Agrégation externe de mathématiques, session 2019
Épreuve de modélisation, option A : Probabilités et statistiques

(*Public2018-A3*)

Résumé : Étant donné un ensemble aléatoire de points dans l'espace, on cherche une méthode pour estimer l'agrégation de ces points ou au contraire leur dispersion.

Mots clefs : Loi uniforme, loi binomiale, loi de Poisson, estimation, tests.

► *Il est rappelé que le jury n'exige pas une compréhension exhaustive du texte. La présentation, bien que totalement libre, doit être organisée et le jury apprécie qu'un plan soit annoncé en préliminaire. L'exposé doit être construit en évitant la paraphrase et mettant en lumière les connaissances, à partir des éléments du texte. Il doit contenir des illustrations informatiques réalisées sur ordinateur, ou, à défaut, des propositions de telles illustrations. Des pistes de réflexion, indicatives et largement indépendantes les unes des autres, vous sont proposées en fin de texte.*

Ce texte vous est fourni avec un fichier `Sinistres.txt` *permettant d'illustrer sur des données réelles certains des résultats présentés. Des instructions pour lire ce fichier à l'aide des logiciels disponibles sont données en dernière page du texte.*

1. Présentation du modèle

Si l'on dispose d'un ensemble $\mathbf{X}$ de points répartis aléatoirement dans l'espace $\mathbb{R}^d$, on peut se demander si cet ensemble présente une propension au regroupement (ou agrégation) des points en sous-groupes relativement éloignés, ou si au contraire ils ont tendance à s'écarter les uns des autres. Le cas autour duquel pivotent ces deux comportements serait un cas dit d'*aléa spatial complet*, que nous présentons dans la section 2.

Ce type de problématique apparaît en dimension $d = 1$ pour l'étude de phénomènes temporels, par exemple pour une compagnie d'assurance il est important de savoir si les sinistres ont une répartition temporelle complètement aléatoire, plutôt agrégée ou dispersée, afin de savoir comment au mieux gérer ses actifs pour disposer de liquidités suffisantes pour les remboursements. Le fichier `Sinistres.txt` contient les dates de sinistres déclarés auprès d'une compagnie d'assurance au cours de 10 années.

En dimension supérieure, par exemple en dimension 2, on peut se poser des questions de positionnement géographique de zones d'activité industrielle ou commerciale au niveau d'un territoire national ou urbain, ou alors des questions sur les positions des troncs d'arbres dans une forêt.

Dans la section 3 nous proposons une mesure de la dispersion/agrégation, pour laquelle un estimateur sera introduit, ainsi qu'une procédure de test pour discriminer ces comportements.

Tout le modèle sera étudié dans un sous-ensemble W de l'espace $\mathbb{R}^d$, que l'on supposera compact d'intérieur non vide, par exemple le cube unité $Q_d = [0,1]^d$. On notera $|W|$ le volume de W, c'est-à-dire sa mesure de Lebesgue, et $\mathrm{int}(W)$ son intérieur.

2. Un modèle pour l'aléa spatial complet

2.1. *Lois uniformes et processus binomial*

La première représentation d'un ensemble complètement aléatoire de points dans W que l'on peut proposer est la suivante : on se donne n points $U_1, \ldots, U_n$ de loi uniforme sur W, indépendants :

$$\forall B \subset W \text{ borélien}, \mathbf{P}(U_i \in B) = \frac{|B|}{|W|}, \tag{1}$$

et l'on considère l'ensemble $\mathbf{X}_n = \{U_1, \ldots, U_n\}$. Cet ensemble vérifie la propriété élémentaire suivante :

Proposition 1. *Pour tout borélien $B \subset W$, le nombre de points de $\mathbf{X}_n$ appartenant à B :*

$$\mathrm{card}(\mathbf{X}_n \cap B) = \sum_{i=1}^{n} \mathbf{1}_{U_i \in B} \tag{2}$$

suit une loi binomiale de paramètres n et $|B|/|W|$.

On appelle alors l'ensemble aléatoire $\mathbf{X}_n$ le *processus binomial* sur W.

Dans toute la suite du texte, nous noterons $N_{\mathbf{Y}}(B) = \mathrm{card}(\mathbf{Y} \cap B)$ le nombre de points d'un ensemble fini $\mathbf{Y}$ qui appartiennent au borélien B.

Dégageons quelques propriétés de ce processus binomial. Pour tout $x \in \mathbb{R}^d$, notons $B(x,\eta)$ la boule ouverte de centre x et de rayon $\eta > 0$.

Proposition 2. *La densité du processus définie en tout point $x \in \mathrm{int}(W)$ par*

$$\lambda(x) = \lim_{\eta \to 0} \frac{\mathbf{E}\left[N_{\mathbf{X}_n}(B(x,\eta))\right]}{|B(x,\eta)|} \tag{3}$$

est constante sur W, égale à $n/|W|$.

Dans la pratique, on considère que la mesure de la fenêtre W est grande, et du même ordre de grandeur que n, de telle sorte que la densité est une constante raisonnable, notée simplement $\lambda \in \mathbb{R}_+^*$.

Un autre résultat intéressant de ce processus binomial est donné par le nombre moyen de voisins d'un point quelconque du processus. La probabilité qu'il y ait un point du processus $\mathbf{X}_n$ en un point donné x est nulle par construction, mais on admet qu'il est possible de donner un sens à l'affirmation suivante : l'espérance conditionnelle du nombre de points de $\mathbf{X}_n$ à distance inférieure à $\eta > 0$ de x sachant qu'il y a un point du processus $\mathbf{X}_n$ en x est égale à $(n-1)|B(x,\eta)|/|W| \simeq \lambda|B(x,\eta)|$, pour tout x tel que la distance de x au bord de W soit strictement supérieure à η.

Pour expliquer cela on peut utiliser les faits suivants :

— presque sûrement tous les points $U_1, \dots, U_n$ sont distincts deux à deux;
— la loi conditionnelle du vecteur $(U_2, \dots, U_n)$ sachant U_1 est la loi uniforme sur W^{n-1}.
Cette dernière propriété nous permettra d'introduire un indicateur de dispersion du processus dans la partie 3.

2.2. *Indépendance et poissonisation du processus binomial*

Hélas, le processus binomial ne satisfait pas totalement à la notion d'aléa spatial complet en ce sens que les quantités $N_{\mathbf{X}_n}(B_1)$ et $N_{\mathbf{X}_n}(B_2)$ ne sont pas des variables aléatoires indépendantes lorsque B_1 et B_2 sont deux sous-ensembles disjoints.

Pour remédier à ce problème, il suffit de procéder de la façon suivante :

Définition 1. *Soit λ un réel strictement positif, N une variable aléatoire de loi de Poisson de paramètre $\lambda|W|$, et $(U_n)_{n\geqslant 1}$ une suite de variables aléatoires indépendantes identiquement distribuées de loi uniforme sur W, indépendante de N. On appelle* processus ponctuel de Poisson de densité λ sur W *l'ensemble* $\mathbf{X}$ *défini par*

$$\mathbf{X} = \{U_1, \dots, U_N\} \text{ si } N \neq 0 \text{ et } \mathbf{X} = \emptyset \text{ si } N = 0. \tag{4}$$

Conditionnellement à $\{N = n\}$ $(n \geqslant 1)$, $\mathbf{X}$ est alors égal au processus binomial $\mathbf{X}_n$.

Ce processus possède la propriété fondamentale suivante :

Proposition 3. *Le processus ponctuel de Poisson* $\mathbf{X}$ *de densité λ vérifie les deux propriétés suivantes :*
— *pour tout borélien borné B de W, la variable aléatoire $N_{\mathbf{X}}(B)$ suit une loi de Poisson de paramètre $\lambda|B|$;*
— *pour tous boréliens bornés B_1 et B_2 disjoints, les variables aléatoires $N_{\mathbf{X}}(B_1)$ et $N_{\mathbf{X}}(B_2)$ sont indépendantes.*

Le premier point est une conséquence de la proposition 1 du paragraphe précédent, en utilisant le conditionnement par les événements $\{N = n\}$, $n \in \mathbb{N}$, le second point se traite de même par conditionnement par ces mêmes événements, en déterminant les probabilités conditionnelles $\mathbf{P}(N_{\mathbf{X}}(B_1) = k, N_{\mathbf{X}}(B_2) = \ell | N = n)$, avec $k, \ell, n \in \mathbb{N}$.

2.3. *Densité et dispersion du processus* $\mathbf{X}$

Si l'on reprend la proposition 2 dans le cas du processus $\mathbf{X}$, on peut vérifier facilement que pour tout $x \in \text{int}(W)$,

$$\lim_{\eta \to 0} \frac{\mathbf{E}\left[N_{\mathbf{X}}(B(x,\eta))\right]}{|B(x,\eta)|} = \lambda, \tag{5}$$

et donc la quantité λ correspond à ce que l'on avait appelé *densité* du processus dans la partie 2.1. Il reste à étudier la dispersion de ce processus. Pour cela fixons $\eta_0 > 0$ et notons $W_0 = \{x \in W \text{ tel que } B(x,\eta_0) \subset W\}$. Nous allons étudier cette dispersion jusqu'à l'échelle η_0.

Fixant $\eta < \eta_0$, nous voulons compter le nombre moyen de voisins d'un point du processus $\mathbf{X}$, *en tenant compte de l'aléa de ce point,* on se propose donc de calculer une moyenne sur

ces points du nombre de voisins, pour cela on est amené à calculer la moyenne empirique suivante :

$$A(\eta) = \frac{1}{N_{\mathbf{X}}(W_0)} \sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1). \tag{6}$$

Le calcul de l'espérance de cette quantité est compliqué, par contre on peut calculer séparément les espérances du numérateur et du dénominateur. En effet, l'espérance de $N_{\mathbf{X}}(W_0)$ est égale à $\lambda |W_0|$. Celle du numérateur est un peu plus compliquée à calculer :

Proposition 4. *L'espérance de* $\sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1)$ *est donnée par*

$$\mathbf{E}\left[\sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1) \right] = \lambda^2 |W_0| \times |B(0,\eta)|. \tag{7}$$

La démonstration de cette proposition se fait par conditionnement selon la famille complète d'événements $\{N = n\}$. En effet, on peut écrire cette espérance de la façon suivante :

$$\begin{aligned} \mathbf{E}\left[\sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1) \right] &= \sum_{n=0}^{+\infty} \mathbf{E}\left[\sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1) \Big| N = n \right] \mathbf{P}(N = n), \\ &= \sum_{n=0}^{+\infty} \mathbf{E}\left[\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbf{1}_{U_i \in W_0} \times \mathbf{1}_{U_j \in B(U_i,\eta)} \right] \frac{(\lambda |W|)^n}{n!} e^{-\lambda |W|}, \\ &= \sum_{n=0}^{+\infty} n(n-1) \int_{W_0} \left(\int_{B(u,\eta) \cap W} \frac{dv}{|W|} \right) \frac{du}{|W|} \frac{(\lambda |W|)^n}{n!} e^{-\lambda |W|}, \end{aligned} \tag{8}$$

cette somme se simplifie alors pour donner le résultat souhaité.

Cette proposition nous permet d'introduire la *fonction K de Ripley* qui mesure (à un facteur multiplicatif près) le nombre moyen de voisins à distance η des points d'un processus ponctuel de densité λ sur W :

Définition 2. *Le quotient*

$$K_{\mathbf{X}}(\eta) = \frac{1}{\lambda^2 |W_0|} \mathbf{E}\left[\sum_{x \in \mathbf{X} \cap W_0} (N_{\mathbf{X}}(B(x,\eta)) - 1) \right] \tag{9}$$

est appelé fonction K de Ripley du processus $\mathbf{X}$.

Ainsi pour un processus ponctuel de Poisson on a la propriété suivante :

Proposition 5. *La fonction K de Ripley d'un processus ponctuel de Poisson de densité* λ *est égale à* $|B(0,\eta)|$ *pour* $\eta < \eta_0$.

3. Estimation de la fonction K

Cette fonction K permet de quantifier la dispersion du processus. En effet, si l'on a un processus de points $\mathbf{Y}$ de densité λ, on peut toujours définir le quotient $K_{\mathbf{Y}}(\eta)$, et si les points ont tendance à être agglomérés on peut l'interpréter sur la fonction $K_{\mathbf{Y}}$ qui sera alors plus

grande à l'échelle de cette agglomération que la fonction correspondante du processus de Poisson, tandis que la dispersion aura tendance à diminuer la valeur de cette fonction dans ces petites échelles et à l'augmenter aux plus grandes échelles.

Pour estimer cette fonction, une façon naturelle consiste à approcher l'espérance de cette somme par la somme sur les points du processus : si l'on note $\mathbf{Y} \cap W_0 = \{V_1, \dots, V_{N_{\mathbf{Y}}(W_0)}\}$, cela revient à introduire l'estimateur

$$\hat{K}_{\mathbf{Y}}(\eta) = \frac{1}{\lambda^2 |W_0|} \sum_{i=1}^{N_{\mathbf{Y}}(W_0)} (N_{\mathbf{Y}}(W \cap B(V_i, \eta)) - 1) = \frac{1}{\lambda^2 |W_0|} \sum_{i=1}^{N_{\mathbf{Y}}(W_0)} \sum_{x \in \mathbf{Y},\ x \neq V_i} \mathbf{1}_{d(V_i,x) \leqslant \eta}. \tag{10}$$

Lorsque l'on propose cet estimateur, deux problèmes se présentent :

— *a priori* λ est aussi une inconnue ;

— les points voisins à distance η peuvent être en dehors de la fenêtre W_0 dans la définition, ce qui rend l'écriture de la somme double compliquée puisque deux types de points x sont possibles selon qu'ils sont dans W_0 ou non.

Le premier point est corrigé en approchant $\lambda|W|$ par le nombre effectif de points $N_{\mathbf{Y}}(W)$ du processus (attention cependant au cas où ce nombre est nul), le second nécessite une étude plus fine, en général cela reviendra à ne considérer dans la seconde somme que les points $x = V_j$, $j \neq i$ de W_0 et à pondérer l'estimateur par un poids $w_{i,j}$:

$$\tilde{K}_{\mathbf{Y}}(r) = \frac{|W_0|}{N_{\mathbf{Y}}(W)^2} \sum_{i \neq j} w_{i,j} \mathbf{1}_{d(V_i,V_j) \leqslant \eta}, \text{ si } N_{\mathbf{Y}}(W) > 0, \text{ 0 sinon,} \tag{11}$$

où $w_{i,j}$ est la proportion de la mesure de surface $(d-1)$-dimensionnelle de la sphère de centre V_i et de rayon $\|V_i - V_j\|$ incluse dans W_0 par rapport à la mesure totale de cette sphère. (Par exemple en dimension 1 c'est la moitié du nombre d'extrémités du segment centré en V_i et de demi-longueur $|V_i - V_j|$ appartenant à la fenêtre W_0).

4. Tests d'agrégation et de dispersion

On souhaite, connaissant la répartition des points de $\mathbf{Y}$ (par exemple à partir des points du fichier `Sinistres.txt`) construire un test permettant de préciser le comportement à l'agrégation ou à la dispersion de notre processus, par exemple en testant l'hypothèse H_0 (le processus $\mathbf{Y}$ est un processus de Poisson de densité λ) contre les hypothèses

— H_1^a : agrégation du processus à la distance $\eta > 0$;

— H_1^d : dispersion du processus à la distance $\eta > 0$.

La détermination de la zone de rejet pour le test (unilatéral) se fait à l'aide de l'inégalité de Tchebychev. En effet sous l'hypothèse H_0, pour la distance $\eta > 0$, l'estimateur $\tilde{K}_{\mathbf{Y}}(\eta)$ admettant une espérance $m(\eta)$ et une variance $\sigma^2(\eta)$ vérifie :

$$\mathbf{P}\left(\tilde{K}_{\mathbf{Y}}(\eta) - m(\eta) \geqslant \sigma(\eta)/\sqrt{\alpha}\right) \leqslant \alpha. \tag{12}$$

Comme la quantité théorique $\sigma^2(\eta)$ n'est pas accessible théoriquement, il faut estimer cette variance sous H_0, ce qui peut se faire en réalisant des simulations du processus de Poisson : on simule m copies indépendantes du processus $\mathbf{X}$ de Poisson de densité λ dans W, qui fournissent m copies de $\tilde{K}_{\mathbf{X}}(\eta)$, notées $\tilde{K}_1(\eta), \dots, \tilde{K}_m(\eta)$. L'estimateur usuel de la variance $\sigma^2(\eta)$ est

alors

$$S_m(\eta) = \frac{1}{m-1} \sum_{j=1}^{m} (\tilde{K}_j(\eta) - \overline{K}_m(\eta))^2, \tag{13}$$

où $\overline{K}_m(\eta) = (\tilde{K}_1(\eta) + \cdots + \tilde{K}_m(\eta))/m$.

La convergence de ces estimateurs de la moyenne et de la variance se traduit dans l'inégalité de Tchebychev, sous l'hypothèse H_0, dans le résultat limite suivant :

$$\limsup_{m \to +\infty} \mathbf{P}\left(\tilde{K}_{\mathbf{Y}}(\eta) - \overline{K}_m(\eta) \geqslant \sqrt{S_m(\eta)}/\sqrt{\alpha}\right) \leqslant \alpha. \tag{14}$$

On obtient donc un test (asymptotique en m) de niveau au moins $1-\alpha$ en rejetant l'hypothèse H_0 contre l'hypothèse H_1^a pour la distance η dès que $\tilde{K}_{\mathbf{Y}}(\eta) \geqslant \overline{K}_m(\eta) + \sqrt{S_m(\eta)}/\sqrt{\alpha}$.

Suggestions et pistes de réflexion

▶ *Les pistes de réflexion suivantes ne sont qu'indicatives et il n'est pas obligatoire de les suivre. Vous pouvez choisir d'étudier, ou non, certains des points proposés, de façon plus ou moins approfondie, mais aussi toute autre question à votre initiative. Vos investigations comporteront une partie traitée sur ordinateur et, si possible, des représentations graphiques de vos résultats. À défaut, si vos illustrations informatiques n'ont pas abouti, il est conseillé d'expliquer ce que vous auriez souhaité mettre en œuvre.*

— *Aspects mathématiques.*
  — On pourra démontrer la proposition 2, ou compléter les démontrations des proposition 3 ou 4.
  — On pourra construire le test de l'hypothèse H_0 contre H_1^d.

— *Programmes et simulations.*
  — On pourra donner la représentation graphique de la fonction $K_{\mathbf{X}}$ pour les points du processus $\mathbf{X}$ donnés dans le fichier `Sinistres.txt` pour $\eta \in [0.01, 1]$ et la fenêtre $W_0 = [1, 9]$.
  — On pourra proposer des tests statistiques pour vérifier les deux points de la proposition 3 à partir de plusieurs réalisations du processus de Poisson dans le carré $[0,1]^2$ (on choisira des boréliens B_1 et B_2 simples).

— *Modélisation.*
  — On pourra donner des exemples de phénomènes dans lesquels on peut s'attendre à obtenir des processus de Poisson, des processus agrégés ou des processus dispersés. Quels types de modèles probabilistes peut-on mettre sur ces phénomènes ?

Lecture des fichiers de données

Indications pour la session 2019

Ce document est agrafé au présent texte car ce dernier est associé à un jeu de données réelles `Sinistres.txt`, qui vous est fourni sous format texte.

La première ligne de ce fichier de données est une ligne de titre, contenant du texte indiquant ce qu'on trouve dans chacune des colonnes ; les lignes suivantes contiennent des données numériques.

Les instructions suivantes expliquent comment charger ce fichier sous différents logiciels de sorte que les données se retrouvent dans une matrice `A`.

Commencez par déplacer `Sinistres.txt` dans votre répertoire personnel :

— ouvrir le répertoire `Données` se trouvant sur le bureau en double-cliquant sur la dernière icône (bleue) de la colonne de gauche

— choisir le fichier `Sinistres.txt` et le déplacer à la souris (ou le copier) dans le `Répertoire personnel` (première icône du bureau de la colonne de gauche).

Lecture sous Scilab.

```
[A,text] = fscanfMat("Sinistres.txt")
```

La matrice `A` contient les donnés numériques, `text` contient la ligne de texte.

Lecture sous R. La commande

```
B<-read.table("Sinistres.txt", header = TRUE)
```

crée une variable `B` qui n'est pas exactement une matrice numérique, mais plutôt une structure. Les lignes suivantes transforment cette structure en matrice numérique si besoin est :

```
D<-dim(B); n<-D[1]; p<-D[2];
A<-matrix(0,n,p);
for (i in 1:p) A[,i]<-B[,i];
```

Lecture sous Octave. Commencer par détruire la première ligne du fichier `Sinistres.txt`, puis exécuter :

```
A = load("Sinistres.txt")
```

Lecture sous Python. On utilise la fonction `loadtxt` de la bibliothèque `numpy` :

```
from numpy import loadtxt
A = loadtxt("Sinistres.txt", skiprows=1)
```

Le résultat obtenu est de type `array`. L'option `skiprows=1` sert à ignorer la ligne de commentaire du fichier.

Description du fichier `Sinistres.txt` **:** Ce fichier a 590 lignes (ligne de titre non comptée) et 1 colonne. Chaque ligne correspond à la date de survenue d'un sinistre.