Rapport de projet tutoré : Etudes des chutes de météorites sur Terre

Yannis Petit, Rassem Djimadoum, Duc-Khoi Nguyen & Garance Malnoë
encadré.e.s par Jean-François Coeurjolly

Janvier-Avril 2025 - M1 SSD

Table des matières

1 Introduction

Ce rapport présente le projet tutoré réalisé de janvier à avril 2025 dans le cadre de notre première année de Master en Statistique et Sciences des Données, sous la direction de Jean-François Coeurjolly. Ce projet se concentre sur l'analyse du jeu de données *Meteorite Landings*, qui recense les météorites tombées au sol. Ce jeu de données est mis à disposition par la Meteoritical Society et est accessible sur l'Open Data de la NASA [NAS].

Au départ, nous n'avons pas d'objectif défini pour ce projet ; l'idée initiale consiste à explorer librement le jeu de données et à identifier les axes d'analyse les plus pertinents. Cependant, Jean-François Coeurjolly nous a suggéré plusieurs pistes lors de notre premier échange :

- Étudier la dimension temporelle des données, notamment la présence d'une saisonnalité ou d'une tendance.

- Analyser les relations entre différentes variables du jeu de données, en particulier l'influence de la masse des météorites.

- Construire un modèle prédictif du nombre de météorites tombées dans un pays en fonction de sa superficie et de sa localisation géographique.

- Réaliser une étude spatiale pour déterminer si certaines régions sont plus touchées et, le cas échéant, identifier les sources des différences.

- Visualiser les chutes de météorites sur un planisphère.

Cependant, en raison de la nature des données, nous n'explorons finalement pas toutes ces pistes. Néanmoins, nous dégageons de nouvelles perspectives que nous détaillons par la suite.

Nous débutons ce rapport par une présentation de notre organisation de travail ainsi que des outils, langages et bibliothèques utilisés. Nous poursuivons avec une analyse exploratoire des données, comprenant des analyses univariées et multivariées, qui nous amènent à discuter des pistes à explorer. Ensuite, nous étudions la modélisation des chutes de météorites à l'aide des processus ponctuels, avant de proposer une visualisation interactive du jeu de données sur un globe. Enfin, nous concluons par une réflexion sur l'impact environnemental du projet.

2 Organisation, Outils, bibliothèques `R` et `Python`

Notre projet a débuté en janvier, contrairement aux autres groupes qui ont commencé en octobre. Ce décalage est du au fait que nous ayons initiallement commencé à travailler sur premier projet d'application Shiny pour le Conseil National des Universités, qui n'a finalement pas abouti. Suite à cela, Jean-François Coeurjolly nous a proposé de nous engager dans ce nouveau projet.

Pour assurer un suivi régulier de notre avancement, nous avons organisé six réunions d'environ 45 minutes, toutes les deux semaines, avec notre encadrant. Ces rencontres nous ont permis de faire le point sur nos progrès, de poser des questions et d'ajuster notre travail en fonction des retours reçus.

En ce qui concerne la répartition du travail, chaque membre de l'équipe a contribué de manière spécifique. Garance a pris en charge l'exploration des données et a également collaboré avec Rassem sur la modélisation par les processus ponctuels, où elle s'est concentrée sur la partie mathématique tandis que Rassem s'est occupé de la simulation. Yannis et Duc-Khoi ont chacun travaillé sur la visualisation en 3D, Duc-Khoi utilisant `Python` pour sa partie, tandis que Yannis s'est concentré sur `R`. Pour la rédaction du rapport, chaque membre a rédigé la section correspondant à sa contribution. Garance a également pris en charge l'introduction, la partie organisation et la conclusion, tandis que Rassem a rédigé la section sur l'impact environnemental.

Nous n'avons pas mis en place d'outils d'organisation formels, tels qu'un diagramme Mermaid ou d'autres outils présentés lors de nos cours de gestion de projet que nous avions alors mis en oeuvre pour notre projet inital. Nous avons privilégié des discussions informelles pour définir les tâches à accomplir et faire le point avant chaque rendez-vous avec notre encadrant, chacun travaillant de manière autonome sur sa partie.

Concernant les langages de programmation, nous avons opté pour un mélange de `R` et de `Python` afin de tirer parti des avantages offerts par les deux langages. Pour l'exploration des données, nous avons utilisé `Python` en raison de ses bibliothèques intéressantes pour la création de graphiques interactifs, telles que `Plotly` et `Shapely`. En revanche, nous avons choisi R pour produire des graphiques esthétiques avec `ggplot2` et pour réaliser des tests d'analyse multivariée. Pour les processus ponctuels, `R` s'est avéré être le choix idéal grâce à la bibliothèque `Spatstat`, présentée dans le livre *Spatial Point Patterns - Methodology and Applications with `R`* [AT15] fourni Jean-François Coeurjolly, qui était spécifiquement axé sur `R`. Pour la visualisation 3D, nous avons décidé de travailler sur `R` et `Python`, ce choix est explicité dans la partie correspondante.

Enfin, pour les outils de développement, nous avons hébergé notre code sur GitHub et utilisé Visual Studio Code (VSCode) ainsi que RStudio pour la rédaction de notre code selon les langages utilisés et les préférences de chacun.

Packages `Python` : `geopandas`, `pandas`,`matplotlib`, `numpy`, `plotly` et `shapely`.
Packages `R` : `corrr`, `dplyr`, `ggplot2`, `leaflet`, `patchwork`,`rnaturalearth`, `scales`, `sf`, `shiny`,`shinydashboard`, `shinyjs`, `spastat`, `threejs`, `xtable` et `patchwork`.

3 Exploration des données

L'analyse exploratoire des données constitue la première étape de notre projet. Notre objectif est de comprendre les distributions des variables de manière individuelle tout en identifiant les données manquantes qui pourraient influencer les possibilités d'exploration pour la suite. Nous examinons également les éventuels liens entre les variables au travers de différents tests. Cette démarche nous permet de sélectionner des axes d'exploration pertinents pour la suite de notre projet.

Pour réaliser les analyses univariées, nous travaillons dans un premier temps avec `Python` avec les bibliothèques numpy, pandas, geopandas, plotly et shapely qui nous permettent notamment d'obtenir des planishphères intéractifs mais nous réalisons également certaines analyses en `R` pour obtenir des graphiques avec ggplot2. Pour les analyses multivariées, nous utilisons uniquement `R`. Le code de ces analyses est organisé dans les fichiers `exploration.ipynb` pour la partie en `Python` et `script_analyses_multivariées.R` pour la partie en `R`. Tous ces fichiers sont disponibles sur le repository GitHub du projet dans le dossier "Exploration des données".

Le jeu de données est composé de 45716 entrées décrites par neuf variables :

- **name** (qualitative nominale) : le nom de la météorite.
- **nametype** (qualitative binaire) : le type d'objet soit "Valid" soit "Relict", Relict signifiant qu'il s'agit d'un objet très déformé considéré comme probablement d'origine météorite.
- **recclass** (qualitative nominale) : la classe de la météorite (ex : L5, H6, L5, ...).
- **mass** (quantitative continue) : la masse de la météorite en grammes.
- **fall** (qualitative binaire) : la nature de l'observation soit si la chute de la météorite a été observée (fell) ou si elle a été trouvée au sol (found).
- **year** (quantitative continue) : l'année où la météorite a été recensée.
- **reclat** (quantitative continue) : latitude où la météorite a été trouvée.
- **reclong** (quantitative continue) : longitude où la météorite a été trouvée.
- **geoLocation** : le couple de la latitude et de la longitude.

3.1 Analyses univariées

Nous commençons par l'analyse univariée de chaque variable en regroupant l'analyse des varaibles **reclat**, **reclong** et **geoLocation**.

Name

Il n'y a pas de données manquantes pour le nom des météorites et tous les noms sont uniques. Par curiosité, nous avons regardé la répartition du choix de la première lettre :

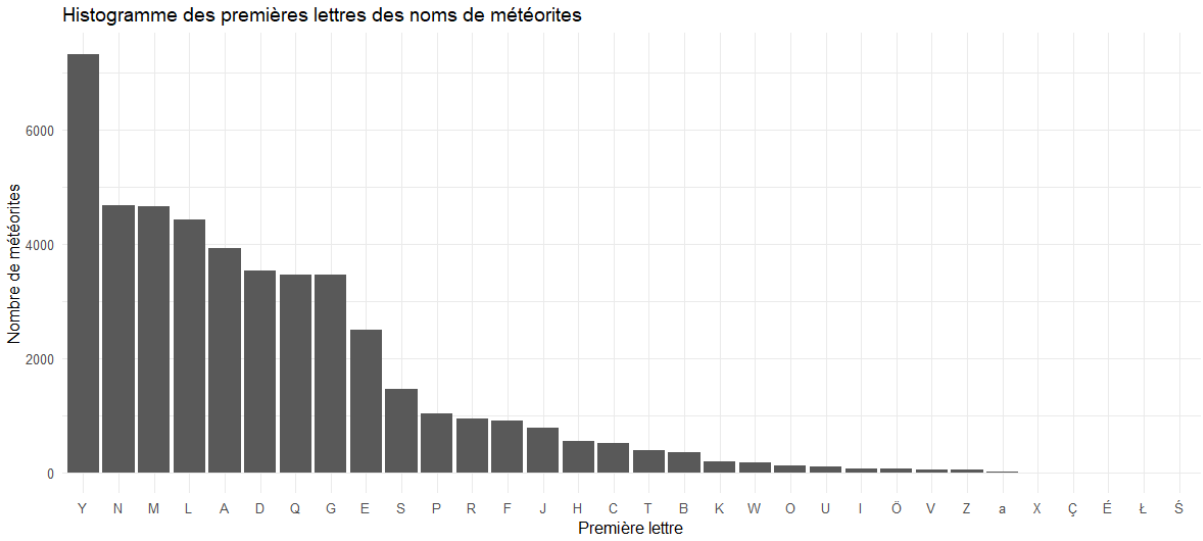


FIGURE 1 – Diagramme en barres du choix de la première lettre du nom des météorites

Nous constatons que la lettre Y ressort nettement plus souvent que les autres. Cela n'est pas une coincidence mais est lié aux conventions de nommage des météorites [Soca] : la grande majorité des météorites sont nommées d'après la localité géographique où elles ont été trouvées avec éventuellement une référence numérique après le nom si de nombreuses météorites sont collectées dans la zone. La popularité de la lettre Y est liée aux nombreuses expéditions japonaises effectuées sur le glacier Yamato en Antartique dont les météorites tiennent leur nom. Sur les 7315 météorites dont le nom commence par la lettre Y, 7269 ont été répertoriées sur le glacier Yamato. De même pour la lettre N, sur 4667 météorites, 4499 météorites commencent par "Northwest Africa" suivi d'un numéro permettant d'identifier la météorite.

Nametype

Cette variable n'a pas de données manquantes. Comme expliqué précédemment, c'est une variable qualitative binaire décrivant si la météorite a bien été identifiée comme valide (Valid) ou s'il s'agit d'un objet fortement déformé qui est probablement d'origine météorite (Relict). Une large majorité de entrées du jeu de données sont considéréees comme valides : 45641 météorites valides soit 99,8% contre 75 "Relict" soit 0,02%.

Recclass

Il n'y a pas de données manquantes pour la variable **recclass** correspondant à la classification de la météorite. Le jeu de données compte 422 classes différentes mais les classes L5, L6, H5, H6, LL5 et LL6 sont nettement majoritaires et représentent près de 74% du jeu de données comme nous pouvons le voir sur la figure ci-dessous :

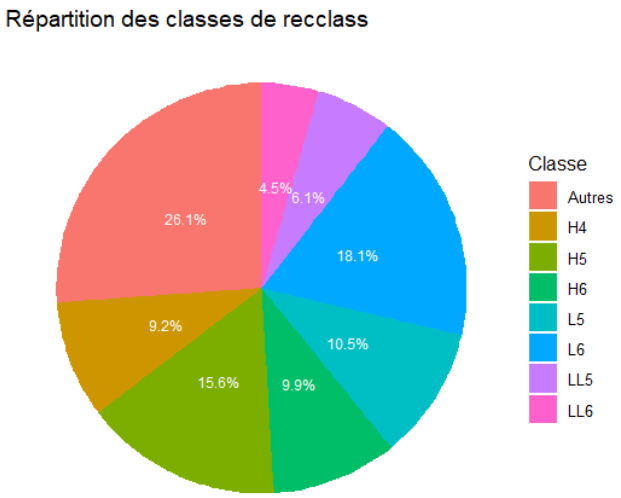
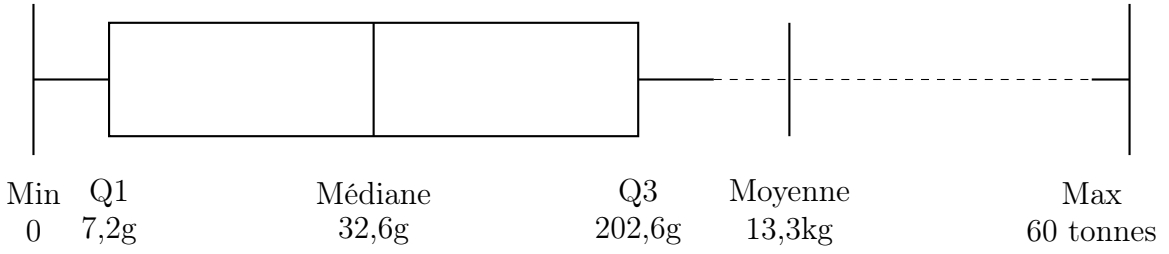


FIGURE 2 – Piechart des classes de météorites

Ces 6 classes sont celles de chondrites ordinaires [SAR]. Les lettres L, H et LL correspondent aux trois sous-groupes de chrondrites ordinaires : H pour "High Iron" pour celles contenant 25-30% de fer et de métal libre, L pour "Low Iron" pour celles contenant 20-25% de fer et moins de métal libre et LL pour celles contenant encore moins de fer (environ 15-20%) et très peu de métal libre. Le numéro qui suit correspond au degré de métamorphisme, compris entre 3 et 7, il indique le niveau des modifications subies par la météorites lors de sa chute dues à la pression et la chaleur modifiant leur composition minéralogique, plus le numéro est élevé plus l'alteration est importante.

Mass

Il manque 131 données pour la variable **Mass** soit moins de 0,3% du jeu de données. L'analyse de cette variable révèle tout d'abord une très grande hétérogénéité dans les valeurs possibles : plusieurs milliers de météorites pèsent moins de cinq grammes tandis que la plus grosse pèse 60 tonnes.



Près de 75% des météorites pèsent moins de 200 grammes mais la moyenne est à 13kg : la très large majorité des météorites rescencées sont très légères mais le jeu de données rescence quelques météorites très massives qui expliquent donc la moyenne très élevée. Dans les faits, seules 1388 météorites font plus de 10kg soit environ 3% du jeu de données. Cela est lié au fait que lors de leur chute les météorites se vaporisent et se fragmentent sous l'effet de la pression et de la chaleur et au fait qu'elles puissent

également se fragmenter lors de l'impact. De plus, en réalité, la proportion de (très) petites météorites est sûrement encore plus importantes puisque les météorites les plus massives ont plus de chances d'être découvertes tandis que les petites météorites passent inapercues et peuvent être aisement confondues avec des roches terrestres empêchant ainsi leur rescencement.

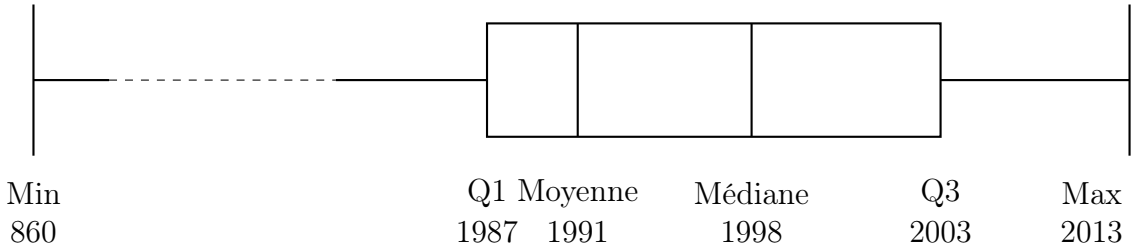
Fall

Il n'y a pas de données manquantes pour cette variable. La très large majorité des météorites ont été trouvées (Found) et peu de météorites ont été obeservées lors de leur chûte (Fell), elles représentent seuelemnt 2,5% du jeu de données.

Year

Pour cette variable, on compte 291 entrée manquantes ainsi qu'une donnée abérrante d'une météorites répertoriée comme tombée en 2101 que nous excluons de notre analyse.

Comme pour la variable **Mass**, il y a de grandes disparités dans la distribution des valeurs :



Nous pouvons tout de suite voir que la grande majorité des valeurs sont très récentes : même si le jeu de données contient des météorites datant du IXème siècle, plus de 75% des météorites rescencées datent d'il y a moins de 50 ans. Cependant, cette variable nous permet de nous rendre compte que le jeu de données n'est pas récent ou pas mis-à-jour récemment puisque la dernière météorite rescencée date de 2013 bien que d'autres météorites soient tombées sur Terre et aient été trouvées depuis.

Intéressons nous alors plus précisément aux 50 dernières années du jeu de données (1963-2013) :

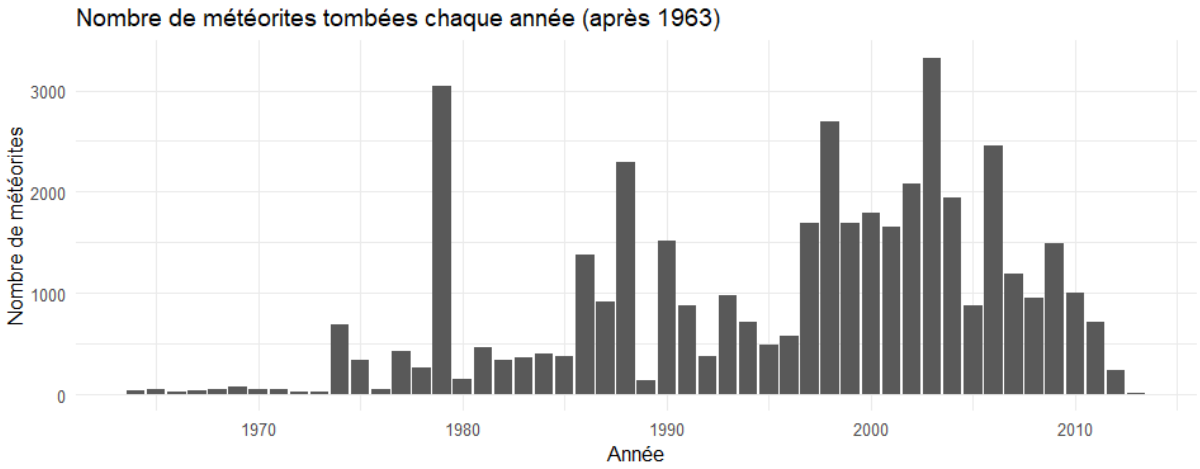


FIGURE 3 – Répartition des météorites tombées entre 1963 et 2013

Nous pouvons observer une nette augmentation du nombre de météorites recensées à partir de 1974. Nous pouvons alors nous poser la question de la source de l'absence de données dans les années précédentes : est-ce que réellement peu de météorites sont tombées ou n'ont-elles juste pas été référencées ? Avant les années 1970, la détection de météorites reposaient principalement sur des découvertes fortuites ou des témoignages. Les années 1970 ont vu l'essor des réseaux de surveillance qui ont permis un suivi plus rigoureux des météorites comme le Meteorite Observation and Recovery Projet [al89] lancé au Canada en 1974 ainsi que la mise en place de bases de données centralisées comme celle de la *Meteoritical Society* [Socb] et de missions de recherche comme le programme *The Antarctic Search for Meteorites* [Uni] lancé au milieu des années 1970. Ainsi, l'hypothèse que les météorites n'étaient pas correctement recensées avant les années 1970 semble assez cohérente.

Location

Cette variable est celle avec le plus de données manquantes : 7315 entrées manquant à chaque fois de la donnée de longitude, de latitude et de geolocation, soit environ 16% du jeu de données.

Dans un premier temps, nous pouvons visualiser sur le planisphère où sont tombées les météorites répertoriées :

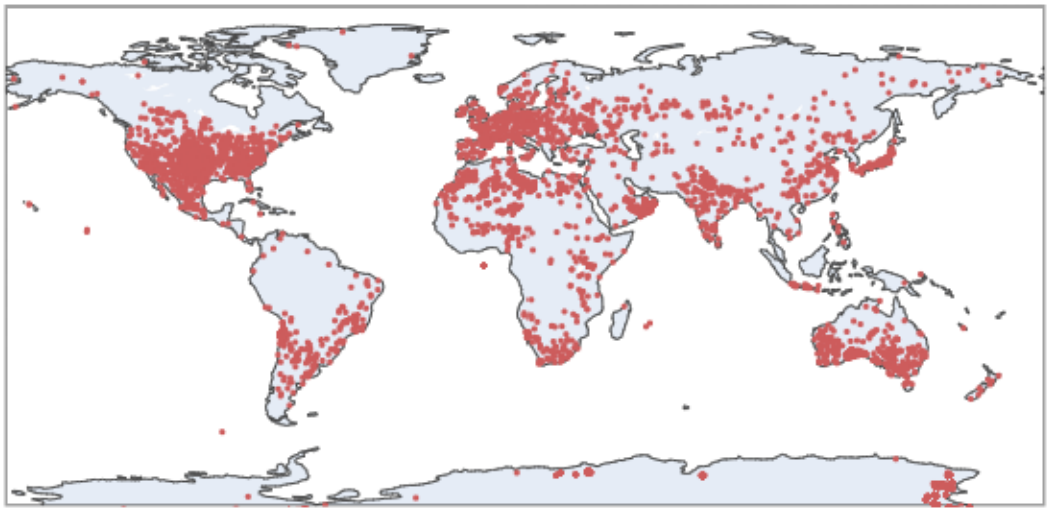


FIGURE 4 – Emplacement des météorites

Nous pouvons observer que certaines régions, comme les États-Unis, l'Europe de l'Ouest, le Japon, l'Afrique du Nord et le sud de l'Australie, concentrent une grande partie des recensements. À l'inverse, d'autres zones, telles que le nord du Canada, la Russie ou la forêt amazonienne, sont presque dépourvues de données. Sans surprise, les zones maritimes sont totalement absentes de notre jeu de données, puisqu'il ne répertorie que les météorites tombées sur des terres émergées.

Cette répartition inégale met en évidence un biais de recensement important : les régions à faible densité de population, comme la forêt amazonienne ou les territoires du nord du Canada, comptent peu ou pas d'observations, contrairement aux zones fortement peuplées, comme l'Europe de l'Ouest, où les découvertes sont bien plus fréquentes. La seule exception notable à cette tendance est l'Antarctique, où de nombreux recensements ont été réalisés. Cette spécificité s'explique par les nombreuses missions de recherche dédiées à l'étude des météorites dans cette région [Uni]. En effet, les déserts glacées de l'Antarctique représentent des conditions idéales pour l'identification des météorites, facilitant leur repérage et leur collecte.

Nous pouvons compléter cette première étude par une étude séparée de la latitude et de la longitude :

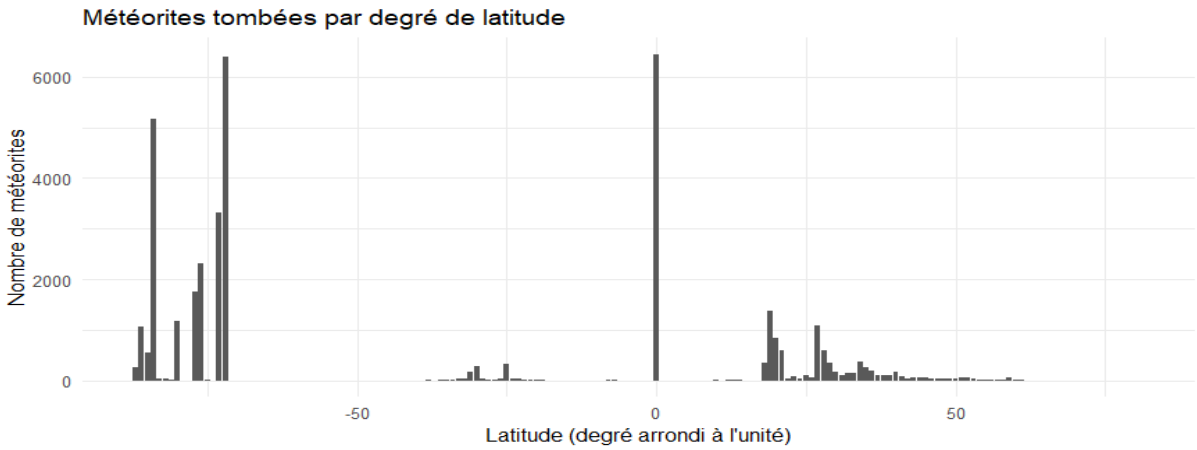


FIGURE 5 – Latitude des météorites

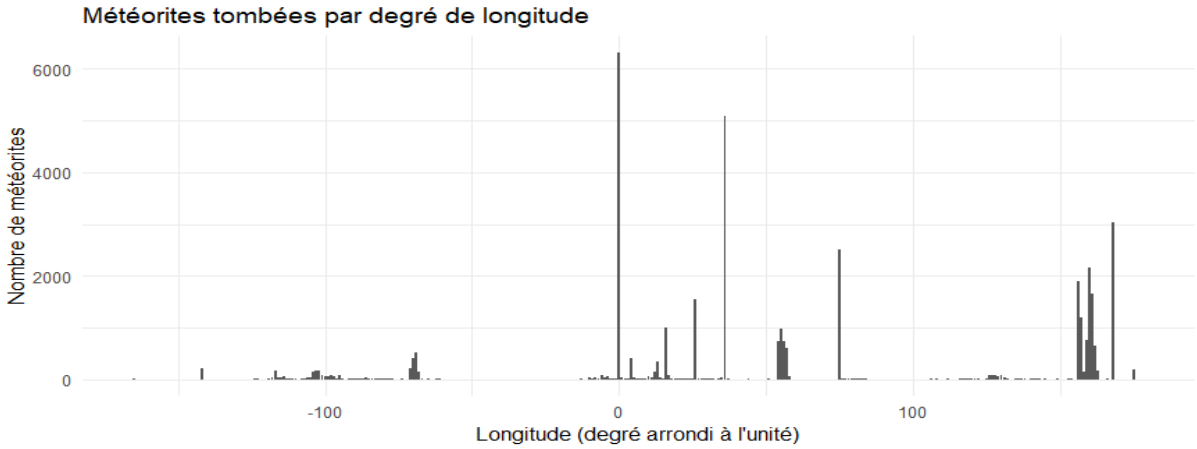


FIGURE 6 – Longitutde des météorites

L'histogramme des latitudes montre une distribution inégale avec plusieurs tendances marquées. On observe un nombre particulièrement élevé d'impacts aux latitudes inférieures à -70°, correspondant aux nombreux points recencés en Antarctique. Un autre pic notable apparaît autour de 0°, ce qui pourrait être associé à l'Afrique du Nord. En revanche, la répartition est plus diffuse entre 15° et 60° Nord, couvrant des régions comme l'Europe de l'Ouest et l'Amérique du Nord.

L'analyse des longitudes révèle également une répartition très hétérogène. Un pic marqué est visible autour de 0°, correspondant notamment à l'Europe de l'Ouest et au nord de l'Afrique. D'autres concentrations apparaissent entre -150° et -162°, suggérant une fréquence plus élevée d'impacts en Australie et en Antarctique.

Nous analysons ensuite la répartition du nombre de météorites par pays à l'aide des contours des pays proposé par le projet *Natural Earth* [Ear] :

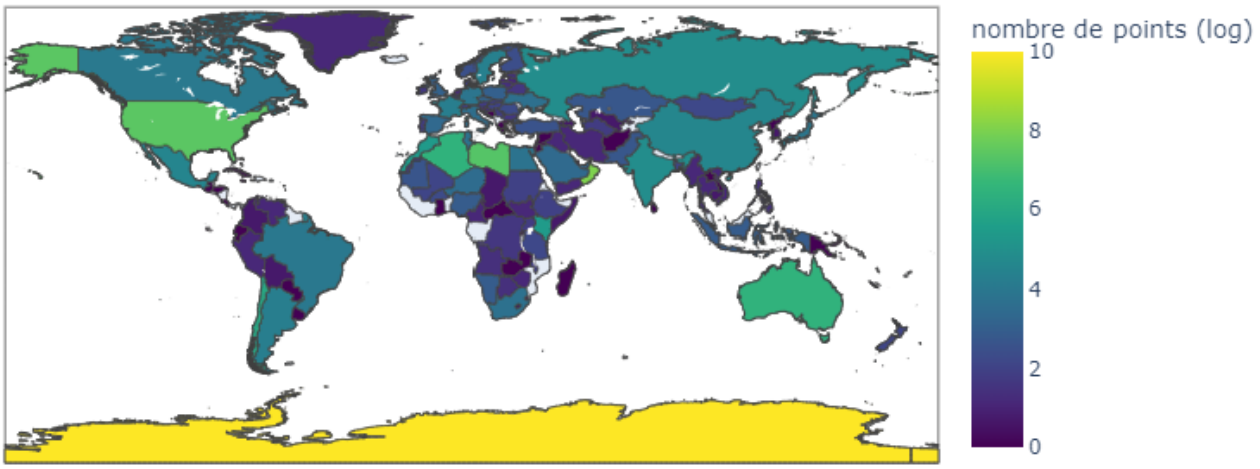


FIGURE 7 – Nombre de météorites recensées par pays

Sans surprise, l'Antarctique est le territoire où le plus grand nombre de météorites a été recensé, suivi notamment des États-Unis, de la Libye, d'Oman, de l'Australie et de l'Algérie qui correspondent bien aux régions identifiées précédemment. Cependant, cette première visualisation ne prend pas en compte la taille des pays qui varie considérablement (par exemple, 7 700 000 km² pour l'Australie contre 309 000 km² pour l'Oman), ce qui peut biaiser l'interprétation des résultats. Afin de mieux appréhender cette distribution, nous calculons le nombre de météorites rapporté à la superficie du pays en km² :

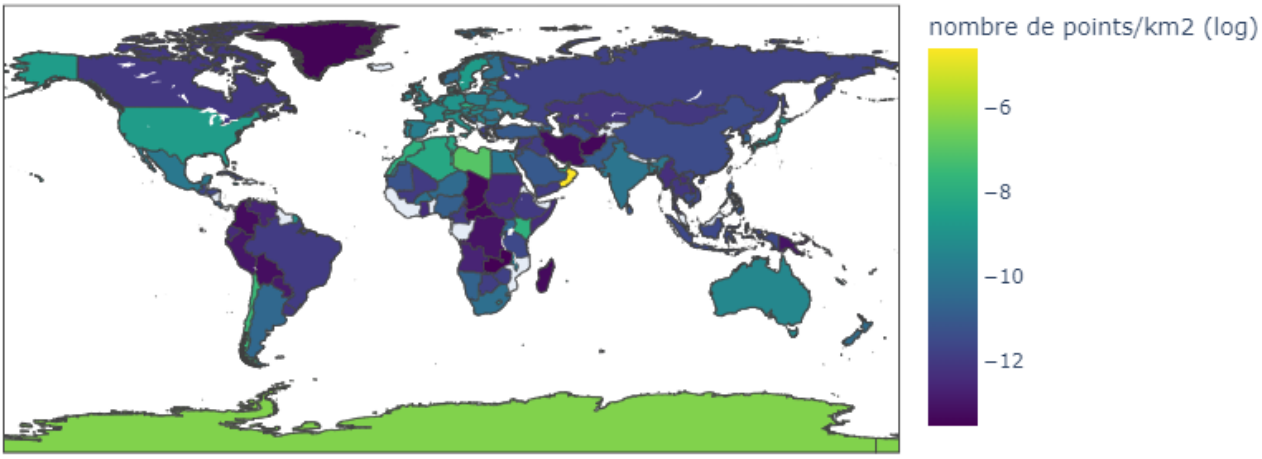


FIGURE 8 – Densité de météorites recensées par km² et par pays

Cette nouvelle représentation met en évidence plusieurs tendances intéressantes : Oman ressort davantage que dans la carte précédente, tandis que l'Australie et l'Antarctique, bien qu'ayant un grand nombre de météorites recensées, apparaissent moins dominantes lorsque nous tenons compte de leur superficie. Par ailleurs, les pays d'Europe de l'Ouest présentent des valeurs comparables à celles des États-Unis et des pays d'Afrique du Nord-Ouest.

Enfin, nous avons également regardé la répartition des météorites dans le monde selon la masse, mais aucune tendance n'est ressortie.

3.2 Analyses multivariées

Nous poursuivons notre analyse du jeu de donnée par des analyses multivariées par le biais de tests sur les croisements des différentes variables entre elles.

Analyse qualitative-qualitative

Pour le croisement de variables qualitatives entre elles, nous effectuons des tests de χ^2. Pour chaque croisement, nous donnons dans le tableau suivant la p-value associée au résultat du test.

	Variables	p_value
1	nametype - recclass	$< 2 \times 10^{-16}$
2	nametype - fall	0.9085
3	recclass - fall	$< 2 \times 10^{-16}$

Analyse quantitative-quantitative

Pour le croisement de variables quantitatives entre elles, nous regardons cette fois le coefficient de corrélation de Pearson. Les résultats sont résumés dans le tableau ci-contre :

	term	mass..g.	year	reclat	reclong
1	mass..g.		-0.1219	0.0292	-0.0219
2	year	-0.1219		-0.1050	0.0903
3	reclat	0.0292	-0.1050		-0.5932
4	reclong	-0.0219	0.0903	-0.5932	

A priori, aucune des variables ne sont très corrélées, la plus forte corrélation est entre la latitude et la longitude à -0.59 qui reste relativement faible et peut s'expliquer par le biais de recensement identifié lors des analyses univariées.

Analyse quantitative-qualitative

Avant de faire les croisement entre variables quantitatives et qualitatives, nous testons l'hypothèse de normalité des données avec un test de Shapiro-Wilk. Pour l'ensemble des variables, le test ressort toujours comme étant très significatif (p-valeur $< 2 \times 10^{-16}$). Nous effectuons donc un test de Kruskal-Wallis ou de Mann-Whitney pour tester le lien, ces deux tests étant robustes à l'absence de l'hypothèse de normalité.

	Variable quantitative	Variable qualitative	p-value
1	mass..g.	nametype	$< 2 \times 10^{-16}$
2	year	nametype	$< 2 \times 10^{-16}$
3	reclat	nametype	$< 2 \times 10^{-16}$
4	reclong	nametype	0.1026
5	mass..g.	recclass	$< 2 \times 10^{-16}$
6	year	recclass	$< 2 \times 10^{-16}$
7	reclat	recclass	$< 2 \times 10^{-16}$
8	reclong	recclass	$< 2 \times 10^{-16}$
9	mass..g.	fall	$< 2 \times 10^{-16}$
10	year	fall	$< 2 \times 10^{-16}$
11	reclat	fall	$< 2 \times 10^{-16}$
12	reclong	fall	$< 2 \times 10^{-16}$

La plupart des croisements sont très significatifs.

3.3 Discussion des limites du jeu de données

L'étude des données manquantes révèle que seules trois variables sont concernées : la masse des météorites, l'année de leur chute et leur localisation. Malgré ces lacunes, en supprimant toutes les entrées comportant au moins une valeur manquante, il reste 38 115 observations, soit environ 83% du jeu de données initial. Ce volume semble a priori suffisant pour mener des analyses pertinentes.

Cependant, l'analyse univariée des variables, en particulier l'emplacement et l'année de chute, met en évidence plusieurs biais significatifs dans les données. D'une part, la présence d'un biais temporel : les météorites récentes sont surreprésentées en raison de l'amélioration du suivi, de la centralisation des bases de données et du développement de missions dédiées à leur recherche depuis les années 1970. D'autre part, un biais géographique est également présent : les météorites sont davantage recensées dans les zones densément peuplées, à l'exception notable de l'Antarctique, mais où le grand nombre de recensements s'explique par des campagnes de recherche intensives menées là bas. Enfin, il est probable qu'un biais en faveur des météorites les plus massives existe, celles-ci étant plus faciles à identifier et à distinguer des roches environnantes. Il semble donc évident que toutes les météorites tombées sur le sol de la Terre ne sont pas répertoriées et que le jeu de données est loin d'être complet.

Un problème majeur réside alors dans le manque d'informations sur la constitution du jeu de données : le site de la NASA où nous avons récupéré nos données (ainsi que d'autres sources en ligne) ne précise pas si les données proviennent uniquement d'observations scientifiques (télescopes, chercheurs, etc) ou si elles incluent des signalements grand public. Cette incertitude complique l'interprétation des analyses et ne nous permet pas de palier aux biais identifiés précèdemment.

Concernant les pistes inialement envisagées, nous en écartons plusieurs :

- Analyse temporelle : Cette approche se révèle impraticable pour plusieurs raisons.

Tout d'abord, les données disponibles ne contiennent que l'année de chute des météorites, sans précision sur le mois ou le jour. De plus, le biais temporel implique que seules les quarante dernières années sont véritablement exploitables. Cela représente une période trop courte pour une analyse temporelle robuste. L'exploration du jeu de données du Natural History Museum [Mus] inculant des informations plus détaillées sur la temporalité (mois et jour) s'avère également problématique : la précision temporelle reste insuffisante (mois parfois indiqués sous des catégories larges comme Printemps ou Juin-Août), et le nombre d'observations réellement exploitables est trop faible en raison de la rareté des données bien documentées. Cette piste est donc abandonnée.

- Modélisation prédictive : Construire un modèle visant à prédire le nombre de météorites tombées dans un pays en fonction de sa superficie et de sa position géographique ne nous semble pas pertinent à cause de la présence d'un fort biais de recensement faussant la relation entre les variables.

Finalement, nous décidons donc de nous concentrer sur deux axes de travail principaux :

- Une analyse du processus derrière la chute des météorites, en testant si leur distribution suit un processus ponctuel de Poisson homogène et inhomogène par le biais de simulations.

- Une visualisation en 3D, afin de se détacher des distorsions engendrées par la projection sur le planisphère et d'obtenir une représentation plus fidèle de la répartition des météorites.

4 Modélisation de processus ponctuels

Nous avons vu lors de l'exploration de données que les points de chutes des météorites ne sont pas répartis de manière uniforme sur le globle. Nous nous demandons si par contre elles sont indépendantes ou si peut-être nous pouvons observer des phénomènes de "repoussement" ou à l'inverse de concentration de météorites dans une espace très restreint.

Pour répondre à ce questionnement, nous simulons des processus ponctuels de Poisson où les points sont répartis aléatoirement dans l'espace mais surtout sont placés de manière indépendante les uns des autres. Nous souhaitons voir si nos données sont similaires aux réalisations d'un processus de Poisson homogène ou inhomogène, ce qui nous permettrait d'avoir une idée sur l'indépendance des chutes de météorites.

Nous commençons par le processus poncutel de Poisson homogène très simple pour prendre en main la méthode puis nous passon au cas inhomogène plus complexe mais qui peut mieux coller à nos données et au biais de rescensement observé précédemment..

Dans ces deux approches, nous commençons par développer la théorie mathématique que nous appliquons ensuite à nos données au travers de simulations. Cette partie du

projet s'appuie sur le livre *Analysing spatial point patterns in R* [AT15].

4.1 Définitions générales des processus ponctuels et distribution uniforme

Pour simplifier notre analyse, nous nous plaçons en dimension 2 et nous travaillons donc sur $\mathbb{R}^2$ pour représenter le planisphère plutôt qu'en dimension 3 sur la surface du globe. Commençons par quelques définitions :

Définition. *Un* ***modèle de points*** $\boldsymbol{x}$ *est un ensemble de points dans* $\mathbb{R}^2$*. On note* $n(\boldsymbol{x})$ *le nombre d'éléments dans cet ensemble et* $n(\boldsymbol{x} \cap B)$ *le nombre d'éléments de l'ensemble dans la région* B*.*

Remarque. *Il est permis d'avoir deux points d'un même modèle de point ayant les mêmes coordonnées.*

Définition. *Un* ***processus ponctuel*** $\boldsymbol{X}$ *est un méchanisme aléatoire dont les réalisations sont des modèles de points.*

Définition. *Un processus ponctuel* ***fini*** *est un méchanisme aléatoire* $\boldsymbol{X}$ *tel que :*

1. *Ses réalisations sont des modèles de points avec un nombre fini d'éléments.*
2. *Pour toute région* B *bornée et fermée, le nombre* $n(\boldsymbol{X} \cap B)$ *de points tombants dans la région* B *est une variable aléatoire bien définie.*

Pour notre analyse, nous travaillerons sur des processus ponctuel de Poisson, nous aurons donc besoin de la notion de processus localement fini.

Définition. *Un modèle de point* ***localement fini*** *est un ensemble* $\boldsymbol{x} = \{x_1, x_2, \dots\}$ *de points dans* $\mathbb{R}^2$ *tel que pour toute région* B*,* $n(\boldsymbol{x} \cap B)$ *est fini même si* $\boldsymbol{x}$ *n'est pas fini.*

Définition. *Un processus ponctuel* ***localement fini*** $\boldsymbol{X}$ *est un méchanisme aléatoire tel que*

1. *Ses réalisations sont des modèles de points* ***localement fini****.*
2. *Pour toute région* B *bornée et fermée, le nombre* $n(\boldsymbol{X} \cap B)$ *de points tombants dans la région* B *est une variable aléatoire bien définie.*

4.2 Processus ponctuel de Poisson

Définition et propriétés

Pour définir les processus ponctuels de Poisson, nous commençons par décrire "avec les mains" les trois propriétés qui le caractérise. Nous donnons la définition plus rigoureuse ensuite.

Nous nous donnons un nombre aléatoire de points, pas forcément fini et nous souhaitons construire un processus ponctuel totalement aléatoire, pour cela, nous le munissons de trois propriétés :

1. Homogénéité :

Nous souhaitons que les points n'aient aucune préférence spatiale. C'est-à-dire que le nombre de points attendus dans une région quelconque B soit être proportionnel à son aire :

$$\mathbb{E}[n(\mathbf{X} \cap B] = \lambda |B| \quad \text{où } \lambda \text{ est une constante.}$$

Dans les faits λ est la moyenne de points par unité d'aire. On l'apppelle **l'intensité du processus ponctuel**.

2. Indépendance :
Nous souhaitons que l'information sur les réalisations d'une région n'ait pas d'influence sur les réalisations dans les autres régions. C'est à dire que pour deux région A et B, $n(\mathbf{X} \cap A)$ et $n(\mathbf{X} \cap B))$ sont deux variables aléatoires indépendantes : la valeur de $n(\mathbf{X} \cap A))$ n'a pas d'influcence sur les probabilités des différentes valeurs de $n(\mathbf{X} \cap B))$. On souhaite que cette hypothèse d'indépendance s'applique à n'importe quelles régions disjointes A et B et à n'importe quel nombre de ses régions.

3. Ordonnancement :
Nous souhaitons que la probas qu'une région A contienne plus d'un point soit négligable lorsque la région est suffisament petite. C'est à dire : $\frac{\mathbb{P}(n(\mathbf{X} \cap A))}{|A|} \underset{|A| \to 0}{\longrightarrow} 0$. Cela correspond en fait au fait que $n(\mathbf{X} \cap A))$ suive une loi de Poisson : $\mathbb{P}(n(\mathbf{X} \cap B) = k) = e^{-\mu} \frac{\mu^k}{k!}$ avec μ la moyenne de la densité de Poisson. Donc $\mathbb{E}[n(\mathbf{X} \cap A)] = \mu = \lambda |A|$

Plus rigoureusement :

Définition. *Un **processus ponctuel de Poisson homogène $\boldsymbol{X}$ d'intensité** $\lambda > 0$ est un processus ponctuel localement fini vérifiant :*

1. ***Homogénéité** : Pour toute région B, la valeur moyenne du nombre de points dans la région B $\mathbb{E}[n(\boldsymbol{X} \cap B)] = \lambda |B|$.*
2. ***Indépendance** : Pour m régions de tests $B_1, B_2, \ldots, B_m$ telles que $\forall i \neq j \in \{1, \ldots, m\} B_1 \cap B_j = \varnothing$, $n(\boldsymbol{X} \cap B_1), \ldots, n(\boldsymbol{X} \cap B_m)$ sont des variables aléatoires bien définies*
3. ***Distribution de Poisson** : pour toute région B, $n(\boldsymbol{X} \cap B$ suit une distribution de Poisson.*

Plusieurs propriétés découlent de cette définition :

Proposition 4.1. *Soit B une région de $\mathbb{R}^2$ telle que $n(\boldsymbol{X} \cap B) = n$. La **propriété conditionnelle** d'un processus de Poisson est que ces n points sont indépendemment et uniformément distribués dans B.*

Proposition 4.2. *Réduire dans le cadre des processus ponctuels signifie que l'on supprime quelques points d'un modèle de points. Dans le cas d'une réduction complètement aléatoire chaque point du modèle de point est aléatoirement supprimé ou conservé avec une probabilité p d'être conservé indépendemment de ce qu'il advient des autres points. La **propriété de réduction** affirme que pour un processus de Poisson homogène d'intensité λ au quel on applique une réduction complètement aléatoire avec une probabilité de conservation p alors les points obtenus suivent un processus de Poisson homogène d'intensité λp.*

Proposition 4.3. *Superposer deux processus ponctuels* $\boldsymbol{X}$ *et* $\boldsymbol{Y}$ *signifie que l'on combine les points des deux processus en un nouveau processus ponctuel* $\boldsymbol{Z} = \boldsymbol{X} \cup \boldsymbol{Y}$. *La* ***propriété de superposition*** *des processus de Poisson affirme que si* $\boldsymbol{X}$ *et* $\boldsymbol{Y}$ *sont des processus de Poisson homogène de d'intensité* $\lambda_{\boldsymbol{X}}$ *et* $\lambda_{\boldsymbol{Y}}$ *alors leur superposition* $\boldsymbol{Z}$ *est aussi un processus de Poisson homogène d'intesité* $\lambda_{\boldsymbol{Z}} = \lambda_{\boldsymbol{X}} + \lambda_{\boldsymbol{Y}}$

On peut aisément simuler un processus de Poisson en utilisant les propriétés précédentes. Pour une région B où l'on souhaite générer des points suivant une intensité λ, on commence par déterminer le nombre total de points en générant un nombre aléatoire N suivant une loi de Poisson de moyenne $\mu = \lambda|B|$. Ces N points sont ensuite placés indépendemment dans B de manière aléatoire.

Bien que les processus ponctuel de Poisson soient très simples, ils restent réalistes pour un certain nombre de phénomènes réels tels que la radioactivité, les évènements rares ou les évènements extrèmes ce qui a motivé leur utilisation pour notre étude sur les chutes de météorites.

Ils sont également utilisés comme modèles de références pour faire des comparaisons à d'autres modèles. Ils sont utilisés comme hypothèse nulle pour les tests statistiques.

Enfin, d'autres modèles sont construits à partir de lui comme le processus ponctuel de Poisson inhomogène que nous verrons par la suite.

Simulations

4.3 Processus ponctuel de Poisson inhomogène

Définition et propriétés

Les processus de Poisson inhomogènes généralisent légèrement les processus homogènes de Poisson en faisant l'hypothèse que la densité de points n'est pas la même sur tout l'espace et qu'il s'agit d'une fonction $\lambda(u)$ d'un point u de l'espace. Pour une région B, le nombre de points attendus dans la région est $\int_B \lambda(u)du$. Cette fonction détermine l'abondance générale de points et leur répartition dans l'espace. Plus rigoureusement :

Définition. *Un* ***processus ponctuel de Poisson inhomogène*** *d'intensité* $\lambda(u)$ *est un processus ponctuel tel que :*

1. ***Fonction d'intensité*** *: Le nombre de points attendus dans une région* B *est* $\mu = \int_B \lambda(u)du$

2. ***Indépendance*** *: Si l'espace est découpé en région ne se croisant pas, les modèles aléatoires au sein de chaque région sont indépendants entre eux.*

3. ***Distribution de Poisson*** *: Le nombre de points tombants dans une région donnée suit une distribution de Poisson.*

C'est un modèle très général comme dans le cas homogène puisqu'il y a très peu de restriction sur $\lambda(u)$ ce qui rend ce modèle très adéquat dans beaucoup de contexte. La

pricipale hypothèse devient donc l'indépendance des points.

Les processus inhomogène de Poisson possèdent des propriétés analogues à celle du cas homogène :

Proposition 4.4. ***Propriété conditionnelle*** *: Soit un processus ponctuel inhomogène de Poisson de fonction d'intensité $\lambda(u)$. Si n points sont tombés dans une région B donnée alors ces points sont indépendants entre eux et chaque point a la même probabilité de distribution sur B : $f(u) = \lambda(u)/\mu$ où $\mu = \int_B \lambda(u)du$.*

Proposition 4.5. ***Propriété de réduction*** *: Soit un processus ponctuel inhomogène de Poisson de fonction d'intensité $\lambda(u)$ qui est aléatoirement réduit avec une probabilité $p(u)$ de conserver un point situé en le point u de l'espace alors le modèle de point obtenu suit également un processus de Poisson inhomogène de fonction d'intensité $\lambda(u)p(u)$.*

Proposition 4.6. ***Propriété de superposition*** *: Soit $\boldsymbol{X}$ et $\boldsymbol{Y}$ deux processus de Poisson inhomogènes et de fonction d'intensité respectives $\lambda_{\boldsymbol{X}}(u)$ et $\lambda_{\boldsymbol{Y}}(u)$ alors leur superposition $\boldsymbol{Z} = \boldsymbol{X} \cup \boldsymbol{Y}$ est également un processus de Poisson inhomogène de fonction d'intensité $\lambda_{\boldsymbol{Z}}(u) = \lambda_{\boldsymbol{X}}(u) + \lambda_{\boldsymbol{Y}}(u)$.*

On peut aisément simuler un processus de Poisson inhomogène de fonction d'intensité $\lambda(u)$ à l'aide des propriétés précédemment présentées et en utilisant la méthode de réduction de Lewis-Shedler. Cette méthode consiste à trouver un majorant M tel que pour tout u dans la région où l'on souhaite générer les points $\lambda(u) \leq M$. On génère alors un processus de Poisson homogène d'intensité M comme décrit dans la section sur les processus de Poisson homogènes. Pour chaque point x_i du modèle de points obtenu, on calcule $p_i = \lambda(x_i)/M$ et on conserve aléatoirement x_i avec une probabilité p_i indépendamment de ce qu'il advient des autres points. On obtient alors une réalisation d'un processus de Poisson de fonction d'intensité $\lambda(u)$. Une autre méthode possible est de diviser l'espace en pixels, de calculer la probabilité que chaque pixel contienne un point and de sélectionner aléatoirement des pixels en utilisant ces probabilités. Cette méthode demande d'avantage de calculs que la méthode de Lewis-Shedler mais nécessite moins de calculs pour chaque point.

Simulations

5 Visualisation en 3 dimensions

Travail de Yannis et Duc-Khoi.

- Choix de R + `Python` : voir les limites des deux langages, les comparer et utiliser leurs forces et faiblesses pour des applications différentes, Jeff est plus à l'aise en R et les ressources qu'il nous a données sont en R, ... - Réexpliquer l'intérêt d'une visualisation end 3D par rapport à une visualisation sur un planisphère (si possible avec des sources). - Pour les deux, mettre *QUELQUES* visuels (utiliser la possibilité de faire des colonnes notamment). - Pour les deux, expliquer les forces et les faiblesses, les limitations rencontrées.

5.1 Visualisation avec `Python`

5.2 Visualisation avec `R`

6 Impact environnemental et sociétal du projet

J'ai remis les consignes du pdf de l'Ensimag. Cette section doit représenter envrion 20% du rapport.

Impact environnemental personnel

Partie moins importante. Estimation de l'impact des trajets domicile-travail, impact de la consommation des équipements utiliés (ordinateurs perso/fixes, temps d'utilisation des serveurs github,...), autres impacts. Expression en exprimé en kg eq. CO2.

Impact global du projet

Dans cette section, nous vous demandons d'évaluer l'impact global du projet sur lequel vous avez travaillé. Si vous avez travaillé sur un produit fini (logiciel, infrastructure...), vous devrez mettre en valeur non seulement l'impact du produit lui-même mais également l'évolution de cet impact entre le début et la fin de votre PFE. Si vous avez travaillé sur une preuve de concept, un avant-projet, un projet de recherche et développement ou un projet de recherche pure, votre évaluation devra tenir compte des possibles utilisations de votre travail dans un contexte applicatif. Cette section sera la plus importante de la partie consacrée à l'impact environnemental et sociétal. Nous ne vous demandons pas une simple évaluation technique, mais une véritable réflexion déclinée sur deux plans : 1. à petite échelle (concernant uniquement votre projet, à court terme) 2. à plus grande échelle (long terme, et dans l'hypothèse où le même type de projet venait à se généraliser et/ou se transposer dans différents secteurs) Nous demandons dans cette section un avis honnête, critique et argumenté sur les impacts positifs et négatifs du projet. Vous ne serez pas évalué sur la quantité ni la qualité des bonnes pratiques sociales et environnementales mises en œuvre dans le cadre de votre PFE : il est donc inutile d'écoblanchir votre discours. Ce qui nous importe est la vision critique que vous adoptez.

Politique de la structure d'acceuil

Dans cette section, nous vous demandons de dresser une liste des actions menées par la structure d'accueil sur les aspects écologiques et sociaux. Cela peut concerner des actions individuelles ou la mise en œuvre d'une véritable politique dans ce domaine. De même, cela concerne à la fois des politiques extérieures éventuelles (fondations, dons à des organismes...), mais également des actions destinées à l'ensemble des collaborateurs de l'entreprise (conditions de travail, mise en œuvre de bonnes pratiques environnementales au quotidien...). Vous mettrez bien entendu en évidence tous les aspects positifs de cette politique. En revanche, si vous estimez qu'il y a des voies d'amélioration possibles en termes de politique de responsabilité sociale et environnementale, nous vous encourageons à proposer une liste d'actions concrètes qui pourraient être mises en œuvre. Cela

montrera non seulement votre capacité à réaliser une analyse critique, mais cela vous permettra également d'être une force de proposition pour votre structure d'accueil.

7 Conclusion

Références

[al89] Ian Halliday et AL. « The flux of meteorites on the Earth's surface ». In : *Meteoritics Planetary Science* (1989). DOI : https://doi.org/10.1111/j.1945-5100.1989.tb00959.x.

[AT15] Ege Rubak ADRIAN BADDELEY et Rolf TURNER. *Analysing spatial point patterns in R*. Chapman Hall, 2015.

[Ear] Natural EARTH. *Natural Earth - Free vector and raster map data*. https://www.naturalearthdata.com/. Visité le 11/04/25.

[Mus] Natural History MUSEUM. *Catalogue of Meteorites (MetCat)*. https://data.nhm.ac.uk/dataset/metcat/resource/96dc3c09-49fd-4af6-b2fb-5a48a76d09ee. Visité le 30/03/25.

[NAS] NASA. *Meteorite Landings*. https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh/about_data. Visité le 30/03/25.

[SAR] The Earth's Memory SARL. *Meteorites.fr - Classification*. http://www.meteorite.fr/en/classification/. Visité le 30/03/25.

[Soca] The Meteoritical SOCIETY. *GUIDELINES FOR METEORITE NOMENCLATURE*. https://www.lpi.usra.edu/meteor/docs/nc-guidelines-2015-february.htm. Visité le 30/03/25.

[Socb] The Meteoritical SOCIETY. *Meteoritical Bulletin*. https://www.lpi.usra.edu/meteor/metbull.php. Visité le 30/03/25.

[Uni] Case Western Reserve UNIVERSITY. *ANSMET, The Antarctic Search for Meteorites*. https://caslabs.case.edu/ansmet/faqs/. Visité le 30/03/25.

8 Annexe