Modélisation des erreurs de position et d'attributs dans les bases de données géographiques

Olivier BONIN

Institut Géographique National, Laboratoire COGIT

Introduction

Les bases de données géographiques contiennent des données géographiques, ou plus généralement des données localisées. Elles représentent une abstraction du monde réel, que l'on appelle terrain nominal.
Techniquement, elles sont composées d'objets qui portent une primitive géométrique : un point (objet ponctuel), une ligne polygonale (objet linéaire), ou un polygone, éventuellement à trou (objet surfacique). Les caractéristiques non spatiales de ces objets sont renseignées par des attributs classiques.

En termes de stockage, les primitives géométriques peuvent être vues comme des attributs particuliers : un triplet *(x,y,z)* pour un point, une suite de points pour une ligne, et une ou plusieurs lignes pour les contours intérieur et extérieur d'une surface. Cette vision ne tient cependant pas compte du fait que les objets géographiques structurent l'espace, et sont en relation les uns avec les autres.

On distingue généralement deux types de relations entre objets : les relations topologiques, et les relations métriques. Les relations topologiques sont liées aux propriétés topologiques des primitives géométriques attachées aux objets : deux polygones peuvent se toucher par exemple. On exploite souvent le graphe associé aux objets linéaires formant un réseau, comme les routes par exemple, pour des applications de parcours de ce réseau : on parle alors de topologie de réseau, ou de carte topologique si on inclut les faces délimitées par les arêtes du graphe. Les relations métriques sont quant à elles liées à la position proprement dite des objets, et à leur organisation spatiale. Deux bâtiments peuvent être proches, éloignés, alignés, etc.

D'un point de vue géographique, on est obligé de s'appuyer sur la géométrie des objets, leurs attributs, et la topologie induite par les réseaux pour pouvoir extraire les relations entre les objets.

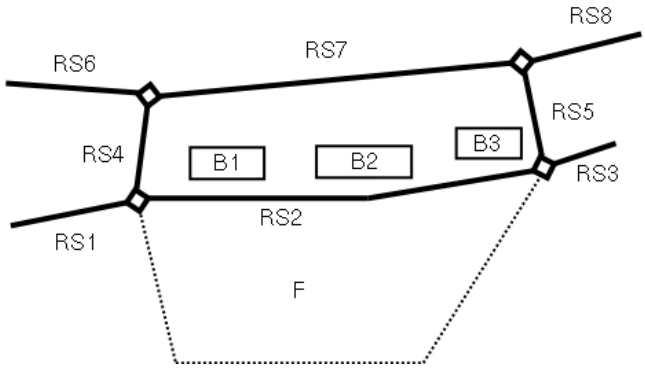

Figure 1 : Exemple de jeu de données

Dans l'exemple de la Figure 1, on est en présence d'un champ F, de trois bâtiments B1, B2 et B3, et d'un certain nombre de tronçons de route numérotés RS1 à RS8. L'exploitation de la carte topologique engendrée par les primitives représentant les routes et les champs permet de déduire que le champ F longe la route : en effet, ces deux objets ont une ligne polygonale en commun, donc une arête commune dans le graphe associé. Le fait que B1, B2 et B3 appartiennent au même îlot est déduit du fait qu'ils dont inclus dans la même face de la carte topologique. En revanche, un analyse plus poussée utilisant leur géométrie doit être menée pour identifier qu'ils sont alignés le long de RS2.

Ces informations (le champ F longe la route, les trois bâtiments appartiennent au même îlot, sont alignés le long de la route, etc.) sont qualifiées d'implicites. Elles ne sont pas présentes directement dans la base de données, mais doivent être recalculées par des méthodes d'analyse spatiale.

Les données géographiques, comme toute source de données, sont par ailleurs entachées d'erreurs. Ces erreurs, heureusement en petit nombre, peuvent porter sur n'importe lequel des aspects évoqués précédemment : la position des objets, leurs attributs, ou les relations entre objets. On évalue ces erreurs lors du contrôle qualité. Pour évaluer la qualité des bases de données géographiques, l'approche naturelle est de mettre au point des modèles d'erreur pour les différents aspects de l'information géographique, en particulier les attributs des objets géographiques, et leurs primitives géométriques. Nous présentons dans cet article des modèles statistiques, pour la plupart développés à l'IGN, et montrons que ces modèles doivent également prendre en compte des informations implicites pour être valides et utilisables. Nous présentons donc successivement des modèles d'erreur de position pour les objets ponctuels et linéaires, et des modèles d'erreurs d'attributs. Nous n'abordons pas les problèmes d'actualité des bases de données géographiques, ni les problèmes de cohérence interne des données propre aux règles de modélisation adoptées pour chaque base de données.

1. Erreurs de position

Les erreurs de position des objets géographiques ont été très tôt abordées, sous la forme de modèles d'erreur de position de points, puis de lignes et de surfaces. Il faut cependant garder en tête le cadre général de la modélisation statistique.

Dans ce cadre, on observe n mesures d'écarts entre les objets de la base de données et leurs références $x_1, x_2, \ldots, x_n$ que l'on suppose être réalisations de n variables aléatoires $X_1, X_2, \ldots, X_n$. Le travail porte alors sur la loi jointe de $(X_1, X_2, \ldots, X_n)$. L'hypothèse classique d'indépendance et d'identité de la distribution des X_i ne peut être avancée dans le cadre géographique qu'après discussion et analyse de la signification physique des écarts x_i mesurés.

1.1. Primitives ponctuelles

Pour les objets représentés par un point, les écarts observés sont simplement les écarts 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. On est dans le cas d'erreurs dues à l'imprécision de la saisie de la base de données, donc d'erreurs de mesure, et l'indépendance de ces erreurs peut être raisonnablement supposée. La loi de ces erreurs est souvent la loi normale, dont l'estimation des paramètres est directe. [4] et [1] présentent des études expérimentales concluant à l'absence de corrélation entre écarts en x et écarts en y, et à la normalité de ces écarts.

Pour ne pas perdre de vue que les données sont localisées, on peut daller la zone contrôlée, et calculer un vecteur d'erreur moyen pour chaque dalle (Figure 2). On peut ainsi mettre en évidence la présence éventuelle de biais spatiaux, et tester statistiquement l'indépendance des séries d'erreurs observées dans chaque dalle.

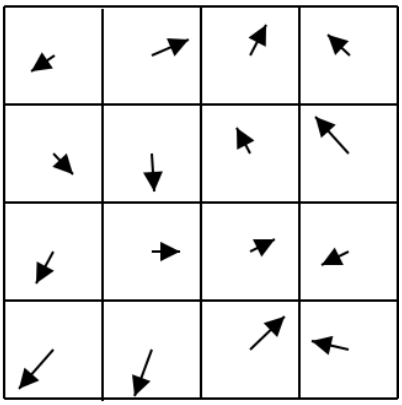

Figure 2 : Vecteurs d'erreur moyens calculés pour chaque dalle de la zone contrôlée

Notons que cette méthode de contrôle de la position de points ne se limite pas aux bases de données géographiques, mais est utilisée également en topographie, photogrammétrie et géodésie.

1.2. Primitives linéaires

Le cas des objets représentés par des lignes polygonales est beaucoup plus complexe, et a donné lieu à différentes approches.

Remarquons tout d'abord que pour des primitives linéaires formant un réseau, telles que celles représentant les routes, on peut se ramener au cas du contrôle ponctuel. En effet, si le réseau est suffisamment dense, le contrôle des intersections de ce réseau suffit à donner une bonne estimation de la position générale du réseau. En revanche, on ne dispose pas d'information sur la qualité de la forme des routes.

L'idée naturelle pour enrichir ce premier contrôle d'un réseau est d'ajouter des points à contrôler. Ces points doivent être définis sans ambiguïté, de manière a être sûr de mesurer un écart entre deux points (base de donnée et terrain nominal) représentant la même réalité géographique. On peut pour cela faire appel à de l'information implicite.

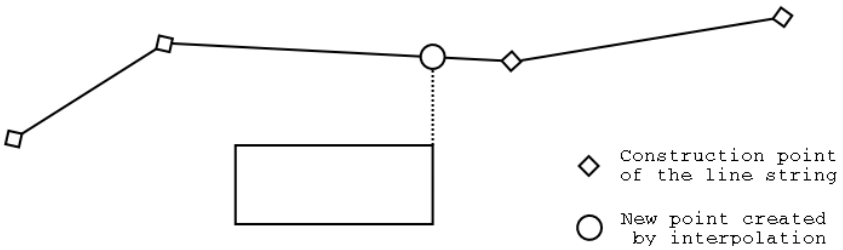

Figure 3 : Point défini implicitement pour le contrôle

La Figure 3 donne un exemple de création de point par explicitation d'information. Le nouveau point crée sur la route dans l'alignement du côté du bâtiment n'appartient pas à la base de données, aussi est-il obtenu par interpolation linéaire sur le segment décrivant la route.

Malheureusement, ces points définis implicitement n'ont pas le même statut que les points définissant la ligne polygonale. En effet, ils ne sont pas résultats de mesures, mais d'interpolations. Les points saisis dans la base de données sont dans la plupart des cas plus précis que ceux interpolés, puisque la base de données donne une représentation schématique de la réalité.

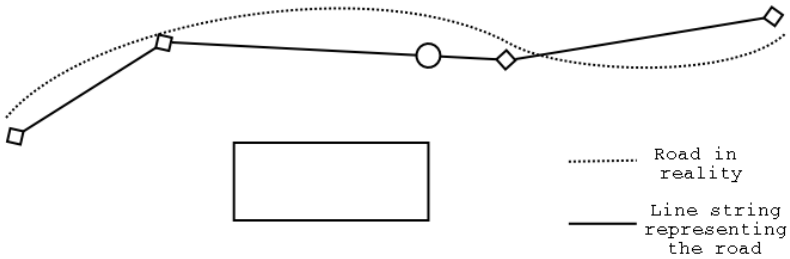

Figure 4 :Erreurs d'interpolation

Par exemple, la figure 4 illustre de manière exagérée la saisie d'une route sous la forme d'une ligne polygonale, pour laquelle les points interpolés sont moins précis que les points saisis. On constate visuellement que les écarts sur la position des points interpolés ne sont pas seulement dus à des erreurs de pointé.

Dès lors, on peut se poser la question de la validité de la loi normale pour modéliser les écarts à la réalité (terrain nominal) de la position des points interpolés. De fait, les estimation de densité de tels écarts montrent généralement des queues de distribution plus lourdes que celle de la loi normale. La deuxième loi de Laplace (exponentielle bilatérale) est un bon candidat dans certains cas. En revanche, l'hypothèse de l'indépendance des erreurs est valide dès que les points recréés n'appartiennent pas au même segment.

Si on veut contrôler des objets linéaires pour lesquels la densité de nœuds est insuffisante, ou pour lesquels il n'est pas possible de recréer des points de contrôle supplémentaires, on peut également mesurer les écarts entre des lignes polygonales (celles du jeu de données et celle d'un référence) à l'aide de mesures adaptées, telles que la distance d'Hausdorff [1] ou de Fréchet [8].

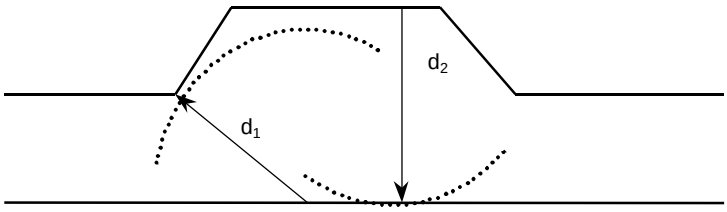

Figure 5 : Calcul de la distance de Hausdorff $d_H = \max(d_1, d_2)$

La distance de Hausdorff entre une ligne L_1 et une ligne L_2 est définie comme $d_H = \max(d_1, d_2)$, d_1 étant la distance de L_1 vers L_2 : $d_1 = \max_{p_1 \in L_1}\left[\min_{p_2 \in L_2}\left[dist(p_1, p_2)\right]\right]$ et d_2 la distance de L_1 vers L_2 : $d_2 = \max_{p_2 \in L_2}\left[\min_{p_1 \in L_1}\left[dist(p_2, p_1)\right]\right]$ (Figure 5).

Lors de la mesure des écarts entre deux lignes polygonales par la distance de Hausdorff, on peut d'attendre à prendre en compte à la fois des erreurs de mesure, et des erreurs d'interpolation si la référence n'a pas un niveau de détail proche de la base de données contrôlée. On trouve dans la littérature (Le Men, cité dans [9]) l'utilisation d'un mélange de loi normale et de loi de Laplace pour modéliser de tels écarts. Notons cependant que ces modèles de mélange sont souvent non identifiables, et qu'il est alors préférable d'utiliser des techniques non paramétriques, comme des estimateurs à noyaux.

L'inconvénient principal de cette approche est que les écarts ont peu de raison d'être indépendants lorsqu'ils s'agit d'écarts successifs le long d'une même route composée de plusieurs tronçons, donc de plusieurs lignes polygonales.

Pour cette raison, nous proposons d'aborder le problème de manière différente selon que l'on s'intéresse à des écarts entre objets ayant une réalité géographique (écarts entre route composées de plusieurs tronçons par exemple), et des écarts le long d'un même objet.

Dans le premier cas, on mesure l'écart par une distance appropriée entre chaque ensemble de tronçons composant la route (ou plus généralement l'objet géographique) et les tronçons de référence dans la réalité. On obtient alors des erreurs de position moyennes entre routes. Ces erreurs peuvent être supposées indépendantes, et on peut estimer des modèles (paramétriques ou non) sans difficulté particulière. On travaille alors à relativement petite échelle. Cette approche nécessite de recréer l'objet route, présent implicitement dans la base de données sous la forme de différents tronçons.

Dans le deuxième cas, on mesure l'écart par une distance appropriée entre chaque tronçon et son tronçon de référence dans la réalité. Alors, les erreurs de position le long d'une route doivent être modélisées par un processus spatial pour rendre compte des corrélations des écarts le long d'une même route. Un modèle non causal, tel que le modèle ARMA bilatéral est alors bien adapté. Un voisinage $J(i)$ pour chaque point i est défini par une matrice de poids $\mathbf{W}$:

$$w_{ii} = 0\,,\ w_{ij} = 0 \text{ si } j \notin J(i) \text{ et } \sum_j w_{ij} = 0\,.$$

Le processus ARMA $\mathbf{X}$ est alors défini par

$$\mathbf{AX} = \alpha + \mathbf{B}\varepsilon$$

avec $\mathbf{A} = \mathbf{I} - \sum_{i=1}^{p} a_i \mathbf{W}^i$, , $\mathbf{B} = \mathbf{I} + \sum_{j=1}^{q} b_i \mathbf{W}^j$ $\mathbf{I}$ étant la matrice identité, ε un bruit blanc, et $\alpha, a_1, ..., a_p, b_1, ..., b_q$ des paramètres inconnus.

Ce modèle est très simple dans notre contexte de lignes car on peut limiter le voisinage à $w_{i,i-1} = w_{i,i+1} = \frac{1}{2}$ dans la plupart des cas. Plusieurs algorithmes existent pour estimer ce modèle, bien répandu en traitement d'images pour modéliser des bruits.

Il est important de préciser qu'un modèle plus simple ne prenant pas en compte les corrélations n'apporte pas d'information significative par rapport au simple contrôle ponctuel de points caractéristiques. En particulier, les simulations par Monte-Carlo d'erreurs de position de lignes nécessitent les corrélations pour éviter l'apparition de problèmes topologiques [5] comme les intersections parasites d'une ligne avec elle-même après bruitage par Monte-Carlo (Figure 6).

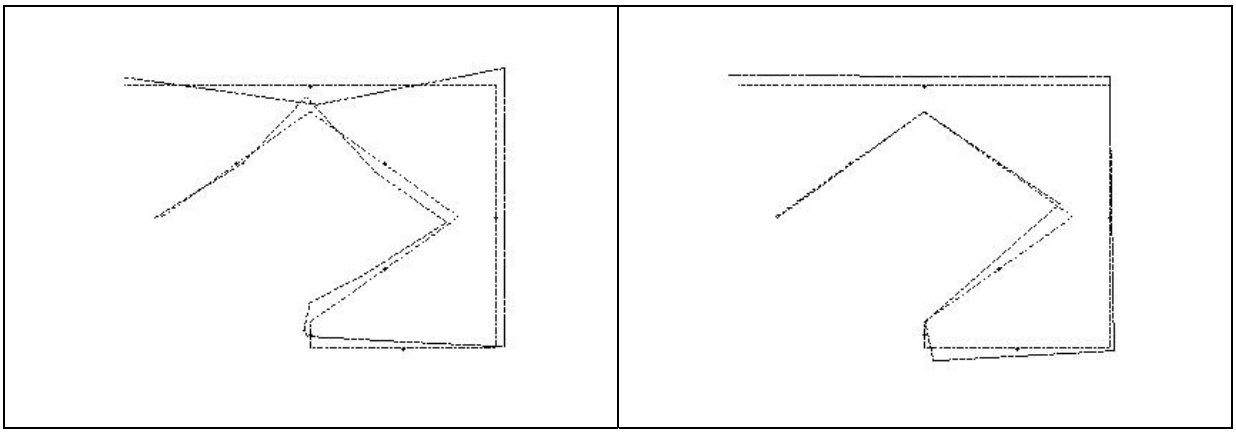

Figure 6 : Ligne originale (trait continu) et ligne bruitée (pointillés) : sans corrélations (gauche) et avec corrélations (droite)

1.3. Synthèse

Nous proposons de synthétiser dans le Tableau 1 le contexte et les particularités des différentes approches pour la modélisation des erreurs de géométrie. Les points importants à souligner sont la nature différente des erreurs de position pour les points des la base de données, et les erreurs pour les points recréés implicitement, ainsi que la nécessité d'introduire des corrélations entre les erreurs le long des lignes.

	Nature des écarts x_i	Origine des écarts	Indépendance des variables Xi ?	Identique distribution ?	Remarques
Objets ponctuels	Erreurs de position de points de la base de données (ex : erreur de positon d'une intersection)	Erreurs de mesure lors de l'acquisition des données	Oui (testable)	Oui (testable)	Estimation par dalles locale et globale, pour pouvoir tester l'absence de dépendance spatiale des erreurs
	Erreurs de position de points recrées implicitement	Erreurs d'interpolation	Oui	Oui, sous réserve d'un degré constant de simplification de la réalité.	Loi normale pas toujours adaptée
Objets linéaires	Ecarts entre objets géographiques représentés par une suite de lignes polygonales (ex : écarts de position entre routes)	Erreurs de mesure et erreurs d'interpolation	Oui, si les lignes sont suffisamment longues	Oui	Estimation non paramétrique
	Ecarts entre lignes polygonales (ex : écarts de position le long d'une route)	Erreurs d'interpolation principalement	Non		Processus spatial non causal de type ARMA

Tableau 1 : Modèle d'erreurs de position des objets géographiques

2. Erreurs d'attributs

Les erreurs d'attributs ont été largement étudiées, mais très peu de modèles statistiques ont été proposés à notre connaissance. On trouve généralement un simple dénombrement de toutes les erreurs rencontrées, éventuellement résumé dans des tableaux croisés.

Comme pour les erreurs de position, il est utile de s'interroger sur la nature de ces erreurs et leur origine. avant de proposer des modèles. Nous avons identifié deux sources principales d'erreur lors de la saisie ou du codage des attributs.

La première source d'erreurs est du domaine de l'erreur purement aléatoire. Un objet peut avoir une mauvaise valeur d'attributs indépendamment de son contexte géographique : par exemple, un opérateur tape le mauvais code, ou choisit la mauvaise valeur dans une liste. Nous proposons pour ce cas deux modèles paramétriques simples, que nous avons utilisé pour étudier les conséquences de telles erreurs sur des applications [6].

La deuxième source d'erreurs est liée à des fautes d'identifications d'objets géographiques. Par exemple, une route est saisie avec un mauvais nombre de chaussées. Dans ce cas, ces erreurs concernent tous les tronçons décrivant la même route, et non pas un tronçon isolé comme dans le cas de l'erreur aléatoire. On peut alors appliquer le même modèle que pour les erreurs aléatoires, mais sur des objets différents : les routes, information implicite qu'il faut reconstruire par analyse de la géométrie des objets, leur topologie, et leurs attributs.

Dans le cas d'erreurs aléatoires, nous proposons deux modèles pour rendre compte de deux cas de figure possibles. Dans le cas d'une erreur dans la saisie d'un code, toutes les autres valeurs possibles de l'attribut ont la même chance d'être choisies : on parle alors de modèle uniforme. Dans le cas de l'erreur dans la saisie d'une valeur dans une liste, ou d'une erreur d'identification, seules les valeurs encadrant la vraie valeur sont susceptibles d'être prises (par exemple, une route à trois voies pourra être codée en route à deux voies ou quatre voies) : on parle alors de modèle tridiagonal.

Pour le modèle uniforme, on note p_{rr} la probabilité pour un attribut d'un objet avec valeur r d'avoir la bonne valeur dans la base de données, et on note p_r la probabilité pour cet attribut d'avoir une autre valeur (donc présence d'une erreur) dans le jeu de données. Ce modèle est paramétré par un $\theta_r \in [0,1[$ et on obtient :

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

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

$\forall r \in \{0,...,K\}$ avec K le nombre de valeurs possibles de cet attribut, N_r le nombre d'objets avec valeur K dans la réalité, et N le nombre d'objets. Notons que 0 dénote l'absence de valeur pour l'attribut. Tous les θ_r peuvent être supposés égaux dans certains cas, simplifiant encore le modèle.

Pour le modèle tridiagonal, on note $p_{rr'}$ la probabilité pour un attribut d'avoir la valeur r' au lieu de la vraie valeur r, et on obtient :

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

$$p_{r(r-1)} = p_{r(r+1)} = \frac{\theta_r}{2}\frac{N_r}{N}$$

$$p_{01} = \theta_0 \frac{N_0}{N}$$

$$p_{(K-1)K} = \theta_K \frac{N_K}{N}$$

avec les même notations. Là encore, les θ_r peuvent être supposés égaux. L'égalité des θ_r est d'ailleurs très simple à tester.

Dans le cas de l'égalité des θ_r, les estimateurs du maximum de vraisemblance du modèle sont respectivement

$$\widehat{\theta} = \frac{N - \sum_{k=0}^{K} N_{kk}}{N}$$

pour le modèle uniforme et

$$\widehat{\theta} = \frac{N_{01} + \sum_{k=1}^{K-1}(N_{k(k-1)} + N_{k(k+1)}) + N_{K(K-1)}}{\sum_{k=0}^{K} N_{kk} + N_{01} + \sum_{k=1}^{K-1}(N_{k(k-1)} + N_{k(k+1)}) + N_{K(K-1)}}$$

pour le modèle tridiagonal.

Comme exemple d'utilisation de ces modèles, nous donnons le cas du contrôle de l'attribut « classement » pour des tronçons de route. Le tableau 2 donne le résumé des erreurs recensées. On lit en colonne les valeurs dans la base de données, et en ligne les valeurs dans la référence. On constate ainsi que dans le jeu contrôlé, 3,25 km de routes non revêtues ont été codée en routes à une voie. L'estimation du modèle tri-diagonal donne $\widehat{\theta} = 0{,}032$ (l'égalité des paramètres ayant été testée). Ce taux d'erreurs assez faible est très représentatif des taux d'erreur généralement rencontrés.

Un modèle prenant en compte erreurs aléatoires et erreurs d'identification est bien sûr possible à construire, mais la difficulté est que les variables ne seront pas indépendantes, du fait des erreurs d'identification affectant tous les tronçons composant une même route par exemple. On se retrouve dans le même cas de figure que pour les erreurs de position le long des tronçons, pour lesquelles il est nécessaire d'utiliser des processus spatiaux.

		Base de données				
		Chemin	Route non revêtue	Route à une voie	Route à deux voies	Route à chaussées séparées
Terrain Nominal	Chemin	98,16	0,75	0	0,26	0
	Route non revêtue	1,25	37,25	3,25	0,66	0
	Route à une voie	1,96	5,42	189,59	3,70	0
	Route à deux voies	0	0,76	12,27	460.89	0
	Route à chaussées séparées	0	0	0	0	29

Tableau 2 : Exemple de résultat de contrôle qualité sur un jeu de données réel (845 km de routes contrôlées)

3. Conclusion et problèmes ouverts

Les modèles que nous présentons pour les erreurs de position et d'attributs dans cet article sont assez simples. Ils ont été utilisés sur des données réelles, et ont fait l'objet d'applications en production pour certains. L'approche proposée est en fait de s'affranchir de la composante spatiale en construisant le plus possible des variables aléatoires indépendantes.

D'un point de vue plus formel, si on veut traiter ces problèmes de dépendances spatiales, il est difficile de se raccrocher directement au domaine de la statistique spatiale (en tous cas dans ces développements classiques [7]) qui recouvre principalement la géostatistique, les champs aléatoire, et les processus ponctuels.

La première approche adaptée aux bases de données géographiques serait de développer des processus indexés par des données géographiques, c'est-à-dire par des points (domaine des processus ponctuels, fort développé), des lignes polygonales et des polygones. Nous proposons un exemple simple d'un tel processus pour des lignes avec l'utilisation des ARMA bilatéraux. De tels processus seraient également d'un grand intérêt pratique pour les simulations géographiques complexes (diffusions par les réseaux géographiques par exemple). De manière optimale, il faudrait concevoir une modélisation à « espace » continu, et non restreinte aux seuls points explicitement représentés dans la base de données. Un état de l'art sur les processus adaptés aux grilles régulières (ARMA spatial causal et non-causal, modèles conditionnels, etc.) montre que la plupart des outils nécessaires sont déjà développés [2]

La deuxième approche, liée à la nécessité de reconstruire de l'information implicite (routes, points interpolés, etc.) consiste à utiliser cette information implicite comme variables cachées dans les modèles, c'est-à-dire des variables non observables. Cette approche n'a pas été suivie pour le moment.

De manière générale, nous avons vu qu'il est important, surtout lorsqu'on traite de données géographiques, de se poser les questions usuelles en modélisation statistique : mes variables sont-elles indépendantes ? Comment puis-je modéliser les dépendances ? Les variables sont-elles de même loi ? Cette loi est-elle simple, paramétrable ou non ? En particulier, des avancées récentes dans le domaine de l'estimation non paramétriques en dimension 2 rendent tout à fait possible l'utilisation de telles techniques [3]. Des tests préliminaires dans ce sens ont donné des résultats satisfaisants dans le cas de réseaux réguliers.

Bibliographie

[1] Abbas I., « Bases de données vectorielles et erreur cartographique. Problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorff », Thèse de doctorat de l'université Paris VII, 1994.

[2] Ah Pine Julien, « Etude des incertitudes lors de calculs de risques d'inondations », Mémoire de DEA de statistique de l'Université Paris VI, encadré par Olivier Bonin, 2003.

[3] Biau, G., « Spatial kernel density estimation », *Mathematical Methods of Statistics*, 2004.

[4] Bolstad P.V, Gessler P. & Lillesand T.M, « Positional uncertainty in manually digitized map data. » *International Journal of Geographical Information Systems*, 4(4), 1990.

[5] Bonin O, « New Advances in error Simulation in Vector Geographical databases », actes du 4ème symposium international *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), 2000.

[6] Bonin O, « Large Deviation Theorems for Weighted Sums Applied to a Geographical Problem », *Journal of Applied Probability*, Volume 39 n°2, 2002.

[7] Cressie, « Statistics for spatial data », Wiley, 1993.

[8] Devogele T., Bonin O., « Using distances for linear accuracy measurements », actes du 4ème symposium international *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*, Heuvelink G.B.M and Lemmens M.J.P.M (Eds), 2000.

[9] Vauglin F. « Modèles statistiques des imprécisions géométriques des objets géographiques linéaires », Thèse de doctorat de l'université de Marne-La-Vallée, 1997.