Traffic accidents risk analysis based on road and land use factors using GLMs and zero-inflated models

Paweenuch Songpatanasilp, Harutoshi Yamada, Teerayut Horanont and Ryosuke Shibasaki

Abstract

Road related factors have been proven to have substantial effects on traffic accidents. However, the influence of land uses has not been studied deeply so far, in spite of its potentially momentous effect on road accidents occurrence. In urbanized cities such as Tokyo, it is essential to consider the effects of land use and urban planning upon mobility and safety. In this paper, the influence of land use factors are analyzed along with that of road related factors using 1 x 1 kilometer grid meshes. Poisson regression, negative binomial regression, zero-inflated Poisson regression, and zero-inflated negative binomial regression models are estimated. By using road related data and detailed land use data along with traffic accidents data occurred in Tokyo 2013, the numbers of traffic accidents in workdays and non-workdays are modeled separately and the influence of those factors upon the occurrence of traffic accidents is investigated.

P. Songpatanasilp (Corresponding author)
Department of Civil Engineering, University of Tokyo, Tokyo, Japan
Email: pawee@iis.u-tokyo.ac.jp

H. Yamada • R. Shibasaki
Central for Spatial Information Science, University of Tokyo, Tokyo, Japan
Email: yamada.hal@csis.u-tokyo.ac.jp

R. Shibasaki
Email: shiba@csis.u-tokyo.ac.jp

T. Horanont
School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Bangkok, Thailand
Email: teerayut@siit.tu.ac.th

1. Introduction

As science and technology have become more and more instrumental, people have changed their inclination towards modernity and expediency. That has led to the rising number of motor vehicles. From 2007 to 2010, the number of registered vehicles increased by 15% in the world (WHO 2010) and had a direct impact on the road traffic injuries and fatalities. The rapid growth of modernized cities is menacing the safety of non-motorized road users, notably cyclists and pedestrians. Road traffic accidents have a great impact not only on human lives but also on the economies. The global loss due to traffic injuries is estimated to be 518 billion US dollars and is approximately equivalent to 1 to 4% of the gross national products (Integrated Conference of Better Air Quality 2014).

In Japan, 629,021 accidents occurred in 2013, involving 781,494 persons injured and 4,373 fatalities (Statistics Bureau of Japan 2013). Japan is one of the countries that has the lowest number of traffic accidents. Nevertheless, the statistical record indicates that the estimated GDP loss due to traffic casualties in 2009 was 1.4% of its GDP (WHO 2013).

These figures emphasize the significance of discovering risk factors influencing the occurrence of traffic accidents, which will be discussed in this paper along with the statistical analysis of traffic accidents in Tokyo. By identifying influential factors, proper and well-established traffic policies will be developed that will realize the improvement of traffic safety for all road users.

In Japan, the positional information of accidents sites, that is, the longitude and latitude, has just began to be measured since 2012. After that, the spatial analysis of traffic analysis finally became possible in a large scale. This paper describes the first step results of the traffic accidents analysis conducted by using the positional information of them.

This paper is broken down into the following sections. Section 2 shows literature reviews of previous work that have studied traffic accidents and influencing factors. Section 3 describes data source and the processing of data. Section 4 introduces the methodology. Section 5 presents and discusses the estimated results. Section 6 concludes the study and indicates future works.

2. Literature Review

Recently, road safety has received more interests from people around the world. Many researchers have gone into details to determine causes of traffic accidents and to improve road safety. The interests were drawn towards investigating various factors, such as weather, temporal factors, driver's factors and road factors. Land uses along a road are known to be one of the most influential factors on traffic accidents. For example, Dissanayake et al. (2009) has investigated the suitability of using land use variables to predict the number of child pedestrian casualties with Generalized Linear Models (GLMs) in Newcastle upon Tyne, UK. The results show secondary retail and high-density residential land use types are associated with all child pedestrian casualties. Furthermore, educational sites, junction density, primary retail and low-density residential land use types are associated with child casualties at different time periods of the day and week.

Pahukula et al. (2014) has examined the effect of temporal factors and road factors by separating crashes occurred in Texas, USA into five time periods; that is, early morning, morning, mid-day, afternoon and evening. Among many road variables, traffic flow, light conditions, surface conditions, time of the year and percentage of trucks on the road have been identified as key factors that make the difference between time periods. In addition, the combination of land uses and road factors was used in predicting vehicle-pedestrian injury collisions in San Francisco, California; that is, street characteristics, land use characteristics, population characteristics, and commuters' behaviors were used.

Wier et al. (2008) showed that the increase in traffic volume, the proportion of arterial streets without public transit, the proportion of land areas designated as neighborhood commercial use and mixed residential/neighborhood commercial use, employees and residents population and the proportion of people living in poverty indicate an increase of the number of pedestrian injury collisions.

3. Data Description

3.1 Traffic accidents data

Road accidents data used in this study were collected by the National Police Agency of Japan. The data consists of information such as the location and the time of accidents, the gender and age of the primarily responsible party and the severity of accidents. The area investigated in this study contains only mainland Tokyo, where more than 40 thousand accidents occurred in 2013. The histogram of traffic accident occurred on general roads (that is, those roads where the access to roadside facilities is not controlled) on workdays in Tokyo is shown in Fig. 1.

3.2 Land use data

The analysis in this paper is based on 1km mesh and accidents related data are aggregated mesh by mesh. The land use data in this study are from the National Land Numerical Information provided by the Ministry of Land, Infrastructure, Transport and Tourism. The traffic volume and travel speed data are from the Road Traffic Census 2010.

Table 1 shows how twelve land use categories were integrated into following eight categories as some of the land use categories have very small number of meshes:

- Commercial area;
- Industrial area;
- Quasi-industrial area;
- Exclusively industrial area;
- Lower-rise residential area;
- Middle-rise residential area;
- Residential area; and
- Mountainous area and others

The categories which are akin with each other are integrated into one as shown in Table 1. Some categories such as *Industrial, Quasi-industrial* and *exclusively industrial* are not integrated together because the number of meshes they cover is not small.

Lower-rise residential areas are designated for low-rise residential buildings including those used as small shops or offices with a floor area up to 150 m^2. On the other hand, middle rise residential areas are designated for medium to high residential buildings and permitted for small shops and offices with the floor area up to 1500 m^2.

Fig 1. Histogram of traffic accident occurred on general roads on workday in Tokyo 2013 aggregated by 1 x 1 km mesh

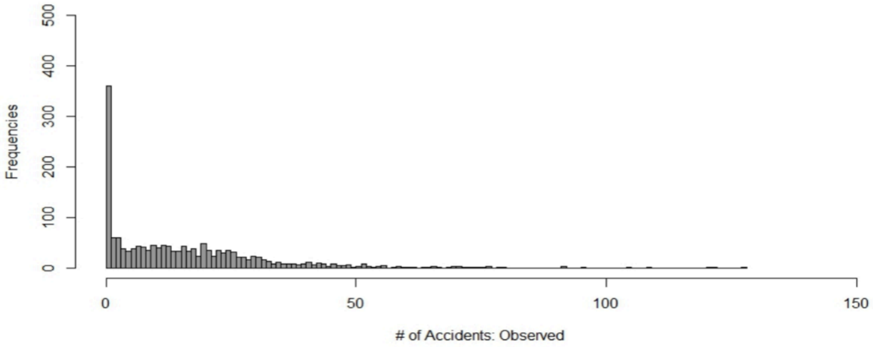


Table 1. Integration table of land use categories

Detailed land uses	Integrated land uses
Low-rise residential 1 Low-rise residential 2	Low-rise residential
Mid-rise residential 1 Mid-rise residential 2	Mid-rise residential
Residential 1 Residential 2 Quasi-Residential	Residential
Neighborhood Commercial Commercial	Commercial
Quasi-Industrial	Quasi-Industrial
Industrial	Industrial
Exclusively Industrial	Exclusively Industrial
Mountainous and others	Mountainous and others

In addition, the POI (Point Of Interest) data were used in this study and divided into ten categories according to the purpose of visit, namely, elementary schools, banks and financial business, shops and convenience stores, department stores, entertainment places such as theater, museums, café and restaurants, etc., hospitals, high schools and universities, sport facilities, railway stations and tourist attractions such as a theme park, zoo, etc.

3.3 Data preparation

The accident data were divided into two sets; one is those accidents occurred in weekdays and the other one is those occurred in weekends and holidays. The traffic accidents data, land use data and road characteristics were imported to Arc GIS program along with the 1km by 1km mesh file of Tokyo in order to calculate the values of explanatory variables mesh by mesh. The basic statistics of objective and explanatory variables are summarized in Table 2. The explanatory variables will be selected among them using statistical testing in the model estimation process. Moreover, a tree model was used to see whether complex interactions among explanatory variables exist (Crawley M. J. 2005). The tree model splits the variables by choosing an appropriate split value, which maximizes the reduction in impurity, until the final nodes are too small to continue the splitting process. Variables that have a correlation coefficient greater than 0.6 with the objective variable were selected as candidates for explanatory variables. If two variables are highly correlated with each other, that is, if the correlation coefficient between them is greater than 0.8, only one of them was chosen to prevent a multi-colinearity issue. The model that gives the best result with the lowest Akaike Information Criterion (AICc) will be used for subsequent analysis.

Table 2. Descriptive statistics of variables

Variable	Min	Mean	Max	S.D.
Objective variables				
# acc. on workdays	0	15.29	127	16.7
# acc. on holidays	0	5.83	58	6.5
Land use elements				
# elementary schools	0	0.8	5	0.9
# banks	0	1.7	98	5.9
# shops	0	80.3	2080	142.9
# dept. stores	0	6.4	324	19.0
# entertain places	0	52.4	2257	147.1
# hospitals	0	0.3	28	1.1
# schools and univ.	0	15.5	331	26.5
# sport facilities	0	2.5	54	4.6
# stations	0	0.4	28	1.4
# attractions	0	6.9	215	16.7
Road related variable				
Len. national rd. (m)	0.0	457.7	7018.2	1028.8
Len. arterial rd. (m)	0.0	1021.4	9072.1	1349.6
Len. municipal rd. (m)	0.0	7118.0	36981.0	5440.0
Len. thin roads (m)	0.0	11656.0	39178.0	8964.4
# intersections	0	462.5	2140	370.2
Traffic vol. (veh/day)	0.0	138.9	688.6	117.2
Traffic speed (km/h)	0.0	23.0	82.00	14.1

	# of meshes
Categorical Variables	
Land use	
• Commercial	69
• Industrial	29
• Quasi-industrial	186
• Exclusive industrial	21
• Lower rise residential	710
• Middle rise residential	199
• Residential	112
• Mountainous and others	333
Densely inhabited district (DID)	
• DID	938
• Partly DID	289
• Non DID	432
Urban area	
• Urbanized	850
• Partly urbanized	438
• Non urbanized	371

Notes
#: Number of acc.: accidents Len.: Length rd.: roads

4. Methods

4.1 Generalized linear models

Generalized linear models (GLM) are widely used in traffic accidents analysis. In this research, two types of GLMs are used; that is, Poisson regression models and negative binomial regression models. Usually, the Poisson distribution is suitable to describe count data such as traffic accidents, because accidents occur only rarely. When traffic accidents are aggregated by geographical meshes, however, there exist excessive meshes which have a zero accident count. This leads to overdispersion. The Poisson distribution that has the same variance value as its mean cannot handle overdispersed data well. There are two solutions to this. One is the adoption of the negative binomial distribution and the other one is the adoption of zero-inflated models. The variance of the negative binomial distribution is greater than its mean and the zero-inflated models can handle the excessive zero counts explicitly. The zero-inflated models, however, do not belong to GLM family and hence will be discussed in the next section.

Let a random variable Y_i be the number of accidents in the i-th mesh, then the probability $\Pr(Y_i = y_i)$ is expressed by Eq. 4.1.1 if Y_i follows the Poisson distribution. If Y_i follows the negative binomial distribution, the probability is expressed by Eq. 4.1.2.

$$P_{Poi}(Y_i = y_i) = \frac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!}, \qquad y_i = 0,1,2\ldots \tag{4.1.1}$$

$$P_{NB}(Y_i = y_i) = \frac{\Gamma(y_i + \tau)}{y_i!\,\Gamma(\tau)}\left(\frac{\tau}{\lambda_i + \tau}\right)^{\tau}\left(\frac{\lambda_i}{\lambda_i + \tau}\right)^{y_i},$$

$$\mathrm{y}_i = 0,1,2\ldots;\ \lambda_i, \tau > 0 \tag{4.1.2}$$

where Y_i represents the number of traffic accidents in each mesh; l_i is the mean of Y_i and t is the overdispersion parameter; and *G()* is the Gamma function. When the value of t approaches zero, the negative binomial distribution approaches the Poisson distribution.

In the framework of GLM, the mean l_i is expressed using a log-link function; that is,

$$E_{Poi}(Y_i) = Var_{Poi}(Y_i) = e^{(X_i'\beta)} \tag{4.1.3}$$

$$E_{NB}(Y_i) = e^{(X_i'\beta)},\ \ Var_{NB}(Y_i) = e^{(X_i'\beta)}(1 + \tau^{-1}e^{(X_i'\beta)}) \tag{4.1.4}$$

where X$_i$ is an explanatory variable vector; b is a coefficient vector and ()' represents transposition of a vector.

The coefficient b is estimated by maximizing the log-likelihood and the appropriate set of explanatory variables are sought for using a log-likelihood ratio test (LRT) and AICc.

4.2 Zero-inflated models

Another method in handling excessive zero observations is zero-inflated models. A zero-inflated model has a mixed distribution function of a binary distribution degenerated to zero and count distribution such as the Poisson or negative binomial distribution, which allows for excessive zero counts.

A zero-inflated Poisson (ZIP) regression model and a zero-inflated negative binomial (ZINB) regression model for a random variable Y_i are shown in Eq. (4.2.5) and Eq. (4.2.6), respectively.

$$P_{Poi}(Y_i = 0) = \omega_i + (1 - \omega_i)e^{-\lambda_i}$$

$$P_{Poi}(Y_i = y_i) = (1 - \omega_i)\left(\frac{e^{-\lambda_i}{\lambda_i}^{y_i}}{y_i!}\right),\ y_i = 1,2,3,\dots \qquad (4.2.5)$$

$$P_{NB}(Y_i = 0) = \omega_i + (1 - \omega_i)\left(1 + \frac{\lambda_i}{\tau}\right)^{-\tau}$$

$$P_{NB}(Y_i = y_i) = (1 - \omega_i)\left(\frac{\Gamma(y_i + \tau)}{y_i!\,\Gamma(\tau)}\right)\left(1 + \frac{\lambda_i}{\tau}\right)^{-\tau}\left(1 + \frac{\tau}{\lambda_i}\right)^{-y_i},$$

$$y_i = 1,2,3,\dots \qquad (4.2.6)$$

where ω_i is a zero-inflation factor and i is the mesh number. From Eq. (4.2.5) and Eq. (4.2.6) it can be seen that, as the value of ω_i approaches zero, ZIP and ZINB regression models approach a Poisson or negative binomial regression model of Eq. (4.1.1) or Eq. (4.1.2). The mean and variance of ZIP and ZINB distributions are presented in Eq. (4.2.7) and Eq. (4.2.8), respectively.

$$E_{Poi}(Y_i) = (1 - \omega_i)\lambda_i$$

$$Var_{Poi}(Y_i) = (1 - \omega_i)\lambda_i(1 + \omega_i\lambda_i) \qquad (4.2.7)$$

$$E_{NB}(Y_i) = (1 - \omega_i)\lambda_i$$

$$Var_{NB}(Y_i) = (1 - \omega_i)\lambda_i \left(1 + \omega_i \lambda_i + \frac{\lambda_i}{\tau}\right) \qquad (4.2.8)$$

The expected value of the Poisson or negative binomial distribution is expressed as an exponential function of a linear combination of explanatory variables, that is, $\lambda_i = e^{X_i'\beta}$, where β is a coefficient vector and x is a covariate vector. ω_i is usually expressed in terms of a logistic function, that is, $\omega_i = \frac{1}{1+e^{-Z_i'\gamma}}$; where, γ is a coefficient vector and Z_i is a zero-inflation covariate vector. From these equations, the log-likelihood functions of ZIP and ZINB can be easily constructed.

4.3 Goodness of fit

The appropriateness, or the *goodness of fit* of estimated models is judged using a log-likelihood ratio test (LRT) (Hoel P. G. 1962) and Akaike Information Criterion with a correction for finite sample sizes (AICc) (Akaike H. 1973) The log-likelihood and AICc are expressed in Eq. (4.3.9) and Eq. (4.3.10), respectively. Original AIC is used to select a more appropriate model for forecast. AICc has the advantage of reducing the possibility of choosing a model having too many parameters.

$$LRT = 2(\log(L_1) - \log(L_0)) \qquad (4.3.9)$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \qquad (4.3.10)$$

where $AIC = 2k - 2\log(L)$; k is the total number of parameters of the model and n is the sample size, $\log(L_0)$ and $\log(L_1)$ are the log-likelihood of the first and second model, respectively.

4.4 The comparison of GLMs and zero-inflated models

LRT is widely used for the comparison of two nested models. However, if two models are not nested LRT is useless. Instead of LRT, the Vuong test is widely used as a test for non-nested models. Wilson, however, suggested that the use of Vuong test to determine whether the zero-inflated model fits the data statistically better than count regression model is incorrect (Wilson P. 2014). In this paper, an alternative approach proposed by Wilson will be used. In the alternative approach, fitted probability values of two models (zero-inflated and non-inflated) are plotted. If points lie approximately along the line y = x, then it is possible to conclude that zero-inflation is not statistically significant.

5. Results And Discussion

Models of workday accidents and those of weekends and holidays accidents were estimated separately. Accidents occurred on expressways, urban expressways or auto-dedicated roads were excluded as access to roadside facilities is fully controlled in these roads. Following four kinds of count models, Poisson regression models, negative binomial regression models, zero-inflated Poisson regression models and zero-inflated negative binomial regression models, were estimated in this study and compared with each other.

The relationship between traffic accidents and the land use along a road is the focus of our interest. Consequently, number of shops, department stores, banks, hospitals etc. in each mesh was adopted as an explanatory variable in addition to such variables as traffic volume, number of intersections and length of roads that represent road and road traffic characteristics.

5.1 Model calibration

After reviewing correlation coefficients among explanatory and objective variables, candidates for explanatory variables were selected. These variables were added to the model one by one in order to monitor the statistical significance of the variables and their effects on AICc. As the number of explanatory variables increases, the value of AICc decreases. Therefore, LRT is conducted for the comparison of the original and the expanded model. If two models are statistically equivalent, the original model was selected. It should be noted that this process has been repeated separately for each type of accidents model.

The term log(Total Length of Road in each mesh) is included as an offset term in all models and the coefficient of the offset term is fixed to one. TLR includes the length of national roads, arterial roads, prefectural and municipal roads, and thin roads within each mesh. Those meshes where the total road length is zero were excluded from the data because we cannot calculate the value of log(0). As no traffic accident occurs in these meshes, this exclusion does not bias the model estimation results.

Most explanatory variables are statistically significant for Poisson or ZIP regression models; and thus, these two models hold the biggest number of parameters. Accidents models of weekends and holidays have less number of significant variables since the numbers of accidents in these days are much smaller than that of workdays; it is approximately one half or one third. Categorical variables such as land use, urban area and DID are highly significant in all of these models.

Categorical variables such as *urban area* and *DID area* are highly correlated with the occurrence of traffic accidents. That is, in urbanized or DID areas, the number of accidents is large while in non-urbanized or non-DID areas the number of accidents is very low and virtually zero. Accordingly, these two categorical variables were selected as a candidate explanatory variable for the zero count part of zero-inflated models. Among them, the *urban area* variable is slightly highly correlated with the occurrence of traffic accidents and was selected for a zero count part explanatory variable. The variable *DID* was used in the count model part.

Traffic volumes and speeds were observed only on workdays and these variables were included only in workday models.

The number of various POIs was found to be significant in all models. However, selected explanatory variables are different among models.

5.2 GLM and zero-inflated models estimation results

The estimation results of Poisson regression and negative binomial regression models for traffic accidents on workdays are summarized in Table 3 and the results of zero-inflated models for traffic accidents on workdays are summarized in Table 4. The estimation results of counterpart models for traffic accidents on weekends and holidays are summarized in Table 5 and Table 6. These tables indicate that the estimated coefficients have same sign and magnitude though the number of explanatory variables of the negative binomial regression model is smaller than that of the Poisson regression model. The same holds true for the estimation results of zero-inflated Poisson regression and zero-inflated negative binomial regression models of workday accidents.

The coefficients of all categories of the variable *land use* are negative. The category commercial is the base case of this categorical variable and the negative coefficients of other categories indicate that traffic accidents occur less frequently in these land use areas. On the other hand, traffic accidents occur most frequently in commercial areas. Traffic accidents occur least frequently in lower-rise residential areas where traffic volume is not so much and land use is restricted to residential buildings. These facts lead to the reduction of traffic conflicts and ultimately to the fewer occurrences of accidents.

Table 3. GLM estimates for traffic accident on workdays

Variable	Poisson regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>|z|)
Intercept	- 8.3500	0.1056	- 79.049	<2e-16
Land use				
Industrial	- 0.4465	0.0566	- 7.893	2.9e-15
Quasi-industrial	- 0.3810	0.0301	- 12.654	<2e-16
Exclusive-industrial	- 0.5607	0.0908	- 6.179	6.5e-10
Lower rise residential	- 0.6959	0.0282	- 24.636	<2e-16
Middle rise residential	- 0.5450	0.0290	- 18.813	<2e-16
Residential	- 0.3825	0.0312	- 12.248	<2e-16
Mountainous	- 0.3647	0.0624	- 5.842	5.2e-09
DID				
DID area	0.6982	0.0802	8.708	<2e-16
Partly DID area	0.5300	0.0794	6.679	2.4e-11
Urban				
Urbanized area	1.1880	0.1244	9.551	<2e-16
Partly urbanized area	1.0110	0.1220	8.291	<2e-16
# of elementary schools	- 0.0292	0.0075	- 3.869	0.0001
# of shops	0.0004	0.0000	6.678	2.4e-11
# of entertain places	- 0.0002	0.0000	- 3.057	0.0022
# of hospitals	0.0098	0.0036	2.768	0.0056
# of sport facilities	0.0056	0.0015	3.724	0.0002
# of attractions	0.0015	0.0004	3.625	0.0003
# of intersections	- 0.0003	0.0000	- 10.805	<2e-16
Avg. traffic volume	0.0022	0.0000	27.932	<2e-16
Avg. traffic speed	- 0.0156	0.0009	- 18.117	<2e-16
				AICc: 11153.29
				Log-likelihood: - 5555.364
				Degree of freedom: 1638

Variable	Negative binomial regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>|z|)
Intercept	- 8.3020	0.1461	- 56.811	<2e-16
Land use				
Industrial	- 0.4009	0.1171	- 3.425	0.0006
Quasi-industrial	- 0.4195	0.0749	- 5.598	<2e-16
Exclusive-industrial	- 0.5448	0.1513	- 3.602	0.0003
Lower rise residential	- 0.6701	0.0713	- 9.393	<2e-16
Middle rise residential	- 0.5407	0.0732	- 7.385	1.5e-13
Residential	- 0.4007	0.0779	- 5.146	2.7e-07
Mountainous	- 0.5675	0.1257	- 4.516	6.3e-06

DID				
DID area	0.7465	0.1047	7.131	10.0e-13
Partly DID area	0.5577	0.1005	5.548	2.9e-08
Urban				
Urbanized area	0.9709	0.1542	6.298	3.0e-10
Partly urbanized area	0.8320	0.1474	5.646	1.6e-08
# of elementary schools	-	-	-	-
# of shops	0.0007	0.0001	6.581	4.7e-11
# of intersections	- 0.0003	0.0000	- 5.460	4.8e-08
Avg. traffic volume	0.0025	0.0002	14.817	<2e-16
Avg. traffic speed	- 0.0157	0.0017	- 9.219	<2e-16
				AICc: 9143.41
				Log-likelihood: - 4554.519
				Degree of freedom: 1643
				Theta: 5.881
				Theta Std. Error: 0.340

Table 4. Zero-inflated model estimates for traffic accident on workdays

Variable	**Zero-inflated Poisson regression model**			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Count component				
Intercept	- 7.3040	0.0778	- 93.818	<2e-16
Land use				
Industrial	- 0.4914	0.0563	- 8.717	<2e-16
Quasi-industrial	- 0.4020	0.0301	- 13.355	<2e-16
Exclusive-industrial	- 0.5965	0.0912	- 6.541	6.1e-11
Lower rise residential	- 0.6812	0.0282	- 24.170	<2e-16
Middle rise residential	- 0.5329	0.0290	- 18.410	<2e-16
Residential	- 0.4034	0.0312	- 12.931	<2e-16
Mountainous	- 0.3721	0.0608	- 6.116	9.6e-10
DID				
DID	0.7939	0.0711	11.163	<2e-16
Partly DID	0.5626	0.0716	7.856	4.0e-15
# of elementary schools	- 0.0262	0.0076	- 3.446	0.0006
# of shops	0.0004	0.0000	7.715	1.2e-14
# of entertain places	- 0.0002	0.0000	- 3.875	0.0001
# of hospitals	0.0122	0.0035	3.471	0.0005
# of sport facilities	0.0070	0.0015	4.670	3.0e-06
# of attractions	0.0012	0.0004	2.959	0.0031
# of intersections	- 0.0003	0.0000	- 11.027	<2e-16
Avg. traffic volume	0.0022	0.0000	28.923	<2e-16
Avg. traffic speed	- 0.0176	0.0009	- 19.002	<2e-16

Zero-inflated component				
Intercept	0.8001	0.1933	4.139	3.5e-05
Urban area				
Urbanized	- 8.1748	2.2028	-3.711	0.0002
Partly urbanized	- 4.4753	0.4264	- 10.496	<2e-16
				AICc: 11128.460 Log-likelihood: - 5541.92 Degree of freedom: 1637

Variable	Zero-inflated negative binomial regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Count component				
Intercept	-7.4530	0.1170	- 63.709	<2e-16
Land use				
Industrial	- 0.4033	0.1150	- 3.506	0.0005
Quasi-industrial	- 0.4184	0.0719	- 5.822	5.8e-09
Exclusive-industrial	- 0.5409	0.1493	- 3.622	0.0003
Lower rise residential	- 0.6446	0.0679	- 9.490	<2e-16
Middle rise residential	- 0.5159	0.0700	- 7.375	1.6e-13
Residential	- 0.4005	0.0749	- 5.346	9.0e-08
Mountainous	- 0.4684	0.1193	- 3.926	8.6e-05
DID				
DID	0.8155	0.0907	8.994	<2e-16
Partly DID	0.5781	0.0890	6.493	8.4e-11
# of elementary schools	-	-	-	-
# of shops	0.0008	0.0001	6.962	3.4e-12
# of intersections	- 0.0003	0.0000	- 4.923	8.5e-07
Avg. traffic volume	0.0025	0.0002	14.706	<2e-16
Avg. traffic speed	- 0.0172	0.0018	- 9.706	<2e-16
Log (theta)	1.8060	0.0583	30.955	<2e-16
Zero-inflated component				
Intercept	0.5458	0.2250	2.425	0.0153
Urban area				
Urbanized	- 17.3135	164.8380	- 0.105	0.9163
Partly urbanized	- 4.8957	0.6593	- 7.426	1.1e-13
				AICc: 9121.593 Log-likelihood: - 4542.588 Degree of freedom: 1641

Table 5. GLM estimates for traffic accident on weekends and holidays

Variable	Poisson regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Intercept	- 9.2660	0.1406	- 65.900	<2e-16
Land use				
Industrial	- 0.2167	0.0848	- 2.554	0.0106
Quasi-industrial	- 0.3800	0.0467	- 8.132	4.2e-16
Exclusive-industrial	- 0.5572	0.1475	- 3.777	0.0002
Lower rise residential	- 0.6545	0.0422	- 15.508	<2e-16
Middle rise residential	- 0.4850	0.0442	- 10.970	<2e-16
Residential	- 0.3603	0.0477	- 7.552	4.3e-14
Mountainous	- 0.2608	0.0970	- 2.689	0.0072
DID				
DID	0.7779	0.1217	6.391	1.6e-10
Partly DID	0.5072	0.1210	4.192	2.8e-05
Urban area				
Urbanized	1.0120	0.1767	5.728	1.0e-08
Partly urbanized	0.7638	0.1722	4.435	9.2e-06
# of banks	- 0.0053	0.0015	- 3.551	0.0004
# of shops	0.0005	0.0002	5.585	2.3e-08
# of sport facilities	0.0137	0.0022	6.282	3.3e-10
# of intersections	- 0.0003	0.0000	- 6.027	1.7e-09
				AICc: 7807.622
				Log-likelihood: - 3887.645
				Degree of freedom: 1643

Variable	Negative binomial regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Intercept	- 9.2900	0.1749	- 53.129	<2e-16
Land use				
Industrial	- 0.0676	0.1387	- 0.488	0.6257
Quasi-industrial	- 0.3117	0.0882	- 3.535	0.0004
Exclusive-industrial	- 0.4841	0.2014	- 2.404	0.0162
Lower rise residential	- 0.5544	0.0825	- 6.717	1.9e-11
Middle rise residential	- 0.3968	0.0850	- 4.668	3.0e-06
Residential	- 0.2753	0.0909	- 3.028	0.0025
Mountainous	- 0.2771	0.1527	- 1.815	0.0695
DID				
DID	0.7819	0.1396	5.602	2.1e-08
Partly DID	0.5078	0.1360	3.733	0.0002
Urban area				
Urbanized	0.8821	0.2020	4.366	1.3e-05
Partly urbanized	0.6754	0.1937	3.487	0.0005

# of shops	0.0006	0.0002	3.545	0.0004
# of sport facilities	0.0157	0.0044	3.467	0.0005
# of intersections	- 0.0002	0.0000	- 3.183	0.0015

AICc: 7118.926
Log-likelihood: - 3543.297
Degree of freedom: 1644
Theta: 5.287
Theta Std. Error: 0.389

Table 6. Zero-inflated model estimates for traffic accident on weekends and holidays

Variable	Zero-inflated Poisson regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Count component				
Intercept	- 8.5140	0.1127	- 75.552	<2e-16
Land use				
Industrial	- 0.2167	0.0847	- 2.558	0.0105
Quasi-industrial	- 0.3691	0.0465	- 7.950	1.9e-15
Exclusive-industrial	- 0.5842	0.1494	- 3.910	9.2e-05
Lower rise residential	- 0.5908	0.0416	- 14.194	<2e-16
Middle rise residential	- 0.4374	0.0436	- 10.026	<2e-16
Residential	- 0.3495	0.0475	- 7.360	1.8e-13
Mountainous	- 0.2884	0.0950	- 3.036	0.0024
DID				
DID	0.9388	0.1072	8.757	<2e-16
Partly DID	0.5686	0.1094	5.200	2.0e-07
# of shops	0.0003	0.0000	5.105	3.3e-07
# of sport facilities	0.0149	0.0022	6.822	9.0e-12
# of intersections	- 0.0002	0.0000	- 5.776	7.7e-09
Zero-inflated component				
Intercept	0.1747	0.2809	0.622	0.5340
Urban area				
Urbanized	- 4.5422	0.4387	- 10.354	<2e-16
Partly urbanized	- 3.3024	0.4575	- 7.218	5.3e-13

AICc: 7801.939
Log-likelihood: - 3884.804
Degree of freedom: 1643

Variable	Zero-inflated negative binomial regression model			
	Coef. Estimates	Std. Error	z-value	Pr(>lzl)
Count component				
Intercept	- 8.6386	0.1406	- 61.454	<2e-16
Land use				
Industrial	- 0.0531	0.1414	- 0.376	0.7071

Quasi-industrial	- 0.3184	0.0858	- 3.712	0.0002
Exclusive-industrial	- 0.5007	0.1995	- 2.509	0.0121
Lower rise residential	- 0.5225	0.0799	- 6.541	6.1e-11
Middle rise residential	- 0.3689	0.0824	- 4.478	7.5e-06
Residential	- 0.2853	0.0887	- 3.215	0.0013
Mountainous	- 0.2747	0.1452	- 1.892	0.0585
DID				
DID	0.9122	0.1234	7.389	1.5e-13
Partly DID	0.5416	0.1232	4.397	1.1e-05
# of shops	0.0006	0.0002	3.472	0.0005
# of sport facilities	0.0171	0.0044	3.845	0.0001
# of intersections	- 0.0002	0.0001	- 2.192	0.0284
Log (theta)	1.6736	0.0798	20.958	<2e-16
Zero-inflated component				
Intercept	- 0.0349	0.3293	- 0.106	0.9157
Urban area				
Urbanized	- 6.0624	1.7796	- 3.407	0.0007
Partly urbanized	- 4.6811	1.3964	- 3.352	0.0008

AICc: 7129.106
Log-likelihood: - 3547.366
Degree of freedom: 1642

The positive signs of the coefficients of DID/partial DID and urbanized/partly-urbanized variables suggest that traffic accidents occur more frequently in these areas in comparison to non-DID/non-urbanized areas.

With regard to POI related variables, the coefficients of # of shops, hospitals, sport facilities, and attractions are positive. These POIs attract many vehicles/people and therefore traffic accidents occur more frequently near them. On the other hand, the coefficients of # of elementary schools, banks and entertainment places are negative. The negative sign of # of elementary schools seems to be contrary to the expectation because there are many pupils near these facilities. However, various traffic safety measures have been already introduced around them and the traffic safety has been improved. The variable # of banks appears only in weekends and holidays model but the coefficient is negative. One explanation for this is that banks are closed on holidays. Therefore, people do not go to banks on holidays. In addition, many banks are located in central business districts in Tokyo. These facts indicate that it is possible to interpret the variable # of banks as a proxy variable for central business districts.

The coefficients of *# of intersections* and *average traffic speed* have negative signs. The slower the traffic speed, the lower the number of traffic accidents. This is a reasonable result. The interpretation of the negative sign of # *of intersections* is not straightforward. But the fact that the number of intersections, particularly small intersections is larger in residential areas in comparison to other land use areas indicates that the negative sign of # *of intersections* is the reflection of the less frequent occurrence of traffic accidents in residential areas. The coefficient of # *of traffic volume* is positive and it is quite reasonable.

5.3 The comparison of GLMs and zero-inflated models using fitted probabilities

The GLMs and zero-inflated models estimated in this paper for workday accidents and non-workday accidents are compared using a graphical method proposed by Wilson (2014). He suggests plotting the individual fitted probabilities of the observed data under the zero-inflated and non-zero inflated model against each other. If the points lie approximately along the line x = y, then zero-inflation is not indicated.

The fitted probabilities of traffic accidents on workdays are shown in Fig. 3 and the fitted probabilities of traffic accidents on non-workdays are shown in Fig. 4. In both figures, graph (a) shows Poisson versus ZIP regression model comparison and graph (b) shows negative binomial vs ZINB regression model comparison.

In all figures, There are three lines; the first line lies approximately on the line x=y while the second and third lines lie under and above the line x=y respectively. Starting from point (0,0) and (1,1), the second and third line deviate from the line x=y as the value of x gets higher or lower respectively. The characteristics of the accidents that are apart from the line x=y have been investigated and found that almost all occurred in mountainous, non-DID and non-urbanized meshes. These figures show the significance of zero-inflated models in comparison to non-inflated models.

6. Summary And Conclusions

Traffic accidents data of the year 2013 in Tokyo were analyzed using land use data and road-related data; and the number of traffic accidents on weekdays and holidays were modeled separately using GLMs and zero-inflated models on the basis of 1 x 1 kilometer meshes. Poisson regression models, negative binomial regression models, ZIP regression models and ZINB regression models were built and compared with each other. In the model selection process, models were compared using AICc and log-likelihood ratio tests. In addition, for the comparison of non-nested models, the graphical approach proposed by Wilson (2014) was adopted.

The results, presented and discussed in section 5, indicate the closeness between GLM estimates and zero-inflated model estimates. Namely, the sign and magnitude of estimated coefficients are approximately the same with each other. The comparison between the Poisson regression model and the negative binomial regression model revealed that the negative binomial regression models have a better fit because they have smaller AICc values for traffic accident on workdays and holidays. The same result was obtained by the comparison between the zero-inflated Poisson regression model and the zero-inflated negative binomial regression model.

In addition, according to the comparison between the Poisson regression model and the zero-inflated Poisson regression model using the fitted probability values, the zero-inflation was indicated due to the influence of land use factor. We have obtained the same results from the comparison between the negative binomial regression model and the zero-inflated negative binomial regression model.

With regard to the relationship between land uses and traffic accidents, it was found that accidents occur more frequently in commercial areas than in other land use areas and that accidents occur least frequently in residential areas. In addition, accidents occur more frequently around shops, hospitals, sport facilities and attractions on workdays. On holidays, accidents occur more frequently around shops and sport facilities.

The accidents analysis in this paper is limited to those occurred in Tokyo. It is necessary to investigate accidents occurred in other areas, in particular in rural areas. Moreover, the estimated models are global one; that is, geographical difference is not taken into account. It is necessary to use, for example, geographically weighted models (GWM) to investigate the geographical difference.

There have been some studies investigating the spatial correlation of traffic accidents. For example, Aguero-Valverde et al. (2005) found the existence of spatial correlation in the injury crash analysis in Pennsylvania using full Bayes hierarchical models. However, no evidence of spatial correlation was found in the fatal crash.

With respect to traffic accident, Quddus et al. (2008) suggested that the Bayesian hierarchical models are more appropriate in developing the relationship between area-wide traffic crashes due to its ability in taking account of both spatial dependence and uncorrelated heterogeneity.

In dealing with count data with overdispersion, Geographically Weighted Negative Binomial Regression (GWNBR) has been proposed by da Silva et al. (2012). Spatial correlation is one of causes of overdispersion. It is, therefore, necessary to investigate if spatial correlation exists in Japanese traffic accidents data. This is one of our future research themes.

In addition, our colleagues have recently found that it is possible to estimate traffic volume of vehicles, public transit users and pedestrians separately using GPS data of mobile phones (Witayangkurn et al. 2013). Therefore, it is now possible to estimate the traffic volume at anytime and anywhere in principle without much difficulty. The estimated traffic volume can be used as an exposure variable to accident risk. We also try to incorporate these exposure variables into our models.

Fig. 3 Fitted probabilities of traffic accidents on workdays under (a) Poisson and ZIP regression models and (b) negative binomial and ZINB regression models

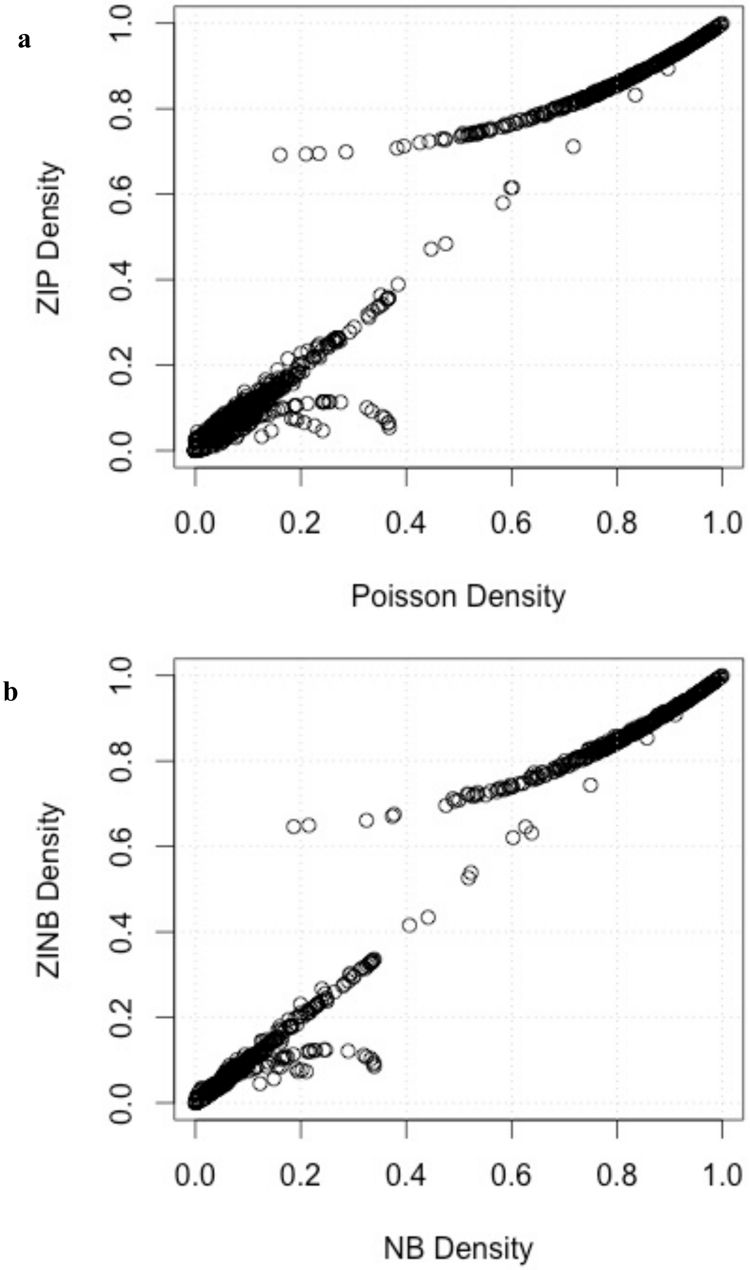
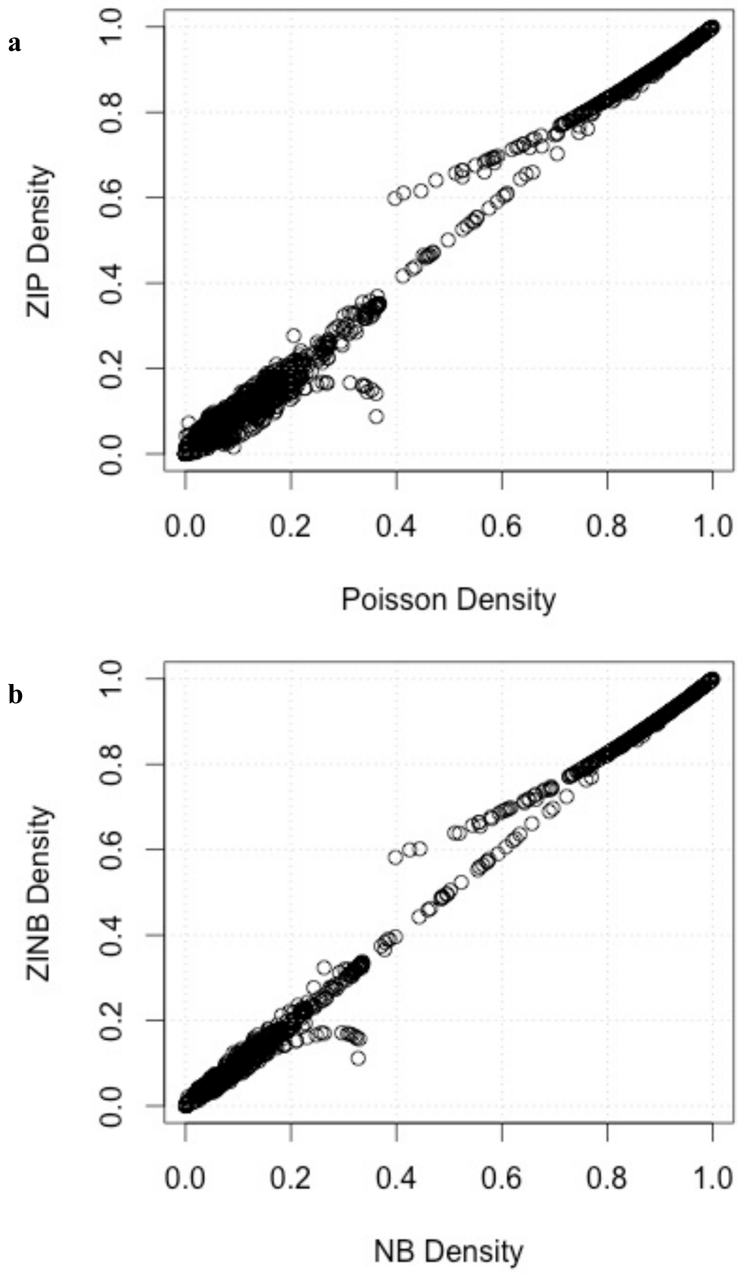


Fig. 4 Fitted probabilities of traffic accidents on weekends and holidays under (a) Poisson and ZIP regression models and (b) negative binomial and ZINB regression models



References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csaki, eds., 2nd International Symposium on Information Theory (Akademia Kiado, Budapest), pp. 267-281.

Aguero-Valverde, J. & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis and Prevention*, *38*(3), 618–25.

Crawley, M J (2005) *Statistics an introduction using R,* John Wiley and Sons, Ltd,.

Crowley, S. (2012) Maximum Likelihood Estimation of the Negative Binomial Distribution, (6), 2–3.

Da Silva, A. R., & Rodrigues, T. C. V. (2014). Geographically Weighted Negative Binomial Regression—incorporating overdispersion. *Statistics and Computing*, 769–783.

Dissanayake, D., Aryaija, J., & Wedagama, D. M. P. (2009). Modelling the effects of land use and temporal factors on child pedestrian casualties. *Accident Analysis and Prevention*, *41*(5), 1016–24.

Fang, R. (2013). Zero inflated negative binomial (ZINB) regression model for over-dispersed count data with excess zeros and repeated measures, an application to human microbiota sequence data.

Hoel, P. G. (1962) Likelihood Ratio Tests. §9.1.4 in *Introduction to Mathematical Statistics, 3rd ed.* New York: Wiley, pp. 220-228.

Ismail, N., Ph, D., Zamani, H., & Ph, D. (2013). Estimation of Claim Count Data using Negative Binomial , Generalized Poisson , Zero-Inflated

Negative Binomial and Zero-Inflated Generalized Poisson Regression Models, (1992), 1–28.

Japan Digital Road Map Association retrieved from http://www.drm.jp/english/drm/e_index.htm

Lord, D., & Park, B. (2005). Negative Binomial Regression Models and Estimation Methods, 1–15.

Mwalili, S. M., Lesaffre, E., & Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*, *17*(2), 123–39.

Note, F. C., & Lanka, S. (2014). Integrated Conference of Better Air Quality (BAQ) 2014 Intergovernmental Eighth Regional Environmentally Sustainable Transport (EST) Forum in Asia Theme : Next Generation Solutions for Clean Air and Sustainable Transport – Towards a Livable Society in Asia, (2010).

Pahukula, J., Hernandez, S., & Unnikrishnan, A. (2015). A time of day analysis of crashes involving large trucks in urban areas. *Accident Analysis and Prevention*, *75C*, 155–163.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. *Accident Analysis and Prevention*, 40 (4), pp. 1486-1497.

Road Traffic Census (2010) http://www.mlit.go.jp/road/census/h22-1/

Shi, X. (2014). A Nondegenerate Vuong Test ∗, 1–40.

Statistic Bureau of Japan (2013) Road Traffic Accident.

Strauss, J., Miranda-Moreno, L. F. & Morency, P. (2014). Multimodal injury risk analysis of road users at signalized and non-signalized intersections. *Accident Analysis and Prevention*, *71*, 201–9.

Vuong, Q. H. (1989). Likelihood ratio test for model selection and non-nested hypotheses. *Econometrica*, Volume 57, Issue 2, 307-333.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis and Prevention*, *41*(1), 137–45.

Wilson, P. (2014). The Misuse of The Vuong Test For Non-Nested Models to Test for Zero-Inflation, Proceeding of the 29th International Workshop on Statistical Modeling, 2, 1-6.

Witayangkurn A.,Horanont T., Ono N., Sekimoto Y., Shibasaki R. (2013) Trip reconstruction and transportation mode extraction on low data rate gps data from mobile phone. In Proceedings of the International Conference on Computers in Urban Planning and Urban Management (CUPUM 2013), pp.1-19.

World Health Organization (2010) World Health Statistics 2010.

World Health Organization (2013) Global Status Report on road safety.

Zone, L. U., Control, B. & Plan, D. (n.d.). Introduction of Urban Land Use Planning System in Japan Outline of Urban Land Use Planning System. http://www.mlit.go.jp/en/index.html