

# A Statistical and Machine Learning Approach to Assess Contextual Complexity of the Driving Environment Using Autonomous Vehicle Data

Final Report

by

Dr. Jennifer Ogle, Ph.D., Clemson University  
Phone: (864) 656-0883  
E-mail: ogle@clemson.edu

Vijay Bendigeri (Clemson University)  
Fengjiao Zou (Clemson University)  
Ahmad Zaki Ghafari (Clemson University)  
Gurcan Comert (Benedict College)

May 2024



Center for Connected Multimodal Mobility (C<sup>2</sup>M<sup>2</sup>)



Benedict College



THE CITADEL  
THE MILITARY COLLEGE OF SOUTH CAROLINA

SCState  
UNIVERSITY



UNIVERSITY OF  
SOUTH CAROLINA

200 Lowry Hall, Clemson University  
Clemson, SC 29634

## DISCLAIMER

*The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by the Center for Connected Multimodal Mobility (C<sup>2</sup>M<sup>2</sup>) (Tier 1 University Transportation Center) Grant, which is headquartered at Clemson University, Clemson, South Carolina, USA, from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.*

Non-exclusive rights are retained by the U.S. DOT.

## ACKNOWLEDGMENT

*The authors would like to acknowledge the U.S. Department of Transportation (USDOT) Center for Connected Multimodal Mobility (C2M2) for the funding to make this research possible. C2M2 is a Tier 1 University Transportation Center headquartered at Clemson University, Clemson, South Carolina, USA.*

**TECHNICAL REPORT DOCUMENTATION PAGE**

1. Report No.		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle Assessment of Contextual Complexity and Risk Using Unsupervised Clustering Approaches with Dynamic Traffic Condition Data Obtained from Autonomous Vehicles				5. Report Date May 2024	
				6. Performing Organization Code	
7. Author(s) Jennifer Ogle, Ph.D.; ORCID: 0000-0003-0521-3104 Vijay Bendigeri Ph.D; ORCID: 0000-0002-5460-0138 Fengjiao Zou; ORCID: 0000-0002-1857-8826 Ahmad Zaki Ghafari; ORCID: 0009-0005-8046-8416 Gurcan Comert, Ph.D.; ORCID: 0000-0002-2373-5013				8. Performing Organization Report No.	
9. Performing Organization Name and Address Glenn Department of Civil Engineering Clemson University 210 Lowry Hall, Clemson, SC 29634				10. Work Unit No.	
				11. Contract or Grant No. 69A3551747117	
12. Sponsoring Agency Name and Address Center for Connected Multimodal Mobility (C <sup>2</sup> M <sup>2</sup> ) Clemson University 200 Lowry Hall, Clemson, SC 29634				13. Type of Report and Period Covered Final Report (August 2021 - May 2024)	
				14. Sponsoring Agency Code	
15. Supplementary Notes					
16. Abstract Traditional road safety assessment methodologies rely heavily on AADT (Annual Average Daily Traffic) data estimates to account for variability in traffic operations. Unfortunately, this approach does not consider the driving environment's fast-changing dynamics, which can influence contextual complexity and risk. This research report presents a method to measure and quantify the contextual complexity of the roadway environment using diverse open-source LiDAR (Light Detection and Ranging) sensor data collected by Waymo autonomous vehicles under dynamic traffic conditions. The proposed Contextual Complexity Factor (CCF) model estimates the driving scene's complexity using the density and proximity of the objects around the vehicle. Besides, an unsupervised machine learning technique using clustering algorithms was used to measure and classify the driving environment's dynamic characteristics (e.g., vehicles, pedestrians, bicycles) into appropriate risk categories to develop a dynamic complexity model. Variables, including velocity, object density of lidar, and object proximity, were selected for k-means and hierarchical clustering analysis. Three clusters were ultimately chosen that categorize the scene into high, medium, and low categories of complexity. Adopting the results from the clustering analysis, the research team further built the complexity ranges for the attributes (i.e., velocity, object density, and object proximity). Both statistical and machine learning models were proficient in predicting the dynamic complexity with justifiable truthfulness. Identifying and predicting high-risk environments in real-time can significantly benefit safety research, driver education, auto-insurance risk assessment, autonomous vehicle route planning, and many more.					
17. Keywords Dynamic Complexity, Autonomous Vehicles, Waymo Open Dataset, Unsupervised Clustering, Machine Learning, Statistical learning, traffic safety.				18. Distribution Statement No restrictions	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 33	22. Price NA

## Table of Contents

DISCLAIMER.....	ii
ACKNOWLEDGMENT.....	iii
TECHNICAL REPORT DOCUMENTATION PAGE.....	iv
LIST OF TABLES.....	vi
LIST OF FIGURES.....	vi
EXECUTIVE SUMMARY.....	1
CHAPTER 1.....	3
Introduction and Background.....	3
1.1 Introduction and Background.....	3
1.2 Projects Highlights and Impacts.....	4
1.3 Research Goals and Objectives.....	4
CHAPTER 2.....	5
Data.....	5
2.1 Data Source.....	5
2.2 Extract Transform Load LiDAR Data.....	7
2.3 Feature Engineering and Transformation.....	7
CHAPTER 3.....	9
Method.....	9
3.1 Contextual Complexity Factor Model.....	9
3.2 Unsupervised clustering analysis.....	9
3.3 Dynamic Complexity Factor Rating.....	11
CHAPTER 4.....	12
Results.....	12
4.1 Statistical Modeling Approach.....	12
4.2 Machine Learning Approach.....	15
4.2.1 Feature selection.....	18
4.2.2 Clustering Analysis.....	20
4.2.3 Cluster Centers and Corresponding Dynamic Complexity.....	22
4.2.4 Dynamic ranges of attributes for complexity categorization.....	23
CHAPTER 5.....	25
Conclusions.....	25
REFERENCES.....	27

## LIST OF TABLES

Table 1 Clustering Techniques Descriptions .....	10
Table 2 Variables extracted after processing the AV data.....	12
Table 3 Critical variables and their complexity class ranges. ....	14
Table 4 Data size at different aggregation distances .....	18
Table 5 PCA analysis results .....	19
Table 6 Correlation coefficients of variables .....	19
Table 7 Rand Index for k-means and hierarchical clustering.....	22
Table 8 Cluster group characteristics and their complexity rank .....	23
Table 9 Attributes and their dynamic complexity ranges .....	23
Table 10 Dynamic complexity ranges for attributes.....	23

## LIST OF FIGURES

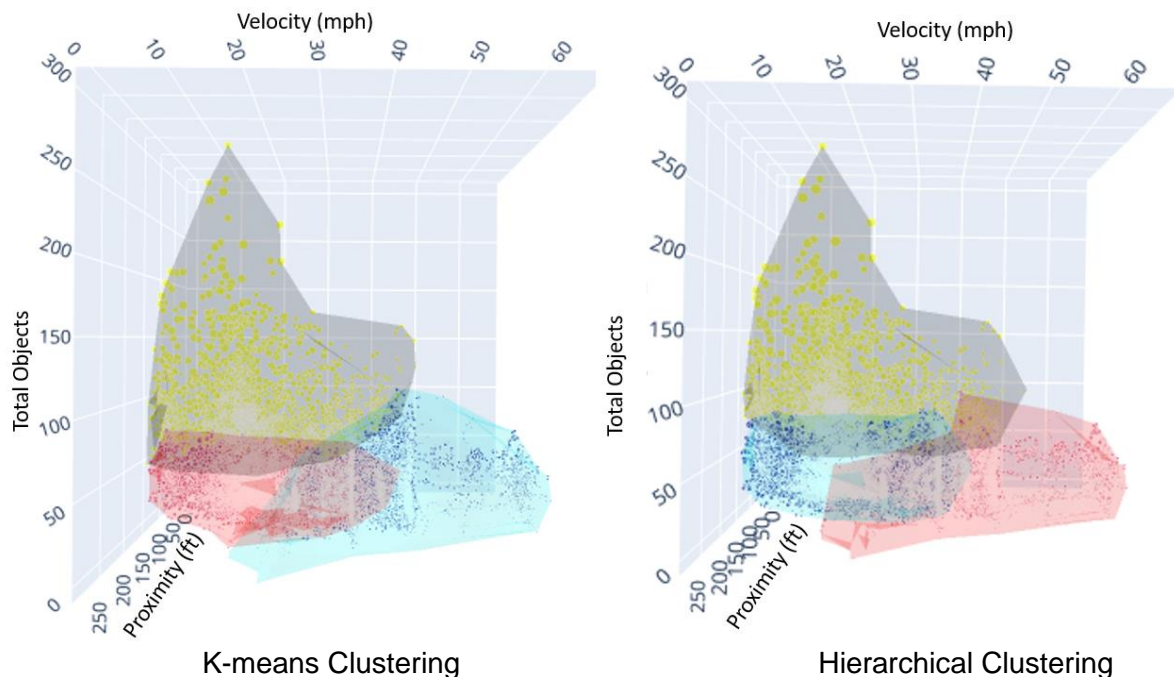
Figure 1 Sensor layout and coordinate system of Waymo autonomous vehicle .....	5
Figure 2 LiDAR 3D bounding box example, Yellow = vehicle, Red=Pedestrian, Blue=sign, Pink = cyclist .....	6
Figure 3 Waymo self-driving car data collection areas.....	6
Figure 4 Data Preprocessing Flowchart.....	6
Figure 5 LiDAR point-cloud, SSD, UFOV, and COV representation with object types.....	7
Figure 6 Method Flowchart .....	11
Figure 7 Statistical distributions of critical variables before the clipping .....	13
Figure 8 Statistical distributions of critical variables after the clipping the frames with speeds greater than 35 mph and less than 0.1 mph.....	13
Figure 9 CCF plots for high, medium, and low-complexity trips (velocity >0.1 mph and <= 35mph).....	14
Figure 10 Histogram of different attributes from Waymo AV Data.....	15
Figure 11 Density plots of different attributes from the AV data .....	16
Figure 12 Normality plots for different aggregate distances .....	17
Figure 13 Density plots for different aggregation distances.....	17
Figure 14 Data size at different aggregation distances .....	18
Figure 15 Distortion Plots from Clustering Analysis .....	20
Figure 16 K-means Vs. Hierarchical Clustering.....	21
Figure 17 K-means vs. Hierarchical clustering - distribution of points .....	21

## EXECUTIVE SUMMARY

Traditional road safety assessment methodologies do not recognize the driving environment's fast-changing dynamics that influence the contextual complexity and, ultimately, accident risk. The advent of autonomous vehicle (AV) open datasets has created new opportunities to measure dynamic complexity and incorporate dynamic interaction metrics into risk estimates and safety assessments.

This research presents a method to measure and quantify the contextual complexity of the roadway environment using diverse open-source LiDAR (Light Detection and Ranging) sensor data, collected by Waymo autonomous vehicles under dynamic traffic conditions. First, a statistical approach (Contextual Complexity Factor Model) was developed during the project. A total of 798 perception data trips, comprising 158,090 LiDAR point cloud frames, were analyzed to develop the Contextual Complexity Factor (CCF) model to measure dynamic complexity. The numerical analysis provided a frame-by-frame comparison of contextual complexity based on the density of objects and their proximity to the autonomous vehicle as represented by the CCF. All trips were categorized as high, medium, or low-complexity trips based on the statistical model of the trip's CCF category.

The statistical modeling approach satisfactorily represents the contextual complexity of the driving environment. However, one impediment of the methodology is that the quartiles do not paint the picture with sufficient granularity. A machine-learning approach using an unsupervised clustering method was tested to overcome this. Specifically, k-means and hierarchical clustering algorithms were used. It's worth noting that the data for different variables are not uniformly represented after a density plot. Since most machine learning algorithms developed for classification were designed to assume close number of samples for each class. The research team considered data normalization. Further, the research team performed principal component analysis (PCA) to identify the most important (impacting) variables. From PCA and correlation results, variables including *object velocity*, *object density of lidar*, and *object proximity* were selected for clustering analysis.



**Figure 1 K-means vs. Hierarchical Clustering**

Figure 1 above shows clustering results for k-means and hierarchical clustering methods for three cluster centers. The cluster groups are labeled zero, one, and two. Velocity is on the x-axis, object density is on the y-axis, and mean proximity is on the z-axis. From the figure, k-means and hierarchical clustering results look identical. The k-means clustering boundaries look smoother compared to the hierarchical clustering boundaries. The edges are sharper in the case of hierarchical clustering. Understanding the parameters of the cluster grouping is essential for assigning a contextual complexity; the authors chose three cluster center models because it would be easier to categorize into three distinct categories: high, medium, and low complexity. Cluster group zero includes locations with low velocity and low density of objects compared to the other two groups, representing a low-complexity environment. Cluster group one includes locations with relatively high velocity, low-medium object density, and low-to-high proximity of objects, representing a "medium-complexity" environment. Cluster group two includes areas with high object density and proximity. They might represent locations in central business districts with increased activity. Compared to the other two groupings, these locations present a relatively complicated driving context. Thus, cluster two represents areas with a "high-complexity" environment. Adopting the results from the clustering analysis, the authors further built the complexity ranges for the attributes (i.e., velocity, object density, and object proximity). Table 1 summarizes the complexity ranges into low, medium, and high obtained from clustering.

**Table 1 Dynamic complexity ranges for attributes.**

Dynamic Complexity	Velocity (mph)	Object count	Object Proximity (feet)
Low	0-28	0-37	0-34
Medium	0-66	0-71	39-189
High	128-209	101-172	115-182

Identifying and predicting high-risk environments in real-time can significantly benefit safety research, driver education, auto-insurance risk assessment, autonomous vehicle route planning, and many more. For example, this research can allow Driving Rehabilitation Specialists (DRSs) to score the dynamic complexity during training and testing to ensure that the driver is competent at all situational levels. The methodology that this project developed utilizing the autonomous vehicle open datasets can aid DRSs in measuring and classifying the contextual complexity of the routes used for on-road driving evaluations for medically-at-risk drivers considering the dynamic variables. The on-road driving evaluation is considered the gold standard for testing and rehabilitating medically at-risk drivers. The product of this research can lay a foundational work to build tools and methodology to measure the roadway context to enhance the consistency and validity of the on-road assessment procedures<sup>1</sup>.

<sup>1</sup>Note that some contents of this report have been published as a Ph.D. Dissertation and in the ASCE International Conference on Computing in Civil Engineering 2021. Here are the citations:  
 Bendigeri, Vijay, "Using Safety Performance Models, Autonomous Vehicle Data, and Machine Learning to Develop Contextual Complexity Criteria to Establish a Standardized Process for On-Road Evaluation of Medically At-Risk Drivers Considering Static and Dynamic Factors of the Roadway Environment" (2022). Ph.D. Dissertation, Clemson University, U.S.A., 2983  
 Bendigeri, V. G., Zou, F., Ogle, J. H., & Kusram, K. Roadway Contextual Risk Assessment Using Dynamic Traffic Conditions Data Obtained from Autonomous Vehicles. In *Computing in Civil Engineering 2021* (pp. 562-569). DOI: <https://doi.org/10.1061/9780784483893.070>



## CHAPTER 1

### Introduction and Background

#### 1.1 Introduction and Background

---

Traditional road safety assessment methodologies rely heavily on historical crash data, static roadway characteristics, and AADT (Annual Average Daily Traffic) data estimates to account for variability in traffic operations. Unfortunately, this approach does not consider the driving environment's fast-changing dynamics (i.e., fast-changing interactions with vehicles, pedestrians, and bicyclists), which can influence contextual complexity and risk. Researchers have conducted case-control studies in the past by returning to crash sites at the same time of day, day of the week, and under similar weather conditions to try to ascertain dynamic operating conditions. The observed conditions are used as a surrogate for the dynamic operating conditions, but still, they could be vastly different from the actual time of the crash. The advent of the autonomous vehicle open datasets has created new opportunities to measure dynamic complexity and incorporate dynamic interaction metrics into risk estimates and safety assessments. Identifying and predicting high-risk environments in real-time can significantly benefit safety research, driver education, auto-insurance risk assessment, autonomous vehicle route planning, and many more. For example, this research could allow driving instructors and rehab specialists to score the dynamic complexity during training and testing to ensure that the driver is competent at all situational levels. Another example is route planning - current autonomous vehicle route planning strategies do not consider scene complexity, making it more challenging for drivers to take control of autonomous vehicles when needed (Bendigari et al., 2022).

Everyday routine trips expose drivers to massive amounts of input that is either static (i.e., roadway configuration and traffic control devices) or dynamic (i.e., movement of surrounding vehicles and other vulnerable road users) (Olson & Farber, 1996). An important concept related to driver information processing is the useful field of view (UFOV). The UFOV is defined as "the total visual field from which target characteristics can be acquired when the head and eye movements are excluded,"; and the extent of the UFOV differs between drivers, depending on how well they select and process relevant information from the environment (Dewar & Olson, 2002). While drivers may scan the whole driving environment while driving, the focus is the view in front, the UFOV. Researchers define all the information that a driver must process to operate a vehicle as the visual demand, including traffic on the road, roadway environment, information in the vehicle, and other inputs (Dewar & Olson, 2002). Human factors experts generally believe that the risk of traffic crashes increases when the visual demand increases (Dewar & Olson, 2002). Prior research determined that more crashes occur on roads with heavy traffic or complicated geometric configurations (Shinar, McDowell, & Rockwell, 1977). Abdel-Aty and Radwan modeled crash occurrence and involvement and found that heavy traffic increases the likelihood of crashes (Abdel-Aty & Radwan, 2000). Crashes increase with the traffic complexity or object density because the driver's cognitive load increases. Cognitive load is believed to be vital in performing complex tasks (Paas, Tuovinen, Tabbers, & Van Gerven, 2003), such as driving. Finally, variation in the speed of the dynamic inputs adds yet another level of complexity that the driver must process. Researchers (Choudhary et al., 2018) have determined that crash rates increase as the speed variations between drivers and other traffic increase, especially at higher traffic volumes. Yet, methods to incorporate complexity into our risk assessments are not currently available.

Complicating matters, UFOV decreases with increases in driver age, vehicle speed, traffic congestion, rain, and any other high-demand tasks (Dewar & Olson, 2002; Rogé et al., 2004).

Researchers estimated that when the drivers are traveling at 30 mph, they can see targets in a visual field of 150 degrees; however, when speed is doubled (60 mph), drivers can only see targets in half of the visual field (approximately 75 degrees) (Dewar & Olson, 2002). As speeds increase, the distance required to perceive hazards and react appropriately increases because drivers need to look further down the road for objects in the potential collision zone - referred to as the stopping sight distance (SSD) (AASHTO, 2018). As the UFOV narrows with speed, it also expands in length due to increased SSD. Research also reveals that the UFOV decreases when the quantity of information processed in the driver's peripheral area increases (Mackworth, 1976), meaning that the level of object density is high and the road scene is complex. As prior research (Choudhary et al., 2018) suggests, specific combinations of static and dynamic parameters increase the likelihood of crash occurrence, not their individual effects.

This research presents a method to measure and quantify the contextual complexity of the roadway environment using diverse open-source LiDAR (Light Detection and Ranging) sensor data collected by Waymo autonomous vehicles under dynamic traffic conditions. An unsupervised machine learning technique using clustering algorithms was used to measure and classify the driving environment's dynamic characteristics (e.g., vehicles, pedestrians, bicycles) into appropriate risk categories to develop a dynamic complexity model. This study proposed a contextual complexity factor (CCF) model that estimated the driving scene's complexity using the density and proximity of the objects around the vehicle (Bendigeri, 2022).

## 1.2 Projects Highlights and Impacts

---

- Traditional road safety assessment methodologies do not recognize the driving environment's fast-changing dynamics that influence the contextual complexity and, ultimately, its risk.
- This research uses diverse open-source sensor data (LiDAR) collected by Waymo autonomous vehicles to estimate the road environment's complexity considering dynamic traffic conditions.
- The proposed machine learning-based contextual complexity factor (CCF) model estimates the driving scene's complexity using the speed, density, and proximity of the objects around the vehicle and classifies high, medium, and low contextual risk categories.
- Identified high-risk environments can significantly benefit safety research, driver education, auto-insurance risk assessment, autonomous vehicle route planning, and many more.

## 1.3 Research Goals and Objectives

---

Goal: Understand the dynamic scene complexity from a driver's perspective and develop a contextual complexity factor (CCF) model using unsupervised clustering that classifies the driving environment's complexity.

Objectives:

- Measure contextual complexity and risk considering the dynamic components of the driving environment.
- Utilize data-rich LiDAR data collected by Waymo autonomous vehicles to reflect dynamic aspects of the environment.
- Apply unsupervised clustering methods to estimate the road environment's dynamic complexity.

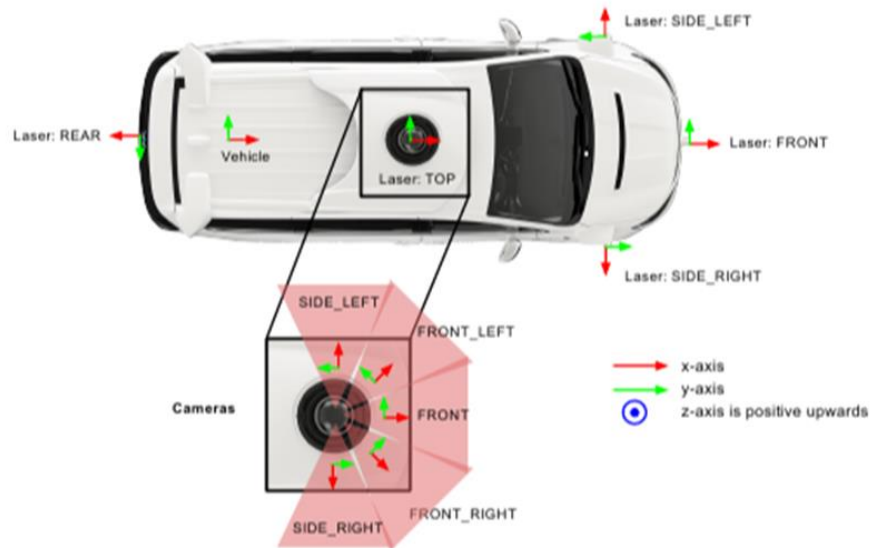
## CHAPTER 2

### Data

#### 2.1 Data Source

The advent of autonomous vehicle open datasets has created new opportunities to measure dynamic complexity to incorporate dynamic interaction metrics into complexity estimates and safety assessments. Several autonomous vehicle datasets have been published in recent years; however, the open dataset published by Waymo in 2019 is the largest, richest, and most diverse self-driving dataset released for research (Waymo, 2019). This chapter describes the data and preprocessing steps.

The raw dataset consists of high-quality LiDAR and video data obtained from multiple sensors mounted on Waymo autonomous vehicles. **Error! Reference source not found.** shows a picture of the Waymo autonomous vehicle with its sensor layout and the relative coordinate systems. The system used five Lidar sensors and five high-resolution pinhole cameras (Sun et al., 2020).



**Figure 1 Sensor layout and coordinate system of Waymo autonomous vehicle (Sun et al., 2020)**

The coordinate system moves with the vehicle with the origin set to the direction of movement. The LiDAR dataset included 3D bounding boxes with object type annotations manually checked for accuracy by trained labelers, see Figure 2 (Sun et al., 2020). The tracked object types include vehicles, pedestrians, bicyclists, and traffic signs. Additionally, vehicle speed vectors in 3-dimensional space for each frame were provided. The data were collected in San Francisco, Phoenix, and Mountain View, see Figure 3 (Sun et al., 2020).

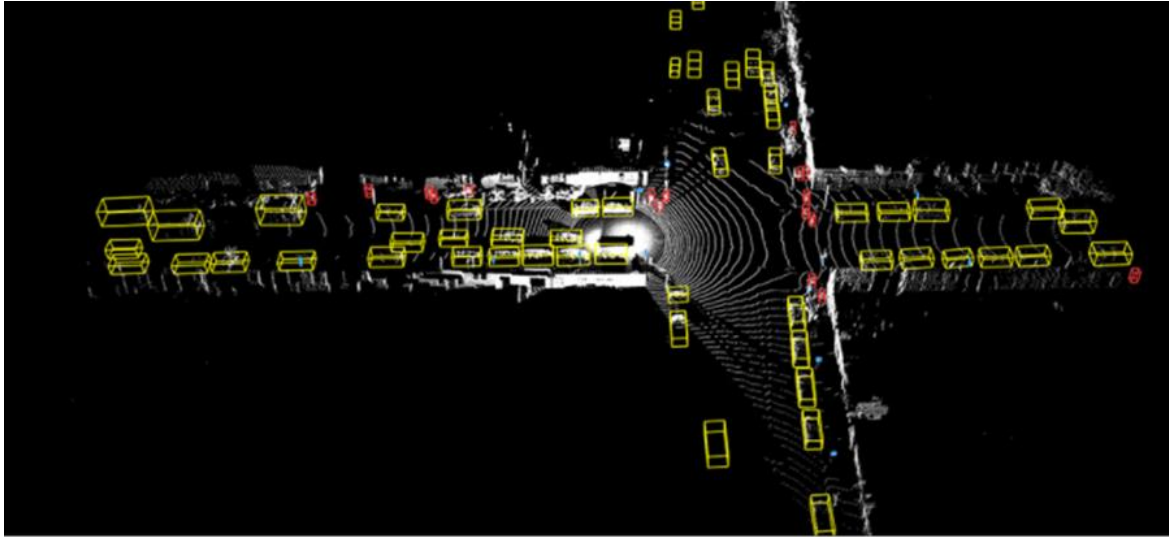


Figure 2 LiDAR 3D bounding box example, Yellow = vehicle, Red=Pedestrian, Blue=sign, Pink = cyclist (Waymo, 2019)



Figure 3 Waymo self-driving car data collection areas (Sun et al., 2020)

Figure 4 shows the workflow with the main tasks and associated sub-tasks in data preprocessing. Work associated with each task is discussed later in this section.

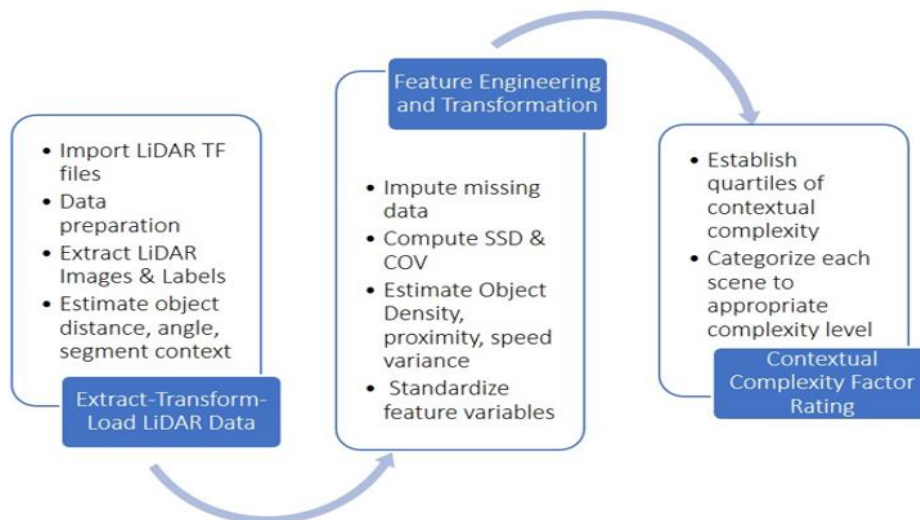


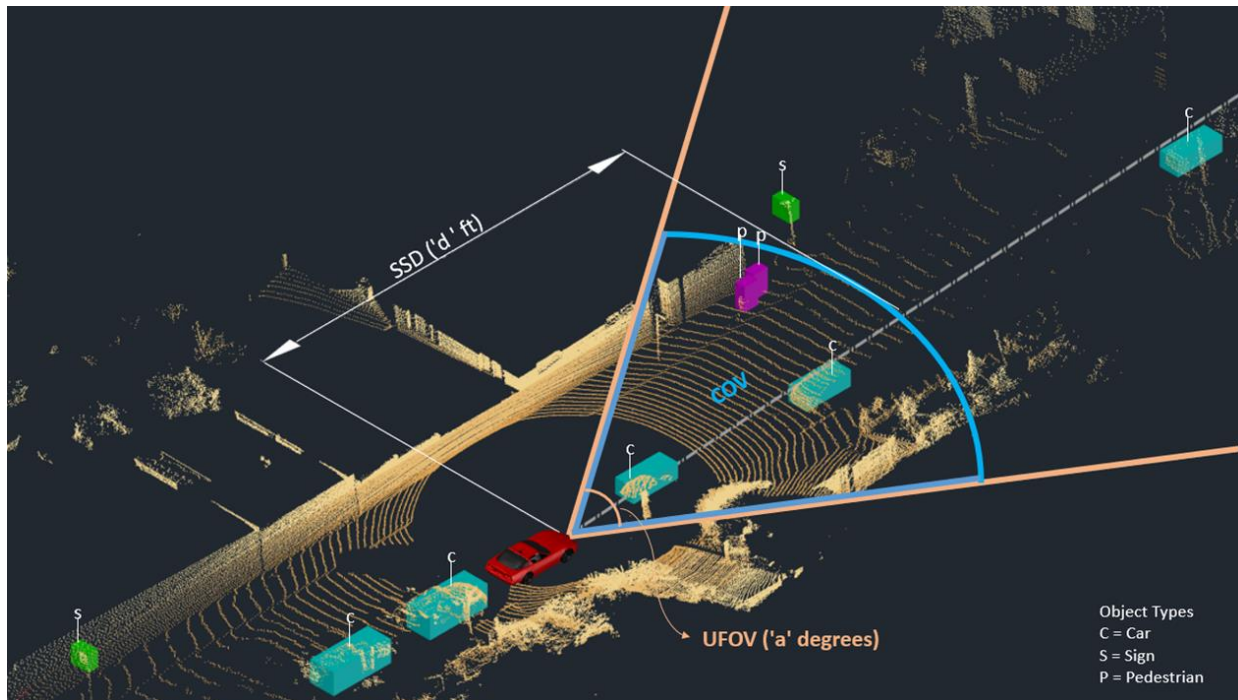
Figure 4 Data Preprocessing Flowchart

## 2.2 Extract Transform Load LiDAR Data

A total of 798 scenes of perception data, each spanning 20 seconds at 10 Hz/second (i.e., ~200 LiDAR frames), were analyzed during this phase. All scenes were stored in a Google cloud bucket in a TensorFlow file format. The files were downloaded, and the raw LiDAR data were extracted. The raw data contains a segment context, LiDAR images, and LiDAR labels. The LiDAR point-cloud data reference each object with x, y, and z coordinates in a three-dimensional space with respect to the autonomous vehicle's origin. The objects' distance and angle from the autonomous vehicle are estimated from x, y, and z coordinates. Vehicle speed for each frame was obtained from the segment context metadata. The driving context assessment was limited to lower speeds (<40 mph) due to the limited range of LiDAR technology, which has a published maximum range of 250 feet, though extended distances were contained in the datasets (Waymo 2019).

## 2.3 Feature Engineering and Transformation

The total number of objects in each LiDAR frame and their proximity to the driver was estimated as a measure of scene complexity. From the literature review, an important concept related to driver information processing is the useful field of view (UFOV) (Dewar & Olson, 2002). As speeds increase, the distance required to perceive hazards and react appropriately increases because drivers need to look further down the road for objects in the potential collision zone or the stopping sight distance (SSD) (AASHTO, 2018). The vehicle's speed was used to derive SSD and select an appropriate UFOV. The SSD and UFOV were then used to construct a 3-dimensional filter cone, the cone of vision (COV), to identify any objects within that cone.



**Figure 5 LiDAR point-cloud, SSD, UFOV, and COV representation with object types**

Figure 5 provides a pictorial representation of the SSD, UFOV, and COV in the LiDAR point cloud. The orange dots represent the LiDAR points. The red car in the center is a representation of the autonomous vehicle. In Figure 5, the blue bounding boxes with the label "c" are the locations of detected cars. The pink bounding boxes with the label "p" represent the location of the pedestrians.

Traffic signs are the green bounding boxes with label "s". And the white dotted line in the center represents the direction of travel of the autonomous vehicle. The UFOV is the angle "a" between the orange lines that extends from the red car. SSD in Figure 5 is a distance that extends from the red car to a distance "d." COV is the 3-dimensional volume of space constructed from SSD & UFOV represented by the blue boundary. Objects within this COV were identified for each frame, along with the total objects in the scene. In Figure 5, there are nine objects in the scene (five cars, two pedestrians, and two signs) and only four objects within the COV (two cars and two pedestrians).

COV is a function of SSD and UFOV. Thus, the SSD and UFOV were first computed to determine the COV for each frame. The SSD of the vehicle for each frame was calculated using Equation 2.1. A standard driver's reaction time of 2.5 seconds (Rogé et.al., 2004) and a flat grade (i.e., grade = 0%) were assumed in all the SSD estimations. UFOV shares an exponential relationship with speed. Following Dewar and Olson (2002), the UFOV is 160 degrees at zero speed, which reduces to 150 degrees at 30 mph speed and further scales down to 75 degrees at 60 mph speed. UFOV was computed using linear interpolation for all fractional speeds that fall in between the speed ranges mentioned above (i.e., 0 mph, 30 mph, and 60 mph) within each scene. The COV boundary was calculated within the LiDAR point cloud using SSD & COV values. Objects within the COV boundary were summarized along with the total objects in the scene.

$$SSD = 1.47st + \frac{s^2}{30 \cdot \frac{a}{g}} \quad \text{Equation 2.1}$$

Where,

*SSD* = Total stopping sight distance for the vehicles (feet)

*s* = speed of the vehicle (mph)

*t* = standard reaction time of the driver (2.5 seconds)

*a* = standard deceleration rate (11.2 ft/s<sup>2</sup>)

*g* = acceleration due to gravity (32.2 ft/s<sup>2</sup>)

The next chapter discusses how the data extracted from the LiDAR data was used to build the models to estimate contextual complexity.

## CHAPTER 3

### Method

This research study sought to create a methodology to measure and quantify the contextual complexity of the driving environment using diverse open-source LiDAR sensor data collected by Waymo autonomous vehicles under dynamic traffic conditions. This chapter details the methods designed to achieve this research objective. Specifically, the methodology discusses a statistical approach (Contextual Complexity Factor Model) and a machine learning approach (Unsupervised clustering analysis).

#### 3.1 Contextual Complexity Factor Model

---

From the literature review, the key variables that measure cognitive load are the density of the objects and their proximity to the vehicle. As the number of objects in the driving environment increases, the amount of information that needs to be processed by the driver also increases, and so does the driver's cognitive load. Near objects present a greater risk to the driver compared to distant objects. A Contextual Complexity Factor (CCF) was estimated for each frame to measure these two important parameters using Equation 3.1.

$$CCF = \Sigma \left( \frac{1}{obj_{dist}} \right) - \text{Equation 3.1}$$

Where,

$Obj_{dist}$  = distance of the object from the autonomous vehicle (feet)

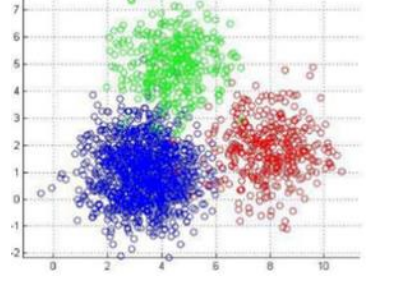
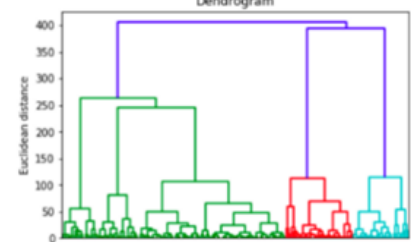
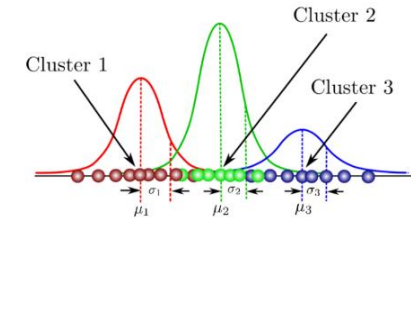
Inverse distance assignments were weighted in descending order, with near objects getting higher and farther objects getting lower weights. The summation of these inverse distances accounted for the total number of objects in the scene, i.e., object density. The scene CCF was estimated for each frame considering all the objects. Additionally, the CCF was estimated from the COV filter in each frame. Statistical quartiles for the total sample size were calculated for the whole scene CCF and CCF within the COV. An individual frame was categorized as high if the CCF > 75th percentile, medium if CCF was within the interquartile range (between the 25th percentile and 75th percentile), and low if the CCF was less than the 25th percentile, respectively. All the frames were assigned a high, medium, or low category based on the scene CCF's respective quartile range.

#### 3.2 Unsupervised clustering analysis

---

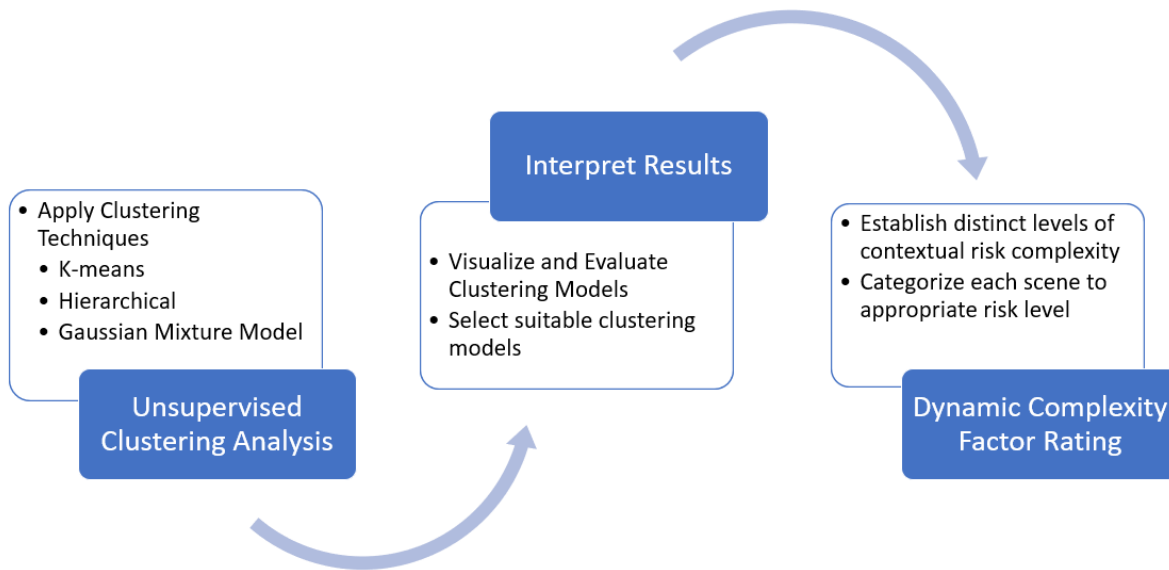
The complexity of multiple variable analysis required a more sophisticated approach to analysis. Thus, unsupervised clustering analysis was carried out to identify natural clusters and identify acceptable boundaries that are impossible by dividing the entire sample size into quartiles. Therefore, clustering analysis was performed on the processed autonomous vehicle data to overcome this design deficiency and get precise. Specifically, k-means, hierarchical, and the Gaussian mixture model clustering algorithms were used to build models. These clustering methodologies have been used for various pattern recognition modeling, such as traffic condition recognition, driver classification, and air pollution hotspot recognition, among others (Montazeri-Gh & Fotouhi, 2011); Govender & Sivakumar, 2020); (Briand, Côme, Mohamed, & Oukhellou, 2016). Table 1 provides a brief description of these algorithms.

**Table 1 Clustering Techniques Descriptions**

Clustering	Method Description	Cluster Representation
K-means clustering (Kanungo et al., 2002)	A simple and effective method of classifying the data into a certain number of clusters. The number of clusters is determined by the value "k." Each point is assigned to the nearest cluster. Different cluster numbers (K) can be applied to classify the scene complexity accurately and choose an optimal number of groups. K-means clustering is used to rank high-crime areas and identify spam emails.	
Hierarchical clustering (Murtagh & Contreras, 2017)	Build a clustering tree by grouping data points closest to each other and further grouping those clusters creating a hierarchy. Hierarchical clustering does not need the specification of several clusters. The number of clusters best fit the data can be chosen by visualizing the tree.	
Gaussian Mixture Model clustering (Liu, Cai, & He, 2010)	This clustering algorithm assumes that the data points are normally distributed. The mean and standard deviation describe the shape of the clusters. The picture shows an example of three gaussian distributions with different mean and standard deviations indicating three distinct clusters. Gaussian mixture model clustering is used in predicting maintenance, classifying handwritten numbers, etc.	

Note: The three figures are from the website: <https://www.mathworks.com/matlabcentral/fileexchange/24616-kmeans-clustering>; <https://www.kdnuggets.com/2019/09/hierarchical-clustering.html>; <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>





**Figure 6 Method Flowchart**

Figure 6 shows the process flowchart to estimate dynamic complexity factor rating. Different clustering techniques were analyzed and compared, including reviewing these results and comparisons and selecting the best fit model. The task consisted of generating visuals of clustering results, recognizing the clustering patterns, and concluding on a choice model. A multi-dimensional cluster visualization tool was created to understand the clusters' boundary division better.

### 3.3 Dynamic Complexity Factor Rating

---

The last step involved identifying the optimal number of cluster classes and classifying each LiDAR scene into the appropriate category. A list of the most influential variables affecting dynamic complexity and their cluster ranges was also identified. The results from the multi-dimensional cluster visuals were converted into a two-dimensional table with ranges defined for each cluster variable to classify a dynamic environment into appropriate complexity categories.

## CHAPTER 4

### Results

The discussion in this section is divided into two parts. The first part presents the analysis and results of the model development using a statistical approach (i.e., Section 3.1 Contextual Complexity Factor Model). The second part explains model development results using an unsupervised clustering approach (i.e., Section 3.2 unsupervised clustering analysis).

#### 4.1 Statistical Modeling Approach

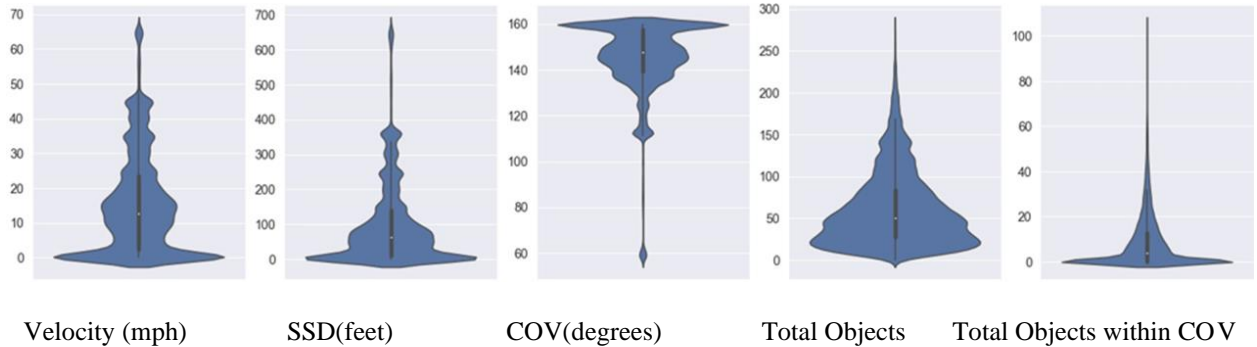
A total of 798 perception data trips, comprising 158,090 LiDAR point cloud frames, were analyzed to develop the contextual complexity factor (CCF) model to measure dynamic complexity. Table 2 lists all the variables available after processing the raw autonomous vehicle data. The first column of Table 2 includes the variable's name, the second column describes the variable, and the third column provides information on the variables derived from one or more combinations of raw variables.

**Table 2 Variables extracted after processing the autonomous vehicle data**

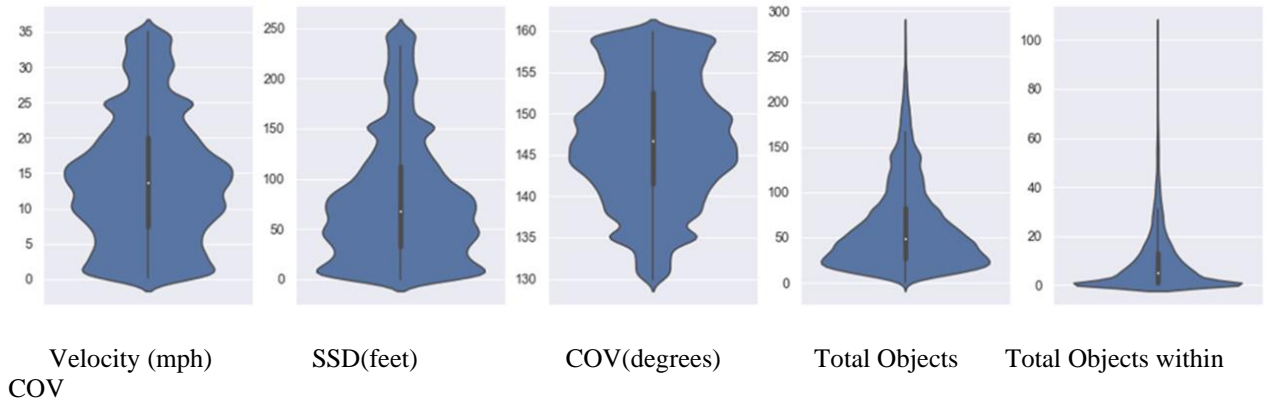
Variable Name	Description	Derived
file_name	This is the waymo trip segment ID	N
frame	Lidar frame number	N
velocity	Waymo vehicle velocity (mph)	N
ssd_2.5	Stopping sight distance of the Waymo vehicle with given velocity and reaction time of 2.5 seconds (feet)	Y
cov	Cone of vision expressed in degrees	Y
tot_objs_video	Total objects identified in the Waymo's 6 camera frames	N
tot_objs	Total objects identified in the Waymo's LIDAR point cloud	N
mean_obj_dist	Average distance of all the objects from Waymo AV (feet)	N
min_dist	Distance of the nearest object from Waymo AV (feet)	N
max_dist	Distance of the farthest object from Waymo AV (feet)	N
dist_sdev	Standard deviation of the object distances from Waymo AV	N
inv_dist_sum	Sum of inverse distance sum of all the objects in the LIDAR frame	Y
objs_within_180	Total objects with in 180 degrees cone of vision of the Waymo AV (1st and 4th quaderant with vehicle driving north)	Y
inv_dist_sum_within_180	Sum of inverse distance of the objects within 180 degree cone of vision	Y
objs_within_ssdcov	Total objects within the cone of vision (COV) and stopping sight distance	Y
inv_dist_sum_within_ssdcov	Sum of inverse distance of objects within COV and SSD	Y
weather_rain	Presence of rain (binary)	N
weather_sunny	Sunny day (binary)	N
location_location_other	Other location (binary)	N
location_location_phx	Phoenix location (binary)	N
location_location_sf	San francisco location (binary)	N
time_Dawn/Dusk	Dawn/Dusk (binary)	N
time_Day	Day (binary)	N
time_Night	Night (binary)	N

Figure 7 provides statistical distributions of the sample size for all the critical variables used to develop the CCF model. The maximum accurate range of the long-range LiDAR mounted on the vehicle was 250 feet, corresponding with a maximum safe operating speed of 35 mph based on human SSD requirements. Objects beyond that range were less likely to be detected or classified. Thus, frames with vehicle speeds exceeding 35 mph were excluded from the analysis (approximately 13% of the total frames). Additionally, there were a substantial number of frames where the vehicle was not moving (zero speed) due to the urban and ultra-urban settings along

with stop-and-go traffic operations. The SSD and the COV were also zero, which skewed the sample towards zero. Thus, frames with speeds less than 0.1 mph were excluded from the analysis. After clipping the frames with speeds greater than 35 mph and less than 0.1 mph, the sample size was reduced to 108,369 frames (68.54% of the total possible frames). Figure 8 shows the distribution of the critical attributes for 68.54% of the data after the trimming process was used in the CCF model building.



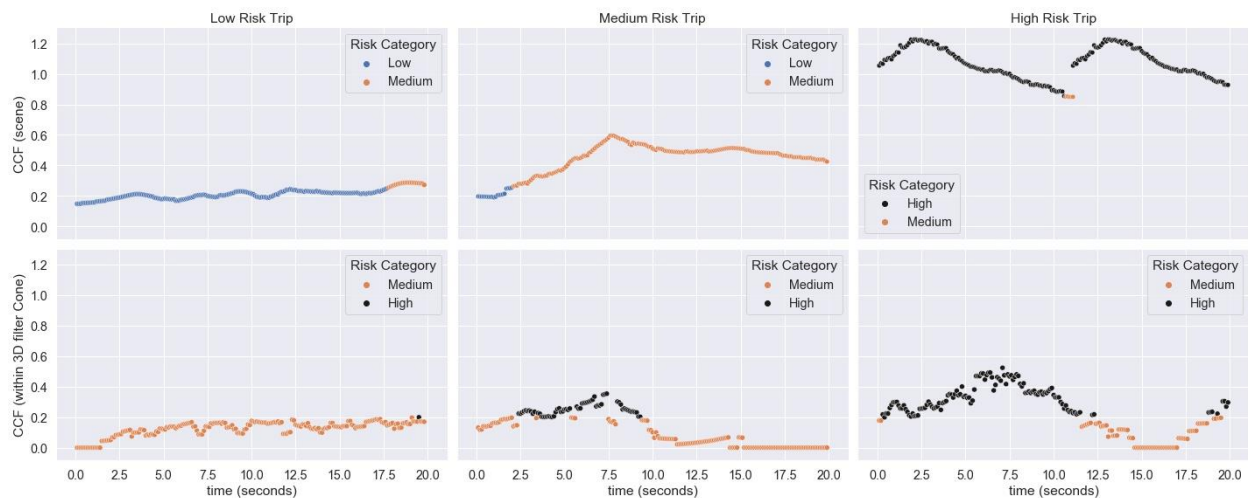
**Figure 7 Statistical distributions of critical variables before the clipping**



**Figure 8 Statistical distributions of critical variables after clipping the frames with speeds greater than 35 mph and less than 0.1 mph.**

The analysis provided a frame-by-frame comparison of contextual complexity based on the density of objects and their proximity to the autonomous vehicle as represented by the CCF. All trips were categorized as high, medium, or low-complexity trips based on the statistical mode of the trip's CCF category.

Figure 9 provides an example of three trips categorized as low, medium, and high-contextual complexity trips. The figure consists of a 2x3 matrix of complexity plots, and each column contains two graphs of an individual trip. The left column is for a low-complexity trip, the middle column is for a medium-complexity trip, and the right column is for a high-complexity trip. The x-axis represents time in seconds. The y-axis describes the CCF. The top row illustrates CCF for the entire scene, and the bottom row displays CCF within the COV. The corresponding video of each of these trips is provided in the respective hyperlinks ([high complexity](#), [medium complexity](#), [low complexity](#))



**Figure 9 CCF plots for high, medium, and low-complexity trips (velocity >0.1 mph and <= 35mph).**

The upper right plot in Figure 9 shows a high-complexity trip on a 2-lane urban road in an ultra-urban area. The trip predominantly consisted of a high density of objects close to the vehicle. The trip started with a medium-complexity context for 3 seconds and transitioned into a high-complexity context for the remainder of the trip. After 3 seconds, the vehicle entered an intersection with many vehicles, pedestrians, and bicyclists, thus elevating the CCF. After traversing the intersection, the vehicle entered another 2-lane urban road with curbside parking, moving vehicles and pedestrians nearby, maintaining an elevated CCF. The bottom right plot shows the resulting CCF within the driver's COV. The CCF within the driver's COV (bottom right plot in Figure 9) and the overall CCF of the scene (top right plot in Figure 9) vary considerably. This is because many objects fall within the driver's COV at the start of the trip as the vehicle traverses the intersection, making it high complexity for the driver. The contextual complexity in the driver's COV diminished to a medium and then a low complexity as the vehicle decelerated and reached a standstill (between 15-17 seconds).

The medium-complexity trip (middle top and bottom plots in Figure 9) consisted of an urban multi-lane highway with a center two-way-left-turn lane. At the trip's start, a few objects were in the scene, making it a low-complexity environment. At the 2-second mark, pedestrians and bicyclists prepared to cross the road and were detected, which elevated the complexity gradually to medium as the vehicle was advanced. This trend is noticeable in the top middle plot. The corresponding CCF within the driver's COV also intensified to a high complexity, which is represented in the bottom middle plot.

The low-complexity trip comprised vehicles driving on a local neighborhood road without moving vehicles, pedestrians, or bicyclists. The entire scene's complexity remained low for most of the trip. On the contrary, the CCF within the COV remained at medium complexity throughout the trip except at the beginning when the vehicle accelerated from standing still.

Based on the visual inspection of the trips, the three examples (high, medium, and low) accurately characterized the contextual complexity of the driving environment. Table 3 below provides the ranges for all critical variables to classify into appropriate complexity categories.

**Table 3 Critical variables and their complexity class ranges.**

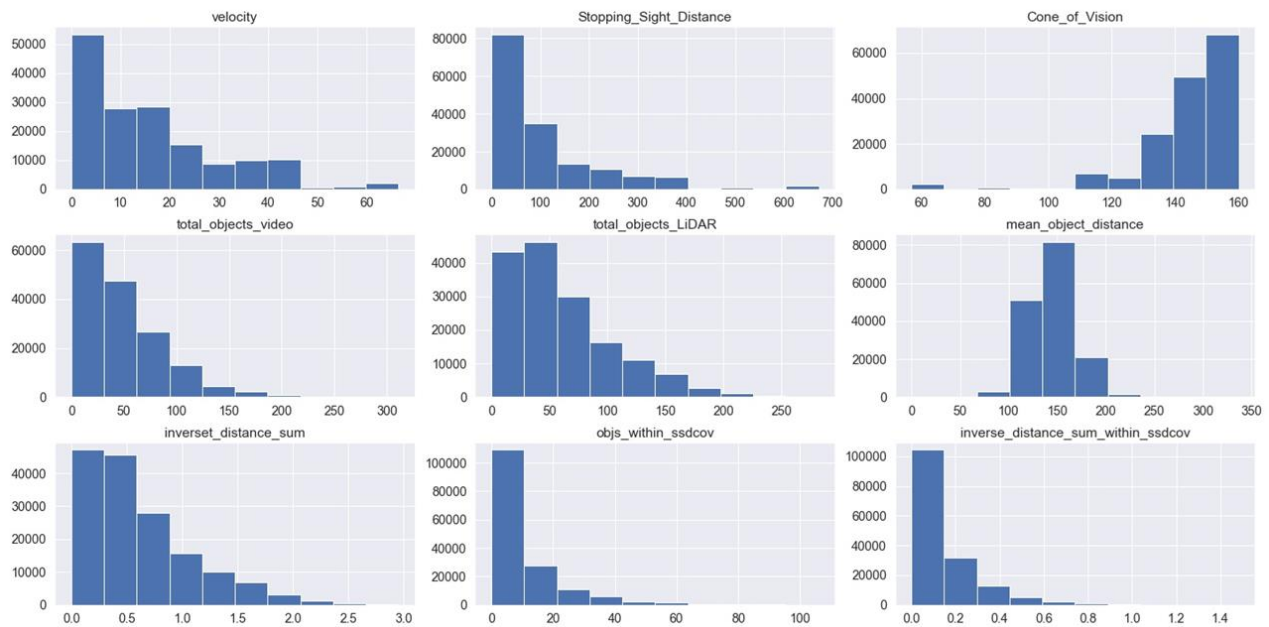
Dynamic Variables	High	Medium	Low
Velocity (mph)	0-40	0-66	0-67

<b>Object Density</b>	44-282	8-64	0-53
<b>Object Distance (feet)</b>	89-196	38-143	0-337

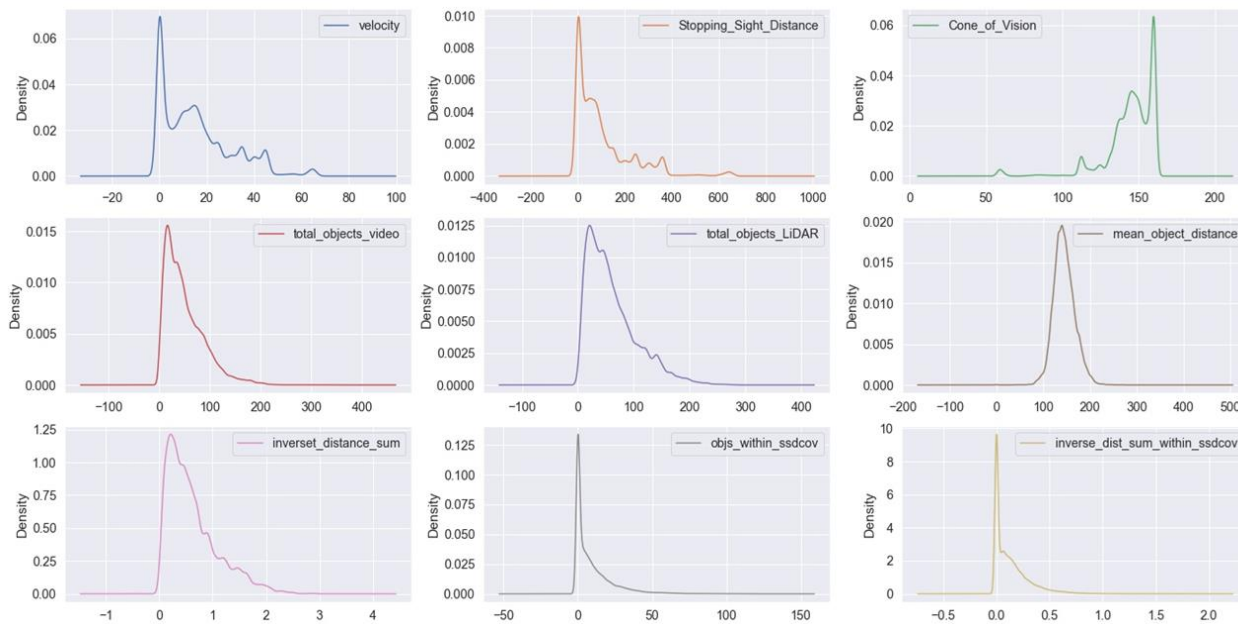
The statistical modeling approach satisfactorily represents the contextual complexity of the driving environment. However, one impediment of the methodology is that the quartiles do not paint the picture with sufficient granularity. It can be seen from Table 3 that the variables overlap between different complexity classes. To overcome this, a machine-learning approach using an unsupervised clustering method was tested, which is discussed in the next section.

## 4.2 Machine Learning Approach

This section explains using an unsupervised clustering analysis to build a model to classify complexity accurately. Specifically, k-means and hierarchical clustering algorithms were used to create the model. The Gaussian mixture model clustering can provide probabilities along with the cluster labels; however, they assume Gaussian subpopulations, do not work well with irregular cluster shapes when data contains categorical features, and are sensitive to initialization and outliers in the data. After initial comparisons with all three models using project data, the Gaussian mixture model was dropped due to poor performance. Analysis in the rest of the paper contains only k-mean and hierarchical clustering methods. Figure 10 and Figure 11 show the statistical distribution of the critical variables chosen for modeling. Figure 10 provides the histograms of the variables, while Figure 11 demonstrates the density of the data points.



**Figure 10 Histogram of different attributes from Waymo Autonomous Vehicle Data**



**Figure 11 Density plots of different attributes from the autonomous vehicle data**

It is evident from the density plots in Figure 11 that the data for different variables are not uniformly represented. For example, velocity, stopping sight distance, and object density are skewed towards the left (i.e., more samples are available). This is because many frames included vehicles at a standstill when the data were collected in urban and ultra-urban settings with stop-and-go traffic. The SSD and COV were also zero at zero speed, which resulted in oversampling. In the statistical approach, this data was clipped to eliminate the bias. However, a consolidation technique was applied in the machine learning approach, which is explained further.

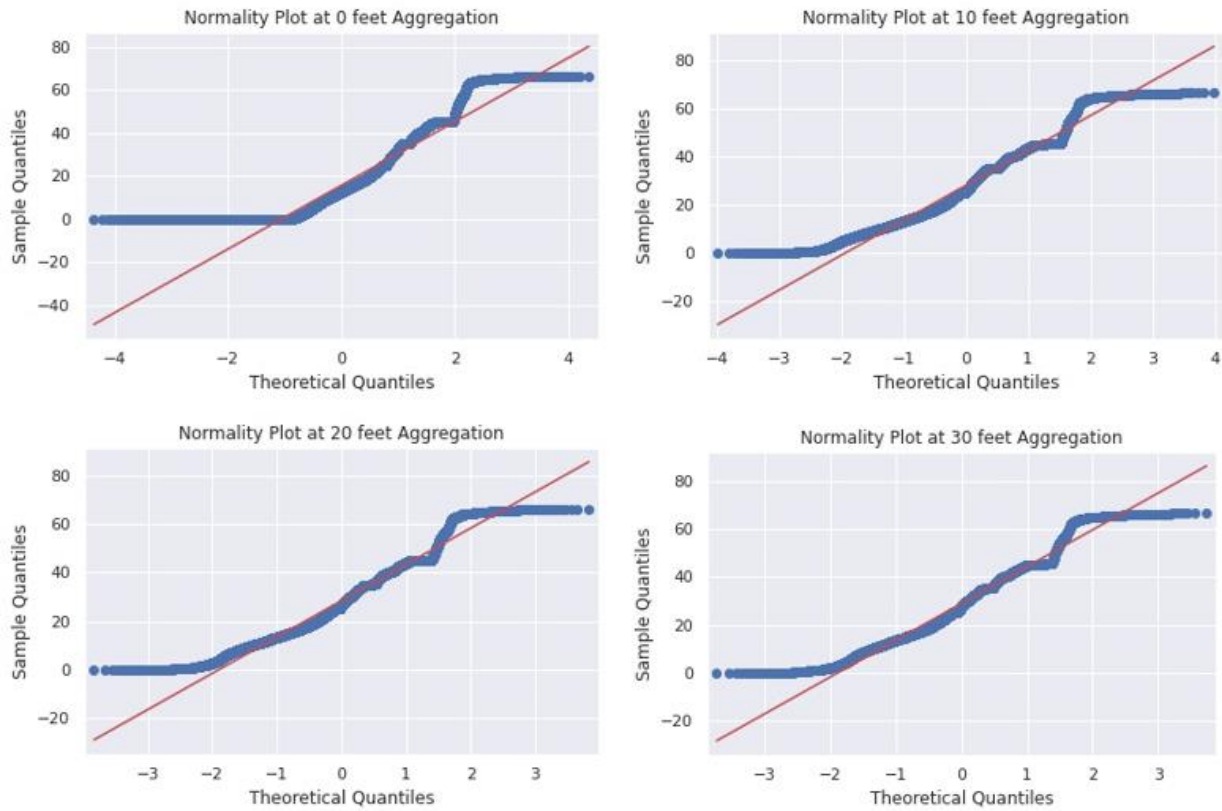
Most machine learning algorithms developed for classification were designed to assume an equal number of samples for each class. Using highly skewed or imbalanced data results in poor classification performance models (Krawczyk, 2016). This is true for k-means and hierarchical clustering algorithms used for clustering models. The skew can be mitigated in two ways:

- Undersampling of the over-represented class
- Oversampling of the under-represented class

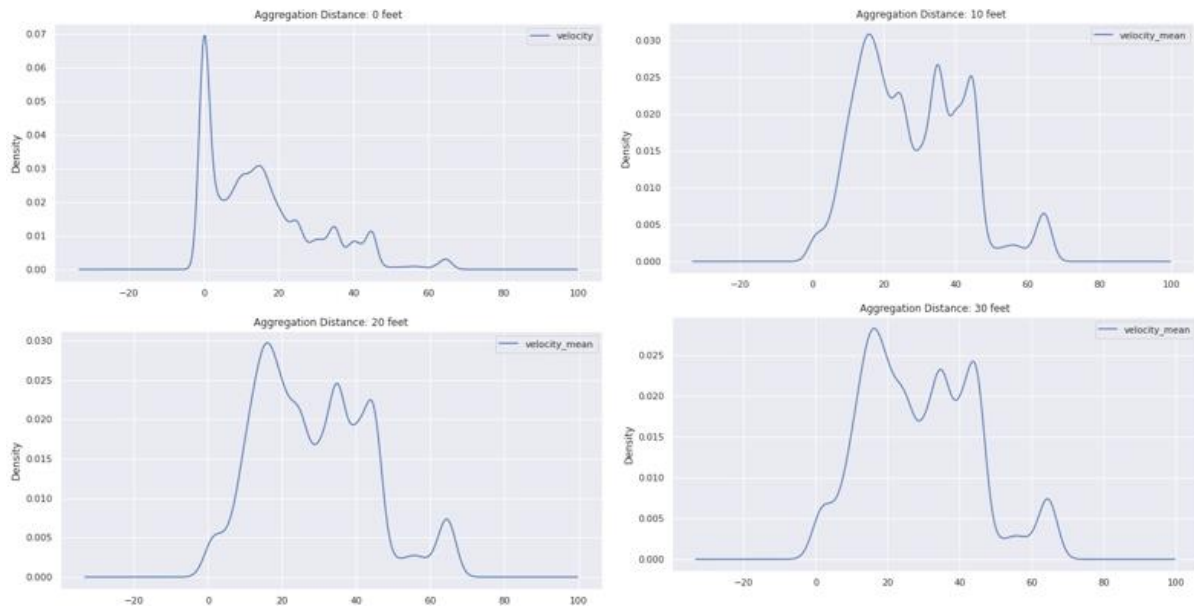
An undersampling approach was considered for these analyses. The LiDAR frames are represented as a factor of time, i.e., the point cloud was collected at a frequency of 10 Hz/Second. Thus, every trip of 20 seconds has 200 frames of LiDAR point cloud data. Since substantial LiDAR frames were collected at a very low or zero speed, a logical solution was to aggregate the data by distance. An aggregation distance of 10, 20, and 30 feet were considered for normalizing the data.

Figure 12 shows the normality plots for the variable velocity at different aggregation distances. The red line that extends diagonally on the chart is a theoretical normal curve plot. The thick blue dots below the theoretical normal curve (see the red line) represent the autonomous vehicle data's normal curve. The top left plot exhibits the normality plot for unaggregated data (i.e., 0 feet aggregation). The top right plot represents 10 feet aggregation, the bottom left displays 20 feet aggregation, and the bottom right shows 30 feet aggregation. It is evident from the graphs that the raw data has a lot of frames clustered at a velocity of zero. At 10 feet aggregation, this improves, and more points move towards the theoretical normal curve. Further, this condition

improves at both 20- and 30- feet aggregation, and the sample more closely resembles a normal curve. Subsequent consolidations did not improve the normality of the dataset. Thus, a consolidation distance of 30 feet was selected for use for the model development.



**Figure 12 Normality plots for different aggregate distances**



**Figure 13 Density plots for different aggregation distances**

Figure 13 demonstrates density curves at different levels of data consolidation. The aggregation reduced the number of zero-velocity frames from the analysis while simultaneously increasing the number of high-velocity frames. There is a noticeable change in the data distribution between

the unaltered dataset (i.e., 0 feet consolidation) and 30 feet consolidation. Although the density distribution of the dataset looks less than the ideal curve characteristic of a normal curve, the data is closer to representing data between all the classes.

Identifying optimal aggregation distance is crucial for model development. Excess consolidation reduces the sample size significantly, rendering it insignificant for model development. On the other hand, insufficient consolidation retains the bias in the data, resulting in poor model performance. Figure 14 shows the plot of the sample size at different consolidation distances. Table 4 represents the same in tabular format. It can be observed that as the consolidation distance increases, the sample size reduces. At 30 feet aggregation, the sample size is 11003 frames. Further consolidation did not yield any improvement in the distribution of the samples. Thus, consolidation of 30 feet and a sample size of 11003 was considered for building the unsupervised clustering model.

Waymo Data Size After Consolidation

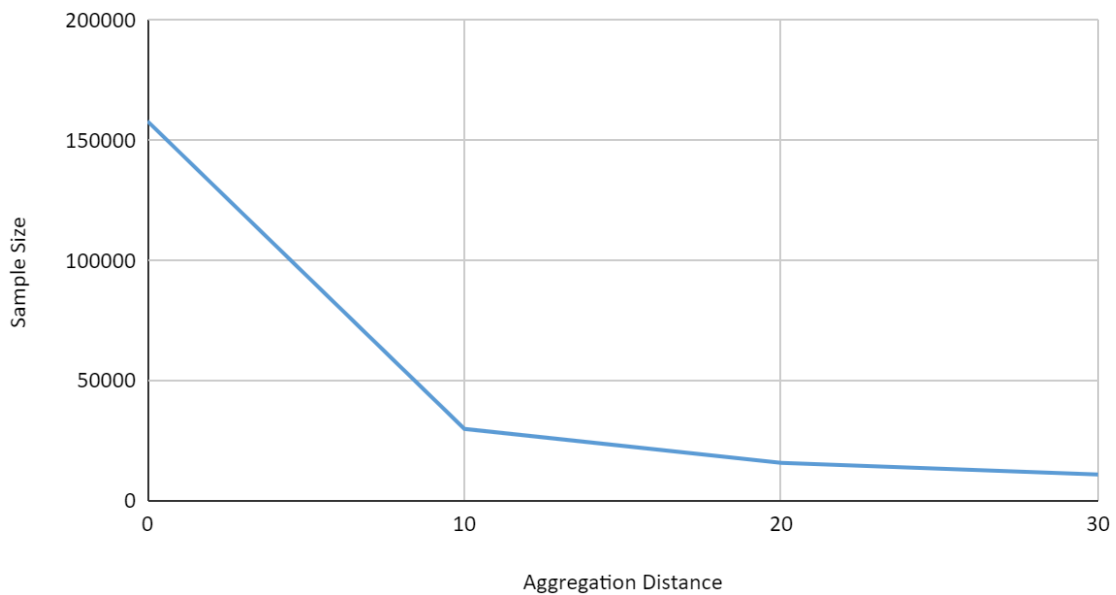


Figure 14 Data size at different aggregation distances

Table 4 Data size at different aggregation distances

Aggregation Distance	Sample Size
0	158090
10	29964
20	15851
30	11003

#### 4.2.1 Feature selection

The variables we use to train the machine learning models significantly influence the performance of the models. Irrelevant or partially relevant features can negatively impact model performance. The author performed principal component analysis (PCA) to identify the most critical variables. Table 5 below presents the variables and their explained variance.



**Table 5 PCA analysis results**

Attributes	Explained variance
velocity	0.8818025176
obj_density_video	0.0047732978
obj_density_lidar	0.0034133763
mean_proximity	0.0011882991
inv_dist_sum	0.0002242688
objs_within_ssdcov	0.0000621344
inv_dist_sum_within_ssdcov	0.0000029339
dist_trav	0.0000023591
cum_dist	0.0000017586
weather_rain	0.0000009873
weather_sunny	0.0000001214
location_other	0.0000001191
location_phoenix	0.0000000364

Based upon the PCA analysis, the top 5 most critical variables are listed below in decreasing order of importance:

- velocity: velocity of the vehicle
- obj\_density\_video: total number of objects captured in the video camera
- obj\_density\_lidar: total number of objects captured in the lidar point cloud
- mean\_proximity: mean distance of all the objects from the vehicle
- objects\_within\_COV: total number of objects captured within the COV

While PCA analysis identified the most important variables, it is also essential to identify the interaction between the variables. Table 6 shows the Pearson correlation coefficients between the variables. Green cells indicate a high positive correlation, and red cells indicate a high negative correlation. The variable "velocity" is highly correlated with variables "obj\_density\_lidar" and "obj\_density\_video." Velocity and total objects in LiDAR are negatively correlated, indicating that the total number of objects decreases as the velocity increases. On the other hand, "velocity" is positively correlated with "mean\_proximity," showing an increase in vehicle speed also increases the proximity of the surrounding vehicles. The proximity and density of the objects are weakly correlated.

**Table 6 Correlation coefficients of variables**

Pearson Correlation Coefficient	Velocity	Object Density (Video)	Object Density (Lidar)	Mean Proximity	Objects within COV
velocity	1	-0.2687	-0.4119	0.4139	0.5682
Object Density (video)	-0.2687	1	0.8036	0.0053	0.0689
Object Density (Lidar)	-0.4119	0.8036	1	-0.1196	0.0927
Mean Proximity	0.4139	0.0053	-0.1196	1	0.2134

From PCA and correlation results following variables are selected for clustering analysis:

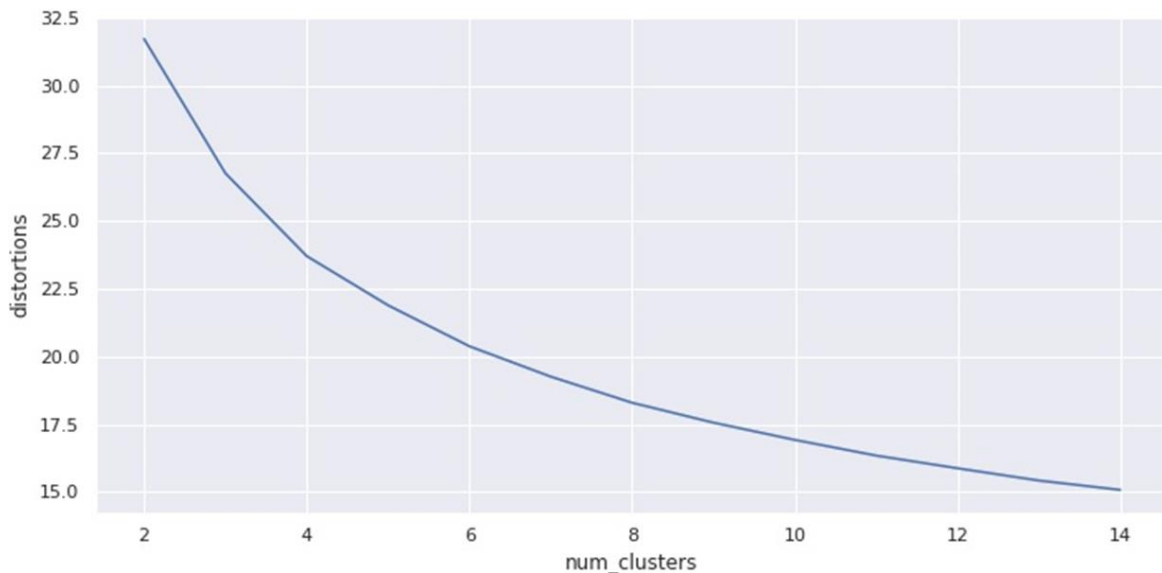
- velocity: This is the most critical variable with the highest variance that can be quantified and explained. Velocity also shows excellent interaction between other vital variables (i.e., object density and proximity)
- obj\_density\_lidar: Although PCA analysis ranked this variable below "obj\_density\_video," it shows a superior correlation with velocity. Adopting the above variable will produce better model results because of its enhanced interaction.
- mean\_proximity: proximity to the nearest object is ranked fourth in the priority list captured from PCA analysis. It also demonstrates a significant correlation with the key variable "velocity."

Based on the inferences mentioned above, the author considered "velocity", "object\_density\_lidar," and "mean\_proximity" to build the clustering model.

#### 4.2.2 Clustering Analysis

Identifying an optimal number of clusters is essential for building a clustering model. However, for the intended audience of this research, the author wanted to know how the clustering model would segregate the trips at different cluster values. The author used the elbow method to identify an ideal number of clusters through the distortion plots. The distortion is the sum of squares of points from cluster centers. It decreases with increasing clusters and becomes zero when the number of clusters equals the number of points.

Figure 15 shows the elbow line plot between cluster centers (x-axis) and the distortion (y-axis). The cluster centers range from a minimum of two to a maximum of 14 clusters.

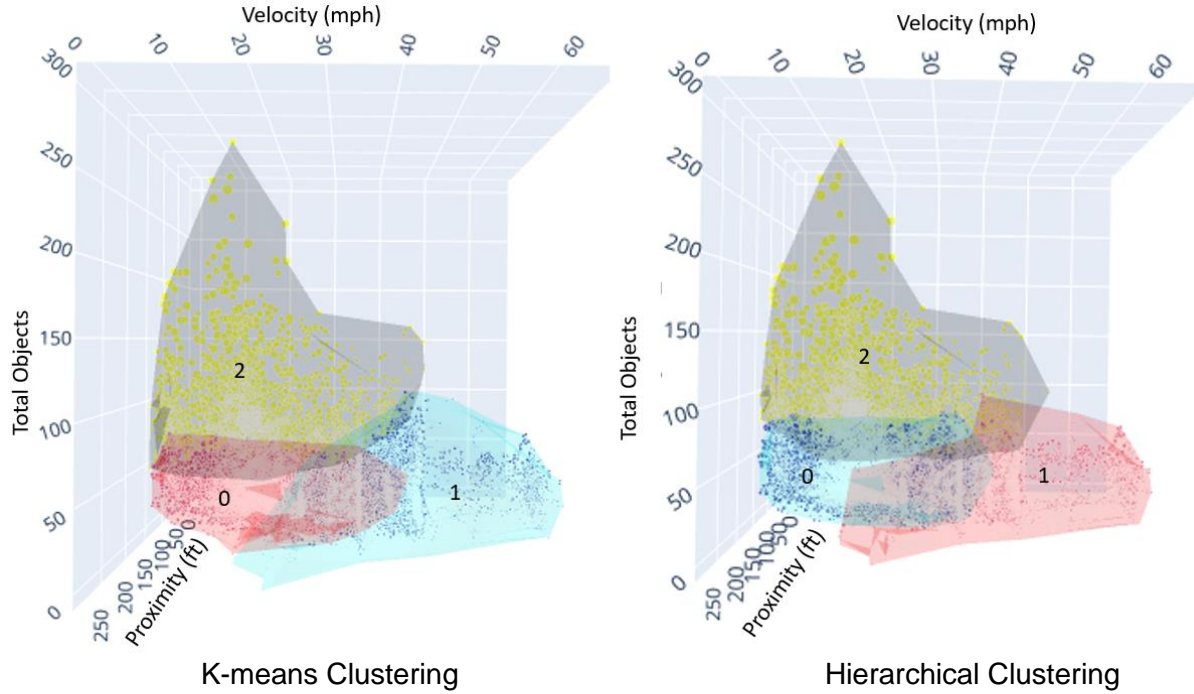


**Figure 15 Distortion Plots from Clustering Analysis**

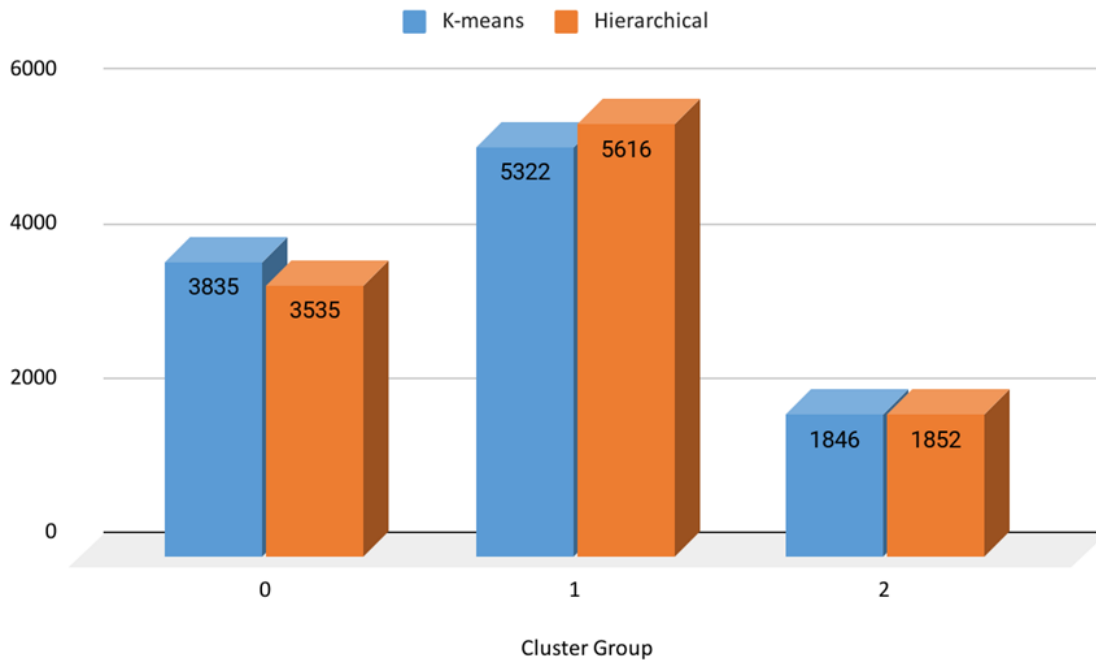
The Elbow method indicates an optimal number of clusters for the model. It is generally identified at locations with an abrupt change in the slope of the line. The first abrupt change is observed at cluster 3; however, the distortion is still very high, indicating more separation possibility. Next, the difference is observed at clusters four, five, and six, after which the slope changes are barely noticeable. Anything less than three does not capture all the distinct grouping due to high distortion. Everything above five leads to too many groups and does not produce a notable reduction in distortion. Thus, the author considers an ideal cluster modeling spectrum ranges between three and six. Despite the ideal cluster range of three to six, the author considered adopting results with three clusters. The sample size available after the consolidation is relatively

small, and increasing the clusters did not yield distinct patterns. Additionally, a value of three is simple to categorize as High, Medium, and Low complexity categories.

Figure 16 below shows clustering results for k-means and hierarchical clustering methods for three cluster centers. The cluster groups are labeled zero, one, and two. Velocity is on the x-axis, object density is on the y-axis, and mean proximity is on the z-axis. **Error! Reference source not found.** compares cluster distribution between k-means and hierarchical clustering.



**Figure 16 K-means vs. Hierarchical Clustering**



**Figure 17 K-means vs. Hierarchical clustering - distribution of points**

The Rand index was estimated to measure the similarity between k-means and hierarchical

clustering models. Rand Index is a ratio of the number of pairs in agreement to the total number of pairs between two clusters and is represented by equation 4.1 and the adjusted rand index is shown in equation 4.2. Table 7 shows the adjusted rand index (ARI) comparison.

$$RI = \frac{\text{Count of Pairs in Agreement}}{\text{Total Number of Pairs}} \quad \text{Equation 4.1}$$

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \quad \text{Equation 4.2}$$

**Table 7 Adjusted Rand Index (ARI) for K-means and hierarchical clustering**

Rand Index	K-Means	Hierarchical
K-Means	1	0.7486
Hierarchical	0.7486	1

From Figure 16 and **Error! Reference source not found.**, it is evident that k-means and hierarchical clustering results look identical. The k-means clustering boundaries look continuous and fluid compared to the hierarchical clustering boundaries. The edges are sharp and wrinkled in the case of hierarchical clustering (Figure 16). From **Error! Reference source not found.**, the number of points in each cluster grouping is indistinguishable, with marginal differences for clusters zero and one. The adjusted rand index (ARI) for k-means and hierarchical is 0.7486, which shows considerable resemblance. Since the models are identical, choosing either would be acceptable. The author looked into the literature to identify methodological nuances that would assist in selecting a model. Hierarchical clustering does not work as well as k-means clustering when the shape of the clusters is hyperspherical, i.e., a circle in 2-dimension or a sphere in 3-dimension. The data we are using for modeling is not spherical in the structure; thus, the k-means clustering model has a superficial edge over the hierarchical clustering model, even though, technically, both are similar. Therefore, the research team considered the k-means clustering model to determine ranges for dynamic complexity determination.

#### 4.2.3 Cluster Centers and Corresponding Dynamic Complexity

Understanding the parameters of the cluster grouping is essential for assigning a contextual complexity. The author chose three cluster center models because it will be easier to categorize into three distinct categories: high, medium, and low complexity. The k-means clustering model in Figure 16 displays three groups with zero, one, and two labels. Velocity is represented on the x-axis, object density on the y-axis, and proximity on the z-axis. Table 8 shows the cluster characteristics and their corresponding complexity rank. Cluster groups and the interpretation behind the assignment of complexity rank are elucidated below:

1. Cluster zero: cluster group zero includes locations with low velocity and low density of objects compared to the other two groups. Cluster zero is also relatively safe due to the low density of objects and low speeds. In other words, these are areas with less traffic and speed. Due to these characteristics, cluster zero represents a low-complexity environment.
2. Cluster one: cluster group one includes locations with relatively high velocity, low-medium object density, and low-to-high proximity of objects. The areas classified in this group are more complex than cluster group zero. Hence, cluster one represents a "medium-complexity" environment.
3. Cluster two: cluster group two includes areas with high object density and proximity. Broadly these locations have increased traffic which is tightly packed. They might represent locations in central business districts with increased activity. Compared to the other two

groupings, these locations present a relatively complicated driving context. Thus, cluster two represents areas with a "high-complexity" environment.

**Table 8 Cluster group characteristics and their complexity rank**

Cluster Group	Characteristics			Complexity Rank
	Velocity	Object Density	Object Proximity	
0	low-to-medium	low-to-medium	low-to-medium	Low
1	medium-to-high	low-to-medium	medium-to-high	Medium
2	low-to-medium	medium-to-high	medium-to-high	High

#### 4.2.4 Dynamic ranges of attributes for complexity categorization

Adopting the results from the clustering analysis, the author further built the complexity ranges for the attributes (i.e., velocity, object density, and object proximity). Table 9 shows the computation of complexity ranges for each variable to categorize into low, medium, and high. Table 10 shows only the complexity ranges without other statics used for calculation.

**Table 9 Attributes and their dynamic complexity ranges**

Attributes		Dynamic Complexity		
		Low	Medium	High
Velocity (mph)	mean	6.1	18.98	16.55
	std	11.07	9.02	8.76
	mean-2*std	0	1	0
	mean+2*std	28	37	34
Object count	mean	27.09	34.78	114.12
	std	19.41	18	37.32
	mean-2*std	0	0	39
	mean+2*std	66	71	189
Object Proximity (feet)	mean	168.72	136.63	148.16
	std	20.27	17.5	16.81
	mean-2*std	128.18	101.63	114.54
	mean+2*std	209.26	171.63	181.78

**Table 10 Dynamic complexity ranges for attributes.**

Dynamic Complexity	Velocity (mph)	Object count	Object Proximity (feet)
Low	0-28	0-37	0-34
Medium	0-66	0-71	39-189
High	128-209	101-172	115-182

It can be observed that the attribute ranges overlap with each other, and this is because all three variables together define the three-dimensional spatial boundary of these clusters. Validation checks were performed considering historical crash data which is available in the dissertation research study by Bendigeri (2022).

## CHAPTER 5

### Conclusions

The research goal of this project was to develop a dynamic contextual complexity model to measure and appropriately categorize the driving environment's complexity from high to low complexity. Traditional road safety assessment methodologies do not recognize the driving environment's fast-changing dynamics that influence the contextual complexity and, ultimately, its risk. The advent of autonomous vehicle open datasets has created new opportunities to measure dynamic complexity and incorporate dynamic interaction metrics into risk estimates and safety assessments.

A total of 798 autonomous vehicle trips, comprising 158,090 LiDAR point cloud frames, were analyzed in this research. The dynamic complexity model was developed using two approaches, i.e., statistical and machine learning. The Contextual Complexity Factor Model developed using the statistical method captures the density and proximity of the objects from the vehicle, which are the key parameters influencing the trip's complexity. The machine learning model included similar key parameters (i.e., object density, proximity, and velocity) and was equally proficient in predicting the dynamic complexity with justifiable truthfulness. This was evident as the dynamic complexity of both models correlated with the historical crash data. )Trips where dynamic contextual complexity was categorized as "high" were also the ones with higher crash totals that included severe injuries. Predominantly, locations with high volumes of pedestrians and bicyclists appeared to tend to be in the high-risk. This is logical as pedestrians and bicyclists take less space and are placed closely, increasing the object density and proximity and consecutively increasing the complexity of the environment. However, this interpretation should be substantiated in a further scientific study with location data to extract historical crash experience information. The dynamic risk ranges were used to develop a numerical rating system that categorizes a given variable into high, medium, or low complexity. Each variable's risk levels and ranges were transformed from a three-dimensional representation to a tabular format.

Identifying and predicting high-risk environments in real-time can significantly benefit safety research, driver education, auto-insurance risk assessment, autonomous vehicle route planning, and many more. For example, this research could allow Driving Rehabilitation Specialists (DRSs) to score the dynamic complexity during training and testing to ensure that the driver is competent at all situational levels. The methodology this project developed utilizing the autonomous vehicle open datasets could aid DRSs to measure and classify the contextual complexity of the routes used for on-road driving evaluations for medically-at-risk drivers considering the dynamic variables. The on-road driving evaluation is the gold standard for testing and rehabilitating medically at-risk drivers. The product of this research could build foundational work to build tools and methodology to measure the roadway context to enhance the consistency and validity of the on-road assessment procedures.

Additionally, this research could assist in the route planning of autonomous vehicles. Current autonomous vehicle route planning strategies do not consider scene complexity, making it more challenging for drivers to take control of the autonomous vehicle when needed. All the highway safety manual models are built on historical data and do not have context associated with them. However, the advent of autonomous vehicles and the technology to process complex sensor fusion data generated from them can assist in building safety models that consider contextual complexity. The addition of context would inform how many cars were there, their proximity, and their arrangement before the crash. Such information is currently missing from safety models.

For future research, it would be interesting to assess human driving data in autonomous vehicle-enabled vehicles to determine potential differences in the distribution of contextual risk between

machine and human driving. Further, the open datasets have redacted location information to protect the identity of the objects measured. Future research should seek permission to connect the historical crash experience and correlate the safety risk associated with varying levels of contextual risk. Identifying conditions surrounding safety risks and complexity could improve our understanding of crash risk and support the development of more efficient safety countermeasures.



## REFERENCES

- AASHTO. (2018). *A Policy on Geometric Design of Highways and Streets, 7th Edition* (7th ed.).
- Abdel-Aty, M. A., & Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5), 633–642.
- Bendigeri, Vijay, "Using Safety Performance Models, Autonomous Vehicle Data, and Machine Learning to Develop Contextual Complexity Criteria to Establish a Standardized Process for On-Road Evaluation of Medically At-Risk Drivers Considering Static and Dynamic Factors of the Roadway Environment" (2022). Ph.D. Dissertation, Clemson University, U.S.A., 2983
- Bendigeri, V. G., Zou, F., Ogle, J. H., & Kusram, K. Roadway Contextual Risk Assessment Using Dynamic Traffic Conditions Data Obtained from Autonomous Vehicles. In *Computing in Civil Engineering 2021* (pp. 562-569).
- Briand et al. (2016). A mixture model clustering approach for temporal passenger pattern characterization in public transport. *International Journal of Data Science and Analytics*, 1(1), 37–50.
- Choudhary P, Imprialou M, Velaga NR, Choudhary A. Impacts of speed variations on freeway crashes by severity and vehicle type. *Accid Anal Prev*. 2018 Dec;121:213-222.
- Dewar, R. E., & Olson, P. L. (2002). *Human Factors in Traffic Accident Litigation in: Human Factors in Traffic Safety*. Lawyers & Judges Publishing Company, Inc.
- Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 11(1), 40–56.
- Kanungo. et al. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 881–892.
- Liu, J., Cai, D., & He, X. (2010). Gaussian mixture model with local consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1).
- Mackworth, N. H. (1976). Stimulus density limits the useful field of view. *Eye Movements and Psychological Processes*, 307–321.
- Montazeri-Gh, M., & Fotouhi, A. (2011). Traffic condition recognition using the k-means clustering method. *Scientia Iranica*, 18(4), 930–937.
- Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86–97.
- Olson, P. L., & Farber, E. (1996). *Forensic aspects of driver perception and response* (Second Edi). Lawyers & Judges Publishing Company, Inc.
- Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71.
- Rogé, J., Pébayle, T., Lambilliotte, E., Spitzenstetter, F., Giselbrecht, D., & Muzet, A. (2004). Influence of age, speed and duration of monotonous driving task in traffic on the driver's useful visual field. *Vision Research*, 44(23), 2737–2744.
- Shinar, D., McDowell, E. D., & Rockwell, T. H. (1977). Eye movements in curve negotiation. *Human Factors*, 19(1), 63–71.
- Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., others, 2020. Scalability in perception for autonomous driving: Waymo open dataset, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2446–2454.
- Waymo\_open\_dataset. (2019). Waymo Open Dataset: An autonomous driving dataset, URL <https://www.waymo.com/open>.