A New Perspective for Modeling Traffic Accidents Considering Unrecorded Data

Amr M. Wahaballa

Abstract— There are many factors that expected to affect traffic accidents are not recorded such as driver reaction time and fatigue. If the effect of these factors on accident rates cannot be considered, any accident model based on these predictions may be inappropriate. However, while observing all accident causes is difficult, the alternate is employing advanced methodologies to extract the effects of unrecorded data from the observed one. The goal of this paper is to model accident rates considering the unrecorded data affecting them using a model that can be handled for use in real-world practice. For this purpose, the suggested method employed the stochastic frontier model that allow estimating two different effects. The effect of the observed factors is related to the frontier and the effect of the unrecorded factors is estimated as the inefficiency of the frontier. The method is applied to a real traffic accidents data as a proof of concept. The cost frontier function is used to represent the relationship between the accident rate as an output and the pavement width, the percent of trucks and the number of access points per kilometer as inputs. Model parameters are estimated by the maximum likelihood method. P-values show that all estimated parameters are statistically significant and the estimation proves a quick convergence. Comparing the accident rate values estimated by the proposed model versus the actual accident rate values shows a goodness-of-fit determination coefficient of more than 95%. The findings reveal that the inefficiency term (which represents the effect of unrecorded factors) has markedly affect accident rate values. This result reflects the usefulness of the proposed model and the importance of considering the data that may be unrecorded.

Keywords— Traffic accidents rate, observed factors, unrecorded factors, modeling, econometric models, stochastic frontier models.

1 Introduction

TRAFFIC accidents causing loss of life, damage to properties, and a notable psychological effect on victims and their families worldwide. Annually, traffic accidents produce more than 50 million injuries and 1.2 million deaths all over the world [1]. Decision-makers need accurate information about the relationships between accidents and their causes. Developing accident prediction models can help in predicting accidents causes effectively and allow transportation authorities to provide accurate road safety measures. Therefore, it is the subject of several studies all over the world to demonstrate accidents causes toward reducing their harmful effects. Traffic accidents modeling has been widely studied in the literature using different techniques during last decades (e.g., [2]-[19]). Although different methodologies have been employed in accident modeling research area, there still numerous basic issues that have not been totally addressed such as considering the effect of the unrecorded factors that affecting accident rates [20]. However, in real-world data, there are many factors that affecting accident rates are not recorded. Many human factors such as driver's age, gender, and other socioeconomic characteristics can be observed. However, drivers perception and reaction time, stress, fatigue, and/or emotional conditions at accident occurrence moment, may not be observable. In addition, some environmental conditions in the place of accident occurrence may not be recorded.

Amr M. Wahaballa is a Ph.D. holder and an assistant professor in the department of civil engineering, Aswan University, Egypt. He is currently a postdoctoral researcher in civil engineering department, Gifu University, Japan. PH-+819044087247. E-mail: amr.whbala@aswu.edu.eg. Gifu-shi, Gifu 501-1193, Japan.

Omitting the effect of these factors may cause incorrect estimations and consequently, lead to improper accident mitigation strategies that may result in more economic and social losses. For example, considering driver age as a representative parameter for the effect of people age on the accident rate is not accurate because of the existence of many embedded factors such as drivers perception and reaction time and body positioning at the time of accident occurrence. Considering driver age as one parameter omit the effect of these factors that vary across people of the same age and affecting accidents because people of the same age are likely to have differences in these unrecorded factors [20]. There are some researchers dealt with considering the effect of unrecorded factors on accidents (e.g., [21], [22], [23]). They used latent-class models that address the effect of unrecorded factors by classifying data into homogeneous characteristics subgroups. However, due to the assumed homogeneous characteristics of each subgroup, these models do not consider the variation among the observations of the same subgroup [20]. Other research such as [13] and [24] have used random-parameters models that allow variating the parameters among observations to consider the effect of unrecorded factors. These models may be more complex and have a large number of parameters which may be costly ineffective in terms of computational time required instead of few-variables models needed by traffic operators. Closing this gap in the literature is the objective of this paper.

In a recent research by the author among others [25], the stochastic frontier model is proved to be a simple and successful tool for modeling some transportation problems. In this regard, this paper suggests a methodology for modeling traffic accidents using stochastic frontier approach. To the best of the author's knowledge, this study is one of the first attempts to do so. The maximum likelihood estimation of the stochastic frontier model is applied to a real-world accident data in Egypt to

illustrate the proposed method. To test the prediction accuracy of our model, a long horizon three years accident data on Aswan-Cairo road was utilized. The statistical analysis will show that the results indicate a satisfactory methodology.

Following this outline, this paper is structured as follows. The next section provides a review of the relevant research. Section 3 presents the methodology used in this paper. Section 4 illustrates the application site, data characteristics and developing the accident model. The frontier model results are discussed in Section 5. Finally, Section 6 outlines the main conclusions of this research.

2 Literature Review

Traffic accidents modeling has been widely investigated in the literature using various modeling approaches. Traditional Poisson regression model has been used to model the factors affecting accident rates (e.g., [2], [3], [4]). The accident rate is the number of accidents per the number of vehicles traveled on a road section during a certain time interval. However, simple Poisson regression model set limits when the mean accident rate is much greater than its variance [20]. To overcome these limitations, many researchers such as [5], [6], [7] have used the negative binomial model or Poisson–Gamma models. A wide methodological progress of traffic accident modeling has been made during many decades. Some researchers (e.g., [3], [6]) have used zero-inflated Poisson and negative binomial models to model traffic accidents in case of no accidents observed in some parts of the road. These models considered observations with zero accident rate by categorizing the road into two different clusters, a zero accident rate category and a category for observations that have accident rate values. Other researchers have used duration models (e.g., [8], [9], [10]); bivariate and multivariate models (e.g., [11], [12], [13]); generalized estimating equation models (e.g., [14], [15]); hierarchical/multilevel models (e.g., [12], [16]) and Poisson-lognormal (or Poisson–Weibull) models such as [17], [18], [19]. Artificial intelligence and machine learning approaches also have been utilized for traffic accident modeling in the literature (e.g., [26], [27], [28], [29]). An extensive review of the various modeling methods that used in traffic accident research area can be found in [20] and [30].

Although these wide research exist in traffic accident modeling, few researchers investigated the consideration of unrecorded factors effect. Some studies have tried to address this problem in the literature using finite-mixture or latent-class models to addressing the effect of unrecorded factors by classifying data into homogeneous characteristics subgroups [21], [22], [23]. Finite-mixture or latent-class models based on determining a finite number of mass points to identify homogeneous subgroups of data not just categorizing road sections as in the case of zero-inflated Poisson models by [3] and [6]. These models can be performed without any distributional assumptions for the variations of parameters across subgroups. However, due to the assumed homogeneous characteristics of each subgroup, these models do not consider the variation among the observations of the same subgroup [20]. Other research such as [13] and [24] have used random-parameters models that allow variating the parameters among observations to consider the effect of unrecorded factors. These models can address the variation of parameters across the observations of different road sections and/or across any number of different groups of the data. Later, [42] developed a model that incorporating random parameters within a finite-mixture model. Although their model considers the variations among subgroups and the heterogeneity within each subgroup, this model may be more complex and has a large number of variables that may be computationally inefficient.

Previous research shows that due to limited data availability of many variables known to be significantly affecting accident rates, and the need to develop a simplified models containing few explanatory variables, advanced statistical methodologies are needed to satisfy the balance between these tradeoffs (simple model provides operators needs versus considering the effect of most affecting factors including unrecorded ones). Recently, the stochastic frontier model as detailed in the next section has used in different transportation models. For example, it was used to define travel time frontiers (or travel time budgets) e.g., [32], and for investigating the relationship between time expenditures and time budgets and its impact on episodic well-being measures using survey data [33]. Canavan et al. [34] have used the stochastic frontier approach to model the effect of the frequencies of delay incidents on the efficiency of a metro rail system. Wahaballa et al. [25] highlighted the superiority of the stochastic frontier model for estimating the platform waiting time on London underground based on smart card data. This paper suggests that the stochastic frontier model may be useful for modeling traffic accidents considering the effect of unrecorded data.

3 Methodology

This paper suggests that the factors affecting accident rates can be classified into two groups: one which is observable, and relates to the observed possible factors affecting accident rates, and another that is based on some factors known to significantly affect accident rates, however, may not be available, or other unexpected factors linked to environmental or human factors. The effect of these two groups of factors on accident rates can be treated differently. From the econometric point of view, this paper employs the stochastic frontier techniques to distinguish these factors. Observed accident causes determine a lower frontier measuring the minimum accident rates representing the unavoidable accident rates under current conditions (without implementing any improvements). While the actual accident rates will exceed that minimum, the difference can be related to unrecorded factors or unexpected human factors, for example. Therefore, that difference is modeled as the "inefficiency" term of the stochastic frontier model (SFM).

The SFM is a model used for the analysis of economic efficiency by estimating the production or the cost [35]. It is an extension of a regression model in which a production (cost) function represents the ideal maximum output attainable (the minimum cost of producing that output) given a set of inputs. The general formulation of the SFM is presented here without detailing the derivation of the corresponding criterion functions, for more details the reader can refer to [36]. The main idea of

the SFM is that no economic agent can exceed the ideal (frontier) and the deviations from this extreme represent the individual inefficiencies. The main advantage of the SFM is the ability to introduce a disturbance term consisting of two different error types, noise, and inefficiency, to separate the effects of random noise from the inefficiency. This is achieved by developing a regression model contains a composite error term in which the measurement error and any other classical noise are included together with a one-sided disturbance error term representing the inefficiency.

In production-related applications, the uncontrollable factors of the production unit, such as faulty machinery and breakdowns, are considered as noise. The errors that come from the non-optimal use of technology are captured by the technical inefficiency term. Therefore, in this application, the SFM can be used to represent the unknown/unrecorded factors that affect accident rates, by a distribution different from the observed possible factors affecting them. The latter distributions can be assumed to be classical noise while the unrecorded factors can be considered a one-sided disturbance error term. The SFM cost minimization function is formulated as follows:

$$y_i = \boldsymbol{\beta^T x_i} + \varepsilon_i \quad (1)$$

where

$$\varepsilon_i = y_i - \boldsymbol{\beta^T x_i} = v_i + u_i \quad (2)$$

y_i and $\boldsymbol{x_i}$ represent the output (cost) and the inputs of the ith productive unit, respectively, and $\boldsymbol{\beta}$ is a vector of the unknown frontier parameters (fixed for all i). The composed error term ε_i is the sum of the random noise (v_i) and inefficiency (u_i). The maximum likelihood method can be used for estimating the frontier model parameters. The estimation of the inefficiency scores is recovered in a second step by applying the estimator developed by Jondrow et al. [37], which is based on the information on u_i contained in the overall residual [38]. This estimation requires some assumptions for the components of the error term as follows [38]:

- u_i and v_i are assumed to be independent and identically distributed (IID) across observations and are independent of each other.
- v_i is a two-sided normal distribution N(0; σ_v^2).
- u_i is a nonnegative random term that follows a one-sided distribution.

The distribution of the inefficiency component has been mostly specified in the relevant literature as being half-normal, exponential or truncated normal. In this paper, the cost frontier function defines the relationship between the accident rate as an output and the pavement width, the shoulder width, the percent of trucks and the number of access points per kilometer as normally distributed inputs. This assumption simplifies estimating the parameters of the unrecorded factors distribution. Given the additive property of the normal distributions, as the observed factors are normally distributed, their sum is also a normal distribution. The variance of the summated normal distribution of the pavement width, the shoulder width, the percent of trucks and the number of access points per kilometer is considered as the noise. For the inefficiency, representing the unrecorded factors, the half normal and the exponential distributions are tested. The inefficiency error term can be estimated based on a vector of variables (Z) as follows [39]:

$$u_i = Z_{\boldsymbol{i}}\,\varphi + \omega_i \quad (3)$$

Where φ is a vector of parameters to be estimated, and ω_i is a random variable which comes from the half normal or exponential distribution. The constraint $\omega_i \geq -Z_i\,\varphi$ should be attained to satisfy a positive value for the random disturbance related to the inefficiency. The derivation of the likelihood function assumes independence between u_i and v_i. Since the composite error term ε_i is $v_i + u_i$, its probability density function is the convolution of the two component densities. The log-likelihood function for the normal/exponential cost frontier based on the output y_i can be obtained from the joint probability density function (pdf) of (u_i, v_i) using the transformation $\varepsilon_i = y_i - \boldsymbol{\beta^T x_i}$ as follows [36]:

$$ln\,L(\beta, \sigma_u) = \sum_{i=1}^{n}\left\{-ln\sigma_u + \frac{\sigma_v^2}{2\sigma_u^2} + ln\Phi\left(\frac{\varepsilon_i - \frac{\sigma_v^2}{\sigma_u}}{\sigma_v}\right) - \frac{\varepsilon_i}{\sigma_u}\right\} \quad (4)$$

Where

β :the unknown parameters to be estimated (fixed for all observations).

σ_u :the standard deviation of the unrecorded factors distribution to be estimated.

σ_v^2 :the sum of variances of the pavement width, the shoulder width, the percent of trucks and the number of access points per kilometer.

$\Phi()$:the cumulative distribution function of the standard normal distribution.

ε_i :the error term for observation i ($= y_i - \boldsymbol{\beta^T x_i}$).

n :the total number of accident observations.

For the normal/half normal cost frontier, the log-likelihood function is [36]:

$$ln\,L(\beta, \sigma_u) = \sum_{i=1}^{n}\left\{\frac{1}{2}ln\left(\frac{2}{\pi}\right) - ln\sigma + ln\Phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right) - \frac{\varepsilon_i^2}{2\sigma^2}\right\} \quad (5)$$

Where $\lambda = \frac{\sigma_u}{\sigma_v}$ and $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$.

The maximum likelihood estimation for the stochastic frontier model allows β, φ, σ_v^2 and σ_u^2 to be estimated jointly.

4 APPLICATION

The illustrated method of the likelihood maximization of the stochastic frontier model is applied to a real traffic accidents data. The characteristics of the study site and data with a detailed description of the developed model's structure are discussed in the following subsections.

4.1 Study Site and Data

Upper Egypt rural roads data is used in this analysis. More specifically, the first hundred kilometers of Aswan-Cairo agricultural rural road are analyzed to model traffic accidents frequency. Accident data used in this study was obtained from the recorded data in the General Authority for Roads, Bridges, and Land Transport (GARBLT) operated by the Egyptian government [40]. The sample size is 108 accidents occurred during three years period. The data records the accident time, location, annual average daily traffic on accident location, weather conditions at the time of accident occurrence, type of accident (single vehicle, front to front, front to back, etc.), type of car and other information. In addition, observations of pavement width, shoulder width, the percentage of trucks on accident location and the number of access roads per kilometer are available. The accident rate (AR) is expressed as accident per million vehicles kilometers (A/mvkm) for a road section as follows:

$$AR = Number\ of\ accidents * 10^6/(AADT * 365 * N * L) \quad (6)$$

Where $AADT$ is the annual average daily traffic, N is the number of years considered and L is the section length in kilometers. The average accident rate among the studied locations during the studied three years is 0.93 accident per million vehicles kilometers. Descriptive statistics of the collected data for all variables are shown in Table 1.

4.2 Modeling

The cost frontier function is used to define the relationship between the accident rate as an output and the pavement width (PW), the shoulder width (SW), the percent of trucks (TR) and the number of access points per kilometer (NA) as inputs. All of these variables were introduced sequentially in order to test the separate effect of variables on accident rates and the significance of different variables on the prediction efficiency. STATA software package [41], is used for estimating the SFM, which is included in the graphical user interface of STATA 13. Maximization of the log-likelihood function is performed by iterating the numerical procedure by switching between the Newton-Raphson (NR) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods up to the convergence of the maximization. Different models are performed. First, the paved width (W) input is represented by one parameter (the sum of PW and SW). Second, the paved width is separated into two inputs PW and SW to consider the difference between the characteristics and the roles of pavement width and shoulder width on accident frequency. Additionally, feeding the model with the noise error (v_i) as a known parameter representing the sum of the variances of PW, SW, TR, and NA is tested. For the inefficiency term, representing the unrecorded factors, the half normal and the exponential distributions are tested. As shown in Table 2, four models for each distribution (the half normal and the exponential distributions) are tested.

TABLE 1
DESCRIPTIVE STATISTICS FOR THE COLLECTED DATA

Variable	Min.	Max.	Mean	St. Dev.
Accident rate (A/mvkm)	0.387	3.048	0.932	0.673
Pavement width (m)	8.25	12.20	10.10	0.844
Shoulder width (m)	0.00	2.75	1.487	0.544
Percent of trucks	15.0%	29.0%	20.6%	5.85%
Number of access points per kilometer	0.00	12.00	1.87	2.67

TABLE 2
SPECIFICATIONS OF THE TESTED SFMS

Model Specifications		Input Noise Error (v_i)	Estimating Noise Error (v_i)
	One independent variable for paved width $X1_i$ =W_i $X2_i$ =TR_i $X3_i$ =NA_i	Model 1 $(\sigma_{vi}^2 = \sigma_W^2 + \sigma_{TR}^2 + \sigma_{NA}^2)$	Model 2
	Two independent variables for pavement width and shoulder width $X1_i$ = PW_i $X2_i$ = SW_i $X3_i$ =TR_i $X4_i$ =NA_i	Model 3 $(\sigma_{vi}^2 = \sigma_{PW}^2 + \sigma_{SW}^2 + \sigma_{TR}^2 + \sigma_{NA}^2)$	Model 4
	One independent variable for paved width $X1_i$ =W_i $X2_i$ =TR_i $X3_i$ =NA_i	Model 1 $(\sigma_{vi}^2 = \sigma_W^2 + \sigma_{TR}^2 + \sigma_{NA}^2)$	Model 2
	Two independent variables for pavement width and shoulder width $X1_i$ = PW_i $X2_i$ = SW_i $X3_i$ =TR_i $X4_i$ =NA_i	Model 3 $(\sigma_{vi}^2 = \sigma_{PW}^2 + \sigma_{SW}^2 + \sigma_{TR}^2 + \sigma_{NA}^2)$	Model 4

5 Results

5.1 Analysis of Estimates

As shown in Table 3, the four developed models (for both normal/exponential and normal/half normal models) are fitted and compared based on the Likelihood Ratio test (LR), Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) methods (the values of the superior model are highlighted in bold). Based on the AIC and BIC values, for the normal/half normal SFMs, Model 2 is better than Models 1, 3 and 4. However, the LR, value suggests that Model 1 is also superior to Models 3 and 4 and having the same LL as Model 2. For normal/exponential SFMs, both AIC and BIC prove that Model 1 is superior to Models 2, 3 and 4. This result (the superiority of Models 1 and 2 over Models 3 and 4) shows that splitting the paved width into two separate inputs has not a strong effect on the model performance. This may be due to the characteristics of the studied road which has a paved shoulder (with an average width of 1.5 m as shown in Table 1) that can be utilized as a paved lane in case of any vehicles defects. In the case of unpaved shoulders or shoulders with a level different from road level, the separation of PW and SW may be more significant.

As shown in Table 4, comparing the two superior models, Model 2 of the normal/half normal SFMs and Model 1 of the normal/exponential SFMs, it is found that Model 1 of the normal/exponential SFMs performs better. This result suggests that assuming the unrecorded factors to follow an exponential distribution is useful for modeling traffic accidents frequency using the SFM. The superior model reached, Model 1 of the normal/exponential SFMs, proved the significance of entering the variability of pavement width, the percent of trucks and the number of access points per kilometer into the frontier model as an expected noise error. Feeding the model with some expected errors are found to be useful for improving model prediction efficiency and decreasing computation time.

The statistical properties and the estimated parameter values for Model 1of the normal/exponential SFMs are shown in Table 5. P-values show that all parameters are statistically significant. It can be noted that the additional mean accident rate due to the unrecorded factors (equals the standard deviation of the inefficiency term exponential distribution shown in the last row of Table 5) is 0.534 A/mvkm. This value highlights the importance of such models that considering this notable effect (given an actual average accident rate of 0.93 A/mvkm as shown in Table 1). This appears reasonable given a variety of unrecorded factors that expected to affect traffic accidents rate. The percent of trucks is proportionally correlated to the traffic accident rates. This matches experts expectations because Aswan city has many important raw materials such as granite and the clay used for the ceramics industry. However, all factories are located in Cairo which develops an increased trucks percent in the studied road section of an average 10 meters width. Whether this factor is indeed significant to be correlated to accident rates with releasing some other factors should be confirmed with further data analysis. Although the coefficients of W and NA are

TABLE 3
Models Comparison

Models		LR chi²(1)	Prob. > chi²	LL	AIC	BIC
	Model 1			-26.25	64.50	75.47
	Model 2			-26.25	**62.50**	**71.64**
	Model 3			-25.95	65.89	78.69
	Model 4			-25.95	63.89	74.87
	Model 1			-17.11	**42.22**	**49.53**
	Model 2			-17.00	43.99	53.13
	Model 3			-16.70	47.39	60.19
	Model 4			-16.70	45.39	56.37

TABLE 4
Comparing Half Normal versus Exponential Models

Model	LR chi²(1)	Prob. > chi²	LL	AIC	BIC
Normal / Half Normal Model 2			-26.251	62.50	71.64
Normal / Exponential Model 1			**-17.110**	**42.22**	**49.53**

TABLE 5
Superior SFM Results

Parameters	Coefficient	p-value (Std. Err.)
Total road width (β_W)	0.0192	0.00 (0.000)
Percent of trucks (β_{TR})	0.8499	0.00 (0.002)
Number of access points (β_{NA})	$5.96*10^{-9}$	0.00 (0.000)
Inefficiency error term (σ_u)	0.534	0.00 (0.079)

small, these factors still statistically significant and improving the model estimation as appears from their P-values shown in Table 5. Noting, however, that this effect is also assured by following the sequential addition of the variables to the SFM as discussed in Section 4.2. The considered variables are all statistically significant, improving the prediction accuracy and providing quick likelihood estimation convergence. The resulted model reached a likelihood maximization convergence after 37 iterations within few seconds.

5.2 Model Validation

The coefficient of determination (R^2) measures the extent of the fluctuation in the variance of the dependent variable based on the values that predicted by a model [42]. The higher this measure (ranging from 0 to 1), the closer the modeled results are to the actual values. The dependent variable (the actual observed accident rate) is predicted from the accident rate values that calculated by the proposed model. Note that the R^2 value does not reflect the extent to which any particular independent variable is correlated with the accident rate. Fig. 1 shows the estimation efficiency of Model 1 of the normal/exponential SFMs by comparing the accident rate values estimated by the model with the observed values. The obtained R^2 of 0.9558 implies that the proposed model explains about 95.6% of the variation in prediction results, which is fairly good. Another validation criterion for the model is the heteroscedasticity test. The heteroscedasticity is the test for homogeneity of variance of the residuals. The heteroscedastic condition is related to a non-constant variance of the residuals. While a well-fitted model is non-heteroscedastic and shows a weak correlation pattern and high scattered plot between the residuals and the fitted values. To demonstrate that the model is non-heteroscedastic, the residuals versus fitted values are plotted utilizing STATA [41]. The heteroscedasticity plot of the model with a reference line y = 0 shows there is no pattern between the residuals and the fitted values, thus the model is not heteroscedastic as shown in Fig. 2.

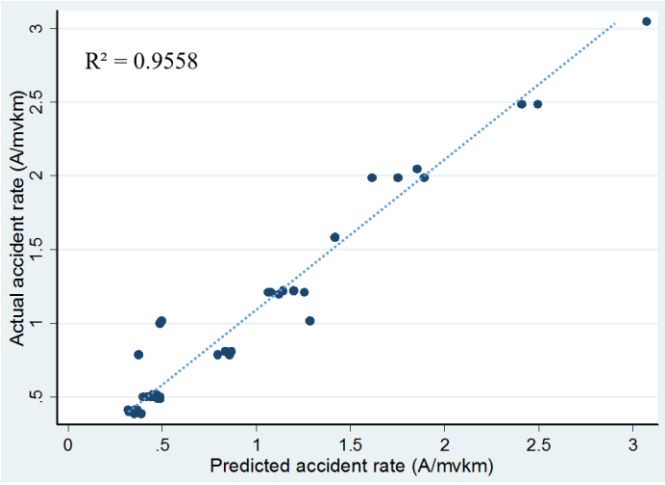

Fig. 1 Actual versus predicted accident rate values

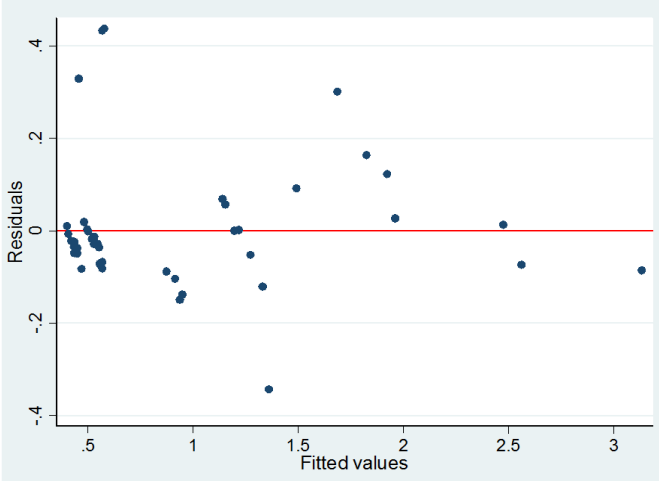

Fig. 2 Heteroscedasticity test

6 Summary and Conclusion

Various approaches have been proposed in the literature to model traffic accidents. However, many factors that expected to affect accident rates are not recorded such as some human and environmental factors. A recent research that extensively reviewed accident analyses, recommended that considering the effect of these unrecorded factors in accident modeling has not been totally addressed in the literature [20]. In another hand, transport authorities operators need simplified models with few possible variables that can be handled for use in real-world practice. In this regard, the main goal of this paper is to model traffic accident rates in a simple statistical framework having a relatively few number of variables and parameter estimates with considering the effects of unrecorded factors. To achieve this, it is suggested that the stochastic frontier approach can be used for modeling traffic accident rates. An advantage of the frontier approach is enabling a prediction of the effect of the observed factors which are linked to the frontier, while the inefficiency of the frontier represents the effect of the unrecorded factors. In the proposed methodology, the reliability of the observed variables is considered as the noise error of the frontier model that modeling their effect on accident rates. In addition, the effect of the unrecorded factors is estimated simultaneously as a separate error term of the frontier model. This method enables differentiating between the characteristics and the effect of those factors.

As a proof of concept, the proposed method is applied to Upper Egypt rural roads using real traffic accidents data during three years period. The model parameters are estimated by the maximum likelihood method. Parameter values are estimated with a reasonable p-value for all the studied variables which indicate the significance of the model with a goodness-of-fit determination coefficient of more than 95%. A notable accident rate related to the unrecorded factors was found which reflects the usefulness of the proposed model that considering such effect. This seems logic according to the possibility of different unrecorded factors that expected to affect traffic accident rate. A notable effect of the percentage of trucks on accident rates

was found which matches the expectation of the experts based on the characteristics of the studied site [40]. The model converges fast, which is an important advantage of this model formulation. This allows certain interesting descriptive statistics that may help practitioners to test the effectiveness of accident mitigation measures in a simple framework. The main disadvantage of the stochastic frontier model is the distributional assumption required to estimate the parameters. Assumed distributions may not fit other data characteristics. Nevertheless, assuming the noise error as a normal distribution and the inefficiency error to follow the exponential distribution is found suitable for the studied data. Whether these distributional assumptions is indeed significant for other datasets should be confirmed with further data analysis. The normality assumption of the observed factors provided a simplified way to estimate the parameters of the unrecorded factors distribution in this paper.

In future work, a larger sample size containing more observed affecting factors is needed. Therefore, a sensitivity analysis of the accident rate records depending on the different factors can be performed. Then, some factors can be excluded from the model and re-estimated to validate the method proposed in this paper. An important further aim is to expand the proposed model to obtain the distributions of both observed and unobserved factors that affecting accident occurrence.

ACKNOWLEDGMENT

Special thanks to Professor Fumitaka Kurauchi for his assistance. Deep thanks also, to the Transport System Design Lab., Gifu University where this work was carried out. General Authority for Roads, Bridges, and Land Transport (GARBLT) provided the data required to achieve this research that funded by the Egyptian government's missions sector.

REFERENCES

[1] World Health Organization, "Global Status Report on Road Safety: Supporting a Decade of Action," World Health Org., Geneva, 2013.

[2] S. C. Joshua and N. J. Garber, "Estimating truck accident rate and involvements using linear and Poisson regression models," Transportation Planning and Technology, vol. 15, no. 1, pp. 41–58, Jun. 1990.

[3] S.-P. Miaou, "The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions," Accident Analysis & Prevention, vol. 26, no. 4, pp. 471–482, Aug. 1994.

[4] Z. Li, W. Wang, P. Liu, J. M. Bigham, and D. R. Ragland, "Using geographically weighted Poisson regression for county-level crash modeling in California," Safety Science, vol. 58, pp. 89–97, Oct. 2013.

[5] V. Shankar, F. Mannering, and W. Barfield, "Effect of roadway geometrics and environmental factors on rural freeway accident frequencies," Accident Analysis & Prevention, vol. 27, no. 3, pp. 371–389, Jun. 1995.

[6] J. Carson and F. Mannering, "The effect of ice warning signs on ice-accident frequencies and severities," Accident Analysis & Prevention, vol. 33, no. 1, pp. 99–109, Jan. 2001.

[7] A. Pirdavani, T. Brijs, T. Bellemans, B. Kochan, and G. Wets, "Evaluating the road safety effects of a fuel cost increase measure by means of zonal crash prediction modeling," Accident Analysis & Prevention, vol. 50, pp. 186–195, Jan. 2013.

[8] F. L. Mannering, "Male/female driver characteristics and accident risk: Some new evidence," Accident Analysis & Prevention, vol. 25, no. 1, pp. 77–84, Feb. 1993.

[9] Y. Chung, "Development of an accident duration prediction model on the Korean freeway systems," Accident Analysis & Prevention, vol. 42, no. 1, pp. 282–289, Jan. 2010.

[10] D. Jovanović, T. Bačkalić, and S. Bašić, "The application of reliability models in traffic accident frequency analysis," Safety Science, vol. 49, no. 8-9, pp. 1246–1251, Oct. 2011.

[11] M. J. Maher, "A bivariate negative binomial model to explain traffic accident migration," Accident Analysis & Prevention, vol. 22, no. 5, pp. 487–498, Oct. 1990.

[12] R.Yu and M. Abdel-Aty, "Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes," Accident Analysis & Prevention, vol. 58, pp. 97–105, Sep. 2013.

[13] S. Narayanamoorthy, R. Paleti, and C. R. Bhat, "On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level," Transportation Research Part B: Methodological, vol. 55, pp. 245–264, Sep. 2013.

[14] D. Lord, A. Manar, and A. Vizioli, "Modeling crash-flow-density and crash-flow-v/c ratio relationships for rural and urban freeway segments," Accident Analysis & Prevention, vol. 37, no. 1, pp. 185–199, Jan. 2005.

[15] X. Wang and M. Abdel-Aty, "Analysis of left-turn crash injury severity by conflicting pattern using partial proportional odds models," Accident Analysis & Prevention, vol. 40, no. 5, pp. 1674–1682, Sep. 2008.

[16] R. Yu and M. Abdel-Aty, "Investigating different approaches to develop informative priors in hierarchical Bayesian safety performance functions," Accident Analysis & Prevention, vol. 56, pp. 51–58, Jul. 2013.

[17] S.-P. Miaou, R. Bligh, and D. Lord, "Part 1: Roadside safety design: Developing guidelines for median barrier installation: Benefit-cost analysis with Texas data," Transportation Research Record: Journal of the Transportation Research Board, vol. 1904, pp. 2–19, Jan. 2005.

[18] J. Aguero-Valverde and P. Jovanis, "Analysis of road crash frequency with spatial models," Transportation Research Record: Journal of the Transportation Research Board, vol. 2061, pp. 55–63, Dec. 2008.

[19] L. Cheng, S. R. Geedipally, and D. Lord, "The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data," Safety Science, vol. 54, pp. 38–42, Apr. 2013.

[20] F. L. Mannering and C. R. Bhat, "Analytic methods in accident research: Methodological frontier and future directions," Analytic Methods in Accident Research, vol. 1, pp. 1–22, Jan. 2014.

[21] Y. Peng and D. Lord, "Application of latent class growth model to longitudinal analysis of traffic crashes," Transportation Research Record: Journal of the Transportation Research Board, vol. 2236, pp. 102–109, Dec. 2011.

[22] Y. Zou, Y. Zhang, and D. Lord, "Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis," Accident Analysis & Prevention, vol. 50, pp. 1042–1051, Jan. 2013.

[23] Y. Zou, Y. Zhang, and D. Lord, "Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models," Analytic Methods in Accident Research, vol. 1, pp. 39–52, Jan. 2014.

[24] M. Castro, R. Paleti, and C. R. Bhat, "A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections," Transportation Research Part B: Methodological, vol. 46, no. 1, pp. 253–272, Jan. 2012.

[25] A.M. Wahaballa, F. Kurauchi, T. Yamamoto and J-D. Schmöcker, " Estimation of platform waiting time distribution considering service reliability based on smart card data and performance reports," Transportation Research Record: Journal of the Transportation Research Board, 2017, (in press).

[26] L. Chang, "Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network", Safety Science, vol. 43, no. 8, pp. 541-557, 2005.

[27] R. Yu and M. Abdel-Aty, "Utilizing support vector machine in real-time crash risk evaluation", Accident Analysis & Prevention, vol. 51, pp. 252-259, 2013.

[28] J. Xiao, B. Kulakowski, and M. El-Gindy, "Prediction of risk of wet-pavement accidents: Fuzzy logic model," Transportation Research Record: Journal of the Transportation Research Board, vol. 1717, pp. 28–36, Jan. 2000.

[29] A.M. Wahaballa, A. Diab, M. Gaber, and A.M. Othman. Sensitivity of Traffic Accidents Mitigation Policies Based on Fuzzy Modeling: A Case Study. In 20th International IEEE Conference on Intelligent Transportation Systems, Yokohama, Japan, 2017 (Accepted).

[30] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," Transportation Research Part A: Policy and Practice, vol. 44, no. 5, pp. 291–305, Jun. 2010.

[31] Y. Xiong and F. Mannering, "The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach", Transportation Research Part B: Methodological, vol. 49, pp. 39-54, 2013.

[32] A. Banerjee, X. Ye and R. Pendyala, "Understanding Travel Time Expenditures Around the World: Exploring the Notion of a Travel Time Frontier", Transportation, vol. 34, no. 1, pp. 51-65, 2006.

[33] S. Ravulaparthy, K. Konduri, and K. Goulias, "Exploring the Role of Activity Time Use Frontiers on Emotional Well Being: An Evidence from Disability and Use of Time Survey," in TRB 96th Annu. Meeting, Washington, DC, Jan. 2017.

[34] S. Canavana, D.J. Graham, P.C. Melo, and R.J. Andersona, "Quantifying the Effects of Delay Incidents on the Performance of Metro Rail Systems using Stochastic Frontier Analysis," in TRB 93rd Annu. Meeting, Washington, DC, Jan. 2014.

[35] D. Aigner, C. A. K. Lovell, and P. Schmidt, "Formulation and estimation of stochastic frontier production function models," Journal of Econometrics, vol. 6, no. 1, pp. 21–37, Jul. 1977.

[36] W. H. Greene, "The Econometric Approach to Efficiency Analysis" in The Measurement of Productive Efficiency and Productivity Growth. H.O. Fried, C.K. Lovell, and S.S. Schmidt, Eds. Oxford University Press, 2008, chapter 2, pp. 92–250.

[37] J. Jondrow, C. A. Knox Lovell, I. S. Materov, and P. Schmidt, "On the estimation of technical inefficiency in the stochastic frontier production function model," Journal of Econometrics, vol. 19, no. 2-3, pp. 233–238, Aug. 1982.

[38] F. Pieri, "Essays on Productivity and Efficiency Analysis," Ph.D. dissertation, Doctoral School of Economics and Management, Univ. of Trento, 2010.

[39] G. E. Battese and T. J. Coelli, "A model for technical inefficiency effects in a stochastic frontier production function for panel data," Empirical Economics, vol. 20, no. 2, pp. 325–332, Jun.1995.

[40] General Authority for Roads, Bridges, and Land Transport, (GARBLT), for the study of "Safety and Protection of Public Transport on the Rural Roads in Egypt", 2003.

[41] StataCorp, "Stata Statistical Software: Release 13" College Station, TX: StataCorp LP, 2013.

[42] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, Statistical and econometric methods for transportation data analysis, Second edition - 2nd edition, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2010.