POLITECNICO DI TORINO
Repository ISTITUZIONALE

Profiling industrial vehicle duties using CAN bus signal segmentation and clustering

Original
Profiling industrial vehicle duties using CAN bus signal segmentation and clustering / Buccafusco, Silvia; Megaro, Andrea; Cagliero, Luca; Vaccarino, Francesco; Salvatori, Lucia; Loti, Riccardo. - ELETTRONICO. - 2841:(2021), pp. 1-6. (Intervento presentato al convegno Workshops of the 24th International Conference on Extending Database Technology/24th International Conference on Database Theory, EDBT-ICDT 2021 tenutosi a Nicosia (Cyprus) nel March 23-26, 2021).

Availability:
This version is available at: 11583/2925804 since: 2021-09-20T19:40:24Z

Publisher:
OpenProceedings

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

12 October 2024

Profiling industrial vehicle duties using CAN bus signal segmentation and clustering

Silvia Buccafusco
Politecnico di Torino
Turin, Italy
silvia.buccafusco@polito.it

Andrea Megaro
Politecnico di Torino
Turin, Italy
andrea.megaro@asp-poli.it

Luca Cagliero
Politecnico di Torino
Turin, Italy
luca.cagliero@polito.it

Francesco Vaccarino
Politecnico di Torino
Turin, Italy
francesco.vaccarino@polito.it

Lucia Salvatori
Tierra spa
Turin, Italy
lsalvatori@topcon.com

Riccardo Loti
Tierra spa
Turin, Italy
rloti@tierratelematics.com

ABSTRACT

Industrial vehicles working in construction sites show rather heterogeneous usage patterns. Depending on its type, model, and context of usage, the vehicle workload may vary from light to heavy with variable periodicity. Duties summarize the current state of a vehicle according to its usage level. They are usually set up manually vehicle by vehicle according to the specifications of the manufacturer. To automate the definition of per-vehicle duty levels, this paper explores the use of clustering techniques applied to CAN bus signals. It first performs a segmentation of the CAN bus signals to identify specific working cycles. Then, it clusters the segments to support the definition of vehicle-specific duty levels. The preliminary results, acquired on real vehicle usage data, show the applicability of the proposed approach.

1 INTRODUCTION

The fleets of industrial vehicles that are commonly employed in construction sites by public and private enterprises show rather variable usage patterns. For example, refuse compactors, which are usually employed in dumps, drive few kilometers per day and work at light workload 24/7 for relatively long periods. Road rollers and tandem rollers, which are frequently used in road maintenance, drive few kilometers per day as well, but work at relatively heavy workload only for short periods. Conversely, forklift trucks, which are employed in warehouses, drive many kilometers per day, work most of the time at light workload, and accomplish specific tasks at heavy workload (e.g., the lift of a heavy pallet).

The advent of Controller Area Network (CAN) bus technology [11] has provided fleet managers with a huge amount of data useful for monitoring and analyzing vehicle usage. The CAN bus allows communication among the electronic control unit devices on board the vehicle. It provides direct access to various signals describing the vehicle state. CAN bus data usually consist of raw time series, which are sampled and aggregated before being transmitted to a central repository. Data regard fuel consumption, vehicle movements (e.g., accelerations and drifts), engine conditions (e.g., RPM, oil and coolant temperature), route characteristics (e.g., slope), and alarms. Domain experts can thus monitor the vehicle state by acquiring, collecting, and analyzing vehicle-specific CAN bus data through data mining and machine learning techniques in order to support fleet managers' decisions (e.g., [16, 17, 22, 23]).

To optimize vehicle usage fleet managers commonly need to monitor the time spent by the vehicles in specific duties. Engine duties describe the current state of a vehicle and are usually classified as (i) *long idle*, which indicates that the vehicle has been stationary and under a minimal workload level for a relatively long period, (ii) *idle*, which indicates that the vehicle has been stationary and under a minimal workload for a short period, (iii) *moving/working*, which indicates non-stationary vehicle usage with light workload, (iv) *light workload*, which indicates non-stationary vehicle usage with light workload, and (v) *heavy workload*, which indicates non-stationary vehicle usage with intensive workload. However, due to the high vehicle heterogeneity over models, types, and context of usage (e.g., ground type, use of vehicle equipment) duties are commonly defined manually by domain experts separately for each vehicle. This is not efficient, particularly time-consuming, and prone to errors.

To make the process of defining per-vehicle duty levels more efficient and effective, we propose to apply a clustering-based approach to the acquired CAN bus signals related to a shortlist of Suspect Parameter Numbers (SPNs). To this end, we make a preliminary attempt to directly cluster the raw SPN series in order to assign approximated, pointwise per-vehicle duties. For instance, Figure 1 shows three examples of SPNs (i.e., engine speed, fuel tank level, engine percent load) corresponding to a representative vehicle. The idle and working states defined by setting the usage level thresholds inferred according to the outcomes of the clustering algorithms are colored in green and red, respectively. Although they discriminate between instants of heavy and light workloads, they do not trace the underlying trends in temporal duty variations. Hence, the results is hardly usable by domain experts.

To get more precise and stable duty state levels, we devise a refined clustering strategy that groups fixed-length segments of CAN bus signals according to ad hoc descriptive features in both the frequency and temporal domains. Segments are produced by applying a motif discovery algorithm on the aligned and synchronized version of the raw SPN series. Figure 2 shows the output of the refined process, which exhibits the newly defined duties. The new states appear to be less susceptible to temporary usage level variations thus becoming usable for profiling vehicle usage.

The results were validated on real vehicle usage data acquired by a multinational company providing telematics services. The validation phase included qualitative and quantitatively analyses. The latter relied on both established clustering validity indices [2] and on a comparison between the assigned duty states and the

© 2021 Copyright for this paper by its author(s). Published in the Workshop Proceedings of the EDBT/ICDT 2021 Joint Conference (March 23–26, 2021, Nicosia, Cyprus) on CEUR-WS.org. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR-WS.org/Vol-2841/DARLI-AP_7.pdf

expected output according to the National Marine Electronics Association (NMEA) 0183 messages data [3].

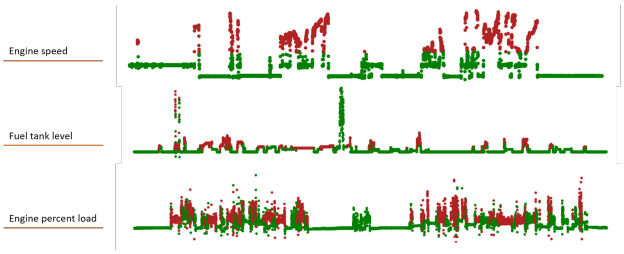

Figure 1: SPNs colored according to the rough duty levels (green=idle, red=working). Point-wise clustering

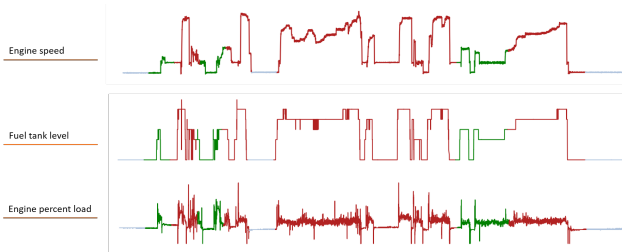

Figure 2: SPNs colored according to the refined duty levels (gray=idle, green=moving, red=working). Segment-wise clustering

The rest of the paper is organized as follows. Section 2 overviews the related literature. Section 3 describes the analyzed data. Sections 4 and 5 present the data preparation and mining phases, respectively. Section 6 summarizes the empirical results, whereas Section 7 draws conclusions and discusses the future developments of this research.

2 RELATED WORK

Clustering techniques have already been applied to analyze CAN Bus data acquired from vehicles. Examples of applications include, amongst other, (i) the optimization of vehicle routes (e.g., [5]), (ii) the identification of driver intentions based on trajectory analysis (e.g., [21]), (iii) the characterization of drivers' behavior (e.g., [7, 15]), (iv) the management of single vehicles and vehicles' fleets (e.g., [9]). The present work belongs to the latter category. To the best of our knowledge, this is the first attempt to automate the process of assigning per-vehicle duty levels based on CAN bus signal clustering.

In [9] the authors focused on explaining clusters mined from multivariate time series data over different time scales and granularity. As application scenario, they analyzed the usage profile of vehicles travelling across urban areas with the aim at planning and supporting maintenance operations. To this purpose, they used a Gaussian mixture model to identify clusters on top of a subset of features extracted from the raw series according to a sliding window strategy. Rather than extracting aggregated statistics for all series over some time window and then identify clusters as indicators of more abstract states, our approach aims at partitioning the series to detect specific vehicle duties. In a nutshell, we cluster segments of SPN series instead of concise series representation. In [10] the authors proposed a reverse engineering strategy to extract undisclosed information about CAN bus configuration. Similar to the present work, the aforesaid research study aims at analyzing vehicle usage via CAN bus signal analysis. However, the research objective is substantially different.

Table 1: Analyzed SPNs.

SPN code (SAE J1939)	Description
81	Engine diesel particulate filter inlet pressure
90	Power takeoff oil temperature
94	Engine fuel delivery pressure
110	Engine coolant temperature
114	Net battery current
123	Clutch pressure
164	Engine injection control pressure
182	Engine trip fuel
183	Engine fuel rate
190	Engine speed
524	Transmission selected gear
975	Estimated percent fan speed
1638	Hydraulic temperature
Custom Number	Engine percent load
Custom Number	Front plow switch
Custom Number	Rear hitch position
Custom Number	Charge pressure
Custom Number	Amount of particulate matter C method
Custom number	Digging depth
Custom number	Fuel Tank Level

3 DATA OVERVIEW

Data were acquired from an experimental CAN bus data logger, which was installed on a test farm tractor working in a construction site. Data were provided by Tierra S.p.A, a multinational company operating in the IoT sector and internationally recognized for providing to their customers sophisticated and reliable telematics solutions for management, maintenance, and remote diagnostics of equipment.

The test vehicle is equipped with a large amount of sensors, which capture CAN parametric messages at a high frequency (up to 100 Hz). Messages were gathered and temporally stored on an SD card which manages data transmission to the cloud infrastructures. Then, raw data were decoded and transformed using the SAE J1939 protocol (https://www.sae.org/), which is established for heavy-duty vehicle manufacturers and provides a shared set of standard messages and conversion rules.

Customers of the telematics service provider can visualize and process in real time the converted data. Among the available vehicle usage indicators, the times spent by the vehicle in each duty (i.e., *long idle*, *idle*, *moving/working*, *heavy workload*) are among the mostly commonly used to optimize vehicle maintenance, production, business, and investments [20]. Unfortunately, the threshold levels used to define the vehicle states are not standardized since they depend on the particular vehicle model, type, and context of usage. Hence, typically, their setup is manually performed by domain experts. This prompts the need for data-driven approaches to automatically inferring the most suitable duty levels separately for each industrial vehicle.

The acquired dataset consists of the SPN series acquired from the test vehicle from November 7, 2019 to April 15, 2020. Each observation is described by SPN name, acquisition timestamp, and measurement. The dataset collects 20 different SPNs describing the state of the vehicle engine such as engine speed, percent load, fuel rate, coolant temperature, fuel delivery pressure (see the full list in Table 1). A more thorough SPN description can be found at www.sae.org/standards.

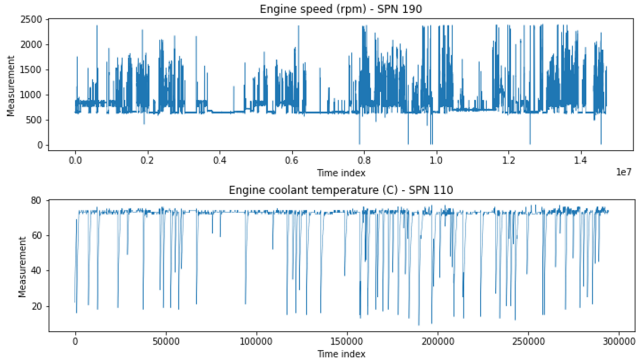

Figure 3: Extracts of the SPN series of engine speed and engine coolant temperature.

The acquired SPN series are highly heterogeneous, not synchronized with each other, and partly noisy. For example, Figure 3 shows two extracts of SPN series, i.e., the engine speed and the engine coolant temperature series. The former has a sampling rate of 50 Hz, whereas the latter 1 Hz. Furthermore, the series periodicity has different granularity over the working cycles.

4 DATA PREPARATION

To prepare the raw CAN bus data to the subsequent analyses we apply the following steps.

Data cleaning. To avoid introducing a bias in the clustering process, we removed missing values (due, for instance, to failures in data acquisition and transmission) and properly managed the presence of noise, decoding errors, and inconsistencies in the SPNs series values. Specifically, for each pair timestamp and SPN we computed an average value to bound data points to the feasible and operative range of the corresponding measured physical quantity.

Working cycle identification. CAN messages are transmitted only when a vehicle is on. To analyze vehicle duties it can be useful to understand whether a vehicle has been turned off at the end of a working cycle or for any other reason. For this reason, at the vehicle restart we analyze the value of the engine coolant temperature as it indicates, to a good approximation, when a vehicle has been turned off for a sufficiently long time.

Series alignment and synchronization. CAN bus messages are asynchronously transmitted at variable rates over the network. For example, SPNs such as engine speed, engine percent load and charge pressure are transmitted quite frequently (sampling rate between 20 Hz and 50 Hz), whereas engine coolant temperature and engine delivery fuel pressure are sent less frequently (rate between 1 Hz to 2 Hz). Hence, to enable SPN series clustering we re-align and synchronize all the analyzed SPN series. To this aim, CAN bus signals are first linearly interpolated in the temporal domain and then down-sampled in frequency domain to the least average sampling rate to align multiple signals (using standard anti-aliasing filters and down-sampling operators).

SPNs selection. We select the subset of SPNs that most likely influence the vehicle duty. To this purpose, we firstly filter out all the SPNs providing less relevant information. For example, according to the manufacturers' specification the engine trip fuel was deemed as irrelevant to our purposes and thus discarded.

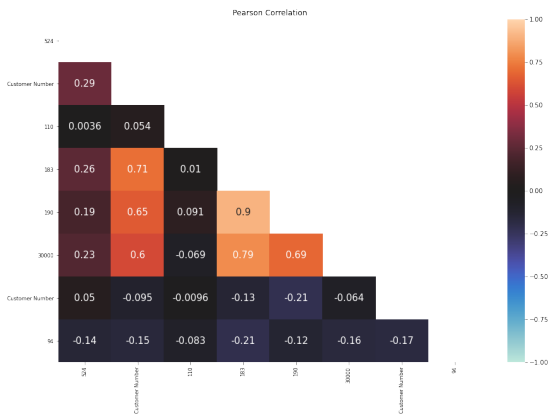

Figure 4: Pearson correlation between SPN pairs.

Next, to reduce the potential bias due to the contemporary presence of correlated SPNs describing related components of the same physical system, we perform also a preliminary correlation analysis of the SPN series.

Figure 4 shows the Pearson's correlation. It clearly indicates the presence of a group of highly correlated SPNs describing the status of the vehicle engine namely engine speed, engine percent load, engine fuel rate and fuel tank level. Hereafter, we will focus our analyses on a group representative, i.e., the engine speed.

Finally, we analyze the spectral content of the SPN signals. Signals characterized by slow variations are disregarded in the following analyses since they incorporate most of their information in correspondence of frequencies close to 0 and their spectral content can be approximated by the temporal average value of the signal. Notice that slow signal variations can be due to either the intrinsic nature of the considered measure (e.g., for the SPN related to the transmission selected gear) or to the limited sensitivity of measurement instrument (e.g., for the SPNs related to charge pressure and engine fuel delivery pressure).

5 PROFILE VEHICLE USAGE

To identify vehicle duties we analyze vehicle usage data by means of clustering techniques. Specifically, the SPN series are first synchronized and segmented into fixed-length intervals. Each segment is described by specific features. Then, segments are clustered into homogeneous groups. The clustering outcomes allow domain experts to empirically set up per-duty levels associated with each SPN.

Time series segmentation. Time series segmentation entails defining a partition of the input series $X(t)$ into into k segments $S_1, S_2, \ldots, S_k$, each one characterized by a distinct time span $[t_{start}, t_{end}]$. Since vehicle usage is described by multiple SPN series, the segmentation problem is extended to a multivariate model, i.e., given the time series $X_1(t), X_2(t), \ldots, X_n(t)$ corresponding to SPNs $SPN_1, SPN_2, \ldots, SPN_n$, respectively, we partition them series into k segments, where the same partition holds for all the considered series. To deal with correlated series, the input series can be preprocessing using Principal Component Analysis [1] with the aim at collapsing the underlying SPN subtrends that are highly correlated with each other into a separate component.

For the sake of simplicity, we address time series segmentation using an established motif discovery algorithm [13]. Motifs are recurring sub-series within a reference time series. The algorithm first splits each of the original time series into fixed-length segments and then compares pairs of segments to select the top most similar pairs. The segment length can vary within a range [L_{min}, L_{max}]. Both the segment length range and the distance measure used to generate the motif are configurable by domain experts. In our experiments, we varied the segment length between 2 minutes and 10 minutes and evaluated the similarity between segments via Euclidean distance.

To empirically identify the most appropriate segment length, we discretized the segment length range into 1-minute bins, counted the number of motifs per length, and selected the length maximizing that count. The lower bound of the segment length range (2 minutes) turned out to be the most appropriate time scale. The corresponding 2405 segments will be hereafter considered in the reported analyses.

Per-segment feature extraction. For each segment we extract a subset of features that describe time series shape and values' distribution. since the aim is to characterize the general shape of each segment in terms of its variations and their corresponding rapidity and amplitudes, SPNs are analyzed in the frequency domain. To this purpose, for each SPN and segment the Fourier transform of the signal is applied by considering only the positive coefficients, as we exploit the symmetry of real signals Fourier coefficients. Then, separately for low, medium, and high frequencies, the signal power value, the signal peaks, and the signal peaks frequencies are computed. Lastly, the signal mean (in the time domain) is considered as well.

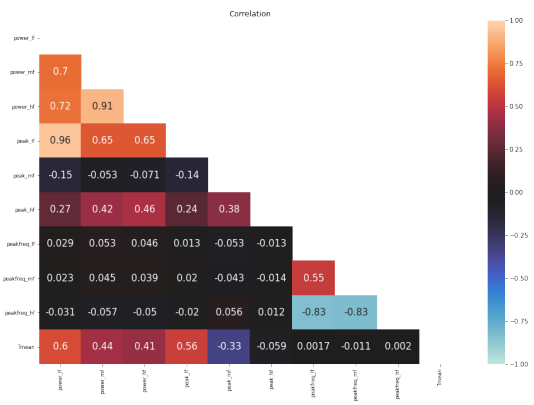

Figure 5: Pearson correlation between pairs of per-segment features.

Figure 5 shows the Pearson's correlation between the pairs of extracted features. According to the correlation values, the feature set is reduced to 3: (i) the power in low frequencies sub-band, (ii) the peak value for high frequencies, and (iii) the peak frequency in the high frequencies sub-band.

Clustering. Clustering aims at grouping data samples that are similar to one another and dissimilar from those assigned to other groups. In this particular context, clustering algorithms are exploited to group the SPN segments into homogeneous groups representing typical vehicle duties (or vehicle states for which duties can be easily inferred). The number of desired clusters is an input parameter, which can be specified by the domain experts. We set up this parameter in an empirical way by assessing the clustering results according to established cluster validity indices. Specifically, to choose the best algorithm and the number of desired clusters we empirically assessed the performance achieved by multiple runs of different clustering algorithms by varying the number of desired clusters k (see Section 6).

Duty level identification. On top of the cluster outcomes the levels associated with each vehicle duty can be identified. To this aim, the SPN segments associated with the same cluster are further split into sub-groups characterizing similar usage patterns. For example, sub-groups allow us to distinguish between vehicles in an idle state and vehicles that keep moving steadily.

6 EXPERIMENTAL RESULTS

We carried out an empirical analysis of the proposed methodology on the real vehicle usage data provided by Tierra SpA. The experiments were run on an Intel(R) Core(TM) i5-8250U machine equipped with 8 GB of RAM and running Windows 10 64-bit.

The summary of the experimental results is organized as follows. Firstly, we compare the performance of different clustering techniques according to the Silhouette validity index [2] and discuss the impact of the number of desired clusters on clustering performance (see Section 6.1). Secondly, we quantitatively evaluate the quality of the clustering outcome against the National Marine Electronics Association (NMEA) 0183 messages data [3] (see Section 6.2). Finally, we report a qualitative analysis of the achieved results (see Section 6.3).

6.1 Comparison between different clustering techniques

We tested various purely partitional clustering algorithms belonging to the following categories: (i) centroid-based, (ii) density-based, (iii) hierarchical, and (iv) shape-based. We considered two of the most renowned algorithms belonging to the centroid-based category (K-Means [14], which exploits the concept of cluster centroid, and Clara [12], based on medoids[1]) Density-based clustering group together samples located in dense regions and well separated from other regions. We exploited a Python implementation of the well-known DBScan algorithm [6]. Hierarchical clustering produces nested clusters, which can be organized in a dendrogram. We exploited an implementation of an agglomerative algorithm [4]. Finally, shape-based clustering is tailored to time series clustering. We considered the K-Shape algorithm [18], which relies on series cross-correlation analyses. Notice that the latter algorithm is designed for univariate time series analysis. It captures the similarities between sub-series independently of the shift of the distinctive segments' properties.

We evaluated clustering performance according to the Silhouette score, which is an established validity index used to measure of how similar a sample is to its own cluster compared to the other clusters [2]. The score ranges from -1 (high separation) to 1 (high cohesion), i.e., the larger the better. For each algorithm we varied the configuration settings to find the best setting.

For all the tested algorithms we achieved the best results by setting the number k of desired clusters to 2. K-Means achieved

[1] Clara is an extension of the k-Medoid algorithm, which is able to scale towards larger and more complex datasets.

the best overall performance (0.67), followed by the hierarchical clustering (0.53), Clara (0.5), DBScan (0.49) and K-Shape (0.23).

6.2 Evaluation based on the National Marine Electronics Association messages

We assessed the quality of the clustering outcomes using, as ground truth, the numerical score provided by the National Marine Electronics Association (NMEA) 0183 messages data [3]. NMEA 0183 is a one-way serial data communication protocol used to send messages from the vehicle to external devices. Unlike CAN bus data, it provides fairly accurate GPS-related information such as the vehicle coordinates (latitude and longitude) and the vehicle speed. Conversely, GPS positions transmitted via CAN bus messages are frequently characterized by relatively high measurement error [8]. Despite the accuracy of NMEA position was not guaranteed overall, we made a preliminary attempt to validate the ability of the proposed method to discriminate between idle states and moving/working ones by comparing the duty labels assigned via clustering against the assignment made based on NMEA message (used as ground truth).

We achieved a 82.37% accuracy score, i.e., we correctly classified approximately 8 duties out of 10. The average recall and precision scores were 82.35% and 82.39%, respectively. The results were quite promising, provided that the segments used for validation purposes were fairly balanced (50.81% of idle segments, 49.19% of moving/working ones).

A deeper analysis of the wrongly labeled segments has shown that, in few cases, there were rapid and multiple changes in the engine speed associated with an idle state. This was probably due to small errors in GPS readings. Therefore, these particular errors seem to be not due to imprecise vehicle duty level assignments.

6.3 Qualitative evaluation

Figure 6 plots two representative segments belonging to three different clusters. With the help of domain experts, we figured out the underlying vehicle usage patterns. Specifically, cluster 1 shows a working or heavy workload state and is characterized by highly variable segments. It indicates an aggressive driving style, with rapid accelerations and breaks. Cluster 2 shows a more stationary vehicle usage. The usage levels are compatible with either a stationary vehicle under medium workload or with a non-stationary vehicle. Finally, cluster 3 contains slow varying segments in which the engine speed is oscillating around minimum levels of use. It likely denotes an idle duty.

Figure 7 shows the number of segments per cluster (corresponding to the previous experiment). According to domain experts' opinion, the distribution is coherent with what expected, since it corresponds to the actual usage of the test vehicle.

7 CONCLUSIONS AND FUTURE WORKS

The paper explores the use of clustering techniques to profile industrial vehicle usage in construction sites. The aim is to define per-vehicle duties, which summarize the current state of the vehicle (e.g., idle, moving, heavy workload). Due to the high heterogeneity of vehicles types, models, and usage contexts, duties are commonly detected by exploiting manually configured threshold. To automate this process, we propose a data-driven approach relying on CAN bus signals segmentation and clustering. The clustering output achieved on real vehicle usage data was validated with the help of domain experts.

The preliminary results leave room for further improvements. First of all, the acquisition of CAN bus data from many test vehicles would allow us to extend the problem of vehicle duty identification from a single vehicle to groups of similar vehicles. Secondly, a deeper analysis of the contextual information related to working site and the vehicle equipment would be useful for further improving the accuracy of the duty level assignments for effectively identifying the driving styles. In addition, as soon as new historical data become available, the framework will be updated basing on results obtained on similar tasks and construction site conditions in the past. Finally, the assigned duties will be exploited to accomplish specific tasks, such predictive maintenance and anomaly detection[19].

8 ACKNOWLEDGMENTS

The research leading to these results has been funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies and by Tierra Spa.

REFERENCES

[1] J. Abonyi, B. Feil, S. Németh, and P. Arva. 2004. Principal Component Analysis based Time Series Segmentation: A New Sensor Fusion Algorithm.

[2] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesus M. Perez, and Inigo Perona. 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* 46, 1 (2013), 243 – 256. https://doi.org/10.1016/j.patcog.2012.07.021

[3] National Marine Electronics Association. [n.d.]. *NMEA 0183.* Retrieved November 11, 2020 from https://www.nmea.org/content/STANDARDS/NMEA_0183_Standard

[4] Maria-Florina Balcan, Yingyu Liang, and Pramod Gupta. 2014. Robust Hierarchical Clustering. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 3831–3871.

[5] Sahar Ebadinezhad, Ziya Dereboylu, and Enver Ever. 2019. Clustering-Based Modified Ant Colony Optimizer for Internet of Vehicles (CACOIOV). *Sustainability* 11, 9 (May 2019), 2624. https://doi.org/10.3390/su11092624

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96).* AAAI Press, 226–231.

[7] Umberto Fugiglando, Paolo Santi, Sebastiano Milardo, Kacem Abida, and Carlo Ratti. 2017. Characterizing the "Driver DNA" Through CAN Bus Data Analysis. In *Proceedings of the 2nd ACM International Workshop on Smart, Autonomous, and Connected Vehicular Systems and Services (CarSys '17).* Association for Computing Machinery, New York, NY, USA, 37–41. https://doi.org/10.1145/3131944.3133939

[8] Yong Heo, Thomas Yan, Samsung Lim, and Chris Rizos. 2009. International standard GNSS real-time data formats and protocols.

[9] Anders Holst, Juhee Bae, Alexander Karlsson, and Mohamed-Rafik Bouguelia. 2019. Interactive Clustering for Exploring Multiple Data Streams at Different Time Scales and Granularity. In *Proceedings of the Workshop on Interactive Data Mining (WIDM'19).* Association for Computing Machinery, New York, NY, USA, Article 2, 7 pages. https://doi.org/10.1145/3304079.3310286

[10] Thomas Huybrechts, Yon Vanommeslaeghe, Dries Blontrock, Gregory Van Barel, and Peter Hellinckx. 2018. Automatic Reverse Engineering of CAN Bus Data Using Machine Learning Techniques. In *Advances on P2P, Parallel, Grid, Cloud and Internet Computing*, Fatos Xhafa, Santi Caballé, and Leonard Barolli (Eds.). Springer International Publishing, Cham, 751–761.

[11] Karl Henrik Johansson, Martin Törngren, and Lars Nielsen. 2005. *Vehicle Applications of Controller Area Network.* Birkhäuser Boston, Boston, MA, 741–765. https://doi.org/10.1007/0-8176-4404-0_32

[12] Leonard Kaufman and Peter J. Rousseeuw. 1987. Clustering by means of medoids. , 405–416 pages.

[13] Michele Linardi, Yan Zhu, Themis Palpanas, and Eamonn Keogh. 2018. Matrix Profile X: VALMOD - Scalable Discovery of Variable-Length Motifs in Data Series. *SIGMOD '18: Proceedings of the 2018 International Conference on Management of Data*, 1053–1066. https://doi.org/10.1145/3183713.3183744

[14] J. MacQueen. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, L. M. Le Cam and J. Neyman (Eds.). University of California Press, Berkeley, CA, USA, 281–297.

[15] C. Marina Martinez, M. Heucke, F. Wang, B. Gao, and D. Cao. 2018. Driving Style Recognition for Intelligent Vehicle Control and Advanced Driver Assistance: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 19, 3 (2018), 666–676. https://doi.org/10.1109/TITS.2017.2706978

[16] Dena Markudova, Elena Baralis, Luca Cagliero, Marco Mellia, Luca Vassio, Elvio Gilberto Amparore, Riccardo Loti, and Lucia Salvatori. 2019. Heterogeneous Industrial Vehicle Usage Predictions: A Real Case. In *Proceedings of the Workshops of the EDBT/ICDT 2019 Joint Conference, EDBT/ICDT 2019, Lisbon,*

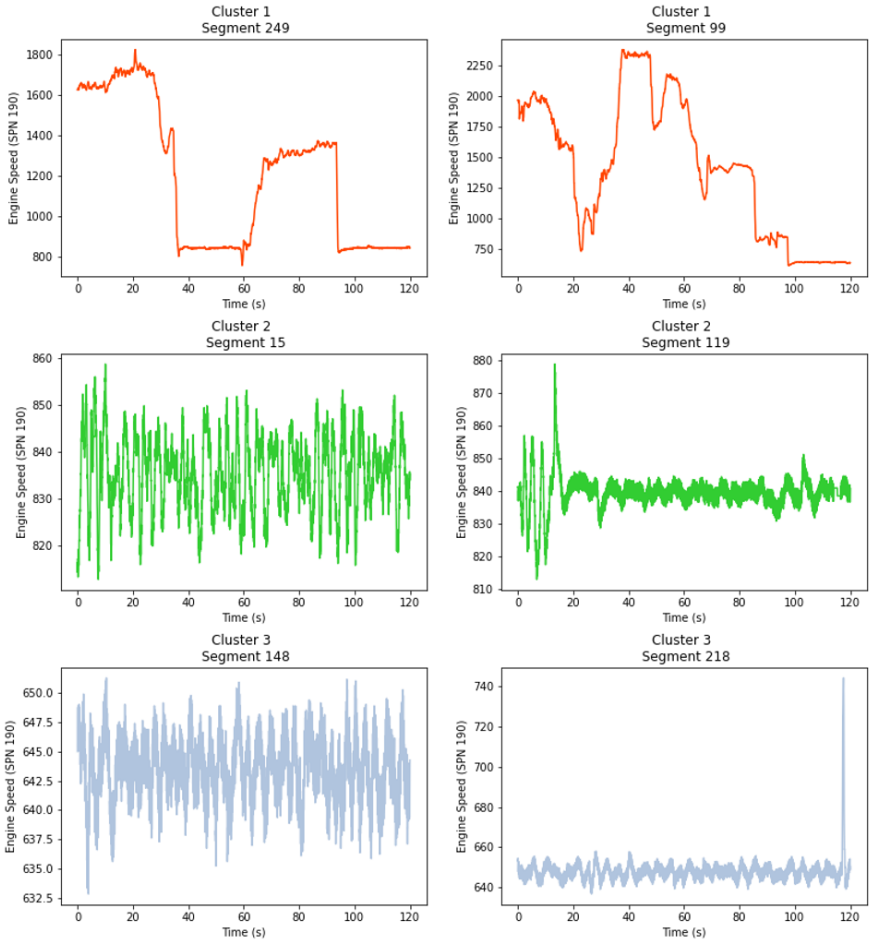

Figure 6: Representative segments for each cluster

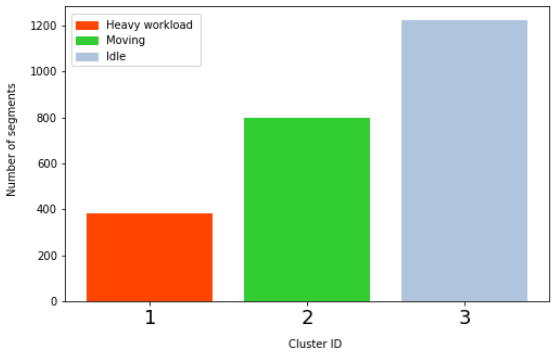

Figure 7: Number of segments per cluster

Portugal, March 26, 2019 (CEUR Workshop Proceedings), Paolo Papotti (Ed.), Vol. 2322. CEUR-WS.org. http://ceur-ws.org/Vol-2322/DARLIAP_13.pdf

[17] Sachit Mishra, Luca Vassio, Luca Cagliero, Marco Mellia, Elena Baralis, Riccardo Loti, and Lucia Salvatori. 2020. Machine Learning Supported Next-Maintenance Prediction for Industrial Vehicles. In *Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, March 30, 2020 (CEUR Workshop Proceedings)*, Alexandra Poulovassilis, David Auber, Nikos Bikakis, Panos K. Chrysanthis, George Papastefanatos, Mohamed A. Sharaf, Nikos Pelekis, Chiara Renso, Yannis Theodoridis, Karine Zeitouni, Tania Cerquitelli, Silvia Chiusano, Genoveva Vargas-Solar, Behrooz Omidvar-Tehrani, Katharina Morik, Jean-Michel Renders, Donatella Firmani, Letizia Tanca, Davide Mottin, Matteo Lissandrini, and Yannis Velegrakis (Eds.), Vol. 2578. CEUR-WS.org. http://ceur-ws.org/Vol-2578/DARLIAP9.pdf

[18] John Paparrizos and Luis Gravano. 2015. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. 1855–1870.

[19] Stefano Proto, Evelina Di Corso, Daniele Apiletti, Luca Cagliero, Tania Cerquitelli, Giovanni Malnati, and Davide Mazzucchi. 2020. REDTag: A Predictive Maintenance Framework for Parcel Delivery Services. *IEEE Access* 8 (2020), 14953–14964. https://doi.org/10.1109/ACCESS.2020.2966568

[20] Lee A Schmidt and Lorenz Riegger. 2012. Automatic Detection of Machine Status for Fleet Management. US Patent App. 13/341,500.

[21] D. Yi, J. Su, C. Liu, and W. Chen. 2019. Trajectory Clustering Aided Personalized Driver Intention Prediction for Intelligent Vehicles. *IEEE Transactions on Industrial Informatics* 15, 6 (2019), 3693–3702. https://doi.org/10.1109/TII.2018.2890141

[22] Weiliang Zeng, Tomio Miwa, Wakita, and Takayuki Morikawa. 2015. Exploring Trip Fuel Consumption by Machine Learning from GPS and CAN Bus Data. *Journal of the Eastern Asia Society for Transportation Studies* 11 (12 2015), 906–921. https://doi.org/10.11175/easts.11.906

[23] W. Zhang, D. Yang, and H. Wang. 2019. Data-Driven Methods for Predictive Maintenance of Industrial Equipment: A Survey. *IEEE Systems Journal* 13, 3 (Sep. 2019), 2213–2227.