Quantifying the impact of driver, vehicle and environment on crash risk using big data

Eva Michelaraki[1]*, Thodoris Garefalakis[1], George Yannis[1]

[1]National Technical University of Athens,
Department of Transportation Planning and Engineering,
5 Heroon Polytechniou str., GR-15773, Athens, Greece
*Corresponding author: evamich@mail.ntrua.gr

Abstract

Human behaviour plays a pivotal role in road safety. Factors such as speeding, distraction or aggressive driving can elevate crash risk. Moreover, the complexity of driving task, such as weather conditions, traffic density or road infrastructure and vehicle conditions have also a significant impact on risk. The aim of this paper was to quantify the impact of driver, vehicle and environment on crash risk using big data. Towards that end, a naturalistic driving experiment was taken place and data from 135 drivers aged 20-65 were collected and analysed. Towards that end, Generalized Linear Models (GLMs) were developed and explanatory variables of risk with the most reliable indicators, such as time headway, distance travelled, speed, time of the day or weather conditions were assessed. Additionally, Structural Equation Models (SEMs) were used to explore how the model variables were inter-related, allowing for both direct and indirect relationships to be modelled. The analyses revealed that drivers, when faced with difficult conditions, tend to regulate well their capacity to apprehend potential difficulties, while driving. It was also found that complex environment conditions led to an increased crash risk due to several reasons. The relationship among environment conditions, driver behaviour and vehicle situation with risk, may depend on the specific context and type of task or activity involved. Authorities may use data systems at population level to plan mobility and safety interventions, set up road user incentives, optimize enforcement and enhance community building on safe travelling.

Keywords driving behaviour; road safety; naturalistic driving experiment; Generalized Linear Models; Structural Equation Models.

1. Introduction

Road traffic crashes result in the deaths of approximately 1.19 million people around the world each year and leave between 20 and 50 million people with non-fatal injuries (World Health Organization, 2023). More than half of all road traffic deaths occur among vulnerable road users, such as pedestrians, cyclists and motorcyclists. Road traffic injuries are the leading cause of death for children and young adults aged 5-29. In addition to the human suffering caused by road traffic injuries, they also incur a heavy economic burden on victims and their families, both through treatment costs for the injured and through loss of productivity of those killed or disabled.

Several factors have a significant impact on road safety. These factors can contribute to the occurrence of road crashes and influence the severity of injuries sustained. For instance, human behaviour plays a critical role in road safety, accounting for 65-95% of road crashes (Salmon et al., 2011). Factors such as speeding, distracted or aggressive driving, and non-compliance with traffic regulations can increase the crash risk (Yannis & Michelaraki, 2024). In addition, socioeconomic factors, such as income level, education, and access to transportation resources, can indirectly influence road safety.

At the same time, the condition and safety features of vehicles also play a critical role in averting crashes and reducing the likelihood of serious. Indicators such as vehicle maintenance, tire condition, brake functionality, and the presence of safety technologies can significantly affect crash outcomes. Similarly, environmental conditions can affect road safety. Factors, such as adverse weather, poor visibility, and

uneven road surfaces can increase the likelihood of crashes. Moreover, the design, condition, and maintenance of roads and infrastructure can impact road safety. Inadequate signage, absence of pedestrian crossings, lack of proper lighting, and insufficient maintenance can contribute to injuries.

The paper is structured as follows. In the beginning, the motivation and the objectives of this study is described. This is followed by the description of the research methodology, encompassing the theoretical foundations of the models utilized. Then, a detailed overview of data collection is presented. Finally, the results of the analysis are presented followed by relevant discussion on key findings.

2. Objectives

The aim of this paper is to quantify the impact of driver, vehicle and environment on crash risk using big data. Towards that end, a naturalistic driving experiment was conducted and a large database consisting of 135 drivers aged 20-65 was collected and analysed. In order to fulfil these objectives, Generalized Linear Models (GLMs) were developed. Explanatory variables of risk and the most reliable indicators, such as time headway, distance travelled, speed, forward collision, time of the day (lighting indicators) or weather conditions were assessed. Structural Equation Models (SEMs) were used to explore how the model variables were inter-related, allowing for both direct and indirect relationships to be modelled.

3. Methodology

3.1 Generalized Linear Models

Generalized Linear Model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value (Hastie & Pregibon, 2017). In a GLM, each outcome Y of the dependent variables is assumed to be generated from a particular distribution in an exponential family, a large class of probability distributions that includes the normal, binomial, Poisson and gamma distributions, among others. The mean, μ, of the distribution depends on the independent variables, X, through:

$$E(Y|\mathrm{X}) = \mu = g^{-1}(X\beta) \quad (1)$$

where: E(Y|X) is the expected value of Y conditional on X; Xβ is the linear predictor, a linear combination of unknown parameters β; g is the link function.

In this framework, the variance is typically a function, V, of the mean:

$$Var(Y|\mathrm{X}) = V(g^{-1}(X\beta)) \quad (2)$$

It is convenient if V follows from an exponential family of distributions, but it may simply be that the variance is a function of the predicted value. The unknown parameters, β, are typically estimated with maximum likelihood, maximum quasi-likelihood, or Bayesian techniques.

GLMs were formulated as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. In particular, Hastie & Tibshirani (1990) proposed an iteratively reweighted least squares method for maximum likelihood estimation of the model parameters. Maximum-likelihood estimation remains popular and is the default method on many statistical computing packages. Other approaches, including Bayesian approaches and least squares fits to variance stabilized responses, have been developed. A key point in the development of GLM was the generalization of the normal distribution (on which the linear regression model relies) to the exponential family of distributions (Collins et al., 2001). Consider a single random variable y whose probability function (if it is discrete) or probability density function (if it is continuous) depends on a single parameter θ. The distribution belongs to the exponential family if it can be written as follows:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \quad (3)$$

where: a, b, s, and t are known functions. The symmetry between y and θ becomes more evident if the equation above is rewritten as follows:

$$f(y;\theta) = \exp\left[\alpha(y)b(\theta) + c(\theta) + d(y)\right] \quad (4)$$

where: s(y)=exp[d(y)] and t(θ)=exp[c(θ)]

3.2 Structural Equation Models

Structural Equation Model (SEM) represent a natural extension of a measurement model, and a mature statistical modelling framework. SEM is widely used for modelling complex and multi-layered relationships between observed and unobserved variables, such as task complexity or coping capacity. Observed variables are measurable, whereas unobserved variables are latent constructs – analogous to factors/components in a factor/principal component analysis. SEMs have two components: a measurement model and a structural model. The measurement model is used to determine how well various observable exogenous variables can measure the latent variables, as well as the related measurement errors. The structural model is used to explore how the model variables are inter-related, allowing for both direct and indirect relationships. In this sense, SEMs differ from ordinary regression techniques in which relationships among variables are direct.

The general formulation of SEM is as follows (Washington et al., 2011; 2020):

$$\eta = \beta\eta + \gamma\xi + \varepsilon \quad (5)$$

where: η is a vector of endogenous variables, ξ is a vector of exogenous variables, β and γ are vectors of coefficients to be estimated, and ε is a vector of regression errors.

The measurement models are then as follows (Chen, 2007):

$$x = \Lambda_x\xi + \delta, \text{ for the exogenous variables} \quad (6)$$

$$y = \Lambda_y\eta + \zeta, \text{ for the endogenous variables} \quad (7)$$

where: x and δ are vectors related to the observed exogenous variables and their errors, y and ζ are vectors related to the observed endogenous variables and their errors, and Λ_x, Λ_y are structural coefficient matrices for the effects of the latent exogenous and endogenous variables on the observed variables. The structural model is often represented by a path analysis, showing how a set of 'explanatory' variables can influence a 'dependent' variable. The paths can be drawn so as to reflect if the explanatory variables are correlated, mediated or independent causes to the dependent variable.

3.3 Evaluation metrics

In the context of model selection, model Goodness-of-Fit measures consist an important part of any statistical model assessment. The Akaike Information Criterion (AIC), which accounts for the number of included independent variables, is used for the process of model selection between models with different combination of explanatory variables. Similarly, the Bayesian Information Criterion (BIC) is used for selecting among a finite set of models. Lower BIC values are generally preferred, suggesting a better model fit. The Comparative Fit Index (CFI) evaluates the model fit by comparing a hypothesized model with an independence model. A CFI value greater than 0.90 indicates very good overall model fit. Moreover, the Tucker Lewis Index (TLI) considers the parsimony of the model. Values above 0.90 are generally accepted as indications of a very good fit. The Root Mean Square Error Approximation (RMSEA) measures the unstandardized discrepancy between the population and the fitted model, adjusted by degrees of freedom. It provides an indication of how well the model fits the population. Lastly, the Goodness of Fit Index (GFI) measures the fit between the hypothesized model and the observed covariance matrix. Values above 0.90 are typically considered to indicate a very good fit.

4. Data overview

In order to achieve the objectives of this study, a naturalistic driving experiment was carried out involving 135 car drivers (with total duration of 4 months) and a large database of 31,954 trips was collected and analysed in order to investigate the most prominent driving behaviour indicators. The experimental design of the on-road study is displayed in Figure 1 and has been subdivided into four consecutive phases. Firstly, phase 1 of the field trials refers to a reference period after the installation of the system inside the vehicle in order to monitor driving behaviour without interventions. Secondly, phase 2 of the field trials refers to a monitoring period during which only in-vehicle real-time warnings were provided using Advanced Driver Assistance Systems (ADAS). Thirdly, in phase 3 of the field trials, feedback via the smartphone app is combined with in-vehicle warnings. Lastly, in phase 4 of the field trials, gamification features are added to the app, with additional support of a web-dashboard.

Figure 1: Overview of the different phases of the experimental design

Explanatory variables of risk and the most reliable indicators of task complexity (e.g. time of the day, weather) and coping capacity, such as average speed, headway, harsh events, distance travelled, duration, forward collision warnings or pedestrian collision warnings were assessed. Table 1 demonstrates the most relevant variables utilized for defining task complexity and coping capacity. These variables are instrumental to this study, essential for capturing complex dynamics of the inter-relationship between the task complexity, operator and vehicle coping capacity, and crash risk.

Table 1: Variables of task complexity, coping capacity and risk

Task complexity	Coping capacity – vehicle state	Coping capacity – operator state		Risk
Car wipers	Vehicle age	Distance	Inter Beat Interval	Speeding levels
Car high beam	First vehicle registration	Duration	Headway	Headway levels
Time indicator	Fuel type	Average speed	Overtaking	Overtaking levels
Distance	Engine Cubic Centimeters	Harsh acceleration/ braking	Fatigue	Fatigue levels
Duration	Engine Horsepower	Forward collision warning (FCW)	Hands on wheel	Harsh acceleration levels
Month	Gearbox	Pedestrian collision warning (PCW)	Gender	Harsh braking levels
Day of the week	Vehicle brand	Lane departure warning (LDW)	Age	Vehicle control events

5. Results

5.1 Generalized Linear Models

A high number of regression model tests were conducted for different combinations of variables. An attempt was made to use the same independent variables in the model applied. For each configuration,

various alternatives were tested through the respective log-likelihood test comparisons. The optimal combination of variables was the one that had a sufficient number of statistically significant independent variables at a 95% confidence level (p-values ≤ 0.05). Moreover, the independent variables were also checked for multicollinearity through the Variance Inflation Factor (VIF). A standard guideline is that VIF values higher than 10 indicate high multicollinearity (Kutner et al., 2004). However, a threshold equal to 5 is also commonly used (Sheather, 2009). Subsequently, the final models were selected as the ones with the independent variable configuration with the lowest AIC and BIC values for the developed model.

One of the major contributors to road crashes is headway, i.e. the distance between two vehicles; when it is too short to allow the following driver to react appropriately to harsh braking by the leading vehicle. The headway between two vehicles can be expressed in terms of time and space. Within this framework, the second GLM investigated the relationship between the headway and several explanatory variables of task complexity and coping capacity (operator state). More specifically, the dependent variable of the developed model is the dummy variable "headway", which is coded with 1 if there is a headway event and with 0 if not. For task complexity, the variables used are time indicator and weather. Concerning coping capacity - vehicle state, the variables used are fuel type, vehicle age and gearbox, while for coping capacity - operator state, the variables used are duration, harsh acceleration, harsh braking, average speed, gender and age. The model parameter estimates are summarized in Table 2.

Findings derived from Table 2 demonstrated that all the explanatory variables were statistically significant at a 95% confidence level. In addition, there was no issue of multicollinearity as the VIF values are much lower than 10. With respect to the coefficients, it was found that time of the day (indicator of task complexity) was negatively correlated with headway, which means that drivers tend to keep safer distances from the vehicle in front of them during the night. This may probably be due to the fact that there is no heavy traffic during night hours; thus, headway events are avoided. The wipers variable was found to have a positive correlation with headway, indicating that there are more headway events during adverse weather conditions, such as rain. This suggests that drivers tend to be more cautious and maintain greater following distances when the windshield wipers are not in use, reflecting the increased need for safety during poor weather conditions.

Table 2: Parameter estimates and multicollinearity diagnostics of the GLM for headway

Variables	Estimate	Std. Error	z-value	Pr(|z|)	VIF
(Intercept)	-0.340	0.002	-151.275	< .001	-
Time indicator	-4.633×10-4	1.467×10-4	-3.158	0.002	1.001
Weather	0.060	0.007	9.026	< .001	1.006
Fuel type - Diesel	-3.430×10-5	1.897×10-6	-18.084	< .001	4.889
Vehicle age	3.318×10-5	1.640×10-6	20.236	< .001	5.995
Gearbox - Automatic	-7.127×10-6	2.303×10-6	-3.095	0.002	3.289
Duration	9.232×10-7	2.569×10-7	3.593	< .001	1.058
Harsh braking	5.703×10-5	1.753×10-6	32.533	< .001	3.397
Harsh acceleration	4.587×10-5	1.819×10-6	25.216	< .001	3.405
Average speed	2.018×10-5	7.686×10-7	26.254	< .001	1.111
Gender - Female	-1.595×10-5	1.818×10-6	-8.775	< .001	1.495
Age	3.891×10-5	1.913×10-6	20.336	< .001	5.342
Summary statistics					
AIC	1.394×10+6				
BIC	1.165×10+6				
Degrees of freedom	822163				

Furthermore, vehicle age appeared to have a positive relationship with the dependent variable, (i.e. headway), indicating that as the vehicle age increases, the likelihood of headway events also increases. This suggests that older vehicles are more frequently involved in headway events, which could be due to various factors, such as the cautious driving habits of owners of older vehicles or the reduced

performance and response times of older vehicles necessitating greater following distances. Interestingly, fuel type and gearbox were negatively correlated with headway. In particular, the negative value of the "fuel type" coefficient implied that when the fuel type was diesel (coded as 1, with hybrid electric coded as 2, and petrol coded as 3), the headway percentage became lower. This suggests that vehicles running on diesel are associated with a lower frequency of headway events compared to those running on hybrid electric or petrol. Similarly, the negative value of the "gearbox" coefficient demonstrated that vehicles with an automatic gearbox experienced fewer headway events. This indicates that vehicles with automatic transmissions are less likely to encounter headway events compared to those with manual transmissions, possibly due to the smoother and more consistent driving patterns facilitated by automatic gearboxes.

Moreover, it was revealed that indicators of coping capacity – operator state, such as duration, harsh acceleration, harsh braking and average speed had a positive impact on headway. This means that longer trip duration, instances of harsh acceleration and braking, and higher average speeds are associated with an increased likelihood of headway events. These factors suggest that longer driving times contribute to more frequent occurrences of maintaining following distances.

Taking into account socio-demographic characteristics, gender was negatively correlated with headway. In particular, the negative value of the “gender” coefficient implied that as the value of the variable was equal to 1 (males coded as 0, females as 1), the headway percentage was lower. This suggests that female drivers perform fewer headway events and tend to be more cautious in maintaining following distances compared to male drivers. On the other hand, age was positively correlated with headway, indicating that as the driver's age increases, the likelihood of headway events also increases. This suggests that older drivers tend to have more headway events, which could be due to various factors, such as slower reaction times or less aggressive driving, leading to a greater need to maintain safe following distances.

5.2 Structural Equation Models

Following the exploratory analysis, the variables associated to the latent variable “task complexity” and “coping capacity” were estimated from the various indicators. This way, the effect of different personal factors on risk was defined and further analysed. Several SEM were applied in order to identify the effect of task complexity and coping capacity on the STZ level, controlling for the above exogenous factors. Risk was measured by means of the STZ levels for headway (level 1 refers to ‘normal driving’ used as the reference case; level 2 refers to ‘dangerous driving’ while level 3 refers to ‘avoidable accident driving’). In particular, positive correlations of risk with the STZ indicators were found.

The latent variable of task complexity is measured by means of the environmental indicator of time of the day and weather. Exposure indicators, such as trip duration was also included in the task complexity analysis. It was revealed that time of the day, weather and duration had a positive correlation with task complexity. In addition, the latent coping capacity is measured by means of both vehicle and operator state indicators. Vehicle state includes variables such as vehicle age and fuel type, while operator state includes indicators, such as average speed, gender and age. Results indicated that vehicle age, fuel type, gender and driver’s age were positively correlated with coping capacity. These factors imply that certain vehicle characteristics and driver demographics contribute to enhanced coping mechanisms in various driving conditions. Interestingly, average speed appeared to have a negative impact on coping capacity. This suggests that as the average speed increases, the ability of drivers to manage and respond to driving demands and challenges effectively decreases. Higher speeds likely reduce the time available for decision-making and increase the complexity of driving tasks, diminishing coping capacity.

The structural model between the latent variables shows some interesting findings: first, task complexity and coping capacity are inter-related with a positive correlation (regression coefficient=0.10). This positive correlation indicates that higher task complexity is associated with higher coping capacity implying that drivers coping capacity increases as the complexity of driving task increases. Overall, the

structural model between task complexity and risk shows a positive coefficient, which means that increased task complexity relates to increased risk according to the model (regression coefficient=13.19). On the other hand, the structural model between coping capacity and risk shows a negative coefficient, which means that increased coping capacity relates to decreased risk according to the model (regression coefficient=-0.05). The respective path diagram of the SEM for headway is presented in Figure 2.

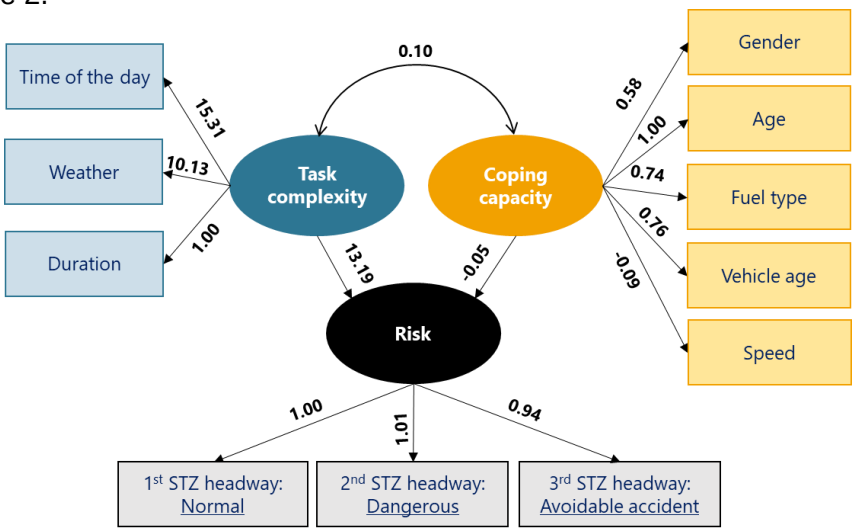

Figure 2: SEM results of task complexity and coping capacity on risk (STZ headway)

In order to gain a clear depiction per each phase, four separate SEM models were estimated in order to explore the relationship between the latent variables of task complexity, coping capacity and risk (expressed as the three phases of the STZ) of headway. Each model corresponds with one of the different experiment phases:

- Phase 1: monitoring (6,940 trips)
- Phase 2: real-time interventions (6,189 trips)
- Phase 3: real-time & post-trip interventions (6,776 trips)
- Phase 4: real-time, post-trip interventions & gamification (7,816 trips)

The latent variable risk is measured by means of the STZ levels for headway (level 1 refers to 'normal driving' used as the reference case, level 2 refers to 'dangerous driving' while level 3 refers to 'avoidable accident driving'), with positive correlations of risk with the STZ indicators.

For the overall model, the CFI of the model is equal 0.945; TLI is 0.927 and RMSEA is 0.106. Table 3 summarizes the model fit of SEM applied for headway.

Table 3: Model Fit Summary for headway

Model Fit measures	Phase 1	Phase 2	Phase 3	Phase 4	Overall
	Value				
CFI	0.977	0.854	0.906	0.943	0.945
TLI	0.965	0.805	0.882	0.923	0.927
RMSEA	0.072	0.217	0.130	0.116	0.106
GFI	0.973	0.723	0.861	0.897	0.921
Hoelter's critical N (α = .05)	295.968	230.407	379.148	303.937	224.059
Hoelter's critical N (α = .01)	349.144	234.542	388.195	318.413	241.364
AIC	2.654×10+6	2.043×10+7	5.445×10+6	6.376×10+6	2.043×10+7
BIC	2.655×10+6	2.043×10+7	5.446×10+6	6.377×10+6	2.043×10+7

Residual variances details for headway are presented in Table 4 that follows.

Table 4: Residual variances for headway

Variable	Estimate	Std. Error	z-value	P(>|z|)
Overall				
Duration	0.996	0.001	670.537	< .001
Time indicator	0.564	8.911×10-4	632.976	< .001
Weather	0.006	6.643×10-4	8.283	< .001
Age	0.035	6.662×10-4	51.797	< .001
Average speed	0.991	0.001	663.324	< .001
Fuel type	0.473	8.009×10-4	590.334	< .001
Vehicle age	0.436	7.641×10-4	570.346	< .001
Gender	0.677	0.001	647.155	< .001
Headway_STZ_level_0	0.055	1.368×10-4	400.312	< .001
Headway_STZ_level_1	0.032	1.188×10-4	265.257	< .001
Headway_STZ_level_2	0.138	2.400×10-4	576.352	< .001
Phase 1				
Duration	1.007	0.004	253.249	< .001
Time indicator	1.415	0.033	42.282	< .001
Average speed	0.999	0.004	255.042	< .001
Fuel type	0.395	0.004	94.850	< .001
Vehicle age	0.667	0.003	197.171	< .001
Gender	0.623	0.003	181.359	< .001
Headway_STZ_level_0	0.072	4.218×10^{-4}	171.222	< .001
Headway_STZ_level_1	0.008	3.340×10^{-4}	23.648	< .001
Headway_STZ_level_2	0.149	6.542×10^{-4}	227.818	< .001
Phase 2				
Weather	-0.499	0.064	-7.755	< .001
Duration	0.998	0.003	306.597	< .001
Distance	0.005	4.939×10^{-4}	9.372	< .001
Average speed	0.003	4.945×10^{-4}	6.652	< .001
Fuel type	0.987	0.003	305.226	< .001
Vehicle age	0.997	0.003	305.282	< .001
Gender	1.000	0.003	305.296	< .001
Age	0.988	0.003	305.234	< .001
Headway_STZ_level_0	0.048	2.272×10^{-4}	211.598	< .001
Headway_STZ_level_1	0.018	1.734×10^{-4}	265.257	< .001
Headway_STZ_level_2	0.085	3.275×10^{-4}	576.352	< .001
Phase 3				
Weather	-8.210×10^{-4}	0.001	-0.616	< .001
Duration	0.996	0.003	327.096	< .001
Time indicator	0.511	0.002	301.354	< .001
Gearbox	0.016	4.210×10^{-4}	38.329	< .001
Average speed	0.992	0.003	324.757	< .001
Overtaking	0.999	0.004	266.060	< .001
Fuel type	0.068	4.479×10^{-4}	151.015	< .001
Vehicle age	0.748	0.002	325.041	< .001
Age	0.437	0.001	319.252	< .001
Gender	0.601	0.002	323.278	< .001
Headway_STZ_level_0	0.058	3.048×10^{-4}	189.597	< .001
Headway_STZ_level_1	0.036	2.706×10^{-4}	131.549	< .001
Headway_STZ_level_2	0.153	5.415×10^{-4}	282.073	< .001
Phase 4				

Variable	Estimate	Std. Error	z-value	P(>|z|)
Duration	0.980	0.003	385.572	< .001
Time indicator	0.571	0.002	369.360	< .001
Weather	0.008	0.001	8.094	< .001
Age	0.076	8.451×10-4	89.674	< .001
Average speed	0.985	0.003	377.599	< .001
Fuel type	0.266	9.625×10-4	276.387	< .001
Vehicle age	0.547	0.002	362.250	< .001
Gender	0.481	0.001	353.588	< .001
Headway_STZ_level_0	0.040	1.793×10-4	221.622	< .001
Headway_STZ_level_1	0.023	1.556×10-4	149.369	< .001
Headway_STZ_level_2	0.109	3.326×10-4	327.770	< .001

Figure 3 shows the graphical structure of the SEM results of the different phases of the experiment. The loadings of the observed proportions of the STZ of headway are not consistent among the different phases, as slight differences were observed between phases 1-2 and 3-4, regarding coping capacity. In particular, coping capacity and risk found to have a positive relationship in phases 1 and 2 of the experiment and a negative relationship in phases 3 and 4.

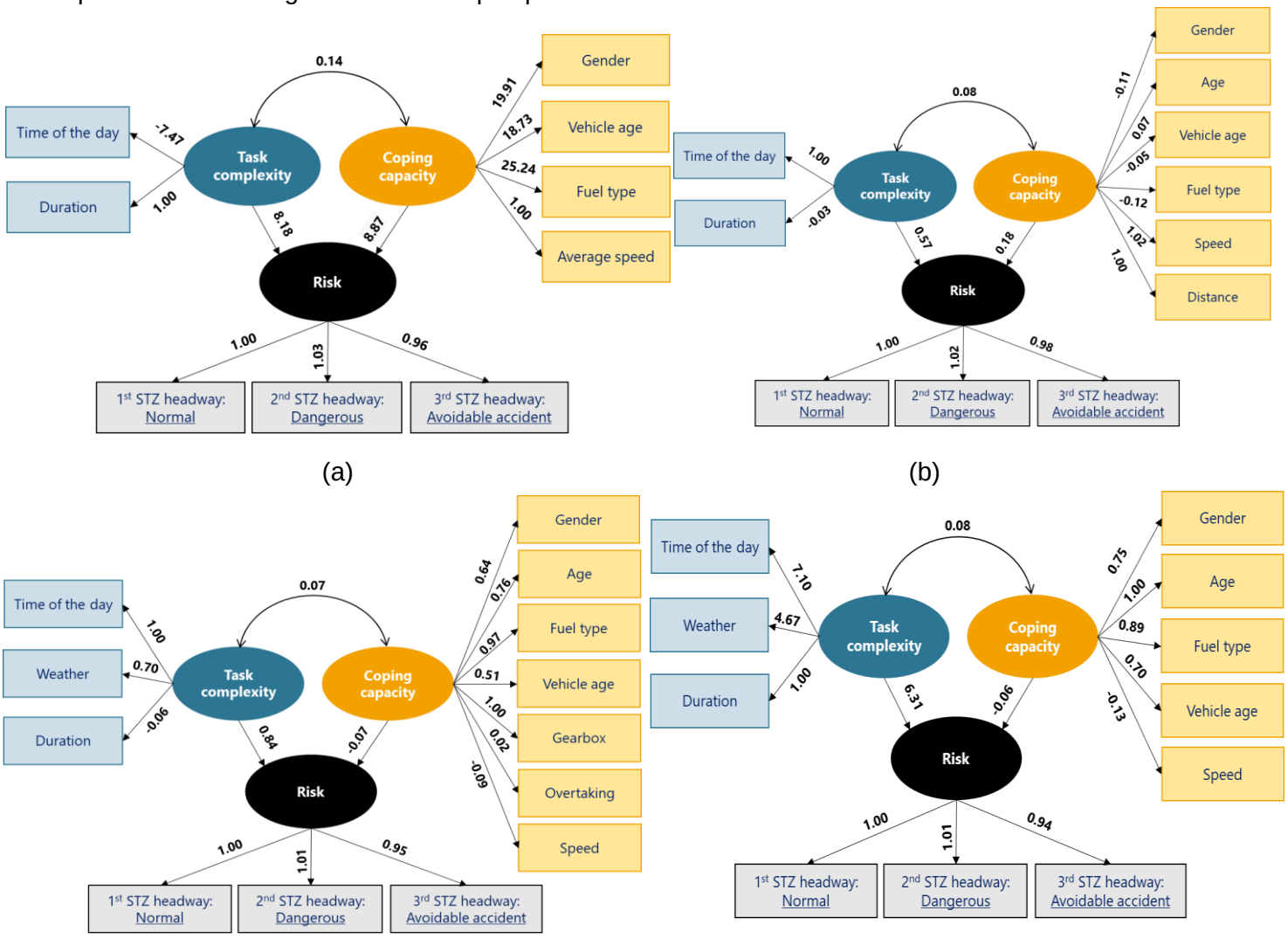

Figure 3: SEM results of task complexity and coping capacity on risk (STZ headway) - Experiment phase 1 (a), 2 (b), 3 (c), 4 (d)

6. Discussion

Within the framework of the regression analysis, the effect of road environment, vehicle state and driver behaviour on crash risk was examined and several significant results were extracted. To begin with, a negative correlation between the time of day and headway was observed. This indicates that drivers tend to maintain safer distances from the vehicle in front of them during night hours. In addition, there are more headway events during the day compared to the night. This trend could be attributed to higher traffic volumes and increased driving activity during daylight hours, leading to more instances where maintaining appropriate following distances (headway) becomes necessary. Interestingly, headway was positively correlated with adverse weather conditions (wipers on), indicating more headway events during rain. This may be because wet and slippery roads make it harder to maintain control, and reduced visibility can obscure obstacles and other vehicles.

With regards to the indicators of coping capacity – vehicle state, a positive correlation between vehicle age and headway was identified. This finding indicates that older vehicles tended to maintain less safe following distances, further compromising road safety. The analysis also revealed that vehicles running on diesel fuel tended to have shorter headways compared to those using other fuel types. One possible explanation for this could be the higher torque and better low-end performance of diesel engines, which allows for quicker acceleration and deceleration. This improved performance might enable drivers to safely maintain shorter distances between vehicles. Gearbox type being automatic was negatively correlated with headway, as vehicles with an automatic gearbox experienced fewer headway events. This finding could be due to the design of automatic transmissions, which shift gears at optimal points for fuel efficiency and smoother acceleration, often resulting in lower peak speeds compared to manual transmissions. This argument is also in line with Török (2022) findings who claimed that when the human driver made the control decisions, the severity of crashes on straight roads was greater compared to when the vehicle system made the control decisions.

Furthermore, it was demonstrated that the majority of the indicators of coping capacity – operator state, such as duration, harsh acceleration and harsh braking had a positive relationship with headway. This means that the longer the duration was, the more likely it was to keep greater distances. This correlation might be due to the fact that drivers becoming more comfortable and confident over longer trips, leading to an increase in speed, or it could reflect the tendency of drivers to speed in order to cover longer distances more quickly. This finding is in line with Fildes et al. (1991), who claimed that drivers in rural areas who were observed travelling above the average speed (and the relative speed limit) were likely to be males travelling over long distances for other than domestic journeys.

Speed had a positive effect on headway which means that as vehicle speed increased, the distance (headway) between vehicles also increased. This relationship likely reflects drivers' tendency to maintain greater distances at higher speeds to ensure safety and allow for adequate reaction time, both in real-world driving and simulated environments. However, this finding contrasts with Brackstone et al. (2009), who claimed that drivers tend to keep longer headways at lower speeds, which decrease as speed increases, stabilizing at higher speeds. Lastly, the GLM applied revealed interesting findings concerning socio-demographic characteristics, particularly gender and age. It was shown that gender influenced driving behaviour, with female drivers exhibiting a negative effect on headway. This means that female drivers tended to maintain larger distances from the vehicle in front of them compared to their male counterparts. Overall, the results indicated that older drivers experienced significantly slower reaction times, drove slower, deviated less in speed and were less able to maintain a constant distance behind a pace car compared to younger people, confirming existing studies (Pavlou & Yannis, 2022).

Within the framework of the latent analysis, five SEMs were implemented aiming to develop an integrated model of driver-vehicle-environment interaction and risk. The ultimate goal of the analyses was to identify the impact of task complexity and coping capacity on crash risk. Through the application of SEM models, the analyses revealed that higher task complexity led to higher coping capacity by the vehicle operators. It was found that when drivers encountered complex tasks, such as driving during

risky hours or adverse weather conditions, they were compelled to engage more deeply with the driving process and tended to regulate well their capacity to react to potential difficulties, while driving.

Results also revealed that task complexity was positively correlated with risk due to several reasons. Firstly, crucial indicators such as the time of day and weather conditions significantly affect crash risk. Driving during night-time or in adverse weather conditions, such as rain or fog can exacerbate the challenges posed by complex tasks, further increasing the likelihood of crashes. Secondly, drivers could become overwhelmed by the demands of complex tasks, leading to reduced attention to the road and other traffic participants. This can result in delayed detection of critical events and inadequate responses. Additionally, complex tasks may require drivers to allocate more mental resources, causing them to divert attention from essential driving activities.

On the other hand, coping capacity was negatively correlated with risk, which means that as coping capacity increases, the crash risk decreases. This relationship can be explained by the fact that drivers with higher coping capacity are better equipped to handle complex and challenging driving situations. They can manage stress, make quicker and more accurate decisions and maintain better control over their vehicles, all of which contribute to safer driving. Thus, their enhanced ability to cope with driving demands reduces the likelihood of crashes and other risky incidents, leading to a lower overall risk.

When looking into the relationship between the interaction of task complexity and coping capacity and its effect on risk, it was shown that the effect of task complexity on risk was greater than the impact of coping capacity on risk. Furthermore, a positive correlation of risk with the STZ indicators was identified in all phases, with the highest values being observed in the normal phase (i.e. STZ level 1), indicating that the latent variable risk could in fact be representing an inverse of risk, more like a normal driving. Lastly, models fitted on data from different phases of the on-road experiment validated that both real-time and post-trip interventions had a positive influence on risk compensation, increasing drivers' coping capacity and reducing dangerous driving behaviour.

This study is not without its limitations. Firstly, regarding task complexity indicators, the research only considered a limited set of variables, such as weather conditions and time of day. To provide a more comprehensive understanding of the role of task complexity on the risk measured by STZ, it would be necessary to include additional variables like road type (highway, rural, urban) and traffic volumes (high, medium, low). Secondly, the study did not account for drivers' demographic characteristics, such as education level or driving experience, which are important for assessing coping capacity. Additionally, the participants' health and medical status were not considered.

Future research could explore additional risk indicators, such as the presence of passengers, drug abuse, alcohol consumption, and seat belt use, as these are significant factors in road crashes. Further research should also incorporate demographic characteristics like education level and driving experience. Increasing the experimental sample size and making comparisons across different countries or transportation modes would be beneficial. Moreover, the models developed in this study could be further refined by including additional task complexity and coping capacity factors, such as road type, personality traits, and driving profiles. Traffic density significantly affects driving complexity, influencing stress levels and reaction times. Therefore, investigating STZ headway in relation to varying traffic conditions (high, medium, low traffic volumes) would be valuable. Additionally, incorporating participants' health and medical parameters, as well as supplementary measurements like electrocardiograms and electroencephalograms, could enhance the data.

7. Conclusions

The objective of this paper was to quantify the impact of driver, vehicle, and environmental factors on crash risk using big data. To achieve this, a naturalistic driving experiment was conducted, collecting and analysing data from 135 drivers aged 20-65. To that end, GLMs were developed to assess risk using reliable indicators such as time headway, distance travelled, speed, time of day, and weather

conditions. Additionally, SEMs were employed to explore the interrelationships among the model variables, enabling the modelling of both direct and indirect relationships.

The results indicated that higher levels of task complexity led to higher coping capacity. This suggests that drivers tend to effectively manage their ability to anticipate potential challenges when faced with difficult conditions while driving. It was found that task complexity and risk were positively correlated throughout all phases of the experiment, meaning that as task complexity increases, so does risk. Conversely, coping capacity and risk were negatively correlated in all phases, indicating that as coping capacity increases, risk decreases. Overall, the interventions positively impacted risk by enhancing the operators' coping capacity and reducing the likelihood of dangerous driving behaviour.

Considering all the aforementioned findings, this study offers valuable guidance and evidence-based recommendations for various levels, including EU, national and local Authorities, industry, and policymakers working to enhance road safety and promote the widespread adoption of effective driver assistance and monitoring systems. By integrating task complexity, coping capacity, and risk, it is possible to improve the behaviour and safety of all travellers through unobtrusive and seamless behaviour monitoring. Additionally, providing feedback and training to travellers can enhance travel behaviour, encourage shifts to safer and eco-friendly modes, and ultimately reduce risk. Authorities can utilize population-level data systems to plan mobility and safety interventions, set up road user incentives, optimize enforcement, and foster community engagement in safe travelling.

Acknowledgments

The research was funded by the EU H2020 i-DREAMS project (Project Number: 814761) funded by European Commission under the MG-2-1-2018 Research and Innovation Action (RIA).

References

Brackstone, M., Waterson, B., & McDonald, M. (2009). Determinants of following headway in congested traffic. Transportation Research Part F: Traffic Psychology and Behaviour, 12(2), 131-142.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. Structural equation modeling: a multidisciplinary journal, 14(3), 464-504.

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. Advances in neural information processing systems, 14.

Fildes, B. N., Rumbold, G., & Leening, A. (1991). Speed behaviour and drivers' attitude to speeding. Monash University Accident Research Centre, Report, 16(186), 104-115.

Hastie, T. J., & Pregibon, D. (2017). Generalized linear models. In Statistical models in S (pp. 195-247). Routledge.

Hastie, T., & Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. Biometrics, 1005-1016.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Wasserman, W. (2004). Applied linear regression models (Vol. 4, pp. 563-568). New York: McGraw-Hill/Irwin.

Pavlou, D., & Yannis, G. (2022). Assessing driving performance of older drivers: a literature review.

Salmon, P., Young, K., Lenne, M., Williamson, A., & Tomasevic, N. (2011). The nature of errors made by drivers (No. AP-R378/11).

Sheather, S. (2009). A modern approach to regression with R. Springer Science & Business Media.

Török, Á. (2020). A novel approach in evaluating the impact of vehicle age on road safety. Promet-Traffic & Transportation, 32(6), 789-796.

Török, Á. (2022). Do Automated Vehicles Reduce the Risk of Crashes–Dream or Reality?. IEEE transactions on intelligent transportation systems, 24(1), 718-727.

Washington, S.P., Karlaftis, M.G., & Mannering, F.L. (2011). Statistical and Econometric Methods for Transportation Data Analysis, second edition. CRC Press.

Washington, S., Karlaftis, M., Mannering, F., & Anastasopoulos, P. (2020). Statistical and econometric methods for transportation data analysis. Chapman and Hall/CRC.

Yannis, G., & Michelaraki, E. (2024). Review of City-Wide 30 km/h Speed Limit Benefits in Europe. Sustainability, 16(11), 4382.