Introduction aux statistiques spatiales
Introduction générale

Denis Allard

Biostatistique et Processus Spatiaux (BioSP), INRA, Avignon

27 novembre 2012

Grands domaines des statistiques spatiales

1. Données continues irrégulièrement espacées.

   Ce sont les données de type géostatistique. Elles peuvent être uni- ou multi-variées. Les données sont en général situées sur une partie du plan ou de l'espace à trois dimensions. Ex : qualité d'un sol ; pollution atmosphérique

2. *Données sur lattice* ne seront pas du tout abordées dans le cadre de ce cours. Les modèles les plus utilisés sont les champs Markoviens. Ces données peuvent évidemment faire aussi l'objet d'un traitement géostatistique. Ex : données sur entité administrative

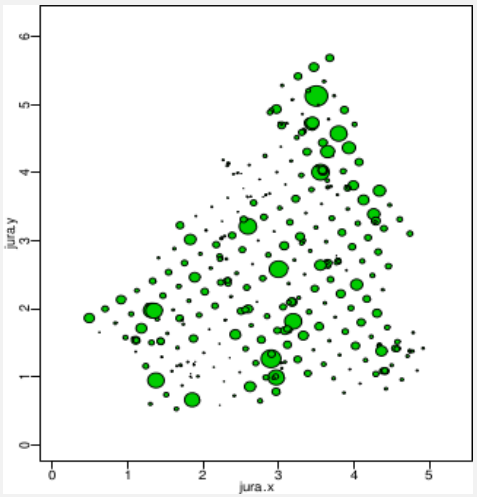
3. *Processus pontcuels*. Les points indiquent seulement la présence d'un événement. Les processus ponctuels peuvent être généralisés aux processus ponctuels marqués, lorsqu'une valeur (une marque) est associée aux points.

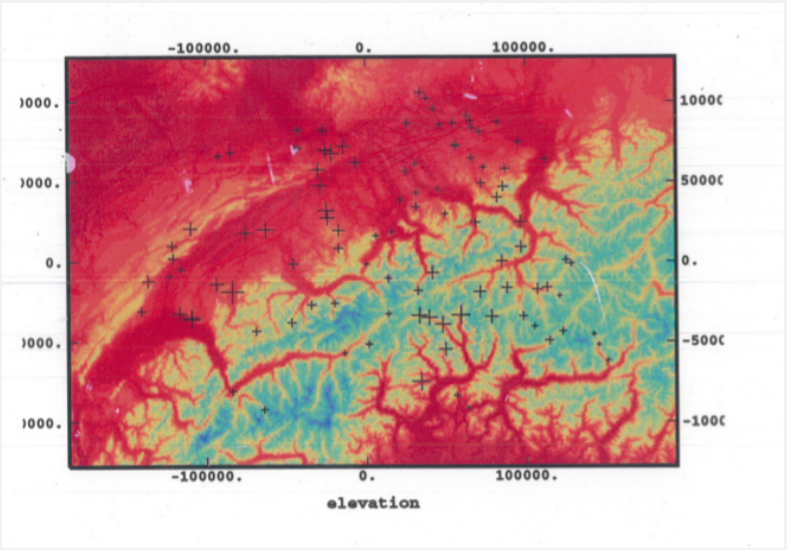
La frontière n'est jamais étanche entre ces trois types de données. Comment traiter des données à support complexe et variable ?
ex. données médicales / dépt.

Données *géostatistiques*

Température, précipitations, qualité d'un sol ; pollution atmosphérique

- caractériser la variabilité spatiale
- interpoler (cartographier) la variable entre les points mesurés
- simuler des variations spatiales du même type
- évaluer l'erreur d'interpolation et la qualité de l'échantillonnage

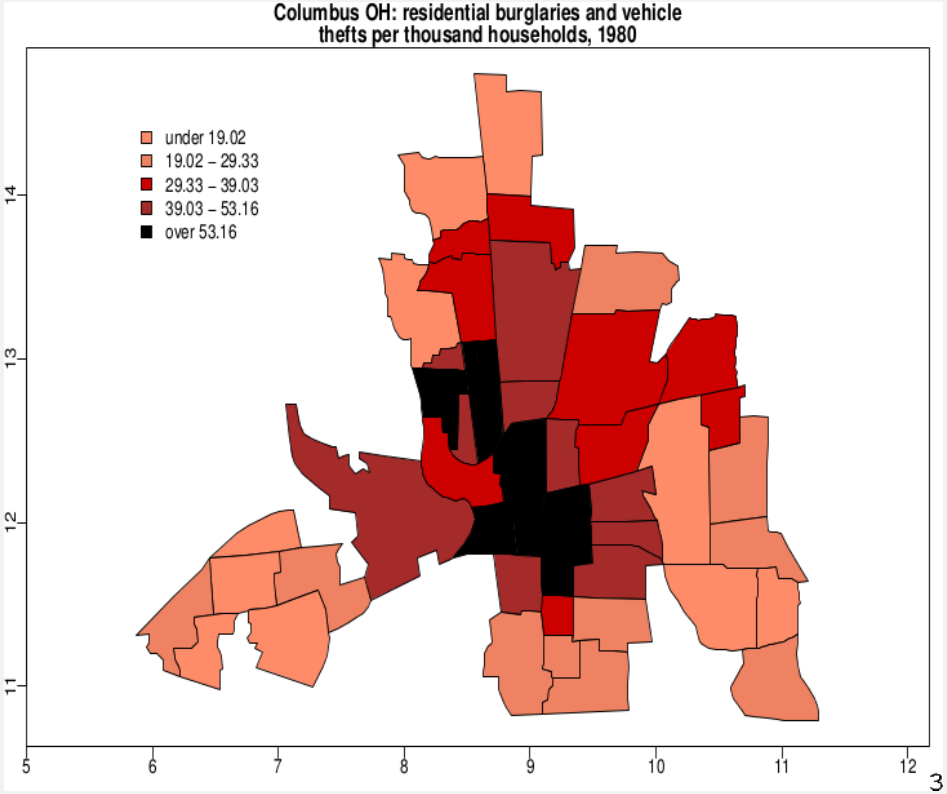




Données sur réseaux (sur lattices)

Données de population, données épidémiologiques sur en ensemble d'entités administratives, image de télédétection

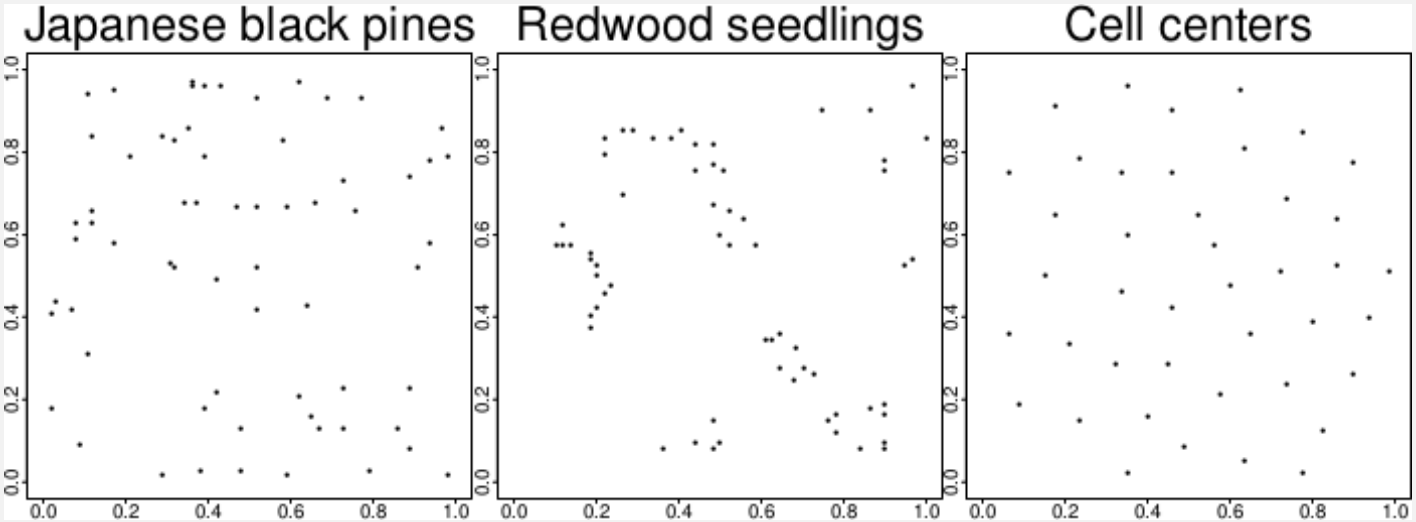
- caractériser la variabilité spatiale : indépendance entre voisins, analyse de résidus
- expliquer la distribution des caractéristiques en fonction des distributions dans un voisinage



Processus ponctuels ou processus d'objets

Arbres d'une forêt, répartition d'espèces végétales, d'animaux.

- Caractériser la distribution spatiale des objets : indépendance, régularité, agrégation ?
- Expliquer la distribution des caractéristiques des objects en fonction de leurs positionnements relatifs
- simuler des distributions spatiales du même type



L'échantillonnage des données

- Échantillon unique : pas de répétition. Pour mener l'inférence à bien, on doit donc faire une hypothèse d'ergodicité, et de stationnarité.
- Région étudiée : parfois la région étudiée est imposée (inventaire forestier, parcelle), mais parfois la région sur laquelle on possède des données est en soi une donnée (banc de poisson, gisement minier).
- L'échantillonnage peut être parfois fortement biaisé (ex : le géologue concentre les échantillons dans les zones susceptibles d'être riches).

Echantillonnages non biaisés

En général, trois types d'échantillonnage non biaisé :

- *l'échantillonnage aléatoire pur*

  échantillonne les faibles distances

  peu rentable car il présente en même temps des redondances et beaucoup de vides

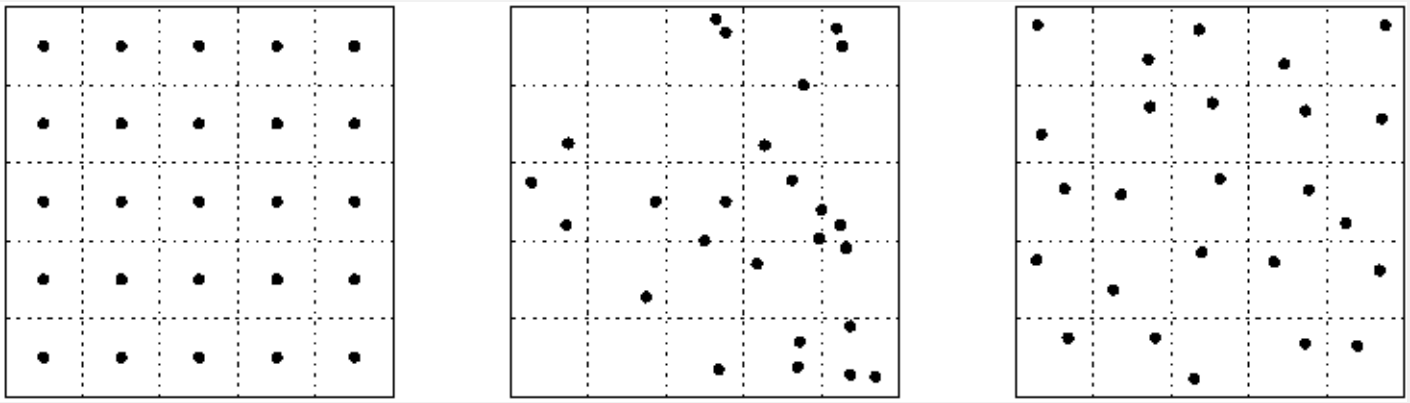
- *l'échantillonnage (régulier)*

  n'échantillonne pas les faibles distances

  risque d'occulter des phénomènes de période égale à la maille

- *l'échantillonnage aléatoire stratifié*

  le bon compromis



Spécifités/difficultés des statistiques spatiales

- Données non indépendantes
- Pas de relation d'ordre dans les espace $d \geq 2$
- Quelle asymptotique ?
- Vraisemblance adapatée ? Calculable ?
- Parfois : effets de bords, dépendance fortes,...

Plan du cours

1. Théorie des champs aléatoires : les différentes hypothèses de stationnarité ; propriété des fonctions de covariance et du variogramme ; lien entre régularité du champ et régularité de la fonction de covariance.
2. Estimation de la fonction de covariance : le variogramme empirique et ses propriétés. Estimation par méthode des moments, par maximum de vraisemblance et ses dérivées (REML, vraisemblance composite)
3. Le krigeage pour la prédiction spatiale : propriétés ; questions pratiques liées à sa mise en œuvre ; erreur de prédiction, validation croisée.
4. Simulation des champs aléatoires ; simulations conditionnelles.
5. Spécificité des données collectées à la fois dans l'espace et dans le temps : statistiques spatio-temporelles.

Organisation du cours

- 5 séances de 4 heures : 10h15 – 12h15 puis 13h15 – 15h15
- Pas de TD
- Un examen écrit (60 %)
- Un projet, par groupes de 2
- Une restitution du projet : rapport + présentation orale (40%)

Régularité