Projet Séries temporelles - Analyse des Données Géologiques

Guillaume POIRIER, Davyd BAYARD

29/12/2023

Résumé

Ceci est un résumé du projet.

Table des matières

Partie 1

Introduction

1.1 Contexte

L'analyse géologique, englobant l'activité sismique et les éruptions volcaniques, joue un rôle crucial dans la compréhension des intrications des processus géologiques de la Terre et, plus essentiellement, dans la prédiction d'événements futurs. À une époque marquée par les avancées technologiques et une sensibilisation mondiale accrue, l'analyse de ces données émerge comme une pierre angulaire dans la préparation et l'anticipation des catastrophes. Ce rapport se plonge dans les méthodologies utilisées pour scruter les données géologiques en séries temporelles, visant non seulement à comprendre les tendances historiques, mais aussi à prévoir des événements sismiques potentiels. La signification 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 proviennent exclusivement du Service Géologique des États-Unis (USGS). La stratégie open-data clairement affichée fournie une base complète pour l'analyse. Le périmètre de l'étude choisi concerne les tremblements de terre à l'échelle mondiale survenus depuis 1990.

Les sections suivantes de ce rapport décrivent l'approche systématique pour traiter les données géologiques en séries temporelles, depuis la prétraitement et l'analyse exploratoire des données jusqu'à la sélection du modèle et la prédiction d'événements.

1.2 Objectifs

Dans une phase préliminaire, l'objectif consiste à acquérir une compréhension, à observer et à constater en vue de partager une première analyse des données. L'importance accordée au prétraitement des données se justifie par le constat que les connaissances spécialisées dans le domaine des activités sismiques ne sont pas universellement répandues. Cette étape s'avère être une véritable aculturation.

Par la suite, la deuxième phase consistera à procéder au nettoyage, à la transformation et à l'enrichissement des données. Étant donné que la collecte se fait dans un format public, des ajustements sont nécessaires pour préparer les données en vue de la modélisation.

Il existe une diversité de choix pour la sélection des variables à modéliser. Ce rapport va exhaustivement décrire nos différentes tentatives de modélisation, aboutissant à notre sujet principal, à savoir la modélisation du nombre de séismes par mois en Alaska.

La modélisation constituera en elle-même une étape significative de ce rapport. Divers modèles seront testés et détaillés. Dans une démarche logique, cette étape sera suivie d'une validation du modèle et de l'évaluation de ses performances, incluant des considérations sur l'explicabilité.

L'objectif ultime de ce processus est le développement de systèmes d'alerte précoce, robustes, capables d'avertir les autorités et les communautés pertinentes dès la détection d'événements géologiques potentiel fréquent sur le mois.

Partie 2

Analyse des Données Géologiques

2.1 Les phénomènes sismiques

D'après une définition donnée sur wikipédia [1], un séisme ou tremblement de terre est une secousse du sol résultant de la libération brusque d'énergie accumulée par les contraintes exercées sur les roches. Cette libération d'énergie se fait par rupture le long d'une faille, généralement préexistante. Plus rares sont les séismes dus à l'activité volcanique ou d'origine artificielle (explosions par exemple). Le lieu de la rupture des roches en profondeur se nomme le foyer ; la projection du foyer à la surface est l'épicentre du séisme. Le mouvement des roches près du foyer engendre des vibrations élastiques qui se propagent, sous la forme de paquets d'ondes sismiques, autour et au travers du globe terrestre. Il produit aussi un dégagement de chaleur par frottement, au point de parfois fondre les roches le long de la faille (pseudotachylites).

La prédominance des séismes se manifeste principalement aux confins des plaques tectoniques, donnant lieu aux séismes inter-plaques. Toutefois, des événements sismiques peuvent également survenir à l'intérieur des plaques, caractérisés comme des séismes intra-plaques. L'explication adéquate de la répartition des ceintures de sismicité à la surface de la Terre réside dans le concept de tectonique des plaques. Les principales ceintures sismiques mondiales, définies par la concentration géographique des activités sismiques, comprennent la ceinture de feu du Pacifique, libérant annuellement 80 % de l'énergie sismique, la ceinture alpine contribuant à hauteur de 15 % de l'énergie annuelle, ainsi que les dorsales océaniques, responsables de 5 % de l'énergie sismique annuelle.

Ci-dessous se trouve un schéma illustrant la dynamique des séismes inter-plaques, incorporant également une explication claire des concepts tels que le foyer, l'épicentre et les ondes sismiques.

Les séismes intra-plaques se produisent à l'intérieur d'une seule plaque tectonique, constituant une forme rare d'activité sismique en contraste avec les séismes inter-plaques, plus fréquents, qui résultent de l'interaction entre au moins deux plaques tectoniques. Bien que moins fréquents, les séismes intra-plaques peuvent avoir des conséquences dévastatrices en raison du manque d'infrastructures parasismiques coûteuses dans des zones généralement considérées comme peu susceptibles de subir de tels événements.

Ces séismes intra-plaques se déroulent principalement dans la partie supérieure de la croûte terrestre et peuvent être classés en deux catégories distinctes : le champ de contraintes régional et le champ de contrainte local, chacun capable de déclencher un séisme. Dans le premier cas, les contraintes globales de la plaque sont uniformes sur sa surface, libérant ainsi l'énergie accumulée dans des failles préexistantes, telles que celles héritées de la dislocation de Pangée. Quant au champ de contrainte local, il concentre l'énergie en un point spécifique, générant ainsi un séisme. Les mécanismes explicatifs de ce dernier cas sont encore incompris, mais deux hypothèses sont en compétition : le modèle de faible force, qui postule que les contraintes se situent dans des zones de faible viscosité, rendant ainsi la croûte plus cassante , et l'hypothèse de l'entraînement

basal, basée sur les courants de convection descendant du manteau qui emportent la croûte avec eux. Dans tous les scénarios, la présence de failles est un prérequis essentiel.

Il est également important de noter que des interactions avec des séismes inter-plaques sont possibles, où l'un peut déclencher l'autre. Un exemple notable est le séisme de 2009 aux Samoa, où un séisme intra-plaque a provoqué deux séismes de subduction.

2.2 Les méthodes de mesures d'amplitudes

Les méthodes de mesure des séismes abordées dans la suite proviennent des références [2] et [3]. Chacune de ces approches vise à quantifier la puissance du phénomène sismique dans les trois cas présentés.

2.2.1 Magnitude locale M_L

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.

La formule originale de Richter pour la magnitude locale (M_L) est une échelle logarithmique simple : $M_L = \log(A) - \log(A_0) + c \times \log(\Delta)$

- A représente l'amplitude maximale,
- A_0 est une amplitude de référence pour un séisme de magnitude 0 à 100 km,
- Δ est la distance épicentrale,
- c 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 $\log(A_0) = 2.48$.

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.

2.2.2 Magnitudes dites d'ondes m_b et M_S

L'échelle de Richter, une mesure locale introduite en 1936, a conduit à l'émergence d'une nouvelle magnitude appelée M_S (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 (M_L) :

$M_S = \log(A_{20}) + b + c \times \log(\Delta)$

- A_{20} est l'amplitude mesurée,

- Δ est la distance épicentrale en degrés,
- b et c sont des constantes d'étalonnage.

Malgré son caractère empirique et les problèmes de saturation, cette mesure 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.

La magnitude des ondes de volume, notée mb (b 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 :

$m_b = \log(A/T) + Q(\Delta, h)$

- A est l'amplitude maximale mesurée,
- Δ est la distance épicentrale en degrés,
- h est la profondeur hypocentrale,
- Q est une fonction d'étalonnage dépendant de ces paramètres.

Cependant, cette mesure présente également des problèmes de saturation rapide avec la magnitude.

D'autres magnitudes sont utilisées à l'échelle locale ou régionale, telles que la magnitude de durée, basée sur la mesure de la durée en secondes du signal sur le sismogramme. La variabilité de ces mesures, due à divers facteurs tels que le type d'onde, le capteur utilisé, la distance et le type de magnitude, explique la difficulté à établir des relations précises entre elles.

2.2.3 Magnitude des Moments

En 1979, Thomas Hanks et Hiroo Kanamori, chercheurs au Caltech en Californie, ont introduit une nouvelle méthode de calcul de la magnitude des séismes, appelée Mw ou magnitude de moment.

Cette approche repose sur le modèle physique de la rupture d'un séisme, prenant en compte la déformation élastique associée à un double-couple de forces de directions opposées et perpendiculaires. Le moment sismique (M_0), exprimé en Newton.mètres (N.m), est une mesure de l'énergie sismique liée au déplacement sur la faille.

La magnitude M_w est calculée à partir du moment sismique selon la formule :

$$M_w = \frac{3}{2} \cdot (\log_{10} M_0 - 9.1)$$

Où M_0 est exprimé en N.m. Par exemple, un séisme de magnitude $M_w \approx 6$ correspond à un moment sismique $M_0 \approx 10^{18}$ N.m.

Le moment sismique M_0 est défini comme :

$$M_0 = \mu \cdot S \cdot \Delta u$$

Où :

- μ est le module 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.

- Δu est le déplacement moyen sur la faille.

Cette approche permet une meilleure évaluation de la magnitude des gros séismes en prenant en compte la répartition temporelle de l'énergie libérée. La contrainte inhérente à cette méthode réside dans l'impératif de présence d'une faille. Ainsi, son efficacité maximale est atteinte uniquement dans le contexte de phénomènes sismiques d'une magnitude significative.

2.3 Collecte des données

Comme précédemment indiqué, les données ont été recueillies à partir du site de l'USGS, un organisme chargé de la surveillance mondiale de la Terre et mettant à disposition une base de données gratuite. C'est une exposition de l'ensemble des informations provenant des laboratoires de surveillances de la terre, rassemblé et mis en commun dans un seul et même endroit. Afin de gérer le volume important de données, le jeu de données présenté et analysé dans la suite du rapport résulte de multiples requêtes jointes.

Il est important de noter l'absence de filtrage des données. Tous les tremblements de terre, survenus à l'échelle mondiale de 1990 à aujourd'hui, sont pris en considération. Une rapide description des colonnes importantes est présenté ci-dessous :

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

Table 2.1: Description des champs des données sur les séismes.

Voici une présentation des premières lignes de notre jeu de données :

2.4 Prétraitement des données

2.4.1 Nettoyage des données

Initialement, nous avons procédé à une vérification de l'intégralité des données relatives aux variables sélectionnées. Heureusement, aucune donnée n'était manquante, attestant ainsi de la haute qualité des données.

Dans une étape subséquente, nous avons examiné la présence éventuelle de doublons. En raison de la nature de plusieurs requêtes associées, nous avons identifié 16869 lignes en double sur un total de 3445751. Par conséquent, un processus de dédoublonnage a été effectué au stade de prétraitement.

2.4.2 Conversion des horodatages

Initialement, étant donné que la variable "time" était exprimée en timestamp, 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 de modélisation répétées, une conversion a été réalisée. Cette conversion 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'agréger les données par mois. En conséquence, la cible et l'objectif du projet ont été modifiés pour refléter cette nouvelle approche.

Partie 3

Analyse descriptive

3.1 Magnitude moyenne et médiane sur différentes périodes

En utilisant le package ggplot2, il est possible de générer une représentation graphique claire des magnitudes annuelles, incluant à la fois la moyenne et la médiane. Cette visualisation permet d'observer une tendance globale, évaluant ainsi la possible corrélation entre ce phénomène naturel et les changements climatiques. Une analyse préliminaire indique cependant qu'aucune relation significative n'est observée.

Cette analyse englobe les données depuis 1990, présentant une perspective sur une période étendue. Ce qui est notable ici, c'est l'absence apparente de saisonnalité, mais plutôt une volatilité significative au sein de la série chronologique. Par ailleurs, il est observé que la magnitude moyenne annuelle des tremblements de terre ne manifeste pas de tendance à la hausse. Par conséquent, à première vue, aucune conclusion ne peut être tirée quant à un éventuel lien avec le réchauffement climatique.

Les diagrammes en boîte fournissent des informations cruciales sur la magnitude moyenne annuelle, tout en mettant en évidence la présence de nombreux tremblements de terre très puissants, qui sont largement ressentis par les êtres humains. C'est encore plus complexe, à mon sens, de dégager une réelle évolution sur le temps.

Ici, pas de saisonnalité non plus. En revanche, une tendance nettement plus stable se manifeste, conformément à nos attentes.

L'année 2023 ne doit pas être prise en considération, car le dataframe s'arrête en juillet. Contrairement à la magnitude, une tendance claire se dégage en termes de fréquence de tremblements, montrant une augmentation au fil des années. Ainsi, pour résumer, il n'y a pas d'augmentation moyenne de la magnitude, mais une fréquence croissante d'apparition de séismes, indiquant une augmentation du nombre de séismes, principalement de faible intensité.

3.2 Cartographie des séismes

Cette section fournira une brève analyse des emplacements fortement affectés par les séismes.

On voit tout de suite deux pays clairement émerger : la Californie et l'Alaska. C'est sur ce dernier que notre modélisation portera.

3.3 Nombre de séisme par année / par endroit

3.4 Les séismes les plus importants

magnitudes Min. 1st Qu. Median Mean 3rd Qu. Max. 0.00 13.00 33.00 74.01 81.00 2910.00

Partie 4

Choix du sujet

4.1 Les essais

4.2 Choix retenu

Partie 5

Modélisation

5.1 ARIMA

5.2 GARCH

5.3 LightGBM

5.4 LSTM

Partie 6

Conclusion

Liste des figures

Bibliographie

[1] *Séisme*. Wikipédia. Disponible à : https://fr.wikipedia.org/wiki/Séisme. Consulté le : Date de consultation (06/01/2024).

[2] *Magnitude (Sismologie*. Wikipédia. Disponible à : https://fr.wikipedia.org/wiki/Magnitude_(sismologie). Consulté le : Date de consultation (07/01/2024).

[3] *Magnitude de moment*. Musée de sismologie. 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.. Consulté le : Date de consultation (07/01/2024).