Research framework for studying public transit bus driver distraction

K. A. D'Souza & S. K. Maheshwari
School of Business, Hampton University, USA

Abstract

Over 3,000 people are killed and 400,000 injured annually in the US due to motor vehicle crashes involving a distracted driver. In the case of passenger vehicles, most of the distraction is within the control of the driver. However, for public transit vehicles, some distractions are caused by factors beyond the driver's control such as operating the fare box or attending to passengers. Research on the distraction of transit bus drivers is very limited, although injuries from transit vehicle accidents are generally higher because buses usually carry many passengers.

This paper proposes a modular research framework for conducting a driver distraction study for transit buses. The research framework provides standardized methodologies structured into four modules – Data Collection, Analysis, Validation and Result Interpretation. The Data Collection module consists of approaches for collecting data from accident databases, surveys, and route observation. The Analysis module provide methods for classification of distracting activities, and development of statistical models that construct relationships between high risk distracting activities and driver attributes and external factors. The Validation module presents simple observation and discussion methods to sophisticated simulation techniques to check the model results. The final module contains guidelines for Results Interpretation and Usage. The framework's standardized techniques are expected to reduce the overall time and cost of conducting a transit bus driver distraction study.
Keywords: transit bus driver distraction, distraction risk index, research framework for bus driver distraction, modelling and predicting driver distraction, model validation, Monte Carlo simulation, route observations.


WIT Transactions on The Built Environment, Vol 130, © 2013 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT130111

1 Introduction

Transit accidents pose significant challenges in metropolitan areas around the world. Analysis of accident databases has found driver distraction to be a significant cause of total motor vehicle crashes [1]. Research on transit bus driver distraction conducted in the US is very limited [2, 3], although risks of distraction are generally higher due to the driver performing added secondary tasks and attending to many passengers. With no established research framework available to study driver distractions, each study is planned and conducted independently at additional time and cost to transit agencies.

This paper proposes a modular research framework for conducting a bus driver distraction study at a transit agency. The objective is to provide an agency with a set of standardized methodologies from data collection to result interpretation and application. The Data Collection module consists of methodology for data extraction from accident database, a survey instrument, and route observations forms. The Analysis module will show how to classify distracting activities, and how to develop statistical models that construct relationships between high risk distracting activities and driver attributes and external factors. The Validation module presents simple route observation and discussion methods as well as simulation techniques to check the model results. The final module contains guidelines for Results Interpretation and Usage.

The modular framework will offer flexibility in choosing one or more tools from the modules while conducting a driver distraction study. The various tools necessary for studying the sources and duration of driver distractions, the risks associated while engaging in potential distracting activities, and visual, manual, and cognitive factors that are believed to be responsible for distraction will be included in the respective modules of the framework. An agency could use these tools to classify the distracting activities into different risk zones. The distracting activities in high risk zones that pose safety concerns could be further analyzed using statistical models to quantify the impact of various factors on driver distraction. Agencies will have the option of validating the results using methods like expert opinions, Monte Carlo simulation, and/or route observations.

The framework can be used for distraction studies that cover a wide range of cost and time intervals such as a low cost, quick study like analysis of existing accident databases maintained by the agencies to relatively higher cost, longer duration study involving field data collection, statistical modelling, analysis, and simulation.

2 Literature review

Driver distraction represents a significant problem in the personal and public transportation sector, and it has been studied by several researchers. A study funded by the American Automobile Association (AAA) Foundation [4] identified the major sources of distraction that cause crashes in personal vehicles, developed a taxonomy of driver distractions in the US, and examined the potential consequences of these distractions on driving performance. The sources

of distractions of bus drivers for a major public transportation company in Australia were investigated using ergonomic methods to develop a taxonomy of the sources of distractions, along with countermeasures to mitigate their effects on drivers' performance [5, 6]. These were one of the foremost studies of distractions that affect transit bus drivers. A taxonomy of the sources of distraction was developed and descriptive statistical analysis followed but the limited sample size of drivers provided insufficient data for inferential statistical analysis. D'Souza and Maheshwari [2, 3] expanded the work of Salmon *et al.* [6] using multivariate statistical models and simulation to draw inferences of driver and external factors on distracting activities.

Various factors such as location, number of hours driven per week, and the driver's age, gender, and experience have impacts on the distraction of transit bus drivers [7]. For example, if the route to be driven is located in a densely populated area, there would be a greater number of passengers and a greater number of external sources of distraction as a result of more frequent stops, traffic congestion, and pedestrians [4]. Studies on the impact of age, gender, driving experience, and driving demands on driving performance suggests that younger (below 20 years) and older (above 60 years) drivers tend to be more vulnerable to the effects of distraction than middle-aged drivers [1]. D'Souza and Maheshwari [2, 3] found that age, gender, weekly driving hours, location, and driving experience have significant relationship with transit bus driver distractions.

Multivariate statistical models are widely used in transportation to study the relationship between the categorical dependent variable and a set of continuous and categorical independent predictor variables [8–10]. A Multinomial Logistic Regression (MLR) model was developed by Morfoulaki *et al.* [11] to identify the factors contributing to service quality and customer satisfaction (*very satisfied, satisfied, somewhat dissatisfied, and very dissatisfied*) with a public transit service in Greece. The impact of age and cognitive functions on driving performance has been studied extensively to predict cognitive distraction with a computational cognitive model and validating the results through simulation [12]. Although a research framework for driver distraction was not reported in the literature, other frameworks related to a driver's mental process [13] and the study of accident causality [14] provided useful inputs for development of the research framework in this paper.

3 Research framework

An outline of the proposed research framework presented in Figure 1 is structured into four modules – Data Collection, Analysis, Validation, and guidelines for Results Interpretation and Usage. Each module consists of relevant tools and steps for studying driver distraction.

3.1 Data collection module

This module contains a set of data collection tools: accident database, driver perception survey, and route observation. The use of accident databases could

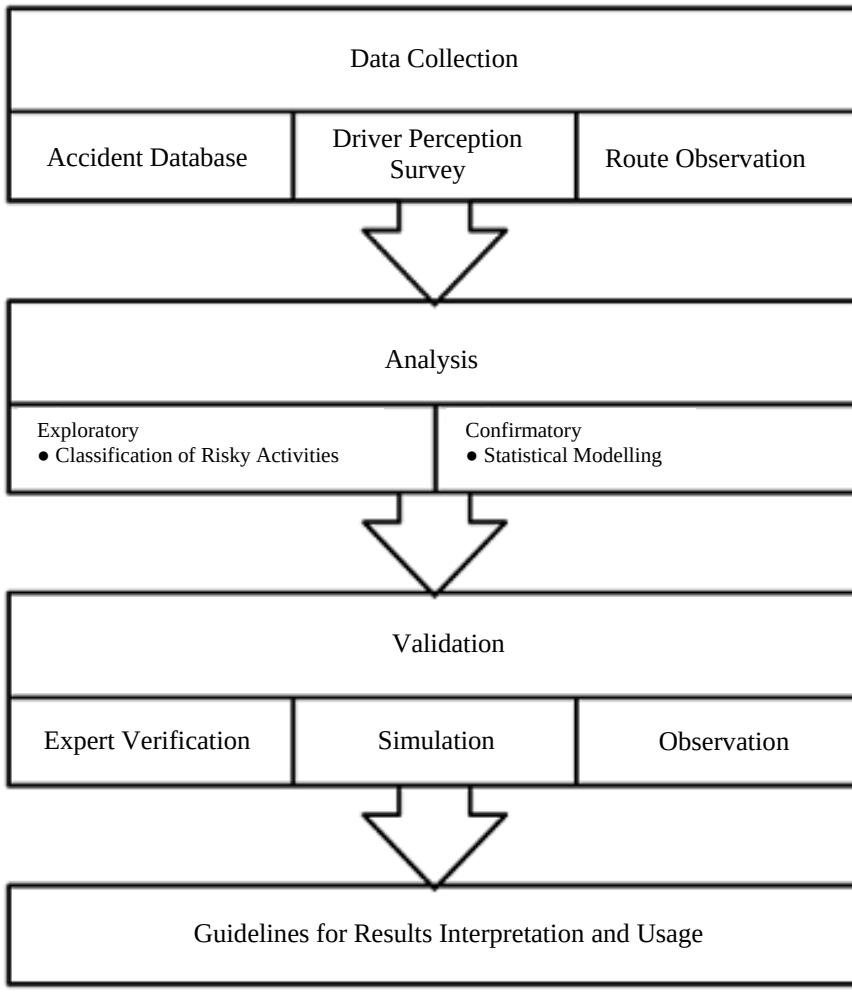

Figure 1: Outline of the proposed research framework.

follow the approach of McEvoy *et al*. [15] who reports that 13.6% of all accidents are caused by driver distraction. The accident database generated from police reports are examined to determine the locations experiencing higher accidents and subsequently establish causes of accidents including distraction related factors. A standard pre-tested survey instrument to study driver perception could be used to determine the factors (external or/and driver attributes) that relate to driver distraction and will be easy to administer and analyze. Data on driver distraction can also be collected via route observations. Analysis of such data can be used in establishing major causes of distraction that would help in developing training and policy guidelines.

3.2 Analysis module

The Analysis module consisting of exploratory and confirmatory steps would assist in classification of distracting activities and development of statistical models for each level of distraction. McEvoy *et al*. [15] reports that 13.6% of all accidents were caused by distracted driving whereas USDOT [1] has estimated 17% of all accidents were caused by distracted driving. Agencies have the option of choosing one of these numbers for their analysis of accidents.

The exploratory steps will develop a system to classify data into risk zones and identify the high risk activities using a standardized distraction risk index. The confirmatory steps will develop an appropriate multivariate statistical model for the high risk distracting activities. The MLR models which have been used in previous studies [2, 3] along with other multivariate techniques could be used by transit agencies to analyze the factors that are related to distractions.

3.2.1 Analysis of accident databases

The accident database analysis could be used in conducting exploratory as well as confirmatory steps to determine the impact of driver distraction. However, the quality and extent of analysis will depend upon type of data collected and available for analysis (not all collected data is publicly available due to legal or other reasons). An analysis of historical accident data for the past two to three years is to be conducted to identify causes of accidents in the city's different locations (for example Northside and Southside). The accidents are to be classified as being either preventable or non-preventable. The non-preventable accidents (the bus maybe hit by another vehicle) are not caused by the bus driver. The preventable accidents (the bus hit another vehicle) could have been avoided if the bus driver had exerted more caution. Some of the preventable accidents are caused by driver distraction but the proportion is unknown.

The relationship between two categorical variables computed by Agresti [16] using a two-way contingency table could be applied by a transit agency to predict the probability of accidents due to driver distraction at two locations in a city listed in Table 1.

Table 1: Contingency table for distracted driving events.

Location of accident	Driver distraction (event B_1)	Other causes (event B_2)	Total
Northside (event A_1)	$n_{11} = 105$	$n_{12} = 663$	$n_{1+} = 768$
Southside (event A_2)	$n_{21} = 227$	$n_{22} = 1442$	$n_{2+} = 1669$
Total	$n_{+1} = 332$	$n_{+2} = 2105$	$n = 2437$

Let X = the explanatory (independent) categorical variable having i levels. i = 2 rows.

Let Y = the response (dependent) categorical variable having j levels. j = 2 columns.

The i, j combinations of outcomes displayed in a tabular form are used to predict probabilites of events. Suppose a driver is selected at random and then classified on the basis of X and Y. The joint probability of X and Y is:

$$p_{ij} = P(X = i, Y = j) \quad (1)$$

$$\text{Where } \Sigma_{i,j}\, p_{i,j} = 1 \quad (2)$$

P_{i+} is the marginal probability representing the row total (i+).
P_{+j} is the marginal probability representing the column total (+j).
n_{ij} = cell count, where total sample size:

$$n = \Sigma_{i,j}\, n_{i,j} \quad (3)$$

$$p_{ij} = (n_{ij}/n) \quad (4)$$

P(Accident in Northside) = $(n_{1+})/(n)$ = 768/2437 = 0.32.
P(Accident in Southside) = $(n_{2+})/(n)$ = 1669/2437 = 0.67.

Using the general rule of multiplication, the probability that a driver from the Northside (Event A_1) will have an accident due to distraction (Event B_1) is:

$$P(A_1 \text{ and } B_1) = P(A_1)\, P(B_1 \mid A_1) \quad (5)$$
$$= (768/2437)(105/768) = 0.043.$$

Using the general rule of multiplication, the probability that a driver from the Southside (Event A_2) will have an accident due to distraction (Event B_1) is:

$$P(A_2 \text{ and } B_1) = P(A_2)\, P(B_1 \mid A_2) \quad (6)$$
$$= (1669/2437)(227/1669) = 0.093.$$

It is clear from the Table 1 data, that the overall probability of the accidents as well as the joint probability of accidents with distractions is higher in the Southside compared to Northside.

In addition to location, the number of accidents is dependent to the days of the week with Fridays having the highest number of accidents in the Southside compared to Northside [17]. The time of the day for the highest number of accidents is between 12:00 PM to 6:00 PM (preventable and non-preventable) [17]. The drivers with the least experience (0 to 5 years) have the highest number of accidents (preventable and non-preventable) and correspondingly a higher number of accidents caused by distracted driving [17].

3.2.2 Analysis of survey data and route observations

In the exploratory analysis, the drivers' response of the various manual, visual and cognitive distracting activities are to be classified to produce the Distraction Risk Index (DRI) that measures the potential risk associated with each risk zone

activity [2]. The DRI considers the rating, duration, and perceptions of each distracting activity in order to classify activities into Risk Zones I and II, III and IV. In the confirmatory analysis, the MLR is suitable to model the high risk distracting activities in Risk Zone I and II using levels of distraction as the dependent variable and correlating it with the factors as independent/predictor variables. For example, categorical dependent variable (driver distraction) had more than two levels: Not Distracted, Slightly Distracted, Distracted, and Very Distracted. The independent variables included categorical variables: gender and location, and continuous variables: age, driving experience, and driving hours per week [2, 3].

The general MLR model proposed by Moutinho and Hutcheson [18] is expressed as

$$log\left(\frac{\Pr(Y=j)}{\Pr(Y=j')}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots . + \beta_k X_k \quad (7)$$

where j is the identified distraction level, and j' is the reference distraction level.

The intercept β_0 is the value of Y when all the independent variables are equal to zero. $\beta_1, \beta_2, \beta_3, \ldots \beta_k$ are the regression coefficients of $X_1, X_2, X_3 \ldots .. X_k$. Each of the regression coefficients describes the size of the contribution of risk factor X_i relative to a reference category. A positive regression coefficient means that the explanatory variable increases the probability of the outcome, while a negative regression coefficient means that the variable decreases the probability of that outcome [9]. Similarly, a large regression coefficient means that the risk factor strongly influences the probability of that outcome, while a near-zero regression coefficient means that that risk factor has little influence on the probability of that outcome [9].

An illustration of the multinomial dependent variable Y_i (*logit*) which measures the total contribution of the five factors (independent variables) is expressed as [2, 3]:

$$Y_i = \beta_0 + \beta_1 * LOCAT + \beta_2 * SEX + \beta_3 * AGE + \beta_4 * EXP + \beta_5 * DRIVING/WK \quad (8)$$

where,

LOCAT: Location of driver, a categorical variable, 1 = Northside, 2 = Southside.

SEX: Gender of driver, a categorical variable, 1 = Male, 2 = Female.

AGE: Reported age of driver in years, a continuous variable.

EXP: Number of years of experience driving a bus, a continuous variable.

DRIVING/WK: Weekly driving hours, a continuous variable.

Statistical packages like SPSS [19] are recommended for solving the MLR model. Each of the top three levels is referenced with the Not Distracted level. In view of space limitations, an illustration of the statistical test ratios and parameter estimates is presented in Table 2 for one Risk Zone I distracting activity (Passengers).

A standardized format to collect route data will help rapid determination of some distraction factors. The frequency distribution will be the primary exploratory step for examining the route observation data to identify major causes of the distraction.

Table 2: Illustration of MLR model outputs for passengers [2].

Model Chi-Square (χ^2) = 36.61 (18)*** Pearson Stat (NS) Deviance Stat(NS)	R^2 = 0.590 (Cox and Snell); 0.649 (Nagelkerke); 0.317(McFadden)	AIC initial/final values: 114.22/104.16 BIC initial/final values: 145.06/140.14		
Independent Variables and Interactions	Coeff β (SE)	Wald Statistic	Odds Ratio Exp (B)	95% CI
Slightly distracted vs. Not distracted				
Intercept	N/S	-		
LOCAT = 1	-2.20 (1.04)**	4.44	0.11	[0.14 – 0.86]
LOCAT = 2	0.00			
SEX =1	16.05 (6.04)**	7.07	9340926	[67.82 – 1.29E12]
SEX = 2	0.00			
AGE	N/S	-	N/A	
EXP	N/S	-	N/A	
DRIVING/WK	0.13 (0.07)*	3.64	1.14	[1.00 – 1.30]
AGE*DRIVING/WK	N/S	-	N/A	
SEX=1*DRIVING/WK	-0.34 (0.13)****	6.87	0.71	[0.55 – 0.92]
AGE*EXP	N/S	-	N/A	
Distracted vs. Not distracted				
Intercept	-224.35 (6.95)****	1042.79		
LOCAT = 1	N/S	-	N/A	
LOCAT = 2	0.00			
SEX =1	235.99 (1.53)****	23736	3.08E102	[1.53E103 – 6.20E103]
SEX = 2	0.00			
AGE	N/S	-	N/A	
EXP	0.20 (0.10)**	3.79	1.22	[1.0 – 1.48]
DRIVING/WK	4.53 (0.10)****	1947	93.15	[76.16 – 113.94]
AGE*DRIVING/WK	N/S	-	N/A	
SEX=1*DRIVING/WK	N/S	-	N/A	
AGE*EXP	N/S	-	N/A	
Very distracted vs. Not distracted				
DRIVING/WK	0.47 (0.21)**	5.00	1.6	[1.06 – 2.41]

*p < 0.10; **p < 0.05; ***p < 0.01; ****p < 0.001. N/S = Not Significant.

3.3 Validation module

The validation module will verify the statistical model results. Expert verification by safety managers in the participating agencies is the starting point for validation. Standardized route observation forms will also be developed for validation purposes. For example, the MLR model for "Passengers" is validated using the statistical outputs from SPSS [19] that are summarized in Table 2. The likelihood ratio test using model fitting information shows that the difference in the Log Likelihood between the intercept only (without any independent variables) and the final model (with all the independent variables) computes the chi-square (χ^2) = 36.61 signifying a good improvement in the model fit. It follows that the independent variables contribute significantly to the outcome of the distraction level. The values of the AIC initial/final values (114..22/104.16); the BIC initial/final values (145.06/140.14) gets smaller during the stepwise process indicating a good fit for the final model.

The model's Goodness of Fit as indicated by multiple statistics such as: the p-values for Pearson and Deviance (both test the same results) chi-square (χ^2) = 1.00 (p = 1) proving no significance. Hence, it can be inferred that the predicted values of the model are not significantly different from the observed values at all outcome levels i.e. the model fits the data well. The measures of Pseudo R^2 (0.59, 0.65, and 0.32) are reasonably similar and high values of R^2 indicating a good fit. The Table 2 further presents outputs from the three binary logistic regression models along with the coefficients, Wald Statistic, and Odds Ratio and 95% CI values which are truncated to < or > 1 and includes or excludes 1 from the 95% CI.

The MLR models could also be simulated using probabilistic distributions to generate driver distraction events that would occur in practice over a range of random factors. Simulation generates average probability values for 1,000 drivers getting Slightly Distracted, Distracted, and Very Distracted. The results for the external factor "Location" is illustrated in Figure 2.

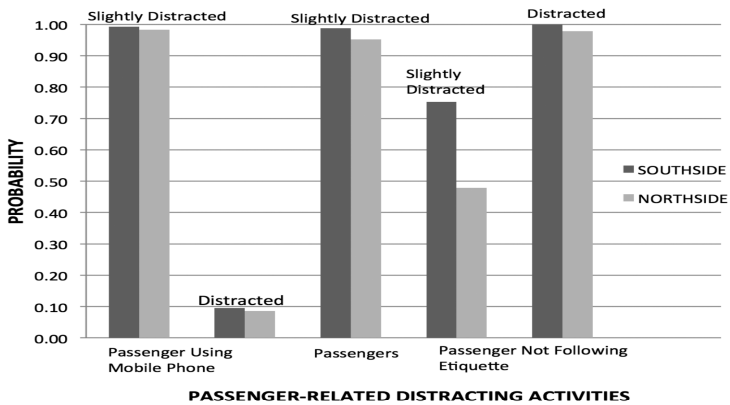

Figure 2: Simulation results for Northside and Southside locations [2].

3.4 Results interpretation and usage module

In the last module, guidelines will be developed for the interpretation of results and application of those results in predicting driver distraction, developing policies, determining training needs, designing driver's cabin, adopting technology, etc.

The results of the MLR models, simulation, and route observations would identify driver attributes and external factors that have a significant impact on high risk distracting activities. It would therefore present a challenge for the agencies to develop effective policies for handling driver behaviour, so that they are less likely to undertake distracted behaviour. Training should focus on drivers who are more likely to get distracted by specific distracting activities. Educational training program on the proper use of technological devices mounted in the cab or issued to the driver, and hazards associated with utilizing these devices while driving should focus on the drivers who are likely to get distracted with technological devices. The control panel and other devices used by the driver must be user-friendly, and not require long glances away from the forward roadway.

4 Conclusions

This paper has attempted to combine independent procedures for studying driver distraction into a comprehensive framework. It is one of only a few studies to consolidate methodologies for data collection, analysis, validation, and interpretation of results into a workable framework. How could a transit agency use the framework proposed in this study? Any transit agencies planning to conduct a driver distraction study could choose relevant tools from the modules according to the time available and budgetary limits such as a quick, low cost study like analysis of existing accident databases maintained by the agencies to a relatively longer duration, higher cost study involving field data collection, statistical modelling, analysis, and simulation.

As additional studies are being conducted in other agencies, the framework can be updated accordingly. The expanded data set can be used for validation as well as further refinement of the proposed framework. The modular structure of this framework permits updating and adding tools in each module as and when required without affecting the other modules. The four modules outlined in this framework is only a start and is expected to grow as more studies are conducted at transit agencies.

References

[1] U.S. Department of Transportation. Traffic safety facts: distracted driving 2009. *DOT HS 811 379*. National Highway Traffic Safety Administration, Washington, DC 20590. September 2010.

[2] D'Souza, K. A. and Maheshwari, S. K. Multivariate Statistical Analysis of Public Transit Bus Driver Distraction. *Journal of Public Transportation: Special Edition: Rural and Intercity Bus*, **15(3)**, pp. 1–23, 2012.
[3] D'Souza, K. A., and Maheshwari, S. K. Improving performance of public transit buses by minimizing driver distraction. *Proc. of the Urban Transport 2012 Conf.,* A Coruña, Spain, eds. J.W.S. Longhurst and C.A. Brebbia. Wessex Institute of Technology Press, Southampton, U. K, pp. 281–293, 2012.
[4] AAA Foundation for Traffic Safety. Distraction in everyday driving. University of North Carolina at Chapel Hill, Highway Safety Research Centre, June, 2003. www.aaafoundation.org.
[5] Salmon, P. M., Young K. L. and Regan, M. A. Distraction 'on the buses': A novel framework of ergonomics methods for identifying sources and effects of bus driver distraction. *Applied Ergonomics*, **42**, pp. 602–610, 2011.
[6] Salmon, P.M., Young, K.L. and Regan, M. A. *Bus driver distraction stage 1:* A*nalysis of risk for State Transit Authority New South Wales bus drivers. Final report*. Monash University Accident Research Centre, Victoria, Australia, 2006.
[7] Young, K. L., Regan, M. A., and Lee, J. D. Factors Moderating the Impact of Distraction on Driving Performance and Safety. *Driver Distraction: Theory, Effects, and Mitigation*. CRC Press, Taylor and Francis Group, pp. 335–351, 2009.
[8] Yan, X., Radwan, E. and Abdel-Aty, M. Characteristics of rear-end accidents at signalized intersections using multiple logistic regression model. *Accident Analysis and Prevention*, **37**, pp. 983–995, 2005.
[9] Washington, S.P., Karlaftis, M.G. and Mannering, F.L. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC, A CRC Press Company, Boca Raton, FL, pp. 303–359, 2011.
[10] Yan, X., Radwan, E. and Mannila, K.K. Analysis of truck-involved rear-end crashes using multinomial logistic regression. *Advances in Transportation Studies: an International Journal*, Section A **17**, pp. 39–52, 2009.
[11] Morfoulaki, M., Tyrinopoulos, Y. and Aifadopoulou, G. Estimation of satisfied customers in public transport systems: a new methodological approach. *Journal of the Transportation Research Forum*, **46 (1),** pp. 63–72, 2007.
[12] Salvucci, D. D., Chavez, A. K., and Lee, F. J. Modelling effects of age in complex tasks: a case study in driving. *26th An. Conf. of the Cognitive Science Society*, 2004.
[13] Wong, Jinn-Tsai and Shah-Hsuan Huang. Modeling driver mental workload for accident causation and prevention. *Jr. of the Eastern Asia Society for Transportation*, **8**, 2009.
[14] Trick, L. M., Enns, J. T., Mills, J. and Vavrik, J. Paying attention behind the wheel: a framework for studying the role of attention in driving. *Theoretical Issues in Ergonomics Science*. **5 (5)**, pp 385–424

[15] McEvoy, S. P., Stevenson, M. R. and Woodward, M. The prevalence of, and factors associated with serious crashes involving a distracted activity. *Accident Analysis and Prevention*. **39**, pp. 475–482, 2007.
[16] Agresti, A. *An Introduction to Categorical Data Analysis.* John Wiley and Sons, Inc., New York, NY 10158, pp. 16–22, 1996.
[17] D'Souza, K, Maheshwari, S. and Banaszak, Z. Research framework for studying driver distraction on Polish city highways. *Workshop on Multimodal Networks Modelling and Design*. Warsaw University of Technology, Warsaw, Poland. June 5, 2012.
[18] Moutinho, L. and Hutcheson, G. *Multinomial Logistic Regression*. The SAGE Dictionary of Quantitative Management Research. SAGE Publications Ltd., pp. 208–212, 2011.
[19] SPSS 17.0. SPSS Inc., Chicago, IL 60606-6412, 2008.