

# MODELISATION DES ERREURS DE POSITION ET D'ATTRIBUTS DANS LES BASES DE DONNEES GEOGRAPHIQUES

*Olivier BONIN*

IGN, Laboratoire COGIT

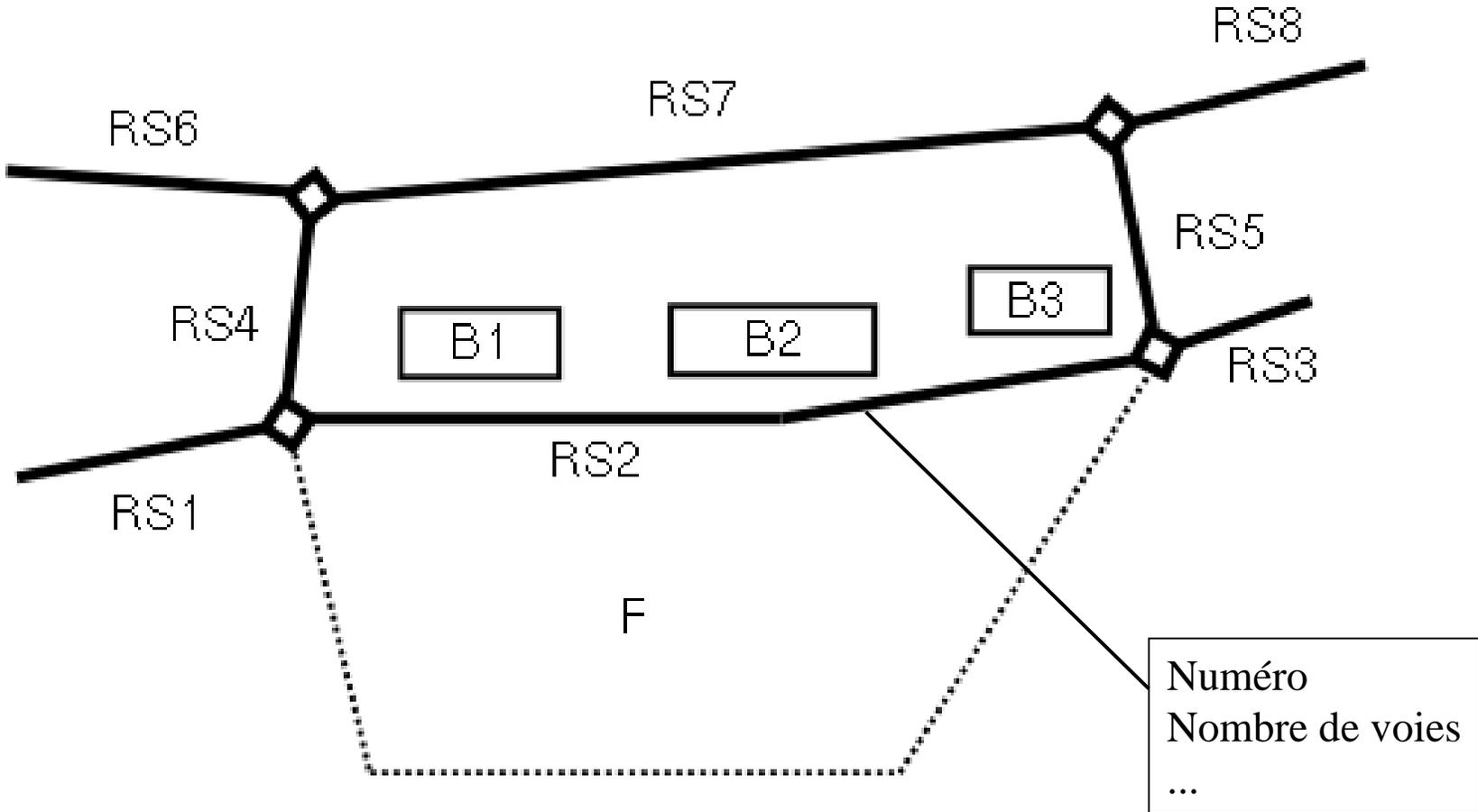


# Bases de données géographiques

---

- Information spatialisée : modélisation de la composante spatiale
  - Modélisation sous forme de grille régulière (approche image) : un attribut par pixel
  - Objets géographiques décrits par des primitives dites « vecteur » : point, ligne polygonale, polygone (éventuellement à trou), et des attributs

# Modélisation « vecteur »



- Erreurs de position des objets
- Erreurs de position relative des objets
- Erreurs de forme
  
- Erreurs d'attributs

# Pourquoi modéliser ?

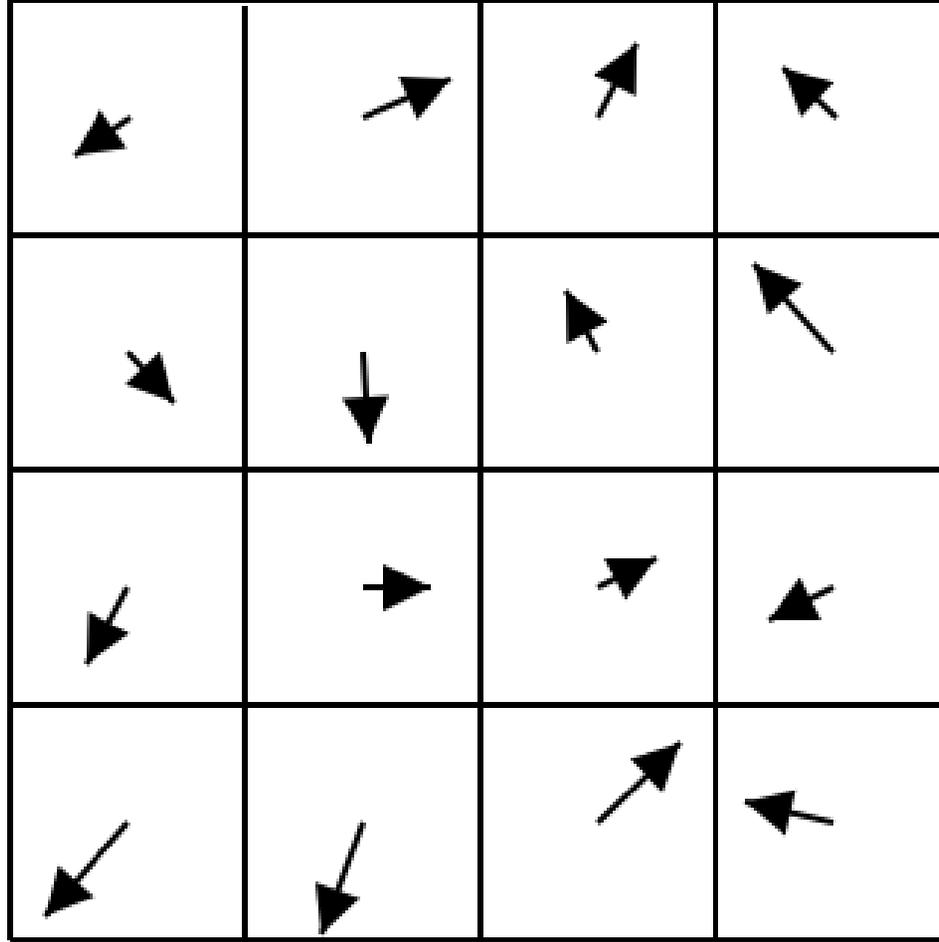
---

- Evaluer synthétiquement les erreurs
- Simuler par Monte-Carlo des erreurs pour des analyses de sensibilité et des analyses d'incertitudes d'applications géographiques
- Mieux connaître la confiance qu'on peut accorder aux résultats d'analyses géographiques et aux prises de décision

- $(x_1, x_2, \dots, x_n)$  réalisations de  $n$  variables aléatoires  $(X_1, X_2, \dots, X_n)$
- Indépendance ? Equidistribution ? Cadre paramétrique pour la loi de  $(X_1, X_2, \dots, X_n)$  ?
- Plus fondamentalement, que représentent réellement les  $x_i$  pour lesquels on construit un modèle ?

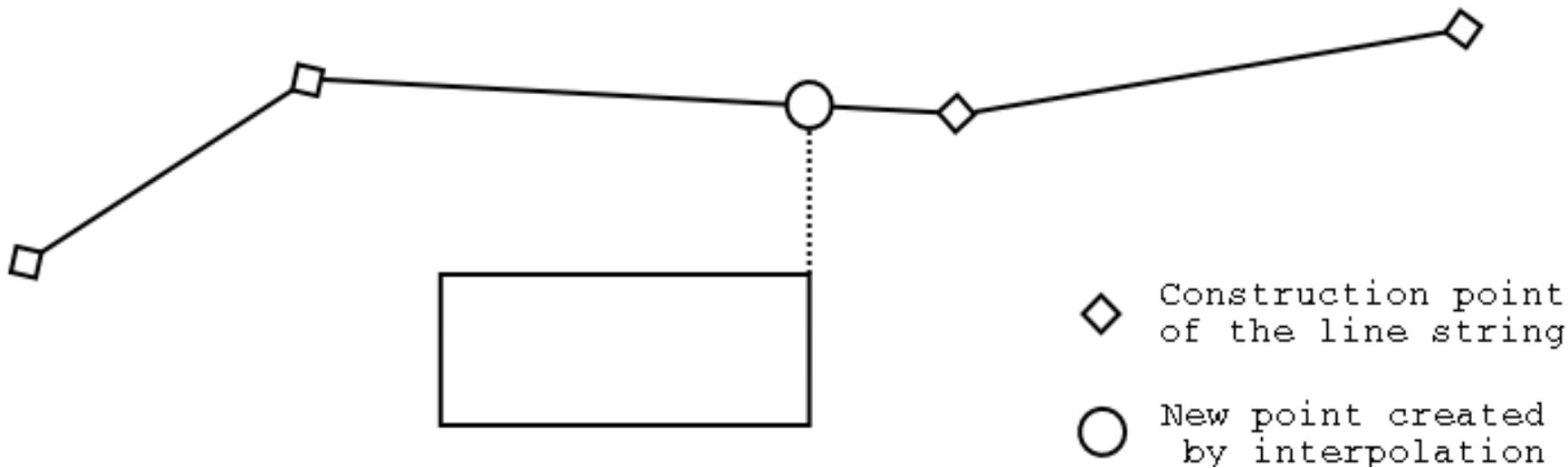
- Ecart observé : écart en  $x$  et en  $y$  entre les points du jeu de la base de données et la position réelle sur le terrain de ces points
- Utilité de régionaliser (daller) la zone d'étude pour mettre en évidence des biais spatiaux
- Modèle Gaussien adapté d'après les études menées (erreurs assimilables à des erreurs de mesure), pas de corrélation entre les écarts en  $x$  et les écarts en  $y$

# Grille de biais régionalisée

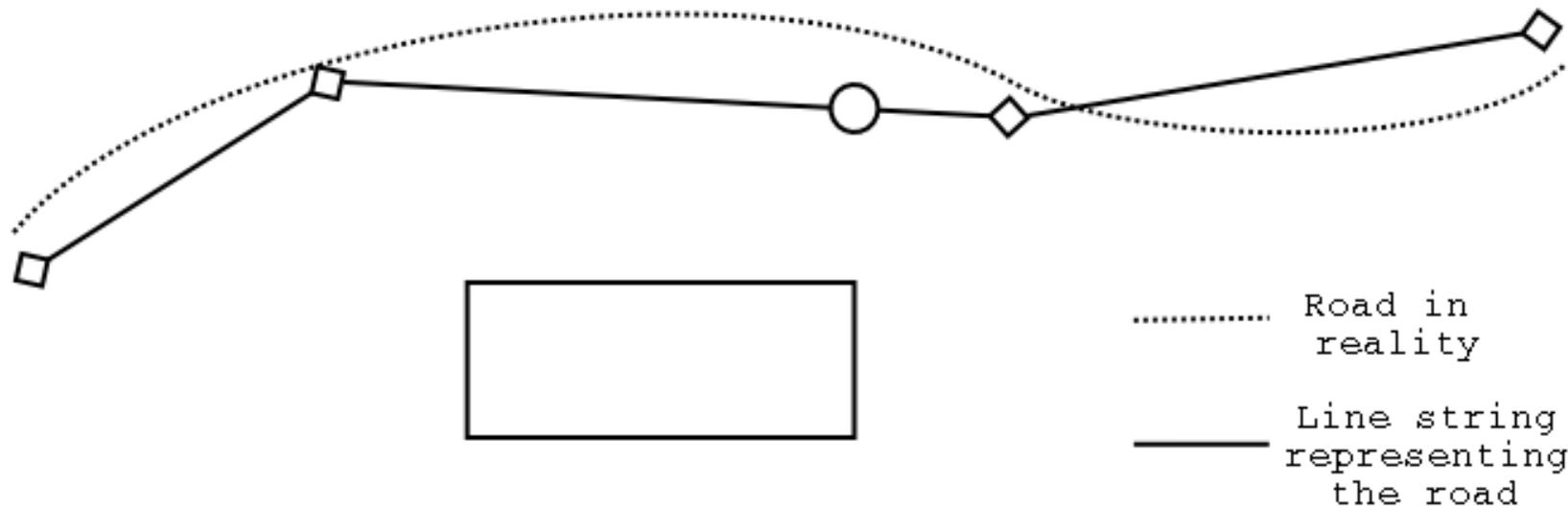


- Géométrie : lignes polygonales
- Idée naturelle : contrôle ponctuel pour les nœuds des réseaux
- Possibilité d'augmenter le nombre de points à contrôler par explicitation d'information

# Explicitation d'information



# Différence de nature des écarts



- Estimer deux séries d'écarts : écarts dus aux erreurs de pointé (points saisis dans la base), et écarts dus à l'interpolation (points explicités)
- Utiliser un modèle de mélange (Gaussien et 2<sup>ème</sup> loi de Laplace par exemple)
- Mais ... les écarts ne sont sans doute pas décorrélés, et les mélanges s'estiment mal

# Approche par processus

---

- Domaine nouveau : processus indexés par des lignes polygonales
- Approches classiques (champs aléatoire, processus ponctuels, etc.) non adaptées
- Modèle proposé : ARMA bilatéral indexé par les lignes (Huang)

# Modèle d'erreurs de position de lignes

$$\mathbf{X}=(X_1, X_2, \dots, X_n)' \quad \boldsymbol{\varepsilon}=(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$$

$$\mathbf{A}\mathbf{X} = \alpha + \mathbf{B}\boldsymbol{\varepsilon} \quad \text{avec}$$

$$\mathbf{A} = \mathbf{I} - \sum_{i=1}^p a_i \mathbf{W}^i \quad \mathbf{B} = \mathbf{I} + \sum_{j=1}^q b_j \mathbf{W}^j$$

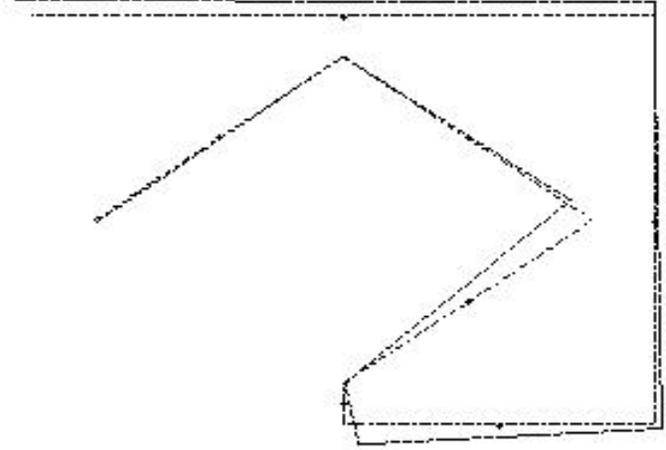
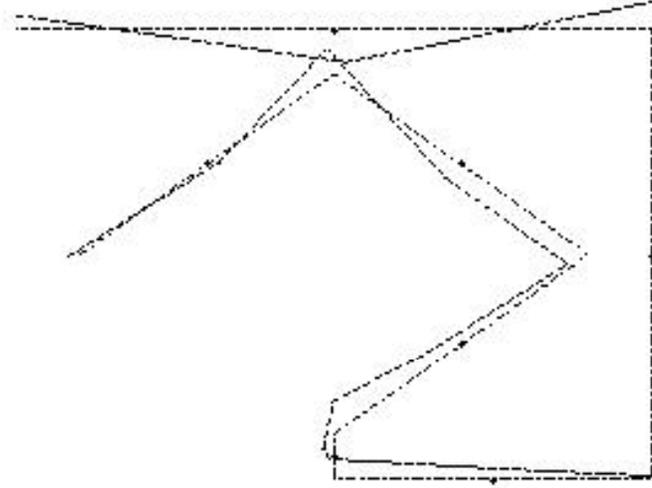
$\mathbf{W}$  matrice de poids :  $w_{ii} = 0$

$$w_{ij} = 0 \quad \text{si } j \notin J(i) \text{ voisinage}$$

$$\sum_j w_{ij} = 0 \quad \text{du point } i$$

$i$  et  $j$  indexent les  $n$  sites : points de la ligne en fonction de l'abscisse curviligne

$w_{ij}$  : degré d'interaction entre le site  $i$  et le site  $j$



- Généralement, seulement du dénombrement
- Deux sources d'erreur : fautes d'identification et erreurs aléatoires
- Modèles paramétriques simples souhaités

# Modèle d'erreurs d'attributs

$P_{rr}$  probabilité pour un attribut d'avoir la valeur correcte  $r$  dans le jeu de données

$P_r$  probabilité pour un attribut d'avoir une valeur incorrecte à la place de la vraie valeur  $r$

$$p_{rr} = (1 - \theta_r) \frac{N_r}{N}$$

$$p_r = \frac{\theta_r}{K} \frac{N_r}{N}$$

$\forall r \in \{0, \dots, K\}$  avec  $\theta_r \in [0, 1[$

$N_r$  nombre d'objets avec valeur  $r$  dans la référence

$N$  nombre total d'objets

- Modèles adaptés aux erreurs rencontrées
- En général, égalité des  $\theta_r$
- Nécessité de reconstruire de l'information implicite pour les fautes d'identification
- Application à l'évaluation d'incertitudes sur des temps de parcours calculés à l'aide d'une base de données géographique (cf. contribution associée)

- Modélisation en points, lignes surfaces proche de la représentation informatique, mais limitée
- Modèle des bases de données vecteur hors du champ actuel de la statistique spatiale : peu d'outils pour modéliser réellement les erreurs sous forme de processus

- Information implicite = information non observée dans les données : approche par modèles à variables cachées
- Statistique paramétrique pas toujours adaptée : nécessité d'utiliser des techniques non paramétriques multi-dimensionnelles