Exploring waste-collection fleet data: challenges in a real-world use case from multiple data providers

Simone Monaco[1], Paolo Bethaz[1], Daniele Apiletti[1], Fabrizio Pio Baldini[2], Carlo Caso[3] and Tania Cerquitelli[1]

[1]Department of control and computer engineering, Politecnico di Torino, Torino, Italy
[2]SEA Soluzioni Eco Ambientali S.r.l., Villanova Canavese (TO), Italy
[3]T2D Transfer to Digital Salerno, Italy

Abstract

In the age of connected vehicles, large amounts of data can be collected while driving through a variety of on-board sensors. The information collected can be used for various types of data-driven analytics that can be of great benefit to both vehicle owners, e.g., to reduce costs by means of predictive maintenance, and to society as a whole, e.g., to optimize mobility behavior. Prior to any real-world data analysis, an investigation and characterization of the available data is of utmost importance in order to evaluate the quality and quantity of the data and to set the right expectations.

In this paper, we focus on the data exploration and characterization step, which is necessary to avoid inconsistencies in the collected parameters and to enable valid, data-driven modeling. The proposed data exploration considers both the frequency of samples and their values for all monitored parameters. A specific cross-provider data comparison is performed to compare values collected for the same vehicle at the same time from different fleet monitoring data providers. The study is applied to a real-world use case with months of data from dozens of vehicles deployed in the waste collection service managed by SEA, Soluzioni Eco Ambientali, in Italy. The analyzes uncover unexpected behaviors in the measurements and lead to their early identification, bringing great benefits to the company operating the fleet by improving data collection and enabling a safe modeling phase.

Keywords

Real-world data characterization, waste-collection

1. Introduction

The recent rise of monitoring systems on modern vehicles is opening a broad range of applications for machine learning (ML) techniques to process this data [1]. Among the most popular use cases, analyses that can be implemented to leverage the collected data may include, for example, estimating the best route for the vehicle, predicting fuel consumption, or predicting a failure on a particular component. However, ML algorithms are effective as long as the available data are sufficiently accurate. Then, it is fundamental to assert the reliability of this data before going into any kind of analysis. This evidence becomes even more important when dealing with data coming from multiple sources, for which collecting strategies, precision, and availability can significantly vary. In this work, we present a real-world use case where waste-collecting vehicles are monitored by multiple data providers, and interventions to fix failures on such vehicles are regularly recorded. The main contribution of this paper are:

1. An exploratory analysis on real data from a waste-collection company, which is tracking its fleet by means of different data providers.
2. An exploratory analysis of the recorded maintenance interventions to enable a future predictive maintenance solution.

The rest of the paper is organized as follows. Section 2 presents the related works, Section 3 describes the real-world use case under analysis, and Section 4 focuses on vehicle-tracking data exploration. Finally, Section 5 draws conclusions and future works.

2. Related work

In recent years, the advent of the Internet of Things and its rapid diffusion has facilitated real-time communication and data exchange between technological devices that can be used in a wide variety of scenarios. When the technologies involved are located on a vehicle (e.g. car, truck), the term Internet of Vehicles (IoV) is used [2, 3]. This term refers to the situation in which data are collected locally on the vehicle and then sent to a remote storage location (cloud) where it can be deeply analyzed.

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK
simone.monaco@polito.it (S. Monaco); paolo.bethaz@polito.it (P. Bethaz); daniele.apiletti@polito.it (D. Apiletti); fabrizio.baldini@seaeco.it (F. P. Baldini); carlo.caso@t2d-digital.com (C. Caso); tania.cerquitelli@polito.it (T. Cerquitelli)
0000-0003-4948-6120 (S. Monaco)

Many infrastructures and different technologies have been implemented to support large-scale, real-time, and reliable information services. Basically, for this purpose, a three-tier architecture [4] is used: the first tier with all the sensors within the vehicle, a second tier representing the communication layer, and the third tier including statistics tools, support for storage, and processing infrastructure. This last tier leverages collected data from sensors to extract information that can bring value in different contexts.

The GPS data which reflects the movement of vehicles can be traced to simulate the mobility model in IoV. Authors in [5] and [6] have presented different machine learning techniques for short-term congestion prediction using vehicle trajectory available for connected vehicles. In another set of works, the goal is to correctly predict fuel consumption under various driving conditions. This is done for example in [7] and [8], where prediction results show the good value of mean absolute percentage errors. Finally, a great deal of research has been done to try to exploit the data collected to be able to do predictive maintenance. This implies the building of a data-driven model that can predict in advance a failure on a given component, thus being able to intervene promptly, avoiding breakdowns during vehicle travel. In this context, in [9] authors try for example to estimate the probability of fuel pump failure, while in [10] the goal is to predict the presence of a Diagnostic Trouble Code ignition on commercial trucks.

In many works, data is collected and managed with a clear purpose in mind. However, in other cases, companies would like to gain general insights of the information they collected, and then drive the decisions on how to effectively reach cost reductions and higher operational efficiency. In such cases, an exhaustive exploration of the available data is required, allowing the company to possibly introduce early changes in the collection process. In this way, our work differs from the above-mentioned ones because our focus is on the data exploration phase, to highlight noteworthy behaviors or anomalies of interest for the data owner. Furthermore, all the above-mentioned works have been conducted on data collected by a single provider. Instead, in our use case, we analyze and compare three different data providers, each with its own specific sensors. A single parameter can be either specific to a single provider, or it can be monitored by several providers (each with its peculiarities). For this reason, we focus on a data comparison step, not covered in previous related works, to highlight the benefits and weaknesses of each data provider.

3. Use case description

We analyze real-world fleet data of garbage collection operated by *SEA, Soluzioni Eco Ambientali*. The company acts in cooperation with the local administration in many Italian municipalities, to handle the collection, transport, and selection of municipal waste. To this aim, the company deploys a fleet of trucks. They are equipped with sensors able to track the vehicle during the trips. Every day, more than 300 vehicles operate in 14 different worksites, located in different cities in the north and center of Italy. An unexpected failure of one of these vehicles would need a quick reorganization of the workload of all the others, to optimally perform the service. Hence, an early detection of the signals leading to possible failure is of primary importance, both from an economic and an organizational point of view.

In this context, we present a preliminary analysis and characterizations of such data. Specifically, we focused on a subset of 40 trucks, which are the most representative and well monitored, i.e., those for which the data was extensively present for most of the time. From the technological point of view, all the data is stored in a data lake, managed by *T2D Transfer to Digital*, which collects the results of regular calls to the REST APIs of the data providers. The different service providers, being either the truck manufacturer or an external supplier, added sensors on the vehicle chassis and the top installations. Hence, data sources are extremely heterogeneous, and also irregular in their timings, which leads to the presence of a variety of anomalies. Among all the available data sources, we can identify the following types of data in the data lake.

- **Tracking.** This data describes the trips each truck performs. We have 3 *providers* (named P1, P2, and P3 in the following), each tracking different features. The only common measures are GPS position, odometer value, total time of engine activity, and average speed. Some providers also include other measures such as acceleration and estimated fuel consumption. In addition to the variability on the tracked features, different providers also have different policies on the sampling of the data, hence leading to very diverse daily measure rates. The raw datasets collected by the providers are time series of a total of 117880 samples of 38 tracked features for 35 vehicles for P1, 68611 samples of 68 features for 8 vehicles for P2, and 91072 samples of 17 features for 15 vehicles for P3.
- **Maintenance.** Among the goals of the current exploratory analysis, we would like to assess the feasibility of a future predictive model aiming to predict failures based on the historical maintenance interventions. To this aim, in the data lake

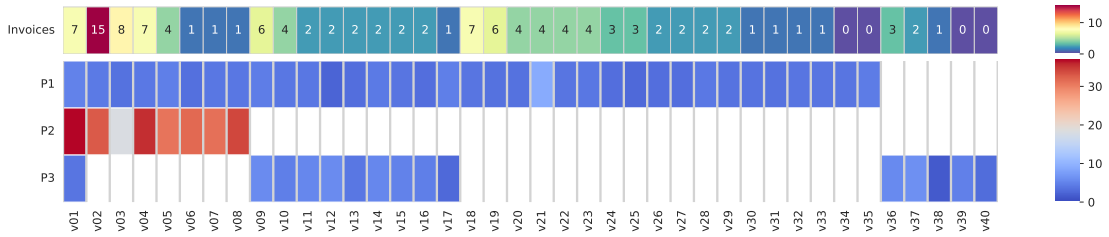

Figure 1: Overview of the invoices and sensors' daily activity on the associated vehicles.

we analyzed the list of all the repairs of the different components of the trucks, for which the corresponding *invoices* are collected. We have a total of 120 invoices associated with 36 vehicles.

4. Exploratory analysis

In this section, we describe the data analysis process and results. The dataset of interest in the current work includes information of 40 vehicles from May 2021 to January 2022. Even though the data providers' REST APIs have been directly used to collect the data at regular and synchronous intervals during the whole period of time, the timestamps of some samples refer to periods sensibly far from the range (May 2021 - January 2022). Furthermore, significant portions of the original data are duplicated, i.e., different calls to the provider APIs return the very same data, also with identical timestamps. Some of these anomalous repeated data are associated with a subsequent period with no collected data. Nevertheless, the overall trend of the vehicles in these situations suggests they where still moving without being tracked. For this reason, we dropped all duplicated values down, since we assumed they are only the result of irregularities of the providers' update systems. In the same way, we then also removed all samples outside of the considered time range.

Since at each API call, all variables are returned by the data provider, but not all variables effectively recorded a change in their value, a feature selection step has been used to remove irrelevant features. After removing duplicate records, we also removed all those variables that we felt would not be helpful for an upcoming data-driven analysis. These variables were identified according to the following criteria: (i) a missing value ratio greater than 50%, (ii) limited variance (e.g., variables whose values are always constant).

As a first indicator of the remaining information cardinality, we defined the average number of samples per vehicle per day as a synthetic and high-level measure. Figure 1 shows how the different providers (P1, P2, and P3 on the rows) supply information across the 40 vehicles (on the columns). The heatmap shows the average daily number of samples, for each vehicle (column) and provider (row). The first row reports the number of invoices for each truck, each corresponding to the maintenance interventions.

Among the 40 vehicles, the first 17 share tracking data from multiple providers at the same time, and are described by more than 30 samples per day per vehicles in some cases, with the average being less than 5 samples per day per vehicle. The number of maintenance interventions (invoices) ranges from zero to 15 in the 9 months under analysis.

In the following, we discuss the results of the exploratory analysis on such data.

4.1. Multi-provider data comparison

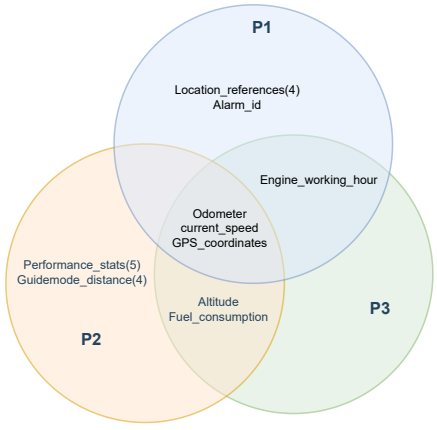

Figure 2: Venn diagram of the high-level semantic aggregation of the features from 3 tracking data providers.

Since each provider has different variables, we executed a semantic association to identify related features from different providers. The resulting features with the corresponding associations are presented in Figure 2,

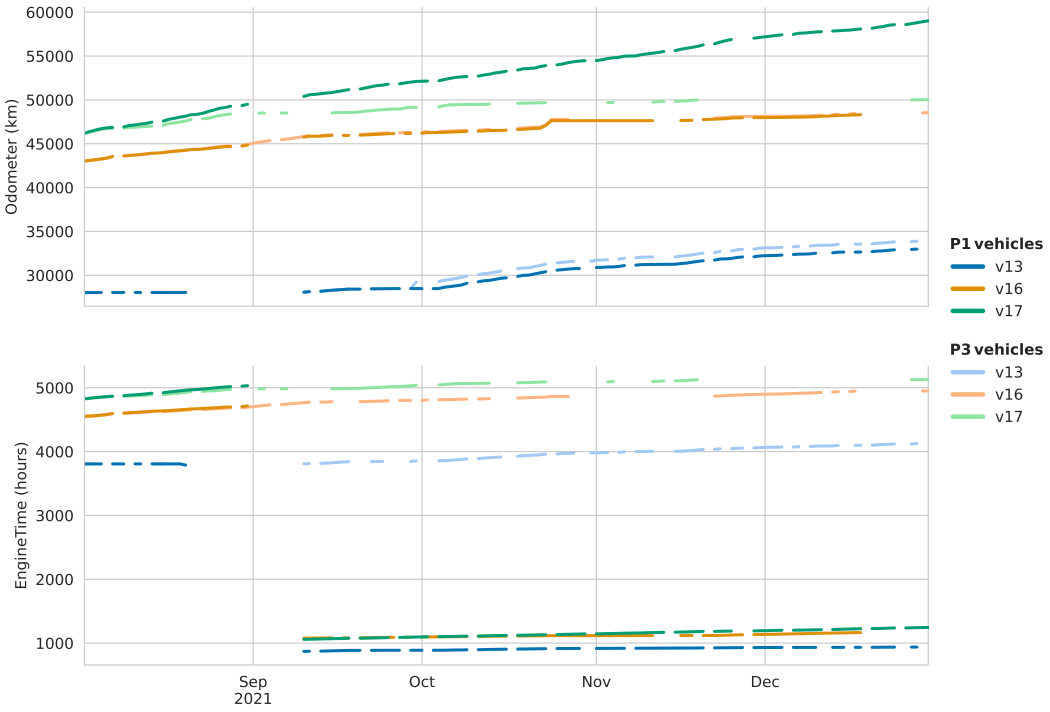

Figure 3: Comparison of the odometer and engine working time for 3 vehicles monitored by both P1 and P3.

where the Venn diagram shows how the semantic groups are shared between the 3 providers.

The features in common to all the providers are the odometer, the speed, and the geographical position of the vehicles. The working time of the engine is available in 2 out of 3 providers. The odometer and the engine working time are of particular interest in our analysis since they give a measure of the mechanical "age" of the truck, and despite some possible differences on the initial offset among different providers, they are supposed to be sampled coherently among them all. Other variables, such as the current speed, fuel consumption, and performance statistics, are also potentially significant to trace the degradation of the vehicles. However, even if they share the same high-level semantics, we noticed that their values were so different that each provider actually tracks a variable on its own, hence they cannot be merged together.

4.1.1. Comparative analysis on monotonic features

The odometer and the engine working time are two monotonic variables shared among different data providers (P1 and P3). Both variables are constant when the vehicle is turned off, whereas both must be strictly increasing while driving.

From Figure 1 we note that 10 vehicles share the P1- and P3-provided odometer and engine working time. In Figure 3 we report the comparison of such variables for 3 overlapping trucks (namely v13, v16, and v17) in the period between August 2021 and November 2021, after a simple conversion of units of measure (P1 uses kilometers and hours, while P3 meters and minutes).

P1 presents a temporal gap in the collected samples, of about one week from September 1st, for all the reported trucks. Since the last measure for each vehicle before the gap and the first one after drastically differ, we can assume that within this period the trucks were still active. This assumption is proven by the measures from P3, confirming that it was associated with a lack of information from the provider itself. The issue is common to all the data of this source, and it was not reported to the knowledge of the company before the current analysis. Moreover, this period, as highlighted by the lower plot, is also associated with a drastic lowering in the value of the engine time. This can correspond to a sort of internal reset of the sensors, which may have made the registered values start from a default position, slightly different from any truck. In this situation, a fix for this error can be obtained by adding the first value from P3 of the correspondent vehicle to all measures of P1. With this correction, each couple of curves after the gap becomes

comparable.

Finally, the upper plot also shows another anomaly concerning the odometer tracking of v17. The traces from the two providers have significantly different trends that continue diverging as time increases. Vehicle v17 traveled with an average speed of 83 km/day and 23 km/day, for P1 and P3, respectively. All providers' average speed distributions are peaked at around 60 km/day, as shown in Fig. 4, but the calculated average values are on the tails of both distributions.

In this situation, as for other trucks monitored by more than one data provider, we experienced that information from different providers can reveal significant differences in the reported values.

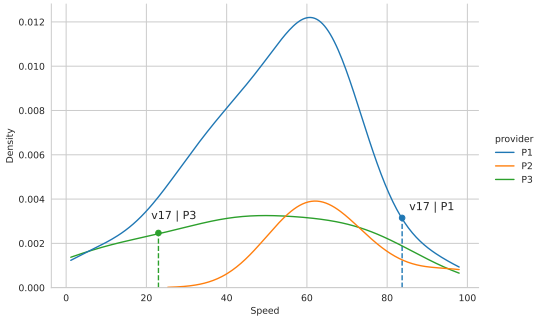

Figure 4: Daily speed distribution for each provider. Two points highlight the measured average speed for v17

4.2. Analysis over time

An important analysis to perform in order to understand effectively the data collected is an exploration of the time period in which the vehicles are monitored. How long and in which manner the data is collected is in fact a useful information for the company that can highlight important differences between the various providers. Below, for each provider we have reported a graph representing the time on the x axis, and the different trucks monitored by the provider on the y axis. In this way, for each truck we can easily see the start and end date of collection and whether there were any anomalies in the monitored period (e.g. missing data or changes in the daily data rate).

4.2.1. P1 provider

The graph shown in Fig 5 shows all trucks monitored by provider P1. The different color shades of the points represent a different sampling frequency on a given day (the darker the point, the more measurements were collected on the same day for that truck). This graph shows that the number of daily measurements collected is almost the same and that all trucks were monitored during the period from May 2021 to January 2022. The anomalies are highlighted by the red rectangles in the figure. The first anomaly that is detected is common to all trucks, and is a lack of data at the beginning of 2021/09 (September). This behavior can be explained by a period of company closure, or by a general failure in the data collection process, affecting all trucks. Other anomalies are instead related to specific trucks, and show rather long periods (even several months) of lack of data. In general, to see if in all of these situations we don't have data because the truck was actually idled or because there was a problem in the collection, we need to look at the cumulative values collected from variables such as the odometer. If the value of this variable in the last sample collected before the period without data is equal to the value of the first sample collected after the period without data, it means that the truck actually did not travel in that period, so no information was lost. Otherwise, if the two values are different, it means that the truck moved but we don't have data that monitored its operations during that period. So there was an error in the data collection phase.

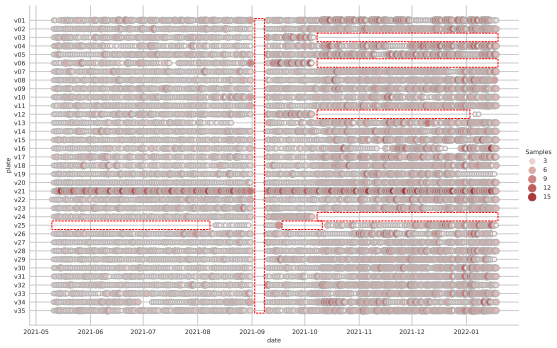

Figure 5: Temporal analysis for P1

4.2.2. P2 provider

A similar analysis was performed for the measurements collected by provider P2 and shown in Figure 6. In this case, the number of trucks monitored is much lower than those monitored by P1. Moreover, all the trucks considered by P2 are also monitored by P1. This is useful in order to make a quick cross-check on anomalies between the two providers. In particular, we can easily notice that in the same periods in which data for P1 was missing (both the period common to all trucks and those for individual trucks), we now have measurements collected by P2. This leads us to assume that the previous lack of data were really due to P1-related failures. The only period in which we have no measurements in either providers is between 2021/12 and 2022/01 for vehicle v03. In this case, the only check we can do to verify if the truck was

really stopped is to monitor the odometer value before and after the lack of data.

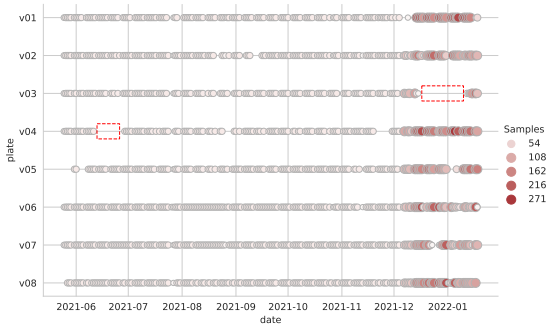

Figure 6: Temporal analysis for P2

Regarding the measurements collected by P2, a particular behavior can be noticed in the sampling frequency. Until 2021/12 there is a very low sampling frequency, equal to around one daily measurement per truck. However, this behavior changes suddenly in the last months of collection, where the sampling frequency increases significantly. This seems to be due to a change in the collection pattern from December onward, probably with the aim of collecting information with a higher level of detail than the single daily information previously collected. The use of the trucks has not changed since December, what has changed is the amount of information describing the behavior of a truck, with up to 200 samples per day.

4.2.3. P3 provider

The same temporal analysis has also been performed for the trucks monitored by provider P3. Five of the trucks in Figure 7 (from v36 to v40) are monitored only by P3, therefore it is impossible to cross-check them with the data of the other providers to try to better understand the anomalies. Regarding instead the remaining trucks, they are all monitored by at least one between P1 and P2 and this allows us to cross-check the collected measurements. In particular, the most evident anomalies in the data collected by P3 are for those five trucks where there are no measurements in the period from 2021/05 to 2021/10. However, for the same trucks and in the same months, the other providers have collected measurements, so it is fair to assume that the lack of data for P3 is a problem related only to that provider.

4.3. Maintenance analysis

Concerning the maintenance data, we analyzed the collection of invoices indicating failures of different components on the trucks. With the help of domain experts,

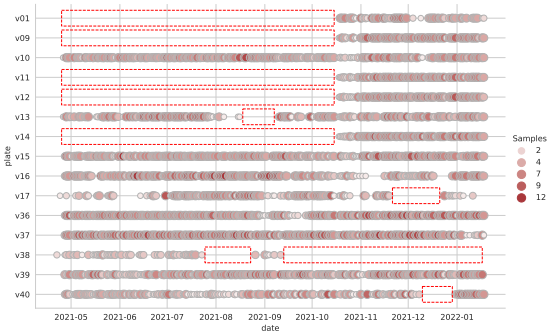

Figure 7: Temporal analysis for P3

we manually excluded all the elements associated with periodical maintenance operations to be done as part of the security enforcement by law, then all the resulting elements have been grouped into meaningful categories based on the failed components. Looking at the remaining invoices, we then observed a significant number of operations from the same category carried out on the same vehicle within a short time (e.g., the day after). We assumed all these cases describe a situation when the repair was insufficient to solve the problem, hence not relevant for a predictive-maintenance analysis. For this reason, we excluded such data from the analysis. The resulting, clean number of failures, grouped by categories, is shown in Figure 8. For each row, besides the total number of failures, also the number of second occurrences, i.e., those which are not the first failures for the vehicle and the category, is reported. The failures coming from this second group are particularly interesting when investigating the possibility of building a predictive model on top of this data since they can give a preliminary measure of the frequencies of the failures.

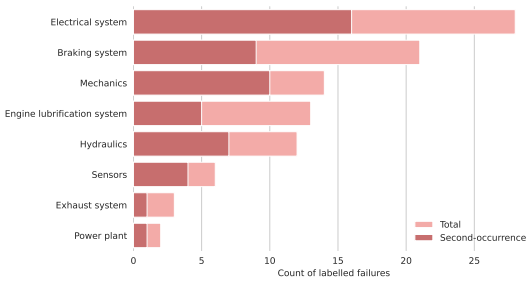

Figure 8: Clean count of the failures, separated by category.

The entire process of cleaning and correctly grouping the available invoices is fundamental whether the collected data will be exploited to address a predictive maintenance analysis, building a predictive model capable of estimating the residual life of a truck before a

maintenance intervention is needed. The possibility of identifying malfunctions or component issues in advance is in fact a key aspect for automotive companies, which can then intervene properly before a failure occurs. This leads to significant cost savings for the company, since a properly implemented predictive model would allow to avoid breakdowns during vehicle travel and to constantly monitor the condition of the various components, scheduling maintenance interventions only when they are really necessary. From Figure 8, we note that for a few categories we do not have enough data to be able to build a robust predictive model for estimating future failures. This is due to the limited time period in which trucks are monitored, which is not long enough to collect enough maintenance interventions for all categories.

5. Conclusions and future works

In this work, we presented an exploratory analysis and preliminary characterization of real-world fleet data coming from different tracking providers. We analyzed and compared the data collected by 3 providers, on 40 vehicles, over 9 months, highlighting issues and challenges. The presence of multi-provider information for the same vehicle, although more expensive, can help recover the frequent issue of missing data. As future work, we plan to extend the current work to integrate data from different sources with the goal of designing a predictive maintenance algorithm, by exploiting the maintenance intervention data. Anomalous information from the vehicle tracing, like the ones we reported in this paper, can lead to unreliable input features and consequently uncertain prediction. Accurate data exploration can instead improve its quality and then overcome the initial cost of implementing redundant monitoring strategies with all the benefits, both in terms of cost and environmental impact reduction, an accurate failure detection could provide.

References

[1] A. Theissler, J. Pérez-Velázquez, M. Kettelgerdes, G. Elger, Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry, Reliability engineering & system safety 215 (2021) 107864.

[2] O. Kaiwartya, A. H. Abdullah, Y. Cao, A. Altameem, M. Prasad, C.-T. Lin, X. Liu, Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects, IEEE Access 4 (2016) 5356–5373.

[3] K. M. Alam, M. Saini, A. El Saddik, Toward social internet of vehicles: Concept, architecture, and applications, IEEE access 3 (2015) 343–357.

[4] K. Golestan, R. Soua, F. Karray, M. S. Kamel, Situation awareness within the context of connected cars: A comprehensive review and recent trends, Information Fusion 29 (2016) 68–83. URL: https://www.sciencedirect.com/science/article/pii/S1566253515000743. doi:https://doi.org/10.1016/j.inffus.2015.08.001.

[5] A. Elfar, A. Talebpour, H. S. Mahmassani, Machine learning approach to short-term traffic congestion prediction in a connected environment, Transportation Research Record 2672 (2018) 185 – 195.

[6] S. J. Kamble, M. R. Kounte, Machine learning approach on traffic congestion monitoring system in internet of vehicles, Procedia Computer Science 171 (2020) 2235–2241. URL: https://www.sciencedirect.com/science/article/pii/S1877050920312321. doi:https://doi.org/10.1016/j.procs.2020.04.241, third International Conference on Computing and Network Communications (CoCoNet'19).

[7] R. Sun, Y. Chen, A. Dubey, P. Pugliese, Hybrid electric buses fuel consumption prediction based on real-world driving data, Transportation Research Part D: Transport and Environment 91 (2021) 102637. URL: https://www.sciencedirect.com/science/article/pii/S1361920920308221. doi:https://doi.org/10.1016/j.trd.2020.102637.

[8] Z. Xu, T. Wei, S. Easa, X. Zhao, X. Qu, Modeling relationship between truck fuel consumption and driving behavior using data from internet of vehicles, Computer-Aided Civil and Infrastructure Engineering 33 (2018). doi:10.1111/mice.12344.

[9] J.-D. Wu, J.-C. Liu, Development of a predictive system for car fuel consumption using an artificial neural network, Expert Systems with Applications 38 (2011) 4967 – 4971. URL: http://www.sciencedirect.com/science/article/pii/S0957417410011127. doi:https://doi.org/10.1016/j.eswa.2010.09.155.

[10] P. Bethaz, S. Cavaglion, S. Cricelli, E. Liore, E. Manfredi, S. Salio, A. Regalia, F. Conicella, S. Greco, T. Cerquitelli, Empowering commercial vehicles through data-driven methodologies, Electronics 10 (2021) 2381.