Développements de grandes déviations pour des sommes pondérées appliqués à un problème géographique

Olivier BONIN

Institut Géographique National, Laboratoire COGIT

1. Introduction

Dans le cadre d'une application de calcul de temps de parcours entre deux localités utilisant une base de données géographiques, nous présentons le problème de l'évaluation de l'impact des erreurs présentes dans la base de données sur les temps de parcours calculés, l'algorithme de calcul de temps de parcours étant supposé exact.

Dans ce but, nous proposons des modèles simples d'erreurs dans les bases de don-nées géographiques, et reformulons le problème comme l'évaluation de la probabilité que l'erreur sur le temps de parcours calculé dépasse un certain seuil (c'est-à-dire l'évaluation des erreurs importantes pour un utilisateur, et non de l'erreur moyenne).

Nous montrons que ce problème correspond à un problème de grandes déviations pour des sommes pondérées de variables aléatoires, et pour des loi composées pondérées. Nous présentons les résultats de grandes déviations précis que nous avons établis pour résoudre ce problème, et montrons qu'ils sont nécessaires pour obtenir numériquement une bonne estimation des probabilités d'intérêt (par rapport au développement du logarithme de ces probabilités). Nous comparons également ces résultats à ceux obtenus par des simulations par la méthode de Monte-Carlo sur des données
réelles.

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

Nous proposons ici un modèle de calcul d'itinéraires en zone urbaine, et un modèle de calcul sur un itinéraire type à travers la France.

2.1 Trajet en zone urbaine

Dans le cas d'une zone urbaine, 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.

Nous faisons l'hypothèse supplémentaire 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.

Le modèle pour un tel trajet 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

2.2 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. Le temps de parcours d'un tel itinéraire s'écrit alors :

$$T = \sum_{i=1}^{N} l_i / V_i$$

avec l_i longueur du tronçon i et V_i sa vitesse de parcours.

3. Critère de qualité des résultats de l'application

3.1 Cas d'un trajet de longueur fixe

Nous disposons 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.

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. En notant avec un indice R les grandeurs calculées dans la référence, et un indice D celles calculées dans 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

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

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

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.

3.2 Cas d'un trajet de longueur aléatoire

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 ième tronçon, cette probabilité s'écrit avec les longueurs l_i et les vitesses V_i :

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

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 d'un modèle d'erreurs d'attributs, la vitesse de chaque tronçon étant calculée à l'aide des valeurs des attributs. Nous pouvons estimer ces

probabilités par des développements de grandes déviations pour des lois composées de variables discrètes pondérées.

Le modèle d'erreurs d'attributs fait l'objet de la section suivante.

4. Modèles d'erreurs d'attributs

Les objets géographiques de la base de données ont leurs caractéristiques décrites par des attributs. On recense tout d'abord les erreurs présentes, pour pouvoir proposer un modèle paramétrique d'erreur.

On s'intéresse à un attribut énuméré a ayant K modalités. On considère N objets correctement appariés entre la référence et le jeu de données, c'est-à-dire des objets décrivant la même réalité. On note alors :

- 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.

Un objet géographique aura pour l'attribut a une valeur $r \in \{1,\dots,K\}$ dans la référence et une valeur $d \in \{1,\dots,K\}$ 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).

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 $(p_{rd})_{0\le r,d\le K}$ avec :

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

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, en remarquant que :

$$P\big((R,D)=(r,d)\big) = P(R=r)\,P(D=d \mid R=r) \qquad \forall (r,d) \in \{0,\dots,K\}^2 \qquad (1)$$

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. La quantité $P(D=d \mid \mathrm{R=r})$ 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.

4.1 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 :

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

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 in-déterminée peut avoir par erreur une valeur, avec équiprobabilité entre les différentes valeurs.

Ce modèle impose de définir $2(K+1)$ coefficients. Un choix naturel est de poser :

$$\begin{cases} p_{rr}=(1-\theta_r)\frac{N_r}{N} \\ p_r=\frac{\theta_r}{K}\frac{N_r}{N} \end{cases},\quad \forall r\in\{0,\dots,K\} \tag{3}$$

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.

4.2 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 voie*s, qui peut prendre les valeurs *Inconn*u, *1 voi*e, *2 voie*s, *3 voie*s, *4 voie*s, *2 voies large*s, *Plus de 4 voie*s. 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 *tridiagonale*s. 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 :

$$p=\begin{pmatrix} p_{00} & p_{01} & 0 & \dots & \dots \\ p_{10} & p_{11} & p_{12} & 0 & \dots \\ 0 & p_{21} & p_{22} & \ddots & \ddots \\ \vdots & 0 & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \tag{4}$$

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.

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 :

$$\begin{cases} p_{rr} = (1-\theta_r)\frac{N_r}{N} & \forall r \in \{0,\ldots,K\} \\ p_{r(r-1)} = p_{r(r+1)} = \frac{\theta_r}{2}\frac{N_r}{N} & \forall r \in \{1,\ldots,K-1\} \\ p_{01} = \theta_0 \frac{N_0}{N} & \\ p_{(K-1)K} = \theta_K \frac{N_K}{N} & \end{cases} \tag{5}$$

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.

4.3 Estimation des modèles pour 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 :

$$\hat{\theta} = \frac{N_{01} + \sum_{k=1}^{K-1}(N_{k(k-1)} + N_{k(k+1)}) + N_{K(K-1)}}{\sum_{k=0}^{K} N_{kk} + N_{01} + \sum_{k=1}^{K-1}(N_{k(k-1)} + N_{k(k+1)} + N_{K(K-1)}} \tag{6}$$

On obtient pour le cas uniforme un résultat similaire :

$$\hat{\theta} = \frac{N - \sum_{k=0}^{K} N_{kk}}{N} \tag{7}$$

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

Soit $X = X_1, X_2, \ldots$ une suite de variables i.i.d non dégénérées centrées définies sur une grille, soit $\{a_{nk} : 1 \le k \le n, 1 \le n < \infty\}$ un tableau triangulaire de réels positifs vérifiant $\sum_{k=1}^{n} a_{nk}^2 = 1$, et soit c un réel strictement positif. Nous étudions ici la probabilité $P(S_n \ge cA_n$, avec $S_n = \sum_{k=1}^{n} a_{nk} X_k$ et $A_n = \sum_{k=1}^{n} a_{nk}$, en supposant 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 $E(X^2) = 1$ et 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 ([5]) :

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

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

Posons $Y_{nk}=a_{nk}X_k-ca_{nk}$, $H_{nk}(y)=F(ya_{nk}^{-1}+c)$, $\phi_{nk}(h)=e^{-hca_{nk}}\phi(ha_{nk})$.

Soit _ $\overline{H}_{nk}$ définie par $d\overline{H}_{nk}(y)=\frac{e^{hy}}{\phi_{nk}(h)}dH_{nk}(y)$ pour tout $0<h<B\sigma_n^{-1}$ et $\overline{Y}_{n1},Y_{n2},\ldots$ une suite de variables aléatoires indépendantes distribuées selon $\overline{H}_{nk}$. Posons $\overline{S}_n=\sum_{k=1}^n\overline{Y}_{nk}$, et $\overline{H}_n(y)=P(\overline{S}_n\le y)$. Avec ces notations, nous obtenons par une extension du cas i.i.d (voir par exemple [1] ou [6]) :

Lemme 1.

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

avec $I_n(h)=h\int_0^\infty e^{-hy}\left[\overline{H}_n^*(y)-\overline{H}_n(0)\right]dy$, $\overline{H}_n(y)$ *et* $\overline{H}_n^*=\frac{1}{2}\left[\overline{H}_n(y)+\overline{H}_n(y-)\right]$ *si* y *est sur un noeud de la grille, et* $\overline{H}_n^*=\overline{H}_n(y)$ *sinon.*

Observons que les conditions I et II ont été introduites pour assurer l'existence d'une suite de réels $\{h_n: 1\le n<\infty\}$ tels que $E(\overline{S}_n(h_n))=0$ et que $\mathrm{Var}(\overline{S}_n(h_n))$ soit uniformément bornée pour tout n. Nous approchons alors $\overline{H}_n^*$, et donc I_n, à l'aide d'un théorème central limite local pour tableau triangulaire de variables aléatoires discrètes réelles :

Théorème 1. *Pour tout* n *soit* $X_{n1},\ldots,X_{nn}$ n *variables aléatoires indépendantes centrées, de fonctions de répartition* F_{nk}. *Soit* $S_n=\sum_{k=1}^n X_{nk}$ *et* $E(S_n^2)=s_n^2$. *Soit* F_n *la fonction de répartition de* S_n/s_n, $\mathcal{N}$ *la fonction de répartition d'une gaussienne centrée réduite et* n *sa densité. Supposons de plus que* S_n *est définie sur une grille et que :*

(i) $cn<s_n^2<Cn$, *avec* $c>0$ *et* $C>0$ *;*

(ii) $E(X_{nk}^4)$ *est uniformément bornée pour tout* n *et pour tout* k *;*

alors

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

avec $\mu_3^{(n)}=\sum_{k=1}^n E(X_{nk}^3)$ *et* $r_n(x)\to 0$ *uniformément en* x *quand* $n\to\infty$, $F_n^{\#}$ *étant la convolution de* F_n *par la distribution triangulaire sur* $\left[-d_n/2,d_n/2\right]$, *avec* d_n *le pas de grille de* S_n.

Remarquons que $\overline{H}_n^{\#}$ et $\overline{H}_n^*$ coïncident aux points $k+\frac{1}{2}d_n$, $\forall k\in\mathbb{Z}$. Nous obtenons ainsi :

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

$$P(S_n\ge cA_n)=\frac{1}{\sqrt{2\pi}}\frac{d_n\,e^{-h_nd_n}}{\overline{\sigma}_n(1-e^{-h_nd_n})}e^{-h_ncA_n}\left[\prod_{k=1}^n\phi(h_na_{nk})\right](1+o(1))$$

La preuve de ce théorème est donnée dans [3].

6. Grandes déviations pour lois composées

Le problème s'écrit à l'aide d'une loi composée de la forme

$$Y = \sum_{i=1}^{N} a_{N_i} X_i \tag{8}$$

avec les X_i des variables aléatoires i.i.d de loi X, et les a_{N_i} des variables aléatoires, *a priori* indépendantes de X_i, pour laquelle il faut établir un développement de grandes déviations. On est dans le cas de la détermination d'une probabilité du type

$$P(Y > y),$$

y étant un seuil donné. 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 $K \in \{0,1,\ldots\}$ $P(N = k) = 0 \quad \forall k > K$ 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.

6.1 Asymptotique $E(N) \to \infty$

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 $F(x) = P(X \le x)$ 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 $0<\alpha\le 1$, $0<\theta\le 1$, tels que pour tout n, au moins αn des a_k, $1\le k\le n$, 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 $I\supseteq(-B,B)$, pour un $B>0$, la fonction $Q=\phi'/\phi$ prend la valeur $\frac{c}{\alpha\theta}$ en un point, et $B_0=\theta^{-1}Q^{-1}\left(\frac{c}{\alpha\theta}\right)\in I$.

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

$$P(S_N>cA_N)=P(S_N>cA_N\mid N>0)\,P(N>0$$
$$=P\left(\sum_{i=1}^{\tilde N}a_iX_i>c\sum_{i=1}^{\tilde N}a_i\right)(1-p_0),$$

avec $$P(\tilde N=k)=(1-p_0)^{-1}p_k$$

pour $k=1,2,\ldots$. Dans ce cas, on pose $\tilde Y=\sum_{i=1}^{\tilde N}a_iX_i$ et $E(e^{t\tilde Y})=(1-p_0)^{-1}(\phi_Y(t)-p_0)$. Dans la suite de la démonstration, le lecteur remplacera Y par $\tilde Y$ 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 :

$$\phi_Y(t)=E(e^{t(S_N-A_N)})=E(E(e^{t(S_N-A_N)}\mid N))=\sum_{k=0}^{\infty}p_k\prod_{i=1}^{k}e^{-ca_it}\phi(a_it)\qquad(9)$$

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 3. *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.

6.2 Asymptotique $y \to \infty$

Nous présentons dans cette section un théorème similaire à celui présenté précédemment, mais pour l'asymptotique $y \to \infty$.

Soit $X = X_1, X_2, \ldots$ une suite de variables aléatoires i.i.d non dégénérées, soit $\{a_k : k = 1, 2, \ldots\}$ 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 $F(x) = P(X \le x)$ 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 $0 < \alpha \le 1$, $0 < \theta \le 1$, tels que pour tout n, au moins αn des a_k, $1 \le k \le n$, 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 $P(Y > 0 = P(Y > y \mid N > 0)\, P(N > 0)$

$$= P\left(\sum_{i=1}^{\tilde{N}} a_i X_i > y\right)(1 - p_0)$$

avec

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

pour $k = 1, 2, \ldots$. Dans ce cas, on pose $\tilde{Y} = \sum_{i=1}^{\tilde{N}} a_i X_i$ et $E(e^{t\tilde{Y}}) = (1 - p_0)^{-1}(\phi_Y(t) - p_0)$. Dans la suite de la démonstration, le lecteur remplacera Y par $\tilde{Y}$ 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} :

$$\phi_Y(t) = E(e^{t(Y)}) = E(E(e^{t(Y)} \mid N)) = \sum_{k=0}^{\infty} p_k \prod_{i=1}^{k} \phi(a_i t) \qquad (10)$$

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 écrirons Y_h sous la forme $Y_h = \sum_{i=1}^{N_h} X_{hi} - y$ et notons ω_{hj} la fonction caractéristique de X_{hj}.

Nous pouvons énoncer nos théorèmes :

Théorème 4. *Supposons que les conditions 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* $|\omega_{hj}(\zeta)| < 1 - \theta(\delta, a)$ $\forall j \; \forall k$ *pour* $as > \zeta > \delta > 0$, *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 .

Les preuves de ces théorèmes sont données dans [4].

7. Mise en oeuvre des modèles

7.1 Itinéraire de longueur fixe

Nous reprenons ici l'application de [2]. Nous avons appliqué notre modélisation et notre théorème à Géoroute, une base de données de l'IGN. Nous avons testé l'égalité des θ_r dans le modèle uniforme pour la zone considérée, et obtenu des valeurs comprises entre 3 et 6 % selon la zone géographique considérée.

Les tronçons de route dans les zones d'étude sont de longueur de l'ordre de 50 mètres, et nous avons posé comme vitesses dans le modèle v_1 = 50 km/h et v_2 = 20 km/h, en fonction de la nature du tronçon de route renseignée par ses attributs. La répartition entre tronçons rapides et tronçons lents a été estimée à $N_1 / N = 0.3$ et $N_2 / N = 0.7$. Ainsi, la loi des variables X_1 et Y_1 est entièrement déterminée :

$$\begin{cases} P(X_1 = -\eta / v_1) = P(Y_1 = -\eta / v_1) = (1-\theta) N_1 / N \\ P(X_1 = -\eta / v_2) = P(Y_1 = -\eta / v_2) = (1-\theta) N_2 / N \\ P(X_1 = (1-\eta)/v_1 - 1/v_2) = P(Y_1 = (-1-\eta)/v_1 + 1/v_2) = \theta N_{1/N} \\ P(X_1 = (1-\eta)/v_2 - 1/v_1) = P(Y_1 = (-1-\eta)/v_2 + 1/v_1) = \theta N_{2/N} \end{cases}$$

Il faut noter que $E(X_1) < 0$ et $E(Y_1) < 0$.

Nous utilisons le théorème de grandes déviations pour des sommes pondérées en réécrivant les longueurs $l_{ni} = l * a_{ni}$ avec $\sum_{i=1}^{n} a_{ni}^2 = 1$, $A_n = \sum_{i=1}^{n} a_{ni}$, et en posant $X = X_1$ et $Y = Y_1$:

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

$$\frac{1}{\sqrt{2\pi}} \frac{d_n^{(X)} e^{-h_n^{(X)} d_n^{(X)}}}{\overline{\sigma}_n^{(X)} (1 - e^{-h_n^{(X)} d_n^{(X)}})} e^{h_n^{(X)}} E(X) A_n \left[\prod_{i=1}^{n} \phi^{(X)}(h_n^{(X)} a_{ni})\right](1 + o(1))$$

$$+ \frac{1}{\sqrt{2\pi}} \frac{d_n^{(Y)} e^{-h_n^{(Y)} d_n^{(Y)}}}{\overline{\sigma}_n^{(Y)} (1 - e^{-h_n^{(Y)} d_n^{(Y)}})} e^{h_n^{(Y)}} E(Y) A_n \left[\prod_{i=1}^{n} \phi^{(Y)}(h_n^{(Y)} a_{ni})\right](1 + o(1)) . \qquad (11)$$

La fonction génératrice des moments $\phi^{(X)}$ of X est donnée par la formule :

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

On obtient une expression similaire pour $\theta^{(Y)}$, et tous les paramètres sont calculés numériquement.

La figure 1 donne les résultats de ce modèle appliqué à un itinéraire comportant 30 tronçons. Dans le graphe du haut, θ (taux d'erreur dans les données) varie de 3% à 7%, avec une seuil d'erreur admissible η = 5%, et dans le graphe du bas η varie de 4% à 10% avec un taux d'erreur dans la base de données θ = 5%. Ces mêmes probabilités ont également été estimées par simulation de Monte-Carlo avec 10000 tirages.

L'estimation logarithmique donne la probabilité :

$$\log P\left(\sum_{i=1}^{n} l_{ni} (X_i - E(X_i)) > -E(X) \sum_{i=1}^{n} l_{ni}\right) \rightarrow h_n^{(X)} E(X) A_n + \sum_{i=1}^{n} \log \phi^{(X)}(h_n^{(X)} a_{ni})$$

qui, bien que très simple, ne convient manifestement pas pour notre application.

Nous avons testé la qualité de notre approximation avec une simulation de Monte-Carlo sur des données réelles. Nous avons calculé avec un algorithme de plus court chemin le temps de parcours de 20 itinéraires avec la base de données, et avec des bases de données bruitées. Pour chaque niveau de bruit θ, nous avons généré 1000 bases de données bruitées indépendantes, calculé le temps de parcours des 20 itinéraires d'intérêt, et estimé ainsi empiriquement la queue de distribution de $(T_R - T_D)/T_R$ pour chacun des 20 itinéraires. Comme les itinéraires comportent entre 30 et 50 tronçons, nous donnons pour chaque niveau d'erreur le résultat de notre approximation de grandes déviations avec n = 30 et tous les tronçons de même longueur. Nous présentons les résultats de cette étude en figure 2. Les erreurs réelles de temps de trajet estimées par Monte-Carlo sont majorées très correctement par notre méthode.

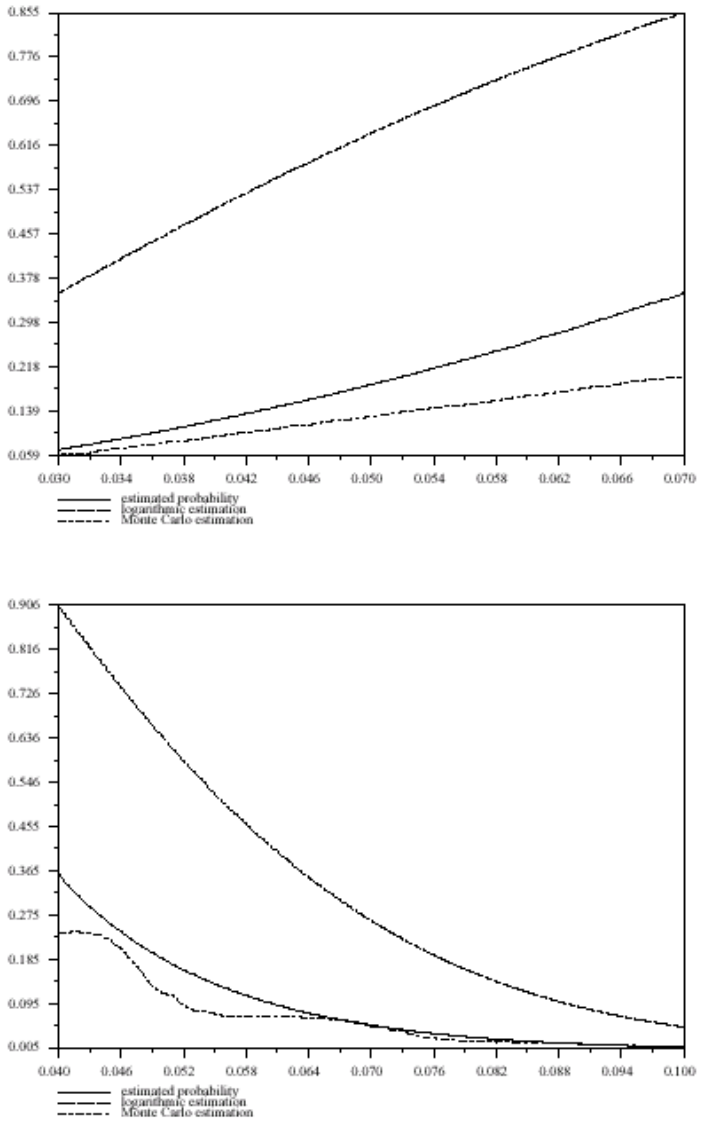


Figure 1 - Probabilité que l'erreur dépasse un seuil η en fonction de θ avec η = 5% (haut) et en fonction de η avec θ = 5% (bas)

7.2 Itinéraire de longueur aléatoire

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. On considère donc que $E(N)\to\infty$, et on peut appliquer directement les résultats de la section précédente.

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.

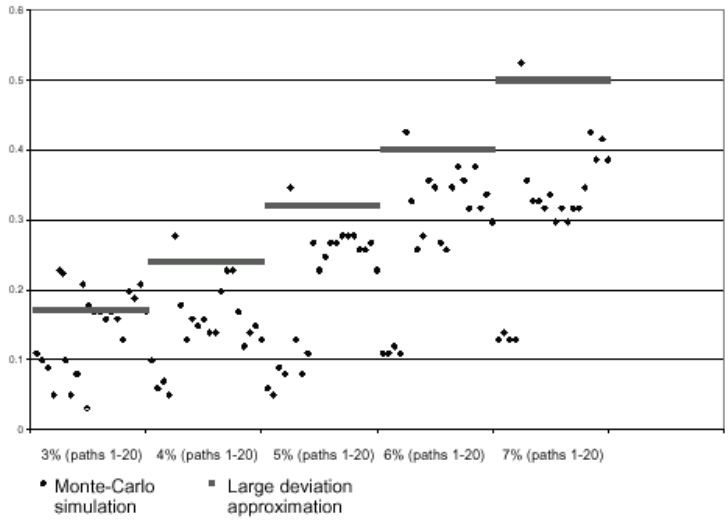


Figure 2 - Probabilité que l'erreur dépasse un seuil η = 5% en fonction de θ, avec θ = 3%, 4%, 5%, 6% et 7% pour 20 itinéraires simulés

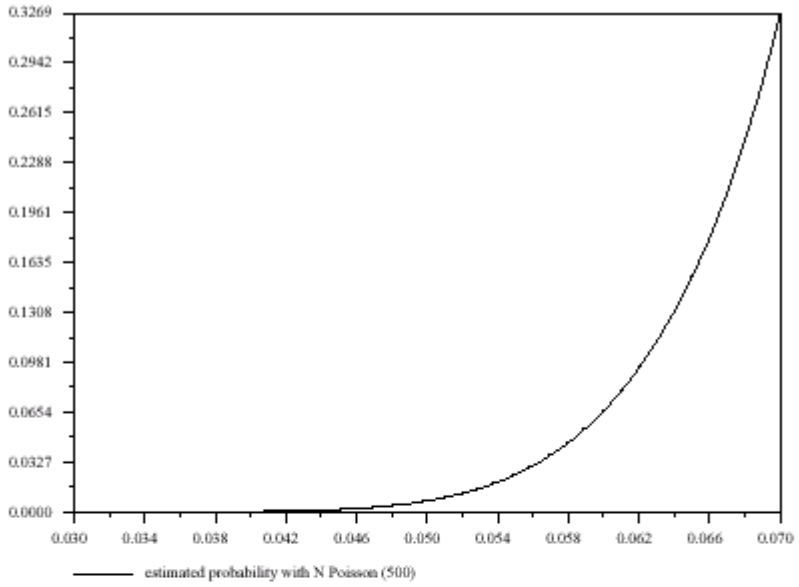
Ce paramètre θ est de l'ordre de 5% sur notre jeu test. Nous avons ensuite choisi le paramètre $E(N)$ = 500, ce qui représente un parcours de 250 km environ, et appliqué le développement de grandes déviations. La vitesse de parcours de chaque tronçon est de 60 km/h sur nationale ou de 120 km/h sur autoroute. 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 3.

Notons toutefois qu'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 $a_{k1},\ldots,a_{kk}$. 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 $l_1,\ldots,l_r$. 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 $p_j = P(a = l_j)$, $1 \le j \le r$, et la loi composée de (8) 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 $N_k, 1 \le k \le r$ suivant des lois de Poisson de paramètres $\lambda_k = \lambda p_k$ avec $\sum_{k=1}^{r} \lambda_k = \lambda$.



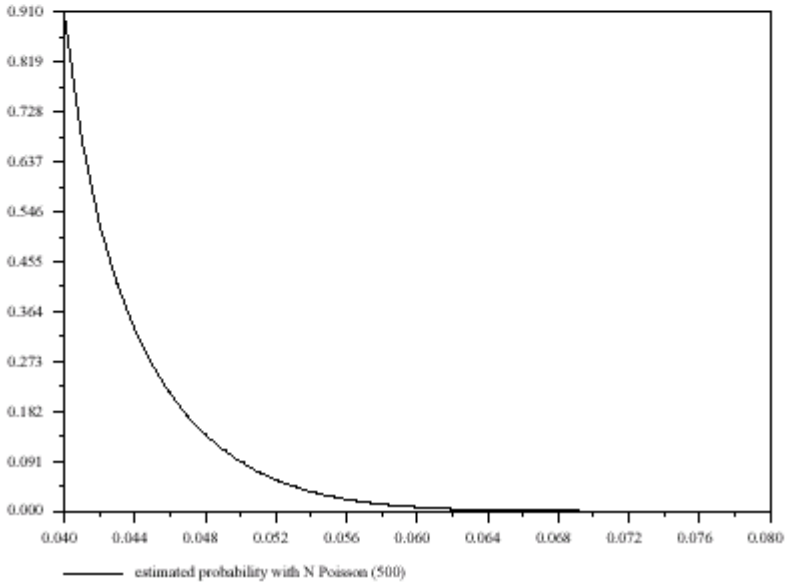


Figure 3 - Probabilité de dépassement d'un seuil d'erreur relative η en fonction de θ avec η = 6% (haut) et en fonction de η avec θ = 5% (bas)

La transformée de Laplace de Y s'écrit alors :

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

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 [7]. Ainsi nous pouvons écrire, quand $y \to \infty$:

$$P\left(\sum_{k=1}^{r}\sum_{i=1}^{N_k} X_{ki} \geq y\right) = \frac{\phi_c(h)\, e^{-hy}}{\sqrt{2\pi}\,\sigma_c(1-e^{-h})}(1+o(1)), \tag{12}$$

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 [7].

8. Approximations de queues de distribution

L'approximation de queue de distribution par la méthode des grandes déviations, dont nous donnons une application géographique dans cet article, n'est bien sûr pas la seule approche possible. Cependant, nous voulons donner pour conclure sur sa qualité un exemple très simple qui illustre les avantages et désavantages de diverses approximations de lois.

Considérons le cas d'une variable aléatoire N de loi binômiale de paramètres n = 10000 et p = 0;0001. On a alors bien sûr $E(N)=1$. Cette variable N peut être approchée de différentes manières : approximations Poisson et Normale, développement de grades déviations (GD), simulation de Monte-Carlo (MC), méthode d'Abramowitz et Segun (AS).

La figure 4 donne les valeurs de $P(N > na)$, calculée par les différentes approximations retenues, avec Le paramètre a a en abscisse. L'approximation de grandes déviations donne nettement les meilleurs résultats (par rapport à Monte-Carlo qui sert de référence), dès que l'on s'éloigne de l'espérance.

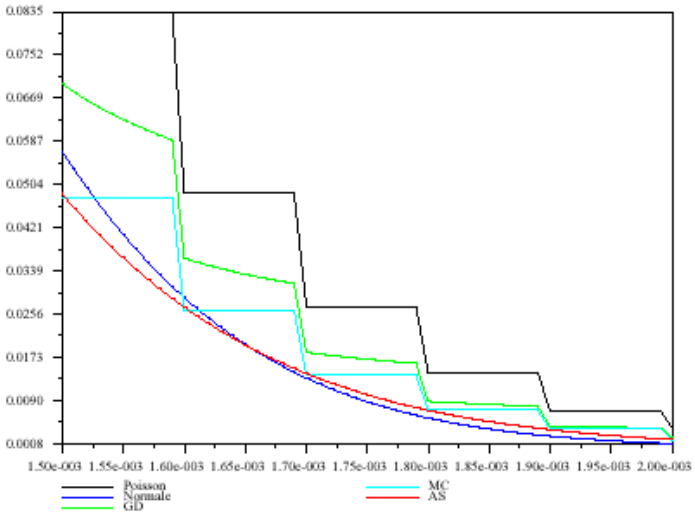


Figure 4 - Approximations de $P(N > na)$, avec N loi binômiale et a variant de 0,00015 à 0,0002

Références

[1] R. R. Bahadur and R. Ranga Rao. On deviations of the sample mean. *Ann. Math. Statist*., **3**1, 1015–1027, 1960.

[2] O. Bonin. Large deviation theorems for weighted sums applied to a geographical problem. *J. Appl. Probab*., **3**9, No. 2, 251–260, 2002.

[3] O. Bonin. *Modèles d'erreurs dans une base de données géographique et grandes déviations pour des sommes pondérées ; application à l'estimation d'erreurs sur un temps de parcour*s. Thèse de doctorat de l'Université Paris VI, 2002.

[4] O. Bonin. Large deviation theorems for weighted comound poisson sums. *Proba-bility and Mathematical Statistic*s, **2**3, No. 2, 357–368, 2003.

[5] Stephen A. Book. A large deviation theorem for weighted sums. *Z. Wahrscheinli-chkeitstheorie verw. Geb*., **2**6, 43–49, 1973.

[6] William Feller. *An Introduction to Probability Theory and Its Applications (vo-lume II*). John Wiley & Sons, 1970.

[7] Jens Ledet Jensen. *Saddlepoint approximation*s. Clarendon Press, Oxford, 1995.