Méthodes quantitatives pour l'identification de sections de route dangereuses - Aspects généraux, approches bayésiennes empiriques

Thierry Brenac

Thierry Brenac
Université Gustave Eiffel (anc. Ifsttar), TS2, LMA
F-13300 Salon de Provence
thierry.brenac@univ-eiffel.fr

Version finale, 5 août 2020

Méthodes quantitatives pour l'identification de sections de route dangereuses

Aspects généraux, approches bayésiennes empiriques

1. Introduction

L'identification des sources de danger le long des infrastructures routières peut prendre différentes voies. Certaines s'appuient directement sur l'identification de configurations connues pour être dangereuses (obstacles fixes massifs à proximité immédiate de chaussées non urbaines, par exemple). D'autres démarches ont recours aux données accidentologiques qui constituent, sous certaines conditions, de bons prédicteurs des accidents à venir en l'absence de mesures de sécurité (nous y reviendrons).

Parmi ces démarches accidentologiques, on peut mentionner d'abord les démarches qualitatives s'appuyant sur l'analyse cas par cas des accidents sur la base des rapports de gendarmerie ou de police (procès-verbaux) complétée par des analyses de terrain. Il s'agit de démarches compréhensives, accordant une large place aux déclarations des impliqués et témoins, et aux particularités des sites et de leur contexte. Elles permettent d'identifier les processus à l'œuvre dans les accidents, la nature et la localisation des configurations d'infrastructure contribuant à ces processus, et constituent une base solide pour la mise au point de mesures de sécurité. Il peut être difficile de mettre en œuvre de telles démarches de façon systématique sur de grands linéaires de route et pour de grands nombres d'accidents.

D'autres démarches accidentologiques sont plutôt quantitatives. Elles visent à identifier des zones susceptibles de contenir des configurations d'infrastructure ou d'environnement contribuant à la production d'accidents. Elles s'appuient sur les nombres d'accidents observés, leur localisation, éventuellement les niveaux de risque mesurés (rapports des nombres d'accidents aux volumes de circulation). Elles ne dispensent pas d'analyses qualitatives ultérieures, sur les zones ainsi repérées, si l'on veut pouvoir identifier précisément les problèmes et mettre au point des mesures de sécurité pertinentes.

C'est à ces démarches quantitatives que nous nous intéressons dans cette note. La section 2 revient rapidement sur les données utilisées et leurs limites. Dans la section 3, nous introduisons quelques considérations sur la nature en partie aléatoire des accidents, sur la prise en compte du nombre d'accidents comme réalisation d'une variable aléatoire de Poisson, et sur la difficulté à connaître précisément le degré d'accidentalité d'un site du fait de ces aspects aléatoires. La section 4 examine les limites des approches conventionnelles d'identification de sections dangereuses s'appuyant sur les nombres d'accidents et taux observés et sur l'utilisation de seuils et d'intervalles de confiance. Les sections 5 et 6 montrent l'apport de l'estimation de la moyenne de la variable de Poisson par une estimation bayésienne empirique, et l'intérêt que peuvent présenter d'autres développements s'appuyant sur l'analyse bayésienne empirique, en présentant le principe de ces méthodes. La section 7 présente un exemple. La section 8 porte sur le cas où on s'intéresse à des taux d'accidents. Dans la section 9, nous revenons sur quelques choix méthodologiques (sectionnement, critères de classement des sections : nombre, taux, gain potentiel, coût d'insécurité…) et leurs implications.

2. Problèmes liés à la qualité des données

Il est connu que les accidents corporels de la circulation ne sont que très incomplètement recensés par les forces de l'ordre (Amoros *et al.*, 2006), en France comme dans la plupart des autres pays. Cela ne poserait qu'un problème relatif si ce sous-enregistrement ne concernait que les accidents les plus bénins, et si le degré de sous-enregistrement était relativement stable dans le temps et dans l'espace. Or ce n'est pas le cas. En particulier, les pratiques peuvent varier d'une circonscription administrative à l'autre, selon le service de gendarmerie ou de police concerné. Et pour un même ressort territorial, le degré de recensement peut varier grandement dans le temps, du moins pour les accidents corporels sans hospitalisation[1]. Il faut donc prendre garde à ce que la sophistication des analyses quantitatives ne conduise pas à masquer la fragilité de leurs résultats, liée entre autres aux limites des données sur lesquelles elles s'appuient.

1. En voici un exemple, concernant l'agglomération du Grand Dijon : « une nouvelle procédure de recensement des accidents a été mise en place [en 2011] par les services de la police nationale, ne tenant plus compte de certaines personnes non hospitalisées », conduisant à une réduction de 75 %, entre 2010 et 2011, du nombre de blessés non hospitalisés recensés dans le fichier BAAC sur le territoire du Grand Dijon (Communauté d'agglomération du Grand Dijon, 2013, p. 34).

D'autre part, la recherche de zones dangereuses sur la base d'une approche quantitative des accidents observés suppose évidemment que les accidents soient bien localisés dans l'espace. Il semble que la localisation par coordonnées GPS du lieu de l'accident soit présente dans le fichier national des accidents corporels, de façon quasi-systématique depuis 2017, du moins pour le secteur d'intervention de la gendarmerie nationale (présence des coordonnées GPS dans 97 % des cas sur le secteur gendarmerie). Un état du pourcentage de cas d'accidents disposant d'une localisation GPS dans ce fichier sur les années 2005-2016 a été publié (El Mansouri et Fournier, 2018). Mais la précision de la localisation GPS de l'accident reste assez mal connue. Sur ce point, une première évaluation sommaire a été faite sur 50 cas d'accidents (en milieu urbain et non urbain) pour lesquels les coordonnées GPS étaient fournies dans le fichier national des accidents. Pour 86 % d'entre eux la distance entre le lieu réel de l'accident – obtenu après des investigations poussées incluant l'analyse du procès-verbal d'accident – et la position GPS donnée dans le fichier est de moins de 100 m. Mais dans 6 % des cas la différence excède 300 m (El Mansouri et Fournier, 2018).

Enfin, certaines techniques d'identification de zones dangereuses utilisent des indicateurs rapportant les nombres d'accidents aux volumes de trafic : taux d'accidents, au sens du nombre d'accidents sur la période d'étude rapporté aux kilomètres parcourus (véhicules kilomètres) sur la même période, ou nombre d'accidents en carrefour rapporté aux volumes de trafic sur les deux routes concernées. Selon les réseaux étudiés, et en particulier sur les réseaux ouverts (donc hors autoroutes et infrastructures similaires), les données peuvent être manquantes ou imparfaites (compteurs en nombre trop limités compte tenu des changements de volume de trafic liés à certains carrefours intermédiaires ; compteurs temporaires impliquant des extrapolations un peu hasardeuses ; compteurs en panne sur certaines périodes, etc.). D'autre part, il est fréquent que, pour les intersections, le niveau de trafic sur la route secondaire ne soit pas connu. Si l'on souhaite prendre en compte plus systématiquement les trafics dans la recherche de zones ou points dangereux, on peut se demander s'il ne faudrait pas accéder à certaines données des opérateurs de télécommunication sur les traces GPS des smartphones pour avoir des informations plus systématiques (la réquisition de ces données pour des usages d'intérêt public ne serait pas nécessairement choquante). Cela ne peut se faire sans certaines précautions (règlementation sur la protection des données personnelles), et cela supposerait aussi quelques calibrages et vérifications par comparaison avec les mesures de compteurs fixes.

3. Aspects aléatoires, utilisation de la distribution de Poisson

La survenue d'un accident est un phénomène en partie aléatoire au sens où, si elle peut être vue comme la conséquence d'un certain nombre de déterminants, d'états, de faits où événements antérieurs, elle dépend néanmoins aussi d'un grand nombre de conditions dont la présence ou la coïncidence dans le temps et dans l'espace peut être considérée comme aléatoire (comme le fait que deux usagers soient présents au même moment dans une intersection, par exemple, ou, plus simplement, la conjonction d'un état de grande fatigue chez le conducteur et de la présence d'une forte pluie limitant la visibilité , etc.).

Du fait de ces facteurs aléatoires, sur un site routier, même lorsque le volume de trafic et le niveau intrinsèque de sécurité de ce site ne varient pas dans le temps, le nombre observé d'accidents subit d'importantes fluctuations aléatoires d'une période d'observation à l'autre.

Pour tenir compte de ces fluctuations aléatoires dans les analyses, il est usuel d'avoir recours à la distribution statistique de Poisson. Le recours à la distribution de Poisson est justifié dans la mesure où elle est particulièrement adaptée à la représentation d'événements rares et indépendants, et où, en outre, divers travaux de recherche ont confirmé la bonne adéquation de cette distribution aux données empiriques sur les accidents survenus sur des sites routiers considérés individuellement (voir par exemple Nicholson et Wong, 1993 ; Jarrett, 1994).

Dans ce cadre, le nombre observé d'accidents sur une période d'étude donnée (de cinq ans, disons, ou tout autre choix), que nous noterons x, est vu comme la réalisation d'une variable aléatoire de Poisson X, de moyenne m. Cette moyenne m (espérance mathématique de X) n'est pas observable mais on considère qu'elle représente le « vrai » niveau d'accidentalité du site au sens où, si on pouvait observer ce site sur une infinité de périodes semblables et dans des conditions identiques, on

observerait en moyenne *m* accidents par période. Dans cette perspective, la probabilité d'observer *x* accidents sur la période d'étude est liée à la moyenne *m* par la relation suivante :

$$P(X = x) = \frac{e^{-m}\ m^x}{x!} \qquad (1)$$

pour tout entier *x* (de zéro à l'infini). La figure 1 en donne une illustration pour le cas où *m* = 3,5.

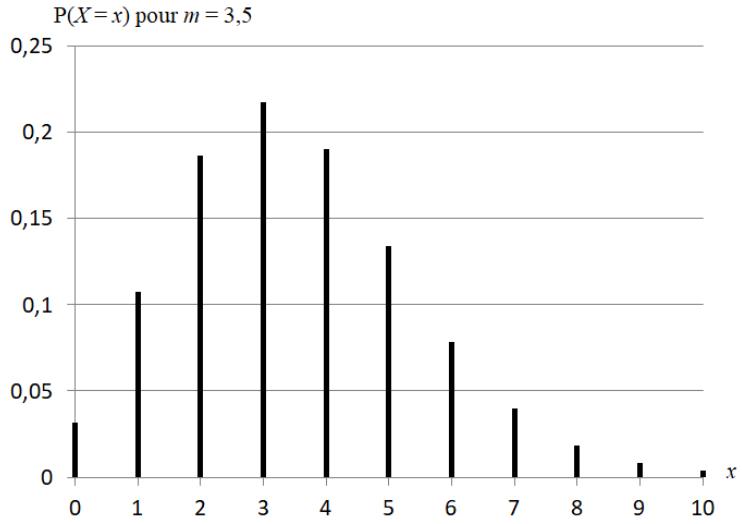


Figure 1. Exemple : probabilité d'observer *x* accidents pour une variable de Poisson *X* de moyenne *m* = 3,5

La moyenne *m* n'est pas observable, mais l'observation *x* à laquelle nous avons accès peut nous renseigner dans une certaine mesure sur les valeurs de *m* les plus compatibles avec l'observation *x*, par exemple au moyen de l'intervalle de confiance de la moyenne *m* que l'on peut associer à l'observation *x*. Prenons l'exemple de l'intervalle de confiance à 95 % (à risques symétriques) de la moyenne *m*, pour une valeur observée *x*. Les bornes supérieure et inférieure de cet intervalle de confiance sont définies par les conditions suivantes[2] :

$$P(X \leq x \mid m = m_{sup}) = 0{,}025$$
$$P(X \geq x \mid m = m_{inf}) = 0{,}025 \qquad (2)$$

où la notation « $\mid m = m_{sup}$ » signifie « sous la condition $m = m_{sup}$ ».

Si l'on a besoin d'une estimation ponctuelle de la moyenne *m*, et non seulement d'un intervalle, et si l'on ne dispose que de l'observation *x*, l'estimation usuelle de *m* est $\hat{m} = x$. Cette estimation[3] ne présente pas de biais[4] au sens où, si l'on répétait la même expérience aléatoire un grand nombre de fois, la moyenne des estimations $\hat{m}$ serait bien égale à *m* : en effet l'espérance de $\hat{m}$ vaut $E(\hat{m}) = E(x) = m$. Il n'y a donc pas d'erreur systématique[5] ; mais lorsqu'on utilise une telle estimation, il y a malgré tout une erreur aléatoire : en effet $\hat{m}$ (= *x*) présente généralement un écart, positif ou négatif, par rapport à *m*, lié aux fluctuations aléatoires de *x*. L'amplitude $|\hat{m} - m|$ de cette erreur, en proportion

2. Autrement dit m_{sup} est la plus grande valeur de *m* pour laquelle l'observation d'un nombre d'accidents inférieur ou égal à *x* n'est pas excessivement improbable (pour une valeur de *m* supérieure, la probabilité d'une telle observation serait inférieure à 2,5 %). Symétriquement, m_{inf} est la plus petite des valeurs de *m* pour laquelle l'observation d'un nombre d'accidents supérieur ou égal à *x* n'est pas excessivement improbable (pour une valeur de *m* inférieure, la probabilité d'une telle observation serait inférieure à 2,5 %).

3. C'est l'estimation du maximum de vraisemblance, c'est-à-dire, en l'occurrence, la valeur de *m* qui maximise $P(X = x \mid m)$. Le symbôle « ^ » au-dessus de la lettre *m* signifie qu'il s'agit d'une estimation de *m*, et non de *m*.

4. Dans le cas courant ; mais dans certains contextes d'étude, où l'on se sert du nombre observé *x* pour sélectionner des sites, l'estimation $\hat{m} = x$ peut présenter un biais résultant directement de cette opération de sélection (biais de sélection) : voir section 4.

5. De façon générale, l'erreur d'estimation sur un paramètre φ, c'est-à-dire $(\hat{\varphi} - \varphi)$, se décompose de la façon suivante : $(\hat{\varphi} - \varphi) = [\hat{\varphi} - E(\hat{\varphi})] + [E(\hat{\varphi}) - \varphi]$, où le premier terme représente l'erreur aléatoire, et le second le biais, ou erreur systématique. En l'absence de biais $(\hat{\varphi} - \varphi) = [\hat{\varphi} - E(\hat{\varphi})]$: l'erreur se réduit à l'erreur aléatoire.

de la valeur de *m*, est en moyenne d'autant plus importante que l'on s'appuie sur de petits nombres d'accidents. La figure 1 ci-dessous représente la moyenne quadratique de cette erreur[6], rapportée à *m*, en fonction de la valeur de *m*.

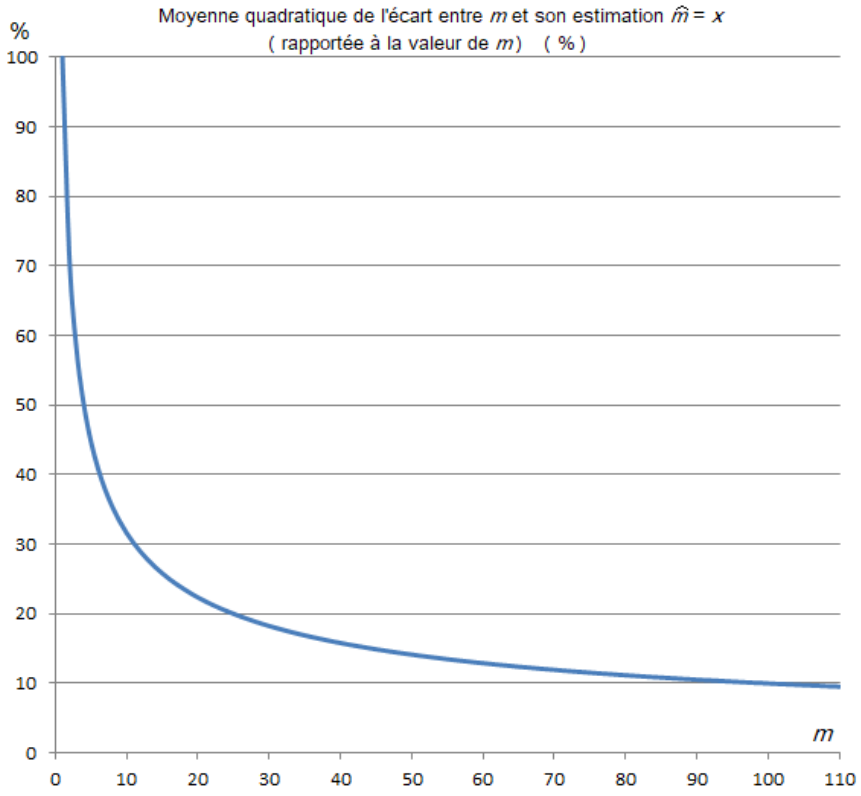


Figure 2. Moyenne quadratique de l'erreur sur *m* (en proportion de la valeur de *m*) lors de l'estimation de *m* par la valeur observée *x*

Analyses portant sur des taux

Dans certaines circonstances, on s'intéresse davantage à un *taux*, c'est-à-dire au nombre d'accidents rapporté à une quantité connue, qu'on appellera l'exposition, et qui peut être selon les cas :

- la circulation (véhicules kilomètres, par exemple, si l'on s'intéresse au risque moyen d'accidents par kilomètre parcouru ou *taux d'accidents*),
- le temps d'observation (si l'on s'intéresse à la fréquence des accidents dans le temps),
- ou encore la longueur d'une section de route (si l'on s'intéresse à la *densité d'accidents* par unité de longueur)...

Dans ces situations, le nombre d'accidents *x* peut être considéré comme une réalisation d'une variable aléatoire de Poisson *X*, de moyenne $m = \lambda t$ où λ représente le taux d'accidents par unité d'exposition et *t* la quantité d'exposition (véhicules kilomètres par exemple), qui est une valeur connue, non aléatoire. Le taux observé (x / t) est alors une réalisation d'une variable aléatoire $\Lambda = (X / t)$ dont la distribution statistique peut être facilement déduite de celle de *X*, puisqu'on peut noter que $P[\Lambda = (x/t)] = P[(X/t) = (x/t)] = P(X = x)$. Les principales caractéristiques et propriétés de cette variable se déduisent aisément de celles de la variable de Poisson *X* qui rend compte du nombre brut d'accidents observé[7].

---

6. C'est-à-dire la racine carrée de l'espérance de $(\hat{m} - m)^2$: $\sqrt{E[(\hat{m} - m)^2]}$. Cette valeur, appelée *root-mean-square error* ou *root-mean-square deviation* en anglais (RMSE ou RMSD), est une valeur positive qui donne une idée de l'amplitude moyenne de l'erreur ; mais chaque erreur $(\hat{m} - m)$ peut être positive ou négative, bien entendu.

7. On notera cependant que $\Lambda = X / t$ n'est pas une variable de Poisson, puisqu'elle prend des valeurs non-entières. De plus, sa variance diffère de son espérance (pour une variable de Poisson, la variance est égale à l'espérance *m*) : la variance de Λ vaut $m/t^2 = \lambda/t$, alors que son espérance vaut $m/t = \lambda$. Cela résulte directement des propriétés de la variance et de l'espérance du produit d'une variable aléatoire (*X*, en l'occurrence) par un scalaire (ici *t*, qui est un réel de valeur fixée, non aléatoire).

En particulier, l'espérance du taux (la moyenne ou valeur « vraie ») vaut $\lambda = m / t$ et les bornes de son intervalle de confiance valent $\lambda_{inf} = (m_{inf} / t)$ et $\lambda_{sup} = (m_{sup} / t)$. Il suffit donc de diviser les bornes de l'intervalle de confiance de m associé au nombre brut d'accidents constatés x, par l'exposition t, pour obtenir l'intervalle de confiance du taux (x représente le nombre d'accidents observé sur l'ensemble de la période d'étude ; t doit être aussi calculé sur l'ensemble de cette période). Si l'on ne dispose que des informations x et t, l'estimation ponctuelle[8] du taux λ est $\hat{\lambda} = x/t$.

4. Limites des approches conventionnelles

Certaines des méthodes employées dans l'analyse quantitative des données d'accidents de la circulation se limitent à une vision naïve (identifiant le niveau d'accidentalité d'un site à la valeur observée x du nombre d'accidents, ou du taux d'accidents). Plus couramment, les méthodes utilisées essayent de prendre en compte les fluctuations aléatoires en utilisant des intervalles de confiance sur la moyenne m, ou en procédant à des tests statistiques pour évaluer des hypothèses relatives à la valeur de la moyenne m compte tenu de la valeur observée x. De telles méthodes peuvent convenir et sont opératoires dans de nombreux cas.

Néanmoins, elles présentent de sérieuses limites et des biais importants dans d'autres circonstances, et en particulier dans les situations où l'on cherche à identifier des zones dangereuses, à classer des sites routiers ou des sections de route par niveau de dangerosité, ou encore lorsque l'on procède à des évaluations avant-après de mesures de sécurité appliquées à des sites routiers.

Biais de sélection (régression vers la moyenne)

Ces limites tiennent principalement à un biais qui fait partie de ce que l'on appelle les « biais de sélection ». Il est connu sous le nom de biais de régression vers la moyenne (*regression to the mean* ou *regression towards the mean*) du fait de ses conséquences dans le cas d'analyses de type avant-après. Ce biais est une conséquence directe du fait que dans l'analyse on applique une sélection basée sur le nombre observé d'accidents, pour classer les sites par niveau de dangerosité, ou pour sélectionner ceux qui seront traités. Son rôle délétère dans le domaine des études d'évaluation des effets des aménagements de points noirs est bien connu depuis de nombreuses décennies (voir par exemple Abbess *et al*., 1981 ; Mountain et Fawaz, 1991 ; Hauer, 1997). Nous expliquons ci-dessous schématiquement la nature de ce biais.

Supposons que l'on s'intéresse, sur une période d'étude donnée P_1, à un grand ensemble de N sites routiers (i = 1 à N), et que, pour chaque site i, on note x_i le nombre observé d'accidents et m_i la moyenne (inconnue) de la variable de Poisson correspondante. Selon le site considéré, x_i peut se situer au voisinage de m_i ou être supérieur à m_i ou encore inférieur à m_i, du fait des fluctuations aléatoires que nous avons évoquées précédemment.

Maintenant supposons en outre que, dans le but d'identifier des sites à plus forte accidentalité (pour décider de priorités d'étude ou de traitement), on sélectionne les sites où le nombre d'accidents observé est supérieur à un seuil S, donc tous les sites vérifiant $x_i > S$. Alors une partie non négligeable des sites sélectionnés aura été sélectionnée du fait des fluctuations aléatoires, ayant conduit par hasard à un nombre observé d'accidents x_i élevé sur la période étudiée (supérieur au seuil), malgré une moyenne m_i faible. Les estimations $\hat{m}_i = x_i$ tendent alors à être globalement supérieures aux valeurs vraies m_i sur les sites sélectionnés[9]. Elles tendent ainsi à surestimer les m_i et sont donc biaisées[10], à la différence du cas général (évoqué dans la section 3). Ce biais est directement lié à l'introduction d'une sélection sur le critère du nombre d'accidents observé.

---

8. C'est l'estimation usuelle du maximum de vraisemblance. L'écart entre cette estimation du taux et le taux vrai λ, en moyenne quadratique et en valeur relative par rapport à λ, ne dépend pas de l'exposition, mais dépend du nombre d'accidents sur la base duquel le taux a été calculé (plus exactement : de l'espérance m de ce nombre). Ainsi, pour une estimation du taux d'accidents $\hat{\lambda} = x/t$ calculée sur la base d'un nombre d'accidents x de l'ordre d'une centaine (correspondant à $m \approx 100$), l'écart entre $\hat{\lambda}$ et λ sera, en moyenne quadratique, voisin de 10 % de λ ; mais il sera d'environ 50 % si $m \approx 4$, d'environ 25 % si $m \approx 16$, d'environ 20 % si $m \approx 25$, d'environ 15 % si $m \approx 44$ (voir figure 2).

9. On peut d'ailleurs montrer que l'espérance conditionnelle de $\hat{m}_i (= x_i)$, conditionnellement à $X_i > S$, est supérieure à m_i.

10. Et à l'inverse, les moyennes des sites pour lesquels x_i est inférieur au seuil sont sous-estimées.

De plus, si on utilise, pour une évaluation avant-après, les valeurs x_i des nombres d'accidents observés sur la période d'étude (P_1) sur les sites sélectionnés, en les comparant aux observations ultérieures x_{i2} sur ces mêmes sites sur une période P_2 semblable à P_1, l'évolution calculée sur cette base est biaisée. On constate en effet, en moyenne, une baisse du nombre d'accidents observé sur les sites sélectionnés, *même en l'absence de tout traitement et de toute évolution réelle* (donc même dans le cas où pour tous ces sites les moyennes m_{i2} sont égales aux moyennes m_i). Cette baisse est en réalité un *artefact statistique* qui traduit le fait que dans la deuxième période, les observations tendent logiquement à être proches globalement des valeurs centrales de leurs distributions (les valeurs les plus probables), donc proches des valeurs moyennes[11], même lorsque dans la première période les observations, les x_i, tendent à être globalement supérieures aux moyennes m_i du fait du processus de sélection. C'est en ce sens qu'on parle de « régression vers la moyenne » (retour vers la moyenne). Le biais peut conduire alors à conclure à tort à l'efficacité d'une mesure qui en réalité, au vu d'une analyse corrigeant ce biais, n'est pas efficace[12], ou à surévaluer parfois fortement les effets d'une mesure modérément efficace, voire à sous-évaluer les effets négatifs d'une mesure contre-productive.

Remarque : Le biais que nous évoquons ici n'est, bien sûr, pas particulièrement lié à la distribution de Poisson. Il peut aussi concerner l'étude de phénomènes relevant de toutes sortes d'autres distributions statistiques, et porter par exemple sur des variables anthropométriques, souvent gaussiennes. Un exemple classique est celui des travaux de Francis Galton (1886) au dix-neuvième siècle, qui a étudié les facteurs héréditaires influant sur la taille des individus. Après avoir mesuré la taille des individus sur un échantillon de jeunes adultes, mesuré la taille de leurs parents, et calculé la taille moyenne des deux parents pour chaque jeune adulte, il a observé que lorsque les parents étaient de grande taille, leur enfant (à l'âge adulte) était en moyenne de taille plus petite que ses parents. Inversement, lorsque les parents étaient de petite taille, leur enfant tendait en moyenne à être plus grand qu'eux à l'âge adulte. Il parlait à ce sujet d'un phénomène de retour à la médiocrité (*regression towards mediocrity*), à l'origine de l'appellation « régression vers la moyenne ». Cependant, il donnait à tort une interprétation génétique à ce constat, alors qu'il ne s'agissait que d'un artefact statistique entièrement dû au biais que nous venons d'évoquer — lié aux fluctuations aléatoires de la taille dans une lignée, et à la segmentation de l'échantillon selon des seuils de taille — et qui ne rend pas compte de la réalité de la transmission héréditaire de la stature.

En conclusion, l'utilisation des estimations usuelles $\widehat{m}_i = x_i$ pour classer des sites en fonction des enjeux en termes d'accidentalité, ou pour évaluer l'effet d'une mesure appliquée à des sites sélectionnés sur la base de tels seuils (en comparant les estimations $\widehat{m}_i = x_i$ aux estimations $\widehat{m}_{i2} = x_{i2}$ correspondant à la période après aménagement[13]) présente des inconvénients notables.

Dans le cas des taux d'accidents, l'utilisation des estimations usuelles $\widehat{\lambda}_i = x_i/t_i$ pour le même type d'usage (classement, évaluation) pose des problèmes similaires.

5. Utilisation d'estimations bayésiennes empiriques : une présentation rapide

Nous présentons ici brièvement cette démarche du point de vue de son intérêt pratique (sans développer les considérations et justifications théoriques, sur lesquelles nous reviendrons dans la section suivante) — dans le cadre d'études visant à classer ou sélectionner des sites routiers en vue de déterminer des priorités d'étude ou de traitement, ou dans le cadre d'évaluations avant-après. Cette démarche est considérée par certains auteurs comme un standard, intégrant les développements de la recherche des dernières décennies, et dont l'application devrait être la règle (voir par exemple Elvik, 2007, qui parle de « *state-of-the-art approach* »). Et de fait, il a été montré qu'elle permet de neutraliser ou de réduire fortement les effets du biais de sélection qui vient d'être évoqué.

L'objectif poursuivi est de disposer d'estimations plus solides des moyennes, plus proches de leurs valeurs réelles m_i et de ce fait moins sujettes au biais décrit dans la section précédente.

L'idée qui sous-tend l'utilisation d'estimations bayésiennes empiriques, dans la perspective qui nous intéresse ici, est la suivante : concernant l'accidentalité d'un site routier donné, nous disposons

11. En effet, dans la période P_2, les observations x_{i2} ne sont conditionnées par aucune présélection, et sont donc globalement centrées autour des moyennes m_{2i}, puisque les valeurs centrales correspondent aux probabilités les plus élevées (voir par exemple la figure 1). La moyenne des x_{i2} constitue alors une estimation non biaisée de celle des m_{i2} (= m_i dans le cas d'une absence de traitement et d'évolution).

12. Voir par exemple les résultats sur l'évaluation des effets d'un programme de radars mobiles dans le Norfolk (données de Jones *et al.*, 2008, ré-analysées par Brenac, 2010). Les moyens de corriger le biais reposent sur des approches bayésiennes telles que présentées dans les sections 5 et 6.

13. Rappelons cependant que l'utilisation des x_{i2} pour l'estimation des m_{i2} n'introduit pas de biais, c'est l'utilisation des x_i pour estimer les m_i qui est à l'origine du biais dans ce type d'analyse, du fait du seuil imposé sur la valeur des x_i.

de deux sources d'information possibles — d'une part l'observation des accidents survenus sur ce site, et d'autre part des connaissances *a priori* (que l'on peut puiser notamment dans les données collectées sur des sites comparables) — qu'il est possible de combiner pour estimer la moyenne *m*. Autrement dit, nous pouvons nous appuyer (*i*) sur l'expérience, l'historique des accidents survenus sur le site d'intérêt sur une période récente, et (*ii*) sur le fait que ce site relève d'une famille de sites similaires, dont on connaît le « comportement », du point de vue de l'accidentalité, au travers notamment de la moyenne et de la variance des nombres d'accidents observés sur ces sites. Et ces deux aspects, (*i*) et (*ii*), peuvent être pris en compte pour estimer *m*.

Concrètement, pour chaque site étudié, il s'agit d'utiliser, au lieu de l'estimation usuelle $\widehat{m} = x$ de la moyenne de la variable de Poisson, une autre estimation (une estimation bayésienne empirique de *m*) qui peut s'exprimer comme une pondération entre le nombre d'accidents observé *x* et la moyenne des nombres d'accidents observés sur un échantillon de sites comparables[14] :

$$\widehat{m}_{BE} = v\,\bar{y} + (1 - v)\,x \qquad (3)$$

où $\bar{y}$ représente la moyenne des nombres d'accidents observés y_j sur un échantillon de sites comparables au site sous examen, et où *v* et $(1 - v)$ représentent les poids affectés aux deux termes. Le poids *v* est défini par $v = \bar{y}/s^2$ où s^2 représente la variance des y_j. Le poids $(1 - v)$ s'en déduit et vaut donc $1 - v = 1 - (\bar{y}/s^2) = (s^2 - \bar{y})/s^2$. Des formulations comparables, pouvant varier dans les notations et la façon de présenter les poids mais identiques sur le fond, figurent dans différentes publications (voir entre autres : Mountain et Fawaz, 1991 ; Mountain *et al*., 1992*a* ; Jarrett, 1994 ; Brenac, 1994 ; Persaud, 1986).

Les nombres d'accidents y_j sont supposés distribués chacun selon une distribution de Poisson, les moyennes de ces distributions de Poisson étant elles-mêmes considérées comme des réalisations d'une variable aléatoire. Mais il n'est pas nécessaire de connaître la distribution de cette variable aléatoire (sa loi de probabilité) pour que l'estimation donnée par l'équation (3) soit valide[15] (Robbins, 1985 ; Jarrett, 1994). Néanmoins, on fait souvent l'hypothèse que ces moyennes sont distribuées selon une distribution Gamma[16]. On montre alors que, dans ce cas, la distribution des y_j sur les différents sites considérés dans leur ensemble — résultant de la combinaison d'une distribution Gamma des moyennes et d'une distribution de Poisson autour de chaque moyenne — est une distribution binomiale négative (nous y reviendrons plus loin) ; $\bar{y}$ et s^2 constituent alors des estimations de l'espérance et de la variance de cette distribution binomiale négative. Différents travaux suggèrent que la distribution binomiale négative est en général en adéquation avec la distribution empirique des nombres d'accidents sur un ensemble de sites (Maher, 1987 ; Abbess *et al*., 1981 ; Elvik, 2007).

On notera que plus les informations apportées par l'échantillon de sites comparables sont vagues (autrement dit plus la variance s^2 est importante par rapport à $\bar{y}$), plus le poids *v* est faible, et donc plus on donne de poids à l'information locale (*x*), dotée alors d'un poids $(1 - v)$ important.

Cas où l'on utilise les résultats d'une modélisation du nombre d'accidents

Il est aussi possible, au lieu de s'appuyer sur un échantillon de sites comparables, d'utiliser un modèle du nombre d'accidents en fonction d'un ensemble de caractéristiques des sites (trafic, type d'aménagement, géométrie, etc.), établi lors de travaux préalables portant sur de grands échantillons

14. Signalons que dans la constitution de l'échantillon de sites comparables, il est extrêmement important de ne pas exclure les sites où aucun accident n'est survenu sur la période d'étude (et plus généralement, aucun critère de sélection de ces sites ne doit être basé sur le nombre d'accidents).

15. En effet, l'équation (3) s'obtient comme *l'estimateur linéaire bayésien empirique* de la moyenne (Robbins, 1985 ; Jarrett, 1994) de la distribution *a posteriori* de *m*, ce qui, dans ce cadre théorique, ne nécessite pas de connaître la forme de la distribution des moyennes de Poisson dans l'échantillon utilisé.

16. L'équation (3) peut être aussi obtenue, dans ce cas, en considérant que cette distribution Gamma constitue la distribution *a priori* de *m*, et en appliquant la démarche bayésienne (voir section suivante) pour déduire, à partir de cette distribution *a priori* et de l'observation *x*, la distribution *a posteriori* de *m*. L'équation (3) donne alors une estimation de la moyenne (espérance mathématique) de cette distribution *a posteriori*, après avoir déduit les paramètres de la distribution binomiale négative des y_j (et donc, également, ceux de la distribution Gamma *a priori*) par la méthode des moments, en se basant sur $\bar{y}$ et s^2. Nous expliciterons ce point dans la section suivante.

de sites[17]. L'application du modèle, lorsque l'on entre les caractéristiques du site d'intérêt, donne alors une estimation ($\hat{\mu}$, valeur prédite par le modèle) de l'espérance $E(y)$ du nombre d'accidents sur des sites présentant de telles caractéristiques, et une estimation de sa variance $Var(y)$ (qui se déduit de $\hat{\mu}$ au moyen de la fonction de variance du modèle). D'une certaine façon, $E(y)$ et $Var(y)$ caractérisent une population virtuelle de sites semblables au site d'intérêt.

Dans le cas général où l'on utilise un modèle s'appuyant sur la distribution binomiale négative[18], la fonction de variance est de la forme $\mu + (1/k)\,\mu^2$, où $(1/k)$ représente un indicateur de surdispersion du modèle[19] (voir par exemple Allain et Brenac, 2001). L'estimation de $Var(y)$ est alors prise égale à $\hat{\mu} + (1/k)\hat{\mu}^2$. La valeur de $1/k$ est obtenue lors de la modélisation.

Dans ce cas, l'estimation bayésienne empirique prend la même forme que dans l'équation (3), en substituant $\widehat{E(y)}$ (c'est-à-dire $\hat{\mu}$) à $\bar{y}$ et en remplaçant s^2 par $\widehat{Var(y)}$ dans l'équation et dans le poids[20] v :

$$\hat{m}_{BE} = v\,\hat{\mu} + (1 - v)\,x \qquad \text{avec } v = \frac{\widehat{E(y)}}{\widehat{Var(y)}} = \frac{\hat{\mu}}{\hat{\mu} + (1/k)\hat{\mu}^2} = \frac{1}{1 + (1/k)\hat{\mu}} \tag{4}$$

<u>*Cas particulier*</u> : Il peut être parfois préférable d'utiliser un modèle s'appuyant plutôt sur une distribution non spécifiée de type « quasi-Poisson » (au lieu de la distribution binomiale négative), dont la fonction de variance est de la forme $\tau\,\mu$, où τ est un indicateur de surdispersion (voir par exemple Brenac et Verne, 2000 ; Allain et Brenac, 2001). Dans ces cas l'estimation bayésienne empirique s'écrit sous une forme semblable mais avec des poids différents :

$$\hat{m}_{BE} = v\,\hat{\mu} + (1 - v)\,x \qquad \text{avec } v = (1/\tau) \text{ et } (1 - v) = 1 - (1/\tau) \tag{5}$$

où $\hat{\mu}$ est l'estimation de l'espérance du nombre d'accidents donnée par le modèle[21]. Le paramètre τ, qui caractérise la surdispersion par rapport à un simple modèle de Poisson, est également obtenu lors de la modélisation.

Portée pratique

Divers travaux suggèrent que les estimations bayésiennes empiriques sont plus proches des moyennes « vraies » — au sens notamment où elles constituent de meilleurs prédicteurs des nombres d'accidents observés sur une période ultérieure (en l'absence de modification des sites et en l'absence d'évolution générale entre les deux périodes, ou après contrôle de cette évolution). Ces travaux montrent aussi que l'utilisation de ces estimations neutralise en général le biais de régression vers la moyenne (voir entre autres Mountain *et al*., 1992*a*, 1992*b* ; Jarrett, 1994 ; Persaud et Lyon, 2007 ; Elvik, 2007), même si dans certains cas le biais semble n'être que partiellement réduit (Elvik, 2007).

---

17. De tels modèles utilisent les techniques des modèles linéaires généralisés, s'appuyant sur la maximisation de la vraisemblance. Pour une présentation, en langue française, des principes et des techniques de modélisation correspondantes, adaptées au cas de l'étude des nombres d'accidents, voir Allain et Brenac, 2001 (voir aussi : Martin, 2000 ; Brenac et Verne, 2000). Pour un ouvrage de référence sur les modèles linéaires généralisés, voir McCullagh et Nelder, 1989.

18. La portée d'un tel modèle dépasse néanmoins le cadre de la distribution binomiale négative. En effet, les recherches sur le pseudo-maximum de vraisemblance (Gouriéroux *et al*., 1984a, 1984b ; Wedderburn, 1974) montrent que l'ajustement d'un modèle linéaire généralisé dépend de la fonction de variance mais pas de la forme précise de la distribution utilisée.

19. Par rapport au modèle de Poisson dans lequel la variance est égale à μ. Le paramètre k est également le paramètre de forme de la distribution Gamma (des moyennes de Poisson) sous-jacente à la distribution binomiale négative utilisée.

20. Le poids v est également parfois donné en fonction de l'espérance $E(\mu)$ et de la variance $Var(\mu)$ de la distribution Gamma des moyennes de Poisson sur la population virtuelle de sites de mêmes caractéristiques (voir par exemple Elvik, 2007), mais on montre sans peine que cette expression, $v = 1/\left(1 + \widehat{Var(\mu)}/\widehat{E(\mu)}\right)$, se ramène à l'expression de ce poids donnée dans l'équation (4). Une expression semblable du poids v peut être aussi rencontrée lorsque l'estimation bayésienne empirique s'appuie sur un échantillon de sites comparables (équation (3)) ; voir par exemple Hauer, 1992, p. 460. On montre qu'elle est équivalente à l'expression du poids v donnée plus haut pour cette équation (3), l'espérance et la variance des moyennes de Poisson sur l'échantillon de sites comparables ayant pour estimations $\bar{y}$ et $(s^2 - \bar{y})$ (Hauer, 1992, p. 461).

21. Il s'agit ici (équation 5) de l'estimateur linéaire bayésien empirique de m. Ce résultat est obtenu en suivant les travaux de Robbins (1983, 1985), appliqués au cas étudié ici (nombres observés distribués selon une loi de Poisson conditionnellement au paramètre d'intérêt, distribution non conditionnelle de type quasi-Poisson). Les informations nécessaires sur $E(y)$ et $Var(y)$ ne sont pas directement tirées d'un échantillon de sites mais sont obtenues par la modélisation. Les estimations de $E(y)$ et $Var(y)$ valent respectivement $\hat{\mu}$ et $\tau\hat{\mu}$.

En consolidant l'estimation de la moyenne par des informations extérieures à x_i et reposant sur des observations plus nombreuses, les estimations bayésiennes empiriques réduisent aussi l'effet des fluctuations aléatoires (qui fragilisent l'estimation). Mais elles n'éliminent pas totalement cet effet.

Dans les pratiques de gestion de la sécurité des routes qui utilisent, dans différents pays, les estimations bayésiennes empiriques, la question des incertitudes autour de ces estimations n'est que rarement abordée. Cela est logique lorsque l'on considère que la distribution des moyennes de Poisson n'est pas précisément connue (au-delà de sa moyenne et de sa variance), ce qui n'empêche pas l'obtention d'une estimation bayésienne empirique (voir équations 3 et 5), mais rend impossible le calcul d'intervalles de crédibilité et la quantification de probabilités *a posteriori*. Néanmoins, il est souvent admis que les moyennes de Poisson sont distribuées selon une loi Gamma, et dans ce cas une application plus complète de la démarche bayésienne est possible, tout en restant dans le cadre bayésien empirique (au sens où les connaissances *a priori* reposent sur des observations), et elle permet de faire de tels calculs. Quelques éléments à ce sujet sont proposés dans la section 6.

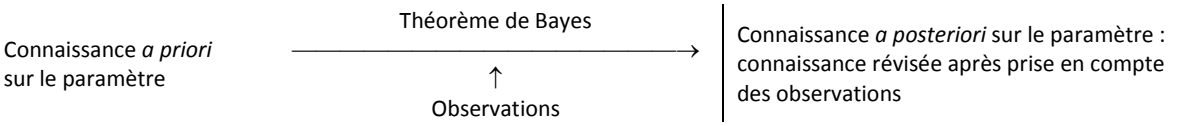
6. Autres développements dans le cadre de l'approche bayésienne

6.1. La démarche bayésienne

La statistique bayésienne admet l'usage de probabilités concernant des paramètres inconnus, tels que, par exemple, la moyenne *m* d'une variable de Poisson, les paramètres (m, σ) d'une distribution de Laplace-Gauss, etc. Ainsi on peut définir des probabilités sur les différentes valeurs possibles du paramètre (ou de l'ensemble de paramètres), de façon à rendre compte de notre incertitude, ou de notre connaissance incertaine, relative à la valeur de ce paramètre. Le paramètre d'intérêt — la moyenne *m* d'une variable de Poisson, par exemple — est alors considéré lui-même comme une instanciation d'une autre variable aléatoire, qui représente en fait l'état de nos connaissances (incertaines) au sujet de ce paramètre.

Avant que les données observées ne soient prises en compte, notre connaissance incertaine au sujet du paramètre est représentée par une *distribution de probabilité a priori* (« *a priori* » signifie ici : avant que les données ne soient prises en compte). Nous y reviendrons au point 6.2.

L'approche bayésienne permet alors, par l'utilisation du théorème de Bayes, de prendre les données en compte et de calculer la *distribution de probabilité a posteriori* du paramètre. Cette distribution *a posteriori* reflète notre connaissance révisée : *ce qui peut être dit au sujet du paramètre étant données les observations*, et étant données les connaissances *a priori*.



Examinons maintenant de façon plus concrète la façon dont les calculs sont conduits dans l'approche bayésienne. Soit x une observation (ou un vecteur d'observations) d'une variable aléatoire X dont la distribution est définie par sa loi de probabilité et son paramètre φ (φ peut être aussi un vecteur de paramètres, si la distribution est définie par plusieurs paramètres). Ce paramètre φ est considéré comme une instanciation d'une variable aléatoire Φ. Dans le cas où la distribution *a priori* de Φ est une distribution continue, définie par sa densité de probabilité a priori $\pi(\varphi)$, le théorème de Bayes[22] donne sa densité de probabilité a posteriori, notée $p(\varphi \mid x)$, sous la forme suivante :

$$p(\varphi \mid x) = \frac{L(x \mid \varphi)\ \ \pi(\varphi)}{\int L(x \mid \varphi)\pi(\varphi)d\varphi} \qquad (6)$$

22. Qui est une conséquence directe des axiomes de la théorie des probabilités (axiomatique de Kolmogorov) et de la définition de la probabilité conditionnelle.

où $L(x \mid \varphi)$ représente la vraisemblance, c'est-à-dire la probabilité d'observer effectivement x pour une valeur φ du paramètre (ou du vecteur de paramètres)[23]. Dans le cas où φ représente un vecteur de plusieurs paramètres, la densité *a priori* est une densité jointe, il en va de même pour la densité *a posteriori*, et l'intégrale du dénominateur est alors une intégrale multiple[24].

Replaçons nous dans le contexte plus particulier de cette note, où la question posée est celle de l'estimation du paramètre m d'une variable de Poisson X pour laquelle nous ne disposons que d'une observation (le nombre d'accidents observé, x). Si l'on considère m comme une instanciation d'une variable aléatoire M, de densité a priori $\pi(m)$, le théorème de Bayes donne accès à sa densité a posteriori :

$$p(m \mid x) = \frac{P(X = x \mid m)\ \pi(m)}{\int P(X = x \mid m)\pi(m)\ dm} \qquad (7)$$

avec $P(X = x \mid m) = \frac{e^{-m}\ m^x}{x!}$. Le choix de la distribution a priori $\pi(m)$ sera traité au point 6.2. L'intégrale au dénominateur est calculée sur l'ensemble du domaine de définition du paramètre.

Une fois déterminée la distribution *a posteriori* de M, sous la forme de sa densité $p(m \mid x)$ ou de sa fonction de répartition $F\left(m \mid x\right) = P(M \leq m \mid x)$, il est possible d'en déduire une estimation ponctuelle, *a posteriori*, de la valeur de m, qui peut être l'espérance (ou éventuellement la médiane) de cette distribution *a posteriori*. Il est possible aussi de déterminer un *intervalle de crédibilité* : par exemple un intervalle dans lequel, connaissant x, il y a 95 % de chances que m se situe[25]. Dans le cas d'un intervalle de crédibilité symétrique à 95 %, par exemple, cet intervalle est défini par les valeurs de m_1 et m_2 telles que $F(m_1 \mid x) = 0{,}025$ et $F(m_2 \mid x) = 0{,}975$. On peut déterminer également, par exemple, pour toute valeur donnée m_0, quelle est la probabilité (x étant connu) que la moyenne m soit supérieure à m_0, en calculant $1 - F(m_0 \mid x)$.

6.2. La distribution a priori *du paramètre d'intérêt*

En statistique bayésienne, le paramètre d'intérêt, par exemple la moyenne m d'une distribution de Poisson, est considéré comme une instanciation d'une variable aléatoire qui représente nos hypothèses sur ce paramètre, sous forme de probabilités subjectives attachées à différentes valeurs qu'il pourrait prendre[26].

La distribution *a priori* peut donc reposer sur des connaissances (points de vue d'experts, synthèse de littérature scientifique, etc.) préexistant aux données et à leur examen.

Pour répondre au cas fréquent où les connaissances *a priori* sont considérées comme insuffisantes voire nulles, les démarches bayésiennes dites « objectives », « non informatives » ou « peu informatives » proposent de prendre des distributions *a priori* du paramètre d'intérêt qui reflètent ce manque de connaissance. Il peut s'agir de distributions non informatives, tentant de représenter l'absence de toute information *a priori* (leur détermination reste cependant débattue et fait l'objet

23. Dans le cas où x représente une observation unique, $L(x \mid \varphi) = \mathrm{P}(X = x \mid \varphi)$. Lorsqu'il s'agit d'un vecteur d'observations $x = (x_1, x_2, \ldots x_n)$, la vraisemblance vaut $L(x \mid \varphi) = \mathrm{P}(X = x_1 \mid \varphi) \times \ldots \times \mathrm{P}(X = x_n \mid \varphi)$.

24. Si $\varphi = (\varphi_1,\varphi_2)$, par exemple, $\pi(\varphi) = \pi(\varphi_1,\varphi_2)$ et de même $p(\varphi \mid x) = p(\varphi_1,\varphi_2 \mid x)$. Et l'intégrale du dénominateur vaut alors $\iint L(x \mid \varphi_1,\varphi_2)\ \pi(\varphi_1,\varphi_2)\ d\varphi_1\ d\varphi_2$.

25. De telles conclusions, portant sur la probabilité que le paramètre d'intérêt se situe dans telle ou telle région de son espace de définition, sont légitimes et correctes dans le cadre bayésien, et répondent aux attentes des utilisateurs. Notons que dans l'analyse statistique plus traditionnelle, de telles interprétations sont erronées : il est faux de dire que l'intervalle de confiance à 95 % des méthodes statistiques conventionnelles contient la valeur du paramètre avec une probabilité de 95 %, de même qu'il est erroné d'affirmer que la *p*-value d'un test représente la probabilité que l'hypothèse nulle soit vraie. Pour davantage d'éléments à ce sujet, voir l'introduction de l'article (Brenac, 2009) et les références qui y sont citées.

26. Les détracteurs de la démarche bayésienne lui reproche d'ailleurs son caractère subjectif. Mais c'est passer un peu vite sur les aspects subjectifs des méthodes statistiques conventionnelles, comme l'arbitraire du choix des seuils d'erreur de première ou de seconde espèce, ou du recours à des hypothèses de répétition indéfinie d'expériences aléatoires virtuelles, nécessaires à la construction d'intervalles de confiance ou de tests statistiques.

d'une littérature scientifique abondante[27]), ou peu informatives (des distributions très « aplaties », par exemple).

Une autre réponse à ce problème de l'insuffisance des connaissances *a priori* est celle que propose la démarche bayésienne empirique, dont le principe consiste à utiliser des données pour apporter une information *a priori* sur le paramètre. Cette démarche est en général considérée comme « hybride », en quelque sorte, entre la démarche bayésienne et la statistique non bayésienne. Elle utilise en effet des méthodes non bayésiennes (comme la maximisation de la vraisemblance) pour estimer une distribution *a priori*. On peut aussi la considérer comme une approximation pratique de la démarche bayésienne (Robert, 2006). Il faut noter que lorsqu'on cherche à établir la distribution *a posteriori* du paramètre d'intérêt sur un site (équation 7), et que ce site est aussi utilisé dans l'échantillon servant à établir la distribution *a priori*, on utilise deux fois l'information issue de ce site : d'une part pour l'ajustement de $\pi(m)$, et d'autre part dans la vraisemblance (égale à $P(X = x \mid m)$ dans l'équation 7). Cela invalide en théorie l'application du théorème de Bayes, mais en pratique cela introduit un biais limité, l'information issue du site d'intérêt ne représentant qu'une très faible part de l'information servant à établir la distribution *a priori* si l'échantillon de sites n'est pas trop réduit (Abbess *et al*., 1981).

Dans la suite de cette section 6 nous nous placerons dans le contexte de **l'approche bayésienne empirique**. En effet, une démarche bayésienne non-informative simple, qui s'appuierait sur une distribution *a priori* non-informative pour le paramètre *m* de la variable de Poisson, ne peut convenir dans le contexte qui nous occupe dans la présente note, puisqu'il s'agit ici, au contraire, d'injecter de *nouvelles informations* (au-delà des observations) pour consolider les estimations des moyennes de Poisson afin de neutraliser le biais de régression vers la moyenne[28].

Nous nous placerons aussi dans l'hypothèse où la distribution *a priori* de la moyenne *m* relève d'une loi Gamma, ce qui permet de compléter l'estimation bayésienne empirique de la moyenne par d'autres résultats, comme des intervalles de crédibilité ou d'autres quantifications des probabilités *a posteriori* répondant aux besoins de l'utilisateur. Le choix de la distribution Gamma a des justifications empiriques, nous l'avons vu dans la section précédente[29].

Pour l'application du théorème de Bayes, la distribution a priori de *m* est alors définie par sa densité *a priori*, celle d'une variable de loi Gamma à deux paramètres. Deux présentations sont possibles — la densité de la distribution Gamma a deux paramètres étant régulièrement présentée sous deux formes différentes (mais strictement équivalentes), ce qui est une source de confusion. Nous présentons les deux formes ci-dessous, mais nous utiliserons ensuite la seconde (équation 9)[30].

Première présentation de la densité d'une distribution Gamma à deux paramètres :

$$\pi(m) = \frac{m^{r-1} e^{\frac{-m}{\theta}}}{\Gamma(r)\theta^r} \qquad (8)$$

où *r* est le paramètre de forme (*shape parameter*) et θ le paramètre d'échelle (*scale parameter*) de la distribution Gamma, et où Γ représente la fonction Gamma.

Seconde présentation de la densité d'une distribution Gamma à deux paramètres :

$$\pi(m) = \frac{\beta^{\alpha} m^{\alpha-1} e^{-\beta m}}{\Gamma(\alpha)} \qquad (9)$$

27. Voir par exemple Kass et Wassermann (1996). Pour un exemple d'application d'une démarche bayésienne non informative ou peu informative dans le champ de la sécurité routière, voir (Brenac, 2009).
28. D'autres démarches seraient néanmoins envisageables dans le cadre de l'approche bayésienne hiérarchique : par exemple la distribution *a priori* de la moyenne *m* de la loi de Poisson peut être considérée comme une distribution Gamma dont les deux paramètres sont eux-mêmes considérés comme relevant de distributions *a priori* non informatives.
29. Il présente aussi un intérêt du point de vue calculatoire, la distribution Gamma étant la loi conjuguée de la distribution de Poisson, ce qui simplifie les calculs dans l'application du théorème de Bayes.
30. La forme de l'équation 9 est celle utilisée par la plupart des auteurs ayant travaillé sur la question traitée dans cette note (voir entre autres : Abbess *et al*., 1981 ; Maher et Mountain, 1988 ; Wright *et al*., 1988 ; Mountain et Fawaz, 1991 ; Hauer, 1997 ; etc.).

où α représente le paramètre de forme (α est égal à r), et où β est aussi d'une certaine façon un paramètre d'échelle (*inverse scale parameter* ; en effet $\beta = 1/\theta$), aussi appelé paramètre d'intensité ou taux[31].

6.3. Calcul de la distribution a posteriori

L'application du théorème de Bayes conduit au calcul suivant, concernant la densité de probabilité *a posteriori* (donc après prise en compte de l'observation x) de la moyenne m de la variable de Poisson, pour une distribution *a priori* Gamma(α,β) :

$$p(m \mid x) = \frac{P(X = x \mid m)\ \pi(m)}{\int P(X = x \mid m)\pi(m)\ dm} = \frac{1}{A}\ \frac{e^{-m}m^{x}}{x!}\ \frac{\beta^{\alpha}m^{\alpha-1}e^{-\beta m}}{\Gamma(\alpha)}$$

$$= \frac{1}{A}\ \frac{\beta^{\alpha}m^{\alpha+x-1}\ e^{-(1+\beta)m}}{x!\ \ \Gamma(\alpha)} = \frac{1}{A}\ \frac{\beta^{\alpha}}{x!\,\Gamma(\alpha)}\ \frac{\Gamma(\alpha+x)}{(\beta+1)^{\alpha+x}}\left[\frac{(\beta+1)^{\alpha+x}\ m^{\alpha+x-1}\ e^{-(1+\beta)m}}{\Gamma(\alpha+x)}\right]$$

où $A = \int P(X = x \mid m)\ \pi(m)dm = \frac{\Gamma(\alpha+x)}{x!\ \Gamma(\alpha)}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{x}$. Le détail du calcul de cette dernière intégrale est donné en annexe 1.

Dans l'équation obtenue ci-dessus, le terme entre crochets correspond à la densité d'une loi Gamma de paramètres ($\alpha + x$) et ($\beta + 1$), et les termes situés en dehors des crochets se simplifient lors de la division par A. Le résultat final est donc :

$$p(m \mid x) = \frac{(\beta+1)^{\alpha+x}\ m^{\alpha+x-1}\ e^{-(\beta+1)m}}{\Gamma(\alpha+x)} = f_{\alpha+x,\beta+1}(m) \qquad (10)$$

où $f_{\alpha+x,\beta+1}(m)$ est la densité de probabilité d'une distribution Gamma de paramètres $\alpha + x$ et $\beta + 1$.

Autrement dit, si m a pour distribution *a priori* une distribution Gamma(α,β), alors sa distribution *a posteriori* — après avoir pris en compte une observation x du nombre d'accidents[32] — est une distribution Gamma($\alpha+x,\beta+1$).

6.3.1. Utilisations de la distribution *a posteriori* de *m*

Les tableurs courants comportent en général une fonction donnant la densité et la fonction de répartition d'une distribution Gamma à deux paramètres (attention toutefois à la forme sous laquelle la distribution est considérée dans le logiciel utilisé ; voir équations 8 et 9).

Si l'on note $F_{\alpha+x,\beta+1}$ la fonction de répartition de la distribution Gamma($\alpha+x,\beta+1$), distribution *a posteriori* de m, les conclusions suivantes, par exemple, peuvent donc être facilement obtenues si l'on connaît α, β et x : la probabilité, après prise en compte de l'observation x, que m soit inférieure ou égale à une valeur quelconque m_0 vaut $P(m \le m_0 \mid x) = F_{\alpha+x,\beta+1}(m_0)$; et en conséquence la probabilité que m soit supérieure à une valeur quelconque m_0 vaut $P(m > m_0 \mid x) = 1 - F_{\alpha+x,\beta+1}(m_0)$.

La fonction réciproque de cette fonction de répartition est également disponible sur la plupart des tableurs. Nous la noterons $F^{-1}_{\alpha+x,\beta+1}$. Elle permet d'obtenir facilement les bornes d'un intervalle de crédibilité symétrique à 95 %, par exemple : $m_1 = F^{-1}_{\alpha+x,\beta+1}(0{,}025)$ et $m_2 = F^{-1}_{\alpha+x,\beta+1}(0{,}975)$. Elle permet d'obtenir aussi n'importe quel centile de la distribution *a posteriori* de m, et entre autres la médiane, $F^{-1}_{\alpha+x,\beta+1}(0{,}5)$. Celle-ci constitue un estimateur possible de m.

---

31. Les notations des différents paramètres (souvent r et θ, ou k et θ pour la forme de l'équation 8, et d'autre part α et β, ou α et λ pour la forme de l'équation 9) varient en outre selon les auteurs ou les logiciels et leurs menus d'aide, ce qui ajoute encore à la confusion. Par exemple Excel, dans sa version française du moins, utilise la forme de l'équation 8 mais note les paramètres α et β, au lieu de r et θ…

32. Dans d'autres types d'application, on peut avoir plusieurs observations, x_1, …, x_i, … x_q pour la même moyenne m. Dans ce cas, on montre sans peine, par des calculs du même genre, que la densité de la distribution *a posteriori* de m connaissant les observations x_1, …, x_i, … x_q est une densité de distribution Gamma($\alpha+\Sigma x_i,\beta+q$).

L'estimateur de la moyenne *m* (*a posteriori*) le plus souvent utilisé est cependant l'espérance de la distribution *a posteriori*. Or l'espérance d'une loi Gamma de paramètres *a* et *b* (avec les notations de l'équation 9) vaut *a*/*b*. Donc l'espérance de la distribution *a posteriori* vaut (α+*x*)/(β+1). Il en résulte que, si α et β sont estimées à partir de données antérieures (échantillon de sites comparables, modèles…), une estimation bayésienne empirique de *m* est :

$$\widehat{m}_{BE} = \frac{\alpha + x}{\beta + 1} = \frac{\alpha}{\beta + 1} + \frac{1}{\beta + 1}\, x \qquad (11)$$

6.3.2. Valeurs de α et β ; autres formulations de l'équation 11

(I) Cas où α et β sont estimés en s'appuyant sur un échantillon de sites comparables au site d'intérêt. Chacun des nombres d'accidents observés y_j dans cet échantillon est distribué selon une distribution de Poisson de moyenne m_j. Si on suppose en outre que ces moyennes sont elles-mêmes distribuées selon une loi Gamma(α,β), la distribution des y_i sur cet ensemble de sites est alors celle d'une loi binomiale négative[33] de paramètres $n = \alpha$ et $p = \beta / (\beta + 1)$. Pour une démonstration voir l'annexe 1. L'espérance et la variance d'une loi binomiale négative de paramètres *n* et *p* valant respectivement $n(1 - p) / p$ et $n(1 - p) / p^2$, ces paramètres peuvent être estimés par la moyenne $\bar{y}$ et la variance s^2 des y_i (méthode des moments[34]) ce qui donne :

$$\frac{n(1-p)}{p} = \bar{y} \qquad \text{et} \qquad \frac{n(1-p)}{p^2} = s^2$$

D'où l'on tire :

$$n = \frac{\bar{y}^2}{(s^2 - \bar{y})} \qquad \text{et} \qquad p = \frac{\bar{y}}{s^2}$$

D'où l'on peut déduire :

$$\alpha = n = \frac{\bar{y}^2}{(s^2 - \bar{y})} \qquad \text{et} \qquad \beta = \frac{p}{(1-p)} = \frac{\bar{y}}{(s^2 - \bar{y})}$$

On peut remarquer qu'en imputant ces valeurs dans l'équation 11, on obtient :

$$\widehat{m}_{BE} = \frac{\bar{y}^2}{s^2} + \frac{(s^2 - \bar{y})}{s^2} x = \left(\frac{\bar{y}}{s^2}\right)\bar{y} + \left(1 - \frac{\bar{y}}{s^2}\right)x$$

On retrouve ici la présentation donnée en section 5 (équation 3) de l'estimation bayésienne empirique, avec les poids $v = \bar{y}/s^2$ et $(1 - v) = 1 - (\bar{y}/s^2)$. Cette estimation reste valable même si les moyennes de Poisson ne sont pas distribuées selon une loi Gamma, mais dans ce dernier cas les autres utilisations de la distribution *a posteriori* ne sont plus légitimes.

(II) Cas où l'on utilise les résultats d'un modèle préexistant reliant les caractéristiques des sites et les nombres d'accidents, s'appuyant sur la distribution binomiale négative. La distribution binomiale négative régissant (par hypothèse) la distribution des nombres d'accidents sur une population virtuelle de sites comparables au site d'intérêt, est alors obtenue par la modélisation. Les estimations de son espérance et de sa variance sont alors $\hat{\mu}$ et $\hat{\mu} + (1/k)\hat{\mu}^2$, où $\hat{\mu}$ est la valeur prédite par le modèle[35], et

33. Dont la distribution de probabilité est donnée par : $P(X = x) = \frac{\Gamma(n+x)}{x!\,\Gamma(n)} p^n (1-p)^x$, où *n* est un réel strictement positif et *p* un réel vérifiant $0 < p < 1$. Il existe cependant d'autres notations et d'autres formes de présentation.

34. On peut aussi, puisque nous nous plaçons dans le cas où la loi de la distribution des m_j est connue, obtenir *n* et *p* (et par conséquent α et β) par la maximisation de la vraisemblance

$$\max_{n,p} \prod_j P_{BN}(Y = y_j)$$

où $P_{BN}(Y = y_j)$ représente la probabilité d'observer y_j accidents selon la distribution binomiale négative de paramètres *n* et *p*. En pratique, cette méthode, plus compliquée, ne semble pas présenter d'avantage décisif par rapport à la méthode des moments, lorsque celle-ci est applicable (Elvik, 2007).

35. La valeur de *k* est aussi obtenue lors de la modélisation (voir l'annexe 2, section A2.3).

les estimations de ses paramètres *n* et *p* sont *k* et $1/(1+\hat{\mu}/k)$. On peut alors en déduire $\alpha = n = k$ et $\beta = p/(1-p) = k/\hat{\mu}$. L'imputation dans l'équation 11 de ces expressions de α et β conduit alors à :

$$\widehat{m}_{BE} = \frac{k}{\frac{k}{\hat{\mu}}+1} + \frac{1}{\frac{k}{\hat{\mu}}+1}\,x = \left(\frac{1}{1+\frac{\hat{\mu}}{k}}\right)\hat{\mu} + \left(\frac{\frac{\hat{\mu}}{k}}{1+\frac{\hat{\mu}}{k}}\right)x = \left(\frac{1}{1+\frac{\hat{\mu}}{k}}\right)\hat{\mu} + \left(1-\frac{1}{1+\frac{\hat{\mu}}{k}}\right)x$$

On retrouve ici l'équation (4) de la section 5, avec les poids $v = \frac{1}{1+(1/k)\hat{\mu}}$ et $1-v = 1-\frac{1}{1+(1/k)\hat{\mu}}$.

Nota : Dans le cas d'un modèle quasi-poissonnien, la distribution des nombres d'accidents sur une population virtuelle de sites comparables au site d'intérêt relève d'une distribution non spécifiée, dont on ne peut estimer que l'espérance et la variance. On ne peut dans ce cas utiliser le théorème de Bayes pour obtenir de façon complète la distribution *a posteriori* de *m*, mais l'estimateur linéaire bayésien empirique (équation 5) reste néanmoins valable.

6.4. Remarque

Les développements que nous venons de présenter dans cette section 6 permettent de tirer un plus grand profit de la démarche bayésienne, par rapport à l'usage aujourd'hui classique d'estimations bayésiennes empiriques ponctuelles de la moyenne (voir notamment le manuel de Hauer, 1997), qui se limite en général au calcul de l'espérance de la distribution *a posteriori*[36].

Mais on reste dans un cas comme dans l'autre dans le cadre de l'approche bayésienne empirique (qui résulte, comme nous l'avons déjà mentionné, d'une sorte d'hybridation entre statistique conventionnelle et statistique bayésienne). D'autres méthodes plus strictement bayésiennes sont possibles, évitant le recours aux données pour l'estimation de distributions *a priori*, dans le cadre de l'approche bayésienne hiérarchique. Ces méthodes, dont la mise en œuvre est plus complexe, ne sont guère utilisées en pratique pour les analyses de la sécurité des réseaux routiers.

7. Exemple d'application sur un jeu de données

Nous reprenons ci-après, à titre purement illustratif et pour donner un exemple de mise en œuvre des calculs, un échantillon de données utilisé dans une recherche publiée en 2000 (Brenac et Verne, 2000), concernant un ensemble de 98 traversées d'agglomération, dans un même département. La plupart de ces agglomérations avaient entre 200 et 3000 habitants, la longueur de traversée était le plus souvent comprise entre 400 m et 2500 m, pour des trafics généralement compris entre 600 et 12000 véhicules par jour (ces fourchettes correspondent aux 5^{e} et 95^{e} centiles des distributions). Les nombres d'accidents corporels, relevés sur la période 1993-1997, variaient de 0 accident (valeur observée sur 27 des 98 traversées) à 33 accidents, mais seuls deux sites avaient un nombre d'accidents supérieur à 16 : l'un avec 28 accidents, l'autre avec 33 accidents.

7.1. Cas où l'information **a priori** *est directement déduite de l'échantillon de sites (I)*

On peut tout d'abord calculer une estimation bayésienne empirique pour chaque site (traversée d'agglomération) en considérant que l'ensemble des 98 sites constitue un échantillon de sites qui lui sont apparentés, et qui peuvent de ce fait nous fournir des informations sur sa sécurité en complément du nombre d'accidents constaté sur ce site. La moyenne du nombre d'accidents observé sur les différents sites de cet échantillon est de 4,949 accidents et sa variance est de 32,15, ce qui permet de calculer les poids v et $(1-v)$ de l'équation (3) : $v = 4{,}949/32{,}15 = 0{,}1539$ et $(1-v) = 1 - 0{,}1539 = 0{,}8461$. Cela permet d'obtenir l'estimation bayésienne empirique, pour chaque site i, en appliquant l'équation (3) : $\widehat{m}_{i\,BE} = 0{,}1539 \times 4{,}949 + 0{,}8461 \times x_i$. Les résultats apparaissent dans le tableau 1.

36. Parmi ces approches qui se limitent à la recherche d'estimations ponctuelles, certaines ont cependant l'avantage de ne pas nécessiter d'hypothèse distributionnelle sur les moyennes *m*, car elles tirent profit des travaux de Robbins (voir notamment Robbins, 1985) sur les estimateurs linéaires bayésiens empiriques.

On remarque que cette estimation bayésienne empirique ne dépend que de constantes et du nombre observé d'accidents. Elle est donc identique pour tous les sites présentant le même nombre d'accidents observé. D'autre part c'est une fonction croissante de ce nombre ; donc le classement des sites selon la valeur de cette estimation est identique à celui basé sur le nombre observé d'accidents. Mais l'enjeu accidentologique est cependant généralement revu à la baisse pour les sites les plus « accidentés » et plutôt à la hausse pour ceux où peu d'accidents ont été observés. En outre, une différence apparaît en cas de sélection des sites à fort enjeu : par rapport à une démarche naïve qui sélectionnerait les sites à partir d'un seuil de nombre d'accidents observé ($x_i \geq 15$, par exemple, ce qui conduit à retenir quatre sites), la sélection réalisée pour un seuil comparable au moyen des estimations bayésiennes empiriques du tableau 1 est différente ($\widehat{m}_{i\,BE} \geq 15$ conduit à ne retenir que deux sites).

Tableau 1. Nombre d'accidents, nombre de sites correspondant, et estimation bayésienne empirique de m_i

Nombre d'accidents observé *x* (1993-1997)	Nombre de sites avec *x* accidents	Estimation bayésienne empirique $\widehat{m}_{i\,BE}$ (I)
0	27	0,76
1	11	1,61
2	4	2,45
3	4	3,30
4	9	4,15
5	6	4,99
6	5	5,84
7	8	6,68
8	3	7,53
9	4	8,38
10	3	9,22
11	5	10,07
12	1	10,91
13	3	11,76
14	1	12,61
15	1	13,45
16	1	14,30
28	1	24,45
33	1	28,68

Si l'on fait en outre l'hypothèse que les moyennes de Poisson sont distribuées selon une distribution Gamma[37], on peut déterminer plus complètement, au-delà des estimations bayésiennes empiriques ponctuelles que nous venons d'évoquer, les distributions *a posteriori* et donc quantifier des intervalles de crédibilité. Pour cela, il faut d'abord calculer α et β à partir de la moyenne et de la variance des nombres d'accidents observés sur les différents sites de l'échantillon (voir plus haut, point 6.3.2) : $\alpha = 4{,}949^2 / (32{,}15 - 4{,}949) = 0{,}9004$ et $\beta = 4{,}949 / (32{,}15 - 4{,}949) = 0{,}1819$. Pour chaque site *i*, la distribution *a posteriori* est une distribution Gamma($\alpha+x_i$,$\beta+1$).

Si l'on calcule son espérance, qui vaut $(\alpha+x_i)/(\beta+1)$, on retrouve l'estimation bayésienne empirique donnée ci-dessus (premier paragraphe du point 7.1).

37. Ce qui implique que la distribution des nombres d'accidents sur l'ensemble des sites suive une distribution binomiale négative. En l'occurrence, sur les données de cet exemple, un test d'ajustement conduirait plutôt à rejeter cette hypothèse ; les éléments fournis ici (colonnes 3 à 6 du tableau 2) sont donc à prendre avec précaution, ils n'ont qu'une visée illustrative.

La fonction réciproque ($F^{-1}_{\alpha+x_i,\beta+1}$) de la fonction de répartition de cette distribution Gamma($\alpha+x_i,\beta+1$), calculée[38] aux points 0,025 et 0,975, donne l'intervalle de crédibilité à 95 % (*a posteriori*) de la moyenne m_i.

La fonction de répartition elle-même ($F_{\alpha+x_i,\beta+1}$) peut permettre par exemple de répondre à la question : quelle est la probabilité (*a posteriori*) que la moyenne m_i soit supérieure à 10 accidents sur la période d'étude[39] ?

Les résultats correspondants pour le jeu de données étudié figurent dans le tableau 2. On remarque que pour la ligne correspondant à 11 accidents observés, l'estimation bayésienne empirique (prise ici égale à l'espérance *a posteriori*) vaut 10,07, mais la probabilité que m_i soit supérieure à 10 est légèrement inférieure à 0,5, ce qui signifie qu'il y a un peu plus de chances que m_i soit inférieure ou égale à 10, que supérieure à 10. Cela s'explique par le fait que la médiane (9,79) ne coïncide pas exactement avec l'espérance et est, elle, légèrement inférieure à 10.

Tableau 2. Quelques points caractéristiques des distributions *a posteriori* des m_i (dans le cas où la distribution *a priori* est une distribution Gamma)

Nombre observé *x*	Quelques points caractéristiques des distributions *a posteriori*			Probabilité (*a posteriori*) que $m_i > 10$
	Espérance [= $\hat{m}_{i\,BE}$ (I)]	Bornes inférieure et supérieure de l'intervalle de crédibilité symétrique à 95 %	Médiane	
0	0,76	0,01 2,94	0,51	0,0000
1	1,61	0,18 4,57	1,34	0,0001
2	2,45	0,49 5,98	2,18	0,0005
3	3,30	0,88 7,29	3,02	0,0023
4	4,15	1,33 8,54	3,87	0,0077
5	4,99	1,81 9,75	4,71	0,0208
6	5,84	2,33 10,93	5,56	0,0471
7	6,68	2,87 12,09	6,40	0,0920
8	7,53	3,43 13,22	7,25	0,1593
9	8,38	4,00 14,34	8,10	0,2486
10	9,22	4,59 15,45	8,94	0,3552
11	10,07	5,19 16,54	9,79	0,4708
12	10,91	5,80 17,63	10,63	0,5856
13	11,76	6,41 18,70	11,48	0,6908
14	12,61	7,04 19,77	12,33	0,7802
15	13,45	7,67 20,83	13,17	0,8512
16	14,30	8,31 21,88	14,02	0,9039
28	24,45	16,36 34,14	24,17	1,0000
33	28,68	19,85 39,11	28,40	1,0000

38. Sous Excel 2010, par exemple, on peut utiliser la fonction LOI.GAMMA.INVERSE.N('probabilité' ; 'alpha' ;'bêta'), en prenant garde à ne pas mettre (β+1), mais 1/(β+1), dans le champ réservé au paramètre « bêta », Excel utilisant la forme de l'équation 8 et non de l'équation 9 pour la distribution Gamma. Dans le champ réservé au paramètre « alpha », par contre, il faut bien mettre ($\alpha+x_i$). Pour le champ 'probabilité', on peut mettre 0,025 pour la borne inférieure de l'intervalle de crédibilité, 0,975 pour la borne supérieure, 0,5 si l'on souhaite obtenir la médiane de la distribution *a posteriori*, etc.

39. Pour cela, sous Excel 2010 par exemple, on peut utiliser la fonction LOI.GAMMA.N('nombre' ; 'alpha' ; 'bêta' ; VRAI), avec les mêmes précautions que celles évoquées dans la note ci-dessus concernant les paramètres alpha et bêta. La spécification 'VRAI' précise que le calcul porte sur la probabilité cumulée (fonction de répartition). Si l'on s'intéresse par exemple au seuil de 10 accidents, z = LOI.GAMMA.N(10 ; ($\alpha+x_i$) ; 1/(β+1) ; VRAI) donne la probabilité *a posteriori* que la moyenne de Poisson m_i soit inférieure ou égale à 10, et (1 – z) donne la probabilité pour que m_i dépasse ce seuil.

7.2. Cas où l'information **a priori** *est déduite d'un modèle (II)*

Sur la base du jeu de données utilisé dans la recherche citée (Brenac et Verne, 2000), il est possible de modéliser le nombre d'accidents attendu, sur la période d'étude de cinq ans, en fonction des variables « population » (POP, en nombre d'habitants), « longueur de la traversée » (LON, en mètres) et « trafic » (TRAF, en véhicules par jour). Pour les besoins de la présente note, nous considérerons le modèle multiplicatif simple ajusté lors de ces travaux antérieurs (Brenac et Verne 2000) sous la forme suivante :

$$\textit{Espérance du nombre d'accidents} = Cste \times POP^{a} \times TRAF^{b} \times LON^{c}$$

où *a*, *b* et *c* sont des coefficients à ajuster par la modélisation. Nous avons ajusté ce modèle aux moyens des techniques des modèles linéaires généralisés (reposant sur la maximisation de la vraisemblance[40]). Pour ce jeu de données, un modèle de type quasi-Poisson est apparu préférable à un modèle s'appuyant sur la distribution binomiale négative[41]. Pour des détails sur l'ajustement de ce modèle, voir la référence citée. Les résultats de cette modélisation permettent de disposer des valeurs prédites par le modèle :

$$\hat{\mu}_i = 5{,}487\ 10^{-5}\ \ POP_i^{\,0{,}4810}\ TRAF_i^{\,0{,}5523}\ LON_i^{\,0{,}4927}$$

où $\hat{\mu}_i$ représente l'estimation de l'espérance du nombre d'accidents sur une population virtuelle de sites présentant les mêmes caractéristiques que le site *i* (POP_i, $TRAF_i$ et LON_i), sur la période 1993-1997. Les résultats donnent aussi accès au paramètre τ de la fonction de variance ($\tau\ \mu_i$), qui pour ce modèle vaut $\tau = 2{,}1446$. Nous donnons davantage d'éléments sur les méthodes de modélisation et leur mise en œuvre en annexe 2.

Il est alors possible de calculer (d'après l'équation 5) les estimations bayésiennes empiriques :

$$\hat{m}_{i\,BE} = \left(\frac{1}{\tau}\right)\hat{\mu}_i + \left(1 - \frac{1}{\tau}\right)\ x_i$$

que nous noterons $\hat{m}_{i\,BE}$ (II) pour les distinguer des estimations $\hat{m}_{i\,BE}$ (I) obtenues au point 7.1.

Dans ce cas, l'estimation bayésienne empirique ne dépend pas seulement du nombre d'accidents observé x_i et de termes ou coefficients indépendants du site, mais elle dépend aussi des caractéristiques de chaque site (population, longueur, trafic). Nous donnons donc dans le tableau 3 (page suivante), pour chaque ligne correspondant à un nombre d'accidents observé *x*, la *moyenne* des estimations $\hat{m}_{i\,BE}$ (II) obtenues pour les sites avec *x* accidents, mise en regard de la valeur $\hat{m}_{i\,BE}$ (I) qui, elle, est commune à l'ensemble des sites avec *x* accidents.

Les estimations bayésiennes empiriques s'appuyant sur le modèle dépendent des caractéristiques de chaque site et ne sont donc plus une fonction nécessairement croissante du nombre d'accidents observé. Cela peut avoir des conséquences sur le classement des sites en fonction de l'enjeu accidentologique. Le tableau 4 (page suivante) présente, pour les dix sites présentant les $\hat{m}_{i\,BE}$ (II) les plus élevés, les rangs de ces sites par ordre d'enjeu accidentologique décroissant, selon que l'on se base sur les estimations bayésiennes empiriques obtenues par la méthode (I) ou la méthode (II). On voit que les classements diffèrent en partie, même si globalement, parmi les dix sites présentant les enjeux accidentologiques les plus élevés au regard de la méthode II, neuf se classent aussi aux dix premiers rangs pour la méthode I.

(Tableaux 3 et 4 : voir page suivante)

40. Ou plus précisément de la quasi-vraisemblance (Wedderburn, 1974 ; Gouriéroux *et al*., 1984*a*, 1984*b*), s'agissant d'un modèle de type quasi-Poisson.

41. Le lecteur trouvera des justifications dans les références suivantes : Brenac et Verne, 2000 ; Allain et Brenac, 2001. En général, dans les démarches de modélisation, on procède d'abord à l'ajustement d'un modèle de Poisson, qui apparaît surdispersé. Un examen graphique (ou une régression) de la relation entre les carrés des résidus standardisés et les valeurs prédites permet alors de suggérer une autre forme de modèle plus appropriée et notamment de s'orienter plutôt vers un modèle de type binomial négatif ou de type quasi-Poisson (voir les références citées).

Tableau 3. Comparaison des estimations bayésiennes empiriques obtenues par les méthodes (I) et (II)

Nombre d'accidents observé *x* (1993-1997)	Estimation bayésienne empirique $\hat{m}_{i\,BE}$ (I) (commune aux sites avec *x* accidents)	Moyenne des estimations bayésiennes empiriques $\hat{m}_{i\,BE}$ (II) pour les sites avec *x* accidents
0 (27 sites)	0,76	0,81
1 (11 sites)	1,61	1,65
2 (4 sites)	2,45	2,31
3 (4 sites)	3,30	3,19
4 (9 sites)	4,15	3,60
5 (6 sites)	4,99	5,26
6 (5 sites)	5,84	5,55
7 (8 sites)	6,68	5,35
8 (3 sites)	7,53	9,05
9 (4 sites)	8,38	8,04
10 (3 sites)	9,22	9,11
11 (5 sites)	10,07	11,75
12 (1 site)	10,91	11,36
13 (3 sites)	11,76	12,02
14 (1 site)	12,61	13,48
15 (1 site)	13,45	15,61
16 (1 site)	14,30	14,27
28 (1 site)	24,45	24,45
33 (1 site)	28,68	27,83

Tableau 4. Différences dans les classements par ordre d'enjeu accidentologique décroissant (pour les dix sites présentant les estimations bayésiennes empiriques les plus élevées selon la méthode II)

Nombre d'accidents observé	$\hat{m}_{i\,BE}$ (I)		$\hat{m}_{i\,BE}$ (II)	
	Valeur	Rang (sur 98)	Valeur	Rang (sur 98)
33	28,68	1	27,83	1
28	24,45	2	24,45	2
15	13,45	4	15,61	3
11	10,07	10	14,56	4
16	14,30	3	14,27	5
13	11,76	6	14,09	6
14	12,61	5	13,48	7
8	7,53	22	13,31	8
13	11,76	6	12,17	9
11	10,07	10	12,06	10
…	…	…	…	…

Le modèle utilisé s'appuie sur une distribution non spécifiée (quasi-poissonnienne), qui permet le calcul d'une estimation bayésienne empirique (l'estimation de l'espérance de la distribution *a posteriori* de m_i connaissant x_i) mais ne permet pas de connaître précisément la forme de la distribution *a posteriori* ; des éléments d'information complémentaires sur cette distribution, du type de ceux rassemblés dans le tableau 2, ne peuvent donc être obtenus dans ce cas. Ce choix de modèle est néanmoins justifié pour ce jeu de données.

Nota : Dans d'autres situations, fréquentes, l'ajustement d'un modèle s'appuyant sur la distribution binomiale négative est préférable à celui d'un modèle de type quasi-Poisson (des éléments de méthode pour en décider figurent dans la référence : Allain et Brenac, 2001). La modélisation donnerait alors accès à des valeurs prédites $\hat{\mu}_i$, et au paramètre k de la fonction de variance [$\mu_i + (1/k)\mu_i^2$]. Les estimations bayésiennes empiriques valent dans ce cas :

$$\widehat{m}_{i_{BE}} = \left(\frac{1}{1+\frac{\hat{\mu}_i}{k}}\right)\hat{\mu}_i + \left(1-\frac{1}{1+\frac{\hat{\mu}_i}{k}}\right)x_i$$

Dans un tel cas, pour un site *i*, la distribution *a priori* de la moyenne m_i est une distribution Gamma de paramètres $\alpha = k$ et $\beta = k/\hat{\mu}_i$ et par conséquent la distribution *a posteriori* de la moyenne m_i est une distribution Gamma de paramètres $(k + x_i)$ et $((k/\hat{\mu}_i) + 1)$. Dans une telle situation, des éléments comparables à ceux présentés dans le tableau 2 pourraient donc être calculés pour chaque site *i*.

8. Cas où l'on s'intéresse plutôt au taux d'accidents

Si l'échantillon soumis à l'analyse est un ensemble de sections de route relativement hétérogènes du point de vue de leur longueur et du trafic qu'elles supportent (et donc sur lesquelles l'exposition, en véhicules kilomètres par exemple, est aussi assez variable), il est important de prendre en compte l'exposition dans l'analyse, et d'estimer les taux d'accidents par unité d'exposition[42]. Deux situations doivent cependant être distinguées : celle où l'on peut s'appuyer sur un modèle du nombre d'accidents, et celle où l'on s'appuie seulement sur un échantillon de sections.

8.1. Situation où l'on s'appuie sur un modèle du nombre d'accidents

De tels modèles intègrent généralement, parmi leurs variables explicatives, les données relatives à l'exposition, telles que la longueur de la section et le volume de trafic. L'estimation bayésienne empirique de la moyenne du *nombre d'accidents* sur chaque site d'intérêt s'appuie alors sur l'information *a priori* que donne le modèle concernant une population virtuelle de sites comparables de même longueur, supportant le même volume de trafic, et présentant en outre un certain nombre d'autres caractéristiques communes. Le rôle de l'exposition est donc bien contrôlé[43].

Dans ce cas, même si l'on s'intéresse *in fine* au taux d'accidents (que nous noterons λ_i, et défini par $\lambda_i = m_i / t_i$ où t_i représente l'exposition sur le site *i*, en véhicules kilomètres par exemple, considérée comme connue et non aléatoire), la démarche consiste d'abord à obtenir des estimations bayésiennes empiriques des moyennes m_i des nombres d'accidents, comme il a été décrit plus haut (voir sections 5.1, 6.3.2.II et 7.2). Ces estimations donnent les espérances des distributions *a posteriori* des moyennes m_i. L'espérance de la distribution *a posteriori* du taux d'accidents se déduit de l'espérance de la distribution de la moyenne, en appliquant la propriété élémentaire de l'espérance d'un produit d'une variable aléatoire par un scalaire. Et de ce fait :

$$\hat{\lambda}_{i\,BE} = \frac{1}{t_i}\,\widehat{m}_{i\,BE} \qquad (12)$$

Dans le cas où l'on peut faire en outre l'hypothèse que les distributions *a priori* des moyennes relèvent de distributions Gamma, la distribution *a posteriori* du taux d'accidents peut être déduite de la distribution *a posteriori* de la moyenne, en utilisant un résultat classique portant sur la loi d'une fonction d'une variable aléatoire (voir par exemple Saporta, 1990, p. 23). Si la moyenne m_i a pour distribution *a posteriori* une loi Gamma(α+x_i,β+1), alors la distribution *a posteriori* du taux ($\lambda_i = m_i/t_i$)

42. Les développements qui suivent concernent principalement le cas de taux d'accidents sur des sections de route (les taux observés sont alors les rapports des nombres d'accidents aux véhicules kilomètres), mais peuvent s'appliquer plus généralement, *mutatis mutandis*, à diverses autres situations (par exemple, cas où le taux observé représente le rapport du nombre d'accidents au trafic total entrant sur une intersection, dans le cas où l'on s'intéresse à des carrefours…).

43. Même si, dans de tels modèles, le trafic intervient à une puissance généralement inférieure à 1, ce qui englobe à la fois le rôle de l'exposition et un rôle modérateur des trafics élevés sur le risque, traduisant notamment des effets liés aux vitesses pratiquées.

relève d'une loi Gamma($\alpha+x_i$, $(\beta+1)t_i$), en particulier[44]. Sur cette base, on peut vérifier notamment que l'estimation bayésienne empirique du taux donnée par l'espérance de sa distribution *a posteriori* vaut :

$$\frac{\alpha + x_i}{(\beta+1)t_i} = \frac{1}{t_i}\left(\frac{\alpha + x_i}{\beta + 1}\right) = \frac{1}{t_i}\widehat{m}_{i\,BE}$$

Il est en outre possible de calculer divers points caractéristiques de la distribution *a posteriori* du taux, comme la médiane ou tel ou tel centile, ou encore :

- la probabilité que ce taux soit inférieur ou égal à une valeur λ_0 donnée : $F_{\alpha+x_i,(\beta+1)t_i}(\lambda_0)$;
- les bornes d'un intervalle de crédibilité à 95 % (symétrique), permettant d'identifier une région où le taux a 95 % de chances de se trouver ; un tel intervalle est défini par ses bornes inférieure et supérieure : $F^{-1}_{\alpha+x_i,(\beta+1)t_i}(0{,}025)$ et $F^{-1}_{\alpha+x_i,(\beta+1)t_i}(0{,}975)$.

8.2. Situation où l'on s'appuie seulement sur un échantillon de sections

Dans ce cas, si l'on applique les méthodes décrites aux sections précédentes, la distribution *a priori* des moyennes est la même pour l'ensemble des sites, et est déduite (en tout ou partie[45]) à partir d'un échantillon de sites ; mais on peut considérer que cette distribution des moyennes n'a guère de sens pour un ensemble de sections de longueur et de trafic hétérogènes puisqu'elle dépend alors en partie des hasards du découpage des sections, ou de la répartition des trafics dans cet échantillon de sites. Raisonner sur les taux d'accidents est alors préférable.

Si l'on ne dispose pas de modèle mais seulement d'un échantillon de sections pour lesquelles on connaît simplement le nombre d'accidents x_i, la longueur de section, et le volume de trafic (ou bien, plus généralement, le nombre d'accidents et la donnée d'exposition — cela peut être le trafic total entrant sur l'intersection, pour un échantillon de carrefours, par exemple), deux possibilités se présentent :

8.2.1. On peut ajuster un modèle simple du *nombre d'accidents* en fonction des seules données d'exposition et se retrouver dans le cas traité au point 8.1 ci-dessus, la première étape étant alors l'obtention de l'estimation bayésienne empirique de chaque moyenne m_i. L'estimation bayésienne empirique du taux est alors aisément déduite de celle de la moyenne de Poisson (équation 12), notamment.

8.2.2. On peut faire porter l'analyse bayésienne empirique non pas sur la valeur m de la moyenne d'une variable de Poisson, mais sur le taux λ d'une variable de Poisson de moyenne λt (= m) où t représente l'exposition, supposée connue et non aléatoire. Cette option n'est pas préférable ni plus commode à mettre en œuvre, mais nous l'explicitons ci-dessous pour l'information du lecteur.

Nous supposerons ici que la distribution *a priori* du taux λ est une distribution Gamma(α,β).

Si $\pi(\lambda)$ représente la densité de probabilité *a priori* du taux, et x le nombre d'accidents observés sur le site étudié (réalisation d'une variable de Poisson de moyenne λt), la densité *a posteriori* du taux λ s'écrit alors :

$$p(\lambda \mid x) = \frac{P(X = x \mid \lambda)\ \pi(\lambda)}{\int P(X = x \mid \lambda)\pi(\lambda)\ d\lambda} = \frac{\dfrac{e^{-\lambda t}(\lambda t)^x}{x!}\dfrac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)}}{\int \dfrac{e^{-\lambda t}(\lambda t)^x}{x!}\dfrac{\beta^\alpha \lambda^{\alpha-1} e^{-\beta\lambda}}{\Gamma(\alpha)} d\lambda} \qquad (13)$$

Par des calculs similaires à ceux présentés en annexe 1, on montre sans peine que l'intégrale du dénominateur vaut :

$$\frac{\Gamma(\alpha + x)}{x!\ \Gamma(\alpha)}\ \left(\frac{\beta}{\beta + t}\right)^\alpha \left(\frac{t}{\beta + t}\right)^x$$

44. Plus généralement le produit d'une variable aléatoire de distribution Gamma(a,b), avec les notations de l'équation 9, par un réel positif ($1/t$) est distribué selon une loi Gamma(a,bt).

45. Dans sa totalité dans l'hypothèse d'une distribution Gamma, ou, dans le cas général, seulement sous la forme de ses moments d'ordre 1 et 2 : espérance et variance.

ce qui permet d'obtenir, après simplification dans l'équation 13, la forme suivante :

$$p(\lambda \mid x) = \frac{(\beta + t)^{\alpha+x} \lambda^{\alpha+x-1} e^{-(\beta+t)\lambda}}{\Gamma(\alpha + x)} \qquad (14)$$

ce qui est la densité de probabilité d'une loi Gamma(α+x,β+t). Autrement dit, si la distribution *a priori* du *taux* (λ) est une distribution Gamma de paramètres α et β, alors sa distribution *a posteriori*, c'est-à-dire après prise en compte de l'observation x, est une distribution Gamma de paramètres α+x et β+t, où t représente l'exposition (connue et non aléatoire)[46]. Une fois que les paramètres α et β de la distribution *a priori* ont été estimés (ce qui soulève ici des difficultés, nous y revenons ci-après), il est donc possible d'obtenir pour chaque site i la distribution *a posteriori* du taux d'accident [de loi Gamma(α+x_i,β+t_i)] ainsi qu'une estimation bayésienne empirique du taux, en prenant l'espérance de cette distribution *a posteriori* , c'est-à-dire ici : $\hat{\lambda}_{i\,BE} = (\alpha + x_i)/(\beta + t_i)$.

Nous nous situons ici dans le cas où les paramètres α et β de la distribution *a priori* du taux sont estimés à partir des données d'un échantillon de sites. L'estimation s'appuyant sur un échantillon de sites soulève davantage de difficultés lorsque l'analyse porte sur le taux, par rapport aux situations où l'on s'intéresse aux nombres d'accidents. En effet, supposons que l'on souhaite obtenir des estimations bayésiennes empiriques des taux λ_i sur un ensemble de sites, pour lesquels nous noterons les nombres d'accidents observés x_i et les expositions t_i. Plaçons-nous en outre dans le cas courant où l'on cherche à estimer α et β sur la base de cet échantillon lui-même et non à partir d'un autre échantillon indépendant. La probabilité d'observer x_i accidents sur un site i, résultant du mélange Gamma-Poisson (x_i distribué selon une loi de Poisson de moyenne $\lambda_i\, t_i$ avec t_i connu, et λ_i distribué selon une loi Gamma de paramètres α et β) relève d'une distribution binomiale négative de paramètres $n = \alpha$ et $p = \beta / (\beta + t_i)$:

$$P(X_i = x_i) = \frac{\Gamma(\alpha + x_i)}{x_i!\ \Gamma(\alpha)} \left(\frac{\beta}{\beta + t_i}\right)^{\alpha} \left(\frac{t_i}{\beta + t_i}\right)^{x_i}$$

Ce résultat peut être aisément obtenu par un calcul similaire à celui présenté en annexe 1. L'ensemble des x_i ne relèvent donc pas d'une seule distribution binomiale négative (dont on pourrait alors estimer les paramètres par la méthode des moments, c'est-à-dire sur la base de la moyenne et de la variance des x_i), contrairement à la situation traitée dans les sections précédentes, mais ils relèvent d'autant de distributions binomiales négatives qu'il y a de valeurs différentes de l'exposition t_i, puisque le second paramètre (p) dépend de t_i. La méthode recommandable pour estimer α et β est alors la maximisation de la vraisemblance[47] :

$$\max_{\alpha,\beta} \prod_i \frac{\Gamma(\alpha + x_i)}{x_i!\ \Gamma(\alpha)} \left(\frac{\beta}{\beta + t_i}\right)^{\alpha} \left(\frac{t_i}{\beta + t_i}\right)^{x_i}$$

Une telle opération nécessite de recourir à des algorithmes itératifs comparables à ceux utilisés par les logiciels de statistique pour l'ajustement de modèles linéaires généralisés (de type Newton-Raphson ou équivalent) et n'est pas réalisable sur un tableur ordinaire. Pour cette raison, cette méthode n'apporte pas d'avantage pratique par rapport à celle évoquée en 8.2.1 ; en outre, elle ne tire pas parti entièrement des informations contenues dans l'échantillon concernant les liens entre les variables liées à l'exposition (trafic, longueur…) et les phénomènes d'accidents. La méthode évoquée en 8.2.1 nous paraît donc préférable.

46. *Nota* : Ce résultat peut apparaître à première vue comme discordant par rapport à la forme de la distribution *a posteriori* du taux d'accident donnée au dernier paragraphe de la section 8.1 : Gamma(α+x,(β+1)t) = Gamma(α+x,βt+t) ; mais en réalité dans ce dernier cas, les hypothèses de base sont différentes puisque α et β caractérisent la distribution *a priori* de la moyenne et non celle du taux (cette distribution de la moyenne de Poisson correspondrait à une distribution *a priori* du taux de paramètres α et βt). Il n'y a donc pas d'incohérence.

47. D'autres méthodes, reposant sur des approximations, ont été étudiées, mais elles présentent de graves défauts de fiabilité, surtout lorsque les effectifs d'accidents par site sont assez faibles (moins de 5 à 10 accidents).

9. Retour sur quelques choix méthodologiques de base : sectionnement, critères de classement

Nous n'évoquerons ici que brièvement ces différents aspects.

Sur la question du sectionnement du réseau

Concernant la recherche de sections dangereuses sur un réseau, il est aujourd'hui reconnu que les méthodes consistant à appliquer des tests statistiques (ou autres critères statistiques) sur les nombres d'accidents pour différentes positions d'une « fenêtre mobile », déplacée progressivement le long d'une infrastructure, sont à déconseiller, dans la mesure où elles conduisent à identifier un grand nombre de « faux positifs » (c'est un problème bien connu lié à la répétition de tests en grand nombre) (Elvik, 2007). La recherche doit donc opérer sur un ensemble de sites clairement définis à l'avance, qui peuvent être énumérés. Si les investigations portent sur des sites particuliers, comme des intersections où des virages, il n'y a pas de difficultés particulières.

Si l'on s'intéresse à des sections de route, la question du découpage et de la longueur de section se pose. Les pratiques à ce sujet sont d'une grande diversité. Dans la mesure où les données d'exposition (trafic) sont contrôlées, il n'y a pas d'objection à considérer des sections de longueurs inégales, du moment que la longueur et le trafic sont pris en compte dans l'analyse. Il paraît par contre discutable de considérer des sections à l'intérieur desquelles le trafic varie notablement, puisque le trafic est toujours la principale variable prédictive du nombre d'accidents, dans le cadre d'une relation souvent non linéaire. En outre, un changement notable de volume de trafic entre deux segments voisins va souvent de pair avec d'autres changements, dans la composition du trafic, les types d'usage de l'infrastructure, l'environnement (et le niveau d'aménagement, dans le cas d'infrastructures non autoroutières du moins) et dans ce cas les enjeux, les niveaux de risque comme les logiques de traitement ont peu de chances d'être les mêmes pour ces deux segments.

Des considérations pratiques conduisent souvent à retenir un découpage préétabli, lié aux sections utilisées pour la gestion des comptages de trafic, par exemple, éventuellement combinées aux données d'inventaire ou de gestion des routes. De telles pratiques paraissent tout-à-fait acceptables. Ainsi par exemple Persaud (1990), dans le cadre d'une investigation sur l'identification de zones dangereuses sur les routes de l'Ontario, étudie 2250 sections de routes, sur un réseau d'une longueur totale de 16 000 km, la longueur moyenne des sections variant de 4 km (routes de classe I : *freeways*, *highways*) à 14 km (routes de classe III : routes secondaires ; cela peut sembler long mais on est dans le cas d'un état canadien avec de grands espaces peu peuplés). Des estimations bayésiennes empiriques ont été ensuite calculées pour ces sections[48].

Si les méthodes d'estimation bayésienne empirique font usage des prédictions de modèles d'accidents, le découpage doit être cohérent avec ces modèles : les sections doivent être homogènes du point de vue des variables explicatives utilisées dans ces modèles. Les modèles déjà établis, visant souvent la mise en évidence des effets de nombreuses variables, utilisent des découpages assez fins assurant l'homogénéité vis-à-vis des différentes variables testées (par exemple, longueur moyenne de section de 750 m chez Chang, 2005, sur des infrastructures autoroutières, de 450 m pour Greibe, 2003, concernant des voies urbaines). Certains modèles du Transport Research Laboratory portent sur des sections de route, dites *'pure' links*, ne contenant aucune intersection (donc délimitées par deux intersections consécutives, quelle que soit l'importance de ces intersections) (Maher et Summersgill, 1996).

Dans le cadre de procédures de routine de gestion de la sécurité des routes, destinées à être répétées régulièrement sur un réseau d'infrastructures, il est peut-être préférable de recourir à des modèles très simples, *ad hoc*, prenant en compte le volume de trafic (sensiblement homogène sur la section), la longueur de la section, le type de route (autoroute, autres routes à deux chaussées en continu, routes à

48. Un autre choix consiste à utiliser des sections courtes de longueur fixe (1 km par exemple ; Elvik, 2007). Sur une proportion souvent importante de ces sections, aucun accident n'est relevé (il faut bien entendu conserver ces sections dans l'échantillon). Des procédures ont été développées pour regrouper ensuite des sections proches dont les estimations bayésiennes empiriques sont élevées. Mais ces procédures (voir notamment Hauer *et al*., 2002) s'appuyant sur des méthodes de fenêtres mobiles de différentes tailles et mettant en œuvre de façon répétée un critère de précision statistique nous semblent susceptibles de produire des agrégations en surnombre, et resteraient à évaluer (en outre, la justification de l'usage du critère de précision statistique et son estimation prêtent à discussion : voir annexe 4).

une chaussée avec ou sans créneaux de dépassement, par exemple), sans introduire de nombreuses variables explicatives, et qui soient ajustés sur le réseau sous examen. Au-delà du fait que cela introduit moins de contraintes sur le découpage et la longueur des sections, cela peut limiter le besoin en expertise, et éviter le problème d'actualisation qui se pose souvent pour les modèles plus élaborés réalisés dans le cadre d'activités de recherche (qui ne peuvent être ré-ajustés régulièrement du fait de la lourdeur du travail correspondant).

Critères de classement

(*A*) Le critère le plus utilisé est celui du nombre d'accidents (sur des sites ponctuels), ou de façon équivalente celui de la densité d'accidents par kilomètre de route (sur des sections de route). Pour éviter une évaluation biaisée, il doit être pris en compte sous la forme de l'estimation bayésienne empirique. Ce critère permet de classer les sites en fonction de l'enjeu qu'ils représentent en termes d'accidentalité, cet enjeu étant approximativement indicatif de l'ampleur des problèmes d'insécurité, pour la collectivité, attachés à ces sites ou à ces kilomètres de voirie.

(*B*) Le taux d'accidents (par véhicule kilomètre, par exemple, s'agissant de sections de route) donne une bonne indication du niveau de risque individuel encourus par les usagers fréquentant ces sites (là encore il est important de s'appuyer sur l'estimation bayésienne empirique du taux). Sur un site à fort taux, si le trafic est faible, l'enjeu pour la collectivité peut être faible, mais le risque reste élevé pour l'individu. Cela peut aussi justifier l'examen d'un tel site ou d'une telle section.

D'autres critères tentent d'approcher les potentialités de gains. Un enjeu accidentologique fort concentré dans l'espace (critère *A*) peut laisser présager des gains élevés, mais ce n'est pas toujours le cas, notamment lorsque l'accumulation d'accidents s'explique par un fort trafic sur un site où le taux est très faible. Un site avec un taux élevé (critère *B*) mais où le trafic est faible laisse présager un « retour sur investissement » faible pour la puissance publique, puisque l'enjeu y est faible et le gain limité de ce fait, même si on réduit fortement le taux ; certes, la sécurité est due à tout citoyen, indépendamment de toute considération de nombre, mais il reste que les mêmes moyens déployés ailleurs permettraient d'éviter davantage de victimes, et que les mesures de sécurité ne sont en pratique jamais mises en œuvre hors de toute considération de limite budgétaire. Des critères visant à rendre compte des potentialités de gain ont donc été avancés :

(*C*) Un critère double, fondé sur l'existence d'un enjeu accidentologique élevé et d'un taux élevé (sur la base des estimations bayésiennes empiriques). Il est difficilement contestable que ce critère double identifie bien des priorités de premier ordre pour des investigations ultérieures, car il répond aux préoccupations majeures évoquées plus haut : un coût d'insécurité élevé pour la collectivité, un risque élevé pour l'individu. Reste qu'un nombre d'accidents élevé et un taux élevé ne garantissent pas nécessairement une possibilité de traitement à un coût qui soit proportionné aux gains.

(*D*) Le critère de l'écart au « comportement » moyen de sites comparables (écart à la prédiction d'un modèle par exemple), ou de l'écart par rapport à tel ou tel centile de la distribution *a posteriori* de l'espérance du nombre d'accidents (ou du taux d'accidents) sur des sites similaires. Là encore, un fort écart au comportement moyen ne signifie pas pour autant que combler cet écart soit possible, ou puisse se faire à faible coût, ou à un coût proportionné aux gains. Un niveau d'accidentalité très élevé dans une traversée d'agglomération (par rapport à d'autres traversées présentant des caractéristiques semblables — population, trafic…) peut être lié par exemple à une autre caractéristique difficilement modifiable (traversée située en descente, problème de visibilité lié au bâti et à une inflexion du tracé, par exemple). Certains auteurs se montrent critiques vis-à-vis de tel critères, notant la relative fragilité des prédictions des modèles (voir par exemple Maher et Mountain, 1988).

Nota : S'agissant des quelques critères qui viennent d'être évoqués, une analyse de leurs performances a été réalisée (Elvik, 2007) en s'appuyant sur la comparaison, pour environ 20 000 sections de routes d'1 km de long en Norvège, entre d'une part une période de référence 1997-2000, où les sections les plus dangereuses ont été identifiées (celles situées au-delà d'un centile fixé de la distribution des sections selon ce critère), et d'autre part la période consécutive 2001-2004, servant d'étalon pour voir si ces sections étaient bien effectivement dangereuses selon le même critère sur cette période ultérieure.

Le tableau 5 ci-après est extrait de cette analyse (Elvik, 2007). Le premier critère, *accident count*, repose sur le nombre brut d'accidents corporels constatés (et non sur une estimation bayésienne empirique) sur la période 1997-2000 : les 1 % (ou bien les 2,5 % ou encore les 5 %) des sections les plus mal classées sur ce critère sont identifiées comme dangereuses. Le deuxième critère, *accident rate*, porte sur le taux d'accidents par véhicule kilomètre (il s'agit ici aussi d'un taux brut et non

d'une estimation bayésienne empirique) et les sections les plus mal classées sur ce critère, sur la période 1997-2000, sont alors identifiées comme les sections dangereuses. Le troisième critère (*accident rate and count*) identifie comme sections dangereuses les plus mal classées selon le premier critère (accident count), sur la période 1997-2000, à condition qu'elles aient aussi un taux d'accidents supérieur à la moyenne (sur l'ensemble des 20 000 sections). Le quatrième critère (*EB-estimate of accidents*) repose sur une estimation bayésienne empirique du nombre d'accidents, s'appuyant sur un modèle, sur la période 1997-2000. Le cinquième critère (*EB-dispersion criterion – potential accident reduction*) repose sur une idée de « potentiel de réduction d'accident » : c'est la différence entre l'estimation bayésienne empirique du taux et la prédiction du modèle, sur la période 1997-2000 ; les sections les plus mal classées selon ce critère (les 5 %, par exemple, de sections avec les écarts les plus importants) sont alors considérées comme les sections dangereuses.

Pour un critère donné, par exemple *accident count*, les sections dangereuses ainsi identifiées qui ne sont plus dangereuses au sens du même critère pour la période 2001-2004 sont considérées comme des « faux positifs » ; celles non identifiées comme dangereuses et qui sont dangereuses pour 2001-2004, au sens du même critère, sont considérées comme des « faux négatifs ». Celles identifiées comme dangereuses pour 1997-2000 et qui restent identifiées comme telles pour 2001-2004 sont les « positifs bien identifiés » (*correct positives*) et celles qui ne sont dangereuses selon ce critère ni pour 1997-2000 ni pour 2001-2004 sont les « négatifs bien identifiés » (*correct negatives*). La sensibilité est la proportion de positifs bien identifiés parmi l'ensemble des sections dangereuses pour la période étalon 2001-2004 : sensibilité = positifs bien identifiés / (positifs bien identifiés + faux négatifs). La spécificité est la proportion de négatifs bien identifiés parmi les sections non dangereuses sur la période étalon 2001-2004 : spécificité = négatifs bien identifiés / (négatifs bien identifiés + faux positifs).

L'hypothèse faite est que sur la quasi-totalité de ce réseau il n'y a pas eu de modification notable (aménagement) entre les deux périodes. D'autre part la tendance globale (*trend*) ne joue pas car on raisonne en pourcentage de sites les plus dangereux. L'existence de faux positifs et de faux négatifs est normale car la période étalon est aussi marquée par les fluctuations aléatoires. Néanmoins, leur proportion est indicative de la performance des différents critères. Les résultats du tableau 5 tendent à montrer que dans cette étude :

- le critère de l'estimation bayésienne du nombre d'accidents est celui qui permet d'obtenir la meilleure sensibilité et la meilleure spécificité ;
- les critères s'appuyant sur les nombres ou taux bruts d'accidents ont une mauvaise sensibilité (c'est encore plus vrai pour le critère s'appuyant sur le seul taux d'accidents) ;
- le critère *EB dispersion criterion (potential accident reduction)* a également une mauvaise sensibilité, ce qui va dans le sens des réserves émises par d'autres auteurs sur ce type d'approche.

Tableau 5. Extrait de (Elvik, 2007), p. 34.

Identification criterion	Correct negatives	Correct positives	False negatives	False positives	Sensitivity	Specificity
Top 1 % of distribution						
Accident count	19272	134	109	108	0.551	0.994
Accident rate	19232	16	188	187	0.078	0.990
Accident rate and count	19340	86	94	103	0.478	0.995
EB-estimate of accidents	19378	130	53	62	0.710	0.997
EB dispersion criterion	19311	62	121	129	0.339	0.993
Top 2.5 % of distribution						
Accident count	18788	285	262	288	0.521	0.985
Accident rate	18726	53	418	426	0.113	0.978
Accident rate and count	18928	186	236	273	0.441	0.986
EB-estimate of accidents	18981	338	152	152	0.690	0.992
EB dispersion criterion	19070	105	195	253	0.350	0.987
Top 5 % of distribution						
Accident count	18065	464	526	568	0.469	0.970
Accident rate	17838	144	805	836	0.152	0.955
Accident rate and count	18308	307	474	534	0.393	0.972
EB-estimate of accidents	18429	692	235	267	0.746	0.986
EB dispersion criterion	18989	136	219	279	0.383	0.986

Source: TØI-report 883/1007

Enfin, tous les accidents ne sont pas équivalents en termes de gravité, de nombres de victimes et *in fine* de conséquences et de coûts pour les individus et la collectivité. Des coûts d'insécurité, monétarisant à des niveaux différents les décès, les blessures graves et les blessures légères, sont donc

parfois pris en compte dans les critères de classement. Les critères reposant sur les coûts d'insécurité sont généralement les suivants :

(*E*) Somme des coûts d'insécurité (liés à l'ensemble des accidents de différents niveaux de gravité) sur le site d'intérêt (ou par kilomètre de route si l'on s'intéresse à des sections).

(*F*) Somme des coûts d'insécurité rapportée aux véhicules kilomètres (par exemple).

(*G*) Critère d'écart entre la somme des coûts d'insécurité sur le site et sa valeur moyenne sur des sites comparables.

Ces différents critères appellent les mêmes commentaires que leurs équivalents évoqués plus haut. En outre, ils ne peuvent être admis que si l'on accepte le principe de la monétarisation de la vie humaine, qui est souvent contesté en dehors du cercle des économistes. Il convient aussi de signaler que s'ils semblent, dans leur principe, plus complets que les précédents critères (le véritable objectif des politiques de sécurité routière étant de réduire les nombres de victimes, surtout les plus gravement touchées, et non les nombres d'accidents en eux-mêmes), ils comportent en pratique une fragilité supplémentaire : du fait du poids largement prédominant des décès, ils sont en réalité très dépendants des nombres d'accidents mortels, en nombre (heureusement) très limité et pour lesquels les estimations bayésiennes empiriques, comme les prédictions des modèles, sont extrêmement fragiles.

—o—

S'il fallait tenter de dégager une orientation générale pour le choix de critères de classement dans une politique de gestion de la sécurité des routes, il serait sans doute pertinent de se référer à l'évolution des méthodes d'évaluation des investissements d'infrastructure depuis une quarantaine d'années, qui tendent à se détourner des critères agrégés et totalisants, au profit d'une approche multicritère, où la quantification séparée de critères simples permet d'éclairer les décideurs sur les différents aspects du problème et préserve leurs prérogatives et leurs marges de manœuvre, s'agissant de l'appréciation du problème comme de la décision. En l'occurrence, le mieux serait sans doute de fournir aux décideurs différents critères complémentaires, de façon séparée, comme le critère du *nombre d'accident*s (tel qu'il peut être appréhendé par son espérance mathématique, au moyen de l'estimation bayésienne empirique), celui du *taux d'accidents* (également quantifié au moyen de l'estimation bayésienne empirique), éventuellement accompagnés de l'information sur la *moyenne pour des sites comparables* (ou la prédiction du modèle pour des sites comparables), et un critère de *gravité des accidents* (proportion d'accidents mortels, par exemple) sur le site ou la section considérée.

10. Conclusion

Nous avons souhaité dans cette note expliciter les bases de méthodes quantitatives qui sont aujourd'hui recommandées pour l'analyse des réseaux routiers du point de vue de l'insécurité routière, à des fins de gestion de la sécurité des infrastructures. En particulier nous avons présenté l'utilisation des estimations bayésiennes empiriques dans de telles méthodes, sans négliger les justifications mathématiques, qui nous semblent nécessaires à la bonne compréhension et à la bonne application ultérieure de ces démarches. L'annexe 2 devrait d'autre part donner des éléments au lecteur pour l'ajustement de modèles prédictifs des nombres d'accidents, dans l'environnement du logiciel libre R, de tels modèles étant souvent utiles pour le calcul d'estimations bayésiennes empiriques ; il est important cependant de compléter ces quelques éléments par la lecture de quelques articles sur ces techniques de modélisation (voir en particulier la référence Allain et Brenac, 2001). Nous espérons que les personnes impliquées dans la mise en œuvre des procédures de gestion de la sécurité des routes impulsées par l'Union européenne pourront tirer quelque profit des contenus de cette note.

ANNEXES

Annexe 1. Mélange Gamma-Poisson et distribution binomiale négative

Si, pour une valeur de *m* fixée, le nombre d'accidents observé *x* est une réalisation d'une variable de Poisson de moyenne *m*, et si cette moyenne *m* est elle-même une instanciation d'une distribution Gamma(α,β), alors, globalement, la probabilité d'observer *x* accidents résultant de ce mélange Gamma-Poisson relève d'une distribution binomiale négative de paramètres $n = \alpha$ et $p = \beta/(\beta+1)$.

En effet, si $\pi(m) = \beta^{\alpha} m^{\alpha-1} e^{-\beta m} / \Gamma(\alpha)$ représente la densité de la distribution Gamma(α,β),

$$P(X = x) = \int_0^{+\infty} P(X = x \mid m)\ \pi(m) dm$$

$$= \int_0^{+\infty} \frac{e^{-m}\ m^x}{x!}\ \frac{\beta^{\alpha} m^{\alpha-1} e^{-\beta m}}{\Gamma(\alpha)}\ dm = \frac{\beta^{\alpha}}{x!\ \Gamma(\alpha)} \int_0^{+\infty} m^{\alpha+x-1} e^{-(\beta+1)m} dm$$

d'où, après le changement de variable $u = (\beta+1)m$ [qui implique $m = u/(\beta+1)$ et $dm = du/(\beta+1)$] :

$$P(X = x) = \frac{\beta^{\alpha}}{x!\ \Gamma(\alpha)} \int_0^{+\infty} \left(\frac{u}{\beta+1}\right)^{\alpha+x-1} e^{-u}\ \frac{du}{(\beta+1)} = \frac{\beta^{\alpha}}{x!\ \Gamma(\alpha)} \left(\frac{1}{\beta+1}\right)^{\alpha+x} \int_0^{+\infty} u^{\alpha+x-1} e^{-u} du$$

L'intégrale terminant cette dernière équation vaut $\Gamma(\alpha+x)$, par définition de la fonction Gamma. D'où :

$$P(X = x) = \frac{\Gamma(\alpha+x)}{x!\ \Gamma(\alpha)} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{x}$$

ce qui correspond à la loi binomiale négative de paramètres $n = \alpha$ et $p = \beta/(\beta+1)$.

Annexe 2. Aspects pratiques de l'ajustement de modèles, quelques illustrations avec R

Nous présentons ci-dessous quelques éléments sur les moyens pratiques d'ajuster des modèles visant à relier les nombres d'accidents observés et un ensemble de variables explicatives. Notons cependant que ces éléments sont loin d'être complets. Ajuster un modèle suppose de l'expertise, concernant l'objet d'étude et les connaissances du domaine, certaines compétences en statistique, et un regard critique sur les variables explicatives candidates et leur codage, comme sur les résultats obtenus, notamment au moyen d'un examen visuel des relations entre les résidus et les valeurs prédites et entre les résidus et les différentes variables explicatives. Nous ne traiterons pas ici de ces divers aspects. Nous rappellerons juste ici un point particulièrement important : dans la constitution de l'échantillon (sur lequel le modèle est ajusté), il ne faut pas écarter les sites où aucun accident ne s'est produit. Plus généralement il faut éviter d'utiliser un critère basé sur le nombre d'accidents dans la constitution de l'échantillon.

Nous nous appuyons dans cette annexe sur l'exemple donné en section 7.2. (issu de Brenac et Verne, 2000), où l'on modélise les liens entre le nombre d'accidents observé dans une traversée d'agglomération, que nous noterons *ACC*, et des variables explicatives : longueur de la traversée (*LON*, en mètres), trafic (*TRAF*, en véhicules par jour) et population de l'agglomération (*POP*, en nombre d'habitants). L'hypothèse sous-jacente est que l'espérance du nombre d'accidents sur un site peut s'exprimer comme une combinaison multiplicative de ces facteurs :

$$\mu = Cste \times POP^a \times TRAF^b \times LON^c \qquad (15)$$

La constante et les coefficients *a*, *b* et *c* étant à déterminer lors de la modélisation. Dans le cadre des modèles linéaires généralisés, le recours à une *fonction de lien*, logarithmique dans notre cas, permet de se replacer dans le cadre linéaire :

$$\eta = \log(\mu) = K + a\log(POP) + b\log(TRAF) + c\log(LON) \qquad (16)$$

où $K = \log(Cste)$ représente une constante. Les conditions pour l'application des méthodes ordinaires de régression aux moindres carrés n'étant pas réunies[49], la modélisation s'appuie sur la maximisation de la vraisemblance (ou quasi-vraisemblance), en suivant les méthodes des modèles linéaires généralisés. Des méthodes itératives permettent, sauf cas particulier de divergence, d'obtenir les valeurs des coefficients du modèle (*K*, *a*, *b*, *c*…, dans notre exemple) qui maximisent la vraisemblance du modèle (pour une présentation plus complète et explicite, voir Allain et Brenac, 2001). Comme nous l'avons mentionné plus haut, il est nécessaire pour cela de spécifier la loi de la distribution des nombres d'accidents, ou du moins la fonction de variance reliant la variance et l'espérance de cette distribution.

Il est utile d'ajuster d'abord un modèle en spécifiant une distribution de Poisson, qui met généralement en évidence une certaine surdispersion, montrant la nécessité de spécifier plutôt une distribution binomiale négative ou de type quasi-Poisson. L'étude de la relation entre les carrés des résidus de Pearson et les valeurs prédites par ce modèle de Poisson, c'est-à-dire, en l'occurrence, entre les valeurs $\frac{(ACC_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$ et les $\hat{\mu}_i$ (où le dénominateur représente la fonction de variance du modèle ; dans l'exemple traité ici : $\hat{\mu}_i = Cste \times POP_i^a \times TRAF_i^b \times LON_i^c$ et, s'agissant d'un modèle de Poisson $V(\hat{\mu}_i) = \hat{\mu}_i$) permet d'orienter le choix vers un modèle binomial négatif ou de type quasi-Poisson (nous ne détaillons pas ici cet aspect : se reporter à la référence Allain et Brenac, 2001).

Nous donnons ci-dessous des exemples de modélisations réalisées avec le logiciel R, qui est un logiciel libre mais constitue aussi un logiciel de référence au niveau international dans le domaine de la statistique. Pour les personnes non familières de R, le manuel d'Emmanuel Paradis, *R pour les débutants* (2005 ; https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf) permet un accès commode à l'usage ce logiciel. Pour faciliter la compréhension, nous mettons en couleur rouge les commandes passées par l'utilisateur, le reste constituant les réponses du logiciel.

49. En particulier, du fait des aspects poissonniens, les termes ($ACC_i - \mu_i$) de la composante aléatoire du modèle n'ont pas une variance identique sur l'ensemble de l'échantillon.

Les exemples ci-dessous traitent du jeu de données `tragdatr.txt` (utilisé dans la publication Brenac et Verne, 2000), dont nous reproduisons ici les premières lignes :

```
POP     LON     ACC     TRAF
200     700     0       700
1000    1500    1       2000
600     2000    4       2500
400     1100    3       5700
800     700     7       6800
2000    1800    11      9300
600     800     5       3000
200     500     6       3800
3000    1400    7       1000
1500    430     7       6600
600     1400    9       4900
…       …       …       …
```

La lecture de ces données par R utilise la commande `read.table` :

```
> tragdata <- read.table("tragdatr.txt",header=TRUE,sep="\t")
```

Pour être lu, le fichier de données (tragdatr.txt) doit être placé sur le répertoire de travail. Pour localiser ce répertoire, faire la commande `getwd()` (pour : *get working directory*). À défaut, il faut compléter le nom du fichier par le chemin d'accès. Le paramétrage `header=TRUE` précise que les données sont précédées d'en-têtes, et `sep="\t"` précise que dans ce fichier le séparateur est de type tabulation. Le logiciel ajoute de lui-même un numéro d'observation à chaque ligne du fichier, comme on peut s'en apercevoir en visualisant les données :

```
> tragdata
    POP  LON ACC  TRAF
1   200  700   0   700
2  1000 1500   1  2000
3   600 2000   4  2500
4   400 1100   3  5700
5   800  700   7  6800
…    …    …    …    …
```

A2.1. Ajustement d'un modèle de Poisson

L'ajustement d'un modèle de Poisson utilise la commande `glm` de R:

```
> modelPoiss <- glm(ACC ~ log(POP)+log(TRAF)+log(LON),family=poisson,data=tragdata)
```

La fonction de lien logarithmique est la valeur par défaut pour le modèle de Poisson, et n'a pas à être spécifiée dans cette commande. Par contre pour les variables explicatives, il est nécessaire de faire apparaître dans la formule du modèle le fait qu'elles interviennent sous la forme de leur logarithme (`log` est la fonction logarithme népérien, dans R). La formule du modèle `ACC ~ log(POP)+log(TRAF)...` est symbolique, la constante (*K*) y est implicitement incluse (on pourrait la faire apparaître en écrivant : `ACC ~ 1 + log(POP)+log(TRAF)...`).

Les principaux résultats concernant le modèle sont obtenus par la commande `summary`, comme suit :

```
> summary(modelPoiss)

Call:
glm(formula = ACC ~ log(POP) + log(TRAF) + log(LON), family = poisson,
    data = tragdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2022  -1.4461  -0.3501   0.7910   3.0641

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.81054    0.77029 -12.736  < 2e-16 ***
log(POP)     0.48096    0.08655   5.557 2.74e-08 ***
log(TRAF)    0.55229    0.07477   7.387 1.51e-13 ***
log(LON)     0.49267    0.10673   4.616 3.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 588.95  on 97  degrees of freedom
Residual deviance: 218.44  on 94  degrees of freedom
AIC: 473.4

Number of Fisher Scoring iterations: 5
```

Les coefficients associés à l' « intercept », à log(*POP*), log(*TRAF*) et log(*LON*) sont, respectivement, la constante *K* et les coefficients *a*, *b* et *c* du modèle décrit plus haut (équation 16). Leurs sont associés, dans ces résultats, les écarts-types de ces estimations et des tests de signification (sur leur différence significative à la valeur zéro). L'exponentielle permet de repasser à la forme pratique du modèle (équation 15) : dans l'équation 15, la constante est donc égale à exp(*K*), et les coefficients *a*, *b* et *c* sont les exposants des variables *POP*, *TRAF* et *LON*, respectivement.

La déviance du modèle est un critère d'ajustement aux données s'appuyant sur la différence de log-vraisemblance entre ce modèle et le modèle saturé, une déviance faible correspondant à un meilleur ajustement aux données. `Null deviance` donne la déviance du modèle ne contenant pas d'autre variable explicative que la constante, et `Residual deviance` donne la déviance du modèle sous examen. Il est généralement recommandable d'ajuster différents modèles en introduisant d'abord une seule variable (en sus de la constante), puis en retenant le modèle à une variable réduisant le plus la déviance (ou augmentant le plus la vraisemblance), d'ajouter ensuite une deuxième variable, et ainsi de suite. Nous ne rendons pas compte ici de ces différentes étapes. Dans le cas présent, *POP*, *TRAF*, et *LON* apparaissent dans l'ordre de leur contribution décroissante à la réduction de la déviance. Introduire de nouvelles variables, même si elles réduisent la déviance, peut conduire à surajuster le modèle aux données (compromettant sa validité générale et sa pertinence pour l'application à d'autres jeux de données) ; pour cela il est utile de vérifier que le critère AIC (critère d'information d'Akaike) diminue bien au fur et à mesure de l'introduction de nouvelles variables (s'il ré-augmente, c'est le signe d'un surajustement) ; d'autres critères plus stricts (pouvant conduire à retenir moins de variables dans les modèles), comme le BIC (critère d'information bayésien) peuvent aussi être calculés.

Les commandes qui suivent permettent d'exporter certains résultats (ici sous forme d'une table contenant les valeurs des $\hat{\mu}_i$ ajustées par le modèle et les observations ACC_i) dans un fichier dans le répertoire de travail :

```
> muPoiss <- fitted(modelPoiss)
> ACC <- tragdata$ACC
> sorties <- data.frame(ACC,muPoiss)
> write.table(sorties,file="exploitN1.txt")
```

Dans les résultats, la mention « `Dispersion parameter for poisson family taken to be 1` » signifie que l'absence de surdispersion est une hypothèse de base pour ce modèle, mais ne signifie pas qu'en réalité il ne soit pas surdispersé. Un estimateur de la surdispersion est le rapport $\chi^2/(n-p)$ où n est le nombre d'observations et p le nombre de paramètres estimés lors de la modélisation, et où χ^2 est le khi-deux généralisé de Pearson $\chi^2 = \sum_i \frac{(ACC_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$ (voir Allain et Brenac, 2001). Dans le présent exemple où l'on étudie 98 traversées d'agglomération et où l'on ajuste un modèle avec 4 paramètres estimés (*K*, *a*, *b* et *c*), $n = 98$ et $p = 4$. Si on calcule cet indicateur de surdispersion dans ce cas, avec $V(\hat{\mu}_i) = \hat{\mu}_i$ s'agissant d'un modèle de Poisson, la valeur obtenue est 2,14 ce qui est nettement supérieur à 1 et indique la présence de surdispersion.

A2.2. Ajustement d'un modèle de type quasi-Poisson

L'analyse de la relation entre les carrés des résidus de Pearson et les valeurs prédites, comme nous l'avons mentionné plus haut, oriente dans l'exemple présenté vers un modèle de type quasi-Poisson, de fonction de variance $V(\hat{\mu}_i) = \tau\hat{\mu}_i$. Un tel modèle peut être obtenu avec la commande `glm` de R :

```
> modeleQP <- glm(ACC ~ log(POP)+log(TRAF)+log(LON),family=quasipoisson, data=tragdata)
> summary(modeleQP)

Call:
glm(formula = ACC ~ log(POP) + log(TRAF) + log(LON), family = quasipoisson,
    data = tragdata)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2022  -1.4461  -0.3501   0.7910   3.0641

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.8105     1.1280  -8.697 1.08e-13 ***
log(POP)      0.4810     0.1267   3.795 0.000262 ***
log(TRAF)     0.5523     0.1095   5.044 2.21e-06 ***
log(LON)      0.4927     0.1563   3.152 0.002176 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 2.144552)

    Null deviance: 588.95  on 97  degrees of freedom
Residual deviance: 218.44  on 94  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Dans ce cas R ajuste le modèle en maximisant la quasi-vraisemblance. Le paramètre de surdispersion n'a pas à être spécifié car il est estimé au cours de l'ajustement. On retrouve ici la valeur du obtenue plus haut en calculant $\chi^2/(n-p)$ pour le modèle de Poisson. La fonction de lien par défaut est la fonction logarithme pour les modèles de type quasi-Poisson, elle n'a pas à être spécifiée.

On observe que les coefficients du modèle quasi-Poisson, et donc les valeurs prédites, sont les mêmes que pour le modèle de Poisson. En revanche, les écarts-types sur les estimations des coefficients sont différents : par rapport au modèle quasi-Poisson, le modèle de Poisson sous-estime ces écarts-types, et donc surestime la précision obtenue sur ces coefficients et sur les valeurs prédites. Dans certaines situations, même si ce n'est pas le cas ici, un modèle de Poisson pourrait inclure une variable explicative dont l'influence n'est en réalité pas significative au regard d'un modèle de type quasi-Poisson qui prendrait en compte la surdispersion.

A2.3. Ajustement d'un modèle binomial négatif

Bien qu'un tel modèle ne soit pas approprié pour les données étudiées ici, il est pertinent dans de nombreux cas, et nous présentons ci-dessous la façon de l'ajuster, dans le seul but de donner des indications méthodologiques. Pour un tel modèle, il faut recourir à la commande `glm.nb` :

```
> library(MASS)
> modeleBN <- glm.nb(ACC ~ log(POP)+log(TRAF)+log(LON),data = tragdata, link = log)
```

L'utilisation de cette commande doit être précédée de la commande `library(MASS)` pour charger ce module de R dans l'espace de travail. Nous avons spécifié ici la fonction de lien logarithmique, mais ce n'est pas nécessaire car c'est la valeur par défaut. Les résultats sont les suivants :

```
> summary(modeleBN)

Call:
glm.nb(formula = ACC ~ log(POP) + log(TRAF) + log(LON), data = tragdata,
    link = log, init.theta = 3.836488373)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5614  -1.1651  -0.2676   0.4322   2.2917

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.4436     1.1870  -8.798  < 2e-16 ***
log(POP)      0.3960     0.1372   2.885 0.003908 **
log(TRAF)     0.6264     0.1125   5.566 2.61e-08 ***
log(LON)      0.5771     0.1690   3.414 0.000641 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.8365) family taken to be 1)

    Null deviance: 291.12  on 97  degrees of freedom
Residual deviance: 129.24  on 94  degrees of freedom
AIC: 452.73
```

```
Number of Fisher Scoring iterations: 1


              Theta:  3.84
          Std. Err.:  1.47

 2 x log-likelihood:  -442.733
```

Les résultats comprennent les valeurs des coefficients, qui diffèrent ici de ceux des modèles quasi-Poisson et Poisson, (0,3960 au lieu de 0,4810 pour *POP* ; 0,6264 au lieu de 0,5523 pour *TRAF* ; et 0,5771 au lieu de 0,4927 pour *LON*). Ils comportent aussi le paramètre de dispersion θ, qui est identique au paramètre k que nous avons utilisé dans l'équation 4 (qui est aussi égal au paramètre α des distributions *a priori* des m_i), et qui intervient dans la fonction de variance de la distribution binomiale négative : $V(\mu) = \mu + (1/k)\,\mu^2$. Il vaut ici 3,84.

On peut alors déduire de ces résultats la distribution *a posteriori* de chaque moyenne m_i, qui est une distribution Gamma de paramètres $(k + ACC_i)$ et $(k/\hat{\mu}_i + 1)$, où $\hat{\mu}_i$ est la valeur prédite par le modèle.

A2.4. Cas de modèles incluant des variables explicatives qualitatives

Dans les exemples présentés ci-dessus, les modèles ne contiennent que des variables quantitatives. En pratique, il est fréquent que les modèles comportent un mélange de variables quantitatives et qualitatives. Nous donnons ci-dessous un exemple avec un autre fichier de données concernant les mêmes traversées d'agglomération, mais complété par la donnée d'une nouvelle variable, *CAT* (« catégorie de route »), prenant deux modalités, *N* (route nationale) ou *D* (route départementale). Nous nous limiterons ici à l'ajustement du modèle de Poisson, l'objectif étant simplement d'illustrer la façon d'intégrer une variable qualitative dans un modèle.

Le modèle étudié, après linéarisation par la transformation logarithmique, est alors le suivant :

$$\eta = \log(\mu) = K + a\log(POP) + b\log(TRAF) + c\log(LON) + d\,CATN \qquad (17)$$

où *CATN* est une variable artificielle prenant la valeur 1 si la traversée d'agglomération est située sur une route nationale (*CAT* = *N*), et 0 si elle est située sur une route départementale (*CAT* = *D*). La modalité « route départementale » constitue le niveau de référence de la variable *CAT*, son effet est contenu dans le terme *K*. Le coefficient *d* est un coefficient à ajuster lors de la modélisation, comme *K*, *a*, *b* et *c*.

On retrouve par exponentiation le modèle sous une forme pratique :

$$\mu = \exp(K) \times POP^a \times TRAF^b \times LON^c \times \exp(d\,CATN) \qquad (18)$$

Autrement dit $\mu = Cste \times POP^a \times TRAF^b \times LON^c \times \exp(d)$ pour une traversée située sur une route nationale, et $\mu = Cste \times POP^a \times TRAF^b \times LON^c$ pour une traversée située sur une route départementale. Le terme $\exp(d) = e^d$ représente donc le facteur multiplicatif à appliquer lorsque la traversée d'agglomération est située sur une route nationale.

Lors de la mise en œuvre avec le logiciel R, il n'est pas nécessaire de créer la variable artificielle *CATN*, le logiciel s'en charge. Il suffit d'introduire la variable *CAT* dans la formule du modèle. Les modalités (*N* et *D*) étant en caractères alphabétiques, une telle variable est reconnue par le logiciel R comme une variable de type *factor* (variable catégorique), avec plusieurs modalités. Le modèle de Poisson peut alors être obtenu par :

```
> modelPoiss <- glm(ACC ~ log(POP)+log(TRAF)+log(LON)+CAT,family=poisson,data=tragdatb)
```

On peut ensuite accéder aux principaux résultats :

```
> summary(modelPoiss)

Call:
glm(formula = ACC ~ log(POP) + log(TRAF) + log(LON) + CAT, family = poisson,
    data = tragdatb)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0964  -1.4292  -0.2894   0.7394   2.9758
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.90257    0.76739 -12.904  < 2e-16 ***
log(POP)     0.49087    0.08718   5.631 1.79e-08 ***
log(TRAF)    0.59894    0.08138   7.359 1.85e-13 ***
log(LON)     0.44693    0.11072   4.036 5.43e-05 ***
CATN        -0.17394    0.11707  -1.486    0.137
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 588.95  on 97  degrees of freedom
Residual deviance: 216.20  on 93  degrees of freedom
AIC: 473.17

Number of Fisher Scoring iterations: 5
```

Pour la variable *CAT*, le logiciel R prend par défaut la modalité *D* comme la modalité de référence (dont l'effet est contenu dans la constante *K*), car elle précéde la modalité *N* dans l'ordre alphabétique, et il crée la variable artificielle *CATN* (valant 1 si *CAT* = *N*, et 0 si *CAT* = *D*)[50].

Dans le cas d'une traversée d'agglomération située sur une route nationale, la valeur prédite par le modèle pour le nombre d'accidents serait donc à multiplier par exp(–0,17394) = 0,84, par rapport à une traversée d'agglomération comparable située sur une route départementale. Cependant l'écart-type de l'estimation du coefficient $d = -0{,}17394$ est important (0,11707) et suggère que *d* n'est pas significativement différent de 0 (la *p*-value de 0,137 correspond à un résultat non significatif).

On voit donc dans ces résultats que la variable *CAT* n'apporte pas de contribution significative. Par rapport au modèle n'intégrant pas cette variable, la déviance résiduelle n'est que très faiblement améliorée (216,20 au lieu de 218,44) et le critère AIC est pratiquement identique (473,17 au lieu de 473,40).

50. Dans le cas d'une variable à plus de deux modalités : par exemple, pour une variable *U* prenant les modalités *A*, *B* ou *C*, le logiciel prendrait la modalité *A* comme modalité de référence, et donnerait les résultats pour deux variables artificielles *UB* et *UC* (*UB* valant 1 si *U* = *B* et 0 sinon, et *UC* valant 1 si *U* = *C* et 0 sinon), sur deux lignes différentes dans le tableau des résultats concernant les coefficients obtenus.

Annexe 3. Cas de sous-dispersion

Lors du calcul d'estimations bayésiennes empiriques s'appuyant sur un échantillon de sites similaires, si l'on est dans le cas très rare où s^2 est inférieur à $\bar{y}$, cela signifie qu'il y a sous-dispersion : les nombres d'accidents sont encore moins dispersés que s'il n'y avait que les fluctuations aléatoires poissonniennes propres à chaque site. Concrètement ce cas est irréaliste et peut être lié au caractère aléatoire des observations ; dans cette situation on fera plutôt l'hypothèse qu'il n'y a ni surdispersion ni sous-dispersion, et on prendra s^2 égal à $\bar{y}$. Cela revient à considérer que le site est pratiquement identique (du point de vue de la sécurité) aux sites de l'échantillon et partage avec eux une même moyenne de Poisson. Cela conduit à une estimation bayésienne empirique égale à $\bar{y}$. Cela reste peu réaliste sauf dans le cas d'un échantillon de sites extrêmement homogène à tous égards.

Une situation analogue peut se produire dans le cas où on s'appuie sur un modèle, lorsque le modèle obtenu est un modèle de Poisson non surdispersé (voire sous-dispersé). L'estimation bayésienne empirique est alors égale à la prédiction du modèle. Cela signifierait soit que l'on a obtenu un modèle parfait, au sens où toutes les variables explicatives pertinentes auraient été prises en compte de façon appropriée, les écarts entre valeurs prédites par le modèle et valeurs observées s'expliquant uniquement par les variations poissonniennes propres à chaque site (ce qui est peu réaliste en général), soit que le modèle a été surajusté : un excès de variables explicatives peut conduire à ajuster excessivement le modèle à l'échantillon et aux aléas des observations propres à cet échantillon (au détriment de la fiabilité de la prédiction sur tout autre site ou échantillon). On peut tenter de se prémunir contre ce dernier cas en se restreignant, dans le cadre d'un modèle de Poisson (ou binomial négatif), à un jeu de variables minimisant le critère d'information d'Akaike (AIC), ou minimisant le critère BIC (*Bayesian information criterion*) qui conduit souvent à retenir moins de variables.

Annexe 4. Erreur liée à l'usage d'estimations bayésiennes empiriques, variance de la distribution *a posteriori*…

Erreur liée à l'usage d'estimations bayésiennes empiriques

Les travaux de Robbins (1985), s'intéressant au cas général où la distribution d'observations x conditionnellement à un paramètre d'intérêt θ n'est pas précisément connue (au-delà du fait que l'on suppose que $E(x|\theta) = \theta$), et où la distribution *a priori* de θ n'est pas non plus spécifiée, permettent notamment de quantifier l'erreur quadratique moyenne (espérance du carré de l'erreur $\hat{\theta}_{BE} - \theta$) liée à l'utilisation de l'estimateur linéaire bayésien empirique.

Si l'on applique ces développements à notre cas, plus spécifique, où le nombre d'accidents est supposé distribué selon une loi de Poisson conditionnellement au paramètre (la moyenne m de cette loi), l'erreur quadratique moyenne, $E[(\hat{m}_{BE} - m)^2]$, peut être estimée par $\bar{y}\left(1 - \frac{\bar{y}}{s^2}\right)$ si l'on s'appuie sur un échantillon de sites comparables, ou bien $\hat{\mu}\,(1 - v)$ si l'on a recours à un modèle. Dans cette dernière expression v représente le poids utilisé soit $\left(\frac{1}{1+\hat{\mu}/k}\right)$ dans le cas d'un modèle binomial négatif et $\frac{1}{\tau}$ dans le cas d'un modèle de type quasi-Poisson.

Cette espérance de l'erreur quadratique est indépendante du nombre d'accidents observé, elle rend compte d'une erreur quadratique moyenne sur l'ensemble du domaine des fluctuations aléatoires du paramètre et des observations (« *the mean squared error of estimate with regard to the random variation of both* θ *and* x », Robbins, 1983, p. 713). Elle traduit donc plutôt un point de vue de statistique conventionnelle, évaluant l'erreur d'estimation moyenne dans l'hypothèse d'une répétition virtuelle illimitée d'une expérience.

Variance de la distribution a posteriori *de la moyenne* m

D'un point de vue bayésien, il paraît plus logique de s'intéresser à la distribution *a posteriori* du paramètre, qui rend compte de notre incertitude concernant la valeur de ce paramètre après prise en compte de l'observation x. Cette distribution est entièrement connue dans le cas où la distribution *a priori* du paramètre est spécifiée. Si la distribution *a priori* est une distribution Gamma(α,β), par exemple, la distribution *a posteriori* est une distribution Gamma(α+x,β+1), d'espérance (α+x)/(β+1) et de variance $Var(m|x) = (\alpha+x)/(\beta+1)^2$.

Hauer (1997, pages 188 et suivantes et 194, 195) a tenté de donner une expression plus générale de la variance $Var(m|x)$ de la distribution *a posteriori*[51], applicable même lorsque la distribution *a priori* n'est pas spécifiée. Mais si l'expression obtenue $Var(m|x) = (1 - v)\,E(m|x)$ (où v est le poids utilisé dans l'expression de l'estimation bayésienne empirique) est bien valable dans le cas d'une distribution *a priori* de type Gamma, elle ne semble pas pouvoir être démontrée dans le cas général (Hauer n'en présente pas une démonstration solide, comme il le reconnaît lui-même, *ibid*.) et elle n'a probablement pas de validité générale.

Autres aspects

Les travaux de Robbins (1985) se sont intéressés d'autre part à l'estimation de la variance de la distribution des x conditionnellement à un paramètre d'intérêt θ, dans le cas où cette distribution n'est pas entièrement spécifiée mais où l'on suppose simplement que $E(x|\theta) = \theta$. Ces développements ne nous concernent pas ici, puisque dans notre cas nous avons supposé que la forme de cette distribution est spécifiée, relevant de la loi de Poisson.

51. Il faut souligner que $Var(m|x)$ est la variance de la distribution *a posteriori* de m, connaissant x, mais n'est pas la variance de $E(m|x)$, contrairement à ce qu'écrit Hauer (1997, p. 188).

Références

Abbess, C., Jarrett, D., Wright, C. C. (1981), "Accidents at blackspots: estimating the effectiveness of remedial treatment, with special reference to the 'regression-to-mean' effect", *Traffic Engineering and Control* , vol. 22(10), p. 535-542.

Allain, E., Brenac, T. (2001), « Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers : le modèle de Poisson et ses extensions », *Recherche Transports Sécurité*, vol. 72, p. 3-18.

Amoros, E., Martin, J.-L., Laumon, B. (2006), "Under-reporting of road crash casualties in France", *Accident Analysis and Prevention*, vol. 38(4), p. 627-635.

Brenac, T. (1994), *Accidents en carrefour sur routes nationales, modélisation du nombre d'accidents prédictible sur un carrefour et exemples d'applications* (rapport INRETS, 185) Arcueil, INRETS, 107 p.

Brenac, T. (2009), "Common before-after accident study on a road site: a low-informative Bayesian method", *European Transport Research Review*, vol. 1(3), p. 125-134.

Brenac, T. (2010), "Safety effects of mobile speed cameras in Norfolk: no more than regression to the mean?" *Journal of Safety Research*, vol. 41(1), p. 65-67.

Brenac, T., Verne, J.-N. (2000), « Niveau d'insécurité routière dans les traversées d'agglomération, modélisation de l'influence du trafic, de la population et de la longueur de traversée », *Bulletin des laboratoires des Ponts et chaussées*, vol. 224, p. 13-24.

Chang, L.-Y. (2005), "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network", *Safety Science*, vol. 43, p. 541-557.

Communauté d'agglomération du Grand Dijon (2013), *Observatoire des mobilités actives 2012*. Dijon, Le Grand Dijon, 67 p.

El Mansouri, S., Fournier, J.-Y. (2018), « Localisation des accidents par coordonnées GPS dans le fichier national des accidents de la circulation, un état de la situation », *Carnets d'accidentologie*, vol. 2018, p. 13-23.

Elvik, R. (2007), *State-of-the-art approaches to road accident black spot management and safety analysis of road networks*, TØI report 883/2007, Oslo, Transportøkonomisk Institutt, 108 p.

Galton, F. (1886), "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute*, vol. 15, p. 246-263.

Greibe, P. (2003), "Accident prediction models for urban roads", *Accident Analysis and Prevention*, vol. 35, p. 273-285.

Gouriéroux, C., Monfort, A., Trognon, A. (1984a) "Pseudo maximum likelihood methods: theory", *Econometrica*, vol. 52, p. 681-700.

Gouriéroux, C., Monfort, A., Trognon, A. (1984b), "Pseudo maximum likelihood methods: application to Poisson models", *Econometrica*, vol. 52, p. 701-720.

Hauer, E. (1992), "Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method", Accident Analysis and Prevention, vol. 24(5), p. 457-477.

Hauer, E. (1997), *Observational before-after studies in road safety*, Oxford, Pergamon.

Hauer, E., Kononov, J., Allery, B., Griffith, M. S. (2002), "Screening the road network for sites with promise", *Transportation Research Record*, vol. 1784(1), 27-32.

Jarrett, D. (1994). "Statistical analysis of road accident frequencies", *Symposium on Safety in the Road Environment*, 21-22 avril 1994, Laboratorio Nacional de Engenharia Civil, Lisbonne.

Jones, A. P., Sauerzapf, V., Haynes, R. (2008), "The effects of mobile speed camera introduction on road traffic crashes and casualties in a rural county of England", *Journal of Safety Research*, vol. 39(1), p. 101-110.

Kass, R. E., Wasserman, L. (1996), "The selection of prior distribution by formal rules", *Journal of the American Statistical Association*, vol. 91, p. 1343-1370.

Maher, M. J. (1987), "Fitting probability distributions to accident frequency data", *Traffic Engineering and Control*, June 1987, p. 356-360.

Maher, M. J., Mountain, L. (1988), "The identification of accident blackspots: a comparison of current methods", *Accident Analysis and Prevention*, vol. 20(2), p. 143-151.

Maher, M. J, Summersgill, I. (1996), "A comprehensive methodology for the fitting of predictive accident models", *Accident Analysis and Prevention*, vol. 28(3), p. 281-296.

Martin, J.-L. (2000), « Utilisation de modèles linéaires généralisés pour tester l'effet sur la sécurité d'une modification d'infrastructure, comparaison des glissières en métal aux barrières en béton en terre-plein central d'autoroute », *Recherche Transports Sécurité*, vol. 68, p. 31-43.

McCullagh, P., Nelder, J. A. (1989), *Generalized linear models*, (2ème édition, "Monographs on statistics and applied probability 37"), Chapman & Hall, 511 p.

Mountain, L., Fawaz, B. (1991), "The accuracy of estimates of expected accident frequencies obtained using an Empirical Bayes approach", *Traffic Engineering and Control*, vol. 32(5), p. 246-251.

Mountain, L., Fawaz, B., Sineng, L. (1992*a*), "The assessment of changes in accident frequencies at treated intersections: a comparison of four methods", *Traffic Engineering and Control*, February 1992, p. 85-87.

Mountain, L., Fawaz, B., Sineng, L. (1992*b*), "The assessment of changes in accident frequencies on link segments: a comparison of four methods", *Traffic Engineering and Control*, July/August 1992, p. 429-431.

Nicholson, A., Wong, Y. D. (1993), "Are accidents Poisson distributed? A statistical test", *Accident Analysis and Prevention*, vol. 25(1), p. 91-97.

Persaud, B. (1986), "Relating the effect of safety measures to expected number of accidents", *Accident Analysis and Prevention*, vol. 18(1), p. 63-70.

Persaud, B. (1990), *Blackspot Identification and Treatment Evaluation*, Ministry of Transportation of Ontario, Research and Development Branch, 32 p.

Persaud, B., Lyon, C. (2007), "Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions", *Accident Analysis and Prevention*, vol. 39(3), p. 546-555.

Robbins, H. (1983), "Some thoughts on empirical Bayes estimation", *The Annals of Statistics*, vol. 11(3), p. 713-723.

Robbins, H. (1985), "Linear empirical Bayes estimation of means and variances", *Proceedings of the National Academy of Sciences*, vol. 82, p. 1571-1574.

Robert, C. (2006), *Le choix bayésien, principes et pratique*, Paris, Springer-Verlag France.

Saporta, G. (1990), *Probabilités, analyse des données et statistique*, Paris, Technip, 493 p.

Wedderburn, R. W. M. (1974), "Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method", *Biometrika*, vol. 54, p. 439-447.

Wright, C. C., Abbess, C., Jarrett, D. (1988), "Estimating the regression-to-mean effect associated with road accident blackspot treatment: towards a more realistic approach", *Accident Analysis and Prevention*, vol. 20(3), p. 199-214.