UNIVERSITY OF VAASA

THE SCHOOL OF TECHNOLOGY AND INNOVATION

INDUSTRIAL DIGITALISATION

Mikko Paaskoski

MASTER'S THESIS

Data-driven approaches to support engine performance characterization.

Master's thesis for the degree of Master of Science in Technology submitted for inspection, Vaasa, 5 December 2019.

Thesis supervisor	Prof. Mohammed Elmusrati
Thesis instructors	D.Sc. Irene Gallici D.Sc. Andrea Greco

PREFACE

This Master's Thesis project was done as an assignment for Wärtsilä Finland Oy's Systems Reliability team. The project has been truly educational project and has provided a lot of new knowledge for me related to internal combustion engines, big data engineering and machine learning.

During this project, I got support from many great individuals. I would like to sincerely thank my Thesis supervisor Mohammed Elmusrati and instructors Irene Gallici and Andrea Greco for their excellent support, guidance and feedback. I would also like to thank the whole Systems Reliability team for their important support during the whole Thesis project. In addition, I am highly grateful for Ilona Söchting, who first hired me to Wärtsilä as trainee in summer 2018 and served as a data science mentor to me.

Finally, I want to address my gratitude for my friends and my family, especially for my mother, for the support they have provided during my studies.

Vaasa, 5.12.2019

Mikko Paaskoski

TABLE OF CONTENTS

ABBREVATIONS

AWS	Amazon Web Services
BDC	Bottom dead centre
CI	Compression-ignition
DF	Dual fuel
EDS	Engine Data Sandbox
ICE	Internal combustion engine
IDE	Integrated development environment
MTBF	Mean time between failures
MTTF	Mean time to failure
MTTR	Mean time to repair
RL	Reinforcement learning
SG	Spark-ignition gas
SHD	Shutdown
SL	Supervised learning
SSL	Semi-supervised learning
SI	Spark-ignition
TDC	Top dead centre
UL	Unsupervised learning

UNIVERSITY OF VAASA
The School of Technology and Innovation

Author:	Mikko Paaskoski
Topic of the Thesis:	Data-driven approaches to support engine performance characterization
Supervisor:	Prof. Mohammed Elmusrati
Instructors:	D.Sc. Irene Gallici D.Sc. Andrea Greco
Degree:	Master of Science in Technology
Major of Subject:	Industrial Digitalisation
Year of Entering the University:	2014
Year of Completing the Thesis:	2019 **Pages:** 70

ABSTRACT

Engine Data Sandbox is a data repository containing sensor data measured from over 1000 different Wärtsilä engines, which have been operating in marine and power plant applications throughout several years. The Engine Data Sandbox comprehends over 10 terabytes of raw sensor data, when data is uncompressed. Considering this huge amount of data, Engine Data Sandbox potentially contains a lot of hidden, valuable information. In this thesis, Engine Data Sandbox content is described and mapped. Furthermore, utilizing contained raw data, four different data-driven approaches were developed in order to support engine performance characterization and reliability engineering analysis. In addition, during the development of these approaches, comprehensive set of different data preparation functionalities were developed in order to preprocess the raw data of Engine Data Sandbox.

Thesis author developed the data-driven approaches relying on the R programming language. Developed data-driven methodologies are:

- Load distribution analysis.
- Automatic shutdown analysis.
- Main feature extraction for relevant sensor signals.
- Anomaly detection of sensor signals.

Obtained results provided the possibility to characterize engines behaviour on field. Furthermore, they allowed to preliminary investigate engines health over the operating lifetime.

The potential usages and limitations, for the data of Engine Data Sandbox, were also identified in this thesis.

KEYWORDS: Reliability engineering, internal combustion engines, big data analysis machine learning, engine health monitoring

VAASAN YLIOPISTO	
Tekniikan ja innovaatiojohtamisen yksikkö	
Tekijä:	Mikko Paaskoski
Diplomityön nimi:	Data-driven approaches to support engine performance characterization
Valvojan nimi:	Prof. Mohammed Elmusrati
Ohjaajien nimet:	TkT Irene Gallici TkT Andrea Greco
Tutkinto:	Diplomi-insinööri
Oppiaine:	Industrial Digitalisation
Opintojen aloitusvuosi:	2014
Diplomityön valmistumisvuosi:	2019 **Sivumäärä:** 70

TIIVISTELMÄ

Engine Data Sandbox on tietovarasto, joka sisältää sensoridataa yli 1000:sta eri Wärtsilän valmistamasta moottorista. Tätä sensoridataa on kerätty eri laiva- ja voimalaitossovelluksista usean eri vuoden ajalta. Engine Data Sandbox käsittää yli 10 teratavua dataa, kun data on pakkaamattomassa muodossa. Ottaen huomion tämän suuren datamäärän, Engine Data Sandbox sisältää potentiaalisesti paljon arvokasta, piilotettua tietoa. Tässä diplomityössä esitellään Engine Data Sandboxin sisältö sekä tutkielman aikana neljä kehitettyä datavetoista sovellusta, jotka hyödyntävät Engine Data Sandboxin raakadataa. Näide datavetoisten sovellusten tarkoituksena on tukea Wärtsilän moottoreiden luotettavuuden analysointia sekä käyttäytymisen karakterisointia. Näiden sovellusten lisäksi tutkimuksen aikana kehitettiin huomattava määrä erilaisia toiminnallisuuksia Engine Data Sandboxin raakadatan preprosessointiin.

Tutkielman aikan kehitettiin R-ohjelmointikielen avulla seuraavat neljä datavetoista sovellusta:

- Sovellus analysoimaan moottorin kuorman jakautumista.
- Sovellus analysoimaan moottorin automaattisten poiskytkentöjen syitä.
- Sovellus merkityksellisten sensorisignaalien pääpiirteiden määrittämiseen.
- Sovellus sensorisignaalien poikkeavan käytöksen löytämiseksi.

Saadut tulokset mahdollistavat käytössä olevien moottoreiden käyttäytymisen karakterisoinnin. Lisäksi tulokset mahdollistavat alustavan moottoreiden kunnon estimoinnin niiden eliniän aikana.

Myöskin Engine Data Sandboxin potentiaalinen käyttötarkoitus sekä rajoitteet tunnistettiin tutkielman aikana.

AVAINSANAT: Luotettavuustiede, polttomoottorit, big data -analyysi, koneoppiminen, moottorin kunnonvalvonta

1 INTRODUCTION

Engine Data Sandbox is a data repository containing sensor data measured from over 1000 different Wärtsilä engines, which have been operating in marine and power plant applications throughout several years. The Engine Data Sandbox (EDS) comprehends over 10 terabytes of raw sensor data, when data is uncompressed. Considering this huge amount of data, Engine Data Sandbox potentially contains a lot of hidden, valuable information. Currently, the EDS data is stored in the Amazon Web Services (AWS) infrastructure.

Wärtsilä has a wide engine portfolio and provides solutions to several applications and for different segments. Data available in EDS grants the opportunity to investigate operating performance of different engines in order to look for peculiar characteristics and to investigate differences in behavior during operations.

EDS data analysis was performed relying both on analytic and machine learning approaches.

1.1 Objective of the Thesis

Main objective of this Master's Thesis can be summarized as follow:

- Characterization of Engine Data Sandbox:
  - Accessibility to data repository.
  - Description of contents.
- Engine performance profiling relying on Engine Data Sandbox:
  - Theoretical approaches.

    - Test case developments.

    - Test case results.

- Investigation of innovative solutions to treat Engine Data Sandbox contents, such as main feature extraction of relevant sensor signals and anomaly detection of sensor signals.

- Definition of potential development and next steps.

Developed approaches and extracted results within this Thesis are limited to dual fuel (DF) and spark-ignited gas (SG) engines. DF and SG engines were selected since they are latest products provided by Wärtsilä in order to reduce emissions levels and they can be seen as a technology-bridge towards hydrogen utilization as main fuel and, therefore, zero carbon emissions.

1.2 Thesis Contributions

All the presented data-driven approaches in this thesis were implemented with R programming language by the thesis author. All the results presented in this thesis were derived from the raw EDS data and the all functionalities required to derive these results were also implemented by the thesis author.

1.3 Structure of the Thesis

This Thesis consists of 6 different chapters:

- Chapter 2 presents relevant foundations and background information related to this Thesis.

- Chapter 3 introduces EDS: the EDS content is overviewed, and accessibility is described.

- Chapter 4 presents the implementation of data-driven approaches and the rules and definitions, which are required to be followed during the implementation of algorithms, for the extraction of information from the raw EDS data.

- Chapter 5 collects the results produced by developed data-driven approaches.

- Chapter 6 concludes the Thesis with discussion about conclusions and possible future developments.

2 FOUNDATIONS

2.1 Reliability engineering

Reliability engineering is an engineering field, which objectives consists of preventing failures or minimizing probability and quantity of those failures in products, identifying causes of occurring failures, defining means to cope with occurring failures in the situations when the cause of failure has not been fixed, and utilizing different approaches to estimate the reliability of new products and designs (O'Connor & Kleyner 2012: 2). Reliability engineering is required to ensure high reliability of different products and equipment during their product lifecycle in addition to high confidence and competitive costs. (Kececioglu 2002: 3). Reliability engineering should be included to support different project activities, concurrent engineering and quality assurance, in order it to be time and cost effective. (Birolini 2013: 1)

In the following subchapter, essential concepts related to reliability engineering are presented.

2.1.1 Concepts of reliability engineering

Before presenting the reliability concepts, differences between non-repairable and repairable systems must be defined.

Non-repairable systems are discarded and replaced, and repairable systems are repaired, when the failure occurs. This does not necessarily mean that non-repairable systems are unrepairable, rather that repairing those systems is not economically reasonable. Repairable systems are repaired when the failure occurs, if replacing or repairing the failed components of the system is economically feasible. (Topuz 2009: 234)

Below, reliability concepts are presented.

Reliability: probability for the event, that during a certain time interval and under certain operating conditions a service will be provided, or product will operate, without a failure (Elsayed 2012: 3). For non-repairable system, when the failure is allowed to occur only once, reliability is the probability for system to survive over its estimated lifetime. For repairable system, when the failure is allowed to occur more than once, reliability is the probability for the event that failure does not occur within certain time interval. (O'Connor & Kleyner 2012: 8)

Failure rate is applicable for both non-repairable and repairable systems. Failure rate is number of occurring failures per certain time unit, when failure is allowed to occur once or more in time continuum. (O'Connor & Kleyner 2012: 8)

Mean time to failure (MTTF), Mean time to repair (MTTR) and Mean time between failures (MTBF):

1. MTTF: Mean time to failure, is applicable for non-repairable systems. MTTF indicates the average operating time of system before failure. (Gnedenko & Ushakov 1995: 87)

2. MTTR: Mean time to repair, is applicable for repairable systems. MTTR informs the needed time to replace or repair the failed hardware module. (Topuz 2009: 234)

3. MTBF: Mean time between failures, is applicable for repairable systems. Can be defined as MTTF of repairable system. In this case, MTBF indicates the average operating time of system before failure. MTBF can also be defined as average time between failures. With this definition, MTBF consist of the average operating time of system before failure (MTTF) and time needed to repair the system (MTTR) (Lazzaroni, Cristaldi, Peretto, Rinaldi & Catelani 2011: 87). Mathematically this can be expressed as:

$$MTBF = MTTF + MTTR$$

Availability: probability for the event, that the system or unit is operational (Topuz 2009: 234). Mathematically this can be expressed as:

$$Availability = \frac{MTBF}{MTBF + MTTR}$$

In the formula above, MTBF is considered as the average operating time of system before the failure.

Maintainability: probability for the event, that for a certain item, repair or preventive maintenance will be performed during a certain time interval with a certain resources and procedures. (Birolini 2013: 8)

2.2 Internal combustion engines

Internal combustion engines (ICE) are engines, which are designed to produce mechanical power from the chemical energy. The chemical energy is released from the fuel residing inside the engine, either by oxidizing or burning the fuel. Required power output is produced by work, which occurs between mechanical components of engine and the working fluids. In the ICE, the working fluids are the burned products following the combustion and the mixture of air and fuel prior the combustion. Design and operating characteristics in ICEs are essentially differing from other engine types due to combustion occurs inside the ICE. (Heywood 1988: 1-2)

ICEs are usually considered to be reciprocating internal combustion engines, which are categorized in two main types: compression-ignition (CI) engines and spark-ignition (SI) engines. Principle of the CI engines is to compress the air to a high pressure and temperature. This supports the combustion to occur spontaneously when the fuel is injected. Operation of SI engines is based on spark plug, which ignites the mixture of air and fuel in the engine. (Foanene 2016: 166)

2.2.1 Classification

Different ICEs can be classified by different means. Below, some of the means listed by John Heywood (1988: 7), are presented:

1. Applications: different applications, in which ICEs are designed to operate, are for instance power generation, marine, locomotive, automobile and light aircraft applications.

2. Basic engine design: different basic engine designs for ICEs are reciprocating engines and rotary engines. Reciprocating engines are classified according to arrangement of engine cylinders, and rotary engines are categorised according to different designs, one of which is, for instance, Wankel design.

3. Working cycle: different working cycles for ICEs include, for instance, four-stroke cycle and two-stroke cycle.

4. Fuel: different fuels for ICEs include fuel oil, diesel oil, gasoline, petrol, natural gas, dual fuel, hydrogen, alcohols (ethanol, methanol) and liquid petroleum gas.

5. Method of ignition: different ignition methods for ICEs include spark ignition and compression ignition.

2.2.2 Four-stroke and two-stroke operating cycles

Both SI engines and CI engines can be designed to operate in two-stroke or in four-stroke operating cycles. (Stone 1999: 1)

Four-stroke operating cycle consists of 4 different phases: in the first phase, which is called the induction stroke, air is drawn in cylinder by the piston traveling down the cylinder while the inlet valve is open. In the second phase (the compression stroke), ignition occurs in the end of the phase, when the piston, which is traveling up the cylinder at this point, reaches the top dead centre (TDC) position, while the both valves are closed. In the third phase, (the working, power or expansion stroke) combustion, which occurs due to ignition, raises temperature and pressure, and forces piston to bottom dead centre (BDC) position from TDC, creating mechanical energy in the process. In the end of the third phase, the exhaust valve opens. In the fourth and last phase (the exhaust stroke), piston travels back from BDC to TDC while the exhaust valve remains open, expelling remaining gases. (Stone 1999: 1-2)

Two-stroke operating cycle consists of 2 different phases (the compression stroke and the power stroke), excluding induction and exhaust strokes, which are included in four-stroke operating cycle. When comparing two-stroke operating cycle to four-stroke operating cycle, two-stroke engines are more powerful, due to two-stroke engines have two times more power strokes than four-stroke engines per unit time. However, the efficiency in four-stroke engines is likely higher. (Stone 1999: 2-3)

2.2.3 Spark-ignition engines

As mentioned before, combustion process in SI engines is based on spark plug, which ignites the mixture of fuel and air. In this subchapter, additional information considering the ignition occurring in SI engines and also some alternatives as fuels for SI engines are presented.

Externally supplied ignition is responsible for starting the combustion process in SI engines by igniting the mixture of fuel and air at the correct time. Ignition is generated by producing electric spark in combustion chamber between electrodes of a spark plug. Reliable ignition under all conditions is required to secure engine operation without faults. Misfiring could lead to low engine output, high consumption or poor exhaust emission figures. By selecting the moment of ignition, the start of the combustion can be controlled in SI engines. Knock limit determines the earliest possible moment of ignition and the latest possible moment of ignition is determined by the maximum allowed exhaust gas temperature. Fuel consumption, exhaust gas emissions and delivered torque are all influenced by moment of ignition. In order to deliver maximum combustion and engine torque, maximum combustion pressure should occur shortly after piston has reached TDC, and this is achieved by timing the ignition to occur before piston reaches TDC and therefore the moment of ignition should be advanced. Advanced moment of ignition reduces fuel consumption and increases power, but also increases nitrogen-oxide and hydrocarbon emissions. Too advanced moment of ignition could cause engine knocking which can damage the engine and too late moment of ignition results higher exhaust gas temperatures which could also damage the engine. (Bosch 2011: 570-572)

SI engine fuels include gasoline, methanol, ethanol, natural gas and hydrogen. Engines operating with gaseous fuels (which include natural gas and hydrogen), are considered to have advantages (which include reduced emissions for instance) over engines operating with gasoline. Natural gas can be used either as compressed natural gas or as liquid natural gas and from these two, compressed natural gas is more common since liquid natural gas is more expensive and more difficult to handle. Major disadvantage related to natural gas is the fact that the gas must be stored in heavy high-pressure tank, which reduces the payload. Hydrogen has many advantages related to combustion process. These advantages include wide flammability limits and high flame speed. However, as in the case of natural gas, major disadvantage is the heavy, expensive tank required to contain the hydrogen. (Najjar 2009: 1-3)

2.2.4 Dual fuel engines

Although the diesel engines are used widely throughout the world, due to their cost-effectiveness, adaptability, reliability and efficiency, they are considered to be one of the main contributors for the environmental pollution. At the same time, the energy demand is increasing, and oil resources are decreasing. When considering the reduction of emissions, increasing energy demand and decreasing oil resources, the usage of alternative fuels is considered as one of the solutions for these challenges. One of these alternative fuels is natural gas, however due to low cetane number and high autoignition temperature compared to diesel fuel, ignition source is required to ignite the natural gas in the cylinder of diesel engine. The way to apply natural gas in diesel engine, is to utilize dual fuel technology (Wei & Geng 2016: 265-266). In this subchapter, three different dual fuel engine concepts are briefly presented and operating principles of one of these concepts, conventional dual fuel combustion engine, is described in more specific.

Compressed natural gas can be utilized in both SI and CI engines. When comparing CI engine and SI engine, CI engine has higher compression ratio which means better thermal efficiency. Due to the high autoignition temperature of natural gas, it will not ignite in conventional CI engines, hence the dual fuel combustion process must be implemented. There are three different dual fuel engine concepts which are derived from three types of dual fuel combustions. First engine concept is high pressure direct injection dual fuel engine, where both fuels are directly injected into the cylinder. In second engine concept, referred as conventional duel fuel engine, diesel is injected through the injector, directly into the cylinder while natural gas is injected into the intake manifold. The third combustion process, which is called dual fuel homogeneous charge compression ignition, both fuels are premixed, and port injected. In this approach, phasing and combustion intensity are controlled by fuel blending, intake conditions (pressure and temperature) and equivalence ratio. (Taritaš, Sremec, Kozarac, Blažić & Lulić 2017: 2-3)

Combustion process in conventional dual fuel engine is a combination of flame propagation (usual in SI engines) and mixing-controlled combustion process (usual in CI engines). Conventional dual fuel process consists of three different phases: premixed combustion of the diesel (1), mixing-controlled combustion of the diesel (2) and flame propagation through the premixed mixture of natural gas and air (3). In conventional dual fuel engine, natural gas, which is injected in the intake manifold, and air, are mixed. This mixture of natural gas and air is directed to cylinder during the induction stroke and compressed during the compression stroke. Due to the high autoignition temperature of natural gas, it does not ignite at the end of the compression stroke and due to this, small amount of diesel fuel is injected in the cylinder. The diesel evaporates, and multiple ignition sources are created by the ignited mixture of evaporated diesel and charge, for the mixture of natural gas and air to be utilized. Finally, when the suitable conditions in combustion chamber are achieved, multiple flames propagate through the mixture of natural gas and air. (Taritaš, Sremec, Kozarac, Blažić & Lulić 2017: 3-4)

2.3 Big data

Big data is a term that is mainly used to describe the huge amount of data in the era when the volume of data has increased considerably. When comparing big data with traditional datasets, big data frequently includes unstructured data which requires more real-time analysis. Also, big data provides new possibilities to discover new information from the data and helps to gain understanding of the new information. In addition to this, big data creates new challenges which include for instance, how to efficiently manage and process these large datasets. (Chen, Mao & Liu 2014: 171)

In this subchapter, big data characteristics are briefly described, and opportunities and challenges generated by big data are reviewed.

2.3.1 Characteristics

Different individuals, organisations and researchers have given various definitions for big data and these definitions include multiple big data characteristics. Below are presented some characteristics listed by Gayatri Kapil, Alka Agrawal and R. A. Khan (2016: 111):

1. Volume (size of the data): describes the quantity of collected and stored data.

2. Velocity (speed of the data): describes the transfer rate of data between the source and the destination.

3. Value (importance of the data): describes the business value to be derived from the data.

4. Variety (type of the data): describes the different types and formats of the data.

5. Veracity (data quality): describes the quality of the data. If data is not trustworthy enough, the data is virtually worthless to be accurately analysed.

6. Validity (data authenticity): describes the correctness of the data which is used to extract information.

7. Volatility (duration of usefulness): describes how long the stored data is useful for the user.

2.3.2 Opportunities and challenges

For the current enterprises, utilization of the valuable information extracted from the big data is a basic competitive strategy. By utilizing the valuable information extracted from the big data, enterprises have possibility to gain multiple advantages, which include improved customer service, improved operational efficiency and new markets. In addition to new possibilities and opportunities, the big data also provides new challenges. These challenges are related to data management: for instance, data storing, sharing, searching, visualization and analysing are challenges that must be overcome in order to maximize the benefits that correct utilization of big data can provide. When considering big data analysis, challenges include data incompleteness, inconsistency, scalability, timeliness and security. Before the data can be analysed, the data must be well constructed. This can be achieved via proper data preprocessing in order to improve data quality. Since the data can be highly incomplete, noisy and inconsistent, various different data preprocessing methods, which include data cleaning, transformation, reduction and integration, should be applied to data preprocessing process in order to remove noise and inconsistencies from the data. (Khan, Yaqoob, Hashem, Inayat, Ali, Alam, Shiraz & Gani 2014: 14)

2.4 Machine learning

Machine learning is an application of artificial intelligence which provides the means for system or machine to learn and improve its performance by utilizing the example or historical data. In the machine learning, execution of the computer program, which utilizes the data, is the learning that optimizes the parameters of the predefined model. This model can be either descriptive to gain information and knowledge from the data or predictive to make predictions after learning from the data. The model can also be predictive and descriptive at the same time. Since the main objective is to make inferences from the data, models mentioned before are mathematical and build by utilizing the theory of statistics. (Alpaydin 2010: 3-4)

In the following subchapters, different machine learning techniques are briefly presented, and concepts of anomaly detection and linear regression are overviewed.

2.4.1 Reinforcement learning

Reinforcement learning (RL) is a machine learning technique that utilizes agents, which learn how to act according to punishments or rewards they receive from the certain environment. This way RL agent learns what is good action and what is bad action in the environment. The goal of these agents is to perform actions which maximizes the amount of rewards and minimizes the amount of punishments (Ravishankar & Vijayakumar 2017: 1). RL algorithms are utilized in applications where the system output is sequence of actions and in these systems the important matter is to execute correct sequence of actions in order to accomplish the objective. For instance, in a game playing, the objective is accomplished with the correct sequence of actions hence one single action is not important by itself. (Alpaydin 2010: 13)

2.4.2 Supervised learning

Supervised learning (SL) is a machine learning technique which utilizes labelled training data set, which consist of input and output values. SL estimates the unknown function of the system, which has provided the values in a training set, and provides the hypothesis function that approximates the true, unknown function. The accuracy of the hypothesis function is estimated with a test data set, which is distinct from the training set but is also provided by the same, true unknown function which has provided the values of the training set. The learning problem is a classification problem when the output value is one of the values in the finite set, and when the output value is a number, the learning problem is called regression. (Russel & Norvig 2010: 695-696)

2.4.3 Semi-supervised learning

Semi-supervised learning (SSL) is a machine learning technique which can be considered as technique between supervised learning and unsupervised learning. SSL uses data which is unlabelled, but also has some labelled information included. For instance, the data set which is utilized by the SSL algorithm could consist of some observations which labels are provided (for example: both input and output values are provided for the observation) and some observations which labels are not provided (for example: only input values are provided for the observation). (Chapelle, Schölkopf & Zien 2006: 2)

2.4.4 Unsupervised learning

Unsupervised learning (UL) is a machine learning technique which utilizes the data which only has input values, excluding the output values. The objective of the UL is to discover regularities from the input values, while the objective of the SL is to learn from the data which has both input and output values, in order to map the output values from the input values. In the input space, there is a structure where certain patterns appear often and finding these patterns can be done with density estimation. One of the methods of density estimation is called clustering, where the objective is to discover groupings or clusters of input values. (Alpaydin 2010: 11)

2.4.5 Anomaly detection

Anomaly detection is the concept to discover patterns from the data that do not follow the expected behaviour, and these patterns are often called as anomalies or outliers. Anomaly detection is crucial since the anomalies in the data can be interpreted as critical information. Anomaly detection is utilized in various different applications, including for instance: fraud detection, fault detection and cyber-security. The formulation of a specific anomaly detection problem is affected by multiple factors, which include the nature of the data and the type of anomalies that have to be detected. Different concepts from the fields such as statistics, data mining, information theory, machine learning and spectral theory, have been applied to these specific anomaly detection problems. (Chandola, Banerjee & Kumar 2009: 15:1-15:4)

2.4.6 Linear regression

Linear regression is an approach to modelling the linear relationship between one or multiple response variables (also called dependent variables and are representing outputs) and one or multiple predictor variables (also called independent variables and are representing inputs). Linear regression is utilized to relate response variables to predictor variables and is considered as estimation of the parameters of the model in a certain system. (Rencher 2002: 322)

Linear regression can be subdivided into three different cases according to number of response and predictor variables. Below, these 3 cases listed by Alvin C. Rencher (2002: 322) are presented:

1. Simple linear regression: includes one response variable and one predictor variable. In this case, the objective is to predict one response variable based on one predictor variable.

2. Multiple linear regression: includes one response variable and multiple predictor variables. In this case, the objective is to predict one response variable based on multiple predictor variables.

3. Multivariate multiple linear regression: multiple response variables and multiple predictor variables, in this case, the objective is to predict multiple response variables based on multiple predictor variables.

3 ENGINE DATA SANDBOX

3.1 EDS description

EDS is a data repository which contains measured sensor data from over 1000 different Wärtsilä engines. The data has been collected from several different engine types, including both marine and power plant applications. Period, for which the data has been collected, is engine specific and could vary from few days to several years. Currently, the EDS data repository is stored in Amazon Web Services (AWS) environment.

EDS was developed in order to provide framework, which allows possibility to learn and test different approaches of big data analytics, with large amount of engine data. Main details of EDS can be summarized as follow:

1. *Data available:* sensor signals are available for relevant systems and main operating parameters of engine. For instance, available signals include engine speed, engine load and different temperatures, pressures and flow rates measured within the engine. However, the amount of measurements is engine specific. At the beginning of this thesis development process, EDS contained installation specific daily files, engine specific raw data files, and also 2-minute aggregated files and running log files for certain engines. These files are covered more in detail in subchapter 3.2.2.

2. *Sampling frequency:* in order to reduce the amount of collected and measured data from the engines, data collection systems of the engines followed the dead banding approach. This means that a value of the signal is only recorded when it varies a certain, predefined amount from last recorded value. Also, signal value is recorded when there has been approximately 10 minutes since the last signal value recording. Sampling frequency and dead banding are covered more in detail in subchapter 3.2.3.

3. *Accessibility:* EDS data can be accessed through S3 Browser (software that is used to interact with AWS data repositories) and Amazon EC2 instance (virtual server in Amazon's Elastic Compute Cloud that is used to run applications in AWS infrastructure). Personal credentials are required in both cases in order to access EDS data. Accessibility is covered more in detail in subchapter 3.4.

3.2 EDS content

This subchapter provides an overview of properties of EDS installations and engines as well as different data files located in EDS. In addition, it presents the structure of the available raw data.

3.2.1 Installations and engines

EDS contains data for 222 different installations. 208 of these installations are identified and the rest 14 installations are unidentified. Reason for inability to identify some of the installations in EDS is covered more in detail in subchapter 3.3.

Out of 208 identified installations, 166 are operating in power plant applications and 42 are operating in marine applications. These 208 installations include 1112 engines (925 in power plant applications and 187 in marine applications).

All identified engines are 4-stroke engines. Following list provides main features of engines whose operating data are collected in the EDS:

- Bore sizes (in millimetres from smallest to largest): 200, 220, 250, 260, 280, 320, 340, 380, 400, 460 and 500.

- Engine configurations: Inline cylinder configuration (L), radial cylinder configuration (R), V-cylinder configuration (V).

- Engine Extensions: Dual Fuel (DF), Spark Gas (SG), etc.

- Number of cylinders per engine: 6, 8, 9, 12, 16, 18 and 20.

- Fuel types: gas, heavy fuel oil, light fuel oil, marine diesel oil and liquid biofuel.

Due to the significant amount of available data, investigated engines in this thesis were limited to engines using either spark-ignition gas (SG) of dual fuel (DF) technology. DF and SG engines were selected since they are latest products provided by Wärtsilä in order to reduce emissions levels and they can be seen as a technology-bridge towards hydrogen utilization as main fuel and, therefore, zero carbon emissions. Total combined number of SG and DF engines in EDS is 473, which comprehends total of 105 installations.

From now on, engines are referred by their respective engine platforms, for instance: W50DF engine. In the abbreviation, W is for Wärtsilä (is included every time in abbreviation as prefix), 50 is bore size of engine in centimetres (2 digits after constant “W”) and last two letters (in this case DF, i.e. Dual Fuel) provide details about engine extension.

3.2.2 Data files

EDS contains 4 main types of data files: installation specific daily files, engine specific raw data files, 2-minute aggregated files and running log files.

Installation specific daily files, which are in .csv format, contain all signal data from single specific day for every engine of the installation. Number of daily files per installation could vary from couple of days to over 1000 days. Also, daily file size could vary since it depends on different features, for instance, the number of engines in the installation, and the number of engine specific signals.

Engine specific raw data files (also in .csv format) were derived from the installation specific daily files. Per each engine within each installation, the engine specific signals from every daily file were extracted, and saved in one, single engine specific raw data file. Below, is the simple figure to describe the process.

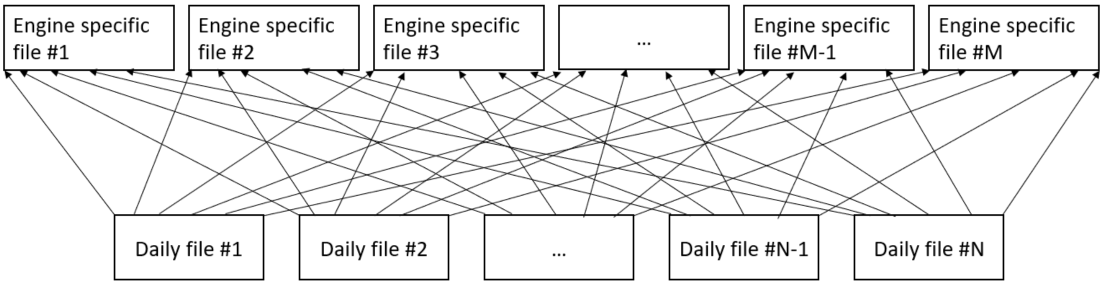
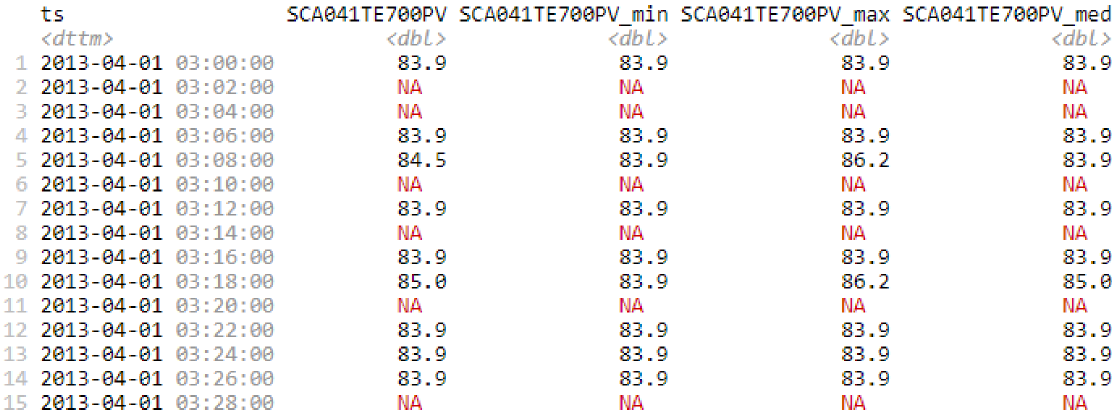


Figure 1. Signal data from N daily files of a certain installation divided into M engine specific raw data files. Here, N and M are the natural numbers with the exception N, M ≠ 0.

2-minute aggregated files (.csv format) were enriched from the engine specific raw data files. In these files, for each signal present in the engine specific raw data file, there are mean, maximum, minimum and median values calculated for every consecutive 2-minute time period. Figure below presents the data sample from 2-minute aggregated data file.

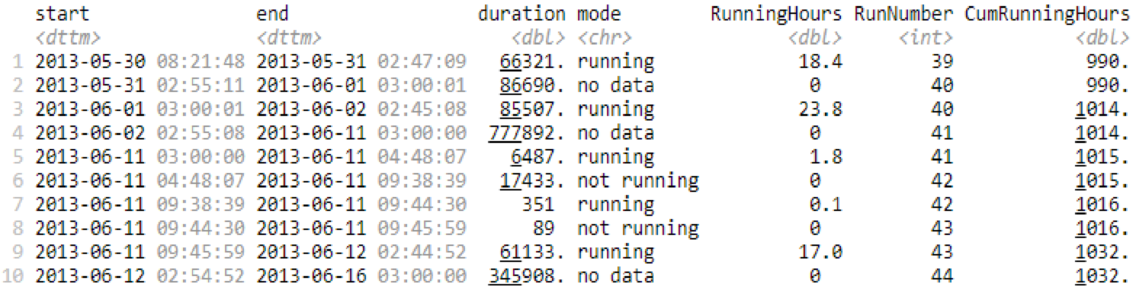


```
   ts                  SCA041TE700PV SCA041TE700PV_min SCA041TE700PV_max SCA041TE700PV_med
   <dttm>                      <dbl>             <dbl>             <dbl>             <dbl>
 1 2013-04-01 03:00:00          83.9              83.9              83.9              83.9
 2 2013-04-01 03:02:00          NA                NA                NA                NA  
 3 2013-04-01 03:04:00          NA                NA                NA                NA  
 4 2013-04-01 03:06:00          83.9              83.9              83.9              83.9
 5 2013-04-01 03:08:00          84.5              83.9              86.2              83.9
 6 2013-04-01 03:10:00          NA                NA                NA                NA  
 7 2013-04-01 03:12:00          83.9              83.9              83.9              83.9
 8 2013-04-01 03:14:00          NA                NA                NA                NA  
 9 2013-04-01 03:16:00          83.9              83.9              83.9              83.9
10 2013-04-01 03:18:00          85.0              83.9              86.2              85.0
11 2013-04-01 03:20:00          NA                NA                NA                NA  
12 2013-04-01 03:22:00          83.9              83.9              83.9              83.9
13 2013-04-01 03:24:00          83.9              83.9              83.9              83.9
14 2013-04-01 03:26:00          83.9              83.9              83.9              83.9
15 2013-04-01 03:28:00          NA                NA                NA                NA  
```

Figure 2. Data sample taken from 2-minute aggregated file.

The example above shows 5 different columns: in the first column there are time stamps between every two minutes, and columns from 2 to 5 are mean, minimum, maximum and median values of single signal for corresponding 2-minute period.

Like the 2-minute aggregated files, also the running log files (.csv format) were enriched from the engine specific raw data files. Running log file informs the time periods when the engine has been running or has not been running, or there has not been any data concerning engine performance. The running log file also informs duration of each engine mission both in seconds and in hours, cumulative running hours and cumulative amount of engine missions. Figure below presents the data sample from running log file.

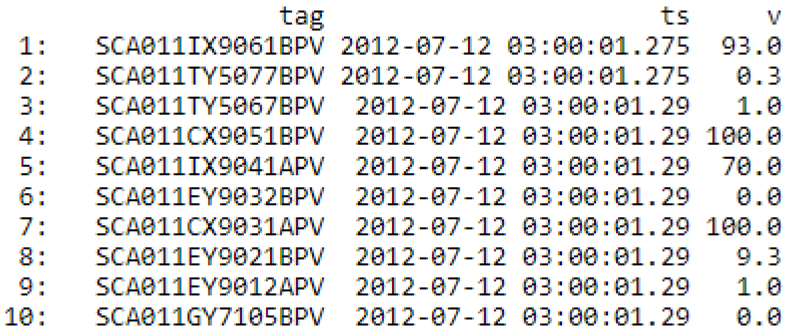


```
   start               end                 duration mode        RunningHours RunNumber CumRunningHours
   <dttm>              <dttm>                 <dbl> <chr>              <dbl>     <int>           <dbl>
 1 2013-05-30 08:21:48 2013-05-31 02:47:09   66321. running             18.4        39            990.
 2 2013-05-31 02:55:11 2013-06-01 03:00:01   86690. no data              0          40            990.
 3 2013-06-01 03:00:01 2013-06-02 02:45:08   85507. running             23.8        40           1014.
 4 2013-06-02 02:55:08 2013-06-11 03:00:00  777892. no data              0          41           1014.
 5 2013-06-11 03:00:00 2013-06-11 04:48:07    6487. running              1.8        41           1015.
 6 2013-06-11 04:48:07 2013-06-11 09:38:39   17433. not running          0          42           1015.
 7 2013-06-11 09:38:39 2013-06-11 09:44:30     351  running              0.1        42           1016.
 8 2013-06-11 09:44:30 2013-06-11 09:45:59      89  not running          0          43           1016.
 9 2013-06-11 09:45:59 2013-06-12 02:44:52   61133. running             17.0        43           1032.
10 2013-06-12 02:54:52 2013-06-16 03:00:00  345908. no data              0          44           1032.
```

Figure 3. Data sample taken from running log file.

3.2.3 Raw data

In this subchapter, structure of the raw data is overviewed. Installation specific daily files and engine specific raw data files have this raw data structure. Figure below presents the data sample from engine specific raw data file.

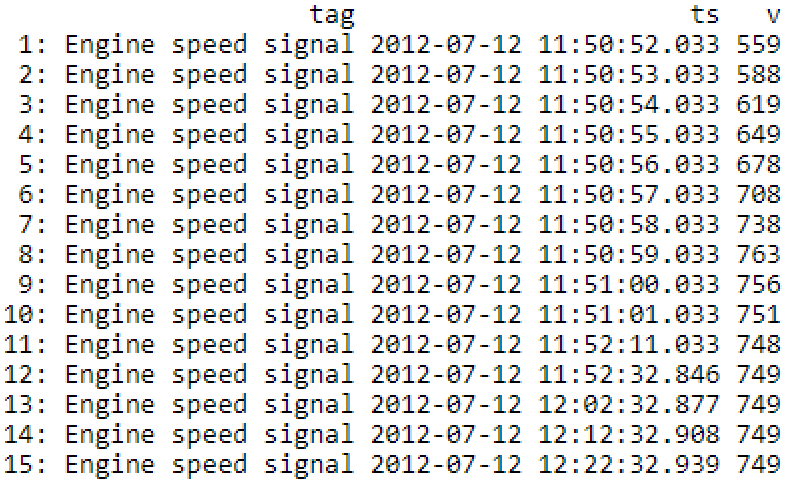


```
                tag                       ts     v
 1:    SCA011IX9061BPV 2012-07-12 03:00:01.275  93.0
 2:    SCA011TY5077BPV 2012-07-12 03:00:01.275   0.3
 3:    SCA011TY5067BPV  2012-07-12 03:00:01.29   1.0
 4:    SCA011CX9051BPV  2012-07-12 03:00:01.29 100.0
 5:    SCA011IX9041APV  2012-07-12 03:00:01.29  70.0
 6:    SCA011EY9032BPV  2012-07-12 03:00:01.29   0.0
 7:    SCA011CX9031APV  2012-07-12 03:00:01.29 100.0
 8:    SCA011EY9021BPV  2012-07-12 03:00:01.29   9.3
 9:    SCA011EY9012APV  2012-07-12 03:00:01.29   1.0
10:    SCA011GY7105BPV  2012-07-12 03:00:01.29   0.0
```

Figure 4. Data sample taken from engine specific raw data file.

In the figure above, every row presents unique sample measured from the engine. First column *"tag"* displays the sensor tag, which provided the signal, second column *"ts"* presents the time when the measurement has taken place, and third column *"v"* displays the value of the measurement.

In the control system, latest output value of the signal is compared to current signal value, by the deadband controller. If the absolute value, taken from difference of these two values, is smaller than predefined value (which defines how much signal value has to change from the latest output value before latest output value is updated), then the current signal value is not updated as latest output value. Otherwise the current value is updated as the latest output value of the signal (Hirche, Hinterseer, Steinbach & Buss 2005: 72). As mentioned before, there is no fixed sampling frequency for the raw data. Data collection systems follow dead band approach, so sampling frequencies of measured signals are defined by dead banding. Figure below presents an example concerning sampling frequency of the engine speed signal.



```
                    tag                      ts   v
 1: Engine speed signal 2012-07-12 11:50:52.033 559
 2: Engine speed signal 2012-07-12 11:50:53.033 588
 3: Engine speed signal 2012-07-12 11:50:54.033 619
 4: Engine speed signal 2012-07-12 11:50:55.033 649
 5: Engine speed signal 2012-07-12 11:50:56.033 678
 6: Engine speed signal 2012-07-12 11:50:57.033 708
 7: Engine speed signal 2012-07-12 11:50:58.033 738
 8: Engine speed signal 2012-07-12 11:50:59.033 763
 9: Engine speed signal 2012-07-12 11:51:00.033 756
10: Engine speed signal 2012-07-12 11:51:01.033 751
11: Engine speed signal 2012-07-12 11:52:11.033 748
12: Engine speed signal 2012-07-12 11:52:32.846 749
13: Engine speed signal 2012-07-12 12:02:32.877 749
14: Engine speed signal 2012-07-12 12:12:32.908 749
15: Engine speed signal 2012-07-12 12:22:32.939 749
```

Figure 5. 15 measurements from engine speed signal.

Time intervals between measurements in above figure show that the sampling frequency is higher when the value of the signal is changing. When engine reaches nominal speed (in this case 749 rpm) and the value is not changing, new measurement is updated approximately in every 10 minutes.

3.3 EDS exceptions and limitations

As mentioned before, out of all 222 installations present in EDS, 14 installations are unidentified. Reason for inability to identify these installations is that their names cannot be found from Wärtsilä master data.

EDS is also practically completely lacking signals registered by automation (e.g. alarms, internal operating modes, etc.). This means, that identifying certain events, which include for instance load reductions and shutdowns, have to be done based on the behavior of analog signals present in the EDS data. Lack of digital signals is a drawback when considering usage of EDS data in big data analytics.

There are also issues related to the quality of data. For instance, there could be signals present in the engine specific raw data files which have incorrect values and signal data present in raw data could be affected by noise.

It is worth highlighting that amount of engine specific signals could vary from couple of dozens to couple of hundreds. In the raw data, these signals are presented in encoded format. Some of these signal codes are mapped and identified, but there are also encoded signals present in the raw data which are not mapped, and this means that these signals cannot be correlated to any sensor tag being useless for data analytics purposes.

3.4 EDS accessibility

All the EDS data is currently stored in Amazon S3. Amazon Simple Storage Service (Amazon S3) is object storage service which provides high scalability, security, data availability and performance opportunities (AWS 2019b). The EDS data can be accessed, for instance, by using Amazon EC2 instances or S3 Browser.

Amazon Elastic Compute Cloud (Amazon EC2) is a web service which provides scalable computing capacity in the AWS cloud. Amazon EC2 provides virtual computing environments known as Amazon EC2 instances. These instances can have various configurations for memory, CPU, storage and networking capacity. These instances can be integrated with various different software, (AWS 2019a). In this thesis for example, Amazon EC2 instances integrated with RStudio (integrated development environment for R programming language) were used to access and process the EDS data.

Other example how to access the EDS data, is the usage of S3 Browser. S3 Browser is freeware for Windows which provides interface for interaction with Amazon S3 and Amazon CloudFront. S3 Browser provides possibility to interact with Amazon S3 data storages by storing and retrieving data. (S3 Browser 2018)

Both of these, Amazon EC2 and S3 Browser, require credentials for the specific Amazon S3 data storages, which user wishes to interact with. Difference between these approaches is that when using Amazon EC2, the data transmission occurs between Amazon S3 and Amazon EC2, in other words, the data resides in the AWS cloud the whole time. In the case of S3 Browser, data transmission occurs between AWS cloud and local computer of the user.

4 IMPLEMENTATION OF RESEARCH

All presented research in this thesis was implemented by utilizing algorithms created with R programming language via RStudio integrated development environment (IDE).

R is programming language for statistical computation and graphics. It is interpreted programming language which allows modular programming using functions, looping and branching (R Project 2018). RStudio is IDE for R. It includes syntax-highlighting editor which supports direct code execution, console, tools for plotting, debugging and workspace management. (RStudio 2018)

Since 2-minute aggregated files and running log files are not available for all the engines within EDS, data-driven approaches were developed relying only on the engine specific raw data files.

4.1 Rules and definitions

In order to extract useful information from EDS data, set of experimental rules and definitions were developed and tested. In the following subchapters, definitions for different events, are overviewed.

4.1.1 Detection of monitoring periods

Analysis was focused on investigation of engine operations through the investigation of time intervals when the value of monitored signal is either above or below predefined limit. For instance, the period when the engine is running, the period when the deviation in exhaust gas temperature occurs due to low temperature and the period when the engine is running in certain load interval.

In order to extract desired data and results it is necessary to extract from the whole time series only the desired time intervals. Two different approaches were used to calculate these monitoring periods. In the first approach, when the end point of monitoring period is unknown, start and end points of monitoring periods are marked to the raw EDS data based on certain conditions. In the second approach, when the end point of monitoring period and length of time interval are known, monitoring period is defined by the known information.

In the first approach, to identify start and end points of monitoring periods, proper detection rules were developed. These rules were utilized in following identification cases.

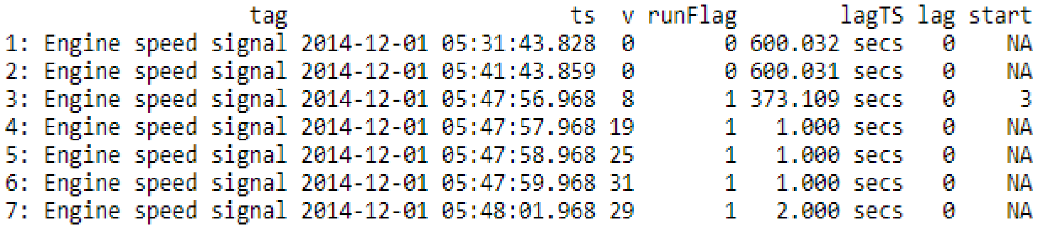
1. Monitoring period - identification of start point (Case 1)

   This identification case identified start points of monitoring periods, when monitored signal actually reaches value above or below predefined limit. The moment is considered as start point when following conditions are fulfilled:

   A) Monitored signal (i.e. speed signal) reaches the value that is above/below predefined limit (current value above/below predefined limit and previous value not above/below predefined limit, i.e. current value of speed signal > 0 rpm and previous value of speed signal = 0 rpm).

   B) Time difference between the moment considered as a start point, and the time stamp of the next sample is below or equal to 3700 seconds.

   C) Previous sample is not marked as the start point of monitoring period.



```
                     tag                      ts  v runFlag        lagTS lag start
1: Engine speed signal 2014-12-01 05:31:43.828  0       0 600.032 secs   0    NA
2: Engine speed signal 2014-12-01 05:41:43.859  0       0 600.031 secs   0    NA
3: Engine speed signal 2014-12-01 05:47:56.968  8       1 373.109 secs   0     3
4: Engine speed signal 2014-12-01 05:47:57.968 19       1   1.000 secs   0    NA
5: Engine speed signal 2014-12-01 05:47:58.968 25       1   1.000 secs   0    NA
6: Engine speed signal 2014-12-01 05:47:59.968 31       1   1.000 secs   0    NA
7: Engine speed signal 2014-12-01 05:48:01.968 29       1   2.000 secs   0    NA
```

Figure 6. Row 3 defined as the start point of monitoring period by marking it with value 3 in column “start” (Case 1 for start point identification)

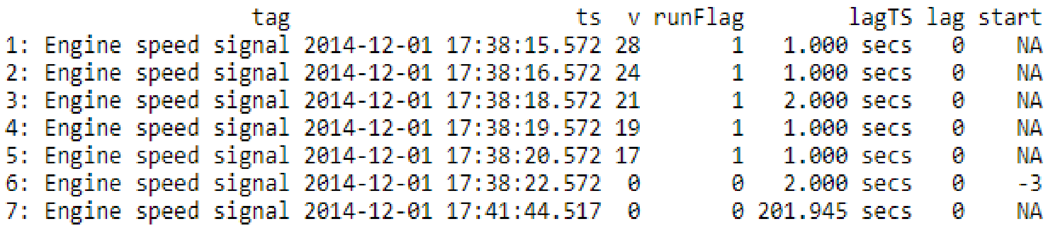
2. Monitoring period – identification of end point (Case 1).

This identification case identified end points of monitoring periods, when monitored signal is not anymore above or below predefined limit but previous sample is. The moment is considered as end point when following conditions are fulfilled:

A) Monitored signal (i.e. speed signal) is not anymore above/below predefined limit but previous sample is (current value of speed signal = 0 rpm and previous value of speed signal > 0 rpm).

B) Time difference between the moment considered as an end point, and the time stamp of the previous sample is below or equal to 3700 seconds.

C) Previous sample is not marked as the end point of monitoring period.



```
                   tag                      ts  v runFlag         lagTS lag start
1: Engine speed signal 2014-12-01 17:38:15.572 28       1   1.000 secs   0    NA
2: Engine speed signal 2014-12-01 17:38:16.572 24       1   1.000 secs   0    NA
3: Engine speed signal 2014-12-01 17:38:18.572 21       1   2.000 secs   0    NA
4: Engine speed signal 2014-12-01 17:38:19.572 19       1   1.000 secs   0    NA
5: Engine speed signal 2014-12-01 17:38:20.572 17       1   1.000 secs   0    NA
6: Engine speed signal 2014-12-01 17:38:22.572  0       0   2.000 secs   0    -3
7: Engine speed signal 2014-12-01 17:41:44.517  0       0 201.945 secs   0    NA
```

Figure 7. Row 6 defined as the end point of monitoring period by marking it with value -3 in column “start” (Case 1 for end point identification).

3700 seconds was set as the boundary between time stamps of consecutive samples. If the time difference between time stamps of consecutive samples is greater than the 3700 seconds, period between those time stamps is deemed as period when there is no data. Periods of no data were not included in monitoring periods.

3. Monitoring period - identification of start point (Case 2)

   This identification case identified start points of monitoring periods, when value of monitored signal is already above or below predefined limit but the time difference to last sample is over 3700 seconds. The moment is considered as start point when following conditions are fulfilled:

A) Value of monitored signal is already above/below predefined limit (current and previous samples have value above/below predefined limit).

B) Time difference between moment considered as a start point, and the time stamp of the previous sample is above 3700 seconds.

C) Time difference between the moment considered as a start point and the time stamp of the next sample is below or equal to 3700 seconds.

D) Previous sample is not marked as the start point of monitoring period.

4. Monitoring period - identification of end point (Case 2)

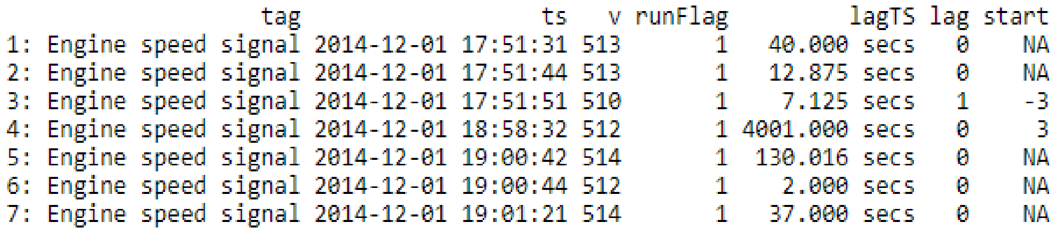
   This identification case identified end points of monitoring periods, when value of monitored signal is already above or below predefined limit but the time difference to next sample is over 3700 seconds. The moment is considered as start point when following conditions are fulfilled:

   A) Value of monitored signal is already above/below predefined limit (current and previous samples have value above/below predefined limit).

   B) Time difference between the moment considered as an end point, and the time stamp of the next sample is above 3700 seconds.

   C) Previous sample is not marked as the end point of monitoring period.

Following figure shows an application of this case. Precisely, row 3 is identified as the end point of monitoring period. It is worth highlighting that row 4 is identified as the start point of the following period since it fulfils the conditions for identification of start point (Case 2).

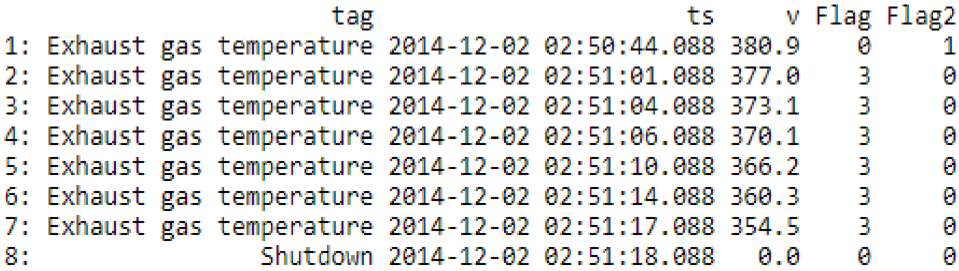


```
                    tag                  ts   v runFlag          lagTS lag start
1: Engine speed signal 2014-12-01 17:51:31 513       1   40.000 secs   0    NA
2: Engine speed signal 2014-12-01 17:51:44 513       1   12.875 secs   0    NA
3: Engine speed signal 2014-12-01 17:51:51 510       1    7.125 secs   1    -3
4: Engine speed signal 2014-12-01 18:58:32 512       1 4001.000 secs   0     3
5: Engine speed signal 2014-12-01 19:00:42 514       1  130.016 secs   0    NA
6: Engine speed signal 2014-12-01 19:00:44 512       1    2.000 secs   0    NA
7: Engine speed signal 2014-12-01 19:01:21 514       1   37.000 secs   0    NA
```

Figure 8. Start and end points (Case 2).

Second approach was used when the end of period of interest and length of time interval were known already (i.e. shutdown and 30 second time interval before shutdown). In this approach, all the samples within the defined time interval were selected to the period of interest. However, due to the dead banding, also the last sample, prior selected samples, must be considered. Following figure shows an example of this case.

If only samples within time interval (from 30 seconds before shutdown to the moment when shutdown occurs) are selected, rows 2-7 are only considered. Also row 1 must be considered but starting only from moment 02:50:48.088.



```
                         tag                      ts     v Flag Flag2
1: Exhaust gas temperature 2014-12-02 02:50:44.088 380.9    0     1
2: Exhaust gas temperature 2014-12-02 02:51:01.088 377.0    3     0
3: Exhaust gas temperature 2014-12-02 02:51:04.088 373.1    3     0
4: Exhaust gas temperature 2014-12-02 02:51:06.088 370.1    3     0
5: Exhaust gas temperature 2014-12-02 02:51:10.088 366.2    3     0
6: Exhaust gas temperature 2014-12-02 02:51:14.088 360.3    3     0
7: Exhaust gas temperature 2014-12-02 02:51:17.088 354.5    3     0
8:                Shutdown 2014-12-02 02:51:18.088   0.0    0     0
```

Figure 9. Example of second approach to identify monitoring period.

4.1.2 Shutdown definition

Once rules were defined to properly detect monitoring periods, an approach was developed to separate unplanned shutdowns from normal stops. In the case of automatic shutdown (SHD), engine control system detects deviation in engine behaviour and causes the SHD to occur. For instance, deviation in exhaust gas temperature could cause engine SHD if it exceeds acceptable thresholds.

In experimental approach to separate SHDs from normal stops, engine load signal, which measures the engine load in kilowatts, was selected as an indicator. In picture below, behaviours of engine load (kW) and engine speed (rpm) signals are plotted as a function of time when engine mission ends with a planned stop.

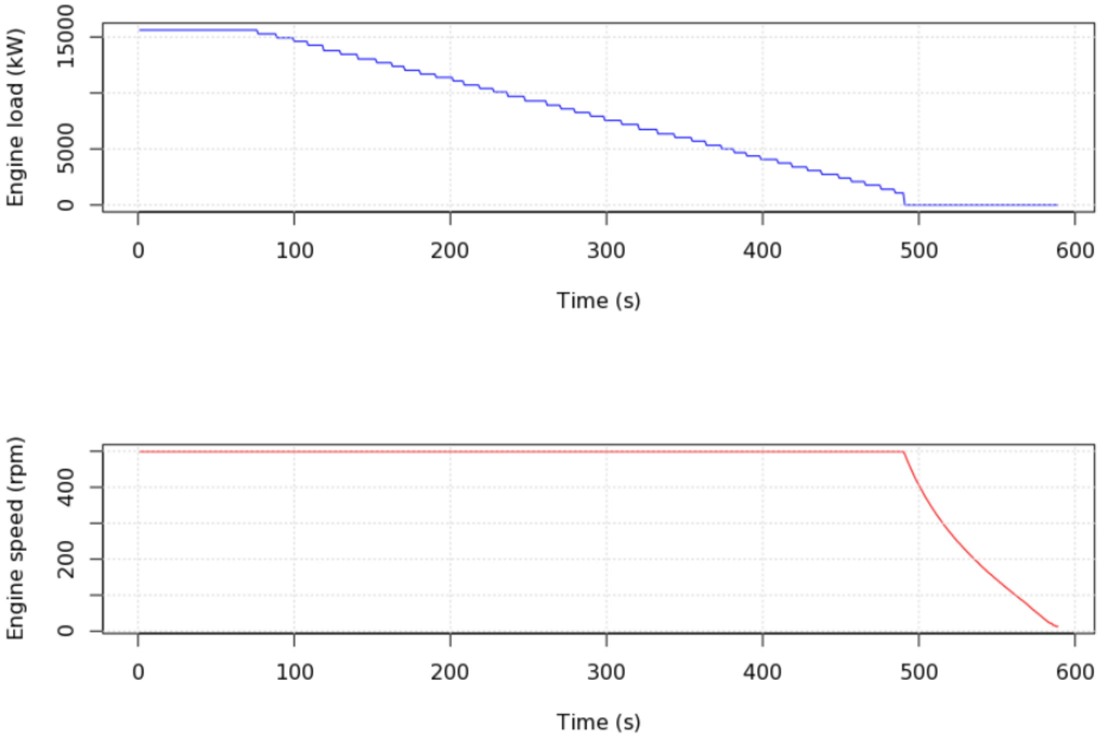


Figure 10. Engine load (kW) and engine speed (rpm), when engine mission is considered to finish with planned stop.

Picture above consist of two plots: first plot has engine load signal (kW) as a function of time (s) and second plot has engine speed (rpm) as a function of time (s). From the first plot it can be observed, that the engine load starts decreasing from nominal load approximately 500 seconds before engine stops, reaching 0 value approximately in 400 seconds. The second plot shows that approximately same time when engine load reaches 0 kW, engine speed starts decreasing from nominal speed, reaching eventually 0 value as well.

The main observation from engine load signal behavior is that the time period to reach 0 kW from nominal load takes some minutes in the case of normal stop. In the next picture, behaviours of engine load (kW) and engine speed (rpm) signals are plotted as a function of time again, but in this case, the engine mission is considered to conclude in SHD.

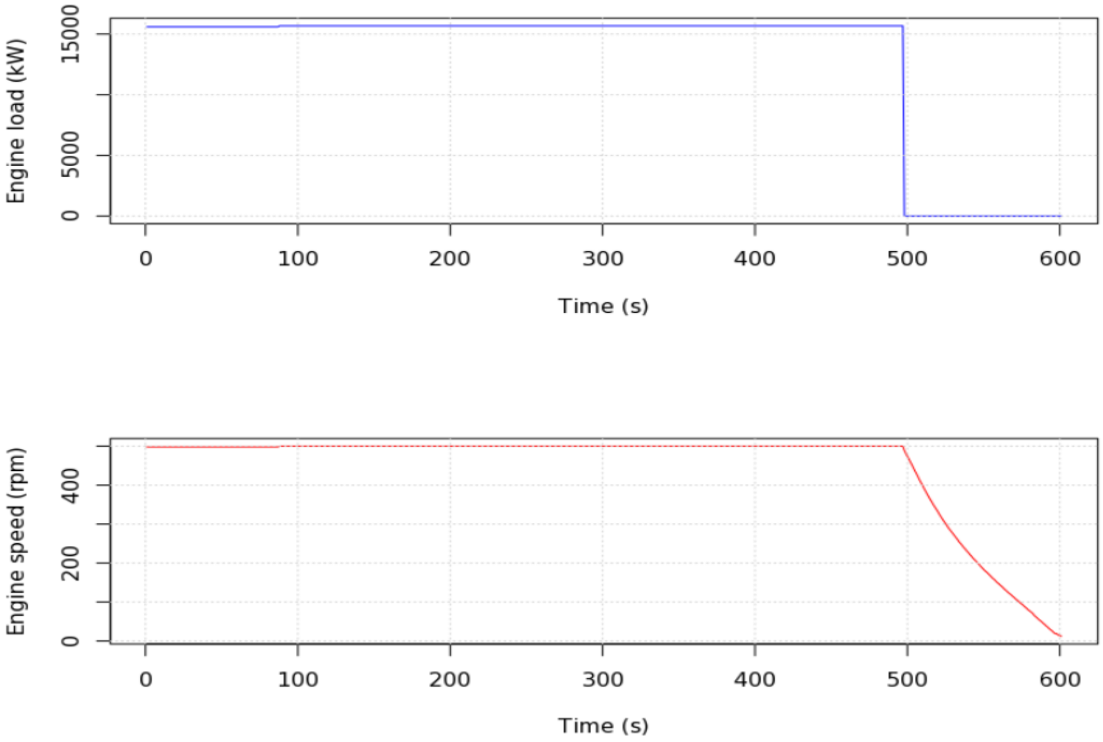


Figure 11. Engine load (kW) and engine speed (rpm) when, engine mission is considered to finish with SHD.

From the load behavior it can observed, that engine load drops instantly from nominal load to 0 kW and approximately at the same time, when the engine load reaches 0 kW, the engine speed starts decreasing from nominal speed to 0.

When comparing the engine speed behavior in both cases (SHD and normal stop), there are no significant differences, therefore engine speed signal is not suitable indicator to deem if engine mission concludes in SHD or in normal stop. However, engine load behavior is significantly different in SHDs than it is in normal stops, hence engine load behavior is the main feature which has to be considered when deciding if engine mission concludes in SHD.

Before defining the conditions, which must be fulfilled in order to classify engine stop as SHD, it has to be considered, that there are differences in maximum operating loads between different engine platforms. Also, the duration for engine load to reach 0 kW from nominal load in the case of normal stop could vary depending on the engine platform and application. SHD is considered to occur when following conditions are fulfilled:

1. The engine load reaches value below minimum acceptable load and remains below that threshold at least 30 seconds. This condition is in place in order to neglect bias in signals and identify only real SHD cases. Also, this has to be the final occasion for engine load to behave this way during the engine mission.

   The engine load is monitored 1-minute-period prior the moment when it reaches value below minimum acceptable load. During that monitoring period engine load has to fulfil conditions 2 and 3.

2. Maximum value from the last 10-seconds must be at least 20% from maximum operating load of the engine.

3. Maximum value from the last 10 seconds must be at least 80 percent from the mean value from first 10 seconds.

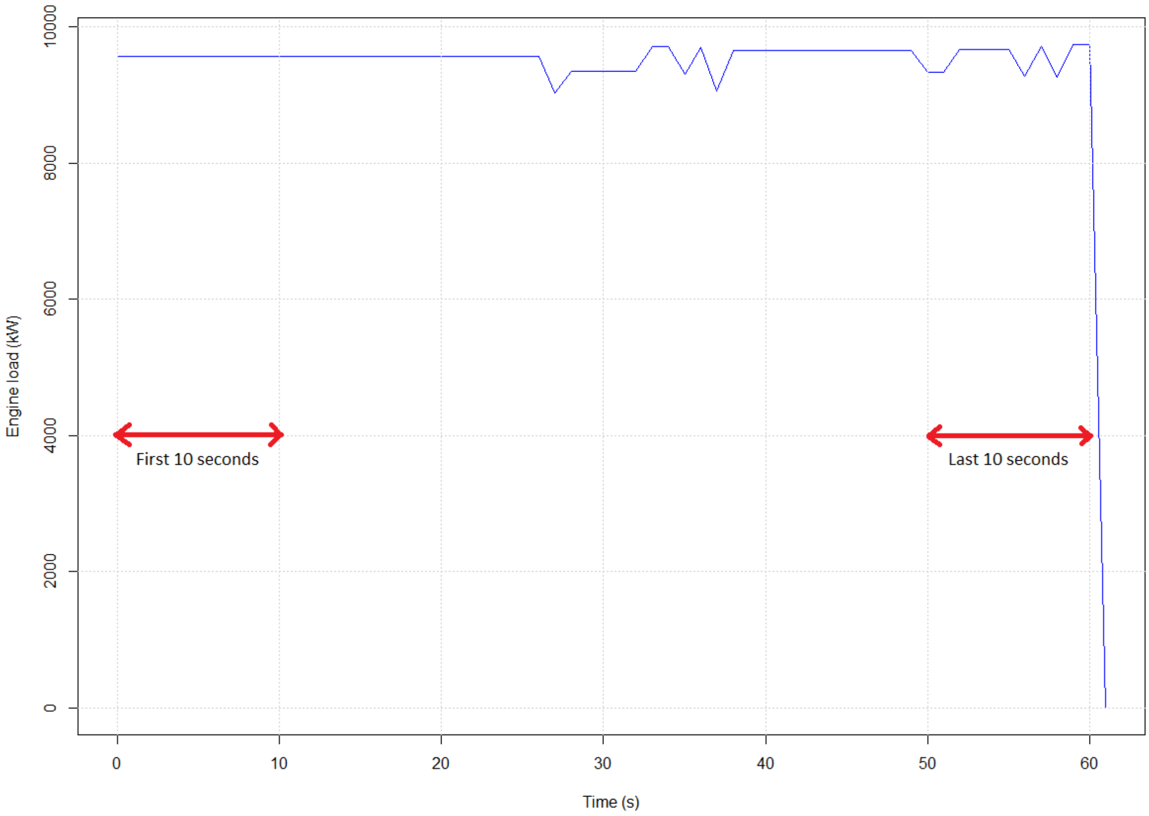


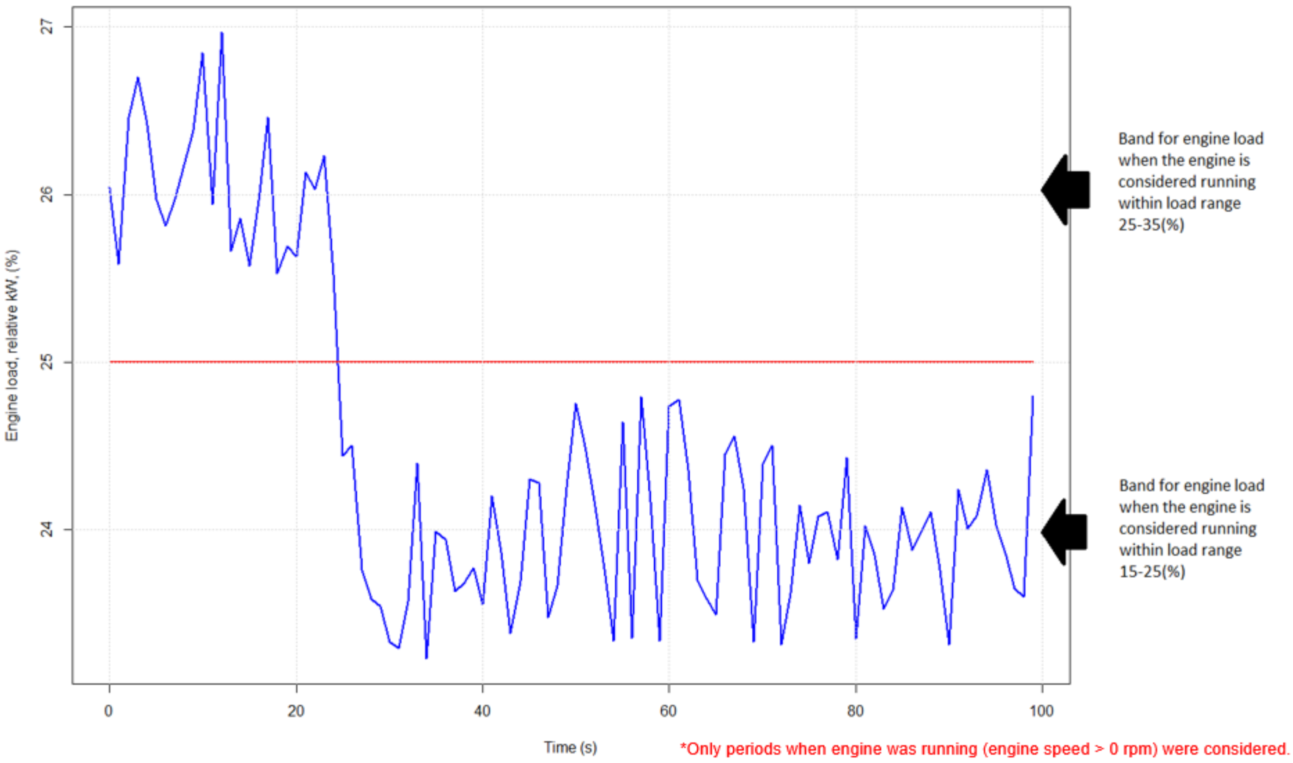
Figure 12. 1-minute monitored period prior the moment when engine load reaches value below minimum acceptable load.

Minimum acceptable load (100 kW) was selected as a threshold instead of 0 kW due to quality of EDS data. For instance, for some installations, engine load signal could have some false values when it is actually 0 kW. These values could be for example small integer numbers or even negative numbers. In this approach, it was presumed, that minimum nominal load for engines is 20 percent from the maximum load of the engine. The objective of the third condition is to ensure that load drops abruptly from nominal load in the end of 1-minute monitoring period. Mean value was selected from first 10 seconds, since the objective was to minimize the effect of possible noise.

This approach also presumes, that engine is running approximately with constant load, hence SHD analysis was only applied to power plant applications, excluding marine applications. This is due to the reason that it is presumed that operating load of marine applications is varying a lot compared to power plant applications.

4.2 Load distribution analysis

In order to characterize engines operation, a data-driven approach oriented to calculate distribution of cumulative running hours of an engine per different load percentage intervals was developed. This solution takes all engine specific raw data files of a certain installation as an input and provides results in .csv format as output. Output contains results for all engines of that specific installation. This solution uses 2 different signals: engine speed and engine load (in percentages). Figure below describes the logic used in this approach.

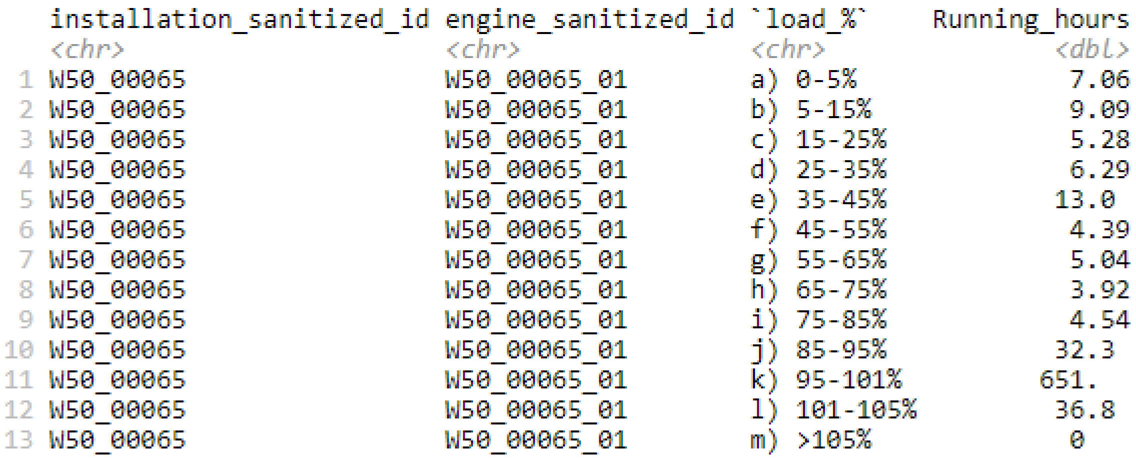


*Only periods when engine was running (engine speed > 0 rpm) were considered.

Figure 13. Load – Analysis: Approach.

The solution is designed to process one installation at a time, which means that results for each engine of a given installation can be provided with a single execution of the algorithm. This solution was developed in order to investigate how different engines are operating with different load percentages and it was applied to DF and SG engines of EDS operating in both power plant and marine applications. Selected load percentage intervals were constant and same for every analysed engine. Load percentage intervals are presented in figure number 14.

By utilizing the speed signal, the solution provides information concerning all the engine missions by defining start and end moment of each engine mission by utilizing rules defined in subchapter 4.1.1. When the start and end moments of each engine mission are defined, that information is used to extract only those engine load percentage signal samples, which are measured during the engine missions. Finally, cumulative running hours for all load percentage intervals are calculated, and the information is stored in installation specific csv-file. In the figure below, example concerning results is presented. First column has installation ID, second column engine ID, third column the load percentage interval and final column cumulative running hours.



```
   installation_sanitized_id engine_sanitized_id `load_%`    Running_hours
   <chr>                     <chr>               <chr>               <dbl>
 1 W50_00065                 W50_00065_01        a) 0-5%              7.06
 2 W50_00065                 W50_00065_01        b) 5-15%             9.09
 3 W50_00065                 W50_00065_01        c) 15-25%            5.28
 4 W50_00065                 W50_00065_01        d) 25-35%            6.29
 5 W50_00065                 W50_00065_01        e) 35-45%           13.0
 6 W50_00065                 W50_00065_01        f) 45-55%            4.39
 7 W50_00065                 W50_00065_01        g) 55-65%            5.04
 8 W50_00065                 W50_00065_01        h) 65-75%            3.92
 9 W50_00065                 W50_00065_01        i) 75-85%            4.54
10 W50_00065                 W50_00065_01        j) 85-95%           32.3
11 W50_00065                 W50_00065_01        k) 95-101%         651.
12 W50_00065                 W50_00065_01        l) 101-105%         36.8
13 W50_00065                 W50_00065_01        m) >105%             0
```

Figure 14. Load Analysis – Example of results.

The solution also works in such a way, that it stops processing the engine specific raw data file if the data file is missing either engine speed or load percentage signal or either of these signals have only 0-values. The solution also stores information to txt-file during the execution of the algorithm. This information includes mean, minimum and maximum values of the load percentage signal to indicate if the provided results are reasonable. Txt-file also includes information for the amounts of engine starts and stops, to express, if the logic defined in subchapter 4.1.1 works without flaws, which means that amounts of engine starts, and stops must be equal.

4.3 Automatic shutdown analysis

Once detection rules for isolation of engine operating time and characterization according to engine load performance were defined, methodology was developed to investigate automatic SHDs due to deviations in selected signals.

This solution detects if engine mission concludes in SHD by following the rules defined in subchapter 4.1.2. After all the SHDs are detected from the raw EDS data, the solution investigates if there have been any deviations in following signals, 30 seconds prior when SHD occurs:

1. Exhaust gas temperature signals.
2. Liner temperature signals.
3. Big end bearing temperature signals.
4. High temperature water pressure and temperature signals.
5. Lube oil pressure and temperature signals.

For all the different engine platforms and engine designs, there are engine specific deviation thresholds and time windows, which indicate how long the signal must exceed or remain below the threshold before the engine control system causes the automatic SHD to occur. These deviation thresholds and time windows are defined in Wärtsilä’s internal document and this solution was built to follow those definitions and rules. Figure below describes how automatic SHD is caused by the deviation in monitored signal.

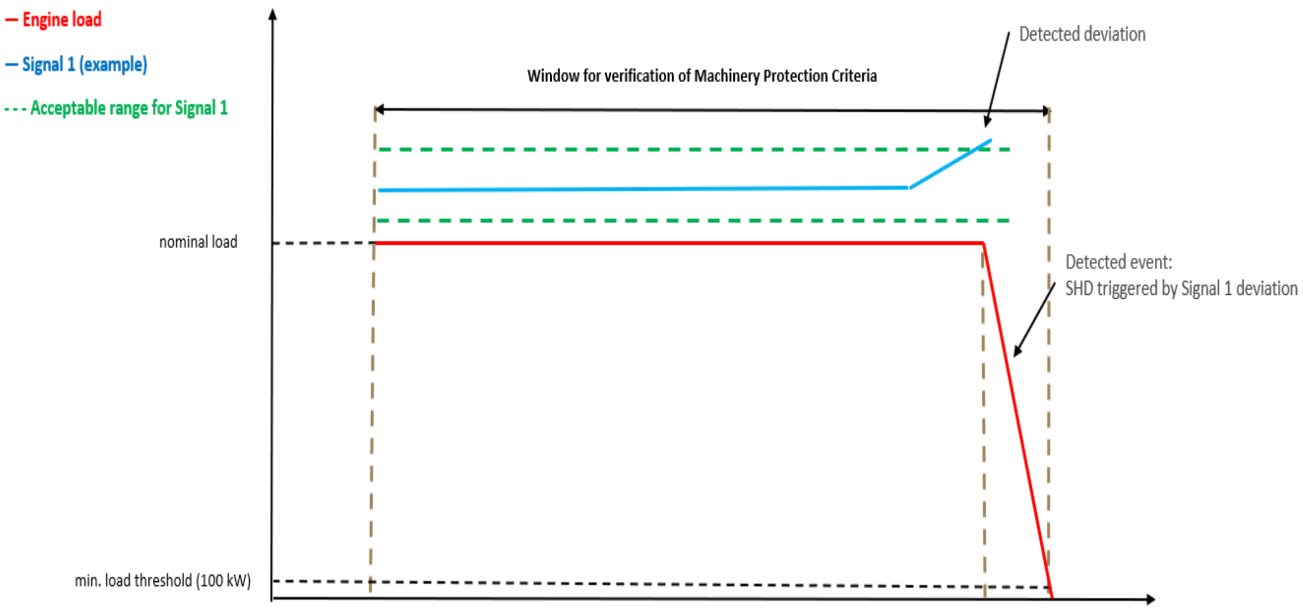

— Engine load
— Signal 1 (example)
- - - Acceptable range for Signal 1
Window for verification of Machinery Protection Criteria
Detected deviation
nominal load
Detected event:
SHD triggered by Signal 1 deviation
min. load threshold (100 kW)

Figure 15. Automatic shutdown caused by deviation in monitored signal.

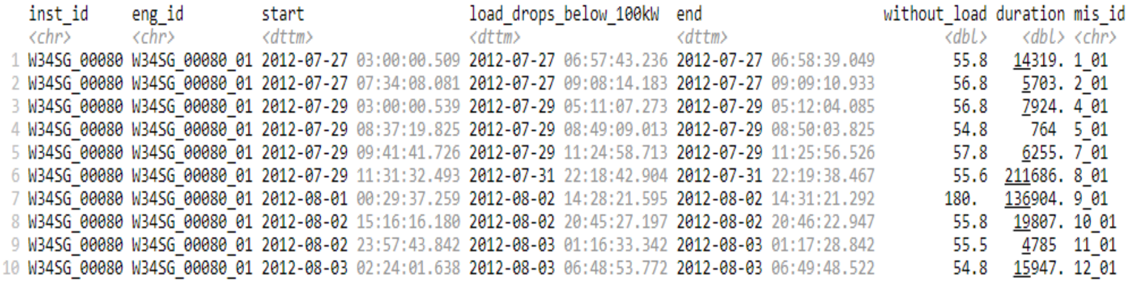
This solution takes all engine specific raw data files of selected installation, and information considering engine platform and engine design of those installation engines, as inputs and provides 4 installation specific files and 2 engine specific files as outputs:

1. Installation specific file #1: it is a csv-file containing information for all engine missions of every installation engine.

2. Installation specific file #2: it is a subset from the first output file, containing only information from the engine missions which conclude in SHD.

   Following figure shows structure of the first two installation specific output files. Each row represents a unique engine mission. The columns are representing following information: first column is the installation ID, second column is the engine ID, third column is the start moment of the mission, fourth column is the moment when the engine load drops under 100 kW last time in that specific mission, fifth column is the end moment of the mission, sixth column is the period how long engine speed remains greater than zero after the load has dropped below

100 kW, seventh column is mission duration in seconds and final column is mission ID. These results are derived from the raw data of a W34SG engine.



```
   inst_id     eng_id         start                   load_drops_below_100kW  end                     without_load duration mis_id
   <chr>       <chr>          <dttm>                  <dttm>                  <dttm>                         <dbl>    <dbl> <chr> 
 1 W34SG_00080 W34SG_00080_01 2012-07-27 03:00:00.509 2012-07-27 06:57:43.236 2012-07-27 06:58:39.049          55.8   14319. 1_01  
 2 W34SG_00080 W34SG_00080_01 2012-07-27 07:34:08.081 2012-07-27 09:08:14.183 2012-07-27 09:09:10.933          56.8    5703. 2_01  
 3 W34SG_00080 W34SG_00080_01 2012-07-29 03:00:00.539 2012-07-29 05:11:07.273 2012-07-29 05:12:04.085          56.8    7924. 4_01  
 4 W34SG_00080 W34SG_00080_01 2012-07-29 08:37:19.825 2012-07-29 08:49:09.013 2012-07-29 08:50:03.825          54.8     764  5_01  
 5 W34SG_00080 W34SG_00080_01 2012-07-29 09:41:41.726 2012-07-29 11:24:58.713 2012-07-29 11:25:56.526          57.8    6255. 7_01  
 6 W34SG_00080 W34SG_00080_01 2012-07-29 11:31:32.493 2012-07-31 22:18:42.904 2012-07-31 22:19:38.467          55.6  211686. 8_01  
 7 W34SG_00080 W34SG_00080_01 2012-08-01 00:29:37.259 2012-08-02 14:28:21.595 2012-08-02 14:31:21.292         180.   136904. 9_01  
 8 W34SG_00080 W34SG_00080_01 2012-08-02 15:16:16.180 2012-08-02 20:45:27.197 2012-08-02 20:46:22.947          55.8   19807. 10_01 
 9 W34SG_00080 W34SG_00080_01 2012-08-02 23:57:43.842 2012-08-03 01:16:33.342 2012-08-03 01:17:28.842          55.5    4785  11_01 
10 W34SG_00080 W34SG_00080_01 2012-08-03 02:24:01.638 2012-08-03 06:48:53.772 2012-08-03 06:49:48.522          54.8   15947. 12_01 
```

Figure 16. Structure of the first two installation specific output files.

3. Installation specific output file #3: it is a csv-file, which contains information concerning deviations of the signals monitored for the identification of automatic SHD. File is stored as a table. Every row represents one deviation of monitored parameters. Columns indicate: the code of the signal in which the deviation has occurred, deviation type (high (in case the signal has exceeded the higher deviation threshold) or low (in case the signal has fell below the lower deviation threshold)), moment of occurred SHD, deviation start moment, deviation end moment and deviation duration prior SHD. In addition, there are columns for installation ID, engine ID and mission ID.

4. Installation specific output file #4: it is a txt-file, which contains information concerning the execution of developed algorithm. The main purpose of txt-file is to store information about input information, deviation thresholds and time windows algorithm has used for each investigated signal, during the execution.

5. Engine specific output file #1: it contains load signal information. This file grants the opportunity to investigate the engine load behavior prior each engine stop.

6. Engine specific output file #2: it contains load signal information for those missions which are labeled as starting failures. Engine missions which last less than 15 minutes (900 seconds) are considered as starting failures. As mentioned in subchapter 4.1.2, the logic, which is defined to classify if engine mission concludes

in SHD or in planned stop, presumes that engine is running with constant load. In the cases of starting failures, engine load can be still increasing towards the nominal load, when the SHD occurs and it is possible that engine load has not reached the nominal load yet. In other words, logic defined in this thesis is not able to detect these SHDs and these output files were stored for the future classification and investigation of these cases.

4.4 Main feature extraction for relevant signals

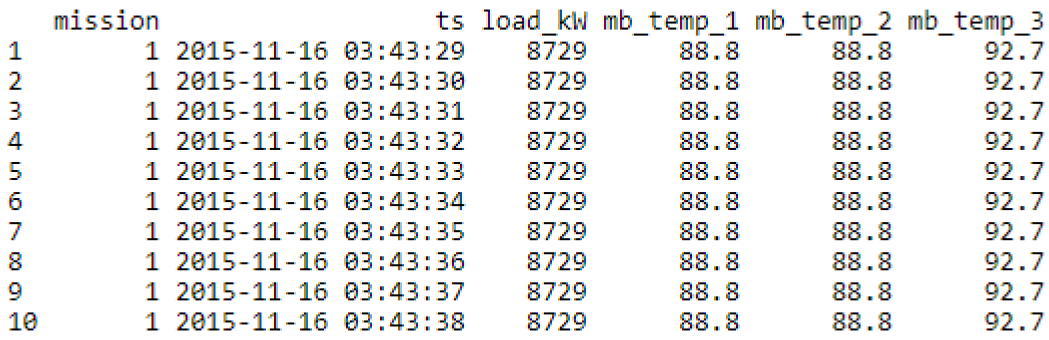
In order to retrieve information as much as possible from EDS contents, algorithm was developed for extracting features for relevant sensor signals of the engine in order to characterize their behaviour to find correlations or trends in data. This activity lays the basis for future machine learning applications, for instance by utilizing these extracted features in the development of a predictive maintenance algorithm. Main bearing temperature signals were selected as the test case for this solution.

This solution takes two inputs. The first input comprehends all the engine specific raw data files of a selected installation and second input is the installation specific file which contains information for all engine missions of every installation engine (i.e. the output file produced in temperature and pressure deviation analysis). The solution utilizes the inputs and produces output file for each investigated signal per every installation engine.

The solution extracts set of statistical features for each investigated signal per each selected engine mission (engine mission duration > 15 minutes), and more specifically, only from the time period which do not include the transient phases. Phase when engine load is increasing from 0 kW towards the nominal load in the beginning of the engine mission and phase when engine load is decreasing from nominal load towards 0 kW in the end of the engine mission, are filtered out from the feature calculation.

After the transient phases are removed and monitored signals are given own columns with fixed sampling frequency (1 Hz), data is in following format showed in the next figure.

First column represents the mission number, second column time stamp, third column load value and columns from fourth to the sixth represent the main bearing temperatures.

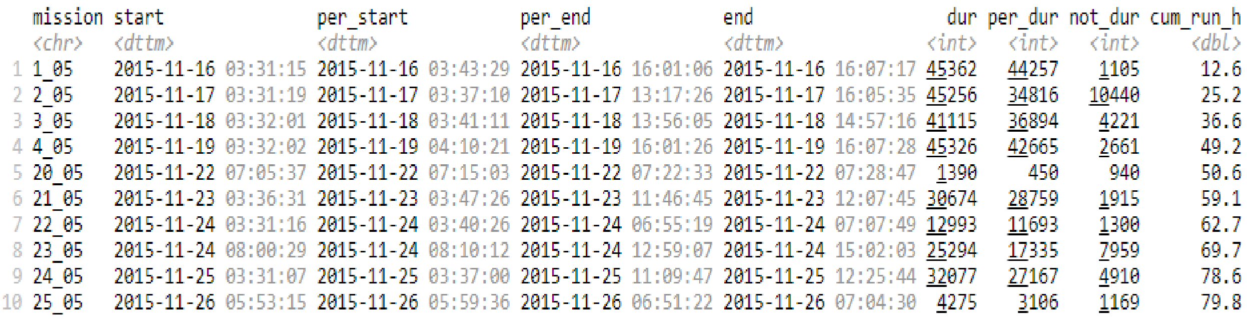


```
   mission                  ts load_kW mb_temp_1 mb_temp_2 mb_temp_3
1        1 2015-11-16 03:43:29    8729      88.8      88.8      92.7
2        1 2015-11-16 03:43:30    8729      88.8      88.8      92.7
3        1 2015-11-16 03:43:31    8729      88.8      88.8      92.7
4        1 2015-11-16 03:43:32    8729      88.8      88.8      92.7
5        1 2015-11-16 03:43:33    8729      88.8      88.8      92.7
6        1 2015-11-16 03:43:34    8729      88.8      88.8      92.7
7        1 2015-11-16 03:43:35    8729      88.8      88.8      92.7
8        1 2015-11-16 03:43:36    8729      88.8      88.8      92.7
9        1 2015-11-16 03:43:37    8729      88.8      88.8      92.7
10       1 2015-11-16 03:43:38    8729      88.8      88.8      92.7
```

Figure 17. Example of prepared data for feature extraction.

Sampling frequencies are fixed for the monitored signals in this solution in order to provide more accurate results when deriving statistical features for the signals. Deriving features from the signals which do not have fixed sampling frequencies could lead to inaccurate results. When fixing the sampling frequencies, missing values were replaced with the previous value of the signal.

After the data is prepared, statistical features for all the investigated signals are derived from each engine mission. Figure below represents the portion about the output file of single signal. Every row represents one engine mission. First column is mission ID, second column is start moment of the mission, third column is moment when the load has reached nominal load first time, fourth column is moment when the load is last time at nominal load, fifth column is the end moment of the mission, sixth column is mission duration in seconds, seventh column is mission duration excluding transient phases, eight column is total time of transient phases and ninth column is cumulative running hours of the engine. In addition, there are columns for the following statistical features of the signal: minimum, maximum, mean, variance, median, standard deviation, standard error, skewness, kurtosis, peak to peak and root mean square values for each engine mission.



```
   mission start               per_start           per_end             end                   dur per_dur not_dur cum_run_h
   <chr>   <dttm>              <dttm>              <dttm>              <dttm>              <int>   <int>   <int>     <dbl>
 1 1_05    2015-11-16 03:31:15 2015-11-16 03:43:29 2015-11-16 16:01:06 2015-11-16 16:07:17 45362   44257    1105      12.6
 2 2_05    2015-11-17 03:31:19 2015-11-17 03:37:10 2015-11-17 13:17:26 2015-11-17 16:05:35 45256   34816   10440      25.2
 3 3_05    2015-11-18 03:32:01 2015-11-18 03:41:11 2015-11-18 13:56:05 2015-11-18 14:57:16 41115   36894    4221      36.6
 4 4_05    2015-11-19 03:32:02 2015-11-19 04:10:21 2015-11-19 16:01:26 2015-11-19 16:07:28 45326   42665    2661      49.2
 5 20_05   2015-11-22 07:05:37 2015-11-22 07:15:03 2015-11-22 07:22:33 2015-11-22 07:28:47  1390     450     940      50.6
 6 21_05   2015-11-23 03:36:31 2015-11-23 03:47:26 2015-11-23 11:46:45 2015-11-23 12:07:45 30674   28759    1915      59.1
 7 22_05   2015-11-24 03:31:16 2015-11-24 03:40:26 2015-11-24 06:55:19 2015-11-24 07:07:49 12993   11693    1300      62.7
 8 23_05   2015-11-24 08:00:29 2015-11-24 08:10:12 2015-11-24 12:59:07 2015-11-24 15:02:03 25294   17335    7959      69.7
 9 24_05   2015-11-25 03:31:07 2015-11-25 03:37:00 2015-11-25 11:09:47 2015-11-25 12:25:44 32077   27167    4910      78.6
10 25_05   2015-11-26 05:53:15 2015-11-26 05:59:36 2015-11-26 06:51:22 2015-11-26 07:04:30  4275    3106    1169      79.8
```

Figure 18. Example of feature extraction results.

4.5 Anomaly detection of sensor signals

Finally, methodology was developed for anomaly detection of sensor signal behaviour. This solution utilizes linear regression and interquartile range method and it was applied to one large bore engine operating in power plant application. Investigated signals in this solution were main bearing temperature signals and exhaust gas temperature signals. This solution utilizes the information produced by different data-driven approaches which are presented earlier in this thesis in addition to raw data.

First, feature extraction presented in subchapter 4.4 is utilized to derive features for engine load percentage signal. After the features are extracted for this signal, median values of this signal from each engine mission, are investigated.

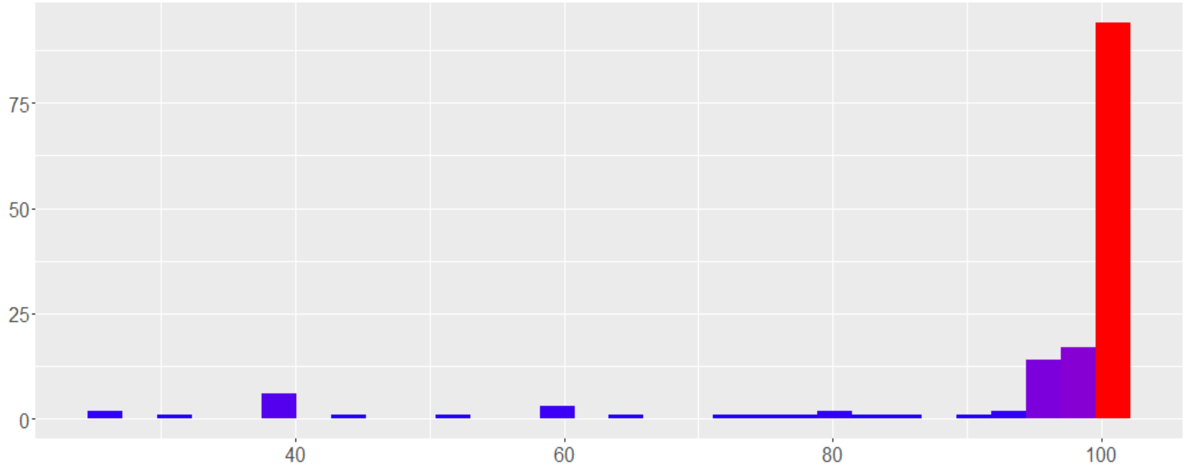


Figure 19. Engine load distribution.

From the histogram above, it can be deemed, that the majority of engine missions have median value of load percentage signal 100 or higher. Based on this observation, investigated load percentage interval is set to 100-105 percentage.

After the load percentage interval is set, main bearing temperature, exhaust gas temperature and load percentage signals are extracted from the raw data file of this engine. When extraction of the signals is done, sampling frequencies of the monitored signals are fixed to same value (1 Hz in this case). After the sampling frequencies are fixed, only samples, that are within selected load percentage interval, are selected. In this case, only samples that have been measured when the engine load percentage has been in closed interval from 100 to 105, are considered.

Proposed solution detects anomalies from the behaviours of selected signals. This is done by selecting one signal among selected ones at a time and treating it as response variable (output). Load percentage signal is treated as predictor variable (input) and simple linear regression is utilized to model linear relationship of these variables. The load percentage signal was used as predictor variable since it is assumed, that other engine parameters are linearly dependent on it. Below, is the equation to model the linear relationship of these variables.

$$y = \alpha x + \beta$$

Where y is the value of the response variable, x is the value of the predictor variable and α and β are parameters of the regression model. The observed values of predictor and response variables are utilized to estimate the regression parameters of the model with simple linear regression. Estimation of regression parameters leads to regression model, that minimizes the sum of squared differences between observed and predicted values. Therefore, the calculated regression model is the model that has the best fit to the given data points of observed values.

After the regression model parameters are calculated, observed values of predictor variable (load percentage signal), and regression model parameters are used to predict values for selected signal. The differences between predicted and observed values of selected signal are then calculated and then they are standardized by calculating z-scores for the population with following formula:

$$z = \frac{x - \mu}{s}$$

Where z is the calculated z-score, x is the sample from the population, μ is mean of the population and s is the standard deviation of the population.

After these z-scores are calculated for each sample, closed interval is defined by utilizing interquartile range value of the z-score population. This closed interval is:

$$[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$$

Where Q_1 is the first quartile of z-score population, Q_3 is the third quartile of the z-score population and *IQR* is the interquartile range value of z-score population (IQR = Q_3-Q_1). Final phase is to classify samples as inliers (so called normal data points) or as outliers (anomalies) with this closed interval: if the z-score value of the sample is within defined closed interval, sample is classified as inlier, otherwise sample is classified as outlier.

This solution provides following information for all monitored signals treated as response variable: plot, where monitored signal is plotted as function of engine load percentage. In the plot, observations of monitored signal are classified as inliers and outliers and the regression line is included as well.

5 RESULTS

5.1 EDS mapping

First mandatory step to be taken before any data analytics approach, is mapping the available data in order to understand how data can be used and which kind knowledge could be gained from their investigation. This subchapter provides this preliminary investigation about EDS contents.

5.1.1 EDS installations: Locations

Figure below shows how EDS installations are distributed around the world. Green color for the country indicates that it has at least one EDS installation operating in it.

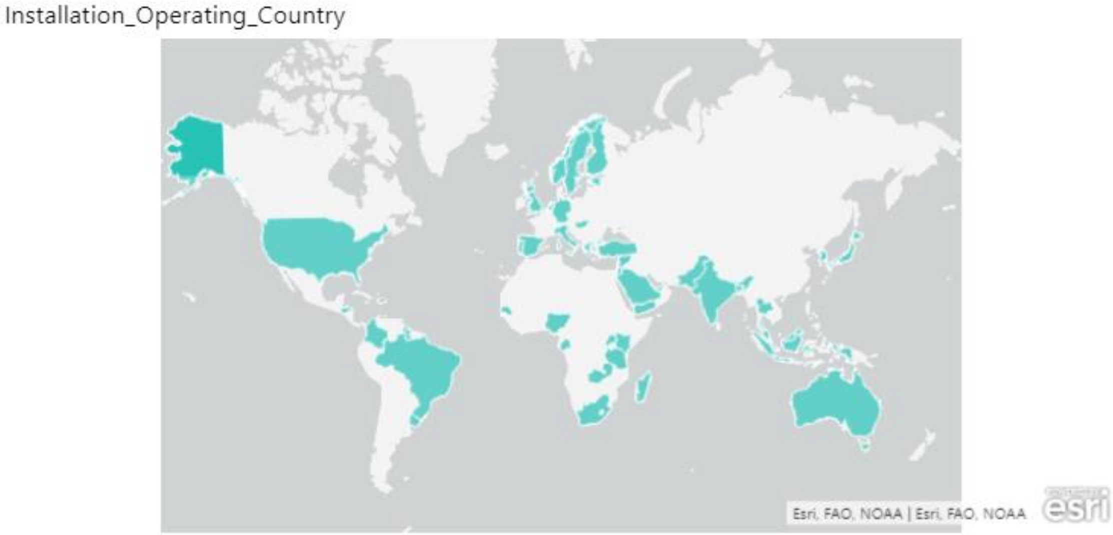

Installation_Operating_Country
Esri, FAO, NOAA | Esri, FAO, NOAA

Figure 20. Operating countries of EDS installations.

5.1.2 EDS installations: Applications and segments

Next figure consists of two different bar charts:

1. First bar chart on the left shows how operating and not operating installations have distributed for both power plant and marine applications.

2. Second bar chart on the right shows how different installation segments have distributed for both power plant and marine applications.

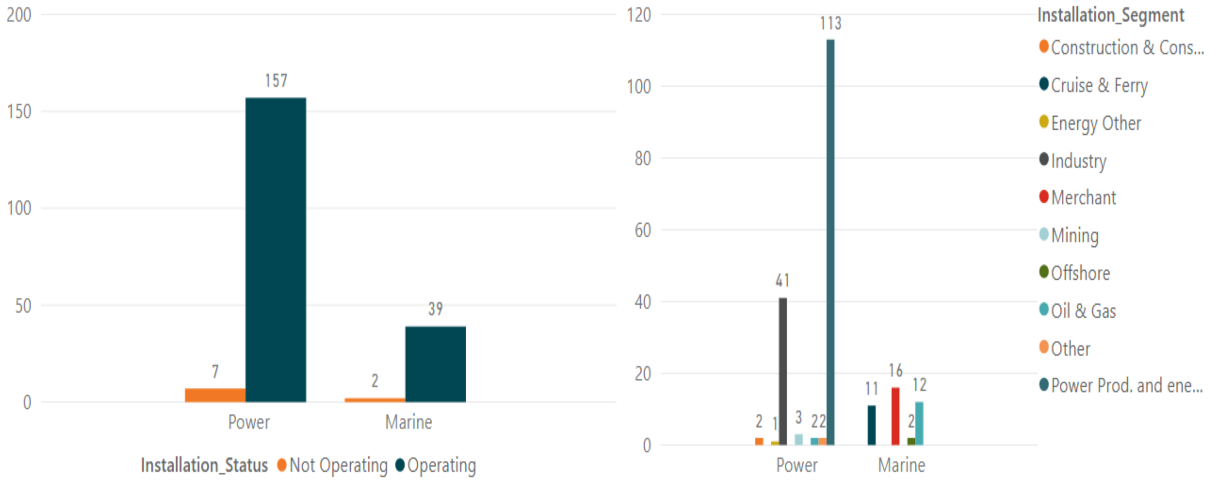


Figure 21. Distributions for operating and not operating installations and installation segments.

5.1.3 EDS engines: Comissioning years

Following figure presents distribution for commissioning years of the EDS engines. It is worth mentioning that commissioning year of the engine does not mean necessarily that EDS holds data for that engine starting from that specific year.

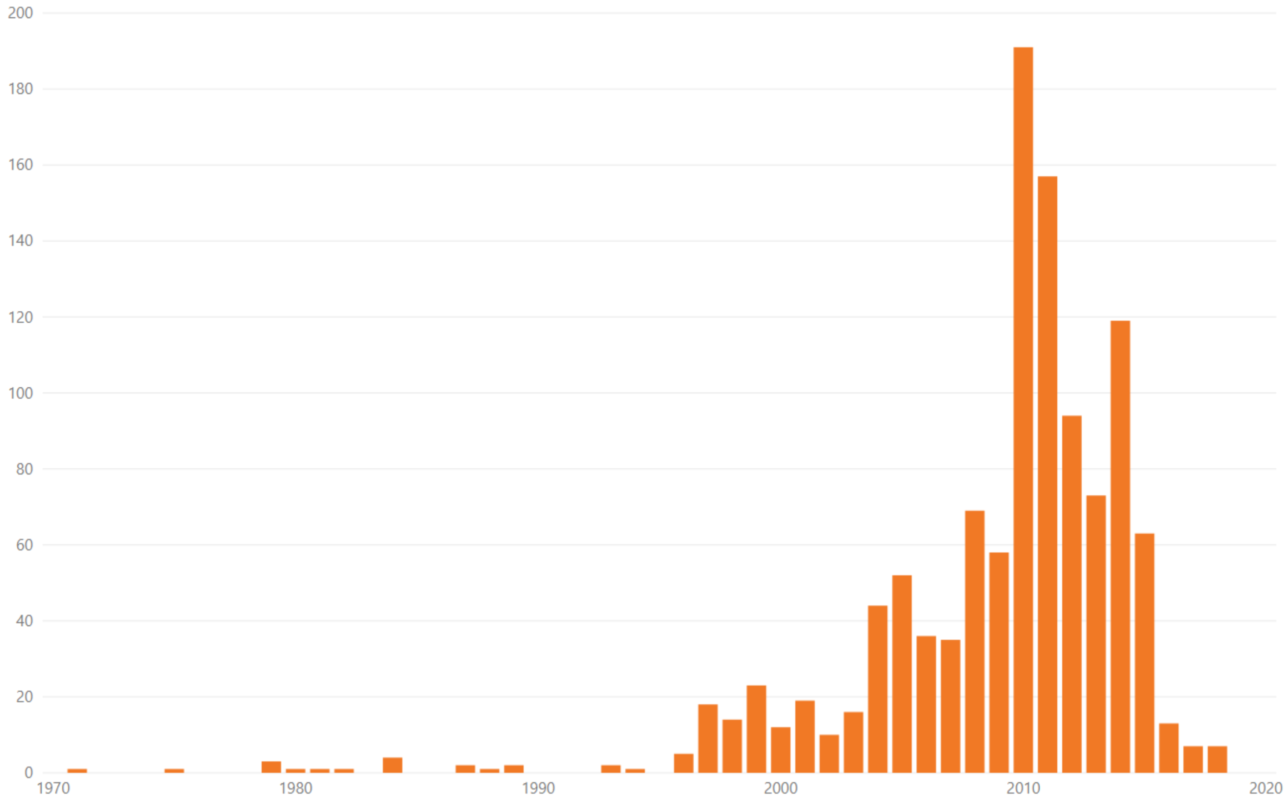


Figure 22. Distribution of commissioning years of the EDS engines.

5.1.4 EDS engines: Cylinder configuration

Next figure displays final example about how EDS content can be viewed and investigated. The figure presents how engines with specific cylinder number are distributed between different product reference types.

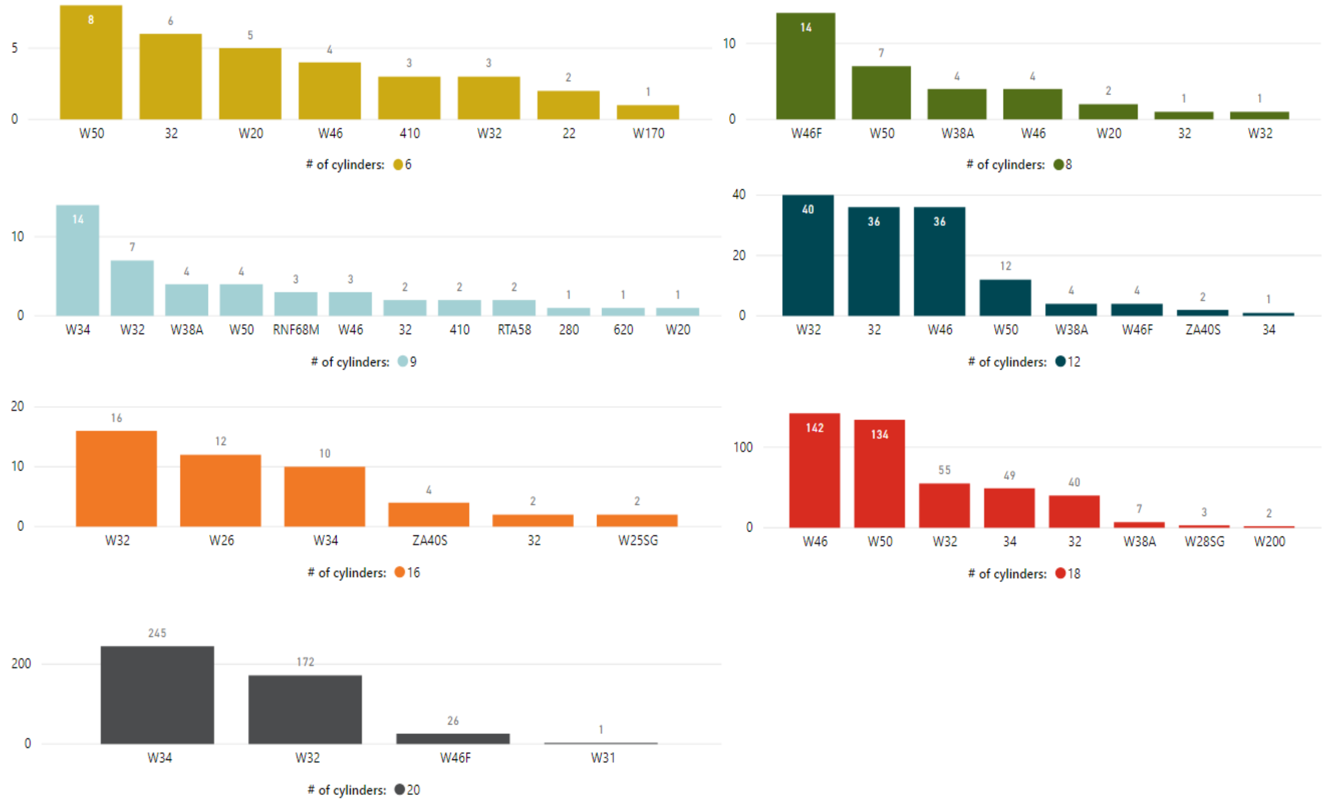


Figure 23. Distributions of product reference types.

5.2 Engine performance

Implemented data-driven approaches were applied for the installations which encoded signal formats were identified. Load distribution analysis was applied for total of 43 installations 277 engines and automatic shutdown analysis for 32 installations and 232 engines. It is worth mentioning, that load distribution analysis was implemented for both power plant and marine applications but automatic shutdown analysis only for the power plant applications. In this subchapter, example results related to performance of EDS engines are presented.

5.2.1 Engine load profiles

Load distribution analysis provides the possibility investigate and compare load profiles of different engines. Following figure shows consists of two different bar charts:

1. Upper bar chart presents distribution of cumulative running hours per different load intervals for power plant engines.

2. Lower bar chart presents distribution of cumulative running hours per different load intervals for marine engines.

From the figure it can be observed, that power plant engines are operating with load close to maximum since they are working usually at fixed speed and load request does not change consistently over time, unless in case of failure. The other observation is that marine engines have more variety in their load since they are working more with variable speed and wider load request.

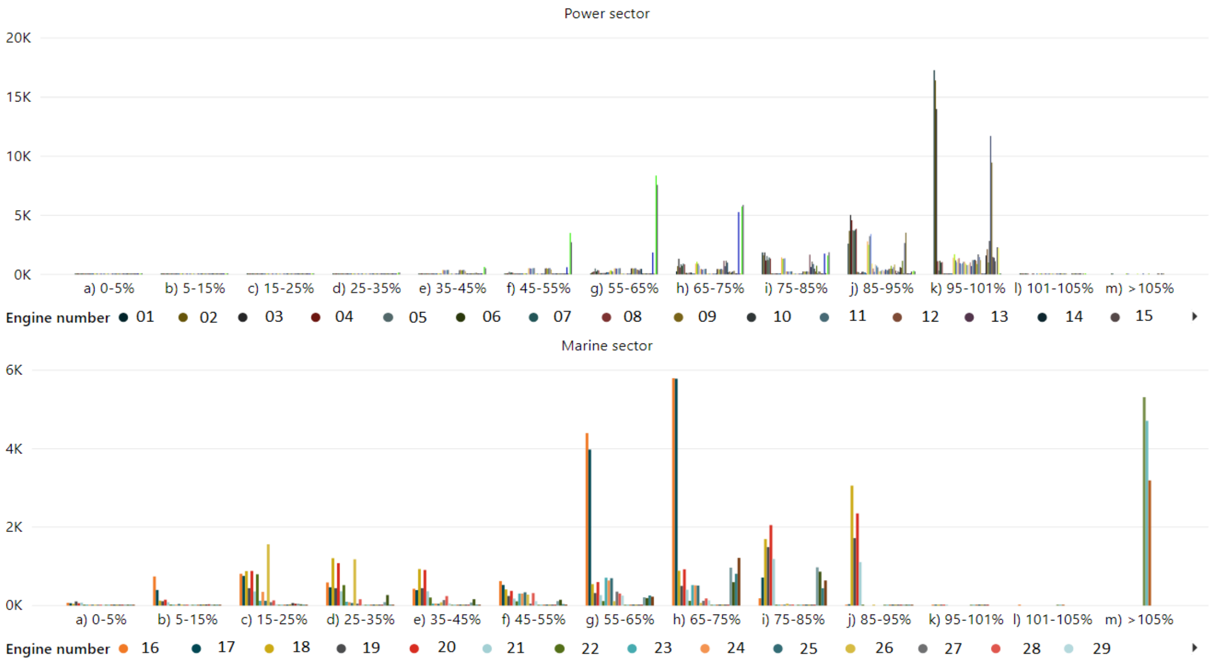


Figure 24. Distributions of cumulative running per different load intervals.

Extracted results allows a load profile comparison between different marine and power plant applications as well. In the next figure, for several applications equipped with large bore engines, distributions of average cumulative running hours per different load intervals are presented. LNG Carriers, Passenger & Cargo Vessels and Combination Tankers are marine applications, whereas Industry, Mining and Power Producer and energy provider are power plant applications.

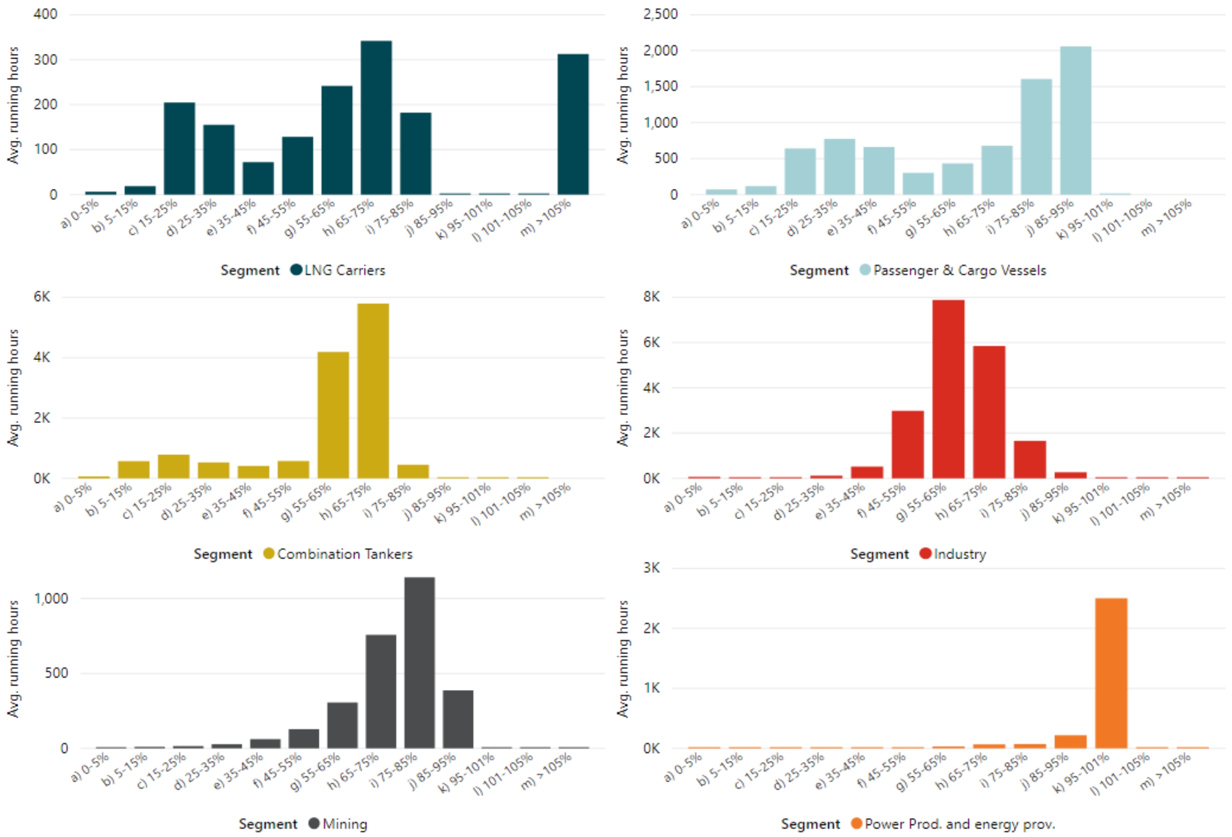


Figure 25. Average cumulative running hours per different load percentage intervals for application equipped with large bore engines.

It is worth mentioning that results extracted from EDS allows load profiles analysis also at single engine level. This ensures a detailed investigation at engine level for all the monitored installations.

5.2.2 Engine running hours

By utilizing information provided by automatic shutdown analysis, engine running hours can be investigated both at installation and engine platform level. Following figure has information about engine running hours for one installation including twelve W34SG engines, and it consists of two different bar charts:

1. Upper bar chart presents total cumulated running hours for installation engines for each monitored year.

2. Lower bar chart presents yearly cumulated running hours for installation engines per each monitored year.

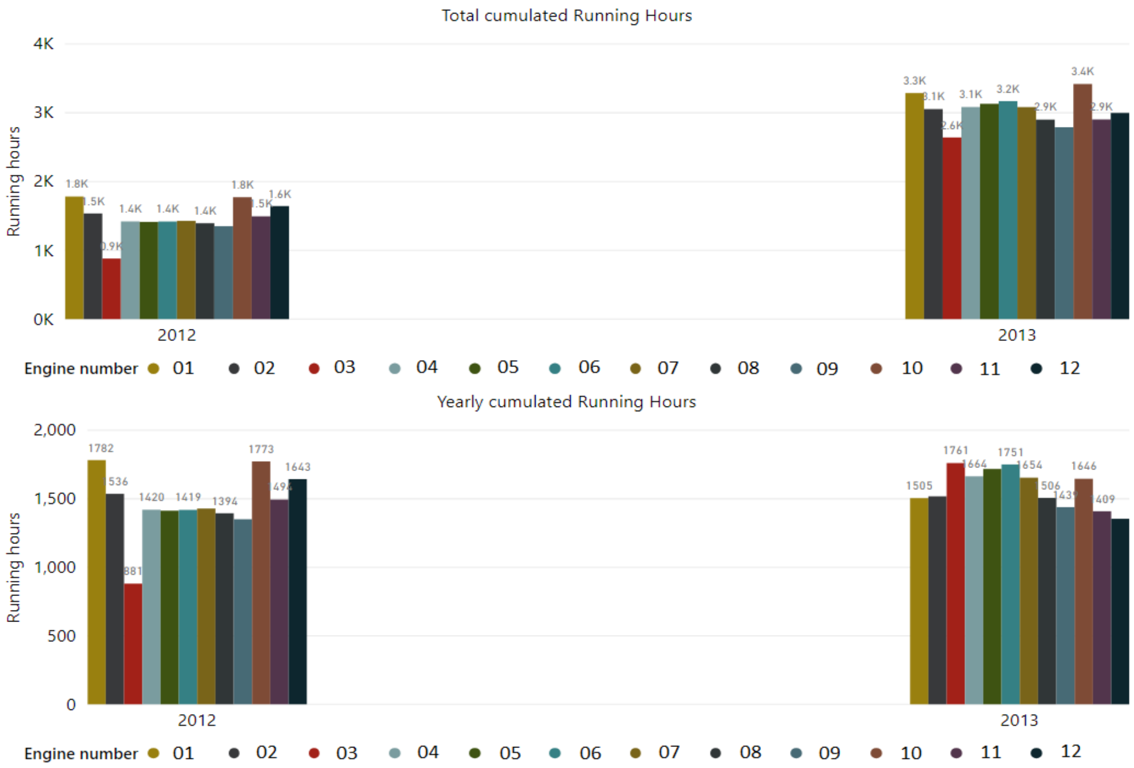


Figure 26. Total and yearly cumulated running hours for one installation including twelve W34SG engines.

5.2.3 Engine inertia

Automatic shutdown analysis also provides information about engine inertia, in other words, how long does it take for the engine speed to reach 0 rpm after the engine load has dropped to 0 kW. Figure below shows median of inertia in seconds for each investigated engine platform, in different scenarios:

- *Exhaust gas temperature deviation*, *Liner temperature deviation*, *LO pressure deviation in TC* and *LO temperature deviation in engine inlet* are considered as automatic shutdowns caused by deviation in that specific monitored signal.

- *Normal stop* means planned stop

- *Starting Failure* means stop occurring when engine mission duration is less than 15 minutes.

- *Unknown* means automatic shutdown, which is not caused by any monitored signal, mentioned in subchapter 4.3. Therefore, the triggering cause is unknown.

From the figure can be observed, that median of inertia is lower for the platforms (W34DF and W34SG) which have smaller bore size than for the platforms (W50DF and W50SG) which have larger bore size. Inertia calculation can be used in order to evaluate normal friction losses.

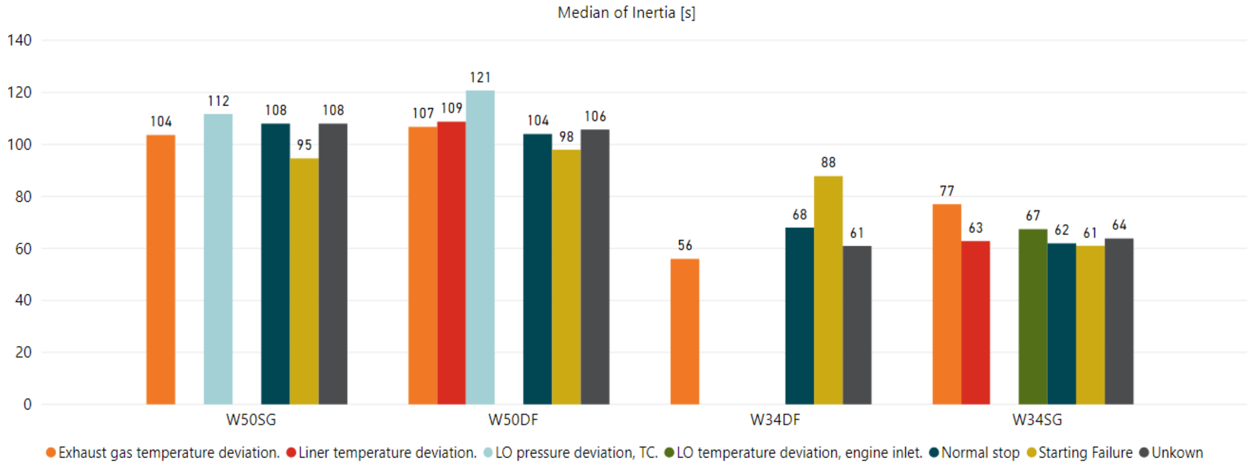


Figure 27. Median of inertia in seconds for each investigated engine platform in different stop scenarios.

5.3 Engine mission analysis

In this subchapter, all presented results are produced by automatic shutdown analysis hence all of them are for power plant engines.

5.3.1 Starting reliability

Engine missions lower than 15 minutes were labelled as starting failures. Starting failure identification allows the calculation of engine starting reliability. This key performance indicator ensures the assessment of engine performance during the starting phase. Starting reliability was calculated for each analysed engine with the following formula:

$$Starting\ reliability = 100\% * \left(1 - \frac{Number\ of\ starting\ failures}{Number\ of\ starting\ attempts}\right)$$

In the next figure, starting reliability for large bore engines is presented per each monitored year.

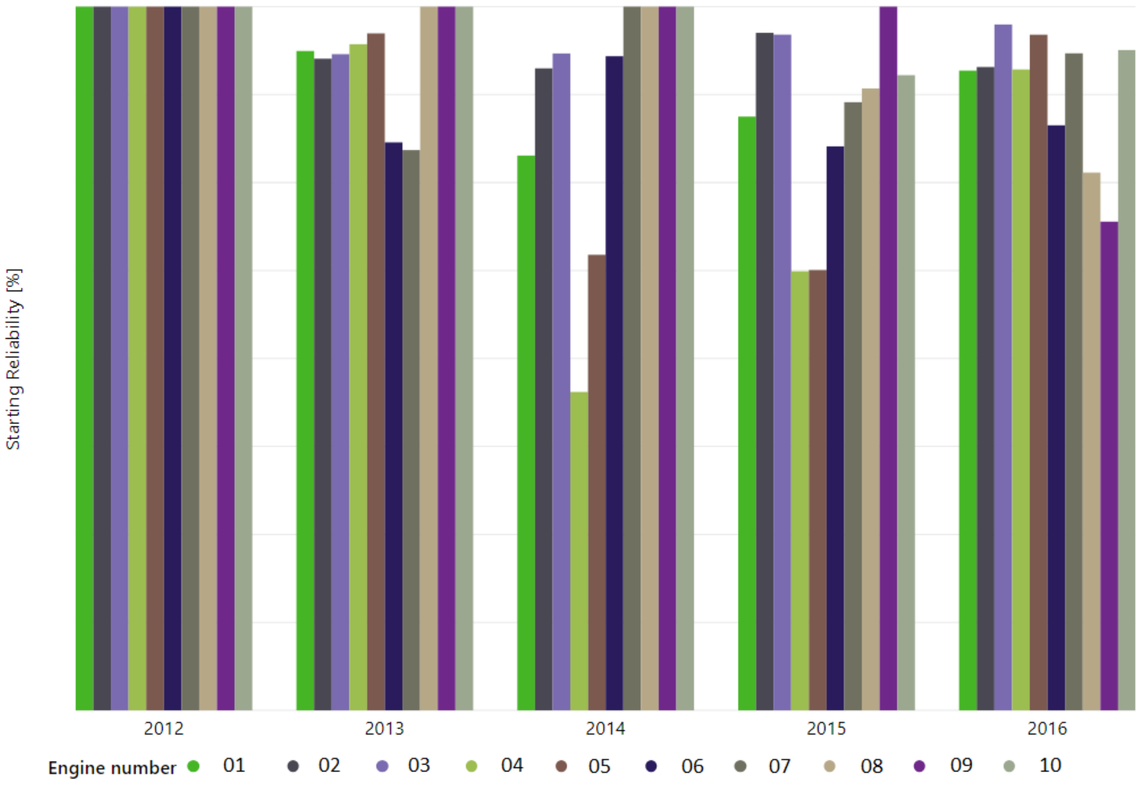


Figure 28. Starting reliabilities of large bore engines for each monitored year.

5.3.2 Automatic shutdown events

Objective of the automatic shutdown analysis was to find reasons for automatic shutdowns, and this was done by identifying deviations in monitored signals prior automatic shutdowns. Results provided by this analysis can be used to investigate distributions of different SHD types at engine, installation and engine platform level. Following figure holds information for one installation with four W34SG engines and consists of two different plots:

1. Upper plot is bar chart which shows distribution of different SHD types for each installation engine.

2. Lower plot is pie chart which shows distribution of SHD types of all installation engines.

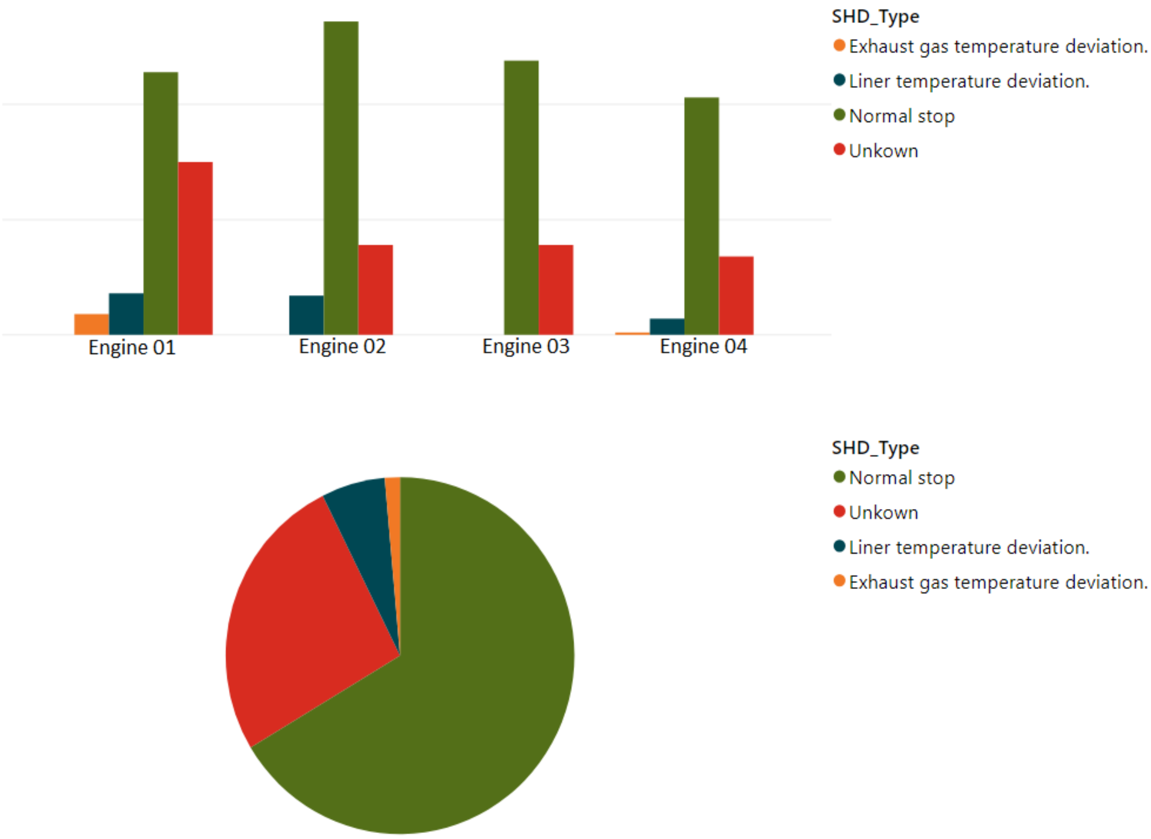


Figure 29. Distributions of SHD types both at engine and installation level for single installation.

As mentioned before, results from automatic shutdown analysis can be investigated also at engine platform level. Next figure consists of 4 different pie charts: one for each investigated engine platform. Each pie chart presents distribution of SHD types for all engines with of that specific engine platform.

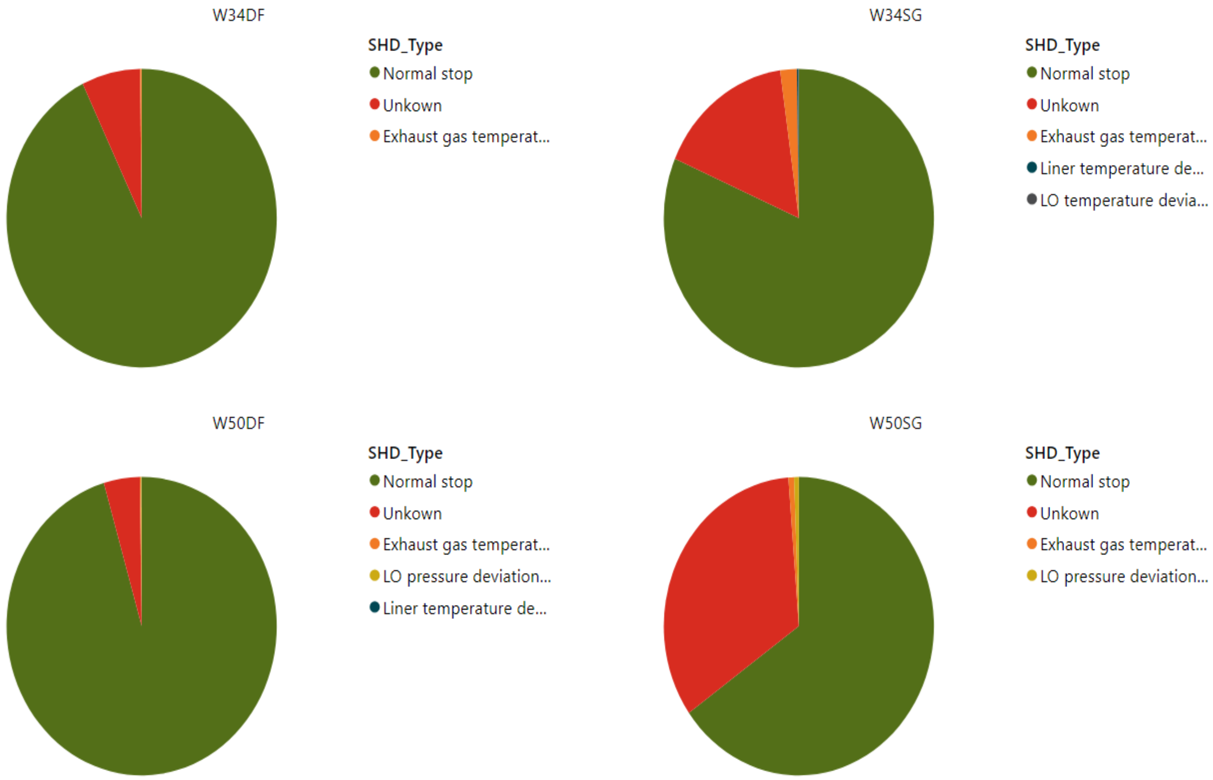


Figure 30. Pie charts of different automatic shutdown events for four different engine platforms.

In figures 29 and 30, SHD type variable indicates what was the reason for automatic shutdown. Planned stops were also included as option Normal stop. Unknown means automatic shutdown which is not caused by any monitored signal mentioned in subchapter 4.3, therefore the cause of automatic shutdown is unknown.

5.4 Sensor data analysis

5.4.1 Main feature extraction for relevant signals

In addition to methodologies for the characterization of engine behavior, algorithm was developed for main feature extraction of relevant signals. Main bearing temperature signals were selected as the test case for this solution. The feature extraction was performed by filtering out all engine starting failures and all the transient phases of selected missions (i.e. engine start up and shut down). This solution was applied to DF and SG engines of EDS operating in power plant applications, precisely 14 installations and 75 engines were processed.

Next figure presents median value of main bearing temperature is changing as a function of cumulative running hours. Each point represents the median temperature value during a specific engine mission.

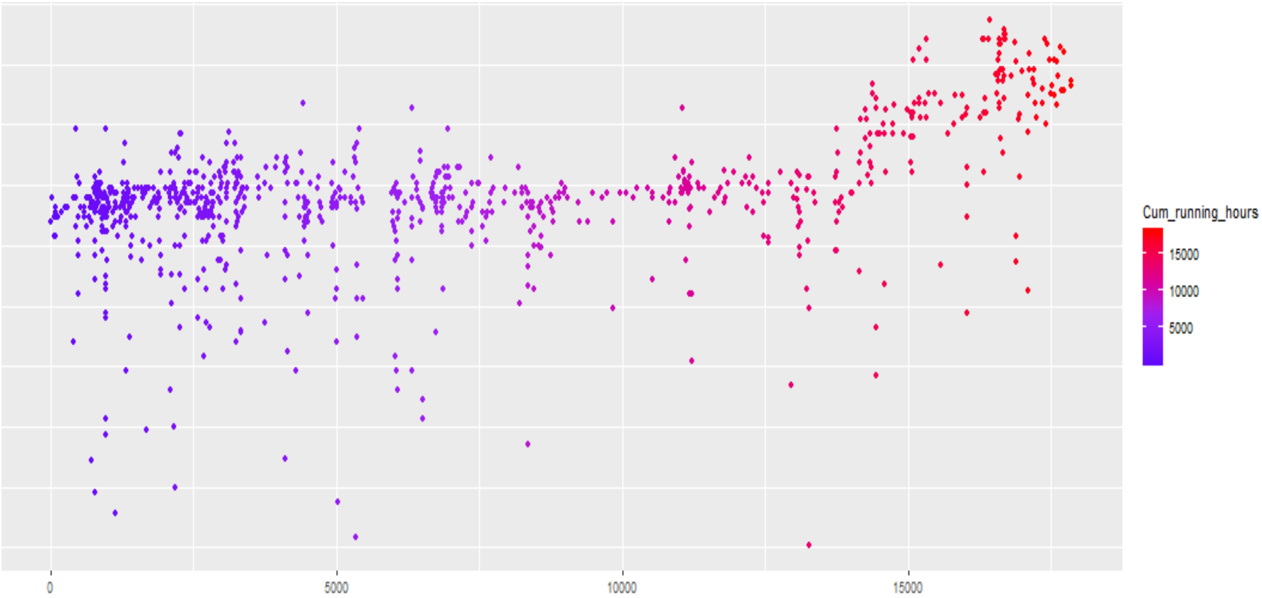


Figure 31. Median value of main bearing temperature as function of cumulative running hours.

From the figure above it can be observed, that the usual median value of main bearing temperature starts to increase approximately after 13 thousand cumulative running hours. The reason for this increasing trend could be faulty sensor, since it was known that monitored engine had some problems with temperature sensors at that time period. In addition to identification of trends, the figure helps to identify anomalies in sensor behavior.

Proposed solution can be modified to derive customized features for any signal present in the data. Also, the investigated time periods can be modified according to different needs and conditions. Extraction of this information is essential for the development of predictive maintenance algorithms and diagnostic tools.

5.4.2 Anomaly detection of sensor signals

EDS data was finally investigated in order to detect anomalies in behaviour of selected sensor signals. The anomaly detection application utilizes linear regression and interquartile range method to deem if samples of preprocessed data are inliers or outliers. Proposed approach was tested to find anomalies among main bearing temperature and exhaust gas temperature signals. The solution was applied to one DF engine of EDS operating in power plant application.

The figure below represents how anomaly detection algorithm has classified samples of main bearing temperature as inliers or as outliers. Main bearing temperature is plotted as function of engine load percentage. Blue and red dots in the plot represent samples of main bearing temperature, and green dots represent the predicted temperature values calculated with regression model, which has been defined by the anomaly detection algorithm. Blue colour for main bearing temperature sample means that sample is classified as inlier and red colour means that sample is classified as outlier.

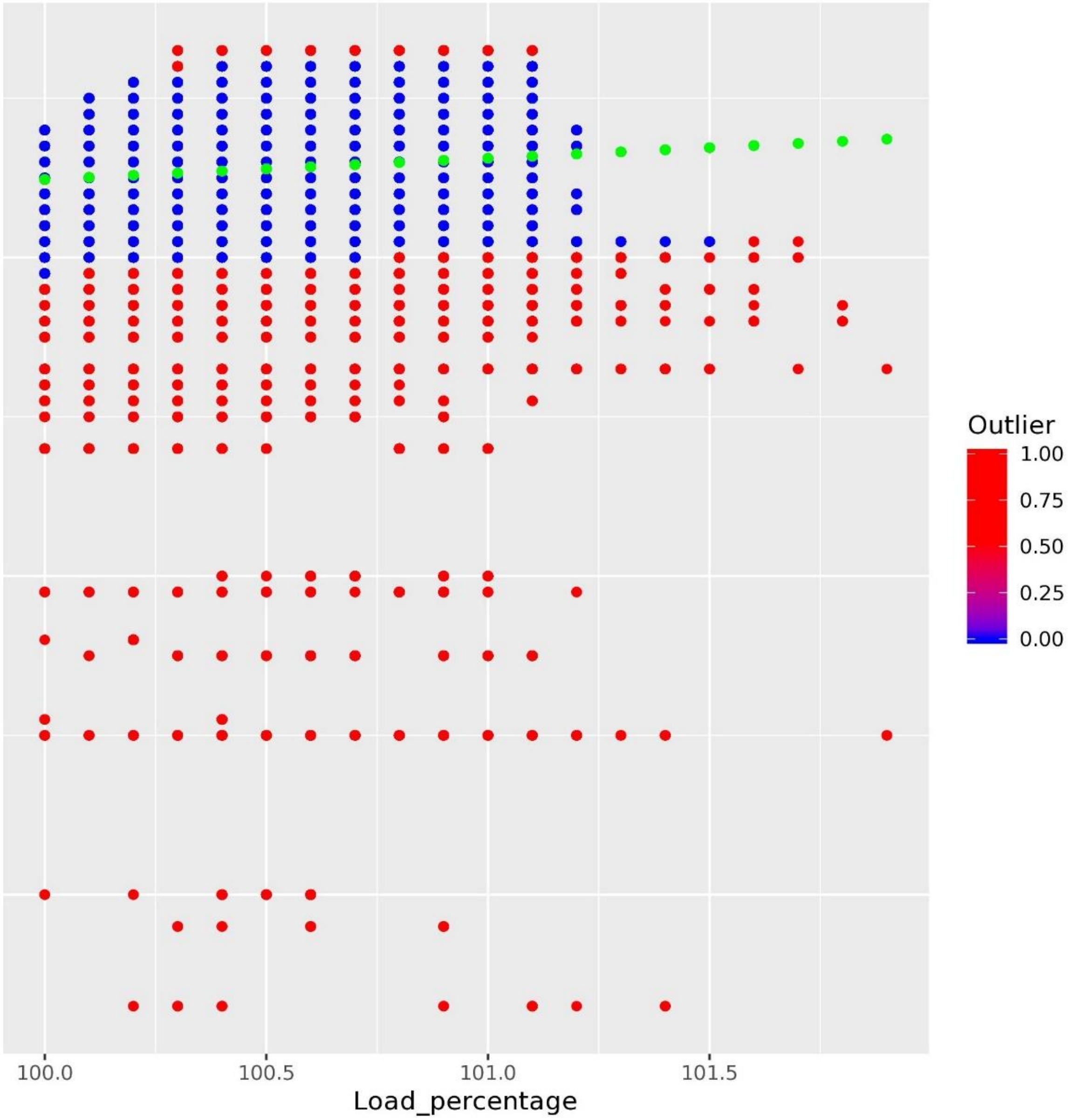


Figure 32. Anomaly detection – Example of results.

Before analyzing the figure above, it should be mentioned, that there are lot of overlapping samples visualized in the figure, hence it is difficult to estimate from the plot their density. As it can be observed, the anomaly detection algorithm classifies values in certain temperature interval as inliers and other values as outliers. Considering adopted approach, this means, that at least 50% of the samples are within this temperature interval in given load percentage interval.

At the moment, this approach is able to identify from the given sensor data if there is any anomaly behavior in a given sensor signal. In the future this approach should be scaled up in order to identify correlations between signal anomalies and events taking place in the engine in order to perform engine health monitoring and diagnostic.

6 CONCLUSIONS AND DISCUSSION

In this thesis, Wärtsilä Engine Data Sandbox (EDS) was investigated and its content was described and mapped. EDS content description and mapping allowed identification of the majority of installations as well as retrieving of information related to engine characteristics, such as type, cylinders configuration, etc. Relying on these data, four different data-driven approaches were developed to support engine performance characterization and analysis. Four developed data-driven approaches were:

- Load distribution analysis, to be used for engine usage profiling and damage accumulation models (e.g. Fatigue analysis).

- Automatic shutdown analysis, for engine health monitoring purposes.

- Main feature extraction of relevant sensor signals as preparatory activity for the development of predictive maintenance algorithm and diagnostic tool. In addition, it allowed the identification of trends in signals for investigation of potential failures.

- Anomaly detection of sensor signals to perform sensors validation.

During the implementation of these four different data-driven approaches, many different functionalities for data preparation were developed. These functionalities include, for instance, fixing the sampling frequencies of selected signals and extracting the required monitoring periods. Therefore, a comprehensive toolbox for data preparation was developed.

Finally, limitations and exceptions of EDS contents were described. EDS data can be utilized to do characterization of engine behaviour and to find anomalies from that behaviour. However, lack of automation system signals is clear disadvantage when trying to implement data-driven approaches with EDS data. In addition to this, more information related to engines is required in order to utilize EDS data for predictive maintenance. This kind of information includes for example, maintenance information of installations and information related to engine operating conditions.

Future activities after this research include retrieving missing EDS related information, creating proper infrastructure for EDS, investigating possible causes for anomalies detected by proposed anomaly detection application and improving the data preparation modules.

The missing EDS information, which should be retrieved, includes installation numbers for unidentified installations and signal codes for unidentified signals. After this, proper infrastructure for EDS should be created in order to grant possibility for Wärtsilä employees to investigate EDS content and implement their own data analytics approaches easily.

At this stage, anomaly detection approach is able to identify from the given sensor data if there is any anomaly behavior in a given sensor signal. In the future, this approach should be scaled up in order to identify correlations between signal anomalies and events taking place in the engine in order to perform engine health monitoring and diagnostic.

Finally, the data preparation functionalities can be regarded as first version of data preprocessing toolbox for data of Engine Data Sandbox. Nevertheless, this tool could be enhanced by adding new functionalities including, for example, proper noise filtering.

REFERENCES

Alpaydin, Ethem (2010). *Introduction to Machine Learning.* 2nd edition. The MIT Press. ISBN 978-0262012430.

AWS (2019a). Amazon EC2 [online]. [Referred on: 30.8.2019] Available at: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

AWS (2019b). Amazon S3 [online]. [Referred on: 30.8.2019] Available at: https://aws.amazon.com/s3/

Birolini, Alessandro (2013). *Reliability Engineering: Theory and Practice.* 7th edition. Springer. ISBN 978-3642395345.

Bosch, Robert (2011), *Automotive Handbook.* 8th edition. John Wiley & Sons. ISBN 978-1-119-97556-4.

Chandola Varun, Arindam Banerjee & Vipin Kumar (2009), Anomaly Detection: A Survey. *ACM Computing Surveys,* Volume 41, Issue: 3, 15:1-15:58.

Chapelle Oliver, Bernhard Schölkopf & Alexander Zien (2006). *Semi-Supervised Learning.* The MIT Press. ISBN 978-0-262-03358-9.

Chen Min, Shiwen Mao & Yunhao Liu (2014). Big data: A survey. *Mobile Networks and Applications,* Volume 19, Issue: 2, 171-209.

Elsayed, Elsayed (2012). *Reliability Engineering.* 2nd edition. John Wiley & Sons. ISBN 978-1-118-13719-2.

Foanene, Adriana (2016). Internal Combustion Engines. *Fiabilitate şi Durabilitate.* Volume 2, Issue: 18, 166-169.

Gnedenko Boris & Igor Ushakov (1995). *Probabilistic Reliability Engineering.* 1st edition. John Wiley & Sons. ISBN 978-0471305026.

Heywood, John (1988). *Internal Combustion Engine Fundamentals.* International edition. McGraw-Hill. ISBN 0-07-100499-8.

Hirche Sandra, Peter Hinterseer, Eckehard Steinbach & Martin Buss (2005). Towards deadband control in networked teleoperation systems. *IFAC Proceedings Volumes,* Volume 38, Issue: 1, 70-75.

Kapil Gayatri, Alka Agrawal & R. A. Khan (2016). *A Study of Big Data Characteristics.* In 2016 International Conference on Communication and Electronics Systems (ICCES), 110-113, October 21-22 2016, Coimbatore, India.

Kececioglu, Dimitri (2002). *Reliability Engineering Handbook.* Volume 1. DEStech Publications, Inc. ISBN 1-932078-00-2.

Khan Nawsher, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mammoud Ali, Muhammad Alam, Muhammad Shiraz & Abdullah Gani (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. *The Scientific World Journal,* Volume 2014.

Lazzaroni Massimo, Loredana Cristaldi, Lorenzo Peretto, Paola Rinaldi & Marcantonio Catelani (2011). *Reliability Engineering: Basic Concepts and Applications in ICT.* 2012 edition. Springer. ISBN 978-3-642-20982-6.

0'Connor Patrick & Andre Kleyner (2012). *Practical Reliability Engineering.* 5th edition. John Wiley & Sons. ISBN 978-0470979815.

Najjar, Yousef (2009). Alternative fuels for spark ignition engines. *Open Fuels and Energy Science Journal,* Volume 2, Issue: 1, 1-9.

R Project (2018). R FAQ [online]. [Referred on: 2.9.2019] Available at: https://cran.r-project.org/doc/FAQ/R-FAQ.html#What-is-R_003f

Ravishankar N.R. & M.V. Vijaykumar (2017). Reinforcement Learning Algorithms: Survey and Classification. *Indian Journal of Science and Technology,* Volume 10, Issue: 1.

Rencher, Alvin C. (2002). *Methods of Multivariate Analysis*. 2nd edition. John Wiley & Sons. ISBN 0-471-41889-7.

RStudio (2018). RStudio [online]. [Referred on: 2.9.2019] Available at: https://www.rstudio.com/products/rstudio/

Russel Stuart & Peter Norvig (2010). *Artificial Intelligence: A Modern Approach*. 3rd edition. Prentice Hall. ISBN 978-0-13-604259-4.

S3 Browser (2018). What is S3 Browser [online]. [Referred on: 30.8.2019] Available at: https://s3browser.com/

Stone, Richard (1999). *Introduction to Internal Combustion Engine.* 3rd edition. Palgrave Macmillan. ISBN 978-0-333-74013-2.

Taritaš Ivan, Mario Sremec, Darko Kozarac, Mislav Blažić & Zoran Lulić (2017). The effect of operating parameters on dual fuel engine performance and emissions – An overview. *Transactions of Famena,* Volume 41, Issue: 1, 1-14.

Topuz, Ercan (2009). Reliability and Availability Basics. *IEEE Antennas and Propagation Magazine,* Volume 51, Issue: 5, 231-236.

Wei, Lijiang & Peng Geng (2016). A review on natural gas/diesel dual fuel combustion, emissions and performance. *Fuel Processing Technology,* Volume 142, 264-278.