OUTILS DE MODÉLISATION MATHÉMATIQUE.

QUELQUES ILLUSTRATIONS

Séminaire IREM-DMI du 29 février 2008

Jean VAILLANT
UAG

MODÈLES ET MODÉLISATION

Définition 1.
Un *modèle* est une représentation symbolique de certains aspects d'un objet ou d'un phénomène du monde réel.
Le modèle donne donc une vision simplifiée ou idéalisée de la réalité.

Définition 2.
La *modélisation* est la démarche qui permet d'élaborer un modèle.

La modélisation prend en compte :

- l'objet et/ou le phénomène à représenter,
- les objectifs, c'est-à-dire l'utilisation que l'on souhaite faire du modèle,
- les données (relatives aux variables) et connaissances (relations entre variables) disponibles ou accessibles par l'expérience ou par l'observation.

La *modélisation mathématique* consiste à proposer une représentation dans le formalisme mathématique d'un objet ou d'un phénomène du monde réel.

Applications très variées : physique, chimie, épidémiologie, écologie, géologie, économie, biologie, sismologie ...

Un modèle a une fonction de représentation qui doit rester un outil d'étude de la réalité : Décrire, Expliquer, Interpréter, Prévoir. Il sert parfois à tester l'écart entre *l'état de la nature* et les suppositions ayant servi à élaborer ce modèle.

Définition 3.
Un *système* est un ensemble d'éléments qui se coordonnent pour concourir à un résultat ou de manière à former un ensemble.

Définition 4.
Un système est dit *isolé* s'il n'entretient aucune relation avec le monde extérieur (il s'agit généralement d'idéalisations de systèmes expérimentaux).

Définition 5.
Un système est dit *clos* s'il n'échange que de l'énergie avec le monde extérieur.

Définition 6.
Un système est dit *ouvert* s'il échange matière et énergie avec son *environnement*.

APPROCHES DÉTERMINISTE, STOCHASTIQUE

Les modèles mathématiques sont de nature

soit *déterministe*
soit *stochastique*.

Dans le premier cas, c'est une valeur bien précise qui est associée aux variables et aux paramètres, alors que dans le second ce sont des distributions de probabilité qui sont associées à ces grandeurs :

- Un modèle est *déterministe* s'il ne fait pas appel au calcul de probabilités.
- Un modèle est *stochastique* s'il fait appel au calcul de probabilités.

Par exemple, pour étudier l'évolution au cours du temps d'un système :

	Approche déterministe	**Approche stochastique**
Temps discret ($t \in \mathbb{N}$)	Equations récurrentes (ex. $u_{t+1} = f(u_t)$)	Processus en temps discrets (ex. processus (X_t) de Markov)
Temps continu ($t \in \mathbb{R}^+$)	Equations différentielles (EDO, EDP,..)	- Processus en temps continu (ex. processus gaussien), - Equations différentielles stochastiques

ETAPES COURANTES D'UNE MODÉLISATION

Etape 1.
Elaborer une vue synthétique de la situation, des connaissances a priori, des données disponibles ou accessibles.
Etape 2.
Caractériser et analyser le phénomène ou l'objet que l'on souhaite représenter.

Pour cela, il est nécessaire de :

1) spécifier si le système est isolé, clos ou fermé.
2) classer les variables agissant sur le système
   - Variables d'état, décrivant l'état du système (taille ou densité d'une population, concentration d'un produit,...)
   - Variables d'action qui modifient l'état du système par action externe (par exemple température par chauffage ou refroidissement, injection ou prélèvement d'un produit, immigration ou émigration contrôlées,...)
   - Variables d'observation ou observables qui permettent d'avoir des informations sur l'état du système (ces observables peuvent être directement des variables d'état).
3) Définir les relations entre les variables.

Etape 3.

Choisir et élaborer un modèle.

Exemple :

Croissance d'une population limitée en ressources, en milieu isolé (pas d'immigration ou d'émigration)

$X(t)$ = biomasse à la date t,
$s(t)$ = ressources à la date t.

La vitesse de croissance peut s'exprimer par une équation différentielle ordinaire (EDO)

$$\frac{dx}{dt} = f(x, s).$$

Si le rendement de la croissance est supposé constant, on a

$$x(t) - x(0) = R(s(0) - s(t)) \qquad \text{où} \;\; R \in \mathbb{R}_+$$

donc

$$\frac{dx}{dt} = -R\frac{ds}{dt}.$$

Le phénomène peut donc être représenté par le système de deux EDO

$$\begin{cases} \dfrac{dx}{dt} = f(x,s) \\ \dfrac{ds}{dt} = -\dfrac{1}{R} f(x,s) \end{cases}$$

avec les valeurs initiales $x(0) = x_0$ et $s(0) = s_0$.

La vitesse de croissance doit vérifier les contraintes :

a) $f(.,.)$ est une fonction positive dans le premier quadrant ($x \geq 0, s \geq 0$)

b) $f(0,s) = 0$ (il n'y a pas de génération spontanée !). Le modèle le plus simple concernant cette contrainte est $\dfrac{dx}{dt} = kxg(s)$ avec $k > 0$ et g une fonction réelle.

c) $f(x,0) = 0$ (s'il n'y a pas de ressources alors la croissance est nulle).

Le modèle le plus simple vérifiant ces trois contraintes est

$$\begin{aligned} \frac{dx}{dt} &= kxs \\ \frac{ds}{dt} &= -\frac{1}{R}kxs \end{aligned} \tag{1}$$

Comme $s = s_0 - \frac{1}{R}(x - x_0)$, on peut ramener ce système de deux EDO à une seule EDO :

$$\frac{dx}{dt} = kx(s_0 - \frac{1}{R}(x - x_0)) = \frac{kx}{R}(Rs_0 + x_0 - x)$$

donc

$$\frac{dx}{dt} = rx(1 - \frac{x}{\rho}) \tag{2}$$

avec $\rho = Rs_0 + x_0$ et $r = \frac{k\rho}{R}$.

<u>Remarque (**Principe de parcimonie**)</u> :
il s'agit d'élaborer le modèle le plus simple possible tout en étant le plus proche possible de la réalité.

Etape 4.
Vérifier les propriétés qualitatives du modèle.

Elles doivent être en accord avec ce que l'expérience nous permet d'observer, de mesurer, ou ce que les connaissances acquises nous disent.

Dans l'exemple de croissance de population, on aura à étudier la courbe représentative de la fonction solution de l'EDO (2).

- $x = \rho$ et $x = 0$ sont des points fixes
- Si $x \in]0, \rho[$, $\dfrac{dx}{dt} > 0$ et $x(t)$ croît avec t.
- Si $x > \rho$, $\dfrac{dx}{dt} < 0$ et $x(t)$ décroît avec t.
- Etude des points d'inflexion :

$$\frac{dx^2}{dt^2} = r(1 - \frac{2x}{\rho})\frac{dx}{dt}$$

$$\frac{dx^2}{dt^2} \quad \text{s'annule pour} \quad \frac{dx}{dt} = 0 \text{ ou } x = \frac{\rho}{2}.$$

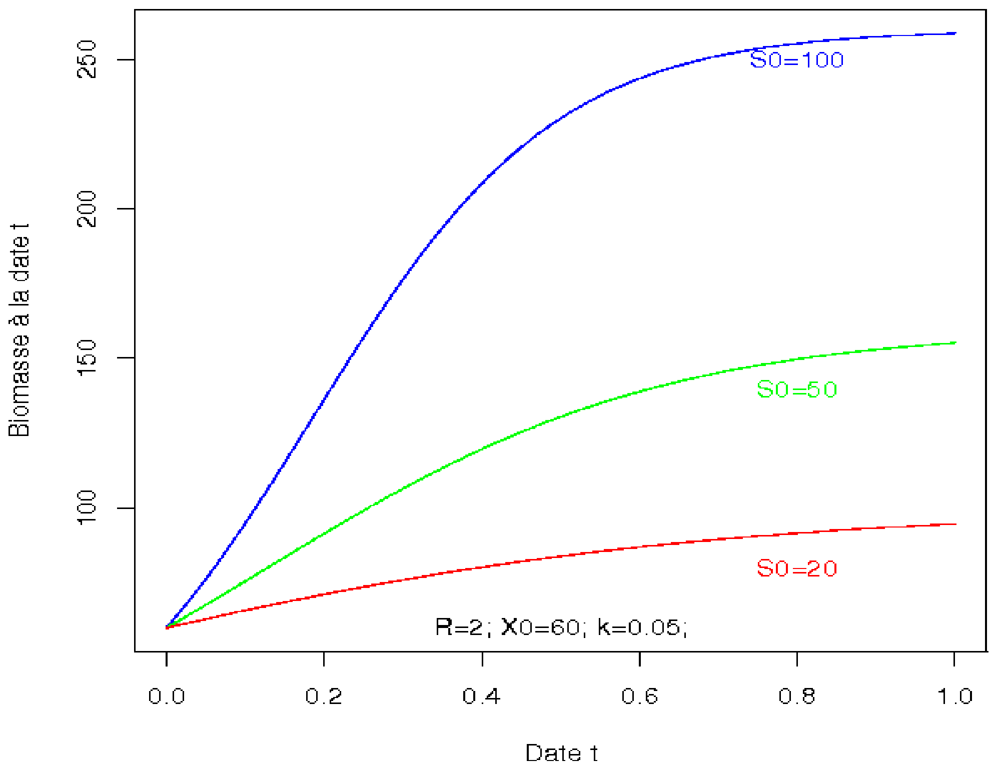
Elle change de signe pour $x = \dfrac{\rho}{2}$ qui est donc un point d'inflexion.

On a une solution explicite de l'EDO :

$$x(t) = \frac{\rho}{1 + \dfrac{\rho - x_0}{x_0} e^{-rt}}.$$

Etape 5.
Identifier le modèle, c'est-à-dire attribuer des valeurs numériques aux paramètres du modèle à partir de données expérimentales.

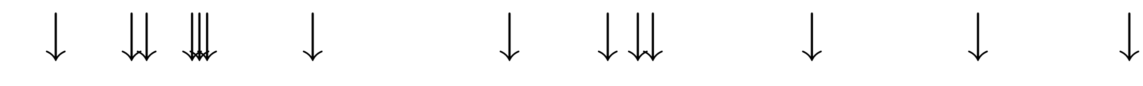
Etape 6.
Valider le modèle à partir de tests statistiques.



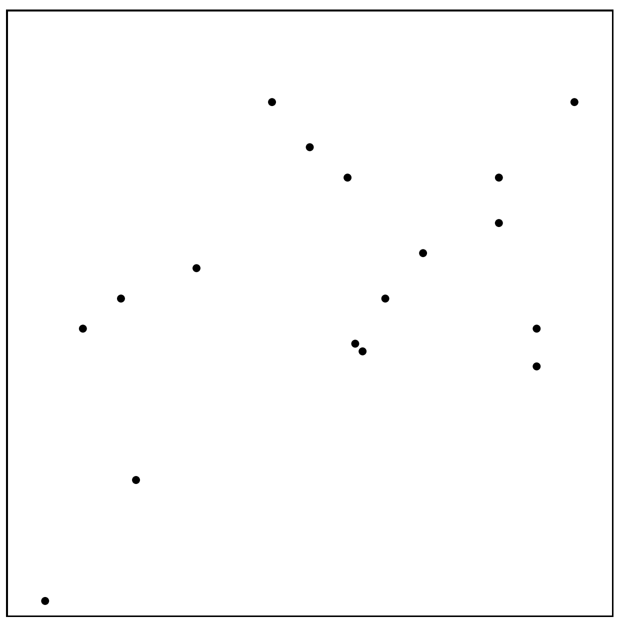
Exemple de trajectoires selon les conditions initiales et paramètres du modèle

QUELQUES MODELES SIMPLES EN ÉCOLOGIE ET EN FIABILITÉ

On s'intéresse à des occurrences d'événements dans le temps et/ou dans l'espace :



Répartition dans le temps



Répartition dans le plan

Soit X une variable aléatoire réelle positive, correspondant à **la durée de vie d'un matériel**.

X est dit sans mémoire (ou sans usure) si

$$\forall (s,t) \in I\!R_+^2, P(X > s+t | X > t) = P(X > s).$$

Autrement dit, sachant le matériel en état de marche à la date t, la loi de la durée de vie future est la même que celle de sa durée de vie initiale.

Le matériel ne s'userait pas !!

Résultat :
X sans mémoire $\Longleftrightarrow$ X suit la loi exponentielle

Définition : Si X de densité f et de fonction de répartition F, on appelle *taux de hasard* la fonction r telle que :

$$\forall t \geq 0, \quad r(t) = \frac{f(t)}{1 - F(t)}.$$

r est aussi appelé *risque instantané*.
Son interprétation ?

Si le matériel fonctionne à la date t, la probabilité d'une panne dans $]t, t+dt[$ est :

$$P(X \in]t, t+dt[\mid X > t) = \frac{P(X \in]t, t+dt[\,,\ X > t)}{P(X > t)}$$
$$= \frac{P(X \in]t, t+dt[)}{P(X > t)} \simeq \frac{f(t)dt}{1 - F(t)} = r(t)dt$$

Pour la loi exponentielle de paramètre λ, on a

$\forall t \geq 0, f(t) = \lambda e^{-\lambda t}$ et $F(t) = 1 - e^{-\lambda t}$ d'où $r(t) = \lambda$.

Cette loi correspond à un risque instantané constant et donc une absence d'usure.

Résultat :
Soit X une variable aléatoire réelle positive admettant une densité, le taux de hasard de X caractérise la loi de cette variable.

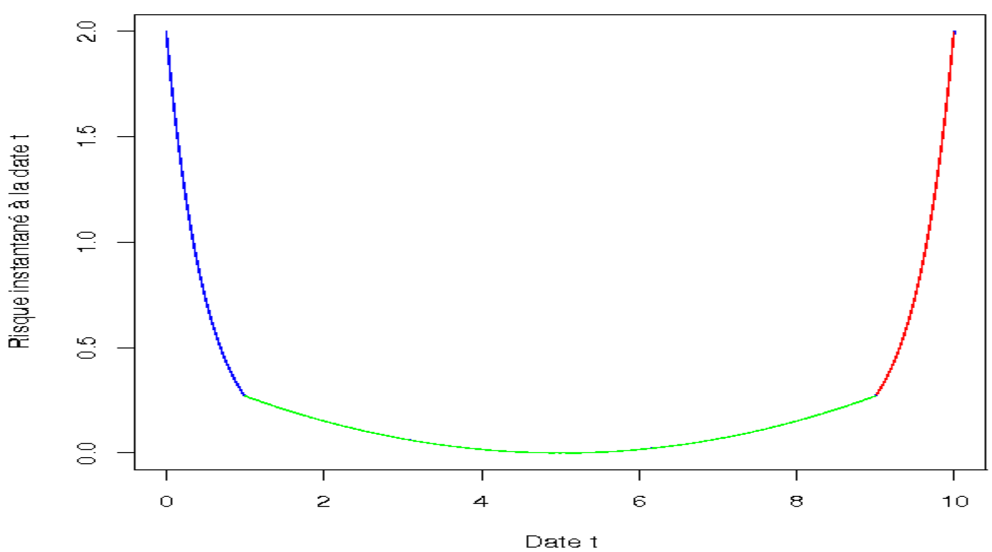
On démontre aisément que

$$F(t) = 1 - \exp(-\int_0^t r(s)ds).$$

avec donc la contrainte $\int_0^{+\infty} r(s)ds = +\infty$.

En général, on a trois phases :

1) Rodage
2) Phase de durée de vie utile
3) Usure, dégénérescence, Vieillissement

- 

Exemple de courbe en baignoire

Test d'ajustement au modèle
Il permet de déterminer si les durées de vie dans une population obéissent au modèle proposé
(avec une **probabilité α de rejeter à tort contrôlée**).

Le plus célèbre est celui dit du χ^2 :

Principe : comparer fréquences observées de survie aux fréquences théoriques fournies par le modèle.

Calcul de

$$C = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}.$$

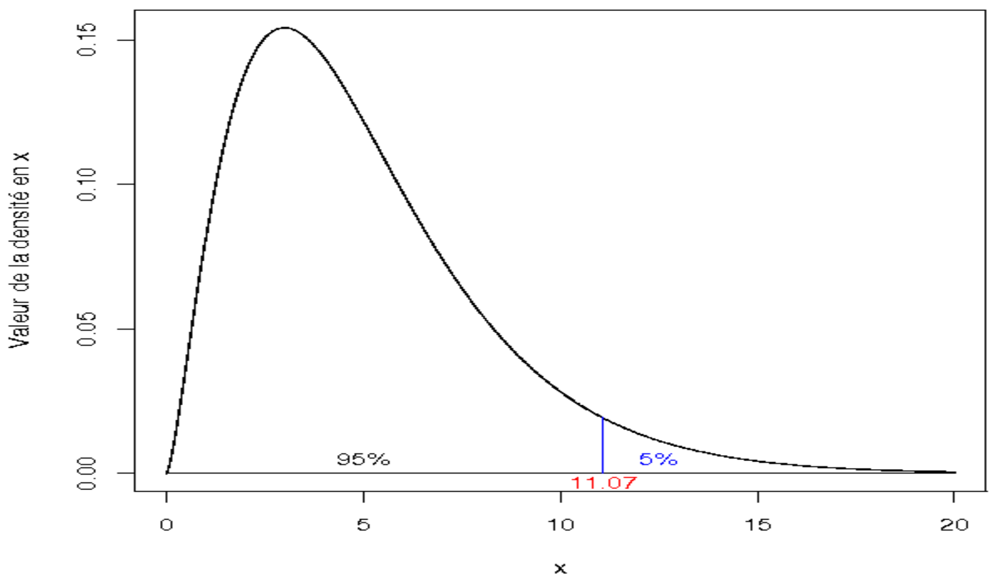
où
k = nombre de classes considérées
n_i = effectif d'individus dans la classe i
p_i = probabilité de la classe i
$n = \sum_{i=1}^{k} n_i$ est le nombre d'individus dans l'échantillon

Règle de décision :

Rejet du modèle si $C > \chi^2_{1-\alpha,\nu}$

Non rejet du modèle si $C \leq \chi^2_{1-\alpha,\nu}$

où $\chi^2_{1-\alpha,\nu}$ est le fractile d'ordre $1-\alpha$ de la loi dite du χ^2 à ν degrés de liberté.



Densité de probabilités de la loi du χ^2
à 5 degrés de liberté

Exemple : observations de durée de vie, 250 individus choisis par tirages indépendants dans une production.

Durée moyenne estimée $\simeq 2$ unités de temps (sous la loi exponentielle)

Classes statistiques	Effectifs observés	Probabilité de classe	Effectifs théoriques	Contributions au χ^2
]0, 1]	52	0,393	98,25	21,77
]1, 2]	71	0,239	59,75	2,12
]2, 3]	53	0,145	36,25	7,74
]3, 4]	25	0,088	22,00	0,41
]4, 5]	32	0,053	13,25	26,53
]5, 6]	7	0,032	8,00	0,12
]6,+∞]	10	0,050	12,50	0,50

$C = 59,196, \quad \chi^2_{0,95;5} = 11,070$

Probabilité critique (p-value) $= P(\chi^2 > C) = 0.00000$

Conclusion : modèle exponentiel largement rejeté

En cas de difficulté pour exprimer les probabilités de classe?

Simulations de Monte Carlo :

Principe :

1) Simuler numériquement N fois un échantillon selon le modèle privilégié
2) Calculer pour chacune des N simulations le critère de test C
3) Comparer la valeur de C calculée à partir des données à la distribution des valeurs simulées de C
4) Rejeter le modèle si la valeur observée est **extraordinaire** par rapport à la distribution simulée.

Remarque : nécessité de savoir simuler le modèle privilégié!

Modèles de répartition de points (Processus ponctuels)

On s'intéresse à la répartition d'individus, d'objets ou d'occurrence d'évènements (points) dans une zone D.

1) Les points sont indiscernables autrement que par leur location temporelle ou spatiale
   (**dans le cas dit sans marque**)
2) Un point peut correspondre à l'occurrence d'un événement, ou à un objet de taille négligeable par rapport à l'échelle d'étude

Exemples

- Instants d'arrivée de clients sur un serveur
- Points d'infestation par un ravageur des cultures
- Succession de pannes d'un système
- Emplacements de lieux classés à risque dans une région
- Positions d'étoiles dans une portion du ciel
- Trafic routier
- Epicentres dans une région au cours d?une année

L'hypothèse souvent testée est celle de :

Répartition complètement aléatoire (RCA) des points.

Elle correspond aux deux suppositions suivantes:

1) **Homogénéité** : Les différentes location dans la zone ont la même probabilité *d'accueillir* un point
2) **Indépendance** : Les points se répartissent indépendamment les uns des autres

On a alors le résultat suivant :

Pour toute partie A mesurable de D, le nombre de points $N(A)$ dans A suit la loi de Poisson de paramètre $\lambda|A|$ où λ est le nombre espéré de points par unité de surface, et $|A|$ la surface de A.

On parle de **processus ponctuel de Poisson homogène**.

$E(N(A) = Var(N(A))$ d'où l'indice dit de dispersion $I_d = \dfrac{Var(N(A)}{E(N(A))}$ qui vaut 1 sous l'hypothèse de répartition complètement aléatoire, et qui permettra de voir l'écart à cette hypothèse.

Test de l'indice de dispersion

Données: comptages $x_1, x_2, \cdots, x_n$ dans n unités d'échantillonnage disjoints de même taille.

Critère de test : $(n-1)\widehat{I_d} = \sum\limits_{i=1}^{n} \dfrac{(x_i - \bar{x})^2}{\bar{x}}$ où

$\bar{x} =$ moyenne des x_i et $\widehat{I_d}$ l'indice de dispersion estimée.

Règle de décision :

Rejet de l'hypothèse RCA si $(n-1)\widehat{I_d} > \chi^2_{1-\alpha,n-1}$

Non rejet de RCA si $(n-1)\widehat{I_d} \leq \chi^2_{1-\alpha,n-1}$

où $\chi^2_{1-\alpha,n-1}$ est le fractile d'ordre $1-\alpha$ de la loi dite du χ^2 à $n-1$ degrés de liberté.

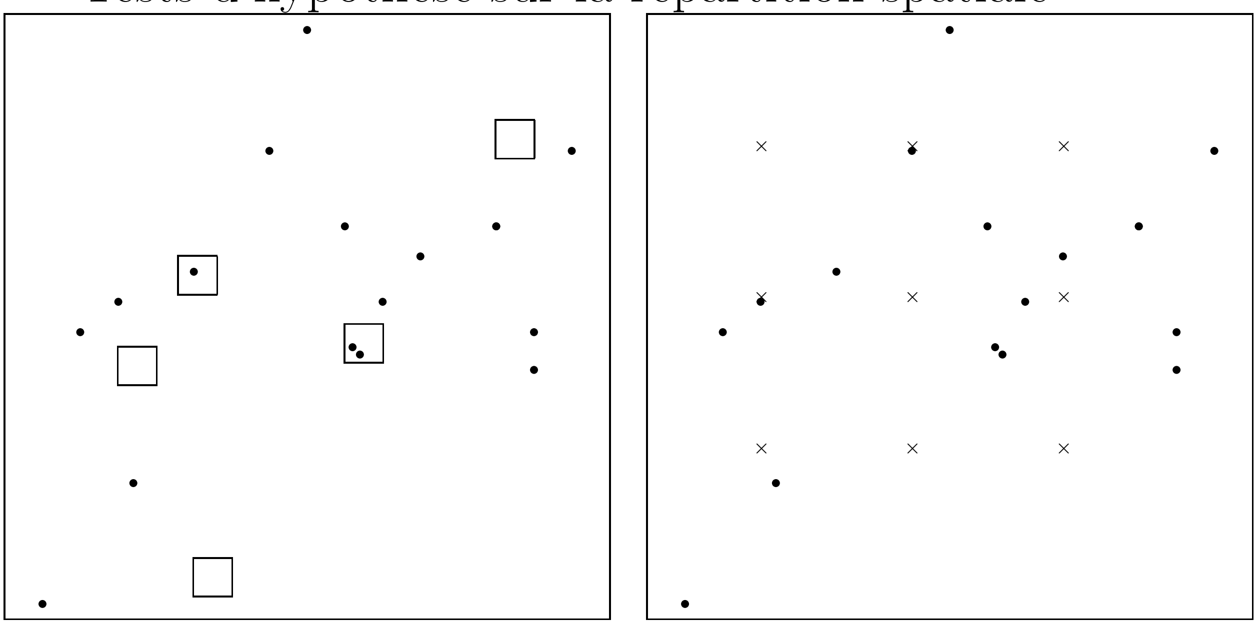
Remarque : 2 façons d'observer partiellement une répartition de points (cas spatial)

- comptage des points de la répartition dans des unités d'échantillonnage spatiale
- mesures de distance entre points de $N(.)$ et points d'échantillonnage

<u>But :</u>

Estimation d'intensité et de fonction de lien

Tests d'hypothèse sur la répartition spatiale



Pour les méthodes basées sur les mesures de distance, on utilise la loi de la distance D au plus proche point de la répartition.

La fonction de répartition vérifie

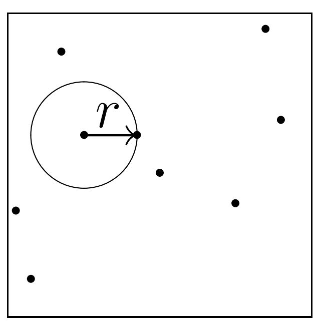
$$F(r) = P(D \leq r) = 1 - P(D > r)$$

Sous l'hypothèse de RCA, on a donc

$$F(r) = 1 - \exp(-\pi r^2) \quad \text{pour} \quad r \geq 0.$$

En dérivant,
on obtient la densité f de la variable aléatoire D :

$$f(r) = 2\pi r \exp(-\pi r^2) \mathbb{1}_{I\!R_+}(r).$$



Quelques processus ponctuels

Modèles mathématiques	**Phénomènes observés**
Processus ponctuel	Répartition aléatoire de points (évènements, petits objets, individus)
Processus de Poisson Homogène (ou stationnaire)	Points répartis indépendamment les uns des autres dans un environnement homogène
Processus de Poisson Hétérogène (ou non stationnaire)	Points répartis indépendamment dans un environnement hétérogène
Processus de groupes	Individus répartis en groupes

Importance de l'information a priori

Présences de facteurs environnementaux (covariables) inconnus ou non observés

$\Downarrow$

SURDISPERSION and AUTOCORRELATION

OUVRAGES CONSEILLÉS

Gilbert DEMENGEL, Paul BENICHOU, Rosine BENICHOU, Norbert BOY.
Probabilités, statistique inférentielle, Fiabilité. Outils pour l?ingénieur.
Edition ELLIPSES, 1997.
ISBN : 2 7298 4720 0

Philippe MICHEL. *Cours de mathématiques pour économistes.*
Edition ECONOMICA, 1999.
ISBN-13 : 978 2 7178 1683 9

Alain PAVÉ. *Modélisation en biologie et en écologie.*
Edition ALEAS, 1994.
ISBN : 2 9080 1632 X