port d'oubli à l'instant $t$
- $\sigma$ est la fonction d'activation sigmoïde $\sigma(z) = \frac{1}{1+e^{-z}}$
- $W_f$ est la matrice de poids porte de l'oubli $(f_t)$
- $U_f$ est la matrice de poids porte de l'oubli $(f_t)$



# IV – Modélisation

## d) LSTM

**Porte d'entrée:**

- Décide quelles unités cellulaires doit être mis à jour

$$i_t = \sigma(W_i x_i + U_i h_{t-1})$$

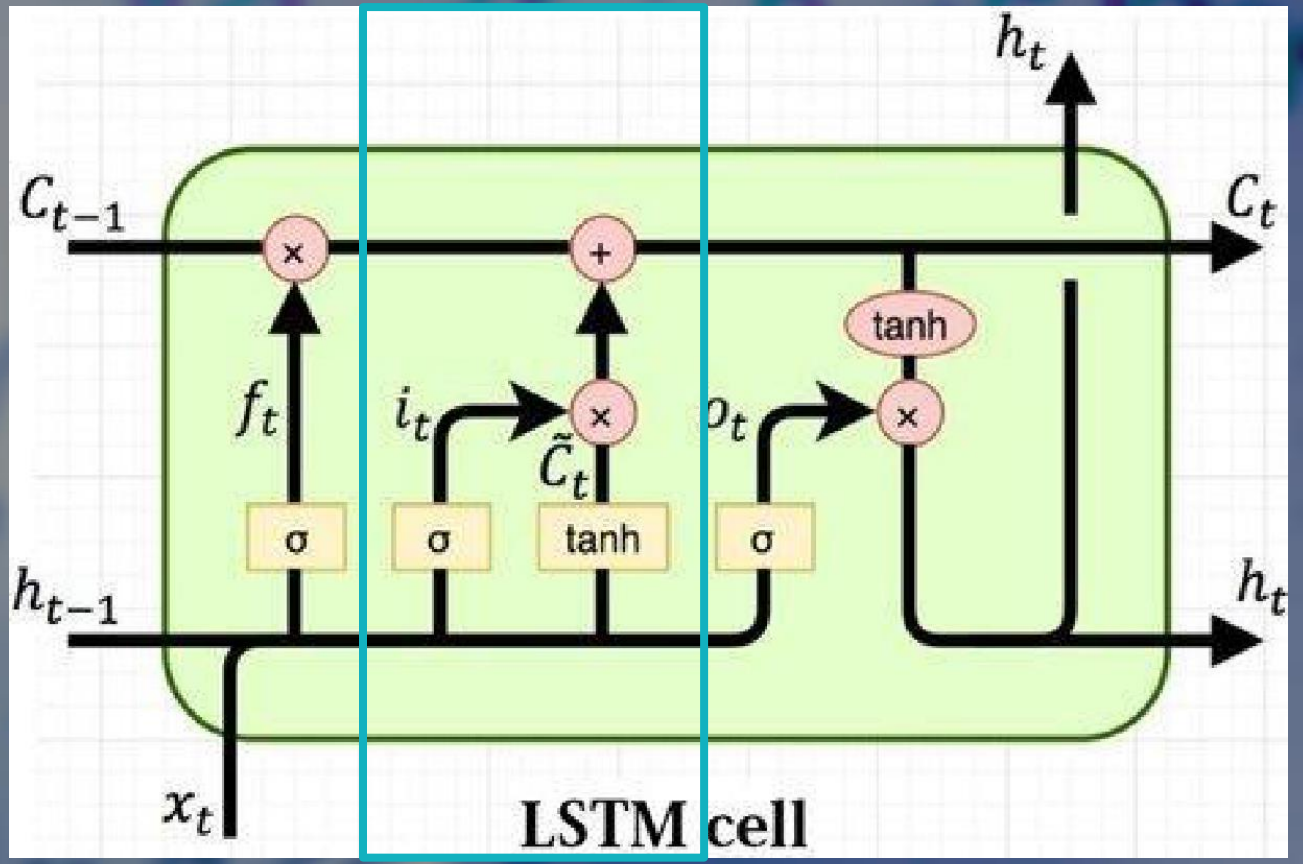
**Porte d'état :**

- Régule la mise à jour de l'état de la cellule

$$\widetilde{C}_t = tanh(W_c x_t + U_c h_{t-1})$$

**Mise à jour:**

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t$$




**Rappel :** *Tanh*

$$f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$

# IV – Modélisation

## d) LSTM

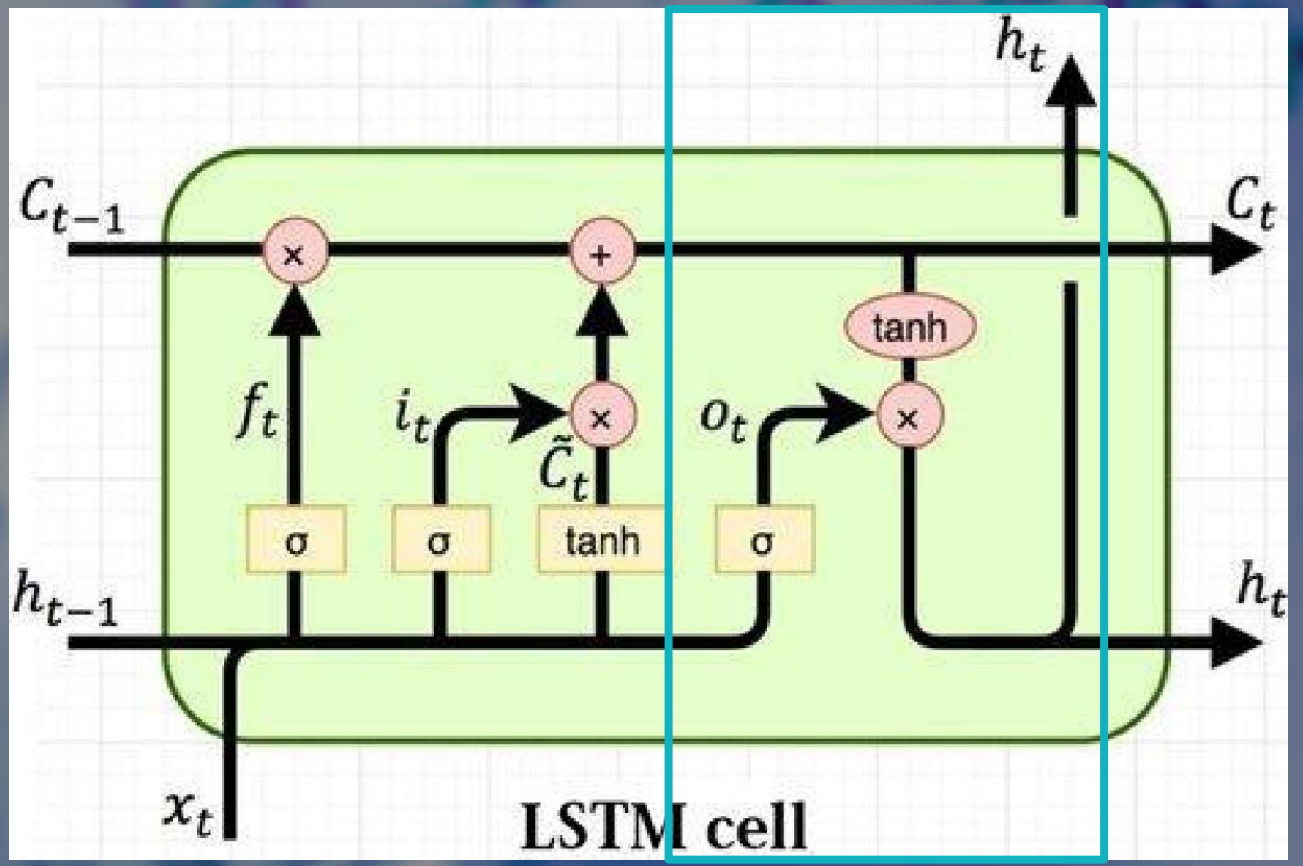
**Porte de sortie :**

- Contrôle la sortie de l'état de la cellule pour l'étape actuelle.

$$o_t = \sigma(W_o x_t + U_o h_{t-1})$$

**Calcul de la prédiction :**

$$h_t = o_t \times tanh(C_t)$$



# IV – Modélisation

## d) LSTM

| 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 |

**Nombre de paramètre:**
522 301 paramètres

**Couche :**
Dropout : éviter le surapprentissage (20%)

**Test:**
BatchNormalization : normaliser les sorties des modèles
GaussianNoise (0,005 ; 0,01 et 0,015)

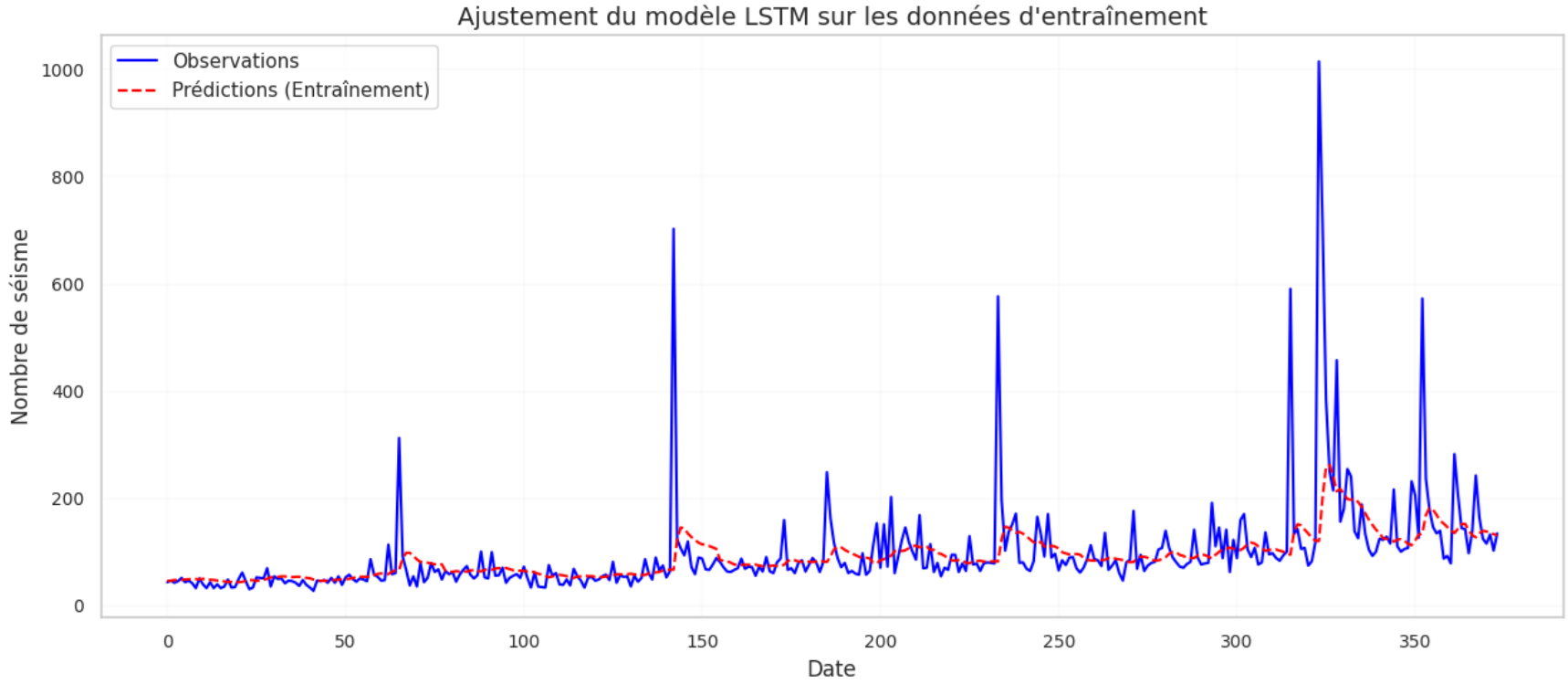
**Compilateur Adam:**
Ajuste les poids et les biais lors de l'entraînement

**Paramètre entraînement:**
50 époch : nb de fois où le modèle parcourt l'intégralité des train data
Batch_size=32 : nb d'échantillon lors d'une itération

**Métriques:**
RMSE : 15,22
MAE : 12,09
MAPE : 9,62%

# IV – Modélisation

## d) LSTM

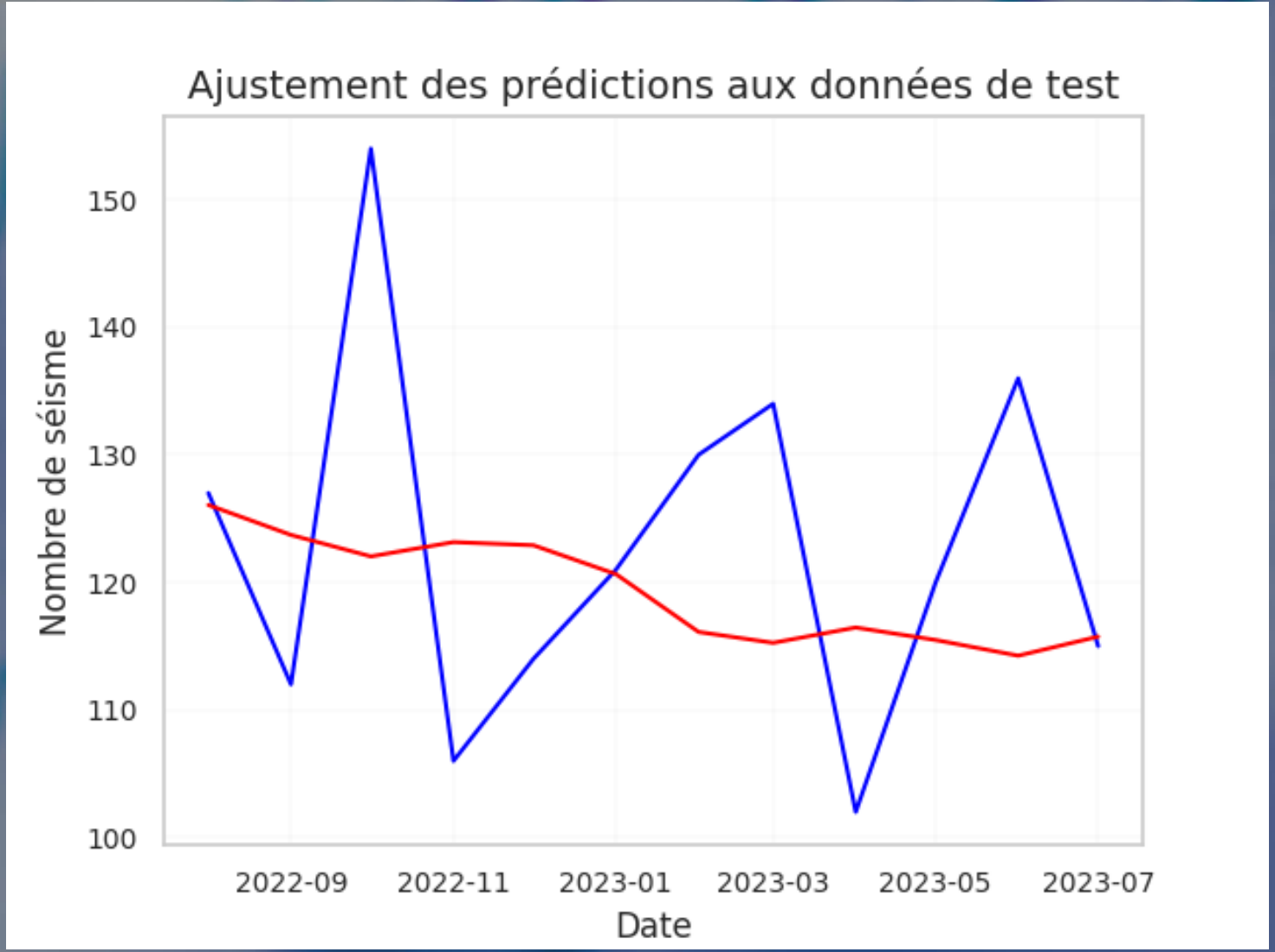



- Suit les tendances générales
- Capture les structures de données complexes

# IV – Modélisation

## d) LSTM

- Bonne adéquation sur les données de test
- Modèle avec une approche conservatrice (volatilité faible)
- Meilleur modèle (robuste, efficace)



# Conclusion

- LSTM a surpassée nos attentes (performances, robustesse, rapidité et efficacité)
- Limites de notre approches : focus sur Alaska (extrapolation) => reproduire l'analyse sur les autres pays / états
- Amélioration des résultats en utilisant des variables exogènes  (nature des sols, données sur les plaques)
- Essais sur la création d'une app streamlit pour utiliser le modèle (problème de version avec streamlit et tensorflow)
- D'autres approches, telles que le modèle FARIMA, auraient pu être évaluées