

Étude de la notion de moyenne dans un ensemble géographique complexe : la Croatie

Estelle ABOU TAYEH, Anais AIT ABDELKADER
Felicia DOSSOU, Arthur LE CAMUS, Elie GAUFFRE

Avril 2024
Supervisé par M. RICHOU

Table des matières

1	Introduction	2
1.1	Problématique	2
1.2	Objectifs, contraintes	2
1.3	Définitions des concepts	2
1.4	Méthodologie	4
2	Approche avec la moyenne euclidienne	5
2.1	Moyenne euclidienne de la population	5
2.2	Introduction à la loi des grands nombres	6
2.2.1	Exemple dans le cas discret	9
2.2.2	Cas des variables aléatoires à densité	9
3	La moyenne de Fréchet	11
3.1	Introduction de la géodésique	12
3.1.1	Approche avec l'algorithme Dijkstra	14
3.1.2	Approche avec l'algorithme des lignes brisées	17
3.2	Approximation numérique de la moyenne de Fréchet	18
3.2.1	Exemple d'un cas des variables aléatoires discrètes	18
3.2.2	Exemple d'un cas des variables aléatoires à densité	19
3.2.3	Comparaison des deux algorithmes	20
4	Conclusion	21

1 Introduction

1.1 Problématique

Comment définir et calculer une moyenne dans un ensemble non convexe ?

1.2 Objectifs, contraintes

Imaginons que l'on cherche à calculer la moyenne d'une variable aléatoire prenant ces valeurs dans la Croatie.

Le caractère fortement non-convexe de la Croatie fait que cette moyenne risque de se retrouver hors du pays ce qui n'est pas un résultat très acceptable. On peut alors se demander s'il est possible de définir une nouvelle notion de moyenne qui soit raisonnable, correspondant à la moyenne habituelle lorsque le pays est convexe par exemple, et qui reste toujours à l'intérieur du pays. Plus généralement on peut se demander comment définir naturellement l'espérance d'une variable aléatoire contrainte à rester à l'intérieur d'un domaine non convexe.

Une solution possible pour ce problème est d'utiliser la moyenne de Fréchet définie ci-après.

1.3 Définitions des concepts

Dans un premier temps, définissons quelques concepts mathématiques importants.

Du fait de sa forme en croissant, la Croatie présente une caractéristique non convexe. Un ensemble C est dit convexe lorsque, pour tous x et y de C , le segment $[x, y]$ est tout entier contenu dans C , c'est-à-dire : pour tout $x, y \in C$, pour $t \in [0; 1]$, $tx + (1 - t)y$ appartient à C .

Les quantités que nous manipulons dans cet ensemble sont des variables aléatoires. Soit (F, B) un espace mesurable. On appelle variable aléatoire définie sur $(\Omega, \mathcal{A}, \mathbb{P})$ et à valeurs dans F , toute application mesurable $X : \Omega \rightarrow F$, c'est-à-dire satisfaisant pour tout $B \in \mathbb{B}$, $X^{-1}(B) \in \mathcal{A}$ avec $X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$.

On appelle variable aléatoire réelle une variable aléatoire à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. On dit que X est une variable aléatoire discrète si son image $X(\Omega)$ est finie ou infinie dénombrable. Dans notre cas, X est une variable aléatoire discrète lorsqu'on travaille sur l'ensemble fini des villes de la Croatie.

On dit que X est une variable aléatoire à densité (ou absolument continue) si sa loi \mathbb{P}_X est absolument continue par rapport à la mesure de Lebesgue, c'est-à-dire que : pour tout $A \in \mathcal{B}(\mathbb{R})$ tel que $\lambda(A) = 0$, on a $\mathbb{P}_X(A) = 0$. Dans notre cas, X est une variable aléatoire à densité lorsqu'on travaille sur l'espace continu géographique de la Croatie.

La moyenne est une mesure statistique qui représente le centre d'un ensemble de données numériques. L'espérance, quant à elle, est le concept de moyenne appliqué aux probabilités, représentant la valeur moyenne qu'on s'attend à obtenir d'une variable aléatoire.

Soit X une variable aléatoire réelle positive. Alors l'espérance, éventuellement infinie, est calculée dans le cas discret comme la somme des produits de chaque issue possible par sa probabilité associée :

$$\mathbb{E}[X] = \sum_{i=1}^n p_i x_i$$

avec x_i les valeurs possibles de la variable aléatoire et p_i leurs probabilités associées. Dans le cas continu, elle est définie comme s'en suit : [4]

$$\mathbb{E}[X] = \int_{\mathbb{R}^2} x f(x) dx$$

avec f correspondant à la fonction de densité associée à X . Une autre mesure de tendance centrale utilisée en statistique et en mathématiques est la moyenne de Fréchet définie comme telle :

$$\tilde{\mathbb{E}}[X] = \operatorname{argmin}_{y \in \bar{D}} [\mathbb{E}[d^2(y, X)]].$$

Dans le cas des variables aléatoires discrètes, la moyenne de Fréchet vaut donc :

$$\tilde{\mathbb{E}}[X] = \operatorname{argmin}_{y \in \bar{D}} \sum_{x_i \in X(\Omega)} p_i d^2(y, x_i)$$

Et dans le cas des variables à densité :

$$\tilde{\mathbb{E}}[X] = \operatorname{argmin}_{y \in \bar{D}} \int_{\mathbb{R}^2} p_i(x) d^2(y, x) dx$$

Avec D un ensemble non convexe, d la distance géodésique que l'on définira juste en dessous, et x et x_i les valeurs que la variable aléatoire X peut prendre respectivement lorsque X est discrète et quand elle est à densité.

Soit D l'ensemble non convexe de la Croatie. La distance, qui est une mesure de l'espace entre deux points, peut être calculée de différentes manières, comme la distance euclidienne, la distance de Manhattan (basée sur un parcours qui suit les axes verticaux et horizontaux, comme les rues d'une ville en grille), ou d'autres méthodes telles que la distance géodésique.

On appelle géodésique la courbe qui représente le chemin le plus court entre deux points sur une surface donnée. Elle peut prendre une allure très différente selon l'espace dans lequel on se trouve.

1.4 Méthodologie

En complément des calculs et démonstrations mathématiques, nous utiliserons des outils informatiques afin de visualiser nos recherches. En particulier, Google Colab nous garantit la possibilité de collaborer en temps réel sur notre projet en ayant l'accès gratuit à des ressources informatiques. Nous avons importé des données cartographiques, plus précisément les contours de carte en fichier GeoJSON, ainsi qu'un support de données datant de 2022 qui inclut la longitude, la latitude et la population de 83 villes croates [1].

Avec l'aide de nombreuses fonctions disponibles dans les librairies Python notamment folium et geopandas, nous manipulons principalement l'objet suivant : la liste des coordonnées de contour des frontières. Par ailleurs, on ne retient seulement que la plus grande composante connexe de la Croatie pour travailler sur un seul ensemble connexe.

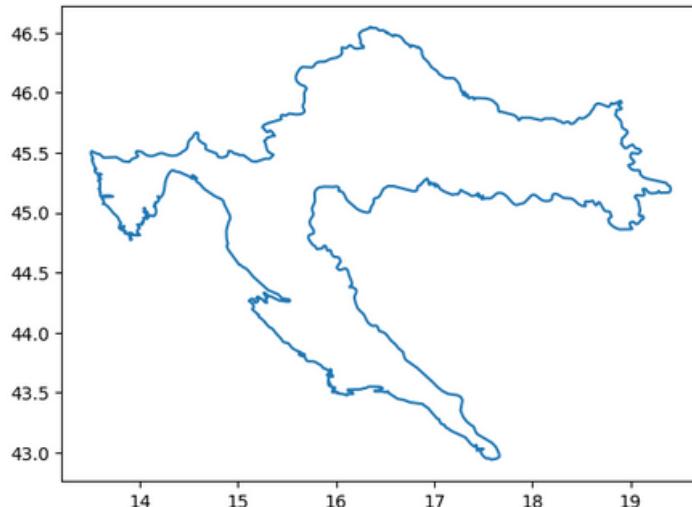


FIGURE 1 – Graphique du contour de la Croatie, illustrant sa frontière géographique par des lignes connectant les points successifs de notre liste de coordonnées.

2 Approche avec la moyenne euclidienne

Le cas de la moyenne euclidienne est intuitif et facile à comprendre. Il consiste à calculer la moyenne en considérant les coordonnées spatiales des points comme des vecteurs dans un espace euclidien, ce qui permet d'obtenir un point central qui minimise la distance euclidienne à tous les autres points.

2.1 Moyenne euclidienne de la population

L'idée étant d'obtenir une moyenne représentative de la répartition de la population à l'intérieur du pays, nous allons prendre en compte l'importance relative de chaque ville dans le calcul de la moyenne, en attribuant un poids à chaque ville basé sur sa population.

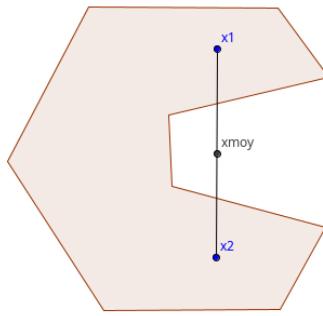


FIGURE 2 – Moyenne euclidienne entre deux points particuliers de même poids

On calcule ainsi le poids de chacune de nos 83 villes. La formule du poids de la $i^{\text{ème}}$ étant la suivante :

$$\text{poids}(i) = \frac{p_i}{p_{\text{totale}}}$$

avec pour toute ville i , p_i la population de la ville, x_i et y_i les coordonnées géographiques de la ville i . p_{totale} représentant la population totale de toutes les villes.

Nous procédons ensuite au calcul du point moyen euclidien pondéré défini par :

$$\text{point_moyen_euclidien} = \left(\sum_{i=1}^{83} x_i p_i, \sum_{i=1}^{83} y_i p_i \right)$$

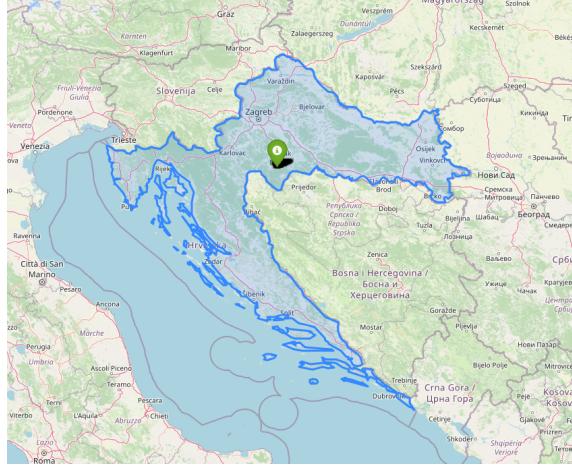


FIGURE 3 – Moyenne euclidienne de la population de la Croatie

Nous allons maintenant étudier comment se comporte la moyenne euclidienne lorsque les variables sont à densité.

2.2 Introduction à la loi des grands nombres

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées selon la loi de X . On suppose que $\mathbb{E}[|X|] < +\infty$ (Sous réserve de l'existence de l'espérance fini).

On introduit la formule de la loi des grands nombres, qui sous ces conditions, converge presque sûrement vers la moyenne.

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

avec X_i nos variables aléatoires correspondant aux coordonnées (latitude, longitude) de nos villes. Ce code peut être simplifié pour une utilisation itérative dans le code,

$$\bar{X}_{n+1} = \frac{n}{n+1} \bar{X}_n + \frac{1}{n+1} X_{n+1}$$

(cf. FIGURE 4) avec n le nombre d'itérations, X_{n+1} la variable aléatoire tirée aléatoirement dans notre ensemble, à chaque nouvelle itération.

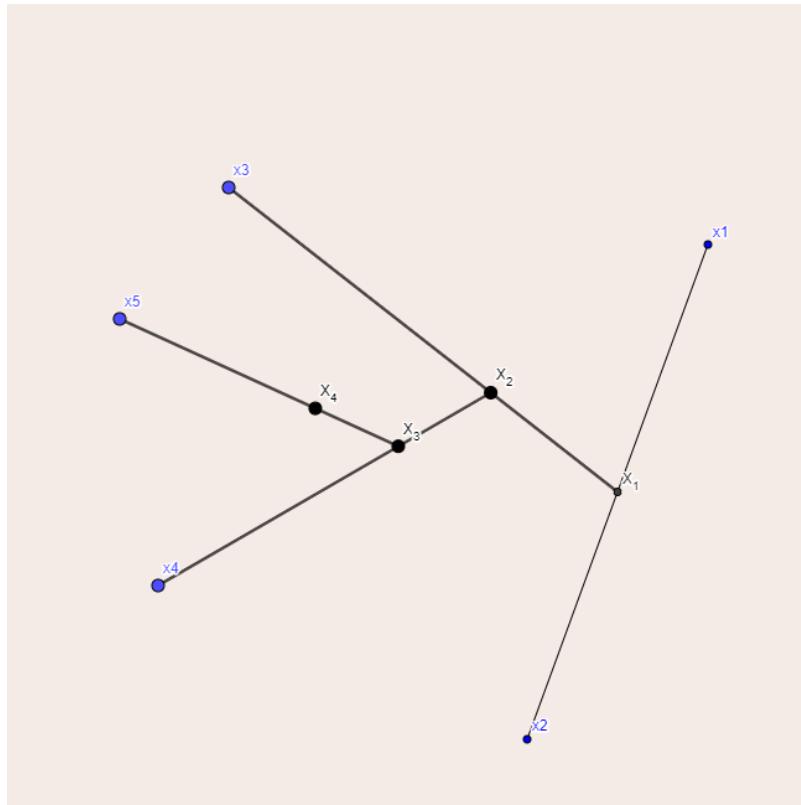


FIGURE 4 – Itérations selon la loi des grands nombres, où chaque point bleu marque l'emplacement d'une valeur tracée existante et chaque point noir indique l'ajout d'une nouvelle valeur, convergeant progressivement vers la moyenne.

Dans le contexte discret, soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires représentant les résultats de tirages itératifs sur différentes villes croates, chacune avec une probabilité associée à sa population. Si nous posons $S_n = X_1 + \dots + X_n$ pour la somme des tirages après n itérations, la moyenne empirique $\frac{S_n}{n}$ tendra, selon la loi des grands nombres, vers la moyenne pondérée des populations des villes concernées à mesure que n augmente.

Dans le scénario continu, les tirages sont effectués sur l'ensemble du territoire croate, considéré comme un continuum. Dans ce cas, $(X_n)_{n \in \mathbb{N}}$ serait une suite de variables aléatoires représentant les valeurs des tirages à différents emplacements, et S_n représenterait une intégrale sur le territoire jusqu'au n -ième tirage. La moyenne empirique $\frac{S_n}{n}$ converge alors vers la vraie moyenne des valeurs territoriales lorsqu'on prend en compte la densité de probabilité de chaque localisation, en augmentant le nombre de tirages.

Dans les deux cas, l'objectif est d'approcher la moyenne réelle, soit en pondérant par population dans le cas discret, soit en tenant compte de la répartition continue des valeurs dans le cas continu.

Proposition (Loi des grands nombres) Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires i.i.d. de même loi que X . On suppose que $\mathbb{E}[|X|^2] < +\infty$. Alors

$$\frac{S_n}{n} \xrightarrow{P} \mathbb{E}[X] \quad \text{lorsque } n \rightarrow +\infty$$

Algorithm 1: Loi des grands nombres

Input: Points pondérés par le nombre d'habitants
Output: Moyenne finale des positions pondérées par le nombre d'habitants

Initialiser moyenne en choisissant aléatoirement un point, pondéré par le nombre d'habitants;
 Initialiser un compteur i à 1;
for i allant de 1 à 999 **do**
 Sélectionner un nouveau point aléatoire $point_aleatoire$, pondéré par le nombre d'habitants;
 Calculer le chemin le plus court entre $point_aleatoire$ et $moyenne$;
 Prendre le point situé au $\frac{1}{i+1}$ -ème du chemin calculé, lequel devient notre nouvelle moyenne;
 Incrémenter i ;
end
return la moyenne qui est notre moyenne finale des positions pondérées par le nombre d'habitants après 999 itérations;

2.2.1 Exemple dans le cas discret

En nous appuyant sur les formules mathématiques, nous avons conçu un algorithme itératif de la loi des grands nombres. Plus on augmente le nombre d'itérations, plus le résultat final obtenu sera proche de la vraie moyenne, mais plus le temps d'exécution augmente. Après plusieurs tests, nous nous sommes rendus compte qu'itérer l'algorithme 10 000 fois était un bon compromis. Afin de faciliter l'observation de la convergence, un dégradé de couleur a été appliqué à un ensemble spécifique de 100 points. Ces points ont été systématiquement sélectionnés avec une fréquence régulière, à chaque centième étape de l'itération.

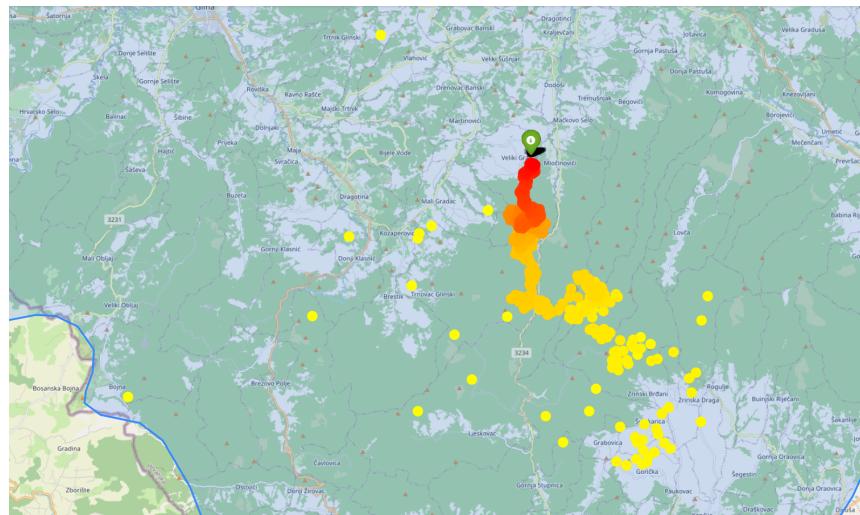
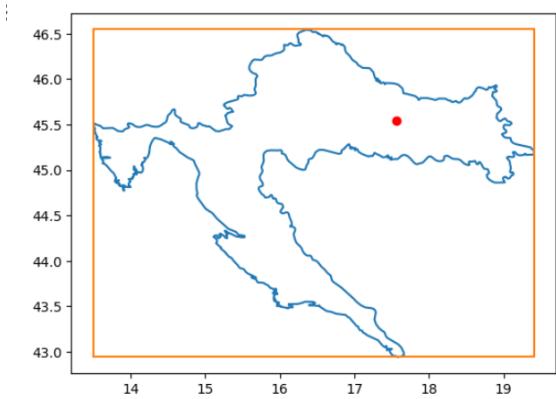


FIGURE 5 – Visualisation de la convergence vers notre moyenne

2.2.2 Cas des variables aléatoires à densité

Dans le cas où la variable aléatoire X suit une distribution de probabilité à densité, la loi des grands nombres permet d'approximer l'espérance $\mathbb{E}[X]$ en utilisant des échantillons de X . Pour cela on utilise une suite de variables aléatoires identiquement et indépendamment distribuées la moyenne empirique converge presque sûrement vers l'espérance théorique lorsque la taille de l'échantillon augmente.

Ici, nous utilisons la méthode du rejet qui consiste à tirer un point aléatoire X dans l'encadrement de la Croatie, en supposant que X suit une loi uniforme. À chaque tirage, nous vérifions si le point X tiré est inclus à l'intérieur des frontières de la Croatie ; si c'est le cas, nous le prenons en compte pour itérer la loi des grands nombres, sinon nous le rejetons et passons au tirage suivant.



Une méthode pour appliquer la loi des grands nombres à notre ensemble dense consiste à effectuer des tirages uniformes dans un rectangle englobant la Croatie.

FIGURE 6 – Encadrement optimal pour délimiter l'espace de tirages uniformes, basé sur les valeurs extrêmes de latitude et longitude.

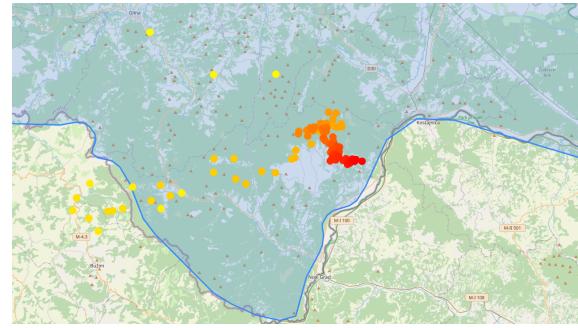


FIGURE 7 – Visualisation des itérations de l'algorithme de la loi des grands nombres appliquée à notre ensemble dense

Néanmoins, notre définition de moyenne actuelle ne garantit toujours pas que le point central moyen se trouve à l'intérieur du pays. Lorsqu'on considère des exemples entre deux villes : la moyenne euclidienne peut ne pas toujours se situer dans l'ensemble non-convexe du pays.

3 La moyenne de Fréchet

Après nous être imprégnés du sujet en définissant et en calculant des moyennes grâce à la moyenne Euclidienne, nous nous sommes rendus compte des problèmes de fiabilité de l'information que pouvait nous renvoyer cette moyenne dans un ensemble non convexe. Pour pallier ces problèmes, nous allons introduire une nouvelle moyenne, la moyenne de Fréchet.

$$\tilde{E}[X] = \operatorname{argmin}_{y \in D} \left[\sum_{i=1}^n d^2(x_i, y) p_i \right]$$

Le principal avantage de cette moyenne est qu'elle préserve la structure géométrique des ensembles de données sur lesquels elle agit (contrairement à la moyenne Euclidienne comme vu précédemment). De plus, malgré son imposante définition, elle est équivalente à la moyenne euclidienne dans \mathbb{R}^2 et plus généralement dans tout espace convexe. Ainsi, la moyenne de Fréchet est un outil précieux pour représenter la tendance centrale dans des ensembles de données non convexes, grâce à sa capacité à prendre en compte la structure géométrique globale des données tout en préservant leur complexité intrinsèque. Ainsi la proposition suivante serait valide pour les variables aléatoires discrètes et à densité.

Proposition *La moyenne de Fréchet vaut la moyenne Euclidienne dans \mathbb{R}^2 avec X une variable aléatoire :*

$$\tilde{\mathbb{E}}[X] = \mathbb{E}[X]$$

Démonstration. Soit X une variable aléatoire réelle.

$$\begin{aligned} \text{On a } \tilde{E}[X] &= \operatorname{argmin}_{y \in \tilde{\mathbb{D}}} [\mathbb{E}(d^2(y, X))] \\ &= \operatorname{argmin}_{y \in \mathbb{R}^2} [\mathbb{E}|X - y^2|] \end{aligned}$$

Soit f appartenant à $\mathcal{C}^2(\mathbb{R}^2)$ tel que $f(y) = \mathbb{E}(|X - y^2|)$

On cherche désormais les points critiques de f , c'est-à-dire, on cherche pour quels y appartenant à \mathbb{R}^2 , $\nabla f(y) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

$$\begin{aligned} \nabla f(y) &= \mathbb{E}[2(y - X)] = 2y - 2\mathbb{E}[X] \\ \nabla f(y) = 0 &\Leftrightarrow 2y - 2\mathbb{E}[X] = 0 \Leftrightarrow y = \mathbb{E}[X] \end{aligned}$$

On a donc $y^* = \mathbb{E}[X]$ comme unique point critique de f .

Montrons maintenant que c'est un minimum global. On a

$$\text{Hess}f(y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} = 2Id$$

Les valeurs propres de $\text{Hess}f(y_1, y_2)$ sont positives, donc $\text{Hess}f(y)$ est définie positive pour tout $(y) \in \mathbb{R}^2$. D'où $y^* = \mathbb{E}[X]$ est le minimum global de f . Et donc,

$$\tilde{\mathbb{E}}[X] = \mathbb{E}[X]$$

□

Cependant, calculer la moyenne de Fréchet entre deux points dans un ensemble non convexe n'est pas simple. Cela implique de trouver la trajectoire qui minimise la distance entre deux points.

3.1 Introduction de la géodésique

Pour calculer la moyenne de Fréchet entre deux points, on a besoin de la géodésique. La géodésique est le chemin le plus court reliant deux points et la distance géodésique est la longueur de ce chemin. Dans le cas euclidien, la géodésique est le segment entre deux points [2].

Proposition *Pour une variable aléatoire X , et sous réserve de son existence, l'espérance de X est sur la géodésique de x_1 et x_2 .*

Démonstration. Montrons que $\tilde{\mathbb{E}}[X] \in G(x_1, x_2)$

Par l'absurde :

Soit $z_1 = \tilde{\mathbb{E}}[X]$ avec $z_1 \notin G(x_1, x_2)$

Soit $z_2 \in G(x_1, x_2)$ tel que $d(x_1, z_2) = d(x_1, z_1)$

$$\begin{aligned} & d(x_1, x_2) < d(x_1, z_1) + d(z_1, x_2) \\ \Leftrightarrow & d(x_1, z_2) + d(z_2, x_2) < d(x_1, z_1) + d(z_1, x_2) \\ \Leftrightarrow & d(z_2, x_2) < d(z_1, x_2) \\ \Rightarrow & d^2(x_1, z_2) + d^2(z_2, x_2) < d^2(x_1, z_1) + d^2(z_1, x_2) \\ \text{Absurde car } & z_1 = \tilde{\mathbb{E}}[X] \end{aligned}$$

$$\tilde{\mathbb{E}}[X] \in G(x_1, x_2)$$

□

Il ne reste désormais plus qu'à trouver la position exacte de la moyenne de Fréchet entre deux points sur la géodésique entre ces deux points. Pour cela, nous allons démontrer que la moyenne de Fréchet entre deux points se situe sur la géodésique entre ces deux points proportionnellement à leurs poids.

Montrons que :

$$\begin{cases} d(x_1, z) = p_2 * d(x_1, x_2) \\ d(x_2, z) = p_1 * d(x_1, x_2) \end{cases}$$

Soit $z = \tilde{E}[X]$

$$z = \tilde{E}[X] \Rightarrow z \in G(x_1, x_2) \Rightarrow$$

$$\begin{cases} d(x_1, z) = a_1 * d(x_1, x_2) \\ d(x_2, z) = a_2 * d(x_1, x_2) \\ a_2 = 1 - a_1 \end{cases}$$

$$\tilde{E}[X] = \underset{y \in \bar{D}}{\operatorname{argmin}} [d^2(x_1, y)p_1 + d^2(x_2, y)p_2] = \underset{y \in \bar{D}}{\operatorname{argmin}} f(y) = z$$

$$\begin{aligned} f(z) &= d^2(x_1, z)p_1 + d^2(x_2, z)p_2 \\ &= a_1^2 d^2(x_1, x_2)p_1 + a_2^2 d^2(x_1, x_2)p_2 \\ &= (a_1^2 p_1 + (1 - a_1)^2 p_2)d^2(x_1, x_2) \\ &= g(a_1)d^2(x_1, x_2) \end{aligned}$$

On sait que z minimise f , donc on cherche a_1 tel que a_1 minimise g .

$$\begin{aligned} g(a_1) &= a_1^2 \cdot p_1 + (1 - a_1)^2 \cdot p_2 \\ g'(a_1) &= 2 \cdot a_1 - 2 \cdot p_2 = 0 \quad \Leftrightarrow \quad a_1 = p_2 \\ a_2 &= 1 - a_1 = 1 - p_2 = p_1 \end{aligned}$$

On a donc bien

$$\begin{cases} d(x_1, z) = p_2 * d(x_1, x_2) \\ d(x_2, z) = p_1 * d(x_1, x_2) \end{cases}$$

Finalement, pour calculer la moyenne de Fréchet entre deux points, il suffit de trouver la géodésique.

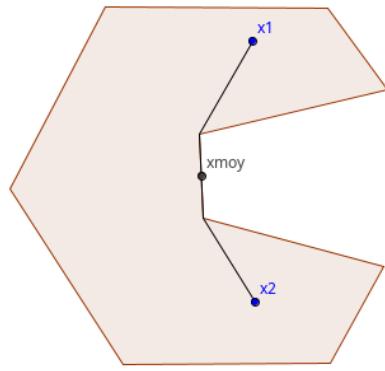


FIGURE 8 – Représentation d'une géodésique entre deux points particuliers de même poids dans un ensemble non convexe et leur moyenne de Fréchet

3.1.1 Approche avec l'algorithme Dijkstra

Dans notre contexte, l'algorithme de Dijkstra[3] est adapté. Il s'agit d'un algorithme de recherche de chemin pour trouver le chemin le plus court entre deux nœuds dans un graphe, ce qui est en général particulièrement utile dans la planification d'itinéraires et la cartographie.

On considère un graphe formé par les intersections d'un quadrillage sur la carte de la Croatie, chaque carreau du quadrillage correspond à une surface de 0.1 degré carré.

Algorithm 2: Algorithme de Dijkstra

Input: Graphe G , nœud de départ s , nœud d'arrivée t
Output: Chemin le plus court de s à t

Initialiser un tableau $d[]$ pour stocker les distances minimales de s à chaque nœud du graphe;
Initialiser un tableau $visited[]$ pour marquer les nœuds déjà visités;
Initialiser un tableau $parent[]$ pour stocker le nœud précédent pour chaque nœud dans le chemin le plus court;
Initialiser une file de priorité Q pour stocker les nœuds à explorer, avec la distance de s à chaque nœud comme priorité;

for v dans G **do**

- | Initialiser $d[v]$ à l'infini, $visited[v]$ à faux et $parent[v]$ à None;

end

Définir $d[s]$ à 0;

Insérer s dans Q avec une priorité de 0;

while Q n'est pas vide **do**

- | Extraire le noeud u de Q avec la plus petite distance $d[u]$;
- | **if** u est égal à t **then**

 - | | Construire et retourner le chemin le plus court en remontant les parents de t ;

- | **end**
- | **if** $visited[u]$ est faux **then**

 - | | Marquer $visited[u]$ comme vrai;
 - | | voisin v de u Calculer la nouvelle distance temporaire $temp_d$ de s à v via u ;
 - | | **if** $temp_d$ est plus petit que $d[v]$ **then**

 - | | | Mettre à jour $d[v]$ avec $temp_d$;
 - | | | Mettre à jour $parent[v]$ avec u ;
 - | | | Insérer v dans Q avec une priorité de $temp_d$;

 - | | **end**

- | **end**

end

if le nœud d'arrivée t n'est pas atteint **then**

- | **return** "Pas de chemin trouvé";

end

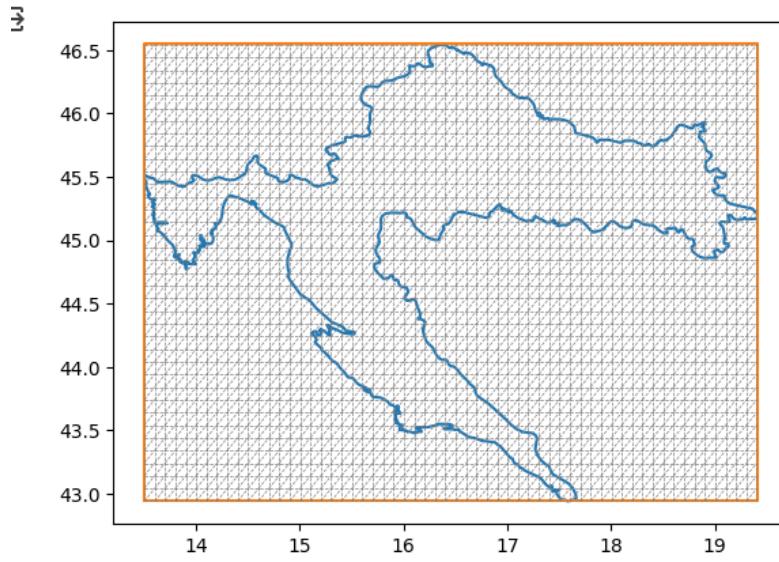


FIGURE 9 – Représentation du quadrillage pris en compte dans l’algorithme de Dijkstra

Le dilemme auquel nous sommes confrontés réside dans la durée d’exécution engendrée par l’application de la loi des grands nombres. Effectivement, pour chaque trajectoire envisagée, notre algorithme se doit de parcourir l’ensemble des possibilités des points du quadrillage. Ainsi, cela revient à examiner chacun des 8 voisins pour chacun des 1 000 100 sommets du graphe, ce qui équivaut à un nombre de possibilités titaniques.

Notre implémentation de l’algorithme de Dijkstra n’étant pas la plus optimale, la complexité de l’algorithme qui s’exprime normalement en $\mathcal{O}(m \log(n))$ avec n le nombre de possibilités, et m le nombre de sommets se voit être beaucoup plus importante. D’autant plus qu’avec l’application de la loi des grands nombres, cet algorithme se voit itérer 1100 fois, ce qui fait un temps d’exécution très long. Par exemple, pour 10 itérations, l’exécution se fait en approximativement un quart d’heure.

Face à un temps de compilation excessivement long, s’étendant sur plusieurs heures lorsque le nombre d’itérations devient important, et considérant le fait que cette démarche devient contre-productive si nous réduisons le nombre d’itérations — puisqu’il s’agit d’une méthode d’approximation —, il devient impératif d’adopter une stratégie plus efficiente en termes de complexité algorithmique.

Nous sommes donc en quête d’un algorithme à la fois plus performant et moins gourmand en temps de calcul.

3.1.2 Approche avec l'algorithme des lignes brisées

L'algorithme des lignes brisées est également une technique conçue pour déterminer le trajet le plus court entre deux points, adapté aux contextes géographiques complexes comme celui de la Croatie. Il commence par tracer une ligne directe entre les points de départ et d'arrivée, examinant si ce trajet reste dans les limites du pays. Lorsque la ligne de frontière dépasse les limites du territoire croate, l'algorithme détermine la distance jusqu'au point de la frontière le plus éloigné du segment. Ce troisième point crée deux nouveaux segments sur lesquels l'algorithme sera réitéré.

Ce processus est répété autant de fois que nécessaire jusqu'à ce qu'un chemin entièrement interne soit tracé, garantissant ainsi que la ligne de frontière reste entièrement contenue dans le territoire croate. En d'autres termes, l'algorithme ajuste la trajectoire de la ligne pour qu'elle reste toujours à l'intérieur des frontières croates, en utilisant les frontières existantes comme guide pour rétablir une limite territoriale cohérente.

Algorithm 3: Algorithme de construction de ligne brisée

Input: Point de départ a , point d'arrivée b
Output: Liste de points formant une ligne brisée
Créer une ligne droite entre les points a et b ;
if *la ligne ne croise pas la frontière de la Croatie* **then**
 | **return** *la liste des points* $[a, b]$;
end
else
 | Trouver le point de "contrôle" ptc qui coupe la Croatie et la ligne (a, b) à
 | l'extérieur du segment $[a, b]$;
 | Trouver l'indice du point le plus proche de c dans la liste des points de la
 | Croatie;
 | Trouver tous les points qui coupent la Croatie et le segment $[a, b]$;
 | **for** *segment formé par ces points* **do**
 | | **if** *ce segment est en dehors de la Croatie* **then**
 | | | Grâce au point de contrôle, choisir la partie de la frontière de la Croatie
 | | | à conserver;
 | | | Parmi les parties de la frontière conservées, trouver le point pt_{dm} le
 | | | plus éloigné du segment $[a, b]$;
 | | **end**
 | **end**
end

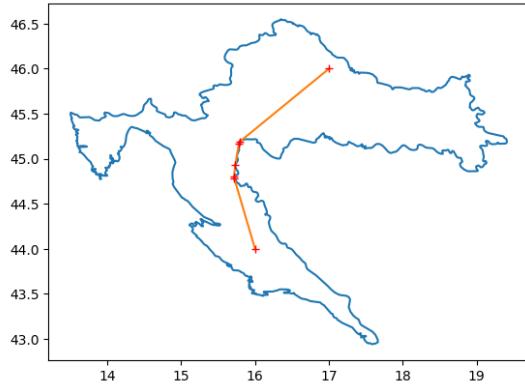


FIGURE 10 – Visualisation de la géodésique

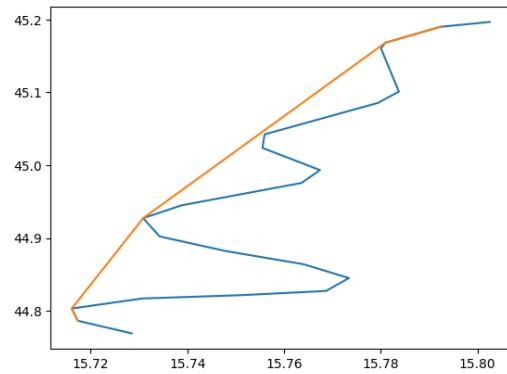


FIGURE 11 – Zoom sur la frontière

En remarquant que la deuxième image est la version zoomée de la première. On voit que la courbe de la géodésique respecte le domaine de la Croatie et donc la caractéristique non convexe de celle-ci.

Ci-dessous un autre exemple d'une géodésique plus complexe :

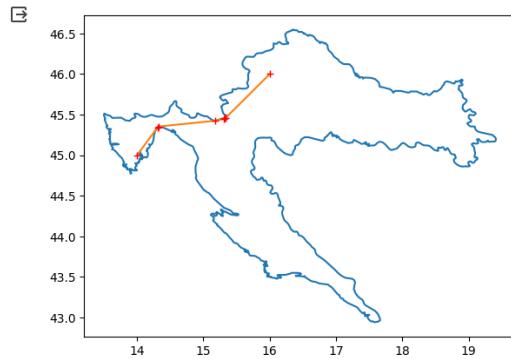
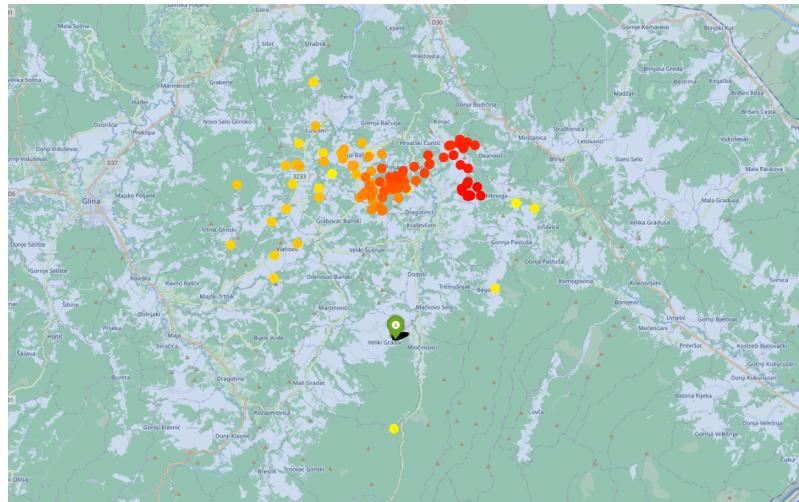


FIGURE 12 – Exemple plus complexe

3.2 Approximation numérique de la moyenne de Fréchet

3.2.1 Exemple d'un cas des variables aléatoires discrètes

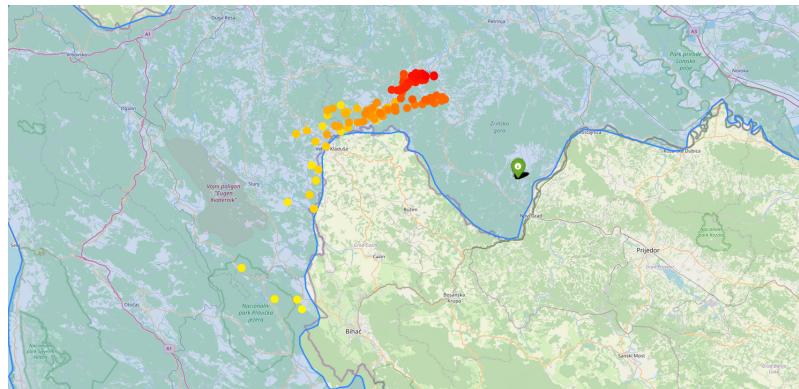
La visualisation ci-dessous donne la représentation des itérations de la loi des grands nombres appliquée à l'algorithme de la ligne brisée sur nos 83 villes sous forme de dégradé, par rapport à la moyenne euclidienne en vert.



On remarque que la moyenne de Fréchet se positionne significativement plus à l'intérieur des terres que la moyenne euclidienne, avec une différence notable d'environ 8 kilomètres. Cette observation met en évidence les insuffisances de la moyenne euclidienne pour représenter de manière fiable les données dans un espace non convexe.

3.2.2 Exemple d'un cas des variables aléatoires à densité

La visualisation ci-dessous donne la représentation des itérations de la loi des grands nombres appliquée à l'algorithme de la ligne brisée sur l'ensemble de points de notre grille sous forme de dégradé, par rapport à la moyenne euclidienne dans le cas des variables à densité approchée par le point en vert.



Dans un espace dense, l'écart de 40 kilomètres entre les moyennes euclidienne et de Fréchet souligne une fois de plus l'incapacité de la moyenne euclidienne à représenter fidèlement les données dans des contextes non convexes.

3.2.3 Comparaison des deux algorithmes

Dans notre analyse comparative des deux algorithmes, nous examinons leurs performances et leur pertinence dans des contextes spécifiques.

Tout d'abord, nous considérons la complexité algorithmique, un aspect crucial pour évaluer l'efficacité des algorithmes. L'algorithme de Dijkstra présente une complexité quasi-linéaire, notée $O(m * \log(n))$, où " n " représente le nombre de nœuds dans le graphe et " m " le nombre d'arêtes. Quant à l'algorithme de la ligne brisée offre une complexité linéaire, notée $O(n)$, où " n " dénote le nombre de segments dans le contour géographique considéré. Cette distinction est cruciale, car une complexité réduite peut représenter un avantage significatif en termes de performance dans certains cas d'utilisation.

4 Conclusion

La recherche approfondie menée sur la notion de moyenne dans des ensembles géographiques complexes, avec un focus particulier sur la Croatie, a abouti à des découvertes significatives qui remettent en question les approches conventionnelles de calcul de la moyenne. La problématique initiale visait à définir et calculer une moyenne dans un contexte non convexe, où les résultats habituels de la moyenne pourraient se retrouver hors des limites géographiques concernées, ce qui s'avère être un défi particulièrement pertinent dans le cas de la Croatie en raison de sa forme géographique unique.

À travers l'exploration de la moyenne euclidienne et son adaptation au contexte géographique de la Croatie, nous avons identifié les limites de cette méthode, en particulier sa tendance à placer la moyenne hors des frontières du pays dans certains cas.

Cette limitation souligne l'importance de développer des approches alternatives qui prennent en compte la géométrie spécifique de l'espace dans lequel elles sont appliquées.

L'introduction de la moyenne de Fréchet constitue une avancée majeure dans cette recherche, offrant une solution élégante au problème posé. En adaptant la définition de la moyenne pour intégrer la structure géométrique de l'espace, la moyenne de Fréchet garantit que le résultat reste toujours à l'intérieur du domaine étudié, tout en conservant les propriétés souhaitables d'une moyenne dans des contextes convexes.

Cette méthode, grâce à son utilisation de la géodésique pour calculer les distances, permet une représentation plus fidèle et significative des données dans des ensembles non convexes.

L'algorithme des lignes brisées, offre des avantages par rapport à l'algorithme de Dijkstra. L'algorithme de Dijkstra, non précis et s'est avéré être moins pratique pour de grandes itérations en raison de sa complexité computationnelle.

À l'inverse, l'algorithme des lignes brisées présente une approche plus simple et plus rapide, suggérant une voie vers des calculs plus efficaces dans des contextes similaires.

En conclusion, cette étude démontre non seulement l'importance de réviser les méthodes traditionnelles de calcul de la moyenne dans des contextes géographiques non convexes mais aussi la valeur de l'innovation algorithmique pour surmonter ces défis.

La moyenne de Fréchet se révèle être un outil précieux dans ce cadre, offrant une nouvelle perspective sur la représentation des données géographiques complexes. Alors que la recherche continue d'évoluer, l'exploration de méthodes encore plus efficientes et précises reste un domaine prometteur pour des avancées futures.

Références

- [1] <https://worldpopulationreview.com/countries/cities/croatia>
- [2] <https://geopy.readthedocs.io/en/stable/>
- [3] <https://www.maths-cours.fr/methode/algorithm-de-dijkstra-etape-par-etape>
- [4] https://moodle.u-bordeaux.fr/pluginfile.php/851940/mod_resource/content/1/Cours2324ProbasL34TMQ601U.pdf