Methodology of Scenario Clustering for Predictive Safety Functions

Hiroki Watanabe[1], Tomáš Malý[2], Johannes Wallner[2],
Tobias Dirndorfer[2], Marcus Mai[1], and Günther Prokop[1]

***Abstract*— Data clustering is recently a common technique to group similar data with certain features. It enables finding the representative in each cluster as well. However, the clustering analysis comprises several challenging tasks, e.g., feature selection, choice among different clustering algorithms, defining the optimal cluster number, clustering with the use of a distance measure dealing with various levels of measurement, cluster validation, and interpretation of results in the end. The objective of this paper is the conceptual design of a scenario catalog including extracted representative near-crash and crash scenarios. Two clustering algorithms based on *k*-covers and *k*-medoids are applied to data in a naturalistic driving study under consideration of aforementioned aspects. Afterwards, the comparison of two clustering algorithms is conducted based on the cluster representativeness, purity, and average silhouette width. Moreover, the clusters are visualized in a two dimensional scenario space by *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE). The derived scenario catalog covers the selected database at best possible rate and enables a cost-efficient development of predictive safety functions.**

I. Introduction

Predictive safety functions support the driver by warning of a critical traffic scenario, assisting him or her during an evasive maneuver, and intervening in longitudinal and lateral vehicle control if necessary to prevent a collision or mitigate its severity. The current announcement of the European New Car Assessment Programme (Euro NCAP) regarding the high scenario diversity in two sectors, i.e., Vulnerable Road Users (VRU) Protection and Safety Assist, reflects the growing customer's concern about predictive safety functions in future road traffic [1]–[3]. This circumstance leads the developers to extend the current system functionality, i.e., active field of existing safety functions, as required to deal with upcoming diverse test scenarios in Euro NCAP. However, there are other scenarios to be addressed by safety functions in real traffic. Thus, there is a need to create a representative scenario catalog which enables an efficient development of highly available predictive safety functions [4]–[6].

II. Related Work

An overview of all the steps towards a representative scenario catalog is depicted in Fig. 1. The clustering approach is selected for the purpose of extracting representative scenarios from real data by quantifying the similarity of existing scenarios in a database. A general overview of clustering algorithms is given in [7]. In the context of traffic scenario analysis, hierarchical agglomerative clustering, partitional clustering, and model-based clustering have been mainly conducted in previous research.

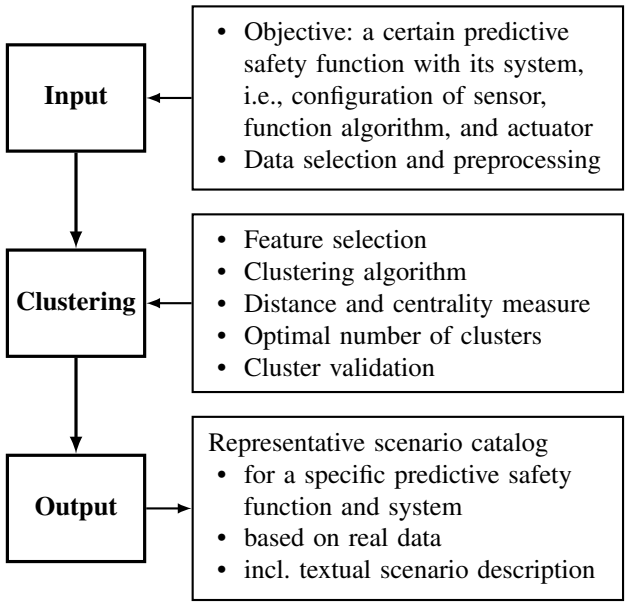


Fig. 1. Workflow from data to a representative scenario catalog

A. Hierarchical Agglomerative Clustering (HAC)

HAC with Manhattan distance and average linkage was applied to describe pedestrian crash scenarios [8], [9], rear-end scenarios [10] and cut-in scenarios [11] while an analysis with Euclidean distance and complete linkage was conducted to inspect run-off-road scenarios [12]. The utilized linkage criterion, i.e., agglomeration method, influences the later cluster structure largely [13, p. 3]. The number of existing clusters in HAC decreases successively as the dendrogram grows [14]. The cut-off value for the required total number of clusters can be defined with the use of coefficients, e.g., inconsistency coefficient [9] or local extrema in the pseudo T- and T^2-statistic as well as cubic clustering criterion [12].

B. Partitional Clustering (PC)

Intersection scenarios were analyzed using partitioning around medoids (PAM) of k-medoids algorithm where the optimal number of clusters k_{opt} was estimated by average silhouette width [13], [15], [16]. In case of analyzing a large data amount, the further derivatives of the *k*-medoids

[1]H. Watanabe, M. Mai, and G. Prokop are with Dresden Institute of Automobile Engineering, Faculty of Transport and Traffic Sciences, Technische Universität Dresden, 01062 Dresden, Germany {hiroki.watanabe, marcus.mai, guenther.prokop}@tu-dresden.de

[2]T. Malý, J. Wallner, and T. Dirndorfer are with AUDI AG, 85045 Ingolstadt, Germany {tomas1.maly, johannes.wallner, tobias.dirndorfer}@audi.de

Corresponding author: hiroki.watanabe@tu-dresden.de

algorithm, e.g., clustering large applications (CLARA) and clustering large applications based on randomized search (CLARANS), enable a less time-consuming clustering [17]. Two options regarding the initial selection of medoids, i.e., random or takeover from HAC, were discussed and the results of k_{opt} and average silhouette width were nearly the same in both options [13].

C. Model-Based Clustering (MC)

Besides aforementioned two distance-based clustering approaches, the latent class clustering (LCC) of MC was applied in [13] where the clustering was done with a probabilistic model describing the distributions in data and the Bayesian information criterion was used to estimate the optimal number of clusters. Moreover, LCC was utilized for cyclist scenarios [18] and highway scenarios [19] as well. However, MC foregrounding the description of underlying data point distributions was considered as a less appropriate approach for creating clusters, each of which contains highly similar scenarios compared to HAC which is a transparent method focusing on quantifying the similarity [12, p. 99].

III. Methodology

In the following, the methodological steps of the workflow in Fig. 1 are presented.

A. Data Selection and Preprocessing

Various classes of test scenarios emerge during the development of predictive safety functions, i.e., true positive (TP), false positive (FP), false negative (FN), and true negative (TN) as shown in Table I [20]. The aim to be attained in this paper is to find representative real-world scenarios which can be assigned to the TP and FP classes.

First, input data are needed, from which representative scenarios can be extracted. Two requirements towards data sources are the representativeness of data on the one hand and the coverage of TP and FP test scenarios on the other hand. The representativeness of a database correlates with its case number [21], [22]. Thus, a database with a large case number is required. Since an accident database does not contain, inter alia, near-crashes which partly belong to the FP class, the data from a naturalistic driving study (NDS) are essential. Due to the fact that the Second Strategic Highway Research Program (SHRP 2) is heretofore the most comprehensive NDS in the United States of America, the following analysis focuses on SHRP 2 [23].

Once a database to be inspected is selected, the variable types, i.e., levels of measurement, are to be explored so that the distinction between nominal, ordinal, and metric variables can be done for the similarity measurements of scenarios. Furthermore, the data restriction is necessary to select the data subset which is relevant regarding a certain predictive safety function and its system, e.g., the single run-off-road (near-)crashes are irrelevant for autonomous emergency braking systems with pedestrian detection and are to be excluded in data preprocessing.

TABLE I
Function-Specific Scenario Classification Based on [20]

		Collision	
		Present	Absent
Intervention	Present	True Positive[i]	False Positive[ii]
	Absent	False Negative[iii]	True Negative[iv]

[i] justified intervention [ii] non-justified intervention
[iii] missing intervention [iv] justified non-intervention

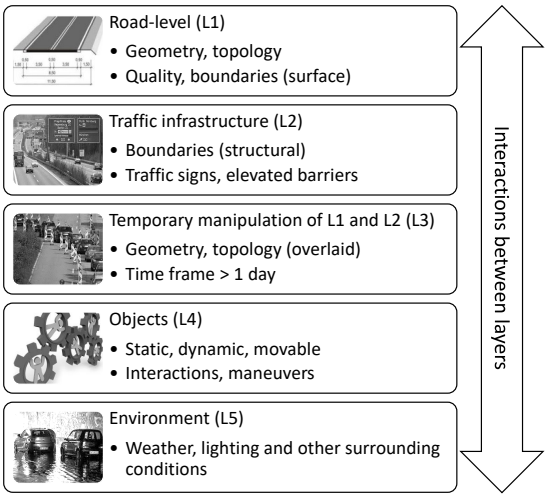


Fig. 2. Five-layer model for scenario description [25, p. 1817]

B. Feature Selection

As a first step in the clustering part, the variables for the cluster analysis, i.e., features $\mathcal{X}_i$, $\{i \in \mathbb{N} \,|\, 1 \leq i \leq \mu\}$, are to be determined, where i is an index and μ represents the number of cluster variables. The selection of cluster variables is an important step in the cluster analysis and only a restricted amount of variables is to be used since even one additional unnecessary variable can affect the clustering results largely [24, p. 350]. For the purpose of describing test scenarios in the development of predictive safety functions, the cluster variables in this study are selected from valid variables in the database based on authors' domain knowledge with a focus on the fourth layer in Fig. 2 [25].

After selecting the cluster variables $\mathcal{X}_i$, the codings of $\mathcal{X}_i$, i.e., variable attributes, are to be inspected and grouped as necessary since some attributes show a similarity due to the given coding scheme in the database, e.g., if the variable *Visual Obstruction* contains {*No Obstruction*, *Sunlight*, *Headlights*, *Building*, *Trees*} and the difference between *Sunlight* and *Headlights* as well as *Buildings* and *Trees* is irrelevant regarding the sensor perception, these attributes can be grouped as {*No Obstruction*, *Lights*, *Statistic Obstacles*}. It is noted that the data containing at least one variable attribute *Unknown* in $\mathcal{X}_i$ are to be filtered since *Unknown* increases the heterogeneity of data subset unnecessarily.

C. Distance and Centrality Measure

In order to quantify the (dis-)similarity of scenarios, a distance measure is needed. The measure used in this research is the mixed similarity measure (MSM) calculating the distance of binary, nominal, ordinal, and metric variables at the same time [26]. Let $\mathcal{S}_a$ and $\mathcal{S}_b$ be two scenarios consisting one attribute from four cluster variables ($\mu = 4$), i.e., $\mathcal{X}_1$, $\mathcal{X}_2$, $\mathcal{X}_3$, and $\mathcal{X}_4$, with different levels of measurement, i.e.,

$$\mathcal{S}_a = \{s_{1a}, s_{2a}, s_{3a}, s_{4a}\}, \tag{1}$$

$$\mathcal{S}_b = \{s_{1b}, s_{2b}, s_{3b}, s_{4b}\}, \tag{2}$$

where $\{s_{ia}, s_{ib}\} \in \mathcal{X}_i$, $\{i \in \mathbb{N} \,|\, 1 \leq i \leq 4\}$. Let $\mathcal{X}_1$ be binary, $\mathcal{X}_2$ be nominal, $\mathcal{X}_3$ be ordinal, and $\mathcal{X}_4$ be metric. The MSM estimates the sub-distance of each cluster variable. For the binary (B) and nominal variable (N), the matching distance d_{BN} from [27] is used, which is defined for the dissimilarity between any two scenarios, i.e., $\mathcal{S}_a$ and $\mathcal{S}_b$, as follows,

$$d_{\text{BN}}(\mathcal{S}_a, \mathcal{S}_b) = \sum_{i=1}^{2} \delta(s_{ia}, s_{ib}), \qquad a, b \in \mathbb{N},$$

$$\text{where} \quad \delta(s_{ia}, s_{ib}) = \begin{cases} 0 & \text{if} \quad s_{ia} = s_{ib}, \\ 1 & \text{if} \quad \text{otherwise}. \end{cases} \tag{3}$$

The sub-distance of ordinal variable d_{O} considers the maximum value in $\mathcal{X}_3$ while the sub-distance of metric one d_{M} takes the difference between the maximum and minimum value in $\mathcal{X}_4$ into account [26].

$$d_{\text{O}}(\mathcal{S}_a, \mathcal{S}_b) = \frac{\mid s_{3a} - s_{3b} \mid}{\max \mathcal{X}_3 - 1}, \tag{4}$$

$$d_{\text{M}}(\mathcal{S}_a, \mathcal{S}_b) = \frac{\mid s_{4a} - s_{4b} \mid}{\max \mathcal{X}_4 - \min \mathcal{X}_4}, \tag{5}$$

where $d_{\text{O}}, d_{\text{M}} \in [0, 1]$. The total distance d_{MSM} between $\mathcal{S}_a$ and $\mathcal{S}_b$ is the sum of sub-distances of cluster variables [26].

$$d_{\text{MSM}}(\mathcal{S}_a, \mathcal{S}_b) = d_{\text{BN}} + d_{\text{O}} + d_{\text{M}}. \tag{6}$$

Using this MSM (6) as a distance measure, two clustering algorithms based on approximate k-covers (AKC) of HAC [28] and PAM of PC [29] are applied to the input dataset. Both algorithms are shown in Algorithm 1 and 2. A measure estimating the centrality of a data point within its cluster based on the Euclidean distance was presented in [30] and is termed centrality measure (CM) in the following. Instead of the Euclidean distance, the aforementioned MSM (6) is utilized for the calculation of CM in this paper. The usage of CM comprises the selection of representative scenarios in both algorithms, the determining of initial medoids in the PAM-based algorithm, and the linkage criterion in the AKC-based algorithm.

Let Ω be the scenario space containing a set of k clusters ($\mathcal{C}$) and the cluster $\mathcal{C}_j$ be a set of m_j scenarios ($\mathcal{S}$), i.e.,

$$\Omega = \{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_k\}, \qquad k \in \mathbb{N}, \tag{7}$$

$$\mathcal{C}_j = \{\mathcal{S}_{j1}, \mathcal{S}_{j2}, ..., \mathcal{S}_{jm_j}\}, \qquad \{j, m_j \in \mathbb{N} \,|\, 1 \leq j \leq k\}. \tag{8}$$

A scenario $\mathcal{S}_{jq}$ where $\{q \in \mathbb{N} \mid 1 \leq q \leq m_j\}$ of cluster $\mathcal{C}_j$ contains μ attributes due to the μ cluster variables used in the analysis. The cluster variable $\mathcal{X}_i$ has ν_i attributes which are grouped in Section III-B, i.e.,

$$\mathcal{X}_i = \{s_{i1}, s_{i2}, ..., s_{i\nu_i}\}, \qquad \{i, \nu_i \in \mathbb{N} \mid 1 \leq i \leq \mu\}. \tag{9}$$

The CM of $\mathcal{S}_{jq}$ in $\mathcal{C}_j$ with respect to the other scenarios in the same cluster, i.e., $d_{\text{CM}}(\mathcal{S}_{jq})$, is expressed as follows [30],

$$d_{\text{CM}}(\mathcal{S}_{jq}) = \sum_{p=1}^{m_j} \frac{d_{\text{MSM}}(\mathcal{S}_{jp}, \mathcal{S}_{jq})}{\sum_{r=1}^{m_j} d_{\text{MSM}}(\mathcal{S}_{jp}, \mathcal{S}_{jr})}, \qquad p, q, r \in \mathbb{N}. \tag{10}$$

The scenario with the lowest $d_{\text{CM}}(\mathcal{S}_{jq})$ among m_j scenarios in $\mathcal{C}_j$ can be considered as the cluster representative in $\mathcal{C}_j$.

While every scenario, i.e., data point, in the hierarchical agglomerative AKC-based algorithm at the first stage can be seen as the initial cluster center, i.e., cluster representative, the PAM-based algorithm requires a method how to identify the initial cluster representatives unless they are selected randomly (see *select_initial_medoids* in Algorithm 2). For this purpose, k scenarios with the lowest $d_{\text{CM}}(\mathcal{S}_{jq})$ values in Ω are selected as initial cluster representative, from which the swap process starts as shown in [30] (see the while loop in Algorithm 2). A full description of the swap process is provided in [29, pp. 103–104] and [31]. In case that there are only two scenarios in a cluster or there are several scenarios with the lowest $d_{\text{CM}}(\mathcal{S}_{jq})$ value in a cluster, the cluster representative is selected randomly among those.

In the agglomeration process of AKC-based algorithm, i.e., while loop in Algorithm 1, a linkage criterion is required, with which two most similar clusters are to be merged (see *d2_min* and *cl_agglom*). As a linkage criterion the cluster representatives are used, each of which is estimated with CM (10) in a cluster. For *d2_min* and *cl_agglom*, the distance *d2_min* between a pair of cluster representatives is measured after each of non-representatives in Ω was assigned to the nearest certain cluster representative, i.e., *cl_assignm*, based on the distance d between the representatives and all points in Ω. In case that there are multiple pairs of cluster representatives with the same *d2_min* values, *cl_agglom* groups those in a single step. The number of clusters after the current agglomerative step is determined in *num_of_cl*. Afterwards, the calculation of the representative in each cluster is carried out using CM (10). It is noted that the current cluster representatives are termed *curr_covers* in Algorithm 1 and *curr_medoids* in Algorithm 2.

D. Optimal Number of Clusters

After defining the distance measure and the way how to identify the cluster representative, the optimal number of clusters k_{opt} needs to be determined systematically. For this purpose, the scenario-specific average silhouette width $\sigma_{\mathcal{S}}$ is utilized at first, which represents the strength of belonging to a certain cluster for each scenario [32]. Afterwards, the k-specific average silhouette width $\sigma_{\Omega k}$ in the whole scenario space Ω is calculated, with which the existence and strength of cluster structures in the scenario space Ω is verified [29]. The number of clusters k is set (see Input in Algorithm 1 and 2) and varied in the range between 2 and N as $\sigma_{\Omega k}$

Algorithm 1 Clustering algorithm based on AKC [28]

Input: A set Ω of N points and the number of clusters k
Output: A set of k covers, average silhouette width σ, and cluster assignments for all N points

1 *num_of_cl* $= N$
2 *curr_covers* $=$ *select_initial_covers*(Ω, *num_of_cl*)
3 **while** (*num_of_cl* $> k$) **do**
4 $d =$ *calc_distances*(Ω, *curr_covers*)
5 *cl_assignm* $=$ *assign_points_to_cl*(Ω, *curr_covers*, d)
6 $d2 =$ *calc_distances*(*curr_covers*)
7 $d2$*_min* $=$ *find_most_similar_covers*($d2$)
8 *cl_agglom* $=$ *merge_cl*(*cl_assignm*, $d2$*_min*)
9 *num_of_cl* $=$ *count_cl*(*cl_agglom*)
10 *curr_covers* $=$ *calc_new_covers*(Ω, *cl_agglom*)
11 $d =$ *calc_distances*(Ω, *curr_covers*)
12 *cl_assignm* $=$ *assign_points_to_cl*(Ω, *curr_covers*, d)
13 $\sigma =$ *calc_silhouette*(Ω, *cl_assignm*)
14 **end while**
15 **return** *curr_covers*, *cl_assignm*, σ

is undefined for $k=1$ [29, p. 85]. In each iteration, $\sigma_{\Omega k}$ is estimated and saved to find the maximal $\sigma_{\Omega k}$ value and its corresponding k value in the end.

In order to verify whether a certain scenario $\mathcal{S}_{jq}$ in $\mathcal{C}_j$ is assigned to the correct cluster, two types of distances are required, i.e., the average distance α_j of $\mathcal{S}_{jq}$ to all scenarios in $\mathcal{C}_j$ and the average distance β_w of $\mathcal{S}_{jq}$ to all scenarios in another cluster $\mathcal{C}_w$ ($w \neq j$) of Ω. They are formulated as

$$\alpha_j(\mathcal{S}_{jq}, \mathcal{C}_j) = \frac{d_{\text{MSM}}(\mathcal{S}_{jq}, \mathcal{C}_j)}{m_j - 1}, \tag{11}$$

$$\text{where} \quad d_{\text{MSM}}(\mathcal{S}_{jq}, \mathcal{C}_j) = \sum_{u=1}^{m_j} d_{\text{MSM}}(\mathcal{S}_{jq}, \mathcal{S}_{ju}). \tag{12}$$

$$\beta_w(\mathcal{S}_{jq}, \mathcal{C}_w) = \frac{d_{\text{MSM}}(\mathcal{S}_{jq}, \mathcal{C}_w)}{m_w}, \quad \{w \in \mathbb{N} \,|\, 1 \leq w \leq k, w \neq j\}. \tag{13}$$

Among all $\beta_w(\mathcal{S}_{jq}, \mathcal{C}_w)$ values in Ω, the minimum distance $\beta_{w\text{min}}(\mathcal{S}_{jq}, \mathcal{C}_w)$ is relevant for the scenario-specific average silhouette width $\sigma_{\mathcal{S}}(\mathcal{S}_{jq})$ that juxtaposes information about how close $\mathcal{S}_{jq}$ is positioned with respect to the scenarios in the same cluster and how close $\mathcal{S}_{jq}$ is positioned with respect to scenarios in the other cluster with minimal distance as follows [32, pp. 55–56],

$$\sigma_{\mathcal{S}}(\mathcal{S}_{jq}) = \frac{\beta_{w\text{min}}(\mathcal{S}_{jq}, \mathcal{C}_w) - \alpha_j(\mathcal{S}_{jq}, \mathcal{C}_j)}{\max(\alpha_j(\mathcal{S}_{jq}, \mathcal{C}_j), \beta_{w\text{min}}(\mathcal{S}_{jq}, \mathcal{C}_w))}. \tag{14}$$

The higher $\sigma_{\mathcal{S}}(\mathcal{S}_{jq})$ is, the better the analyzed scenario $\mathcal{S}_{jq}$ fits into the cluster $\mathcal{C}_j$ [13]. The mean of all N scenario-specific $\sigma_{\mathcal{S}}(\mathcal{S}_{jq})$ values in Ω is the final σ value for the preset k, i.e.,

$$\sigma = \sigma_{\Omega k} = \frac{1}{N} \sum_{j=1}^{k} \sum_{q=1}^{m_j} \sigma_{\mathcal{S}}(\mathcal{S}_{jq}), \tag{15}$$

Algorithm 2 Clustering algorithm based on PAM [29]

Input: A set Ω of N points and the number of clusters k
Output: A set of k medoids, average silhouette width σ, minimum costs, and cluster assignments for all N points

1 *curr_medoids* $=$ *select_initial_medoids*(Ω, k)
2 $d =$ *calc_distances*(Ω, *curr_medoids*)
3 *cl_assignm* $=$ *assign_points_to_cl*(Ω, *curr_medoids*, d)
4 *min_costs* $=$ *calc_costs*(d)
5 *repeat* $=$ true
6 **while** (*repeat* $=$ true) **do**
7 **for** $h = 1$ to $N - k$ **do**
8 *new_medoids* $=$ *calc_new_medoids*(Ω, *curr_medoids*, h)
9 *new_d* $=$ *calc_distances*(Ω, *new_medoids*)
10 *new_cl_assignm* $=$ *assign_points_to_cl*(Ω, *new_medoids*, *new_d*)
11 *new_costs* $=$ *calc_costs*(*new_d*)
12 **if** (*min_costs* $>$ *new_costs*) **then**
13 *min_costs* $=$ *new_costs*
14 *cl_assignm* $=$ *new_cl_assignm*
15 $d =$ *new_d*
16 *curr_medoids* $=$ *new_medoids*
17 **break**
18 **end if**
19 **if** ($h = N - k$) **then**
20 *repeat* $=$ false
21 **end if**
22 **end for**
23 **end while**
24 $\sigma =$ *calc_silhouette*(Ω, *cl_assignm*)
25 **return** *curr_medoids*, *cl_assignm*, σ, *min_costs*

which is calculated in line 13 in Algorithm 1 and line 24 in Algorithm 2. The meaning of σ is depicted in Table II [13], [29]. The optimal number of clusters k_{opt} is estimated when σ achieves its maximum [29], i.e.,

$$k_{\text{opt}} = \underset{\sigma \in [-1,1]}{\arg\max}\, \sigma \quad \text{s.t.} \quad k \in [2, N], \tag{16}$$

where $\sigma = 0$ for $k = N$ [29, p. 85].

E. Cluster Validation

The cluster validation concludes the cluster analysis. The output of cluster algorithms is a set of representative scenarios. The representative $\mathcal{S}_{j\text{rep}}$ in its cluster $\mathcal{C}_j$ can be described with μ cluster variables (9) as follows (see Table III),

$$\mathcal{S}_{j\text{rep}} = \{s^j_{1\text{rep}}, s^j_{2\text{rep}}, ..., s^j_{\mu\text{rep}}\}. \tag{17}$$

Besides the average silhouette width σ, the purity measures, i.e., ρ_1 and ρ_2, are used to quantify the quality of all clusters [28]. The measure ρ_1 represents the clustering ability to assign scenarios with dissimilar characteristics in different clusters, while the measure ρ_2 indicates the ability to group scenarios with identical characteristics in a cluster. Further information about ρ_1 and ρ_2 are provided in [28, pp. 14–15].

TABLE II

AVERAGE SILHOUETTE WIDTH σ BASED ON [13], [29]

Range*	Description
$-1.00 \leq \sigma \leq 0.25$	No substantial structure has been found.
$0.26 \leq \sigma \leq 0.50$	A weak structure has been found that could be artificial.
$0.51 \leq \sigma \leq 0.70$	A reasonable structure has been found.
$0.71 \leq \sigma \leq 1.00$	A strong structure has been found.

*rounded to two decimal places

TABLE III

SCENARIO SPACE Ω

$\mathcal{C}$	$\mathcal{S}$	$\mathcal{X}_1$	$\mathcal{X}_2$	...	$\mathcal{X}_i$	...	$\mathcal{X}_\mu$
$\mathcal{C}_1$	$\mathcal{S}_{1\text{rep}}$	$s^1_{1\text{rep}}$	$s^1_{2\text{rep}}$	...	$s^1_{i\text{rep}}$	...	$s^1_{\mu\text{rep}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathcal{C}_j$	$\mathcal{S}_{j\text{rep}}$	$s^j_{1\text{rep}}$	$s^j_{2\text{rep}}$	...	$s^j_{i\text{rep}}$	...	$s^j_{\mu\text{rep}}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\mathcal{C}_k$	$\mathcal{S}_{k\text{rep}}$	$s^k_{1\text{rep}}$	$s^k_{2\text{rep}}$	...	$s^k_{i\text{rep}}$	...	$s^k_{\mu\text{rep}}$

Moreover, the representativeness of the representative $\mathcal{S}_{j\text{rep}}$ in its cluster $\mathcal{C}_j$ and the representativeness of k clusters in Ω are inspected. Those are termed $\gamma_{\mathcal{C}_j}$ and γ_Ω. For each variable $\mathcal{X}_i$, the number of scenarios with representative attribute $s^j_{i\text{rep}}$ in the cluster $\mathcal{C}_j$ is defined as η_{ji}. The representativeness of the representative scenario $\mathcal{S}_{j\text{rep}}$ with its attributes (17) in $\mathcal{C}_j$ is formulated as

$$\gamma_{\mathcal{C}_j} = \frac{1}{\mu} \sum_{i=1}^{\mu} \frac{\eta_{ji}}{m_j}. \tag{18}$$

The representativeness of all k clusters in Ω is expressed as the mean of $\gamma_{\mathcal{C}_j}$ values weighted according to the cluster content m_j, i.e.,

$$\gamma_\Omega = \sum_{j=1}^{k} \frac{m_j}{N} \gamma_{\mathcal{C}_j}. \tag{19}$$

In this paper, ρ_1, ρ_2, $\gamma_{\mathcal{C}_j}$, and γ_Ω are calculated for the binary and nominal cluster variables since the ordinal ones, e.g., level of service, are diverse and the metric ones, e.g., velocity with a decimal place, are even more.

IV. EXEMPLARY ANALYSIS AND RESULTS

A. Data Preprocessing

The SHRP 2 dataset [33] contains 1465 crashes and 2710 near-crashes, and 20 000 balanced sample baseline events. The data are restricted by criteria in Table IV and the remaining 973 rear-end striking events between two motorists are analyzed. A full description of variables and attributes is provided in [34]. The selected cluster variables $\mathcal{X}_i$ and the number of attributes ν_i are listed in Table V.

TABLE IV

FILTER CRITERIA

1. *eventSeverity1*: Crash or Near-crash
2. *vehicle1SubjectConfig*: Same trafficway and same direction
3. *vehicle3Config*: Unknown accident type or No impact×
4. *incidentType1*: Rear-end striking
5. *motorist2Location*:
   - In front of the subject vehicle
   - In front and to the immediate left of the subject vehicle
   - In front and to the immediate right of the subject vehicle
6. *motorist2Type*: Automobile, SUV, Van, or Pickup
7. *preIncidentManeuver*†:
   - Going straight (constant speed, accelerating, with unintentional drifting) or starting in traffic lane
   - Decelerating in traffic lane or stopped in traffic lane
8. *precipitatingEvent*:
   - Other vehicle ahead - stopped on roadway more than 2 seconds
   - Other vehicle ahead - slowed and stopped 2 seconds or less
   - Other vehicle ahead - at a slower constant speed
   - Other vehicle ahead - decelerating
   - Other vehicle lane change - left in front of subject
   - Other vehicle lane change - right in front of subject
9. *vehicle1EvasiveManeuver1*†:
   - No reaction
   - Braked
   - Steered to left or right
   - Braked and steered left or right
   - Accelerated
   - Accelerated and steered left or right

× For only two motorists † Subject vehicle

TABLE V

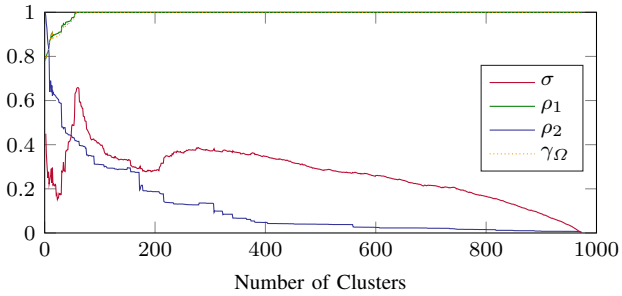
CLUSTER VARIABLES

i	$\mathcal{X}_i$	ν_i	Description
1	*visualObstructions*[B]	2	If driver's vision obscured
2	*motorist2Location*[N]	3	Position of other vehicle
3	*preIncidentManeuver*[N]	2	Vehicle maneuver, ca. 2-6 s prior to critical event
4	*precipitatingEvent*[N]	6	Critical event
5	*vehicle1EvasiveManeuver1*[N]	6	Subject driver's reaction
6	*trafficDensity*[O]	7	Level of service
7	*v_mean_precEv*[M]		Mean ego-velocity in km/h between critical event and 2 s before critical event
8	*v_mean_coll*[M]		Mean ego-velocity in km/h between (near-)crash and 0.5 s before (near-)crash

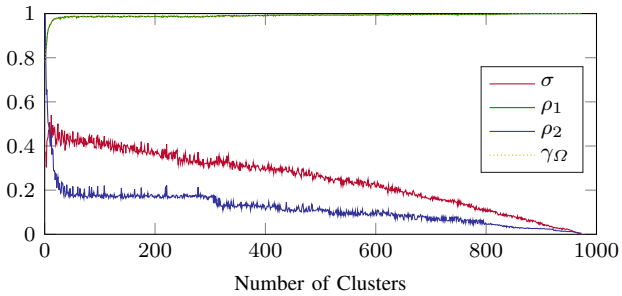
[B] binary, [N] nominal, [O] ordinal, [M] metric

B. Results

Two clustering algorithms presented in Section III are applied to the aforementioned restricted dataset. The optimal number of clusters k_{opt} is estimated at the maximum of the average silhouette width σ in Fig. 3. Both of the maximal values, i.e., $\sigma^{\text{AKC}}_{\text{max}}$ and $\sigma^{\text{PAM}}_{\text{max}}$, represent that reasonable structures exist in Ω according to Table II. Thus, AKC-based algorithm creates 60 clusters at $\sigma^{\text{AKC}}_{\text{max}}$ while the PAM-based algorithm generates 12 clusters at $\sigma^{\text{PAM}}_{\text{max}}$. The analysis deduces that an agglomerative bottom-up approach seldom groups outliers with the other faraway scenarios, which is reflected in the



(a) AKC-based algorithm



(b) PAM-based algorithm

Fig. 3. Cluster validation with quality measures

result that k_{opt} of the AKC-based algorithm is higher than k_{opt} of the PAM-based algorithm. Therefore, many outliers have their own clusters in the AKC-based algorithm as illustrated in Fig. 4. Consequently, the AKC-based algorithm can be considered as less robust towards outliers than the PAM-based algorithm. Moreover, the analysis shows that ρ_2 is around 0.4 at σ_{max}, which means the algorithms are barely capable of grouping the equivalent scenarios in the same cluster while they are able to separate the different scenarios at σ_{max} as ρ_1 outlines. This result indicates that there are many almost identical scenarios in several clusters. γ_Ω shows a similar tendency to ρ_1 in the whole range while ρ_1 correlates with ρ_2 inversely. It is noted that only binary and nominal variables are considered for ρ_1, ρ_2, and γ_Ω. The noisy curves of PAM-based algorithm can result from its swap process in each iteration.

Each cluster representative of the largest clusters in the AKC-based algorithm, which cover at least $75\,\%$ of N scnearios with m_j in total, is exemplary depicted in Table VI. Furthermore, the distribution of metric variables is illustrated in Fig. 5 where a high diversity of velocity values in several clusters is apparent.

V. Discussion

The clustering of rear-end striking near-crashes and crashes is conducted with eight variables by applying the AKC-based and PAM-based algorithm in combination with MSM and CM.

AKC is an agglomerative clustering approach and requires less computing time than the optimal *k*-covers algorithm (OKC) to achieve the same quality of clustering [28, p. 16]. To authors' knowledge, there is no research dealing with AKC or OKC in traffic scenario analysis. The implemented linkage criterion measures the distance between cluster representatives calculated in each agglomerative iteration. Other criteria, e.g., average linkage or Ward's method, are to be applied as well since the agglomeration method affects the results significantly [13, p. 3]. An overview of linkage criteria is given in [7, p. 647].

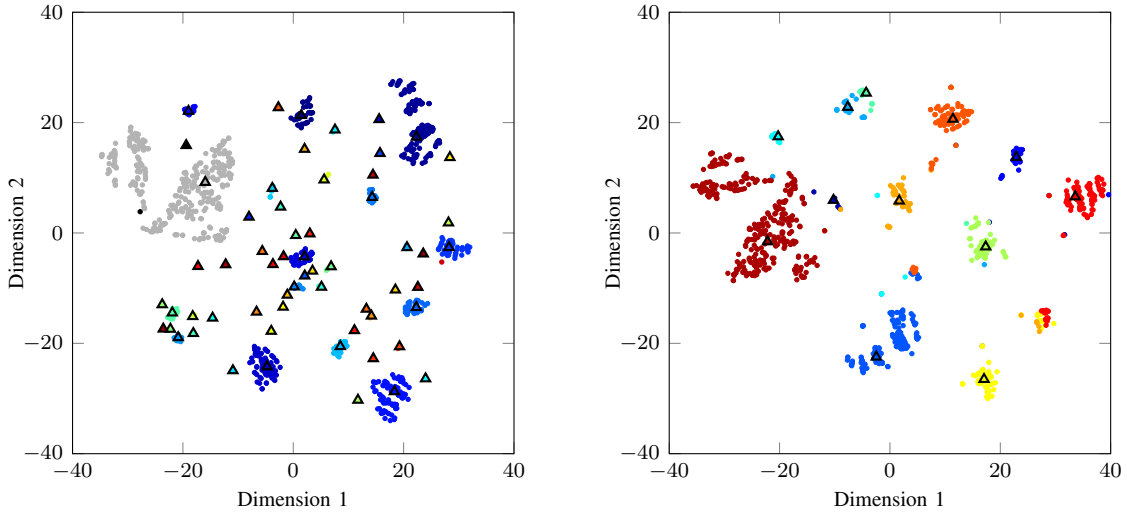
In comparison to AKC, PAM is a commonly used algorithm in previous studies, e.g., [13], [15], [16]. Another less time-consuming derivative of the k-medoids algorithm, i.e., CLARA, considers only a part of the input data, which can lead to a local minimum during the distance calculation in the swap process [17]. Thus, the PAM was preferred in this study. The initial medoids are selected not randomly but based on CM (10). A further option is to utilize the final cluster representatives of HAC, e.g., AKC-based algorithm, to analyze the influence of the initial medoids on the swap process and final cluster structures in a similar way to [13].

The results regarding k_{opt} indicate that the PAM-based algorithm is robust towards outliers while the AKC-based algorithm is bottom-up and seeks the most similar data points globally in each step. Since there can be relevant rare scenarios, i.e., outliers, in the traffic occurrence or a database, which are to be considered in the development of predictive safety functions, the outliers are also important. Thus, the merging of rare scenarios and a large cluster is to be avoided. Alternatively, a weighting scheme for rare scenarios is needed to prevent the clustering of outliers.

The results of AKC-based and PAM-based algorithm using MSM and CM show that a high ρ_1 value can be achieved at σ_{max} while ρ_2 is still low. The next step is to explore how to optimize both of the purity measures at σ_{max}. Due to the lack of consideration of ordinal and metric variables in the calculation of purity measures, a new measure is required, which takes into account all the levels of measurement in cluster variables. Moreover, the entropy renders the inhomogeneity of clusters measurable [36]–[38]. The comparison between entropy value before and after the clustering can indicate the ability of a cluster algorithm to reduce the inhomogeneity in input data.

The sensitivity analysis of cluster variables, i.e., features, is an essential step to better understand the influence of a single cluster variable, its type, and its distribution. The diverse velocity values in clusters in Fig. 5 indicate that d_{M} (5) in MSM must be weighted and prioritized since the velocity values show a high granularity and d_{M} achieves the maximal distance, i.e., $d_{\text{M}} = 1$, less often than the matching distance d_{BN} (3) of binary and nominal variables. As another option, the clustering can be conducted with binary, nominal, and ordinal variables in the initial dataset at first and with metric variables in the clustered data subsets afterwards.

The feature selection can be supported by the principle component analysis as shown in [13]. As an alternative approach, a set of risk-inducing variables in [5], [6] can be used as features. Furthermore, the similarity measurement of time series data is to be conducted as presented in [39]. However, it is noted that the number of variables needs to be low with respect to the case number in the input data [40].



(a) AKC-based algorithm ($k_{\text{opt}} = 60$) (b) PAM-based algorithm ($k_{\text{opt}} = 12$)

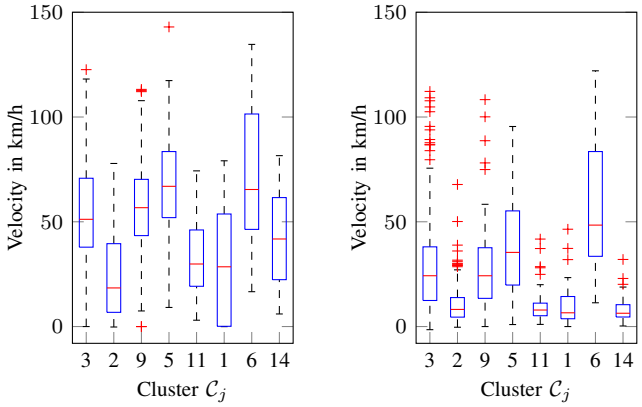
Fig. 4. Visualization of cluster structures in Ω using t-distributed stochastic neighbor embedding (t-SNE) [35] (△: representative)

TABLE VI

8 OF 60 REPRESENTATIVE SCENARIOS IN AKC-BASED ALGORITHM

$\mathcal{C}_j$	m_j	Description
3	311	Subject: Going straight, Vehicle in front: Decelerating while *v_mean_precEv* = 50.83 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 23.39 km/h
2	130	Subject: Going straight, Vehicle in front: Slowed and stopped 2 seconds or less while *v_mean_precEv* = 18.67 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 9.70 km/h
9	83	Subject: Slowing in lane, Vehicle in front: Decelerating while *v_mean_precEv* = 53, 32 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 28.48 km/h
5	68	Subject: Slowing in lane, Vehicle in front: Decelerating while *v_mean_precEv* = 61.45 km/h, Evasive maneuver: braked and steered, No visual obstacles, Traffic density: B*, *v_mean_coll* = 34.55 km/h
11	48	Subject: Slowing in lane, Vehicle in front: slowed and stopped 2 seconds or less while *v_mean_precEv* = 30.67 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 8.02 km/h
1	44	Subject: Going straight, Vehicle in front: stopped on roadway more than 2 seconds while *v_mean_precEv* = 42.66 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 9.36 km/h
6	36	Subject: Going straight, Vehicle in front right: lane change while *v_mean_precEv* = 60.48 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 47.02 km/h
14	36	Subject: Slowing in lane, Vehicle in front: - stopped on roadway more than 2 seconds while *v_mean_precEv* = 39.31 km/h, Evasive maneuver: braked, No visual obstacles, Traffic density: B*, *v_mean_coll* = 4.68 km/h

*Level of service B: Flow with some restrictions



(a) *v_mean_precEv* (b) *v_mean_coll*

Fig. 5. Distribution of metric variables in clusters shown in Table VI

VI. CONCLUSIONS

In order to reduce the test effort during the development of predictive safety functions, representative scenarios are needed, which cover a database at best possible rate. For this purpose, two distance-based clustering methods, i.e., approximate k-covers and partitioning around medoids of k-medoids, were implemented using the mixed similarity measure and centrality measure to quantify the distances of four different levels of measurement and identify the representative in each cluster. As a linkage criterion, the distances between multiple cluster representatives were utilized. The selection of initial medoids was done based on the centrality measure. The algorithm based on the approximate k-covers delivered a higher number of clusters and did not merge outliers with the closest large clusters. The outliers, i.e., rare scenarios, in a database can be relevant for a certain predic-

tive safety function and system. Judging from the average silhouette width, the algorithm based on the approximate k-covers was considered as a better suited approach for the given dataset even though both of the algorithms showed the similar purity measures and representativeness of cluster representatives. However, a broad distribution regarding vehicle velocity values within the same clusters was noticed. In further steps, the influence of selected features, especially metric features, on clustering results is to be inspected by conducting a sensitivity analysis. Besides metric features, the similarity measurement of time series data is also necessary for the distinct scenario description in the future.

ACKNOWLEDGMENT

The research project is financially supported by AUDI AG. Hiroki Watanabe, Tomáš Malý, and Johannes Wallner possess a valid data use license for the SHRP 2 dataset [33]. The findings and conclusions of this paper are those of the authors and do not necessarily represent the views of VTTI, the Transportation Research Board or the National Academies.

REFERENCES

[1] Euro NCAP, *Test Protocol - AEB VRU Systems*, 3rd ed., July 2019.
[2] ——, *Test Protocol - AEB Car-to-Car Systems*, 3rd ed., July 2019.
[3] ——, *Test Protocol - Lane Support Systems*, 3rd ed., July 2019.
[4] G. Prokop and M. Köbe, "Methodology for effectiveness assessment of road safety functions," in *IEDAS - Active safety and automated driving / 3rd interdisciplinary expert dialogue*, Oct. 2017, pp. 134–147.
[5] H. Watanabe, L. Tobisch, T. Laudien, J. Wallner, and G. Prokop, "A method for the estimation of coexisting risk-inducing factors in traffic scenarios," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, June 2019.
[6] H. Watanabe, L. Tobisch, J. Rost, J. Wallner, and G. Prokop, "Scenario mining for development of predictive safety functions," in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, Sept. 2019.
[7] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, May 2005.
[8] J. Lenard, R. Danton, M. Avery, A. Weekes, D. Zuby, and M. Kühn, "Typical pedestrian accident scenarios for the testing of autonomous emergency braking systems," in *22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV)*, June 2011.
[9] J. Lenard, A. Badea-Romero, and R. Danton, "Typical pedestrian accident scenarios for the development of autonomous emergency braking test protocols," *Accident Analysis & Prevention*, vol. 73, pp. 73–80, 2014.
[10] L. Liu, X. Zhu, and Z. Ma, "Study on test scenarios of environment perception system under rear-end collision risk," in *WCX SAE World Congress Experience*. SAE International, Apr. 2018.
[11] X. Ma, Z. Ma, X. Zhu, J. Cao, and F. Yu, "Naturalistic driving behavior analysis under typical normal cut-in scenarios," in *WCX SAE World Congress Experience*. SAE International, Apr. 2019.
[12] D. Nilsson, M. Lindman, T. Victor, and M. Dozza, "Definition of run-off-road crash clusters - for safety benefit estimation and driver assistance development," *Accident Analysis & Prevention*, vol. 113, pp. 97–105, 2018.
[13] U. Sander and N. Lubbe, "The potential of clustering methods to define intersection test scenarios: Assessing real-life performance of AEB," *Accident Analysis & Prevention*, vol. 113, pp. 1–11, 2018.
[14] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999.
[15] P. Nitsche, P. Thomas, R. Stuetz, and R. Welsh, "Pre-crash scenarios at road junctions: A clustering method for car crash data," *Accident Analysis & Prevention*, vol. 107, pp. 137–151, 2017.
[16] P. Nitsche, R. H. Welsh, A. Genser, and P. D. Thomas, "A novel, modular validation framework for collision avoidance of automated vehicles at road junctions," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, Nov 2018, pp. 90–97.
[17] R. T. Ng and J. Han, "CLARANS: a method for clustering objects for spatial data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5, pp. 1003–1016, 2002.
[18] S. Kaplan and C. G. Prato, "Cyclist-motorist crash patterns in denmark: A latent class clustering approach," *Traffic Injury Prevention*, vol. 14, no. 7, pp. 725–733, 2013.
[19] J. de Oña, G. López, R. Mujalli, and F. J. Calvo, "Analysis of traffic accidents on rural highways using latent class clustering and bayesian networks," *Accident Analysis & Prevention*, vol. 51, pp. 1–10, 2013.
[20] T. Helmer, *Development of a Methodology for the Evaluation of Active Safety using the Example of Preventive Pedestrian Protection*, ser. Springer Theses. Springer International Publishing Switzerland, 2015.
[21] H. Johannsen, *Unfallmechanik und Unfallrekonstruktion - Grundlagen der Unfallaufklärung*, 3rd ed. Wiesbaden: Springer Vieweg, 2013.
[22] N. Andricevic, "Robustheitsbewertung crashbelasteter Fahrzeugstrukturen," Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg im Breisgau, 2016.
[23] J. M. Hankey, M. A. Perez, and J. A. McClafferty, "Description of the SHRP 2 naturalistic database and the crash, near-crash, and baseline data sets," Virginia Tech Transportation Institute, Tech. Rep., Apr. 2016.
[24] G. W. Milligan and M. C. Cooper, "Methodology review: Clustering methods," *Applied Psychological Measurement*, vol. 11, no. 4, pp. 329–354, 1987.
[25] G. Bagschik, T. Menzel, and M. Maurer, "Ontology based scene creation for the development of automated vehicles," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, June 2018, pp. 1813–1820.
[26] D. S. Ali, A. Ghoneim, and M. Saleh, "Data clustering method based on mixed similarity measures," in *6th International Conference on Operations Research and Enterprise Systems (ICORES)*, Feb. 2017, pp. 192–199.
[27] S. Aranganayagi and K. Thangavel, "Improved k-modes for categorical clustering using weighted dissimilarity measure," *International Journal of Computer and Information Engineering*, vol. 3, no. 3, 2009.
[28] A. Watve, S. Pramanik, S. Jung, B. Jo, S. Kumar, and S. Sural, "Clustering non-ordered discrete data," *Journal of Information Science and Engineering*, vol. 30, no. 1, pp. 1–23, 2014.
[29] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, New Jersey: John Wiley & Sons, 1990.
[30] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3336–3341, 2009.
[31] E. Schubert and P. J. Rousseeuw, "Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms," in *12th International Conference on Similarity Search and Applications (SISAP)*, Oct. 2019.
[32] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
[33] C. Witcher, J. Wallner, S. Engel, D. Wittmann, P. Feig, A. Schneider, and M. A. Perez. (2016) Predictive Safety - Analysis, whether statistical data can be used for driving safety purposes. [Online]. Available: https://doi.org/10.15787/VTT1/O5TEIX [Accessed: Nov. 1, 2018].
[34] *SHRP 2 Researcher Dictionary for Video Reduction Data*, 3rd ed., Virginia Tech Transportation Institute, Blacksburg, Virginia, Feb. 2015.
[35] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.
[36] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.
[37] Y. Zhao and G. Karypis, "Criterion functions for document clustering: Experiments and analysis," University of Minnesota, Department of Computer Science / Army HPC Research Center, Tech. Rep. 01-40, Feb. 2001.
[38] E. Aldana-Bobadilla and A. Kuri-Morales, "A clustering method based on the maximum entropy principle," *Entropy*, vol. 17, no. 1, pp. 151–180, 2015.
[39] C. A. Ratanamahatana and E. J. Keogh, "Everything you know about dynamic time warping is wrong," in *Third Workshop on Mining Temporal and Sequential Data*, 2004.
[40] S. Dolnicar, "A review of unquestioned standards in using cluster analysis for data-driven market segmentation," in *Australian and New Zealand Marketing Academy Conference (ANZMAC)*, Dec. 2002.