# Enlever les valeurs extrêmes de concentration: pour quoi faire?

Jacques.Rivoirard@ensmp.fr

CG Fontainebleau

Journées de Géostatistique 18-19 Septembre 2003

# Les valeurs extrêmes de concentration (métaux précieux, poisson, pollution…)

- un problème *extrêmement* sérieux, vue l'importance des valeurs extrêmes en abondance globale ou en dépassement de seuil

  … mais forte instabilité des statistiques et des outils

- Besoin de méthodes adéquates, avec hypothèses mesurées (développements en cours pour or et poisson)

# Une « technique » très répandue: la suppression des valeurs extrêmes

- … notamment du variogramme, pour calculer une « structure »
  … laquelle pourra servir à un krigeage… excluant ou non les valeurs extrêmes
- Problème abordé ici: que peut-on dire d'une telle technique de façon « théorique »?
  est-il possible de construire des modèles géostatistiques qui la légitiment?
  par exemple dans lesquels le variogramme « entier » est identique au variogramme sans extrêmes, ou s'en déduit simplement (ex: addition de pépite)

# Hypothèses de base

- seront considérées comme valeurs extrêmes: les valeurs au-dessus d'un seuil z donné, supposées peu fréquentes et beaucoup plus fortes que les autres
- ces valeurs extrêmes ne sont pas des valeurs erronées
- a priori elles peuvent se trouver n'importe où dans le champ étudié
- on ignore les incertitudes sur le variogramme calculé sans les extrêmes: celui-ci est supposé parfaitement connu

# Le variogramme conditionnel

# Le variogramme conditionnel, en théorie…

$$\gamma_{-z}(h) = \frac{1}{2} E\left[ \left( Z(x+h) - Z(x) \right)^2 \mid Z(x) < z, Z(x+h) < z \right]$$

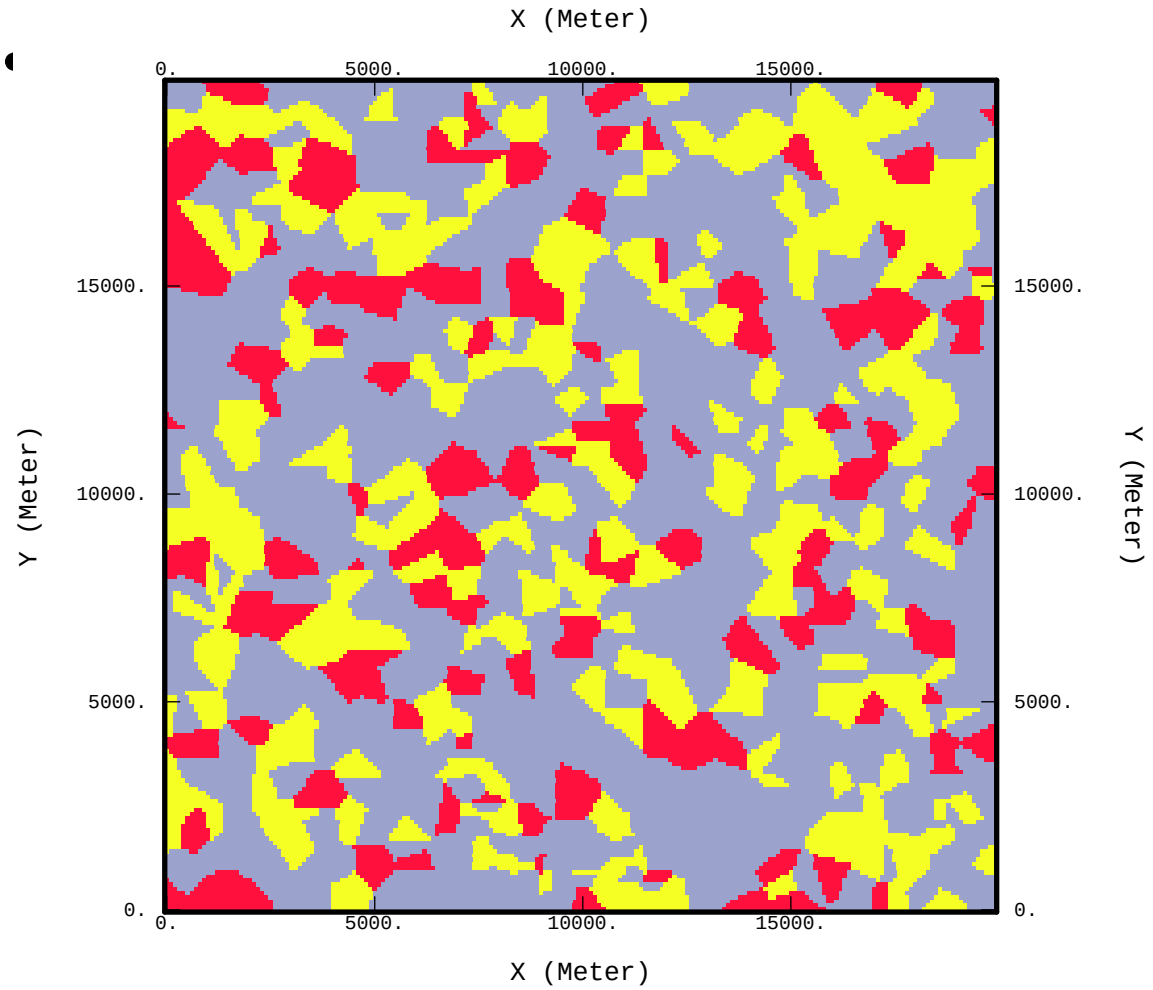
- est-ce nécessairement une fonction de type variogramme (conditionnellement définie négative)?

- Lien avec $\gamma(h) = \frac{1}{2} E\left[ \left( Z(x+h) - Z(x) \right)^2 \right]$ théoriquement possible connaissant les lois bivariables (Z(x), Z(x+h)) (géostat non-linéaire, lourde en hypothèses!)

# Un cas simple: le modèle mosaïque à valuations indépendantes

- Espace partitionné en compartiments
- On value (tous les points de) chaque compartiment indépendamment des autres, et selon la même loi
- Deux points à distance h:
  - appartiennent à un même compartiment (et ont donc la même valeur) avec proba r(h)
  - appartiennent à des compartiments différents (et ont donc des valeurs indépendantes, éventuellement égales) avec proba $\gamma(h) = 1 - r(h)$

# Modèle mosaïque à valuations indépendantes



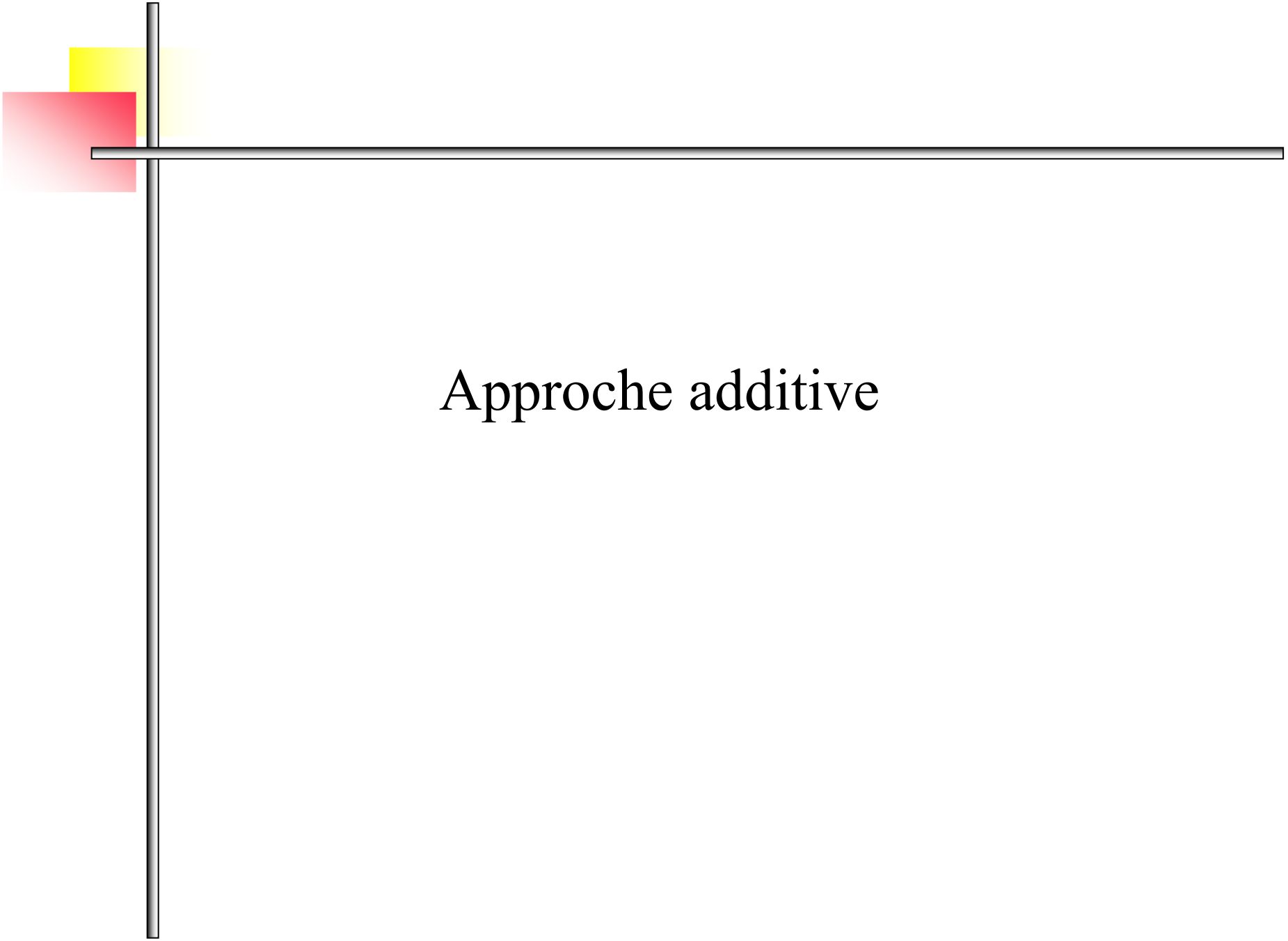
# Modèle mosaïque à valuations indépendantes

- La variable Z(x), de même que ses transformées, ont toutes une structure identique à $\gamma(h) = 1 - r(h)$

- Cependant le variogramme (?) sans les extrêmes est différent (plus régulier à petites distances, car on chevauche moins souvent les compartiments):

$$\gamma_{-z}(h) = \frac{\mathrm{var}(Z \mid Z < z)\ P(Z < z)\ \ \gamma(h)}{1 - \left[1 - P(Z < z)\right]\gamma(h)}$$

$$\gamma(h) = \frac{\gamma_{-z}(h)}{\mathrm{var}(Z \mid Z < z)\ P(Z < z) + \left[1 - P(Z < z)\right]\gamma_{-z}(h)}$$



# Approche additive

# Approche additive

- L'élimination des fortes valeurs suggère le modèle additif:

$$Z(x) = Z_1(x) + Z_2(x)$$

somme de deux FA $\geq 0$ indépendantes:

  - fond $Z_1$, inférieur au seuil z
  - $Z_2$ responsable du dépassement de seuil

- ce qui permet le calcul de statistiques concernant $Z_1(x)$ à partir des seuls points de $S_1$ où $Z(x) < z$

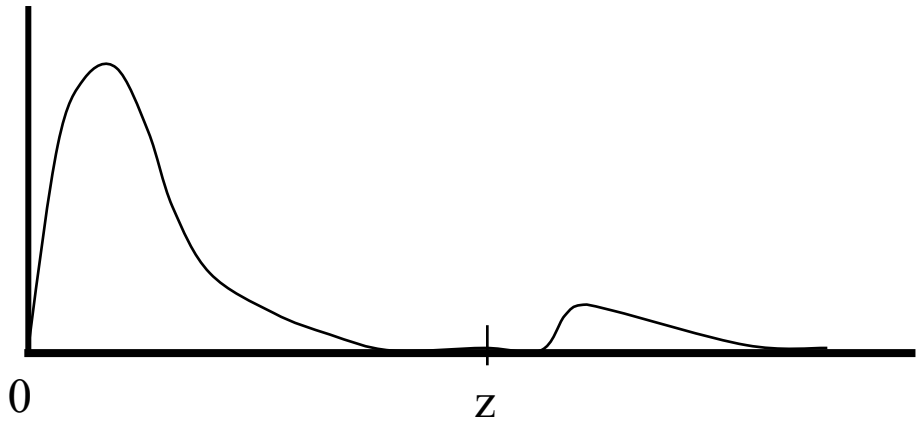
# Approche additive: distribution

- $Z_2(x) > 0 \Leftrightarrow Z(x) = Z_1(x)+Z_2(x) \geq z \Leftrightarrow Z_2(x) \geq z - Z_1(x)$

- $Z_1(x)$ et $Z_2(x)$ étant indépendants:

  $Z_2(x) > 0 \Rightarrow Z_2(x) \geq z\text{-}\min(Z_1) \geq \max(Z_1) - \min(Z_1)$

  rôle très particulier du seuil dans la distribution de Z (bimodale par ex )



# Approche additive: structure

- $Z(x) = Z_1(x) + Z_2(x)$ => Modèle structural simple:

$$\gamma_Z(h) = \gamma_1(h) + \gamma_2(h)$$

$$\gamma_{Z,Z_1}(h) = \gamma_1(h)$$

$$\gamma_{Z,Z_2}(h) = \gamma_2(h)$$

- $\gamma_1(h)$ étant supposé connu: $\gamma_Z(h)$ ?
  - $\gamma_2(h)$ composante supplémentaire, pépite par ex
  - rescaling si $\gamma_2(h) \equiv \gamma_1(h)$

# Approche additive: estimation

- $\gamma_Z(h)$

  permet calcul de variance et krigeage sur Z(x)

- modèle bivariable permet calcul de variances et cokrigeage à partir des données:
  - $Z_1(x)$ et $Z_2(x)=0$ connus sur $S_1$
  - $Z(x) = Z_1(x)+Z_2(x)$ connu sur les autres points $S_2$

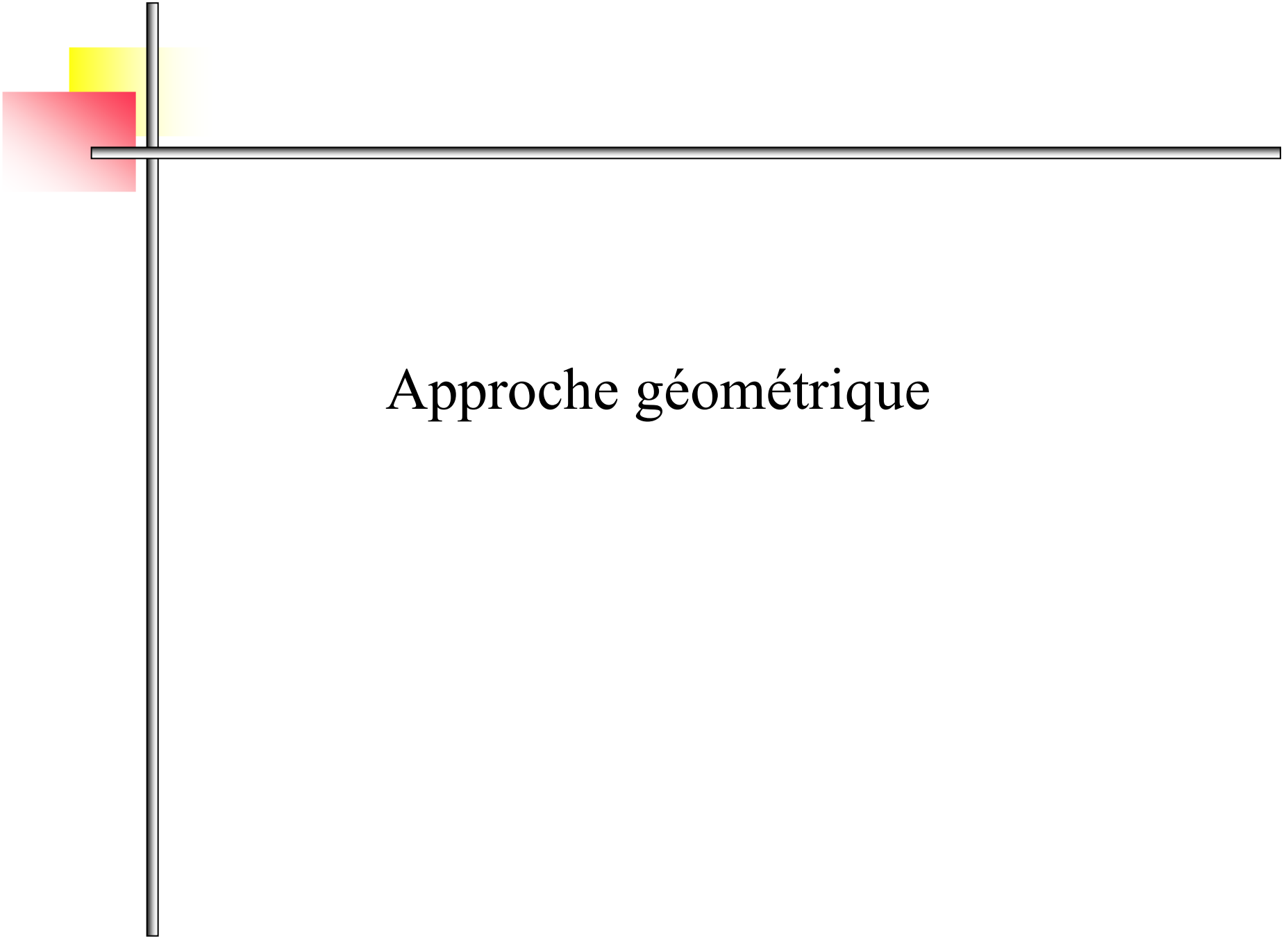
  (bien que $Z_1(x)$ et $Z_2(x)$ soient indépendantes, leur cokrigeage n'est pas leur krigeage)

# Approche additive: estimation

- Alternative au cokrigeage (non optimale, mais privilégiant la structure de $Z_1$, la mieux connue):
  - kriger $Z_1$, y compris sur $S_2$, à partir des seules valeurs non extrêmes $Z=Z_1$ sur $S_1$
  - en déduire valeurs estimées de $Z_2=Z-Z_1$ sur $S_2$
  - estimer $Z_2$ à partir de ces valeurs estimées de $Z_2$ sur $S_2$ et des valeurs nulles de $Z_2$ sur $S_1$ (saupoudrage uniforme de la moyenne $m_2$ de $Z_2$ si celle-ci est pépitique)

# Approche additive: conclusions

- intérêt: simplicité du modèle structural et de l'estimation
- inconvénients:
  - hypothèse très particulière sur la distribution de valeurs de $Z$
  - $Z_1$ et $Z_2$ ne sont connus individuellement qu'aux points de données où $Z<z$,
    l'estimation ne tient pas explicitement compte du fait que $Z_2$ est $> 0$ aux points de données où $Z \geq z$



# Approche géométrique

# Approche géométrique: base

- distinction selon l'appartenance aux ensembles

$$A = A_z = \{x \mid Z(x) \geq z\}$$

$$A^c = A_z^c = \{x \mid Z(x) < z\}$$

par exemple:

$$Z(x) = Z(x)1_{Z(x)<z} + Z(x)1_{Z(x)\geq z}$$

$$= Z(x)1_{x\in A^c} + Z(x)1_{x\in A}$$

# Approche géométrique

- Eléments à considérer: structures et relations entre
  - ensemble A des fortes valeurs
  - Z dans A
  - Z dans $A^c$
  - Z au passage de $A^c$ à A

# Approche géométrique: modèle simple

- $$Z(x) = Y_1(x)1_{x\in A^c} + Y_2(x)1_{x\in A}$$

avec $1_{x\in A}$

$Y_1(x)$, *entre* 0 *et* $z$

$Y_2(x)$, $\geq z$

indépendants
pouvant faire l'objet d'estimations séparées
(mais hétérotopiques)

- Structure de Z = combinaison des structures des 3 variables

# Approche géométrique: modèle simple

si $1_{x \in A}$ et $Y_2(x)$ (supposés) quasi-pépitiques:

- Variogramme $\gamma_Z(h) \equiv \gamma_{-z}(h)\left[P(Z < z)\right]^2 + pépite$

  avec $pépite = \operatorname{var} Z - \operatorname{var}(Z \mid Z < z)\left[P(Z < z)\right]^2$

- Estimation de $Y_1$ , soit $Z \mid Z<z$, se complète par saupoudrage de $m_2 = E[Y_2(x)] = E[Z(x) \mid Z(x) \geq z]$ en proportion $P(Z \geq z)$

# Approche géométrique: modèle simple

- Un peu plus général (suppose absence d'effets de bord, simples et couplés, de Z dans A et $A^c$): factorisation par:

$$1_{Z(x)\geq z} \qquad [Z(x)-m_1]1_{Z(x)<z} \qquad [Z(x)-m_2]1_{Z(x)\geq z}$$

avec $\quad m_1 = E[Z(x) \mid Z(x) < z]$

$$m_2 = E[Z(x) \mid Z(x) \geq z]$$

- estimation par krigeage séparé (isotopique) des facteurs

# Approche géométrique: modèle simple

- Le modèle simple légitime un krigeage séparé des valeurs de Z inférieures au seuil z
- Il fait jouer un rôle très particulier au seuil: ce qui se passe au-dessus ne dépend pas de ce qui se passe au-dessous. Difficile en particulier d'étendre à un modèle multi-seuil.

# Conclusions

- Dans les cas envisagés, la suppression des valeurs extrêmes permet des techniques effectivement simples, mais dans des modèles où le seuil joue un rôle très particulier.

- D'autres approches, plus sophistiquées semblent nécessaires sinon…