Document généré le 19 oct. 2024 10:25

L'Actualité économique

Modèle bayésien généralisé pour l'identification des sites routiers dangereux

Denis Bolduc et Sylvie Bonin

Volume 73, numéro 1-2-3, mars–juin–septembre 1997

L'économétrie appliquée

URI : https://id.erudit.org/iderudit/602223ar
DOI : https://doi.org/10.7202/602223ar

Aller au sommaire du numéro

Éditeur(s)

HEC Montréal

ISSN

0001-771X (imprimé)
1710-3991 (numérique)

Découvrir la revue

Citer cet article

Bolduc, D. & Bonin, S. (1997). Modèle bayésien généralisé pour l'identification des sites routiers dangereux. *L'Actualité économique*, *73*(1-2-3), 81–98. https://doi.org/10.7202/602223ar

Résumé de l'article

Dans le présent article, nous décrivons une méthodologie générale à information complète pour analyser la dangerosité des sites routiers. La technique proposée, de type bayésienne, permet de traiter simultanément les problèmes d'hétérogénéité déterministe et aléatoire ainsi que celui de la corrélation spatiale attribuable à la proximité ou l'environnement similaire caractérisant les sites à l'étude. Notre cadre méthodologique englobe des approches bayésiennes de pratique courante qui mettent l'accent sur l'analyse des fréquences d'accidents et d'autres du même type qui étudient les proportions d'accidents impliquant une caractéristique donnée. Les propriétés et l'intérêt de la nouvelle méthode sont démontrés à l'aide d'un exemple concret basé sur des données de la région de Québec.

Tous droits réservés © HEC Montréal, 1997

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

L'Actualité économique, Revue d'analyse économique, vol. 73, nos 1-2-3, mars, juin, septembre 1997

MODÈLE BAYÉSIEN GÉNÉRALISÉ POUR L'IDENTIFICATION DES SITES ROUTIERS DANGEREUX*

Denis BOLDUC
GREEN
Département d'économique
Université Laval
Sylvie BONIN
Département d'aménagement
Université Laval

RÉSUMÉ – Dans le présent article, nous décrivons une méthodologie générale à information complète pour analyser la dangerosité des sites routiers. La technique proposée, de type bayésienne, permet de traiter simultanément les problèmes d'hétérogénéité déterministe et aléatoire ainsi que celui de la corrélation spatiale attribuable à la proximité ou l'environnement similaire caractérisant les sites à l'étude. Notre cadre méthodologique englobe des approches bayésiennes de pratique courante qui mettent l'accent sur l'analyse des fréquences d'accidents et d'autres du même type qui étudient les proportions d'accidents impliquant une caractéristique donnée. Les propriétés et l'intérêt de la nouvelle méthode sont démontrés à l'aide d'un exemple concret basé sur des données de la région de Québec.

ABSTRACT – In this paper, we describe a general full information Bayesian methodology to analyze road accident sites. The technique allows for the presence of deterministic and random heterogeneity together with spatial autocorrelation among neighboring sites. The suggested framework contains as subcases the Bayesian approaches currently used to study accidents frequencies and those intended for the analysis of accidents proportions of accidents with a given characteristic. To demonstrate the feasibility and the usefulness of the suggested approach, we apply it on accidents data taken from the Québec city database.

* Cette recherche a été réalisée grâce au Programme d'Action concertée de soutien à la recherche en sécurité routière financé conjointement par le ministère des Transports du Québec, la Société de l'assurance automobile du Québec et le Fonds FCAR. Nous aimerions remercier M. Ben Heydecker pour ses commentaires et suggestions judicieuses. L'approche de base considérée dans le texte lui est attribuée.Nous remercions finalement un évaluateur anonyme pour ses remarques pertinentes.

INTRODUCTION

Une composante importante dans la recherche en sécurité routière concerne le développement d'outils permettant d'identifier les sites routiers à plus hauts risques. Dans un contexte d'optimisation sous contraintes budgétaires, l'organisme en charge du système routier veut idéalement concentrer ses interventions visant à améliorer la sécurité, pour les sites considérés les plus dangereux (ou coûteux pour la société). La façon la plus immédiate, bien qu'incorrecte, de retenir ces sites consiste à se limiter à ceux qui ont connu le plus grand nombre d'accidents au cours de l'année précédente. Le problème lié à cette démarche vient du fait que même si on fait aucune intervention sur un site où il s'est produit un très grand nombre d'accidents, le nombre total d'accidents observé cette année pour ce même site, va naturellement décroître vers la moyenne d'accidents du site. C'est le phénomène de régression vers la moyenne. L'approche empirique de Bayes (AEB) suggérée initialement par Hauer (1986) tire sa grande popularité du fait qu'elle permet l'identification des sites dangereux en corrigeant explicitement le phénomène de régression vers la moyenne. Cette technique permet de plus d'introduire dans l'étude d'un site donné, l'information concernant d'autres sites à caractéristiques similaires.

Pour mettre en oeuvre l'AEB, l'analyste doit exercer un bon jugement dans la création des groupes de sites dits comparables ou homogènes. Ce problème a été récemment discuté dans Hauer (1992). D'un côté, si on utilise une définition trop serrée du concept de similitude lors de la sélection des sites à étudier, la population de référence devient alors trop petite pour obtenir des estimations fiables. D'autre part, chaque site a des caractéristiques qui lui sont propres et de ne pas en tenir compte lors de l'analyse peut mener à des conclusions erronées sur le risque relatif des sites. La solution à ce dilemme telle que prônée par cet auteur consiste à définir des groupes de référence plus grands en s'assurant toutefois de contrôler pour la variabilité inter-sites à l'aide d'une régression multivariée exprimée en termes des caractéristiques spécifiques des sites. Cette procédure permet ainsi de conserver un nombre de degrés de liberté suffisants afin de procéder à une analyse statistique fiable des sites. Techniquement, cette approche est dite à hétérogénéité déterministe, car elle explique les différences entre sites à l'aide d'une relation fonctionnelle bien déterminée.

Les différentes variables importantes permettant le traitement du problème d'hétérogénéité ne sont pas toujours accessibles à l'analyste. Il y a donc un risque qu'une partie de cette variabilité inter-sites soit non expliquée par la relation déterministe. Cette composante résiduelle caractérise l'hétérogénéité aléatoire. Cette dernière est généralement modélisée via le terme d'erreur. Dans le présent texte, nous décrivons une méthodologie à information complète de type AEB qui traite simultanément les problèmes d'hétérogénéité déterministe et aléatoire. Notre approche qui constitue une alternative à celle de Hauer (1992) englobe les travaux de Higle et Witkowsky (1989) qui mettent l'accent sur l'analyse des fréquences d'accidents et de Heydecker et Wu (1991) qui étudient

les proportions d'accidents impliquant une caractéristique donnée. Une autre particularité du modèle proposé est qu'il permet de prendre en considération les liens entre les sites spatialement corrélés dû à leur proximité ou à un environnement similaire. Maher (1990), de même que Loveday et Jarrett (1992) démontrent que les sites ne devraient pas être modélisés sous l'hypothèse qu'ils sont indépendants entre eux. La première section décrit l'approche proposée pour l'analyse des fréquences d'accidents, alors que la deuxième section est consacrée au cas des proportions d'accidents. La dernière partie du document présente les résultats d'applications démontrant l'utilité de prendre en compte à la fois les problèmes d'hétérogénéité et de corrélation spatiale entre sites routiers.

1. ANALYSE DES FRÉQUENCES D'ACCIDENTS

Dans un premier temps, cette section présente l'approche bayésienne de base à information complète permettant l'analyse des fréquences d'accidents. Nous procédons ensuite aux extensions du modèle visant à prendre en compte l'hétérogénéité déterministe, l'hétérogénéité aléatoire et la corrélation spatiale entre sites d'accidents.

1.1 *L'approche de base*

Soit $i = 1, ..., N$ l'indicateur du site étudié et soit N la taille de la population de référence. Le modèle bayésien basé sur les fréquences d'accidents suppose que le nombre d'accidents n_i enregistrés au site i au cours d'une période de temps donnée[1] t_i suit une distribution de Poisson de moyenne μ . On écrit :

$$f(n_i|t_i,\mu) = \frac{(\mu t_i)^{n_i}\, exp(-\mu t_i)}{n_i!}. \tag{1}$$

Pour sa part le taux moyen d'accidents μ est interprété comme une variable aléatoire qui suit par hypothèse une loi de distribution gamma de densité :

$$g_b(\mu|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\mu^{\alpha-1}\, exp(-\beta\mu), \tag{2}$$

ce que l'on dénote également $P_b(\mu \mid \alpha,\beta) \sim$ gamma(α,β), où l'indice b (pour *before*) fait ressortir le contexte *a priori*. Les équations (1) et (2) s'expriment toutes deux en termes d'un taux moyen inconnu μ. Ces deux expressions peuvent être combinées, de façon à obtenir une fonction de distribution marginale de n_i exprimée uniquement en termes des paramètres α et β de la distribution gamma. L'évaluation de l'intégrale suivante :

$$\int f(n_i \mid t_i,\mu) g_b(\mu \mid \alpha,\beta) d\mu, \tag{3}$$

1. Il est possible de regrouper à des fins d'analyse des observations correspondant à des périodes de temps différentes.

conduit à la densité marginale :

$$f(n_i|t_i, \alpha, \beta) = \frac{\Gamma(\alpha+n_i)}{n_i!\Gamma(\alpha)} \frac{t_i^{n_i}\beta^{\alpha}}{(\beta+t_i)^{\alpha+n_i}}, \tag{4}$$

ce qui correspond à la fonction de densité d'une distribution binomiale négative. L'approche bayésienne empirique suggère de donner à α et β les valeurs qui maximisent la probabilité conjointe des observations $(n_1, ..., n_N)$, concernant les N sites étudiés. Dans un second temps, l'application du théorème de Bayes, qui dans notre notation s'écrit :

$$g_a(\mu|n_i, t_i, \alpha, \beta) = \frac{f(n_i|t_i, \mu)\ g_b(\mu|\alpha, \beta)}{f(n_i|t_i, \alpha, \beta)}, \tag{5}$$

permet d'établir la distribution *a posteriori* de μ laquelle combine de façon optimale l'information *a priori* et celle apportée par l'échantillon en main. On peut ainsi déduire que la loi *a posteriori* de μ correspond alors à une distribution gamma révisée :

$$P_a(\mu \mid n_i, t_i, \alpha, \beta) \sim \text{gamma}(\alpha + n_i, \beta + t_i),$$

selon laquelle la densité s'écrit :

$$g_a(\mu|\alpha+n_i, \beta+t_i) = \frac{(\beta+t_i)^{\alpha+n_i}}{\Gamma(\alpha+n_i)} \mu^{\alpha+n_i-1}\, exp(-(\beta+t_i)\mu). \tag{6}$$

La forme de la loi *a posteriori* de μ dépend de façon critique des hypothèses faites concernant les distributions du nombre d'accidents n_i et du taux moyen d'accidents μ. Les distributions de Poisson et de gamma sont généralement utilisées conjointement, car elles se combinent de manière pratique et élégante. L'identification des sites routiers dangereux par l'approche AEB requiert l'évaluation de la densité *a posteriori* $g_a(\mu \mid \alpha + n_i, \beta + t_i)$, où l'indice *a* (pour *after*) fait ressortir le caractère *a posteriori* de la distribution.

Pour procéder à l'identification des sites, les valeurs des paramètres α et β doivent d'abord être déterminées. Les approches bayésiennes présentées dans Higle et Witkowsky (1989) et dans Heydecker et Wu (1991) utilisent les valeurs de α et β qui maximisent la fonction de vraisemblance des observations $(n_1, ..., n_N)$. La fonction de densité définie en (4) et qui correspond à la fonction au dénominateur de la relation de Bayes à l'équation (5) est employée pour effectuer la maximisation. Sur la base de l'efficacité statistique, la méthode du maximum de vraisemblance est également utilisée dans le présent document comme méthode permettant d'attribuer à α et à β des valeurs sensées.

1.1.1 *Analyse bayésienne*

Tel que mentionné précédemment, la distribution *a posteriori* décrit notre niveau de connaissance du taux moyen d'accidents μ , après que les observations $(n_1, ..., n_N)$ aient été combinées à l'information *a priori*. L'estimateur bayésien du nombre d'accidents enregistrés à un site *i* est donné par la moyenne *a posteriori* qui s'écrit :

$$E_a(\mu|n_i, t_i, \alpha, \beta) = \frac{\alpha + n_i}{\beta + t_i}. \tag{7}$$

D'autres mesures peuvent aussi être calculées pour aider à déterminer le niveau de risque des sites à l'étude. Soit μ^m le taux médian d'accidents associé à la distribution *a priori*. Cette valeur peut être obtenue en procédant à l'intégration suivante sur l'équation (2) :

$$\int_{\mu=\mu^m}^{\infty} g_b(\mu|\alpha, \beta)d\mu = 0{,}5. \tag{8}$$

En utilisant la valeur de la médiane de la loi *a priori* ainsi que la distribution *a posteriori*, deux mesures de risque peuvent être formulées. L'expression :

$$\begin{aligned} B_{1i} &= \int_{\mu=\mu^m}^{\infty} g_a(\mu|\alpha + n_i, \beta + t_i)d\mu \\ &= Pr(\mu > \mu^m), \end{aligned} \tag{9}$$

fournit la probabilité que, par rapport au groupe de référence, le site *i* présente un plus grand risque d'accidents que la normale (valeur médiane). Un site peut être considéré comme dangereux si la valeur de la probabilité B_{1i} est supérieure à un seuil critique de 0,8, par exemple. L'interprétation précédente souligne à nouveau l'importance de bien définir la population de sites similaires. Une seconde mesure de risque plus conservatrice est obtenue comme suit :

$$\begin{aligned} B_{2i} &= \int_{\mu'=0}^{\infty} \left[\int_{\mu=\mu'}^{\infty} g_a(\mu|\alpha + n_i, \beta + t_i)d\mu \right] g_b(\mu'|\alpha, \beta)d\mu' \\ &= E_{\mu'}[Pr(\mu > \mu')]. \end{aligned} \tag{10}$$

La valeur de B_{2i} peut être interprétée comme la probabilité que le taux moyen d'accidents au site *i* soit plus grand qu'aux autres sites du même genre[2].

---

2. La première probabilité, B_1, est calculée en utilisant comme critère la valeur médiane de la loi *a priori*, alors que B_2 peut être vue comme une moyenne des valeurs de B_1, lorsque $B_1 = \Pr(\mu > \mu')$ est calculée pour toutes les valeurs possibles de μ', non seulement pour la valeur $\mu' = \mu^m$. Pour cette raison, B_2 est numériquement plus près de 0.5 que B_1. Ceci explique son caractère plus conservateur.

Ceci complète les grandes lignes nécessaires à l'implantation de l'approche bayésienne de base pour l'analyse des sites d'accidents. Notons qu'en postulant une distribution aléatoire pour le taux moyen d'accidents μ, une part d'hétérogénéité des sites est prise en compte de façon implicite dans le modèle. Cette hétérogénéité est limitée par le fait que la population de sites définie pour l'analyse doit être petite pour assurer une certaine similitude des lieux étudiés. Dans le cadre d'analyse défini par Hauer (1992), en permettant l'utilisation d'échantillons plus grands, l'hétérogénéité peut être explicitement décrite à la fois de manière déterministe et aléatoire. Tel qu'expliqué à la sous-section suivante, nous prenons en considération dans le modèle l'hétérogénéité déterministe en introduisant les caractéristiques propres aux sites dans la spécification des paramètres α et β de la distribution *a priori* $g_b(\mu \mid \alpha,\beta)$.

En comparant deux sites, il est important de pouvoir contrôler l'effet de variables telles que le débit de trafic, par exemple. L'approche que nous décrivons maintenant comporte cette importante caractéristique. Le principal avantage de cette méthodologie est de rendre le processus de définition d'une population de sites similaires moins critique. Pour tenir compte des différences non expliquées entre les sites, une composante aléatoire est également introduite dans la spécification des paramètres α et β. De cette façon, étant donnée la non disponibilité de certaines variables, on tient compte des facteurs constituant des sources de variation additionnelles entre les sites ne pouvant être pris en compte explicitement dans le modèle.

1.2 *L'approche avec hétérogénéité et corrélation spatiale*

Les deux premiers moments de la distribution *a priori* (2) prennent la forme :

$$E_b(\mu) = \frac{\alpha}{\beta} \text{ et } V_b(\mu) = \frac{\alpha}{\beta^2}. \tag{11}$$

Pour prendre en compte les deux sources d'hétérogénéité mentionnées précédemment, les paramètres généraux α et β sont remplacés par des paramètres α_i et β_i spécifiques au site i à l'étude :

$$\alpha_i = \alpha(x_i,\varphi_\alpha,\varepsilon_i) = x_i\varphi_\alpha + \sigma_\alpha\varepsilon_i \text{ et} \tag{12}$$

$$\beta_i = \beta(z_i,\varphi_\beta,\eta_i) = z_i\varphi_\beta + \sigma_\beta\eta_i\,, \tag{13}$$

où $x_i\varphi_\alpha$ et $z_i\varphi_\beta$ sont les composantes servant à expliquer l'hétérogénéité déterministe (voir Bolduc et Bonin, 1995), alors que ε_i et η_i constituent la partie non observée. Nous discutons maintenant de la loi de distribution qui gouverne ε_i et η_i. Ce sont des variables aléatoires de moyenne nulle via lesquelles se transmet l'interdépendance entre les sites. Les paramètres σ_α et σ_β correspondent à des écarts-types, qui nécessitent d'être estimés simultanément avec les autres paramètres. Plus précisément, α est fonction d'un vecteur x_i de dimension $(k \times 1)$, contenant les caractéristiques propres au site i, associé au vecteur de paramètres

φ_α et β est fonction d'un vecteur de caractéristiques z_i de dimension $(p \times 1)$ associé au vecteur de paramètres φ_β. Nous supposons que les termes ε_i et η_i peuvent être affectés par la corrélation spatiale entre les sites. Les composantes individuelles ε_i et η_i sont définies par un processus spatial autorégressif d'ordre un comme suit :

$$\varepsilon = \rho W\varepsilon + \xi = (I_N - \rho W)^{-1}\xi, \tag{14}$$

$$\eta = \rho W\eta + \nu = (I_N - \rho W)^{-1}\nu, \tag{15}$$

où ε et η représentent des vecteurs de dimensions $(N \times 1)$, ρ est un paramètre de corrélation tel que $-1 < \rho < 1$, I_N est la matrice identité de dimension $(N \times N)$ et W correspond à une matrice de poids décrivant les relations entre les sites ; W est de dimension $(N \times N)$. Pour simplifier, on fait l'hypothèse que le même processus autorégressif caractérise ε et η. Un modèle plus général attribuerait à chaque vecteur d'erreurs un processus autorégressif particulier.

Une première façon de définir les éléments de la matrice de poids W est la suivante : $w_{ij} = 1$ lorsque les sites i et j sont voisins et $w_{ij} = 0$ sinon, pour $1 \leq i, j \leq N$. Une seconde possibilité serait d'exprimer les w_{ij} comme une fonction inverse de la distance séparant les sites i et j. La proximité de deux sites leur confère habituellement des caractéristiques similaires en termes d'environnement et de débit de circulation. Deux sites peuvent également présenter des conditions semblables dû à un environnement similaire même si la distance qui les sépare est plus grande. On peut penser que deux intersections localisées près d'un bar ou d'une école se rapprochent au niveau du risque d'accidents qu'elles représentent. Une formulation générale pour tenir compte de ces liens géographiques serait :

$$w_{ij} = d_{ij}^{-\tau} \cdot p_i^{-\gamma} \cdot p_j^{-\delta},$$

où d_{ij} correspond à la distance séparant les sites i et j , et p_k est la distance moyenne entre le site k, $k = i, j$, et les bars situés à proximité, par exemple. Dans cette spécification, τ, γ et δ sont des paramètres supplémentaires à estimer. Ci-dessous, nous considérons les paramètres présents dans w_{ij} comme étant connus.

Pour clarifier la notation, posons $s = (\xi,\nu)$. Sur la base des hypothèses faites précédemment, s est un vecteur de composantes aléatoires normales standardisées indépendantes, ce que l'on écrit $s \sim N(0,I_{2N})$. La moyenne de la loi *a priori* variera d'un site à l'autre en fonction de l'information incluse dans la spécification des paramètres donnée par les équations (12) et (13). En permettant ainsi à chaque site d'avoir un taux d'accident moyen μ_i qui tient compte de ses différentes caractéristiques, la notation pour le calcul de la moyenne devient :

$$E_b(\mu_i) = \underset{s}{E}[E_b(\mu_i|s)], \text{ où } s = (\xi, \nu),$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} E_b(\mu_i|s)\, h(s)ds,$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\alpha(x_i, \varphi_\alpha, \xi)}{\beta(z_i, \varphi_\beta, \nu)} h(s)ds, \tag{16}$$

où $h(.)$ est la fonction de densité associée à $N(0,I_{2N})$. Rappelons que s est le vecteur conjoint des composantes ξ et ν, ce qui implique que l'intégrale en (16) est de dimension $2N$. Il est intéressant de noter que lorsque $\rho = 0$, la formulation se ramène au cas avec hétérogénéité seulement. Dans le cas particulier où σ_α et σ_β sont tous les deux fixés à zéro dans les équations (12) et (13), on obtient une spécification du modèle qui permet de prendre en compte uniquement l'hétérogénéité déterministe. Dans ce cas, l'hétérogénéité aléatoire et la corrélation spatiale ne sont plus traitées par le modèle ; les niveaux d'intégration disparaissent de la dernière équation qui s'exprime alors par :

$$E_b(\mu_i) = \frac{\alpha(x_i, \varphi_\alpha)}{\beta(z_i, \varphi_\beta)} = \frac{x_i, \varphi_\alpha}{z_i, \varphi_\beta}$$

L'implantation de l'approche bayésienne avec hétérogénéité et corrélation spatiale requiert l'estimation des hyper-paramètres φ_α et φ_β et du vecteur $\gamma = (\sigma_\alpha, \sigma_\beta, \rho)$ qui contient les écarts-types σ_α et σ_β et le paramètre ρ, par la maximisation de :

$$f(n_i|\varphi_\alpha, \varphi_\beta, \gamma) = \underset{s}{E}[f(n_i|\varphi_\alpha, \varphi_\beta, \gamma, s)], \text{ où } s = (\xi, \nu),$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(n_i|\varphi_\alpha, \varphi_\beta, \gamma, s) h(s)ds, \tag{17}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\Gamma(n_i + \alpha_i, (\xi))}{n_i! \Gamma(\alpha_i(\xi))} \frac{t_i^{ni} \beta_i(\nu)^{\alpha_i(\xi)}}{(t_i + \beta_i(\nu))^{n_i + \alpha_i(\xi)}} h(s)ds, \tag{18}$$

où $\alpha_i(\xi)$ et $\beta_i(\nu)$ sont calculés à partir des relations définies par les équations (12) à (15).

Pour implanter l'algorithme d'optimisation, les intégrales de l'équation (18) représentant des espérances mathématiques, et qui on le rappelle sont de dimension $2N$, sont remplacées par des moyennes empiriques telles que :

$$\tilde{f}(n_i|\varphi_\alpha, \varphi_\beta, \gamma) = \frac{1}{R} \sum_{r=1}^{R} f(n_i|\varphi_\alpha, \varphi_\beta, \gamma, s_r), \tag{19}$$

où s_r représente un tirage particulier de s parmi un total de R différents tirages effectués pour simuler $f(n_i \mid \varphi_\alpha, \varphi_\beta, \gamma)$. L'estimation du maximum de vraisemblance où la fonction $f(.)$ est remplacée par $\tilde{f}(.)$ est connue sous le nom de maximum de vraisemblance simulée (MVS). Plusieurs applications du MVS ont été réalisées au cours des dernières années et elles démontrent bien que ce type d'estimation donne de bons résultats.

Pour tenir compte de la composante aléatoire introduite dans la spécification des paramètres α_i et β_i, des ajustements doivent être apportés au niveau des formules (8) à (10) employées pour l'analyse bayésienne des fréquences d'accidents. étant donnés ξ et ν , ce qui attribue une valeur à $\varepsilon_i, \eta_i, \alpha_i$ et β_i, la fonction de densité *a posteriori* conditionnelle de μ_i s'écrit :

$$\begin{aligned} g_a(\mu_i|\alpha_i + n_i, \beta_i + t_i) &= g_a(\mu_i|\alpha_i(x_i, \varphi_\alpha, \varepsilon_i) + n_i, \beta_i(z_i, \varphi_\beta, \eta_i) + t_i), \\ &= \frac{(\beta_i + t_i)^{\alpha_i + n_i}}{\Gamma(\alpha_i + n_i)} \mu_i^{\alpha_i + n_i - 1} exp(-(\beta_i + t_i)\mu_i). \end{aligned} \tag{20}$$

Le taux médian pour chaque site, μ_i^m, est alors obtenu en solutionnant l'équation :

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{\mu=\mu_i^m}^{\infty} g_b(\mu|\alpha_i(\xi), \beta_i(\nu)) d\mu \right] h(s) ds = 0,5. \tag{21}$$

Étant donné μ_i^m, les deux probabilités suivantes peuvent être calculées :

$$\begin{aligned} B_{1i} &= \underset{s}{E}[B_{1i}(\alpha_i, \beta_i|s)], \text{ où } s = (\xi, \nu), \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} B_{1i}(\alpha_i, \beta_i|s)\, h(s) ds, \end{aligned} \tag{22}$$

et

$$\begin{aligned} B_{2i} &= \underset{s}{E}[B_{2i}(\alpha_i, \beta_i|s)], \text{ où } s = (\xi, \nu), \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} B_{2i}(\alpha_i, \beta_i|s)\, h(s) ds, \end{aligned} \tag{23}$$

où les termes $B_{1i}(\alpha_i, \beta_i \mid s)$ et $B_{2i}(\alpha_i, \beta_i \mid s)$ sont respectivement décrits par les équations (9) et (10). En pratique, B_{1i} et B_{2i} sont simulés à l'aide de moyennes empiriques. Ainsi, pour identifier les sites routiers dangereux, on utilise

$$\tilde{B}_{1i} = \frac{1}{R} \sum_{r=1}^{R} B_{1i}(\alpha_i, \beta_i|s_r), \text{ et}$$

$$\tilde{B}_{2i} = \frac{1}{R} \sum_{r=1}^{R} B_{2i}(\alpha_i, \beta_i|s_r).$$

2. ANALYSE DES PROPORTIONS D'ACCIDENTS

Cette section présente une adaptation de l'approche décrite précédemment pour l'analyse des proportions d'accidents. Le modèle est une extension directe de l'outil bayésien d'analyse utilisé dans Heydecker et Wu (1991). Leur méthodologie étudie les proportions d'accidents à un site qui impliquent une certaine caractéristique (e.g. proportion d'accidents se produisant la nuit, durant la fin de semaine, collisions à angle droit, ...). Ce type d'analyse a l'intérêt d'être complémentaire à l'analyse des fréquences d'accidents de la section précédente. Notons que les deux sources d'hétérogénéité et la corrélation spatiale des sites d'accidents sont également prises en compte dans ce modèle en incorporant dans la spécification des paramètres de la distribution *a priori* de l'information propre aux sites. Les hypothèses concernant les distributions appropriées pour l'analyse des proportions d'accidents sont maintenant décrites.

Soit x_i le nombre d'accidents enregistrés au site i impliquant une caractéristique donnée et n_i le nombre total d'accidents enregistrés au site. Le modèle suppose que les observations x_i sont distribuées suivant une loi binomiale[3] de moyenne θ_i. On écrit :

$$f(x_i|n_i,\theta_i)=\binom{n_i}{x_i}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i}, \quad 0\le x_i\le n_i. \tag{24}$$

La moyenne *a priori* θ_i est par hypothèse distribuée suivant une loi bêta de densité :

$$g_b(\theta_i|\alpha_i,\beta_i)=\frac{\theta_i^{\alpha_i-1}(1-\theta_i)^{\beta_i-1}}{B(\alpha_i,\beta_i)}, \quad 0<\theta_i<1, \tag{25}$$

où $B(\alpha_i, \beta_i)$ correspond à la fonction bêta. Les fonctions α_i, et β_i prennent des valeurs qui diffèrent d'un site à l'autre, selon les relations définies aux équations (12) et (13). Les deux précédentes expressions sont des fonctions de densité conditionnelles aux valeurs de α_i et β_i. Comme nous l'avons vu précédemment, pour obtenir les fonctions de densité non conditionnelles, il faut procéder à une intégration de dimension $2N$. La moyenne non conditionnelle de la distribution *a priori* s'exprime alors par[4] :

---

3. Au lieu de traiter l'occurrence d'un événement par rapport à sa non-occurrence, on pourrait traiter simultanément celle de plusieurs événements. Pour ce faire, il faudrait faire appel à des lois polytomiques. C'est d'ailleurs un aspect sur lequel nous travaillons présentement.

4. Le calcul de la variance non conditionnelle s'avère plus complexe. Sur la base du résultat bien connu $V(X) = E[V(X \mid Y)] + V[E(X \mid Y)]$, on obtient :

$$V_b(\theta_i)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\alpha_i\beta_i}{(\alpha_i+\beta_i)^2(\alpha_i+\beta_i+1)}h(s)ds$$
$$+\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\frac{\alpha_i}{\alpha_i+\beta_i}-E_b(\theta_i)\right]^2 h(s)ds.$$

$$E_b(\theta_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\alpha_i}{\alpha_i + \beta_i} h(s)ds, \tag{26}$$

où, pour simplifier la notation, nous omettons les arguments (ξ) et (ν) associés à α_i et β_i respectivement. En combinant les deux précédentes distributions, en appliquant des résultats connus, on obtient la distribution bêta-binomiale non conditionnelle de x_i suivante :

$$f(x_i|n_i, \varphi_\alpha, \varphi_\beta, \gamma) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \binom{n_i}{x_i} \frac{B(\alpha_i + x_i, \beta_i + n_i - x_i)}{B(\alpha_i, \beta_i)} h(s)ds. \tag{27}$$

Pour obtenir la distribution *a posteriori* de θ_i conditionnelle aux valeurs de α_i, et β_i, on applique le théorème de Bayes aux équations (24) et (25). Le résultat est une distribution bêta révisée en fonction de l'information spécifique au site étudié :

$$g_a(\theta_i|\alpha_i + x_i, \beta_i + n_i - x_i) = \frac{\theta_i^{\alpha_i + x_i - 1}(1-\theta_i)^{\beta_i + n_i - x_i - 1}}{B(\alpha_i + x_i, \beta_i + n_i - x_i)}, 0 < \theta_i < 1. \tag{28}$$

Selon le principe AEB, on maximise ensuite la log-vraisemblance de l'échantillon défini selon la distribution bêta-binomiale en (27) par rapport aux hyper-paramètres φ_α et φ_β, utilisés dans la spécification de α_i et β_i, et par rapport au vecteur γ qui caractérise les termes d'erreur.

Par la suite, l'identification des sites dangereux se fait à l'aide de la relation non conditionnelle de l'équation (28). Ainsi, les deux probabilités B_{1i} et B_{2i} peuvent se calculer de la façon suivante :

1. On calcule pour chaque site la proportion médiane d'accidents θ_i^m ajustée en fonction de l'information propre au site contenue dans x_i, z_i :

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\int_{\theta=\theta_i^m}^{1} g_b(\theta|\alpha_i, \beta_i) d\theta \right] h(s)ds = 0,5. \tag{29}$$

Étant donné θ_i^m, les deux mesures de risque s'écrivent :

$$\begin{aligned} B_{1i} &= \underset{s}{E}[B_{1i}(\alpha_i, \beta_i|s)], \text{ où } s = (\xi, \nu) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} B_{1i}(\alpha_i, \beta_i|s)\, h(s)ds \end{aligned} \tag{30}$$

et

$$\begin{aligned} B_{2i} &= \underset{s}{E}[B_{2i}(\alpha_i, \beta_i|s)], \text{ où } s = (\xi, \nu) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} B_{2i}(\alpha_i, \beta_i|s)\, h(s)ds, \end{aligned} \tag{31}$$

où :

$$B_{1i}(\alpha_i, \beta_i|s) = \int_{\theta=\theta_i''}^{1} g_a(\theta|\alpha_i + x_i, \beta_i + n_i - x_i)d\theta, \text{ et}$$

$$B_{2i}(\alpha_i, \beta_i|s) = \int_{\theta=0}^{1} \left[\int_{\theta=\theta'}^{1} g_a(\theta|\alpha_i + x_i, \beta_i + n_i - x_i)d\theta \right] g_b(\theta'|\alpha_i, \beta_i)d\theta'.$$

En pratique, les espérances mathématiques seront remplacées par des moyennes empiriques $\tilde{B}_{1i}$ et $\tilde{B}_{2i}$ tel que discuté à la section précédente. Il est important de noter que les résultats des deux approches proposées dans ce document peuvent, sous certaines conditions particulières, être comparées directement. Soit n_i la fréquence d'accidents au site i avec $n = \sum_{i=1}^{N} n_i$ représentant le nombre total d'accidents de la population. Si les proportions n_i/n sont utilisées dans le modèle exprimé en termes de proportions, des estimations très semblables à celles obtenues à l'aide du modèle pour l'analyse des fréquences seront produites et donc des conclusions similaires concernant l'identification des sites routiers dangereux seront obtenues.

3. APPLICATIONS

Nous avons appliqué les différentes versions de la méthodologie proposée pour l'analyse des fréquences aux données d'accidents de la ville de Québec. Les 224 intersections à quatre branches formées de routes de types comparables et ayant enregistré des accidents au cours de la période 1990-1993 servent de base pour la présente analyse. Parmi ces 224 sites, nous en avons retenu 90, pour constituer notre groupe de référence, qui comptent au moins 16 accidents au cours de la période d'observation de 4 ans, soit un minimum de 4 accidents par année en moyenne. Le tableau 1 présente les résultats d'estimation de la distribution non conditionnelle des fréquences d'accidents n_i. Rappelons que la fonction de densité utilisée à cet effet est la binomiale négative donnée à l'équation (4). Le tableau 2 montre les résultats de l'analyse bayésienne réalisée à l'aide des formules (9) et (10) pour les 10 sites affichant les plus grandes valeurs de probabilité B_1.

TABLEAU 1

APPROCHE DE BASE : RÉSULTATS D'ESTIMATION

Paramètre	Estimé	Écart-type	Stat. *t*
α	9,52	2,16	4,41
β	1,44	0,28	5,23
Fonction de vraisemblance : -329,674			

TABLEAU 2

APPROCHE DE BASE : LES 10 SITES AYANT LES PLUS FORTES PROBABILITÉS B_1

Numéro du site	Moyenne observée	Moyenne *a priori*	Moyenne *a posteriori*	B_1	B_2
20	18,25	6,59	15,16	1,00000	0,99736
22	16,25	6,59	13,69	1,00000	0,99194
88	15,75	6,59	13,32	1,00000	0,98945
125	15,75	6,59	13,32	1,00000	0,98945
91	11,50	6,59	10,20	0,99934	0,91669
187	11,25	6,59	10,02	0,99891	0,90721
118	11,00	6,59	9,83	0,99825	0,89684
129	10,75	6,59	9,65	0,99721	0,88553
214	10,50	6,59	9,46	0,99565	0,87321
101	10,00	6,59	9,10	0,98992	0,84537

Les résultats d'estimation et l'analyse bayésienne présentés aux tableaux 3 et 4 se rapportent au cas avec hétérogénéité déterministe. Un gain de 11 unités pour la valeur de la log-vraisemblance est observé, ce qui est appréciable compte tenu de l'ajout de deux paramètres seulement. Le modèle de base sans hétérogénéité déterministe est obtenu en fixant à zéro les coefficients $\varphi_\alpha(2)$ et $\varphi_\alpha(3)$ associés à la superficie de l'intersection et au type de contrôle (e.g. panneau d'arrêt, feu de circulation)[5]. Le fait d'inclure ces deux caractéristiques dans l'analyse modifie

5. À ce point-ci, il est peut-être nécessaire de revenir sur le concept de sous-population homogène. La variable type de contrôle, par exemple, au lieu de jouer le rôle de régresseur comme dans le cas présent, pourrait être utilisée afin de scinder ou définir les sous-populations. C'est à l'utilisateur qu'incombe la tâche d'employer judicieusement les variables spécifiques aux sites.

de manière notoire le rang des sites sur la base des valeurs de B_1 et B_2. La distribution *a posteriori* du taux d'accidents se trouve affectée par l'information spécifique au site. Certains sites considérés comme dangereux par l'approche de base, selon le critère B_1 ne font plus partie du groupe à traiter en priorité suite à l'application de l'approche proposée.

TABLEAU 3

APPROCHE AVEC HÉTÉROGÉNÉITÉ DÉTERMINISTE : RÉSULTATS D'ESTIMATION

Paramètre	Régresseur	Estimé	Écart-type	Stat. *t*
$\varphi_\alpha(1)$	Terme constant	3,53	0,52	6,72
$\varphi_\alpha(2)$	Superficie de l'intersection	0,36	0,08	4,32
$\varphi_\alpha(3)$	Type de contrôle	-0,28	0,16	-1,72
$\varphi_\beta(1)$	Terme constant	1,46	0,17	8,60

Fonction de vraisemblance : -318,407

TABLEAU 4

APPROCHE AVEC HÉTÉROGÉNÉITÉ DÉTERMINISTE :
LES 10 SITES AYANT LES PLUS FORTES PROBABILITÉS B_1

Numéro du site	Moyenne observée	Moyenne *a priori*	Moyenne *a posteriori*	B_1	B_2
88	15,75	7,41	12,86	1,00000	0,98497
22	16,25	7,75	13,31	1,00000	0,98494
101	10,00	5,12	8,31	0,99957	0,94371
129	10,75	5,69	9,00	0,99949	0,94179
118	11,00	5,93	9,24	0,99940	0,93933
187	11,25	6,54	9,62	0,99812	0,91889
125	15,75	10,51	13,94	0,99505	0,89638
91	11,50	7,45	10,10	0,99137	0,87715
20	18,25	12,98	16,43	0,99062	0,87627
167	9,75	6,83	8,74	0,96708	0,81689

Les quatre derniers tableaux donnent les estimations et l'analyse bayésienne pour les extensions du modèle de base discutées à la section 1.2. Au tableau 5, le modèle prend en compte les deux sources d'hétérogénéité et le tableau 7 concerne

les résultats d'estimation pour la version la plus générale du modèle qui incorpore également les liens entre les sites spatialement corrélés[6]. Notons que, pour les estimations de ces deux tableaux, les valeurs des paramètres σ_α et σ_β ont été fixées à 1. De plus, nous avons utilisé la première définition décrite précédemment pour la spécification des w_{ij} de la matrice de poids, soit celle où $w_{ij} = 1$ lorsque les sites i et j sont voisins, et $w_{ij} = 0$ dans le cas contraire. Notre objectif présent est de proposer et de décrire une approche générale pour l'étude des sites routiers. Basé sur les résultats d'estimation obtenus, de prendre en compte l'hétérogénéité déterministe et aléatoire lors d'une analyse améliore clairement la fonction de log-vraisemblance. Malheureusement, dans les données utilisées pour cette analyse, la corrélation spatiale ne semble pas présente. Afin de s'assurer que ce résultat ne soit pas attribué à la définition des poids w_{ij} utilisés lors de l'analyse, nous avons tenté une nouvelle spécification, exprimée en termes de la distance séparant les sites : $w_{ij} = d_{ij}^{-1}$, où d_{ij} correspond à la distance entre les sites i et j. Une spécification tenant compte d'un environnement similaire, comme la proximité d'un bar, n'était pas possible dans le présent cas. Les données disponibles rapidement n'incluaient pas de telles informations. L'utilisation à venir d'une base de données plus complète incluant de meilleurs descripteurs des sites et de leur environnement devrait nous permettre de produire des exemples convaincants où la corrélation spatiale sera présente de manière significative.

TABLEAU 5

APPROCHE AVEC HÉTÉROGÉNÉITÉ DÉTERMINISTE ET ALÉATOIRE

Paramètre	Régresseur	Estimé	Écart-type	Stat. t
$\varphi_\alpha(1)$	Terme constant	13,50	2,06	6,57
$\varphi_\alpha(2)$	Superficie de l'intersection	0,92	0,34	2,73
$\varphi_\alpha(3)$	Type de contrôle	-1,06	0,52	-2,05
$\varphi_\beta(1)$	Terme constant	5,44	0,76	7,19
Fonction de log-vraisemblance simulée : -310,225				
Nombre de tirages : 50				

6. Pour les analyses bayésiennes présentées aux tableaux 6 et 8, associées aux estimations des tableaux 5 et 7 respectivement, notons d'abord que seules les probabilités B_1 sont données dans les tableaux. Pour justifier l'omission des probabilités B_2, nous signalons tout d'abord que leur calcul est très laborieux. De plus, il faut observer que dans ces versions plus générales, les valeurs de la probabilité B_1, sont plus faibles et offrent plus de discernement que dans les versions de base ou à hétérogénéité déterministe. Ceci s'explique par la formulation des équations pour l'analyse bayésienne sous forme de valeurs moyennes, menant ainsi à des valeurs plus conservatrices. Mentionnons cependant que, bien que plus faibles, les valeurs obtenues sont cohérentes avec l'identification des sites faite par les précédentes versions du modèle, puisque les mêmes sites sont retenus comme ceux présentant un plus grand risque d'accidents routiers.

TABLEAU 6

APPROCHE AVEC HÉTÉROGÉNÉITÉ DÉTERMINISTE ET ALÉATOIRE : LES 10 SITES AYANT LES PLUS FORTES PROBABILITÉS B_1

Numéro du site	Moyenne observée	Moyenne *a priori*	Moyenne *a posteriori*	B_1
101	10,00	5,02	5,56	0,80245
88	15,75	7,37	8,34	0,73542
22	16,25	7,52	8,54	0,73541
129	10,75	5,60	6,18	0,71560
187	11,25	6,50	7,00	0,71529
118	11,00	6,32	6,86	0,69167
125	15,75	9,57	10,10	0,67786
20	18,25	11,12	11,85	0,67112
167	9,75	6,57	6,93	0,64895
91	11,50	7,51	7,89	0,62966

TABLEAU 7

APPROCHE AVEC HÉTÉROGÉNÉITÉ ET CORRÉLATION SPATIALE

Paramètre	Régresseur	Estimé	Écart-type	Stat. *t*
$\varphi_\alpha(1)$	Terme constant	13,51	2,08	6,51
$\varphi_\alpha(2)$	Superficie de l'intersection	0,92	0,34	2,74
$\varphi_\alpha(3)$	Type de contrôle	-1,06	0,52	-2,04
$\varphi_\beta(1)$	Terme constant	5,45	0,76	7,16
ρ	Paramètre de corrélation	0,02	0,75	0,03
Fonction de log-vraisemblance simulée : -310,222				
Nombre de tirages : 50				

TABLEAU 8

APPROCHE AVEC HÉTÉROGÉNÉITÉ ET CORRÉLATION SPATIALE : LES 10 SITES AYANT LES PLUS FORTES PROBABILITÉS B_1

Numéro du site	Moyenne observée	Moyenne *a priori*	Moyenne *a posteriori*	B_1
101	10,00	5,02	5,55	0,80024
22	16,25	7,53	8,54	0,73505
88	15,75	7,38	8,34	0,73313
129	10,75	5,60	6,18	0,71536
187	11,25	6,51	7,00	0,71423
118	11,00	6,32	6,86	0,69135
125	15,75	9,58	10,10	0,67749
20	18,25	11,12	11,86	0,67075
167	9,75	6,57	6,93	0,65025
91	11,50	7,51	7,89	0,62737

CONCLUSION

Dans ce document, nous avons décrit une méthodologie générale pour l'analyse des sites d'accidents de la route qui permet de prendre en compte l'hétérogénéité des sites et la corrélation spatiale dans un cadre d'analyse bayésienne à information complète. Deux types d'analyses complémentaires sont présentés : approche en termes de fréquences et approche en termes de proportions d'accidents. Comme l'exemple empirique l'a démontré, l'hétérogénéité peut affecter de manière significative les estimations. Pour la banque de données que nous avons utilisée, la corrélation spatiale n'était pas présente de manière significative. Nous attribuons ce dernier résultat à la non disponibilité dans la base de données de certaines informations utiles à l'analyse. La généralisation proposée dans ce rapport présente un grand potentiel pour l'aide à la prise de décision en sécurité routière concernant l'identification des sites routiers dangereux.

BIBLIOGRAPHIE

BOLDUC, D., et S. BONIN (1995), « Bayesian Analysis of Road Accidents : Accounting for Deterministic Heterogeneity », *Compte-rendus de la IX^e Conférence canadienne multidisciplinaire sur la sécurité routière, Montréal, Canada.*

HAUER, E. (1992), « Empirical Bayes Approach to the Estimation of Unsafety : the Multivariate Regression Method », *Accident Analysis & Prevention*, Vol. 24, No 5 : 457-477.

HAUER, E. (1986), « On the Estimation of the Expected Number of Accidents », *Accident Analysis & Prevention*, Vol. 18, No 1 : 1-12.

HEYDECKER, B., et J. WU (1991), Using the Information in Road Accident Records », For presentation at the 19th Summer Annual Meeting of PTRC, University of Sussex, England.

HIGLE, J.L., et J.M. WITKOWSKI (1989), « Bayesian Identification of Hazardous Locations », *Transportation Research Record*, 1185 : 24-36.

LOVEDAY, J., et D. JARRETT (1992), « Spatial Modelling of Road Accident Data », *Mathematics in Transport Planning and Control*, Institute of Mathematics and its Applications Conference, Series 38 : 433-446.

MAHER, M.J. (1990), « A Bivariate Negative Binomial Model to Explain Traffic Accident Migration », *Accident Analysis & Prevention*, Vol. 22, No. 5 : 487-498.