Identification of factors contributing to the occurrence of crashes at high-risk locations

RA-MOW-2011-027

T. De Ceunynck, E. De Pauw, S. Daniels, T. Brijs, E. Hermans & G. Wets

Onderzoekslijn Infrastructuur

DIEPENBEEK, 2013.
STEUNPUNT MOBILITEIT & OPENBARE WERKEN
SPOOR VERKEERSVEILIGHEID

Documentbeschrijving

Rapportnummer:	RA-MOW-2011-027
Titel:	Identification of factors contributing to the occurrence of crashes at high risk locations
Auteur(s):	T. De Ceunynck, E. De Pauw, S. Daniels, T. Brijs, E. Hermans & G. Wets
Promotor:	Prof. dr. Tom Brijs
Onderzoekslijn:	Infrastructuur
Partner:	Universiteit Hasselt
Aantal pagina's:	40
Projectnummer Steunpunt:	**Error! Reference source not found.**
Projectinhoud:	The purpose of this study is to identify the most important underlying factors that determine the number of injury crashes at dangerous intersections. Furthermore, the models in this report will be used to correct for the regression-to-the-mean effect in a future before-after evaluation study of dangerous locations

Uitgave: Steunpunt Mobiliteit & Openbare Werken – Spoor Verkeersveiligheid, mei 2012.

Steunpunt Mobiliteit & Openbare Werken
Spoor Verkeersveiligheid
Wetenschapspark 5
B 3590 Diepenbeek

T 011 26 91 12
F 011 26 91 99
E info@steunpuntmowverkeersveiligheid.be
I www.steunpuntmowverkeersveiligheid.be

Samenvatting

Aan de hand van ongevalgegevens van 1997-1999 heeft de Vlaamse overheid 1014 locaties geïdentificeerd als gevaarlijke punten. Dit rapport geeft een aantal verklarende modellen weer die gebouwd werden aan de hand van een dataset van 601 van deze locaties. De bedoeling van het rapport is om de belangrijkste onderliggende factoren te bepalen die het aantal letselongevallen op deze locaties beïnvloeden. Ook zijn de modellen in dit rapport gebruikt in een recente voor-en-na studie over gevaarlijke punten van het Steunpunt Mobiliteit & Openbare Werken, Spoor Verkeersveiligheid om te corrigeren voor het regressie-naar-het-gemiddelde effect.

Negatief binomiale modellen worden gebruikt omdat ze kunnen omgaan met de overdispersie die aanwezig is in de dataset. De afhankelijke variabele is het aantal letselongevallen of het aantal zware letselongevallen dat op het kruispunt gebeurd is in de periode 2000-2003. Deze periode is gekozen om het regressie-naar-het-gemiddelde effect op te vangen dat aanwezig is in de ongevalgegevens van 1997-1999. Ongevaldata tot en met 2003 kunnen gebruikt worden omdat de eerste van deze gevaarlijke punten in 2004 werd heringericht in het kader van een grootschalig project van de Vlaamse overheid, waarin 800 gevaarlijke punten worden heringericht om de verkeersveiligheid te verbeteren. Daardoor is overeenstemming tussen de verzamelde ongevallendata en kruispuntkenmerken verzekerd. Informatie over 38 potentieel belangrijke onafhankelijke variabelen werd verzameld. De modellen zijn gebouwd volgens een iteratief proces om problemen met ontbrekende waarden zo goed mogelijk op te vangen. De kwaliteit van de modellen wordt beoordeeld aan de hand van de AIC-waarde. De eindmodellen worden gecontroleerd op multicollineariteitsproblemen met de type II tolerance test.

Een aantal modellen en submodellen werden gefit. Twee algemene modellen voor alle ongevallen werden gefit, waarvan één op basis van alle letselongevallen en één enkel gebaseerd op ongevallen met ernstig gewonde of dodelijke slachtoffers. Submodellen werden gefit voor voorrangsgeregelde en lichtengeregelde kruispunten, voor driearmige en vierarmige kruispunten, voor kruispunten binnen de bebouwde kom en buiten de bebouwde kom, en voor drie verschillende types categorisering van de hoofdweg. Hoewel er een aantal verschilpunten zijn tussen de modellen, kunnen een aantal algemene conclusies getrokken worden.

Vele ongevalsvoorspellingsmodellen geven aan dat expositie het grootste deel van de structurele variatie in de data verklaart. Ook in deze studie zijn de expositievariabelen LOGVOLMAJTOTAL en LOGVOLMINTOTAL zeer belangrijke verklarende variabelen. De modellen in deze studie tonen een positief maar inelastisch verband aan tussen het aantal ongevallen en de voertuigintensiteiten.

Wat betreft de geometrische kruispuntkenmerken, blijkt de aanwezigheid van een middenberm op de hoofd- of onderliggende weg te correleren met een hoger aantal ongevallen. De variabele SIGNALS is enkel aanwezig in het algemene model voor ongevallen met zwaargewonde en dodelijke slachtoffers. De variabele wijst aan dat er een groter aantal ernstige letselongevallen gebeuren op de lichtengeregelde kruispunten in deze dataset dan op niet-lichtengeregelde kruispunten. Driearmige kruispunten blijken een lager aantal letselongevallen te hebben dan kruispunten met meer dan drie armen. Verder blijken kruispunten op wegen met een snelheidslimiet van 50 km/u vaak een significant hoger aantal ongevallen te hebben dan de andere categorieën. Het is daarom aanbevolen dat gevaarlijke kruispunten op wegen met een snelheidslimiet van 50 km/u speciale aandacht krijgen van wegontwerpers en beleidsmakers. Variabelen die de aanwezigheid van voorzieningen voor zwakke weggebruikers beschrijven verschijnen ook in een aantal modellen, maar wijzen niet consistent in dezelfde richting. De bestaande wetenschappelijke literatuur is echter ook niet eenduidig over de invloed van voorzieningen voor zwakke weggebruikers op het aantal ongevallen. Een aantal modellen geven een lager aantal ongevallen aan wanneer één of meerdere kruispunttakken niet loodrecht aansluit in vergelijking met kruispunten waar wel alle takken loodrecht aansluiten. Hoewel dit eerder onverwacht lijkt, komen een aantal andere studies ook tot

gelijkaardige bevindingen. Het aantal rijstroken van de kruispunttakken is aanwezig in enkele modellen, maar is vaak niet aanwezig in het eindmodel omwille van multicollineariteitsproblemen.

De functionele wegencategorisering is een aspect dat zelden wordt opgenomen in studies die gebruik maken van ongevalsvoorspellingsmodellen. De variabelen RDCATMAJ en RDCATMIN verschijnen in deze studie echter in verschillende eindmodellen. Kruispunten op primaire wegen blijken vaak een significant hoger aantal letselongevallen te tellen dan andere wegcategorieën. Deze kruispunten verdienen dan ook speciale aandacht. Over het algemeen lijkt de tendens dat kruispunten van een hogere wegcategorie een hoger aantal ongevallen hebben. Alleen de categorie MAIN volgt dit patroon niet, wat verklaard kan worden door het feit dat dit in feite kruispunten zijn met op- en afritten van hoofdwegen. Één submodel geeft een hoger aantal ongevallen aan wanneer het kruispunt tot het functionele en/of recreatieve fietsroutenetwerk behoort, maar de variabele zou ook een proxy kunnen zijn voor een hogere intensiteit van fietsers.

In een aantal model geven de grondgebruikvariabelen PUBLIC, RESIDENTIAL en ECONOMIC aan dat er een hoger aantal letselongevallen gebeurt op kruispunten waar het grondgebruik bestaat uit openbare voorzieningen, woongebied of commerciële activiteiten respectievelijk. Deze variabelen fungeren mogelijk echter als proxy voor de expositie van zwakke weggebruikers. BUILTUP is aanwezig in het submodel voor secundaire wegen. De variabele wijst op een hoger aantal ongevallen wanneer bebouwing aanwezig is rond het kruispunt, wat de veiligheidsproblemen die veroorzaakt worden door de aanwezigheid van lintbebouwing benadrukt.

Het is aanbevolen dat toekomstig onderzoek gegevens over de expositie van andere types weggebruikers dan enkel motorvoertuigen ook in rekening brengt. Ook kunnen nauwkeurigere telgegevens van de voertuigintensiteiten worden verzameld, of kan een meer gesofisticeerde expositiemaatstaf worden gemeten, zoals bijvoorbeeld het aantal ontmoetingen tussen weggebruikers. Tenslotte kan toekomstig onderzoek ook focussen op het verder onderzoeken van de causaliteit van de correlaties die werden gevonden in de dataset.

English summary

Using crash data of the years 1997-1999, the Flemish government identified 1014 road locations as dangerous locations. This report presents a number of explanatory models, fitted on a dataset of 601 of these locations. The purpose of the report is to identify the most important underlying factors that determine the number of injury crashes at these dangerous intersections. Furthermore, the models in this report have been used to correct for the regression-to-the-mean effect in a before-after evaluation study of dangerous locations by the Flemish Policy Research Centre Mobility & Public Works, track Traffic Safety.

Negative binomial models are used to deal with the overdispersion that is present in the dataset. The dependent variable is the number of injury crashes or severe injury crashes that has occurred at the location in the 2000-2003 period. The period has been chosen to account for the regression-to-the-mean effect that is known to be present in the 1997-1999 period that is used to select the locations. Crash data up to and including 2003 could be used because the first of these locations were reconstructed in 2004 in a large scale project of the Flemish government to reconstruct 800 dangerous locations to improve their safety performance. This way, full correspondence between the collected crash data and the intersection characteristics is assured. Information about 38 possible independent variables is collected. The models are fit using an iterative process to overcome missing data issues. The fit of the models is judged using the AIC-value. The final models are checked for multicollinearity using the type II tolerance test.

A number of models have been fit. Two general models for all intersections have been fit; one using all injury crashes and one using only serious injury and fatal crashes. Submodels are fit for priority, signalized, four-leg and three-leg intersections, for intersections inside and outside built-up area, and for three different types of major road category intersections. There are a number of differences between the models, but a number of general conclusions can be drawn.

A lot of crash prediction model studies indicate that exposure explains most of the structural variance. In this study, the exposure variables LOGVOLMAJTOTAL and LOGVOLMINTOTAL are also in this study highly important explanatory variables. The models in this study show a positive but inelastic correlation between the number of crashes and the vehicle volumes.

Concerning the geometric variables, the presence of a median on the major or minor road tends to correspond with a higher number of crashes. The variable SIGNALS is only present in the general model for serious injury and fatal crashes. The variable indicates a higher number of serious crashes in case the intersection is signalized. Three-leg intersections on the other hand have a lower number of injury crashes than other types of intersections. Furthermore, intersections with a speed limit of 50 km/h tend to show a significantly higher number of crashes than the other categories. It is therefore recommended that these intersections receive special attention from road designers and policy makers. The presence of facilities for vulnerable road users at the intersections appears in some of the presented models. The patterns are however not consistent. Literature is however not conclusive on this subject either. Furthermore, some of the models indicate a safety benefit for non-perpendicular intersections compared to regular 90° intersections. Although this is rather unexpected, some literature comes to similar findings. The number of lanes of the intersection legs is present in some models, but is often not included in the end model due to multicollinearity issues.

Functional road classification is a variable that is not often included in crash prediction model literature. The variables RDCATMAJ and RDCATMIN appear in quite a number of models in this study however. Intersections of primary roads tend to have a significantly higher number of injury crashes than the other categories. These intersections deserve special attention. Generally, the trend seems to be that a higher road category corresponds with a higher number of crashes. Only the category MAIN does not follow this pattern, which can be explained by the fact that these are actually intersections with

on/off ramps of main roads. One submodel indicates a higher crash rate at intersections that belong to the cycle route network, but the variable could be a proximate for a higher volume of bicyclists.

In a number of models, the land use variables PUBLIC, RESIDENTIAL and ECONOMIC indicate a higher number of injury crashes in case the land use is public facilities, residential or commercial activities respectively. These variables are however likely to function as a proxy for the exposure of vulnerable road users. BUILTUP is present in the submodel for secondary roads. The variable indicates a higher number of crashes when buildings are present, which stresses the safety issues of ribbon development.

It is recommended that future research includes exposure data from other types of road users than only motor vehicles. Furthermore, more accurate motor vehicle counts could be collected, or a more sophisticated exposure measure such as the number of encounters. Also, future research should focus on establishing the causality of the correlations that have been found in the dataset.

Inhoudsopgave

1. INTRODUCTION

Road crashes have become one of the most important problems of modern society. Road crashes are highly undesirable, both from a societal perspective as from an economic perspective. Road crashes do not only harm the ones that are directly involved in the incident, but also indirectly affect the social networks the individual belongs to, and have therefore far-reaching negative influences to society. Furthermore, road crashes have been estimated to constitute an economic loss of about 2-2.5% of the gross national product of developed countries (Elvik, 2000a; Jacobs, Aeron-Thomas, & Astrop, 2000; WHO, 2004). Therefore, it is not surprising that policy makers are highly motivated to improve road safety.

By making use of crash data of the years 1997-1999, the Flemish government identified locations exceeding a certain threshold value as being dangerous. This way, 1014 locations with an exceptionally high number and/or severity of crashes have been identified as dangerous locations. These are all locations with a priority code equal to or higher than 15 during this period. The priority code is defined as follows:

$$PC = 5 * D + 3 * SE + 1 * SL$$

Where D is the number of deadly injured road users, SE the number of seriously injured, and SL the number of slightly injured. Most of these dangerous locations are intersections, which is not surprising because the interactions between road users are more complex at intersections than on road sections, corresponding with a higher risk of errors and crashes (Geurts, 2006).

The Flemish government intended to reconstruct 800 of these locations in the period of 2003-2008. Each of these dangerous locations has been analyzed individually to identify possible safety issues, and a reconstruction plan was designed based on these findings, following a fixed procedure (AWV, 2009). Despite some delays in the program, at this point, most of the locations have been adjusted.

Because of the individual approach of these dangerous locations, little is known about infrastructural or other characteristics that have a general impact on the safety of these locations. In other words, it is not known what characteristics of the dangerous locations generally have a high impact on the number of crashes. However, road authorities and policy makers need a good insight in these variables that explain the crash levels at their roads, because this will allow them to design or redesign roads more safely in the future. The influence of design characteristics on the level of safety are most often investigated by the fitting of cross-sectional risk models. These models explain the variation in road safety performance of the sample of locations by using regression modelling techniques (Daniels et al., 2010). Currently, the models that are most commonly used in road safety research are Poisson or negative binomial regression models (Daniels et al., 2010; Lord & Mannering, 2010; Reurings et al., 2006).

The purpose of this report is to identify the most important factors that explain the variation in the number of crashes between different types of dangerous intersections. To this purpose, state-of-the-art cross-sectional risk models are used, based on injury crash data, traffic flow data and infrastructural properties for a sample of 601 dangerous locations. The study incorporates 38 different independent variables that could have an impact on road safety, some of which rarely have been used in prior studies making use of crash prediction models.

Furthermore, the models that are estimated in this study can be used to estimate the expected crash count of dangerous locations. The expected number of crashes is necessary to be able to correct before-after studies for the so-called regression-to-the-mean effect (Hauer, 1997). Regression-to-the-mean implies that, because of the rare nature of road crashes, an abnormally high number of recorded crashes can result from random fluctuations (Elvik, 1997). In other words, the number of crashes is expected to lower again to the long-term average in the subsequent period without a need for

intervention to achieve this reduction. The risk models presented in this report will therefore be used in a future before-after study evaluating the effectiveness of the reconstruction program of dangerous locations that will be executed by the Policy Research Centre Mobility and Public Works, track Traffic Safety (De Pauw, Daniels, Brijs, Hermans, & Wets, 2012).

The rest of the report is structured as follows. Section 2 describes the data that are used in this study. Section 3 describes the underlying theory behind the study and presents methodological issues. Section 4 presents the study results. These results are discussed and interpreted into more detail in section 5. Section 6 summarizes the main conclusions of the study.

2. DATA COLLECTION

For the purpose of this study, data about 38 variables of 601 dangerous intersections in the region of Flanders are collected. It is not possible to incorporate all 800 dangerous locations because of three reasons. The first reason is that, for some locations, too little information was available to incorporate them in the analyses. Secondly, some of the 800 dangerous locations are road sections instead of intersections. Since it is not possible to analyze them in the same way as intersections (e.g. because they have other geometric characteristics and because they cannot be geocoded as a point-location) it is decided to omit the road sections from the analyses. And thirdly, for other intersections it was not possible to geocode them because it is not possible to determine to which intersection the file from the provincial audit commission, from which the information about the dependent variables has been retrieved, is referring to (e.g. some intersections with on- and off-ramps of main roads).

Data are collected about accident counts, motor vehicle volumes, geometrical characteristics of the road, the road environment and the transportation planning context. The incorporated values and their descriptive statistics are presented in Table 1. With the exception of traffic volume data and crash counts, all variables are dummy variables or categorical variables.

2.1 Traffic volume data

The traffic volumes in the data are the sum of the busiest morning peak hour (7.00 am – 9.00 am) and the busiest evening peak hour on one day (4.00 pm – 6.00 pm). Therefore, the values must be divided by two to obtain an hour estimate. Shifts of 15 minutes were taken. This implies that for the morning peak, the possibilities are 7.00-8.00, 7.15-8.15, 7.30-8.30, 7.45-8.45 and 8.00-9.00. The possibilities for the evening peak are analogous.

The use of a relatively unusual two-hour value instead of a more traditional value for one hour or for a 24-hours period does not have an impact on the parameters because of the model form. For both the main and the minor road, total motor vehicle flow values are collected, as well as the number of vehicles going straight ahead, turning left and turning right.

2.2 Data describing intersection and intersection environment

A literature review showed two reports (Nambuusi, Brijs, & Hermans, 2008; Reurings et al., 2006) that provide a good overview of the most relevant geometrical and environmental characteristics as they appear from previous crash prediction model studies. Geometrical characteristics describe the road-technical attributes of the infrastructure. They include for instance the number of lanes on the roads, the presence of a median, the type of bicycle infrastructure, the right-of-way regulation, the number of legs of the intersection,… Data about the intersection environment describes the land-use of the surrounding area of the intersection. Spatial planning features include the road categorization of the leading roads, and whether the intersection is part of the Flemish bicycle network.

It should be noted that some of the variables (speed limits, number of lanes) that are included as a categorical value could also be treated as a continuous variable. For instance, literature treats the number of lanes sometimes as a continuous variable (e.g. Chin & Quddus, 2003) and sometimes as a categorical variable (e.g. Greibe, 2003). Therefore, both approaches seem valid for modeling the number of crashes. The latter option is chosen, because the authors think there is no strong evidence to expect a

continuous relationship (i.e. systematic raising or lowering) between the number of crashes and the number of lanes, or between speed limits. Since continuous is considered to be a higher measurement level than categorical (Baarda & de Goede, 2001), the authors think that, in the absence of strong evidence for a continuous relationship, the most conservative choice is to treat the variable as categorical.

2.3 Crash data

For the crash data, geo-coded recordings from all reported injury crashes since 1996 are available from the federal government department of economics, FOD ADSEI. The 601 dangerous intersections are first geo-coded by using Google Earth. Using the GIS program TransCAD, the geo-coded intersections are linked with the injury crash data. All injury crashes within 100m from the centre of the intersection are included. Accident data from the years 2000-2003 are used because of the following reasoning. The dangerous intersections are selected based on crash data of the years 1997-1999. However, the selection of dangerous locations based on a relatively short period of time is vulnerable to the so-called "regression-to-the-mean effect", which means that the high number of crashes can be caused by chance (Elvik, 1997). To overcome this issue, it is decided not to use the crash data from 1997-1999. By using a different time period of crash data than the time period that is used to identify the dangerous locations, the regression-to-the-mean effect should be of no influence because a set of locations with an unusually high crash count in one period is expected to return to its long term mean value in the next period (Hauer, 1997). This implies that no systematic chance effects should be present in the models.

The first of these dangerous locations has been reconstructed in the year 2004, which implies that the crash data from the years 2000-2003 can be used for the analyses, because all of them still have the "dangerous" configuration during this period. This way, one can be sure that the collected crash data exactly match the intersection characteristics data. Of course, for some locations that are reconstructed relatively late, data from more years could have been used. However, then the crash counts of the different locations could not be compared. Configuring the measure to, for instance, the number of crashes per year instead of the total number of crashes would only partly solve this issue, because there are confounding trends in transportation in general, and road safety in particular. For instance, there has been a trend of an annual increase in the number of vehicle kilometers, and a trend of an annual reduction in the total number of severe road crashes (Belgian Federal Government, 2010). In total, 7122 injury crashes are included in the dataset for the 601 dangerous intersections. The description of the variables is presented in Table 1.

Table 1 - Variables description

Variable (ABBREVIATION) (additional remarks)	***Descriptive statistics***
Crash data (dependent variable)	
Number of injury crashes at intersection (ACCCOUNT) (4-years total, 2000-2003)	Mean: 11,85; S.D.: 8,80; missing=0
Number of severe injury crashes at intersection (ACCCOUNT_SER) (4-years total, 2000-2003)	Mean: 2,01; S.D.: 1,85; missing=0
Traffic Volume variables	
Traffic volume on the major road – total of turning traffic and traffic straight ahead (VOLMAJTOTAL)	Mean: 3016;S.D.: 1528; missing=87
Traffic volume on the major road – left turning traffic (VOLMAJLEFT)	Mean: 323;S.D.: 354; missing=87
Traffic volume on the major road – right turning traffic (VOLMAJRIGHT)	Mean: 337;S.D.: 358; missing=87
Traffic volume on the major road – traffic driving straight ahead (VOLMAJSTR)	Mean: 2366;S.D.: 1453; missing=87
Traffic volume on the minor road – total of turning traffic and traffic straight ahead (VOLMINTOTAL)	Mean: 1074;S.D.: 1063; missing=115
Traffic volume on the minor road – left turning traffic (VOLMINLEFT)	Mean: 323;S.D.: 354; missing=115
Traffic volume on the minor road – right turning traffic (VOLMINRIGHT)	Mean: 323;S.D.: 325; missing=115
Traffic volume on the minor road – traffic driving straight ahead (VOLMINSTR)	Mean: 428;S.D.: 666; missing=115
Geometry variables	
Number of lanes on major road (sum of different directions) (LANEMAJ)	1=4; 2=312; 3=20; 4=220; 5=1; 6=4; missing=40
Number of lanes on minor road (sum of different directions) (LANEMIN)	1=36; 2=462; 3=6; 4=49; missing=48
Traffic signals at the intersection? (SIGNALS)	Yes=269; No=303; missing=29
Three-leg intersection? (3LEGS)	Yes=130; No=451; missing=20
Four-leg intersection? (4LEGS)	Yes=414; No=167; missing=20
More than four legs intersection? (56LEGS)	Yes=29; No=552; missing=20
Is the intersection a roundabout? (RNDBT)	Yes=7; No=574; missing=20
Is the shape of the 4-leg intersection diagonal? (DIAGON)	Yes=68; No=513; missing=20
Is the 3-leg intersection Y-shaped? (3LEGY)	Yes=12; No=569; missing=20
Presence of median on major road? (MEDIMAJ) (in case of different situations at different legs: "yes")	Yes=225; No=322; missing=54

Presence of median on minor road? (MEDIMIN) (in case of different situations at different legs: "yes")	Yes=89; No=459; missing=53
Allowed speed on major road (SPEEDMAJ) (in case of different situations at different legs: highest speed limit is used)	30=1; 50=128; 70=202; 90=211; 120=6; missing=53
Allowed speed on minor road (SPEEDMIN) (in case of different situations at different legs: highest speed limit is used)	30=2; 50=230; 70=137; 90=153; 120=11; missing=68
Type of cycle facilities (CYCLFAC) (in case of different situations at different legs: "highest" type is used; 0 = no facilities; 1 = cycle lanes = max. 1m away from roadway; 2= separate cycle paths > 1m away from roadway; 3 = grade-separation (bridge or tunnel); "bicycle suggestion lanes" are considered as "no specific facilities" because of their limited legal meaning)	0=110; 1=272; 2=119; 3=54; missing=46
Footpaths present? (PEDFAC)	Yes=309; No=193; missing=99
Public transport stop present at the intersection? (PTFAC)	Yes=189; No=258; missing=154
Vehicle parking possible in immediate environment of intersection? (PARKING)	Yes=313; No=198; missing=90
Crossing facilities present for pedestrians? (PEDCROSS) (in case of different situations at different legs: "yes")	Yes=288; No=234; missing=79
Crossing facilities present for bicyclists? (CYCCROSS) (in case of different situations at different legs: "yes")	Yes=392; No=132; missing=77
Legal status variables	
On functional/recreational cycle route network? (CYCLNETW) (in case of different situations at different legs: "yes")	Yes=487; No=74; missing=40
Road category of the major road according to the existing spatial structure plans (RDCATMAJ) (in case of different situations at two sides: highest category used (MAIN>PRIM>SEC>LOC))	MAIN=16; PRIM=160; SEC=276; LOC=102; missing=47
Road category of the minor road according to the existing spatial structure plans (RDCATMIN) (in case of different situations at two sides: highest category used (MAIN>PRIM>SEC>LOC))	MAIN=14; PRIM=33; SEC=76; LOC=385; missing=93

Land use variables	
Is the location inside built-up area? (INSIDE)	Yes=177; No=380; missing=44
Is the land use at the intersection residential? (RESIDENTIAL)	Yes=405; No=166; missing=30
Is the land use at the intersection nature, wood, agrarian? (RURAL)	Yes=174; No=397; missing=30
Is the land use at the intersection industry, offices, companies? (ECONOMIC)	Yes=84; No=487; missing=30
Is the land use at the intersection public services, schools,…? (PUBLIC)	Yes=24; No=547; missing=30
Are there any buildings present around the intersection? (BUILTUP)	Yes=469; No=70; missing=62

3. METHODOLOGY

3.1 Model form

In regression models, a set of independent variables is used to estimate or predict the value of a dependent variable (Anderson et al., 2005). In this study, the total number of crashes at the intersection (N=601) is the dependent variable, and the intersection characteristics are the independent variables. Traditionally, multiple linear regression (MLR) models have often been used to model road crashes in function of geometric and traffic factors. However, researchers have become aware of some undesirable statistical properties of these models, such as the possibility of estimating negative crash counts (Chin & Quddus, 2003). Because road crashes are in fact discrete, nonnegative and to a certain extent random events, the family of Poisson regression models is thought to be more suitable to model them (Lord & Mannering, 2010).

The traditional Poisson loglinear model has the advantage of simplicity, but some analysis limitations have to be carefully taken into account to use it in a valid way. Preliminary analyses showed that at least one of the limitations is violated. A basic assumption of the Poisson model is that the variance has to be more or less equal to the mean (Washington et al., 2003). In the study sample, this is not the case. Preliminary analyses showed that the variance significantly exceeds the mean for all models. This implies that the data are overdispersed, which is a common phenomenon in studies that include crash data (Daniels et al., 2010; Lord & Mannering, 2010; Nambuusi et al., 2008). Overdispersion does not affect the coefficient estimates, but results in an underestimation of their standard errors (Greibe, 2003). In other words, the significance of the variables is systematically overestimated when a Poisson model is applied in the presence of overdispersion. A possible reason for the arising of overdispersion could be an excess number of zeros in the data (Lord et al., 2005). This is an issue that occurs relatively often in a set of random road locations, and the use of a zero-inflated model can solve this issue (Lord & Mannering, 2010; Lord et al., 2005; Washington et al., 2003). However, since this study makes use of a set of dangerous locations instead of a set of random locations, the number of zeros in the dataset is very low (n=5) and therefore, zero-inflation cannot be the cause of the overdispersion.

Many references (Chin & Quddus, 2003; Greibe, 2003; Kulmala, 1995; Lord, 2006; Lord et al., 2005; Poch & Mannering, 1996; Sawalha & Sayed, 2006; Shankar, Albin, Milton, & Mannering, 1998; Washington et al., 2003) state that, in this case, a good way to deal with overdispersed data is to use a negative binomial model (also known as Poisson-gamma model). The negative binomial model is a generalization of the Poisson regression model that assumes that unobserved crash variation across sites is gamma distributed, while crashes within sites are Poisson distributed (Mitra & Washington, 2007; Washington et al., 2003). The negative binomial model relaxes the condition of mean equal to variance of the standard Poisson model by introducing a stochastic component, the dispersion parameter, that explicitly models the dispersion of the data (Chin & Quddus, 2003). In the models presented in this report, the negative binomial dispersion parameter is estimated using maximum likelihood (SAS Institute Inc., 2008), which is the technique that is most often used to estimate this parameter (El-Basyouny & Sayed, 2006). A study by Lord (2006) indicates that maximum likelihood often provides a slightly more correct estimation of the dispersion parameter and a smaller standard deviation than other commonly used estimators, such as the method of moments or weighted regression. If the value of the dispersion parameter is between 0 and 1, the data are overdispersed. If the value is between 0 and -1, the data are underdispersed. In case the value of the dispersion parameter equals 0, the negative binomial dispersion is exactly equal to the traditional Poisson loglinear model.

Dispersion parameter estimation problems may occur in case the data are characterized by low sample-mean values and/or small sample sizes (Lord & Mannering, 2010). Since this study makes use of crash data from dangerous intersections rather than random

intersections, this problem is of less influence. The average crash count of each location for all injury crashes is 11,85 crashes per intersection, which is not considered as a low sample-mean. The model that only takes into account the serious injury and fatal crashes has a sample-mean of 2,01 crashes per intersection, which can be considered as a relatively low mean. Lord (2006) recommends a sample size of about 500 locations in case of a sample mean of 2 crashes per location to avoid estimation problems with the dispersion parameter. This model is built using 424 locations, so this guideline is not fully met. Therefore, the risk that the dispersion parameter estimate in this model is unreliable cannot be completely eliminated. However, Lord (2006) indicates that usually the problem is that the dispersion parameter erroneously indicates pure Poisson characteristics, i.e. a dispersion parameter that does not significantly differ from zero. The fact that the dispersion parameter still indicates overdispersion, combined with the fact that the dataset is only slightly too small to completely avoid the issue, indicates that the chance of an unreliable dispersion parameter estimation is relatively small.

All exposure variables are transformed to their natural logarithm for the analyses. Models have also been fitted without transforming the exposure measures, but the transformed measures provide a better fit and are therefore preferred. Furthermore, transforming exposure measures to the natural logarithm is common practice in crash prediction modeling (Reurings et al., 2006). Therefore, the functional form of the estimated models is as follows:

$$E(\lambda) = e^{\alpha} Q_{Maj}^{\beta_1} Q_{Min}^{\beta_2} e^{\sum_{i=1}^{n} \gamma_i x_i} \quad (1)$$

Where E(λ) = expected annual number of crashes (dependent variable)

Q_{maj} = traffic volume of major road

Q_{min} = traffic volume of minor road

x_i = other explanatory variables (independent variables)

α, β_1, β_2, γ = model parameters

e = natural logarithm = 2,718

3.2 Missing data

As can be seen from section 2, this study makes use of an elaborate dataset of a relatively high number of records with a large number of variables. Inevitably, some data are missing, as can be seen in Table 1. Three possible ways of dealing with this problem are distinguished. The first possibility is to estimate the missing values by using statistical imputation techniques (Little & Rubin, 2002). A second possibility is to explicitly incorporate "missing" as a separate category for categorical variables. The third possibility is to omit all records with one or more missing values from the dataset, and only estimate the models based on the records that have a known value for all possible variables (Allison, 2001; Little & Rubin, 2002). This is a so-called complete case analysis.

The authors prefer not to opt for the first possibility, i.e. estimating the missing values by using statistical imputation techniques. It is preferred to solely use the truly observed values to avoid any introduction of biases because of incorrect estimations. The risk of biased estimates in case basic assumptions for the estimation techniques are violated is shown by Allison (2001).

Modeling "missing" as a separate category is a method to deal with missing data that is not often mentioned in literature, but could be an option for this study because most variables are categorical. This method has the risk that the "missing" category has a high explanatory power in the model, which can make the interpretation of the models cumbersome. In cases where a particular category is systematically expected to be

missing more than other categories, a high explanatory power for the category "missing" is less of a problem. In these cases, the attribute is in some way significant to the model, and the results can usually be interpreted (Witten & Frank, 1999). However, as is explained below, no systematic patterns for missing data are expected in the dataset. Furthermore, the method increases the number of degrees of freedom of each variable.

The third possibility, omitting records with missing data, has the advantages of being applicable to any kind of statistical analysis and requires no special computational methods (Allison, 2001). Therefore, it is the simplest solution to missing data, and it is often the default setting in statistical software packages (Little & Rubin, 2002; SAS Institute Inc., 2008). However, the method has some disadvantages as well. It can lead to an inefficient use of the available data because of excluding a large number of records, and the method could lead to biased estimates in case data is not missing at random; i.e. in case there are some values of the variable that are systematically missing more often than others (Allison, 2001; Creemers, 2011; Little & Rubin, 2002). Nevertheless, Allison (2001) indicates that, in regression analysis, complete case analysis is more robust to violations of the missing at random assumption than more sophisticated methods to deal with missing data.

Because of the previous arguments, the third possibility, i.e. performing complete case analyses, is preferred. Although a full assessment of the possibility of a violation of the missing at random assumption is beyond the scope of the research, and the risk of non-random missing data can therefore not fully be excluded, the authors see no plausible reasons to assume a violation of this assumption for most variables.

The problem of inefficient data use because of a large number of excluded records is indeed present in the dataset. Only 279 of the total of 601 dangerous intersections are complete cases, having no missing values for any of the variables. The problem is partly taken care of by using a specific procedure for building the models. This procedure is explained in the next section.

3.3 Modeling strategy

First, the data are explored by building a model, stepwise adding variables to the model. This is called the "stepwise forward" procedure. The models are fitted making use of the log link function in the GENMOD-procedure of the statistical program SAS (SAS Institute Inc., 2008). In case a variable is significant at the 95% confidence interval ($P<0,05$), the variable is kept in the model. The type III-test is used to evaluate the significance of the variables because the results of the test indicate the significance of the variable as a whole. This is not necessarily the same as the significance of the individual estimates of the different categories for categorical variables; even if the estimates of some categories do not significantly differ from the reference category, the total variable can still be significant to the model in case some of the other categories differ significantly from each other. The significance of the internal differences between the individual categories of a categorical variable are calculated using the ESTIMATE statement, which provides a p-value for the significance of the difference between each two individual categories of a variable (SAS Institute Inc., 2008).

If the variable is not significant, the variable is omitted from the model again. In the stepwise forward procedure, variables that are identified in literature as having a significant impact in previous crash prediction model studies are added to the model first. These are primarily the volume variables and geometrical variables. Variables describing the legal status or the land-use of the intersection are not often identified as strongly influential variables by literature. Furthermore, the characteristics of the major road generally have a stronger influence than the characteristics of the minor road. The variables are therefore added to the model in the order they are presented in Table 1. When the stepwise forward procedure is completed, any variables that have become

insignificant (P>0,05) throughout the process are omitted from the model again. Then, all variables are reinserted one by one to this preliminary model. This reduces the impact of the order of adding the variables.

Next, the data are explored using a stepwise backward procedure instead of a stepwise forward procedure to see if this results in a different end model. The procedure therefore starts from an "all-in model", containing all variables, and deletes variables stepwise, starting with the least significant. When the procedure is completed, the variables not included in the preliminary end model are reinserted stepwise to check for any variables that are significant when added to this model.

It is important to mention that the number of observations is not held constant in the previous analyses. This has the disadvantage that each model that is generated in every step has a different number of observations, making it impossible to objectively compare the models with each other because the measures to evaluate the quality of the model (e.g. AIC-value) are sensitive to the number of records included. However, the advantage of this approach is that it makes optimal use of the available data to explore the importance of the variables. During both procedures, the significance of each variable is carefully monitored. Variables that reach $p \leq 0{,}10$ at some point of any of both procedures, are considered as potentially important.

Next, the actual complete case analysis phase starts. A complete case dataset is extracted from the full database, only including records that have no missing data for the variables that appear potentially important from the previous steps. From this dataset, the previous analyses (stepwise forward and backward procedure) are repeated, only including the variables identified as potentially important. Since these variables have no missing values in this dataset, the number of records in each step remains constant. Therefore, the models in each step can be objectively compared to see whether adding/deleting the variable improves the model or not. In this study, the Akaike Information Criterion (AIC) is used to compare the models. The measure indicates the relative goodness-of-fit of the model, but imposes a penalty on models with larger numbers of parameters (Akaike, 1987; Washington et al., 2003). In other words, the criterion provides an estimate of the tradeoff between the accuracy and the complexity of the model. The model with the lowest AIC model is considered to be the best model. The only restriction that is incorporated is that the traffic volume of the major and minor road (LOGVOLMAJTOTAL and LOGVOLMINTOTAL) is always included in the model, irrespective of its impact on the AIC value. This is done because it is widely acknowledged in crash prediction model and other literature that exposure is an important determining factor for the number of crashes at most locations and should therefore always be included (Carroll, 1973; Janssen, 2004; Miranda-Moreno, Morency, & El-Geneidy, 2011; Qin & Ivan, 2001; Qin, Ivan, & Ravishanker, 2004; Reurings et al., 2006; Zhang, 2008). In many studies, exposure accounts for most of the systematic variance in crash counts (e.g. Greibe, 2003; Mitra & Washington, 2007; Reurings et al., 2006).

From the complete case analysis, it becomes more clear which of the potentially important variables actually have the highest importance. Now, the complete case analysis is repeated, omitting the variables that did not appear as important variables from the first complete case analysis. The number of records increases in this second step, because less variables are required to be known for the analysis. Based on the AIC-value, the best model is again determined. If necessary, this procedure is repeated until the number of records is maximized. Using this procedure, the final models are based on complete case analysis, allowing the use of an objective measure (AIC-value) to evaluate the model, while making optimal use of the available data by including as many records as possible.

One final potential issue with these final models can be multicollinearity. Multicollinearity implies that two or more of the independent variables in the model are correlated with each other, which is a common issue in crash prediction modeling (Kulmala, 1995). Multicollinearity is a serious issue because, when two or more variables are strongly correlated, one can theoretically never be certain which variables should be included

(Verbeek, 2008). The variance inflation factor (VIF) is a measure that is often used to detect multicollinearity in a model (Washington et al., 2003). The VIF can be calculated by performing the "type II tolerance" test of the SAS GLM-procedure (SAS Institute Inc., 2008). The type II tolerance values are simply the inverse of the variance inflation factor (VIF). There is no formal boundary value for determining the presence of multicollinearity in a dataset. Generally, the lower the VIF the better, because lower VIF values indicate a lower multicollinearity between the variables. A (conservative) rule of thumb is that the VIF should be no higher than 4 (O'brien, 2007). In case an end model shows multicollinearity issues, the variable with problematic VIF-values is removed from the end model.

4. RESULTS

The results are provided in the following tables. In Table 2, a model is fitted for all the intersections together. Table 3 presents models for three-leg intersections and four-leg intersections separately. No model is fitted for intersections with more than four legs because the number of records (N=29) is too low to be able to draw valid conclusions. Table 4 presents separate models for intersections with traffic lights and for intersections with a fixed priority regulation. Models for roundabouts are not fitted because of the low number of records (N=7) in the database. Readers who are interested in crash prediction models for (Flemish) roundabouts are referred to Daniels et al. (2010). Table 5 presents separate models for intersections with a different major road category. Table 6presents submodels for intersections inside and outside the built-up area. The models presented are considered the best models based on the AIC-value.

The model for all intersections using all injury crashes contains 11 variables: LOGVOLMAJTOTAL, LOGVOLMINTOTAL, 3LEGS, MEDIMAJ, MEDIMIN, SPEEDMAJ, SPEEDMIN, RDCATMAJ, RDCATMIN, ECONOMIC and PUBLIC. LOGVOLMAJTOTAL and LOGVOLMINTOTAL indicate that the number of injury crashes increases with higher vehicle volumes. The negative value for the variable 3LEGS indicates that the number of injury crashes is lower at three-leg intersections compared to intersections with more than three legs. MEDIMAJ and MEDIMIN indicate that a higher number of injury crashes occurs at locations with a median on any of the intersection legs. SPEEDMAJ indicates that intersections with a major speed limit of 50 km/h and 70 km/h have the highest crash counts. SPEEDMIN indicates that intersections with a minor road speed limit of 50 km/h have a higher number of injury crashes than other categories, and intersections with a minor road speed limit have a lower number of injury crashes. For both speed limit variables it is difficult to draw conclusions concerning the 30 km/h category due to the low number of records for this category. Both the variables RDCATMAJ and RDCATMIN indicate that intersections with primary roads have significantly more injury crashes than the other road categories. For the major road category, this finding is significant at the 95% confidence level (CL), for the minor road category at the 90%-CL. Finally, the model shows that intersections with adjacent land use from either the ECONOMIC or the PUBLIC type have significantly more crashes than intersections with other types of surrounding land use.

The model for all intersections using only severe injury crashes incorporates 8 variables: LOGVOLMAJTOTAL, LOGVOLMINTOTAL, SIGNALS, 3LEGY, MEDIMAJ, SPEEDMAJ, PEDCROSS and CYCCROSS. The influence of LOGVOLMAJTOTAL, LOGVOLMINTOTAL and MEDIMAJ is similar to the model with all injury crashes. Furthermore, the variable SIGNALS indicates that the number of severe injury crashes is higher at signalized intersections compared to other types of intersections. 3LEGY indicates a lower number of severe injury crashes in case the intersection is a Y-shaped three-leg intersection. SPEEDMAJ is also a significant variable in the model. More precisely, this variable indicates that the number of severe injury crashes at intersections with a major road speed limit of 120 km/h is significantly lower than for intersections with a 70 km/h (90% CL) and a 50 km/h speed limit (95% CL). Furthermore, the variable PEDCROSS indicates that intersections with a pedestrian crossings have a lower number of serious injury crashes than intersections without a pedestrian crossing. CYCCROSS, on the other hand gives an indication that intersections with bicycle crossings have a higher number of serious injury crashes than intersections without bicycle crossings, but the effect is not significant at the 90%-CL and should therefore be interpreted with caution.

Table 2 - Parameter estimates for models for all intersections combined

Variables[1]	All intersections, all injury crashes (N=404)	All intersections, only serious/fatal injury crashes (N=424)
Intercept	-1.103 (SE=0.495) *(p=0.038)***	-2.546 (SE=0.706) *(p<0.001)****
LOGVOLMAJ-TOTAL	0.219 (SE=0.056) *(p<0.001)****	0.234 (SE=0.079) *(p=0.003)****
LOGVOLMIN-TOTAL	0.173 (SE=0.029) *(p<0.001)****	0.150 (SE=0.045) *(p<0.001)****
SIGNALS		0.242 (SE=0.121) *(p=0.046)***
3LEGS	-0.235 (SE=0.078) *(p=0.003)****	
3LEGY		-0.833 (SE=0.414) *(p=0.028)***
MEDIMAJ	0.214 (SE=0.063) *(p<0.001)****	0.216 (SE=0.092) *(p=0.020)***
MEDIMIN	0.164 (SE=0.076) *(p=0.031)***	
SPEEDMAJ	30: -1.032 (SE=0.665) (p=0.121)° 50: 0.447 (SE=0.181) (p=0.014)** 70: 0.316 (SE=0.177) (p=0.074)* 90: 0 120: 0.021 (SE=0.259) (p=0.936)° *(p=0.059)**	30: -0.095 (SE=0.862) (p=0.913)° 50: 0.027 (SE=0.241) (p=0.913)° 70: 0.272 (SE=0.235) (p=0.246)° 90: 0 120: -0.514 (SE=0.386) (p=0.182)° *(p=0.027)***
SPEEDMIN	30: 0.066 (SE=0.323) (p=0.838)° 50: 0.283 (SE=0.102) (p=0.006)*** 70: 0.048 (SE=0.100) (p=0.634)° 90: 0 120: -0.403 (SE=0.181) (p=0.026)** *(p=0.001)****	
PEDCROSS		-0.201 (SE=0.104) *(p=0.054)**
CYCCROSS		0.152 (SE= 0.095) *(p=0.108)*°
RDCATMAJ	LOC: -0.064 (SE=0.072) (p=0.376)° SEC: 0 PRIM: 0.196 (SE=0.060) (p=0.001)*** MAIN: -0.194 (SE=0.126) (p=0.123)° *(p=0.010)***	
RDCATMIN	LOC: -0.079 (SE=0.068) (p=0.244)° SEC: 0 PRIM: 0.172 (SE=0.091) (p=0.059)* MAIN: -0.199 (SE=0.143) (p=0.163)° *(p=0.025)***	
ECONOMIC	0.135 (SE=0.079) (p=0.086)*	
PUBLIC	0.324 (SE=0.126) (p=0.010)***	
Dispersion[2]	0.181 (SE=0.020)***	0.144 (SE=0.043)***
AIC (smaller is better)	2538.80	1530.21

[1] values present the parameter estimates of the negative binomial model. For categorical variables with more than 2 categories, the category is indicated. P-values of individual categories and Standard Errors between (). *P-values of the variables in total are in italics between ()*.
[2] values higher than 0 indicate overdispersion.
*** P≤0,01 (significant at 99% CI)
** P≤0,05 (significant at 95% CI)
* P≤0,10 (significant at 90% CI)
° P>0,10 (not significant at 90% CI)

Submodels for four-leg and for three-leg intersections are presented in Table 3. The submodel for four-leg intersections is quite comparable to the model for all intersections. The impact of LOGVOLMAJTOTAL, LOGVOLMINTOTAL, MEDIMAJ, MEDIMIN, SPEEDMAJ, SPEEDMIN, RDCATMAJ, ECONOMIC and PUBLIC is strongly comparable. Furthermore, the variable PEDFAC indicates a (non-significant) lower number of injury crashes at four-leg intersections with sidewalks. Furthermore, an extra land use variable (i.e. RESIDENTIAL) corresponds with a higher crash count in this model.

In the submodel for three-leg intersections, LOGVOLMAJTOTAL is included in its centered form. This has been done to solve a multicollinearity between LOGVOLMAJTOTAL and the intercept. This transformation does not change the estimate or significance of the variable, nor those of other variables in the model. Only the estimate of the intercept changes because of the transformation. LOGVOLMINTOTAL also has a positive influence on the number of injury crashes, but the significance is lower than in the model for all intersections (i.e., significant at the 90% CL instead of the 99% CL). The variable LANEMIN indicates that intersections with a one-lane minor road have a significantly higher number of crashes than all other categories at the 90% CL. MEDIMIN indicates a higher number of injury crashes when a median is present on the minor road. SPEEDMIN indicates the highest crash count at intersections with a minor road with speed limit of 50 km/h. The difference is significant at the 90%-CL for all categories except 30 km/h. The variable CYCCROSS indicates that intersections with bicycle crossings have a significantly lower number of injury crashes. On the other hand, CYCLNETW indicates that three-leg intersections that belong to the cycle route network have a higher number of injury crashes than intersections that do not belong to the cycle route network. Intersections with a local minor road (RDCATMIN) have a significantly lower number of crashes than intersections with a secondary minor road. The other differences between categories are not significant. Finally, PUBLIC indicates that intersections near public services have a higher number of injury crashes.

Table 3 - Parameter estimates, split by number of legs

Variables[1]	4LEGS intersections (N=312)	3LEGS intersections (N=65)
Intercept	-1.155 (SE=0.529) *(p=0.029)***	1.076 (SE=0.447) *(p=0.016)***
LOGVOL-MAJTOTAL	0.188 (SE=0.062) *(p=0.003)****	
LOGVOL-MAJTOTAL (centered)		0.513 (SE=0.141) *(p<0.001)****
LOGVOL-MINTOTAL	0.217 (SE=0.034) *(p<0.001)****	0.092 (SE=0.052) *(p=0.079)**
LANEMIN		1: 0.655 (SE=0.239) (p=0.006)*** 2: 0.005 (SE=0.161) (p=0.976)° 3: -0.298 (SE=0.310) (p=0.336)° 4: 0 *(p=0.048)***
MEDIMAJ	0.271 (SE=0.073) *(p<0.001)****	
MEDIMIN	0.251 (SE=0.091) *(p=0.006)****	0.627 (SE=0.201) *(p=0.003)****

SPEEDMAJ	30: -0.917 (SE=0.671) (p=0.172)° 50: 0.512 (SE=0.187) (p=0.006)*** 70: 0.291 (SE=0.181) (p=0.107)° 90: 0 120: -0.079 (SE=0.279) (p=0.777)° *(p=0.014)***	
SPEEDMIN	 50: 0.230 (SE=0.084) (p=0.006)*** 70: 0.059 (SE=0.080) (p=0.460)° 90: 0 120: -0.330 (SE=0.181) (p=0.067)* *(p=0.057)**	30: -0.006 (SE=0.411) (p=0.989)° 50: 0.491 (SE=0.188) (p=0.009)*** 70: -0.021 (SE=0.179) (p=0.907)° 90: 0 120: -0.562 (SE=0.486) (p=0.247)° *(p=0.012)***
PEDFAC	-0.130 (SE=0.081) (p=0.109)°	
CYCCROSS		-0.487 (SE=0.200) *(p=0.017)***
CYCLNETW		0.734 (SE=0.250) *(p=0.004)****
RDCATMAJ	LOC: -0.144 (SE=0.083) (p=0.084)* SEC: 0 PRIM: 0.193 (SE=0.069) (p=0.006)*** MAIN: -0.100 (SE=0.146) (p=0.495)° *(p= 0.015)***	
RDCATMIN		LOC: -0.320 (SE=0.159) (p=0.044)** SEC: 0 PRIM: 0.041 (SE=0.197) (p=0.835)° MAIN: -0.101 (SE=0.346) (p=0.770)° *(p=0.003)****
RESIDENTIAL	0.142 (SE=0.083) *(p=0.087)**	
ECONOMIC	0.166 (SE=0.091) *(p=0.068)**	
PUBLIC	0.438 (SE=0.140) *(p=0.001)****	0.588 (SE=0.244) *(p=0.018)***
Dispersion[2]	0.191 (SE=0.023)***	0.071 (SE=0.032)***
AIC (smaller is better)	2004.52	377.83

[1] values present the parameter estimates of the negative binomial model. For categorical variables with more than 2 categories, the category is indicated. P-values of individual categories and Standard Errors between (). *P-values of the variables in total are in italics between ()*.
[2] values higher than 0 indicate overdispersion.
*** P≤0,01 (significant at 99% CI)
** P≤0,05 (significant at 95% CI)
* P≤0,10 (significant at 90% CI)
° P>0,10 (not significant at 90% CI)

Table 4Table 4 presents separate submodels for signalized intersections and for priority intersections. The model for signalized intersections includes the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL, 3LEGY, MEDIMAJ, PEDFAC and RESIDENTIAL. The influence of the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL, MEDIMAJ, SPEEDMAJ and RESIDENTIAL are in line with the model for all intersections. 3LEGY indicates a lower injury crash count at Y-shaped three-leg intersections. The estimates of the categories of SPEEDMAJ indicate that intersections where the major road speed limit is 50 km/h have a significantly higher (99% CL) crash count than the other categories. Intersections where the major road speed limit is 120 km/h have a significantly lower (90% CL) crash count than the other categories. The variable PEDFAC indicates that signalized intersections with sidewalks have a significantly lower crash count than signalized intersections without sidewalks.

The model for priority intersections consists of the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL, MEDIMAJ, SPEEDMAJ, RDCATMAJ and PUBLIC. The influence of LOGVOLMAJTOTAL, LOGVOLMINTOTAL and MEDIMAJ is in line with the model for all intersections. SPEEDMAJ indicates that intersections with major road category 50 km/h have a significantly higher (90% CL) injury crash count than the other categories. RDCATMAJ indicates that priority intersections with a main major road have a significantly lower (99% CL) crash count than the other categories. Furthermore, intersections with major road category primary have a significantly higher crash count than the categories MAIN and SEC (99% CL); the difference with category LOC is not statistically significant. PUBLIC indicates a significantly higher crash count at intersections near public land use.

Table 4 - Parameter estimates, split by right-of-way regulation

Variables[1]	SIGNALS intersections (N=197	PRIORITY intersections (N=220)
Intercept	-2.509 (SE=0.692) *(p<0.001)****	-0.654 (0.676) *(p=0.334)*°
LOGVOL-MAJTOTAL	0.310 (SE=0.075) *(p<0.001)****	0.164 (SE=0.078) *(p=0.038)***
LOGVOL-MINTOTAL	0.339 (SE=0.055) *(p<0.001)****	0.159 (SE=0.037) *(p<0.001)****
3LEGY	-0.571 (SE=0.276) *(p=0.045)***	
MEDIMAJ	0.332 (SE=0.081) *(p<0.001)****	0.259 (SE=0.098) *(p=0.009)****
SPEEDMAJ	50: 0.336 (SE=0.089) (p<0.001)*** 70: 0.040 (SE=0.079) (p=0.610)° 90: 0 120: -0.364 (SE=0.175) (p=0.037)** *(p=0.002)****	30: -0.916 (SE=0.643) (p=0.154)° 50: 0.619 (SE=0.224) (p=0.006)*** 70: 0.236 (SE=0.221) (p=0.285)° 90: 0 *(p<0.001)****
PEDFAC	-0,197 (SE=0.098) *(p=0.046)***	
RDCATMAJ		LOC: 0.187 (SE=0.119) (p=0.117)° SEC: 0 PRIM: 0.372 (SE=0.120) (p=0.002)*** MAIN: -0.813 (SE=0.270) (p=0.003)*** *(p=0.001)****
RESIDENTIAL	0.160 (SE=0.089) *(p=0.076)**	
PUBLIC		0.840 (SE=0.245) *(p=0.001)****
Dispersion[2]	0.174 (SE=0.025)***	0.229 (SE=0.035)***
AIC (smaller is better)	1319.51	1297.54

[1] values present the parameter estimates of the negative binomial model. For categorical variables with more than 2 categories, the category is indicated. P-values of individual categories and Standard Errors between (). *P-values of the variables in total are in italics between ()*.
[2] values higher than 0 indicate overdispersion.
*** P≤0,01 (significant at 99% CI)
** P≤0,05 (significant at 95% CI)
* P≤0,10 (significant at 90% CI)
° P>0,10 (not significant at 90% CI)

Table 5 presents separate models for the different major road categories. The MAIN category has a low number of records, and is in fact relatively similar to the category of primary roads. Therefore, it is decided to fit one model for both categories combined. The model for MAIN/PRIM intersections contains the variables LOGVOLMAJTOTAL,

LOGVOLMINTOTAL, 3LEGS, DIAGON, 3LEGY, SPEEDMAJ, PEDFAC and RESIDENTIAL. The impact of LOGVOLMAJTOTAL, LOGVOLMINTOTAL, 3LEGS and RESIDENTIAL is in line with the model for all intersections. DIAGON indicates that non-perpendicular four-leg intersections have a lower crash count than other intersections. 3LEGY indicates a higher crash count for Y-shaped three-leg intersections. SPEEDMAJ indicates a strongly significant (99% CL) higher crash count at intersections where the major road has a 50 km/h speed limit compared to other speed limits. PEDFAC indicates a significantly lower number of injury crashes at intersections where pedestrian facilities are present.

In the model for SEC intersections, the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL and PUBLIC show a comparable influence on the number of injury crashes as in the model for all intersections. SPEEDMAJ indicates a significantly higher number of injury crashes at locations with a speed limit of 50 km/h (90% CL). The variable BUILTUP indicates that the intersections where buildings are present have a significantly higher crash count than intersections without buildings.

The model for LOC intersections only contains four variables, LOGVOLMAJTOTAL, LOGVOLMINTOTAL, LANEMIN and PEDCROSS. The model shows a lower importance of the major road volume (LOGVOLMAJTOTAL) compared to the other models. The variable is not statistically significant in this model. The effect of LOGVOLMINTOTAL is however in line with the general model. LANEMIN indicates that intersections with a one lane major road have a significantly higher (90% CL) crash count than the other categories, while two-lane intersections have a significantly lower crash count. Only for the three-lane category can no significant conclusions be drawn since the category only contains one record. The variable PEDCROSS indicates a strongly significantly higher number of injury crashes when pedestrian crossings are present compared to intersections without pedestrian crossings.

Table 6 shows submodels for dangerous intersections inside and outside built-up area. The model for intersections inside built-up area contains the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL, RDCATMAJ and PUBLIC, which are in line with the model for all intersections. For RDCATMIN, the category PRIM is in line with the general model. The category LOC indicates that intersections with a local minor road have a significantly lower crash count than the other categories. Furthermore, the model also contains the variable RESIDENTIAL, which indicates a significantly higher number of crashes in residential areas.

The model for locations outside built-up area contains the variables LOGVOLMAJTOTAL, LOGVOLMINTOTAL, 3LEGS, MEDIMAJ, SPEEDMIN, RDCATMIN and ECONOMIC. The estimates for LOGVOLMAJTOTAL, LOGVOLMINTOTAL, 3LEGS, MEDIMAJ and SPEEDMIN are in line with the model for all intersections. RDCATMIN indicates a significantly lower number of injury crashes at intersections with a main minor road compared to intersections with a secondary minor road; other differences between categories are not significant in this submodel. Like in the model for all intersections, ECONOMIC indicates a higher number of crashes in the presence of economic land use, but in this submodel the significance is stronger (99% CL in this submodel compared to 90% CL in the general model).

Table 5 - Parameter estimates, split by major road category

Variables[1]	MAIN or PRIM only (N=138)	SEC only (N=208)	LOC only (N=86)
Intercept	-3.672 (SE=0.801) *(p<0.001)****	-1.890 (SE=0.760) *(p=0.013)***	0.282 (SE=0.671) *(p=0.674)*°
LOGVOL-MAJTOTAL	0.555 (SE=0.093) *(p<0.001)****	0.263 (SE=0.089) *(p=0.004)****	0.125 (SE=0.083) *(p=0.136)*°
LOGVOL-MINTOTAL	0.289 (SE=0.041) *(p<0.001)****	0.246 (SE=0.034) *(p<0.001)****	0.145 (SE=0.056) *(p=0.011)***
LANEMIN			1: 0.564 (SE=0.289) (p=0.051)* 2: -0.278 (SE=0.161) (p=0.083)* 3: -0.045 (SE=0.399) (p=0.910)° 4: 0 *(p=0.079)**
3LEGS	-0.470 (SE=0.155) *(p=0.003)****		
DIAGON	-0.267 (SE=0.123) *(p=0.032)***		
3LEGY	0.734 (SE=0.345) *(p=0.034)***		
SPEEDMAJ	50: 0.561 (SE=0.141) (p<0.001)*** 70: -0.069 (SE=0.093) (p=0.457)° 90: 0 120: -0.247 (SE=0.178) (p=0.165)° *(p<0.001)****	30: -0.927 (SE=0.646) (p=0.152)° 50: 0.541 (SE=0.224) (p=0.016)** 70: 0.216 (SE=0.222) (p=0.330)° 90: 0 *(p<0.001)****	
PEDFAC	-0.354 (SE=0.093) *(p<0.001)****		
PEDCROSS			0.485 (SE=0.126) *(p<0.001)****
RESIDENTIAL	0.263 (SE=0.090) *(p=0.004)****		
PUBLIC		0.466 (SE=0.201) *(p=0.017)***	
BUILTUP		0.326 (SE=0.158) *(p=0.042)***	
Dispersion[2]	0.147 (SE=0.027)***	0.237 (SE=0.033)***	0.129 (SE=0.039)***
AIC (smaller is better)	898.08	1319.99	500.29

[1] values present the parameter estimates of the negative binomial model. For categorical variables with more than 2 categories, the category is indicated. P-values of individual categories and Standard Errors between (). *P-values of the variables in total are in italics between ()*.
[2] values higher than 0 indicate overdispersion.
*** P≤0,01 (significant at 99% CI)
** P≤0,05 (significant at 95% CI)
* P≤0,10 (significant at 90% CI)
° P>0,10 (not significant at 90% CI)

Table 6 - Parameter estimates, inside and outside built-up area

Variables[1]	INSIDE only (N=115)	OUTSIDE only (N=277)
Intercept	-0.530 (SE=0.834) *(p=0.525)*°	-2.047 (SE=0.539) *(p<0.001)****
LOGVOL-MAJTOTAL	0.278 (SE=0.102) *(p=0.008)****	0.329 (SE=0.062) *(p<0.001)****
LOGVOL-MINTOTAL	0.109 (SE=0.051) *(p=0.035)***	0.256 (SE=0.034) *(p<0.001)****
3LEGS		-0.204 (SE=0.095) *(p=0.033)***
MEDIMAJ		0.148 (SE=0.074) *(p=0.046)***
SPEEDMIN		30: 0.144 (SE=0.448) (p=0.748)° 50: 0.228 (SE=0.130) (p=0.081)* 70: 0.014 (SE=0.128) (p=0.916)° 90: 0 120: -0.407 (SE=0.206) (p=0.048)** *(p=0.019)***
RDCATMAJ	LOC: -0.144 (SE=0.145) (p=0.321)° SEC: 0 PRIM: 0.327 (SE=0.158) (p=0.039)** MAIN: -0.277 (SE=0.324) (p=0.391)° *(p=0.077)**	
RDCATMIN	LOC: -0.405 (SE=0.122) (p<0.001)*** SEC: 0 PRIM: 0.650 (SE=0.219)(p=0.003)*** *(p=0.003)****	LOC:-0.018 (SE=0.073) (p=0.809)° SEC: 0 PRIM: 0.052 (SE=0.096) (p=0.588)° MAIN: -0.247 (SE=0.144) (p=0.087)* *(p=0.037)***
RESIDENTIAL	0.467 (SE=0.227) *(p=0.045)***	
ECONOMIC		0.218 (SE=0.096) *(p=0.009)****
PUBLIC	0.568 (SE=0.210) *(p=0.006)****	
Dispersion[2]	0.215 (SE=0.039)***	0.160 (SE=0.023)***
AIC (smaller is better)	772.90	1700.04

[1] values present the parameter estimates of the negative binomial model. For categorical variables with more than 2 categories, the category is indicated. P-values of individual categories and Standard Errors between (). *P-values of the variables in total are in italics between ()*.
[2] values higher than 0 indicate overdispersion.
*** P≤0,01 (significant at 99% CI)
** P≤0,05 (significant at 95% CI)
* P≤0,10 (significant at 90% CI)
° P>0,10 (not significant at 90% CI)

5. DISCUSSION

5.1 Discussion of the analyzed variables

5.1.1 Exposure

As explained before, the exposure variables LOGVOLMAJTOTAL and LOGVOLMINTOTAL are forced into all models for theoretical reasons. However, even without this intervention, the variables would have appeared in all end models anyway, as all of them are important enough to improve the AIC-value in all models.

LOGVOLMAJTOTAL is significant in nearly all presented models. Only in the submodel for intersections with a local major road, the variable is not statistically significant (p=0.136). In the other models, the variable is significant at the 99% CL, except the submodel for priority intersections where the variable is only significant at the 95% CL. In most models, the estimate of LOGVOLMAJTOTAL is around 0.2. The variable has the highest estimate in the submodels for three-leg intersections and for primary/main major road intersections, where the estimate is around 0.5. All estimates are therefore significantly lower than 1, which indicates a rather strong inelasticity between the number of injury crashes and the vehicle volume. Most other studies confirm indeed a positive but inelastic relationship between crash count and traffic volume. The estimate of the major road volume is usually in the range 0.5-1.0 (e.g. Oh, Lyon, Washington, Persaud, & Bared, 2003; Reurings et al., 2006), which is somewhat higher than the estimates found in this study.

The variable LOGVOLMINTOTAL is significant at the 99% CL in all models, except the submodel for three-leg intersections (significant at 90% CL) and the submodel for local major road intersections (significant at 95% CL). The estimate is usually also around 0.2. This is in line with existing literature; for instance, the report by Reurings et al. (2006) shows values in the range of 0.2-0.5 for the minor road vehicle volume.

Data are also collected for the different moving directions of the total vehicle volume. The possibility to use one of these directional flows instead of the total vehicle flow is investigated in the explorative models, as described in section 3. The use of volumes for specific movements on the intersection rather than general volumes is not common practice in literature. Occasionally, some preliminary models showed a better fit for the left-turning flow or for the straight flow. However, this is not a systematic tendency. Furthermore, using different exposure measures in the different models would make it more difficult to mutually compare the presented models. Therefore it is decided to systematically use the total vehicle volumes to fit the models.

5.1.2 Geometric variables

Some studies identify the number of driving lanes (i.e. LANEMAJ and LANEMIN) as an important variable in crash prediction models for intersections. Most of them indicate that a higher number of crashes is present at intersections with a lower number of lanes (Bauer & Harwood, 1999, 2000; Reurings et al., 2006). The variable LANEMIN is present in the submodel for three-leg intersections and the submodel for local major road intersections only. In both models, the variable indicates that intersections with a one-lane minor road have a significantly higher crash count than other categories, which is in line with literature. The submodel for local major road intersections also indicates a lower crash count at intersections with two lanes on the minor road, which is not in line with literature. The explanation for this finding is unclear. The variable LANEMAJ is not present in any of the end models. The variable originally appeared in a number of submodels, but has been removed from the end models because of multicollinearity issues.

Signalized intersections are often characterized by a higher number of crashes than other types of intersections (Abdel-Aty & Keller, 2005). In this study, the variable SIGNALS is only present in the model for all intersections using only the serious/fatal injury crashes. The variable indicates a significantly higher number of serious/fatal injury crashes at signalized intersections. This might seem surprising, since one of the goals of placing traffic lights is to reduce the number of crashes by reducing/avoiding conflicts between vehicles by separating their passage in time. However, some conflicts can still occur, for instance when one of the road users violates the red light. Therefore, interactions between road users at signalized intersections will be less prevalent in general, but often of a higher severity than at other types of intersections (Svensson & Hydén, 2006). This might partly explain why the variable only contributes to the number of severe injury crashes and not to all injury crashes.

In summary, most of the presented models do not show a higher number of crashes at signalized intersections compared to other intersections. However, the submodels for different types of right-of-way regulation show some differences between the model for signalized intersections and for priority intersections. This indicates that, even though the number of crashes is comparable between both types of locations, the characteristics contributing to this number of crashes differ from each other.

Generally, three-leg intersections have a lower number of crashes than four-leg intersections (Elvik, Høye, Vaa, & Sørensen, 2009; Janssen, 2004; Kulmala, 1995; O'Cinneide & Troutbeck, 1998). This is mainly because four-leg intersections put higher demands on road user alertness and behaviour than three-leg intersections because of the higher number of potential conflict points between streams of traffic (Elvik, Høye, et al., 2009). The results of this study confirm this. The variable 3LEGS corresponds with a highly significantly lower (99% CL) number of crashes than intersections with more legs. The variable is also included in the submodels for MAIN/PRIM intersections and intersections outside built-up area. It is remarkable that the model that only uses serious injury and fatal crashes does not include the variable 3LEGS, which indicates that especially the number of minor injury crashes is lower at three-leg intersections.

The variable 4LEGS can sometimes be a substitute for 3LEGS, which is not surprising because these variables are nearly perfectly collinear (i.e. nearly all intersections that have a "0" for 3LEGS have a "1" for 4LEGS, and vice versa) but the fit is usually slightly worse for 4LEGS than for 3LEGS. Furthermore, although it could be expected that intersections with more than four legs have a higher crash rate because of their very complex nature, the variable 56LEGS does not appear in any of the models. This is probably because of the relatively low number of 56LEGS intersections in the dataset (N=29). Also the number of roundabouts in the sample is too small to make a contribution to any of the models (N=7).

The variables 3LEGY and DIAGON indicate intersections that have non-perpendicular angles for the approaching roads. Intuitively, it is anticipated that these intersections could be more unsafe than perpendicular intersections because the view on some of the intersection legs might not be optimal. However, there is disagreement in literature on the optimal angle of approach at intersections (O'Cinneide & Troutbeck, 1998). The Transportation Research Board (1987) recommends perpendicular intersections. Kulmala (1995) found a higher crash count for certain types of non-perpendicular intersections angles, but also a lower crash count for others. A report from the National Cooperative Highway Research Program also finds that non-perpendicular four-leg intersections experience lower crash rates than perpendicular intersections (NCHRP Report 197, 1987). So existing literature is not conclusively on this issue. Since most models in which 3LEGY or DIAGON are included indicate a lower crash rate at locations with non-perpendicular angles, this study associates with the literature that indicates safety benefits for non-perpendicular intersections. DIAGON is present in the submodel for MAIN/PRIM intersections and indicates a lower crash count at non-perpendicular four-leg intersections. The variable 3LEGY is present in a number of models. The variable indicates a significantly lower number of injury crashes in the model for all intersections

with only serious/fatal injury crashes and the submodels for signalized intersections. Possibly, this is related to the finding that three-leg intersections in general often have a lower number of crashes (Elvik, Høye, et al., 2009; Janssen, 2004; Kulmala, 1995). The group of Y-shaped three-leg intersections is a subgroup of all three-leg intersections. Therefore, this variable indicates a lower number of crashes at Y-shaped three-leg intersections, compared to all other types of intersections (including four-leg intersections etc.), and therefore it does not necessarily imply that the Y-shaped intersection has a better safety performance than the perpendicular layout. The fact that the variable is not present in the submodel for three-leg intersections seems to confirm indeed that the Y-shape itself is not responsible for the safety effect. On the other hand, in the submodel for MAIN/PRIM intersections, the variable indicates a significantly higher number of crashes for Y-shaped three-leg intersections. It should also be noted that the total number of Y-shaped three-leg is relatively low (N=12), so the findings should be interpreted with caution. The influence of non-perpendicular intersection angles on the occurrence of crashes should be investigated in future research.

The presence of a median (MEDIMAJ and MEDIMIN) on one or more of the approaching legs correlates in all models in which the variables are included with a higher number of injury crashes. MEDIMAJ is included in both models for all intersections and the submodels for four-leg intersections, signalized intersections, priority intersections and intersections outside built-up area. MEDIMIN is present in the model for all intersections using all injury crashes, and the submodels for four-leg and three-leg intersections. These findings are not in line with the expectation that, ceteris paribus, the presence of a median could improve road safety, because it can help to avoid head-on collisions, which are among the most dangerous types of crashes (Mao et al., 1997). A possible explanation might be that medians are often applied to roads of a higher functional class, generally with higher traffic volumes, driving speeds, etc. which can lead to a generally higher number of crashes.

Intuitively, it seems apparent that, ceteris paribus, crash risk and injury severity increase with increasing speed (Evans, 2004; O'Cinneide & Troutbeck, 1998). However, this does not automatically imply that roads with a higher speed limit have a higher number of crashes. Usually, the design of the road is matched to the allowed speed limit, and roads with a higher speed limit will often be better equipped (e.g. wider lanes, guard rails, etc.). Furthermore, these roads tend to have a lower number of vulnerable road users and be located in sparsely built-up areas (Greibe, 2003). Also, the actual driving speeds do not always correspond with the formal speed limits. This can result in a lower number of crashes at locations with higher speed limits. In this study, SPEEDMAJ and SPEEDMIN are present in the models for all intersections (SPEEDMIN only in the model for all injury crashes) and a number of submodels. Most models indicate that the intersections with a speed limit of 50 km/h have a significantly higher number of injury crashes than the other categories (except the model for all intersections that only includes serious injury and fatal accidents). Although this finding does not indicate that the lower speed limit is the cause of the higher number of crashes, it does indicate that these types of intersections require special attention. It should also be noted that the models for all intersections also indicate a higher number of injury crashes in case the major road speed limit is 70 km/h, but this finding is not confirmed by the submodels.

The variable PEDFAC (the presence of a sidewalk) is included in the submodels for four-leg intersections, signalized intersections and MAIN/PRIM intersections. The variable indicates a lower number of crashes in case sidewalks are present at these locations. The presence of sidewalks is a variable that is not often found in previous crash prediction studies. Only one crash prediction model by Kulmala (1995) for three-leg intersections includes the variable. The model indicates a lower number of injury crashes in case pedestrian facilities are present on the minor road. General literature is not entirely conclusive about the safety effects of sidewalks. Knoblauch et al. (1988) indicate a strong reduction in the number of vehicle crashes when sidewalks are applied. The overview by Elvik, Høye et al. (2009) on the other hand only finds a slightly lower

number of crashes involving vulnerable road users, but also a higher number of vehicle crashes for intersections with sidewalks.

Crossing facilities for vulnerable road users can have an important impact on the number of crashes because they guide the interaction between vulnerable road users and motor vehicles. It is important to keep in mind that pedestrians have priority over motor vehicles at pedestrian crossings, but that bicyclists do not always have priority over motor vehicles. The model for all intersections that uses only serious injury/fatal crashes includes both variables. However, while the presence of pedestrian crossings corresponds with a significantly (90% CL) lower number of serious injury/fatal crashes, the presence of bicycle crossings seems to correspond with a higher number of serious injury/fatal crashes. Although the latter finding is not significant at the 90% CL and should therefore be interpreted cautiously, this contrast is quite interesting. A possible explanation could be that the rules at pedestrian crossings are unambiguous and clear. This may lead to a relatively low number of violations against the formal priority rules. For bicyclist crossings on the other hand, the rules differ from location to location and are even not always uniform for comparable locations. This could lead to a higher number of severe crashes because of failure-to-yield crashes. However, this general trend is not extended to the submodels. The submodel for three-leg intersections indicates a significantly lower (95% CL) number of crashes when bicycle crossings are present at three-leg intersections. The submodel for intersections with a local major road indicates a significantly higher (CL 99%) number of crashes in case pedestrian crossings are present. The presence of crossing facilities for vulnerable road users is not included in crash prediction model literature. General literature is not conclusive on the safety effects of crossings for vulnerable road users. Elvik, Høye et al. (2009) provide a summary of existing studies about pedestrian crossings that correct for chance, and they conclude that regular marked crossings for pedestrians result in a non-significant higher number of crashes. Only in case the crosswalk is raised, a significantly lower number of crashes is found.

The variables CYCLFAC, PARKING and PTFAC are rarely included in existing crash prediction model literature, and do not appear in any of the end models of this study either.

5.1.3 Legal status variables

The variable CYCLNETW indicates the most important connections for bicyclists. The variable has not been found in crash prediction literature, nor in general literature, possibly because the concept of bicycle route networks is not common practice in many countries. The variable is only present in the three-leg intersections submodel. The variable indicates a significantly higher (99% CL) number of crashes in case one or more of the intersection legs belongs to the bicycle route network. It is plausible that the variable CYCLNETW is a proxy for the number of bicyclists passing the intersection. Therefore, the higher exposure of bicyclists may be the cause of the higher crash count, rather than the actual legal status.

RDCATMAJ and RDCATMIN refer to the functional class of the major road and the minor road respectively. RDCATMAJ is present in the general model for all intersections (all injury crashes), and in the submodels for four-leg intersections, priority intersections and intersections inside built-up area. RDCATMIN is present in the general model for all injury crashes at all intersections, and in the submodels for three-leg intersections, the submodel for locations inside built-up area and outside built-up area. The models show that intersections with a primary major road usually have a higher number of crashes than the other categories. Some models also indicate a lower number of crashes at intersections with local or main roads compared to the reference category of secondary roads, but this pattern is not always statistically significant. So the general tendency seems to be that the number of crashes is higher at intersections with higher road categories, with the exception of intersections with main roads. This can be explained by the fact that intersections with main roads are actually intersections with an on and/or off ramp of the main roads. The findings are difficult to compare with international literature,

because the concept of road classification differs between countries. For instance, The Netherlands only distinguish between three types of functional road classes (Dijkstra, 2003, 2010). One study was found in literature that includes the major road classification (Bauer & Harwood, 2000). However, the study remains inconclusive; it shows both models where a higher number of crashes occurs at roads from a lower class, and models where a lower number of crashes occurs at roads from a lower class. The findings of this study are in line with the latter.

5.1.4 Land use variables

Land use variables are not often included in crash prediction studies (Reurings et al., 2006). Only one crash prediction study by Harnen et al. (2006) included a land use variable in a crash prediction model for motorcycle crashes in Malaysia; the study indicated a higher number of motorcycle crashes in case the land use category was "commercial".

The variables PUBLIC, ECONOMIC and RESIDENTIAL are present in a number of submodels. In case one or more of the variables are included in one of the models, their impact is always similar, i.e. a significantly higher number of injury crashes in case the land use around the intersection belongs to the category.

BUILTUP refers to the fact whether there are any buildings present around the intersection or not. The variable is only significant in the submodel for secondary major road intersections, and indicates a significantly higher number of crashes when buildings are present. This is an important point of attention since ribbon development is ubiquitous at Flemish secondary roads (Albrechts, 1999). This finding confirms the safety issues that result from having buildings at roads that have an important traffic function. Therefore, it is recommended not to have buildings at roads with an important traffic function, which is one of the basic ideas behind the "Sustainable Safety" principle (Dijkstra, 2003).

However, the presence and type of land-use near the location can have a strong influence on the interactions that can occur at the location, and can therefore have an important impact on the number of crashes that occur. However, this implies that the land use is indirectly linked with the number of injury crashes, because the land use that is present will influence the risk exposure of mainly vulnerable road users (Miranda-Moreno et al., 2011; Ukkusuri, Miranda-Moreno, Ramadurai, & Isa-Tavarez, 2012).

The variable INSIDE does not appear in any of the models. So the absolute number of injury crashes seems not to be significantly influenced by whether the location is inside built-up area or not. However, separate submodels are fit for locations inside and outside built-up area, as is sometimes done in crash prediction model literature (Reurings et al., 2006). Both models show some clear differences, which indicates that, even though the number of crashes between both does not differ significantly, the characteristics that lead to this number of crashes differ between both types of locations.

5.2 Study limitations

In statistics, one of the basic requirements to be able to generalize conclusions from analyses of a sample from the population to the full population (i.e. external validity), is that the sample is randomly drawn from the population (Anderson et al., 2005). It is clear that the sample of intersections in this study is not random, but exists exclusively of dangerous locations. Therefore, it is not possible to generalize the conclusions of this study to all intersections. The models are useful to identify intersection characteristics that contribute to the number of crashes at dangerous intersections. However, for a new intersection or for an intersection that will be rebuilt, it is a priori not clear whether the

intersection is a dangerous intersection or not. Therefore, the models cannot claim predictive power, but only explanatory power for the current dataset. Using the models from this report to predict future crash counts would probably result in an overestimation of the number of injury crashes, because the records in the dataset of this study have a higher-than-expected number of injury crashes. The term "crash prediction model" is in that respect quite misleading in this report, although even crash prediction models based on a random dataset are only based on historical data, so that they can only describe or explain past events, and there is no guarantee that the patterns found will remain valid towards the future (Reurings et al., 2006). Nevertheless, despite the lack of predictive power in absolute terms, these models can provide valuable insights to road designers and policy makers that will help them to handle dangerous intersections more effectively and efficiently.

Another limitation is that the variables included in the study are partly based on variables that were used in comparable studies, but also partly based on the practical limitations to collect data about them. Therefore, the risk of omitted-variable bias cannot be excluded (Lord & Mannering, 2010). It implies that data could not be collected about all potentially useful variables. Especially the absence of exposure data for vulnerable road users is a limitation. As became clear from section 5.1 some of the variables with a high importance, especially the land use variables, are expected to function as a proximate for exposure. It would therefore be preferable to include these variables instead of their proxies. Some studies already presented models where separate exposure measures for different transport modes showed a very good fit (e.g. Daniels, Brijs, Nuyts, & Wets, 2010, 2011). Furthermore, the fact that the exposure variables are based on two-hour counts brings some additional uncertainty into the model. The same hours are used for all intersections, and the data is collected in a relatively short time period, which reduces a number of biases and trend influences. However, some biases cannot be excluded. For instance, Cools (2009) shows an impact of weather conditions and holiday periods on the traffic volume. The exposure data of this study are not corrected for such influences.

Another point of discussion is that the method of building crash prediction models is a relatively rough technique to analyze data. Many characteristics of the intersections needed to be standardized to be able to perform quantitative analyses. Inevitably, this has lead to a number of simplifications. For instance, the variable PEDCROSS only tells us whether pedestrian crossings are present at the intersection or not. It does not tell how many crossings are present, at what type of legs they are located, whether they are raised, whether they are pronounced with a warning sign, how far they are away from the intersection plane, etc. Such micro-level differences are erased by standardizing the reality. These simplifications and generalizations are likely to conceal valuable insights at a more detailed intersection design level.

Furthermore, the presence of a correlation between the number of crashes (the dependent variable) and a certain intersection characteristic (the independent variables) is not sufficient to conclude that there is a causal connection between the intersection characteristic and the number of crashes (i.e. the characteristic causes the number of crashes) (Elvik, 2000b, 2011; Hauer, 2010). The found correlations between the used data and the number of crashes are unarguable, the interpretation of these correlations however is not. The interpretations in this discussion chapter are carefully considered, based on existing literature as much as possible, but can nevertheless not be directly proved by the data. Further research should verify the causality of the relationships that have been found, especially relationships that are not supported by existing literature, or that are not in line with it.

5.3 Recommendations for further research

It is suggested that future research could build a model that is able to calculate the expected crash count for random intersections. Therefore, information from a random

sample of intersections should be collected to be able to claim predictive power. Data availability will be an issue for developing such a model.

Furthermore, a more sophisticated model could be fit that takes into account time effects.

It is also recommended to use more accurate exposure data, i.e. traffic volume counts over a larger time period than the two-hour period of this study, and separate exposure values for different types of road users. Another possibility is to use a more sophisticated and accurate measure of exposure to risk instead of the aggregate measures used in this study, such as the number of encounters or simultaneous arrivals (Elvik, Erke, & Christensen, 2009). Also the possibility to use vehicle volumes of only certain movements (e.g. only left-turning traffic) could be explored in further research.

The findings of this research should be analyzed into more detail, and confirmed by other crash prediction model research in other countries. Especially the causality of some of the correlations that have been found should be studied into more detail.

6. CONCLUSIONS

A number of models and submodels have been fit to identify the most important underlying factors that determine the number of injury crashes at dangerous intersections. Data about 38 variables of 601 dangerous intersections in Flanders are collected. By using crash data from a different period than the one that is used to select the dangerous locations, the regression-to-the-mean effect is eliminated. Negative binomial models are fit to predict the number of injury crashes based on a number of independent variables. This type of model has been chosen because it can deal with the overdispersion that is present in the dataset.

Two general models for all intersections have been fit; one using all injury crashes and one using only serious injury and fatal crashes. Submodels are fit for priority, signalized, four-leg and three-leg intersections, for intersections inside and outside built-up area, and for three different types of major road category intersections. Because the dataset is not random but exists of dangerous locations only, the models cannot claim predictive power but only explanatory power. These models can however provide valuable insights to road designers and policy makers that will help them to handle dangerous intersections more effectively and efficiently. Although all models show some differences from each other, a number of general conclusions can be drawn.

As expected from existing literature, the exposure variables LOGVOLMAJTOTAL and LOGVOLMINTOTAL are important contributing factors to the number of injury crashes that occur. In all models the variables have an estimate lower than one, indicating a positive but less than proportional relationship between the traffic volume and the crash rate.

From the geometric variables, the presence of a median on the major or minor road tends to correspond with a higher number of crashes. Signalized intersections appear to have a higher number of serious injury and fatal crashes. Three-leg intersections on the other hand have a lower crash rate than other types of intersections. Regarding the speed limits, intersections where the approaching legs have a speed limit of 50 km/h tend to have a significantly higher number of crashes than the other categories. Therefore, dangerous locations on 50 km/h roads seem to deserve special attention. The presence of facilities for vulnerable road users has an influence in some models, but the findings are not always consistent. Also the influence of non-perpendicular intersection layouts (DIAGON and 3LEGY) is not completely conclusive, although most models wherein the variables are included seem to indicate a lower number of crashes in case the intersection layout is non-perpendicular. Literature is however ambiguous about both latter findings either. The number of lanes of the intersection legs is only present in a few of the models, mainly due to multicollinearity with other variables.

Regarding the legal status variables, the functional road classification variables (RDCATMAJ and RDCATMIN) are present in the general model and a number of submodels. Most models indicate a significantly higher number of crashes for intersections on primary roads. These intersections deserve special attention from road designers and policy makers. Generally, the trend seems to be that a higher number of crashes occur at intersections from a higher road category. Only the category MAIN does not follow this pattern. One of the submodels indicates a significantly higher number of injury crashes in case the intersection belongs to the bicycle route network, but the variable could be a surrogate for a higher volume of bicyclists.

In a number of models, the land use variables PUBLIC, RESIDENTIAL and ECONOMIC indicate a higher number of injury crashes in case the land use is public facilities, residential or commercial activities respectively. However, these variables are likely to be a proxy for the exposure of vulnerable road users. Furthermore, the variable BUILTUP indicates in the submodel for secondary roads a significantly higher crash count when buildings are present. This stresses the safety problems of ribbon development, which is present at many secondary roads.

It is recommended that future research includes exposure data from other types of road users than only motor vehicles. Furthermore, future research should focus on establishing the causality of the correlations that have been found in the dataset.

7. BIBLIOGRAPHY

Abdel-Aty, M., & Keller, J. (2005). Exploring the overall and specific crash severity levels at signalized intersections. *Accident Analysis & Prevention*, *37*(3), 417–425. doi:10.1016/j.aap.2004.11.002

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332. doi:10.1007/BF02294359

Albrechts, L. (1999). Planners as catalysts and initiators of change. The new structure plan for flanders. *European Planning Studies*, *7*(5), 587–603. doi:10.1080/09654319908720540

Allison, P. D. (2001). *Missing data*. Thousand Oaks, California, USA: SAGE.

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2005). *Statistiek voor economie en bedrijfskunde* (4th ed.). Den Haag: Sdu Uitgevers bv.

AWV. (2009). *Vademecum Veilige Wegen en Kruispunten*. Brussel: Agentschap Wegen en Verkeer.

Baarda, D. B., & de Goede, M. P. M. (2001). *Basisboek Methoden en Technieken* (3rd ed.). Groningen: Wolters-Noordhoff.

Bauer, K. M., & Harwood, D. W. (1999). *Statistical Models of at-Grade Intersection Accidents* (No. FHWA-RD-96-125). Washington D.C., USA: Federal Highway Administration.

Bauer, K. M., & Harwood, D. W. (2000). *Statistical Models of At-Grade Intersection Accidents - Addendum* (No. FHWA-RD-99-094). Washington D.C., USA: Federal Highway Administration.

Belgian Federal Government. (2010). Statistics Belgium - Verkeer. Retrieved June 12, 2011, from http://statbel.fgov.be/nl/statistieken/cijfers/verkeer_vervoer/verkeer/

Carroll, P. S. (1973). Classifications of Driving Exposure and Accident Rates for Highway Safety Analysis. *Accident Analysis & Prevention*, *5*, 81–94.

Chin, H. C., & Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, *35*(2), 253–259. doi:10.1016/S0001-4575(02)00003-9

Cools, M. (2009). *Inter- and Intraday Variability of Flemish Travel Behavior* (Doctoral dissertation). Hasselt University, Diepenbeek.

Creemers, A. (2011). *Models for Incomplete and Clustered Data with Applications in Clinical Trials and Health Economic Studies* (Doctoral dissertation). Hasselt University, Diepenbeek, Belgium.

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2010). Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention*, *42*(2), 393–402. doi:16/j.aap.2009.08.019

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2011). Extended prediction models for crashes at roundabouts. *Safety Science*, *49*(2), 198–207. doi:16/j.ssci.2010.07.016

De Pauw, E., Daniels, S., Brijs, T., Hermans, E., & Wets, G. (2012). *Het programma voor de herinrichting van de gevaarlijke punten op gewestwegen in Vlaanderen: een effectevaluatie* (No. RA-MOW-2011-021). Diepenbeek, Belgium: Policy Research Centre for Traffic Safety.

Dijkstra, A. (2003). *Kwaliteitsaspecten van duurzaam-veilige weginfrastructuur - Voorstel voor een stelsel van DV-eisen waarin alle DV-principes zijn opgenomen* (No. R-2003-10). Leidschendam, The Netherlands: SWOV.

Dijkstra, A. (2010). *Welke aanknopingspunten bieden netwerkopbouw en wegcategorisering om de verkersveiligheid te vergroten? Eisen aan een duurzaam veilig wegennet* (No. R-2010-3). Leidschendam, The Netherlands: SWOV.

El-Basyouny, K., & Sayed, T. (2006). Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transportation Research Record: Journal of the Transportation Research Board*, *1950*(-1), 9–16. doi:10.3141/1950-02

Elvik, R. (1997). Evaluations of road accident blackspot treatment: A case of the iron law of evaluation studies? *Accident Analysis & Prevention*, *29*(2), 191–199. doi:16/S0001-4575(96)00070-X

Elvik, R. (2000a). How much do road accidents cost the national economy? *Accident Analysis & Prevention*, *32*(6), 849–851. doi:10.1016/S0001-4575(00)00015-4

Elvik, R. (2000b). Evaluating the Effectiveness of Norway's "Speak Out!" Road Safety Campaign: The Logic of Causal Inference in Road Safety Evaluation Studies. *Transportation Research Record: Journal of the Transportation Research Board*, *1717*(-1), 66–75. doi:10.3141/1717-09

Elvik, R. (2011). Assessing causality in multivariate accident models. *Accident Analysis & Prevention*, *43*(1), 253–264. doi:10.1016/j.aap.2010.08.018

Elvik, R., Erke, A., & Christensen, P. (2009). Elementary Units of Exposure. *Transportation Research Record: Journal of the Transportation Research Board*, *2103*(-1), 25–31. doi:10.3141/2103-04

Elvik, R., Høye, A., Vaa, T., & Sørensen, M. (2009). *Handbook of Road Safety Measures* (2nd ed.). Bingley, UK: Emerald Group Publishing Limited.

Evans, L. (2004). *Traffic Safety*. Bloomfield Hills, USA: Science Serving Society.

Geurts, K. (2006). *Ranking and Profiling Dangerous Accident Locations using Data Mining and Statistical Techniques* (Doctoral dissertation). Hasselt University, Diepenbeek.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention*, *35*(2), 273–285. doi:16/S0001-4575(02)00005-2

Harnen, S., Umar, R., Wong, S. V., & Hashim, W. W. (2006). Motorcycle accident prediction model for junctions on urban roads in Malaysia. *Advances in Transportation Studies*, *8*. Retrieved from http://trid.trb.org/view.aspx?id=786942

Hauer, E. (1997). *Observational Before-After Studies in Road Safety*. Kidlington, Oxford, UK: Emerald Group Publishing Limited.

Hauer, E. (2010). Cause, effect and regression in road safety: A case study. *Accident Analysis & Prevention*, *42*(4), 1128–1135. doi:16/j.aap.2009.12.027

Jacobs, G., Aeron-Thomas, A., & Astrop, A. (2000). *Estimating global road fatalities* (No. TRL report 445). Crowthorne: Transport Research Laboratory.

Janssen, S. T. M. C. (2004). *Veiligheid op kruisingen van verkeersaders binnen de bebouwde kom - Vergelijking van ongevallenrisico's* (No. R-2003-36). Leidschendam, The Netherlands: SWOV.

Knoblauch, R. L., Tustin, B. H., Smith, S. A., & Pietrucha, M. T. (1988). *Investigation of exposure based pedestrian accident areas: crosswalks, sidewalks, local streets and major arterials. Final report* (No. FHWA/RD-87-038). Washington D.C., USA: Federal Highway Administration. Retrieved from http://trid.trb.org/view.aspx?id=287673

Kulmala, R. (1995). *Safety at rural three- and four-arm junctions - Development and application of accident prediction models* (Doctoral dissertation). Helsinki University of Technology, Espoo, Finland.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition* (2nd ed.). Hoboken, New Jersey, United States of America: Wiley-Interscience.

Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, *38*(4), 751–766. doi:10.1016/j.aap.2006.02.001

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. doi:10.1016/j.tra.2010.02.001

Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis & Prevention*, *37*(1), 35–46. doi:16/j.aap.2004.02.004

Mao, Y., Zhang, J., Robbins, G., Clarke, K., Lam, M., & Pickett, W. (1997). Factors Affecting the Severity of Motor Vehicle Traffic Crashes Involving Young Drivers in Ontario. *Injury Prevention*, *3*(3), 183–189. doi:10.1136/ip.3.3.183

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. *Accident Analysis & Prevention*, *43*(5), 1624–1634. doi:10.1016/j.aap.2011.02.005

Mitra, S., & Washington, S. P. (2007). On the nature of over-dispersion in motor vehicle crash prediction models. *Accident Analysis & Prevention*, *39*(3), 459–468. doi:10.1016/j.aap.2006.08.002

Nambuusi, B. B., Brijs, T., & Hermans, E. (2008). *A review of accident prediction models for road intersections* (No. RA-MOW-2008-004). Diepenbeek, Belgium: Steunpunt Mobiliteit & Openbare Werken, Spoor Verkeersveiligheid.

NCHRP Report 197. (1987). *Cost and Safety Effectiveness of Highway Design Elements*. Washington D.C., USA: Transportation Research Board, National Research Council.

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. doi:10.1007/s11135-006-9018-6

O'Cinneide, D., & Troutbeck, R. J. (1998). At-grade inersections/worldwide review. *Transportation Research Circular*, (E-C003). Retrieved from http://trid.trb.org/view.aspx?id=656813

Oh, J., Lyon, C., Washington, S., Persaud, B., & Bared, J. (2003). Validation of FHWA Crash Models for Rural Intersections: Lessons Learned. *Transportation Research Record: Journal of the Transportation Research Board*, *1840*(-1), 41–49. doi:10.3141/1840-05

Poch, M., & Mannering, F. (1996). Negative Binomial Analysis of Intersection-Accident Frequencies. *Journal of Transportation Engineering*, *122*(2), 105–113.

Qin, X., & Ivan, J. (2001). Estimating Pedestrian Exposure Prediction Model in Rural Areas. *Transportation Research Record: Journal of the Transportation Research Board*, *1773*(-1), 89–96. doi:10.3141/1773-11

Qin, X., Ivan, J., & Ravishanker, N. (2004). Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis & Prevention*, *36*(2), 183–191. doi:10.1016/S0001-4575(02)00148-3

Reurings, M., Janssen, T., Eenink, R., Elvik, R., Cardosa, J., & Stefan, C. (2006). *Accident Prediction Models and Road Safety Impact Assessment: a state-of-the-art*.

SAS Institute Inc. (2008). *Version 9.2 of the SAS System for Windows*. Cary, NC: SAS Institute Inc.

Sawalha, Z., & Sayed, T. (2006). Traffic accident modeling: some statistical issues. *Canadian Journal of Civil Engineering*, *33*(9), 1115–1124.

Shankar, V., Albin, R., Milton, J., & Mannering, F. (1998). Evaluating Median Crossover Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. *Transportation Research Record: Journal of the Transportation Research Board*, *1635*(-1), 44–48. doi:10.3141/1635-06

Svensson, Å., & Hydén, C. (2006). Estimating the severity of safety related behaviour. *Accident Analysis & Prevention*, *38*(2), 379–385. doi:10.1016/j.aap.2005.10.009

Transportation Research Board. (1987). *Designing Safer Roads, Practices for Resurfacing, Restoration and Rehabilitation* (No. Special Report 214). Washington D.C., USA: National Research Council.

Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G., & Isa-Tavarez, J. (2012). The role of built environment on pedestrian crash frequency. *Safety Science*, *50*(4), 1141–1151. doi:10.1016/j.ssci.2011.09.012

Verbeek, M. (2008). *A guide to modern econometrics* (3rd ed.). West Sussex, England: John Wiley and Sons.

Washington, S. P., Karlaftis, M. G., & Mannering, F. L. (2003). *Statistical and Econometric Methods for Transportation Data Analysis, Second Edition*. New York, USA: Chapman and Hall/CRC.

WHO. (2004). *World report on road traffic injury prevention*. Geneva, Switzerland: World Health Organization.

Witten, I. H., & Frank, E. (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations* (1st ed.). San Francisco, USA: Morgan Kaufmann.

Zhang, C. (2008). Defining new exposure measures for crash prediction models by type of collision. *Dissertations Collection for University of Connecticut*, Paper AAI3293728.