Nicolas OMONT 23/12/2002
Cédric AULIAC

Diffusion sur un graphe

Projet de modélisation dans le cadre du
D.E.A. « Application des Mathématiques et de l'Informatique à la Biologie »

Introduction

On considère le processus suivant. Sur chaque nœud d'un graphe arbitraire G, on place un jeton. A chaque pas de temps n, pour chaque jeton se trouvant sur chacun des nœuds g, on tire au hasard un des successeurs g' de g. Le jeton se retrouve ainsi en g' au temps n+1. On note J_i le vecteur des nombre de jetons sur chaque sommet au temps i.

On veut savoir si $\sum_{i=0}^{n-1} J_i$ possède une limite lorsque n tend vers l'infini. On veut ensuite vérifier cette prédiction à l'aide d'une simulation informatique. D'où le plan de ce rapport : en première partie, une rapide étude mathématique et en deuxième partie, les résultats de la simulation.

1 Etude Mathématiques

1.1 Remarque préliminaire

Il y a une remarque à faire qui permet de simplifier considérablement le problème mathématique en le ramenant à un domaine parfaitement balisé, celui des chaînes de Markov dans des espaces d'états finis.

Elle consiste dans le fait qu'il n'y a aucune interaction entre les jetons. Le processus peut donc être considéré comme la somme de processus indépendants de « diffusion » d'un jeton sur le graphe.

La conséquence est la suivante : le processus à un jeton correspond à une chaîne de Markov où la variable de l'espace d'état est le nœud sur lequel se trouve le jeton, la matrice de transition est la matrice de connectivité du graphe normée à 1 ligne par ligne. Il y a tout de même une exception : il faut que tous les nœuds aient au moins un successeur, sinon, on ne peut pas transformer la matrice de connectivité en matrice markovienne. Dans ce cas, le bon sens fait ajouter un arc du nœud sur lui même dans le graphe, ce qui corrige le problème.

Nous avons donc montré qu'étudier le processus revenait à étudier une chaîne de Markov. Voyons maintenant ce que cette étude peut nous livrer.

1.2 Rappels sur les chaînes de Markov

Les théorèmes cités proviennent du cours de maîtrise de mathématiques de Dominique Bacry à l'université Paul Sabatier de Toulouse.

On considère que la matrice de Markov P représente une application linéaire de l'espace des probabilités associés aux N états du processus dans lui même, soit de R^N dans R^N. On peut rechercher les valeurs propres de cette matrice et les vecteurs associées. La somme des valeurs

des vecteurs propres sera toujours égale à 1, afin de faciliter leur utilisation éventuelle comme probabilité de présence sur chacun des nœuds, si cela a un sens.

1.2.1 Mesure Invariante

Parmi les valeurs propres intéressantes, la plus importante est 1. En effet, le vecteur propre associé à 1 correspond à une mesure invariante. C'est-à-dire que si la probabilité de présence en chaque noeud au temps t est la mesure invariante, la probabilité de présence au temps t+1 sera toujours la mesure invariante. C'est un point fixe du processus.

Le théorème de Perron-Frobenius affirme que si une matrice est markovienne alors elle possède au moins une mesure invariante.

Mais elle peut en posséder plusieurs. Un autre théorème affirme qu'elle en possède exactement autant que de classes de récurrence. Une classe de récurrence est un ensemble d'états tels qu'il existe une probabilité non nulle de « partir » d'un état arbitraire de l'ensemble vers un deuxième état arbitraire de l'ensemble puis de « revenir » au premier en un temps fini. Dans le cas des graphes, cela correspond à savoir s'il existe un chemin du nœud g au noeud g' et autre de g' à g. Notons aussi qu'il existe des états qui n'appartiennent à aucune classe de récurrence. Une fois qu'on y est passé, on ne peut plus y revenir : ce sont des états transitoires.

Un théorème intéressant est que les composantes des mesures invariantes sont nulles sur tous les états transitoires.

1.2.2 Périodicité

Les valeurs propres complexe de module 1 sont aussi importantes.

La périodicité d'une chaîne de Markov traduit le fait qu'on ne puisse revenir en un nœud g après y être passé qu'en un temps T multiple d'un entier strictement supérieur à 1. Par exemple, un graphe non orienté correspond à une chaîne de Markov de période 2.

Un théorème affirme qu'être de période k équivaut à posséder toutes les racines k-ièmes de l'unité comme valeur propre.

Enfin, les autres valeurs propres sont de module inférieur à 1. D'où les théorèmes de convergence suivants.

1.2.3 Convergence

On s'intéresse d'abord au cas où il n'y a qu'une classe de récurrence. Dans ce cas, un théorème nous affirme la convergence presque sûrement de la suite suivante :

$$Eq(1) : \forall m_0 : m_n = \frac{m_0}{n} \sum_{i=0}^{n-1} P^i \xrightarrow[n\to\infty]{} m$$

On a même dans le cas où le processus est apériodique :

$$Eq(2) : m_0 P^n \xrightarrow[n\to\infty]{} m$$

μ_0 : Distribution initiale
μ : Mesure Invariante
P : Matrice markovienne

Ce qui signifie que la probabilité de présence du jeton sur chaque nœud tend vers la mesure invariante.

Dans le cas où il y a plusieurs classes de récurrences, le problème est plus difficile. En effet la limite de (1) est une combinaison linéaire de toutes les mesures invariantes et dépend de l'état initial. Intuitivement, on comprend que le jeton est bloqué dans une classe de récurrence. Si le jeton est initialement sur un état récurrent, alors les probabilités de présence asymptotiques correspondent à la mesure invariante de cette classe. Mais si le jeton commence sur un état transitoire, il peut aboutir sur plusieurs classes de récurrences. Dans ce cas, le processus a plusieurs attracteurs. Cependant, on peut supposer que la probabilité limite associée est la combinaison linéaire des mesures invariantes de chaque classe de récurrence pondérée par la probabilité que le jeton parvienne dans cette classe de récurrence à partir de sa position initiale. Par la suite, on se contentera de prédire les différentes mesures invariantes correspondant à chaque classe de récurrence, et on ne calculera pas la probabilité de se retrouver dans chaque classe de récurrence.

1.2.4 Estimateurs

On considère dans cette partie des chaînes irréductibles.

Revenons maintenant au problème initial. En faisant la simulation pour m jetons en même temps, c'est comme si on réalisait plusieurs fois le processus en parallèle et de manière indépendante. Le fait qu'on ne sache pas faire la différence entre deux jetons n'a aucune importance, d'après la propriété de Markov forte.

Il y a donc deux manières de voir le problème : soit comme la réalisation de m processus markoviens, soit comme la réalisation d'un seul processus markovien dont l'espace d'état est le produit des m espaces d'état correspondant aux processus à un jeton. On note :

$$X_{i,n} = \left[d_{j,k} \right]_{j \in [1,N]} \in \{0,1\}^N , i \in [1,m]$$

le vecteur dont toutes les composantes sont nulles sauf la k^{ème,} où k est le nœud sur lequel le jeton i est présent à l'instant m, qui vaut 1. N est le nombre de nœuds du graphe. La seconde manière de voir le processus comme un unique processus markovien revient à considérer la variable $X_n = X_{1,n} \otimes X_{2,n} \otimes ... \otimes X_{m,n}$ et l'espace d'état associé.

On peut alors utiliser le théorème de la limite centrale des chaînes de Markov avec ces deux représentations, en tenant compte de l'indépendance des processus associé à chaque jeton:

$$Eq(3) : \forall i \in [1,m], \sqrt{n} \left(\frac{1}{n} \sum_{i=0}^{n-1} X_{i,n} - m \right) \xrightarrow[n \to \infty]{} N(0, \Sigma)$$

$$Eq(4) : \sqrt{n} \left(\frac{1}{n} \sum_{i=0}^{n-1} X_n - m \otimes m \otimes ... \otimes m \right) \xrightarrow[n \to \infty]{} N(0, \Sigma \otimes \Sigma \otimes ... \otimes \Sigma)$$

On remarque que $J_n = f(X_n) = 1/m . \sum X_{i,n}$. Ainsi on peut utiliser la delta–méthode avec l'équation 4 pour obtenir :

$$Eq(5) : \sqrt{n} \left(\frac{1}{n.m} \sum_{i=0}^{n-1} J_i - m \right) \xrightarrow[n \to \infty]{} N\left(0, \frac{\Sigma}{m}\right)$$

Ainsi, on a un estimateur non biaisé de m pour n grand. On voit qu'augmenter le nombre de jetons permet de réduire la variance de l'estimateur au même titre que n. La variance de l'estimateur est donc inversement proportionnelle au produit m.n. Toutefois, n et m ne sont pas

équivalent car le théorème de la limite centrale montre que, à n fixé, lorsque m tend vers l'infini, l'estimateur précédent tend vers $m_0 P^n$ et non vers la mesure invariante.

2 Simulation

2.1 Implémentation

La simulation du processus de diffusion nécessite l'implémentation de différentes fonctions : un générateur de graphes, un analyseur de chaînes de Markov, le simulateur proprement dit, et un script d'analyse des résultats. La programmation est fait dans un premier temps en langage Scilab afin de pouvoir utiliser les outils d'analyse markovienne et les outils de représentation graphique du logiciel. L'inconvénient majeur est la lenteur du programme. A titre indicatif, une simulation de 400 jetons pendant 15000 pas sur un graphe de 400 nœuds et environ 1600 liens prend environ 45mn sur un Athlon 1.2GHz.

2.1.1 Générateur de graphe

On réalise d'abord un générateur de graphes aléatoires (`randgm`). On lui donne le nombre de nœuds n souhaité et il remplit la matrice de connectivité du graphe en tirant selon une loi binomiale de paramètre fixé pour avoir le nombre moyen de liens voulu. Si il existe une ligne qui n'a aucun 1, on en place 1 sur cette ligne à l'intersection avec la diagonale. Cela revient à ajouter un lien d'un nœud vers lui-même si ce nœud n'a aucun lien sortant. Ce générateur ne garantit pas le nombre de liens, ce qui n'est pas important pour les études que nous voulons faire avec le simulateur.

Ensuite, comme on s'intéresse à un phénomène de diffusion, on réalise des générateurs de graphes permettant de représenter des maillages d'espace de dimension 1 et 2 (`mesh1d` et `mesh2d`). En dimension 1, chaque nœud est connecté à ses deux voisins directs, sauf les 2 qui sont aux extrémités du segment. En dimension deux, les nœuds sont disposés sur les sommets d'un pavage carré du plan. Ils ont donc 4 voisins.

2.1.2 Analyseur de chaînes de Markov

On souhaite connaître les valeurs propres, et les mesures invariantes du processus. Les fonctions `spec` (spectre de la matrice), `eigenmarkov` (mesures invariantes), et `classmarkov` (classes de récurrences et états transitoires) permettent de répondre simplement et rapidement à ces questions. A l'occasion, on teste aussi le fait que la matrice fournie est bien markovienne.

2.1.3 Simulateur

En plus de la simulation du processus, le simulateur permet d'échantillonner $\sum_{i=0}^{n-1} J_i$ à différents instants lors de la simulation pour étudier sa convergence. C'est la fonction `graphsimulator`.

2.1.4 Script d'analyse

Dans le cas où le graphe étudié est aléatoire, s'il n'est pas irréductible, il n'y a pas d'analyse à faire : on peut seulement vérifié que le processus a convergé vers une des classes de récurrence prévue lors de l'analyse. S'il est irréductible, on peut étudier la vitesse de convergence à l'aide de l'erreur moyenne commise par nœud et par jeton :

$$\frac{1}{N}\sqrt{\left(\frac{1}{m.n}\sum_{i=0}^{n-1} J_i - m\right).\left(\frac{1}{m.n}\sum_{i=0}^{n-1} J_i - m\right)^t}$$

L'équation (4) prévoit que ce terme d'erreur soit en $o\left(1/\sqrt{n}\right)$. Le script qui réalise cette fonction ainsi que la coordination générale est `rangnsimulation`.

Dans les cas des maillages d'un droite ou d'un plan, les scripts sont `mesh1dsimulation` et `mesh2dsimulation`. En plus du calcul du terme d'erreur, ils permettent des représentations graphiques des résultats, comme on le verra dans la partie suivante.

2.2 Résultats

2.2.1 Graphe aléatoire

Lorsque l'on a fixé le nombre N de nœuds du graphe, il existe un rapport E/N, où E est le nombre de liens au dessus duquel le graphe est très souvent irréductible et sans états transitoires. En dessous de ce rapport, la majorité des états sont transitoires et il existe une ou plusieurs classes de récurrence qui compte en général 1 élément (ceci vient du fait que l'on ajoute des 1 sur la diagonale lors de la génération du graphe).

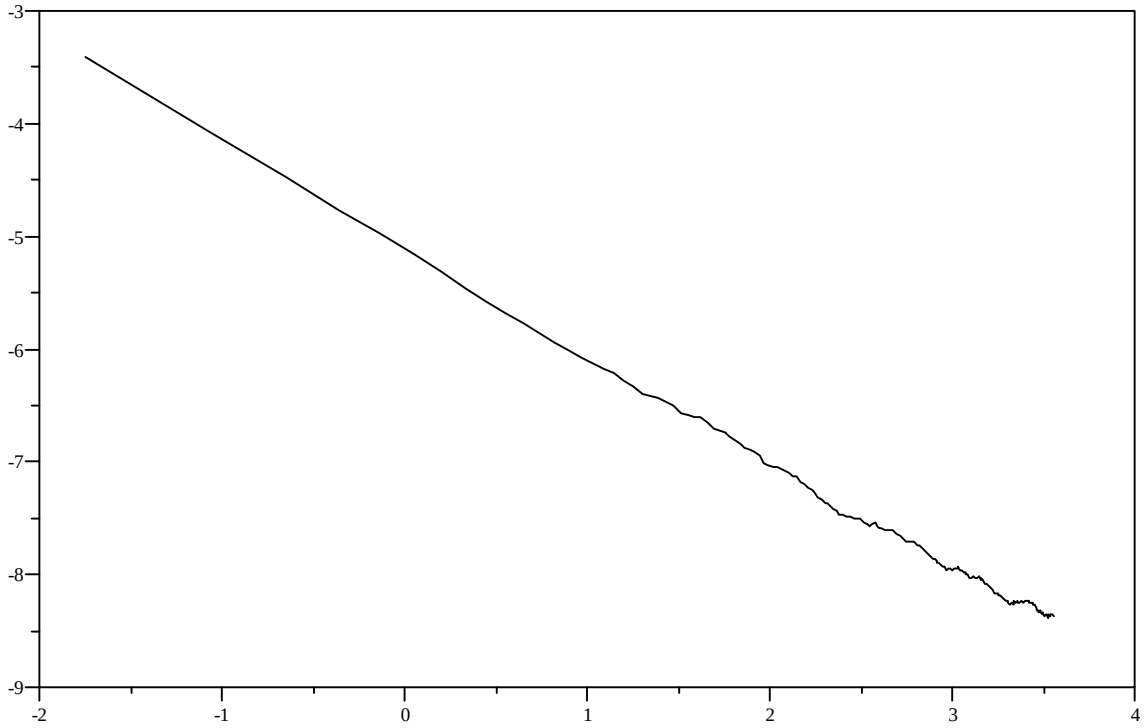
La simulation présentée est celle d'un graphe de 35 nœuds et 136 liens. L'état initial est d'un jeton par sommet. L'analyse donne une classe de récurrence comptant 34 états et un état transitoire. Les résultats se trouvent dans le tableau ci-dessous. Chaque ligne représente un sommet. La première colonne est J_n calculé par la simulation. La deuxième, J_n estimé par la mesure invariante et la troisième l'écart. On retrouve l'état transitoire à la première ligne de la deuxième colonne. On prévoit qu'il ne devrait pas passer de jetons sur ce sommet et c'est le cas puisque le seul qui y est passé, c'est celui qu'on y a mis au début.

```
res  =

!   26970.    27065.    95.  !
!   16683.    16799.    116. !
!   13890.    14078.    188. !
!   43806.    43576.  - 230. !
!   21197.    21156.  - 41.  !
!   14833.    14988.    155. !
!   11509.    11410.  - 99.  !
!   24148.    24374.    226. !
!   20878.    20965.    87.  !
!   15313.    15406.    93.  !
!   13434.    13532.    98.  !
!   16374.    16378.    4.   !
!   30335.    30536.    201. !
!   12959.    12855.  - 104. !
!   2362.     2399.     37.  !
!   41985.    42211.    226. !
!   20004.    20236.    232. !
!   1.        0.      - 1.   !
!   45498.    45717.    219. !
!   6694.     6732.     38.  !
!   3937.     3951.     14.  !
!   23880.    24153.    273. !
!   51241.    51679.    438. !
!   25907.    26042.    135. !
!   21254.    21225.  - 29.  !
!   15013.    15192.    179. !
!   12294.    12548.    254. !
!   10382.    10509.    127. !
!   14377.    14472.    95.  !
!   7551.     7678.     127. !
!   39110.    39211.    101. !
!   20705.    20828.    123. !
!   11678.    11806.    128. !
!   14218.    14345.    127. !
!   26080.    25932.  - 148. !
```

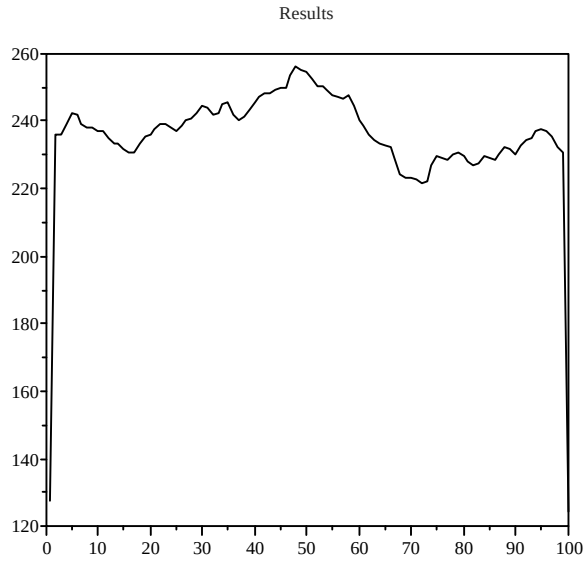
On trace le graphe log-log de l'erreur moyenne en fonction de n. On s'attend à trouver une fonction dont la dérivée est bornée supérieurement passé un certain temps par -1/2. C'est bien ce qu'on trouve. On trouve même une droite de pente -0,92. C'est un indice pour dire que, pour ce graphe avec ces conditions initiales, la convergence empirique est en $0\left(n^{-0.92}\right)$.

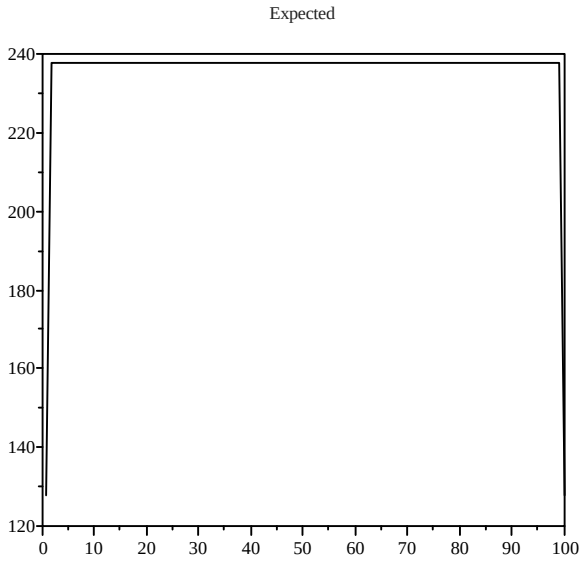
Graphe log-log de l'erreur moyenne



2.2.2 Maillage d'une droite

On réalise une simulation de longueur 30000 sur 100 nœuds. Au départ, les 100 jetons sont sur le $50^{ème}$ nœud. L'analyse donne -1 comme valeur propre, ce qui est normal vu que le graphe est non orienté, sans lien d'un nœud vers lui-même, donc de période 2. L'analyse donne aussi des valeurs propres très proches de 1, ce qui laisse supposer une convergence très lente vers l'équilibre. C'est bien ce que l'on observe. Le résultat est mauvais. Ceci rappelle que la diffusion est un phénomène physique pour lesquelles les échelles de distances sont inversement proportionnelles à la racine du temps de diffusion. C'est-à-dire que pour diffuser 2 fois plus loin, il faut 4 fois plus de temps.





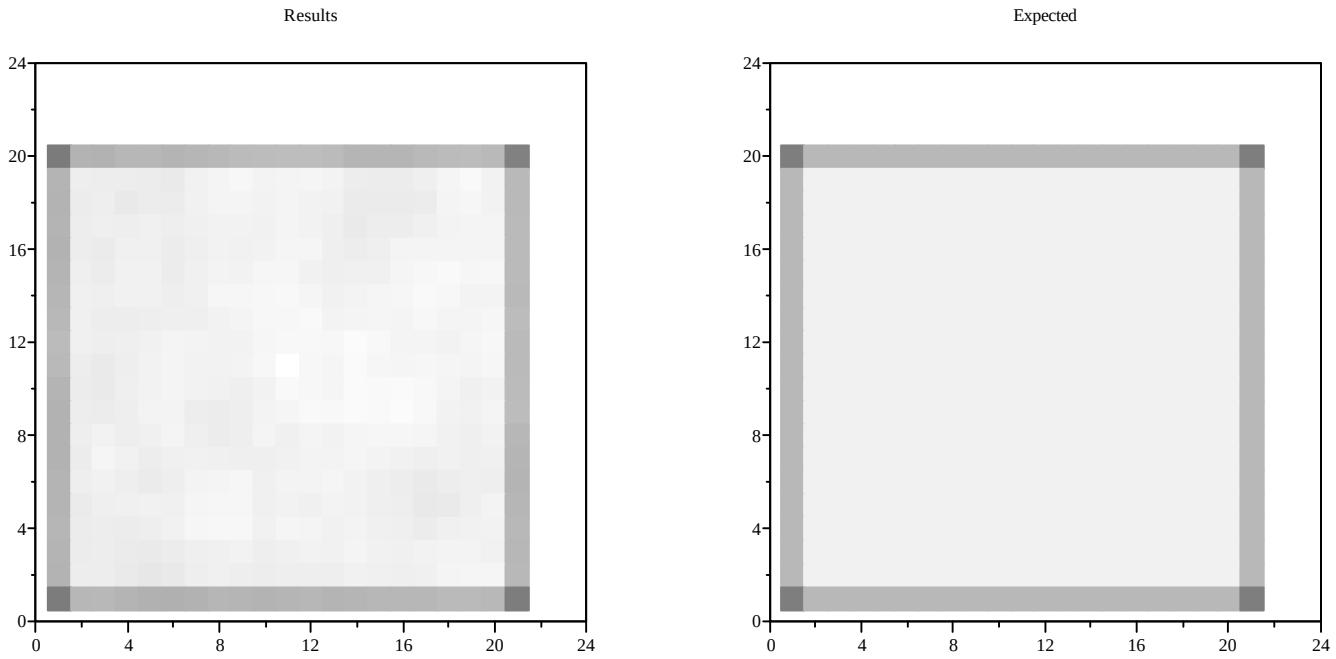
2.2.3 Maillage d'un plan

On réalise une simulation de longueur 15000 sur un maillage 21x20. Au départ les 420 jetons sont sur le nœud (11,11).

Grâce à la fonction Matplot, on représente le nombre de jetons sur chaque sommet par un niveau de gris proportionnel au nombre de jetons présents (noir : pas de jeton, blanc : beaucoup de jetons). Pour garder un contraste fort pour chaque échantillon, la couleur noire est affectée à la case qui compte le moins de jetons et le blanc à celle qui en compte le plus.

La première figure compare les résultats aux résultats attendus. On remarque que la distribution est homogène, aux effets de bords près. En effet, les nœuds qui se trouvent au bord ont moins de voisins, ce qui les rend plus difficilement atteignables, d'où une couleur plus sombre. Ce n'est pas ce qui se passe dans la réalité, où les parois du récipient utilisé ont tendance à adsorber les molécules qui diffusent, ce qui augmente la probabilité de présence en ces points (cf ménisque dans une éprouvette).

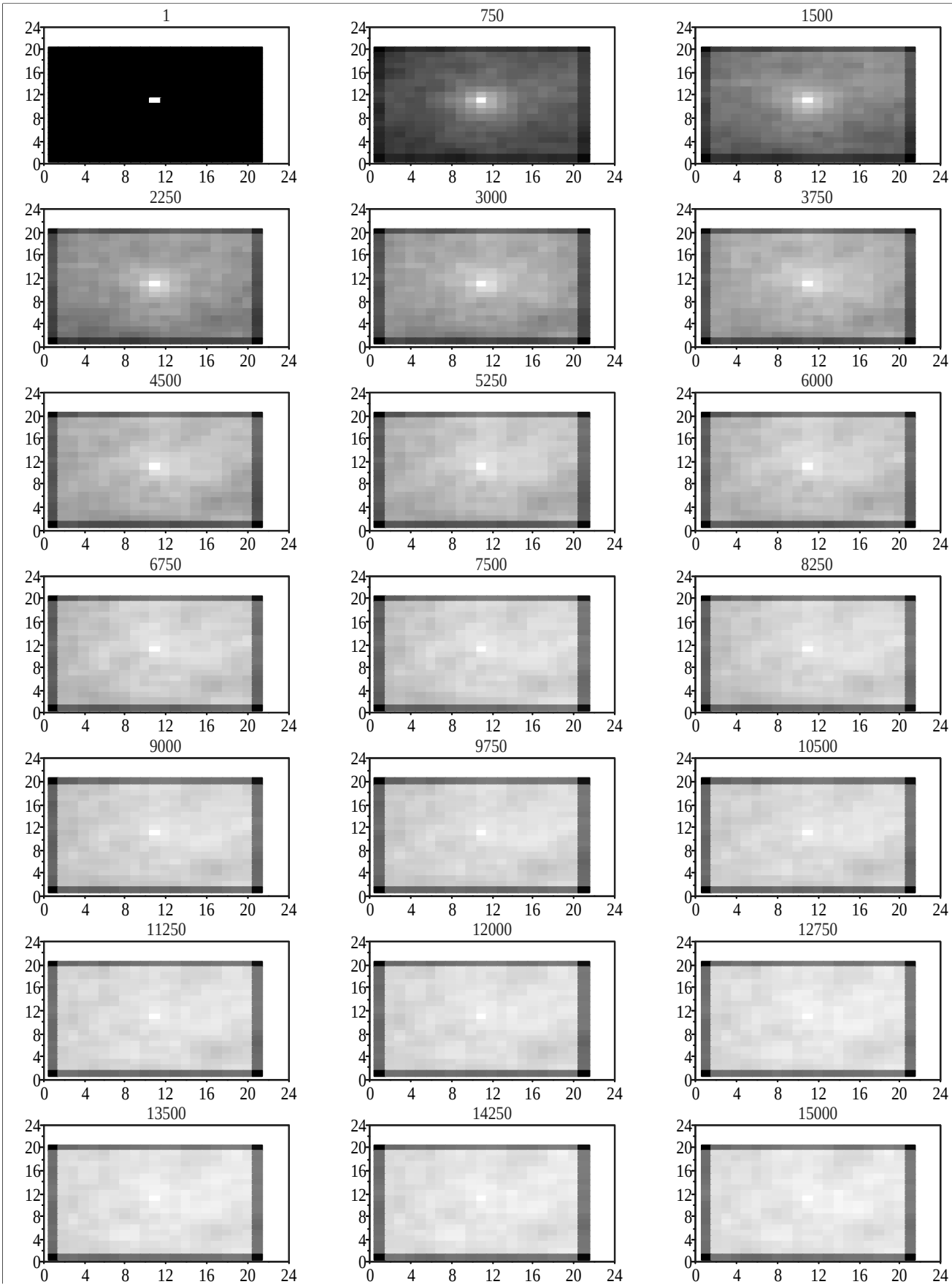
La seconde représente les échantillons de J_n aux instants indiqués au dessus des vignettes.



Conclusion

La réalisation de ce simulateur permet d'illustrer la théorie sous-jacente. La simulation en elle-même est intéressante pour étudier les échelles de temps et la vitesse de convergence. La programmation en Scilab ne permet pas une étude pleinement satisfaisante, car la simulation est trop lente pour pouvoir réaliser un grand nombre de processus à partir de la même matrice de transition. Ceci permettrait d'étudier la variance de l'estimateur et de vérifier qu'il converge bien en loi vers une gaussienne.

D'autre part, un autre axe de développement est l'étude mathématique puis « expérimentale » de l'adéquation de ce modèle à l'équation de la diffusion en milieu continu.



Echantillons de $\sum J_i$ pour un pavage du plan de taille 20x21