December, 2024

IA703

Algorithmic Information & Artificial Intelligence

Micro-study

teaching.dessalles.fr/FCI

Auteurs : Luca Hachani & Robin Guiavarch

La Complexité comme Features pour la Segmentation d'Arbres Fractals

Résumé

Cette étude explore les propriétés des arbres fractals de Mandelbrot-Vicsek. Leurs encodages binaires sont obtenus grâce aux marches de Harris, qui transforment les arbres en séquences binaires capturant l'information de leurs motifs. L'étude propose un compresseur spécifique exploitant les répétitions de ces motifs et atteignant des taux de compression comparables à ceux d'outils classiques comme bzip2. Une analyse des distances NCD permet de segmenter les arbres en fonction de leurs multiplicités et leurs maturités.

1 Arbres fractals de Mandelbrot-Vicsek : Définition, Construction et 1^{er} Encodage

1.1 Définition générale d'une fractale

Une fractale F peut être vue comme l'objet résultant de l'application répétée à l'infini d'une transformation S à un ensemble de départ E.

Cette transformation S peut être définie comme une ou la somme de m contractions S_i qui généreraient une ou m copies transformées à partir d'un ensemble donné $E \subset \mathbb{R}^n$. On définit ainsi S ces procédures qui, répétées à l'infini, construiront les fractales telles que :

$$S = \sum_{i=1}^{m} S_i$$

Ainsi, pour chaque fractale que l'on voudra construire, il existe un unique support $F \subset \mathbb{R}^n$ de celle-ci pour lequel la procédure S est une involution :

$$\exists! F \subset \mathbb{R}^n, \quad F = \bigcup_{i=1}^{m} S_i(F)$$

Ce support F pourra être approché à l'infini par n'importe quel sous-ensemble compact (fermé) non vide $E \subset \mathbb{R}^n$ tel que [1] :

$$d(F, S^k(E)) \xrightarrow{k \to +\infty} 0, \quad \text{avec} \quad S^k = S \circ S \circ \cdots \circ S \quad (k \text{ fois})$$

1.2 Définition d'une pré-fractale

Dans la réalité physique, les objets que l'on considère comme étant fractals sont en fait pré-fractals. La définition formelle du support d'une pré-fractale est l'ensemble $S^k(E)$. La distance d'Hausdorff permettrait donc de quantifier la maturité d'une pré-fractale à une étape k de sa construction. Plus cette distance $d(F, S^k(E))$ tend vers 0, plus le support de cette pré-fractale pourrait être considéré comme "mature". Dans la suite de ce travail, on considèrera qu'une pré-fractale obtenue à une étape de construction k est plus ou moins mature en fonction d'un critère de compressibilité.

1.3 Construction des arbres fractals de Mandelbrot-Vicsek

Dans ce travail, nous nous intéresserons tout particulièrement à une famille d'arbre fractal introduite par B.B.Mandelbrot & T.Vicsek dans [2]. Ces arbres fractals uniformes déterministes relativement simples à construire ont l'avantage d'être de très bons exemples sur lesquels illustrer nos modèles.

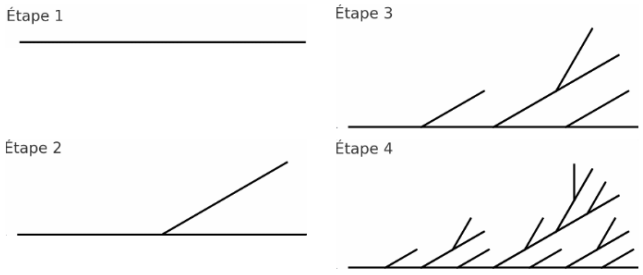


FIGURE 1 – Construction - Arbre fractal

Comme montré sur la figure 1, un arbre fractal de Mandelbrot-Vicsek de multiplicité $m = 3$ se construit par contraction de son 1er motif - un lien (deux nœuds) - en $m = 3$ liens (un nœud racine et trois nœuds feuilles). On applique après chaque étape cette similitude S sur chacun des liens de la pré-fractale résultante.

1. $\forall A \in D \subset \mathbb{R}^n$, $A_\delta = \{x \in D \mid \exists a \in A, |x - a| \leq \delta\}$ est l'épaississement de A d'un rayon $\delta \in \mathbb{R}^+$. On définit alors la distance d'Hausdorff entre deux sous-ensembles $(A, B) \subset D^2$ telle que :

$$\forall (A, B) \subset D^2, \ d(A, B) = \inf\{\delta \in \mathbb{R}^+ \mid A \subset B_\delta \text{ et } B \subset A_\delta\}.$$

1.4 Un premier encodage naïf et lourd

Une première approche pour encoder un arbre fractal est de partir de sa définition. À savoir, d'encoder la similitude S que l'on répète une infinité de fois. Mais il faut pour ceci définir la classe d'un arbre, ce qu'est un nœud, attribuer une valeur aux nœuds pour encoder l'arbre. Tout ceci donne un programme long et un encodage lourd.

2 Les marches de Harris : Une méthode d'encodage simple & éfficace

2.1 Définition d'une marche de Harris

La marche de Harris est une courbe fractale W qui nous permettra d'encoder dans une séquence binaire X l'information contenue dans un arbre. Cette marche consiste à noter la génération des nœuds rencontrés en suivant le contour de l'arbre par un parcours en profondeur. Pour plus de simplicité, notre encodage consistera à enregistrer à chaque pas d'une marche si celui-ci avait permis de monter (1) ou de descendre (0) d'une génération. Ainsi, notre encodage peut être défini à partir d'une marche de Harris W tel que :

$$\forall t \in [\![1, n-1]\!],\ X(t) = \frac{1 + (W(t+1) - W(t))}{2} \in \{0,1\}^{n-1}, \quad \text{avec } |W| = n$$

La marche de Harris capture efficacement la morphologie d'un arbre, mais ne tient pas compte des labels des nœuds qui le composent. Si l'information des labels des nœuds nous était nécessaire, nous aurions plutôt opté pour la notation de neveu [2]. Cependant, il y aurait eu un coût supplémentaire à payer en matière de complexité. Or dans notre étude, cette information ne nous intéresse pas. Ainsi, nous avons choisi la marche de Harris.

Exemple : Considérons un arbre fractal simple $T(V, E)$ représenté sur la figure 2. La marche de Harris pour cet arbre est donc $W = [0, 1, 2, 1, 0, 1, 2, 1, 2, 1, 0]$ ce qui donne l'encodage $X = [1, 1, 0, 0, 1, 1, 0, 1, 0, 0]$ pour $T(V, E)$.

2.2 Propriété des marches de Harris pour les arbres fractals

Pour un arbre fractal, il est toujours possible d'intuiter sa marche de Harris. Rappelons qu'un objet fractal peut être défini par son ou ses motifs. Mathématiquement, cela se traduit par le fait que la transformation S à l'origine de la construction d'une fractale suffit à la caractériser. Or, en capturant l'information d'un arbre fractal, la

2. La notation de neveu conserve pour chaque nœud son indice dans sa fratrie ainsi que l'indice de ses ascendants dans leur propre fratrie. la racine est désignée par l'ensemble vide $\emptyset$, (les index de la marche de Harris représenté sur la figure 2 sont construits à partir de la notation de neveu)

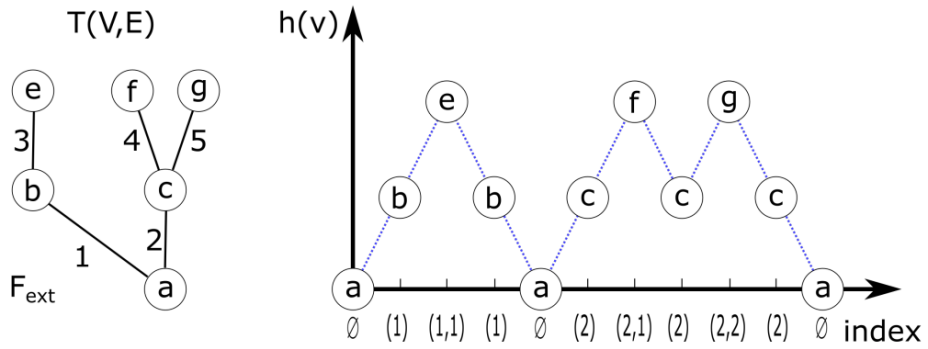


FIGURE 2 – (gauche) Représentation d'un arbre $T(V, E)$. (droite) Marche de Harris de l'arbre $T(V, E)$ vec en abscisse la notation de neveu des noeuds.

marche de Harris capturera aussi l'information de ses motifs. On verra ainsi apparaître des motifs dans les marches de Harris correspondant à ceux observés dans nos arbres.

Aussi, il est possible d'intuiter ces motifs et leurs récurrences (cf figure 3) et ainsi construire une marche de Harris sans que cela nécessite qu'on parcourt le contour de l'arbre fractal que l'on cherche à encoder. Un avantage d'autant plus appréciable que le niveau de détail d'une pré-fractale croît exponentiellement à mesure où l'on répète la transformation S et finit par tendre vers l'infini [3]

3 Approche de la complexité des arbres fractals de Mandelbrot-Vicsek

Les arbres fractals, bien que comportant un niveau de détail infini, ont une complexité faible et finie. Cette complexité peut être approchée en analysant les étapes de construction des fractals. Nous allons ainsi pour une machine de Turing hypothétique proposer une structure décrivant la séquence binaire encodant nos arbres fractals. Nous nous appuierons sur la classe des arbres fractals de Mandelbrot-Vicsek pour illustrer la haute compressibilité des objets fractals.

3.1 Marche de Harris des arbres fractals de Mandelbrot-Vicsek

Ainsi, nous avons su intuiter la marche de Harris d'un arbre fractal de Mandelbrot-Vicsek de multiplicité $m > 2$ et cela à toutes les étapes k de sa construction (cf figure 3). L'algorithme est plutôt simple et peut être décomposé pour chaque étape de construction en 3 temps pour une séquence d'encodage binaire X_k donnée :

— On identifie dans la séquence binaire X_{k-1} encodant l'étape $k-1$ de construction de l'arbre fractal, les m motifs présents. On isole alors dans la séquence X_{k-1}, le premier des m motifs de longueur $2 \cdot m^{k-1}$

3. Notons que par définition, nous ne pourrons jamais tracer la marche de Harris d'une fractale, mais seulement d'une pré-fractale. Rappelons qu'une pré-fractale (cf section 1.2) est un objet dont le support est défini par $S^k(E)$ et que pour $k \neq +\infty$ ce-dernier se rapproche du support F de la véritable fractale à mesure que $k \to +\infty$.

— On insère alors à la position $t = 2^{k-1}$ du premier motif de la séquence X_{k-1}, $m-1$ motifs et obtient ainsi 1 des m motifs de la séquence binaire encodant l'étape N de construction de l'arbre fractal
— On réplique m fois cette nouvelle séquence pour obtenir la séquence binaire complète X_k encodant l'étape k de construction de l'arbre fractal.

En bouclant cet algorithme à l'infini, on pourrait théoriquement tracer la marche de Harris d'une authentique fractale de Mandelbrot-Vicsek de multiplicité $m > 2$. Évidemment, dans un temps fini, autrement dit en réalité, nous ne pourrons qu'approcher son support F par une pré-fractale $S^k(E)$ à une étape donnée k. Sur la figure 3, on retrouve les marches de Harris pour un arbre fractal de Mandelbrot-Vicsek de multiplicité $m = 3$ respectivement lors de la 1^{re}, 2^{me} et 3^{me} étape de sa construction.

3.2 Complexité et niveau de détail

Voici un programme[4] que pourrait interpréter une machine de Turing donnée U pour construire la marche de Harris d'un arbre fractal de Mandelbrot-Vicsek de multiplicité m à l'étape $k = N$ de sa construction :

```
boucle(1,
       X_0,
       repetition(insert(patern(X, 2*m**{k-1}),
                         patern(X, 2*(m-1)*m**{k-1}),
                         2**{k-1}),
                  m),
       N)
```

Traduite en binaire, cette structure peut être représentée par un code binaire délimité par des espaces tel que :

```
0 00 11 <X_0> 1 1 0 11 0 10 0 01 11 <X> 0 1 1 10 0 00 1 <m> 0 0 1 <k>
1 1 0 101 11 <X> 0 1 1 10 0 1 0 0 1 <m> 1 1 0 00 1 <m> 0 0 1 <k> 1 1
0 0101 1 10 0 0 1 <k> 1 1 1 <m> 1 <N>
```

Pour faciliter la lecture du code ci-dessus, on peut se référer pour les choix de codification des différents éléments présents dans cette structure au tableau 1

Dès lors, on constate que ce code binaire est relativement court alors qu'il contient l'information nécessaire pour construire l'encodage binaire X_k d'un objet $S^k(E)$ dont le niveau de détail lui tend vers l'infini. Ainsi, on peut approcher la complexité d'un arbre fractal de Mandelbrot-Vicsek par la longueur de ce code binaire. Cette longueur est paramétrée par la longueur en binaire de l'entier m et N suivant une

4. Dans ce programme, l'opérateur "boucle" incrémente de 1 à N la valeur de k et applique récursivement le reste de la structure à la nouvelle chaine X_k obtenue entre chaque itération à partir de la chaine initiale X_0

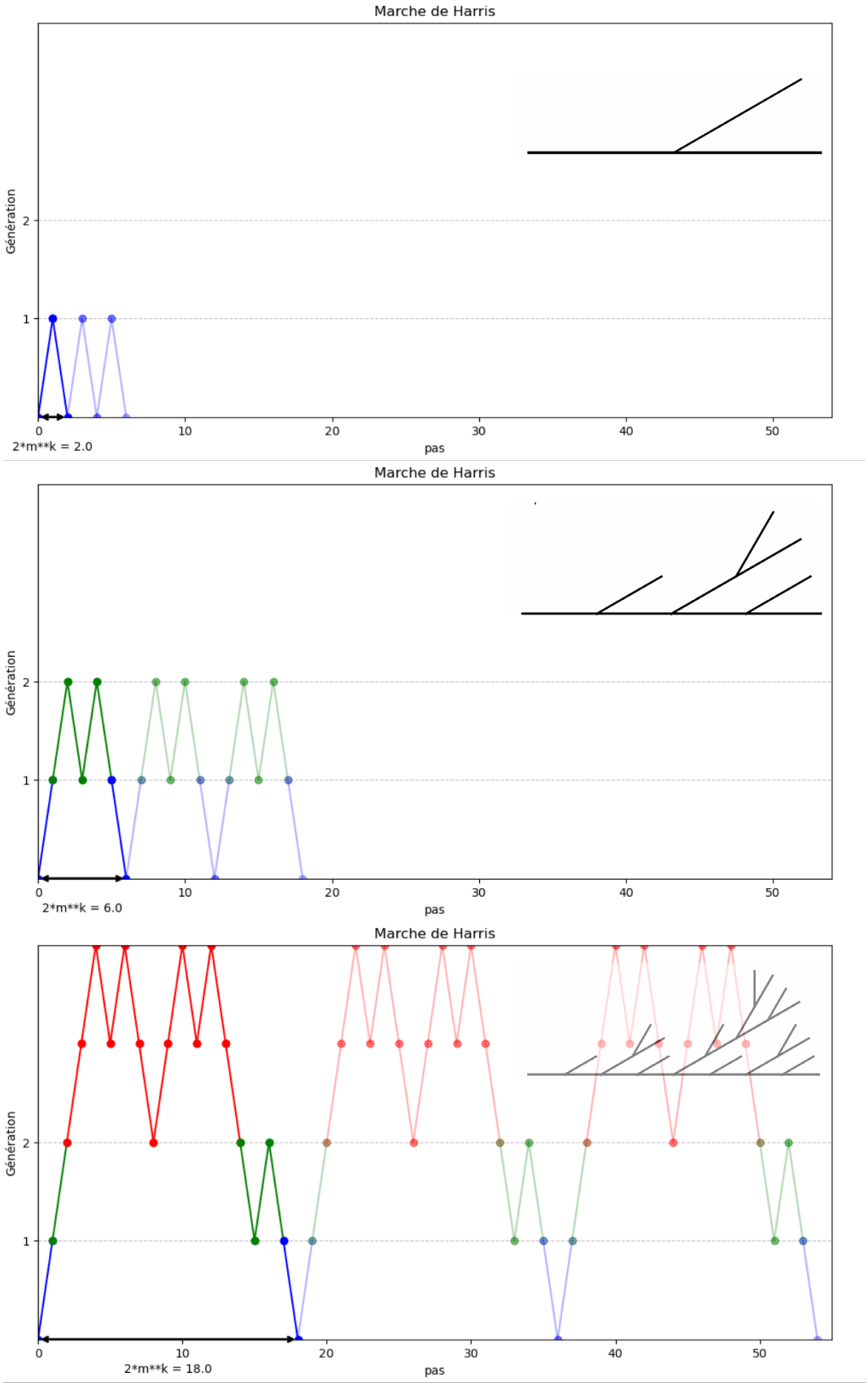


FIGURE 3 – Représentation sous forme de marche des Harris des $k = \{1, 2, 3\}$ premières étapes de construction d'un arbre fractal de Mandelbrot-Vicsek de multiplicité $m = 3$.

Élément	Description	Code Binaire
opérateur	**boucle**	`0 00`
opérateur	**repetition**	`0 11`
opérateur	**insert**	`0 10`
opérateur	**patern**	`0 01`
opérateur	Puissance (**)	`0 00`
opérateur	Multiplication (*)	`0 1`
opérateur	Soustraction (-)	`0 0`
string	Séquence binaire **X**	`11 <X>`
entier	Entier **n**	`1 <n>`

TABLE 1 – Récapitulatif des codes binaires utilisés pour les différents éléments présents dans la structure de notre programme.

croissance logarithmique alors que le niveau de détail de notre pré-fractale lui croît exponentiellement.

4 Compression des marches de Harris

Objectifs :

- Construire un compresseur spécifique aux arbres fractals qui tient compte des spécificités de leur marche de Harris dans le but d'obtenir un taux de compression similaire à celui des compresseurs classiques comme zip, gzip, bzip2.
- Choisir le meilleur compresseur pour notre clustering de fractales selon les distances NCD.

4.1 Notre compresseur - construction

Nous avons conçu un compresseur spécifique aux arbres fractals basé sur leur représentation binaire obtenue via la marche de Harris. Cette approche exploite les répétitions et les duplicats miroirs comme illustrés sur la figure 5) pour réduire la taille de la représentation tout en préservant la structure hiérarchique des arbres. Les étapes principales de ce compresseur sont articulées sur la figure 4.

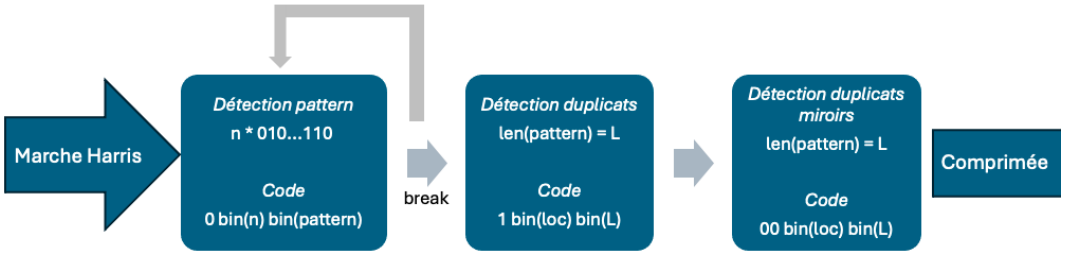


FIGURE 4 – Pipeline de compression.

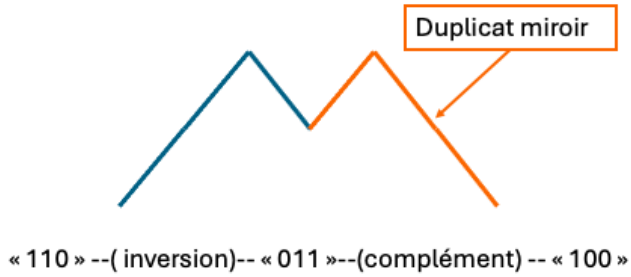


FIGURE 5 – Duplicats miroirs.

4.2 Performances des compresseurs

Afin d'évaluer les performances, nous avons comparé notre compresseur avec les compresseurs classiques tels que zip, gzip, et bzip2. Les résultats (voir tableau 2) montrent que notre compresseur atteint un taux de compression compétitif sur les arbres fractals. Voici les résultats :

— **X3_3** : Arbre fractal de multiplicité $m = 3$, de nombre d'itérations $k = 3$ et de taille 162.

— **X4_5** : Arbre fractal de multiplicité $m = 4$, de nombre d'itérations $k = 5$ et de taille 8192.

Signal	Compressor	Original Length	Compressed Length	Compression Rate
X3_3	Pipeline	162	73	0.55
X3_3	ZIP	162	25	0.85
X3_3	GZIP	162	37	0.77
X3_3	BZIP2	162	55	0.66
X4_5	Pipeline	8192	192	0.98
X4_5	ZIP	8192	129	0.98
X4_5	GZIP	8192	109	0.99
X4_5	BZIP2	8192	74	0.99

TABLE 2 – Performance des différents compresseurs.

Résultat 1 : On obtient bien des taux de compression équivalents pour les 2 signaux. Par ailleurs, on montre aussi le caractère hautement compressible de nos arbres fractals.

4.3 Qualité des compresseurs

On évalue la qualité de nos compresseurs selon 6 critères, 6 tests que nous leur faisons passer :

— Idempotence : $Z(Z(x)) \approx Z(x)$

— Symétrie : $Z(x, y) \approx Z(y, x)$

— Détection de régularités : $Z(100 \cdot x) \ll Z(\text{signal random})$

— Détection de dépendances : Si $y = 10 \cdot x$ alors $Z(x, y) \leq Z(x) + Z(y)$
— Consistance : $\frac{Z(\text{signal long})}{Z(\text{signal court})} \approx \frac{\text{len(signal long)}}{\text{len(signal court)}}$
— Auto-concaténation : $Z(x, x) \approx Z(x)$

Compresseur	Idempotence	Symétrie	Détection régularités	Dépendances	Consistance	Auto-concatenation
ZIP	Échoué	Passé	Passé	Passé	Échoué	Échoué
GZIP	Échoué	Passé	Passé	Passé	Échoué	Échoué
BZIP2	Passé	Passé	Passé	Passé	Échoué	Passé
Maison	Échoué	Échoué	Passé	Passé	Échoué	Échoué

FIGURE 6 – Tests compresseurs

Résultat 2 : On gardera notre compresseur pour calculer les distances NCD entre fractales de différentes multiplicités m. Pour l'accompagner, nous choisirons le compresseur bzip2 qui passe le plus de tests (tous sauf celui de consistance).

5 Distance NCD et clustering

Objectif : L'idée est de compresser les marches de Harris binaires d'arbres fractals en faisant varier leur multiplicité m ainsi que le nombre d'itérations k. En s'inspirant en partie du travail réalisé par Ming Li et al. [1], on veut montrer qu'à partir des distances NCD des différentes marches de Harris, on est capable de regrouper les arbres fractals en fonction de leur multiplicité m quelque soit le nombre d'itérations.

5.1 Première tentative de clustering sur des arbres fractals "immature"

Pour rappel, on nomme Xm_k la marche de Harris binaire de l'arbre fractal de multiplicité m à l'itération k. Sur la figure 7, deux matrices répertorient les distances NCD mesurées pour différents couples d'arbres de Mandelbrot-Vicsek et cela respectivement pour notre compresseur maison et le compresseur bzip2. Les couples sont choisis en faisant varier leur multiplicité entre 3 et 5 et leurs étapes de construction entre 2 et 4.

Interprétations des résultats de la figure 7 :
— Notre compresseur maison qui échoue à 4 tests sur 6 donne des résultats aberrants. On le discrédite pour cette étude. Nous n'utiliserons plus que le compresseur bzip2.

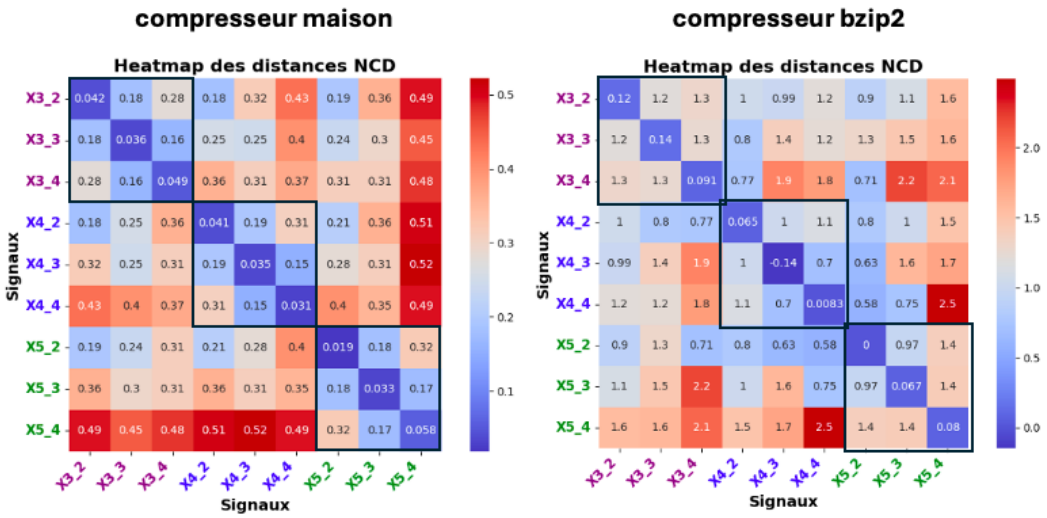


FIGURE 7 – Distance NCD - compresseurs maison et bzip2

— En ce qui concerne la matrice de distance NCD du compresseur bzip2, les résultats sont convaincants. On aperçoit bien la baisse de la distance NCD (dans les carrés noirs) entre arbre fractals de même multiplicité. Mais il semble tout de même y avoir une plus grande distance NCD entre des arbres fractals qui ont une différence d'itérations de 2. X4_4 avec X4_2, etc.

5.2 Maturité/immaturité d'une fractale du point de la compression

La figure 8 représente la croissance rapide du taux de compression de la marche de Harris à mesure que k l'étape de construction de l'arbre pré-fractal augmente. On observe un coude pour $k = 4$. Dans la suite, on statuera que notre arbre pré-fractal est immature si k est inférieur à 4 et mature sinon. La maturité représente l'état d'avancement d'une pré-fractale vis-à-vis de son état fractal final (cf section 1.2).

5.3 Deuxième tentative de clustering sur des arbres fractals "mature"

La figure 8 nous montre aussi la matrice de résultat des distances NCD réalisée avec le compresseur bzip2, mais cette fois-ci, on fait varier k de 6 à 8 pour comparer des pré-fractales matures. Nous allons voir que pour des pré-fractales matures la NCD définit une feature très pertinente pour faire de la segmentation.

Résultats - Discussion :

— La figure 8 met en évidence que le clustering fonctionne très bien pour des multiplicités $m \leq 4$. Pour des fractales matures en choisissant un seuil NCD = 0.3 on peut aisément les segmenter en fonction de leur multiplicité m.

— La figure 9 illustre l'extrême variabilité des tailles des signaux étudiés, allant de seulement 4 000 bits à plus de 80 millions. Pour mieux visualiser cette di-

versité, la matrice des distances NCD a été reconstituée en classant les marches de Harris par taille croissante. Cette représentation met en évidence une limite au-delà de laquelle le compresseur bzip2 ne parvient plus à fonctionner correctement lors de la concaténation des signaux (Z(x,y)). Cette limite se manifeste par une frontière en forme de quart de cercle, d'un rayon d'environ 10 millions de bits, observable notamment pour des combinaisons comme X4_7+X5_7 ou X6_6+X4_8. Cela explique pourquoi la NCD échoue à segmenter les pré-fractales de multiplicité $m > 4$.

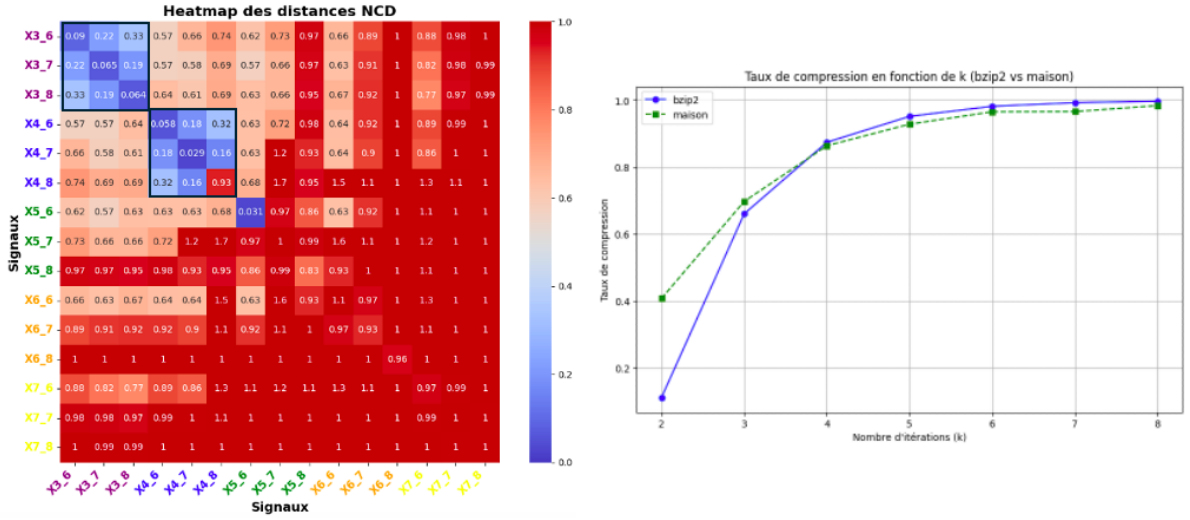
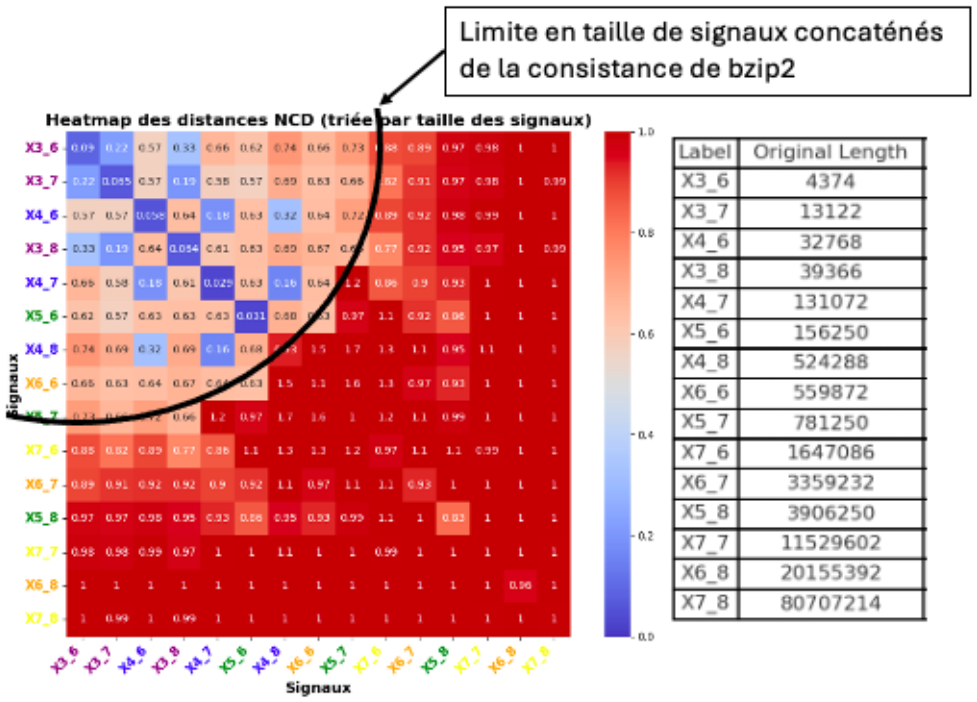


FIGURE 8 – Distances NCD pré-fractales Matures ; Taux de compression - itérations k



Label	Original Length
X3_6	4374
X3_7	13122
X4_6	32768
X3_8	39366
X4_7	131072
X5_6	156250
X4_8	524288
X6_6	559872
X5_7	781250
X7_6	1647086
X6_7	3359232
X5_8	3906250
X7_7	11529602
X6_8	20155392
X7_8	80707214

FIGURE 9 – Inconsistance compresseur bzip2

6 Conclusion

L'objectif principal de cette étude fut d'estimer au plus près la complexité de Kolmogorov d'un arbre fractal. Cette complexité peut être approximée par la longueur du code binaire du programme générant la marche de Harris. La brièveté de ce code reflète le caractère hautement compressible des objets fractals.

Dans un second temps, nous avons cherché à regrouper les arbres fractals en fonction de leur multiplicité m, en calculant la distance NCD entre leurs marches de Harris. Pour des valeurs faibles de m et un nombre limité d'itérations k, le clustering est efficace. Cependant, lorsque k et m deviennent trop grands, et que les signaux atteignent des tailles trop élevées, le calcul des distances NCD perd en pertinence en raison de l'inconsistance du compresseur bzip2 utilisé.

Références

[1] Ming Li et al. "The Similarity Metric". In : *IEEE Transactions on Information Theory* 50.12 (déc. 2004), p. 3250-3264. ISSN : 1557-9654. DOI : 10.1109/TIT.2004.838101. (Visité le 20/12/2024).

[2] B. B. Mandelbrot et T. Vicsek. "Directed Recursion Models for Fractal Growth". In : *Journal of Physics A : Mathematical and General* 22.9 (mai 1989), p. L377-L383. ISSN : 0305-4470. DOI : 10.1088/0305-4470/22/9/005. (Visité le 16/07/2021).