Journal of Mathematics and Statistics Studies
ISSN: 2709-4200
DOI: 10.32996/jmss
Journal Homepage: www.al-kindipublisher.com/index.php/jmss

Data Geometry and Extreme Value Distribution

Mamadou Cisse [1] **Aliou Diop** [2], **Souleymane Bognini** [3], **Nonvikan Karl-Augustt Alahassa**[4]
[13] *National School of Statistics and Economical Analysis, University of Cheikh Anta Diop, Dakar, Senegal*
[2] *Gaston Berger University of Saint Louis, Saint Louis, Senegal*
[4] *Department of Mathematics and Statistics, University of Montreal, Montreal, Canada*
✉ **Corresponding Author**: Nonvikan Karl-Augustt Alahassa, E-mail: alahassan@dms.umontreal.ca

ARTICLE INFORMATION

Received: June 08, 2021
Accepted: July 05, 2021
Volume: 2
Issue: 2
DOI: 10.32996/jmss.2021.2.2.2

KEYWORDS

Mobile ad hoc Network, Clifford Algebra, Grassmann Algebra, Stochastic Geometric Graph, Blade groups, hypercube, Routing protocol, GEV distribution

ABSTRACT

In extreme values theory, there exist two approaches about data treatment: block maxima and peaks-over-threshold (POT) methods, which take in account data over a fixed value. But, those approaches are limited. We show that if a certain geometry is modeled with stochastic graphs, probabilities computed with Generalized Extreme Value (GEV) Distribution can be deflated. In other words, taking data geometry in account change extremes distribution. Otherwise, it appears that if the density characterizing the states space of data system is uniform, and if the quantile studied is positive, then the Weibull distribution is insensitive to data geometry, when it is an area attraction, and the Fréchet distribution becomes the less inflationary.

1. Introduction

En théorie des valeurs extrêmes, il existe deux approches fondamentales, largement utilisées: la méthode du block maxima, et le peaks-over-threshold[2] (POT) qui tient compte des observations dépassant un seuil donné (de Haan et Ferreira, 2010). La méthode du block maxima regroupe les données en blocs de longueur égale et adapte les données aux maximums de chaque bloc, par exemple, en calculant les maxima annuels des quantités quotidiennes de précipitations. Le choix de la taille des blocs est critique: des blocs qui sont trop petits peuvent conduire à des biais et des blocs qui sont trop grands génèrent trop peu de maxima, ce qui conduit à une grande dispersion (Coles, 2001).

Cependant, la théorie des valeurs extrêmes telle que développée avec ces deux approches connaît des limites. Voici un problème standard en réseaux qui oblige à revoir ces approches. Considérons un réseau mobile ad hoc (mobile ad hoc network ou MANET), continûment auto-configuré, constitués de terminaux mobiles connectés sans fil. Chaque appareil dans un MANET est libre de se déplacer indépendamment dans une direction quelconque, et donc peut changer ses liens vers d'autres appareils fréquemment. Chacun doit transférer le trafic sans rapport avec son propre usage, et donc être un routeur. A chaque instant t, le réseau ad-hoc peut contenir un ou plusieurs émetteurs-récepteurs qui se trouvent dispersés sur une vaste région, et où chaque émetteur-récepteur ne peut communiquer avec quelques autres à proximité. Il en résulte une topologie autonome très dynamique. On suppose que les caractéristiques de chaque routeur présent dans le réseau à un moment t donné sont aléatoires (le nombre de mise à jour d'états de lien, le taux d'insatisfaction3, le taux de perte4, etc.). Un défi principal dans la construction d'un MANET est d'équiper chaque dispositif pour maintenir en permanence les informations nécessaires pour bien acheminer le trafic. On suppose que le protocole de routage introduit une surcharge si le taux d'insatisfaction pour un routeur dépasse un seuil z_0. En fonction de

[2] Le POT a été mis au point par Pickands (1975) qui a fourni le cadre théorique et les outils statistiques appropriés.
[3] Ce taux représente le pourcentage de clients n'obtenant pas les ressources demandées. Plus ce taux est faible, meilleure est la performance du réseau.
[4] Ce taux représente le pourcentage de paquets perdus lors de leur transition à travers le routeur.

la dynamique et de la distribution aléatoire des routeurs du réseau, l'on cherche la probabilité qu'à un instant t_k fixé, aucun routeur n'introduise une surcharge.

Pour répondre à ce genre de problème, il est nécessaire d'intégrer la dynamique et la topologie des données dans la modélisation des extrêmes. Comment modifier les distributions GEV (Generalized Extreme value) pour intégrer la géométrie et la dynamique du système complet des données ? D'autres outils statistiques sont nécessaires pour y arriver.

2. GÉOMÉTRIE ET GRAPHES

Une façon de modéliser les extrêmes sans perte d'information sur l'ensemble des données (i.e sans recourir uniquement à la technique des block maxima), est de tenir compte de la géométrie des données de départ. Dans ce cadre, le statisticien pourra faire recourt aux graphes grâce aux propriétés qu'ils offrent. Un graphe est caractérisé par un ensemble de sommets, dans un espace de dimension d, $\mathbb{R}^d$ en l'occurrence, et un ensemble de liens qui relient ces sommets.

Définition 1.1. Soit $\|\cdot\|$ une norme[5] sur $\mathbb{R}^d$ (par exemple la norme euclidienne), et soit r un réel positif. Soit $X \subset \mathbb{R}^d$ un sous-ensemble, avec un ensemble d'arcs (ou arêtes) connectant chacun deux éléments *{x, y}* de X à la condition$\|x - y\| < r$. Le graphe obtenu est appelé graphe géométrique[6], et est noté $G^{(d)}(X; r)$. L'ensemble E de ses arcs est défini par :

$$\{X_i - X_j\} \epsilon E <=> 0 < \|X_i - X_j\| < r$$

Ce à quoi l'on s'intéresse le plus est la connectivité des points (sommets), la distribution des degrés[7].

Soit *f* une densité de probabilité sur R^d, et $X_1, X_2, \ldots$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de densité f. On pose $X = \{X_1, X_2, \ldots, X_n\}$. Notre centre d'intérêt ici est $G^{(d)}(X; r)$.

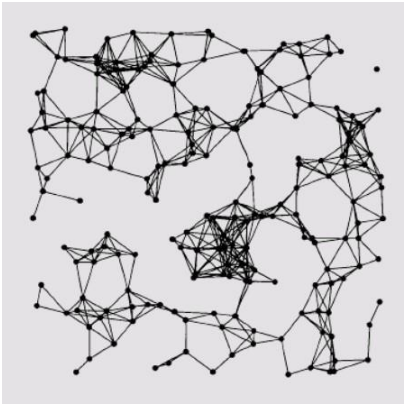

Figure 1– Un exemple de graphe géométrique avec (d = 2, n = 200, r = 0, 11), et f une densité uniforme sur [0, 1]²

Un célèbre graphe est celui de Erdõs et Rényi (1959) à partir duquel l'on montre que pour n assez grand, la probabilité qu'un graphe (non orienté) à n sommets avec une probabilité d'arc entre deux sommets de $\frac{2lnn}{n}$ soit connecté vaut 1. Les graphes de Erdõs et Rényi possèdent la propriété d'indépendance pour deux arcs distincts, ce qui n'est pas le cas pour un graphe géométrique. Dans un graphe géométrique, si X_i est proche de X_i ($\|X_i - X_j\| \leq r$) et si X_j est proche de X_k, alors X_i est également proche de X_k ($\|X_i - X_k\| \leq \|X_i - X_j\| + \|X_j - X_k\| \leq 2r$). Alors qu'il est juste nécessaire d'exploiter les outils combinatoires pour comprendre le graphe de Erdõs et Rényi, les outils de la géométrie stochastique sont plus que nécessaires pour appréhender un graphe géométrique.

Plus généralement, les points (X_i, $1 \leq i \leq n$) représentent des données spatiales (données multivariées), la mesure de d attributs de l'individu i dans un ensemble à n individus[8]. Nous supposons de plus que chacun des attributs mesurés est une variable continue[9]. Lorsque l'on construit des séries temporelles de ces données, à chaque instant l'on peut entreprendre une analyse de

[5] Ici le choix de la norme ne modifie pas l'analyse : sur $\mathbb{R}^d$, toutes les normes sont équivalentes.
[6] Graphe de proximité.
[7] En théorie des graphes, le degré (ou valence) d'un sommet d'un graphe est le nombre de liens (arêtes ou arcs) reliant ce sommet, avec les boucles comptées deux fois. Le degré d'un sommet s est noté deg(s).
[8] Une utilisation concrète de la théorie des graphes géométriques est la *classification*.
[9] Ce qui n'est pas forcément le cas dans la pratique, mais cette hypothèse facilite la présente étude.

clusters avec l'ensemble de points X, en déduire des groupes (ou clusters), grâce à quelques hypothèses sur la similarité entre deux individus. Cependant, ce qui nous intéresse dans ce document est l'aspect aléatoire du graphe.

2.1 Quelques définitions

La suite de l'analyse requiert l'usage des notions de groupes et algèbres, que nous introduisons dans cette section. Pour toute la section, soit $n \in \mathbb{N}$, et p et q deux entiers non nuls tels que $p + q = n$.

Les groupes de blade $B_{p,q}$

Définition 1.2. Soit $B = \{e_1, e_2, ..., e_n\}$. Soit $B_{p,q}$ le groupe multiplicatif généré par B à travers les éléments $\{e_\varphi, e_\alpha\}$ vérifiant les relations suivantes pour tout $x \in B \cup \{e_\varphi, e_\alpha\}$:

$$\begin{cases} e_\alpha x = x e_\alpha \\ e_\varphi x = x e_\varphi = x \\ e_\varphi^2 = e_\alpha^2 = e_\varphi \end{cases}$$

$$e_i e_j = \begin{cases} e_\alpha e_i e_j & si\ i \neq j \\ e_\varphi x & si\ i = j \leq p \\ e_\alpha & si\ p < i = j \end{cases}$$

Le couple (p, q) est la signature du groupe $\boldsymbol{B_{p,q}}$, dit groupe de blade. Le groupe multiplicatif $\boldsymbol{B_{p,q}}$ est déterminé par l'ensemble multi-indexé $\{e_I, e_\alpha e_I, I \in 2^{[n]}\}$, avec $[n] = \{1, 2, ..., n\}$:

$$e_I \coloneqq \prod_{i \in I} e_i, e_I e_J = v(I,J) e_{I \Delta J}, I \Delta J = (I \cup J)(I \cap J) e_I^{-1} = v(I,I) e_I$$

$$v(I,J) = e_\alpha^{\mu_p(I \cap J) + \sum_{i \in J} \mu_j(I)}, \mu_j \coloneqq |\{i \in I :\ i > j\}|, e_I v(I,I) e_I = v(I,I) e_{I \Delta I} = e_\varphi$$

$$e_I \prec e_J \iff \mathrm{I} \prec \mathrm{J} \iff \sum_{i \in I} 2^{i-1} < \sum_{j \in J} 2^{j-1}$$

Où $2^{[n]}$ est l'ensemble des parties de $[n]$.

Le n-hypercube $(Q_n, E(Q_n))$

Définition 1.3. Un n-hypercube $(Q_n, E(Q_n))$, est un graphe dont les sommets $s \in Q_n$ sont des n-uplets de l'ensemble $\{0, 1\}$, et dont les arcs (ou arêtes) sont définis par la relation :

$$\{v_1, v_2\} \in E(Q_n) \iff d_H(v_1 \oplus v_2) = 1$$

avec d_H la distance de Hamming définie par :

$$d_H(a,b) = |\{i : 1 \leq i \leq n, a_i \neq b_i\}|, a, b \in \{0,1\}^n$$

Algèbre de Clifford, algèbre de Grassmann

Les algèbres de Clifford sont une généralisation des nombres complexes, et possèdent de multiples applications en physique quantique.

Définition 1.4. Soit V un espace de dimension n, de même structure algébrique que $\mathbb{R}^n$ ($V \cong \mathbb{R}^n$), et de base orthonormée B. L'algèbre de Grassmann sur V est définie comme l'algèbre associative définie par B, avec la multiplication extérieure $\wedge$, vérifiant :

$$x \wedge y = \begin{cases} 0 & s'il\ existe\ \alpha \in \mathbb{R}\ tel\ que\ x = \alpha y \\ -y \wedge x & sinon \end{cases}$$

Le produit extérieur $x \wedge y$ représente un parallélogramme orienté déterminé par x et y. En posant $e_\varphi = 1$, on pose :

$$e_I = e_{I_1} \wedge e_{I_2} \wedge \ldots \wedge e_{I_m}, I = \{I_1, \ldots, I_m\} \subseteq [n]$$

Définition 1.5. L'algèbre de Clifford de signature (p, q), noté $Cl_{p,q}$ est définie comme l'algèbre réelle associative engendré par les vecteurs $\{e_i\}$, et le scalaire $e_\varphi = 1$, vérifiant :

$$[e_i, e_j]_+ = \begin{cases} 2\ si\ 1 \le i = j \le p \\ -2\ si\ p+1 \le i = j \le n \\ 0\ si\ i \ne j \end{cases}$$

Avec le scalaire $e_\varphi = 1$. Le produit extérieur de Clifford vérifie :

$$x \wedge y = xy - \langle x, y \rangle, \langle x, y \rangle = \sum_{I \in 2^{[n]}} x_I y_J\ , x = \sum_{I \subseteq [n]} x_I e_I\ , e_I = \prod_{i \in I} e_i$$

Définition 1.6. L'algèbre de Clifford symétrique $Cl_{p,q}^{sym}$ de signature *(p, q)* est définie comme l'algèbre associative générée par la collection $\{\varsigma_i : 1 \le i \le n\}$, avec le scalaire $\varsigma_\varphi = 1$, vérifiant :

$$\varsigma_i^2 = \begin{cases} 1\ si\ 1 \le i \le p \\ -1\ si\ p < i \le n \end{cases}$$

et pour tout $i \ne j$, on a :$\varsigma_i \varsigma_j = \varsigma_j \varsigma_i$. Ici, nous avons l'extension :

$$\varsigma_I \varsigma_J = (-1)^{\mu_p(I \cap J)} \varsigma_{I \Delta J}$$

L'application $Cl_n \to \mathbb{R}(\bigoplus_{i=1}^n \mathbb{Z}_2)$, définit un isomorphisme d'algèbre, grâce à l'extension linéaire : $a\varsigma_I \longmapsto a z_I$, avec la correspondance $Cl_{n,0} = Cl_n$et :

$$z_I \leftrightarrow \left(a_1 a_2 \dots a_{|I|}\right) \Leftrightarrow a_i = \begin{cases} 1\ si\ i \in \mathrm{I} \\ 0\ sinon \end{cases}$$

$|I|$ est le cardinal de *I*.

Lemme 1.1. $\mathcal{B}_{p,q}/\langle e_\alpha \rangle$ *est isomorphe à un hypercube* Q_n *de dimension n.*

Lemme 1.2. *Une algèbre de Clifford* $Cl_{p,q}$ *est isomorphe à une algèbre de blade* $\mathcal{B}l_{p,q} \equiv \mathbb{R}\mathcal{B}_{p,q}/\langle e_\alpha + 1 \rangle$

Proposition 1.3. *L'algèbre de Clifford symétrique,* $Cl_{p,q}^{sym}$ *est isomorphe à une algèbre de Clifford* $Cl_{p+2q,q}$.

2.2 Les processus stochastiques algébriques

A présent, abordons la dimension spatio-temporelle. Un processus stochastique $(G_i^{(d)}(X;r))_{t \le 0}$ de graphes géométriques est une suite de graphes géométriques$G_i^{(d)}(X;r)$, dont les points (sommets) sont ceux d'une région de l'espace $R_0 \subset \mathbb{R}^d$ (avec une infinité de points éventuellement). Pour faire simple, nous allons supposer que le temps est discret, i.e $t \in \mathbb{N}$.

Puisque dans la pratique l'on pourra se ramener à des séries temporelles de graphes géométriques, l'on peut se ramener à un cas plus simple en discrétisant[10] R_0, et en fixant R_0 isomorphe à un hypercube $\mathbb{Q}_{N^d}$, de taille N^d. Dans ce cas, on considère l'algèbre de Clifford symétrique Cl_V^{sym}, avec *V* l'ensemble des sommets[11] de l'hypercube $\mathbb{Q}_{N^d}$. Autre hypothèse : nous supposons le cas d'indépendance des points de l'hypercube (hypothèse d'indépendance spatiale). Soit $\mathcal{F}$ la *σ-algèbre* des sous-ensembles de *V*, pour un sous-ensemble $U \subseteq V$. La topologie du graphe géométrique dont les sommets sont des éléments de *U*, est uniquement déterminée par la relation :

$$v_1 \sim v_2 \iff 0 < \| v_1 - v_2 \| \le r, \qquad v_1, v_2 \in U$$

On définit également la mesure de probabilité :

$$\mathbb{P}(U) = \prod\nolimits_{v \in U} \mathbb{P}(v) \prod\nolimits_{w \notin U} (1 - \mathbb{P}(w))$$

[10] Confère méthodes des éléments finis, pour la discrétisation de l'espace $\mathbb{R}^2$ par maillage, car les individus sont des éléments de $\mathbb{R}^2$ avec d attributs observés. On pourra prendre N aussi grand que l'on désire selon la finesse de la discrétisation désirée.

[11] Il suffit de considérer un d-cube unitaire $[0,1]^d$ de $\mathbb{R}^d$. En divisant chaque côté en N intervalles de taille d, l'on obtient un graphe géométrique dont les sommets appartenant à l'ensemble $(\mathbb{Q} \cap [0,1])^d$ sont:

$$V = \left\{ \left(\frac{2j_1 - 1}{2N}, \dots, \frac{2j_d - 1}{2N} \right) : 1 \le j_1, \dots, j_d \le N \right\}$$

L'espace de probabilité $(V, \mathcal{F}, \mathbb{P})$ induit une probabilité sur la collection de graphes géométriques $\mathcal{G} = (G_U)_{U \subseteq V}$ possibles avec V. Alors, un processus stochastique de graphes géométriques est une marche[12] aléatoire sur l'hypercube $\mathbb{Q}_{N^d}$, induit par des processus algébriques sur Cl_V^{sym}. A chaque blade $\mathcal{G}_U$ de Cl_V^{sym}, est associé un unique graphe géométrique G_U , avec $|U| = n_U$ ($cardinal\ de\ U$).

Un résultat fondamental montré dans ce document est la relation de dépendance suivante :

$$P_{G_t^{(d)}=G_U}\left(\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(X_j^t\right)<\rho\right)\right)=\frac{P(\max_{1\le j\le|V|}\left(\prod_i\left(X_j^t\right)<\rho\right),G_t^{(d)}=G_U)}{P(G_t^{(d)}=G_U)}\neq P(\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(X_j^t\right)<\rho\right))$$

avec $\prod_i$ l'application définie pour $i\epsilon[1;d] par\ x \longmapsto \prod_i(x) = x_i$, une projection de $\mathbb{R}^d$sur $\mathbb{R}$.

2.2 Les processus de Poisson

Soit $(E, \mathcal{E})$ un espace d'état. Nous allons considérer *E* comme un espace localement compact à base dénombrable, et $\mathcal{E}$ une tribu associée.

Définition 1.7. Soit $(x_n)_{n\geq 1}$ une suite de points de *E*. On appelle mesure ponctuelle sur *E*, une mesure *m* de la forme :

$$m = \sum_{n=1}^{\infty} \delta_{x_n}$$

où δ_{x_n} est la mesure de Dirac en x_n.

Nous allons désigner par $M_p(E)$ l'espace de toutes les mesures ponctuelles définies sur *E*, et le munir de $\mathcal{M}_p(E)$, la plus petite tribu qui, $\forall\ F\epsilon\ \mathcal{E}$, rend mesurable l'application :

$$(M_p(E), \mathcal{M}_p(E)) \longmapsto ([0,\infty], \mathcal{B}_{[0,\infty]})$$

$$m \longmapsto m(F)$$

Définition 1.8. On appelle processus ponctuel, toute application mesurable *N* de la forme :

$$N\colon (\Omega, F, \mathbb{P}) \ \rightarrow\ (M_p(E), \mathcal{M}_p(E))$$

Définition 1.9. Soient Λ une mesure sur $\mathcal{E}$, finie sur tout compact de *E*, et *N* un processus ponctuel. On dit que *N* est un processus ponctuel de Poisson de mesure d'intensité μ si :

→ Pour tout F ∈ E, on a :

$$N(F\ )\ \sim\ P(\Lambda(F\ ))$$

→ Pour une famille d'ensemble disjoints $(F_k)_{k\geq 1}$ de $\mathcal{E}$, $(N(F_k))_{k\geq 1}$ forme une famille de variables aléatoires indépendantes. On notera :

$$N\ \sim\ P\ P\ P\ (\Lambda)$$

Soit *N* un processus ponctuel de Poisson sur une région $A \subset \mathbb{R}^d$ d'intensité non-homogène $\lambda(\bullet, \theta)$, appartenant à une famille paramétrique spécifique, θ un paramètre à estimer. On a :

$$E(N(A))\ =\ \Lambda(A,\theta)\ = \int_A \lambda(x,\theta)dx$$

Théorème 1.4. Soit N_1, N_2, ... une suite de processus de Poisson sur *A*. La suite est dite convergente en distribution vers *N*, *i.e* :

$$N_n \xrightarrow{d} N$$

si pour tout choix de *m* sous-ensembles bornés A_1, A_2, ..., A_m tels que $\mathbb{P}\big(N(\partial A_j) = 0\big) = 1, j = 1,\dots,m, \partial A_j$ étant la frontière de A_j, la distribution jointe $(N_n(A_1), N_n(A_2), \dots, N_n(A_j))$ converge en distribution vers $(N(A_1), N(A_2), \dots, N(A_j))$.

[12] Une k-marche $\{v_0, v_1, \dots, v_k\}$ sur un graphe (G,E) est une séquence de sommets dans G, suivant une règle de déplacement aléatoire (probabiliste).

On pose

$$A0 \ = \ \{x \ \in \ \mathbb{R}^d, \Pi_i(x) \ > \ \rho\}$$

Pour des raisons de convenance, on a nécessairement :

$$\Lambda(A_0, \theta) \ = \int_{A_0} \lambda(x, \theta) dx \ = \int_{A_0} f(x) dx$$

Finalement on déduit :

$$P \ (\ \max_{1 \leq j \leq n} \left(\prod_i \left(X_j^t \right) < \rho \right)) \ = \ P \ (N(A_0) \ = \ 0) \ = \ e^{-\int_{A_0} f(x) dx}, \text{ t fixé}$$

2.4 La Preuve

Soit $G^{(d)}(V \ ; \ r; \ E)$ un graphe géométrique formé de $|V| = N^d$ points en dimension *d*, et *E* l'ensemble des arêtes du graphe défini par :

$$\left(v_i, v_j \right) \in \ E \ \Leftrightarrow \ v_i \ \sim \ v_j \ \Leftrightarrow \ 0 \ < \| \ v_i \ - \ v_j \ \| \leq \ r, \qquad v_i, v_j \ \in \ V$$

Considérons la transformation de l'espace associant à chaque sous-ensemble de sommets $U \subset V$, l'ensemble des voisins les plus proches de chacun de ses sommets :

$$S(U) \ = \ \{w \ \in \ V \ | v \ \in \ U, v \ \sim \ w\}$$

Nous notons $S_t(U)$ la transformation[13] consécutive de *U* à la période $t \geq 0$.

Si l'on fixe un point initial v_0 dans *V*, les itérations $(S_t(v_0))$ de la transformation S engendrent un chemin (une trajectoire) sur le graphe $G^{(d)}(V \ ; \ r; \ E)$, et décrit un processus de Markov discret $\{v_t\}_{t \in \mathbb{N}}$, dont la probabilité de transition est définie par :

$$T_{ij} = Pr[v_t + 1 \ = \ j | v_t \ = \ i] \ > \ 0 \ \Leftrightarrow \ i \ \sim \ j$$

$$= D^{-1} A, \qquad D \ = \ diag(deg(1), \ldots, deg(|V|))$$

$$= 1/deg(i) \ si \ (i, j) \ \in \ E$$

où *A* est la matrice adjacente du graphe, et la probabilité de transition de *i* à *j* en $t > 0$ étapes est donnée par :

$$p_{ij}^{(t)} \ = \ (T^t)_{ij}$$

Soit g une densité discrète, $(v) \ \geq \ 0, v \ \in \ V$, définie sur $G^{(d)}(V \ ; \ r; \ E)$ tel que :

$$\sum_{v \epsilon V} g(v) = 1$$

$$\Gamma(U) = \sum_{u \epsilon U} g(u)$$

Chaque état du système à l'instant t est caractérisé par g. Γ définit une mesure sur *V*, et si $\|V\|$ représente l'ensemble des parties de *V*, alors $(V, \|V\|, \Gamma)$ est un espace mesurable[14]. En indexant chacun des éléments de *V* par *j*, avec $1 \leq j \leq N^d$, on peut caractériser plus explicitement g à chaque instant *t* :

$$\sum_{j=1}^{N^d} g_j(S_t(v_j)) = 1$$

Pour $n \ \geq \ 1, soit \ \{Y_1, \ldots, Y_n\}$ une suite de variables aléatoires (i.i.d.) prenant leurs valeurs dans l'ensemble $\{1, \ldots, N^d\}$ avec les probabilités :

[13] S_t est une loi dynamique sur un système thermodynamique dont l'espace des phases est *V* (Mackey,1992).

[14] Un système thermodynamique dont on désire étudier la dynamique.

$$P\ (Y_t\ =\ l)\ =\ g_l(S_t(v_l))$$

V étant isomorphe à l'ensemble $\{e_1, e_2, \dots, e_{N^d}\}$, on s'intéresse à la dynamique de la marche alétoire sur le graphe $G^{(d)}(V\ ;\ r;\ E)$; elle correspond à une marche aléatoire sur l'algèbre de Clifford Cl_V^{sym}, définie pour un élément $v_j \in V$ et à l'étape *n* par :

$$\Upsilon_n^j = \prod_{k=1}^{n} \varsigma Y_k$$

où ςY_k est un blade aléatoire de Cl_V^{sym}. Pour généraliser la marche aléatoire à tout le graphe, il suffit de considérer l'état du système à chaque instant t en fonction de tous les blades $\varsigma_j, 1 \le j \le N^d$, et caractériser son évolution de la première période à la période n à travers un vecteur algébrique de Clifford définit par :

$$\prod_{t=1}^{n}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)$$

Par ailleurs, chaque graphe G_U est caractérisé par des blades ς_U uniques,

On en déduit[15] ainsi la probabilité :

$$P\left(G_t^{(d)} = G_U\right) = \langle\prod_{i=1}^{t}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)\ \ ,\varsigma_U\rangle$$

Par ailleurs, à chaque point $v_{ij}\ , 1 \le j \le k\ de\ U$, est associé des attributs $x_{ij} = (x_{ij}^1, x_{ij}^2, \dots, x_{ij}^d) \in \mathbb{R}^d$. Soit G_Z le graphe résultant de G_U, à travers la contrainte :

$$\left\{\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(x_{i_j}\right) < \rho\right)\right\}$$

avec Π_i l'application définit pour $i \in [1;\ d]\ par\ x \mapsto \Pi_i(x) = x_i$, une projection de $\mathbb{R}^d\ sur\ \mathbb{R}$.

On trouve donc la formule finale :

$$P_{\boldsymbol{G}_t^{(d)}=G_U}\left(\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(X_j^t\right) < \rho\right)\right) = \frac{P(\max_{1\le j\le|V|}\left(\prod_i\left(X_j^t\right) < \rho\right), \boldsymbol{G}_t^{(d)} = G_U)}{P(\boldsymbol{G}_t^{(d)} = G_U)}$$

$$= \frac{P\left(\boldsymbol{G}_t^{(d)} = G_Z\right) * P(\max_{1\le j\le|V|}\left(\prod_i\left(X_j^t\right) < \rho\right)}{P(\boldsymbol{G}_t^{(d)} = G_U)}$$

$$= P(\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(X_j^t\right) < \rho\right) * \frac{\langle\prod_{i=1}^{t}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)\ \ ,\varsigma_Z\rangle}{\langle\prod_{i=1}^{t}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)\ \ ,\varsigma_U\rangle}$$

$$= e^{-\int_{A_0} f(x)dx} * \frac{\langle\prod_{i=1}^{t}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)\ \ ,\varsigma_Z\rangle}{\langle\prod_{i=1}^{t}(\sum_{j=1}^{N^d} g_j\left(S_t(v_j)\right)\varsigma_j)\ \ ,\varsigma_U\rangle}$$

$$\neq P(\max_{1\le j\le|V|}\left(\prod\nolimits_i\left(X_j^t\right) < \rho\right) = e^{-\int_{A_0} f(x)dx}$$

3. Application à une distribution GEV : Gumbel ($\Lambda(x)$)

Soit f une densité de probabilité sur $\mathbb{R}^2, et\ X_1, X_2, \dots, X_8$ une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) de densité f. On pose $X = \{X_1, X_2, \dots, X_8\}$. Notre centre d'intérêt ici est le graphe $G^{(d)}(X\ ;\ r;\ E)$, où E est l'ensemble des arrêtes du graphe. Dans notre exemple, nous prenons :

$$f(x,y) = \frac{1}{x}\exp\left(-\frac{y}{x} - x\right)1_{\mathbb{R}_+^*\times\mathbb{R}_+^*}(x,y)$$

[15] Voir aussi Schott et Staples (2012).

$$f_x(x) = \int_0^\infty \frac{1}{x} \exp\left(-\frac{y}{x} - x\right) dy = e^{-x}, x > 0$$

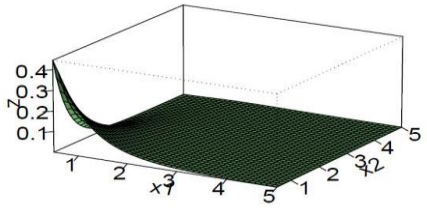

Figure 2 – Tracé de f_{xy} sur [0, 5; 5]

Ici *V* contient $2^3 = 8$ points $\{v_1 = 000, v_2 = 001, v_3 = 101, v_4 = 100, v_5 = 010, v_6 = 011, v_7 = 111, v_8 = 110\}$ dont les attributs associés à vi est un vecteur alatoire bidimensionnel *Xi*. Les éléments de la base de l'algèbre de Clifford symétrique associée Cl_V^{sym} sont des blades de la forme $\varsigma_I, I \subset [3] = \{1, 2, 3\}$, canoniquement ordonnés grâce à la relation :

$$\varsigma_I \prec \varsigma_J \Leftrightarrow I \prec J \Leftrightarrow \sum_{i \in I} 2^{i-1} < \sum_{j \in J} 2^{j-1}$$

Soit :

$$\varsigma_\Phi \prec \varsigma_1 \prec \varsigma_2 \prec \varsigma_{12} \prec \varsigma_3 \prec \varsigma_{13} \prec \varsigma_{23} \prec \varsigma_{123}$$

Le 3-hypercube Q_3 associé à l'algèbre de blade quotienté $\mathbb{R}\mathcal{B}_{3,0}/\langle e_\alpha + 1\rangle$ isomorphe ici à Cl_V^{sym} est représenté[16] ci-dessous (figure 3).

La densité g caractérisant l'état du système à chaque instant t est uniforme, avec : $g_j\left(S_t(v_j)\right) = 1/8$, et

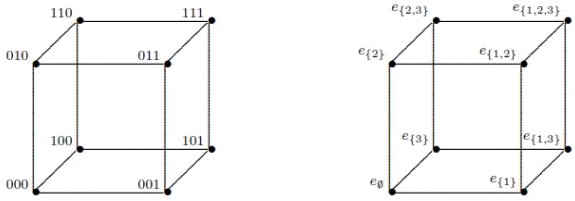

Figure 3 – $\mathbb{R}\mathcal{B}_{3,0}/\langle e_\alpha + 1\rangle$ à droite et Q_3 à gauche

Soit $U \subset V$ un sous-graphe de *V*, tel que $U = \{v_1 = 000, v_2 = 001, v_3 = 101\}$.

$$P\left(G_2^{(d)} = G_U\right) = \langle \prod_{i=1}^{2} (\sum_{j=1}^{8} g_j\left(S_t(v_j)\right)\varsigma_j) \ , \varsigma_U \rangle$$

$$\prod_{i=1}^{2} (\sum_{j=1}^{8} g_j\left(S_t(v_j)\right)\varsigma_j) = \frac{1}{64}\{8\varsigma_\Phi + 8\varsigma_1 + 8\varsigma_2 + 8\varsigma_{12} + 8\varsigma_3 + 8\varsigma_{13} + 8\varsigma_{23} + 8\varsigma_{123}$$

$$P\left(G_2^{(d)} = G_U\right) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Par ailleurs, si les attributs de $v_1, v_2 \; et \; v_3$ sont respectivement (0, 66; 0, 45), (1, 59; 3, 45), (2, 88; 0.51), alors la restriction

$$\{\max_{1 \le j \le 8} (\prod\nolimits_1 (x_j^2) < \rho)\}, \rho = 2$$

[16] Cette représentation est dite de Cayley.

engendre le graphe d'ensemble de sommets $Z = \{v_1, v_2\}$, avec Π_1 l'application définit par $x \mapsto \Pi_1(x) = x_1$, une projection de $\mathbb{R}^2 \; sur \; \mathbb{R}$ $(ici \; t = 2)$.

$$P\left(G_2^{(d)} = G_Z\right) = \frac{1}{8} + \frac{1}{8} = \frac{1}{4}$$

$$P\left\{\max_{1\leq j\leq 8}\left(\prod\nolimits_1\left(x_j^2\right) < \rho\right)\right\} = P\left\{\max_{1\leq j\leq 8}\left(\prod\nolimits_1\left(x_j^2\right) - \log 8\right) < \rho - \log 8\right\} \longrightarrow \Lambda(\rho - \log 8) = \exp\left(-e^{-(\rho - \log 8)}\right)$$

$$P_{G_2^{(d)} = G_U}\left(\max_{1\leq j\leq 8}\left(\prod\nolimits_1\left(X_j^8\right) < \rho\right)\right) = \frac{2 * \exp\left(-e^{-(\rho - \log 8)}\right)}{3} \simeq 0{,}23 \neq 0{,}34 = \exp\left(-e^{-(\rho - \log 8)}\right)$$

4. ESTIMATION D'UN COEFFICIENT DE RÉDUCTION DE L'INFLATION

Dans le calcul de la probabilité $P_{G_t^{(d)} = G_U}\left(\max_{1\leq j\leq |V|}\left(\prod_i\left(X_j^t\right) < \rho\right)\right)$, on retrouve le terme :

$$\frac{\langle \prod_{i=1}^{t}\left(\sum_{j=1}^{N^d} g_j\left(S_t\left(v_j\right)\right)\varsigma_j\right) \; , \varsigma_Z\rangle}{\langle \prod_{i=1}^{t}\left(\sum_{j=1}^{N^d} g_j\left(S_t\left(v_j\right)\right)\varsigma_j\right) \; , \varsigma_U\rangle}$$

Or, dans la pratique, on ne connaît pas les lois dynamiques d'un système thermodynamique dont les données sur les individus sont résumées dans une base de données obtenue à la période t_k fixée. Autrement dit, on ne connaît pas g, la fonction de densité caractérisant l'état du systéme à chaque instant, et on n'ignore la pré-pondérance d'un processus stochastique de graphes géométriques régissant la trajectoire des observables. Pour pallier à cette situation, l'idéal est de construire un invariant temporel, lorsque le graphe est considéré vérifiant certaines conditions. L'objectif est d'estimer pour n'importe quelle distribution dont la classe de convergence en type en connue (cf. théorème de convergence en type de Gnedenko), un **coefficient de réduction de l'inflation identifiée.**

Approximation de $P\left(G_t^{(d)} = G_U\right)$

Comme nous l'avons remarqué dans la section précédente, l'approximation de la quantité $P\left(G_t^{(d)} = G_U\right)$ est facile en supposant que g est uniforme (on uniformise la densité) : si le cardinal $n_U \; de \; U$ est connu, on a bien dans ce cas :

$$P\left(G_t^{(d)} = G_U\right) = n_U \times \left[\frac{n_V^{t-1}}{n_V^t}\right] = \frac{n_U}{n_V}$$

où n_V est ici le cardinal de *V*.

Approximation de $P\left(G_t^{(d)} = G_Z\right)$

Nous construisons ici une approximation de $P\left(G_t^{(d)} = G_Z\right)$, où G_Z le graphe résultant de G_U , à travers la contrainte :

$$\left\{\max_{1\leq j\leq |V|}\left(\prod\nolimits_i\left(x_{i_j}\right) < \rho\right)\right\}$$

avec Π_i l'application définit pour $i \in [1;\; d] \; par \; x \longmapsto \Pi_i(x) = x_i$, une projection de $\mathbb{R}^d \; sur \; \mathbb{R}$.

On suppose que le domaine d'attraction $D(H_i)$ de la distribution marginale H_i des observables X_j^i pour tout individu *j* est de celui de l'une des distributions suivantes :

- **Fréchet (α >0)** :

$$\Phi_\alpha = \begin{cases} 0 & si \; x \leq 0 \\ e^{-x^{-\alpha}} & si \; x > 0 \end{cases}$$

- **Weibull (α >0)** :

$$\Psi_\alpha = \begin{cases} e^{-|x|^\alpha} & si\ x \leq 0 \\ 1 & si\ x > 0 \end{cases}$$

- **Gumbel : $\Lambda = e^{-e^{-x}}, x \epsilon \mathbb{R}$**

Alors, on peut approcher n_Z , le cardinal de Z par :

$$\hat{n}_Z = \begin{cases} \Phi_\alpha(\rho) n_U & pour\ Fréchet \\ \Psi_\alpha(\rho) n_U & pour\ Weibull \\ \Lambda(\rho) n_U & pour\ Gumbel \end{cases}$$

Par suite, le coefficient de réduction de l'inflation est calculé :

$$\hat{n}_Z = \begin{cases} \frac{\Phi_\alpha(\rho) n_U}{n_V} \left[\frac{n_U}{n_V}\right]^{-1} = \Phi_\alpha(\rho) & pour\ Fréchet \\ \frac{\Psi_\alpha(\rho) n_U}{n_V} \left[\frac{n_U}{n_V}\right]^{-1} = \Psi_\alpha(\rho) & pour\ Weibull \\ \frac{\Lambda(\rho) n_U}{n_V} \left[\frac{n_U}{n_V}\right]^{-1} = \Lambda(\rho) & pour\ Gumbel \end{cases}$$

Ainsi, si $g\ est\ uniforme, et\ \rho > 0$, la distribution insensible à la géométrie des données est la distribution de Weibull, et la distribution la moins inflationniste est celle de Fréchet.

5. Conclusion

Nous avons montré que la géométrie des données modélisée grâce à l'aide d'un processus de graphes géométriques déflate les probabilités calculées avec des distributions GEV. Autrement dit, tenir compte du système de données dans son ensemble pour évaluer des extrêmes, donne des probabilités différentes de celles que l'on obtient de par les techniques de *block maxima* ou POT.

Une des limites principales du résultat obtenu est l'indépendance des points du graphe. Ce travail pourra toutefois être étendu aux modèles spatiaux, en intégrant une interdépendance au niveau des points des graphes géométriques utile à une analyse globale des relations géométriques liant différents marchés boursiers en finance, le déplacement des robots avec des systèmes d'adjustement automatiques aux chocs, le traitement des réseaux ad-hoc, l'évolution dynamique de particules dans un champ quantique, etc.

Funding: This research received external funding from Benin Government and Fondation Vallet - Fondation de France.
Acknowledgments: The authors would like to thank M. Bocar Toure (ex-Director of the ENSAE). We acknowledge all the administrative and technical support received from the National School of Statistics and Economical Analysis (Dakar, Senegal), and University of Montreal while completing the manuscript.
Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Erdős, P. ; Rényi, A. (1959). On Random Graphs, Publicationes Mathematicae 6 : 290297.
[2] Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London.
[3] de Haan, L.,(1984), A spectral representation for maxstable processes, The Annals of Probability, 12(4) :11941204.
[4] de Haan, L. et Ferreira, A. (2006). Extreme value theory : An introduction. Springer Series in Operations Research and Financial Engineering.
[5] de Haan, L. et Ferreira, A. (2010). On the block maxima method in extreme value theory.
[6] Mackey, C. M. (1992). Time's arrow : the origin of thermodynamique behavior, Springer-Verlag New York, Inc.
[7] Penrose, M. D. (1992). Semi-min-stable processes. Annals of Probability, 20(3) :14501463.
[8] Pickands, J. III (1975). Statistical inference using extreme order statistics. Ann. Statist. 3, 119-131.
[9] Schott, R., Staples, G., S. (2012), OPERATOR CALCULUS ON GRAPHS : Theory and Applications in Computer Science, Imperial College Press, ISBN-10 1-84816-876-4.