NREL/CP-5400-70996. Posted with permission. Presented at WCX 18: SAE World Congress Experience, 10-12 April 2018, Detroit, Michigan.

2018-01-1199 Published 03 Apr 2018

Leveraging Big Data Analysis Techniques for U.S. Vocational Vehicle Drive Cycle Characterization, Segmentation, and Development

Adam Duran, Caleb Phillips, Jordan Perr-Sauer, Kenneth Kelly, and Arnaud Konan National Renewable Energy Laboratory

Citation: Duran, A., Phillips, C., Perr-Sauer, J., Kelly, K. et al., "Leveraging Big Data Analysis Techniques for U.S. Vocational Vehicle Drive Cycle Characterization, Segmentation, and Development," SAE Technical Paper 2018-01-1199, 2018, doi:10.4271/2018-01-1199.

Abstract

Under a collaborative interagency agreement between the U.S. Environmental Protection Agency and the U.S. Department of Energy (DOE), the National Renewable Energy Laboratory (NREL) performed a series of in-depth analyses to characterize on-road driving behavior including distributions of vehicle speed, idle time, accelerations and decelerations, and other driving metrics of medium- and heavy-duty vocational vehicles operating within the United States. As part of this effort, NREL researchers segmented U.S. medium- and heavy-duty vocational vehicle driving characteristics into three distinct operating groups or clusters using real-world drive cycle data collected at 1 Hz and stored in NREL's Fleet DNA database. The Fleet DNA database contains millions of miles of historical drive cycle data captured from medium- and heavy-duty vehicles operating across the United States. The data encompass existing DOE activities as well as contributions from valued industry stakeholder participants. For this project, data captured from 913 unique vehicles comprising 16,250 days of operation were drawn from the Fleet DNA database and examined. The Fleet DNA data used as a source for this analysis has been collected from a total of 30 unique fleets/data providers operating across 22 unique geographic locations spread across the United States. This includes locations with topographies ranging from the foothills of Denver, Colorado, to the flats of Miami, Florida. This paper includes the results of the statistical analysis performed by NREL and a discussion and detailed summary of the development of the vocational drive cycle weights and representative transient drive cycles for testing and simulation. Additional discussion of known limitations and potential future work is also included.

Introduction

In August of 2011, the U.S. Environmental Protection Agency (EPA) and the National Highway Traffic Safety Administration (NHTSA) adopted a set of national standards to reduce greenhouse gas (GHG) emissions and improve fuel efficiency of medium-duty (MD) and heavy-duty (HD) trucks [1, 2, 3, 4]. The jointly developed under the Energy Independence and Security Act, and the Clean Air Act, the GHG standards included the establishment on limits for carbon dioxide, nitrous oxide, and methane emissions. These limits would be enforced on model year 2014-2018 vehicles and have come to be known as Phase 1 of the national GHG regulatory program. The development of the Phase 1 standards provided EPA, NHTSA, and the state of California with a set of fully aligned regulations, allowing manufacturers to build a single fleet of vehicles and engines for the U.S. market.

Soon after implementation of the Phase 1 regulations, in response to the President's Climate Action Plan in February of 2014, President Obama announced efforts to update existing MD/HD vehicle regulations in Phase 2 of EPA's national GHG program [5, 6, 7]. As part of the EPA's proposed Phase 2 rulemaking, the U.S. Department of Energy (DOE) and EPA partnered to support a targeted project to refine and evaluate appropriate duty cycles for tractor-trailers and vocational vehicles to be used as part of MD/HD vehicle certification procedures for GHG emission standards. The National Renewable Energy Laboratory (NREL) has provided technical support utilizing DOE Vehicle Technologies Office-supported data, tools, and expertise to assist these efforts.

NREL's experience with large transportation database projects, including Fleet DNA [8] and the Transportation Secure Data Center [9], provide the prerequisite capabilities for tackling data-intense problems. Additionally, NREL's data analysis tools including the Drive Cycle Rapid Investigation, Visualization, and Evaluation tool (DRIVE) [10, 11], are used to distill large volumes of on-road vehicle data into statistically representative subsets suitable for testing and evaluation purposes. While the data contained in both Fleet DNA and the Transportation Secure Data Center may not be statistically representative of the entire U.S. population of commercial vehicles across all vocations, vehicle builds, and applications, NREL researchers believe this to be the most extensive database used to date to generate representative drive cycles. These skill sets combined with NREL's long-standing efforts in evaluating the on-road performance of conventional and advanced technology MD/HD vehicles for large commercial fleets are valuable tools to provide enhanced information for the EPA Phase 2 GHG rulemaking.

Segmenting the U.S. Vocational Vehicle Population

NREL researchers utilized a set of eight metrics describing each drive cycle to define representative segments through cluster analysis. Principal components analysis (PCA) and cross-correlation analysis were used to identify which of these eight metrics provide the greatest amount of information on which to support segmenting vehicle drive patterns. Cluster analysis was used to find an optimal grouping of vehicle drive cycles given these metrics. Finally, a resampling scheme was developed to explore possible impacts of sample bias in the Fleet DNA data set on the resulting segmentation.

The metrics in this study were chosen given their role in previous NREL research to characterize impacts of drive cycle on vehicle fuel economy and emissions production. The metrics used in the analysis were:

- Aerodynamic Speed (ft/s) (AS) - Describes the positive tractive energy required to overcome aerodynamic drag per unit distance over a given drive cycle. It is defined as:

$$AS = \frac{\sum_{i=1}^{N-1} pos(\frac{1}{2}\times\left(velocity_{i+1}-velocity_{i}\right)+g\times\left(h_{i+1}-h_{i}\right)}{D}$$

- Where: D = drive cycle cumulative distance (ft)
- g = gravitational constant (ft/s^2)
- h = height of vehicle indexed from start of drive cycle in time by i (ft)
- velocity = speed of vehicle indexed from start of drive cycle in time by i (ft/s)
- pos = positive only values

- Characteristic Acceleration (ft/s^2) (CA) - Describes the positive tractive energy required to accelerate/raise a vehicle per unit distance over a given drive cycle. It is defined as:

$$CA = \frac{\sum_{i=1}^{N-1} \bar{v}_{i,i+1}^{3}\times\left(t_{i+1}-t_{i}\right)}{D}$$

$$\bar{v}_{i,i+1}^{3} = \frac{v_{i+1}^{3}+v_{i+1}^{2}+v_{i}+v_{i}^{2}+v_{i}^{3}}{4}$$

- Where: D = drive cycle cumulative distance (ft)
- t = drive cycle time indexed from start of drive cycle in time by i (s)
- v = speed of vehicle indexed from start of drive cycle in time by i (ft/s)

- Percent of total cycle distance accumulated at speeds below 55 mph
- Percent of total cycle time duration accumulated at vehicle speeds of 0 mph
- Number of vehicle stops per mile
- Mean (nonzero) driving speed (mph)
- Maximum driving speed (mph)
- Standard deviation of (nonzero) driving speed (mph).

For this analysis, NREL utilized a collection 16,250 daily drive cycles from 913 vehicles. Among those vehicles, 108 (5,071 cycles) are long-haul trucks with sleeper, and 754 (10,765) are vocational vehicles as shown in Figure 1. Also contained in the database are 51 vehicles (414 cycles) classified as unknown vocation but possessing drive cycle data. Results for the entire sample population, as well as detailed results for the vocational segment, will be presented in the following subsections. Additionally, for the vocational segment, resampling of the database based on reliable estimates of the true vehicle population in the United States was performed to ensure the sample data set possessed a composition that is representative.

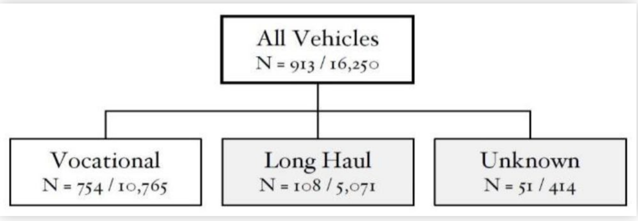
Performing Statistical Analysis

To explore the variability in the eight-metric data set, NREL researchers performed a pairwise correlation and a PCA on the Fleet DNA data set.

Linear Dependence and Correlation

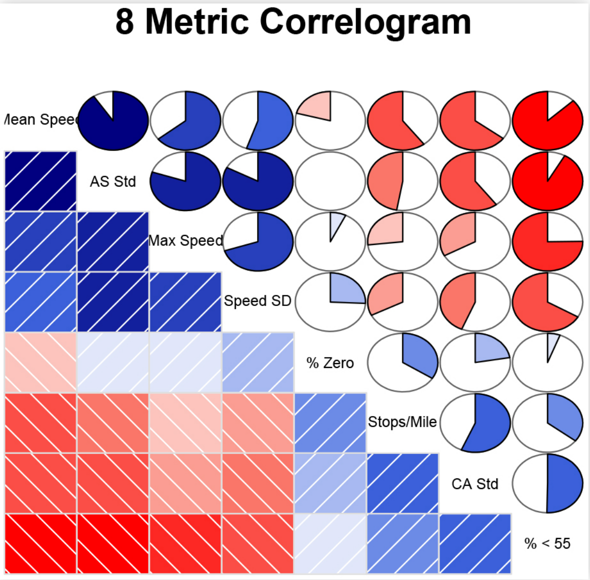
A correlogram (Figure 2) gives a visual indication of the degree of linear correlation (dependence) between each combination of variables and can be used to select an independent subset among a large number of potential variables. The variables examined include average driving speed (mean speed), aerodynamic speed (AS Std), maximum driving speed (max speed), standard deviation of driving speed (Speed SD), percentage of total cycle time at zero speed (% Zero), number of stops per mile (Stop/Mile), characteristic acceleration (CA Std), and percentage of total drive cycle mileage accumulated below 55 mph (% < 55). For instance, one can see from this plot that mean speed and percentage below 55 mph are strongly inversely correlated as shown by the dark red box and pie chart, while the mean speed and the aerodynamic speed (AS Std) are highly positively correlated as shown by the dark blue box and pie chart.

FIGURE 1 Sample composition



© SAE International

FIGURE 2 A correlogram for the eight metrics



© SAE International

TABLE 1 Principal component loadings (all vehicles)

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AS Std	0.44	−0.1	0.13	−0.09	−0.02	−0.26
CA Std	−0.34	−0.12	0.54	0.37	0.54	−0.4
Percent below 55 mph	−0.42	−0.04	−0.21	0.19	0.07	0.39
Percent at 0 mph	−0.24	−0.57	−0.11	−0.72	0.29	−0.04
Stops per mile	−0.34	0	0.58	−0.25	−0.68	−0.07
Mean driving speed (mph)	0.43	0.13	0.1	−0.2	0.11	−0.29
Max driving speed (mph)	0.36	−0.21	0.53	0	0.19	0.71
Driving speed SD (mph)	0.17	−0.76	−0.14	0.45	−0.34	−0.13

© SAE International

TABLE 2 Principal component loadings (vocational vehicles)

Variable	PC1	PC2	PC3	PC4	PC5	PC6
AS Std	0.44	0.1	0.1	0.06	−0.02	0.24
CA Std	−0.33	0.13	0.53	−0.55	0.36	0.41
Percent below 55 mph	−0.42	0.01	−0.26	−0.22	−0.02	−0.39
Percent at 0 mph	−0.2	0.72	−0.22	0.43	0.44	0.1
Stops per mile	−0.32	0.1	0.59	0.49	−0.51	0.04
Mean driving speed (mph)	0.43	−0.15	0.07	0.19	0.23	0.3
Max driving speed (mph)	0.35	0.23	0.48	−0.07	0.25	−0.72
Driving speed SD (mph)	0.27	0.6	−0.13	−0.42	−0.55	0.1

© SAE International

Principal Component Analysis

The PCA is a dimensionality reduction process that allows us to describe a higher dimensional data set with a smaller number of dimensions that can be easily visualized (e.g., in two dimensions). In PCA, each observation can be described by a weighted sum of orthogonal loadings (Eigen vectors; principal components). As its focus is dimensionality reduction, the orthogonal loadings are always fewer than the number of starting dimensions.

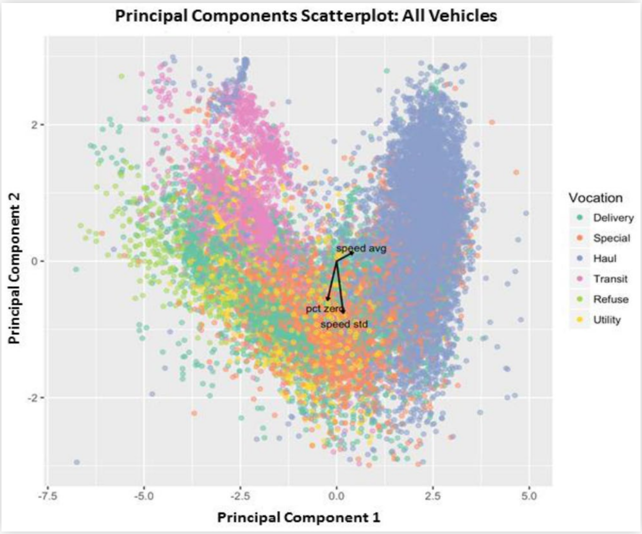
A PCA on the eight metric drive cycle data set suggests that two components are able to describe 75% of the variance in the complete data set. The first four components are able to describe 91% of the variance, and the first six components describe 99%. As each component is a weighted sum of the eight metrics, each component contains information from some combination of the metrics and some are weighted more heavily than others. All eight components together would reproduce the original eight-dimensional data set exactly.

Tables 1 and 2 give the PCA loadings (rotation) for all vehicles and for vocational vehicles, respectively. The first principal component (PC1) describes the greatest amount of variance in the data using primarily driving speed and standard deviation. The second principal component (PC2) most heavily weights the percentage of time spent at zero mph. These metrics are heavily weighted in the first two components because they account for the greatest degree of variance among the samples. Figures 3 and 4 show how the population of drive cycle characteristics is distributed in a two-dimensional PCA space for all vehicles and for vocational vehicles, respectively. Each point describes the position in the feature space and hence describes the driving characteristics for one day of driving for one vehicle. In this figure, some vocational categories are more distinct in their characteristics than others-Transit (which includes school buses) and Haul vehicles are most similar within their categories (i.e., are most tightly and consistently clustered) although the variance within those groups is relatively high. Refuse and transit vocations stand out in their usage as compared to the rest of the sample. In both plots the rotation of a subset of underlying variables is given at the center of the plot. Note that with vocational vehicles, the optimal PCA loadings are slightly different and the second component (PC2) changes sign, which causes the resulting plots to appear flipped relative to one another. In both plots, there are two modalities visible in this data set that appear to be largely differentiated by maximum speed and variability in speed, with a large cluster of mixed-mode driving somewhere in between.

Drive Cycle Clustering

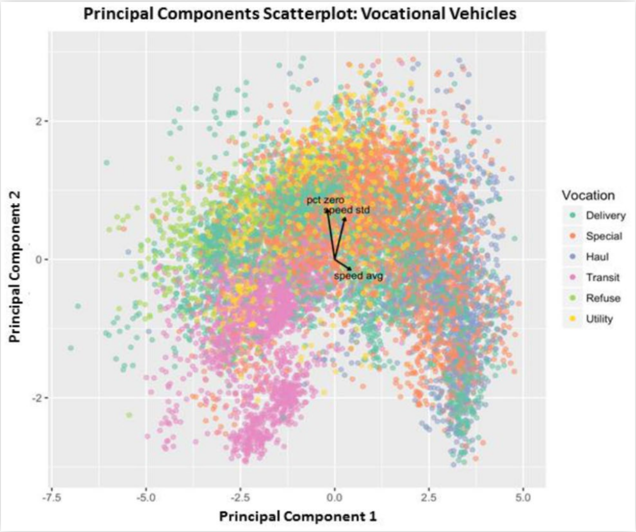
NREL researchers aimed to group similar drive cycles and derive representative drive cycle characteristics based on

FIGURE 3 All drive cycles for all vehicles visualized in the space defined by the first two principal components. Long-haul vehicles are present in a high-speed grouping. Color indicates vocation.



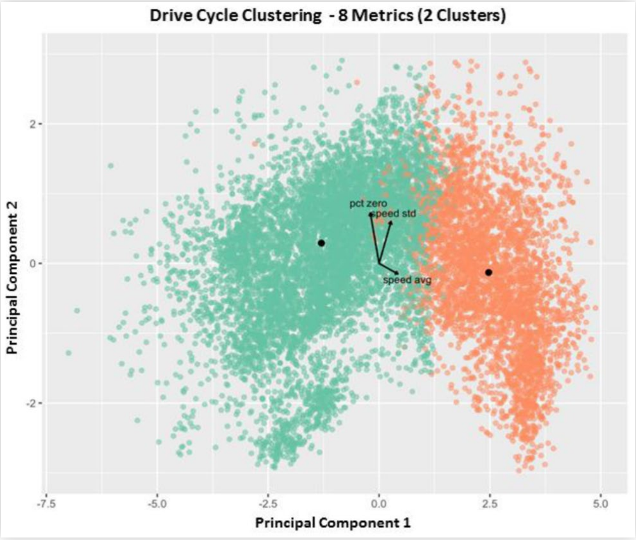
© SAE International

FIGURE 4 Scatter plot of all vocational drive cycles. Refuse and transit vocations stand out in their usage as compared to the rest of the sample. Note that PC2 changes signs in this plot, which makes it appear inverted relative to Figure 3.



© SAE International

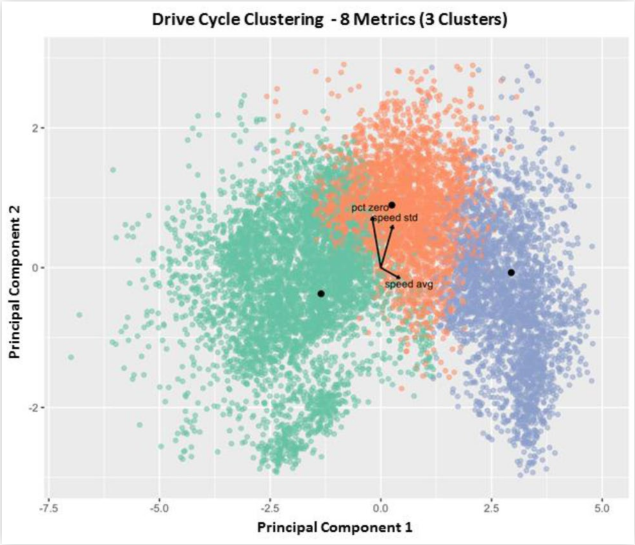
FIGURE 5 Two clusters using k-medoids algorithm with cluster centers marked in black. Only vocational vehicles are drawn in this plot.



© SAE International

the central tendency of each cluster using a variety of data clustering methods. After some work evaluating various clustering methods for the data (K-means, hierarchical, etc.), the K-medoids algorithm was selected as the best candidate. K-medoids functions well on large data sets by optionally clustering random subsamples. This algorithm works by first randomly selecting a number of cluster centers specified a priori. It then assigns all points to the closest cluster. New cluster centers are chosen, and the operation repeats until a convergent set of optimal clusters is found. Unlike K-means that provide metric averages for each cluster, the K-medoids algorithm selects a most-representative data point that improves the interpretability of the results [12, 13]. To determine an optimal number of clusters in the data, we utilize the silhouette method described in [14] where clustering is performed sequentially starting with two clusters and iterating with additional clusters. At each iteration, the silhouette analysis describes the ratio between tightness (within cluster variance) and separation (between cluster variance). The optimal clustering is one that provides the greatest separation between clusters while being robust to small numbers of outlier points and keeping the number of clusters as small as possible. Based on this analysis, we select two clusters as the optimal clustering in the metric space. Figure 6 shows

FIGURE 6 Three clustering using k-medoids algorithm with cluster centers marked in black. Only vocational vehicles are shown in this plot.



© SAE International

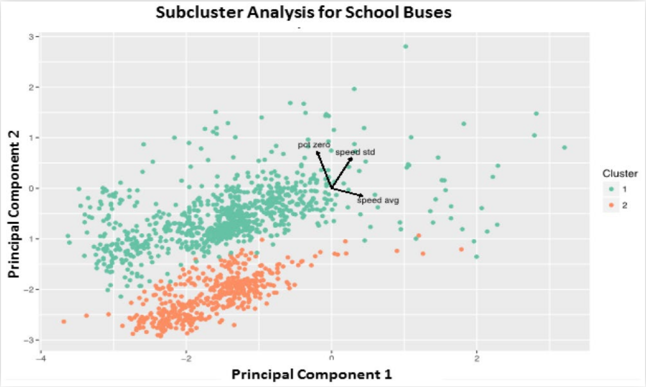
the resulting two clustering for vocational vehicles, which finds the underlying bimodal structure and splits the space logically. A small section of detached vehicles at the bottom of the left cluster are school buses with higher percentage of time at zero speed compared to the other vehicles in the cluster. Figure 8 focuses on school buses specifically with a sub-cluster analysis. Here it can be seen that there is additional structure present: school buses fall into two separate clusters within their own data. The first cluster is typified by its center point with 36.6% of time spent at zero speed while the center point of the lower cluster spends 3.3% of time at zero speed. As the vehicles in these two groups do not seem characteristically different, more investigation is needed to understand why school bus driving characteristics are partitioned this way. The balance of vocations appears relatively homogenous within the two dominant modalities.

For the sake of clarity, in the following sections we will refer to the left cluster, which contains slower speed cycles with more stops, as the "Slow" cluster and the right cluster, which contains higher speed cycles with fewer stops, as the "Fast" cluster. In this plot, as in the prior plots, each point represents one day of driving in the entire data set. Points are colored according to their optimized cluster placement. A single vehicle may have any number of drive cycles, which may over- or under-represent individual vehicles in these plots.

For regulatory purposes where it may be useful to consider three classes of vehicles, an optimal solution with three clusters was calculated for the vocational vehicle drive cycle sample. Figure 7 shows this data. The first two clusters are joined in this plot by a middle cluster that contains those traces that do not clearly fall into either the left (slower) or right (faster) cluster. Tables 4 and 5 contain the drive cycle characteristics for the center point of each cluster along with the vocation of the vehicle located at the medoid center.

To utilize cluster centers to define a representative drive cycle for testing/modeling purposes, the drive cycle characteristics from the top 50 days of data closest to these centers as ranked by multivariate least squares distance from the centroid were combined to determine the cluster averages for each metric.

FIGURE 7 Sub-cluster analysis for school buses, which themselves fall into two clusters.



© SAE International

© 2018 National Renewable Energy Laboratory.

TABLE 3 Vocational drive cycles - 2-clustering centers

	Left (Slow) Cluster	Right (Fast) Cluster
AS Std (ft/s)	54.59	81.21
CA Std (ft/s²)	0.48	0.28
Percent mileage below 55 mph	91.08	26.62
Percent time at 0 mph	50.62	22.11
Mean stops/mile	2.68	0.24
Mean driving speed (mph)	21.26	43.76
Max driving speed (mph)	59.22	65.01
Driving speed Std Dev. (mph)	16.28	21.06

© SAE International

[a] Note: The medoid vehicle may be representative for the overall cluster, while being abnormal for its own vocation. For instance, in the three-cluster solution the school bus that is selected as the medoid vehicle spends far less time at zero mph than other school buses. Nevertheless, this one vehicle is most representative of the entire class of vehicles irrespective of normality or abnormality for its own vocation.

TABLE 4 Vocational drive cycles - 3-clustering centers

	Left (Slow) Cluster	Middle Cluster	Right (Fast) Cluster
Medoid Cycle Vocation	School Bus	Towing	Freight
AS Std (ft/s)	48.88	69.27	84.85
CA Std (ft/s²)	0.5	0.68	0.2
Percent mileage below 55 mph	97.85	64.46	25.38
Percent time at 0 mph	47.87	52.68	25.43
Mean stops/ mile	1.22	0.98	0.26
Mean driving speed (mph)	24.12	34.5	47.7
Max driving speed (mph)	62.61	67.86	70.79
Driving speed Std Dev. (mph)	13.03	18.75	20.48

© SAE International

TABLE 5 Vocational resampling by MOVES categories

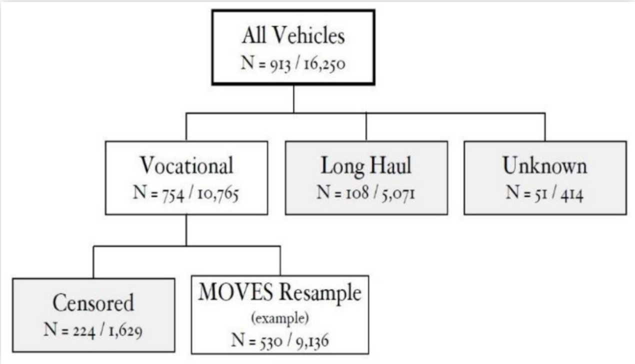
Vocation	Entire Fleet DNA Sample Population	MOVES Resampled Fleet DNA Population
Short Haul	441 (58.9%)	441 (84.1%)
School Bus	240 (23.6%)	57 (10.9%)
Refuse	50 (5.3%)	20 (3.9%)
Transit Bus	23 (2.1%)	6 (1.1%)

© SAE International

Resampling Data - Comparison to Existing Models

Although the Fleet DNA database provides detailed data for a large number of vehicles, to draw broad conclusions about how vehicles behave we must address sources of potential bias

FIGURE 8 Resampling flow for vocational vehicles



in this data set. This bias may arise simply because those fleets most willing to contribute data to Fleet DNA may not be perfectly representative of the entire population of vehicles in the United States. As shown in Figure 8, in this section we perform resampling based on EPA's MOtor Vehicle Emission Simulator (MOVES) study categorization, compute new clusters based on the new vehicle population, and then measure the difference in cluster center [15].

Using this analysis, it is possible to determine the degree to which the entire Fleet DNA sample population differs from the population of vehicles proportioned according to estimated U.S. population statistics. It is important to note that due to differences in vehicle classification systems between the Fleet DNA and MOVES databases, it was necessary for NREL researchers to aggregate Fleet DNA vehicle types into broader categorical groupings to match those of the MOVES and MOVES subcategory designations for resampling. In the case of a global MOVES resampling, the detailed list of vehicle types and vocations discussed in Section 1 were aggregated into four major vehicle categories as shown in Table 5.

Table 5 gives the MOVES-equivalent vehicle counts and proportions for resampling as compared to the entire Fleet DNA data. The four classes listed here are common to both the Fleet DNA data and the MOVES data. The Fleet DNA data appear to have a smaller fraction of short-haul vocational vehicles while having relatively more school buses, refuse trucks, and transit buses. The resulting resampled population would contain 230 fewer vehicles if resampled according to these proportions. A MOVES-equivalent sub-category resampling, which includes vehicle weight class, is given in Table 6. To match proportions of each category within weight classes, the total population of vehicles must be decreased by 306 (nearly half of all vocational vehicles).

To evaluate the impact of resampling the underlying data on the extant drive cycle clusters, a 10-fold evaluation was performed, where 10 random subsamples of daily driving cycles that are consistent with the MOVES proportions were selected and used for cluster analysis. For instance, starting by selecting a random sample of vehicles from the Fleet DNA database that has the same categorical and fractional breakdown as given in Table 6, cluster analysis and statistical characterization were then performed. This process was repeated 10 times (folds) to obtain a notion of how much variability there was among random subsamples.

TABLE 6 Vocational resampling by MOVES sub-categories

Vocation	Entire Fleet DNA Sample Population	MOVES Fleet DNA Resampled Population
Short Haul - Class 4/5	68 (6.3%)	53 (34.6%)
Short Haul - Class 6/7	155 (14.4%)	29 (19.0%)
Short Haul - Class 8	224 (20.9%)	21 (12.4%)
Refuse - Class 6/7	2 (0.2%)	0 (0%)
Refuse - Class 8	55 (5.1%)	4 (2.6%)
School Bus - Class 6/7	212 (19.8%)	12 (7.8%)
School Bus - Class 8	27 (25.2%)	1 (0.7%)
Transit Bus - Class 6/7	3 (0.3%)	0 (0%)
Transit Bus - Class 8	20 (1.9%)	1 (0.6%)
N/A	306 (28.5%)	

Figures 9 and 10 show the center value of each of the eight metrics both with and without resampling for MOVES categories and subcategories. Figures 11 and 12 illustrate how quickly the values stabilize when averaging results from iterative resamplings for both MOVES categories and subcategories. These plots show that after just five resamplings there is not a meaningful degree of additional variability.

Development of a Representative Transient

Drive cycle has been shown to dramatically impact fuel consumption and emissions production for MD and HD vehicles [16, 17, 18, 19, 20, 21]. As such, controlled laboratory test procedures representative of real-world operating

FIGURE 9 Difference between resampled cluster center statistics and original Fleet DNA (entire population) statistics for two clusters. The left pane shows the median value for the slow (cluster 1) population for each metric as a maroon dot. The right pane shows the same for the fast (cluster 2) population. In each pane, the boxplots give the distribution of median values with random MOVES-based resampling. Black dots represent outliers.

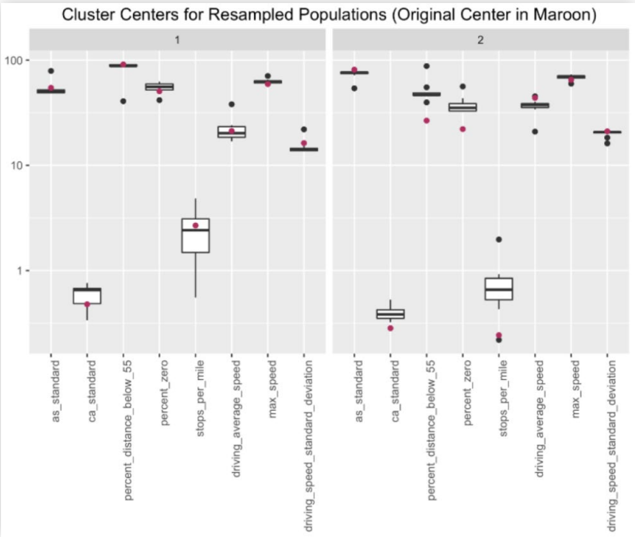
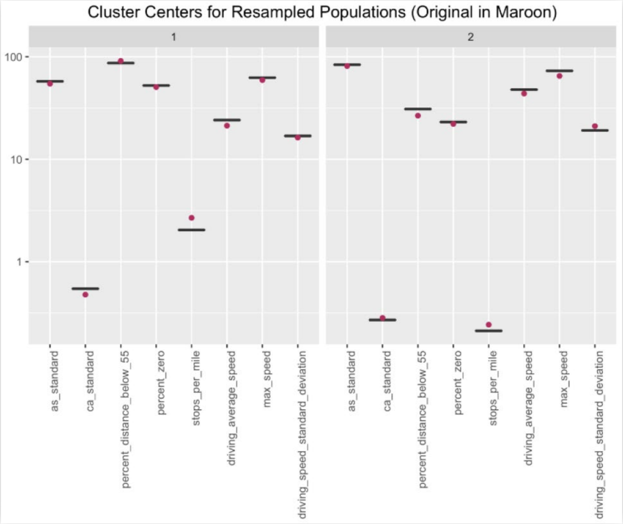
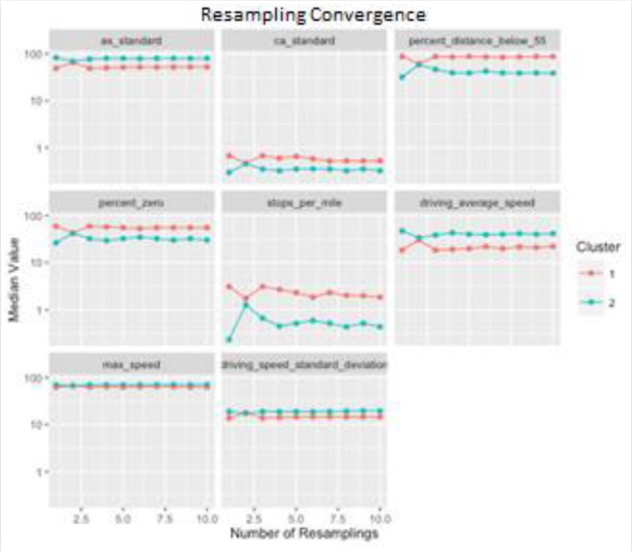


FIGURE 10 Difference between resampled cluster center statistics and original Fleet DNA (entire population) statistics for two clusters. This plot uses MOVES subcategory resampling. The left pane shows the median value for the slow (cluster 1) population for each metric as a red dot. The right pane shows the same for the fast (cluster 2) population. In each pane, the boxplots give the distribution of median values with random MOVES-based resamplings. The subcategory resampling has smaller variance in the metrics because of a smaller sample size.



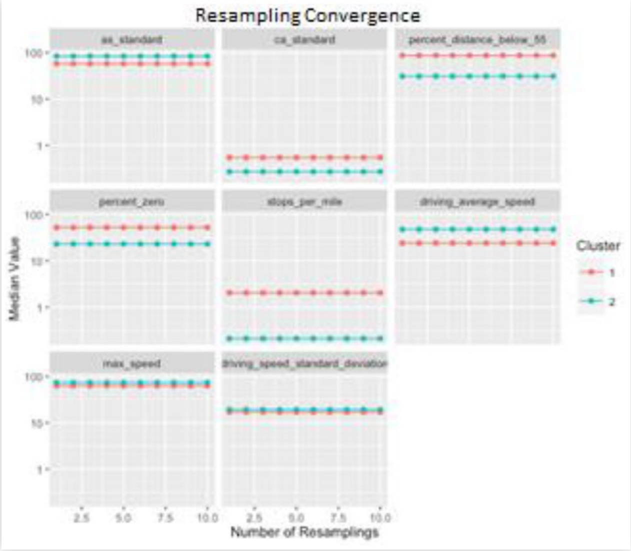
© SAE International

FIGURE 11 Convergence on final cluster center metrics as a function of successive resamplings



© SAE International

FIGURE 12 Convergence on final cluster center metrics as a function of successive MOVES subcategory resamplings



© SAE International

conditions are necessary to accurately quantify these parameters. Numerous approaches have been developed to generate representative drive cycles from real-world driving data [22, 23, 24, 25, 26], including NREL's DRIVE tool. Having already completed segmentation of the U.S. MD and HD commercial vehicle data into a collection of three distinct clusters based on a multivariate drive cycle clustering analysis and development of a logistic model to predict cluster participation, NREL researchers then applied the results of the clustering analysis towards the development of both a representative low- and high-speed transient drive cycle representative of MD and HD vocational vehicle use. Deploying NREL's DRIVE tool within its high-performance computing environment, researchers condensed thousands of hours of on-road driving data down into representative speed-time drive cycles approximately 12 and 20 minutes in duration.

Applying DRIVE™

To develop the representative low-speed transient cycle, an iterative method was deployed where drive cycles from each of the three clusters described in section 3 were ranked by root mean squared distance from the medoid calculated using the eight key metrics and then fed through DRIVE to generate candidate representative cycles which were then compared to the statistics for the low-speed cluster to identify an ideal representative cycle. The procedure was as follows:

1. The top 50 cycles from each of the three clusters were fed into DRIVE Space to generate candidate representative cycles. The top 50 cycles were chosen as a means of limiting overall computation time while still capturing the overall representative behavior of the cluster. This decision also allowed researchers to explore and optimize the generated drive cycles by adjusting final drive cycle duration.
2. Having identified the top 150 total drive cycles most representative of their respective clusters, a variety of input parameters such as the desired and minimum cycle durations were then adjusted to generate over 100 unique cycles.
3. The representative cycles were then compared to the average values for the top 50 cycles in the low speed cluster using a non-weighted least squares approach using the eight drive cycle metrics from the clustering

analysis described in section 3. The low-speed cluster was chosen as the ideal target due to its inherent low-speed transient behavior. This is especially true when compared to the mixed mode and high-speed clusters.

Several additional constraints were applied when running the DRIVE tool including:

- A cycle duration of 668 seconds was targeted to match the duration of the California Air Resources Board Heavy Heavy-Duty Diesel Truck (HHDDT) Transient Cycle.
- A minimum allowable cycle duration of 300 seconds was established to ensure sufficient test duration.
- Any drive cycles with a maximum speed in excess of 55 mph were excluded to ensure the generated drive cycle is representative of low-speed transient operation only.

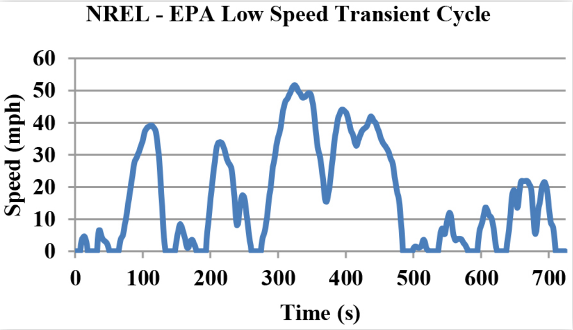
To examine the effects of population makeup on the resultant transient cycle, the drive cycle generation process was repeated a total of three times: one for each of the unique populations explored in the cluster analysis. The full Fleet DNA drive cycle population, the MOVES resampled drive cycle population, and the MOVES subcategory resampled drive cycle data were all examined - see section 3 for more details on each sample set. This was done to examine the sensitivity of representative transient cycle generation to source population and minimize the impact of any potential biasing as a result of source data composition (i.e., more school bus data than nationally representative). It was found that there were minimal differences between the weighting, and the MOVES resampled population was chosen because it produced the most representative drive cycle of the three populations. Additional detailed information regarding resampling and its impact on cycle generation can be found in the appendix.

The transient cycle shown in Figure 13 was developed using 150 drive cycles drawn from the MOVES resampled Fleet DNA database following the procedure described previously.

Key statistics for the final low-speed transient cycle include:

- 724 seconds in duration
- Total of 10 microtrips
- Maximum speed of ~52 mph
- Average driving speed of ~21 mph
- ~22% of total cycle duration is at zero speed.

FIGURE 13 Speed-time trace for representative low-speed EPA transient cycle



© SAE International

The target zero speed duration for the representative transient cycle was less than 24.5%. This value was chosen as it represents the percentage of zero speed time observed from the high-speed cluster identified in section 3 (cluster 3-3). The high-speed cluster possesses the lowest amount of zero speed time of all the clusters; thus, it was used as a lower bound when developing the low speed transient cycle because for any of the other driving conditions one can simply add an additional idle time segment to achieve an appropriate overall idle time weighting.

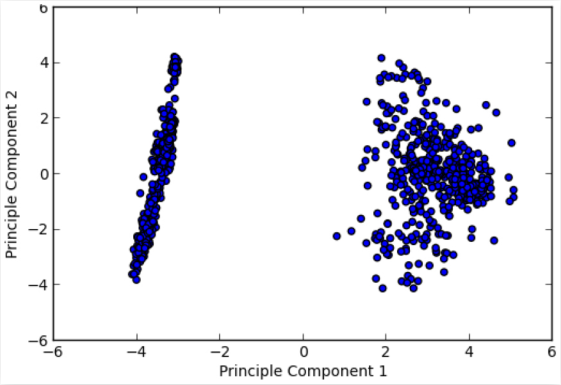
Known Limitations and Ongoing Work

The analysis presented here investigates the dynamics and modalities of common drive cycle characteristics within U.S. commercial fleets. While this analysis was performed with an extensive data set and care was taken to analyze potential impacts of sample/selection bias, there are some limitations and opportunities for future work that should be discussed.

Beyond the population of vehicles, these results are also dependent on the metrics or features used in clustering. We observe two strong modalities in drive patterns, but cluster separation is not strong between these two modalities. This result belies the difficulty in performing a cut between the two segments. One possible solution discussed here is the introduction of a third cluster that better represents the dynamics of those vehicles "between" the two clusters. However, an open question remains as to whether a different set of metrics may produce a result with greater separation.

To understand the effect of the choice of metrics on the resulting categorization, we have begun to investigate "domain-agnostic" feature extraction as a method for determining the best features for partitioning the vehicle drive dynamics. Using the approach, hundreds of candidate statistical features are extracted from the underlying time series data and are used as the basis of clustering [26, 27, 28]. These features are generic statistical descriptors and may not relate to industry notions of driving behavior. Figure 14 shows the result of one of these clustering methods when applied to the same subset of data shown in our results. These preliminary results show this method appears to segment the data with more separation. We

FIGURE 14 All vehicles visualized by the first two principal components in a space of domain agnostic parameters



© SAE International

believe this approach may hold promise for future clustering applications with massive data sets. However, this may come at the cost of direct interpretability. Approaching this challenge, among others, is an area of ongoing work.

During this study, the Fleet DNA database has continued to grow. Performing analyses like those presented here in a way that scales to arbitrary quantities of data requires some care. To support future analyses, we have begun development of the Fleet DNA Big Data application programming interface, which provides streamlined storage, aggregation, and analysis functions using an industry standard big data framework that utilized distributed processing libraries Spark and Hadoop distributed filesystem.

Conclusion

In this study we demonstrated a new method for analyzing the structure and modalities of drive cycle dynamics. In doing so, we have leveraged a scalable data framework and a large and diverse data set of 913 unique commercial vehicles' drive cycle data comprising 16,250 days of operation. Our method utilizes aggregate feature extraction, k-medoids clustering, and the NREL DRIVE tool to find the fundamental modalities within the driving characteristics of this population of vehicles. Our results describe representative drive cycles for vocational vehicles in two (or three) classes that can be used for regulatory and testing applications.

References

1. U.S. Environmental Protection Agency, "EPA and NHTSA Adopt First-Ever Program to Reduce Greenhouse Gas Emissions and Improve Fuel Efficiency of Medium- and Heavy-Duty Vehicles," EPA-420-F-11-031, Office of Transportation and Air Quality, 2011, http://www.epa.gov/otaq/climate/documents/420f11031.pdf, accessed Feb. 6, 2016.
2. U.S. Environmental Protection Agency, "Final Rulemaking to Establish Greenhouse Gas Emission Standards and Fuel Efficiency Standards to Medium- and Heavy-Duty Engines and Vehicles: Regulatory Impact Analysis," EPA-420-R-11-901, Office of Transportation and Air Quality and National Highway Traffic Safety Administration, 2011, http://www.epa.gov/otaq/climate/documents/420r11901.pdf, accessed Jan. 13, 2016.
3. Federal Register, "Greenhouse Gas Emissions Standards and Fuel Efficiency Standards for Medium- and Heavy-Duty Engines and Vehicles: Final Rule," Environmental Protection Agency and Department of Transportation, 76(179), 2011, http://www.gpo.gov/fdsys/pkg/FR-2011-09-15/pdf/2011-20740.pdf, accessed Feb. 16, 2016.
4. U.S. Environmental Protection Agency, "Transportation and Climate: Regulations & Standards: Heavy-Duty," Office of Transportation and Air Quality, Feb. 10, 2015, http://www.epa.gov/otaq/climate/regs-heavyduty.htm, accessed Feb. 16, 2015.
5. The White House, Office of the Press Secretary, "Remarks by the President on Fuel Efficiency Standards of Medium and Heavy-Duty Vehicles," Feb. 18, 2014, http://www.whitehouse.gov/the-pressoffice/2014/02/18/remarks-president-fuel-efficiency-standards-medium-and-heavyduty-vehicl, accessed Feb. 16, 2015.
6. The White House, Office of the Press Secretary, "FACT SHEET: Opportunity for all: Improving the Fuel Efficiency of American Trucks - Bolstering Energy Security, Cutting Carbon Pollution, Saving Money and Supporting Manufacturing Innovation," Feb. 18, 2014,http://www.whitehouse.gov/the-press-office/2014/02/18/fact-sheet-opportunity-allimproving-fuel-efficiency-american-trucks-bol, accessed Feb. 6, 2016.
7. U.S. Environmental Protection Agency, "EPA and NHTSA Propose Standards to Reduce Greenhouse Gas Emissions and Improve Fuel Efficiency of Medium- and Heavy-Duty Vehicles for Model Year 2018 and Beyond," EPA Fact Sheet EPA-420-F-15-901, 2015, http://www3.epa.gov/otaq/climate/documents/420f15901.pdf.
8. Walkowicz, K., Kelly, K., Duran, A., and Burton, E., "Fleet DNA Project Data," National Renewable Energy Laboratory, 2014, http://www.nrel.gov/fleetdna.
9. Transportation Secure Data Center, National Renewable Energy Laboratory, 2014, www.nrel.gov/tsdc, accessed Oct. 31, 2014.
10. National Renewable Energy Laboratory, "DRIVE Analysis Tool Generates Custom Vehicle Drive Cycles Based on Real-World Data," NREL/FS-5400-54507 Golden, CO: National Renewable Energy Laboratory, 2013, http://www.nrel.gov/docs/fy13osti/54507.pdf, accessed Feb. 16, 2015.
11. NREL, "NREL Vehicle Drive Cycle Tool, User Guide, Alliance for Sustainable Energy," LLC, 2009.
12. Kaufman, L. and Rousseeuw, P.J., "Partitioning Around Medoids (Program PAM)," . In: *Finding Groups in Data: An Introduction to Cluster Analysis*. (Hoboken, John Wiley & Sons, Inc., 1990), doi:10.1002/9780470316801 Chapter 2.
13. Kaufman, L. and Rousseeuw, P.J., "Clustering Large Applications (Program CLARA)," . In: *Finding Groups in Data: An Introduction to Cluster Analysis*. (Hoboken, John Wiley & Sons, Inc., 1990), doi:10.1002/9780470316801 Chapter 3.
14. Rousseeuw, P.J., "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* 20:53-65, 1987 ISSN 0377-0427, http://dx.doi.org/10.1016/0377-0427(87)90125-7.
15. U.S. Environmental Protection Agency, "MOVES (Motor Vehicle Emissions Simulator)," Office of Transportation and Air Quality, Dec. 1, 2015, https://www.epa.gov/moves/moves2014a-latest-version-motor-vehicle-emission-simulator-moves, accessed Feb. 4, 2016.
16. Sharer, P., Leydier, R., and Rousseau, A., "Impact of Drive Cycle Aggressiveness and Speed on HEVs Fuel Consumption Sensitivity," SAE Technical Paper 2007-01-0281, 2007, doi:10.4271/2007-01-0281.
17. Fellah, M., Singh, G., Rousseau, A., Pagerit, S. et al., "Impact of Real-World Drive Cycles on PHEV Battery Requirements," SAE Technical Paper 2009-01-1383, 2009, doi:10.4271/2009-01-1383.
18. Sun, Z. and Andreae, M., "Vehicle Duty Cycle Characteristics for Hybrid Potential Evaluation," SAE Technical Paper 2012-01-2023, 2012, doi:10.4271/2012-01-2023.

19. Stichter, J., "Investigation of Vehicle and Driver Aggressivity and Relation to Fuel Economy Testing," Master's thesis, 2012, http://ir.uiowa.edu/cgi/viewcontent.cgi?article=3542&context=etd, accessed Feb. 6, 2016.
20. Berry, I., "The Effects of Driving Style and Vehicle Performance on the Real-World Consumption of U.S. Light-Duty Vehicles," Master's thesis, 2012, http://web.mit.edu/sloan-auto-lab/research/beforeh2/files/IreneBerry_Thesis_February2010.pdf, accessed Feb, 2, 2016.
21. Reinhart, T.E., "Commercial Medium- and Heavy-Duty Truck Fuel Efficiency Technology Study - Report #1," Report No. DOT HS 812 146, Washington, DC: National Highway Traffic Safety Administration, June 2015.
22. Lee, T. and Filipi, Z., "Synthesis of Real-World Driving Cycles Using Stochastic Process and Statistical Methodology," *Int. J. Vehicle Design* 57(1):17-36, 2011, doi:10.1504/IJVD.2011.043590.
23. Tong, H. and Hung, W., "A Framework for Developing Driving Cycles with On-Road Driving Data," *Transport Reviews* 30(5):589-615, 2010, doi:10.1080/01441640903286134.
24. Park, J., Lee, J., and Lee, J., "Development of Driving Cycle for CO2 Emission Test of Heavy-Duty Vehicles," SAE Technical Paper 2013-01-2520, 2013, doi:10.4271/2013-01-2520.
25. Moore, W., Finch, T., and Sutton, M., "Development of Heavy Duty Diesel Real World Drive Cycles for Fuel Economy Measurements," SAE Technical Paper 2013-01-2568, 2013, doi:10.4271/2013-01-2568.
26. Kulkarni, A., Sapre, R., and Sonchal, C., "GPS Based Methodology for Drive Cycle Determination," SAE Technical Paper 2005-01-1060, 2005, doi:10.4271/2005-01-1060.
27. Wang, X., Smith, K., and Hyndman, R., "Characteristic-Based Clustering for Time Series Data," *Data Mining and Knowledge Discovery* 13(3):335-364, 2006.
28. Christ, M., "TSFresh," https://github.com/blue-yonder/tsfresh.

Contact Information

Adam Duran is a senior research engineer working with the Transportation and Hydrogen Systems Center at the National Renewable Energy Laboratory. Adam's work focuses primarily in the areas of drive cycle analysis and characterization, custom drive cycle development, and medium/heavy-duty fleet evaluations. He may be reached at Adam.Duran@nrel.gov.

Dr. Caleb Phillips is a data scientist with the Data Analysis and Visualization Group within the Computational Sciences Center at the National Renewable Energy Laboratory. Caleb comes from a background in computer science systems, applied statistics, computational modeling, and optimization. His work at NREL spans the breadth of renewable energy technologies and focuses on applying modern data science techniques to data problems at scale. He may be reached at Caleb.phillips@nrel.gov.

Jordan Perr-Sauer is an RPP intern with the Data Analysis and Visualization Group within the Computational Sciences Center at the National Renewable Energy Laboratory. Jordan hopes to use his professional background in Software Engineering and his academic training in Applied Mathematics to solve the challenging problems facing America and the world. He may be reached at Jordan.Perr-Sauer@nrel.gov

Ken Kelly currently manages NREL's Commercial Vehicle Technologies team partnering with government and industry to develop advanced vehicle technologies for medium- and heavy-duty applications. While at NREL, Ken has managed and conducted research in a wide array of advanced vehicle and transportation technologies including: medium- and heavy-duty vehicle technologies; alternative fuel vehicles; hybrid electric vehicle systems; advanced heat transfer technologies; computer aided engineering and robust design methods. He holds M.S. and B.S. degrees in Mechanical Engineering from Ohio University. He may be reached at Kenneth.kelly@nrel.gov.

Acknowledgments

The authors would like to thank the United States Environmental Protection Agency for its support in funding the research presented in this paper. Without their feedback this work would not have been possible. The authors would also like to thank Robert Prohaska and Eric Miller for their contributions to the work presented.

The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

Definitions/Abbreviations

AS - aerodynamic speed
CA - characteristic acceleration
DOE - U.S. Department of Energy
DRIVE™ - Drive Cycle Rapid Investigation and Visualization Tool
EPA - U.S. Environmental Protection Agency
GHG - greenhouse gas
HD - heavy duty
MD - medium duty
MOVES - MOtor Vehicle Emission Simulator
NHTSA - National Highway Transportation Safety Association
NREL - National Renewable Energy Laboratory
PCA - principal components analysis
TSDC - Transportation Secure Data Center

Positions and opinions advanced in this paper are those of the author(s) and not necessarily those of SAE International. The author is solely responsible for the content of the paper.

ISSN 0148-7191