

ESTIMATEUR À NOYAU (KDE) SUR RÉSEAUX : UNE APPLICATION AUX ACCIDENTS DE LA ROUTE BELGES

David DABIN, Christiane DICKENS et Paul WOUTERS

Résumé

Le problème de la détection des zones de haute concentration des accidents de la route est un sujet de première importance tant pour les décideurs que pour les gestionnaires des infrastructures routières. Cependant, la notion même de point noir reste sujette à de vives discussions entre les experts du domaine bien que des définitions fonctionnelles existent dans plusieurs pays (Elvik, 2008). Les différents outils utilisés par les autorités en Belgique négligent plusieurs dimensions importantes des données d'accident notamment l'aspect réseau dont elles proviennent, l'aspect stochastique des accidents, les éventuelles erreurs de localisation de ceux-ci et rendent impossible la définition de zones noires d'accidents de la route. Dans ce travail, nous proposons une méthodologie mixte de trois étapes: (i) la fonction de densité des événements ponctuels sur le réseau est évaluée par un KDE, (ii) la significativité des valeurs lissées observées est calculée par des simulations MC et (iii) des outils d'analyse des réseaux permettent de classifier les éléments significatifs en *hotspots* et *hotzones*. La méthodologie est testée sur l'entièreté des routes numérotées belges pour la période 2006-2009.

Mots clés

accident de la route - point noir - zone noire - estimateur de densité à noyau - réseau

Abstract

The detection of high accident concentrations on roads is of vital importance for stakeholders but also for road safety managers. However experts in the field do not agree about the notion of black spots itself despite the existence of different practical definitions in many countries (Elvik, 2008). The different tools currently in use at the Belgian scale neglect some important dimensions of the accidents data such as the fact that accidents occur only on the road network, the spatial random component of some accidents, the localization problem and the impossibility to define black zones of road accidents. We propose here a new methodology based on three steps: (i) the density function of point event is estimated using a Kernel Density Estimator (KDE), (ii) the p-value of observed values against Complete Spatial Randomness (CSR) is computed through Monte-Carlo simulations and (iii) the significant items are classified into hotspots and hotzones by network analysis tools. The methodology is then evaluated on all road accidents with casualties on the Belgian numbered road network.

Keywords

road accident - hotspot - hotzone - KDE - network

1. INTRODUCTION

Les États Généraux de la Sécurité Routière de 2011 ont fixé un objectif de 620 tués maximum sur les routes belges en 2015 et de 420 tués en 2020. Un effort conséquent doit donc encore être consenti en matière de sécurité routière afin de tenir ces objectifs d'autant que le travail devient de plus en plus complexe à mesure que les progrès sont engrangés car plus le nombre de tués se réduit, plus les nouvelles diminutions de mortalité sont difficiles à obtenir (Geurts et Wets, 2003 ; Cassteels *et al.*, 2010). Afin de porter ses fruits, cet effort supplémentaire doit viser au minimum chacune des 3

composantes clés de la sécurité routière que sont (i) les infrastructures, (ii) le niveau de sécurité et de sûreté des véhicules et, finalement, (iii) le conducteur (Iversen et Rundmo, 2002 ; Castellà et Pérez, 2004). Dans le contexte de ressources budgétaires limitées, la question est de savoir où doivent être portés ces efforts pour obtenir un effet maximal.

En Belgique, plusieurs méthodes sont actuellement utilisées par les autorités afin d'identifier les éléments ponctuels problématiques nommés classiquement points noirs ou *hotspots* dans le domaine de la circulation routière.

- Comptage du nombre absolu d'accidents par km. Toutes les bornes hectométriques dont la fréquence des accidents dépasse une valeur seuil sont alors considérées comme des points noirs.
- Indice de dangerosité pondéré par la gravité des blessures encourues défini par la relation

$$S(i) = LI(i) + 3 SI(i) + 5 DI(i)$$

où $S(i)$ est l'indice de dangerosité de la borne i , LI est le nombre de blessés légers, SI est le nombre de blessés graves et DI le nombre de tués (Geurts et Wets, 2003). Cet indice est calculé pour toutes les bornes hectométriques dont le nombre d'accidents de roulage avec lésions corporelles est de minimum 3 pour les 3 dernières années. Les bornes hectométriques qui présentent une valeur supérieure à 15 sont alors considérées comme dangereuses et prioritaires.

- Indice de risque spatio-temporel (Romano et Heuchenne, 1996 ; Romano, 1997 ; Antoine, 2010). Route par route, un indice mensuel est d'abord calculé sur une matrice espace-temps qui représente une route sur la période de temps donnée. Cet indice se construit par une fenêtre mobile de 200 m et de 5 mois c'est-à-dire que l'indice est fonction du nombre d'accidents à la borne hectométrique elle-même et de ses deux voisins de part et d'autre mais aussi en fonction des accidents sur les 5 mois précédents et des 5 mois suivants. L'indice de risque pour une borne hectométrique est ensuite calculé comme la moyenne des indices mensuels à cette borne. Des valeurs supérieures à 2,4 sont classifiées en tant que zones à haut risque alors que des valeurs comprises entre 1,2 et 2,3 sont considérées comme des zones à risque moyen.

L'exploitation des résultats de ces analyses *hotspots* démontre cependant plusieurs limitations.

- L'impossibilité de définir des zones noires ou *hotzones*. Les autorités responsables des infrastructures ainsi que celles répondant des services de police visent un maximum d'efficacité. Comme il est plus facile de traiter des segments de route moins nombreux mais plus longs, une approche par zone noire semble plus appropriée que celle par point noir.
- La négation du réseau routier. En effet, les indices de dangerosité de Geurts et Wets (2003)

et de risque de Romano et Heuchenne (1996) sont calculés route par route en se basant sur les bornes hectométriques mais en négligeant totalement les carrefours. Les relations entre les routes disparaissent donc totalement.

- La nature stochastique des accidents de roulage. Un certain nombre d'accidents sont provoqués par des causes qui ne dépendent pas directement du milieu environnant. Le site de l'accident et ses caractéristiques propres telles que les infrastructures ne permettent pas d'expliquer entièrement ou en partie le sinistre. Les accidents revêtent alors un caractère purement aléatoire du point de vue spatial. Ils ont été observés à la borne hectométrique i cette année mais pourraient très bien se produire à la borne $i+1$ ou $i-1$ l'année prochaine. Par une approche purement fréquentiste, il est impossible de tenir compte de cette nature aléatoire.
- La prise en compte des erreurs d'allocation. La qualité de l'encodage du lieu des accidents laisse à désirer. En effet, sur la période 2006-2009, ce sont près de 1/5 des accidents qui n'ont pas pu être géocodés. Plusieurs raisons expliquent ce défaut d'encodage des coordonnées spatiales des accidents dont notamment les difficultés à trouver la borne hectométrique (BH) ou l'adresse la plus proche lors de l'enregistrement (BH cachée, absente, accidents en rase campagne sans maison proche), la méconnaissance du lieu et du nom des rues, une modification de la structure du réseau routier par l'introduction de nouvelles routes, le manque de temps en opération, l'absence de motivation à remplir ces données...

Des développements récents mettent en avant de nouvelles possibilités afin de dépasser ces limitations. Citons entre autres, les travaux de Flahaut *et al.* (2003), Elvik (2008), Xie et Yan (2008), Moons *et al.* (2009), Okabe *et al.* (2009), Steenberghen *et al.* (2010), Manepalli *et al.* (2011) et Truong et Somenahalli (2011). Parmi toutes ces méthodologies, l'estimateur de densité à noyau (Kernel Density Estimator, KDE) est couramment utilisé pour étudier la distribution d'évènements ponctuels et identifier les zones de hautes concentrations ou *hotspots* au sein d'un espace homogène par l'ajustement d'une surface de densité (Xie et Yan, 2008 ; Anderson, 2009). Cette technique est même considérée comme la carte la plus populaire en analyse criminelle après la *pin map* (Smith et Bruce, 2008) et est évaluée comme une des méthodes les plus performantes pour prédire les patterns spatiaux de crimes futurs (Chainey *et al.*, 2008).

Cependant, de nombreux phénomènes ponctuels ne se déroulent pas librement dans l'espace mais sont naturellement contraints à un réseau. Au niveau de la problématique de sécurité, citons par exemple les vols de cuivre sur le réseau ferroviaire, le problème des pick-pockets dans les transports publics, le vandalisme dans les espaces publics, les vols de voiture ou encore les accidents de la circulation. Pour ces exemples, l'utilisation de méthodes bidimensionnelles classiques qui supposent l'homogénéité de l'espace dans toutes les directions semble donc hasardeuse (Xie et Yan, 2008). Un réseau est en effet un espace particulier dont la dimension de 1,5 est comprise entre la ligne (1 D) et le plan (2 D) (Steenberghen *et al.*, 2010).

Dans ce cadre, les objectifs de ce travail sont triples: (i) établir une méthodologie pour détecter les *hotzones* d'accidents de la circulation sur un réseau routier complexe, (ii) déterminer les paramètres optimaux de cette méthodologie pour l'analyse au niveau de la Belgique et (iii) tester cette méthodologie sur base des Accidents de Roulage avec Lésions Corporelles (AccRLC) sur les routes principales belges.

II. DONNÉES

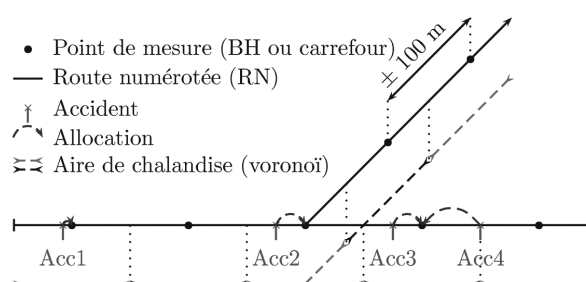
L'analyse porte sur l'entièreté du réseau des routes principales de Belgique, c'est-à-dire toutes les routes du réseau numéroté. Ce réseau représente environ 16.718 km répartis en autoroutes, routes nationales et routes régionales. Notons finalement que les routes numérotées disposent en Belgique d'un système de référencement nommé « borne hectométrique » car il prend la forme de bornes placées approximativement tous les 100 m.

Les données d'accidents de roulage sont issues des Formulaires d'Accidents de la Circulation (FAC) rédigés par les forces de Police pour tous les AccRLC. Dans ces FAC, les accidents de roulage sont localisés selon un des 3 systèmes suivants illustrés en Figure 1:

- Borne Hectométrique (BH): commune, numéro de route, borne hectométrique (BH) la plus proche;
- Adresse: commune, nom de rue et numéro de maison la plus proche;
- Carrefour: commune, nom de rue 1 et nom de rue 2.

Cependant parmi ces accidents, ceux localisés par l'adresse la plus proche sont alloués lors du post-traitement à la borne hectométrique ou au carrefour le plus proche pour des raisons de cohérence et de consistance des données. Le référentiel spatial de localisation des AccRLC sur le réseau des routes numérotées belges est alors composé de 185.475 bornes hectométriques et 11.404 carrefours soit 196.879 points de mesure. Le processus d'allocation au point de mesure le plus proche fait apparaître des aires de chalandise autour

de chacun de ces points. Chaque aire de chalandise, également connue sous le nom de segment « Voronoï », correspond à la portion du réseau qui verra tous les accidents s'y déroulant être alloués au point de mesure central (Figure 1). Notons que ce phénomène est constaté pour tous les processus nécessitant un géocodage. D'un phénomène purement continu pouvant se dérouler partout sur le réseau, le géocodage retourne seulement un nombre fini de points passant d'un pro-



blème continu à un problème discret.

Figure 1. Système de localisation des accidents par les bornes hectométriques (BH) et les carrefours et segments « Voronoï » induits sur les routes nationales (RN)

III. MÉTHODE

Considérons un réseau $G=(N, L)$ composé de N noeuds (*vertices*) et L lignes (*edges*) illustré en Figure 2. Chaque noeud i appartenant à l'ensemble N possède des caractéristiques propres telles que son aire de chalandise (segment « Voronoï » sous-jacent) et le nombre d'évènements ponctuels qui y ont pris place. Chaque noeud se trouve connecté à ses voisins j par des lignes l_{ij} de longueur $|l_{ij}|$. Le parallèle avec la localisation des accidents par les points de mesure est évident.

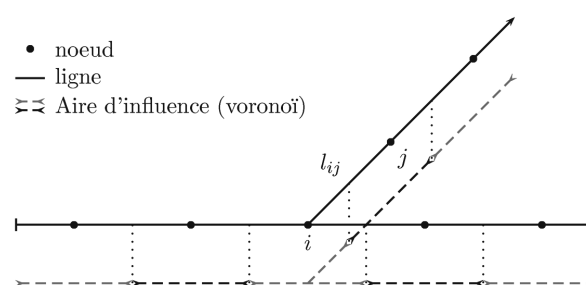


Figure 2. Description des éléments constitutifs du réseau

Pour un noeud i quelconque, nous pouvons définir un voisinage V_i qui contient tous les noeuds se trouvant à une distance inférieure ou égale à h de ce noeud. Dans un estimateur à noyau de densité (KDE pour *Kernel Density Estimator*) classique, cette distance est de type euclidien alors qu'elle sera calculée le long du réseau dans ce cas-ci. L'impact de ce changement de distance est évident pour l'exemple de la Figure 3 où le point p

fait partie du voisinage selon la distance euclidienne et en est exclu selon la distance le long du réseau. La distance le long du réseau paraît également plus pertinente avec une application sur des routes européennes qui présentent un tracé moins régulier que le réseau routier en damier des Etats-Unis par exemple.

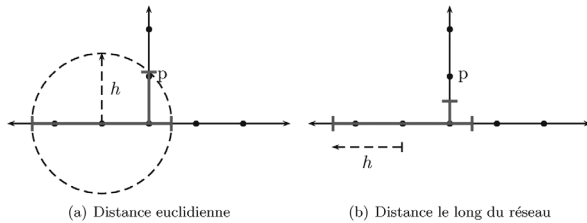


Figure 3. Calcul du voisinage sur base des (a) distances euclidiennes et des (b) distances le long du réseau

Nous pouvons également définir une fonction $k(x_i)$ quelconque qui satisfait à deux conditions

$$k(x_i) \begin{cases} \geq 0 & \forall x \in V_i \\ = 0 & \forall x \notin V_i \end{cases} \quad (1)$$

et,

$$\int_{x \in V_i} k(x_i) dx = 1 \quad (2)$$

Classiquement, la distance h qui définit le voisinage V est dénommée fenêtre ou *bandwidth* et $k(x_i)$ noyau ou *kernel function*. L'estimation de la densité à un point j est alors la somme sur l'ensemble du voisinage V_j des noyaux multipliés par le nombre d'occurrences n_i .

$$K(x_j) = \sum_{i \in V_j} k(x_i) n_i \quad (3)$$

Au delà de sa formulation mathématique, un estimateur à noyau peut se concevoir simplement comme le remplacement des événements ponctuels par un noyau dont la masse vaut 1 mais répartie sur une fenêtre de largeur h . À chaque point du réseau, la somme des noyaux définit la densité lissée. Dans l'Équation (3), 2 paramètres apparaissent: (i) la forme de la fonction $k(x_i)$ et (ii) la fenêtre h qui définit le voisinage V .

A. Choix du noyau

Les noyaux classiques incluent notamment le noyau uniforme, le noyau d'Epanechnikov, le noyau gaussien ou encore le noyau triangulaire. À fenêtre égale, le noyau gaussien présente la densité la plus élevée à l'origine par rapport au noyau triangulaire et au noyau d'Epanechnikov. Cette relation s'inverse logiquement

à une certaine distance en raison de la contrainte de l'Équation (2). Le noyau gaussien donne plus de poids aux points très proches alors que l'Epanechnikov tend à distribuer la densité plus loin et donc à lisser plus fort les données. Cependant bien que différents, de nombreux auteurs (Silverman, 1986; Bailey et Gattrell, 1995; O'Sullivan et Unwin, 2002; Schabenberger et Gotway, 2005; O'Sullivan et Wong, 2007 cités par Xie et Yan, 2008) suggèrent que la forme du noyau est asymptotiquement de moindre importance que la fenêtre sur le résultat final.

En plus de la distance mesurée le long du réseau, l'utilisation d'un noyau classique nécessite toutefois une seconde correction pour une application sur réseau. En effet, l'aire totale sous la courbe devient une fonction du nombre de branches du carrefour comme illustré en Figure 4. L'intégrale de la fonction de densité vaut 1 pour les sections de routes sans carrefour mais descend jusqu'à 0,5 pour les voies sans issue et monte à 1,5 pour un carrefour à 3 routes. L'Équation (3) est donc modifiée par un facteur de correction basé sur la longueur l des segments sous-jacents pour assurer que l'aire sous la courbe reste à 1.

$$K^*(x_j) = \frac{1}{c(x_j)} \sum_{i \in V_j} k(x_i) n_i \quad (4)$$

avec

$$c(x_j) = \sum_{i \in V_j} l_i k(x_i)$$

Une comparaison des valeurs $K(x)$ et $K^*(x)$ est présentée en Figure 4 pour des simulations de réseaux avec 1 accident par point de mesure. Bien que les accidents soient uniformément distribués à une distance constante de 100 m, l'image fournie par l'estimateur à noyau $K(x)$ tend à exagérer la densité aux carrefours et à sous-estimer la densité sur les voies sans issue. Cette exagération de la densité aux carrefours est d'autant plus forte que le nombre de voies qui se croisent est élevé. À l'inverse, l'estimateur $K^*(x)$ corrigé par le facteur d'échelle se comporte plus conformément aux attentes et permet de mieux rendre compte de la distribution uniforme des accidents. Notons finalement que $c(x)$ correspond à la somme de Riemann ou aire sous la courbe estimée par la somme des aires des rectangles de longueur égale à la longueur des segments et de largeur égale à la valeur de densité à une distance d .

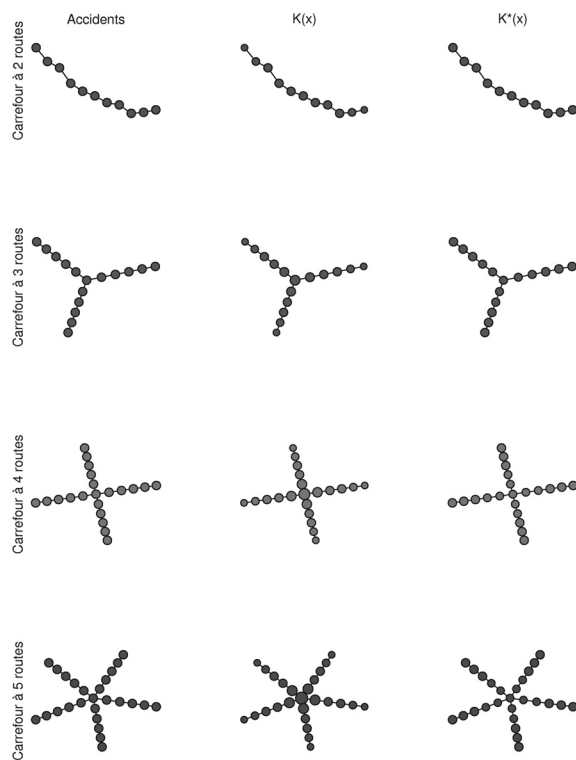


Figure 4. Fréquence des accidents, $K(x)$ et $K^*(x)$ sur base d'un noyau gaussien et une fenêtre de 1,5. Les accidents présentent une distribution uniforme et sont répartis de manière équidistante le long des réseaux simulés. Les bouts de branches du réseau représentent des voies sans issues. La surface du symbole est proportionnelle linéairement à la valeur représentée.

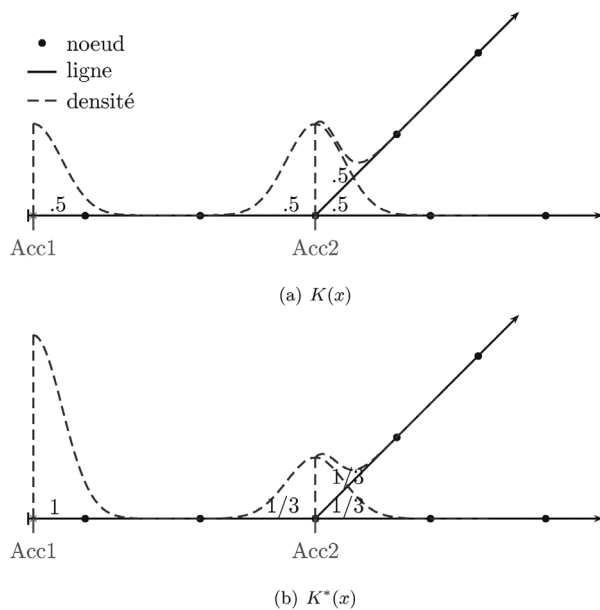


Figure 5. Noyau gaussien lors de l'application à un réseau avec carrefour et voie sans issue. Les aires sous la courbe de l'accident 1 sur la voie sans issue et de l'accident 2 au carrefour de 3 routes sont respectivement de 0,5 et 1,5 pour l'estimateur $K(x)$ alors que les valeurs sont de 1 dans les deux cas pour $K^*(x)$.

Cette correction pose toutefois problème. Okabe *et al.* (2009) démontrent en effet que les estimateurs à noyau sur réseau appartenant à la famille des *similar shape kernel function* dont fait partie $K^*(x)$, sont tous biaisés mais qu'ils présentent des caractéristiques intéressantes dont l'unimodalité, l'égalité de la densité à distance égale, l'isotropie et la symétrie par rapport au centre du noyau. Les alternatives proposées par Okabe *et al.* (2009) pour obtenir un estimateur non biaisé consistant en la définition de deux nouvelles fonctions: (i) l'*equal split kernel* et (ii) l'*equal split continuous kernel*.

L'*equal split kernel* voit sa densité divisée par le nombre de routes présentes à chaque carrefour. Il présente donc des discontinuités et des asymétries aux carrefours avec des sauts dans la densité. Comme nous ne disposons pas d'informations sur le risque d'accident sur chaque bras du carrefour ou encore du nombre de véhicules sur chacun des éléments du carrefour, nous n'avons aucune raison d'allouer des densités différentes à des points équidistants de l'accident. Une solution pourrait être trouvée en estimant l'importance du trafic selon la hiérarchie des routes et constitue certainement une évolution possible de notre méthode.

L'*equal split continuous kernel* est une version corrigée de l'*equal split kernel* pour atteindre la continuité. Cette correction conduit à la définition d'un estimateur dont le mode n'est pas nécessairement le point d'occurrence de l'accident, dont la forme n'est pas symétrique par rapport au site de l'accident et surtout dont l'implémentation est très lourde en termes de temps de calcul. Pour l'instant, cette solution n'a pas encore été appliquée de manière extensive sur des données et est difficilement implémentable techniquement à notre niveau pour l'entière du réseau numéroté belge. Cependant, elle mériterait certainement d'être investiguée plus avant dans des évolutions futures de notre méthodologie.

B. Choix de la fenêtre

Plusieurs éléments bibliographiques permettent d'identifier un intervalle de variation pour la fenêtre optimale. Pour des accidents de la route, Xie et Yan (2008) testent des valeurs entre 20 et 2.000 m mais n'avancent pas de choix optimal. Okabe *et al.* (2009) optent pour une valeur de 200 m sans aucune justification ni motivation pour ce choix. Finalement, Steenberghen *et al.* (2010) travaillent sur des données belges avec des valeurs de 25 à 500 m et concluent que tout dépend de l'application suivant qu'elle soit locale ou plus globale. L'utilisateur est donc laissé libre de choisir dans l'intervalle [25, 500] m selon l'échelle de travail.

Des éléments logiques tels que la distance d'arrêt sur route humide ou encore la distance utilisée en conception routière sont également informatifs pour déterminer la fenêtre optimale. La distance d'arrêt est fonction

non seulement de la vitesse du véhicule qui détermine la distance de freinage mais aussi du temps de réaction du conducteur avant d'actionner le système de freinage et de plusieurs autres variables telles que notamment les conditions de la chaussée, les réflexes et l'attention du conducteur ou encore la qualité et l'usure de la gomme des pneus. Plusieurs relations empiriques existent pour déterminer ces différentes distances, citons notamment les relations utilisées par le Sétra (Vertet et Giausserand, 2006) pour des routes planes

$$\begin{aligned} D_{\text{réaction}} &= 2v \\ D_{\text{freinage}} &= \frac{v^2}{2gf} \\ D_{\text{arrêt}} &= D_{\text{réaction}} + D_{\text{freinage}} \end{aligned} \quad (5)$$

où v est la vitesse exprimée en [m/s], g est l'accélération de la pesanteur et f le coefficient de frottement longitudinal de la route. Par convention en conception routière, une route humide peu adhérente avec des pneumatiques usagés est considérée (Vertet et Giausserand, 2006) et f prend alors une valeur comprise entre 0,31 et 0,46. Le Tableau 1 donne un aperçu de ces distances de freinage, des temps de réaction du conducteur et finalement de la distance d'arrêt selon différentes conditions de chaussée. Une distance d'arrêt de 300 m sur autoroute n'est donc pas totalement inconcevable si le temps de réaction du conducteur augmente à cause d'inattentions diverses telles que GPS et GSM, que la route est détrempée et en mauvaise état et que le couple freins/pneus du véhicule est en piètre état également.

Finalement, le variogramme constitue également une indication pour choisir la taille de la fenêtre à partir d'éléments empiriques. Pour rappel, le variogramme décrit la dépendance spatiale d'une variable par le biais d'une mesure de dissimilitude entre toutes les paires de points séparés par un intervalle de distance de 0 à l'infini. Dans notre cas, nous étudierons donc la différence du nombre d'accidents observés entre

deux points de mesure en fonction de la distance. L'estimateur classique du semi-variogramme proposé par Matheron (1963) est défini par

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i,j \in N(h)} (Z(s_i) - Z(s_j))^2$$

avec $N(h)$ le nombre de paires de points séparés par une distance h et $Z(s_i)$ la valeur de la variable Z mesurée au site s_i .

L'estimateur du semi-variogramme pour les AccRLC est donné en Figure 6. Avant tout, les points de mesure se trouvant à moins de 100 m l'un de l'autre sont, par définition, relativement rares pour des bornes hectométriques et induisent donc un saut entre 0 et 100 m sur l'axe des abscisses. Pour vérifier la robustesse du variogramme, nous avons estimé sa valeur pour des intervalles de distance de 50 à 150 m par pas de 25 m. A première vue, la variance entre deux sites de mesure augmente jusqu'à atteindre un palier aux alentours de 300 m. Notons également que l'effet pépète, c'est-à-dire la valeur du variogramme à une distance nulle, semble très important.

Qualité de la chaussée	f	70 km/h	90 km/h	120 km/h
Béton bitumeux propre et sec	0.8	63.0	89.8	137.5
Revêtement sec de qualité moyenne	0.7	66.4	95.5	147.6
Pavé sec	0.6	71.0	103.1	161.1
Béton bitumeux humide	0.4	87.1	129.6	208.2
Revêtement humide de qualité moyenne	0.35	93.9	141.0	228.5
Pavé humide	0.3	103.1	156.2	255.4

Tableau 1. Distance d'arrêt [m] en fonction de la vitesse pour un temps de réaction moyen de 2 s et des conditions variables de chaussée

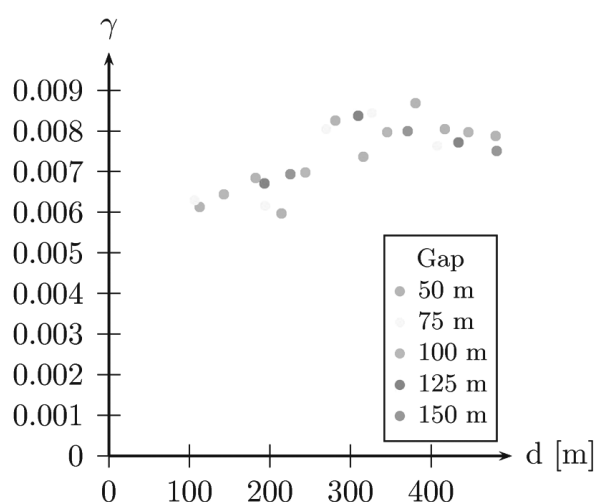


Figure 6. Variogramme empirique

De l'ensemble des éléments, qu'ils soient empiriques, bibliographiques ou bien théoriques, une distance de 300 m semble pertinente dans le cas de l'étude des AccRLC sur les routes numérotées.

C. Simulations de Monte-Carlo

Comme nous l'avons déjà avancé dans ce travail, les accidents de la circulation peuvent présenter un comportement stochastique. En outre, comme le soulèvent Xie et Yan (2008), Truong et Somenahalli (2011) et Chainey *et al.* (2008), le KDE présente le désavantage de ne pas estimer de significativité des valeurs calculées et par là même ne pas offrir de seuils naturels au-dessus desquels une valeur induit un *hotspots*. Pour vaincre cette faiblesse et tester la significativité des valeurs de K observées vis-à-vis d'une situation spatialement aléatoire (Complete Spatial Randomness, CSR), des simulations de Monte-Carlo (MC) sont utilisées. Ces simulations permettent de générer des situations dans lesquelles les accidents respectent l'hypothèse aléatoire spatiale sur le réseau. Il devient alors possible de confronter les valeurs observées de densité lissée avec les valeurs attendues sous l'hypothèse de distribution aléatoire des accidents. À chaque itération du processus :

- le nombre total d'accidents est réparti aléatoirement sur le maillage des points de mesure par un algorithme d'échantillonnage aléatoire avec remise, pondéré selon la longueur des segments;
- les valeurs $K^*(x)$ sont calculées pour tous les points de mesure et enregistrées.

Ensuite, la valeur observée est comparée à la distribution des simulations pour chaque point et fournit ainsi une p-valeur. Le choix du seuil de significativité est fonction premièrement de l'aversion au risque de l'ana-

lyste mais aussi d'un compromis entre d'autres paramètres dont le nombre de points significatifs désirés par les décideurs en fonction des moyens disponibles, du nombre d'accidents inclus dans ces éléments et de la longueur du réseau couverte par les points noirs.

D. Chaînage des points noirs

Le calcul de significativité via les simulations de Monte-Carlo permet d'identifier les points de mesure dont la valeur observée est significativement différente d'une distribution aléatoire des accidents. Ces éléments portent classiquement le nom de *hotspots*. Ils sont indépendants les uns des autres même si le calcul utilise un estimateur à noyau qui inclut le voisinage de chaque point. Via l'analyse de réseau, l'information topologique des données spatiales sert à unir les points significatifs contigus les uns aux autres en un seul et même objet que nous nommerons *hotzone*. Les points de mesures avec une valeur significative seront donc désormais dénommés *hotzones* s'ils sont contigus alors que les points isolés porteront le nom de *hotspots*.

Cette étape de chaînage des éléments utilise un algorithme de reconnaissance des composantes du réseau formé par l'ensemble des points de mesures avec une valeur significative. Les composantes de minimum 2 éléments significatifs définissent les *hotzones* alors que les éléments significatifs isolés sont les *hotspots*.

IV. RÉSULTATS

Dans cette analyse au niveau belge sur les routes numérotées, 79.182 AccRLC de 2006 à 2009 correctement géolocalisés sont utilisés. Cette période de 4 ans est conseillée par Elvik (2008) pour obtenir une image fiable des concentrations d'accidents. Après allocation au point le plus proche, le minimum observé par point de mesure est de 0 alors que le maximum est de 47 accidents avec lésions corporelles. La distribution du nombre d'accidents par point de mesure suit approximativement une distribution de Poisson avec une densité concentrée sur les effectifs faibles proches de 0.

Les résultats présentés en Figures 7 et 8 sont calculés avec un noyau gaussien de 300 m de fenêtre et un calcul de significativité des valeurs de K estimées sur base de 1.000 simulations de Monte-Carlo. Au niveau national, 10.768 points de mesure ressortent comme significatifs au seuil $\alpha < 0,001$ (voir Tableau 2). Selon nos définitions, ces points significatifs se répartissent en 1.315 *hotzones* et 2.172 points de mesures significatifs isolés. Les 1.315 *hotzones* sont donc composées de 8.596 points de mesures, couvrent 532 km du réseau routier et contiennent 21.810 accidents. En terme relatif, cela représente près de 27,54% des accidents pour seulement 3,18% du réseau routier des routes numé-

Signif.	Points significatifs		Hotzones						Hotspots					
			Longueur couverte		AccRLC		Longueur couverte		AccRLC					
[α]	[n]	[n]	[n points]	[km]	[% Total]	[n]	[% Total]	[n]	[km]	[% Total]	[n]	[% Total]		
<0,001	10.768	1.315	8.596	532	3,18	21.810	27,54	2.172	134	0,80	8.336	10,53		
0,001	12.680	1.500	10.328	645	3,86	24.581	31,04	2.352	148	0,89	8.587	10,84		
0,002	13.862	1.641	11.359	713	4,26	26.168	33,05	2.503	168	0,96	8.818	11,14		
0,003	14.731	1.742	12.140	765	4,58	27.296	34,47	2.591	167	1,00	8.877	11,21		
0,004	15.486	1.811	12.801	810	4,85	28.235	35,66	2.685	175	1,05	8.989	11,35		
0,005	16.143	1.887	13.338	846	5,06	28.977	36,60	2.805	184	1,10	9.194	11,61		
0,006	16.659	1.948	13.774	877	5,25	29.587	37,37	2.885	190	1,13	9.279	11,72		
0,007	17.185	1.998	14.248	911	5,45	30.218	38,16	2.937	194	1,16	9.342	11,80		
0,008	17.637	2.042	14.636	938	5,61	30.721	38,80	3.001	200	1,19	9.399	11,87		
0,009	18.072	2.091	15.021	964	5,77	31.222	39,43	3.051	205	1,22	9.401	11,87		
0,010	18.452	2.148	15.369	988	5,91	31.687	40,02	3.083	208	1,24	9.382	11,85		

Tableau 2. Seuil α de significativité et classification des points significatifs en hotzones et hotspots

rotées. A propos des *hotspots*, ce sont 8.336 AccRLC (soit 10,53%) qui sont concernés et 134 km (soit 0,8%) du réseau routier numéroté. Le Tableau 2 décrit l'évolution de ces différents paramètres pour des valeurs de significativité α comprises entre 0 et 0,01. Conformément aux attentes, ces chiffres illustrent bien la concentration spatiale des AccRLC. Un effet Pareto apparaît même pour les *hotzones* avec 3,18% du réseau routier qui concentrent 27,54% des accidents. Ensuite, le nombre de zones noires calculées est en relation directe avec le seuil de significativité sur l'intervalle [0-0,01]. Plus celui-ci augmente, plus le nombre de *hotspots* et de *hotzones* est important.

L'interprétation globale de la Figure 7 est conforme aux grandes tendances avancées par l'Institut Belge pour la Sécurité Routière (IBSR) (Casteels *et al.*, 2010) avec:

- des concentrations (*hotzones* et *hotspots*) plus nombreuses à l'entrée des grandes agglomérations belges et dans les zones urbanisées qu'en rase campagne;
- une différence nette entre Wallonie et Flandre traduisant le risque supérieur d'accident de la Flandre bien que la gravité soit supérieure en Wallonie;
- une majorité des *hotzones* et *hotspots* prennent place sur les routes nationales et régionales;

À l'échelle locale de la Figure 8, un rapport est rédigé pour chaque *hotzone* et *hotspot* avec une analyse détaillée des données des accidents. Une fiche descriptive est ainsi produite avec le nombre de victimes par catégorie (tués, blessés graves, blessés légers), l'âge, le sexe, et

le type de chacun des usagers impliqués, la longueur de la zone, les conditions particulières observées lors des accidents (lumières, qualités de la chaussée, etc.) et une carte de l'environnement immédiat. Ces fiches constituent l'aboutissement de notre méthodologie et se présentent alors comme un outil pratique pour le choix et la mise en oeuvre de mesures correctrices.

V. CONCLUSIONS

La sécurité routière est un des points majeurs d'attention en Belgique tout comme en Europe. Les moyens budgétaires limités actuels imposent des choix de la part des décideurs et des gestionnaires du réseau. Il convient d'identifier les portions du réseau routier les plus dangereuses et problématiques nommées communément points noirs. Cependant, la notion même de point noir reste sujette à de vives discussions entre les experts du domaine bien que des définitions fonctionnelles existent dans plusieurs pays (Elvik, 2008).

Plusieurs outils utilisés par les autorités en Belgique négligent plusieurs dimensions importantes des données d'accidents de la route, notamment l'aspect réseau dont elles proviennent, l'aspect stochastique des accidents, les éventuelles erreurs de localisation de ceux-ci, et rendent impossible la définition de zones noires d'accidents de la route. Des évolutions et de nouvelles méthodes apparaissent régulièrement dans la littérature mais aucune n'a été appliquée, à notre connaissance, à l'échelle d'un pays pour une fenêtre temporelle de 4 ans. Ainsi, Elvik (2008) propose de travailler avec des modèles bayésiens et décourage l'utilisation des fenêtres mobiles mais n'applique nullement ses idées et concepts à un cas pratique. Flahaut *et al.* (2003) et Manepalli *et al.* (2011) travaillent avec des mesures

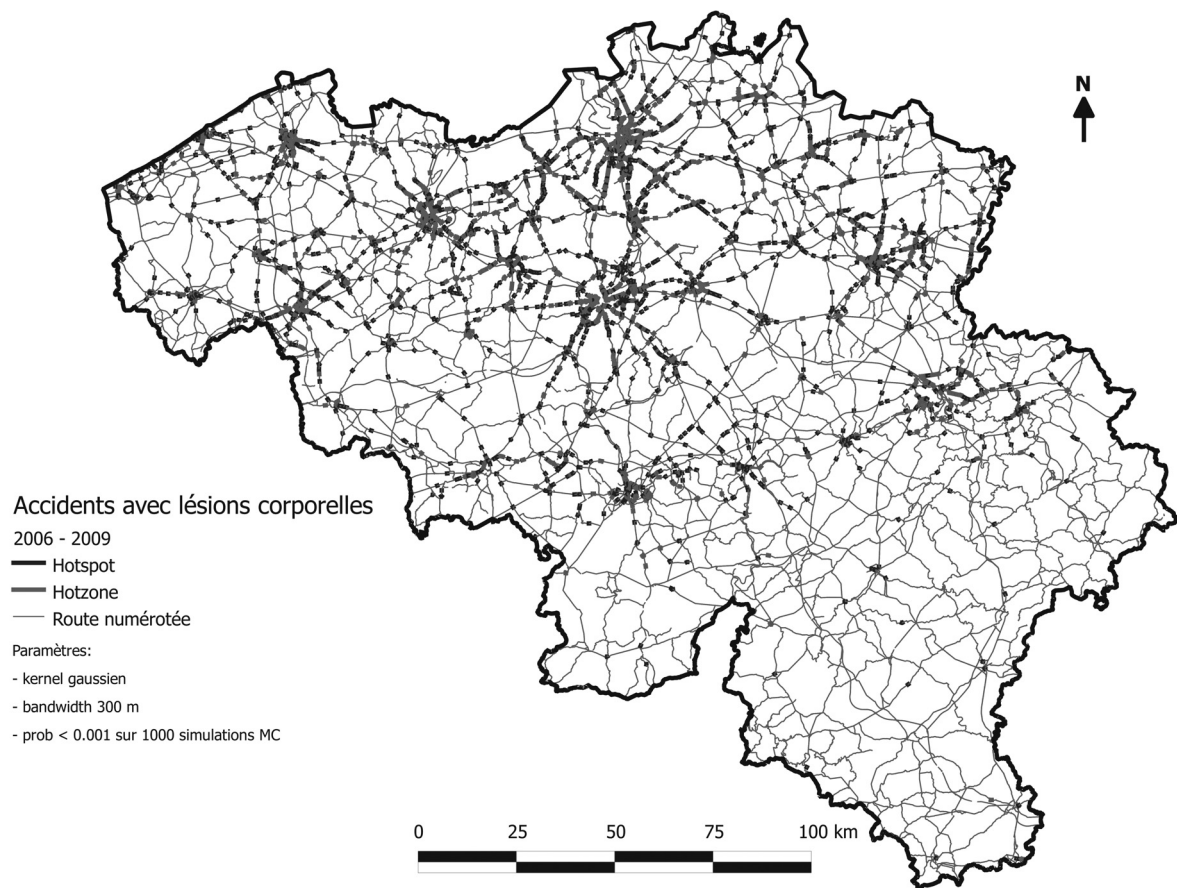


Figure 7. Aperçu global de la distribution des hotspots et hotzones sur les routes numérotées belges

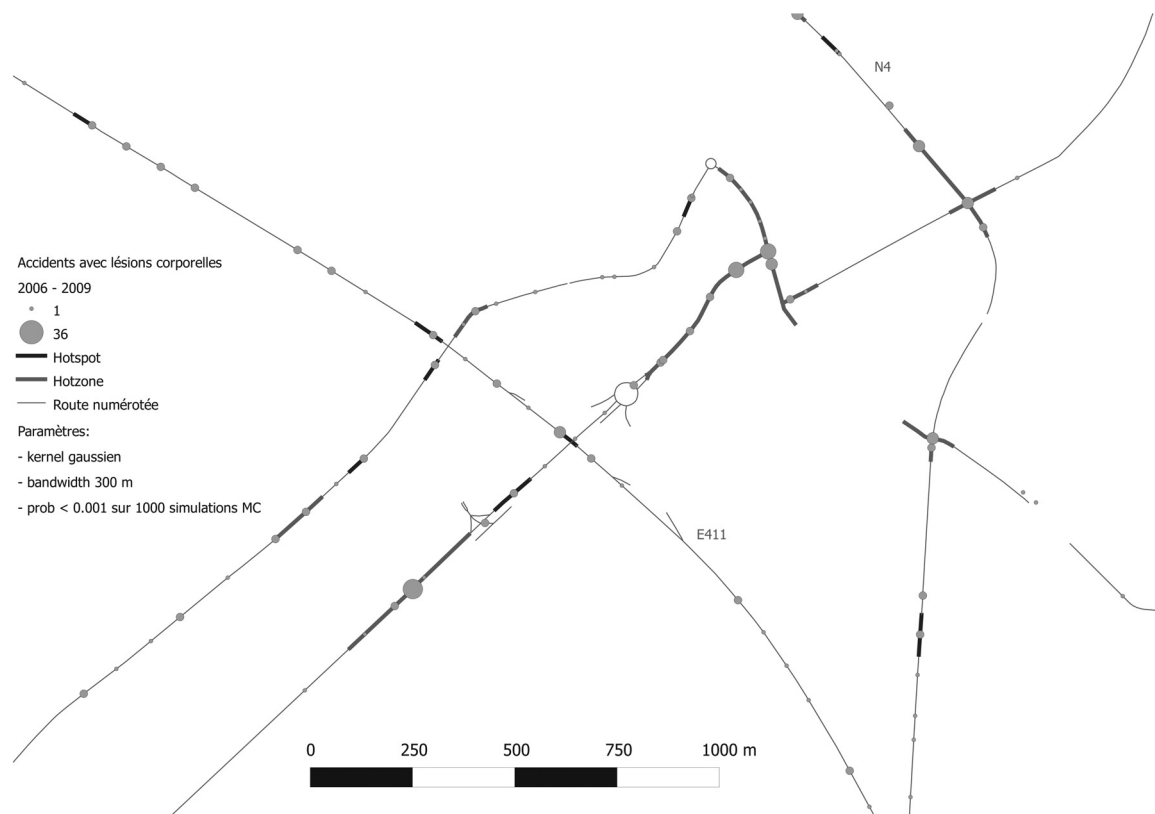


Figure 8. Détails de la distribution des hotspots et hotzones aux alentours du croisement de la E411 et la N4 sur la commune de Wavre

d'autocorrélation spatiale mais au niveau d'une seule route. Moons *et al.* (2009) étudient une province belge mais uniquement au niveau de 2 autoroutes. Finalement, Okabe *et al.* (2009), Xie et Yan (2008) et Steenberghen *et al.* (2010) analysent une seule ville. Ici, notre attention se porte sur le développement d'une solution permettant de considérer l'entière du réseau routier principal d'un pays. Pour plusieurs raisons motivées dans ce travail, nous proposons une méthodologie mixte en trois étapes: (i) la fonction de densité des événements ponctuels sur le réseau est évaluée par un KDE modifié pour tenir compte de l'aspect réseau des accidents de la route dont le choix des paramètres est justifié tant empiriquement que théoriquement, (ii) la significativité des valeurs lissées observées est calculée par des simulations MC et (iii) des outils d'analyse de réseaux permettent de classer les éléments significatifs en *hotspots* et les *hotzones*.

La méthodologie est ensuite testée sur l'entière des routes numérotées belges pour la période 2006-2009. Les résultats se montrent conformes aux constatations globales des autres partenaires actifs dans le domaine de la sécurité routière en Belgique. La méthode est également opérationnalisée par la mise en place de fiches descriptives pour chaque point et zone noire.

Naturellement, plusieurs évolutions de la méthode sont possibles dont l'intégration d'information du trafic routier ou encore la considération de nouvelles fonctions par le KDE. De même, la considération d'une seconde période d'accidents (2010-2013) permettrait d'évaluer la stabilité des résultats et le pouvoir prédictif des *hotspots* et *hotzones* passés pour les configurations futures d'accidents de la route.

BIBLIOGRAPHIE

- Anderson, T. K. (2009). Kernel density estimation and k-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41 (3): 359–364.
- Antoine, D. (2010). *Zones à risque et tronçons dangereux 2005 - 2009*. Rapport technique. Namur : Service Public de Wallonie.
- Bailey, T. C. et Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. Essex : Longman.
- Casteels, Y., Martensen, H., Merckx, F., Nuyttens, N., Riguelle, F. et Thijs, R. (2010). *Satistiques de sécurité routière 2008*. Bruxelles : IBSR, Observatoire pour la sécurité routière.
- Castellà, J. et Pérez, J. (2004). Sensitivity to punishment and sensitivity to reward and traffic violations. *Accident Analysis & Prevention*, 36(6): 947–952.
- Chainey, S., Tompson, L. et Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21: 4–28.
- Elvik, R. (2008). A survey of operational definitions of hazardous road locations in some european countries. *Accident Analysis & Prevention*, 40: 1830–1835.
- Flahaut, B., Mouchart, M., Martin, E. S. et Thomas, I. (2003). The local spatial autocorrelation and the kernel method for identifying black zones a comparative approach. *Accident Analysis & Prevention*, 35: 991–1004.
- Geurts, K. et Wets, G. (2003). *Black spot analysis methods: literature review*. Diepenbeek : Steunpunt Verkeersveiligheid bij stijgende mobiliteit.
- Iversen, H. et Rundmo, T. (2002). Personality, risky driving and accident involvement among norwegian drivers. *Personality and Individual Differences*, 33: 1251–1263.
- Manepalli, U. R. R., Bham, G. H. et Kandada, S. (2011). Evaluation of hotspots identification using kernel density estimation (k) and getis-ord (gi*) on i-360. *3rd International Conference on Road Safety and Simulation, Indianapolis, 14-16 September 2011*.
- Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58: 1246–1266.
- Moons, E., Brijs, T. et Wets, G. (2009). Improving moran's index to identify hot spots in traffic safety. In Murgante, B., Borruoso, G. et Lapucci, A., editors, *Geocomputation and Urban Planning*, volume 176 of *Studies in Computational Intelligence* (p. 117–132). Berlin - Heidelberg : Springer.
- Okabe, A., Satoh, T. et Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a gis-based tool. *International Journal of Geographical Information Science*, 23 (1): 7–32.
- O'Sullivan, D. et Unwin, D. J. (2002). *Geographic Information Analysis*. New Jersey : John Wiley.
- O'Sullivan, D. et Wong, D. W. S. (2007). A Surface-Based Approach to Measuring Spatial Segregation. *Geographic Analysis*, 39 (2): 147–168.
- Romano, U. (1997). *Atlas de l'insécurité routière 1990-1994*. Namur : Ministère Wallon de l'Équipement et des Transports (MET).
- Romano, U. et Heuchenne, D. (1996). *Modèle mathématique d'évaluation de l'insécurité routière*. Namur : Ministère Wallon de l'Équipement et des Transports (MET).
- Schabenberger, O. et Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton : Chapman & Hall/CRC.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London : Chapman Hall.
- Smith, S. C. et Bruce, C. W. (2008). *CrimeStat III User-book*. Washington: The National Institute of Justice.
- Steenberghen, T., Aerts, K. et Thomas, I. (2010). Spatial clustering of events on a network. *Journal of transport geography*, 18: 411–418.
- Truong, L. T. et Somenahalli, S. V. C. (2011). Using gis to identify pedestrian-vehicle crash hot spots and unsafe bus stops. *Journal of Public Transportation*, 14(1): 99–114.

- Vertet, M. et Giausserand, S. (2006). *Comprendre les principaux paramètres de conception géométrique des routes*. Rapport Sétra. Bagnaux : République Française - Ministère des Transports, de l'Équipements, du Tourisme et de la Mer - Service d'Études techniques des routes et autoroutes (Sétra). Consultable sur <http://www.setra.equipement.gouv.fr>.
- Xie, Z. et Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32 (5): 396–406.

Coordonnées des auteurs :

David DABIN, Christiane DICKENS
& Paul WOUTERS,
Analystes-stratégiques, Police Fédérale
- Direction de l'Information Policière Opérationnelle
- Service d'Analyse Stratégique.
Rue Fritz Toussaint 8, 1080 Ixelles,
david.dabin.5968@police.be.