Identifying Safety-Critical Heavy-duty Vehicles in Fleets with Complementary Vehicle Inspection Datasets through Cross-Database Clustering Analysis

Chenyu Yuan
Department of Civil and Environmental Engineering
Carnegie Mellon University, Pittsburgh, PA, 15213
Email: chenyuyu@andrew.cmu.edu
ORCiD: 0000-0002-3821-3314

Ying Shi
Department of Civil and Environmental Engineering
Carnegie Mellon University, Pittsburgh, PA, 15213
Email: yingshi@andrew.cmu.edu
ORCiD: 0000-0003-2244-5791

Ruoxin Xiong
Department of Civil and Environmental Engineering
Carnegie Mellon University, Pittsburgh, PA, 15213
Email: ruoxinx@andrew.cmu.edu
ORCiD: 0000-0001-8273-8276

Pingbo Tang[1]
Department of Civil and Environmental Engineering
Carnegie Mellon University, Pittsburgh, PA, 15213
Email: ptang@andrew.cmu.edu
ORCiD: 0000-0002-4910-1326

Word Count: 5490 words + 5 table (250 words per table) = 6,740 words

Submitted July 31st, 2022

[1] Corresponding Author: 5000 Forbes Ave, BPH123F, Pittsburgh, PA 15213

ABSTRACT

Defective truck equipment is a significant cause of many truck accidents and incidents. Conducting safety inspection programs with a minimum impact on the mobility of commercial vehicle fleets is a practical challenge for maintaining fleet safety and efficiency. Planned or unannounced inspections are critical for ensuring truck safety while greatly hindering fleet operational efficiency and mobility. Vehicle deterioration models concluded by historical inspection records can realize targeted inspection with priority. Unfortunately, vehicles can have different deterioration trends including different failure rates of components – called “failure modes.” A single deterioration model could hardly capture all failure modes of diverse vehicles and achieve reliable failure predictions. In addition, various inspection databases could capture different deterioration-related information. So cross-database analysis is essential to overcome these challenges for comprehensive failure mode analysis.

This research examines commercial heavy-duty vehicles’ safety-efficiency tradeoffs by analyzing two historical inspection data sources to comprehensively capture and synthesis failure modes. Two algorithms, K-means clustering, and Latent Dirichlet Allocation collectively analyzed different temporal-spatial failure modes among vehicles and carriers. The identified component failure modes could prioritize inspection and maintenance plans for inspectors, drivers, and fleet managers, which help avoid repetitive out-of-service violations and improve fleet operational strategies with less mobility reduction.

Keywords: Vehicle Safety Inspection, Failure Modes, Safety-efficiency Tradeoffs, Clustering Analysis, Natural Language Processing

INTRODUCTION

Commercial trucks and trailers must perform different types and levels of inspections annually. These inspections include periodic inspections (e.g., annual or semi-annual) and random roadside inspections, which focus on vehicle components such as brakes, tires, and lights. However, the current inspection strategies are almost performed randomly by individual inspectors, which may fail to identify the high-risk vehicles from massive truck fleets. Based on the investigation by Keall and Newstead in 2013, vehicle defects are evident in many crashes, contributing to about 13.5% (1). They also found that if we reduce the annual inspection interval from 12 months to 6 months, the injury crash rate decreases by 8%. However, frequent inspection intervals will constrain mobility and increase unnecessary operating costs.

In spite deterioration model identification is important to provide a threshold set as a buffer for different vehicles to be alerted when "safe level" is reached (2). But it is complicated to generalize all the situations of hundreds or even thousands of vehicles regarding totally different deterioration trends. Failure modes identification becomes important and straightforward to navigate and instruct legislators and motor carriers with simpler and more concise clusters rather than threshold sets with time and mileage factors. Suppose we have a list of failure modes that point to component defects or operation problems in vehicles and carriers with certain background features and driving behaviors. In that case, inspectors can inspect vehicles customized and strategically with a more efficient pipeline. Meanwhile, drivers and fleet managers from carriers can also benefit from this failure mode identification because they can pay more attention to sensitive and fragile components.

Another shortcoming comes from limited data sources information. For example, Motor Carrier Management Information System (MCMIS) Catalog contains detailed descriptions of the violations found during vehicle inspections while having no detailed mileages of vehicles (3). On the other hand, some commercial vehicle inspection companies maintain databases that capture detailed mileage while only mentioning the problematic vehicle components without detailed descriptions of the violations (4). So cross-database analysis can overcome information absence problems and embodies failure mode analysis in a more complementary and comprehensive way.

The research presented in this paper aims at a more comprehensive failure mode identification from two databases that contain complementary inspection records for capturing different information related to the deterioration trends of various commercial vehicles. The research team used two historical inspection datasets to summarize violation patterns among various vehicle features or components. We introduce K-means clustering and Latent Dirichlet Allocation models to identify failure modes based on information integration cross-database. Finally, we used violation counts or probability as performance metrics to evaluate failure modes' effectiveness in identifying groups of vehicles of high risk.

LITERATURE REVIEW

While legislators are trying to simplify and humanize the inspection process of inspections, motor carriers should also focus on self-inspection and real-time monitoring to avoid being cited or given a score below average on FMCSA Safety Measurement System. Besides the argument of the effectiveness of inspection programs, identifying each component's violation probability and crash risk probability can improve carriers' safety and efficiency performance. Randhawa et al. (5) found the most often cited component in incidence reports. They reviewed 3,600 selected police reports from six states, and brakes are reported as a major cited mechanical factor with 1.7% of involvements. Then comes components such as tires, wheels, coupling, and load securement, all at about 0.4%. Daniel Blower et al also examined the relationship between the mechanical condition of heavy trucks and crash involvement (6). They used the Large Truck Crash Causation Study (LTCCS) to test if trucks with defects and out-of-service (OOS) conditions were statistically more likely to be involved than trucks without these conditions. They also found that violations in the brake system (36% of all) and the lighting system (19%) were the most frequent, and violations related to brake adjustment increased the odds of the truck's being the striking vehicle by 1.8 times. Above all the discussion focusing on mechanical factors, researchers emphasize the importance of component healthy conditions with brakes, lights, and tires. But how valuable it is for different makes of vehicles and carriers with different operation patterns to schedule self-pre-trip inspections or install real-time monitoring devices like telematics remains unknown.

Failure mode identification can provide a tool for drivers and fleet managers to navigate through different combinations of critical vehicle components in various vehicles to avoid high-risk vehicle operation scenarios. Researchers used statistical approaches to identify individual high-risk vehicles from annual safety inspection records. Zheng et al. (7) tried a gradient boosting data mining model to evaluate several factors' relationship with crash injury severity. They classified the crash severity into four different categories. They concluded that wet road surface, bad visualization (dark or low light conditions, or fog/poor weather conditions), a strong crosswind, heavy gross vehicle weight, and collisions with opposite traffic would increase the likelihood of more severe outcomes. Liang et al. (8) tested the effectiveness of safety roadside inspections by exploring accidents caused by reduced caution in driving and lack of vehicle maintenance. They also applied a classical case in economics by Becker's research (9) to point out that if motor carriers or fleet managers are aware of this regulation, such practices will undermine the effectiveness of the regulation by reducing their compliance. Unfortunately, these studies have not yet traced how vehicle component defects interact with other features such as age, mileage, and vehicle properties, leading to high-risk operation scenarios and crashes.

The contributions of the paper include: 1) generalizing failure modes from millions of vehicle inspection records; 2) revealing distributions of different background features (such as age, mileage, and urbanity) in each mode; 3) synthesizing text recording into failure topics that represent a specific failure mode found during random roadside inspections.

METHODOLOGY

In this paper, we utilized historical truck inspection data from different sources to explore potential failure modes behind historical inspection records. The first sub-section below describes data preprocessing pipelines defining reasonable time ranges and validating correct inspection records. The second sub-section introduces clustering methods, such as K-means clustering and latent Dirichlet allocation methods applied to different datasets to cluster multiple failure modes based on descriptions and topics extracted from inspection records. Figure 1 shows the overall framework of the proposed method.

Data Sources and Preparation

This research uses two vehicle inspection databases. The first is a database maintained by a privately owned IT contractor in Pennsylvania. In many states, such as Pennsylvania, inspection data are collected by the state government and privately owned IT contractors and inspection companies. CompuSpections, LLC (CompuSpections) is a privately owned IT service company incorporated in 2003. Their work includes over 30 years of performing State Inspections and creating record management software services for inspection stations. Their software service, SIRPAWeb, is designed for Pennsylvania vehicle safety inspection stations for recording and printing accurate and uniform MV-431/480 safety inspection forms.

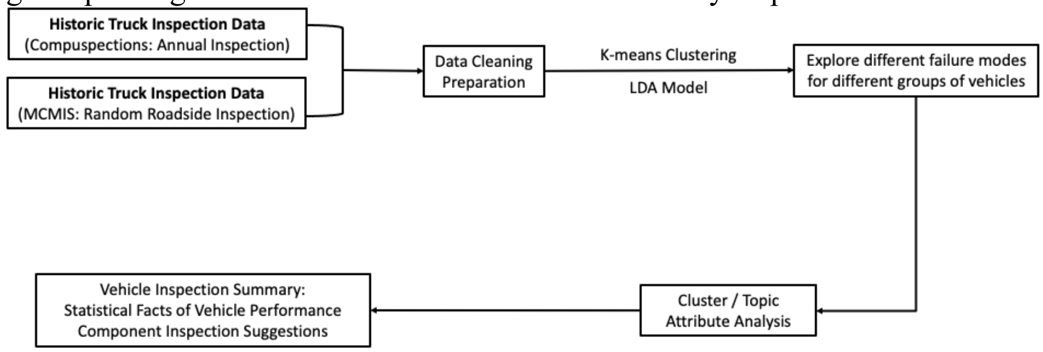

Figure 1. Research Process Designation.

MCMIS (Motor Carrier Management Information System), maintained by FMCSA. MCMIS is a source for FMCSA inspection, crash, compliance review, safety audit, and registration data (10). From that database, multiple tables are used to extract useful information for each inspection with violations. These tables include the INSPECTION table, UNIT table, VIOLATION table, and INSP SUPP VIOLATION table.

Because different inspection stations and inspection agencies have their naming and recording regulations, dataset checks, transformation, and loading processes are essential for further analysis. Checking regulations will be introduced in the validation experiment design section to clean all invalid inspection records and filter commercial vehicles that are heavy-duty tractors or trailers. A dataset attribute summary is provided in Table 1.

TABLE 1 Data Summary for Two Different Sources

	Compuspections	**MCMIS**
Dataset Description	Inspection records that using Compuspections software service in Pennsylvania	Inspection records conducted by state personnel under the Motor Carrier Safety Assistance Program (MCSAP)
Data Source	Collected by Compuspections software service, SIRPA Web	Captured by FMCSA through SAFETYNET
Inspection Type	Annual Periodic Inspection	Random Roadside Inspection
Data Range	200– - 2021	2021
Data Type	Vehicle identification number (VIN), make, model, model year, binary inspection geographic information, inspection overall result and component results, vehicle odometer reading	Vehicle identification number (VIN), make, model, model year, non-binary inspection geographic information, inspection overall result and component results, inspection defect descriptions

As for urbanity classification for registered vehicles in the Compuspections dataset, this research used the Urban-Rural classification scheme provided by The Center for Disease Control's National Center for Health Statistics (NCHS) (11). This scheme distinguishes urban and rural areas into six categories, from Type 1 as most urban to Type 6 as most rural. After that, we used 2010 Census data to assign the NCHS classification to all the counties shown in the dataset (12).

Inspected Vehicles Failure Mode Identification

Given that historical inspection records are high-dimensional and have unstructured values for some attributes (e.g., text descriptions of violations), generalizing thousands of inspection records into failure mode clusters is necessary but challenging. According to research by D. Peck et al.(13), the inspection failure rate is related to three parameters such as urban/rural county classification, age, and odometer reading. M. Beydoun (14) also suggested that mileage, age, weight, and vehicle makes such as Chrysler, Ford, GM, Hyundai, and Mazda have significantly impacted estimations for testing emission failure on passenger vehicles. Based on all the recent research, this research decided to organize different clusters based on vehicle information (e.g., mileage, age) and usage contexts (urban/rural county).

This research considered all component inspection results in the CompuSpections dataset to divide datasets into different clusters and considered violation descriptions in the MCMIS database to divide datasets into various topics. Perr-Sauer et al.'s research (15) about commercial vehicle time-series data analysis with K-means clustering shows three steps of K-means clustering. These steps include 1) extracting the overall and each historical component inspection results in the Compuspections dataset; 2) applying the elbow method to find the best performance k values for the components inspection dataset. The Silhouette coefficient assisted in evaluating the performance of the clustering method; and 3) calculating the difference between each cluster's average violation counts and the whole dataset's violation count, summarizing the failure modes behind them.

Regarding the fact that the MCMIS database has an individual file that records violation descriptions on the roadside, topic modeling is another technique that can help cluster inspection records specifically. This research adopted topic modeling techniques demonstrated in Subasish Das et al. (16) for processing the FARS database and NHTSA vehicle complaint database to test the effectiveness of state vehicle inspection. In the MCMIS database, the steps of establishing topic modeling in this research include 1) data prepossessing to clean violations unrelated to vehicle maintenance information. 2) tokenizing each paragraph, cleaning stop words, stemming, and lemmatizing words to get a final analyzable dataset about violation descriptions; 3) Calculating the TF-IDF value to evaluate each word's frequency and importance; 4) distinguishing each vehicle's failure mode by the recorded descriptions and LDA topic modeling. Figure 2 shows the proposed method combining the K-means clustering method and the LDA model for identifying each vehicle's specific failure clusters/topics.

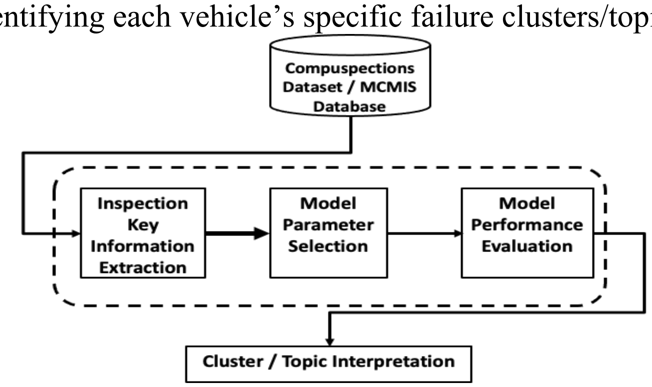

Figure 2. K-Means Clustering Method and LDA Model Design for Compuspections Annual Inspection Dataset and MCMIS Random Roadside Inspection Database.

VALIDATION EXPERIMENT DESIGN

Data Cleaning and Preprocessing

Cross-analysis of historical inspection data from different inspection types is essential to estimate the optimal inspection timing interval for drivers and fleet managers. According to FMCSA (Federal Motor Carrier Safety Administration), four types of inspection are daily driver inspections, periodic/annual inspections, roadside inspections, and onsite compliance reviews. From all the inspection types above, periodic/annual inspections and roadside inspections are required by the federal or state departments of transportation and have relatively uniform inspection standards.

Because different inspection stations and agencies have their naming and recording regulations, various checks, transformation, and loading processes are essential for further analysis. For cleaning all invalid inspection records and filtering commercial vehicles, heavy-duty tractors, and trailers, a checking regulation pipeline is designed, as shown below in Figure 3:

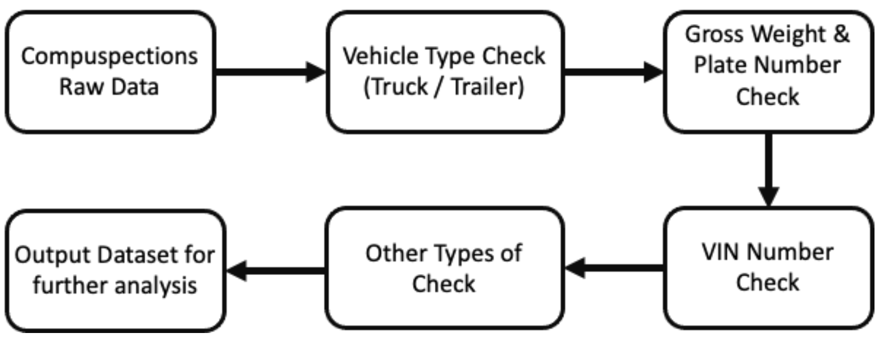

Figure 3. Compuspections Dataset Data Preprocessing Flow Chart.

As for vehicle checks, this research first filtered commercial trucks and trailers and then used gross vehicle weight and plate number to exclude vehicles that were not heavy-duty tractors and trailers. There are also naming regulations for VIN numbers to check, such as total length, security check digit, and model year digit. All these naming regulations exclude invalid VIN numbers and corresponding illegal inspection records from further analysis. The last step is to exclude inspection records that are not correct. The algorithm excludes outliers from further analysis depending on the “passorfailedinspection” column, component test columns, odometer reading columns, brake thickness, and tire tread columns.

The authors also performed a similar data preprocessing flow for the MCMIS database compared to Compuspections data. We use “INSPECTION_ID” as a key to join the inspection table, unit table, and violation table, and “INSP_VIOLATION_ID” as a key to join the violation table with the violation supplement table so that we can combine each violation record with text descriptions. In addition, a similar VIN naming regulation check was manipulated as Compuspections data to exclude incorrect VIN numbers from further analysis. For vehicle types and gross weight, the further investigation also only kept heavy-duty trucks and trailers. Only vehicle types related to trucks and trailers are kept according to “INSP_UNIT_VEHICLE_ID_NUMBER” column. What differs from the Compuspections dataset is that the MCMIS database also recorded vehicle violations unrelated to component defects. So only maintenance violation codes that related to components are chosen here.

K-Means Clustering Algorithm

Clustering algorithms can use various vehicle attributes or background features (e.g., mileage, age, and urbanity) to identify similar vehicles in multiple aspects. A good set of vehicle attributes or background features can lead to clustering results with clear boundaries and fewer overlaps between clusters, where some vehicles fall into both categories and hard to tell the differences between two clusters. Hopkins statistics measure the clustering tendency of a dataset (17). This metric aims to measure how different the distances are among the data points in a real dataset from their neighbors, comparing the distances of a uniformly distributed dataset. A Hopkins Statistic greater than 0.9 indicates a dataset far different from random uniformly distributed dataset, with highly clusterable performance. From the results in Table 2, every feature sets are highly clusterable (Hopkins Statistic > 0.9). Another value used for measuring the quality of clustering and selecting proper features/attributes is the "Silhouette Coefficient" (18). The Silhouette Coefficient value closer to 1 means that clusters have clear boundaries and not too much-overlapped area among them. Besides measuring how suitable clusters can be as failure modes identification, it is also important to determine what background features are critical enough to influence failure modes identification results. The authors divided all vehicle features into five feature sets to check if any lead to clear clustering boundaries for identifying the vehicle's failure modes.

TABLE 2. Summary of the Clusterability Analysis for each Feature Set.

Feature Set	Information Included	Hopkins Statistic	Silhouette Coefficient
Feature Set 1	$O^{(1)} + EC^{(2)}\ IR^{(3)}$	0.9907	k = 4, $SC^{(4)}$ = 0.8765; k = 7, $SC^{(4)}$ = 0.8860;
Feature Set 2	Mileage + $O^{(1)} + EC^{(2)}\ IR^{(3)}$	0.9269	0.6305
Feature Set 3	Age + $O^{(1)} + EC^{(2)}\ IR^{(3)}$	0.9863	0.5652
Feature Set 4	Urbanity + $O^{(1)} + EC^{(2)}\ IR^{(3)}$	0.9902	0.6279
Feature Set 5	All Feature Included	0.9219	0.4027

(1) O = Overall
(2) EC = Each Component
(3) IR = Inspection Result
(4) SC = Silhouette Coefficient

After checking that each feature set is suitable to proceed with the K-Means clustering method, determining K, the number of clusters, is vital to find failure modes. The elbow method is the first criterion to identify K and Silhouette coefficient assisted in evaluating if clusters have a clear boundary with fewer overlaps. Based on the performance and evaluation by both elbow method and Silhouette coefficient, only feature set 1, with overall inspection result and each component inspection result has two selections for K value. Clustering feature set 1 with K equaling 7 has a slightly better Silhouette coefficient performance than K equaling 4. However, Data Version 5 has a Silhouette coefficient that is below 0.5, which shows uncertain boundaries with clusters and overlapped areas. So, in this case, feature set 5 is not considered further for clustering analysis.

Topics Modeling for Failure Modes Identification

MCMIS, a roadside inspection database established by FMCSA, has a different pattern of inspection recording compared to Compuspections annual inspection dataset. It has a detailed description of each violation on commercial trucks and trailers to illustrate the current conditions of component defects. Based on the information provided, the authors used topic modeling, such as latent Dirichlet allocation modeling, to explore topics rather than clusters.

Before measuring word importance by TF-IDF for each document, text cleaning is performed before measurement. A full text cleaning step includes:

- *Message Clearance*: remove numbers and punctuations and transform all letters to lower cases.
- *Message tokenized*: splitting a text object into words from whitespaces.
- *Stopword removals*: remove all words that have no semantic relevance to the document. For example, words such as articles, pronouns, and prepositions are stopwords that need to be removed.
- *Stemming and Lemmatization*: stemming refers to the process of reducing each word to its root or base. For example, words such as "warning," "warned," and "warner" are all reduced to the stem "warn." However, there are still words such as "good," "better," and "best" that cannot be solved by stemming. Lemmatization is introduced to operate on a single word with knowledge of the context. Lemmatization can discriminate between words with different meanings depending on the part of speech.

Based on all the text cleaning processes above, a "word list" was generated for each vehicle's inspection documents, and their word importance (TF-IDF) is measured from there on to implement the LDA model. LDA model is a popular way to convert an unstructured and complex textual dataset into topics (19). In this method, LDA model assigns each document with different probabilities of topics, and also assigns each topic with different probabilities of words. When topics with a sets of words are listed, LDA model gives a parameter (per-topic-per-word probability) to each word in a certain topic. This parameter shows how likely this word can be generate in this topic. All these processes can be done by many open-source tools such as NLTK (20).

After text cleaning and TF-IDF calculation, we should define the exact number of topics for the LDA model. In general, the number of topics, K, can adjust the granularity of the topic model. The more topics accepted, the more narrow results it will get, or vice versa. According to the nature of the LDA model and previous studies, we used a grid search method to assign the best performance value for each parameter (21). Finally, the best number of topics is eight.

RESULTS

K-Means Clustering Results

This research interprets clusters into different failure modes by failure rate analysis. It took groups of vehicles in each cluster and calculated the average number of violation counts for them. If some indicators or components' average number of violation counts are significantly different from baseline average violation counts, then we can identify this cluster with a specific failure mode. Based on this logic, Figures 4 – 8 show how significantly different each cluster's average violation counts are from baseline (the whole dataset average) average violation counts and interpretations about what failure modes can conclude from the data.

After cluster interpretation analysis, k-means clustering can divide the whole vehicle fleet into four or seven clusters based on Figures 4 and 5. From there, it shows that groups of vehicles with lighting, brake, and tire problems are significantly above average. This conclusion suggests that lighting, brakes, and tires can be key inspection components during annual inspection processes.

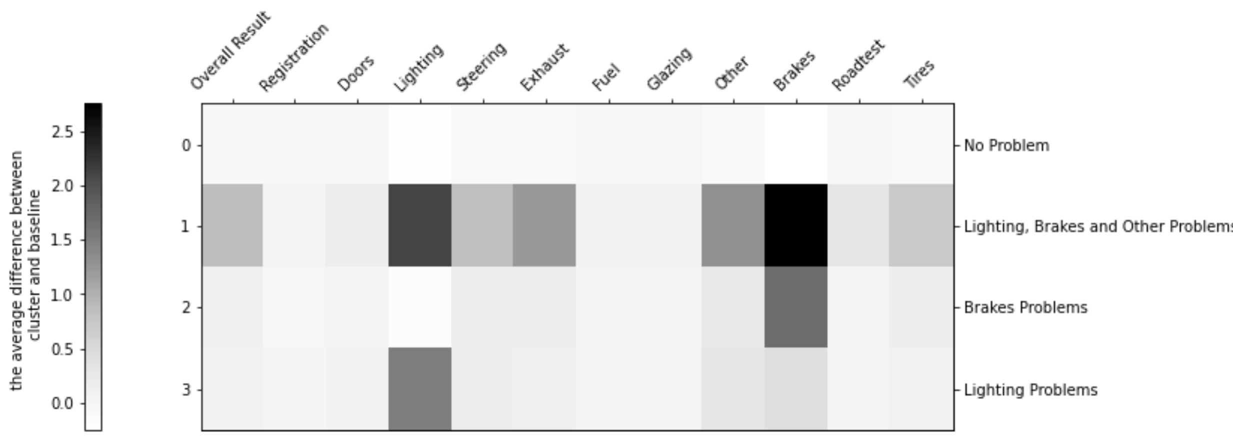

Figure 4. Failure Mode Heatmap Summary of Feature Set 1 (k = 4).

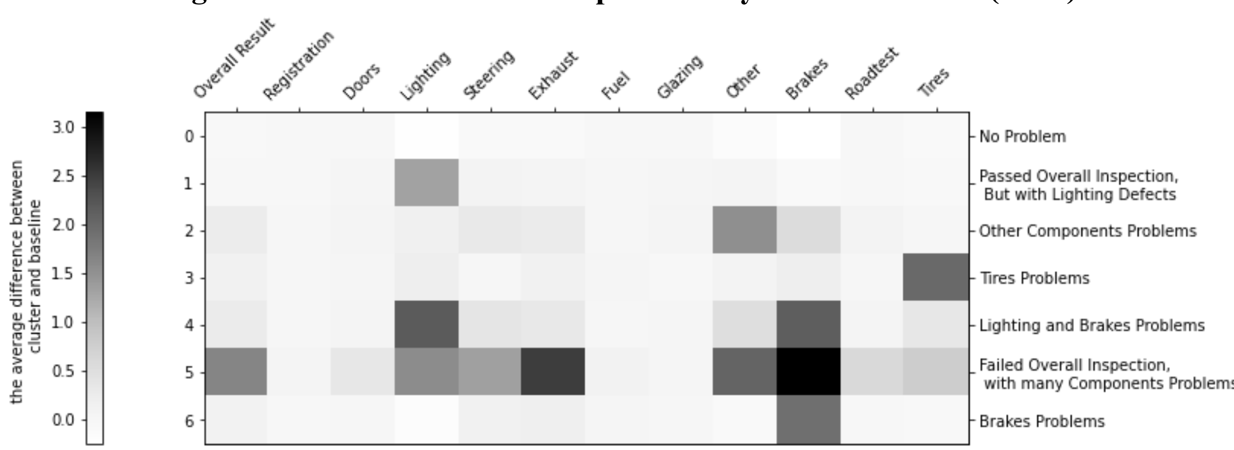

Figure 5. Failure Mode Heatmap Summary of Feature Set 1 (k = 7).

The K-means clustering that uses feature set 2 (overall plus milage) divides all vehicles into four groups depending on mileage driven per year. Figure 6 shows the clustering results using feature set 2. The clustering result shows that the vehicle group with slightly above average mileage (2988.23 miles) has the most significant lighting and brakes problems. It indicates that vehicles with average mileage driven per year are the most noticeable cluster if inspected, especially with lighting and brake components. Identical results are found by adding age and urbanity features (Figures 7 and 8). Medium age generation and vehicles registered at the large fringe and medium metro area also have significant problems with lighting and brakes problem, compared to other age groups and urbanity areas. Overall, brakes, and tire problems are the most common failure mode when annual inspections are performed based on different vehicle properties. While talking about background information such as mileage, age, and urbanity, vehicles with certain features can be key important features to give extra attention to when doing annual inspections, such as vehicles with average mileage driven, medium vehicle age and vehicles from large fringe and medium metro area.

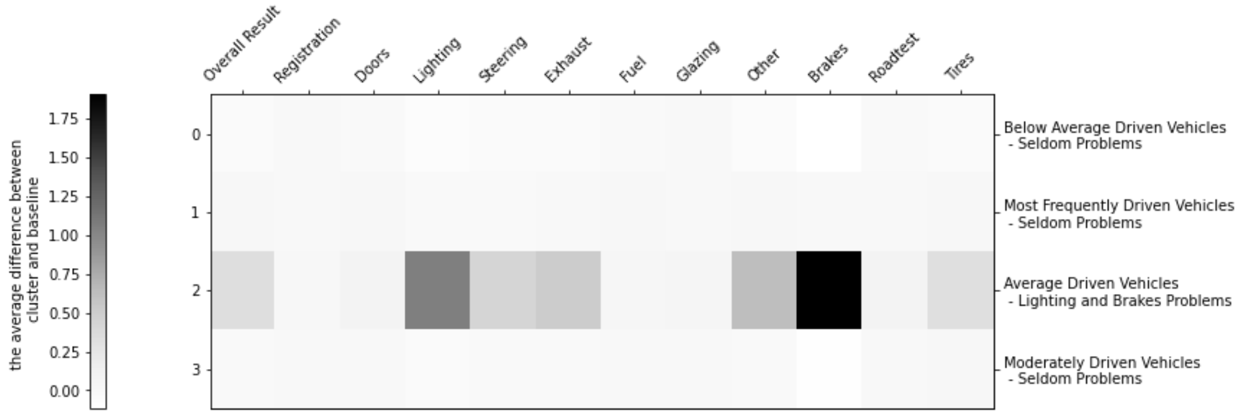

Figure 6. Failure Mode Heatmap Summary of Feature Set 2 (with mileage, k = 4).

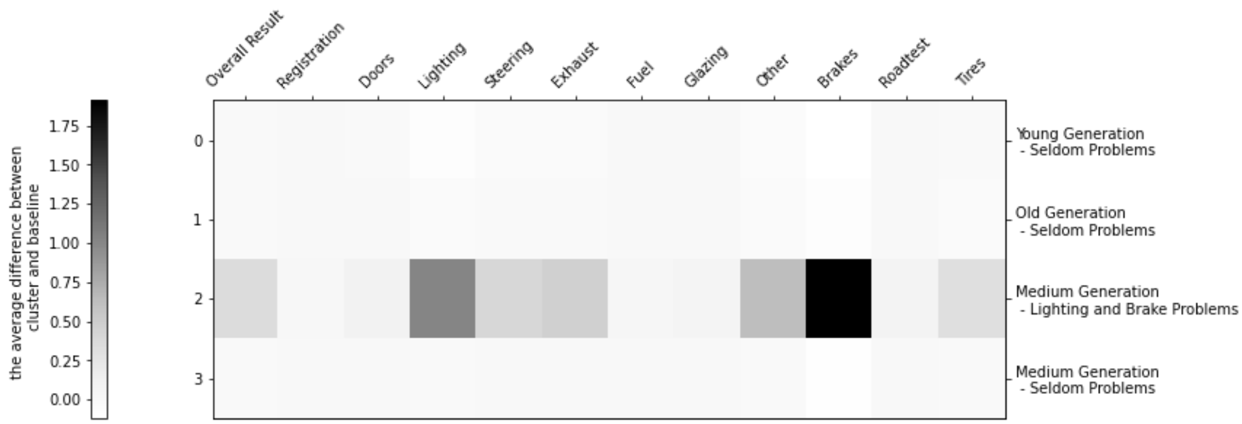

Figure 7. Failure Mode Heatmap Summary of Feature Set 3 (with age, k = 4).

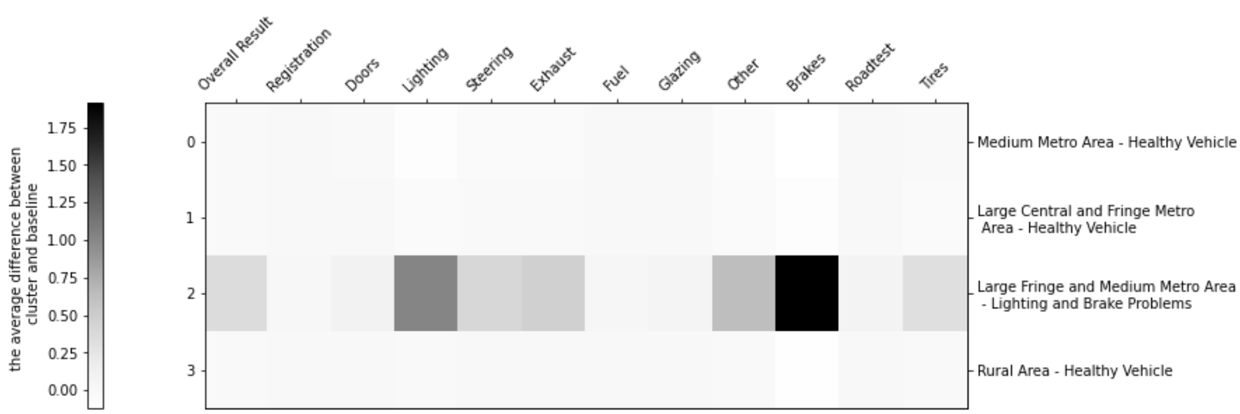

Figure 8. Failure Mode Heatmap Summary of Feature Set 4 (with urbanity, k = 4).

LDA Topic Modeling Results

By training the LDA model and selecting the best parameters, we can obtain the list of topics and analyze the meaning of each failure mode. Table 3 shows the list of topics and the top 10 words of each topic, ranking words by per-topic-per-word probability. For example, the probability of term “lamp” is generated in topic 1 is 0.045.

TABLE 3. Top 8 topics with ten keywords by LDA model from the MCMIS Database.

Topic 1 Word: 0.051*"inop" + **0.045*"lamp"** + 0.034*"inoper" + 0.031*"rear" + 0.030*"turn" + **0.029*"signal"** + 0.026*"front" + 0.026*"right" + 0.026*"left" + **0.025*"light"**
Topic 2 Word: **0.034*"air"** + **0.024*"leak"** + 0.024*"axl" + **0.021*"brake"** + 0.019*"hose" + 0.016*"x" + 0.015*"l" + 0.014*"chamber" + 0.014*"r" + 0.013*"v"
Topic 3 Word: **0.051*"tire"** + 0.050*"axl" + 0.036*"psi" + 0.035*"right" + 0.031*"left" + 0.027*"side" + 0.026*"insid" + 0.021*"outsid" + 0.021*"inop" + 0.021*"flat"
Topic 4 Word: **0.027*"display"** + **0.026*"number"** + 0.025*"name" + 0.024*"usdot" + 0.023*"dot" + 0.022*"carrier" + 0.022*"lb" + 0.017*"vehicl" + 0.016*"compani" + 0.015*"truck"
Topic 5 Word: 0.021*"none" + **0.020*"trailer"** + 0.019*"secur" + **0.019*"chain"** + 0.018*"breakaway" + 0.016*"cabl" + 0.015*"unit" + 0.015*"attach" + 0.013*"strap" + **0.012*"connect"**
Topic 6 Word: **0.016*"oil"** + 0.015*"miss" + **0.014*"leak"** + 0.014*"rear" + **0.014*"engin"** + 0.012*"right" + 0.012*"side" + 0.011*"left" + 0.010*"inop" + 0.009*"cover"
Topic 7 Word: **0.049*"expir"** + 0.035*"" + 0.034*"registr" + 0.019*"current" + 0.016*"plate" + 0.016*"inspect" + **0.014*"proof"** + **0.014*"insur"** + 0.013*"card" + 0.013*"display"
Topic 8 Word: **0.027*"window"** + **0.024*"windshield"** + 0.023*"tint" + 0.021*"fluid" + 0.018*"washer" + 0.017*"measur" + 0.016*"crack" + 0.016*"driver" + 0.014*"side" + 0.013*"adjust"

In Table 3, each topic represents a specific failure mode based on the words selected. For example, Topic 1 is related to lighting violation because it includes words such as “lamp,” “rear,” “turn,” “signal,” and so on, which represents problems such as signal light problems and inoperable lights detected during the roadside inspection. Topic 2 refers to another major violation category, brake problems, because “air,” “leak,” and “hose” are all components related to the brake system. Topic 3 can also be interpreted as “tire problems” since tire violation terminology such as “tire,” “psi,” and “flat” is included. Topic 4 and Topic 7 are related topics that both refer to registration and equipment problems. Topic 4, with words such as “display” and “usdot,” shows that vague display numbers on vehicle bodies can be a major cause of registration violations. Topic 7, with the words “expir” and “insur,” discloses another important insurance proof issue for the registration violation. Other topics, such as topics 5, 6, and 8, also have specific keywords in their content. Topic 5 implies tractor-trailer connection issue, topic 6 implies engine oil leak issue, and topic 8 implies windshield problem.

Any vehicles from our database can be assigned to the most probable topic based on the LDA model. Generally, the LDA model assigns a probability vector to each vehicle. We select the most probable topic for each vehicle and categorize it to that failure mode. Figure 9 shows how popular each topic is, and how many vehicles are in there.

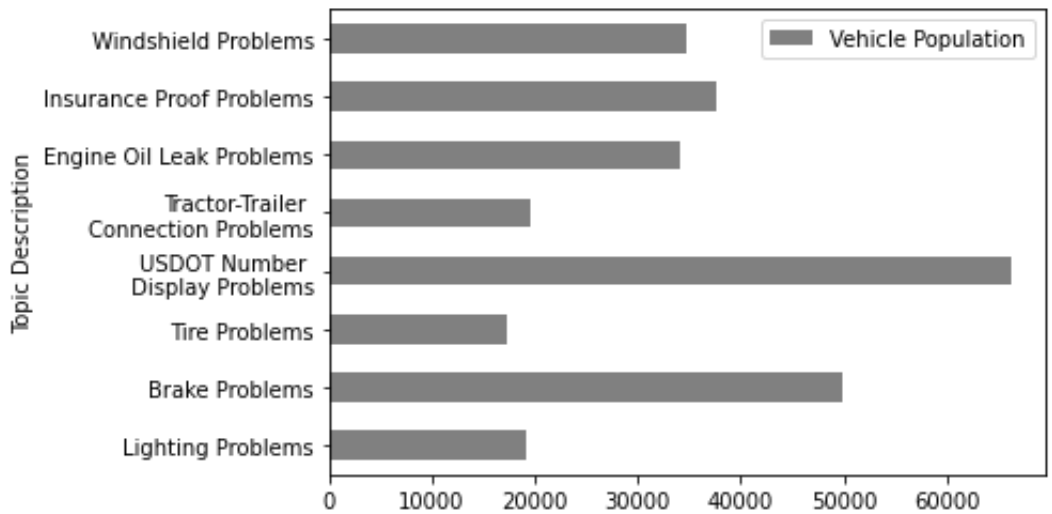

Figure 9. Vehicle Population Counts for Each Topic.

Figure 9 reveals that USDOT Number Display Problems are the most popular. The vague and incomplete USDOT numbers on the body of vehicles could be a common question for motor carriers. The following comes to brake, insurance proof, and windshield problems. That indicates brake and windshield problems lead to failure-prone components during roadside inspections.

Combination Comparison Between Two Failure Modes Identifications

From clustering analysis and topic model analysis based on previous studies, there are some possible failure modes that historical inspection data can define. But how much they overlapped and why certain makes of vehicles can be categorized into these modes remained unknown. This part of the analysis aims to compare and correlate the failure modes found from different databases' records for a potential cross-database analysis that reveals more comprehensive failure mode information of various vehicles. Here the authors select the most popular vehicle makes that exist in both vehicle datasets and analyze their information as a combination comparison. This research adopted feature set 1 with k equaling 7 (since its accuracy among other feature sets) compared with MCMIS topic modeling results to see if the results are similar.

TABLE 4 Population Percentage(%) for Each Make and Each Failure Mode from Compuspections Dataset.

Make / Failure Modes*	0	1	2	3	4	5	6
Make 1	51.2	**13.0**	1.3	0.6	**32.9**	0.0	0.9
Make 2	74.1	4.8	**6.7**	1.1	**6.0**	4.9	2.3
Make 3	69.9	**10.6**	**6.1**	0.8	6.0	1.5	5.1
Make 4	84.0	**6.0**	2.1	0.1	2.3	1.6	**3.8**
Make 5	90.4	**2.1**	**2.1**	0.2	**2.2**	1.8	1.2
Make 6	74.2	4.3	**5.4**	**5.5**	5.2	1.7	3.5

* For Failure mode codes, refer to Figure 5. Bold fonts indicate top 2 largest percentage of the mode for a given make without failure mode code 0 (no problem).

From the percentages tables above with popular makes among data sources, two failure mode identification methods have good performance in identifying lighting defects. For K-means clustering with Compuspections dataset, Make 1, 2 and 3 have a high percentage of lighting defects with failure modes code 1(lighting problems) and failure modes 4 (lighting and brake problems). And this conclusion is also reflected by LDA topic modeling with the MCMIS database, which shows a higher percentage of code 1(lighting problems) from the total population. Besides that, since there are different recording contents in each type of inspection, some failure modes' tendencies in one method are not reflected in another. For example, Table 5 shows a different distribution pattern of Make 4 with failure mode codes 3(tires problems) and 5(tractor–trailer connection problems) from other makes, which are failure mode codes 1(Lighting problems) and 6(engine oil leak problems). In addition, failure mode code 5 (tractor-trailer connection problems) in the MCMIS database is not an item detected by Compuspections annual inspection. However, this finding suggests a correlation analysis between truck makes and failure modes in the future to explore any potential correlation between truck makes and component defects.

TABLE 5 Population Percentage (%) for Each Make and Each Failure Mode from MCMIS Database.

Make / Failure Modes**	0	1	2	3	4	5	6	7
Make 1	6.9	**35.4**	6.1	13.1	5.7	5.7	**13.7**	13.4
Make 2	6.0	**35.1**	5.2	**16.6**	5.2	6.7	14.0	11.2
Make 3	6.5	**34.3**	5.7	10.8	6.3	7.8	**16.5**	12.2
Make 4	6.9	5.4	4.8	**35.1**	7.2	**18.7**	11.3	10.6
Make 5	7.1	**29.4**	7.5	11.9	7.6	9.1	**14.0**	13.5
Make 6	6.5	**38.4**	6.9	10.6	4.9	5.0	12.1	**15.6**

** For Failure mode codes, refer to Table 3. Bold fonts indicate the top 2 largest percentage of the mode for a given make.

DISCUSSION

Previous research only discussed possible optimization strategies based on individual vehicles on a statistical level. This paper proposes a new way to generalize different vehicles operated by carriers into groups, showing that potential groups of vehicles need extra attention when inspected. By exploring potential failure modes with different formats of inspection recording datasets, the inspection process can be optimized by targeting and strategic plans. This study considered how to categorize inspection records into groups of failure modes, and if carriers own similar conditions vehicles, how to make preventive maintenance ahead to avoid unnecessary risks. For annual inspection, we derive failure-prone components from Compuspections Dataset, which indicates failure-prone components are brakes, lighting, and tires. When features such as age, mileage, and urbanity are involved, groups like middle-age generation, average mileage driven groups, and large fringe and medium metro areas are highly attention groups to check if there are any unsafe components. These results are consistent with previous research about brake pad and tire tread deterioration because all these components are perishable if age and mileage get older and longer.

When it comes to roadside inspection with the MCMIS database, a topic model indicates that mechanical component problems are not only popular topics, but some registration problems such as USDOT number display and insurance proof can also be trivial but critical violations that influence carriers' performance in the FMCSA rating system. If motor carriers concentrate on improving their rating scores on the FMCSA website, these mistakes should be prevented. Besides that, high probability also makes brake, windshield, and engine violations very popular. That result suggests that motor carrier workers such as drivers and fleet managers include more precise and detailed pre-trip inspections or install real-time monitoring devices such as telematics.

CONCLUSION

From Compuspections Dataset (annual periodical inspection), this research concludes that there are approximately four different failure modes, most of which point to brake and light failures. When background information is included, these feature sets also correlate with component inspection results. For example, from Figures 6 – 8, vehicle groups with medium mileage driven, middle age, and from the large fringe metro and medium metro areas have significant differences compared to the baseline overall average model (more than 1.9 violation cases). When inspectors inspect vehicles with these features, they should pay extra concern with key components. From MCMIS Database, eight topics are not only related to component failures but also to registration and insurance proof problems. That means a basic pre-trip check is essential for basic display and paperwork materials to prevent the negative influence of tiny mistakes and ignorance, such as a reduction in CSA safety score and ranking. Both results indicate that brakes, lights, and tires are failure-prone components that form obvious failure modes.

ACKNOWLEDGEMENTS

The authors thank Bernie Elder of Compuspections and FMCSA for providing the raw data and approving its use for this project. The authors thank Paul Sniegocki and Roger Elmer of Clarience Technologies for discussing heavy-duty vehicle inspection and vehicle telematics for this project. We also thank the US DOT Mobility21 Center and the Department of Civil and Environmental Engineering at Carnegie Mellon University for funding. 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.

AUTHOR CONTRIBUTIONS

CY developed algorithms and wrote Python code to identify failure modes. YS and RX provided feedback on research methods and conclusions. PT devised the project and provided critical feedback on methods and analyses. All authors wrote and edited the manuscript.

REFERENCES

1. Keall, M. D., & Newstead, S. (2013). An evaluation of costs and benefits of a vehicle periodic inspection scheme with six-monthly inspections compared to annual inspections. *Accident; analysis and prevention*, 58, 81–87. https://doi.org/10.1016/j.aap.2013.04.036

2. Krishnappa, V., Matthews, H. S., & Liu, Y. (2019). Data-Driven Analysis to Support Revised Tire Tread Inspection Standards. Transportation Research Record, 2673(11), 517–528. https://doi.org/10.1177/0361198119851456

3. United States Department of Transportation. (n.d.). *MCMIS Catalog and Documentation.* Federal Motor Carrier Safety Administration. Retrieved July 30, 2022, from https://ask.fmcsa.dot.gov/app/mcmiscatalog/d_inspection2
4. *Pennsylvania Safety Inspection Record Solution.* CompuSpections. (n.d.). Retrieved July 30, 2022, from https://www.compuspections.com/
5. Randhawa, S. U., Miller, S. G., Bell, C. A., & Montagne, P. E. (1998). A study of commercial vehicle safety alliance's out-of-service criteria. *Accident; analysis and prevention*, 30(1), 61–67. https://doi.org/10.1016/s0001-4575(97)00062-6
6. Blower, D., Green, P. E., & Matteson, A. (2010). Condition of Trucks and Truck Crash Involvement: Evidence from the Large Truck Crash Causation Study. *Transportation Research Record,* 2194(1), 21–28. https://doi.org/10.3141/2194-03
7. Zheng, Z., Lu, P., & Lantz, B. (2018). Commercial truck crash injury severity analysis using gradient boosting data mining model. *Journal of safety research*, 65, 115–124. https://doi.org/10.1016/j.jsr.2018.03.002
8. Liang, Yuanning, Do Safety Inspections Improve Safety? Evidence from the Roadside Inspection Program for Commercial Vehicles (May 16, 2021). Available at SSRN: https://ssrn.com/abstract=3847396 or http://dx.doi.org/10.2139/ssrn.3847396
9. Becker, G. S. (2000). Crime and Punishment: an Economic Approach BT - The Economic Dimensions of Crime (N. G. Fielding, A. Clarke, & R. Witt (eds.); pp. 13–68). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-62853-7_2
10. FMCSA *vehicle maintenance basic factsheet - transportation.* (n.d.). Retrieved July 28, 2022, from https://csa.fmcsa.dot.gov/Documents/FMC_CSA_12_009_BASICs_VehMaint.pdf
11. Centers for Disease Control and Prevention. (2017, June 1). Data Access - Urban Rural classification scheme for Counties. *Centers for Disease Control and Prevention.* Retrieved July 19, 2022, from https://www.cdc.gov/nchs/data_access/urban_rural.htm
12. Bureau, US Census. "Data." Census.gov, December 17 2021, https://www.census.gov/programs-surveys/popest/data.html.
13. Peck, D., Scott Matthews, H., Fischbeck, P., & Hendrickson, C. T. (2015). Failure rates and data driven policies for vehicle safety inspections in Pennsylvania. *Transportation Research Part A: Policy and Practice,* 78, 252–265. https://doi.org/10.1016/j.tra.2015.05.013
14. Beydoun. (2006). Vehicle characteristics and emissions: Logit and regression analyses of I/M data from Massachusetts, Maryland, and Illinois. *Transportation Research.*, (1), 59–76. https://doi.org/info:doi/
15. Perr-Sauer, Jordan, Duran, Adam W., & Phillips, Caleb T. Clustering Analysis of Commercial Vehicles Using Automatically Extracted Features from Time Series Data. United States. https://doi.org/10.2172/1597242
16. Das, S., Geedipally, S. R., Dixon, K., Sun, X., & Ma, C. (2019). Measuring the Effectiveness of Vehicle Inspection Regulations in Different States of the U.S. *Transportation Research Record,* 2673(5), 208–219. https://doi.org/10.1177/0361198119841563
17. Hopkins, B., & Skellam, J.G. (1954). A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany, 18,* 213-227.
18. Aggarwal, C. C. and Reddy, C. K. (eds.). *Data Clustering: Algorithms and Applications.* CRC Press, 2014. ISBN 978-1-46-655821-2.
19. Silge, J., and D. Robinson. *Text Mining with R. O'Reilly*, 2018. https://www.tidytextmining.com/. Accessed July 27, 2018.

20. Steven Bird. 2006. NLTK: The Natural Language Toolkit. In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

21. Daniel Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, B. Pfetsch, G. Heyer, U. Reber, T. Häussler, H. Schmid-Petri & S. Adam (2018) Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, *Communication Methods and Measures*, 12:2-3, 93-118, DOI: 10.1080/19312458.2018.1430754