Exploring driving behavior as a latent variable in safety modeling. A preliminary analysis from a driving simulator study.

Christelle Al Haddad[a], Kui Yang[a], George Yannis[b], Constantinos Antoniou[a]

[a] *Technical University of Munich*
[b] *National Technical University of Athens*
contact information: email:christelle.haddad@tum.de/mobile phone:+4915224251582

Abstract

Driver behavior, among other factors, plays an important role in road safety. To better understand it, driving simulator studies and field operational studies and/or naturalistic driving studies are often used for analyzing human factors in road safety. In an attempt to eliminate road fatalities and serious injuries on European roads, the EU is adoping a “Zero Vision” (European Commission, 2011), setting targets and investigating the factors affecting road crashes, including infrastructure, vehicle safety, driver behavior, and emergency response. The on-going EU project i-DREAMS (https://idreamsproject.eu/wp/) aims to develop a tool for monitoring driving safety and bringing in the appropriate interventions (real–time and post–trip gamification). The mathematical modeling of a safety tolerance zone as postulated in i–DREAMS is a complex dynamic problem, benefiting from advanced technologies including OBDII, smart phone application (OSeven), Mobileye (ADAS advanced features using a single camera mounted on the windshield), CardioWheel (physiological based on electrocardiogram (ECG)). The proposed methodology in this paper is a data–driven approach using discrete choice models with safety levels as a dependent variable and driving behavior as a latent explanatory variable in these models; this could be formulated as well as “normal”/“abnormal” driving, or in a broader sense aggressiveness indicator. The developed discrete choice models will be tested under different forms (multinomial, nested, but also ordered); the overall methodology also focuses on how to implement these models dynamically. A preliminary step for this is to cluster driving observations into safety levels, based on several clustering techniques, including k–means, hierarchical clustering, and Gaussian Mixture Models. An application of the proposed methods on a dataset from a previous Greek driving simulator study demonstrated that the optimal number of clusters would be four using k–means (elbow) method. In the hEART 2020 conference, the developed discrete choice models will be presented using the safety clusters showcased in this paper.

Keywords: Driving behavior, Discrete choice analysis, Latent variable model, Driving simulator, Data–driven

1. Introduction

Research in road safety often focuses on certain accident types (head–on collision, read–end collision, roll–over accident, etc.) and the risk factors contributing to such scenarios. An interest in driving behavior has also oriented research in identifying human factors in relation to risky behavior. Few studies however have aimed to develop models solely presenting different safety levels, with driving behavior as a latent variable. In the scope of the on–going i-DREAMS project, an interest arises in developing models for monitoring road/drivers’ safety, to bring in the right interventions. These models would include different variables, among which road variables, environmental, operator–related, but also physiological, measured

[*] Corresponding author
Email address: christelle.haddad@tum.de (Christelle Al Haddad)

with advanced sensory systems. The aim of this research is to propose an alternative for modeling safety levels using discrete choice models, to identify the latent variable associated with driving behavior. As part of a bigger methodology, in which the aim is to use dynamic discrete choice models, this paper presents findings using solely static discrete choice models: particularly a latent variable model. In order to do so, a clustering of driving events is necessary. As the i-DREAMS project is in its experimental design phase, the methodology is being applied on a previous dataset from a driving simulator study in Athens, Greece. In this short paper, preliminary findings on driving safety clusters are given.

2. Methodology

The overall methodology for exploring drivers' behavior follows a data–driven approach as presented in Figure 1, based on Antoniou et al. (2013).

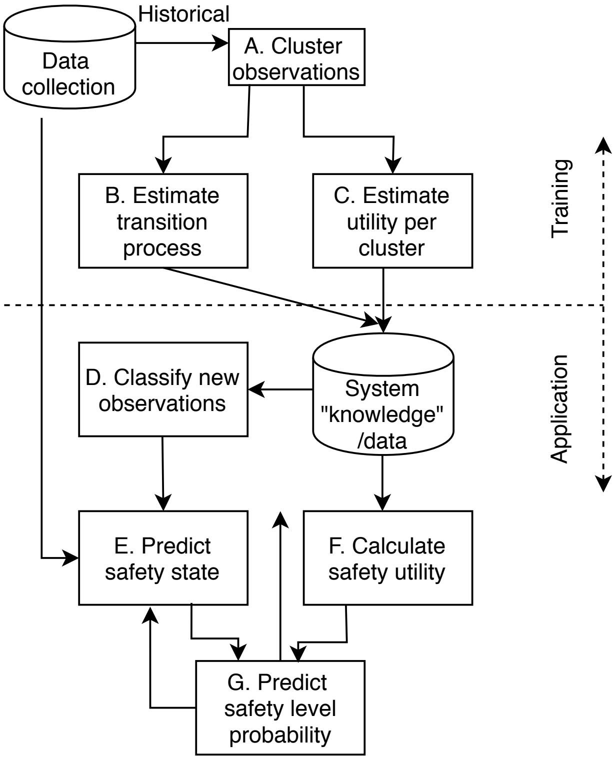

Historical
Data collection
A. Cluster observations
B. Estimate transition process
C. Estimate utility per cluster
Training
System "knowledge" /data
D. Classify new observations
Application
E. Predict safety state
F. Calculate safety utility
G. Predict safety level probability

Figure 1: Data–driven safety level prediction framework, based on Antoniou et al. (2013)

Essentially, after collecting sufficient driving observation events, the dataset can be classified into an optimal number of clusters (Step A), then a transition process can be estimated by training a Markov process (Step B). A utility per cluster, i.e. safety level, can be estimated (since the proposed models are discrete choice models, Step C). The training process can be then used to predict new observations, as explained by steps D to G. In this paper, the focus will be on the Step C and the development of suitable discrete choice models for safety levels.

An interest in exploring Dynamic Discrete Choice models (DDCMs) arises as the safety models (driving simulator and field trials) can be formulated as a dynamic problem. In a review by Cirillo & Xu (2011), DDCMs are introduced as computationally very expensive and possibly not feasible real–time. Moreover, a review of their applications to transport problems (Cirillo & Xu, 2011) presented cases on short–to–medium term vehicle–holding decisions or other dynamic decisions in travel behavior. To the best of the authors' knowledge, these models haven't been applied to similar real–time assessment problems. Therefore, due to the complexity of such problems, and until the start of the data collection, this paper will present discrete choice models (static) with one or more latent variables, applied to an existing (previous) dataset. The proposed latent variable(s) will be introduced as a behavioral construct on driving behavior; this is to be explored as “driving style”, “driving behavior”, or even “driving aggressiveness”, measured as an “aggressiveness indicator”, somehow similar to a Driving Anger Scale by Deffenbacher et al. (1994)).

To formulate the latent variable model in the i-DREAMS context, the following variables are of interest: environment parameters (time headway, forward collision warning, speed limit, etc.), vehicle parameters (vehicle speed, GPS data, RPM), driver parameters (ECG signal from the CardioWheel, as presented by Lourenço et al. (2015), mobile phone use, etc.). These are mostly dynamic; socio–demographic parameters are also of interest and are obtained from questionnaires, aiming as well to get attitudes and perceptions of drivers. The presented problem in i–DREAMS (for both simulator and field trial experiments) can be formulated as an ordered model, where the dependent variable of the safety level or zone can be ordered from a lower to a higher number of safety levels. To start however, a Multinomial Logit Model can be estimated.

Still, in all cases, the indicated variables can be part of an integrated choice and latent variable model, as presented by Ben-Akiva et al. (2002). Assuming one latent variable for an agent n for alternative (here safety level) i, the structural equations of this model are:

$$U_{ni} = X_{ni}\boldsymbol{\beta_1} + Z^*_{ni}\boldsymbol{\beta_l} + \varepsilon_{ni} \tag{1}$$

$$Z^*_{ni} = Y_{ni}\boldsymbol{\lambda}_{\boldsymbol{l}+}\omega_{ni} \tag{2}$$

$$I_{ni} = Z^*_{ni}\boldsymbol{\alpha} + \delta_{ni} \tag{3}$$

U_{ni} is the utility for agent n of alternative i

X_{ni} and Y_{ni} are subsets of explanatory variables (for the utility and latent variables, respectively).

Z^*_{ni} is the set of latent variables (one or more).

$\boldsymbol{\beta_1}$, $\boldsymbol{\beta_l}$, $\boldsymbol{\lambda_l}$, and $\boldsymbol{\alpha}$ are coefficient vectors to estimate.

ε_{ni}, ω_{ni}, δ_{ni} are error terms that are assumed to follow a normal distribution; ε_{ni} is assumed to follow a standard logistic regression.

I_{ni} are a set of indicators given from questionnaires on agents' (here drivers) attitudes and perceptions.

Applied to the studied problem (latent variable within a model for safety levels), equations 1 and 2 can be rewritten as given below.

$$U_{safety,ni} = X_{ni}\boldsymbol{\beta_1} + Z^*_{ni}\boldsymbol{\beta_l} + \varepsilon_{ni} \tag{4}$$

$$Z^*_{ni} = DrivingAggressiveness^*_{ni} = X_{ni}\boldsymbol{\lambda_l} + \omega_l \tag{5}$$

Equation 3 can similarly be written by substituting Z^* to *DrivingAggressiveness*, similarly to Equation 5. The estimated latent variable is the driving behavior or as presented here the aggressiveness index, as an explanatory variable for the safety models. It is measured by the indicators in the measurement equation in Equation 3. The latent variable itself also explains a a set of explanatory variables as given in Equation 2. For the latent variable models explained above, a full path diagram for safety is drawn in Figure 2.

3. Case study setup

As the i–DREAMS project is in the design phase and since the data collection has not yet started, the methodology is being applied to a dataset from two nationally (Greek) funded projects on road safety using driving simulator: DistrAct (modeling the impact of distraction on driver behavior) and DriverBrain (modeling the impact of mild cognitive impairment, MCI, on driver behavior; results can be found in several publications including Pavlou et al. (2015); Papantoniou et al. (2015). The methodology follows the one in Figures 1 and 2; of course, the parameters are adapted to the ones of the case study.

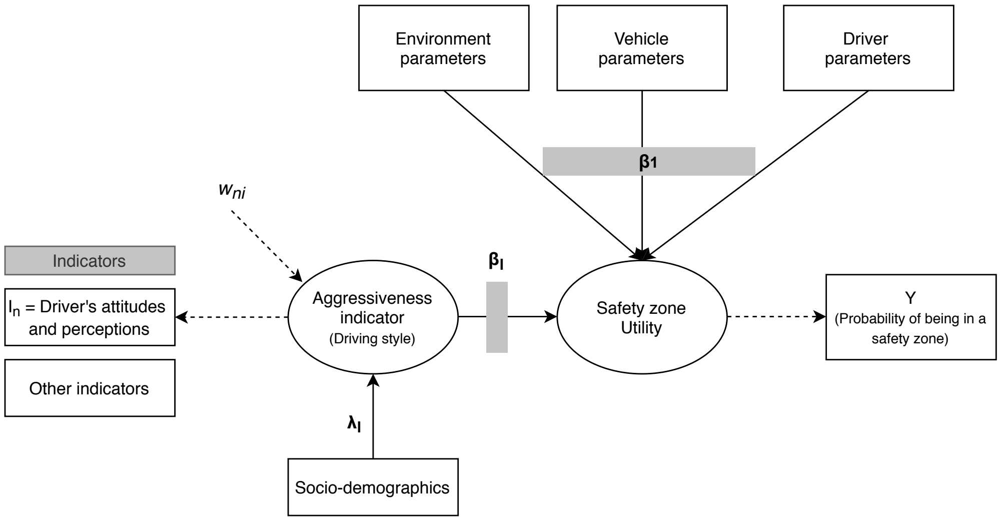


Figure 2: Full Path Diagram for Latent Variable Safety Model (*own illustration*)

The simulator study was conducted at the Department of Transportation Planning and Engineering of the School of Civil Engineering of the National Technical University of Athens (NTUA), using the FOERST Driving Simulator. Both projects were based on medical assessments, questionnaires, and a series of simulator studies. Out of the 316 participants, 192 were found to have symptoms of Mild Cognitive Impairment (MCI) or Alzheimers Disease (AD) or Parkinsons Disease (PD), while the rest did now show any symptoms and were designated as control group. Participants were asked to fill out two questionnaires, one regarding their driving behavior and one concerning a self-assessment of their memory; the study was eventually completed by 225 participants as some dropped out.

After an initial familiarization and practice sessions (20 min.), drivers were asked to participate in 12 trials each; combinations of rural/urban, low/high traffic and undistracted/conversation/mobile phone scenarios were assessed. During the driving trials, unexpected incidents took place such as an animal appearing on the street or a vehicle suddenly exiting a parking lot, aiming to capture the reaction time of the drivers. The simulator collected data at intervals of 33 to 50 milliseconds (ms), including the lateral position of the vehicle, velocity and speed limit violations, time-to-collision with the preceding vehicle, headway average, incidents of sudden braking, etc. Therefore, each second measured values for the same variable up to 30 times. The resulting database consists of three variable categories: variables on the driver characteristics (age, gender, education, medical state), driver errors (frequency of engine stops, speed limit violations, number of accidents, hit of the side bars during the simulation experiments), driving performance (driving time, average speed, reaction times and average headway). Questionnaire variables included memory questions (about speed limits, animals shown on the road, etc.), self–assessment forms (aimed at knowing drivers' own perceptions of their driving), behavioral questions (on the driving experience, driving and drinking, concentration/attention), and questions on driving characteristics (dangerous driving, angry driving, accident history, etc.).

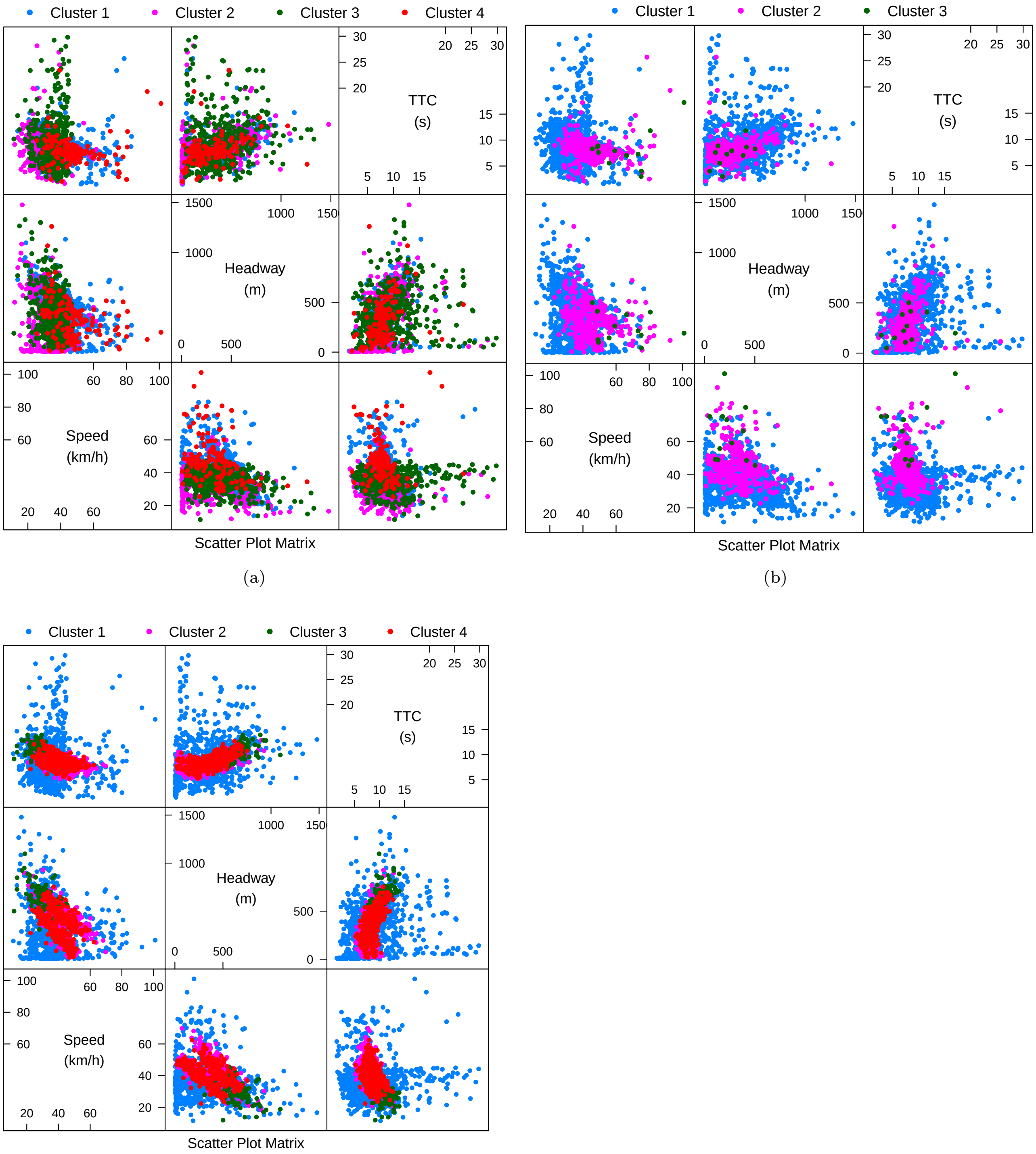
In the scope of the DriverBrain/DistrAct projects, an exploratory factor analysis aimed at investigating the variables behind driving performance, driving error, cognitive fitness and motor skills. Four structural equations models (SEM) were then developed in order to estimate the influence of the aforementioned variables on driving performance, driving errors, reaction time and accident probability. The analysis through the SEM methodology indicated the importance of cognitive fitness as a predictor of driving competence. Mobile phone use had a significant negative effect on driving task in general. Driving performance seems to deteriorate with age, although its impact on driving errors is not statistically significant.

The nature of the developed models are however different than an integrated choice and latent variable model as in Ben-Akiva et al. (2002). This is as well a main motivation to explore this dataset and apply the proposed methodology, as a first iteration (serving as an initial validation) for i–DREAMS. Although the variables collected in the Greek simulator study do not exactly correspond to the ones envisaged for i–DREAMS, the overall framework is the same. The path flow diagram presented in Figure 2 could therefore be slightly adapted to the existing dataset, with a slight difference on the collected measurements (not all technologies were used).

4. Preliminary results

In this section, preliminary results (Step A in Figure 1) applied to the case study described in Section 3 are presented (as introduced in Yang et al. (2020)). In order to identify the ideal number of clusters, three clustering algorithms were used: k–means and elbow method (Bholowalia & Kumar, 2014), hierarchical clustering method (Banfield & Raftery, 1993), and the Gaussian mixture models (GMMs) (Reynolds, 2009). The presented results in Figures 3(a) to 3(c) represent an example of the clustering algorithm outputs applied to three variables (speed, headway, and time–to–collision). The optimal number of clusters for each algorithm were found by using the elbow method (for k–means), cluster dendogram (for the hierarchical clustering), and the Bayesian Information Criterion (BIC, for the GMM); these resulted in four, three, and four clusters, respectively.

Using the t–distributed Stochastic Neighbor Embedding (t-SNE) by Maaten & Hinton (2008), the obtained clusters are visualized, and presented in Figures 4(a) to 4(c). For the clustering algorithms presented in Figure 4, all variables were used, except the ones from the questionnaires, and aggregated (average and standard deviation) for each trial (6 trials per individual and three events per trial: two incident–related events and one normal driving event). At a later stage, the included observations will be improved by adding as well questionnaire variables. The obtained results are presented in Figure 4.

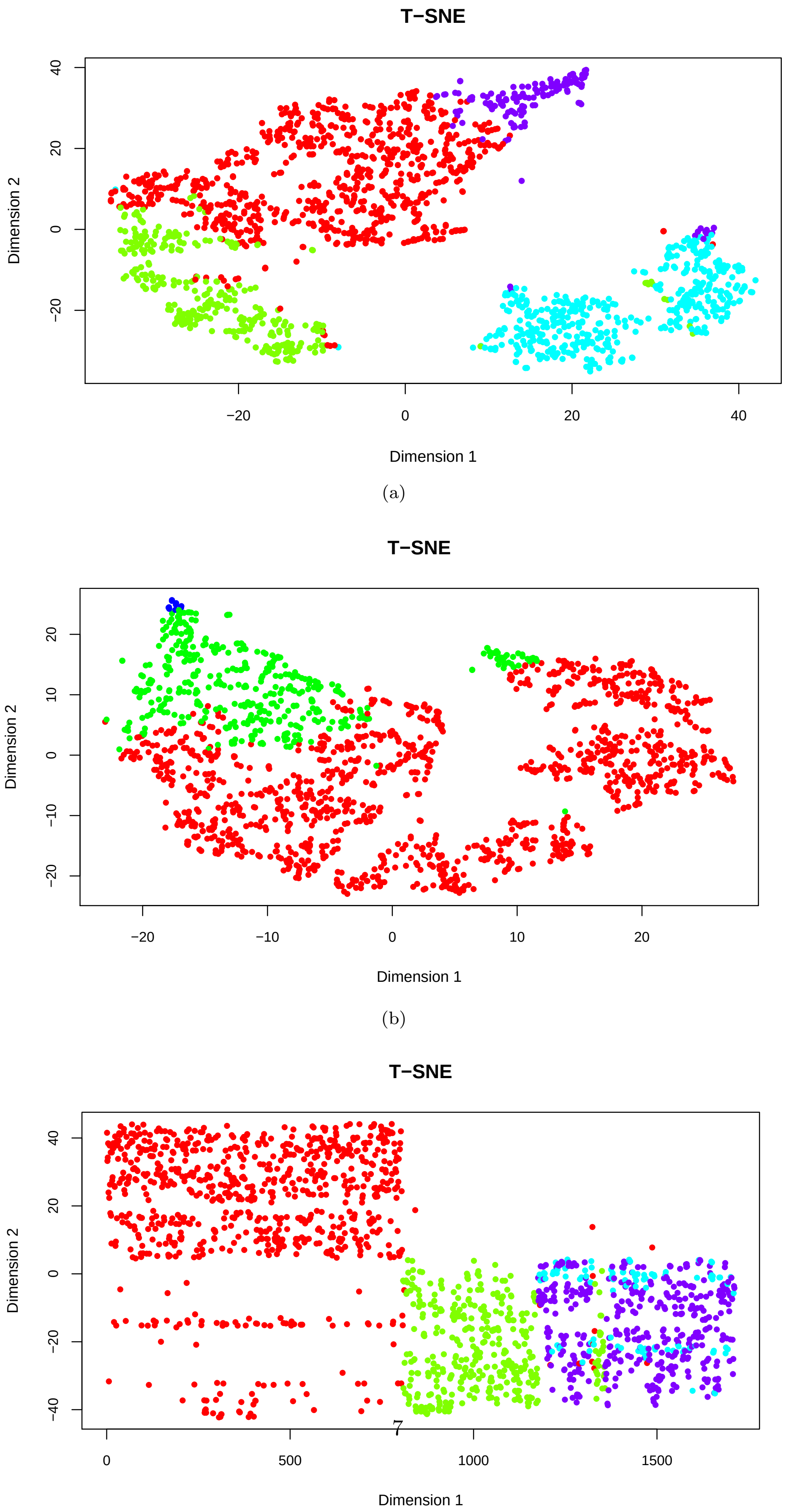


(a)

(b)

(c)

Figure 3: Clustering methods using: (a) k–means, (b) hierarchical clustering, (c) Gaussian Mixture Models



(a)

(b)

(c)

Figure 4: t–SNE visualizations for: (a) k–means, (b) hierarchical clustering, and (c) Gaussian Mixture Models (GMM)

5. Discussion and conclusions

The obtained results from the three clustering algorithms were compared in terms of their performance by computing two parameters: within.cluster.ss and avg.silwidth, as presented in Table 5, and using the "fpc" package in R by Hennig & Imports (2015). The former shows how closely related objects are within clusters; the smaller its value, the more closely related. The latter is the silhouette value (usually ranging from 0 to 1), a measurement that additionally looks how clusters are separated from each other; a value closer to 1 suggests the data is better clustered.

The obtained results suggest that the k–means clustering performed best with four optimal clusters, since the within-cluster sum of squares cluster is the smallest, and the average silhouette width closest to 1.

	Kmeans: Elbow method	K-means: Hierarchical Cluster	Mixture of Gaussians Model (GMM)
within.cluster.ss	6.2E8	11.3E8	11.6E8
avg.silwidth	0.341	0.176	-0.0239

In the hEART 2020 conference, the developed latent discrete choice models will be presented, using the obtained clusters as a dependent variable for safety models; developed models will include multinomial logit models, and ordered models, all of which introduce a latent variable on drivers' behavior. Developed models will be further improved if initial sample datasets from the i–DREAMS project are collected until then, from the integrated interface between the driving simulator, and the remaining technologies.

Acknowledgements

This work has been funded from the European Union's Horizon 2020 research and innovation programme i–DREAMS under grant agreement No 814761.

References

Antoniou, C., Koutsopoulos, H. N., & Yannis, G. (2013). Dynamic data-driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, *34*, 89–107.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, (pp. 803–821).

Ben-Akiva, M., Walker, J., Bernardino, A. T., Gopinath, D. A., Morikawa, T., & Polydoropoulou, A. (2002). Integration of choice and latent variable models. *Perpetual motion: Travel behaviour research opportunities and application challenges*, (pp. 431–470).

Bholowalia, P., & Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, *105*.

Cirillo, C., & Xu, R. (2011). Dynamic discrete choice models for transportation. *Transport Reviews*, *31*, 473–494.

Deffenbacher, J. L., Oetting, E. R., & Lynch, R. S. (1994). Development of a driving anger scale. *Psychological reports*, *74*, 83–91.

European Commission (2011). White paper. roadmap to a single european transport area towards a competitive and resource efficient transport system. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52011DC0144&from=EN.

Hennig, C., & Imports, M. (2015). Package fpc. *URL: http://cran. r-project. org/web/packages/fpc/fpc. pdf (available 08.07. 2017)*, .

Lourenço, A., Alves, A. P., Carreiras, C., Duarte, R. P., & Fred, A. (2015). Cardiowheel: Ecg biometrics on the steering wheel. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 267–270). Springer.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*, 2579–2605.

Papantoniou, P., Antoniou, C., Papadimitriou, E., Yannis, G., & Golias, J. (2015). Exploratory analysis of the effect of distraction on driving behaviour through a driving simulator experiment. In *Proceedings of the 6th Pan-hellenic Road Safety Conference, Hellenic Institute of Transportation Engineers, National Technical University of Athens, Athens*.

Pavlou, D., Papadimitriou, E., Antoniou, C., Papantoniou, P., Yannis, G., Golias, J., & Papageorgiou, S. (2015). Driving behaviour of drivers with mild cognitive impairment and alzheimers disease: a driving simulator study. In *Proceedings of the 94th Annual meeting of the Transportation Research Board, Washington*.

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*.

Yang, K., Al Haddad, C., Yannis, G., & Antoniou, C. (2020). Driving behaviour safety state classification and estimation. *Under review*, .