THÈSE de DOCTORAT de l'UNIVERSITÉ PARIS 6

Spécialité MATHÉMATIQUES

Option : Statistique

Présentée par

Olivier Bonin

Pour obtenir le grade de DOCTEUR de l'UNIVERSITÉ PARIS 6

Sujet de la thèse :

Modèle d'erreurs dans une base de données géographiques et grandes déviations pour des sommes pondérées ; application à l'estimation d'erreurs sur un temps de parcours

Soutenue le 1er mars 2002

devant le jury composé de :

Monsieur Gérard d'Aubigny	Rapporteur
Monsieur Michel Broniatowski	Examinateur
Monsieur Paul Deheuvels	Examinateur
Monsieur Hervé Le Men	Examinateur
Monsieur Thomas Mikosch	Rapporteur
Monsieur Daniel Pierre-Loti-Viaud	Directeur de thèse
Monsieur Gabriel Ruget	Président

Table des matières

Table des figures

Liste des tableaux

A Alice

Remerciements

Ce travail est le fruit d'une collaboration entre le laboratoire COGIT de l'Institut Géographique National et le Laboratoire de Statistique Théorique et Appliquée de l'Université Paris VI, sous la direction de Monsieur le professeur Daniel Pierre-Loti-Viaud.

Je suis particulièrement reconnaissant à Monsieur le professeur Gabriel Ruget, directeur de l'ENS, de m'avoir fait l'honneur de présider le jury, et d'avoir porté un regard extérieur et éclairé sur mes travaux.

Je suis très reconnaissant à Messieurs les professeurs Gérard d'Aubigny et Thomas Mikosch d'avoir accepté d'examiner mes travaux, et d'être les rapporteurs de cette thèse. Je remercie tout particulièrement Thomas Mikosch d'avoir affronté un texte partiellement en français, dans un temps record.

Je remercie également Monsieur le professeur Paul Deheuvels, directeur du LSTA, d'avoir donné la chance à un ingénieur en géographie scientifique de faire un DEA, puis une thèse, et de l'avoir encouragé dans cette voie.

Je remercie Monsieur le professeur Michel Broniatowski et Monsieur Hervé Le Men, directeur technique adjoint de l'IGN, d'avoir accordé un grand intérêt à mes travaux, et de m'avoir ouvert un certain nombre de perspectives par leurs questions.

Je remercie enfin Monsieur le professeur Daniel Pierre-Loti-Viaud, directeur adjoint du LSTA, d'avoir encadré mes recherches. Il a fait preuve à mon égard d'une grande disponibilité, d'un conseil scientifique efficace, et m'a beaucoup apporté par sa grande rigueur et son esprit critique. Je lui dois aussi de m'avoir encouragé à persévérer dans un travail appliqué faisant ressurgir des thématiques déjà bien explorées.

Je remercie naturellement les membres du laboratoire COGIT, et en particulier sa directrice Anne Ruas, et Vincent Beauce qui a partagé mon bureau.

J'ai une pensée toute particulière pour ma femme Alice et mon fils Gonzague, qui m'ont soutenu, encouragé et distrait pendant ces années de thèse.

Je remercie mes parents, qui m'ont toujours encouragé dans la voie scientifique.

Je dédie cette thèse à Alice.

Introduction

Les Bases de Données Géographiques contiennent sous forme numérique toute l'information géographique présente dans les cartes traditionnelles, enrichie d'informations complémentaires. Elles offrent un large éventail d'applications, et sont le support naturel de l'analyse de problèmes faisant intervenir des données localisées. La gestion de ressources naturelles, des risques naturels, de flottes de taxis ou de camions, la détermination du lieu d'implantation d'un hôpital en termes de facilité d'accès sont des exemples d'applications nécessitant une analyse géographique.

La qualité des résultats d'une application géographique dépend de la modélisation proposée pour résoudre le problème, des algorithmes utilisés, mais également de la qualité des bases de données géographiques employées. Cette dernière est généralement mesurée par le producteur de données, et résumée sous forme d'indicateurs de qualité, mais ces indicateurs ne permettent pas directement de déterminer l'impact de la qualité des données sur les résultats d'une application géographique. Cette thèse propose des méthodes pour mesurer et prévoir l'influence de la qualité des données géographiques sur une application de calcul d'itinéraires. L'application de calcul d'itinéraires a été retenue car elle est de plus en plus présente dans la vie courante, et présente une réelle dépendance aux éventuels problèmes de qualité des données.

Nous avons adopté deux approches pour résoudre le problème posé : une par simulation, et une par le calcul. Pour ces deux approches, nous avons tout d'abord proposé des modèles statistiques d'erreurs dans les bases de données géographiques, pour pouvoir résumer les différents indicateurs fournis par le contrôle qualité de production (chapitre I). Ces modèles ont été testés et validés sur les bases de données de l'IGN. Nous pouvons ainsi caractériser par quelques paramètres synthétiques la qualité d'un jeu de données.

Nous avons ensuite mené une étude par simulation pour répondre au problème posé (chapitre II). Nous avons construit des algorithmes permettant d'introduire un bruit réaliste et contrôlé dans une base de données, écrit un logiciel de calcul d'itinéraires, et étudié numériquement la dégradation des résultats fournis par le logiciel lors de l'utilisation de données de qualité décroissante.

L'étude par simulation étant difficile à mettre en place (écriture de nombreux algorithmes), et très consommatrice de temps de calcul, nous avons étudié des modèles simplifiés de calcul de temps de parcours, et avons déterminé la probabilité que l'erreur relative en temps dépasse un certain seuil fixé, en fonction du taux d'erreur présent dans la base de données utilisée. Nous avons étudié pour un itinéraire type cette probabilité d'erreur en fonction des erreurs d'attributs dans la base de données (chapitre III), et en fonction des erreurs d'attributs et de géométrie (chapitre IV). Nous avons ensuite ajouté un aléa supplémentaire sur l'itinéraire emprunté (chapitre V). Ces études utilisent des développe-

ments de grandes déviations pour des sommes pondérées de variables aléatoires discrètes. L'utilisation de résultats asymptotiques impose cependant d'utiliser des développements exacts pour obtenir une bonne précision d'approximation. Nous illustrons la qualité des développements en les confrontant sur des exemples à des estimations des mêmes probabilités par simulation de Monte-Carlo. L'estimation des probabilités d'intérêt fait intervenir des sommes pondérées de variables treillis i.i.d (chapitre III), de variables treillis non équidistribuées (chapitre IV), et enfin des sommes aléatoires de variables treillis pondérées (chapitre V).

Les contributions de cette thèse sont de plusieurs ordres :

- les modèles statistiques d'erreurs dans les bases de données géographiques, en particulier d'attributs, et les méthodes de bruitage associées, sont d'un grand intérêt pratique, de par leur nature synthétique (résumé de la qualité en quelques paramètres), et les applications en découlant (études par simulation) ;
- l'estimation des probabilités d'erreurs sur des temps de parcours à l'aide de développements de grandes déviations est une application originale des techniques de grandes déviations, et la précision des approximations obtenues est remarquable ;
- les développements de grandes déviations utilisés pour des sommes pondérées sont des extensions nouvelles des résultats existants au cas des variables treillis, treillis non équidistribuées, et des lois composées pondérées.

Les développements de grandes déviations remontent aux travaux de Chernoff [Che52], et ont été largement étudiés depuis, sur la base de l'article de Bahadur et Ranga Rao [BR60]. Le problème des grandes déviations est le suivant. Étant donnée une suite de variables $X_1, X_2, \ldots$, que peut-on dire de la probabilité :

$$p_n = P\left(\sum_{i=1}^{n} X_i > c_n + a_n\right),$$

avec $c_n = E\left(\sum_{i=1}^{n} X_i\right)$, et $a_n \to \infty$ quand $n \to \infty$? La probabilité p_n étant de l'ordre de grandeur exponentiel, on peut en première approximation étudier $\log p_n$, quand $n \to \infty$ (nous parlerons de *développements logarithmiques*). Dans le cas d'une suite de variables i.i.d, et pour $a_n = na$, Chernoff a établi que

$$\sqrt{n} \log p_n \to -\log \rho,$$

en donnant l'expression de ρ. Bahadur et Ranga Rao ont obtenu le développement suivant :

$$p_n = \frac{\rho^n}{(2\pi n)^{1/2}} b_n (1 + o(1)),$$

avec ρ identique au coefficient de Chernoff, et $\log b_n = O(1)$ quand $n \to \infty$ (nous parlerons de *développement au premier ordre exact*). Le gain en précision apporté par le terme exact

est important pour une utilisation non asymptotique, en particulier quand ρ est proche de 1 ou quand n est grand.

Différents travaux s'écartent du cas i.i.d. On trouve dans [CS85] et [CS93] des extensions dans des cas non i.i.d, et dans [Ste78] une résolution dans le cas multidimensionnel. Des extensions existent également pour des objets probabilistes plus complexes, ou pour $P(S_n \in nA)$ avec A borélien arbitraire [BB].

En modélisant et étudiant la probabilité que l'erreur relative sur le temps de parcours d'un itinéraire type dépasse un seuil fixé, en fonction des erreurs d'attributs présentes dans la base de données utilisée, nous avons fait apparaître des développements de grandes déviations pour des sommes pondérées de variables i.i.d discrètes (chapitre III). L'écart à l'équidistribution, contrairement à l'écart à l'indépendance, a été peu abordé dans la littérature. Book ([Boo73]) a obtenu un théorème de type Chernoff pour des sommes de variables pondérées, mais n'a obtenu un théorème de type Bahadur et Ranga Rao ([Boo72]) que dans le cas de variables pondérées possédant une densité. Or, la contrainte du n fixé relativement faible rend insuffisamment précis les théorèmes de type Chernoff, comme le montrent les applications numériques du chapitre III.

Nous avons donc étendu les résultats de Book ([Boo72]) pour des variables quelconques, y compris treillis, en utilisant une méthode de preuve légèrement différente. Nous précisons dans la dernière section du chapitre III en quoi notre méthode diffère de celle de Book.

Nous avons obtenu sous des conditions de régularité des poids le résultat suivant, pour des variables définies sur une grille. Soit $X = X_1, X_2, \dots$ une suite de variables i.i.d non dégénérées centrées définies sur une grille, soit $\{a_{nk} : 1 \leqslant k \leqslant n, 1 \leqslant n < \infty\}$ un tableau triangulaire de réels positifs vérifiant $\sum_{k=1}^{n} a_{nk}^2 = 1$, et soit c un réel strictement positif. Notons $S_n = \sum_{k=1}^{n} a_{nk} X_k$ et $A_n = \sum_{k=1}^{n} a_{nk}$, et supposons que $a_{nk}/a_{nl} \in \mathbb{Q}$ pour tout k et pour tout l, ce qui assure que S_n est définie sur une grille. Nous supposons aussi que $E(X^2) = 1$, notons $F(x)$ la fonction de répartition de X, et $\phi(t) = E(e^{tX})$ sa fonction génératrice des moments. Les deux conditions suivantes sont classiques ([Boo73]) :

Condition I. *Il existe deux réels α et θ, $0 < \alpha \leqslant 1$, $0 < \theta \leqslant 1$, tels que, pour tout n assez grand, au moins αn des a_{nk} sont supérieurs ou égaux à $\theta\sigma_n$, où $\sigma_n = \max\{a_{nk} : 1 \leqslant k \leqslant n\}$.*

Condition II. *$\phi(t)$ est finie sur $\mathcal{I} \supseteq (-B, B)$, pour un $B > 0$, la fonction $Q = \phi'/\phi$ prend la valeur $\frac{c}{\alpha\theta}$ en un point, et $B_0 = \theta^{-1} Q^{-1}(\frac{c}{\alpha\theta}) \in \mathcal{I}$.*

Posons $Y_{nk} = a_{nk} X_k - c a_{nk}$, $H_{nk}(y) = F(y a_{nk}^{-1} + c)$, $\phi_{nk}(h) = e^{-hca_{nk}} \phi(h a_{nk})$. Soit $\bar{H}_{nk}$ définie par $d\bar{H}_{nk}(y) = \frac{e^{hy}}{\phi_{nk}(h)} dH_{nk}(y)$ pour tout $0 < h < B\sigma_n^{-1}$, et $\bar{Y}_{n1}, \bar{Y}_{n2}, \dots$ une suite de variables aléatoires indépendantes distribuées selon $\bar{H}_{nk}$. Posons $\bar{S}_n = \sum_{k=1}^{n} \bar{Y}_{nk}$, et $\bar{H}_n(y) = P(\bar{S}_n \leqslant y)$.

Nous obtenons le théorème suivant :

Théorème. *Supposons I et II. Soit h_n solution de l'équation $E(\bar{S}_n(h_n)) = 0$. Posons $\bar{\sigma}_n^2 = \operatorname{Var} \bar{S}_n(h_n)$ et d_n le pas de la grille de S_n. Alors, quand $n \to \infty$,*

$$P(S_n \geqslant cA_n) = \frac{1}{\sqrt{2\pi}} \frac{d_n e^{-h_n d_n}}{\bar{\sigma}_n (1 - e^{-h_n d_n})} e^{-h_n c A_n} \left[\prod_{k=1}^{n} \phi(h_n a_{nk}) \right] (1 + o(1)).$$

L'apport du développement au premier ordre exact est le terme

$$\frac{1}{\sqrt{2\pi}} \frac{d_n e^{-h_n d_n}}{\bar{\sigma}_n (1 - e^{-h_n d_n})}$$

qui permet à l'approximation d'avoir une précision suffisante. Dans notre contexte applicatif, sur un exemple avec $n = 30$, le développement exponentiel donne une probabilité de 50% et le développement au premier ordre exact de 12%, à comparer aux 8% obtenus par méthode de Monte-Carlo.

La figure 0.1 représente la probabilité estimée au premier ordre exact que l'erreur de temps de parcours calculé pour un itinéraire type dépasse un seuil fixé η, en fonction du taux d'erreur θ présent dans la base de données (en ligne continue). Nous comparons cette probabilité estimée par un développement au premier ordre exact à celle estimée par un développement logarithmique (ligne tiretée), et à la probabilité estimée par une méthode de Monte-Carlo (ligne mixte). Le développement logarithmique est inadapté à ces conditions d'application.

Nous avons ensuite généralisé ce résultat pour pouvoir aborder l'étude des répercussions des erreurs de géométrie conjointement aux erreurs d'attributs sur les temps de parcours calculés (chapitre IV). Cette extension s'écrit

$$P(S_n \geqslant cA_n) = \frac{1}{\sqrt{2\pi}} \frac{d_n e^{-h_n d_n}}{\bar{\sigma}_n (1 - e^{-h_n d_n})} e^{-h_n c A_n} \left[\prod_{k=1}^{n} \phi_{nk}(h_n a_{nk}) \right] (1 + o(1)),$$

les $\phi_{nk}(t)$ étant des fonctions de t et de a_{nk}.

Enfin, nous avons abordé le cas où l'itinéraire n'est pas fixé, mais introduit un aléa supplémentaire (chapitre V). Le cas du tirage au sort d'un itinéraire parmi tous les itinéraires possibles peut être résolu avec des développements pour des mélanges de lois composées, qui figurent par exemple dans [Aas85]. On utilise alors une approximation de la probabilité $P(Y > y)$ quand $y \to \infty$, Y étant un mélange fini de lois de Poisson composées :

$$Y = \sum_{k=1}^{r} \sum_{i=1}^{N_k} X_{ki},$$

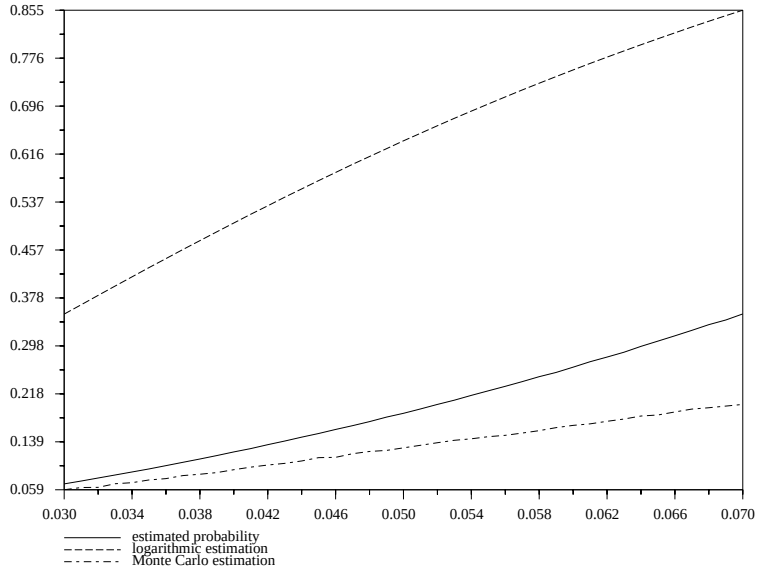

FIG. 0.1.: Probabilité de dépassement du seuil $\eta = 5\%$ en fonction de θ

les X_{ki} étant des variables aléatoires discrètes indépendantes et de loi P_i et les $N_k, 1 \leqslant k \leqslant r$ suivant des lois de Poisson de paramètres $\lambda_k = \lambda p_k$, avec $\sum_{k=1}^{r} \lambda_k = \lambda$. Et on obtient alors le développement quand $y \to \infty$:

$$P\left(\sum_{k=1}^{r}\sum_{i=1}^{N_k} X_{ki} \geqslant y\right) = \frac{\phi_c(h)e^{-hy}}{\sqrt{2\pi}\sigma_c(1-e^{-h})}(1+o(1)), \tag{0.1}$$

avec $\phi_c(t) = E(e^{tY})$, et h et σ_c connus. Ce résultat est aussi valide pour l'asymptotique $\lambda \to \infty$.

Un autre cas est celui d'un trajet connu, mais de destination probabiliste (un automobiliste se rendant dans le sud de la France par exemple). Il se résout à l'aide d'une extension des développements de grandes déviations pour des lois composées au cas des lois composées pondérées.

Soit $X = X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d non dégénérées, soit $a_1, a_2, \ldots$ une suite de réels positifs, avec $\sigma = \max\{a_k\} < \infty$, et soit c une constante réelle positive. Nous considérons $S_N = \sum_{i=1}^{N} a_i X_i$, où N est une variable aléatoire discrète suivant une loi de Poisson de paramètre λ, nous notons $A_N = \sum_{i=1}^{N} a_i$, et étudions le comportement de la probabilité $P(S_N > cA_N)$ quand $E(N) \to \infty$. Nous supposons que $E(X) = 0$ et que $E(X^2) = 1$. Nous notons $F(x) = P(X \leqslant x)$ la fonction de répartition de X, $\phi(t) = E(e^{tX})$ la fonction génératrice des moments de X, et $\phi_{S_N}(t) = E(e^{tS_N})$ la fonction génératrice des moments de S_N. Nous notons $\xi(t) = E(t^N)$. Soit $Y = S_N - A_N$, et $\phi_Y(t) = E(e^{tY})$.

Condition I' Il existe α et θ avec $0 < \alpha \leqslant 1$, $0 < \theta \leqslant 1$, tels que pour tout n, au moins αn des a_k, $1 \leqslant k \leqslant n$, sont supérieurs ou égaux à $\theta\sigma$.

Condition III Les a_i sont tels que S_N est une variable treillis de pas d.

Nous effectuons enfin la transformation exponentielle suivante, en notant H_{c0} la fonction de répartition de Y :

$$\frac{dH_{ch}}{dH_{c0}}(x) = \frac{e^{hx}}{\phi_Y(h)},$$

pour $0 < h < B\sigma^{-1}$. Soit Y_h une variable aléatoire distribuée suivant H_{ch}.

Nous pouvons énoncer le théorème suivant :

Théorème. *Supposons que les conditions I' et II sont vérifiées. Soit $c > 0$ fixé et h solution de l'équation $E(Y_h) = 0$. Posons $s^2 = \mathrm{Var}(Y_h)$ et $\mu_3 = E(Y_h^3)$. Supposons de plus que $\mu_3/s^3 \to 0$. Alors, quand $E(N) \to \infty$, $s \to \infty$ et*

$$P(S_N > cA_N) = \frac{1}{\sqrt{2\pi}}\frac{1}{sh}(\phi_Y(h) - p_0)(1 + o(1))$$

lorsque X n'est pas treillis, et

$$P(S_N > cA_N) = \frac{1}{\sqrt{2\pi}}\frac{de^{-hd}}{s(1 - e^{-hd})}\phi_Y(h)(1 + o(1)),$$

lorsque X est treillis et la condition III est vérifiée, avec d pas de la grille de S_N.

Cette thèse a donné lieu à quatre publications dans le domaine de l'Information Géographique : une revue internationale [Bon00a], une conférence avec comité de lecture [Bon98], et deux conférences [Bon99] [Bon00b], et deux articles de revues de Mathématiques : [Bon01] [Bon02]. De plus, deux articles présentant les résultats du chapitre V sont en préparation.

Chapitre I.

Qualité des Bases de Données géographiques et applications géographiques

L'objet de ce chapitre est de présenter les bases de données géographiques vectorielles telles qu'elles sont constituées à l'IGN, et de définir la notion de qualité des données géographiques. Nous détaillons en particulier les indicateurs de qualité mesurés par les producteurs de données géographiques. Nous proposons un modèle d'erreur de classification et d'attributs pour pouvoir manipuler la qualité des données géographiques. Nous présentons ensuite un point de vue «utilisateur» sur la qualité, en nous intéressant à l'impact de la qualité sur les résultats d'une application géographique.

1. Qualité des données géographiques

1.1. Information géographique

On appelle *information géographique* toute information ayant un sens géographique, c'est-à-dire décrivant le monde et structurant le paysage. La notion d'information géographique est proche mais distincte de celle d'information localisée, également appelée information géoréférencée, c'est-à-dire liée à une position à la surface de la terre. L'information présente dans une carte est de l'information géographique, alors que des données issues d'un recensement sont localisées, mais ne sont pas géographiques. L'information géographique est aujourd'hui présente sous forme numérique dans des bases de données, de contenu et de structuration très variables. Ces bases de données géographiques sont souvent enrichies d'information localisée.

1.1.1. Objets et zones

On distingue en général deux approches différentes pour structurer l'information géographique, l'approche *objet* et l'approche *zone*. Notons que ces approches sont souvent confondues avec les notions d'information vectorielle et maillée (raster) dans les systèmes d'information géographique, mais c'est là restreindre un peu les notions d'objet et de zone.

Dans une conception objet de l'information géographique, l'univers est décomposé en objets possédant une géométrie souvent simple (des points, des lignes et des surfaces), des attributs pour porter l'information descriptive, et des relations topologiques avec les autres objets. On peut penser aux objets *château d'eau*, *route* et *champ* comme exemples d'objets ponctuels, linéaires et surfaciques.

Dans une conception zone, l'univers est représenté par une carte segmentée en différentes zones, chacune des zones portant des attributs. Les zones décrivent des propriétés géographiques, sans recourir à la notion d'objets. Un exemple type est une carte géologique, où les zones représentent la structure géologique du terrain, et n'ont pas forcément d'appui sur des objets géographiques.

1.1.2. Vectoriel et maillé

L'information géographique est maintenant stockée sous forme de données numériques dans des bases de données. Pour gérer le caractère localisé des données, et les relations to-

pologiques, des logiciels spécifiques sont nécessaires. De tels logiciels sont appelés *Systèmes d'Information Géographique* (SIG). Il existe deux grandes catégories de SIG, les SIG *vectoriels* et les SIG *maillé (raster)*, selon le type de représentation que l'on choisit (objets ou zones). De plus en plus de logiciels sont capables de gérer les deux types de représentation simultanément.

Un SIG vectoriel manipule des bases de données géographiques vectorielles. Les éléments de la base de données sont des objets, possédant une position, une forme et des attributs descriptifs. Nous parlons de géométrie de l'objet pour désigner sa position et sa forme, et de sémantique de l'objet pour désigner ses attributs. La géométrie des objets dans les bases de données géographiques est décrite à l'aide de primitives géométriques, généralement le point, le segment, et la surface. Une route par exemple dispose d'une géométrie et d'une sémantique. La géométrie d'une route est sa représentation à la surface de la terre, et sa sémantique est l'information complétant sa description. Une représentation usuelle pour une route est une ligne constituée de plusieurs segments dont la géométrie est stockée de façon vectorielle (une suite de points dont les coordonnées sont exprimées dans un système de référence). Chacun des segments porte des attributs décrivant les caractéristiques du tronçon de route associé. De la même façon, une surface est décrite par les segments constituant le polygone la représentant.

Un SIG maillé adopte une représentation semblable à celle des images sur écran. L'information géographique est représentée par une matrice, constituée d'éléments qui forment un pavage (les pixels dans une image). Chaque élément de la matrice porte un ou plusieurs attributs. Une image satellite, sur laquelle on renseigne l'occupation du sol par exemple, est le genre de données manipulées par les SIG maillé. C'est une image numérique portant des attributs qui complètent les seules données de radiométrie.

1.2. Cadre de l'étude

Nous nous intéressons dans la suite uniquement à l'approche objet de l'information géographique, et donc à des données vectorielles. En effet, ce type de représentation est celui principalement choisi à l'IGN, en particulier pour sa base de données de référence (la BDTopo). De plus, les données vectorielles offrent un champ d'application très large du fait de leur richesse géométrique, topologique et sémantique.

Nous allons travailler sur des bases de données vectorielles. Ces bases de données contiennent les objets géographiques, avec une *géométrie* implémentée de façon vectorielle, une *topologie* et une *sémantique*.

Typiquement, un objet ponctuel aura des coordonnées dans un ou plusieurs systèmes de référence, un objet linéaire est représenté par un ou plusieurs segments de droite, avec un sommet initial, un sommet final, et des points intermédiaires, et un objet surfacique par un polygone, pour ce qui est de la géométrie. Un aspect intéressant des bases de données vectorielles est la *topologie*, calculée à partir des primitives géométriques. On peut

déterminer aisément, grâce à la modélisation vectorielle de la géométrie, des relations d'inclusion, d'adhérence, de connexion, etc. Cette topologie est très importante dans nombre d'applications, puisque c'est elle qui va déterminer la façon dont on peut se déplacer dans l'espace (explorer la carte) en respectant les contraintes géographiques.

Nous allons particulièrement étudier l'aspect sémantique des bases de données géographiques. Nous entendons par sémantique le sens porté par l'information géographique (les objets géographiques dans notre cas), et donc l'information portée par les attributs décrivant ces objets (toute l'information hors la géométrie).

Les attributs ont plusieurs natures possibles ([DF97]). La terminologie retenue en géographie est différente de la terminologie mathématique. Dans le contexte de l'information géographique, on pourra faire la distinction entre attributs *qualitatif* et *quantitatif*, selon que l'on peut évaluer numériquement les valeurs de cet attribut ou non. Dans tous les cas, on pourra également faire la distinction entre attribut énuméré et non énuméré si l'ensemble de ses valeurs est un ensemble infini. Donnons un exemple d'attribut de chaque type. La population d'une ville est un attribut quantitatif non énuméré. Le nombre de chaussées sur une route est un attribut quantitatif énuméré. Le nom d'une commune (toponyme) est un attribut qualitatif non énuméré. Le type d'une route (nationale, départementale, etc.) est un attribut qualitatif énuméré.

Remarquons ici que la nature d'un attribut peut varier selon le traitement que l'on souhaite lui appliquer. Notamment, un attribut non énuméré peut toujours être vu comme un attribut énuméré, en regroupant ses valeurs en un nombre fini de classes.

1.3. Qualité d'un base de données géographiques

Les bases de données géographiques sont de qualité variable — en prenant qualité dans son acception la plus courante — dépendant de leur producteur, du soin apporté à leur constitution, des sources de données utilisées, etc. Pour pouvoir parler de qualité d'une base de données géographiques, il faut donc d'abord définir le terme qualité. C'est une tâche un peu délicate ici, car il est difficile de séparer le *produit* «base de données géographiques» des *services* qui l'accompagnent (livraison de mises à jour, extraction de thèmes particuliers dans la base, conversions de formats de données, etc.)

L'ISO (International Standardisation Organisation) donne la définition suivante du terme qualité :
«Ensemble des propriétés et caractéristiques d'un produit, ou d'un service qui lui confère l'aptitude à satisfaire des besoins exprimés ou implicites.» (norme ISO 8402)

Cette définition suppose la donnée d'un produit et d'une application de ce produit, pour pouvoir mesurer la qualité. Ce n'est pas le cas des bases de données de l'IGN, généralistes, destinées à répondre à un ensemble de besoins. Ces bases peuvent être déclinées sous un grand nombre de produits dérivés, et sont de qualité variable selon les zones (zones plus difficiles à cartographier, moins souvent mises à jour, par exemple). En outre, les besoins

des utilisateurs (exprimés ou implicites) sont très différents les uns des autres, y compris dans la même classe d'applications.

On distingue donc deux étapes dans l'évaluation de la qualité d'une base de données : d'une part l'évaluation de l'adéquation des spécifications de la base à un besoin exprimé ou implicite, et d'autre part l'évaluation de la conformité de la base à ses spécifications. Nous nous concentrons sur la deuxième étape, et nous considérons à partir de maintenant que le terme *qualité* d'une base de données géographiques décrit la conformité de la base à ses spécifications de produit. La mesure et le contrôle de la qualité d'une base de données, dans cette acception de qualité, sont donc indépendants des applications utilisant cette base.

La qualité d'une base de données est très complexe, car les données géographiques localisées doivent être correctement référencées sur la terre, leurs caractéristiques doivent être correctement décrites, et leur agencement les unes par rapport aux autres doit également être conforme à la réalité. La qualité d'une base de données géographique est donc évaluée, vu sa complexité, en composantes distinctes. Un consensus est apparu pour proposer un certain nombre de critères ; nous en retiendrons cinq, comme le préconise la Direction Technique de l'IGN. Ce sont les suivants, que nous détaillons plus loin :

- la *généalogie* (lineage) ;
- l'*actualité* (temporal accuracy) ;
- la *cohérence logique* (logical consistency) ;
- la *précision géométrique* (positional accuracy) ;
- la *précision sémantique* et l'*exhaustivité* (attribute accuracy).

Nous allons tout d'abord donner la définition de la notion de *terrain nominal*, utile dans la suite.

1.4. Terrain nominal

Le *terrain nominal* est défini ainsi :
«Image de l'univers à travers le filtre constitué par l'ensemble des spécifications de la base de données géographiques.» [DF97]

C'est donc le contenu parfait et idéal de la base de données, sans fautes ni omissions, parfaitement à jour, respectant toutes les spécifications. Les spécifications définissent le contenu de la base de données, et la façon dont sont représentés les objets. Donnons deux exemples tirés des spécifications de la BDTopo pour illustrer la notion de terrain nominal. «Les arbres isolés ne sont représentés que s'ils portent un nom», ce qui définit le contenu de la base de données pour ce thème. «Une rivière de largeur inférieure à 7,5 mètres est représentée par une ligne, et par une surface dans le cas contraire». Le même objet géographique, une rivière, aura donc deux représentations possibles dans la base, selon sa largeur courante ([Vau97]).

Il faut noter que, quelle que soit la précision des spécifications de contenu, il n'y a pas unicité du terrain nominal. En effet, on ne peut pas prévoir tous les cas particuliers, et les spécifications peuvent être imprécises. Les surfaces représentant des forêts, par exemple, ont forcément des limites dépendant de l'appréciation de l'opérateur de saisie. On introduit donc la notion d'incertitude du terrain nominal, qui est l'écart qui peut exister entre deux occurrences possibles du terrain nominal de la même réalité, avec des spécifications identiques.

Le terrain nominal est une abstraction ; en général, il n'est pas observable. Or, par définition, il sert de référence quand on cherche à évaluer la qualité d'un jeu de données, puisque c'est le contenu de la base exactement accordé à ses spécifications. Évaluer la qualité d'un jeu de données est donc comparer ce jeu de données au terrain nominal. Le nouveau problème qui apparaît est de construire ou d'estimer ce terrain nominal.

Dans la plupart des applications, on estime le terrain nominal par le contenu d'une base de données plus précise, et on compare la base dont on cherche à évaluer la qualité à cette base de référence. Il faudra mettre en correspondance les objets des deux bases qui représentent la même réalité ; en géographie, on appelle cette opération : «appariement». On peut aussi acquérir de façon indépendante, par exemple par saisie directe sur le terrain, un jeu de données plus précis, qu'on espère plus proche du terrain nominal.

On pourra rarement comparer les jeux de données entiers, pour des raisons de coût. On procède donc à la sélection d'un échantillon, qu'on suppose représentatif de la base entière, et on travaille sur cet échantillon. Cette opération d'échantillonnage n'est pas sans incidence sur le contrôle qualité ; on peut cependant contrôler ses répercussions moyennant certaines conditions. Un étude sur ce sujet a été menée par l'intermédiaire d'un stage [Gon01], et le problème de définir une méthode d'échantillonnage adapté aux données géographiques est encore largement ouvert. Il dépasse le cadre de ce travail.

1.5. Composantes de la qualité

Nous allons détailler les composantes de la qualité citées plus haut, sans détailler les indicateurs permettant de les mesurer.

1.5.1. Généalogie

Elle regroupe une description de l'histoire des données, en particulier des informations concernant le producteur, les sources d'acquisition des données, les méthodes utilisées, et les opérations appliquées sur ces données.

Cette information est considérée comme indispensable, car elle permet de se faire rapidement une idée a priori du contenu de la base de données, et de la qualité moyenne qu'on est en droit d'attendre, au vu de son mode de fabrication et de son producteur. Elle est en particulier très utile pour les transferts de données, puisqu'elle est une sorte de carte de visite de la base.

1.5.2. Actualité

C'est le décalage entre le jeu de données à une date T_1 et le terrain nominal à une date de référence T_2. Elle décrit la «fraîcheur» des données.

Les informations utiles à ajouter sont la date de dernière mise à jour, la date de validité des données, etc. On se contente en général de vérifier la présence des bons éléments dans la base de données, mais en toute rigueur il faudrait également évaluer leur bonne position (une route peut avoir été déviée de son trajet initial, par exemple).

1.5.3. Cohérence logique

C'est le degré de cohérence interne des données selon les règles de modélisation et de spécifications du jeu de données.

On distingue les règles de formatage des données des contraintes d'intégrité. Les premières rendent utilisable le jeu de données au sens informatique (lecture de fichiers), et les secondes découlent des spécifications de la base. On trouve des règles explicites, écrites dans les spécifications, et des règles implicites, liées à la nature des objets représentés (des courbes de niveau ne se croisent pas par exemple).

1.5.4. Précision géométrique

Elle décrit l'écart de géométrie entre un objet du terrain nominal et son homologue dans la base de données.

On distingue la précision de position (des objets ponctuels, linéaires et surfaciques), de la précision de forme. Ces deux notions sont cependant liées. La précision de position donne une information sur les écarts de géométrie entre les primitives, et la précision de forme prend en compte la nature des objets.

1.5.5. Précision sémantique et exhaustivité

C'est la différence entre les valeurs descriptives des objets du jeu de données (attributs) et les valeurs de leurs homologues dans le terrain nominal (précision sémantique), et l'absence ou la présence d'éléments du jeu de données par rapport au terrain nominal (exhaustivité).

La précision sémantique et l'exhaustivité portent sur le classement des objets, leurs attributs, et les liens entre les objets. Cette composante est très regardée par les utilisateurs, car elle leur donne l'assurance que la base contient bien les objets qu'elle est censée contenir, et que ceux-ci sont bien codés et bien référencés. C'est un critère essentiel pour beaucoup d'applications.

Ces composantes de la qualité, en particulier la précision géométrique et, dans une moindre mesure, la cohérence logique ont été largement étudiées (voir par exemple [GJ98] et [GG98] pour un état de l'art sur le sujet), et ont donné lieu à de nombreuses études à l'IGN ([Vau97] [bHA97] [Abb94] [Fas94] [Pen94] [Rav96]). Le problème de la visualisation

des informations de qualité a également été abordé ([You96] [Fai96]). Notons cependant que l'étude de la précision sémantique est relativement peu abordée.

1.6. Indicateurs de la qualité sémantique

Cette partie présente les indicateurs de qualité sémantique définis à l'IGN et mesurés par le contrôle qualité en production [DF97]. Ils sont adoptés également par le CEN (Comité Européen de Normalisation). Les définitions sont exposées brièvement ; on pourra se rapporter aux documents de référence pour plus de détails.

La qualité sémantique se divise elle-même en deux composantes. En effet, un objet géographique appartient à une *classe*, dans la base de données, et est décrit par des *attributs*. Un objet peut donc être classé dans la mauvaise classe (erreur de classement), ou avoir un de ses attributs mal renseigné (erreur d'attribut). Nous appelons *référence* une base de données (conceptuelle) constituée des objets du terrain nominal, et *jeu de données* la base de données dont nous mesurons la qualité sémantique par le biais des indicateurs.

1.6.1. Classement des objets

On définit pour un ensemble d'objets géographiques du jeu de données quatre principaux indicateurs décrivant le classement de ces objets : le *taux de déficit*, le *taux d'excédent*, le *taux d'accord* et le *taux de confusion*. Soient :

- N_i (resp. n_i) le nombre d'objets de la classe C_i dans la référence (resp. dans le jeu de données) ;
- N_{0j} (resp. N_{i0}) le nombre d'objets de la classe C_j du jeu de données sans homologue dans la référence (excédents) (resp. nombre d'objets de la classe C_i de la référence sans homologue dans le jeu de données (déficits)) ;
- N_{ij} le nombre d'objets de la classe C_i de la référence avec pour homologue un objet de la classe C_j du jeu de données.

Les taux de déficit et d'excédent sont calculés de la façon suivante :

- Le taux de déficit de la classe C_i, noté $T_{C_{i0}}$, est le rapport N_{i0}/N_i (si $N_i = 0$ alors $N_{i0} = 0$ et on pose par convention $T_{C_{i0}} = 0$) ;
- le taux d'excédent de la classe C_j, noté $T_{C_{0j}}$, est le rapport N_{0j}/n_j (si $n_j = 0$ alors $N_{0j} = 0$ et par convention $T_{C_{0j}} = 0$).

Les taux de confusion et les taux d'accord sont définis ainsi :

- Le taux de confusion entre les classes C_i et C_j, noté $T_{C_{ij}}$, est le rapport N_{ij}/N_i (si $N_i = 0$ alors $N_{ij} = 0$ et on pose par convention $T_{C_{ij}} = 0$) ;
- le taux d'accord de la classe C_i, noté $T_{C_{ii}}$, est le rapport N_{ii}/N_i (si $N_i = 0$ alors $N_{ii} = 0$ et par convention $T_{C_{ii}} = 1$).

1.6.2. Codification d'un attribut

On définit pour un attribut énuméré a ayant K modalités des notions de *taux de confusion*, de *taux d'accord*, de *taux de déficit* et de *taux d'excédent* de la même façon que précédemment. On considère N objets correctement appariés entre la référence et le jeu de données, c'est-à-dire des objets décrivant la même réalité. Soient :

- N_i (resp. n_i) le nombre d'objets ayant la modalité i dans la référence (resp. dans le jeu de données), avec $i = 0$ pour les attributs non renseignés ;
- N_{0j} (resp. N_{i0}) le nombre d'objets ayant la modalité j dans le jeu de données et étant non renseignés dans la référence (resp. nombre d'objets ayant la modalité i dans la référence et étant non renseignés dans le jeu de données) ;
- N_{ij} le nombre d'objets ayant la modalité i dans la référence et j dans le jeu de données.

Les taux de déficit et d'excédent sont calculés de la façon suivante :

- Le taux de déficit de la modalité i de l'attribut a, noté t_{i0}, est le rapport N_{i0}/N_i ;
- le taux d'excédent de la modalité j de l'attribut a, noté t_{0j}, est le rapport N_{0j}/n_j.

Les taux de confusion et les taux d'accord sont définis ainsi :

- Le taux de confusion entre les modalités i et j de l'attribut a, noté t_{ij}, est le rapport N_{ij}/N_i ;
- le taux d'accord de la modalité i de l'attribut a est le rapport N_{ii}/N_i.

1.6.3. Matrices de confusion

Les indicateurs d'erreurs de classement, et les indicateurs d'erreurs d'attributs sont regroupés dans des tableaux appelés *matrices de confusion*. Pour une base de données ne contenant que des objets du même thème (voies de communication par exemple), on donne une matrice de confusion pour les classements, et une matrice de confusion pour les attributs de chaque classe.

Nous présentons en table 1.1 un exemple fictif de matrice de confusion pour le classement des objets. Nous lisons en première ligne et deuxième colonne que 10% des 110 chemins de la référence (donc 11 chemins) ont été classés à tort en allée dans le jeu de données. Remarquons que la somme des valeurs ligne par ligne est bien égale à 100%. Ce n'est pas le cas pour les colonnes, puisque les taux sont calculés par rapport à la référence.

1.7. Modèles d'incertitude

La qualité des données géographiques est décrite par de nombreux indicateurs, mais ceux-ci sont de natures très diverses. Cette hétérogénéité empêche de les combiner et de les agréger. Nous relevons au moins trois types d'indicateurs :

Référence \ Jeu de données	Chemin 95	Allée 50	Piste cyclable 41	Escalier 8
Chemin 110	90%	10%	0%	0%
Allée 50	10%	70%	0%	20%
Piste cyclable 40	0%	0%	100%	0%
Escalier 10	0%	4%	6%	90%

TAB. 1.1.: Exemple de matrice de confusion de classement

- commentaires purement descriptifs (nom et coordonnées du fournisseur de données par exemple) ;
- indicateurs calculés à l'aide de dénombrement (indicateurs de qualité sémantique par exemple) ;
- paramètres de modèles probabilistes d'incertitude (moyenne et écart-type de la position d'un carrefour par exemple).

Les indicateurs fondés sur des modèles probabilistes sont utilisables tels quels. Ils peuvent être combinés et propagés, bien que ces opérations deviennent très vite complexes [Rav96]. En revanche, le fait que les indicateurs de qualité sémantique soient uniquement du dénombrement rend impossible leur agrégation, et leur trop grand nombre empêche de les manipuler tous ensemble. Nous établissons donc des modèles probabilistes d'erreur d'attributs, pour pouvoir ensuite estimer les paramètres de ces modèles, et propager les modèles dans les applications. De tels modèles sont nouveaux, et peuvent également s'appliquer à des bases de données classiques.

2. Modélisation d'erreurs d'attributs dans une base de données géographiques

Nous présentons maintenant un modèle décrivant les erreurs d'attributs présentes dans une base de données géographiques. Nous étudions en particulier un attribut A qualitatif énuméré possédant K modalités notées $1, 2, \ldots, K$. Ce modèle permet de synthétiser les indicateurs de qualité fournis par le contrôle qualité de l'IGN, et de les manipuler, ce qui se révèle indispensable dans la suite de notre travail. Il est à notre connaissance original.

2.1. Cadre du modèle

2.1.1. Observations

Pour mesurer la qualité d'une base de données géographiques, la méthode la plus courante est de comparer cette base à une base de meilleure qualité, qui devient notre référence, assimilée au terrain nominal. Nous disposons donc de deux bases de données géographiques que nous décrivons par le terme *jeu de données*, et par le terme *référence*.

Ces bases de données contiennent un certain nombre d'objets géographiques portant l'attribut A. Dans une base de données routières par exemple, les tronçons de route (objets) ont un attribut A qui renseigne le nombre de voies du tronçon. Un objet commun aux deux bases aura pour l'attribut a une valeur $r \in \{1, \ldots, K\}$ dans la référence et une valeur $d \in \{1, \ldots, K\}$ dans le jeu de données. Nous décrirons cet état par (r, d). Il faut ajouter la possibilité pour un objet d'avoir son attribut non renseigné dans l'une des deux bases. On parle d'excédent si la valeur de l'attribut est déterminée dans le jeu de données et non dans la référence, et de déficit dans le cas contraire. Ces excédents (resp. déficits) sont notés par $(0, d)$ si l'attribut n'est pas renseigné dans la référence et vaut d dans le jeu de données (resp. $(r, 0)$ si l'attribut vaut r dans la référence et n'est pas renseigné dans le jeu de données).

2.1.2. Lois

Les observations dont on dispose sont les couples (r, d) pour les objets communs aux deux bases et correctement appariés, c'est-à-dire les objets décrivant la même réalité. Ce sont des réalisations d'une variable aléatoire $X = (R, D)$. Nous supposons qu'il y a

indépendance entre les erreurs commises sur tous les objets. La loi de probabilité décrivant la valeur de l'attribut A dans les deux bases (loi de X) est une loi discrète p définie par une matrice $(p_{rd})_{0\leqslant r,d\leqslant K}$ avec :

$$P(X=(r,d))=p_{rd} \qquad \forall (r,d)\in\{0,\dots,K\}^2.$$

2.2. Estimation des paramètres du modèle

Le modèle étant maintenant posé, nous allons vouloir estimer ses paramètres. Nous proposons des hypothèses simplificatrices pour réduire le nombre de paramètres ($(K+1)^2-1$ pour le moment), et calculons la vraisemblance (voir [Bor87]) dans chacun des cas étudiés.

2.2.1. Vraisemblance

Écrivons la vraisemblance pour notre modèle. Nous notons les N réalisations de X $x_1,\dots,x_N$. La vraisemblance s'écrit :

$$L(x_1,\dots,x_N)=\prod_{i=1}^{N}\left(\sum_{r=0}^{K}\sum_{d=0}^{K}\mathbf{1}_{(x_i=(r,d))}p_{rd}\right) \tag{2.1}$$

le signe $\mathbf{1}$ désignant la fonction indicatrice de l'événement placé en indice. Pour chaque i fixé, un seul des termes de la somme double est non nul. Cette expression se transforme donc en :

$$L(x_1,\dots,x_N)=\prod_{r=0}^{K}\prod_{d=0}^{K}p_{rd}^{\sum_{i=1}^{N}\mathbf{1}_{(x_i=(r,d))}} \tag{2.2}$$

2.2.2. Lien avec le contrôle qualité

Le contrôle qualité nous donne un certain nombre d'indicateurs de qualité, que nous avons présentés précédemment. En particulier, $\sum_{i=1}^{N}\mathbf{1}_{(x_i=(r,d))}$ est ce que le contrôle qualité appelle N_{rd} (nombre d'objets ayant la valeur r dans la référence et d dans le jeu de données). La vraisemblance 2.2 peut donc être réécrite très simplement :

$$L(x_1,\dots,x_N)=\prod_{r=0}^{K}\prod_{d=0}^{K}p_{rd}^{N_{rd}} \tag{2.3}$$

2.3. Hypothèses simplificatrices

Nous pouvons avancer un certain nombre d'hypothèses simplificatrices pour réduire le nombre de paramètres du modèle. Ces hypothèses, indispensables, doivent avoir une justification géographique. Nous nous ramenons ainsi à un cadre paramétrique raisonnable.

Nous proposons deux hypothèses différentes, qui nous serviront de base de travail, en fonction de la nature de l'attribut étudié. Nous écrivons ensuite une paramétrisation des modèles.

2.3.1. Cas uniforme

Supposons qu'une erreur dans le jeu de données pour une valeur d'un attribut se répartisse uniformément parmi les valeurs possibles de l'attribut. Par exemple un objet o de valeur d'attribut r dans la référence voit sa valeur d'attribut dans le jeu de données mal codée, avec une égale probabilité d'erreur entre les autres valeurs possibles, y compris l'absence de valeur notée 0.

Cette hypothèse se traduit par :

$$P((R,D)=(r,d)) = \begin{cases} p_{rr} & \text{si } d=r \\ p_r & \text{sinon} \end{cases}, \forall (r,d) \in \{0,\dots,K\}^2. \tag{2.4}$$

Les coefficients d'une même ligne sont tous égaux, à l'exception de la diagonale. Remarquons que cette hypothèse pour $r=0$ signifie qu'un attribut dont la valeur est indéterminée peut avoir par erreur une valeur, avec équiprobabilité entre les différentes valeurs.

2.3.2. Cas tridiagonal

L'hypothèse précédente n'est pas forcément pertinente pour tous les attributs, en particulier lorsqu'on a remarqué que les valeurs d'un attribut sont généralement ordonnées de façon logique. Prenons l'exemple de l'attribut *Nombre total de voies*, qui peut prendre les valeurs *Inconnu*, *1 voie*, *2 voies*, *3 voies*, *4 voies*, *2 voies larges*, *Plus de 4 voies*. On constate qu'une route à 3 voies par exemple, en cas d'erreur, est plus probablement codée en 2 voies ou 4 voies qu'une autre valeur. Cela nous amène à proposer des structures de matrices concentrées sur la diagonale, que nous nommons *tridiagonales*. Ces structures sont assez fidèles aux matrices de confusion estimées par le contrôle qualité effectué en production.

Écrivons la structure d'une loi de probabilité à structure tridiagonale. On ne va garder que la diagonale, et les coefficients juste au-dessus et au-dessous de cette diagonale. On obtient donc :

$$p = \begin{pmatrix} p_{00} & p_{01} & 0 & \dots & \dots \\ p_{10} & p_{11} & p_{12} & 0 & \dots \\ 0 & p_{21} & p_{22} & \ddots & \ddots \\ \vdots & 0 & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix}. \tag{2.5}$$

Si on ajoute une hypothèse semblable à celle du cas uniforme, la relation $p_{k(k-1)} = p_{k(k+1)}$ $\forall k \in \{1,\dots,K-1\}$ réduit encore le nombre de paramètres.

Remarquons qu'un tel modèle ne permet pas de prendre en compte les déficits et les excédents. On peut toutefois étendre le modèle tridiagonal en ajoutant une colonne et une ligne pour les intégrer.

2.4. Paramétrisation

Nous allons maintenant paramétrer le modèle, dans le cadre des deux hypothèses uniforme et tridiagonale.

Nous réécrivons le modèle en nous inspirant de la formule de Bayes :

$$P((R,D)=(r,d)) = P(R=r)P(D=d \mid R=r) \qquad \forall (r,d) \in \{0,\dots K\}^2. \tag{2.6}$$

La quantité $P(R=r)$ décrit la répartition des valeurs d'attribut dans la référence. Cette probabilité peut être estimée par la valeur N_r/N. Nous rappelons que N désigne le nombre d'objets total de la référence, et N_r le nombre d'objets ayant pour valeur d'attribut r dans la référence. La quantité $P(D=d \mid R=r)$ est la probabilité que l'attribut ait la valeur d dans le jeu de données alors qu'il a la valeur r dans la référence.

2.4.1. Hypothèse uniforme

Dans le cas de l'hypothèse uniforme, il faut définir $2(K+1)$ coefficients. Un choix naturel en tenant compte de la remarque précédente est de poser :

$$\begin{cases} p_{rr} = (1-\theta_r)\frac{N_r}{N} \\ p_r = \frac{\theta_r}{K}\frac{N_r}{N} \end{cases} , \forall r \in \{0,\dots,K\} \tag{2.7}$$

Le nombre de paramètres peut encore être extrêmement réduit en faisant par exemple l'hypothèse de l'égalité des θ_r (un seul paramètre), ou en retenant un paramètre pour la ligne r pour laquelle N_r est le plus grand (valeur de l'attribut la plus représentée), et un paramètre pour les autres lignes.

2.4.2. Hypothèse tridiagonale uniforme

Le cas tridiagonal uniforme nécessite aussi $2(K+1)$ coefficients. Un raisonnement identique au précédent, avec un aménagement pour la première et la dernière ligne conduit à proposer :

$$\begin{cases} p_{rr} = (1-\theta_r)\frac{N_r}{N} & \forall r \in \{0,\dots,K\} \\ p_{r(r-1)} = p_{r(r+1)} = \frac{\theta_r}{2}\frac{N_r}{N} & \forall r \in \{1,\dots,K-1\} \\ p_{01} = \theta_0\frac{N_0}{N} \\ p_{(K-1)K} = \theta_K\frac{N_K}{N} \end{cases} \tag{2.8}$$

Nous pouvons de même réduire le nombre de paramètres en faisant par exemple l'hypothèse de l'identité des θ_r, ou en conservant deux paramètres comme nous le proposons pour le cas uniforme.

2.5. Calcul d'estimateurs

Nous allons détailler le calcul d'estimateurs dans le cas de la dimension 2, ce qui revient à étudier un attribut à deux modalités sans s'intéresser aux déficits ni aux excédents. Nous étendrons ensuite les résultats à des dimensions quelconques.

2.5.1. Cas d'un attribut à deux modalités

La particularité de la dimension 2 est que les hypothèses uniforme et tridiagonale uniforme coïncident. Nous allons écrire un modèle à un seul paramètre. La loi p s'écrit :

$$p = \begin{pmatrix} (1-\theta)\frac{N_1}{N} & \theta\frac{N_1}{N} \\ \theta\frac{N_2}{N} & (1-\theta)\frac{N_2}{N} \end{pmatrix}$$

La vraisemblance s'écrit d'après 2.3 :

$$L(x_1, \dots, x_N) = \left((1-\theta)\frac{N_1}{N}\right)^{N_{11}} \left((1-\theta)\frac{N_2}{N}\right)^{N_{22}} \left(\theta\frac{N_1}{N}\right)^{N_{12}} \left(\theta\frac{N_2}{N}\right)^{N_{21}} \tag{2.9}$$

En regroupant les termes constants, l'expression devient :

$$L(x_1, \dots, x_N) = K \times (1-\theta)^{N_{11}+N_{22}} \times \theta^{N_{12}+N_{21}} \tag{2.10}$$

K étant une constante réelle. Cette fonction admet clairement son maximum pour le point tel que $\frac{L'}{L} = 0$:

$$\theta(N_{11} + N_{22}) = (1-\theta)(N_{12} + N_{21})$$

soit

$$\theta = \frac{N_{12} + N_{21}}{N_{11} + N_{22} + N_{12} + N_{21}} = \frac{N_{12} + N_{21}}{N}$$

L'estimateur du maximum de vraisemblance est donc dans ce cas :

$$\hat{\theta} = \frac{N_{12} + N_{21}}{N} \tag{2.11}$$

Si on décide maintenant de faire dépendre la loi de deux paramètres θ_1 et θ_2, c'est-à-dire :

$$p = \begin{pmatrix} (1-\theta_1)\frac{N_1}{N} & \theta_1\frac{N_1}{N} \\ \theta_2\frac{N_2}{N} & (1-\theta_2)\frac{N_2}{N} \end{pmatrix} \tag{2.12}$$

on obtient exactement de la même façon d'après 2.3 :

$$L(x_1, \dots, x_N) = \left((1-\theta_1)\frac{N_1}{N}\right)^{N_{11}} \left((1-\theta_2)\frac{N_2}{N}\right)^{N_{22}} \left(\theta_1\frac{N_1}{N}\right)^{N_{12}} \left(\theta_2\frac{N_2}{N}\right)^{N_{21}} \tag{2.13}$$

soit en regroupant les termes :

$$L(x_1, \dots, x_N) = K \times (1-\theta_1)^{N_{11}} \times \theta_1^{N_{12}} \times (1-\theta_2)^{N_{22}} \times \theta_2^{N_{21}} \tag{2.14}$$

K étant une constante réelle. Cette fonction admet son maximum en

$$\theta_1 = \frac{N_{12}}{N_{11} + N_{12}} = \frac{N_{12}}{N_1}$$

et

$$\theta_2 = \frac{N_{21}}{N_{22} + N_{21}} = \frac{N_{21}}{N_2}$$

L'estimateur du maximum de vraisemblance est donc ici :

$$\begin{cases} \hat{\theta}_1 = \dfrac{N_{12}}{N_1} \\ \\ \hat{\theta}_2 = \dfrac{N_{21}}{N_2} \end{cases} \tag{2.15}$$

2.5.2. Cas d'un attribut à K modalités

Nous reprenons maintenant le cadre général, avec les déficits et les excédents. Nous allons donner les estimateurs du maximum de vraisemblance des paramètres des lois sans détailler les calculs, car ils sont identiques à ceux du cas à deux modalités présenté plus haut.

Dans le cas d'une loi à un paramètre θ, en faisant l'hypothèse tridiagonale, on obtient, avec les mêmes notations :

$$\hat{\theta} = \frac{N_{01} + \sum_{k=1}^{K-1} \left(N_{k(k-1)} + N_{k(k+1)} \right) + N_{K(K-1)}}{\sum_{k=0}^{K} N_{kk} + N_{01} + \sum_{k=1}^{K-1} \left(N_{k(k-1)} + N_{k(k+1)} \right) + N_{K(K-1)}} \tag{2.16}$$

On obtient pour le cas uniforme un résultat similaire :

$$\hat{\theta} = \frac{N - \sum_{k=0}^{K} N_{kk}}{N} \tag{2.17}$$

2.6. Étude de contrôles qualité sur des données réelles

Nous avons appliqué nos modèles au thème routier et au thème bâti de la base de données BDTopo. Nous avons utilisé un modèle uniforme à 1 paramètre, ainsi qu'un modèle tridiagonal à 1 et 2 paramètres, et estimé ces paramètres, pour une dizaine de feuilles BDTopo. Pour chaque feuille, les paramètres estimés reflètent correctement la qualité de la base, telle qu'elle a été qualifiée par le contrôle qualité de l'IGN. Le caractère très synthétique du modèle proposé ne semble pas induire de perte d'information importante. Soulignons que la qualité des feuilles est très bonne, et qu'ainsi les paramètres du modèle sont assez petits (de 3% à 5%).

Nous reproduisons ici en figure 2.1 une matrice réelle de contrôle qualité, et donnons la valeur des estimateurs des paramètres des différents modèles.

Chemin	Route empierrée	Route à 1 voie	Route à 2 voies ou +	Chaussées séparées
93,3	0,71	0	0,24	0
2,85	87,33	7,42	1,37	0
0,97	2,68	93,74	1,83	0
0	0,16	2,57	96,54	0
0	0	0	0	100

FIG. 2.1.: Matrice de confusion déterminée en contrôle qualité

Les longueurs d'objets contrôlés en kilomètres ayant pour valeur d'attribut *Chemin*, *Route empierrée*, *Route à 1 voie*, *Route à 2 voies ou +*, *Chaussées séparées* sont 105,21 km, 43,8 km, 205,25 km, 477,41 km et 29 km.

Le modèle uniforme à un paramètre donne un paramètre $\theta = 0,0353$, et le modèle tridiagonal à deux paramètres donne $\theta = 0,0317$.

3. Impact de la qualité des données sur une application

La qualité est estimée et mesurée par le producteur de données ; elle est résumée sous la forme d'indicateurs qui sont fournis avec les jeux de données. Un des problèmes qui se pose pour un utilisateur de ces données est de pouvoir prévoir et mesurer les répercussions de la qualité des données sur l'utilisation qu'il compte en faire. Cela se traduit souvent par la question : « les données répondent-elle à mon besoin en termes de qualité ? ».

Pour répondre à cette question, il faut définir ce qu'est une *application géographique*, et estimer les répercussions des problèmes de qualité sur une application géographique.

3.1. Application géographique

On appelle *application géographique* tout processus utilisant une ou plusieurs bases de données géographiques vectorielles (puisque le cadre de notre étude est l'information géographique objet), et produisant une nouvelle information géographique.

Ce terme décrira donc aussi bien un logiciel exploitant ou gérant des bases de données géographiques (un SIG par exemple), que le processus de réalisation d'une carte à l'aide d'un logiciel de dessin.

Les résultats d'une application géographique peuvent prendre des formes très variées. On peut trouver des valeurs, c'est-à-dire des nombres (temps de parcours entre deux villes par exemple), des graphes valués (réseau hydrographique portant l'indication de débit des différents cours d'eau), des cartes numériques (vectorielles ou maillées), des cartes papier, des modèles numériques de terrain, etc.

Au vu de ces quelques exemples, il est difficile d'établir une typologie des différentes applications géographiques en fonction de leur résultats, et cette typologie serait totalement arbitraire, bien qu'il existe des tentatives dans ce sens [Alb95].

Une étape importante pour l'étude d'une application géographique est la caractérisation de ses résultats. Il s'agit de quantifier, ou tout au moins de rendre comparables différents résultats d'une même application géographique. Prenons l'exemple d'une application de calcul de zones inondables. Les résultats sont les différentes zones inondables ; ils peuvent être caractérisés par la superficie des zones inondables, la population des agglomérations inondables, etc.

Caractériser de cette façon les résultats d'une application géographique est en général un travail complexe. On doit garder présent à l'esprit que la caractérisation doit être orientée vers l'analyse qu'on fait subir aux résultats de l'application.

Dans le cas d'un image maillée, on dispose d'outils bien connus de traitement d'image (citons entre autres moyens d'analyse l'étude du contraste, de la résolution, de la transformée de Fourier). Ces outils permettent de comparer des images, et de les caractériser. Il faut définir des critères de caractérisation des résultats d'une application géographique dans le même esprit.

3.2. Exemple : calcul d'itinéraires

Nous allons décrire ici un prototype d'application de calcul d'itinéraires, utilisant la base de données Géoroute de l'IGN. De telles applications sont largement répandues dans le grand public, et embarquées à bord de véhicules. Nous avons développé cette application à des fins de simulation, elle est donc dépourvue d'interface utilisateur.

3.2.1. Description de l'application

L'application utilise une base de données géographiques pour déterminer selon certains critères (rapidité, distance), le «meilleur» itinéraire pour rallier une destination à partir d'un lieu donné.

Présentons brièvement la base de données utilisée. Géoroute est une base de données vectorielles, à vocation routière, réalisée en majorité à l'aide de deux autres bases de données géographiques de l'IGN, la BDCarto et la BDTopo. C'est une base de données particulière, car elle comporte des données de résolution différente en zone urbaine (échelle de saisie de la BDTopo), et en zone moins dense (échelle de saisie de la BDCarto). En plus de ce caractère bi-échelle, qui permet d'avoir le même niveau de détail quelle que soit la densité de la zone, sa très grande richesse sémantique, en particulier sur le réseau routier, la rend toute désignée pour des applications routières.

Nous représentons en figure 3.1 un schéma simplifié d'une partie de cette base de données. Une *route* (classe non représentée) est composée de *tronçons de routes*. Les *tronçons de routes* ont leur géométrie décrite par un sommet initial et un sommet final (liens à la classe *nœuds routiers*), et leurs caractéristiques décrites par la valeur des attributs qu'ils portent. Les *noeuds routiers* possèdent des coordonnées dans un système de référence permettant de les situer à la surface de la terre.

L'application de calcul d'itinéraires calcule le plus court chemin entre un point de départ et un point d'arrivée spécifiés par l'utilisateur, et donne la longueur et le temps de parcours de cet itinéraire. La notion de *plus court chemin* au sens strict est celle de plus court chemin en distance sur le réseau, mais par abus de langage, elle peut désigner également le chemin le plus rapide. Les attributs permettent d'attribuer à chaque tronçon une vitesse de parcours possible (éventuellement nulle si le tronçon est impraticable). Trouver le plus

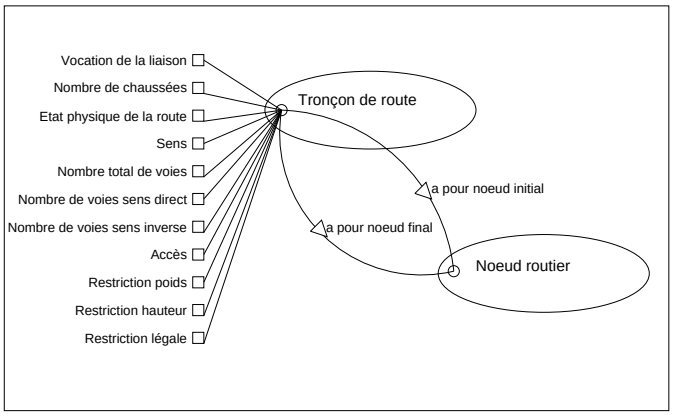

FIG. 3.1.: Schéma simplifié de la structure de Géoroute

court chemin entre deux points revient donc à trouver le plus court chemin sur un graphe valué. Le graphe peut être valué par les longueurs des tronçons (pour déterminer le plus court chemin), ou par les temps de parcours de ces tronçons (pour déterminer le chemin le plus rapide). Une littérature abondante existe sur la détermination du plus court chemin ; nous utilisons l'algorithme de Dijkstra [Dij59].

Nous présentons figure 3.2 la zone étudiée (Lagny), et un exemple de plus court chemin (en noir).

3.2.2. Caractérisation des résultats

Nous proposons plusieurs indicateurs pour évaluer la qualité des résultats de cette application. Le but de l'étude n'est pas de mettre en évidence les imperfections de l'application de calcul d'itinéraires, mais de mesurer les répercussions d'éventuels problèmes de qualité de la base de données sur les résultats de l'application. Les indicateurs mesurent donc seulement les erreurs provoquées par une dégradation de la qualité de la base utilisée.

La détermination de tels indicateurs suppose que l'on se dote d'une référence. C'est dans notre cas tous les résultats de l'application obtenus avec une base de données qui serait conforme au terrain nominal.

Nous pouvons ensuite calculer des indicateurs de qualité des résultats en comparant les résultats obtenus aux résultats de référence. Des indicateurs envisageables sont :

- l'écart en temps sur un itinéraire donné ;
- l'écart en distance sur un itinéraire donné ;
- les moyennes et écart-types des précédents écarts sur tous les itinéraires possibles ;
- le nombre de tronçons communs entre un itinéraire donné et l'itinéraire de référence ;
- le nombre de tronçons non praticables (sens interdit, ou restriction de hauteur ou de poids) retenus pour un itinéraire donné ou pour l'ensemble des itinéraires.

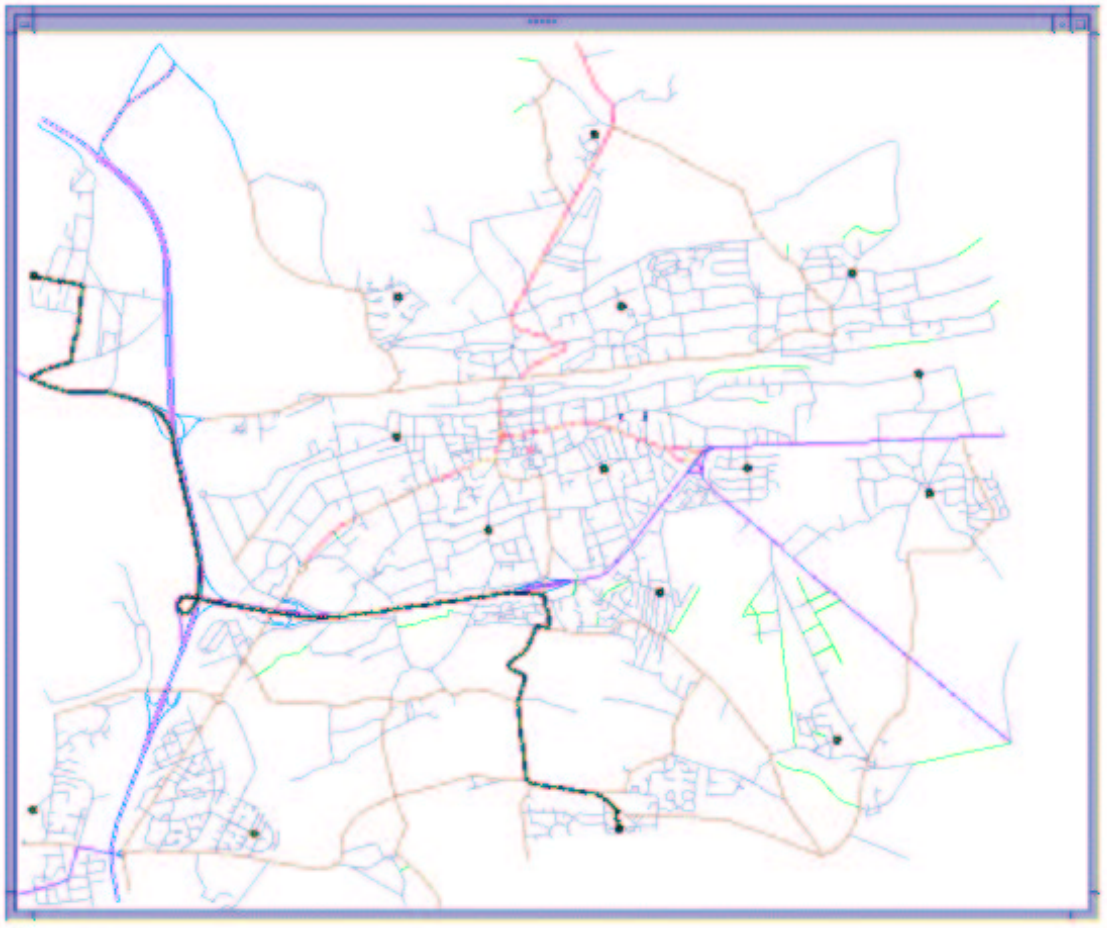

FIG. 3.2.: Exemple de plus court chemin sur la zone étudiée (Lagny)

Cette liste n'est bien entendu pas exhaustive. Le choix de l'un ou l'autre indicateur dépendra essentiellement des exigences de l'utilisateur. Une compagnie de taxis voudra prévoir des temps de parcours les plus exacts possibles, alors qu'un transporteur routier pourra également être intéressé par l'itinéraire le plus court pour économiser les camions et le gazole. Le dernier indicateur proposé est important pour des particuliers, car ils veulent éviter de prendre des sens interdits, mais aura peu d'importance pour des véhicules prioritaires comme ceux des pompiers par exemple.

3.3. Influence de la qualité sur un calcul d'itinéraires

L'application de calcul d'itinéraires ne s'exprime pas de façon analytique en fonction des attributs et des longueurs des tronçons de route, car elle repose sur des algorithmes non linéaires. Deux approches peuvent être envisagées pour étudier l'impact d'erreurs

d'attributs sur l'application. La première consiste à simuler des erreurs contrôlées dans la base de données, à des taux de plus en plus élevés, et d'étudier la dégradation des résultats. Cette approche, quoique complexe, donne des résultats intéressants. Nous la présentons dans le chapitre II. La seconde consiste à construire un modèle analytique simple de l'application, à l'aide d'un certain nombre d'hypothèses soigneusement choisies. Elle est décrite dans le chapitre III, avec des extensions dans les chapitres IV et V.

Chapitre II.

Étude par simulation

Nous présentons dans ce chapitre une étude par simulation de l'impact des erreurs dans une base de données sur une application. Puisque les applications géographiques sont en général très complexes, il est rarement possible de pouvoir les considérer autrement que comme des boîtes noires, ce qui motive l'utilisation de techniques de simulation pour mener l'étude.

Nous proposons d'utiliser une technique originale inspirée de l'analyse de sensibilité géographique (*geographical sensitivity analysis*), conçue pour répondre à un problème similaire dans le cas des SIG maillés (voir par exemple [McM96] [LMS90] [Fis91] [GW94] [Heu93] [HB93]). Nous l'avons transposée au cas des données vectorielles, et appliquée avec succès à l'étude d'une application de calcul d'itinéraires [Bon98].

Cette technique ne nécessite pas de connaître le fonctionnement de l'application géographique. En revanche, elle impose de disposer d'un certain nombre d'outils que nous allons présenter [Bon99] :

- un outil de bruitage *contrôlé* de la base de données ;
- l'application géographique elle-même ;
- un outil de calcul d'un ou plusieurs critères de qualité des résultats de l'application.

Elle nécessite enfin de mener une étude statistique pour dépouiller les résultats.

Un attention particulière est accordée à l'outil de bruitage [Bon00b] [Bon00a], qui a également d'autres applications, en contrôle qualité par exemple.

1. Principe de l'analyse de sensibilité géographique

L'analyse de sensibilité géographique consiste à considérer l'application géographique comme une boîte noire, ou comme un filtre en traitement du signal, avec en entrée la base de données, et en sortie les résultats de l'application. La qualité des données en entrée et en sortie est connue et mesurée, et on détermine par simulation la relation entre qualité en sortie et qualité en entrée à l'aide de méthodes statistiques telles que les modèles linéaires.

La qualité de la base de données est décrite par des paramètres fournis par le producteur. Rappelons que ces paramètres sont hétérogènes, très nombreux et impossibles à agréger tels quels. Les résultats de l'application sont qualifiés par l'utilisateur, expert dans son domaine. La difficulté est de quantifier cette qualification de la qualité des résultats.

Nous choisissons de mener une étude dynamique, c'est-à-dire d'analyser la dégradation de qualité des résultats quand on utilise des bases de données de moins en moins bonne qualité. Une telle étude est intéressante, car la qualité d'une base de données est toujours estimée sur un échantillon qu'on espère représentatif. Grâce à une telle étude, on peut prédire la qualité des résultats de l'application sachant que la qualité du jeu de données se situe dans un intervalle connu.

Comme il est impossible de disposer d'une base de données parfaite (référence), notre jeu de données devient notre référence, et nous créons à l'aide de cette référence de nouveaux jeux de données de moins bonne qualité en introduisant des erreurs par des techniques de Monte-Carlo. Ces erreurs doivent être introduites de la façon la plus réaliste possible, pour obtenir des jeux de données semblables à ceux observés en production.

Nous calculons à l'aide de notre application géographique les résultats pour chaque jeu de données utilisé, et mesurons à chaque fois la qualité des résultats de l'application. La dernière étape est de chercher la relation statistique entre la qualité en entrée et la qualité en sortie. Le principe de l'analyse de sensibilité géographique est illustré par la Figure 1.1.

Remarquons que cette étude dépend non seulement de l'application étudiée, mais aussi des critères retenus pour mesurer la qualité des résultats de l'application. Elle doit aussi être guidée par les exigences de l'utilisateur en terme de qualité. Pour une application de calcul d'itinéraires, l'utilisateur peut par exemple désirer étudier l'impact des erreurs de sens interdits. Il faut donc que l'outil de bruitage soit suffisamment souple pour permettre ce type d'étude.

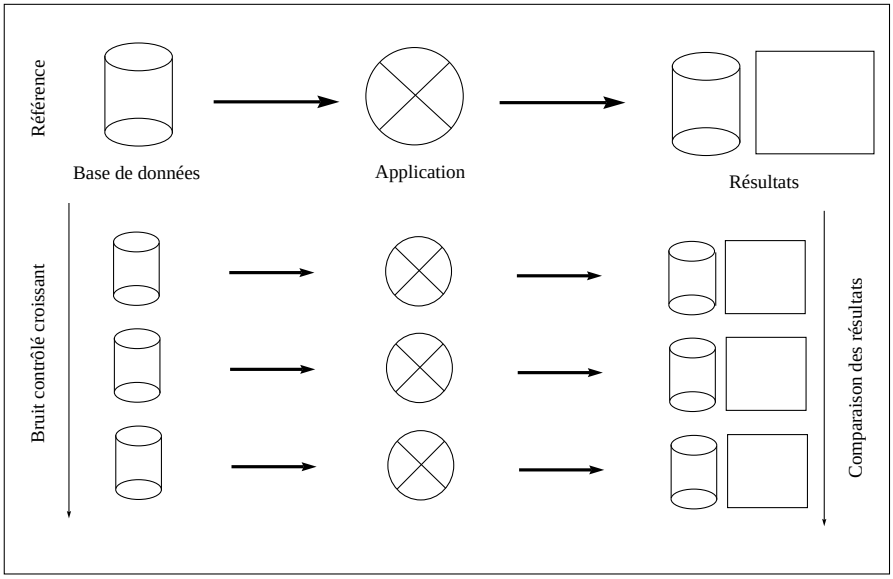

FIG. 1.1.: Principe de l'analyse de sensibilité

Nous présentons maintenant les méthodes de bruitage que nous avons mises au point dans le cadre de cette étude. Elles permettent de perturber les attributs et la géométrie.

2. Bruitage contrôlé d'une base de données géographiques

2.1. Bruitage des attributs

L'introduction d'erreurs d'attributs dans une base de données est fortement contrainte par la répartition très hétérogène des valeurs d'attributs dans la base. La plupart des attributs admettent une valeur extrêmement sur-représentée, comme le montre l'exemple du tableau 2.1 qui donne la répartition des valeurs de l'attribut «restriction de poids» du jeu de données Géoroute couvrant la zone urbaine que nous avons choisie pour notre étude.

Modalité	**Répartition dans la base de données**
Aucune	84.0 %
3.5 t	6.5 %
6 t	0.7 %
9 t	4.7 %
16 t	0.7 %
19 t	3.4 %

TAB. 2.1.: Répartition des modalités de l'attribut «restriction de poids»

En présentant notre modèle d'erreurs d'attributs (chapitre I), nous avons signalé que, selon l'attribut, les valeurs sont ordonnées logiquement ou non, ce qui rend certaines erreurs plus probables que d'autres lors de l'acquisition des données. Ainsi, nous avons proposé un modèle uniforme et un modèle tridiagonal, selon la nature de l'attribut.

Nous allons également distinguer entre deux composantes d'erreurs d'attributs, correspondant à deux types d'erreurs observées lors de la constitution d'une base de données :

- les fautes d'identification ;
- les erreurs informatiques.

Cette distinction impose de faire la différence entre les *objets géographiques* de la base de données (une route par exemple), et les *objets informatiques* (un tronçon de route, soit un segment de droite localisé portant des attributs).

Les *fautes d'identification* correspondent à une mauvaise interprétation des sources de données (photographies aériennes par exemple), et se répercutent sur la totalité des objets

géographiques concernés. Une piste cyclable confondue avec une contre-allée par exemple sera mal codée dans son intégralité. On pourra modéliser les fautes d'identification par une variable aléatoire dont les réalisations seront les objets géographiques, avec le modèle uniforme ou le modèle tridiagonal selon la nature de l'attribut.

Les *erreurs informatiques* correspondent à des erreurs ne concernant qu'un seul objet informatique, et sont d'origines diverses : bogue dans le logiciel ou faute d'inattention de l'opérateur en tapant un code par exemple. Les erreurs informatiques pourront être modélisées par une variable aléatoire dont les réalisations seront des objets informatiques. Nous utilisons le modèle uniforme, le modèle tridiagonal n'étant pas évidemment pas pertinent dans ce cas.

Nous introduisons du bruit dans la base de données de la manière suivante. Nous définissons tout d'abord un taux de bruit global pour la base de données. Nous répartissons ensuite ce taux global entre les deux composantes d'erreur que nous venons de présenter. La deuxième composante est la plus simple à introduire, puisque les objets sont directement présents dans la base. Pour chaque attribut, nous définissons des *coefficients de répartition*, que nous déduisons de la répartition des modalités de l'attribut. Nous pouvons utiliser la répartition telle quelle, ou amplifier les erreurs sur une modalité pour étudier son importance particulière. Supposons que l'attribut étudié ait quatre modalités a, b, c, et d, et que nous soyons intéressés par l'impact des erreurs sur la modalité b. Nous présentons dans le tableau 2.2 un exemple de coefficients de répartition pour un tel attribut.

Modality	**Répartition dans la base de données**	**Coefficients de répartition**
a	10 %	8 %
b	20 %	26 %
c	60 %	58 %
d	10 %	8 %

TAB. 2.2.: Exemple de coefficients de répartition

Dans l'hypothèse d'un bruit global de 10%, dont 90% se répartit sur la composante *erreurs informatiques*, nous tirons au sort 9% des objets informatiques de la base de données. Pour chacun de ces objets tirés au sort, nous tirons un nombre N uniformément entre 0 et 100 et choisissons ainsi la modalité à modifier :

- si $0 \leqslant N < 8$ alors perturber a ;
- si $8 \leqslant N < 34$ alors perturber b ;
- si $34 \leqslant N < 92$ alors perturber c ;
- si $92 \leqslant N \leqslant 100$ alors perturber d.

Supposons que $N = 32$. Nous devons perturber la modalité b. Si l'attribut est ordonné, nous tirons à pile ou face entre a et c, sinon nous tirons uniformément entre a, c et d.

La composante *fautes d'identification* nécessite de connaître les objets géographiques

avant de pouvoir bruiter. L'identification des objets géographiques peut s'appuyer sur les attributs (numéro de la route), sur la géométrie ou sur la topologie (parcours de graphe).

Cette méthode de bruitage a été implantée et testée [Fou99]. Les résultats sont tout à fait conformes à la réalité, et les contrôles qualité de production effectués sur les bases ainsi bruitées sont semblables aux contrôles qualité réels.

2.2. Bruitage de la géométrie

Le bruitage de la géométrie est assez simple en première approche, puisqu'on dispose de modèles validés à l'IGN [Abb94] [Vau97] [Rav96].

Dans le cas d'objets ponctuels, l'incertitude sur la position de l'objet est représentée par un disque gaussien, et dans le cas d'objets linéaires, l'incertitude sur les points de ces objets est représentée par une loi appelée GES à l'IGN (mélange de Gaussienne et de loi de Laplace). Remarquons que la loi GES suffit à recouvrir les deux cas. Avec ce modèle, un point centré sur l'origine dans la référence a une probabilité de présence en (x, y) définie par la densité :

$$f(x,y) = \alpha \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2+y^2}{2\sigma^2}} + (1-\alpha)\frac{\lambda}{2} e^{-\lambda\sqrt{x^2+y^2}}.$$

Si l'on suppose que les erreurs commises sur tous les points de la base de données sont indépendantes et de même loi, la simulation d'erreur suivant une loi GES est très simple. Chaque point appartenant à un objet linéaire de la base de données est déplacé dans une direction θ choisie uniformément, à une distance r de sa position initiale, le paramètre r étant tiré selon une loi GES (Figure 2.1).

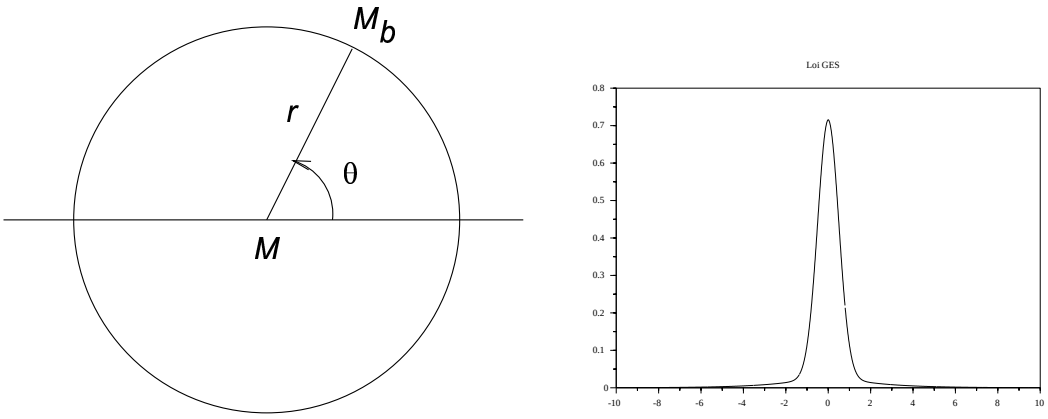

FIG. 2.1.: Déplacement d'un point suivant une loi GES

Une telle simulation d'erreurs conduit très vite à des problèmes topologiques, dès que les taux d'erreur deviennent modérés (au delà de 10 mètres dans la base de données Géoroute).

Nous présentons en figure 2.2 un exemple d'un tel problème ; après bruitage, une polyligne peut s'intersecter elle-même, créant ainsi des intersections parasites.

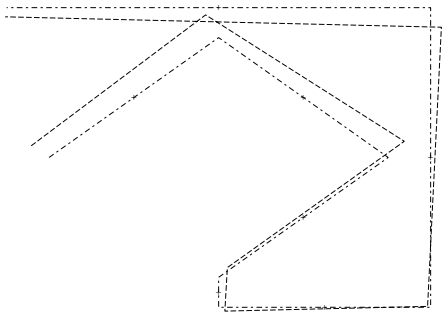

FIG. 2.2.: Problèmes topologiques induits par l'introduction d'erreurs

Ces problèmes topologiques ne sont pas réalistes, car même si la saisie de la base de données se fait avec des problèmes importants de précision, les problèmes topologiques sont en général évités par les opérateurs.

Il est donc nécessaire d'introduire des corrélations le long des polylignes pour obtenir des jeux bruités plus réalistes. Pour ce faire, nous proposons une réflexion originale sur le modèle GES qui, bien que très schématique, permet d'améliorer considérablement les résultats obtenus.

Le modèle GES est un mélange de Gaussienne et de loi de Laplace. Ce mélange s'est révélé être nécessaire pour permettre de bons ajustements des observations dans de nombreux jeux de données. Au vu des travaux de Vauglin [Vau97], nous faisons l'hypothèse que la composante gaussienne et la composante Laplace correspondent à deux réalités distinctes. Dans cette optique, la composante gaussienne modélise des imprécisions de mesure ou de pointé au moment de l'acquisition des données. La composante Laplace correspond aux erreurs de représentation, c'est-à-dire au problèmes de la représentation par polygonales d'objets plus réguliers. Rappelons que le tracé des routes est établi pour des raisons de facilité de circulation à partir de segments de droite, de clothoïdes et d'arcs de cercle, ce qui fait qu'une représentation par polylignes d'une route induit des écarts entre son tracé dans la base de données et sa géométrie réelle. Nous illustrons ces deux hypothèses par la figure 2.3.

La composante de représentation est moins importante pour les points de la base de données, pointés sur les routes, que pour les points intermédiaires des segments, constitués par interpolation linéaire. En revanche, la composante d'imprécision présente souvent la caractéristique inverse, puisque les routes présentent une régularité et une «rigidité» im-

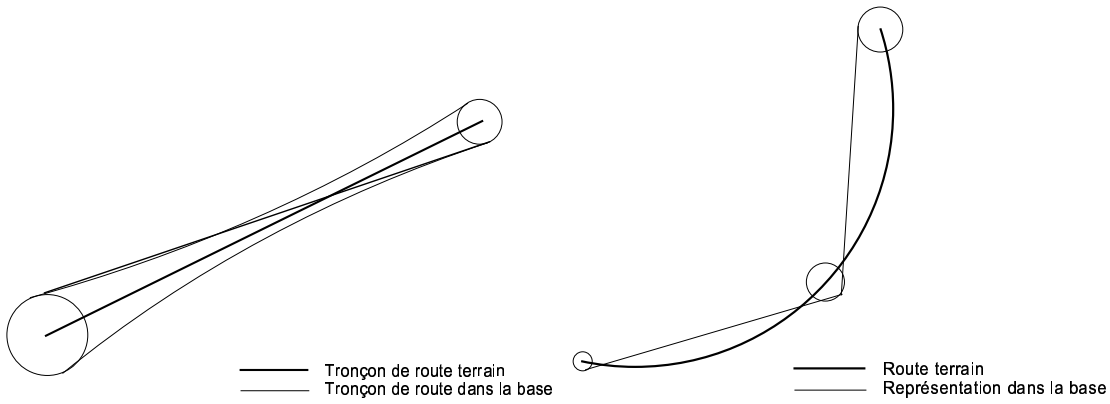

FIG. 2.3.: Composantes d'imprécision et de représentation

portante. Nous en déduisons que, selon notre hypothèse, deux effets s'opposent quand on analyse les erreurs de position le long des routes. Les points présents dans la base de données (extrémités des segments composant les polylignes) présentent des erreurs de pointé, mais pas d'erreur de représentation, alors qu'un point choisi sur un segment présente des erreurs de représentation d'autant plus importantes qu'il est éloigné des deux extrémités des segments, alors que les erreurs de pointé auront tendance à se compenser.

Nous travaillons dans le cadre de la comparaison de deux bases de données, puisque nous générons à partir d'un jeu de données de nouveaux jeux de qualité moindre. Notre référence est ainsi une base de données, avec des élément linéaires constitués de polygonales. La composante de représentation est absente dans notre contexte. Nous sommes dans le cas de figure où la «rigidité» des éléments linéaires induit une compensation des erreurs des extrémités des segments le long du segment lui-même. Dans la suite, nous nous intéressons à la composante imprécision, et nous réduirons les lois GES à leur composante gaussienne.

Pour introduire les corrélations entre les points d'une même polyligne, nous utilisons une méthode très simple qui utilise les variogrammes (voir [Fou99]). Le choix de s'appuyer sur les variogrammes est guidé par le fait qu'ils sont utilisés à l'IGN. Nous perturbons de façon indépendante les noeuds du graphe, c'est-à-dire les carrefours, et prenons en compte les corrélations pour les points intermédiaires. Nous présentons un exemple très simple illustrant notre méthode, pour une polyligne composée de seulement trois points (deux noeuds et un point intermédiaire) (figure 2.4), mais la méthode se généralise à un nombre de points quelconque.

Les points A et C sont déplacés indépendamment selon une loi gaussienne, et le point B est perturbé en fonction de A, C, et des longueurs l_{AB} et l_{BC}. Supposons pour plus de simplicité que les erreurs sont centrées. Les points A et C sont perturbés suivant des normales $N(0, \sigma^2)$, et le point B suivant une normale $N(0, \sigma'^2)$, avec σ'^2 calculé à l'aide du variogramme. Dans le cas d'un variogramme linéaire et de points équidistants, nous obtenons pour la fonction σ'^2/σ la courbe de la figure 2.5.

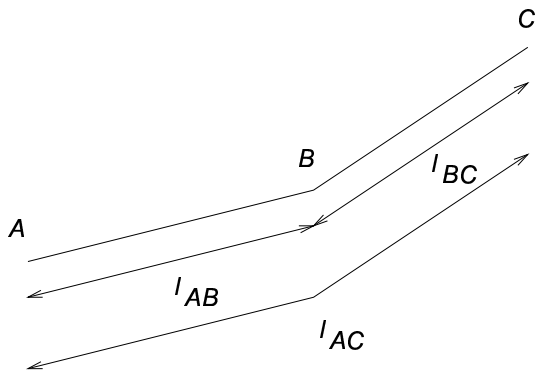

FIG. 2.4.: Exemple de polyligne élémentaire

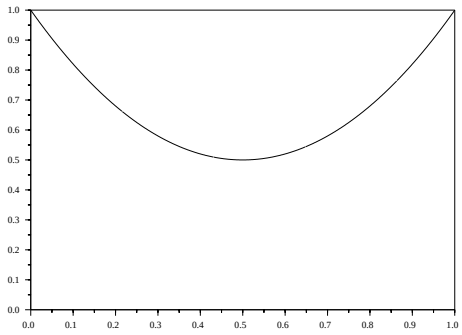

FIG. 2.5.: Variation de σ'^2/σ en fonction de la position de B (l'axe des abscisses représente la polyligne avec A en 0 et C en 1)

Cette approche réduit notablement les problèmes topologiques qui apparaissent lors de simulations sans corrélations. Les problèmes présents sur l'exemple précédent ont disparu (figure 2.6).

Cette méthode a été implantée, et a été testée sur des petits exemples [Fou99], mais n'a pas encore été évaluée sur des bases de données complètes. Cependant, les premiers résultats semblent très prometteurs. Nous avons constaté également que cette méthode a tendance à réduire les décalages qui peuvent apparaître lors de l'introduction d'erreurs sans corrélations (figure 2.7).

La méthode de bruitage des objets linéaires avec prise en compte des corrélations est simple, et donne des résultats satisfaisant. Les jeux de données bruités sont suffisamment réalistes pour être employés à de l'analyse de sensibilité.

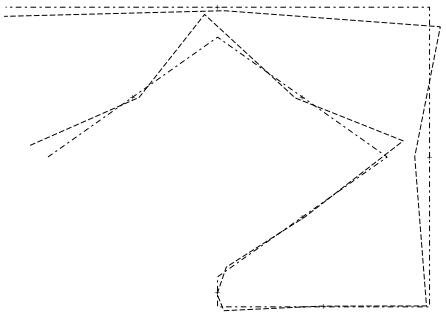

FIG. 2.6.: Résolution de problèmes topologiques par l'introduction de corrélations

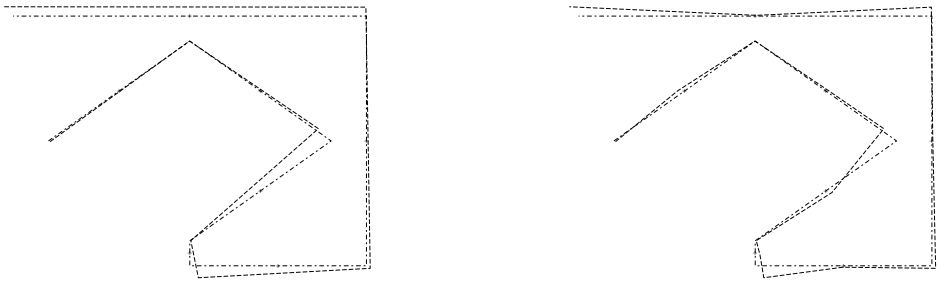

FIG. 2.7.: Suppression du biais parasite (à gauche : sans corrélation, à droite : avec corrélation)

3. Étude d'une application de calcul d'itinéraires

Nous reproduisons ici un article présenté à SSDBM'98 (10th International Conference on Scientific and Statistical Database Management) [Bon98], qui expose une étude par simulation de la sensibilité d'une application de calcul d'itinéraires à la qualité de la base de données utilisée.

3.1. Introduction

Geographic information data quality has been widely studied in the last fifteen years ([GG98]). This paper considers *geographic information* in spatial databases used in vector Geographic Information Systems.

A large number of parameters have been proposed to evaluate different aspects of geographic information quality. These aspects are gathered in *components*; we use the classification of the CEN (Comité Européen de Normalisation) [SC97]. We will distinguish between *lineage*, *temporal accuracy*, *logical consistency*, *positional accuracy* and *attribute accuracy*. Each component can be evaluated by several parameters. These components and parameters are defined and determined by the data producer. We will focus in this paper on attribute accuracy because this component has great impact on many applications and it has been less studied than positional accuracy for vector databases. Attribute accuracy deals with the classification of objects in the database and their presence or absence compared with *nominal ground*. We call *nominal ground* the abstraction consisting of the perfect database (with neither mistakes nor omissions) according to the specifications of the database [DF97].

As the need for quality highly depends on the use of geographic information, it is necessary to assess the impact of data uncertainty in geographical applications. Uncertainty in the database will propagate in applications and cause results to be biased. The aim of this paper is to evaluate the accuracy of the results of a geographical application knowing the quality parameters of the input database (Figure 3.1).

We present here a geographical sensitivity analysis on a vector road database. Known amounts of noise are introduced in the database by simulation — only on attributes for this study, but it could be extended to geometry —, and the impact of this noise on the

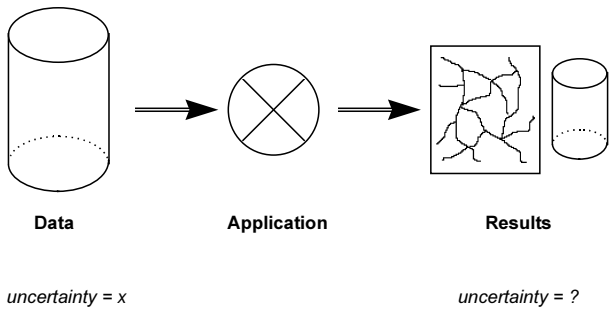

Figure 3.1.: How does quality propagate ?

application, which is an itinerary computation program, is assessed. The idea is to predict the uncertainty on the results given the input uncertainty. Next section defines more precisely sensitivity analysis and the quality parameters of a road database that we will handle. Then are described the geographical application itself and the simulation program used to create noise in the database. Last section presents results of the analysis and emphasizes problems that rise when performing such analysis.

3.2. Methodology

3.2.1. Strategy for the study

Attribute accuracy in vector databases is different from positional accuracy because little mathematical formalisation is possible. There are three main kinds of parameters: *deficit ratio*, *excess ratio* and *confusion ratio* on attributes [DF97]. There is a deficit when an object does not have an attribute that exists in nominal ground. There is an excess if an object has an attribute that does not exist in nominal ground. Confusion occurs when an attribute has a wrong value in the database compared with nominal ground. More precisely deficit ratio of attribute A from the geographic class C is the number of objects whose value of A is defined in nominal ground but undefined in the database, divided by the number of objects of the class C in nominal ground. A similar definition is given for excess ratio. Confusion ratio between attributes A_i and A_j is the number of objects from the nominal ground that have attribute A_j in the database instead of A_i, with $i \neq j$, divided by the number of objects in the class with the attribute A_i in the nominal ground [DF97]. These three parameters are gathered in a matrix called *confusion matrix* for each attribute. Figure 3.2 presents an example of confusion matrix.

We study the impact of noise in attributes on the results of the application. As most applications cannot be expressed mathematically, we decided to use a real application, and

Data set

Nominal ground	Number	Path 94	Roadslip 54	Stairs 47	Footbridge 16
Path	100	90%	5%	0%	0%
Roadslip	50	2%	90%	0%	4%
Stairs	50	6%	4%	90%	0%
Footbridge	10	0%	0%	0%	100%

Figure 3.2.: Confusion matrix for the attribute "Road type"

to study the evolution of results according to the noise amount in the database. Different data sets are created with gradual amounts of errors. The original data set in which noise is introduced becomes the reference. The results of the application are computed for each data set and are compared. This technique of analysis makes sense provided that the number of data sets is large enough. It has been called "geographical sensitivity analysis" by Lodwick et al. They define geographical sensitivity analysis as the "study of the effects of imposed perturbations (variations) on the inputs of a geographical analysis on the outputs of that analysis" [LMS90].

Geographical sensitivity analysis has been used by lots of researchers including Fisher [Fis91] and McMaster [McM96] on raster data and applications based on map overlay. These studies showed very interesting results and pointed out the difficulties that rise in such an approach of the problem, especially the mixing of continuous and discrete parameters. They made clear that semantics has various degrees of influence on results and that there is no isotropy of the problem. The results obtained for raster data encouraged us to apply sensitivity analysis to vector data.

Our application is an itinerary computation program on a vector road database. There is a real need for quality estimation of such applications [GJ98]. They are used as well by taxi drivers as by car constructors to design navigation tools, and should become as common and useful as maps. But it is necessary to know the quality of the results and the confidence one can have in it. For instance traffic-jam regulation requires minimum result precision to be efficient.

Road applications are adapted to our goals because they rely on databases with large numbers of different attributes (discrete, continuous, nominal, ordinal ...) and they intensively use these attributes. It allows investigating various aspect of attribute accuracy.

3.2.2. Data description

The database chosen for this study is Géoroute®, which was designed especially for road applications and itinerary computation. It is a vector database structured in road sections that exhaustively describes the French road network. Figure 3.3 partially illustrates the structure of the base. Each road section has many attributes, so that itinerary computations may take into account accessibility, condition of the road, number of roadways, average speed, speed limits, etc. The resolution of the database is higher in urban areas, to ensure the same level of detail everywhere: the acquisition scale is about 1/20,000 in urban zones and 1/50,000 elsewhere.

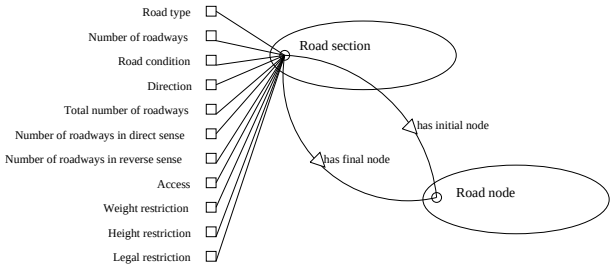

Figure 3.3.: Simplified structure of the road network of Géoroute®

The test area is heterogeneous. It covers urban areas, and sparser zones, with transition zones. There are motorways as well as dead ends, and some complicated roadslips. This diversity is required because the incidence of noise on calculation depends on the kind of landscape and many real applications have to deal with long itinerary passing through cities.

3.3. Implementation

3.3.1. Itinerary computation

A computer program was written to determine the shortest path between two points of the network. It is an adaptation of Dijskra's algorithm [Dij59]: two virtual vehicles (a car and a lorry) are put on the network and the shortest path between two points is computed in function of the attributes of each road section. The speed (or transport time) of each road section depends on its type, number of roadways, condition. Only sections whose attributes (weight or height restrictions ...) are compatible with the vehicle are taken into account [Cou97].

We get the shortest paths between any points of the network as well as travel time on any path. It allows producing new results. Given a node as an origin we can compute

travel time to any other point of the network. All the nodes at a time t of the origin constitute what is called an *isochrone*. These isochrones are useful because the conversion of travel time into altitudes enables to handle results as a relief. Isochrones are the basis for visualization and analysis.

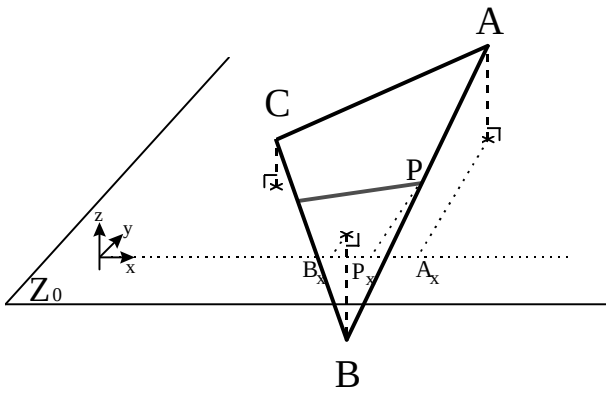

Figure 3.4.: Intersection of a facet and a plan

The strategy adopted to compute isochrones from the travel times given by Dijskra's algorithm is to approach the "relief" (with our analogy between travel time and altitude) by triangular facets, and to compute the intersections between these facets and an horizontal plan which are segments (Figure 3.4). We triangulate the network by a Delaunay triangulation [BY95], and draw continuous isochrones by joining the segments with an appropriate method.

The program was written in C++. It relies on the data structure of GéO2 (the Object-Oriented SIG developed at COGIT lab upon O2 DBMS) [DRS95] and uses its visualization facilities. The calculations are independent from GéO2, in order to be faster. The performance is good, and allows numerous runs with diverse data sets in reasonable time (less than a minute on 5,000 road sections).

3.3.2. Error simulation

A program was created then to simulate noise in the geographical database. The key point is to introduce controlled and realistic noise. As we study attribute accuracy noise is introduced in the attributes of the database.

We chose to disturb neither the completeness of the attribute *Road type*, so that the speed of each road section can be computed, nor the completeness of the class *Road section* which would modify the graph. This prevents from interacting with positional accuracy.

The strategy for introducing noise is quite difficult to determine. To introduce totally random noise is not a suitable strategy because it would not pass elementary controls of

the production process so it is not realistic. We clearly see that mistaking a motorway and a principal road is more likely than mistaking a motorway and a staircase, even if this last possibility actually occurs (it can be an unpredictable bug in the database for instance). We have to find probability distributions describing actual uncertainty. But errors on attributes seldom follow statistical laws and such discrete laws a re merely impossible to estimate.

Fisher [Fis91] suggested that noise on attributes should depend on the values of the attributes of the area. This is close to reality. However quality controls show that anything can happen as previously mentioned hence this strategy is not suitable for every attribute.

One can think to introduce noise proportionally to the number of elements that have the same attribute value. But counting the number of attributes in Géoroute® shows that a value is largely dominant, and that other values are rarely taken (Figure 3.5).

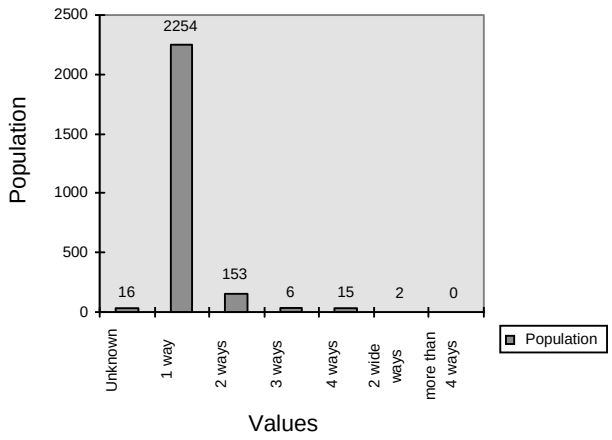

Figure 3.5.: Repartition of the values of the attribute "Number of roadways" in Géoroute®

To introduce noise this way prevents less common attributes from being changed when one value is too predominant. This situation is poorly adapted to our study, because the purpose is to take into account all attribute errors and not only the error on the dominant attribute.

We finally adopted a mix of two simple techniques. We introduce noise on attributes, with a global amount and different confusion ratios (fixed) for the attribute values. This way we can have instances of less common values, and we can measure the impact of these phenomena on results. If we do not specify the ratios between the values, the ratios are computed according to the population (we speak of uniform noise).

Let us take an example. The attribute *Direction* has four values (we indicate the number of road sections for each value in the parenthesis): *undetermined*(56), *both ways*(2033),

one-way direct(195), *one-way indirect*(146). One value is largely predominant, so we may want to decrease its effect to see the effect of the other values. A 10% noise on this attribute (global amount) will be introduced following four ratios, say 20%, 40%, 20% and 20% (the sum must be 100%). It means than 20% of the noise (20% of 10%) will be put on the first value of the attribute, 40% of the noise on the second value, and so on. The 40% ratio on the dominant value is smaller than the ratio calculated according to its population (uniform noise) that would be $\frac{2033}{56+2033+195+146} = 83,6\%$. This method increases the impact of other values.

3.4. Data analysis

3.4.1. Results characterization

Sensitivity analysis requires comparing the results obtained with various degrees of noise in the database. This comparison is quantitative and must fulfill conditions: the parameters chosen for comparison have to be significant from a geographical point of view and they have to be statistically discriminating. As results can be either maps or numbers or topological graphs there is no universal parameter.

The choice of parameters is determinant because it is impossible to draw any conclusion if we use inappropriate ones. If we try to model a relation between the quality parameters of the database and the results of an application this relation must have a geographical interpretation.

In our example results can be described by the number of road sections common to the shortest paths computed with each data set, by the difference between isochrones area, or by the variation of length of a given itinerary. We cannot determine *a priori* the best comparison parameter. The choice will depend on the model chosen for the analysis as well as on the user's requirements. A slight error on a travel time can be less critical than a longer (in term of distance) itinerary for a Fire Department but it could be the opposite for a taxi company for instance. Each model is designed for discrete and/or continuous parameters so we must adapt our parameters to fit the requirements of the analysis.

3.4.2. Dealing with discrete data

Most parameters of attribute accuracy are discrete (confusion ratio and agreement ratio), and some results are too. The evolution of a quality parameter when we increase noise in the data set is difficult to measure as Figure 3.6 illustrates. It represents the amount of confusion on an attribute depending on the global noise amount (noise increasing from 1% to 40% with uniform distribution on all attributes).

We see that we cannot distinguish between 10% noise and 30% noise looking at this data quality parameter.

Many results of the application are also discrete ones. The two main strategies to handle

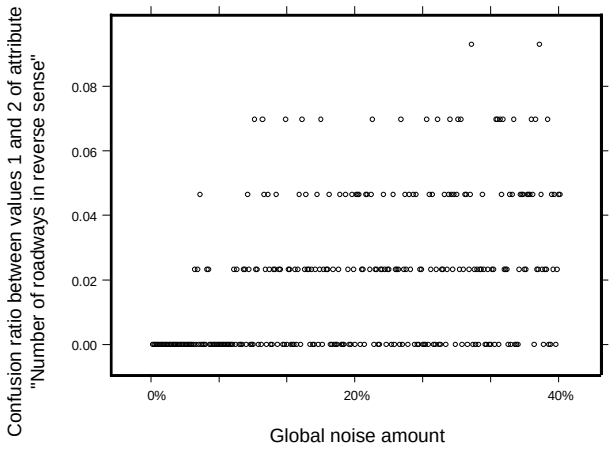

Figure 3.6.: Evolution of a discrete parameter with increasing noise

discrete data are to create continuous variables from the discrete ones (by calculating means on intervals for instance), or to use specific methods as logistic regression. The explanatory variables are imposed by the objectives of the study, which are to deduce the quality of the results from the existing parameters. These parameters are mainly discrete. The evaluation of the results has to be done with continuous variables when possible, because it facilitates the comparison between results and makes interpretation easier. Figure 3.7 shows that there is no apparent link between a discrete parameter (agreement ratio for attribute *number of roadways*) and a discrete result (ratio of common road sections between the different data sets).

3.4.3. Results of the simulation

We now present the results of our study deduced from the simulation. We computed 300 data sets with noise increasing from 1% to 40% (global amount) and distributions on attributes close to those given by quality control on actual data sets. We then ran the application with each data set.

The choice of this noise range is not realistic but it facilitates the interpretation of small phenomena that would be invisible otherwise, as confusions between less common attributes. Let us recall that a global amount of 10% on an attribute with 4 values at 25% ratio each means actually a 2.5% noise on each value.

When we analyze continuous variables (e.g. variation of an isochrone's area) computed with the results we note that the graphs have the same aspect for different parameters, with various dispersions. Figure 3.8 is a typical illustration of these graphs.

To study the imprecision on a path length the 300 data sets are gathered in 30 groups

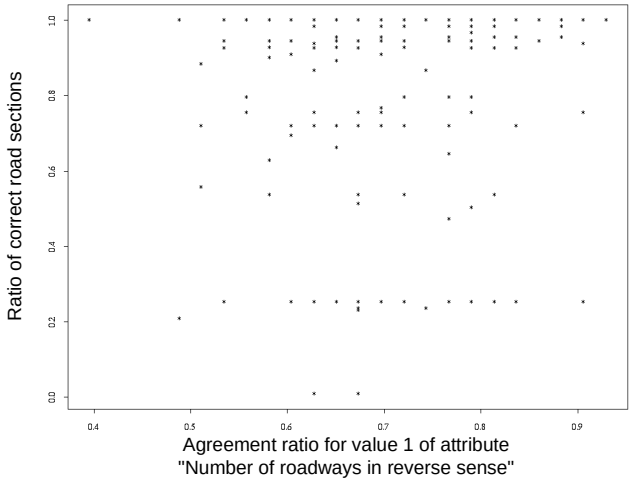

Figure 3.7.: Influence of noise on discrete results

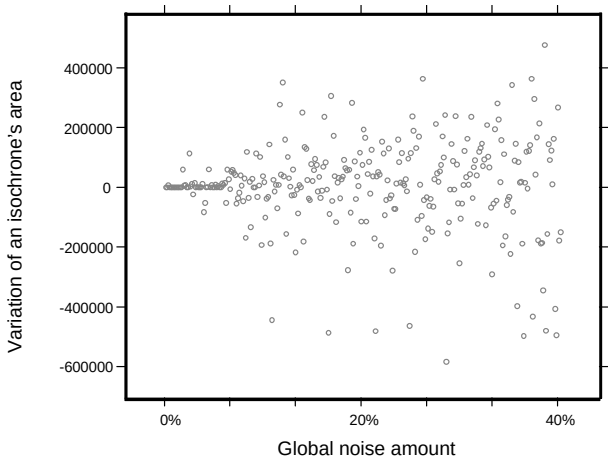

Figure 3.8.: Variation of an isochrone's area with increasing noise

according to their noise ratio. Then the median of the absolute value of the length discrepancy (compared with the reference) is calculated for each group. This imprecision measure can be explained with the agreement ratio of the dominant value of a critical attribute (road direction) as Figure 3.9 illustrates it.

Such graphs can be obtained for every quality parameter. It is possible to explain most of the imprecision by two or three quality parameters, because they are often correlated.

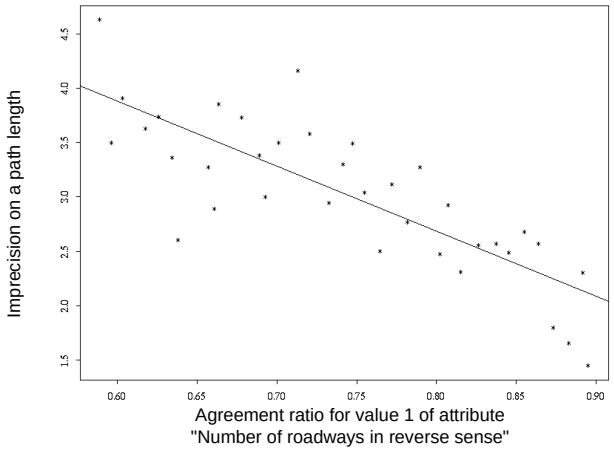

Figure 3.9.: Imprecision on path length depending on an attribute accuracy parameter

However the selection of the significant parameters is sometimes difficult. The main one is the agreement ratio of the most numerous value of the preponderant attribute (depending on the application). Other ones must be determined by comparing the slopes of the regressions.

3.5. Conclusion

Attribute accuracy in vector database is difficult to study but essentially because it has an important impact on geographical applications. There is a real need for quality estimation of the results of geographical applications and few tools to perform it. The question is: Can the quality of an application be predicted knowing the quality parameters of the geographic database?

We use for our research a road application. This choice was made because such applications use complex geographic databases with many attributes. We wrote a computer program to find the shortest path between two points, the length of any path and to calculate isochrones. This gave us various results to study.

We then wrote another program to simulate noise in a geographic database. This program lets us precisely control the amount of noise we introduce. We obtain that way many data sets with various levels of noise.

The idea for the analysis is to perform geographical sensitivity analysis. We try to see the evolution of quality in the application results when we increase the amount of noise in the database.

We must quantify the results in order to compare them. This can be done by defining

adequate parameters for the results. With a quasi continuous one we succeeded in linking the imprecision of itinerary computation results to the agreement ratio of the main value of a critical attribute for the application. It is however difficult to refine the analysis and to discover the other main significant parameters.

Chapitre III.

Étude des erreurs d'attributs

Nous présentons dans ce chapitre une étude probabiliste de l'impact des erreurs d'attributs sur une application de calcul d'itinéraires. Nous définissons un critère de qualité des résultats de l'application, et estimons ce critère en fonction des paramètres du modèle d'erreurs d'attributs présenté au chapitre I.

Nous proposons un modèle d'application de calcul d'itinéraires, et explicitons le critère de qualité des résultats dans le cadre de ce modèle (section 1). La modélisation retenue fait intervenir des développements de grandes déviations pour des sommes pondérées de variables discrètes. Après une brève présentation des résultats bien connus de grandes déviations (section 2), nous étendons un résultat de Book [Boo72] pour résoudre notre problème (section 3). Nous montrons sur des exemples la bonne qualité de notre approximation, en confrontant nos approximations à des simulations de Monte-Carlo . Nous précisons la position de nos résultats par rapport aux théorèmes de la littérature à la fin de cette section. Ces résultats font l'objet de deux articles [Bon01] [Bon02].

Les erreurs de géométrie seront considérées dans le chapitre suivant.

1. Modèle de l'application et critère de qualité des résultats

Nous reprenons l'étude de notre application géographique de calcul d'itinéraires en étudiant sa sensibilité aux erreurs d'attributs. Pour cela, nous proposons un modèle simple de cette application, et formulons un critère de qualité des résultats que nous mettons en relation avec les paramètres de qualité sémantique de la base de données géographiques.

1.1. Modèle de déplacement en zone urbaine

Nous proposons ici un modèle de calcul d'itinéraires en zone urbaine. Dans le cas d'une zone urbaine dense, la vitesse est officiellement limitée à 50 km/h, à l'exception de certaines voies rapides limitées à 70 km/h. Cependant, en centre-ville, ou dans des quartiers très résidentiels, la vitesse effective de parcours d'un tronçon est souvent très inférieure à 50 km/h (présence de passages protégés pour les piétons, ralentisseurs, priorités à droite, stops). En conséquence, les facteurs déterminants pour calculer un itinéraire sont le sens de circulation des rues, et le quartier qu'elles traversent. Le modèle peut être très simple, et affecter deux vitesses en fonction de l'emplacement de la rue ; une lente en zone densément peuplée, et une rapide ailleurs.

Pour pouvoir poursuivre la modélisation, nous devons envisager le cas d'un trajet relativement long, et celui d'un trajet court.

Dans le cas d'un trajet long, nous faisons l'hypothèse que le plus court chemin entre deux points éloignés (trajet long) comporte approximativement le même nombre de tronçons quels que soient les taux d'erreur de sens de circulation et de classification (tronçon rapide ou lent), pourvu qu'ils restent faibles. En effet, dans la plupart des cas, les détours imposés par la présence d'erreurs de sens de circulation sont relativement courts, et de longueurs comparables aux trajets originels, par rapport aux longueurs totales des trajets. On peut pour s'en convaincre imaginer une ville américaine, sous la forme d'une grille régulière. Une erreur sur un sens interdit ne change pas le nombre de tronçons parcourus. En effet, quand on se déplace sur une grille régulière entre deux points de cette grille, le nombre de tronçons parcourus ne dépend pas de l'itinéraire emprunté pourvu qu'on se déplace toujours dans la direction du point à atteindre. (Remarquons toutefois que se déplacer toujours dans la direction du point à atteindre n'est pas possible dans le cas

de très nombreux sens interdits.) Ces hypothèses ont été vérifiées expérimentalement par simulation.

Le modèle pour un trajet long consiste ainsi à écrire que le temps de parcours T_{AB} du plus court chemin entre A et B, mettant en jeu k tronçons, s'écrit :

$$T_{AB} = \sum_{i=1}^{k} l_i/V_i,$$

avec l_i longueur du tronçon i et V_i sa vitesse de parcours, k ne dépendant pas de la qualité de la base. Les erreurs d'attributs dans la base se traduisent par des erreurs sur les vitesses V_i.

Dans le cas d'un trajet court, un détour peut modifier de façon significative le nombre de tronçons parcourus. Nous sommes conduits à écrire le modèle pour un trajet court de la façon suivante. Le temps de parcours T_{AB} du plus court chemin entre A et B, s'écrit :

$$T_{AB} = \sum_{i=1}^{K} l_i/V_i,$$

avec l_i longueur du tronçon i et V_i sa vitesse de parcours, K étant une variable aléatoire distribuée selon la loi estimée du nombre de tronçons de l'itinéraire, loi dépendant des erreurs présentes dans la base de données. Ce modèle est étudié dans le chapitre V.

Remarque. Nous pouvons adapter ces modèles au cas des zones rurales. Pour un trajet en zone rurale, relativement long, le nombre de tronçons parcourus pourra être supposé constant. En revanche, tous les attributs seront utilisés pour déterminer la vitesse de parcours de chacun des tronçons de route, ce qui conduira à des vitesses plus variées que dans le cas urbain. La loi des vitesses aura alors plus de deux modalités.

1.2. Critère de qualité des résultats de l'application

Pour simplifier le problème, nous supposons que la détermination de l'itinéraire le plus rapide ne dépend pas de la qualité de la base de données utilisée, vus les faibles taux d'erreurs constatés en contrôle qualité (de 3% à 5%). Nous disposons donc pour relier deux destinations d'un itinéraire fixé unique composé d'un certain nombre de tronçons de route k. Le nombre de tronçons et les longueurs de chaque tronçon sont des paramètres déterministes. En revanche, la vitesse de parcours de chaque tronçon est déterminée à l'aide des valeurs des attributs du tronçon, qui sont l'objet de notre modèle d'erreurs d'attributs. Nous pouvons supposer pour simplifier les notations que les vitesses de chaque tronçon sont enregistrées dans un attribut unique (figure 1.1).

Nous étudions la qualité des résultats de l'application en considérant la probabilité que l'erreur relative en temps sur le trajet complet dépasse un seuil fixé $\eta > 0$. En notant avec un indice R les grandeurs calculées dans la référence, et un indice D celles calculées dans

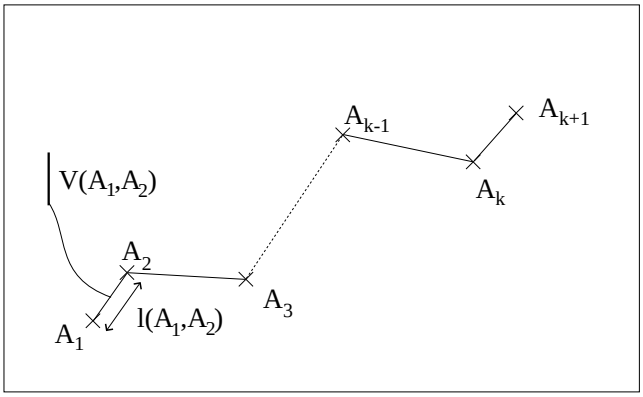

FIG. 1.1.: Temps de parcours d'un chemin

le jeu de données, l'erreur relative en temps s'écrit :

$$\left|\frac{T_R - T_D}{T_R}\right|,$$

soit, en faisant apparaître les k tronçons et les longueurs et les vitesses :

$$\left|\frac{\sum_{i=1}^{k} l_i(1/V_{Ri} - 1/V_{Di})}{\sum_{i=1}^{k} l_i/V_{Ri}}\right|.$$

Or nous cherchons à estimer la probabilité

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right),$$

qui est la somme des deux probabilités

$$P\left(\sum_{i=1}^{k} l_i\left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} - \eta\frac{1}{V_{Ri}}\right) > 0\right) + P\left(\sum_{i=1}^{k} l_i\left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} + \eta\frac{1}{V_{Ri}}\right) < 0\right).$$

Chacune des deux probabilités fait apparaître les sommes de k variables indépendantes et identiquement distribuées $X_i = (1/V_{Ri} - 1/V_{Di} - \eta \times 1/V_{Ri})$ et $Y_i = (1/V_{Ri} - 1/V_{Di} + \eta \times 1/V_{Ri})$, pondérée par les longueurs. Si nous centrons les variables, notre probabilité d'erreur devient

$$P\left(\sum_{i=1}^{k} l_i(X_i - E(X_i)) > -E(X_1)\sum_{i=1}^{k} l_i\right) + P\left(\sum_{i=1}^{k} l_i(Y_i - E(Y_i)) < -E(Y_1)\sum_{i=1}^{k} l_i\right),$$

ce qui fait apparaître des probabilités de grandes déviations. Pour estimer chacune de ces deux probabilités, nous proposons d'utiliser des résultats asymptotiques : nous supposons que le nombre de tronçons k tend vers l'infini, et utilisons des développements de

grandes déviations. Comme dans la pratique k est fixé, cette approche nécessite que le développement utilisé converge suffisamment vite pour qu'il soit précis avec des valeurs de k relativement petites. Nous verrons que cette contrainte nécessite d'utiliser des développements exacts, par opposition aux développements du logarithme de ces probabilités.

2. Introduction aux développements de grandes déviations

2.1. Principe de la méthode

Le but des techniques de grandes déviations est d'obtenir une estimation de la probabilité que la somme de n variables aléatoires dépasse largement sa moyenne, probabilité en principe petite. Le contexte est donc une suite de variables $X_1, X_2, \ldots, X_n$. On note $S_n = \sum_{i=1}^{n} X_i$, $c_n = E(S_n)$, et on veut estimer

$$P(S_n \geqslant c_n + a_n),$$

avec $a_n \to \infty$ quand $n \to \infty$.

Une littérature extrêmement abondante existe sur le sujet. Le cas le plus simple est celui de la somme de variables i.i.d, et a d'abord été résolu pour $a_n = na$, avec a réel positif fixé, par Chernoff [Che52] et Bahadur et Ranga Rao [BR60]. On pourra se référer à [CS85] et [CS93] pour des extensions au cas non i.i.d, et à [Ste78] pour la résolution dans le cas multidimensionnel. De nombreux livres couvrent le sujet de façon assez large ([Var84] [DS89] [SS91] [Pet75] [Pet95] [Cra70] [Jen95] parmi d'autres). L'étude de grandes déviations pour a_n tendant vers ∞ moins vite que n (on parle aussi de moyennes déviations), en particulier $a_n = a\sqrt{n}$, avec a réel positif fixé, fait également l'objet d'un grand nombre d'articles. On pourra se référer à [Nag79] pour une revue de la question. Un traitement unifié du problème quelle que soit la vitesse est développé dans [Bro87] et [BM94]. Des extensions existent également pour des objets probabilistes plus complexes, ou pour $P(S_n \in nA)$ avec A borélien arbitraire [BB].

On distingue généralement entre deux types de résultats de grandes déviations, que nous désignons respectivement par *développement logarithmique* et *développement au premier ordre exact*. (On trouve dans la littérature russe les termes de *rough large deviation* et de *precise (sharp) large deviation*.) Les développements logarithmiques, à la suite de l'article de Chernoff [Che52], sont obtenus en établissant que

$$1/n \log P(S_n > c_n + na) \to \log \rho,$$

et donnent l'expression de ρ en fonction de a. Les développements au premier ordre exact,

présentés pour la première fois par Bahadur et Ranga Rao [BR60], s'écrivent

$$P(S_n > c_n + na) = \frac{\rho^n}{(2\pi n)^{1/2}} b_n(1 + o(1)),$$

avec ρ identique au coefficient de Chernoff, et $\log b_n = O(1)$ quand $n \to \infty$.

Nous utilisons dans la suite des développements au premier ordre exact, et montrons qu'ils sont nécessaires pour obtenir une précision raisonnable sur l'estimation de la probabilité de grande déviation.

La méthode utilisée pour obtenir un développement de grandes déviations est d'utiliser une approximation normale d'une nouvelle distribution, transformée exponentielle de la distribution étudiée. En effet, les approximations normales sont généralement précises au centre des distributions. La transformée exponentielle est choisie de telle sorte qu'un développement au voisinage de sa moyenne donne une bonne approximation du développement de La queue de la distribution étudiée. Cette technique est couramment appelée *approximation de point-selle* [Jen95] [Kol97].

Les résultats présents dans la littérature utilisent, après une transformation exponentielle pour faire apparaître le terme ρ^n, un développement de type théorème central limite local, appelé également développement d'Edgeworth, ou une formule d'inversion de type Fourier. Nous avons retenu la première méthode, qui présente à notre avis l'avantage d'une grande polyvalence.

2.2. Transformation exponentielle

Pour établir les résultats de grandes déviations, on a recours en général à une transformation exponentielle. La transformation exponentielle d'une variable aléatoire X de mesure de probabilité P consiste à définir une famille exponentielle de mesures de probabilité P_θ, $\theta \in \Theta$ vérifiant :

$$\frac{dP_\theta}{dP}(\omega) = \frac{e^{\theta X(\omega)}}{E(e^{\theta X})}.$$

Le paramètre θ est ensuite choisi de manière appropriée au problème. Le lecteur pourra se référer à [Jen95] pour une étude détaillée des propriétés de cette famille exponentielle.

2.3. Développements d'Edgeworth

Le principe de ces développements est de renforcer le théorème central limite dans le cas d'existence de moments d'ordre supérieur. Soient $X_1, X_2, \ldots, X_n$ n variables i.i.d de fonction de répartition F. L'outil de base pour ces décompositions est la fonction caractéristique, que nous notons φ. Elle est définie ainsi :

$$\varphi(\zeta) = \int_{-\infty}^{\infty} e^{i\zeta x} F\{dx\}$$

et s'écrit dans le cas d'existence d'une densité :

$$\varphi(\zeta) = \int_{-\infty}^{\infty} e^{i\zeta x} f(x)dx.$$

Nous notons μ_k le $k^{\text{ième}}$ moment défini par :

$$\mu_k = \int_{-\infty}^{\infty} x^k F\{dx\}.$$

Les développements étant du type théorème central limite, nous aurons besoin de la densité $\mathfrak{n}$ de la loi normale centrée réduite $N(0,1)$, et de sa fonction de répartition $\mathfrak{N}$:

$$\mathfrak{n}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad \mathfrak{N}(x) = \int_{-\infty}^{x} \mathfrak{n}(t)dt$$

Notons que sa transformée de Fourier est $e^{-\frac{1}{2}\zeta^2}$.

Nous supposons dans la suite que $\mu_1 = 0$ et $\mu_2 = \sigma^2$. Nous notons $F_n(x)$ la fonction de répartition de $(X_1 + \ldots + X_n)/\sqrt{n}$, et notons f_n la densité de F_n si elle existe.

Nous citons quelques théorèmes utiles sans donner leurs preuves. Le lecteur pourra se référer à [Fel70], et trouvera dans la suite la preuve d'extensions de ces théorèmes dans des cas non i.i.d.

Théorème 1. *Supposons que μ_3 existe et que $|\varphi|^\nu$ est intégrable pour un $\nu \leqslant 1$. Alors f_n existe pour $n \geqslant 1$ et*

$$f_n(x) - \mathfrak{n}(x) - \frac{\mu_3}{6\sigma^3\sqrt{n}}(x^3 - 3x)\mathfrak{n}(x) = o\left(\frac{1}{\sqrt{n}}\right) \tag{2.1}$$

uniformément en x quand $n \to \infty$.

Nous pouvons écrire le même type de développement que 2.1 pour des fonctions de répartition. Nous obtenons ainsi

Théorème 2.

$$F_n(x) - \mathfrak{N}(x) - \frac{\mu_3}{6\sigma^3\sqrt{n}}(1 - x^2)\mathfrak{n}(x) = o\left(\frac{1}{\sqrt{n}}\right) \tag{2.2}$$

En fait, cette expression est vraie même si F n'a pas de densité, à l'exception des distributions sur un sous-ensemble discret de $\mathbb{R}$ (variables treillis). Pour de telles distributions, une légère modification du théorème 2 est nécessaire :

Théorème 3. *Pour des variables treillis, on a le développement 2.2 en remplaçant F_n (définie sur une grille de pas d_n) par $F_n^{\#}$ convolution de F_n par la distribution triangulaire sur $[-d_n/2, d_n/2]$].*

3. Grandes déviations pour des sommes pondérées de variables i.i.d

Nous reproduisons ici (sections 3.1 à 3.4) un article en cours de révision au *Journal of Applied Probability*. Cet article présente l'application de notre technique à un calcul d'itinéraires en zone urbaine avec la base de données Géoroute de l'IGN. Dans ce cas précis, l'utilisation d'un développement asymptotique avec un nombre de termes relativement petit impose d'utiliser un développement au premier ordre exact, comme l'illustrent les applications numériques. La comparaison des résultats obtenus par notre méthode à l'estimation du critère de qualité des résultats par une méthode de Monte-Carlo avec un logiciel de calcul d'itinéraires illustre la qualité de notre modèle et la précision de notre approximation.

Les preuves des résultats de grandes déviations qui sont présentées sommairement dans l'article sont données dans la section 3.5. La correspondance avec le cas i.i.d est établie dans la section 3.6. Il est aussi précisé de manière détaillée en quoi nos résultats sont nouveaux dans la section 3.7.

3.1. Introduction and statement of the problem

The aim of this paper is to present an application of a large deviation expansion to the following problem. Geographical databases are typically subject to small errors which accumulate when they are used to evaluate travel times between two locations. Our goal is to evaluate the impact of these errors on such computations.

To this end, we model the errors via a simple probabilistic model and evaluate the probability that the relative travel time computation error exceeds a threshold. As will become obvious later on, this approach leads to a large deviation problem for weighted sums of independent identically distributed (i.i.d) discrete random variables. Unfortunately no appropriate result is available in the literature to allow us to solve this problem. This motivates the need of generalizing the large deviation results of Book [Boo72] to the case of non-absolutely continuous random variables (including the case of lattice variables).

Our results turn out to improve on the usual logarithmic rate for large deviations [Boo73]. Several results in the literature come close to our theorems. We mention [CS93], [Hög79], [Hwa98], [Sau80], [SS76], [Wol80] among others. For example, observe that con-

ditions (a–b) of theorems 3.3 and 3.4 in [CS93] do not apply in the present setup.

The remainder of our paper is organized as follows. The next section is devoted to the exposition of our geographical error model and to the derivation of the subsequent large deviation probabilities. Section 3 presents an application on real life data. Section 4 contains our large deviation and local central limit theorems.

3.2. Geographical model and reduction to a large deviation problem

Geographical databases contain topographic and geographical information represented by objects. These objects (e.g., roads) have a spatial location and attributes describing their main characteristics (e.g., number of lanes, road name or number, etc.) Due to the measurement errors, geographical databases are far from being error-free. There is a considerable amount of literature on techniques describing and modelling the corresponding uncertainty. However, little is known about the influence of data quality on the results of spatial analysis using such databases. We are concerned here with the evaluation of the impact of attribute errors upon the results of travel time computation. Therefore, we need to model the attribute errors as well as to assess error propagation in travel time computation.

The most common method to assess the quality of a geographical database (called *dataset*) is to compare it to a more accurate database considered to be the *reference*. For an attribute with K modalities denoted by $1, 2, \ldots, K$, a geographical object has value $r \in \{1, \ldots, K\}$ in the reference and $d \in \{1, \ldots, K\}$ in the dataset.

Observations are samples (r, d) of a random variable $Z = (R, D)$. Our model assumes that the errors which correspond to distinct objects are independent. The distribution of Z is then a discrete law defined through the matrix

$$P(Z = (r, d)) = p_{rd} \quad \forall (r, d) \in \{1, \ldots, K\}^2.$$

Let us assume now that the distribution of the error for an attribute value is uniform on the set of possible alternative values. This corresponds for instance to the case of encoding errors. This assumption may be reformulated into:

$$P((R, D) = (r, d)) = \begin{cases} p_{rr} & \text{if } d = r \\ p_r & \text{otherwise} \end{cases} \quad \forall (r, d) \in \{1, \ldots, K\}^2. \tag{3.1}$$

To reduce the number of parameters of this model, we make use of the multiplication law:

$$P((R, D) = (r, d)) = P(R = r) P(D = d \mid R = r) \quad \forall (r, d) \in \{1, \ldots, K\}^2.$$

The probability $P(R = r)$ corresponds to the choice of the attribute value r in the reference. It can be approximated by N_r/N, where N_r denotes the number of objects with attribute

value r in the reference and N the total number of objects in the reference. The probability $P(D = d \mid R = r)$ is the probability that an object has the attribute value d in the dataset given the corresponding reference value r. In view of the above formula, we impose that $p_{rr} + (K-1)p_r = N_r/N$, and we let $\theta_r = 1 - Np_{rr}/N_r$, $\forall r \in \{1, \ldots, K\}$. For $\theta \in]0,1[$, we obtain the following model:

$$P((R,D) = (r,d)) = \begin{cases} p_{rr} = (1-\theta_r)\frac{N_r}{N} \\ p_r = \frac{\theta_r}{K-1}\frac{N_r}{N} \end{cases} \quad \forall (r,d) \in \{1, \ldots, K\}^2,$$

where θ_r is an indicator of the fraction of errors pertaining to the attribute value r. This approach leads naturally to a parametric probabilistic model for geographical database errors that has been tested on real databases. The number of parameters can be further reduced by making the simplifying assumption of equality of all θ_r's, $r \in \{1, \ldots, K\}$.

We now propose a model for travel times computation. Let us consider a fixed itinerary between two locations. This travel path is summarized in the geographical database by a set of n short straight road sections. Each road section has a position on the map, and some attributes (numbers of lanes, road type, ...). We denote by T_R the travel time computed from the reference, and by T_D the travel time computed from the dataset. As road sections are usually short, we make the simplifying assumption that speed is uniform on each road section. Given this condition, the speed remains as the only attribute of interest.

We will estimate the relative accuracy of the travel time T_D with respect to the reference travel time T_R through

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right) = P\left(T_R - T_D - \eta T_R > 0\right) + P\left(T_R - T_D + \eta T_R < 0\right), \tag{3.2}$$

where η is a specified threshold value.

For each road section of length l with speed attribute V, the travel time is $T = l/V$. Thus, if we index the quantity that corresponds to the ith road section by i, the probability (3.2) can be rewritten in terms of lengths l_{ni} and speeds V_i as follows:

$$P\left(\sum_{i=1}^{n} l_{ni}\left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} - \eta\frac{1}{V_{Ri}}\right) > 0\right) + P\left(\sum_{i=1}^{n} l_{ni}\left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} + \eta\frac{1}{V_{Ri}}\right) < 0\right). \tag{3.3}$$

Now, for both probabilities in (3.3), our error model states that the random variables $X_i = (1/V_{Ri} - 1/V_{Di} - \eta \times 1/V_{Ri})$, respectively $Y_i = (-1/V_{Ri} + 1/V_{Di} - \eta \times 1/V_{Ri})$, are i.i.d. The probabilities in (3.3) can therefore be interpreted as large deviation probabilities of weighted sums of the form:

$$P\left(\sum_{i=1}^{n} l_{ni}(X_i - E(X_i)) > -E(X_1)\sum_{i=1}^{n} l_{ni}\right) + P\left(\sum_{i=1}^{n} l_{ni}(Y_i - E(Y_i)) > -E(Y_1)\sum_{i=1}^{n} l_{ni}\right), \tag{3.4}$$

and evaluated as such.

3.3. Results and discussions

We have applied this approach to Géoroute, a vector road database of IGN (Institut Géographique National – National Geographic Institute of France). The equality of all θ_r's to some $\theta \in]0,1[$ has been successfully tested on homogeneous areas. The estimation of θ depends upon the geographical area, and has been found to vary between 3% and 6%. We will assume in the following that $\theta \in]0.03, 0.07[$.

We computed travels within an urban area. Road section lengths are in general close to 50 metres. The speed is supposed to be either v_1 =50 km/h or v_2 =20 km/h, depending on the nature of the road section determined by its attributes. Note that the computation may be refined by considering more than two speeds, but the methodology remains essentially the same. The repartition between quick road sections and slow road sections has been estimated in the reference to be close to $N_1/N = 0.3$ and $N_2/N = 0.7$. Then X_1 and Y_1 have four modalities, and are defined by:

$$\begin{cases} P(X_1 = -\eta/v_1) = P(Y_1 = -\eta/v_1) = (1-\theta)N_1/N \\ P(X_1 = -\eta/v_2) = P(Y_1 = -\eta/v_2) = (1-\theta)N_2/N \\ P(X_1 = (1-\eta)/v_1 - 1/v_2) = P(Y_1 = (-1-\eta)/v_1 + 1/v_2) = \theta N_1/N \\ P(X_1 = (1-\eta)/v_2 - 1/v_1) = P(Y_1 = (-1-\eta)/v_2 + 1/v_1) = \theta N_2/N \end{cases}$$

It is noteworthy that we so obtain $E(X_1) < 0$ and $E(Y_1) < 0$.

We use our large deviation approximation for weighted sums of lattice random variables (see Section 4) to estimate by (3.4) the probability (3.2) that the relative error on travel time exceeds a threshold η. Lengths l_{ni} being rewritten by $l_{ni} = l * a_{ni}$, with $\sum_{i=1}^n a_{ni}^2 = 1$, $A_n = \sum_{i=1}^n a_{ni}$, and letting $X = X_1$ and $Y = Y_1$, Theorem 1 in Section 4 allows us to write, as $n \to \infty$,

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right) = \frac{1}{\sqrt{2\pi}} \frac{d_n^{(X)} e^{-h_n^{(X)} d_n^{(X)}}}{\bar{\sigma}_n^{(X)} (1 - e^{-h_n^{(X)} d_n^{(X)}})} e^{h_n^{(X)} E(X) A_n} \left[\prod_{i=1}^n \phi^{(X)}(h_n^{(X)} a_{ni})\right] (1 + o(1)) + \frac{1}{\sqrt{2\pi}} \frac{d_n^{(Y)} e^{-h_n^{(Y)} d_n^{(Y)}}}{\bar{\sigma}_n^{(Y)} (1 - e^{-h_n^{(Y)} d_n^{(Y)}})} e^{h_n^{(Y)} E(Y) A_n} \left[\prod_{i=1}^n \phi^{(Y)}(h_n^{(Y)} a_{ni})\right] (1 + o(1)). \quad (3.5)$$

The moment-generating function $\phi^{(X)}$ of X is given by the formula:

$$\phi^{(X)}(t) = (1-\theta)(N_1/N)e^{(-\eta/v_1)t} + (1-\theta)(N_2/N)e^{(-\eta/v_2)t} + \theta(N_1/N)e^{((1-\eta)/v_1 - 1/v_2)t} + \theta(N_2/N)Ne^{((1-\eta)/v_2 - 1/v_1)t}.$$

A similar expression is obtained for $\phi^{(Y)}$, and the parameters $h_n^{(X)}$, $\bar{\sigma}_n^{(X)}$, $d_n^{(X)}$, $h_n^{(Y)}$, $\bar{\sigma}_n^{(Y)}$, $d_n^{(Y)}$ are computed numerically.

Observe that the logarithmic rates of [Boo73] for the probabilities in (3.4) are given when $n \to \infty$ by

$$\log P\left(\sum_{i=1}^{n} l_{ni}(X_i - E(X_i)) > -E(X)\sum_{i=1}^{n} l_{ni}\right) \to$$

$$h_n^{(X)} E(X) A_n + \sum_{i=1}^{n} \log \phi^{(X)}(h_n^{(X)} a_{ni})$$

and by the same expression with X replaced by Y.

Figure 3.1 presents the results of this analysis for an itinerary with 30 road sections. In the upper graph θ varies from 3% to 7 % with a fixed error threshold $\eta = 5\%$, while in the lower graph η varies from 4% to 10% with a fixed dataset error $\theta = 5\%$. We have also estimated the tails of the distributions of X and Y by a Monte-Carlo simulation with 10,000 repetitions. We have added on each graph the results of the Monte-Carlo simulation and of the logarithmic approximation. Both graphs illustrate the good precision of the approximation given by our large deviation methodology. An approximation with better precision than that following from logarithmic rates is required for small values of n, as shown on Figure 3.2 (n varying from 20 to 200).

To compare our results to those of an actual geographical analysis, we have written an itinerary computation program using Dijkstra's algorithm [Dij59]. This software computes the shortest path between any chosen locations in a geographical database. We want to estimate the probability (3.2) with this software via Monte-Carlo methods. We consider now our database as the reference, and we generate by simulation perturbed datasets with error level θ according to our attribute error model. With each perturbed dataset we determine the shortest path for a given itinerary, and then we compute $(T_R - T_D)/T_R$. To estimate the tail of the empirical distribution of $(T_R - T_D)/T_R$ for a given itinerary, we generate 1000 independent datasets for each error level θ.

We have simulated perturbed datasets for different values of θ. We have fixed θ equal to 3, 4, 5, 6 and 7%, according to the remark at the beginning of this section. We have studied the relative travel time error for 20 itineraries. These itineraries include between 30 and 50 road sections. We present the results in figure 3.3, and we compare them to our large deviation approximation computed with $n = 30$, and all road sections of the same length. On this example, the approximation is quite correct because it provides an accurate majoration of travel time errors estimated by Monte-Carlo simulation.

3.4. Large deviation theorems

Let $X = X_1, X_2, \ldots$ be a sequence of i.i.d nondegenerate random variables, let $\{a_{nk} : 1 \leqslant k \leqslant n, 1 \leqslant n < \infty\}$ be a double array of non-negative real numbers such that

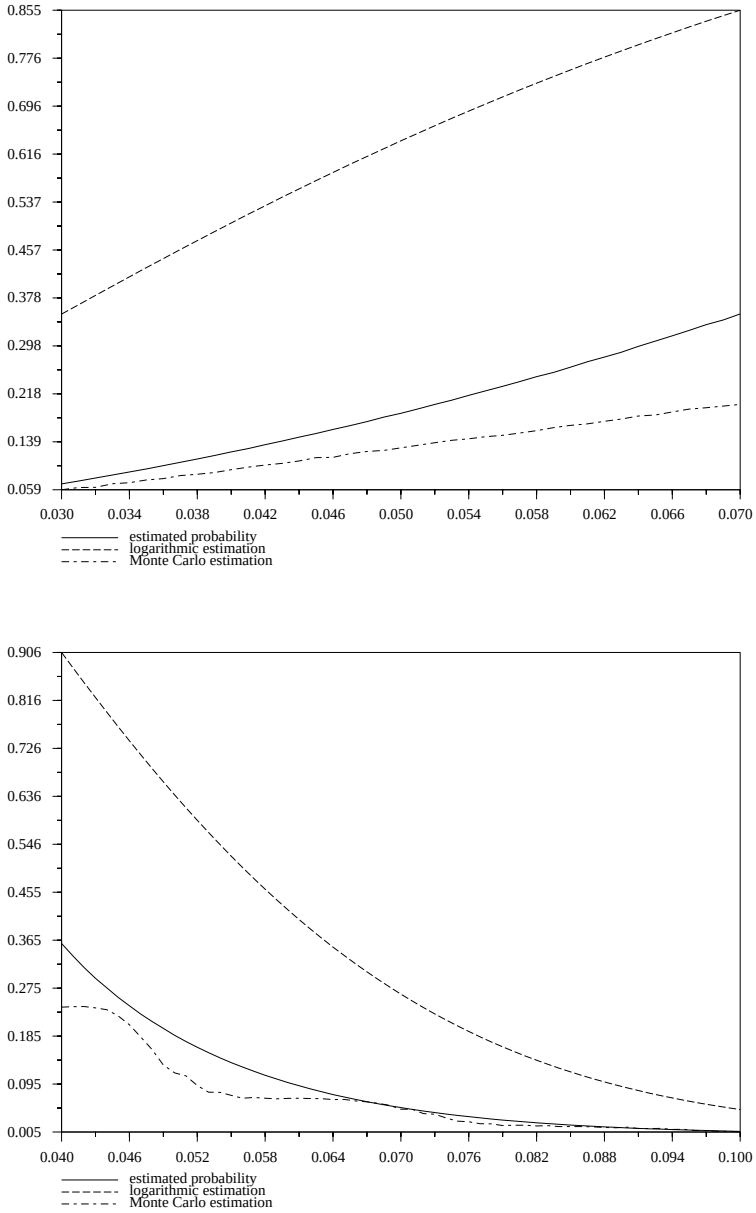

Figure 3.1.: Probability of the error exceeding the threshold η in terms of θ with $\eta = 5\%$ (top) and in terms of η with $\theta = 5\%$ (bottom)

$\sum_{k=1}^{n} a_{nk}^2 = 1$ and let c be a positive real constant. We want to study the asymptotic behaviour of $P\left(\sum_{k=1}^{n} a_{nk}X_k > c\sum_{k=1}^{n} a_{nk}\right)$. Let $S_n = \sum_{k=1}^{n} a_{nk}X_k$ and $A_n = \sum_{k=1}^{n} a_{nk}$. We assume that $E(X) = 0$ and $E(X^2) = 1$. We denote by $F(x) = P(X \leqslant x)$ the

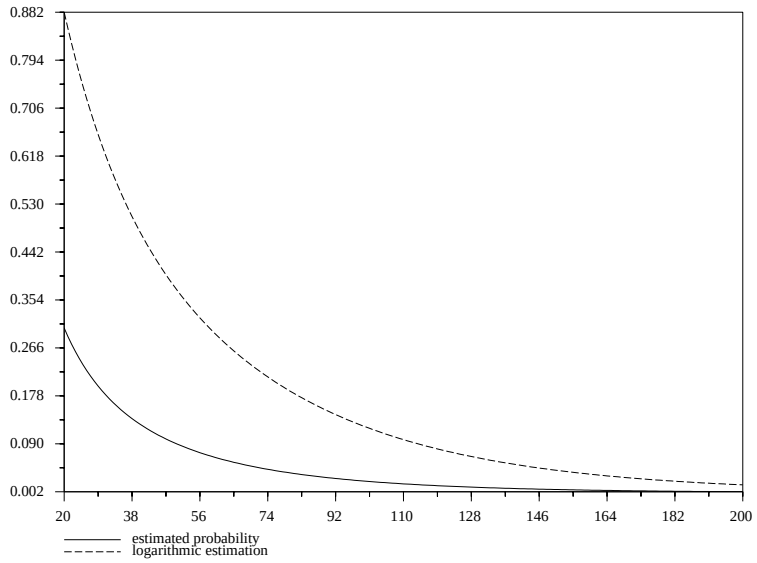

Figure 3.2.: Probability of error exceeding the threshold η=5% in terms of n, with $\theta = 5\%$

distribution function of X and by $\phi(t) = E(e^{tX})$ the moment-generating function of X.

Our results are established under the same assumptions as those used by Book [Boo72]. Let $\sigma_n = \max\{a_{nk} : 1 \leqslant k \leqslant n\}$.

Condition I There exist α and θ with $0 < \alpha \leqslant 1$, $0 < \theta \leqslant 1$, such that, for all n sufficiently large, at least αn of the a_{nk}'s exceed or equal $\theta\sigma_n$.

The second condition is a natural extension of Chernoff's condition. We denote by $Q(t) = \phi'(t)/\phi(t)$, for $t \in \mathbb{R}$, the cumulant of X. We note that Q is increasing and denote by Q^{-1} the Cramér's index of X (the inverse of Q with respect to composition).

Condition II $\phi(t)$ is finite on a $\mathcal{I} \supseteq (-B, B)$ for some $B > 0$, Q assumes the value $\frac{c}{\alpha\theta}$ at some point, and $B_0 = \frac{1}{\theta}Q^{-1}(\frac{c}{\alpha\theta}) \in \mathcal{I}$.

To state our results, let us introduce the random variables $Y_{nk} = a_{nk}X_k - ca_{nk}$. We observe that $P(S_n > cA_n) = P(\sum_{k=1}^{n} Y_{nk} > 0)$. Moreover, the distribution function of Y_{nk} is $H_{nk}(y) = F(ya_{nk}^{-1} + c)$ and the moment-generating function of Y_{nk} is $\phi_{nk}(h) = e^{-hca_{nk}}\phi(ha_{nk})$, where under Condition I $\phi_{nk}(h)$ exists for $|h| < B\sigma_n^{-1}$. Let us define an "associated" distribution function $\bar{H}_{nk}(h)$ by

$$d\bar{H}_{nk}(h)(y) = \frac{e^{hy}}{\phi_{nk}(h)} dH_{nk}(y)$$

for $0 < h < B\sigma_n^{-1}$. Let $\bar{Y}_{n1}(h), \bar{Y}_{n2}(h), \ldots$ be a sequence of real random variables distributed according to $\bar{H}_{nk}(h)$, and let $\bar{S}_n(h) = \sum_{k=1}^{n} \bar{Y}_{nk}(h)$.

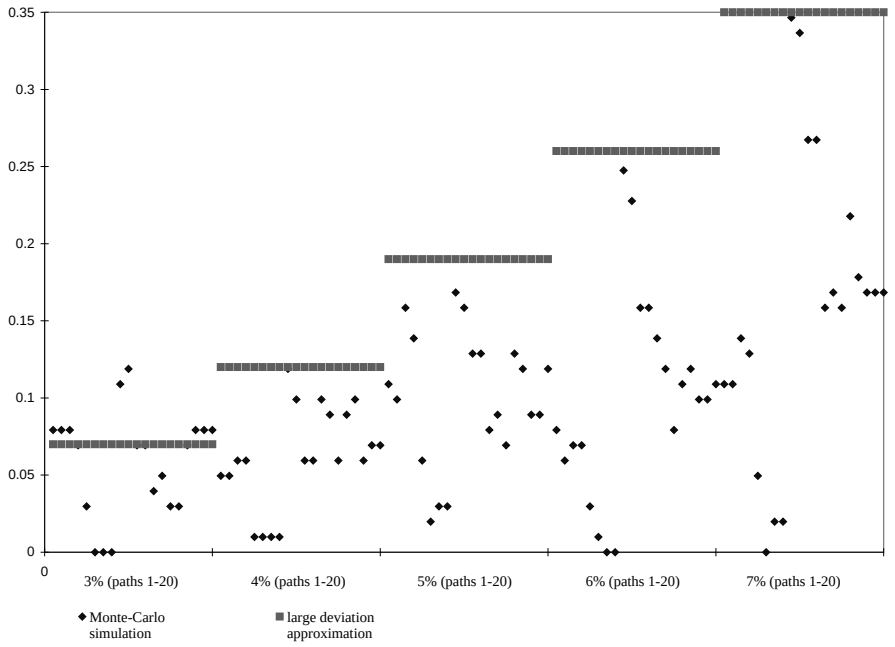

Figure 3.3.: Probability of error exceeding the threshold η=5% in terms of θ, with θ=3%, 4%, 5%, 6% and 7% for 20 simulated itineraries

We can now state our theorem:

Theorem 1. *Assume the fulfilment of Conditions I and II. Fix $c > 0$, and let h_n be solution of the equation $E(\bar{S}_n(h_n)) = 0$. Set $\bar{\sigma}_n^2 = \mathrm{Var}(\bar{S}_n(h_n))$. Then, as as $n \to \infty$,*

$$P(S_n > cA_n) = \frac{1}{\sqrt{2\pi}} \frac{1}{\bar{\sigma}_n h_n} e^{-h_n c A_n} \left[\prod_{k=1}^{n} \phi(h_n a_{nk}) \right] (1 + o(1)),$$

if X_1 is not lattice, and

$$P(S_n > cA_n) = \frac{1}{\sqrt{2\pi}} \frac{d_n e^{-h_n d_n}}{\bar{\sigma}_n (1 - e^{-h_n d_n})} e^{-h_n c A_n} \left[\prod_{k=1}^{n} \phi(h_n a_{nk}) \right] (1 + o(1)),$$

if S_n is lattice, with d_n the span of S_n.

Remark 1. Notice that if X_1 is not lattice, then S_n is not lattice. On the other hand, if X_1 is lattice, then S_n is lattice unless there exists at least a_{ni} and a_{nj} such as $a_{ni}/a_{nj} \in \mathbb{R} - \mathbb{Q}$. Then this last case is covered by our theorem.

A local central limit theorem is required to prove theorem 1. Our version is very close to Theorem 7 found in chapter 6 of [Pet75]. For each $n = 1, 2, \ldots$, let $X_{n1}, \ldots, X_{nn}$ be a sequence of independent random variables with distribution functions F_{nk} and moment-generating functions $\omega_{nk}(\zeta)$. Let $S_n = \sum_{k=1}^{n} X_{nk}$, and $E(S_n^2) = s_n^2$. Let us denote by F_n

the distribution function of S_n/s_n, by $\mathfrak{N}$ the distribution function of a standard Gaussian variable and by $\mathfrak{n}(x)$ its density.

Theorem 2. *If $X_{n1}, \ldots, X_{nn}$, for $n = 1, 2, \ldots$, satisfy:*

(i) *$E(X_{nk}) = 0$ for all n and for all k;*

(ii) *$|\omega_{n1}(\zeta) \cdots \omega_{nn}(\zeta)| \leqslant \frac{\varepsilon_n(\delta,a)}{\sqrt{n}}$ for $a > \zeta > \delta > 0$ with $\varepsilon_n(\delta, a) \to 0$ as $n \to \infty$;*

(iii) *$cn < s_n^2 < Cn$, where $c > 0$ and $C > 0$;*

(iv) *$E(X_{nk}^4)$ is uniformly bounded;*

then

$$F_n(x) = \mathfrak{N}(x) + \frac{\mu_3^{(n)}}{6s_n^3}(1 - x^2)\mathfrak{n}(x) + n^{-\frac{1}{2}} r_n(x)$$

with $\mu_3^{(n)} = \sum_{k=1}^n E(X_{nk}^3)$ and $r_n(x) \to 0$ uniformly with respect to x as $n \to \infty$.

If S_n is defined on a lattice of span d_n, let $F_n^{\#}$ be the convolution of F_n by the distribution function of the triangular distribution on $[-d_n/2, d_n/2]$.

Theorem 3. *If S_n is defined on a lattice, Theorem 2 holds under assumptions (i), (iii), and (iv), with F_n replaced by $F_n^{\#}$.*

Sketch of the proof of Theorem 1. We use the same technique as Bahadur and Ranga Rao [BR60] to reduce the large deviation problem to an expansion of distribution functions.

Lemma 1. *If $\bar{H}_n(h)(y) = P(\bar{S}_n(h) \leqslant y)$ then*

$$P(S_n > cA_n) = e^{-hcA_n} \left[\prod_{k=1}^n \phi(ha_{nk}) \right] I_n(h),$$

with

$$I_n(h) = h \int_0^\infty e^{-hy} [\bar{H}_n(h)(y) - \bar{H}_n(h)(0)] dy$$

if $\bar{S}_n(h)$ is not lattice, and

$$I_n(h) = h \int_0^\infty e^{-hy} [\bar{H}_n^*(h)(y) - \bar{H}_n(h)(0)] dy$$

if $\bar{S}_n(h)$ is lattice, $\bar{H}_n^*(h)$ denoting the following function:

$$\begin{aligned} \bar{H}_n^*(h)(y) &= \frac{1}{2}[\bar{H}_n(h)(y) + \bar{H}_n(h)(y-)] && \text{if } y \text{ is on a vertex of the lattice} \\ &= \bar{H}_n(h)(y) && \text{otherwise.} \end{aligned}$$

Conditions I and II ensure the existence of a sequence of real numbers $\{h_n : 1 \leqslant n < \infty\}$ replacing h in the definition of the "associated" variables such that $E(\bar{S}_n(h_n)) = 0$ and $\mathrm{Var}(\bar{S}_n(h_n))$ is uniformly bounded for all n. For simplicity of notation, we omit the h_n in the sequel.

Then we use Theorem 2 and Theorem 3 to approximate $\bar{H}_n$. Consider the real random variables $X_{nk} = A_n[\bar{Y}_{nk} - E(\bar{Y}_{nk})]$. These new variables satisfy $E(X_{nk}) = 0$ and $\sum_{k=1}^n E(X_{nk}^2) = A_n^2\bar{\sigma}_n^2$. We observe that $\bar{H}_n(x) = P((\sum_{k=1}^n X_{nk})/s_n \leqslant x\bar{\sigma}_n^{-1})$, and so the local central limit theorem gives us the required approximation:

$$\bar{H}_n(x) = \mathfrak{N}(x\bar{\sigma}_n^{-1}) + \frac{\mu_3^{(n)}}{6s_n^3}(1 - (x\bar{\sigma}_n^{-1})^2)\mathfrak{n}(x\bar{\sigma}_n^{-1}) + n^{-\frac{1}{2}}r_n(x\bar{\sigma}_n^{-1})$$

with $r_n(x) \to 0$ uniformly with respect to x as $n \to \infty$ and with $\bar{H}_n$ being replaced by $\bar{H}_n^{\#}$ in the case of lattice variables.

Introducing the expansion of $\bar{H}_n$ in I_n, we obtain that:

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{1}{h_n\bar{\sigma}_n}(1 + o(1)),$$

in the non-lattice case, and that:

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{e^{-h_nd_n}d_n}{\bar{\sigma}_n(1 - e^{-h_nd_n})}(1 + o(1)),$$

in the lattice case, which concludes the proof. □

3.5. Preuves des théorèmes

Preuve du lemme 1. En se rappelant que $P(S_n > cA_n) = P(\sum_{k=1}^n Y_{nk} > 0)$ nous avons :

$$\begin{aligned}
P(S_n > cA_n) &= \int\cdots\int_{x_1+\cdots+x_n>0} dH_{n1}(x_1)\cdots dH_{nn}(x_n) \\
&= e^{-hcA_n}\left[\prod_{k=1}^n \phi(ha_{nk})\right]\int\cdots\int_{z_1+\cdots+z_n>0} e^{-h(z_1+\cdots+z_n)}d\bar{H}_{n1}(z_1)\cdots d\bar{H}_{nn}(z_n) \\
&= e^{-hcA_n}\left[\prod_{k=1}^n \phi(ha_{nk})\right]\int_{0<y\leqslant\infty} e^{-hy}d\bar{H}_n(y) \\
&= e^{-hcA_n}\left[\prod_{k=1}^n \phi(ha_{nk})\right]\int_{0<y\leqslant\infty}\int_{y\leqslant z\leqslant\infty} he^{-hz}dzd\bar{H}_n(y).
\end{aligned}$$

L'application du théorème de Fubini complète la preuve si X n'est pas sur une grille. Si X est sur une grille, nous pouvons écrire :

$$P(S_n > cA_n) = e^{-hcA_n}\left[\prod_{k=1}^{n}\phi(ha_{nk})\right]\frac{1}{2}\Bigg[\int_{0<y\leqslant\infty}\int_{y\leqslant z\leqslant\infty} he^{-hz}dzd\bar{H}_n(y) + \int_{0<y\leqslant\infty}\int_{y<z\leqslant\infty} he^{-hz}dzd\bar{H}_n(y)\Bigg]$$

$$P(S_n > cA_n) = e^{-hcA_n}\left[\prod_{k=1}^{n}\phi(ha_{nk})\right]\frac{1}{2}\Bigg[h\int_0^\infty e^{-hy}[\bar{H}_n(y) - \bar{H}_n(0)]dy + h\int_0^\infty e^{-hy}[\bar{H}_n(y-) - \bar{H}_n(0)]dy\Bigg]$$

$$P(S_n > cA_n) = e^{-hcA_n}\left[\prod_{k=1}^{n}\phi(ha_{nk})\right]h\int_0^\infty e^{-hy}[\bar{H}_n^*(y) - \bar{H}_n(0)]dy.$$

□

Remarquons qu'une majoration de $P(S_n \geqslant cA_n)$ s'obtient directement avec cette écriture puisque $0 \leqslant \bar{H}_n(y) - \bar{H}_n(0) \leqslant 1$ pour tout n et pour tout $y \geqslant 0$ et donc $I_n \leqslant 1$.

Nous allons d'abord contrôler à l'aide du choix de la suite h_n les quantités $E(\bar{S}_n)$ et $\mathrm{Var}(\bar{S}_n)$. Nous pouvons calculer la fonction génératrice des moments de $\bar{Y}_{nk}$, que nous notons $\bar{\phi}_{nk}$:

$$\bar{\phi}_{nk}(t) = \int e^{ty}d\bar{H}_{nk}(y) = \frac{1}{\phi_{nk}(h)}\int e^{(t+h)y}dH_{nk}(y) = \frac{\phi_{nk}(t+h)}{\phi_{nk}(h)}.$$

La fonction génératrice des moments permet de calculer les deux premiers moments de $\bar{S}_n$ car $E(\bar{S}_n) = \sum_{k=1}^n \bar{\phi}'_{nk}(0)$ et $\mathrm{Var}(\bar{S}_n) = \sum_{k=1}^n[\bar{\phi}''_{nk}(0) - (\bar{\phi}'_{nk}(0))^2]$. Il suffit de calculer les dérivées pour obtenir :

Lemme 1.

$$E(\bar{S}_n) = \sum_{k=1}^{n} a_{nk}Q(ha_{nk}) - cA_n$$

et

$$\mathrm{Var}(\bar{S}_n) = \sum_{k=1}^{n} a_{nk}^2 Q'(ha_{nk}).$$

Les conditions I et II ont été introduites pour assurer l'existence d'une suite de réels $\{h_n : 1 \leqslant n < \infty\}$ remplaçant h dans la définition des variables «associées», tels que $E(\bar{S}_n)$ soit égal à 0 pour tout n et que $\mathrm{Var}(\bar{S}_n)$ soit uniformément bornée pour tout n. Ces deux propriétés sont l'objet des deux lemmes suivants :

Lemme 2. *Sous les conditions I et II, pour tout entier positif n, il existe une solution $h = h_n$ de l'équation $E(\bar{S}_n) = 0$, et cette solution vérifie les inégalités suivantes :*

$$b_0 = Q^{-1}(c\alpha\theta^2) \leqslant h_n\sigma_n \leqslant \theta^{-1}Q^{-1}(\frac{c}{\alpha\theta}) = B_0.$$

Preuve. Notons $Q_n^*(h) = \sum_{k=1}^n a_{nk}Q(ha_{nk})$. Ainsi $E(\bar{S}_n) = Q_n^*(h) - cA_n$. La fonction Q_n^* vérifie $Q_n^*(0) = 0$ et est continue sur son ensemble de définition. Comme la condition I assure que $Q_n^* \geqslant \alpha n\theta\sigma_n Q(h\theta\sigma_n)$ il suffit de prouver qu'il existe un h pour lequel $\alpha n\theta\sigma_n Q(h\theta\sigma_n) = cA_n$ pour avoir par continuité l'existence d'un h_n tel que $E(\bar{S}_n) = 0$. Cette condition est équivalente à $Q(h\theta\sigma_n) = \frac{cA_n}{\alpha\theta n\sigma_n}$. Or $\frac{cA_n}{\alpha\theta n\sigma_n} \leqslant \frac{cn\sigma_n}{\alpha\theta n\sigma_n} = \frac{c}{\alpha\theta}$ et cette dernière valeur est supposée être prise dans la condition II donc l'existence du h et donc du h_n est établie. La majoration découle du fait que $cA_n = Q_n^*(h_n) \geqslant \alpha n\theta\sigma_n Q(h_n\theta\sigma_n)$ et donc $c \geqslant \alpha\theta n\sigma_n A_n^{-1}Q(h_n\theta\sigma_n)$. En remarquant que $n\sigma_n A_n^{-1} \geqslant n\sigma_n(n\sigma_n)^{-1} = 1$ nous obtenons $c \geqslant \alpha\theta Q(h_n\theta\sigma_n)$. Le fait que Q soit croissante donne $h_n\sigma_n \leqslant \theta^{-1}Q^{-1}(\frac{c}{\alpha\theta})$. Pour la minoration, nous pouvons écrire que $cA_n = Q_n^*(h_n) \leqslant n\sigma_n Q(h_n\sigma_n)$ et donc que $c \leqslant n\sigma_n A_n^{-1}Q(h_n\sigma_n) \leqslant (\alpha\theta^2)^{-1}Q(h_n\sigma_n)$, puisque $\sigma_n A_n \geqslant \sum_{k=1}^n a_{nk}^2 \geqslant \alpha n\theta^2\sigma_n^2$. □

Lemme 3. *Sous les conditions I et II, pour tout entier positif n, pour $h = h_n$ solution de l'équation $E(\bar{S}_n) = 0$, il existe deux réels $d_0^2 > 0$ et $D_0^2 < \infty$ tels que*

$$d_0^2 \leqslant \operatorname{Var}\bar{S}_n \leqslant D_0^2.$$

Preuve. Comme X_1 est une variable non dégénérée, $Q^{'}(t)$, qui est la variance de la variable «associée» $\bar{X}_1$, est strictement positive pour tout t. Ceci assure que $d_0^2 = \min\{Q^{'}(z) : 0 \leqslant z \leqslant B_0\}$ est strictement positif et donc que $\operatorname{Var}(\bar{S}_n) = \sum_{k=1}^n a_{nk}^2 Q^{'}(h_n a_{nk}) \geqslant d_0^2 \sum_{k=1}^n a_{nk}^2 = d_0^2$. Le fait que $\phi(0) = 1$ et que $\phi(B_0) < \infty$ assure la finitude de $D_0^2 = \max\{Q^{'}(z) : 0 \leqslant z \leqslant B_0\}$ et donc $\operatorname{Var}(\bar{S}_n) \leqslant D_0^2$. □

Nous notons par la suite $\operatorname{Var}\bar{S}_n = \bar{\sigma}_n$.

Preuve du théorème 2. La preuve de ce théorème repose sur l'étude de la transformée de Fourier de $D_n(x) = F_n(x) - G_n(x)$ avec $G_n(x) = \mathfrak{N}(x) - \frac{\mu_3^{(n)}}{6s_n^3}(1-x^2)\mathfrak{n}(x)$. En notant ω_{kn} la fonction caractéristique de X_{kn}, la transformée de Fourier de D_n s'écrit :

$$e^{v_n\left(\frac{\zeta}{s_n}\right)} - e^{-\frac{1}{2}\zeta^2} - \frac{v_n^{'''}(0)}{6s_n^3}i^3\zeta^3 e^{-\frac{1}{2}\zeta^2}, \tag{3.6}$$

avec

$$v_n(\zeta) = \sum_{k=1}^n \log\omega_{nk}(\zeta).$$

Nous pouvons montrer aisément que $v_n(0) = 1$, $v_n^{'}(0) = 0$, $v_n^{''}(0) = i^2 s_n^2$ et $v_n^{'''}(0) = i^3\mu_3^{(n)}$. Remarquons que $|G_n^{'}(x)| \to 0$ uniformément en x quand $n \to \infty$. En effet, $E(X_{nk}^3) \leqslant M$ car par hypothèse $E(X_{nk}^4)$ est uniformément bornée donc $\mu_3^{(n)} \leqslant nM$. Comme $cn < s_n^2 < Cn$, $\frac{\mu_3^{(n)}}{6s_n^3} \leqslant M/6\sqrt{n}$.

Soit $\varepsilon > 0$ fixé. Nous choisissons une constante a suffisamment grande pour que $|G_n^{'}(x)| < \varepsilon a$ pour pour tout x et pour tout n. D'après un théorème de lissage (page 538 de [Fel70]), nous pouvons écrire :

$$|D_n(x)| \leqslant \int_{-as_n}^{as_n} \left| \frac{e^{v_n\left(\frac{\zeta}{s_n}\right)} - e^{-\frac{1}{2}\zeta^2} - \frac{v_n^{'''}(0)}{6s_n^3} i^3\zeta^3 e^{-\frac{1}{2}\zeta^2}}{\zeta} \right| d\zeta + \frac{24\varepsilon}{\pi s_n}. \tag{3.7}$$

Nous séparons cette intégrale en deux domaines d'intégration. Le premier est le domaine défini par $\delta s_n \leqslant |\zeta| \leqslant as_n$, et le second par $|\zeta| < \delta s_n$, δ étant un réel positif fixé que nous précisons plus loin. La contribution des intervalles $\delta s_n \leqslant |\zeta| \leqslant as_n$ à l'intégrale (3.7) est inférieure à

$$\log\left(\frac{a}{\delta}\right) \frac{\varepsilon_n(\delta, a)}{\sqrt{n}} + \int_{\delta s_n \leqslant |\zeta| \leqslant as_n} \frac{e^{-\frac{1}{2}\zeta^2}}{\zeta} \left(1 + \left| \frac{v_n^{'''}(0)}{6s_n^3}\zeta^3 \right|\right) d\zeta. \tag{3.8}$$

car par hypothèse $|\omega_{1n}(\zeta)\cdots\omega_{nn}(\zeta)| = \frac{\varepsilon_n(\delta,a)}{\sqrt{n}}$ uniformément en ζ pour $\zeta > \delta > 0$. Or $cn < s_n^2 < Cn$ donc le terme intégral de (3.8) est un petit o de n'importe quelle puissance de n. Nous en déduisons que (3.8) s'écrit

$$\log\left(\frac{a}{\delta}\right) \frac{\varepsilon_n(\delta, a)}{\sqrt{n}} + o(1/\sqrt{n}). \tag{3.9}$$

Pour le domaine $|\zeta| < \delta s_n$, en posant

$$\psi(\zeta) = v_n(\zeta) + \frac{1}{2}s_n^2\zeta^2$$

l'intégrande de (3.7) se réécrit

$$\frac{e^{-\frac{1}{2}\zeta^2}}{\zeta} \left| e^{\psi\left(\frac{\zeta}{s_n}\right)} - 1 - \frac{v_n^{'''}(0)}{6s_n^3} i^3\zeta^3 \right| d\zeta \tag{3.10}$$

et nous allons l'estimer à l'aide de l'inégalité suivante tirée de [Fel70] :

$$|e^\alpha - 1 - \beta| = |(e^\alpha - e^\beta) + (e^\beta - 1 - \beta)| \leqslant (|\alpha - \beta| + \frac{1}{2}\beta^2)e^\gamma, \tag{3.11}$$

avec $\gamma \geqslant \max(|\alpha|, |\beta|)$, pour α et β arbitraires, réels ou complexes. Nous pouvons faire un développement de Taylor au troisième ordre de ψ, en développant chacun des n termes de la somme. Pour obtenir un développement indépendant de n, nous utilisons le fait que $E(X_{nk}^4)$ est uniformément bornée, ce qui assure que $v_n^{'''}$ a une dérivée uniformément bornée au voisinage de l'origine. Ceci nous permet de déduire l'existence d'un δ tel que

$$\left| \psi\left(\frac{\zeta}{s_n}\right) - \frac{v_n^{'''}(0)}{6s_n^3} i^3\zeta^3 \right| < \sum_{k=1}^{n} \varepsilon_k \left|\frac{\zeta}{s_n}\right|^3 < \varepsilon n \left|\frac{\zeta}{s_n}\right|^3 \tag{3.12}$$

pour $|\zeta| < \delta s_n$. Nous prenons δ suffisamment petit pour avoir également

$$\left|\psi\left(\frac{\zeta}{s_n}\right)\right| < \frac{1}{4}\zeta^2, \qquad \left|\frac{v_n'''(0)}{6s_n^3}\zeta^3\right| \leqslant \frac{1}{4}\zeta^2$$

pour $|\zeta| < \delta s_n$. Avec ce choix de δ nous majorons l'intégrale (3.7) sur le domaine $|\zeta| < \delta s_n$ en utilisant la formule (3.11) par :

$$\int_{|\zeta|<\delta s_n} e^{-\frac{1}{4}\zeta^2}\left|\frac{\varepsilon n}{s_n^3}|\zeta|^2 + \frac{(\mu_3^{(n)})^2}{72 s_n^6}|\zeta|^5\right| d\zeta. \tag{3.13}$$

Comme $cn < s_n^2 < Cn$ et $\varepsilon_n(\delta, a) \to 0$ pour tout $\delta > 0$ nous pouvons choisir n assez grand pour avoir $\log\left(\frac{a}{\delta}\right)\frac{\varepsilon_n(\delta,a)}{\sqrt{n}} < \frac{\varepsilon}{\sqrt{n}}$ et l'intégrale (3.13) inférieure à $1000\varepsilon/\sqrt{n}$. Nous avons ainsi montré que pour tout x

$$|D_n(x)| \leqslant \frac{24\varepsilon}{\pi c\sqrt{n}} + \frac{\varepsilon}{\sqrt{n}} + \frac{1000\varepsilon}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

et comme ε est arbitraire, nous en concluons que $D_n(x) = o(1/\sqrt{n})$ uniformément en x. □

Preuve du théorème 3. Supposons que les F_{nk} sont concentrées sur les grilles $a_{nk} \pm kd_{nk}$, $k \in \mathbb{Z}$. Alors F_n est concentrée sur une grille $a_n \pm kd_n$, et comme s_n est de l'ordre de $\sqrt{n}$, d_n est de l'ordre de $1/\sqrt{n}$. Notons $G^{\#}$ la convolution de G par la distribution triangulaire sur $[-d_n/2, d_n/2]$, soit

$$G^{\#}(x) = \frac{2}{d_n}\int_{-d_n/2}^{d_n/2}\left(1 - \frac{2|y|}{d_n}\right)G(x-y)dy.$$

En notant M le maximum de $|G''|$, un développement de Taylor à l'ordre 2 de G au point x permet d'écrire que

$$|G^{\#}(x) - G(x)| < \frac{1}{24}Md_n^2 = O(1/n)$$

puisque d_n est de l'ordre de $1/\sqrt{n}$, et donc pour prouver le théorème il suffit d'établir que

$$|F_n^{\#}(x) - G^{\#}(x)| = o(1/\sqrt{n}).$$

Comme une convolution correspond à une multiplication pour les transformées de Fourier, l'équation (3.7) permet d'écrire que

$$|F_n^{\#}(x) - G^{\#}(x)| \leqslant \int_{-as_n}^{as_n}\left|\frac{e^{v_n\left(\frac{\zeta}{s_n}\right)} - e^{-\frac{1}{2}\zeta^2} - \frac{v_n'''(0)}{6s_n^3}\zeta^3 e^{-\frac{1}{2}\zeta^2}}{\zeta}\right||\nu_n(\zeta)|d\zeta + \frac{24\varepsilon}{\pi s_n}. \tag{3.14}$$

avec $\nu_n(\zeta) = \frac{\sin^2(\frac{1}{2}d_n\zeta)}{\left(\frac{1}{2}d_n\zeta\right)^2}$ fonction caractéristique de la loi triangulaire. Nous pouvons appliquer tous les arguments de la démonstration précédente, en ajoutant l'argument supplémentaire

$$\int_{\delta s_n}^{as_n} \frac{|e^{v_n\left(\frac{\zeta}{s_n}\right)}\nu_n(\zeta)|}{\zeta} d\zeta = o(1/\sqrt{n}). \tag{3.15}$$

Or

$$\int_{\delta s_n}^{as_n} \frac{|e^{v_n\left(\frac{\zeta}{s_n}\right)}\nu_n(\zeta)|}{\zeta} d\zeta = \frac{4}{(d_n s_n)^2} \int_{\delta}^{a} \frac{\left|e^{v_n(y)} \sin^2\left(\frac{d_n s_n y}{2}\right)\right|}{y^3} dy. \tag{3.16}$$

Or la fonction $e^{v_n(y)}$ a pour période $\frac{2\pi}{d_n s_n}$, de même que $\sin^2\left(\frac{d_n s_n y}{2}\right)$, donc il suffit de prouver que

$$\int_{\delta}^{\delta+\frac{2\pi}{s_n d_n}} \frac{\left|(\omega_{n1}(y)\cdots\omega_{nn}(y)) \sin^2\left(\frac{d_n s_n y}{2}\right)\right|}{y^3} dy = o(1/\sqrt{n}). \tag{3.17}$$

Écrivons tout d'abord que

$$\int_{\delta}^{\delta+\frac{2\pi}{s_n d_n}} \frac{\left|(\omega_{n1}(y)\cdots\omega_{nn}(y)) \sin^2\left(\frac{d_n s_n y}{2}\right)\right|}{y^3} dy \leqslant \frac{1}{\delta^3} \int_{\delta}^{\delta+\frac{2\pi}{s_n d_n}} \left|(\omega_{n1}(y)\cdots\omega_{nn}(y)) \sin^2\left(\frac{d_n s_n y}{2}\right)\right| dy$$

Le problème se situe en $\frac{2\pi}{d_n s_n}$, où le produit des fonctions caractéristiques vaut 1. Décomposons cette intégrale en deux domaines, l'intervalle $I_n = \left[\frac{2\pi}{d_n s_n} - \frac{1}{2n^{1/4}}, \frac{2\pi}{d_n s_n} + \frac{1}{2n^{1/4}}\right]$, et le complémentaire de cet intervalle. Dans l'intervalle I_n, nous majorons le produit des fonctions caractéristiques par 1 et ainsi

$$\int_{I_n} \sin^2\left(\frac{d_n s_n y}{2}\right) dy = \int_{I_n} (y^2 + (o(y^2))dy = o(1/\sqrt{n}).$$

En dehors de cet intervalle, nous majorons le sinus au carré par 1 et chacun des ω_{nk} par $1 - \frac{\sigma_{nk} y}{2}$, et l'intégrale tend donc vers 0 plus vite que n'importe quelle puissance de n. Le théorème est ainsi établi. □

Nous allons maintenant appliquer le théorème central limite local pour estimer $\bar{H}_n$. Normalisons d'abord les variables $\bar{Y}_{nk}$ de façon à pouvoir appliquer correctement le théorème. En effet, si on se ramène au cas i.i.d en prenant tous les a_{nk} égaux à $1/\sqrt{n}$ (et donc $A_n = \sqrt{n}$), nous constatons que les variables Y_{nk} sont égales à $\frac{1}{\sqrt{n}}(X_k - c)$ et donc qu'elles tendent toutes vers 0. La normalisation correcte est en fait $A_n a_{nk}(X_k - c a_{nk})$. Nous allons écrire le théorème central limite local pour les variables $X_{nk} = A_n[\bar{Y}_{nk} - E(\bar{Y}_{nk})]$. Ces

nouvelles variables vérifient $E(X_{nk}) = 0$ et $\sum_{k=1}^{n} E(X_{nk}^2) = A_n^2 \operatorname{Var} \bar{S}_n = A_n^2 \bar{\sigma}_n^2 = s_n^2$. Nous notons ω_{nk} la fonction caractéristique de X_{nk} et F_{nk} sa fonction de répartition. Remarquons que $\bar{H}_n(x) = P(\bar{S}_n \leqslant x) = P((\sum_{k=1}^{n} X_{nk})/s_n \leqslant x\bar{\sigma}_n^{-1})$, et que le théorème précédent va bien nous permettre d'approcher la quantité voulue.

Pour l'appliquer, nous devons vérifier une condition sur s_n.

Lemme 4. *Sous les conditions I et II, pour tout entier positif n, il existe deux réels positifs c et C tels que*

$$cn < s_n^2 < Cn.$$

Preuve. Par définition, $s_n^2 = A_n^2 \bar{\sigma}_n^2$. Le lemme 3 nous permet d'écrire l'encadrement suivant :

$$d_0^2 \left(\sum_{k=1}^{n} a_{nk} \right)^2 \leqslant s_n^2 \leqslant D_0^2 \left(\sum_{k=1}^{n} a_{nk} \right)^2.$$

La condition I donne l'encadrement d'A_n suivant :

$$\alpha n \theta \sigma_n \leqslant A_n \leqslant n\sigma_n.$$

Or $\sum_{k=1}^{n} a_{nk}^2 = 1$ donc

$$\alpha n (\theta \sigma_n)^2 \leqslant 1 \leqslant n\sigma_n^2.$$

Nous en déduisons l'encadrement :

$$\alpha\theta\sqrt{n} \leqslant A_n \leqslant \frac{\sqrt{n}}{\theta\sqrt{\alpha}},$$

et donc

$$d_0^2 \alpha^2 \theta^2 n \leqslant s_n^2 \leqslant \frac{D_0^2}{\alpha\theta^2} n,$$

et nous avons établi ainsi que $cn < s_n^2 < Cn$. □

Nous devons également vérifier que $E(X_{nk}^4)$ est uniformément bornée.

Lemme 5. *Sous les conditions I et II, pour tout n et pour tout k, $E(X_{nk}^4)$ est uniformément bornée.*

Preuve. Nous sommes dans les conditions de validité du lemme 2, donc $b_0 \leqslant h_n \sigma_n \leqslant B_0$. Nous obtenons par le calcul que

$$E(X_{nk}^4) = A_n^4 a_{nk}^4 G(h_n a_{nk}),$$

avec

$$G(t) = -3 \left[\frac{\phi'(t)}{\phi(t)} \right]^4 + 6 \left[\frac{\phi'(t)}{\phi(t)} \right]^2 \left[\frac{\phi''(t)}{\phi(t)} \right] - 4 \left[\frac{\phi'(t)}{\phi(t)} \right] \left[\frac{\phi^{(3)}(t)}{\phi(t)} \right] + \left[\frac{\phi^{(4)}(t)}{\phi(t)} \right].$$

Comme c'est le quatrième moment d'une distribution non dégénérée, nous savons que $G(h_n a_{nk}) > 0$ pour $a_{nk} > 0$. Or $G(t)$ est continue sur le fermé $0 \leqslant t \leqslant B_0$, et atteint donc son maximum G_0 sur cet intervalle. Nous pouvons donc écrire que

$$E(X_{nk}^4) \leqslant \frac{n^2}{\theta^4 \alpha^2} \sigma_n^4 G_0 \leqslant \frac{1}{\alpha^4 \theta^8} G_0,$$

ce qui conclut la preuve. □

Toutes les conditions du théorème étant vérifiées, nous pouvons l'appliquer pour les variables X_{nk} et nous obtenons :

$$\bar{H}_n(x) = \mathfrak{N}(x\bar{\sigma}_n^{-1}) + \frac{\mu_3^{(n)}}{6s_n^3}(1 - (x\bar{\sigma}_n^{-1})^2)\mathfrak{n}(x\bar{\sigma}_n^{-1}) + n^{-\frac{1}{2}} r_n(x\bar{\sigma}_n^{-1}) \tag{3.18}$$

avec $r_n(x) \to 0$ uniformément en x quand $n \to \infty$, en remplaçant $\bar{H}_n$ par $\bar{H}_n^{\#}$ dans le cas de variables sur des grilles.

Preuve du théorème 1. Étudions d'abord le cas où X n'est pas définie sur une grille. Si nous notons $K_n(x\bar{\sigma}_n^{-1}) = \mathfrak{N}(x\bar{\sigma}_n^{-1}) + \frac{\mu_3^{(n)}}{6s_n^3}(1 - (x\bar{\sigma}_n^{-1})^2)\mathfrak{n}(x\bar{\sigma}_n^{-1})$, nous obtenons pour l'intégrale du lemme 1 :

$$I_n = h_n \int_0^\infty e^{-h_n y}[K_n(y\bar{\sigma}_n^{-1}) - K_n(0)]dy + o(1/\sqrt{n}),$$

soit, avec le changement de variables $x = y\bar{\sigma}_n^{-1}$

$$I_n = h_n \bar{\sigma}_n \int_0^\infty e^{-h_n \bar{\sigma}_n x}[K_n(x) - K_n(0)]dx + o(1/\sqrt{n}).$$

En faisant une intégration par parties, nous pouvons écrire que :

$$I_n = \int_0^\infty e^{-h_n \bar{\sigma}_n x} K_n'(x)dx + o(1/\sqrt{n}).$$

Étant donné que $K_n'(x) = \mathfrak{n}(x) + \frac{\mu_3^{(n)}}{6s_n^3}(x^3 - 3x)\mathfrak{n}(x)$, nous allons d'abord étudier la contribution de $\frac{\mu_3^{(n)}}{6s_n^3}(x^3 - 3x)\mathfrak{n}(x)$ à l'intégrale. Cette contribution est égale à :

$$\frac{\mu_3^{(n)}}{6s_n^3} \int_0^\infty e^{-h_n \bar{\sigma}_n x}(x^3 - x)\mathfrak{n}(x)dx = \frac{\mu_3^{(n)}}{6s_n^3} \frac{1}{\sqrt{2\pi}} \int_0^\infty (x^3 - x)e^{-h_n \bar{\sigma}_n x} e^{-\frac{x^2}{2}} dx.$$

Nous avons vu que dans les conditions d'application du théorème central limite local $\frac{\mu_3^{(n)}}{6s_n^3}$ est uniformément borné par $M/6\sqrt{n}$ donc l'intégrale précédente est égale à $o(1/\sqrt{n})$. La contribution de $\mathfrak{n}(x)$ se calcule en posant le changement de variables $y = h_n \bar{\sigma}_n + x$, ce qui

donne :

$$\begin{aligned}\int_0^\infty e^{-h_n\bar{\sigma}_n x}\mathfrak{n}(x)dx &= \frac{1}{\sqrt{2\pi}}\int_{h_n\bar{\sigma}_n}^\infty e^{-h_n\bar{\sigma}_n(y-h_n\bar{\sigma}_n)}e^{-\frac{(y-h_n\bar{\sigma}_n)^2}{2}}dy\\ &= \frac{1}{\sqrt{2\pi}}e^{\frac{(h_n\bar{\sigma}_n)^2}{2}}\int_{h_n\bar{\sigma}_n}^\infty e^{-\frac{y^2}{2}}dy\\ &= e^{\frac{(h_n\bar{\sigma}_n)^2}{2}}\int_{h_n\bar{\sigma}_n}^\infty \mathfrak{n}(y)dy\\ &= e^{\frac{(h_n\bar{\sigma}_n)^2}{2}}[1-\mathfrak{N}(h_n\bar{\sigma}_n)]\end{aligned}$$

En utilisant le premier terme du développement asymptotique suivant :

$$1-\mathfrak{N}(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\{x^{-1}-x^{-3}+3x^{-5}+O(x^{-7})\} \text{ quand } x\to\infty,$$

nous obtenons :

$$\int_0^\infty e^{-h_n\bar{\sigma}_n x}\mathfrak{n}(x)dx = \frac{1}{\sqrt{2\pi}}\frac{1}{h_n\bar{\sigma}_n}+o(1/\sqrt{n}).$$

Nous avons ainsi montré que

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{1}{h_n\bar{\sigma}_n}+o(1/\sqrt{n}).$$

Or, d'après le lemme 2, $b_0 \leqslant h_n\sigma_n \leqslant B_0$, et nous avons montré que le fait que $\sum_{k=1}^n a_{nk}^2 = 1$ implique que $\frac{1}{\sqrt{n}} \leqslant \sigma_n \leqslant \frac{1}{\theta\sqrt{\alpha}\sqrt{n}}$. Nous en déduisons que $o(1/\sqrt{n}) = \frac{1}{h_n}o(1)$, et donc que

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{1}{h_n\bar{\sigma}_n}(1+o(1)),$$

ce qui conclut la preuve du théorème dans ce cas.

Si X est définie sur une grille, nous pouvons écrire que

$$I_n(h) = h_n\int_0^\infty e^{-h_n y}[\bar{H}_n^*(y)-\bar{H}_n(0)]dy.$$

Nous pouvons également écrire cette intégrale

$$I_n(h) = \sum_{k=0}^\infty h_n\int_{kd_n}^{(k+1)d_n} e^{-h_n y}[\bar{H}_n^*((k+\frac{1}{2})d_n)-\bar{H}_n(0)]dy.$$

Or, aux points milieux de la grille, $\bar{H}_n^*$ et $\bar{H}_n^\#$ coïncident donc nous pouvons appliquer le théorème central limite local. Nous remarquons que $\bar{H}_n(0) = \bar{H}_n^\#(d_n/2)$ et nous obtenons que

$$I_n(h) = \sum_{k=0}^\infty h_n\int_{kd_n}^{(k+1)d_n} e^{-h_n y}\left[K_n\left((k+\frac{1}{2})\frac{d_n}{\bar{\sigma}_n}\right)-K_n\left(\frac{d_n}{2\bar{\sigma}_n}\right)\right]dy+o(1/\sqrt{n}),$$

soit, en intégrant l'exponentielle :

$$I_n(h) = \sum_{k=0}^{\infty} (e^{-h_n k d_n} - e^{-h_n (k+1) d_n}) \left[K_n \left((k + \frac{1}{2}) \frac{d_n}{\bar{\sigma}_n} \right) - K_n \left(\frac{d_n}{2\bar{\sigma}_n} \right) \right] dy + o(1/\sqrt{n}),$$

et en écrivant la différence sur K_n comme une intégrale :

$$I_n(h) = \sum_{k=0}^{\infty} (e^{-h_n k d_n} - e^{-h_n (k+1) d_n}) \int_{\frac{d_n}{2\bar{\sigma}_n}}^{\frac{(k+\frac{1}{2})d_n}{\bar{\sigma}_n}} K_n^{'}(y) dy + o(1/\sqrt{n}).$$

De même que précédemment, $K_n^{'}(x) = \mathfrak{n}(x) + \frac{\mu_3^{(n)}}{6 s_n^3}(x^3 - 3x)\mathfrak{n}(x)$, et la contribution de $\frac{\mu_3^{(n)}}{6 s_n^3}(x^3 - 3x)\mathfrak{n}(x)$ à l'intégrale vaut $\frac{1}{h_n} o(1)$. Celle de $\mathfrak{n}(x)$ vaut

$$\sum_{k=0}^{\infty} (e^{-h_n k d_n} - e^{-h_n (k+1) d_n}) \frac{1}{\sqrt{2\pi}} \int_{\frac{d_n}{2\bar{\sigma}_n}}^{\frac{(k+\frac{1}{2})d_n}{\bar{\sigma}_n}} e^{-\frac{x^2}{2}} dx.$$

Nous développons $e^{-\frac{x^2}{2}} = 1 - \frac{x^2}{2} + o(x^2)$ et ainsi l'intégrale précédente vaut

$$\sum_{k=0}^{\infty} (e^{-h_n k d_n} - e^{-h_n (k+1) d_n}) \frac{k d_n}{\sqrt{2\pi}\bar{\sigma}_n} + o(1/\bar{\sigma}_n)$$

Nous sommons les deux séries :

$$\sum_{k=0}^{\infty} k e^{-h_n k d_n} = \frac{e^{-h_n d_n}}{(1 - e^{-h_n d_n})^2}$$

et

$$\sum_{k=0}^{\infty} k e^{-h_n (k+1) d_n} = \frac{e^{-2 h_n d_n}}{(1 - e^{-h_n d_n})^2}$$

et ainsi nous obtenons

$$I_n = \frac{1}{\sqrt{2\pi}} \frac{e^{-h_n d_n} d_n}{\bar{\sigma}_n (1 - e^{-h_n d_n})} (1 + o(1)),$$

ce qui conclut la preuve du théorème. □

3.6. Cas i.i.d

Dans le cas de variables i.i.d, c'est-à-dire de poids a_{nk} tous égaux à $n^{-\frac{1}{2}}$, nous retrouvons le théorème de Bahadur – Ranga Rao ([BR60]). Nous obtenons les résultats correspondant aux trois cas exposés dans leur article (absolument continu, treillis et mixte) avec la même technique de preuve, alors que les auteurs ont utilisé trois techniques différentes (respectivement un théorème de Cramér [Cra70], et deux théorèmes d'Esseen [Ess37]).

Théorème 4 (Théorème de Bahadur – Ranga Rao). *Il existe deux nombres positifs ρ et b, avec $0 < \rho < 1$, tels que*

$$P\left(\sum_{k=1}^{n} X_k \geqslant nc\right) = \frac{1}{\sqrt{2\pi n}}\rho^n b(1 + o(1)),$$

avec ρ le minimum de $\psi(t) = e^{-ct}\phi(t)$ atteint au point τ et $b = \frac{1}{\sigma\tau}$ si X n'est pas sur une grille et $b = \frac{d}{\sigma(1-e^{-\tau d})}$ si X est sur une grille de pas d.

Preuve. En posant $a_{nk} = n^{-\frac{1}{2}}$, nous avons $S_n = n^{-\frac{1}{2}}\sum_{k=1}^{n} X_k$ et $cA_n = cn^{\frac{1}{2}}$ donc $P(\sum_{k=1}^{n} X_k \geqslant nc) = P(S_n > cA_n)$. La condition $E(\bar{S}_n) = 0$ du lemme 2 se réécrit

$$n^{\frac{1}{2}}\left[\frac{\phi'(hn^{-\frac{1}{2}})}{\phi(hn^{-\frac{1}{2}})}\right] - cn^{\frac{1}{2}} = 0,$$

c'est-à-dire $\frac{\phi'(hn^{-\frac{1}{2}})}{\phi(hn^{-\frac{1}{2}})} = c$. En prenant $h_n = \tau n^{\frac{1}{2}}$, avec τ la valeur de t qui minimise $\psi(t) = e^{-ct}\phi(t)$ avec minimum ρ, nous calculons dans notre théorème

$$e^{-h_n cA_n}\prod_{k=1}^{n}\phi(h_n a_{nk}) = \left(e^{-c\tau}\phi(\tau)\right)^n = \rho^n.$$

À l'aide du lemme 1 nous obtenons

$$\bar{\sigma}_n = \frac{1}{b\tau},$$

et donc en remplaçant ces valeurs dans le théorème 1 nous concluons la démonstration. Nous identifions ainsi que si X n'est pas sur une grille,

$$b = \frac{1}{\sigma\tau}$$

et que si X est sur une grille de pas d,

$$b = \frac{de^{-\tau d}}{\sigma(1 - e^{-\tau d})}$$

car alors $d_n = \frac{d}{\sqrt{n}}$. □

3.7. Commentaires sur les résultats obtenus

Observons que les résultats de grandes déviations exposés dans ce chapitre sont des variantes classiques des résultats de la littérature. Nous souhaitons préciser ici leur nouveauté par rapport aux théorèmes existants. Le théorème pour des sommes pondérées de variables **absolument continues** est identique à celui de Book ([Boo72]), mais utilise une méthode de preuve légèrement différente et plus simple. Le résultat pour des sommes pondérées de

variables **treillis** est, à notre connaissance, nouveau. Book fait d'ailleurs remarquer dans un article où il présente des théorèmes de type Chernoff ([Boo73]) qu'il n'a pas réussi à l'obtenir.

La technique utilisée est elle-aussi bien connue, mais rarement exposée dans tous ses détails en dehors du cadre de variables i.i.d. Nous utilisons pour toutes les démonstrations des développements d'Edgeworth, dans la lignée des théorèmes d'Esseen ([Ess37]), exposés en particulier dans le livre de Feller ([Fel70]). Nous reprenons l'articulation générale des théorèmes et lemmes de la section 3.4 et commentons leurs différences par rapport aux résultats de la littérature.

Lemme 1. Ce lemme est très classique, semble-t-il. Il est indispensable pour l'exposé du résultat et présente cependant deux particularités par rapport aux versions présentes dans la littérature :

- Pour introduire l'intégrale $I_n(h)$, nous utilisons le théorème de Fubini au lieu d'une intégration par parties exigeant des conditions de régularité ; nous n'avons pas trouvé de démonstration rigoureuse de ce lemme dans la littérature pour les variables treillis ;
- Nous introduisons la fonction $\bar{H}_n^*$ dans le cas treillis (qui est différente de $\bar{H}_n$) ; ce lemme est donc légèrement différent des lemmes figurant dans la littérature. Cette modification permet d'aborder la preuve du théorème de grandes déviations plus simplement dans le cas treillis.

Théorème 1. Ce théorème est une extension du théorème de Book, qui ne traitait que le cas des variables **absolument continues**, au cas des variables **treillis**. Book n'avait obtenu pour les variables treillis que le terme logarithmique. Il est donc nouveau.

Théorème 2. Ce théorème, bien que très similaire, est plus général que celui présent dans le livre de Petrov [Pet75], chapitre 6, paragraphe 4, utilisé généralement dans la littérature ([BR60] par exemple). La différence essentielle réside dans la condition (III) de Petrov , page 173 :

$$\sqrt{n}\int_{t>\varepsilon}|t|^{-1}\prod_{j=1}^{n}|v_j(t)|dt\to 0$$

pour tout $\varepsilon>0$ fixé, qui n'est pas strictement identique à notre condition (ii) :

$$\prod_{j=1}^{n}|v_j(t)|dt\leqslant\frac{\varepsilon_n(\delta,a)}{\sqrt{n}}$$

pour tout $a>t>\delta>0$, avec $\varepsilon_n(\delta,a)\to 0$ quand $n\to\infty$. Le fait de ne pas exiger l'uniformité de la propriété en t sur $]\delta,+\infty[$ mais sur $]\delta,a[$ assure que la condition (ii) est vérifiée pour toutes les variables non treillis, ce qui n'est pas le cas de la condition de Petrov. En outre, les fonctions

$$v_j(t)=\frac{1}{1+log(1+log(1+t))},$$

qui vérifient le critère de Polya et sont donc des fonctions caractéristiques, vérifient (ii) et pas (III). En revanche, indiquons qu'avec les conditions (I) et (II) de Petrov, page 173, nous n'avons pas réussi à exhiber de fonction vérifiant (ii) et pas (III). La condition (III) est cependant d'une manière générale extrêmement difficile à vérifier, notamment dans notre contexte après un changement de variable exponentiel, à la différence de notre condition (ii). Cela présente un intérêt pratique important.

Théorème 3. Ce théorème repose sur l'application d'un théorème classique de lissage (livre de Feller, équation (3.13) page 538 par exemple), et étend le théorème de Feller (page 540, théorème 2) au cas de variables treillis indépendantes mais non identiquement distribuées. Cette extension n'est pas exposée dans la littérature à notre connaissance. Nous ajoutons que, pour le cas i.i.d, les éléments de preuve données par Feller dans son livre ne permettent pas de conclure la démonstration (une convolution par une distribution uniforme n'est pas suffisante, ce qui nous a amené à proposer une convolution par une distribution triangulaire).

Enfin, comme l'illustrent les applications numériques, l'obtention du premier ordre exact, nécessaire pour atteindre une précision satisfaisante, justifie l'exposition d'une variante nouvelle, quoique classique, des résultats de la littérature. Nous verrons aux chapitres IV et V que notre technique de preuve s'étend facilement à des contextes légèrement plus généraux.

Chapitre IV.

Étude des erreurs d'attributs et de géométrie

Nous avons étudié dans le chapitre III les erreurs d'attributs, et montré que le problème pouvait être résolu à l'aide de développements de grandes déviations pour des sommes pondérées de variables discrètes. Nous étudions maintenant l'impact des erreurs d'attributs et de position. La prise en compte des erreurs de position conduit à modifier légèrement les résultats du chapitre précédent, puisque les erreurs de position ont un impact sur les longueurs des tronçons, et donc sur les temps de parcours calculés.

Nous définissons en premier lieu des modèles d'erreurs de longueurs des tronçons, à partir des modèles d'erreur de position présentés au chapitre II (section 1). Nous montrons ensuite que l'estimation du critère de qualité des résultats de l'application se fait de façon très similaire à celle du chapitre III (section 2). Enfin, nous présentons des applications numériques illustrant la performance de notre méthode (section 3).

1. Modèles d'erreurs de longueurs des tronçons

1.1. Modèle fondé sur les erreurs de position

Nous avons présenté dans le chapitre II des modèles d'erreur de position dans les bases de données géographiques, ainsi que des méthodes de bruitage. Notre application de calcul d'itinéraires, et son modèle simplifié présentés au chapitre III ne sont sensibles aux erreurs de position que par le biais des longueurs des tronçons de route. Il faut donc avant tout définir un modèle d'erreurs de longueurs à l'aide du modèle GES présenté au chapitre II. Rappelons qu'avec ce modèle, un point A centré sur l'origine dans la référence a une probabilité de présence en (x, y) définie par la densité :

$$f(x,y) = \alpha \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2+y^2}{2\sigma^2}} + (1-\alpha)\frac{\lambda}{2} e^{-\lambda\sqrt{x^2+y^2}}.$$

Le paramètre de mélange α est souvent de l'ordre de $0,75$ dans les estimations de contrôle qualité.

La longueur d'un tronçon est la distance entre un point $A_1(X_1, Y_1)$ et un point $A_2(X_2, Y_2)$ dont on connaît les positions de référence (x_1, y_1) et (x_2, y_2), et les densités de présence $f_1(\alpha_{1x}, \lambda_{1x}, \sigma_{1x})$, $g_1(\alpha_{1y}, \lambda_{1y}, \sigma_{1y})$, $f_2(\alpha_{2x}, \lambda_{2x}, \sigma_{2x})$, et $g_2(\alpha_{2y}, \lambda_{2y}, \sigma_{2y})$. On calcule la loi de la variable $D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$, sous l'hypothèse d'indépendance des variables X_1, Y_1, X_2, Y_2, ainsi que des variables $(X_1 - X_2)$ et $(Y_1 - Y_2)$. Un calcul élémentaire montre que la densité de f de $(X_1 - X_2)$ vaut :

$$f(x) = \int f_1(v) f_2(v - x) dv$$

et que la densité g de $(Y_1 - Y_2)$ vaut :

$$g(y) = \int g_1(v) g_2(v - y) dv.$$

Pour calculer h densité de $D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$, un passage en coordonnées polaires permet d'obtenir :

$$h(r) = r \int_\theta \left(\int_u f_1(u) f_2(u - r\cos(\theta)) du \right) \left(\int_u g_1(u) g_2(u - r\sin(\theta)) du \right) d\theta$$

Cette densité a été calculée numériquement, mais n'a pas d'expression simple. Comme dans le chapitre II, nous allons nous intéresser au cas particulier où les lois GES se réduisent à des gaussiennes, et $\sigma_{1x}^2 = \sigma_{2x}^2 = \sigma_{1y}^2 = \sigma_{2y}^2$. Dans ce cas, $(X_1 - X_2)$ et $(Y_1 - Y_2)$ sont des gaussiennes et D^2 suit une loi du χ^2.

1.2. Modèle simplifié

La modélisation fondée sur les erreurs de position ayant l'inconvénient d'une relative lourdeur, nous proposons également d'estimer directement la loi η des écarts de longueur, en supposant que les longueurs s'écrivent $L_i = l_i + \eta_i$, l_i longueur dans la référence et η_i de loi η. Remarquons que cette écriture suppose que les erreurs de longueur sur les tronçons ne dépendent pas de la longueur de ces tronçons, ce qui est raisonnable. Dans la pratique, nous avons ajusté η à une loi normale centrée de variance s^2 sur la base de données étudiée.

2. Calcul de temps de parcours et critère de qualité

Nous utilisons le modèle d'application de calcul d'itinéraires du chapitre III, en incorporant les erreurs de position à l'aide de notre modèle simplifié, et considérons toujours le critère de l'erreur relative en temps :

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right).$$

Pour un itinéraire composé de k tronçons de route, le temps de parcours calculé dans la référence s'écrit comme précédemment :

$$T_R = \sum_{i=1}^{k} \frac{l_i}{V_{Ri}},$$

avec l_i longueur du i-ème tronçon et V_{Ri} sa vitesse déterminée à l'aide des attributs de la base de référence, mais le temps de parcours calculé à l'aide du jeu de données (comportant des erreurs par rapport à la référence) s'écrit :

$$T_D = \sum_{i=1}^{k} \frac{l_i + \eta_i}{V_{Di}},$$

η_i gaussienne centrée de variance σ^2.

Le critère retenu devient

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right),$$

qui se réécrit :

$$P\left(\sum_{i=1}^{k} l_i \left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} - \eta\frac{1}{V_{Ri}}\right) - \frac{\eta_i}{V_{Di}} > 0\right) + P\left(\sum_{i=1}^{k} l_i \left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} + \eta\frac{1}{V_{Ri}}\right) - \frac{\eta_i}{V_{Di}} < 0\right).$$

Pour se ramener à une somme de variables pondérées, nous écrivons :

$$\varepsilon_i = \frac{1}{l_i}\eta_i + 1$$

et nous obtenons ainsi :

$$P\left(\sum_{i=1}^{k} l_i \left(\frac{1}{V_{Ri}} - \frac{\varepsilon_i}{V_{Di}} - \eta\frac{1}{V_{Ri}}\right) > 0\right) + P\left(\sum_{i=1}^{k} l_i \left(\frac{1}{V_{Ri}} - \frac{\varepsilon_i}{V_{Di}} + \eta\frac{1}{V_{Ri}}\right) < 0\right).$$

Chacune des deux probabilités fait intervenir la somme de k variables aléatoires indépendantes pondérées. Nous notons $X_i = \frac{1}{V_{Ri}} - \frac{\varepsilon_i}{V_{Di}} - \eta\frac{1}{V_{Ri}}$ et $Y_i = \frac{1}{V_{Ri}} - \frac{\varepsilon_i}{V_{Di}} + \eta\frac{1}{V_{Ri}}$. Les variables X_i et Y_i sont indépendantes, et les X_i (resp. Y_i) ont même espérance, mais des variances différentes.

En réécrivant $l_i = l * a_{ki}$, avec $\sum_{i=1}^{k} a_{ki}^2 = 1$, en notant $A_k = \sum_{i=1}^{k} a_{ki}^2 = 1$, nous pouvons réécrire la somme des deux probabilités

$$P\left(\sum_{i=1}^{k} a_{ki}(X_i - E(X_i)) > -A_k E(X_1)\right) + P\left(\sum_{i=1}^{k} a_{ki}(Y_i - E(Y_i)) > -A_k E(Y_1)\right).$$

Dans le cas de tronçons de route à deux vitesses possibles, la loi de X_i s'écrit

$$\begin{cases} P(X_i = (1-\eta-\varepsilon_i)/v_1) = (1-\theta)N_1/N \\ P(X_i = (1-\eta-\varepsilon_i)/v_2) = (1-\theta)N_2/N \\ P(X_i = (1-\eta)/v_1 - \varepsilon_i/v_2) = \theta N_1/N \\ P(X_i = (1-\eta)/v_2 - \varepsilon_i/v_1) = \theta N_2/N, \end{cases} \tag{2.1}$$

avec $\varepsilon_i \sim N(1, s_i^2,\ s_i^2 = s^2/l_i^2)$.

Nous pouvons utiliser un développement très similaire à celui de du chapitre III pour estimer chacune des deux probabilités. Nous obtenons, quand $k \to \infty$, sous les conditions du théorème 1,

Théorème 5.

$$P(S_n \geqslant cA_n) = \frac{1}{\sqrt{2\pi}} \frac{d_n e^{-h_n d_n}}{\bar{\sigma}_n(1 - e^{-h_n d_n})} e^{-h_n c A_n} \left[\prod_{k=1}^{n} \phi_{nk}(h_n a_{nk})\right](1 + o(1)), \tag{2.2}$$

les $\phi_{nk}(t)$ étant des fonctions de t et de a_{nk}.

Pour notre problème, l'erreur relative en temps peut s'écrire alors :

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right) =$$
$$\frac{1}{\sqrt{2\pi}} \frac{1}{\bar{\sigma}_k^{(X)} h_k^{(X)}} e^{h_k^{(X)} E(X_1) A_k} \left[\prod_{i=1}^{k} \phi_i^{(X)}(h_k^{(X)} a_{ki})\right](1 + o(1))$$
$$+ \frac{1}{\sqrt{2\pi}} \frac{1}{\bar{\sigma}_k^{(Y)} h_k^{(Y)}} e^{h_k^{(X)} E(Y_1) A_k} \left[\prod_{i=1}^{k} \phi_i^{(Y)}(h_k^{(Y)} a_{ki})\right](1 + o(1)).$$

La fonction génératrice des moments $\phi_i^{(X)}$ de X_i se calcule en écrivant que $\phi_i^{(X)} = E(e^{tX_i}) = E(E(e^{t\varepsilon_i|X_i}))$:

$$\phi_i^{(X)}(t) = (1-\theta)\frac{N_1}{N}e^{-\frac{\eta}{v_1}t+\frac{\sigma_i^2}{2v_1^2}t^2} + (1-\theta)\frac{N_2}{N}e^{-\frac{\eta}{v_2}t+\frac{\sigma_i^2}{2v_2^2}t^2} + \theta\frac{N_1}{N}e^{(\frac{1-\eta}{v_1}-\frac{1}{v_2})t+\frac{\sigma_i^2}{2v_2^2}t^2} + \theta\frac{N_2}{N}e^{(\frac{1-\eta}{v_2}-\frac{1}{v_1})t+\frac{\sigma_i^2}{2v_1^2}t^2}. \quad (2.3)$$

Une formule similaire donne $\phi_i^{(Y)}$, et les paramètres $h_k^{(X)}$, $\bar{\sigma}_k^{(X)}$, $h_k^{(Y)}$, $\bar{\sigma}_k^{(Y)}$ se calculent numériquement. Nous constatons que la prise en compte des erreurs de position introduit des termes supplémentaires dans les fonctions génératrices dépendant de t^2.

Preuve du théorème 5. La ligne de preuve de ce théorème est décalquée sur celle du théorème 1. Remarquons que la transformation exponentielle utilisée produit des nouvelles variables indépendantes mais non équidistribuées, et que les théorèmes 2 et 3 peuvent s'appliquer directement dans notre cas.

Le lemme 1 est valide, en remplaçant ϕ par ϕ_i. En posant $Q_i = \phi_i'/\phi$, on obtient de même le lemme 1.

Nous définissons $Q_1 = \max_{i=1,\dots,k}\{Q_i\}$ et $Q_2 = \min_{i=1,\dots,k}\{Q_i\}$. Remarquons que la condition I assure l'existence de Q_1 et de Q_2. Nous modifions légèrement la condition II :

Condition II $\phi_i(t)$ est finie pour tout i sur $\mathcal{I} \supseteq (-B, B)$ avec $B > 0$, Q_1 prend la valeur $\frac{c}{\alpha\theta}$, et $B_0 = \frac{1}{\theta}Q_1^{-1}(\frac{c}{\alpha\theta}) \in \mathcal{I}$.

Les lemmes 2 et 3 s'écrivent alors en introduisant Q_1 et Q_2. Le lemme 2 devient :

Lemme 6. *Sous les conditions I et II, pour tout entier positif n, il existe une solution $h = h_n$ de l'équation $E(\bar{S}_n) = 0$, et cette solution vérifie les inégalités suivantes :*

$$b_0 = Q_2^{-1}(c\alpha\theta^2) \leqslant h_n\sigma_n \leqslant \theta^{-1}Q_1^{-1}(\frac{c}{\alpha\theta}) = B_0.$$

Le reste de preuve découle naturellement de ces lemmes.

□

3. Applications numériques

Nous avons comparé notre estimation de l'erreur relative en temps sur un trajet à celle obtenue sans prise en compte des erreurs de position (figure 3.1). Comme au chapitre III, nous constatons que cette fois encore l'utilisation d'une formule asymptotique dépendant du nombre de tronçons k avec k fixé relativement petit impose un développement au delà du terme logarithmique pour obtenir une précision suffisante. Nous illustrons l'influence de la prise en compte des erreurs de position figure 3.2, en faisant varier l'écart-type de la loi des erreurs de longueurs.

Nous avons confronté nos estimations à des estimations obtenues par des simulations de Monte-Carlo sur des trajet réels (figure 3.3)

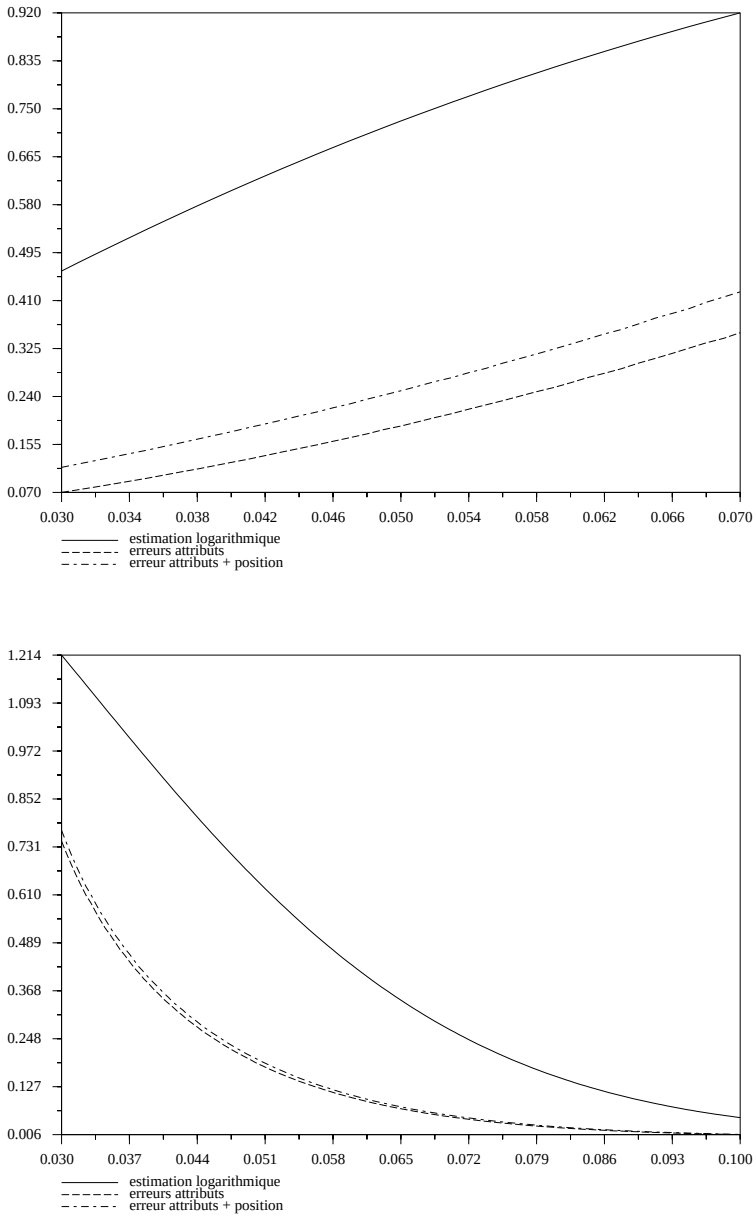

FIG. 3.1.: Probabilité de dépassement du seuil d'erreur relative η en fonction de θ avec $\eta = 5\%$ (haut) et en fonction de η avec $\theta = 5\%$ (bas)

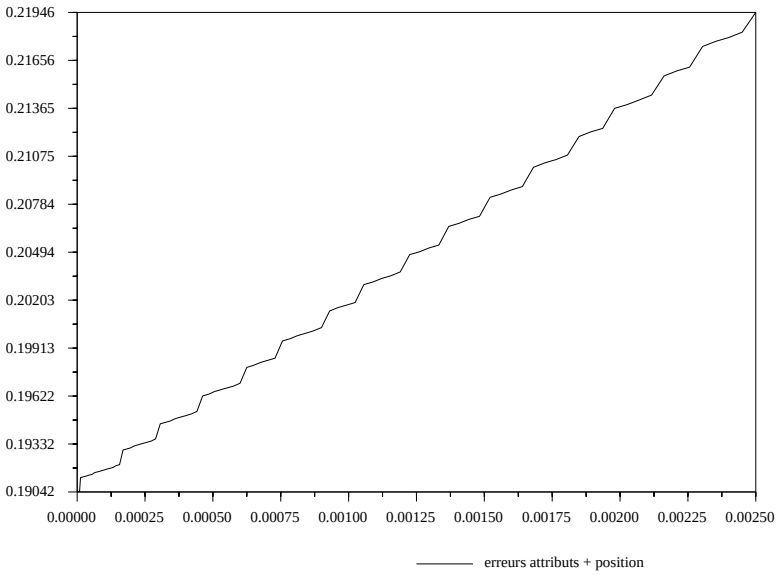

FIG. 3.2.: Probabilité de dépassement du seuil d'erreur relative η=5%, avec $\theta = 5\%$, en fonction de σ

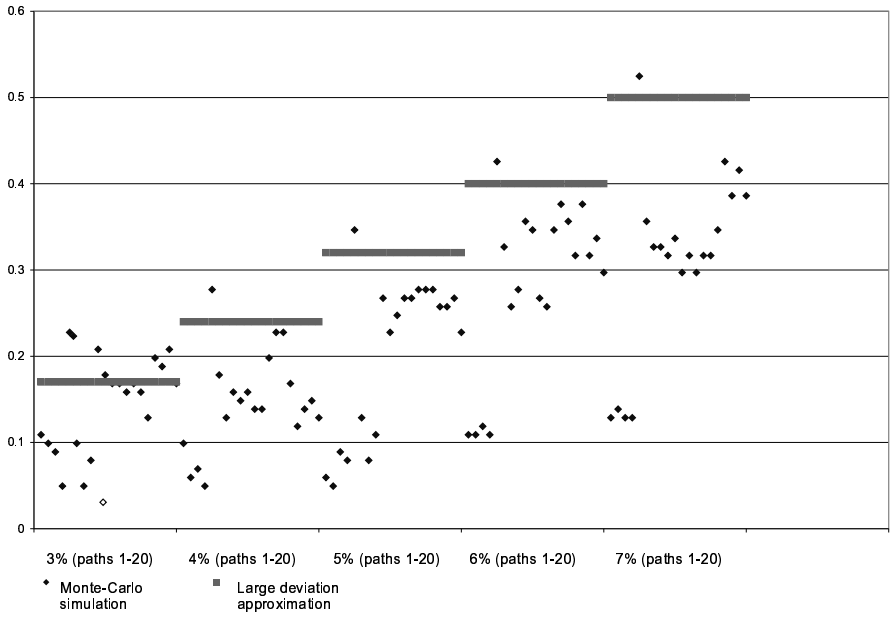

FIG. 3.3.: Probabilité de dépassement du seuil d'erreur relative η=5%, avec $\theta = 3\%, 4\%, 5\%, 6\%, 7\%$, avec $\sigma = 0$ (trait continu) et $\sigma = 0.025$ (tireté)

Chapitre V.

Étude de l'influence du choix de l'itinéraire, et erreurs sur des parcours de longueur aléatoire

Nous présentons dans ce chapitre deux extensions des résultats précédents, qui correspondent à des généralisations du problème posé.

Nous étudions d'abord l'erreur commise sur le temps de parcours d'un itinéraire tiré au sort parmi tous les itinéraires possibles. La modélisation retenue conduit à des grandes déviations pour des lois composées par un mélange de lois de Poisson, et nous appliquons les résultats classiques de la littérature ([Aas85]).

Nous nous intéressons ensuite à un itinéraire typique (Paris – Marseille par l'autoroute A6 par exemple), et étudions l'erreur commise sur le temps de parcours d'un trajet de Paris à une destination des environs de Lyon, c'est-à-dire l'erreur commise sur le temps de parcours pour un automobiliste dont on ne connaît que statistiquement la destination. Ce problème se résout à l'aide d'un développement de grandes déviations pour des lois composées pondérées, qui étend les résultats classiques de [EJMT85].

Notons que dans ces deux études deux asymptotiques sont envisageables. Nous donnons dans chaque cas les deux résultats.

1. Influence du choix de l'itinéraire

Nous nous intéressons ici au problème de l'estimation de l'erreur de temps de parcours commise pour un itinéraire tiré au hasard parmi tous les itinéraires possibles. La modélisation précédente est valide, mais le nombre de tronçons mis en jeu par un itinéraire tiré au hasard est une variable aléatoire discrète N, dont on peut ajuster la loi à l'aide de tests sur la base de données étudiées. Nous conservons le critère de l'erreur relative en temps pour l'itinéraire considéré pour mesurer l'impact des erreurs présentes dans la base de données, soit :

$$P\left(\frac{\sum_{i=1}^{N}(T_{Ri}-T_{Di})}{\sum_{i=1}^{N}T_{Ri}}>\eta\right)$$

η étant un seuil d'erreur admissible. Le problème s'écrit donc à l'aide d'une loi composée de la forme

$$Y=\sum_{i=1}^{N}a_{Ni}X_i, \tag{1.1}$$

avec les X_i des variables aléatoires i.i.d de loi X, et les a_{Ni} des variables aléatoires, *a priori* indépendantes de X_i, pour laquelle il faut établir un développement de grandes déviations.

Le problème des développements de grandes déviations pour des lois composées de la forme $\sum_{i=1}^{N}X_i$ est un problème ancien, dont la première résolution a été obtenue en 1932 par Esscher dans le cas où N suit une loi de Poisson et avec les X_i i.i.d [Ess32]. Ce type de développement pose des problèmes d'uniformité discutés dans [Jen88]. Parmi les extensions classiques, le cas d'un mélange fini de lois composées de Poisson a été résolu par Aase [Aas85], et celui du modèle avec inflation sur les X_i par Willmot [Wil89]. Enfin dans le cas où N est une loi de type Polya (binômiale négative par exemple), on trouve une résolution complète présentée par Embrechts, Jensen, Maejima et Teugels [EJMT85].

Ces développements ont été initialement motivés par des applications en assurance. Dans ce cas, on cherche à estimer la probabilité qu'un ensemble de sinistres (modélisé par une loi composée) dépasse un certain montant, soit

$$P\left(\sum_{i=1}^{N}X_i>y\right),$$

N étant le nombre de sinistres, X_i le coût de chaque sinistre, et y le seuil choisi. On peut alors considérer les asymptotiques $y\to\infty$, ou bien $E(N)\to\infty$. C'est généralement la

première des deux qui fait l'objet des études sur le sujet. Pour comprendre l'asymptotique $y \to \infty$, il faut s'intéresser à la façon dont est établi le développement. Comme dans le cas d'une somme ordinaire, on peut effectuer une transformation de point-selle, et obtenir une nouvelle somme de la forme

$$\sum_{i=1}^{N_h} X_{hi}.$$

Pour l'asymptotique $y \to \infty$, $h \to \tau_2 = \sup\{t : \phi(t) < \infty\}$, ce qui nous permet de distinguer quatre cas :

1. Si N est à support fini, soit pour un $K \in \{0, 1, \dots\}$ $P(N = k) = 0$ $\forall k > K$ et $P(N = K) > 0$, alors quand $y \to \infty$, N_h devient concentrée en K ;
2. Si N est à support infini, et $\xi(t) = E(e^{tN}) < \infty$ $\forall t$, alors $E(N_h) \to \infty$ quand $y \to \infty$, et on est dans le cas où un effet TCL a lieu (identique au cas classique) ;
3. Si N est à support infini et $\xi(t) = \infty$ $\forall t > t_0$ pour un $0 < t_0 < \infty$ (avec $\xi(t) < \infty$ pour $t < t_0$), alors quand $y \to \infty$, $h \to h_0$ défini par $\log(\phi(h_0)) = t_0$, et il n'y a pas d'effet TCL.

L'asymptotique $E(N) \to \infty$ provoquera dans tous les cas un effet TCL.

Pour notre application, le problème s'écrit :

$$P\left(\sum_{i=1}^{N} a_{Ni} X_i > c \sum_{i=1}^{N} a_{Ni}\right) + P\left(\sum_{i=1}^{N} a_{Ni} X'_i > c' \sum_{i=1}^{N} a_{Ni}\right),$$

et donc l'asymptotique s'impose d'elle même. Sauf mention contraire, on considére dans le suite que $E(N) \to \infty$ pour établir les différents résultats.

Au prix d'hypothèses sur les variables a_{Ni}, nous pouvons nous ramener dans le cadre de résultats classiques. A toute valeur k prise par la variable aléatoire N correspond un tableau de variables aléatoires $a_{k1}, \dots, a_{kk}$. La taille de ces tableaux est une variable aléatoire de loi N. Dans notre application géographique, les a_{ki} sont proportionnelles aux longueurs des tronçons de route. Nous supposons dans cette section que le nombre de tronçons N suit une loi de Poisson de paramètre λ.

Nous pouvons envisager le cas où les a_{Ni} sont indépendantes de N et i.i.d de même loi qu'une variable a, et ne prennent qu'un nombre fini r de valeurs $l_1, \dots, l_r$. Notons que cette dernière condition est toujours vérifiée dans la pratique, car les tronçons de la base de données ne peuvent mesurer qu'un nombre fini de longueurs, puisque les longueurs dans une base de données sont stockées avec une précision finie (au mètre près par exemple). Ces hypothèses reviennent à dire que quel que soit l'itinéraire considéré et quelle que soit sa taille, il contient en moyenne la même proportion de tronçons de chaque longueur. Nous notons $p_j = P(a = l_j)$, $1 \leqslant j \leqslant r$, et la loi composée de (1.1) s'écrit :

$$Y = \sum_{k=1}^{r} \sum_{i=1}^{N_k} X_{ki},$$

les X_{ki} étant des variables aléatoires discrètes indépendantes et de loi P_k (on a $X_{ki} = l_k X'_{ki}$, les X'_{ki} étant i.i.d $\forall(i,k)$), et les $N_k, 1 \leqslant k \leqslant r$ suivant des lois de Poisson de paramètres $\lambda_k = \lambda p_k$, avec $\sum_{k=1}^{r} \lambda_k = \lambda$.

La transformée de Laplace de Y s'écrit alors :

$$\begin{aligned}\phi_c(t) = E(e^{tY}) &= e^{\sum_{k=1}^{r} -\lambda_k(1-\phi(l_k t))} \\ &= e^{-\lambda\left[1-\sum_{k=1}^{r} \frac{\lambda_k}{\lambda}\phi(l_k t)\right]},\end{aligned}$$

On constate que Y a la même distribution que $\sum_{i=1}^{N} X_i$, avec N variable suivant une loi de Poisson de paramètre λ et X_i mélange fini de loi $P = \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} P_k$.

Les résultats de la littérature pour les lois composées Poisson sont applicables tels quels. Il faut vérifier que la queue de distribution de P vérifie certaines propriétés. Dans notre cas, P est une loi treillis à support borné et donc vérifie la proposition 7.2.5 page 197 de [Jen95]. Ainsi nous pouvons écrire, quand $y \to \infty$:

$$P\left(\sum_{k=1}^{r}\sum_{i=1}^{N_k} X_{ki} \geqslant y\right) = \frac{\phi_c(h)e^{-hy}}{\sqrt{2\pi}\sigma_c(1-e^{-h})}(1+o(1)), \tag{1.2}$$

avec h solution de l'équation

$$\lambda \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} l_k \phi'(l_k h) = y$$

et

$$\sigma_c^2 = \lambda \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} l_k^2 \phi''(l_k h).$$

Nous obtenons le même résultat pour l'asymptotique $\lambda \to \infty$, comme le note Jensen page 193 [Jen95].

2. Erreurs sur un itinéraire type

Nous nous intéressons maintenant à l'erreur relative commise sur le temps de parcours d'un itinéraire type de la base de données, pour un automobiliste dont nous ne connaissons la destination que de façon probabiliste (par exemple, un automobiliste partant de Paris se rendant dans le sud est de la France). Cet itinéraire emprunte N tronçons, N étant une variable aléatoire discrète que nous supposons suivre une loi de Poisson. Pour un tel itinéraire composé de N tronçons, nous notons T_R (resp. T_D) le temps de parcours calculé à l'aide de la *référence* (resp. du *jeu de données*). Nous voulons étudier la probabilité que l'erreur relative en temps dépasse un seuil η. Indiçant par i les grandeurs se rapportant au $i$ème tronçon, cette probabilité s'écrit avec les longueurs l_i et les vitesses V_i :

$$P\left(\left|\frac{T_R - T_D}{T_R}\right| > \eta\right) =$$
$$P\left(\sum_{i=1}^{N} l_i \left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} - \eta\frac{1}{V_{Ri}}\right) > 0\right) + P\left(\sum_{i=1}^{N} l_i \left(\frac{1}{V_{Ri}} - \frac{1}{V_{Di}} + \eta\frac{1}{V_{Ri}}\right) < 0\right).$$

Pour chacune de ces deux probabilités, les lois des N variables indépendantes et équidistribuées $(1/V_{Ri} - 1/V_{Di} \pm \eta \times 1/V_{Ri})$ se calculent à l'aide de notre modèle d'erreurs, la vitesse de chaque tronçon étant calculée à l'aide des valeurs des attributs. Nous pouvons ainsi estimer ces probabilités par des développements de grandes déviations pour des lois composées de variables discrètes pondérées, en considérant l'asymptotique $E(N) \to \infty$, puis l'asymptotique $y \to \infty$.

2.1. Grandes déviations pour lois composées

Soit $X = X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d non dégénérées, soit $a_1, a_2, \ldots$ une suite de réels positifs, soit $\sigma = \max\{a_k\} < \infty$, et soit c une constante réelle positive. Nous considérons $S_N = \sum_{i=1}^{N} a_i X_i$, où N est une variable aléatoire discrète suivant une loi de Poisson de paramètre λ, nous notons $A_N = \sum_{i=1}^{N} a_i$, et étudions le comportement de la probabilité $P(S_N > cA_N)$ quand $E(N) \to \infty$. Nous supposons que $E(X) = 0$ et que $E(X^2) = 1$. Nous notons $F(x) = P(X \leqslant x)$ la fonction de répartition de X, $\phi(t) = E(e^{tX})$ la fonction génératrice des moments de X, et $\phi_{S_N}(t) = E(e^{tS_N})$ la fonction génératrice des moments de S_N. Nous notons $\xi(t) = E(t^N)$. Soit $Y = S_N - A_N$ et $\phi_Y(t) = E(e^{tY})$.

Nous imposons une condition de régularité sur la suite a_i.

Condition I' Il existe α et θ avec $0 < \alpha \leqslant 1$, $0 < \theta \leqslant 1$, tels que pour tout n, au moins αn des a_k, $1 \leqslant k \leqslant n$, sont supérieurs ou égaux à $\theta\sigma$.

Posons $Q(t) = \phi'(t)/\phi(t)$ pour $t \in \mathbb{R}$. Remarquons que Q est croissante et que l'image de Q est l'enveloppe convexe du support de X. Nous imposons une extension naturelle de la condition de Cramer :

Condition II $\phi(t)$ est finie sur $\mathcal{I} \supseteq (-B, B)$, pour un $B > 0$, la fonction $Q = \phi'/\phi$ prend la valeur $\frac{c}{\alpha\theta}$ en un point, et $B_0 = \theta^{-1}Q^{-1}(\frac{c}{\alpha\theta}) \in \mathcal{I}$.

Notons que si X est une loi absolument continue, S_N n'est pas continue à cause d'une masse $p_0 = P(N = 0)$ en 0. On écrira dans ce cas que

$$\begin{aligned} P(S_N > cA_N) &= P(S_N > cA_N \mid N > 0)P(N > 0) \\ &= P\left(\sum_{i=1}^{\widetilde{N}} a_i X_i > c\sum_{i=1}^{\widetilde{N}} a_i\right)(1 - p_0), \end{aligned}$$

avec

$$P(\widetilde{N} = k) = (1 - p_0)^{-1} p_k,$$

pour $k = 1, 2, \ldots$. Dans ce cas, on pose $\widetilde{Y} = \sum_{i=1}^{\widetilde{N}} a_i X_i$ et $E(e^{t\widetilde{Y}}) = (1-p_0)^{-1}(\phi_Y(t)-p_0)$. Dans la suite de la démonstration, le lecteur remplacera Y par $\widetilde{Y}$ si Y n'est pas treillis.

Si X est une variable treillis et il existe au moins un rapport a_i/a_j non rationnel, alors S_N n'est pas une variable treillis. Nous imposons donc la condition suivante :

Condition III Les a_i sont tels que S_N est une variable treillis de pas d.

Nous calculons une expression explicite de $\phi_Y(t)$ en fonction des $p_k = P(N = k)$, de ϕ et des a_i :

$$\phi_Y(t) = E(e^{t(S_N - A_N)}) = E(E(e^{t(S_N - A_N)} \mid N)) = \sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} e^{-ca_i t}\phi(a_i t), \tag{2.1}$$

et constatons que, sous les conditions I et II, $\phi_Y(h)$ est finie pour tout h vérifiant $|h| < B\sigma^{-1}$.

Nous effectuons la transformation exponentielle suivante, en notant H_{c0} la fonction de répartition de Y :

$$\frac{dH_{ch}}{dH_{c0}}(x) = \frac{e^{hx}}{\phi_Y(h)},$$

pour $0 < h < B\sigma^{-1}$. Soit Y_h une variable aléatoire distribuée suivant H_{ch}.

Nous pouvons énoncer notre théorème :

Théorème 6. *Supposons que les conditions I' et II sont vérifiées. Soit $c > 0$ fixé et h solution de l'équation $E(Y_h) = 0$. Posons $s^2 = \mathrm{Var}(Y_h)$ et $\mu_3 = E(Y_h^3)$. Supposons de plus que $\mu_3/s^3 \to 0$. Alors, quand $E(N) \to \infty$, $s \to \infty$ et*

$$P(S_N > cA_N) = \frac{1}{\sqrt{2\pi}}\frac{1}{sh}(\phi_Y(h) - p_0)(1 + o(1))$$

si X n'est pas treillis, et

$$P(S_N > cA_N) = \frac{1}{\sqrt{2\pi}} \frac{de^{-hd}}{s(1-e^{-hd})} \phi_Y(h)(1+o(1)),$$

si X est treillis et la condition III est vérifiée, avec d pas de la grille de S_N.

Preuve du théorème 6. Nous avons classiquement :

Lemme 7.

$$P(S_n > cA_N) = \phi_Y(h) I(h),$$

avec

$$I(h) = h \int_0^\infty e^{-hx} [H_{ch}(x) - H_{ch}(0)] dx$$

si Y_h n'est pas treillis, et avec

$$I(h) = h \int_0^\infty exp(-hx) [H^*_{ch}(x) - H^*_{ch}(0)] dx$$

si Y_h est treillis, H^*_{ch} étant définie ainsi :

$$\begin{aligned} H^*_{ch}(x) &= \frac{1}{2}[H_{ch}(x) + H_{ch}(x-)] && \text{si } x \text{ est sur un noeud du treillis} \\ &= H_{ch}(x) && \text{sinon.} \end{aligned}$$

Preuve. C'est la même que celle du cas d'une somme classique (Cf. lemme 1). □

Nous allons ensuite choisir un h adapté au problème, et établir un développement de $I(h)$ quand $E(N) \to \infty$.

Calculons d'abord la fonction génératrice des moments de Y_h, que nous notons ϕ_h :

$$\phi_h(t) = E(e^{tY_h}) = \int e^{ty} dH_{ch}(y) = \frac{1}{\phi_Y(h)} \int e^{(t+h)y} dH_{c0}(y) = \frac{\phi_Y(t+h)}{\phi_Y(h)}.$$

En introduisant l'expression (2.1) de $\phi_Y(t)$, nous obtenons :

$$\phi_h(t) = \frac{\sum_{k=0}^\infty p_k \prod_{i=1}^k e^{-ca_i(t+h)} \phi(a_i(t+h))}{\sum_{k=0}^\infty p_k \prod_{i=1}^k e^{-ca_i h} \phi(a_i h)}. \tag{2.2}$$

Nous écrivons maintenant Y_h sous la forme $Y_h = \sum_{i=1}^{N_h} X_{hi}$, où les X_{hi} sont des v.a. indépendantes, et calculons la fonction génératrice des moments de Y_h en fonction de ϕ_{hi} fonction génératrice des moments de X_{hi} :

$$\phi_h(t) = \sum_{k=0}^\infty \prod_{i=1}^k \phi_{hi}(t) P(N_h = k). \tag{2.3}$$

En identifiant les termes des expressions (2.2) et (2.3), nous obtenons les relations

$$q_k = P(N_h = k) = \frac{\prod_{i=1}^k e^{-ca_i h} \phi(a_i h)}{\sum_{k=0}^\infty p_k \prod_{i=1}^k e^{-ca_i h} \phi(a_i h)} p_k, \tag{2.4}$$

et

$$\phi_{hi}(t) = e^{-ca_i t}\frac{\phi(a_i(t+h))}{\phi(a_i h)}.$$

Ainsi les X_{hi} ont pour loi F_{hi} définie par

$$\frac{dF_h}{dF}(x) = \frac{e^x}{e^{-ca_i h}\phi(a_i h)}.$$

Remarque. Les X_{hi} sont les variables associées aux variables $a_i(X_i - c)$, comme au chapitre III.

Notons que $E(N_h)$ varie avec $E(N)$. En particulier, nous avons pour $E(N_h)$ et $\mathrm{Var}(N_h)$ les encadrements suivants :

Lemme 8.

$$\lambda e^{-c\sigma h} \leqslant E(N_h) \leqslant \lambda\phi(\sigma h) \tag{2.5}$$

$$\lambda e^{-c\sigma h} + \lambda^2(\phi(\sigma h)^2 - e^{-2c\sigma h}) \leqslant \mathrm{Var}(N_h) \leqslant \lambda\phi(\sigma h) + \lambda^2(-\phi(\sigma h)^2 + e^{-2c\sigma h}). \tag{2.6}$$

Preuve. En utilisant l'expression 2.4, nous obtenons

$$E(N_h) = \frac{\sum_{k=0}^{\infty} k p_k \prod_{i=1}^{k} e^{-ca_i h}\phi(a_i h)}{\sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} e^{-ca_i h}\phi(a_i h)} = \frac{\sum_{k=0}^{\infty} k p_k \psi_k(h)}{\sum_{k=0}^{\infty} p_k \psi_k(h)},$$

en posant $\psi_k(h) = \prod_{i=1}^{k} e^{-ca_i h}\phi(a_i h)$. Comme N suit une loi de Poisson de paramètre λ, on obtient :

$$E(N_h) = \frac{\sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)} = \lambda\frac{\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{k-1!}\psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)} = \lambda\frac{\sum_{k=0}^{\infty} p_k\psi_{k+1}(h)}{\sum_{k=0}^{\infty} p_k\psi_k(h)}.$$

Or, avec la condition I',

$$e^{-c\sigma h}\psi_k(h) \leqslant \psi_{k+1}(h) \leqslant \psi_k(h)\phi(\sigma h)$$

puisque $0 < a_i \leqslant \sigma\ \forall i$ et $\phi(0) = 1$. Nous en déduisons directement l'encadrement de $E(Y_h)$. L'encadrement $\mathrm{Var}(N_h)$ s'obtient par des calculs extrêmement similaires. □

Lemme 9.

$$E\left((1-\theta)^{N_h}\right) \leqslant e^{-\theta\lambda\phi(\sigma h)}.$$

Preuve. Nous utilisons l'expression (2.4) de la loi de N_h, et introduisons le développement du binôme de $(1-\theta)^k$ pour obtenir

$$E(1-\theta)^{N_h} = \frac{\sum_{k=0}^{\infty}\sum_{i=0}^{k}\frac{k!}{i!(k-i)!}(-\theta)^{k-i}e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)}.$$

Par le théorème de Fubini, nous avons

$$\begin{aligned} E(1-\theta)^{N_h} &= \frac{\sum_{i=0}^{\infty}\sum_{k=i}^{\infty}\frac{1}{i!(k-i)!}(-\theta)^{k-i}e^{-\lambda}\lambda^k\psi_k(h)}{\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)} \\ &= \frac{\sum_{i=0}^{\infty}\sum_{l=0}^{\infty}\frac{1}{l!i!}(-\theta)^{l}e^{-\lambda}\lambda^{l+i}\psi_{l+i}(h)}{\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)} \\ &= \frac{\sum_{i=0}^{\infty}e^{-\lambda}\frac{\lambda^i}{i!}\sum_{l=0}^{\infty}(-\theta)^{l}\frac{\lambda^l}{l!}\psi_{l+i}(h)}{\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)}. \end{aligned}$$

Or, avec la condition I',

$$\psi_{l+i}(h) \leqslant \psi_i(h)\phi^l(\sigma h),$$

et donc

$$E(1-\theta)^{N_h} \leqslant \frac{\sum_{i=0}^{\infty}e^{-\lambda}\frac{\lambda^i}{i!}\psi_i(h)\sum_{l=0}^{\infty}(-\theta)^{l}\frac{\lambda^l}{l!}\phi^l(\sigma h)}{\sum_{k=0}^{\infty}e^{-\lambda}\frac{\lambda^k}{k!}\psi_k(h)} = e^{-\theta\lambda\phi(\sigma h)}.$$

□

Maintenant, sous les conditions I' et II, nous choisissons h tel que $E(Y_h) = 0$.

Lemme 10. *Sous les conditions I et II, il existe un h solution de l'équation $E(Y_h) = 0$, et il vérifie les inégalités suivantes :*

$$\frac{Q^{-1}(\alpha\theta c)}{\sigma} \leqslant h \leqslant \frac{\theta^{-1}Q^{-1}\left(\frac{c}{\alpha\theta}\right)}{\sigma}.$$

Preuve. Calculons $E(Y_h)$:

$$E(Y_h) = E\left(E\left(\sum_{i=1}^{N_h}X_{hi} \mid N_h\right)\right) = \sum_{k=0}^{\infty}q_k\left(\sum_{i=1}^{k}E(X_{hi})\right).$$

Or

$$E(X_{hi}) = a_i\frac{\phi'(a_ih)}{\phi(a_ih)} - ca_i,$$

d'où

$$E(Y_h) = \sum_{k=0}^{\infty}q_k\left(\sum_{i=1}^{k}a_i(Q(a_ih) - c)\right).$$

La condition I' permet d'écrire que

$$E(Y_h) \geqslant \sum_{k=0}^{\infty}q_k\left[\alpha k\theta\sigma Q(\theta\sigma h) - k\sigma c\right],$$

soit

$$\left[\alpha\theta Q(\theta\sigma h) - c\right]E(N_h) \leqslant E(Y_h),$$

et de même que

$$0 = E(Y_h) \leqslant \sum_{k=0}^{\infty} q_k \left[k\sigma Q(\sigma h) - \alpha k\theta\sigma c\right],$$

soit

$$\left[Q(\sigma h) - \alpha\theta c\right] E(N_h) \geqslant E(Y_h).$$

La condition II assure l'existence d'un h tel que $E(Y_h) = 0$.

Comme Q est croissante, avec h vérifiant $E(Y_h) = 0$, on a

$$\sigma h \leqslant \theta^{-1} Q^{-1}\left(\frac{c}{\alpha\theta}\right).$$

et

$$\sigma h \geqslant Q^{-1}(\alpha\theta c),$$

ce qui donne l'encadrement de h. □

Nous encadrons ensuite $\mathrm{Var}(Y_h)$.

Lemme 11. *Soit h tel que $E(N_h = 0)$. Alors, il existe deux réels s_0 et s_1 tels que*

$$s_0^2 \alpha\theta^2\sigma^2 E(N_h) \leqslant \mathrm{Var}(Y_h) \leqslant s_1^2\sigma^2 E(N_h) + (\sigma Q(\sigma h) - \alpha\theta\sigma c)^2 \mathrm{Var}(N_h). \tag{2.7}$$

Preuve. Nous obtenons par définition

$$\mathrm{Var}(Y_h) = E\left[\sum_{i=1}^{N_h} \mathrm{Var}(X_{hi})\right] + \mathrm{Var}\left[\sum_{i=1}^{N_h} E(X_{hi})\right].$$

En introduisant

$$\mathrm{Var}(X_{hi}) = a_i^2 Q'(a_i h),$$

nous obtenons

$$\mathrm{Var}(Y_h) = \sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} a_i^2 Q'(a_i h) + \mathrm{Var}\left[\sum_{i=1}^{N_h} a_i (Q(a_i h) - c)\right].$$

Avec la condition I', nous pouvons encadrer $\mathrm{Var}(Y_h)$. Comme $Q'(h)$ est la variance de la transformée exponentielle de X, Q' est strictement positive, et donc $s_0^2 = \min\{Q'(z) : 0 \leqslant z \leqslant B\}$ est strictement positif. Ainsi, $\mathrm{Var}(Y_h) \geqslant s_0^2\alpha\theta^2\sigma^2 E(N_h)$, puisque $\mathrm{Var}\left[\sum_{i=1}^{N_h} E(X_{hi})\right] > 0$. Comme $\phi(0) = 1$ et $\phi(B_0) < \infty$, $s_1^2 = \max\{Q'(z) : 0 \leqslant z \leqslant B\}$ est fini et donc $\mathrm{Var}(Y_h) \leqslant s_1^2\sigma^2 E(N_h) + (\sigma Q(\sigma h) - \alpha\theta\sigma c)^2 \mathrm{Var}(N_h)$. Nous en déduisons l'encadrement de $\mathrm{Var}(Y_h)$. □

Enfin, nous établissons une approximation de $H_{ch}(sx)$ par la fonction

$$G(x) = \mathfrak{N}(x) + \frac{\mu_3}{6s^3}(1 - x^2)\mathfrak{n}(x)$$

avec $\mu_3 = E(Y_h)^3$ et $s^2 = \mathrm{Var}(Y_h)$.

Théorème 7. *Soit Y_h définie précédemment, avec $E(Y_h) = 0$. Notons ω_{hj} la fonction caractéristique de X_{hj}. Si Y_h vérifie :*

(i) *$|\omega_{hj}(\zeta)| < 1 - \theta(\delta, a)$ $\forall j$ pour $a > \zeta > \delta > 0$, $0 < \theta(\delta, a) < 1$;*

(ii) *$(\mu_4 - 3s^4)/s^4$ est uniformément bornée $\forall h$;*

(iii) *$\mu_3/s^3 \to 0$ quand $s \to \infty$;*

alors

$$H_{ch}(sx) = \mathfrak{N}(x) + \frac{\mu_3}{6s^3}(1 - x^2)\mathfrak{n}(x) + \frac{1}{s}r_s(x),$$

avec $r_s(x) \to 0$ uniformément en x quand $s \to \infty$.

Preuve du théorème 7. On étudie la transformée de Fourier de $D(x) = H_{ch}(sx) - G(x)$. Nous écrivons $\omega(\zeta) = \phi_h(i\zeta)$ fonction caractéristique de Y_h sous la forme $\sum_{k=0}^{\infty} q_k e^{v_k(\zeta)}$, en posant $v_k(\zeta) = \sum_{j=1}^{k} \log(\phi_{hj}(i\zeta))$. Notons que, puisque ω est une fonction caractéristique, $\omega(0) = 1$, $\omega'(0) = E(Y_h) = 0$, $\omega''(O) = i^2 s^2$ et $\omega'''(0) = i^3\mu_3$. Dans ce cas, $\sum_{k=0}^{\infty} q_k v_k(0) = 0$, $\sum_{k=0}^{\infty} q_k v_k'(0) = 0$, $\sum_{k=0}^{\infty} q_k v_k''(0) = i^2 s^2$ et $\sum_{k=0}^{\infty} q_k v_k'''(0) = i^3\mu_3$. Remarquons que $|G'(x)| = O(\mu_3/s^3)$ quand $y \to \infty$.

Soit $\varepsilon > 0$ fixé. Nous choisissons une constante a suffisamment grande pour que $|G'(x)| < \varepsilon a$ pour pour tout x et pour tout y. D'après un théorème de lissage (page 538 de [Fel70]), nous pouvons écrire :

$$|D(x)| \leqslant \frac{1}{\pi}\int_{-as}^{as} \left| \frac{\sum_{k=0}^{\infty} q_k \left[e^{v_k\left(\frac{\zeta}{s}\right)} - e^{-\frac{1}{2}\zeta^2} - \frac{v_k'''(0)}{6s^3} i^3 \zeta^3 e^{-\frac{1}{2}\zeta^2} \right]}{\zeta} \right| d\zeta + \frac{24\varepsilon}{\pi s}. \tag{2.8}$$

Nous séparons cette intégrale en deux domaines d'intégration. Le premier est le domaine défini par $\delta s \leqslant |\zeta| \leqslant as$, et le second par $|\zeta| < \delta s$, δ étant un réel positif fixé que nous précisons plus loin. Sur le domaine $\delta s \leqslant |\zeta| \leqslant as$, tous les $|\phi_{hj}(i\zeta)| < 1 - \theta$ $\forall j$ $\forall k$, avec $\theta > 0$. Remarquons que θ ne dépend pas de h, puisque h est uniformément borné. Alors, par le lemme 9, $|\omega(\zeta/s)| < E(1-\theta)^{N_h} < e^{-\theta\lambda\phi(\sigma h)}$. Ainsi, la contribution de ce domaine à l'intégrale (2.8) est inférieure à

$$\log\left(\frac{a}{\delta}\right) e^{-\theta\lambda\phi(\sigma h)} + \int_{\delta s \leqslant |\zeta| \leqslant as} \frac{e^{-\frac{1}{2}\zeta^2}}{\zeta}\left(1 + \left|\frac{\mu_3}{6s^3}\zeta^3\right|\right) d\zeta.$$

Le terme de droite de la formule précédente est un petit o de n'importe quelle puissance de s.

Pour le domaine $|\zeta| < \delta s$, en posant

$$\psi_k(\zeta) = v_k(\zeta) + \frac{1}{2}s^2\zeta^2,$$

l'intégrande de (2.8) se réécrit

$$\frac{e^{-\frac{1}{2}\zeta^2}}{\zeta}\left|\sum_{k=0}^{\infty} q_k \left[e^{\psi_k\left(\frac{\zeta}{s}\right)} - 1 - \frac{v_k'''(0)}{6s^3} i^3\zeta^3 \right]\right| d\zeta \tag{2.9}$$

et nous allons l'estimer à l'aide de l'inégalité suivante tirée de [Fel70] :

$$|e^{\alpha} - 1 - \beta| = |(e^{\alpha} - e^{\beta}) + (e^{\beta} - 1 - \beta)| \leqslant (|\alpha - \beta| + \frac{1}{2}\beta^2)e^{\gamma}, \tag{2.10}$$

avec $\gamma \geqslant \max(|\alpha|, |\beta|)$, pour α et β arbitraires, réels ou complexes. Nous pouvons faire un développement de Taylor au troisième ordre de ψ. Comme $v^{(4)} = (\mu_4 - 3s^4)/s^4$ est uniformément bornée par hypothèse, ceci nous permet de déduire l'existence d'un δ tel que pour tout k

$$\left|\psi_k\left(\frac{\zeta}{s}\right) - \frac{v_k'''(0)}{6s^3}i^3\zeta^3\right| < \varepsilon k\left|\frac{\zeta}{s}\right|^3 \tag{2.11}$$

pour $|\zeta| < \delta s$. Nous prenons δ suffisamment petit pour avoir également

$$\left|\psi_k\left(\frac{\zeta}{s}\right)\right| < \frac{1}{4}\zeta^2, \qquad \left|\frac{v_k'''(0)}{6s^3}\zeta^3\right| \leqslant \frac{1}{4}\zeta^2$$

pour $|\zeta| < \delta s$. Avec ce choix de δ nous majorons l'intégrale (2.8) sur le domaine $|\zeta| < \delta s$ en utilisant la formule (2.10) par :

$$\int_{|\zeta|<\delta s} e^{-\frac{1}{4}\zeta^2}\left|\frac{\varepsilon E(N_h)}{s^3}|\zeta|^2 + \frac{\mu_3^2}{72s^6}|\zeta|^5\right| d\zeta. \tag{2.12}$$

En choisissant $E(N)$ assez grand pour que $\log\left(\frac{a}{\delta}\right)e^{-\theta\lambda\phi(\sigma h)} \leqslant \frac{\varepsilon}{s}$ (ce qui est toujours possible vu l'encadrement (2.7) de s et les expressions (2.5) et (2.6)), et pour que l'intégrale (2.12) soit inférieure à $\frac{1000\varepsilon}{s}$, nous avons montré que pour tout x

$$|D(x)| \leqslant \frac{24\varepsilon}{\pi s} + \frac{\varepsilon}{s} + \frac{1000\varepsilon}{s} + o\left(\frac{1}{s}\right),$$

et comme ε est arbitraire, nous en concluons que $D(x) = o(1/s)$ uniformément en x. □

Théorème 8. *Si Y_h est définie sur une grille, le théorème 7 s'applique sous les conditions (ii) et (iii), en remplaçant H_{ch} par $H_{ch}^{\#}$ convolution de H_{ch} par la distribution triangulaire sur $[-d/2, d/2]$, d étant le pas de la grille de Y_h.*

Preuve du théorème 8. Nous prenons les convolutions de H_{ch} et de G par une distribution triangulaire sur $[-d/2, d/2]$, avec d pas de la grille sur laquelle est définie Y_h. Notons $G^{\#}$ la convolution de G par la distribution triangulaire sur $[-d/2, d/2]$, soit

$$G^{\#}(x) = \frac{2}{d}\int_{-d/2}^{d/2}\left(1 - \frac{2|y|}{d}\right)G(x-y)dy.$$

On remarque que, en notant M le maximum de $|G''|$, qu'un développement de Taylor à l'ordre 2 de G au point x permet d'écrire que

$$|G^{\#}(x) - G(x)| < \frac{1}{24}Md^2.$$

Puisque d est de l'ordre de $1/s$, pour prouver le théorème il suffit d'établir que

$$|H_{ch}^{\#}(sx) - G^{\#}(x)| = o(1/s).$$

Comme une convolution correspond à une multiplication pour les transformées de Fourier, l'équation (2.8) permet d'écrire que

$$|H_{ch}^{\#}(sx) - G^{\#}(x)| \leqslant \int_{-as}^{as} \left| \frac{e^{v\left(\frac{\zeta}{s}\right)} - e^{-\frac{1}{2}\zeta^2} - \frac{v'''(0)}{6s^3}\zeta^3 e^{-\frac{1}{2}\zeta^2}}{\zeta} \right| |\nu(\zeta)| d\zeta + \frac{24\varepsilon}{\pi s}. \tag{2.13}$$

avec $\nu(\zeta) = \frac{\sin^2(\frac{1}{2}d\zeta)}{\left(\frac{1}{2}d\zeta\right)^2}$ fonction caractéristique de la loi triangulaire. Nous pouvons appliquer tous les arguments de la démonstration précédente, en ajoutant l'argument supplémentaire

$$\int_{\delta s}^{as} \frac{|e^{v\left(\frac{\zeta}{s}\right)}\nu(\zeta)|}{\zeta} d\zeta = o(1/s). \tag{2.14}$$

Or

$$\int_{\delta s}^{as} \frac{|e^{v\left(\frac{\zeta}{s}\right)}\nu(\zeta)|}{\zeta} d\zeta = \frac{4}{(ds)^2} \int_{\delta}^{a} \frac{\left|e^{v(y)} \sin^2\left(\frac{dsy}{2}\right)\right|}{y^3} dy. \tag{2.15}$$

Or la fonction $e^{v(y)}$ a pour période $\frac{2\pi}{ds}$, de même que $\sin^2\left(\frac{dsy}{2}\right)$, donc il suffit de prouver que

$$\int_{\delta}^{\delta+\frac{2\pi}{sd}} \frac{\left|E(\omega_{h1}(y)\cdots\omega_{hN_h}(y)) \sin^2\left(\frac{dsy}{2}\right)\right|}{y^3} dy = o(1/s), \tag{2.16}$$

ou encore que

$$\int_0 \frac{\pi}{ds} \left|E_{N_h}(\omega_{h1}(y)\cdots\omega_{hN_h}(y))\right| y dy = O(1/s^2),$$

ce qui est vrai puisque dans un voisinage de l'origine,

$$|E_{N_h}(\omega_{h1}(y)\cdots\omega_{hN_h}(y))| < e^{-s^2y^2/2},$$

et on choisit $E(N)$ assez grand pour que $\log\left(\frac{a}{\delta}\right) e^{-\theta\lambda\phi(\sigma h)} \leqslant \frac{\varepsilon}{s}$ (ce qui est toujours possible vu l'encadrement (2.7) de s et les expressions (2.5) et (2.6)). □

Nous complétons maintenant la preuve du théorème 6. Nous approchons H_{ch} à l'aide du théorème 7 si S_n n'est pas treillis, ou du théorème 8 si S_n est treillis, et obtenons :

$$H_{ch}(x) = \mathfrak{N}(x/s) + \frac{\mu_3}{6s^3}(1 - (x/s)^2)\mathfrak{n}(x/s) + s^{-1}r(x/s) \tag{2.17}$$

avec $r(x) \to 0$ uniformément en x quand $E(N) \to \infty$, en remplaçant H_{ch} par $H_{ch}^{\#}$ dans le cas de variables sur des grilles.

Étudions d'abord le cas où X n'est pas définie sur une grille. Si nous notons $K(x/s) = \mathfrak{N}(x/s) + \frac{\mu_3}{6s^3}(1-(x/s)^2)\mathfrak{n}(x/s)$, nous obtenons pour l'intégrale du lemme 7 :

$$I = h\int_0^\infty e^{-hy}[K(y/s) - K(0)]dy + o(1/s),$$

soit, avec le changement de variables $x = y/s$

$$I = hs\int_0^\infty e^{-hsx}[K(x) - K(0)]dx + o(1/s).$$

En faisant une intégration par parties, nous pouvons écrire que :

$$I = \int_0^\infty e^{-hsx}K^{'}(x)dx + o(1/s).$$

Étant donné que $K^{'}(x) = \mathfrak{n}(x) + \frac{\mu_3}{6s^3}(x^3-3x)\mathfrak{n}(x)$, nous allons d'abord étudier la contribution de $\frac{\mu_3}{6s^3}(x^3-3x)\mathfrak{n}(x)$ à l'intégrale. Cette contribution est égale à :

$$\frac{\mu_3}{6s^3}\int_0^\infty e^{-hsx}(x^3-x)\mathfrak{n}(x)dx = \frac{\mu_3}{6s^3}\frac{1}{\sqrt{2\pi}}\int_0^\infty (x^3-x)e^{-hsx}e^{-\frac{x^2}{2}}dx,$$

et donc est égale à $\frac{1}{hs}o(1)$. La contribution de $\mathfrak{n}(x)$ se calcule en posant le changement de variables $y = hs + x$, ce qui donne :

$$\begin{aligned}\int_0^\infty e^{-hsx}\mathfrak{n}(x)dx &= \frac{1}{\sqrt{2\pi}}\int_{hs}^\infty e^{-hs(y-hs)}e^{-\frac{(y-hs)^2}{2}}dy\\ &= \frac{1}{\sqrt{2\pi}}e^{\frac{(hs)^2}{2}}\int_{hs}^\infty e^{-\frac{y^2}{2}}dy\\ &= e^{\frac{(hs)^2}{2}}[1-\mathfrak{N}(hs)]\end{aligned}$$

En utilisant le premier terme du développement asymptotique suivant :

$$1-\mathfrak{N}(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\{x^{-1}-x^{-3}+3x^{-5}+O(x^{-7})\} \text{ quand } x\to\infty,$$

nous obtenons :

$$\int_0^\infty e^{-hsx}\mathfrak{n}(x)dx = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}(1+o(1)).$$

Nous avons ainsi montré que

$$I = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}(1+o(1)) + \frac{1}{s}o(1).$$

Comme h est uniformément borné, nous obtenons

$$I = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}(1+o(1)),$$

ce qui conclut la preuve du théorème dans ce cas.

Si X est définie sur une grille, nous pouvons écrire que

$$I(h) = h \int_0^\infty e^{-hy}[H^*_{ch}(y) - H_{ch}(0)]dy.$$

Nous pouvons également écrire cette intégrale

$$I(h) = \sum_{k=0}^{\infty} h \int_{kd}^{(k+1)d} e^{-hy}[H^*_{ch}((k+1/2)d) - H_{ch}(0)]dy.$$

Or, aux points milieux de la grille, H^*_{ch} et $H^{\#}_{ch}$ coïncident donc nous pouvons appliquer le théorème central limite local. Nous remarquons que $H_{ch}(0) = H^{\#}_{ch}(d/2)$ et nous obtenons que

$$I(h) = \sum_{k=0}^{\infty} h \int_{kd}^{(k+1)d} e^{-hy} \left[K\left((k+1/2)\frac{d}{s}\right) - K\left(\frac{d}{2s}\right)\right] dy + o(1/s),$$

soit, en intégrant l'exponentielle :

$$I(h) = \sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \left[K\left((k+1/2)\frac{d}{s}\right) - K\left(\frac{d}{2s}\right)\right] dy + o(1/s),$$

et en écrivant la différence sur K comme une intégrale :

$$I(h) = \sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \int_{\frac{d}{2s}}^{\frac{(k+1/2)d}{s}} K'(y)dy + o(1/s).$$

De même que précédemment, $K'(x) = \mathfrak{n}(x) + \frac{\mu_3}{6s^3}(x^3 - 3x)\mathfrak{n}(x)$, et la contribution de $\frac{\mu_3}{6s^3}(x^3 - 3x)\mathfrak{n}(x)$ à l'intégrale vaut $\frac{1}{s}o(1)$. Celle de $\mathfrak{n}(x)$ vaut

$$\sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \frac{1}{\sqrt{2\pi}} \int_{\frac{d}{2s}}^{\frac{(k+1/2)d}{s}} e^{-\frac{x^2}{2}} dx.$$

Nous développons $e^{-\frac{x^2}{2}} = 1 - \frac{x^2}{2} + o(x^2)$ et ainsi l'intégrale précédente vaut

$$\sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \frac{kd}{\sqrt{2\pi}s} + o(1/s).$$

Nous sommons deux les séries :

$$\sum_{k=0}^{\infty} k e^{-hkd} = \frac{e^{-hd}}{(1 - e^{-hd})^2}$$

et

$$\sum_{k=0}^{\infty} k e^{-h(k+1)d} = \frac{e^{-2hd}}{(1 - e^{-hd})^2}$$

et ainsi nous obtenons

$$I = \frac{1}{\sqrt{2\pi}} \frac{e^{-hd} d}{s(1 - e^{-hd})}(1 + o(1)),$$

ce qui conclut la preuve du théorème. □

2.2. Application à une base de données routières

Pour appliquer cette méthodologie à la base de données routières Géoroute de l'IGN, nous avons travaillé sur un itinéraire autoroutier type. Nous avons testé l'égalité des θ_r et estimé le paramètre θ à l'aide de l'estimateur du maximum de vraisemblance. Ce paramètre θ est de l'ordre de 5% sur notre jeu test. Nous avons ensuite choisi le paramètre $E(N) = 500$, ce qui représente un parcours de 250 km environ, et appliqué le développement de grandes déviations. La vitesse de parcours de chaque tronçon est de 60 km/h sur nationale ou de 120 km/h sur autoroute. Enfin, la répartition entre nationales et autoroutes dans la base de référence a été estimée à 9/10 et 1/10. Les résultats sont présentés dans les deux graphiques de la figure 2.1.

2.3. Développement de l'asymptotique $y \to \infty$

Nous présentons dans cette section un théorème similaire à celui de la section précédente, mais pour l'asymptotique $y \to \infty$. Nous redonnons de manière détaillée l'ensemble des notations et des hypothèses du problème, car elles diffèrent légèrement de celles de la section 2.1.

Soit $X = X_1, X_2, \dots$ une suite de variables aléatoires i.i.d non dégénérées, soit $\{a_k : k = 1, 2, \dots\}$ une suite de réels positifs, et soit y une constante réelle positive. Nous considérons $Y = \sum_{i=1}^{N} a_i X_i$, où N est une variable aléatoire discrète, et étudions le comportement de la probabilité $P(Y > y)$ quand $y \to \infty$. Nous supposons que $E(X) = 0$ et que $E(X^2) = 1$. Nous notons $F(x) = P(X \leqslant x)$ la fonction de répartition de X, $\phi(t) = E(e^{tX})$ la fonction génératrice des moments de X, et $\phi_Y(t) = E(e^{tY})$ la fonction génératrice des moments de Y. Nous notons $\xi(t) = E(t^N)$. Soit $Z = Y - y$, et notons $\phi_Z(t) = E(e^{tZ}) = e^{-ty}\phi_Y(t)$.

Nous imposons la même condition de régularité sur les a_{ki} que dans le cas de sommes classiques. Nous notons $\sigma_1 = \min\{a_k\}$ et $\sigma_2 = \max\{a_k\}$

Condition I' Il existe α et θ avec $0 < \alpha \leqslant 1$, $0 < \theta \leqslant 1$, tels que pour tout n, au moins αn des a_k, $1 \leqslant k \leqslant n$, sont supérieurs ou égaux à $\theta\sigma_2$. De plus, $\sigma_1 > 0$.

Posons $Q(t) = \phi'(t)/\phi(t)$ pour $t \in \mathbb{R}$. Remarquons que Q est croissante et que l'image de Q est l'enveloppe convexe du support de X.

Condition II' X_1 est telle que $\sup\{t : \phi(t) < \infty\} = +\infty$, et son support contient des valeurs strictement positives.

Notons que si X est une loi absolument continue, Y n'est pas continue à cause d'une masse $p_0 = P(N = 0)$ en 0. On écrira dans ce cas que

$$\begin{aligned} P(Y > 0) &= P(Y > y \mid N > 0)P(N > 0) \\ &= P\left(\sum_{i=1}^{\widetilde{N}} a_i X_i > y\right)(1 - p_0), \end{aligned}$$

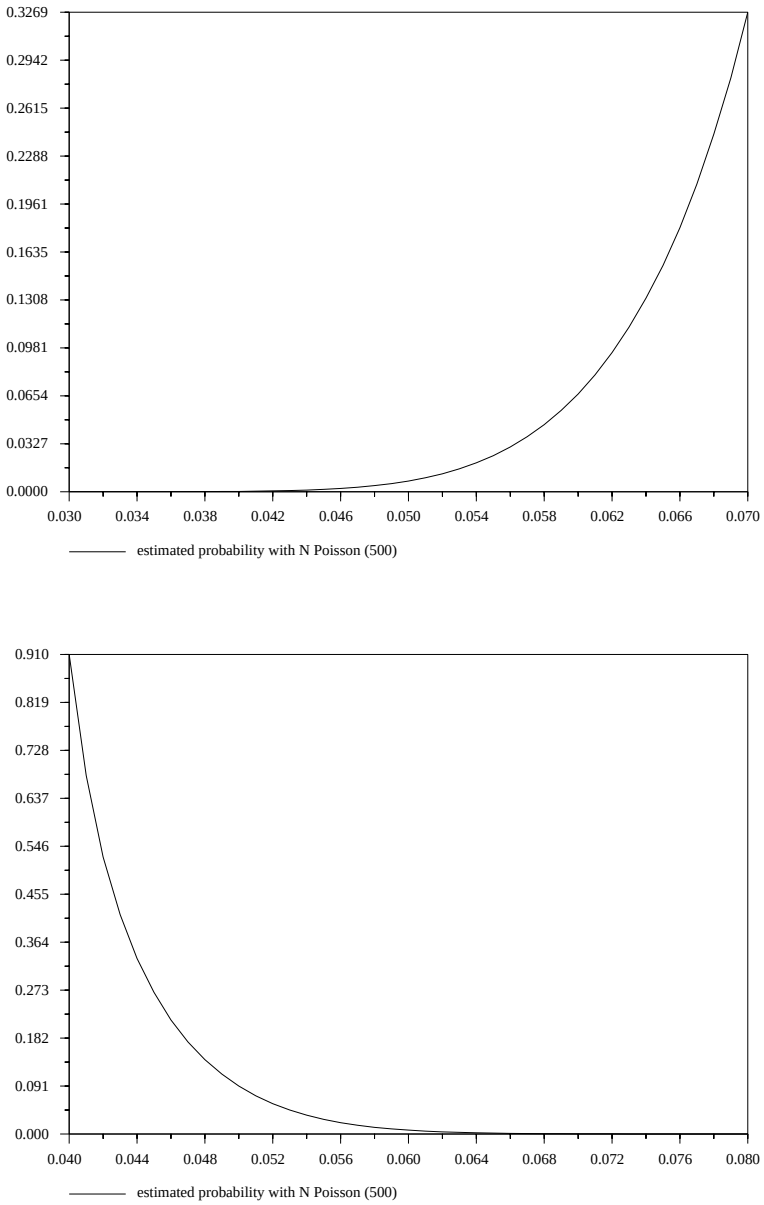

FIG. 2.1.: Probabilité de dépassement d'un seuil d'erreur relative η en fonction de θ avec $\eta = 6\%$ (haut) et en fonction de η avec $\theta = 5\%$ (bas)

avec

$$P(\widetilde{N} = k) = (1 - p_0)^{-1} p_k,$$

pour $k = 1, 2, \ldots$. Dans ce cas, on pose $\widetilde{Y} = \sum_{i=1}^{\widetilde{N}} a_i X_i$ et $E(e^{t\widetilde{Y}}) = (1-p_0)^{-1}(\phi_Y(t) - p_0)$.

Dans la suite de la démonstration, le lecteur remplacera Y par $\widetilde{Y}$ si Y n'est pas treillis.

Si X est une variable treillis et il existe au moins un rapport a_i/a_j non rationnel, alors Y n'est pas une variable treillis. Nous imposons donc la condition suivante :

Condition III Les a_i sont tels que Y est une variable treillis de pas d.

Nous calculons une expression explicite de $\phi_Y(t)$ en fonction des $p_k = P(N = k)$, de ϕ et des a_{ki} :

$$\phi_Y(t) = E(e^{t(Y)}) = E(E(e^{t(Y)} \mid N)) = \sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i t), \tag{2.18}$$

et constatons que, sous les conditions I et II, $\phi_Y(h)$ est finie pour tout h assez grand. Nous effectuons la transformation exponentielle suivante, en notant H_{c0} la fonction de répartition de Z :

$$\frac{dH_{ch}}{dH_{c0}}(x) = \frac{e^{hx}}{\phi_Z(h)}.$$

Soit Y_h une variable aléatoire distribuée suivant H_{ch}.

Nous pouvons énoncer nos théorèmes :

Théorème 9. *Supposons que les condition I' et II sont vérifiées.Soit h solution de l'équation $E(Y_h) = 0$. Posons $s^2 = \mathrm{Var}(Y_h)$, $\mu_3 = E(Y_h^3)$ et $\mu_4 = E(Y_h^4)$. Supposons de plus que $(\mu_5 - 10\mu_3 s^2)/s^5 = O(1/s^3)$, que $(\mu_4 - 3s^4)/s^4 = O(1/s^2)$, et que $h/s \to 0$ quand $y \to \infty$. Alors, quand $y \to \infty$,*

$$P(Y > y) = \frac{1}{\sqrt{2\pi}} \frac{1}{sh} e^{-hy} (\phi_Y(h) - p_0)(1 + o(1)),$$

si X n'est pas treillis et la condition (i) du théorème 10 est vérifiée, et

$$P(Y > y) = \frac{1}{\sqrt{2\pi}} \frac{de^{-hd}}{s(1 - e^{-hd})} e^{-hy} \phi_Y(h)(1 + o(1)),$$

si X est treillis et la condition III est vérifiée, avec d pas de la grille de Y.

Nous pouvons écrire :

Lemme 12.

$$P(Y > y) = e^{-hy} \phi_Y(h) I(h),$$

avec $I(h) = h \int_0^\infty e^{-hx}[H_{ch}(x) - H_{ch}(0)]dx$ si Y_h n'est pas treillis, et avec $I(h) = h \int_0^\infty exp(-hx)[H^*_{ch}(x) - H^*_{ch}(0)]dx$, si Y_h est treillis, H^*_{ch} étant définie ainsi :

$$\begin{aligned} H^*_{ch}(x) &= \frac{1}{2}[H_{ch}(x) + H_{ch}(x-)] && \text{si } x \text{ est sur un noeud du treillis} \\ &= H_{ch}(x) && \text{sinon.} \end{aligned}$$

Preuve. C'est la même que celle du cas d'une somme classique. □

Nous allons ensuite choisir un h adapté au problème, et établir un développement de $I(h)$ quand $y \to \infty$.

Calculons d'abord la fonction génératrice des moments de Y_h, que nous notons ϕ_h :

$$\phi_h(t) = E(e^{tY_h}) = \int e^{ty} dH_{ch}(y) = \frac{1}{\phi_Z(h)} \int e^{(t+h)y} dH_{c0}(y) = e^{-ty} \frac{\phi_Y(t+h)}{\phi_Y(h)}.$$

En introduisant l'expression (2.18) de $\phi_Y(t)$, nous obtenons :

$$\phi_h(t) = e^{-ty} \frac{\sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i(t+h))}{\sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i h)}. \tag{2.19}$$

Nous écrivons Y_h sous la forme $Y_h = \sum_{i=1}^{N_h} X_{hi} - y$, et calculons la fonction génératrice des moments de Y_h en fonction de ϕ_{hi} fonction génératrice des moments de X_{hi} :

$$\phi_h(t) = e^{-ty} \sum_{k=0}^{\infty} \prod_{i=1}^{k} \phi_{hi}(t) P(N_h = k). \tag{2.20}$$

En identifiant les termes des expressions (2.19) et (2.20), nous obtenons

$$P(N_h = k) = \frac{\prod_{i=1}^{k} \phi(a_i h)}{\sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i h)} p_k,$$

et

$$\phi_{hi}(t) = \frac{\phi(a_i(t+h))}{\phi(a_i h)},$$

donc les X_{hi} ont pour loi F_{hi} définie par

$$\frac{dF_h}{dF}(x) = \frac{e^x}{\phi(a_i h)}.$$

Notons que $E(N_h)$ varie avec y. On peut écrire :

$$E(N_h) = \frac{\sum_{k=0}^{\infty} k p_k \prod_{i=1}^{k} \phi(a_i h)}{\sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i h)}.$$

Posons $\psi_k(h) = \prod_{i=1}^{k} \phi(a_i h)$. Alors $E(N_h)$ s'écrit :

$$E(N_h) = \frac{\sum_{k=0}^{\infty} k p_k \psi_k(h)}{\sum_{k=0}^{\infty} p_k \psi_k(h)}.$$

Dans le cas où N suit une loi de Poisson de paramètre λ, on obtient :

$$\begin{aligned} E(N_h) &= \frac{\sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \psi_k(h)} = \frac{\sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \psi_k(h)} = \lambda \frac{\sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{k-1!} \psi_k(h)}{\sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \psi_k(h)} \\ &= \lambda \frac{\sum_{k=0}^{\infty} p_k \psi_{k+1}(h)}{\sum_{k=0}^{\infty} p_k \psi_k(h)}. \end{aligned}$$

Or, avec la condition I',

$$\phi(\sigma_1)\psi_k(h) \leqslant \psi_{k+1}(h) \leqslant \psi_k(h)\phi(\sigma_2 h)$$

puisque $\sigma_1 < a_{ki} \leqslant \sigma_2 \ \forall i \ \forall k$, donc on obtient pour $E(N_h)$ l'encadrement :

$$\lambda\phi(\sigma_1 h) \leqslant E(N_h) \leqslant \lambda\phi(\sigma_2 h). \tag{2.21}$$

Remarquons que par des calculs extrêmement similaires, nous obtenons pour $\mathrm{Var}(N_h)$ l'encadrement suivant :

$$\lambda\phi(\sigma_1 h) + \lambda^2(\phi(\sigma_1 h)^2 - \phi(\sigma_2 h)^2) \leqslant \mathrm{Var}(N_h) \leqslant \lambda\phi(\sigma_2 h) + \lambda^2(-\phi(\sigma_1 h)^2 + \phi(\sigma_2 h)^2). \tag{2.22}$$

Lemme 13.

$$E(1-\theta)^{N_h} \leqslant e^{-\theta\lambda\phi(\sigma_2 h)}.$$

Preuve. Elle est identique à celle du lemme 9. □

Moyennant des conditions sur X et sur les a_i, nous choisissons h tel que $E(Y_h) = 0$. Comme $E(Y_h) = \phi'_h(0) = \phi'_Y(h)/\phi_Y(h)$, h existe pour tout y puisque $Im(\phi'_Y/\phi_Y)$ est l'enveloppe convexe du support de Y (on suppose que N suit une loi de Poisson et a ainsi un support infini).

Calculons $E(Y_h)$:

$$E(Y_h) = E\left(E\left(\sum_{i=1}^{N_h} X_{hi} \mid N_h\right)\right) - y = \sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} E(X_{hi}) - y.$$

Or

$$E(X_{hi}) = a_i \frac{\phi'(a_i h)}{\phi(a_i h)},$$

d'où

$$E(Y_h) = \sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} a_i Q(a_i h) - y.$$

L'équation $E(Y_h) = 0$ est équivalente à $\sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} a_i Q(a_i h) = y$. La condition I' permet d'écrire que

$$\sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} a_i Q(a_i h) \geqslant \alpha\theta\sigma_2 Q(\theta\sigma_2 h) E(N_h).$$

qui tend vers ∞ quand $h \to \infty$ (la condition II' assure que Q est positive pour h assez grand et (2.21) prouve que $E(N_h) \to \infty$ quand $h \to \infty$). Ainsi, l'existence de h solution de l'équation $E(N_h) = 0$ est établie. De même, on peut écrire que

$$0 = E(Y_h) \leqslant \sum_{k=0}^{\infty} q_k k\sigma_2 Q(\sigma_2 h) - y,$$

soit

$$Q(\sigma_2 h) \geqslant \frac{y}{\sigma_2 E(N_h)}. \tag{2.23}$$

En utilisant les inégalités (2.21) et (2.23), on peut écrire que :

$$\frac{y}{\sigma_2}\frac{\phi(\sigma_2 h)}{\phi'(\sigma_2 h)} \leqslant E(N_h) \leqslant \lambda\phi(\sigma_2 h),$$

et donc nous obtenons :

$$\phi'(\sigma_2 h) \geqslant \frac{y}{\sigma_2 \lambda},$$

ce qui prouve que quand $y \to \infty$, $\phi'(\sigma_2 h) \to \infty$ et donc $h \to \infty$. Vu l'encadrement (2.21), on a ainsi établi que $E(N_h) \to \infty$ quand $y \to \infty$.

Nous étudions ensuite $\mathrm{Var}(Y_h)$. Dans notre cas, comme h est choisi tel que $E(Y_h) = 0$, $\mathrm{Var}(Y_h) = E(Y_h^2)$. Nous obtenons alors :

$$\mathrm{Var}(Y_h) = E\left[\sum_{i=1}^{N_h} \mathrm{Var}(X_{hi})\right] + \mathrm{Var}\left[\sum_{i=1}^{N_h} E(X_{hi})\right].$$

En introduisant

$$\mathrm{Var}(X_{hi}) = a_i^2 Q'(a_i h),$$

nous obtenons

$$\mathrm{Var}(Y_h) = \sum_{k=0}^{\infty} q_k \sum_{i=1}^{k} a_i^2 Q'(a_i h) + \mathrm{Var}\left[\sum_{i=1}^{N_h} a_i Q(a_i h) - y\right].$$

Par la condition I' nous en déduisons l'encadrement suivant, pour tout h solution de l'équation $E(Y_h = 0)$:

$$Q'(\theta\sigma_2 h)\alpha\theta^2\sigma_2^2 E(N_h) \leqslant \mathrm{Var}(Y_h) \leqslant Q'(\sigma_2 h)\sigma_2^2 E(N_h) + (\sigma_2 Q(\sigma_2 h))^2 \mathrm{Var}(N_h). \tag{2.24}$$

On remarque que, quand $y \to \infty$, $\mathrm{Var}(Y_h) \to \infty$ puisque $E(N_h) \to \infty$.

Nous établissons ensuite une approximation de $H_{ch}(sx)$ par la fonction

$$G(x) = \mathfrak{N}(x) + \frac{\mu_3}{6s^3}(1-x^2)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(-15x + 10x^3 - x^5)\mathfrak{n}(x) + \frac{\mu_4 - 3s^4}{24s^4}(3x - x^3)\mathfrak{n}(x),$$

avec $\mu_i = E(Y_h)^i$, $i = 3, 4$ et $s^2 = \mathrm{Var}(Y_h)$.

Théorème 10. *Soit Y_h définie précédemment. Notons $s^2 = \mathrm{Var}\, Y_h$ et ω_{hj} fonction caractéristique de X_{hj}. Si Y_h vérifie :*

(i) $|\omega_{hj}(\zeta)| < 1 - \theta(\delta, a)$ $\forall j$ $\forall k$ *pour as* $> \zeta > \delta > 0$;

(ii) $(\mu_5 - 10\mu_3 s^2)/s^5 = O(1/s^3)$;

(iii) $(\mu_4 - 3s^4)/s^4 = O(1/s^2)$;

alors

$$H_{ch}(sx) = \mathfrak{N}(x) + \frac{\mu_3}{6s^3}(1-x^2)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(-15x+10x^3-x^5)\mathfrak{n}(x) + \frac{\mu_4 - 3s^4}{24s^4}(3x-x^3)\mathfrak{n}(x) + \frac{1}{s^2}r_s(x),$$

avec $r_s(x) \to 0$ *quand* $y \to \infty$.

Preuve. On étudie la transformée de Fourier de $D(x) = H_{ch}(sx) - G(x)$. Nous écrivons $\omega(\zeta) = \phi_h(i\zeta)$ fonction caractéristique de Y_h sous la forme $\sum_{k=0}^{\infty} q_k e^{v_k(\zeta)}$, soit posons $v_k(\zeta) = \sum_{j=1}^{k} \log(\phi_{hj}(i\zeta))$. Notons que, puisque ω est une fonction caractéristique, $\omega(0) = 1$, que $\omega'(0) = E(Y_h) = 0$, que $\omega''(O) = i^2s^2$, que $\omega^{(3)}(0) = i^3\mu_3$ et que $\omega^{(4)}(0) = i^4\mu_4$ Dans ce cas, des dérivations successives montrent que $\sum_{k=0}^{\infty} q_k v_k(0) = 0$, $\sum_{k=0}^{\infty} q_k v_k'(0) = 0$, $\sum_{k=0}^{\infty} q_k v_k''(0) = i^2s^2$, $\sum_{k=0}^{\infty} q_k v_k'''(0) = i^3\mu_3$ et $\sum_{k=0}^{\infty} q_k v_k^{(4)}(0) = i^4(\mu_4 - 3s^4)$. Remarquons que $|G'(x)| = O(\mu_3/s^3)$ quand $y \to \infty$. Le polynôme $G(x)$ peut être vu comme un polynôme en x ou en $1/s$. Plutôt que d'étudier la transformée de Fourier de $D(x)$, nous allons étudier la transformée de Fourier d'un polynôme $E(x)$ dont la troncature en $1/s$ des termes plus petits que $1/s$ est égale à $D(x)$. A cet effet, on pose $\beta_k = \frac{v_k'''(0)}{6s^3}i^3\zeta^3 + \frac{v_k^{(4)}(0)}{24s^4}i^4\zeta^4$, et $\beta = \sum_{k=0}^{\infty} q_k\beta_k$.

Soit $\varepsilon > 0$ fixé. Nous choisissons une constante a suffisamment grande pour que $|G'(x)| < \varepsilon a$ pour pour tout x et pour tout y. D'après un théorème de lissage (page 538 de [Fel70]), nous pouvons écrire :

$$|E(x)| \leqslant \frac{1}{\pi}\int_{-as^2}^{as^2} \left| \frac{\sum_{k=0}^{\infty} q_k \left[e^{v_k\left(\frac{\zeta}{s}\right)} - e^{-\frac{1}{2}\zeta^2} - e^{-\frac{1}{2}\zeta^2}\left(\beta_k + \beta_k^2/2\right)\right]}{\zeta} \right| d\zeta + \frac{24\varepsilon}{\pi s^2}. \qquad (2.25)$$

Nous séparons cette intégrale en deux domaines d'intégration. Le premier est le domaine défini par $\delta s \leqslant |\zeta| \leqslant as^2$, et le second par $|\zeta| < \delta s$, δ étant un réel positif fixé que nous précisons plus loin. Sur le domaine $\delta s \leqslant |\zeta| \leqslant as^2$, on a tous les $|\phi_{hj}(it)| < 1 - \theta \ \forall j \ \forall k$, avec $\theta > 0$. Alors $|\omega(\zeta/s)| < E(1-\theta)^{N_h}$, que nous majorons à l'aide du lemme 13. Ainsi, la contribution de ce domaine à l'intégrale (2.25) est inférieure à

$$\log\left(\frac{a}{\delta}\right) e^{-\theta\lambda\phi(\sigma_2 h)} + \int_{\delta s \leqslant |\zeta| \leqslant as} \frac{e^{-\frac{1}{2}\zeta^2}}{\zeta}\left(1 + \left|\beta + \beta^2/2\right|\right) d\zeta.$$

Le terme de droite de l'équation précédente est un petit o de n'importe quelle puissance de s.

Pour le domaine $|\zeta| < \delta s$, en posant

$$\psi_k(\zeta) = v_k(\zeta) + \frac{1}{2}s^2\zeta^2$$

l'intégrande de (2.25) se réécrit

$$\frac{e^{-\frac{1}{2}\zeta^2}}{\zeta}\left|\sum_{k=0}^{\infty} q_k\left[e^{\psi_k\left(\frac{\zeta}{s}\right)} - 1 - \beta_k - \beta_k^2/2\right]\right| d\zeta \qquad (2.26)$$

et nous allons l'estimer à l'aide de l'inégalité suivante tirée de [Fel70] :

$$|e^{\alpha} - 1 - \beta - \beta^2/2| \leqslant (|\alpha - \beta| + \frac{1}{6}|\beta|^3)e^{\gamma}, \tag{2.27}$$

avec $\gamma \geqslant \max(|\alpha|, |\beta|)$, pour α et β arbitraires, réels ou complexes. Nous pouvons faire un développement de Taylor au quatrième ordre de ψ. Comme $v^{(5)} = (\mu_5 - 10\mu_3 s^2)/s^5$, la condition (iv) permet de déduire l'existence d'un δ tel que

$$\left|\psi k\left(\frac{\zeta}{s}\right) - \frac{v_k'''(0)}{6s^3} i^3\zeta^3 - \frac{v_k^{(4)}(0)}{24s^4} i^4\zeta^4\right| < \varepsilon k \left|\frac{\zeta}{s}\right|^4 \tag{2.28}$$

pour $|\zeta| < \delta s$. Nous prenons δ suffisamment petit pour avoir également

$$\left|\psi_k\left(\frac{\zeta}{s}\right)\right| < \frac{1}{4}\zeta^2, \qquad \left|\frac{v_k'''(0)}{6s^3}\zeta^3 - \frac{v_k^{(4)}(0)}{24s^4}\zeta^4\right| \leqslant \frac{1}{4}\zeta^2$$

pour $|\zeta| < \delta s$. Avec ce choix de δ nous majorons l'intégrale (2.25) sur le domaine $|\zeta| < \delta s$ en utilisant la formule (2.27) par :

$$\int_{|\zeta|<\delta s} e^{-\frac{1}{4}\zeta^2} \left|\frac{\varepsilon E(N_h)}{s^4}|\zeta|^3 + \frac{1}{6\zeta}\left|\frac{\mu_3^2}{s^6}\zeta^5 + \frac{\mu_4 - 3s^4}{24s^4}\zeta^4\right|^2\right| d\zeta. \tag{2.29}$$

En choisissant y assez grand pour que $\log\left(\frac{a}{\delta}\right) e^{-\theta\lambda\phi(\sigma_2 h)} \leqslant \frac{\varepsilon}{s^2}$ et que l'intégrale (2.29) soit inférieure à $\frac{1000\varepsilon}{s^2}$, nous avons montré que pour tout x

$$|E(x)| \leqslant \frac{24\varepsilon}{\pi s^2} + \frac{\varepsilon}{s^2} + \frac{1000\varepsilon}{s^2} + o\left(\frac{1}{s^2}\right),$$

et comme ε est arbitraire, nous en concluons que $E(x) = o(1/s^2)$ uniformément en x. Nous obtenons le développement annoncé en considérant le polynôme comme un polynôme en $1/s$, et en négligeant les termes plus petits que $1/s^2$. □

Théorème 11. *Si Y_h est définie sur une grille, le théorème 10 s'applique sous les conditions (ii), (iii) et (iiv), en remplaçant H_{ch} par $H_{ch}^{\#}$ convolution de H_{ch} par la distribution triangulaire sur $[-d/2, d/2]$, d étant le pas de la grille de Y_h, et avec un reste de la forme $O(1/s^2)$.*

Preuve. Nous prenons les convolutions de H_{ch} et de G par une distribution triangulaire sur $[-d/2, d/2]$, avec d pas de la grille sur laquelle est définie Y_h. Notons $G^{\#}$ la convolution de G par la distribution triangulaire sur $[-d/2, d/2]$, soit

$$G^{\#}(x) = \frac{2}{d}\int_{-d/2}^{d/2}\left(1 - \frac{2|y|}{d}\right) G(x-y)dy.$$

On remarque que, en notant M le maximum de $|G''|$, qu'un développement de Taylor à l'ordre 2 de G au point x permet d'écrire que

$$|G^{\#}(x) - G(x)| < \frac{1}{24}Md^2.$$

Puisque d est de l'ordre de $1/s$, pour prouver le théorème il suffit d'établir que

$$|H_{ch}^{\#}(sx) - G^{\#}(x)| = O(1/s^2).$$

Nous allons poser $D^{\#} = H_{ch}^{\#}(sx) - G^{\#}(x)$, et étudier comme précédemment un polynôme $E^{\#}$ dont la troncature en $1/s$ est égale à $D^{\#}$. Comme une convolution correspond à une multiplication pour les transformées de Fourier, l'équation (2.25) permet d'écrire que

$$|E^{\#}(x)| \leqslant \int_{-as^2}^{as^2} \left| \frac{e^{v\left(\frac{\zeta}{s}\right)} - e^{-\frac{1}{2}\zeta^2} - e^{-\frac{1}{2}\zeta^2}\left(\beta + \beta^2/2\right)}{\zeta} \right| |\nu(\zeta)| d\zeta + \frac{24\varepsilon}{\pi s^2}. \tag{2.30}$$

avec $\nu(\zeta) = \frac{\sin^2(\frac{1}{2}d\zeta)}{\left(\frac{1}{2}d\zeta\right)^2}$ fonction caractéristique de la loi triangulaire. Nous pouvons appliquer tous les arguments de la démonstration précédente, en ajoutant l'argument supplémentaire

$$\int_{\delta s}^{as^2} \frac{|e^{v\left(\frac{\zeta}{s}\right)}\nu(\zeta)|}{\zeta} d\zeta = O(1/s^2). \tag{2.31}$$

Or

$$\int_{\delta s}^{as^2} \frac{|e^{v\left(\frac{\zeta}{s}\right)}\nu(\zeta)|}{\zeta} d\zeta = \frac{4}{(ds)^2} \int_{\delta}^{as} \frac{\left|e^{v(y)} \sin^2\left(\frac{dsy}{2}\right)\right|}{y^3} dy. \tag{2.32}$$

Or la fonction $e^{v(y)}$ a pour période $\frac{2\pi}{ds}$, de même que $\sin^2\left(\frac{dsy}{2}\right)$, donc il suffit de prouver que

$$\int_{\delta}^{\delta + \frac{2\pi}{sd}} \frac{\left|E(\omega_{h1}(y) \cdots \omega_{hN_h}(y)) \sin^2\left(\frac{dsy}{2}\right)\right|}{y^3} dy = O(1/s^2). \tag{2.33}$$

ou encore que

$$\int_0 \frac{\pi}{ds} \left|E_{N_h}(\omega_{h1}(y) \cdots \omega_{hN_h}(y))\right| y dy = O(1/s^2),$$

ce qui est vrai puisque dans un voisinage de l'origine,

$$|E_{N_h}(\omega_{h1}(y) \cdots \omega_{hN_h}(y))| < e^{-s^2y^2/2},$$

et on choisit y assez grand pour que $\log\left(\frac{a}{\delta}\right) e^{-\theta\lambda\phi(\sigma h)} \leqslant \frac{\varepsilon}{s^2}$. □

Nous approchons H_{ch} avec le résultat précédent, et obtenons :

$$H_{ch}(x) = \mathfrak{N}(x/s) + \frac{\mu_3}{6s^3}(1 - (x/s)^2)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(-15(x/s) + 10(x/s)^3 - (x/s)^5)\mathfrak{n}(x/s) + \frac{\mu_4 - 3s^4}{24s^4}(3(x/s) - (x/s)^3)\mathfrak{n}(x/s) + \frac{1}{s^2} r_s(x),$$

avec $r_s(x) \to 0$ uniformément en x quand $y \to \infty$, en remplaçant H_{ch} par $H_{ch}^{\#}$ dans le cas de variables sur des grilles.

Étudions d'abord le cas où X n'est pas définie sur une grille. Si nous notons $K(x/s) = \mathfrak{N}(x/s) + \frac{\mu_3}{6s^3}(1-(x/s)^2)\mathfrak{n}(x/s)$, nous obtenons pour l'intégrale du lemme 12 :

$$I = h\int_0^\infty e^{-hz}[K(z/s) - K(0)]dz + o(1/s^2),$$

soit, avec le changement de variables $x = z/s$

$$I = hs\int_0^\infty e^{-hsx}[K(x) - K(0)]dx + o(1/s^2).$$

En faisant une intégration par parties, nous pouvons écrire que :

$$I = \int_0^\infty e^{-hsx}K^{'}(x)dx + o(1/s^2).$$

Étant donné que $K^{'}(x) = \mathfrak{n}(x) + \frac{\mu_3}{6s^3}(x^3-3x)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(x^6-15x^4+45x^2-15)\mathfrak{n}(x) + \frac{\mu_4-3s^4}{24s^4}(x^4-6x^2+3)\mathfrak{n}(x)$, nous allons d'abord étudier la contribution de $\mathfrak{n}(x)$ à l'intégrale. Elle se calcule en posant le changement de variables $y = hs + x$, ce qui donne :

$$\begin{aligned}\int_0^\infty e^{-hsx}\mathfrak{n}(x)dx &= \frac{1}{\sqrt{2\pi}}\int_{hs}^\infty e^{-hs(y-hs)}e^{-\frac{(y-hs)^2}{2}}dy\\ &= \frac{1}{\sqrt{2\pi}}e^{\frac{(hs)^2}{2}}\int_{hs}^\infty e^{-\frac{y^2}{2}}dy\\ &= e^{\frac{(hs)^2}{2}}[1-\mathfrak{N}(hs)].\end{aligned}$$

En utilisant le premier terme du développement asymptotique suivant :

$$1-\mathfrak{N}(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\{x^{-1}-x^{-3}+3x^{-5}+O(x^{-7})\} \text{ quand } x\to\infty,$$

nous obtenons :

$$\int_0^\infty e^{-hsx}\mathfrak{n}(x)dx = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}(1+o(1)).$$

La contribution de $\frac{\mu_3}{6s^3}(x^3-3x)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(x^6-15x^4+45x^2-15)\mathfrak{n}(x) + \frac{\mu_4-3s^4}{24s^4}(x^4-6x^2+3)\mathfrak{n}(x)$ est égale à $\frac{1}{hs}O\left(\frac{\mu_3^2}{s^6}+\frac{\mu_4-3s^4}{s^4}\right)$. Nous avons ainsi montré que

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}(1+o(1)) + \frac{1}{hs}O\left(\frac{\mu_3^2}{s^6}+\frac{\mu_4-3s^4}{s^4}\right) + o\left(\frac{1}{s^2}\right).$$

Comme $1/s^2 \leqslant 1/hs$, nous obtenons

$$I_n = \frac{1}{\sqrt{2\pi}}\frac{1}{hs}\left(1+O\left(\frac{\mu_3^2}{s^6}+\frac{\mu_4-3s^4}{s^4}\right)\right),$$

ce qui conclut la preuve du théorème dans ce cas.

Si X est définie sur une grille, nous pouvons écrire que

$$I(h) = h \int_0^\infty e^{-hy}[H^*_{ch}(y) - H_{ch}(0)]dy.$$

Nous pouvons également écrire cette intégrale

$$I(h) = \sum_{k=0}^{\infty} h \int_{kd}^{(k+1)d} e^{-hy}[H^*_{ch}((k+1/2)d) - H_{ch}(0)]dy.$$

Or, aux points milieux de la grille, H^*_{ch} et $H^{\#}_{ch}$ coïncident donc nous pouvons appliquer le théorème central limite local. Nous remarquons que $H_{ch}(0) = H^{\#}_{ch}(d/2)$ et nous obtenons que

$$I(h) = \sum_{k=0}^{\infty} h \int_{kd}^{(k+1)d} e^{-hy} \left[K\left((k+1/2)\frac{d}{s}\right) - K\left(\frac{d}{2s}\right)\right] dy + o(1/s^2),$$

soit, en intégrant l'exponentielle :

$$I(h) = \sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \left[K\left((k+1/2)\frac{d}{s}\right) - K\left(\frac{d}{2s}\right)\right] dy + o(1/s^2),$$

et en écrivant la différence sur K comme une intégrale :

$$I(h) = \sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \int_{\frac{d}{2s}}^{\frac{(k+1/2)d}{s}} K^{'}(y)dy + o(1/s^2).$$

De même que précédemment, $K^{'}(x) = \mathfrak{n}(x) + \frac{\mu_3}{6s^3}(x^3 - 3x)\mathfrak{n}(x) + \frac{\mu_3^2}{76s^6}(x^6 - 15x^4 + 45x^2 - 15)\mathfrak{n}(x) + \frac{\mu_4 - 3s^4}{24s^4}(x^4 - 6x^2 + 3)\mathfrak{n}(x)$. La contribution de $\mathfrak{n}(x)$ vaut

$$\sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \frac{1}{\sqrt{2\pi}} \int_{\frac{d}{2s}}^{\frac{(k+1/2)d}{s}} e^{-\frac{x^2}{2}} dx.$$

Nous développons $e^{-\frac{x^2}{2}} = 1 - \frac{x^2}{2} + o(x^2)$ et ainsi l'intégrale précédente vaut

$$\sum_{k=0}^{\infty} (e^{-hkd} - e^{-h(k+1)d}) \frac{kd}{\sqrt{2\pi}s} + o(1/s^2)$$

Nous sommons deux les séries :

$$\sum_{k=0}^{\infty} k e^{-hkd} = \frac{e^{-hd}}{(1 - e^{-hd})^2}$$

et

$$\sum_{k=0}^{\infty} k e^{-h(k+1)d} = \frac{e^{-2hd}}{(1 - e^{-hd})^2}$$

La contribution de $\frac{\mu_3}{6s^3}(x^3-3x)\mathfrak{n}(x)+\frac{\mu_3^2}{76s^6}(x^6-15x^4+45x^2-15)\mathfrak{n}(x)+\frac{\mu_4-3s^4}{24s^4}(x^4-6x^2+3)\mathfrak{n}(x)$ à l'intégrale vaut $\frac{e^{-hd}d}{s(1-e^{-hd})}\left(-15\frac{\mu_3^2}{76s^6}+3\frac{\mu_4-3s^4}{24s^4}+O\left(\frac{\mu_4-3s^4}{s^4}\right)\right)$ et ainsi, puisque $O(1/s^2)=1/hsO((\mu_4-3s^4)/s^4))$, nous obtenons

$$I(h)=\frac{1}{\sqrt{2\pi}}\frac{e^{-hd}d}{s(1-e^{-hd})}\left(1-15\frac{\mu_3^2}{76s^6}+3\frac{\mu_4-3s^4}{24s^4}+O\left(\frac{\mu_4-3s^4}{s^4}\right)\right),$$

ce qui conclut la preuve du théorème.

Bibliographie

[Aas85] K. K. Aase. Accumulated claims and collective risk in assurance : Higher order asymptotic approximations. *Scand. Actuarial J.*, 65–85, 1985.

[Abb94] Iyad Abbas. *Base de données vectorielles et erreur cartographique : Problèmes posés par le contrôle ponctuel ; une méthode alternative fondée sur la distance de Hausdorff : le contrôle linéaire.* Thèse de doctorat de l'université Paris 7, 1994.

[Alb95] Jochen Albrecht. *Universal Analytical GIS Operations.* PhD Thesis, University of Vechta, Germany, 1995.
(`http://www.geog.ucsb.edu/~jochen/diss/gruen4.html`).

[BB] P. Barbe et M. Broniatowski. A sharp petrov type deviation formula. *article soumis.*

[bHA97] Atef bel Hadj Ali. *Appariement géométrique des objets géographiques et étude des indicateurs de qualité.* Mémoire de stage de DEA SIG, ENSG, 1997.

[BM94] Michel Broniatowski et David M. Mason. Extended large deviations. *Journ. Theoret. Prob*, **7**(3), 647–666, 1994.

[Bon98] Olivier Bonin. Attribute uncertainty propagation in vector geographic information systems : Sensitivity analysis. *IEEE / Computer Society, Maurizio Rafanelli and Mathias Jarke editors*, 254–259, 1998.

[Bon99] Olivier Bonin. Sensibilité des applications géographiques aux incertitudes : lien avec le contrôle qualité. *Bulletin d'information de l'IGN*, **70**, 71–76, 1999.

[Bon00a] Olivier Bonin. Error simulation in road databases. *GIM International*, **15**, n^{O} 3, 2000.

[Bon00b] Olivier Bonin. New advances in error simulation in vector geographical databases. *Accuracy 2000, G.B.M. Heuvelink and M.J.P.M. Lemmens editors*, 59–65, 2000.

[Bon01] O. Bonin. Grandes déviations pour des sommes pondérées de variables aléatoires i.i.d. appliquées à un problème géographique. *C. R. Acad. Sci. Sér. I Math*, t. 333, Série I, 369–372, 2001.

[Bon02] O. Bonin. Large deviation theorems for weighted sums applied to a geographical problem. A paraître au *Journal of Applied Probability*, 2002.

[Boo72] Stephen A. Book. Large deviation probabilities for weighted sums. *The Annals of Mathematical Statistics*, **43**, n^{o} 4, 1221–1234, 1972.

[Boo73] Stephen A. Book. A large deviation theorem for weighted sums. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **26**, 43–49, 1973.

[Bor87] A. A. Borovkov. *Statistique mathématique*. Moscou, Ed. Mir., 1987.

[BR60] R. R. Bahadur et R. Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist.*, **31**, 1015–1027, 1960.

[Bro87] M. Broniatowski. Grandes, très grandes et petites déviations pour des suites de variables aléatoires indépendantes équidistribuées. *C. R. Acad. Sci. Sér. I Math*, **305**, 627–630, 1987.

[BY95] Jean-Yves Boissonnat et Mariette Yvinec. *Géométrie algorithmique*. Ediscience International, 1995.

[Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, **23**, 493–507, 1952.

[Cou97] Pierre Couget. *Étude d'un outil de bruitage de la qualité de données géographiques*. Rapport de stage de DESS AIST, Paris VI, 1997.

[Cra70] Harald Cramér. *Random Variables and Probability Distributions*. Cambridge at the University Press, 1970.

[CS85] Narasinga R. Chaganty et J. Sethuraman. Large deviation local limit theorms for arbitrary sequences of random variables. *The Annals of Probability*, **13**, n^{o} 1, 97–114, 1985.

[CS93] Narasinga Rao Chaganty et Jayaram Sethuraman. Strong large deviation theorems and local limit theorems. *The Annals of Probability*, **21**, n^{o} 3, 1671–1690, 1993.

[DF97] Benoît David et Pascal Fasquel. *Qualité d'une base de données géographique : concepts et terminologie*. BI de l'IGN n^{o} 67, 1997.

[Dij59] E. W. Dijkstra. A note on two problems in connection with graphs. *Nümerische Mathematik*, **1**, 269–271, 1959.

[DRS95] Benoît David, Laurent Raynal, et Guylaine Schorter. Building an oo-gis prototype : experiments with géo2. *ACSM/ASPRS AutoCarto 12*, **4**, 137–146, 1995.

[DS89] Deuschel et Strook. *Large deviations*. Academic Press, 1989.

[EJMT85] Paul Embrechts, Jens L. Jensen, Makoto Maejima, et J. L. Teugels. Approximations for compound poisson and polya processes. *Adv. Appli. Prob.*, **17**, 623–637, 1985.

[Ess32] F. Esscher. On the probability function in the collective theory of risk. *Skand. Akt. Tidskr.*, 78–86, 1932.

[Ess37] Carl Gustav Esseen. Fourier analysis of distribution functions. *Acta Mathem.*, **77**, 1–125, 1937.

[Fai96] Sami Othman Faiz. *Modélisation, exploitation et visualisation de l'information qualité dans les Bases de Données Géographiques.* Thèse de l'université de Paris-Sud, UFR scientifique d'Orsay, 1996.

[Fas94] Pascal Fasquel. *Expression de contrôle de cohérence géographique dans un langage indépendant des SIG.* Mémoire de stage de DEA SIG, ENSG, 1994.

[Fel70] William Feller. *An Introduction to Probability Theory and Its Applications (volume II).* John Wiley & Sons, 1970.

[Fis91] Peter F. Fisher. Modelling soil map-unit inclusions by monte carlo simulation. *Int. J. Geographical Information Systems*, **5**, n^{O} 2, 193–208, 1991.

[Fou99] Lucie Fouqué. *Simulation d'erreurs dans une base de données géographique.* Mémoire de stage de DESS de Mathématiques Appliquées, (Statistique et modèles stochastiques), Université de Rennes I, 1999.

[GG98] Michael Goodchild et Sucharita Gopal. *Accuracy of Spatial Databases.* Hermes, 1998.

[GJ98] Michael Goodchild et Robert Jeansoulin. *Data Quality in Geographic Information.* Hermes, 1998.

[Gon01] Carlos Goncalves. *Contrôle de la qualité d'une base de données géographique.* Mémoire de stage de DA de Statistique, Université Paris VI, 2001.

[GW94] Sucharita Gopal et Curtis Woodcock. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric engineering and remote sensing*, **LX**, n^{O} 2, 181, 1994.

[HB93] Gerard B. M. Heuvelink et Peter A. Burrough. Error propagation in cartographic modelling using boolean and continuous classification. *Int. J. Geographical Information Systems*, **7**, n^{O} 4, 231–246, 1993.

[Heu93] Gerard B. M. Heuvelink. *Error propagation in quantitative spatial modelling, applications in GIS.* Netherlands Geographical Studies, 1993.

[Hög79] T. Höglund. A unified formulation of the central limit theorem for small and large deviations from the mean. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **49**, 105–117, 1979.

[Hwa98] Hsien-Kuei Hwang. Large deviations of combinatorial distributions. ii : Local limit theorems. *The Annals of Applied Probability*, **8**, n^{O} 1, 163–181, 1998.

[Jen88] J. L. Jensen. Uniform saddlepoint approximations. *Adv. Appli. Prob.*, **20**, 622–634, 1988.

[Jen95] Jens Ledet Jensen. *Saddlepoint approximations.* Clarendon Press, Oxford, 1995.

[Kol97] John E. Kolassa. *Series Approximation Methods in Statistics.* Lecture Notes in Statistics – Springer-Verlag, 1997.

[LMS90] Weldon A. Lodwick, William Monson, et Larry Svoboda. Attribute error and sensitivity analysis of map operations in GIS : suitability analysis. *Int. J. Geographical Information Systems*, **4**, n$^{\text{o}}$ 4, 413–428, 1990.

[McM96] Susanna McMaster. *Assessing the impact of data quality on forest management decisions using geographical sensitivity analysis*. GISDATA'96 Summer Institute, to appear, 1996.

[Nag79] S. V. Nagaev. Large deviations of sums of independant random variables. *The Annals of Probability*, **7**, n$^{\text{o}}$ 5, 745–789, 1979.

[Pen94] Eric Pennors. *La qualité des données géographiques : une méthode de contrôle des objets linéaires*. Mémoire de stage de DEA SIG, ENSG, 1994.

[Pet75] Valentin V. Petrov. *Sums of Independent Random Variables*. Springer–Verlag, 1975.

[Pet95] Valentin V. Petrov. *Limit Theorems of Probability Theory : Sequences of Independant Random Variables*. Clarendon Press – Oxford, 1995.

[Rav96] Benoît Ravel. *Modélisation des imprécisions géométriques dans les Bases de Données Géographiques*. Rapport de stage de deuxième année, ENSAE, IGN/COGIT, 1996.

[Sau80] L. Saulis. Large deviations for sums of independent weighted random variables. *Lith. Math. J.*, **19**, 277–287, 1980.

[SC97] Stephen V. Stehman et Raymond L. Czaplewski. Basic structures of a statistically rigorous thematic accuracy assessment. *ACSM/ASPRS Annual Convention & Exposition*, **3**, 543–553, 1997.

[SS76] L. Saulis et V. Statulevicius. Large deviations in weighted sums of random variables. *Lith. Math. J.*, **16**, 243–250, 1976.

[SS91] Saulis et Statulevicius. *Limit theorems for large deviations*. Kluwer, 1991.

[Ste78] Josef Steinebach. Convergence rates of large deviation probabilities in the multidimensional case. *The Annals of Probability*, **6**, n$^{\text{o}}$ 5, 751–759, 1978.

[Var84] S. R. S. Varadhan. *Large deviations and applications*. CBMS-NSF, 1984.

[Vau97] François Vauglin. *Modèles statistiques des imprécisions géométriques des objets géographiques linéaires*. Thèse de doctorat de l'université de Marne-la-Vallée, 1997.

[Wil89] G. E. Willmot. The total claims distribution under inflationary conditions. *Scand. Actuarial J.*, 1–12, 1989.

[Wol80] W. Wolf. Some remarks on large deviations for weighted sums if cramér's condition is not satisfied. *Mathematical Statistics, Banach Center Publications*, **6**, 347–352, 1980.

[You96] Khalid Yousfi. *Mesure et représentation de la qualité des Bases de Données géographiques*. Mémoire de troisième cycle, Institut Agronomique et vétérinaire Hassan II, Rabat, 1996.