POLITECNICO DI TORINO

Corso di Laurea Magistrale in
Ingegneria Energetica e Nucleare

Tesi di Laurea Magistrale

Diagnostic and prognostic maintenance decision making. A case study regarding the main equipment of a cogeneration plant

Relatore:
Prof. Andrea Carpignano
Correlatori:
Ing. Paolo Tarasco
Ing. Raffaella Gerboni

Candidato:
Andrea Vercellotti

A.A. 2018/19

Alla mia famiglia che mi ha sempre supportato

Index

Introduction

Objective

The present work is framed in the field of maintenance engineering. Its primary aim is the application of a statistical technique and the evaluation of its effectiveness on monitoring performance and health condition of complex equipment of a cogeneration plant. This plant is serving food industry. More specifically, the case study regards one internal combustion diesel engine. Nowadays, maintenance plays a decisive role in the path towards sustainable manufacturing and data-driven predictive maintenance strategies are expected to yield multiple benefits: minimization of negative environmental impacts, increase in the life span of facilities and reduction in maintenance costs. For example, the consumption of hydraulic oil of a machine tool can be decreased. By measuring crucial parameters like the concentration of particles and the water content in the hydraulic oil, it is possible to change the fluid only when necessary. Whenever the lubrication oil is substituted later than the conservative maintenance schedule suggested by the supplier, the annual quantity of waste lubricating oil is reduced. As a further example, if the degradation process of a SCR system is monitored, then appropriate maintenance tasks will be executed prior to component failure. These actions permit to use the component more efficiently, with concurrent cuts in the emissions of NO_x pollutants.

Expected results

This research aspires to develop a method for the gathering, the rationalization and the validation of the available operational data. The first part of the study is dedicated to data processing and to the creation of a reference model for the internal combustion engine, on the basis of the available database. The monograph contains the definition of computational codes and verification of model robustness and accuracy, with the support of maintenance history for the components under examination. Moreover, output results analysis is offered, for the purpose of identifying incipient conditions of drop in operational performance and/or of operational problems that are dissimilar from those due to regular working cycles of the endothermic engine, ICE.

State of the art

Before 1950, maintenance was essentially unplanned, taking place only when breakdowns occurred. Between 1950 and 1960, preventive maintenance (PM), also named planned maintenance, was introduced. The idea behind this technique is to establish periodic intervals for machine inspections and maintenance regardless of its health condition. Even though this policy sometimes reduces equipment failures, it is labour-intensive, it does not remove catastrophic failures and it results in unnecessary maintenance. In this context, condition-based maintenance (CBM) steps in. Condition based-maintenance suggests maintenance interventions either founded on information gathered via online monitoring or based on a non real-time signal processing, like vibrational monitoring. CBM tries to avoid superfluous maintenance duties by performing maintenance actions only when an incipient failure

condition is observed. A CBM program, if properly defined and successfully implemented, can significantly lessen maintenance costs [1] [2]. Furthermore, CBM is an efficient method for switching from classical “fail and fix” practices to a “predict and prevent” methodology. Generally, the target of CBM is the informed judgement that supports maintenance decisions and the success of CBM relies upon three connected processes: monitoring, diagnostics and prognostics. Diagnostics is the process of estimating the health status and the equipment degradation exploiting information delivered by the condition-monitoring system. The most important objectives of diagnostics are:

a) fault detection, which indicates that an unwanted event is impending;
b) fault isolation, which locates the faulty component;
c) fault identification, which facilitates the ascertainment of the root cause of the fault.

Prognostics is the skill to forecast the evolution of engine deterioration [3]. In other words, prognostic aids in foretelling how much time is left before a failure (or a fault) arises, given the current state of the system and its past operating profile. The time left previous to observing a failure is called remaining useful life [1]. Lastly, recent years are witnessing the rapid development of prognostics and health management (PHM), that aims to supply users with an integrated view of the health state of a machine or an overall system. This is achievable thanks to the growth of information technology, IT. Health management is the process of performing opportune and prompt maintenance actions and making precise logistics decisions based on outcomes from diagnostics and prognostics, available resources and operational request [4].

The present thesis addresses the aforementioned topics, with a particular attention to diagnostic and prognostic in maintenance decision making.

Methodology and case study

In order to fulfil these intentions, it is necessary to process a big amount of data. A large raw database concerning a whole year of equipment operation was supplied. The preliminary step consists in defining, for endothermic engine, operating conditions, such as “full load”, “derated”, “stand-by”, etc. In case of several operating conditions, the creation of as many databases as the number of operating conditions is needed. Then the issue of rationalization of the entire data set is coped with. This procedure requires two stages. The first is on the single parameter, with the use of a control criterion based upon standard deviation. The second demands a comparison between subsets of parameters. A choice of significant variables is realized and the selection is grounded in Failure Modes and Effects Analysis, FMEA. FMEA is a qualitative and inductive methodology that allows highlighting the failure modes of different components that could affect the system’s functionality. The approach in FMEA seeks to point out a rational strategy for maintenance. In fact, FMEA involves reviewing systems to discover the mode of failure that may occur and its effect. The FMEA audit produces information on the range of variables to be quantified for specific failure modes. The various parameters to be considered are usually those which connote a fault condition, by either a decrease or an increase in the characteristic measured value [5] [6]. After the selection

of most relevant variables, z-score normalization is carried out. At this stage a trend of each important variable, with respect to time, is shown. When the percentage of retained data is suitable for a statistical analysis, the last step comprehends the application of one distinct statistical technique, Principal Component Analysis. Output results are essential for drawing conclusions, according to the prearranged goal. Flowchart in Figure 1 outlines the logical scheme followed throughout this experimental research. Blue blocks are within the scope of data pre-processing and they are described in the first half of Chapter 1 and Chapter 2; red blocks refer to statistical investigation of data and it is detailed in the second half of Chapter 1 and Chapter 2.

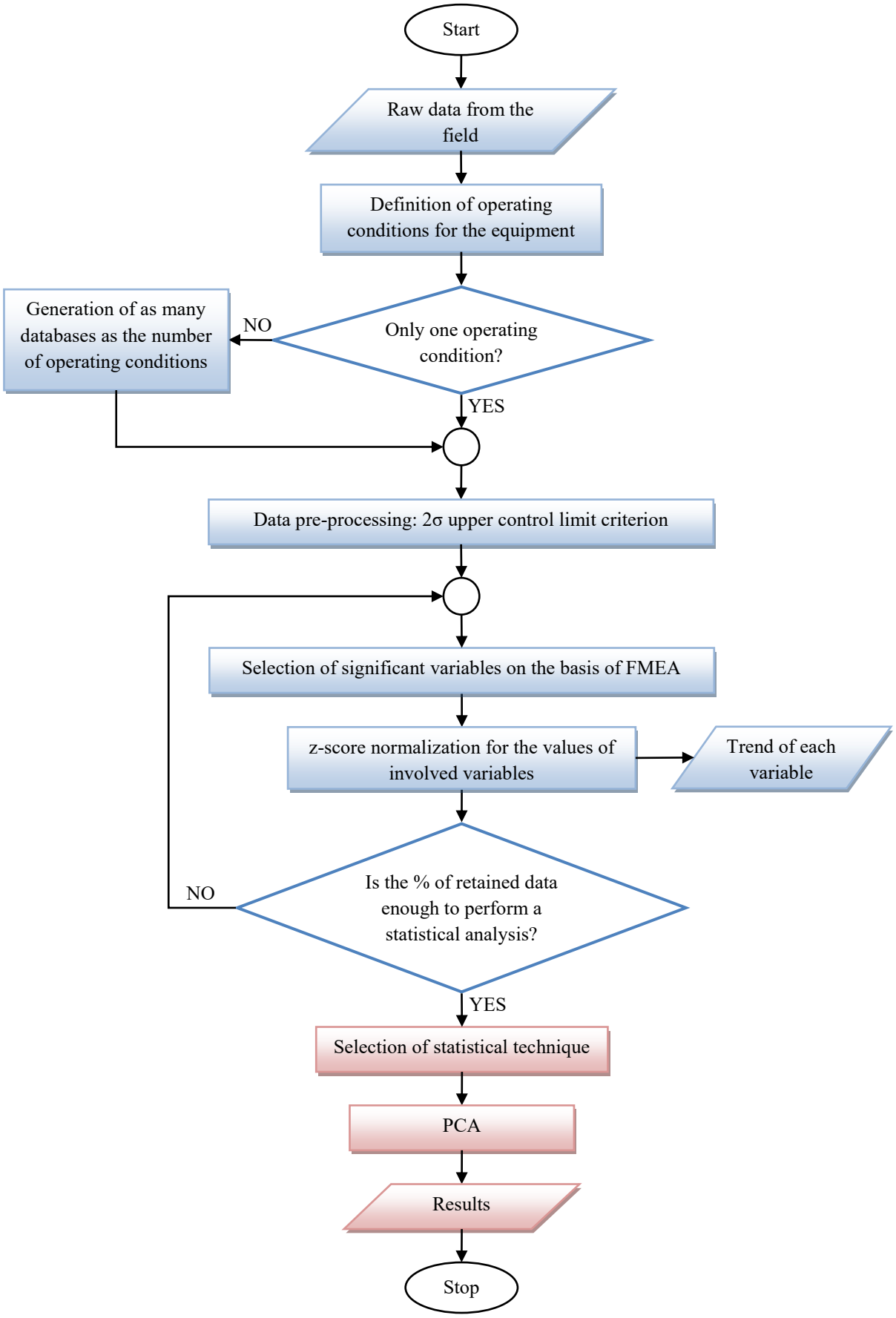


Figure 1. An overview of the methodology

Chapter 1 – Theoretical aspects

This chapter starts with an exposition of the tools developed for data pre-processing. Some alternative options of data processing are sketched, since the procedure followed was not the only possible solution. Subsequently, the chapter establishes the mechanics and properties of the data-analytic technique that was named Principal Component Analysis, PCA. A paragraph is then devoted to the multivariate quality control procedure which was employed in the current thesis, the Hotelling T^2 statistic. Lastly, another method is introduced, the Q-statistic, and the reasons that promoted the choice of the former test will be explained.

Chapter 1 attains the objective of covering the theoretical framework underlying this case study.

Tools developed for data pre-processing

Data acquisition is a series of steps that comprehend the collection, the conversion and the recording of useful data from a physical asset. The hardware of data acquisition systems usually comprises sensors, an amplifier circuit, an analogue-to-digital (A/D) converter, a data transmission device and a data recording circuit. A sensor is a converter that measures a physical quantity and converts it into a signal which can be read by an operator or by an electronic device. An electronic amplifier is an electronic instrument that elevates the power of a signal, ensuring the output matches the input signal shape, with a greater amplitude. An A/D converter is a tool which enables the conversion of a continuous physical quantity (generally voltage) into a digital number that represents the quantity amplitude. Digital signals bringing information on the health state of the component have to be conveyed to the control computer.

Data processing plays a pivotal role in machinery prognostics and maintenance management and decision making. One of the first stages of data processing is data cleaning because data always contain outliers and errors. The scope of data cleaning is enhancing the chances that error-free data are employed for study and modelling. Without data cleaning step, the analyst could run into the “garbage in garbage out” situation. Incorrect data are due to several causes, such as human mistakes or sensor faults. There is not usually a simple way to clean data [1].

The abundance of data, coupled with the necessity for powerful data analysis tools, can be depicted as a data rich but information poor situation. In fact, a tangible risk is that this vast amount of data, collected and stored in large data repositories, could be seldom used to take important decisions. Efforts to extract the knowledge embedded in the huge amount of data are then extremely valuable. Therefore, data pre-processing, which is a branch of the broad subject called data mining, aims at improving the quality of a database, to prepare the basis for a reliable statistical research. The actions required for the knowledge discovery process are:

1) data cleaning, to remove noise and inconsistent data;
2) data integration, where multiple data sources may be merged;

3) data selection, where data suitable for the examination task are retrieved from the database;
4) data transformation, where data are consolidated and turned into fitting forms for mining by conducting summary or aggregation and/or reduction operations;
5) data mining, a key process where brilliant methods are executed to extract data patterns;
6) pattern evaluation, to recognize the patterns which are really interesting;
7) knowledge presentation, where reports and visualization techniques are utilized to show mined knowledge to operators and users.

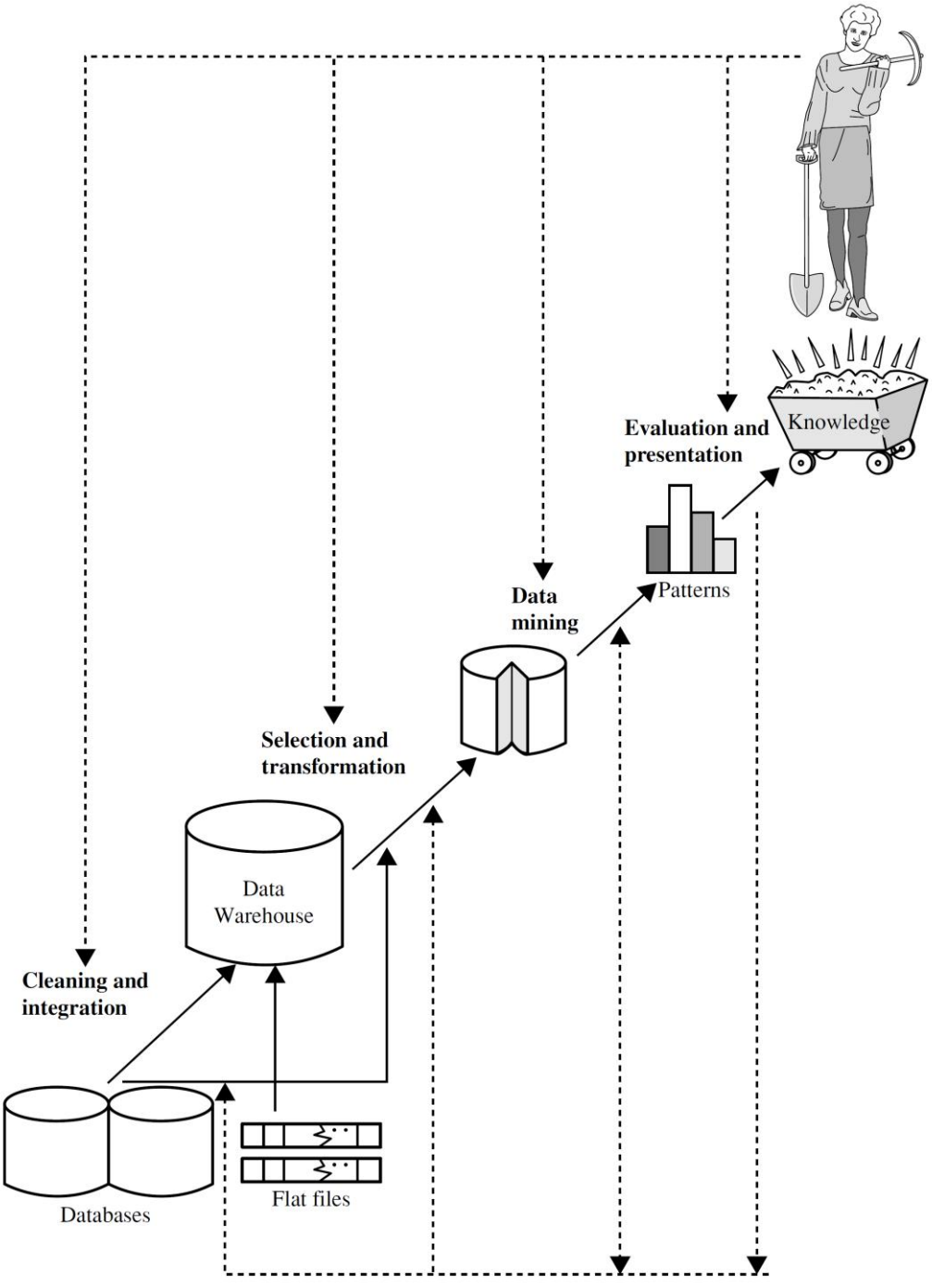


Figure 2. Knowledge discovery process as an iterative sequence of steps [7]

Figure 2 summarizes the knowledge discovery process as an iterative sequence of stages. Stages 1 through 4 are different types of data pre-processing, where data are prepared for mining.

For a successful data pre-processing it is indispensable to have an overall view of the available database. A statistical description aids in identifying features of the data and underlining which data values should be treated as outliers. Statistics studies the collection, analysis, interpretation and presentation of data. Data mining has an intrinsic and strong link with statistics. A statistical model is a set of mathematical functions that outline the behaviour of the elements in a target class in terms of random variables and their related probability distributions. Quantities such as sensor measurements are thus considered random variables. Applying statistical techniques in data mining is far from trivial. A significant challenge is how to scale up a statistical method over a big database. Various statistical methods have a certain degree of complexity in computation and algorithms should be cautiously structured and adapted to avoid unacceptable computational costs.

Another fast-growing discipline, closely related with data mining, is machine learning; it investigates how computers can learn from data to identify complicated patterns and to make clever decisions. Anyway, the above-mentioned matter goes beyond the scope of this thesis [7].

Outliers may be detected adopting statistical tests that assume a probability model for data, hence the reason for a look at a statistical quality control method. A method deriving from Shewhart control charts has been chosen. Control charts are mainly determined by two values: the upper and lower control limits. To use these charts, data generated for a single monitored parameter are separated into subgroups and subsets statistics, like the subgroup average and standard deviation, are computed. When the subset statistic does not fall within the two limits, the inference is that there is at least one outlier in that subgroup, which is discarded [8].

Delving into the details, let suppose to have N independent and normally distributed observations of a parameter X(t), which is interpreted as a random variable with mean μ_0 and variance σ_0^2. These N observations can be seen like a sequence of m random samples $\{X_{ij}; i = 1,...,n; j = 1,...,m\}$ with n = N/m elements for each sample. The mean of a specific sample is:

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}\,; \quad j = 1, \dots, m \tag{1}$$

This statistic has mean and variance that are connected with those belonging to X:

$$E\{\overline{X}_j\} = \mu_0; \quad Var\{\overline{X}_j\} = \frac{\sigma_0^2}{n} \tag{2}$$

If the process is under a statistical control condition, 1 - α is the probability that any sample mean will fall between

$$\mu_0 - \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2}; \quad \mu_0 + \frac{\sigma_0}{\sqrt{n}} u_{\alpha/2} \tag{3}$$

It is customary to replace $u_{\alpha/2}$ by 3, yielding $\alpha = 0.0027$ and a probability equal to 99.73%, so that three-sigma limits are employed. For the current work, a two-sigma rule was adopted, only with upper control limit, because it proved to be a good compromise between the risk of pointing out a false outlier and the risk of missed outlier; in the latter drawback the control method is not able to detect an outlier. Two-sigma rule reduces the indicated probability to 95.46%.

The values for μ_0 and σ_0^2 have to be estimated on the basis of available data samples:

$$\hat{\mu}_0 = \overline{X} = \frac{1}{m}\sum_{j=1}^{m} \overline{X}_j \tag{4}$$

$$\hat{\sigma}_0^2 = \frac{1}{m}\sum_{j=1}^{m} S_j^2 \tag{5}$$

with:

$$S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_{ij} - \overline{X}_j\right)^2; \quad j = 1, \dots, m \tag{6}$$

The result coming from equation (4) defines the central line of $\overline{X}$ chart while ±2 times the result given by the square root of (5) specifies the two distances, from central line, at which upper and lower control limits for $\overline{X}$ chart have to be drawn. The space delimited by the upper control limit (UCL) and the lower control limit (LCL) is called control band. The statistic to insert into $\overline{X}$ chart, for each *m*th sample, is the one provided by (1).

On $\overline{S}$ chart, the statistic S_j, gained with the square root of (6) is represented. In order to assess the central line and control limits for $\overline{S}$ chart, it is necessary to compute average and variance of S_j:

$$E\{S_j\} = \sigma_0 c_4; \quad Var\{S_j\} = \sigma_0^2(1 - c_4^2) \tag{7}$$

where:

$$c_4 = \frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\Gamma\left(\frac{n-1}{2}\right)} \tag{8}$$

The coefficient c_4 is frequently reported in existing literature on the subject, arranged in tabular form, since it varies according to sample size n. Instead of utilizing equations in (7), the following two excellent approximations were chosen:

$$E\{S_j\} \cong \sigma_0\sqrt{(2n-3)/(2n-2)} \tag{9}$$

$$Var\{S_j\} \cong \sigma_0^2/[2(n-1)] \tag{10}$$

Thanks to this second option, after the selection of n, MATLAB code can automatically evaluate last equations.

Equation (9) brings to the central line of $\bar{S}$ chart, whereas the square root of (10) multiplied by 2 produces the distance, from central line, where control limits have to be plotted [9]. The final assumption for this particular case study is that control chart related with average, $\bar{X}$ chart, is not significant and may lead to the wrong elimination of data samples.

An alternative method for statistical quality control is represented by $\bar{X}$ and R charts. R denotes the range of the sample, in other words, the difference between the largest and smallest observations inside the sample. Both methodologies should be based on at least 20 to 25 subgroups or samples and this constraint is amply honoured, since every parameter is made of thousands of measurements. Generally, $\bar{X}$ and $\bar{S}$ charts are preferable to their counterparts, $\bar{X}$ and R charts, when the sample size n is moderately large, namely $n > 10$. The range method for estimating the variance of a variable X(t) loses statistical efficiency for moderate to large samples. In view of the fact that in this study $n = 20$, $\bar{X}$ and R charts are clearly not a compatible possibility.

As stated above, a parameter X(t) located in the database is seen like a random variable with normally distributed observations. Broadly speaking, if x is a normal random variable, it follows the normal probability distribution:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}; \quad -\infty < x < \infty \tag{11}$$

The mean of the normal distribution is μ and the variance is $\sigma^2 > 0$. The visual appearance of the normal distribution is a symmetric and bell-shaped curve, displayed in Figure 3. There is a straightforward interpretation of the standard deviation of a normal distribution. For example, the 68.26% of the population values fall in the interval determined by the mean minus and plus one standard deviation.

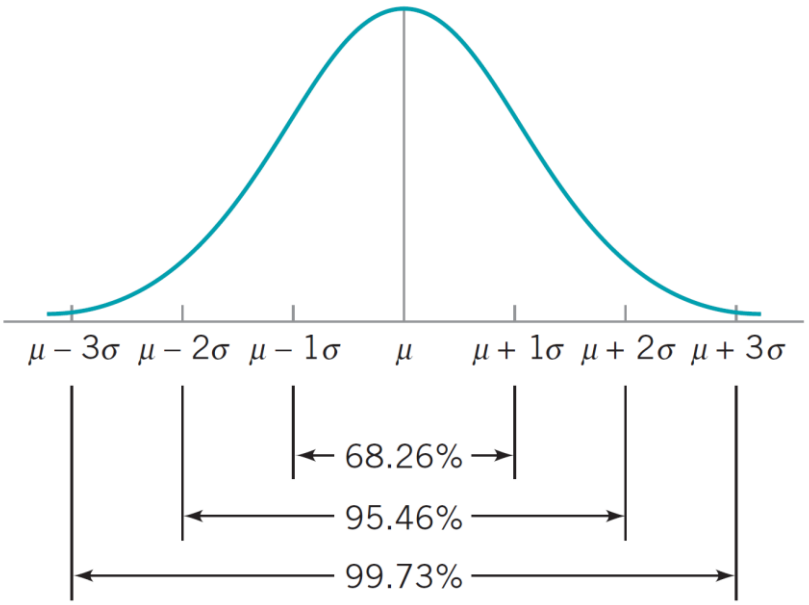


Figure 3. Areas under the normal distribution [10]

Considering that a uniform and organized database, the output of data pre-processing, is a fundamental requirement for a truthful statistical analysis, it is common practice to apply this transformation:

$$z = \frac{x - \mu}{\sigma} \tag{12}$$

The aforementioned operation is known as z-score normalization or zero-mean normalization or simply standardization, because it converts a normally distributed random variable, denoted by $N(\mu,\sigma^2)$, into an N(0,1) random variable [10]. It is worth underlining that parameters standardization entails the removal of measurement unit from every variable, since standard deviation is expressed in the same units of measurement. The measurement unit employed may affect data analysis. For instance, a variable could seem to contribute to a great extent to the overall variability of the system, just because its scale of measurement has larger magnitudes than the other variables. Standardization finds a nice answer to this issue, trying to give all variables an equal weight. An alternative for data normalization is min-max normalization, which performs a linear transformation on the original data. Let suppose that min_Q and max_Q are the minimum and maximum values of a variable Q. Min-max normalization maps a value v_i of Q to v_i' in the range (new_min_Q,new_max_Q) by calculating

$$v_i' = \frac{v_i - min_Q}{max_Q - min_Q}\left(new\_max_Q - new\_min_Q\right) + new\_min_Q \tag{13}$$

Min-max normalization maintains the relationships among the original data values [7].

Z-score normalization is the last mosaic piece of data pre-processing and it follows temporal alignment of data, a problem connected with this peculiar case study whose solution will be discussed in the respective section.

Methodology of Principal Component Analysis

Maintenance activity merges different techniques and methods to decrease maintenance costs while augmenting availability, reliability and safety of equipment. As a consequence of this, one generally speaks about failures diagnostic, development of strategic responses at management level and scheduling of these actions. The above-stated steps are in line with the need of perceiving and understanding the phenomena, with the aim of acting accordingly. However, instead of comprehending a failure event which has just occurred, predicting or anticipating its manifestation seems more convenient in order to adopt suitable countermeasures; this is defined as the prognostic process. Prognostic permits to boost safety, reduce maintenance costs and down time, owing to an improvement in maintenance organization. Prognostic is based on assessment criteria, whose limits are dependent on the system itself and on the desired performance. Strictly speaking, the accuracy of a prognostic system is connected to its capability to approximate and forecast the equipment degradation. Prognostic methods can be associated with one of the following two approaches, each of them with pros and cons: model-based and data-driven approach. The model-based method supposes that a precise mathematical model can be formulated from first principles. For

instance, fatigue models founded on physics have been extensively used to characterize the genesis and the propagation of structural anomalies. The main benefit of model-based approaches is their ability to include a physical comprehension of the monitored system. The strong correlation with a mathematical model may also turn out to be a weak point: it is sometimes difficult or even impossible to simulate system behaviour. Data-driven method exploits real data to recognize features revealing the deterioration of components and to foresee the global system functioning. Data-driven approaches can be categorized into two distinct classes: artificial intelligence (AI) techniques, such as artificial neural networks (ANNs), and statistical techniques like multivariate statistical methods. The strength of data-driven procedures consists in their ability to convert high-dimensional and noisy data into pieces of information which are crucial to diagnostic and prognostic decisions. A major disadvantage of data-driven approaches is their heavy reliance on the quantity and quality of system operational data [11].

The technique that best suits the purpose of this work is the Principal Component Analysis, a multivariate statistical method in which a number of related variables are transformed into a smaller set of uncorrelated variables. PCA method dates back to Karl Pearson in 1901, even though the widespread procedure used nowadays had to wait for Harold Hotelling who made a fundamental contribution to the technique development.

A basic knowledge of matrix algebra is indispensable for the grasp of the present section. Some essential definitions and operations linked to matrix algebra will be briefly reported, whereas theorems are not germane to this dissertation and will not be included. In this paragraph, the method of principal components is explained by dint of a small hypothetical two-variable example, involving a couple of parameters that are part of the case study. These variables are the coolant Δp in Charge-Air Cooler for banks A and B. The example encompasses 15 pairs of observations which were obtained with a 10 minutes time span between two pairs of measurements, as displayed in Table 1.

Minutes of the month	Coolant Δp bank A [mbar]	Coolant Δp bank B [mbar]
35860	81.9	78.9
35870	78.6	73.5
35880	75.2	68.1
35890	71.9	62.8
35900	72.2	62.4
35910	75.2	67.5
35920	78.2	72.6
35930	81.2	77.7
35940	84.2	82.8
35950	85.6	87.0
35960	82.4	81.6
35970	79.2	76.2
35980	76.1	70.7
35990	72.9	65.3
36000	71.1	62.8

Table 1. Data for PCA example

The preparation of Figure 4 is one of the first things to be considered, because a scatter diagram for such a limited number of samples would easily point out any outliers or other aberrations in the data as well as indicate the relationship between the two variables.

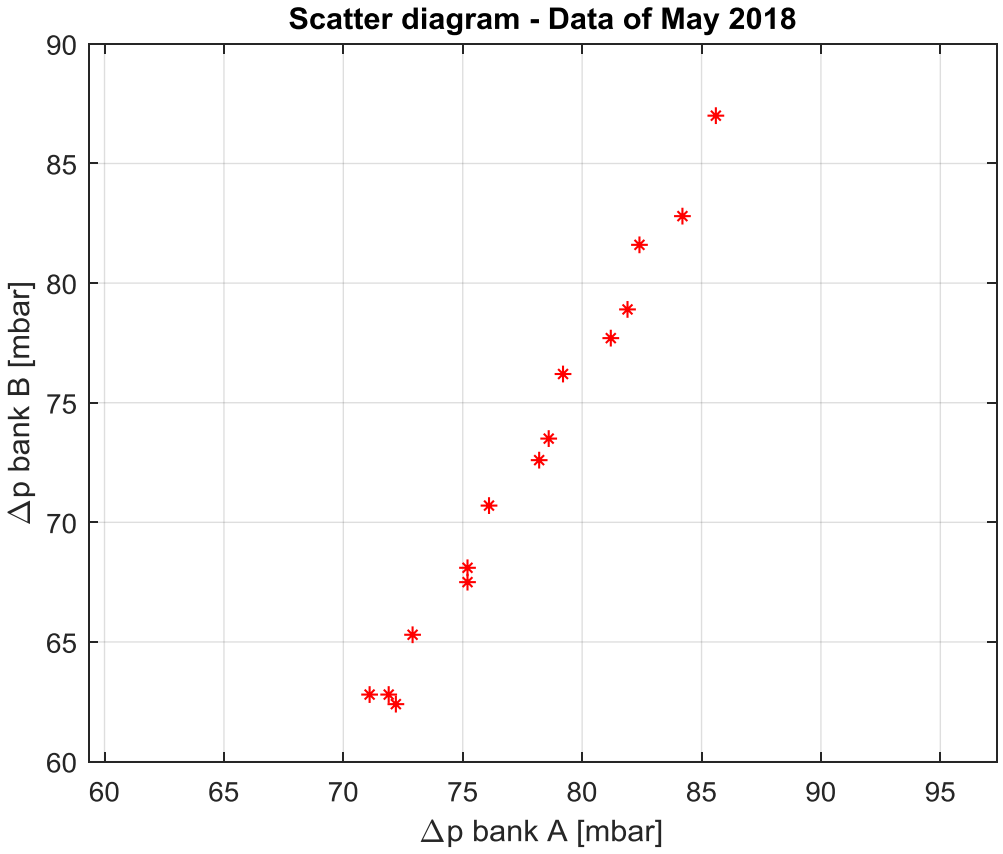


Figure 4. Scatter plot of PCA example data

The compulsory prerequisite for the clarification of PCA method is to determine the sample means, variances and the covariance between the two variables for the data in Table 1. Let x_1 be the variable coolant Δp bank A and the variable coolant Δp bank B be denoted by x_2. The vector of sample means is:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 77.73 \\ 72.66 \end{bmatrix}$$

and the sample covariance matrix is:

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{bmatrix} = \begin{bmatrix} 21.82 & 36.61 \\ 36.61 & 62.22 \end{bmatrix}$$

where the main diagonal elements of **S** are the variances and the off-diagonal elements are the covariances. Although the correlation coefficient between x_1 and x_2 is not required, it is often convenient to employ it, rather than the covariance. The correlation coefficient is defined as:

$$r_{12} = \frac{s_{12}}{s_1 s_2}$$

where s_1 and s_2 are the standard deviations of the two variables (i.e. just the square roots of the respective variances). The correlation coefficient r lies in the range [-1,1]. A vanishing covariance, viz. r = 0, is only a necessary condition for independence, but it is not a sufficient condition.

The principal components procedure takes advantage of a noteworthy result from matrix algebra. A p x p symmetric, non-singular matrix, like the covariance matrix **S**, may be reduced to a diagonal matrix **L** by premultiplying and postmultiplying it by a particular matrix **U** such that:

$$\mathbf{U'SU} = \mathbf{L} \tag{14}$$

The diagonal elements of **L**, l_1, l_2,..., l_p are called the characteristic roots or eigenvalues of **S**. Matrix **U** is orthonormal (i.e. both orthogonal and normalized) and its columns, $\mathbf{u}_1$, $\mathbf{u}_2$,..., $\mathbf{u}_p$ are termed the characteristic vectors or eigenvectors of **S**. The eigenvalues are obtained from the solution of the following characteristic equation:

$$|\mathbf{S} - l\mathbf{I}| = 0 \tag{15}$$

where **I** is the identity or unit matrix. More specifically, **I** is a square matrix that has ones on the diagonal and zeros elsewhere. The equation (15) produces a *p*th degree polynomial in l from which the values l_1, l_2,..., l_p are calculated. In this instance there are p = 2 variables and hence

$$|\mathbf{S} - l\mathbf{I}| = \begin{vmatrix} 21.82 - l & 36.61 \\ 36.61 & 62.22 - l \end{vmatrix} = l^2 - 84.0418l + 17.3310 = 0$$

The values that satisfy this equation are $l_1 = 83.84$ and $l_2 = 0.21$. The characteristic vectors are given by the solution of the equations

$$[\mathbf{S} - l_i\mathbf{I}]\mathbf{t}_i = 0 \tag{16}$$

and

$$\mathbf{u}_i = \frac{\mathbf{t}_i}{\sqrt{\mathbf{t}_i'\mathbf{t}_i}} \tag{17}$$

for i = 1, 2,..., p. For the proposed example, i = 1 yields:

$$[\mathbf{S} - l_1\mathbf{I}]\mathbf{t}_1 = \begin{bmatrix} 21.82 - 83.84 & 36.61 \\ 36.61 & 62.22 - 83.84 \end{bmatrix}\begin{bmatrix} t_{11} \\ t_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

These are two homogeneous linear equations in two unknowns. Let $t_{11} = 1$ and merely use the first equation:

$$36.61t_{21} - 62.01 = 0$$

The solution is $t_{21} = 1.6939$. These results are then placed in the normalizing equation (17) to obtain the characteristic vector $\mathbf{u}_1$:

$$\mathbf{u}_1 = \frac{\mathbf{t}_1}{\sqrt{\mathbf{t}_1'\mathbf{t}_1}} = \frac{1}{\sqrt{3.8692}}\begin{bmatrix} 1.0 \\ 1.6939 \end{bmatrix} = \begin{bmatrix} 0.5084 \\ 0.8611 \end{bmatrix}$$

In similar fashion, using $l_2 = 0.21$ and letting $t_{22} = 1$, the second characteristic vector $\mathbf{u}_2$ becomes:

$$\mathbf{u}_2 = \frac{\mathbf{t}_2}{\sqrt{\mathbf{t}_2'\mathbf{t}_2}} = \frac{1}{\sqrt{3.8692}}\begin{bmatrix} -1.6939 \\ 1.0 \end{bmatrix} = \begin{bmatrix} -0.8611 \\ 0.5084 \end{bmatrix}$$

These characteristic vectors form the matrix

$$\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2] = \begin{bmatrix} 0.5084 & -0.8611 \\ 0.8611 & 0.5084 \end{bmatrix}$$

U is orthonormal, that is

$$\mathbf{u}_1'\mathbf{u}_1 = 1; \quad \mathbf{u}_2'\mathbf{u}_2 = 1; \quad \mathbf{u}_1'\mathbf{u}_2 = 0$$

In general terms, an orthonormal p x p matrix is a square matrix with the following properties:

1) $|\mathbf{A}| = \pm 1$, where $|\mathbf{A}|$ is the determinant of **A**;
2) $\sum_{i=1}^{p} a_{ij}^2 = \sum_{j=1}^{p} a_{ij}^2 = 1$ for all i, j and this implies that the sum of squares of any column or row is equal to unity;
3) $\sum_{i=1}^{p} a_{ij}a_{ik} = 0$ for all $j \neq k$, to wit, the sum of crossproducts of any two columns is zero and implicates that the coordinate axes, represented by these two columns, intersect at an angle of 90°.

This signifies that $\mathbf{A}\mathbf{A}' = \mathbf{I}$; consequently, for an orthonormal matrix **A**, $\mathbf{A}' = \mathbf{A}^{-1}$, where $\mathbf{A}^{-1}$ is the inverse of **A**.

Furthermore

$$\mathbf{U'SU} = \begin{bmatrix} 0.5084 & 0.8611 \\ -0.8611 & 0.5084 \end{bmatrix} \begin{bmatrix} 21.82 & 36.61 \\ 36.61 & 62.22 \end{bmatrix} \begin{bmatrix} 0.5084 & -0.8611 \\ 0.8611 & 0.5084 \end{bmatrix} = \begin{bmatrix} 83.84 & 0 \\ 0 & 0.21 \end{bmatrix} = \mathbf{L}$$

verifying equation (14).

From a geometrical point of view, the routine just described can be interpreted as a principal axis rotation of the original coordinate axes x_1 and x_2 about their means. The elements of the characteristic vectors are nothing more than the direction cosines of the new axes related to the old. The Figure 5 highlights the angle θ_{11} between the x_1-axis and the first new axis and the angle between this new axis and the x_2-axis, θ_{21}; u_{11} and u_{21} are the cosines of θ_{11} and θ_{21}, respectively.

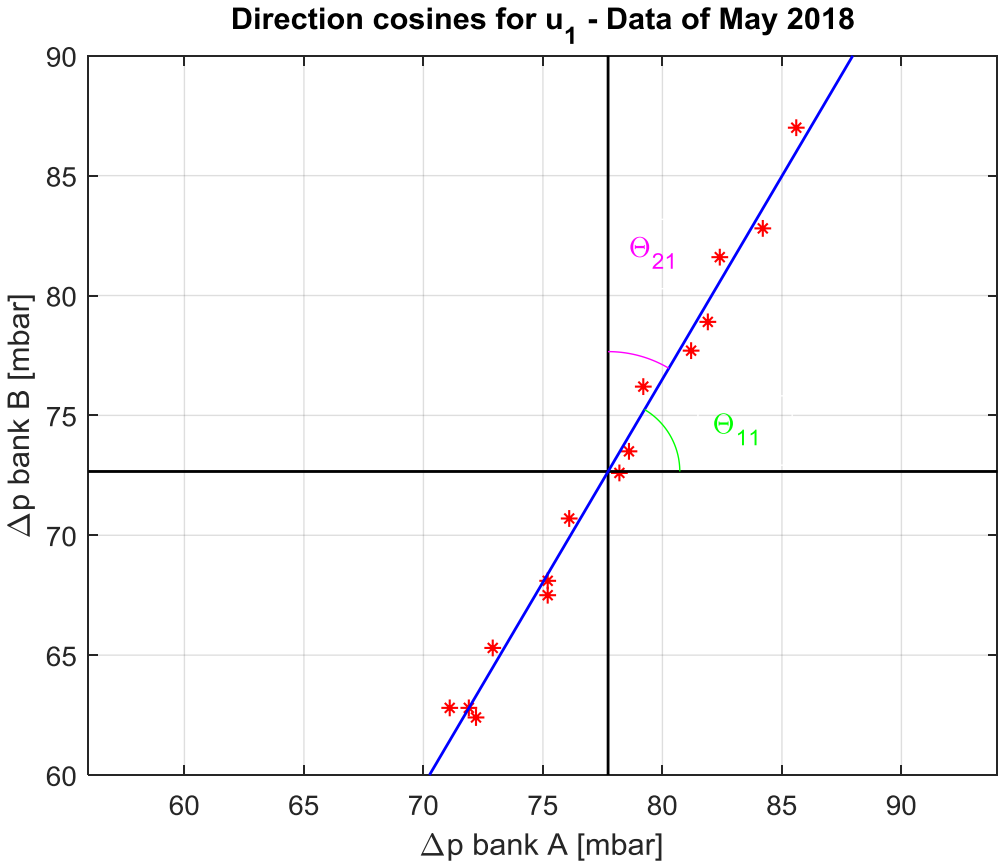


Figure 5. Direction cosines for u_1 of PCA example data

Figure 6 contains the same relationships for $\mathbf{u}_2$, which locates the second new axis. In particular, $\theta_{11} = \theta_{22} = 59.44°$, $\theta_{21} = 90° - \theta_{11} = 30.56°$ and $\theta_{12} = 90° + \theta_{11} = 149.44°$. Except for $p = 2$ or $p = 3$, equation (15) is not used in practice as the resulting equations become unwieldy. Iterative and numerical procedures, such as the power method, are available for obtaining both the characteristic roots and vectors. Most computer packages have currently substituted the power method with more efficient techniques; MATLAB library function pca, widely exploited in Chapter 2, utilizes, by default, the Singular Value Decomposition (SVD) algorithm.

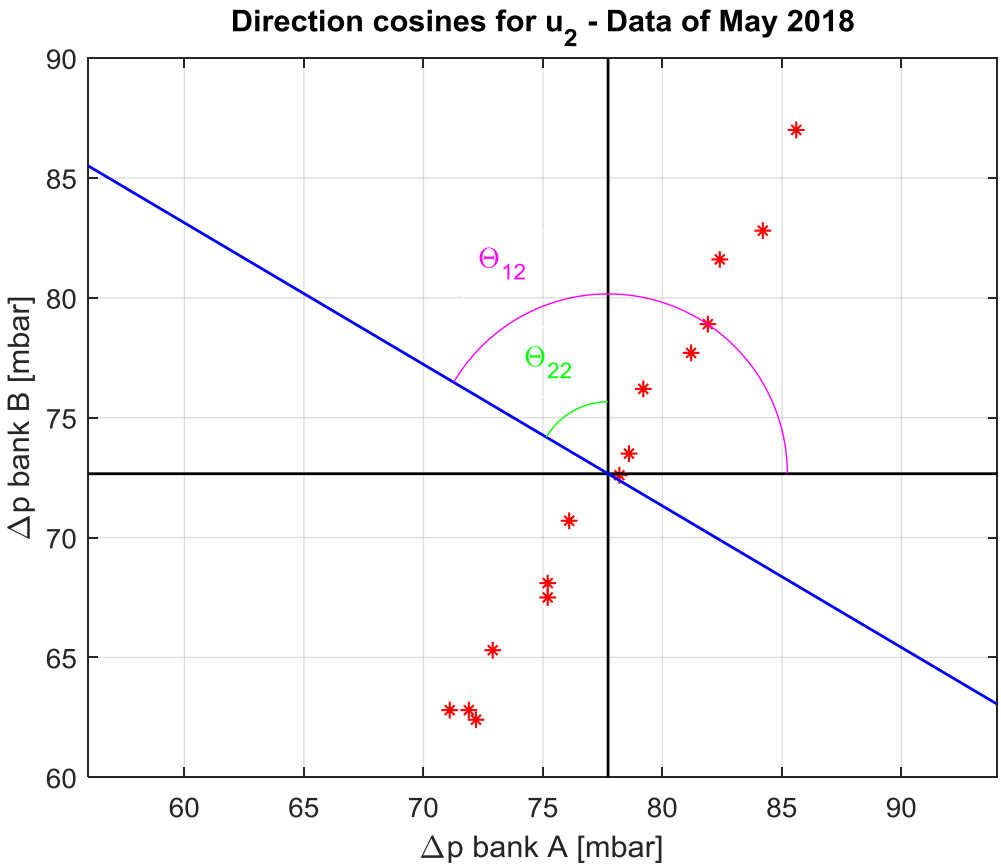


Figure 6. Direction cosines for u_2 of PCA example data

The previously attained principal axis transformation enables to transform p correlated variables x_1, x_2,..., x_p into p new uncorrelated variables z_1, z_2,..., z_p. The coordinate axes of the aforesaid new variables are identified by the characteristic vectors $\mathbf{u}_i$ which constitute the matrix $\mathbf{U}$:

$$\mathbf{z} = \mathbf{U}'[\mathbf{x} - \bar{\mathbf{x}}] \tag{18}$$

The term in square brackets in (18) is the difference between the p x 1 vectors of observations on the original variables and their corresponding means. In the present example, $[\mathbf{x} - \bar{\mathbf{x}}]$ is a matrix with 2 rows and 15 columns. The transformed variables are named the principal components of $\mathbf{x}$ or PCs for short. The *i*th principal component is

$$\mathbf{z}_i = \mathbf{u}_i'[\mathbf{x} - \bar{\mathbf{x}}] \tag{19}$$

and has mean equal to zero and variance equal to l_i, the *i*th characteristic root. To distinguish between the transformed variables and the individual transformed observations, the latter ones are called z-scores. Substituting only the first two couples of example data in (18) produces

$$\mathbf{z} = \begin{bmatrix} 0.5084 & 0.8611 \\ -0.8611 & 0.5084 \end{bmatrix} \begin{bmatrix} 81.9 - 77.73 & 78.6 - 77.73 \\ 78.9 - 72.66 & 73.5 - 72.66 \end{bmatrix} = \begin{bmatrix} 7.495 & 1.167 \\ -0.421 & -0.325 \end{bmatrix}$$

so the z-scores for the first observation are $z_{1,1}$ = 7.495 and $z_{2,1}$ = -0.421. The variance of z_1 is equivalent to l_1 = 83.84, while the variance of z_2 is equal to l_2 = 0.21. Table 2 includes, for the considered observations, the deviations from their means and the principal components along with a quantity that will be properly introduced later [12].

Minutes of the month	$x_1 - \bar{x}_1$ [mbar]	$x_2 - \bar{x}_2$ [mbar]	z_1	z_2	T^2
35860	4.2	6.2	7.495	-0.421	1.529
35870	0.9	0.8	1.167	-0.325	0.527
35880	-2.5	-4.6	-5.211	-0.142	0.422
35890	-5.8	-9.9	-11.453	0.005	1.565
35900	-5.5	-10.3	-11.645	-0.457	2.627
35910	-2.5	-5.2	-5.728	-0.447	1.360
35920	0.5	-0.1	0.189	-0.438	0.929
35930	3.5	5.0	6.106	-0.429	1.334
35940	6.5	10.1	12.023	-0.419	2.575
35950	7.9	14.3	16.351	0.510	4.448
35960	4.7	8.9	10.074	0.521	2.522
35970	1.5	3.5	3.797	0.531	1.536
35980	-1.6	-2.0	-2.515	0.404	0.866
35990	-4.8	-7.4	-8.792	0.415	1.754
36000	-6.6	-9.9	-11.860	0.694	4.006

Table 2. Results of PCA illustrative example

The concept of principal components is graphically illustrated in Figure 7. The first principal component z_1 accounts for most of the variability in the two initial variables x_1 and x_2. The information embedded in the whole set of all p PCs is perfectly equivalent to the information in the complete set of all original process variables. More generally, the basic intent of principal components is to find a new set of orthogonal directions that identify the maximum variability in the original data; this leads to a database representation which requires fewer than the initial p variables. Therefore, if a set of p (> 2) variables has substantial correlations among them, then the first few k (< p) PCs will hopefully provide a satisfactory description of the process [10] [13]. The larger k is, of course, the better the fit of the PCA model; conversely, the smaller k is, the more simple the model will be [12]. There are different criteria to choose the optimum number of retained PCs. The eigenvalues associated with each principal component reveal how much information, viz. variation, each PC explains. One can look at the cumulative percent variance captured by the first few PCs and select a number of PCs that accounts for a target percentage of the variability of data [14]. The proportion of the total variability in the original data explained by the *i*th principal component is given by the ratio

$$\lambda_{\%i} = \frac{l_i}{l_1 + l_2 + \cdots + l_p} * 100 \tag{20}$$

Thus, one can easily assess how much variability is explained by keeping just a few (k) of the p principal components by computing the sum of the eigenvalues for those k components and

comparing that amount to the sum of all p eigenvalues [10]. This is the criterion adopted in the current study. For the treated example, the (20) yields $\lambda_{\%1} = 99.75\%$ and $\lambda_{\%2} = 0.25\%$.

Another common technique is to accept the PCs whose eigenvalues exceed the average characteristic root. The rationale here is that any deleted PC has a root smaller than the average.

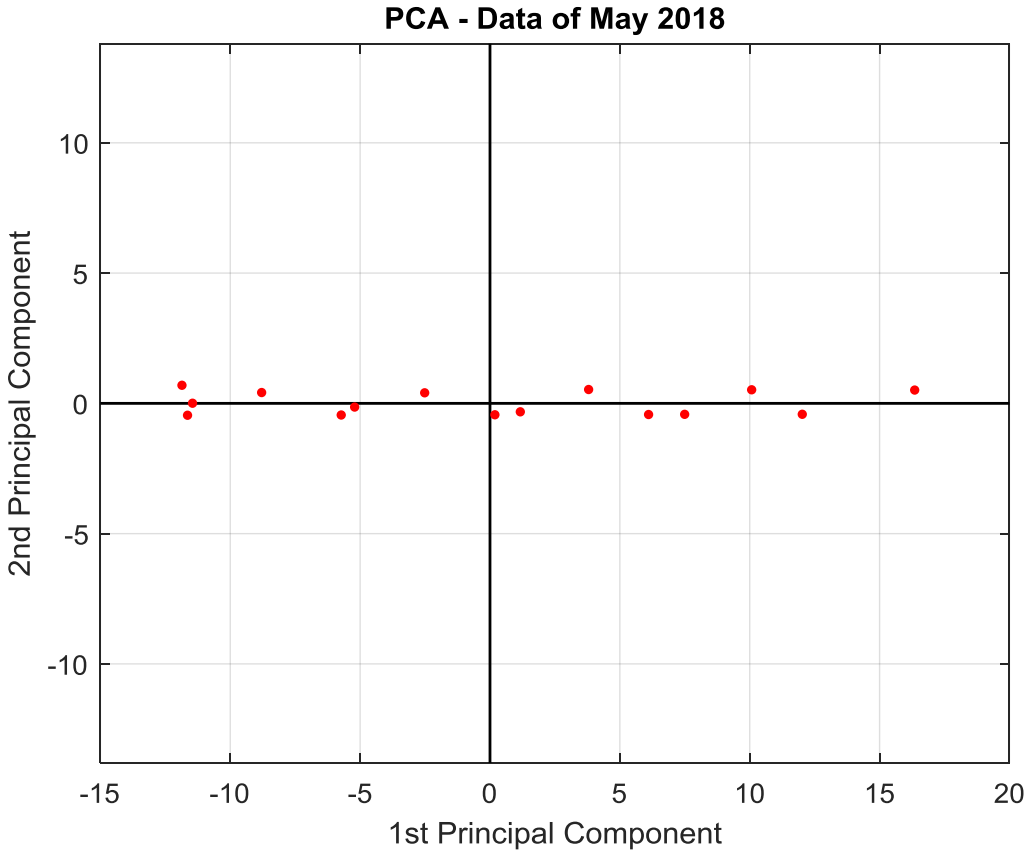


Figure 7. Plot of the example observations with respect to their PCs

The quantity shown in Figure 8 indicates the overall conformance of an individual observation vector to its mean or an established standard. This quantity, due to Hotelling, is a multivariate generalization of the Student t-test and it is used to determine if the process is in control. The original form of $\mathbf{T}^2$ is:

$$\mathbf{T}^2 = [\mathbf{x} - \bar{\mathbf{x}}]' \mathbf{S}^{-1} [\mathbf{x} - \bar{\mathbf{x}}] \tag{21}$$

The equation (21) does not involve PCA and is a statistic often employed in multivariate quality control. In the context of this research, Hotelling T^2 control chart is exploited to highlight the presence of outliers that might be helpful to predict an impending failure event. The above-mentioned chart is a direct analogue of the Shewhart $\overline{X}$ chart and it needs an upper control limit, which is:

$$T^2_{p,n,\alpha} = \frac{p(n-1)}{n-p} F_{p,n-p,\alpha} \tag{22}$$

where n is the number of observation vectors and $F_{p,n-p,\alpha}$ is the F-distribution, connected with a level of significance α. The suitable value of $F_{p,n-p,\alpha}$ can be found in tabular form or with the aid of MATLAB finv command, which returns the inverse of the F cumulative distribution function. Obviously, Hotelling T^2 control chart has only an upper control limit because T^2 is a squared quantity. For the same reason, the ordinate scale is sometimes logarithmic. In the example examined in this chapter, p = 2, n = 15, $F_{2,13,0.05} = 3.8056$, so $T^2_{2,15,0.05} = 8.1966$. The last column of Table 2 reports the results of (21), represented in Figure 8 by green dots: there are no points out of control on the chart shown.

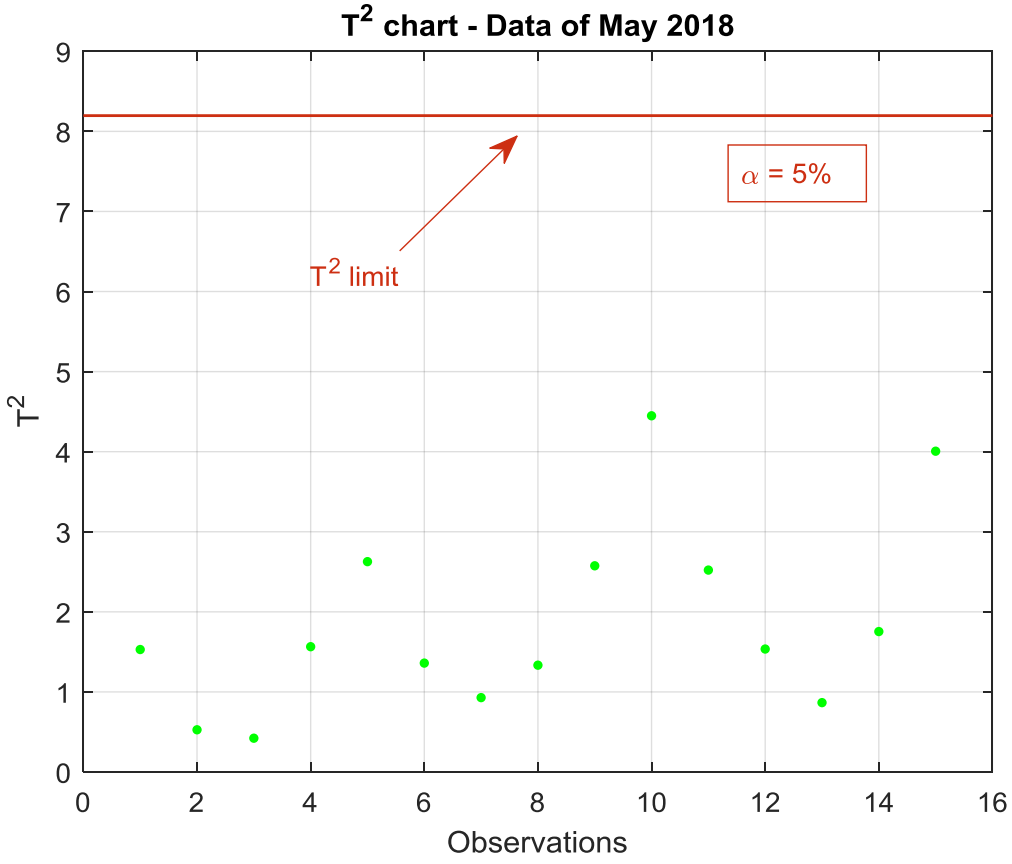


Figure 8. Hotelling T^2 control chart for PCA example data

A further interesting property of PCA is the fact that equation (18) may be inverted so that the original parameters can be stated as a function of the full set of principal components:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{U}\mathbf{z} \tag{23}$$

This implies that, given the z-scores, the values of the original variables can be uniquely determined. However, **x** will be perfectly recreated only if all the PCs are used. In cases where k < p PCs are retained, a simple estimate $\hat{\mathbf{x}}$ of **x** will be produced. The Q-statistic is based on

the sum of squares of the distance of $\mathbf{x} - \hat{\mathbf{x}}$ from the k-dimensional space that the PCA model defines [12]. Hotelling T^2 statistic is a measure of the variation within the PCA model, while the Q-statistic measures the lack of model fit for each sample. The latter one is evaluated as the difference between the data point and its projection on the PC model [14]. It has been the experience of several practitioners, particularly but not restricted to the physical sciences and engineering, that the greatest advantage of the Q-statistic is its ability to detect bad data, measurement errors, and the like [12]. A similar task pertains to data pre-processing and has already been carried out at this stage. In light of such considerations, Hotelling T^2 method was chosen, since it does not seem strictly related to the specific subset of retained PCs and it is intrinsically independent of the PCA.

Chapter 2 – Case study

The case study section explains the pathway to the transformation of a raw database, coming from a real cogeneration plant, into a data set which is ready for a statistical investigation; this transformation was possible thanks to a MATLAB code, built from scratch by the author. Details about the script are provided, in order to satisfy a curious reader. Then, the selection process of parameters subsets is delineated and the Principal Component Analysis is applied to the data pertaining to three months of 2017 plus May 2018. Finally, the focus of the statistical research concentrates on a larger group of variables, with a three-dimensional representation.

The goal of the present part is the disclosure of the most relevant variables involved in the study and of the modalities that allowed the construction of a "tidy" and organized database. Moreover, the most appreciable results deriving from the synergy between PCA and Hotelling T^2 statistic are included.

Data pre-processing

The methodology subsection contained in the Introduction touches upon FMEA key role in identifying failure modes of components. FMEA also suggests to measure some parameters because they potentially indicate the presence of faults. The aforesaid systematic approach provides information on the range of parameters that need to be measured for particular failure modes [15].

Failure Modes and Effects Analysis applied to the endothermic engine recommended plant manager to monitor all the 37 variables reported in Table 3. The large raw database coming from the field covers a period of 13 months, namely the 2017 calendar year plus May 2018. The latest month will turn out to be useful in the last part of this thesis. As shown in the first column of the table, each variable can be associated with a component, whose acronym is clarified at the bottom of the table (see legend to Table 3). Unfortunately, not all the sensors measurements were correctly recorded. For example, data sets on fuel oil flow rate at engine inlet were only available for November 2017, December 2017 and May 2018. Engine operating load is at the top of variables list because it is essential for the definition of ICE operating condition, which is the proper initial step of data pre-processing. The lack of engine operating load data set, of course, would have made worthless a full month of data. ICE database consisted of more than 20.5 millions of numbers, split into roughly 150 comma separated value files, CSV, on a monthly basis. Comma separated value files are a plain text format utilized for storing data in a tabular structure; the CSV format is popular, largely because of its versatility and because it can be loaded as a spreadsheet in software packages such as Microsoft Excel. Each file began with a header row, usually reporting sensors tags with each field separated by a comma. Observations were recorded in contiguous rows and fields were once again separated by commas. Every variable was registered in a distinct column. The total amount of numbers comprising the ICE database was determined with the help of Excel COUNTA function, a function that counts the quantity of cells that are not

empty in a specified range. The estimate is conservative, for sure, since it does not take into consideration indispensable pieces of information: timestamps linked with signals.

		A = Available data set; N/A = Not Available data set												
		Year												
		2017												2018
Component	Parameter	January	February	March	April	May	June	July	August	September	October	November	December	May
G	*Engine operating load*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*
FC	Fuel oil flow rate at engine inlet	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	A	A	A
E	Temperatures of cylinders A1, A2, A3, A4, A5 exhaust gases	A	A	A	A	A	A	A	A	A	A	A	A	A
E	Temperatures of cylinders A6, A7, A8, A9, A10 exhaust gases	A	A	A	A	A	A	A	A	A	A	A	A	A
E	Temperatures of cylinders B1, B2, B3, B4, B5 exhaust gases	A	A	A	A	A	A	A	A	A	A	A	A	A
E	Temperatures of cylinders B6, B7, B8, B9, B10 exhaust gases	A	A	A	A	A	A	A	A	A	A	A	A	A
E	Coolant Δp in Charge-Air Cooler (banks A and B)	A	A	A	A	A	A	A	N/A	A	A	A	A	A
TC	Environmental temperature and humidity	A	A	A	A	A	A	A	A	A	A	A	A	A

TC	Revolutions per minute (banks A and B)	A	A	A	A	A	A	A	A	A	A	A	A	A
TC	Inlet temperatures (banks A and B)	A	A	A	A	A	A	A	A	A	A	A	A	A
TC	Outlet temperatures (banks A and B)	A	A	A	A	A	A	A	A	A	A	A	A	A
SCR	Backpressure at SCR system inlet	A	A	A	A	A	A	A	A	A	A	A	A	A
SCR	Differential pressure drop in SCR system	A	A	A	A	A	A	A	A	A	A	A	A	A
SCR	Specific consumption of NH_3	A	A	A	A	A	A	A	A	A	A	A	A	A
SCR	Inlet and outlet temperatures of exhaust fumes	A	A	A	A	A	A	A	A	A	A	A	A	A
Legend	G	Generator												
	FC	Fuel Circuit												
	E	Engine												
	TC	Turbocharger												
	SCR	Selective Catalytic Reduction System												

Table 3. Parameters included in the work and their availability during the 13 months

Usual operating pressures, temperatures and mass flow rates of a 10 MW ICE are displayed in Figure 9. Modern combustion engines are equipped with turbochargers, with the aim of raising the output and improving the efficiency.

The database brought along some critical issues, joined to its real field origin and outlined in the current paragraph. The first problem occurs when a measuring device was unable to record an observation; this translates into void cells, like described in Table 4. A similar situation is given by the complete absence of data for a considerable portion of a day or for whole days. For instance, Table 5 outlines the dearth of data for exhaust gases temperatures of cylinders B1, B2, B3, B4 and B5, in the lapse between 18:29:28 of 22nd June 2017 and 17:46:26 of 26th June 2017.

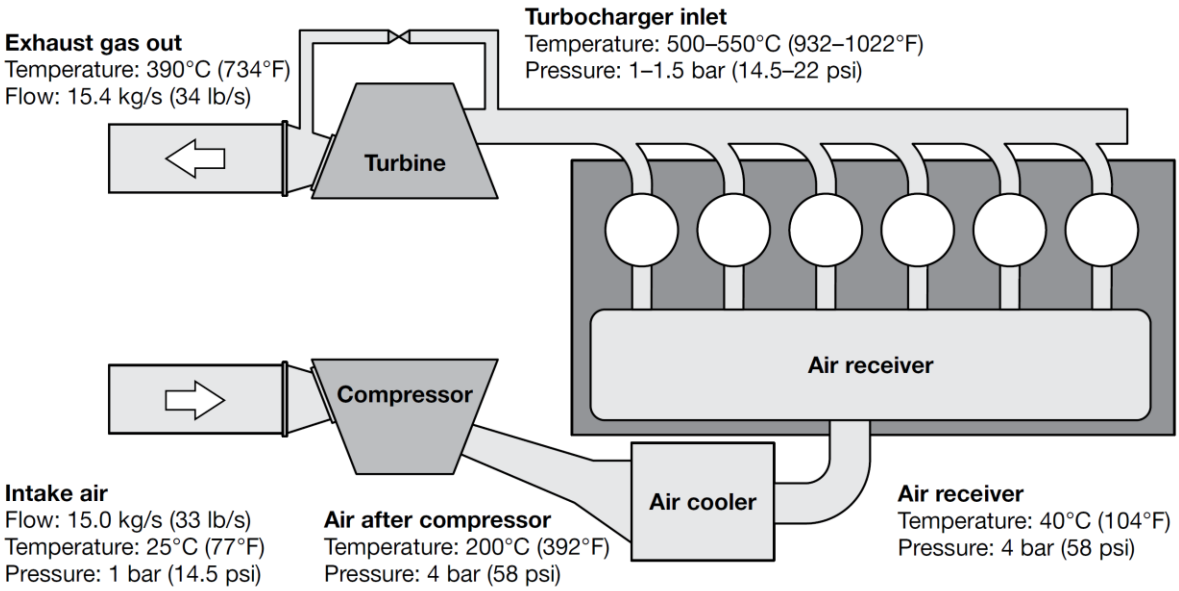


Figure 9. Typical temperatures, pressures and mass flows from a 10 MW combustion engine [16]

Date	Time	Temperature of cylinder A6 exhaust gases [°C]	Temperature of cylinder A7 exhaust gases [°C]	Temperature of cylinder A8 exhaust gases [°C]	Temperature of cylinder A9 exhaust gases [°C]	Temperature of cylinder A10 exhaust gases [°C]
06/01/2017	11:18:16	398	413	431	411	425
06/01/2017	11:19:00	398	413	431	411	425
06/01/2017	11:19:44					
06/01/2017	11:20:28					
06/01/2017	11:21:12					
06/01/2017	11:21:56	399	414	431	412	426
06/01/2017	11:22:40	399	414	431	412	426

Table 4. Short sequence of missing data

Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]	Temperature of cylinder B3 exhaust gases [°C]	Temperature of cylinder B4 exhaust gases [°C]	Temperature of cylinder B5 exhaust gases [°C]
22/06/2017	18:28:00	483	441	473	476	482
22/06/2017	18:28:44	483	441	473	476	482
22/06/2017	18:29:28	483	441	473	476	482
26/06/2017	17:46:26	487	445	477	493	485
26/06/2017	17:47:07	487	445	477	493	485
26/06/2017	17:47:48	486	445	476	493	484

Table 5. Unavailability of data for several days

Table 6 highlights another issue, already recognizable in the preceding table. Some parameters were irregularly sampled, although they are part of the same month. The example reveals that at the beginning of July data capture was executed with a frequency of one measure per 41 seconds, while towards the end of July the time span between two values is 26 seconds. Moreover, this non-uniformity in data sampling also appears for larger amount of time elapsed for two consecutive signals, as shown in Table 7.

Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]	Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]
02/07/2017	04:27:00	469	429	25/07/2017	05:42:00	471	428
02/07/2017	04:27:41	469	429	25/07/2017	05:42:26	471	428
02/07/2017	04:28:22	469	429	25/07/2017	05:42:52	471	428
02/07/2017	04:29:03	469	429	25/07/2017	05:43:18	471	428
02/07/2017	04:29:44	469	429	25/07/2017	05:43:44	471	428
02/07/2017	04:30:25	469	429	25/07/2017	05:44:10	471	428

Table 6. Irregular signals sampling for short time span

Date	Time	Backpressure at SCR system inlet [mbar]	Date	Time	Backpressure at SCR system inlet [mbar]
14/02/2017	00:49:43	42.5	12/10/2017	14:57:54	27.5
14/02/2017	01:49:43	33.8	12/10/2017	15:02:54	28.8
14/02/2017	02:49:43	36.2	12/10/2017	15:07:54	30.0
14/02/2017	03:49:43	40.2	12/10/2017	15:12:54	31.2
14/02/2017	04:49:43	30.7	12/10/2017	15:17:54	32.4
14/02/2017	05:49:43	39.9	12/10/2017	15:22:54	33.6

Table 7. Uneven data collection for long time span

Data acquisition was sometimes not synchronized, with variables recorded starting from disparate instants of the same month. The case just mentioned is typified by Table 8, where parameters belonging to the same nature, such as temperatures of cylinders exhaust gases, were registered beginning from different days and times of January 2017.

Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]	Date	Time	Temperature of cylinder B6 exhaust gases [°C]	Temperature of cylinder B7 exhaust gases [°C]
03/01/2017	17:29:32	445	404	05/01/2017	16:31:28	409	433
03/01/2017	17:30:16	445	404	05/01/2017	16:32:12	410	434
03/01/2017	17:31:00	445	404	05/01/2017	16:32:56	410	434
03/01/2017	17:31:44	445	404	05/01/2017	16:33:40	410	434
03/01/2017	17:32:28	446	405	05/01/2017	16:34:24	410	434
03/01/2017	17:33:12	446	405	05/01/2017	16:35:08	410	434

Table 8. Temporal misalignment of data sampling

An additional peculiarity found in the database is the presence of redundancy of data; a CSV file occasionally held data pertaining to the prior or to the following month. As illustrated by Table 9, extra data share an almost equal timestamp but there is no overlap in acquisition time. MATLAB code managed this kind of supplementary information appropriately. A further obstacle to address, preparatory to data pre-processing, was the incorrect structure of recorded values in some CSV files. Those files contained measurements in the configuration "integer part – comma – decimal digit – comma". Thus, comma character was employed in delimiting fields and in dividing the integer parts from the fractional parts. The concurrence of the above-stated wrong format and negative numbers starting with a zero (Table 10) gave rise to a thorny issue.

Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]	Date	Time	Temperature of cylinder B1 exhaust gases [°C]	Temperature of cylinder B2 exhaust gases [°C]
02/02/2017	17:25:48	456	415	02/02/2017	17:25:40	456	415
02/02/2017	17:26:32	456	415	02/02/2017	17:26:24	456	415
02/02/2017	17:27:16	456	415	02/02/2017	17:27:08	456	415
02/02/2017	17:28:00	456	415	02/02/2017	17:27:52	456	415
02/02/2017	17:28:44	455	414	02/02/2017	17:28:36	455	414
02/02/2017	17:29:28	455	414	02/02/2017	17:29:20	455	414

Table 9. Redundancy of data in CSV files

Date-Time,Backpressure at SCR system inlet,Differential pressure drop in SCR system,Specific consumption of NH_3,Exhaust fumes temperature at outlet,Exhaust fumes temperature at inlet
22-05-2017 11:49:43,-0,9,0,5,0,0,30,6,27,5,
22-05-2017 12:49:43,-0,8,0,8,0,0,31,2,27,7,
22-05-2017 13:49:43,-0,7,1,0,0,0,31,8,27,9,
22-05-2017 14:49:43,27,2,1,9,130,4,319,1,76,3,
22-05-2017 15:49:43,19,8,5,4,97,8,248,5,143,8,
22-05-2017 16:49:43,12,3,8,9,65,2,177,9,211,4,

Table 10. Anomaly in CSV file format

Splitting blindly comma separated values into different columns would have led to an automatic conversion of "-0" into "0", since Excel is not capable of distinguishing cells with "-0" from cells with "0", unless these data are stored like words. A workaround consists in selecting "Text" for column data format, while exploiting the Text to Columns Excel function. It was then necessary to use Excel CONCATENATE function that joins two or more text strings into one string. Specifically, three types of cell were joined: the cell with measurement integer part, a cell with a dot and the cell holding observation fractional part. The cells resulting from the described operation were ultimately formatted as numbers.

Last but not least, there are outliers inside the database; outliers are observations that differ considerably from the bulk of the data and Table 11 provides an example. It seems highly improbable that fuel oil flow rate drops to zero in few minutes before jumping suddenly to 1856 liters per hour. Additional examples of outliers or "bad" values can be full scale values.

Date	Time	Fuel oil flow rate at engine inlet [L/h]
31/05/2018	23:46:00	1855
31/05/2018	23:46:44	1854
31/05/2018	23:47:28	
31/05/2018	23:48:12	
31/05/2018	23:48:56	0
31/05/2018	23:49:40	1856

Table 11. Outlier in a data set

All these difficulties arisen from the case study contributed to the need of a model which rationalizes and validates available operational data. The problem was handled by a MATLAB code, whose details are now explained. An intuition of paramount importance allowed the generation of an artificial seconds counter, equally applied to all observations. This temporal reference system works on a monthly basis, like the entire data pre-processing, and has its origin at time 00:00:00 of the first day of the month under analysis. For instance, 00:00:59 of 1st March 2017 corresponds to 60 seconds. Time scale construction was carried out in every Excel worksheet that collected raw data pertaining to a precise month. The implementation of this idea demands little effort, namely the creation of two new cells for each timestamp. One cell reports the number representing the day of the month, obtained making use of a couple of Excel functions, i.e. DATE and YEAR. In the other cell, seconds counter is calculated with the aid of the following Excel functions: HOUR, MINUTE and SECOND. HOUR function returns the hour of a time value; the hour is given as an integer, ranging from 0 to 23. MINUTE and SECOND functions yield the minutes and the seconds of a time value, respectively. Both minutes and seconds are integers in the interval [0,59]. The formula adopted to determine seconds counter (SC) was:

$$SC = (DoM - 1) * s_D + HOUR(tv) * s_H + MINUTE(tv) * s_M + SECOND(tv) + 1 \qquad (24)$$

where DoM stands for day of the month and tv is the cell with time value. The remaining terms of (24) are the amount of seconds in a day s_D, the quantity of seconds per hour s_H and the number of seconds per minute s_M. In order to enhance the effectiveness of MATLAB code exposition, a practical example is provided. It involves the temperature of cylinder A1 exhaust gases, covering the period from the 07:00:08 of 19th January 2017 to the 10:30:52 of 20th January 2017, as summarized in Table 12.

Date	Time	Seconds counter	Engine operating load [kW]	Date	Time	Seconds counter	Temperature of cylinder A1 exhaust gases [°C]
19/01/2017	07:00:08	1580409	6972	19/01/2017	07:00:08	1580409	385
19/01/2017	07:00:52	1580453	6972	19/01/2017	07:00:52	1580453	385
19/01/2017	07:01:36	1580497	6972	19/01/2017	07:01:36	1580497	385
19/01/2017	07:02:20	1580541	6972	19/01/2017	07:02:20	1580541	385
19/01/2017	07:03:04	1580585	6972	19/01/2017	07:03:04	1580585	385
19/01/2017	07:03:48	1580629	7889	19/01/2017	07:03:48	1580629	385
19/01/2017	07:04:32	1580673	8305	19/01/2017	07:04:32	1580673	406
...	...		...	...	...		...
20/01/2017	10:29:24	1679365	0	20/01/2017	10:29:24	1679365	108
20/01/2017	10:30:08	1679409	0	20/01/2017	10:30:08	1679409	108
20/01/2017	10:30:52	1679453	0	20/01/2017	10:30:52	1679453	108

Table 12. Initial and final observations of data pre-processing example

The choice of these temporal limits was focused on checking, confirming and showing the behaviour of proposed model. Maintenance history of internal combustion engine recounts a failure event that concerned fuel oil pump of cylinder A1, with engine shutdown at 09:20 of 20th January 2017. Fuel oil seepage occurred and it entailed a reduced fuel flow rate in

combustion chamber. Temperature of cylinder A1 exhaust gases inevitably decreased before plant stoppage, while the overall engine power output remained constant with small oscillations.

MATLAB script loads solely two data columns per variable from the Excel worksheet associated with a specific month: the column of signals and the column of seconds counter, which compresses date and time of a captured signal into a single number. Some parameters, like the backpressure at SCR system inlet, were sampled every 60 minutes (see Table 7) but the overwhelming majority of data acquisition frequencies was higher than one measure per minute. With the aim of unifying data representation within the same time step, a time scale of one minute was settled on. In accordance with this choice, a special care was required by parameters whose collection timestamps were a multiple of the minute. Each observation was extended for a certain amount of minutes and two options were examined. In case of variables recorded every hour, for example, the first possibility assumes that the acquired value lasts for the successive 59 minutes; the second alternative is founded on the principle that observation holds good for the 30 minutes that precede and the 29 minutes that follow the moment the measurement was captured. The first option prevailed over the other one because it only implies a guess about the upcoming values, i.e. the ensuing 59 minutes. The second idea, conversely, involves a supposition about both the future and the past. The method is graphically illustrated in Figure 10 that pertains to a short part of January 2017.

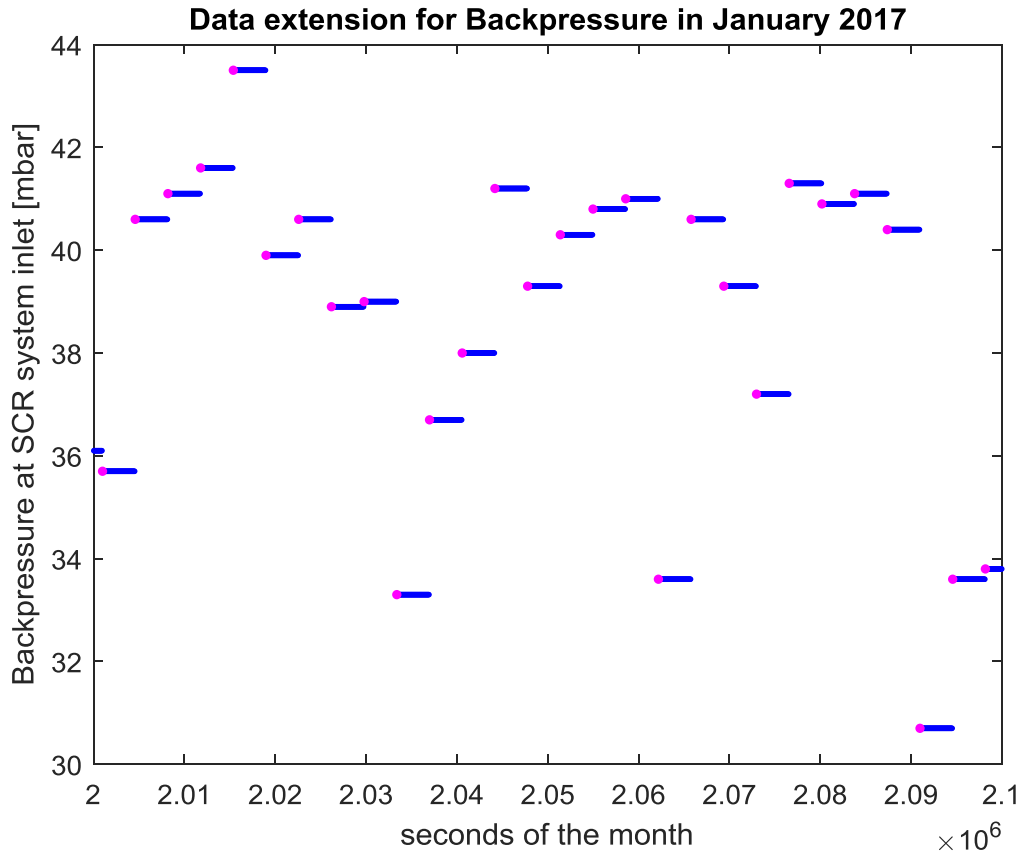


Figure 10. Example of data extension

Magenta dots indicate original backpressure data, whereas each blue line is composed of 59 neighbouring dots, to wit, the added data. The procedure was not repeated on the last observation recorded, owing to the fact that the value might have been obtained close to the end of the month; this would have caused an inexact extension until the early minutes of the succeeding month. The parameter selected for the example which accompanies the text does not require the technique just detailed, as one can infer from Table 12. Figure 11 shows the trend of the temperature of cylinder A1 exhaust gases, prior to data pre-processing. Looking at x axis, the lack of raw data for a sizeable portion of January 2017 is evident. Furthermore, a handful of data points are far from the regular working range.

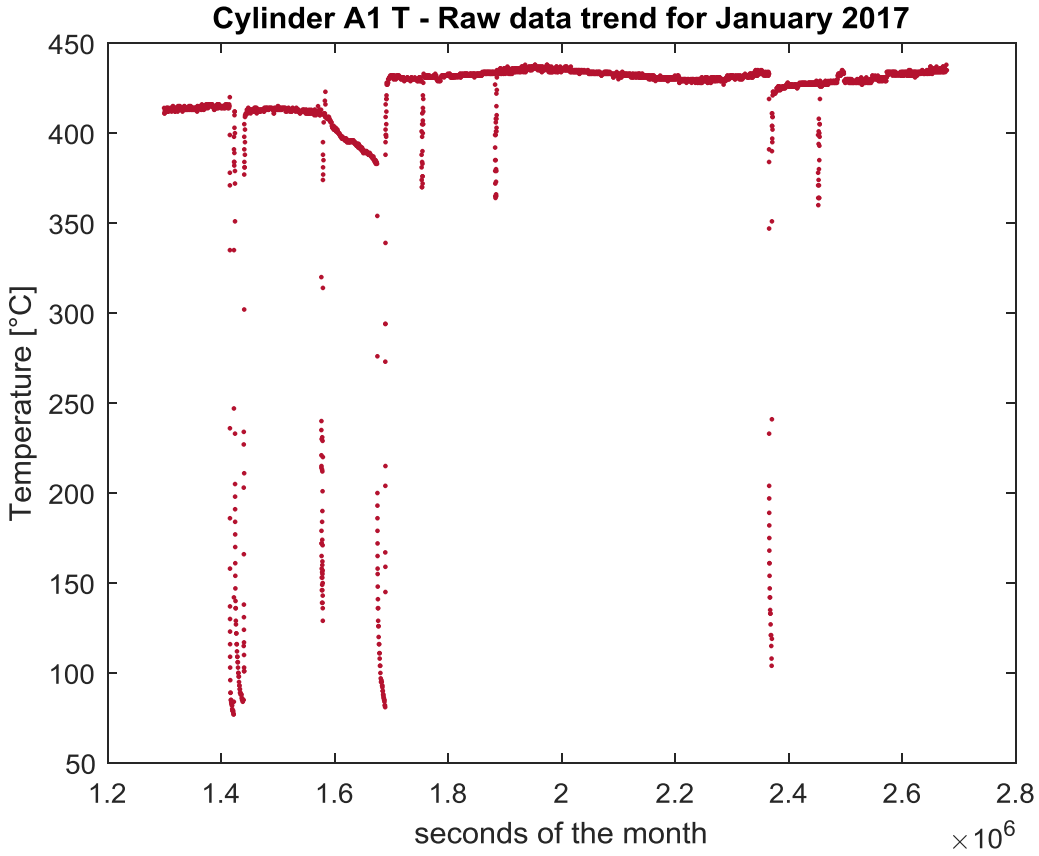


Figure 11. Raw data trend of the temperature of cylinder A1 exhaust gases

One salient issue of the case study that needed to be tackled before the actual data pre-processing was the presence of missing data. When MATLAB reads data columns from Excel, it stores empty cells with NaN, namely Not-a-Number. NaN elements bring trouble during calculations and had to be erased. The first logical step was the substitution of data values equal to 0 with NaN. This was feasible thanks to MATLAB standardizeMissing function and it was performed on the following variables: fuel oil flow rate, operating temperatures and revolutions per minute. An overview of the database induced the elimination of data below or equal to 0 from the parameters related to SCR system, since such values do not have any physical meaning. The last stage was the utilization of MATLAB rmmissing command, which removes NaN values together with their linked seconds counters. At this point, all matrices were exclusively filled with numbers.

The only interesting operating mode for the endothermic engine was the full load working condition. From a conceptual perspective, it is important to reveal that data extension procedure was never applied to the signals reporting engine power, regardless of the month under analysis. This was possible because the frequency of data sampling for engine operating load was higher than one measure per minute, when data were available. In addition, the power was considered to be described in a Boolean way and therefore not subjected to statistical variations. In light of the above, the power is free from constraints imposed by statistical quality control and had the fundamental purpose of defining the full load operating condition. In this regard, a threshold was set at 7800 kW. A matrix with 3 columns was built. Each row of the matrix refers to a specific minute of the month, hence the sum of the rows is equal to the number of minutes in the month. The first column keeps track of the number of minutes elapsed from the beginning of the month, in ascending order. The second column contains the values of power produced by the engine when system was running at full load. The algorithm was designed to check each 60 seconds time span, one by one. If two or more data of power higher than 7800 kW were found, the arithmetic mean of those values was written in the second column; if only one valid value of power was found, that number was directly stored in the second column; in all other cases, the code recorded NaN in the second column. The third column is composed of 0 and 1: 1 denotes a minute with an interesting engine operating mode, otherwise 0 is assigned to the fitting row. Then, a significant quantity was computed, namely the percentage ratio between the number of minutes associated with the full load operating condition and the total number of minutes in the month:

$$\mathrm{FL} = \frac{\text{minutes of the month at full load operating condition}}{\text{minutes in the month}} * 100 \tag{25}$$

Obviously, FL = 100% in an ideal situation, while FL = 49.2% in January 2017. The closing program instructions for power permitted the writing of first two columns of the aforesaid matrix onto a Microsoft Excel spreadsheet file, by virtue of a combination of MATLAB xlswrite and num2cell functions. The latter function allows to convert a numeric array into a cell array with consistently sized cells and without alterations to numeric values.

Data pre-processing that pertain to the other 36 variables is now addressed; from now on, the actions performed and explained for one parameter, the temperature of cylinder A1 exhaust gases, extend to all variables. The knowledge of the minutes characterized by the absence of information about power or by a negligible operating condition granted the possibility of discarding data that fall within those minutes. With this intent, MATLAB code transformed each uninteresting minute into its respective interval of seconds, according to the aforementioned temporal reference system. Afterward, the program scanned all the variable column that retained the seconds counters in order to delete the rows holding both data and seconds that were part of useless time spans. This operation did not require seconds counters (with their respective signals) to be in chronological order, bypassing the obstacle of data redundancy in CSV files (see Table 9). Indeed, before data pre-processing, a supplementary string of data can always be added at the top or at the bottom of raw data columns, as long as it concerns the same parameter and month. Figure 12 shows the rationalization process applied to the observations set of the example. Baby blue dots represent the values of internal

combustion engine power, whereas dark blue horizontal line is the established limit for the operating condition. Magenta circles draw attention to the values of temperature that were rejected. The right side of the graph depicts the behaviour of the model when the engine was shut down.

The algorithm evaluated the percentage ratio between the number of data left after working condition control and the total number of available input data for the actual data pre-processing:

$$\mathrm{OC} = \frac{\text{number of data left after operating condition filter}}{\text{total number of available input data for data pre} - \text{processing}} * 100 \qquad (26)$$

For instance, OC = 95.5% for the temperature of cylinder A1 exhaust gases in January 2017.

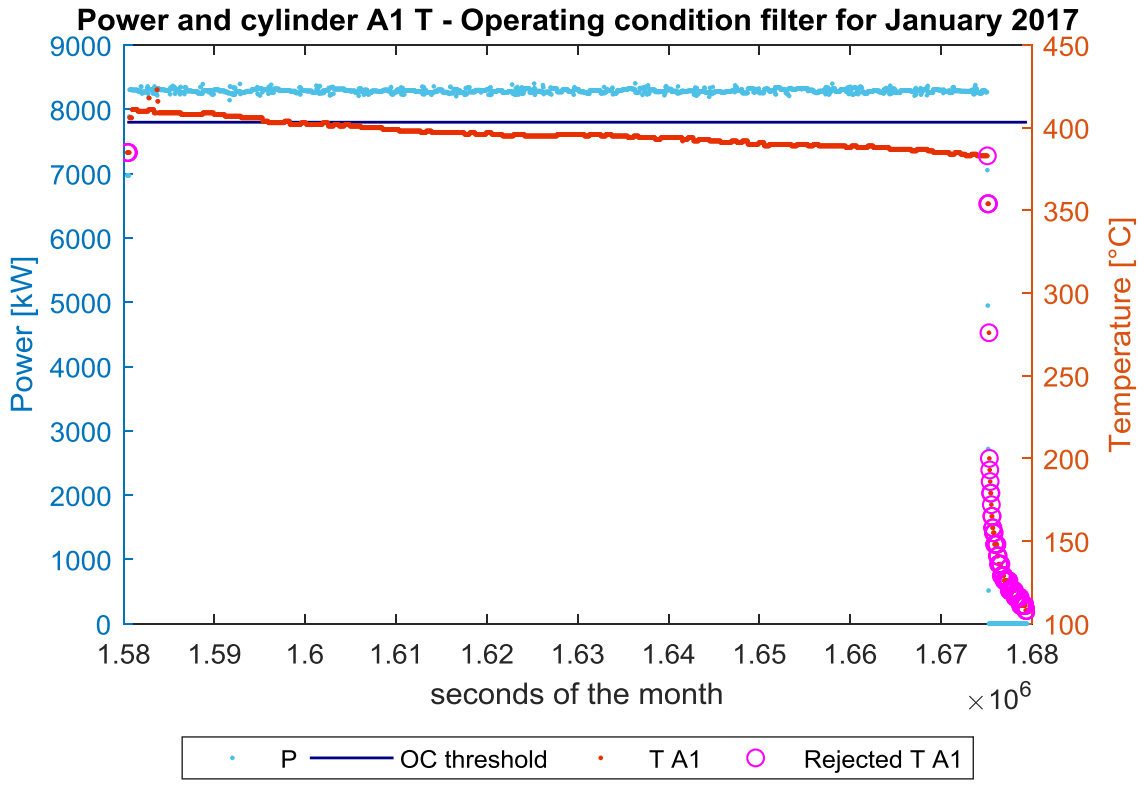


Figure 12. Operating condition control applied to the example data

For the purpose of achieving an accurate statistical quality control, parameter observations were sorted in ascending order with respect to time or, more precisely, seconds counters. This was carried out with the help of MATLAB sortrows function and managed data additions, where they occurred. As stated in the subsection devoted to the theoretical aspects, a method having its origin in Shewhart control charts was employed, with a sample size n = 20. As a consequence, at most only 19 observations were left out from statistical quality control. Considering that each variable was made up of thousands of measurements, 19 signals can

really be deemed an irrelevant amount. The number of samples m was computed by means of MATLAB floor function:

$$m_{samples} = \text{floor}\left(\frac{\text{number of data left after operating condition filter}}{n}\right) \quad (27)$$

The floor function rounds the argument to the nearest integer towards minus infinity. At this point, a n-by-m matrix was generated; each of its columns stored one of the m samples or subgroups. Quantities defined in (1), (2), (4), (5), (6), (9) and (10) were estimated with the aid of MATLAB mean and var commands. It is worth reminding the reader that the 2σ upper control limit criterion on $\bar{S}$ chart was adopted. Figure 13 illustrates the statistical check implemented on the temperature of cylinder A1 exhaust gases. For consistency's sake, Figure 13 includes exclusively the subgroups connected with the period of time that concerns the example.

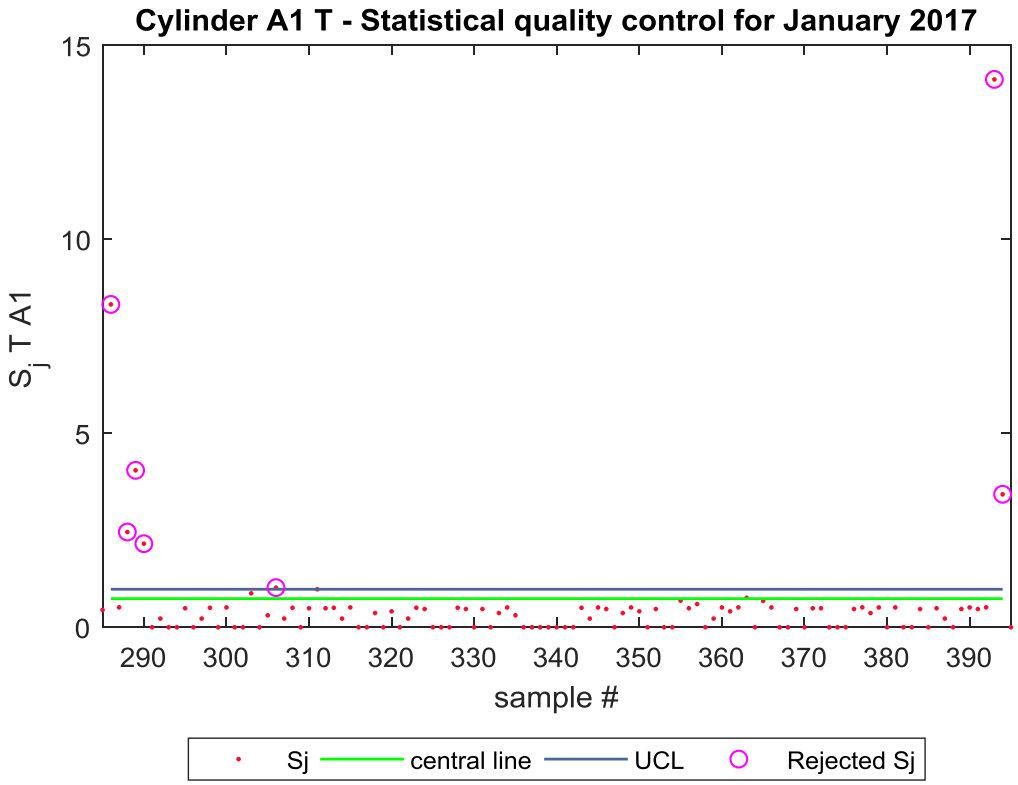


Figure 13. Execution of the statistical quality control on the example data

Red dots represent samples statistic S_j, while bright green and blue horizontal lines correspond to the central line and upper control limit, respectively. Magenta circles emphasize small data clusters that did not pass the statistical test, like some subgroups at the beginning and at the end of example time span. Once the samples not in compliance with statistical requirement were identified, the column vector of variable observations was paired with the corresponding column vector of seconds counters. However, it has to be noted that a degree of caution was exercised. In order to ensure an error-free combination of the two vectors,

uninteresting data were replaced by NaN elements. Then, NaN values, together with their seconds counters, were removed through rmmissing function at a stroke. Again, a meaningful quantity was determined. Specifically, the percentage ratio between the amount of data left after statistical quality test and the total number of data available at the start of data pre-processing:

$$\mathrm{SQ} = \frac{\text{number of data left after statistical quality filter}}{\text{total number of available input data for data pre} - \text{processing}} * 100 \qquad (28)$$

SQ = 89.5% for the temperature of cylinder A1 exhaust gases in January 2017.

The objective of the last part of MATLAB code is the temporal alignment of data. The developed strategy called for the conversion of each minute of the month into an interval of 60 seconds, according to the temporal reference system introduced. For example, the first minute of the month is equivalent to the interval [1,60]; the second minute is defined in [61,120] and so on up to the last minute of the month. In this way, minute by minute, the program searched for parameter observations associated with seconds counters that fell in the interval of the minute under review. If two or more data were found, the average of those data was recorded in the *k*th row of a new column vector, where k is the number of the minute involved in the iteration. In case of a month with 31 days, k goes from 1 to 44640 and the new vector containing data has, of course, a length of 44640. If only one value was detected, that measurement was written on the suitable vector position; NaN was registered whenever no data were found. Afterward, the number TA was calculated as follows:

$$\mathrm{TA} = \frac{\text{minutes of the month with valid data after all filters}}{\text{minutes in the month}} * 100 \qquad (29)$$

The numerator is the amount of data per minute that are useable in a statistical assessment; hence, NaN elements were not taken into account for TA evaluation. On the other hand, the denominator represents the amount of data that would be available in a perfect case. TA is a percentage ratio, like all the other values given by (25), (26) and (28). By the way, a direct comparison between (25) and (29) is allowed and the relation TA ≤ FL always holds. For instance, TA = 46.1% for the temperature of cylinder A1 exhaust gases.

Figure 14 outlines the impact of temporal alignment on the example data. The vertical dash-dotted line indicates the exact minute during which the engine was shut down: the *27921*th minute of January 2017. As one can clearly deduce from Figure 14, the algorithm excluded few samples that were really close to the plant stoppage. If a sharp spike in the temperature occurred, MATLAB code would have eliminated the subgroups connected with the spike. The logical sequence of steps followed for the implementation of the different filters turned out to be an effective model.

The last operation on data was the z-score normalization for each of the variables here considered. The methodology that supports the above-mentioned procedure has already been covered in the theoretical section of Chapter 1. In the end, the program wrote all the valid data on a Microsoft Excel spreadsheet file, as it did for engine operating load. Since the case study

comprehended 13 months, 13 Microsoft Excel files were generated. A uniform database was therefore accessible for use in any further statistical research and Table 13 gives an idea of how the observations have been organized.

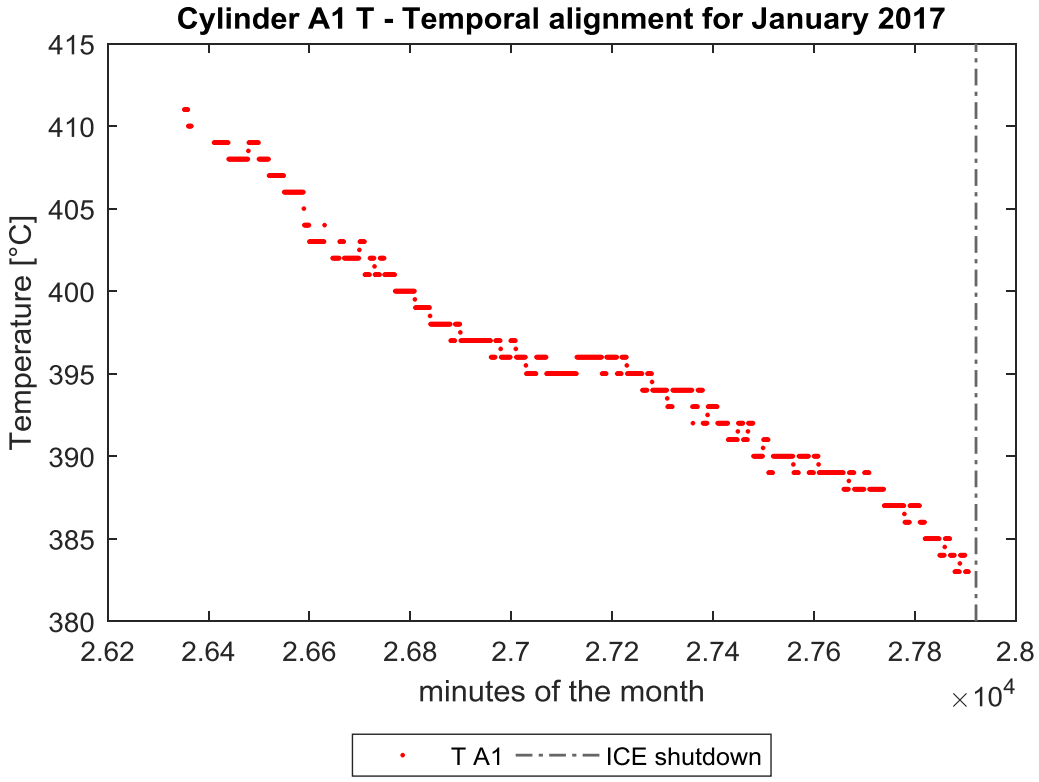


Figure 14. Temporal alignment performed on the example data

May 2018						
Minute of the month	Power [kW]	Fuel oil flow rate [l/h]	Normalized fuel oil flow rate	...	Temperature exhaust at SCR inlet [°C]	Normalized temperature exhaust at SCR inlet
1				...		
2				...		
...	...	...		...	...	
44639	8269	1856	0.44766	...		
44640	8275	1855	0.24732	...		

Table 13. Structure of Excel worksheet containing processed data

Annex A comprises one of data pre-processing expected results (see Figure 1), the trends of all the variables in May 2018. May 2018 has been chosen because it is representative of a month of plant optimal functioning.

The MATLAB code expressly built for data pre-processing has several advantages:

- ✓ it has been improved in terms of computational time, clearing items that lost their utility during the code execution from the system memory;

- ✓ it accepts data that are not in chronological order;
- ✓ it has a flexible structure, because new variables can be added to the program with a few changes;
- ✓ it is possible to distinguish between more than two operating conditions, modifying and increasing the number of thresholds set for working conditions but keeping the same basis;
- ✓ it is adaptable to any month of the year;
- ✓ it works also in case of leap years, needing a simple correction on the number of days in February.

Table 14 is the complete summary of data pre-processing on the internal combustion engine database. The information provided in the table will be commented in the second part of Chapter 2, since the output of data pre-processing constitutes the starting point of Principal Component Analysis.

		Year												
		2017												**2018**
Parameter	Filters	**January**	**February**	**March**	**April**	**May**	**June**	**July**	**August**	**September**	**October**	**November**	**December**	**May**
Engine operating load	*FL*	***49.2***	***93.9***	***89.5***	***43.0***	***30.0***	***30.8***	***16.7***	***0.0***	***33.5***	***45.4***	***28.4***	***17.5***	***78.1***
Fuel oil flow rate at engine inlet	OC	-	-	-	-						-			91.4
	SQ	-	-	-	-						-			84.9
	TA	**-**	**-**	**-**	**-**						**-**			**72.6**
Temperature of cylinder A1 exhaust gases	OC	95.5	99.2	87.8	44.5						46.9			83.2
	SQ	89.5	82.6	78.2	39.0						45.0			72.5
	TA	**46.1**	**78.5**	**79.7**	**37.7**						**43.6**			**68.2**
Temperature of cylinder A2 exhaust gases	OC	95.5	99.2	87.8	44.5						46.9			83.2
	SQ	91.4	80.3	82.4	39.0						45.0			72.2
	TA	**47.1**	**76.3**	**84.1**	**37.8**						**43.5**			**67.9**
Temperature of cylinder A3 exhaust gases	OC	95.5	99.2	87.8	44.5						46.9			83.2
	SQ	88.6	86.7	81.6	36.2						43.8			72.4
	TA	**45.6**	**82.3**	**83.4**	**35.1**						**42.4**			**68.1**

Temperature of cylinder A4 exhaust gases	OC	95.5	99.2	87.8	44.5						46.9			83.2
	SQ	89.2	84.3	81.6	39.0						44.8			72.5
	TA	**45.9**	**80.1**	**83.2**	**37.8**						**43.3**			**68.1**
Temperature of cylinder A5 exhaust gases	OC	95.5	99.2	87.8	44.5						46.9			83.2
	SQ	84.8	82.8	81.9	38.8						44.8			72.7
	TA	**43.7**	**78.7**	**83.6**	**37.6**						**43.4**			**68.3**
Temperature of cylinder A6 exhaust gases	OC	57.7	95.8	90.7	41.3						46.9			83.2
	SQ	51.3	76.3	79.1	34.4						43.8			73.5
	TA	**43.7**	**74.1**	**77.8**	**35.7**						**42.4**			**69.1**
Temperature of cylinder A7 exhaust gases	OC	57.7	95.8	90.7	41.3						46.9			83.2
	SQ	50.7	80.0	79.5	36.5						44.8			71.6
	TA	**43.3**	**78.1**	**78.0**	**37.9**						**43.4**			**67.3**
Temperature of cylinder A8 exhaust gases	OC	57.7	95.8	90.7	41.3						46.9			83.2
	SQ	51.3	87.5	86.6	36.2						44.5			73.5
	TA	**43.7**	**85.6**	**85.2**	**37.6**						**43.1**			**69.1**
Temperature of cylinder A9 exhaust gases	OC	57.7	95.8	90.7	41.3						46.9			83.2
	SQ	51.6	82.4	84.7	37.0						44.9			72.6
	TA	**44.0**	**80.5**	**83.2**	**38.4**						**43.4**			**68.3**
Temperature of cylinder A10 exhaust gases	OC	57.7	95.8	90.7	41.3						46.9			83.2
	SQ	50.6	79.8	79.0	36.4						45.9			71.3
	TA	**43.2**	**77.9**	**77.4**	**37.8**						**44.4**			**67.0**
Temperature of cylinder B1 exhaust gases	OC	54.4	95.9	90.5	40.3						46.9			83.2
	SQ	49.4	82.5	82.1	36.2						46.2			74.0
	TA	**44.8**	**80.6**	**80.8**	**38.6**						**44.7**			**69.5**
Temperature of cylinder B2 exhaust gases	OC	54.4	95.9	90.5	40.3						46.9			83.2
	SQ	47.7	81.3	81.1	36.1						43.0			80.6
	TA	**43.3**	**79.4**	**79.8**	**38.5**						**41.6**			**75.7**
Temperature of cylinder B3 exhaust gases	OC	54.4	95.9	90.5	40.3						46.9			83.2
	SQ	48.3	78.9	77.5	34.7						44.8			72.0
	TA	**43.8**	**76.9**	**76.0**	**37.0**						**43.4**			**67.7**

Temperature of cylinder B4 exhaust gases	OC	54.4	95.9	90.5	40.3						46.9			83.2
	SQ	47.2	79.6	77.5	33.6						44.0			73.9
	TA	**42.8**	**77.6**	**76.1**	**35.9**						**42.6**			**69.4**
Temperature of cylinder B5 exhaust gases	OC	54.4	95.9	90.5	40.3						46.9			83.2
	SQ	49.7	80.0	81.8	35.9						42.4			72.8
	TA	**45.1**	**78.0**	**80.9**	**38.3**						**41.1**			**68.5**
Temperature of cylinder B6 exhaust gases	OC	57.9	95.8	90.6	41.5						46.9			83.2
	SQ	50.8	80.3	79.7	36.4						44.8			73.9
	TA	**43.2**	**78.3**	**78.4**	**37.6**						**43.4**			**69.4**
Temperature of cylinder B7 exhaust gases	OC	57.9	95.8	90.6	41.5						46.9			83.2
	SQ	50.7	80.1	82.9	37.4						44.2			73.3
	TA	**43.1**	**78.2**	**81.5**	**38.6**						**42.8**			**68.9**
Temperature of cylinder B8 exhaust gases	OC	57.9	95.8	90.6	41.5						46.9			83.2
	SQ	50.6	79.9	82.0	40.6						44.1			73.2
	TA	**43.0**	**78.1**	**80.7**	**42.1**						**42.7**			**68.8**
Temperature of cylinder B9 exhaust gases	OC	57.9	95.8	90.6	41.5						46.9			83.2
	SQ	49.8	79.7	81.8	35.7						44.1			72.1
	TA	**42.4**	**77.6**	**80.2**	**36.9**						**42.7**			**67.8**
Temperature of cylinder B10 exhaust gases	OC	57.9	95.8	90.6	41.5						46.9			83.2
	SQ	51.0	80.4	77.7	35.7						42.9			72.8
	TA	**43.3**	**78.4**	**76.2**	**36.9**						**41.6**			**68.5**
Coolant Δp in Charge-Air Cooler, bank A	OC	50.2	93.9	89.4	43.1						45.4			78.2
	SQ	48.0	89.4	84.0	40.6						41.7			66.1
	TA	**46.8**	**89.2**	**83.0**	**40.4**						**41.7**			**66.0**
Coolant Δp in Charge-Air Cooler, bank B	OC	50.2	93.9	89.4	43.1						45.4			78.2
	SQ	47.5	89.1	84.5	40.4						41.4			66.3
	TA	**46.3**	**88.9**	**83.5**	**40.2**						**41.3**			**66.2**
Environmental humidity	OC	57.2	95.9	89.8	46.6						46.9			83.2
	SQ	54.0	86.3	80.0	39.9						40.7			70.7
	TA	**45.7**	**84.2**	**79.2**	**37.0**						**39.5**			**66.4**

Environmental temperature	OC	57.2	95.9	89.8	46.6						46.9			83.2
	SQ	49.6	82.2	77.9	39.8						40.3			72.2
	TA	**40.8**	**79.6**	**77.2**	**36.5**						**39.1**			**67.9**
Revolutions per minute, bank A	OC	53.0	95.9	98.0	47.5						57.1			91.4
	SQ	51.4	91.8	94.0	45.8						55.7			90.1
	TA	**47.5**	**90.0**	**85.7**	**41.4**						**44.3**			**77.0**
Revolutions per minute, bank B	OC	52.0	95.9	90.2	45.7						57.1			91.5
	SQ	50.4	92.3	86.5	43.8						55.8			90.7
	TA	**47.5**	**90.3**	**85.7**	**41.2**						**44.4**			**77.3**
Turbocharger inlet temperature, bank A	OC	52.0	99.1	85.0	40.3						46.9			83.2
	SQ	48.6	88.8	77.1	35.7						45.3			71.4
	TA	**45.6**	**42.7**	**34.9**	**38.1**						**43.8**			**67.1**
Turbocharger outlet temperature, bank A	OC	52.0	95.9	90.0	40.3						46.9			83.2
	SQ	46.0	83.3	80.4	36.2						45.0			66.8
	TA	**43.2**	**81.2**	**79.7**	**38.6**						**43.5**			**62.9**
Turbocharger inlet temperature, bank B	OC	-	-	84.8	40.3						46.9			83.2
	SQ	-	-	77.6	35.4						45.0			70.0
	TA	**-**	**-**	**35.2**	**37.8**						**43.6**			**65.9**
Turbocharger outlet temperature, bank B	OC	52.0	95.9	90.0	40.3						46.9			83.2
	SQ	46.0	81.1	79.5	36.2						45.3			67.1
	TA	**43.3**	**79.2**	**78.7**	**38.7**						**43.9**			**63.1**
Backpressure at SCR system inlet	OC	98.4	93.9	89.5	43.0						55.9			88.1
	SQ	87.7	84.3	83.8	40.3						48.9			79.0
	TA	**32.3**	**84.2**	**82.8**	**40.2**						**39.6**			**69.9**
Differential pressure drop in SCR system	OC	50.2	93.9	89.3	43.3						55.6			78.7
	SQ	50.2	93.9	86.9	41.5						52.5			75.5
	TA	**49.1**	**93.8**	**85.4**	**41.1**						**42.8**			**74.8**
Specific consumption of NH_3	OC	50.2	93.9	89.3	50.3						56.0			88.2
	SQ	50.2	93.9	86.9	47.1						53.0			86.1
	TA	**49.1**	**93.8**	**85.4**	**40.1**						**42.9**			**76.1**

Exhaust fumes temperature at SCR system outlet	OC	50.2	93.9	89.3	43.2						46.2			78.1
	SQ	50.2	93.9	87.8	41.5						44.8			76.5
	TA	**49.1**	**93.8**	**86.3**	**41.2**						**44.0**			**76.3**
Exhaust fumes temperature at SCR system inlet	OC	50.2	93.9	89.3	43.2						46.2			78.1
	SQ	50.2	93.9	88.1	42.1						45.3			77.1
	TA	**49.1**	**93.8**	**86.6**	**41.8**						**44.5**			**76.9**
Legend	FL	Full Load												
	OC	Operating Condition												
	SQ	Statistical Quality												
	TA	Time Alignment												

Table 14. A summary of the whole data pre-processing

Application of Principal Component Analysis

The criteria that led to the ICE parameters selection and to the identification of appropriate data samples, in order to study engine performance, are now expounded. Three factors influenced months and parameters subsets choice for the application of PCA technique:

- FL, i.e. the fraction of minutes of the month with a full load operating condition (see (25) in Chapter 2), equal to or larger than 40%;
- Subdivision of parameters in conformity with FMEA grouping of variables by component (see Table 3 in Chapter 2);
- The percentage of minutes where data of each and every variable within a subset were simultaneously available had to be around 20%.

Translating the first and third sentences into numbers, a month of 31 days that meets the above-stated requirements has at least 17856 minutes with the plant working at full load; moreover, any individual subset, which underwent the statistical evaluation here reported, comprised approximately 8900 minutes retaining a comprehensive collection of validated measures. According to the Table 14, the months that satisfied the constraint on FL were: January 2017, February 2017, March 2017, April 2017, October 2017 and May 2018. However, measurements about turbocharger inlet temperature (bank B) were not properly recorded during the first two months of 2017 and consequently these months were excluded from the statistical investigation.

The second criterion suggested the association of parameters that are principally related to possible system anomalies. At the component level, a parameter deviation can truly be linked to a possible failure. The case study considered in this thesis involved three major components: the engine, the turbocharger and the Selective Catalytic Reduction system. Figure 15, Figure 16 and Figure 17 provide a representation of the aforementioned components.

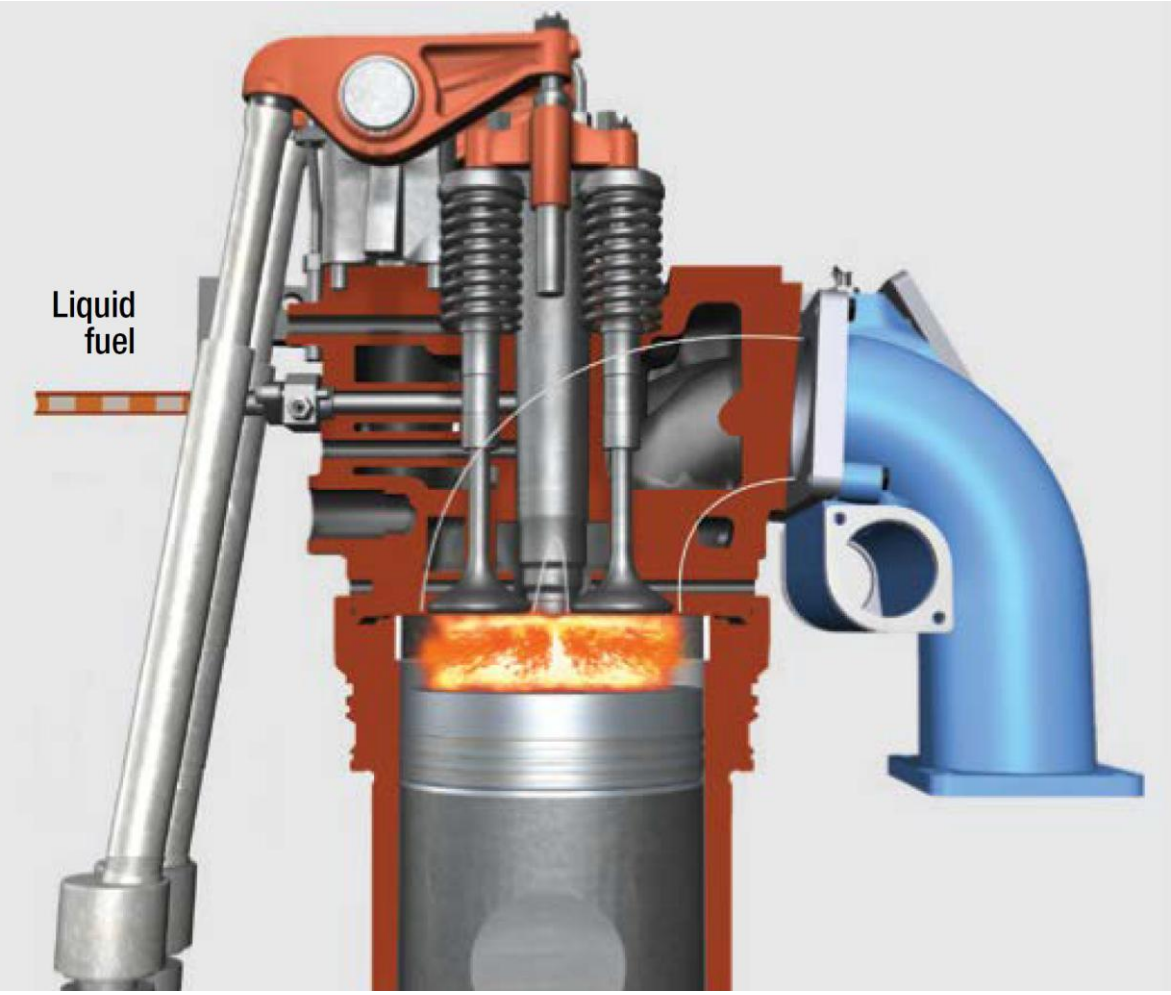


Figure 15. Cylinder and fuel injection system of the ICE [16]

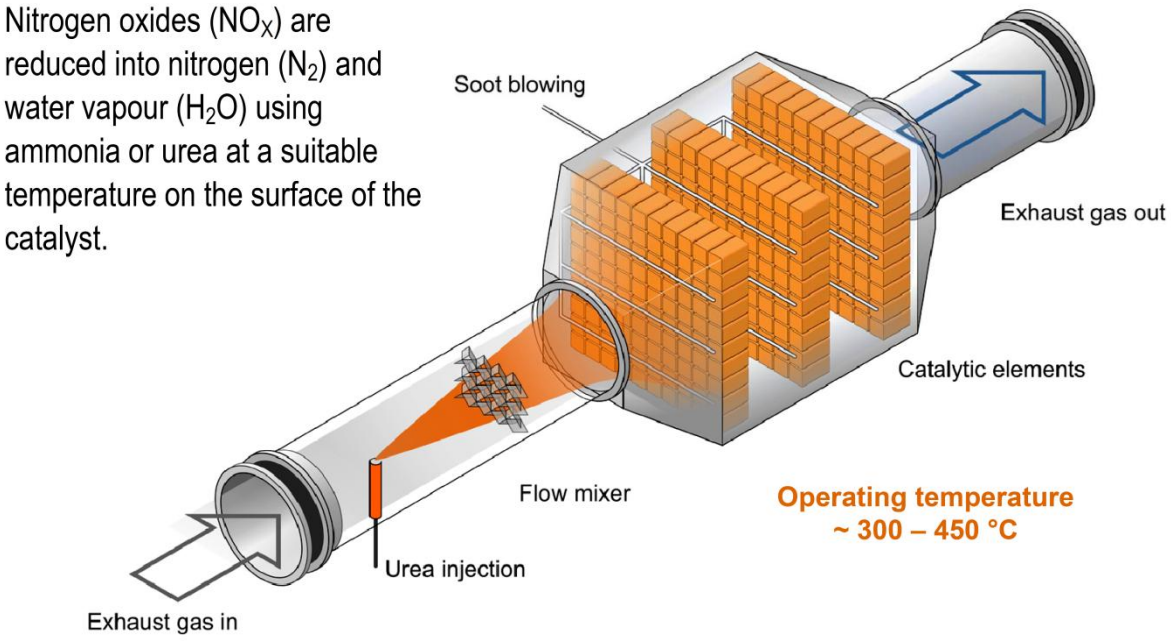
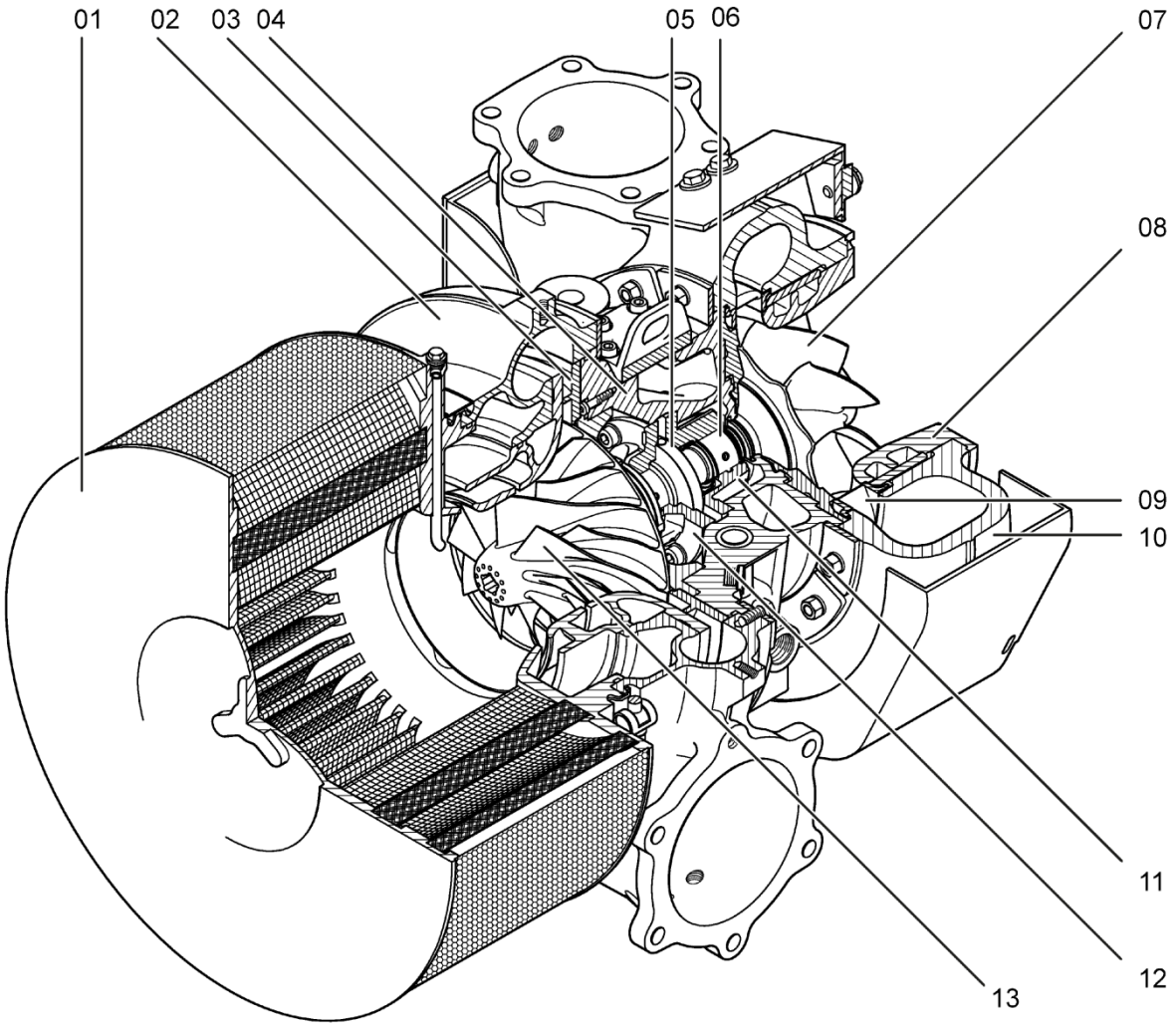


Figure 16. Layout of the Selective Catalytic Reduction system [16]



01	Filter silencer / air suction branch	08	Gas outlet flange
02	Compressor casing	09	Nozzle ring
03	Diffuser	10	Turbine casing
04	Bearing casing	11	Turbine-end bearing flange
05	Axial thrust bearing	12	Compressor-end bearing flange
06	Radial plain bearing	13	Compressor wheel
07	Turbine		

Figure 17. Cutaway drawing of the turbocharger [17]

Four groups of variables were therefore established:

1) temperatures of cylinders exhaust gases (bank A), coolant Δp in Charge-Air Cooler (bank A) and backpressure at SCR system inlet;
2) temperatures of cylinders exhaust gases (bank B), coolant Δp in Charge-Air Cooler (bank B) and backpressure at SCR system inlet;
3) revolutions per minute (banks A and B), turbocharger inlet and outlet temperatures (banks A and B), environmental temperature and humidity;
4) backpressure at SCR system inlet, differential pressure drop in SCR system, specific consumption of NH_3, inlet and outlet temperatures of exhaust fumes in SCR system.

Principal Component Analysis and Hotelling T^2 statistic were applied to the four groups; the complete set of figures can be found in Annex B, whereas Table 15 shows the most salient information that was obtained from the utilization of the two techniques.

		2017			2018
Groups	**Type of retrieved information**	**March**	**April**	**October**	**May**
First subset (Engine)	**% of useful data**	50.3	19.1	30.9	33.3
	Number of useful data	22465	8230	13788	14844
	% of total variance explained by first two PCs	64.29	67.71	91.02	72.11
	Number of observations that exceed T2 limit	1161	581	670	969
Second subset (Engine)	**% of useful data**	44.7	21.6	28.4	36.4
	Number of useful data	19937	9332	12691	16251
	% of total variance explained by first two PCs	62.31	67.55	90.76	69.06
	Number of observations that exceed T2 limit	2160	690	1086	836
Third subset (TC)	**% of useful data**	25.1	26.4	32.5	44.1
	Number of useful data	11225	11393	14496	19702
	% of total variance explained by first two PCs	74.32	77.92	76.40	92.61
	Number of observations that exceed T2 limit	512	834	658	944
Fourth subset (SCR system)	**% of useful data**	81.4	39.2	35.8	66.2
	Number of useful data	36345	16936	15982	29539
	% of total variance explained by first two PCs	70.21	79.01	80.77	87.96
	Number of observations that exceed T2 limit	1565	1096	911	1972

Table 15. Relevant information retrieved during the application of PCA and Hotelling T^2 statistic (2D approach)

A comparison between the four months is presented in Figure 18, Figure 19, Figure 20 and Figure 21.

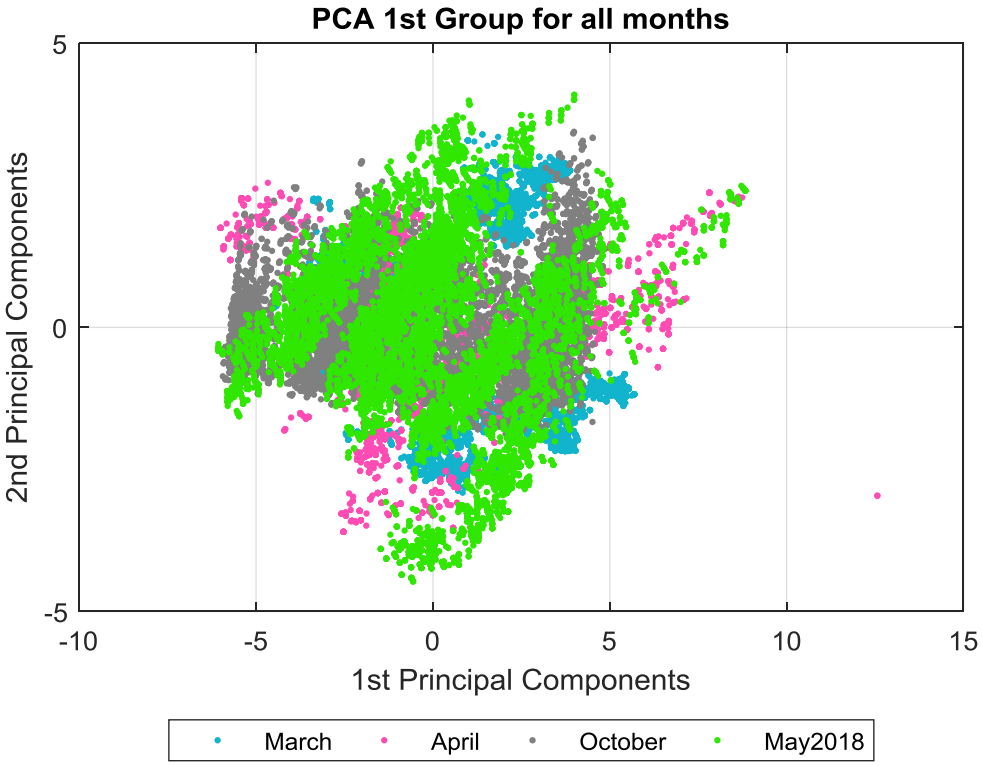


Figure 18. Plot of the first two PCs for the first set of parameters (all months)

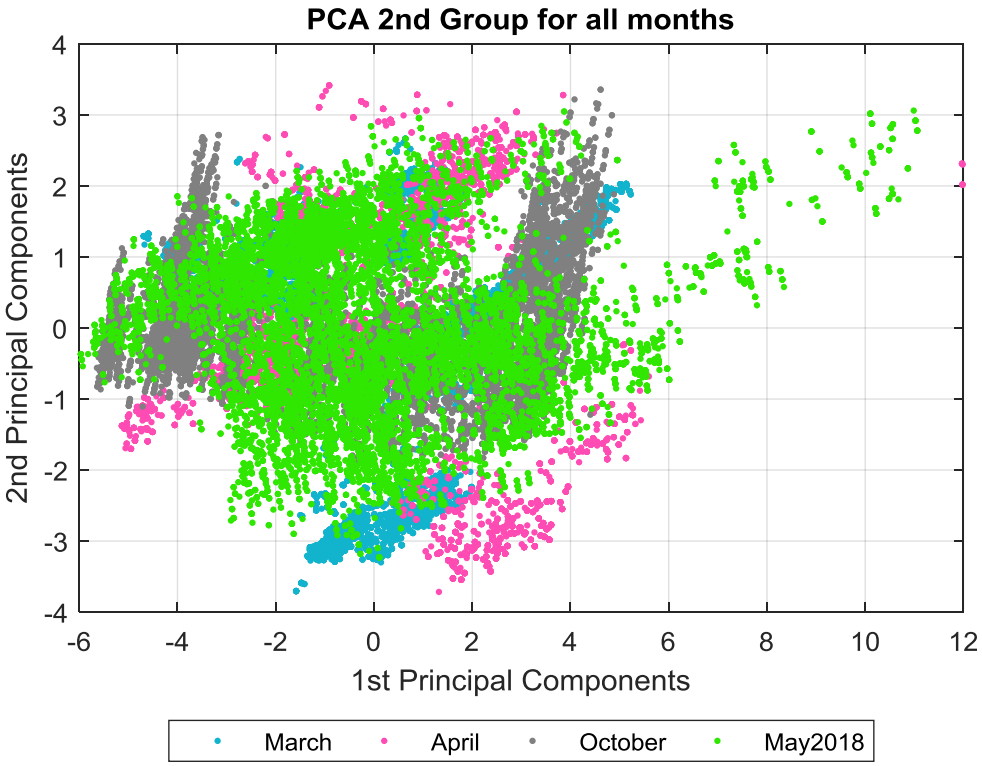


Figure 19. Plot of the first two PCs for the second set of parameters (all months)

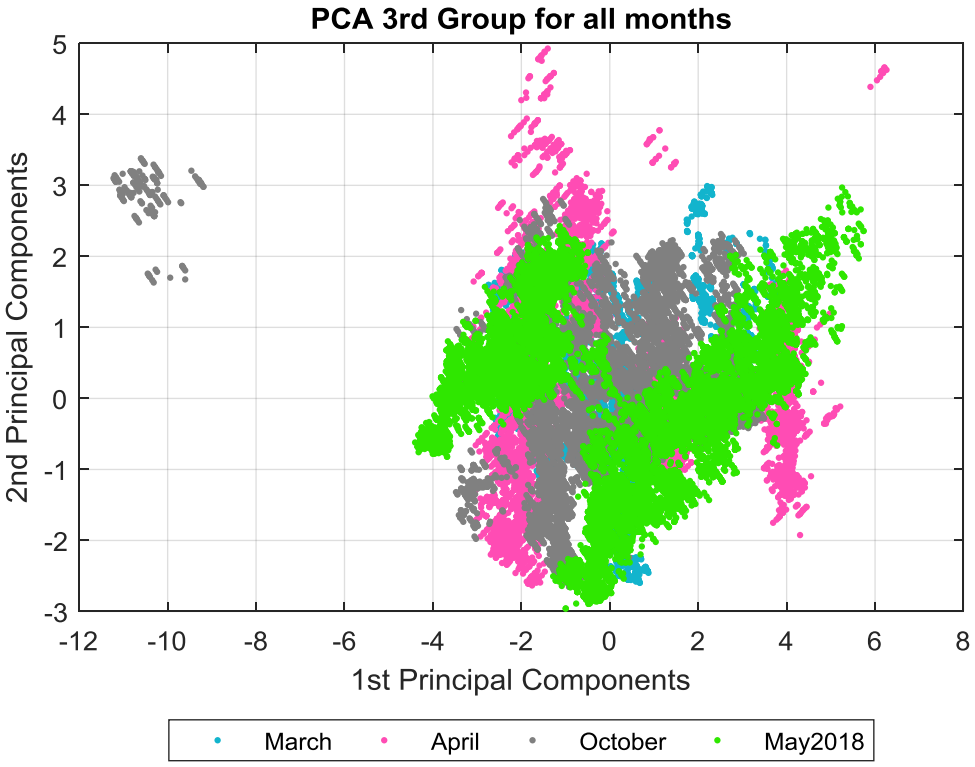


Figure 20. Plot of the first two PCs for the third set of parameters (all months)

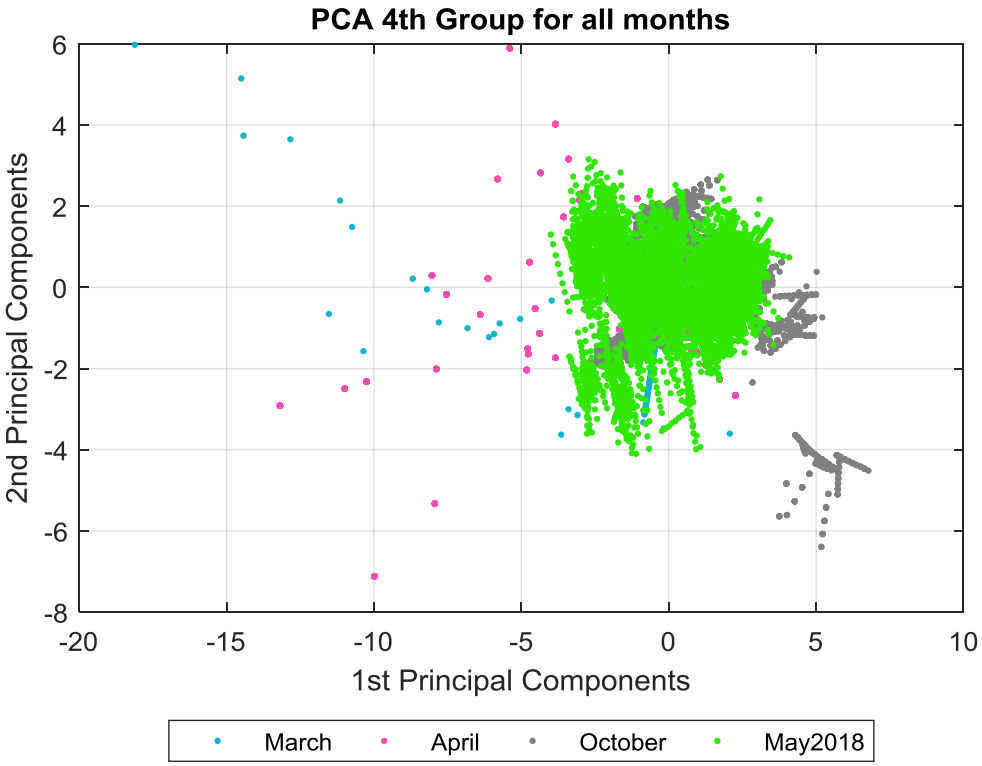


Figure 21. Plot of the first two PCs for the fourth set of parameters (all months)

		Stop		Restart		
		Day	**Time**	**Day**	**Time**	**Failed equipment**
2017	**March**	9	09:30	9	12:44	Engine (cylinders A2, B1, B2)
		16	10:05	16	11:12	Engine (cylinder A2)
		17	07:00	19	13:00	Engine (all cylinders)
		24	04:00	24	05:00	Engine (cylinder A8)
		27	08:50	27	09:50	Engine (cylinder A8)
		28	08:22	28	09:31	Engine (cylinders A9, B7)
		30	12:00	30	13:38	Engine (cylinder B5)
	April	8	17:35	8	19:56	Engine (cylinder B8)
		12	20:00	12	21:10	Engine (cylinder A9)
		14	06:40	18	18:15	Engine (cylinder B8)
	October	1	00:00	4	18:15	Engine (all cylinders), turbocharger, SCR system
		6	03:44	6	04:35	Engine (cylinder A10)
		8	08:25	8	10:45	Engine (cylinder B1)
		12	23:12	13	01:10	Engine (cylinder B1)
		15	12:15	15	13:03	Engine (cylinder A7)
		16	21:54	17	02:10	Engine (cylinder B1)
		29	19:58	31	15:00	SCR system
2018	**May**	1	00:00	**4**	**09:45**	SCR system
		6	**08:00**	**6**	**08:40**	**Engine (cylinder A2)**
		8	**15:00**	**8**	**16:15**	**Engine (cylinder A10)**
		9	**19:20**	**9**	**20:03**	**Engine (cylinder B4)**
		25	**07:42**	**25**	**11:50**	**Engine (cylinders B4, B5), turbocharger**

Table 16. ICE maintenance history

Even though the amount of useful data was enough to perform a reliable statistical study (see Table 15), the results were hardly interpretable and not explanatory, according to engine's maintenance history that is reported in Table 16. The number of considered principal components was increased by one and, at the same time, one (and only one) large group of relevant variables has been examined. The group encompasses 33 parameters: temperatures of cylinders exhaust gases (banks A and B), coolant Δp in Charge-Air Cooler (banks A and B), revolutions per minute (banks A and B), turbocharger inlet and outlet temperatures (banks A and B), backpressure at SCR system inlet, inlet temperature of exhaust fumes in SCR system, fuel oil flow rate at engine inlet, environmental temperature and humidity. Table 17 has the same kind of information of Table 15 but it concerns the three-dimensional approach.

		2017			2018
	Type of retrieved information	**March**	**April**	**October**	**May**
All the relevant variables	**% of useful data**	12.8	10.0	19.5	17.3
	Number of useful data	5720	4325	8684	7711
	% of total variance explained by first three PCs	67.25	69.71	89.58	75.64
	Number of observations that exceed T2 limit	611	443	611	659

Table 17. Relevant information retrieved during the application of PCA and Hotelling T^2 statistic (3D approach)

In the latter approach, the quantity of valuable data is lower than in the former one but PCA often reveals relationships that were not previously suspected and this was the case. Figure 22 outlines different operating zones. As an example, dark blue dots represent the period from the 09:45 of 4th May 2018 (just after the restoration of SCR system) to the 08:00 of 6th May 2018 (see Table 16). Plant stoppage was due to the failure of engine's cylinder A2. An optimal functioning zone for the ICE can be identified by magenta dots. Figure 23, Figure 24 and Figure 25 show that points corresponding to a good operating condition for the system are located in the same area, with the exception of October 2017. The problem still had to be translated because this piece of information was not ready for use. There is the necessity of providing human operators with warnings and Hotelling T^2 statistic was applied with this intent. As illustrated in Figure 26, Figure 27, Figure 28 and Figure 29, research outcomes regarding the months of March 2017 and May 2018 are interesting, although warnings are not always given with adequate advance notice. In addition, these kinds of warnings are not specific: they merely offer an indication of forthcoming problems and the information is at the machine level. Table 18 includes the final results deriving from the observation of Hotelling T^2 charts. Confidence limits associated with output results should be calculated for the purpose of reaching a prognosis.

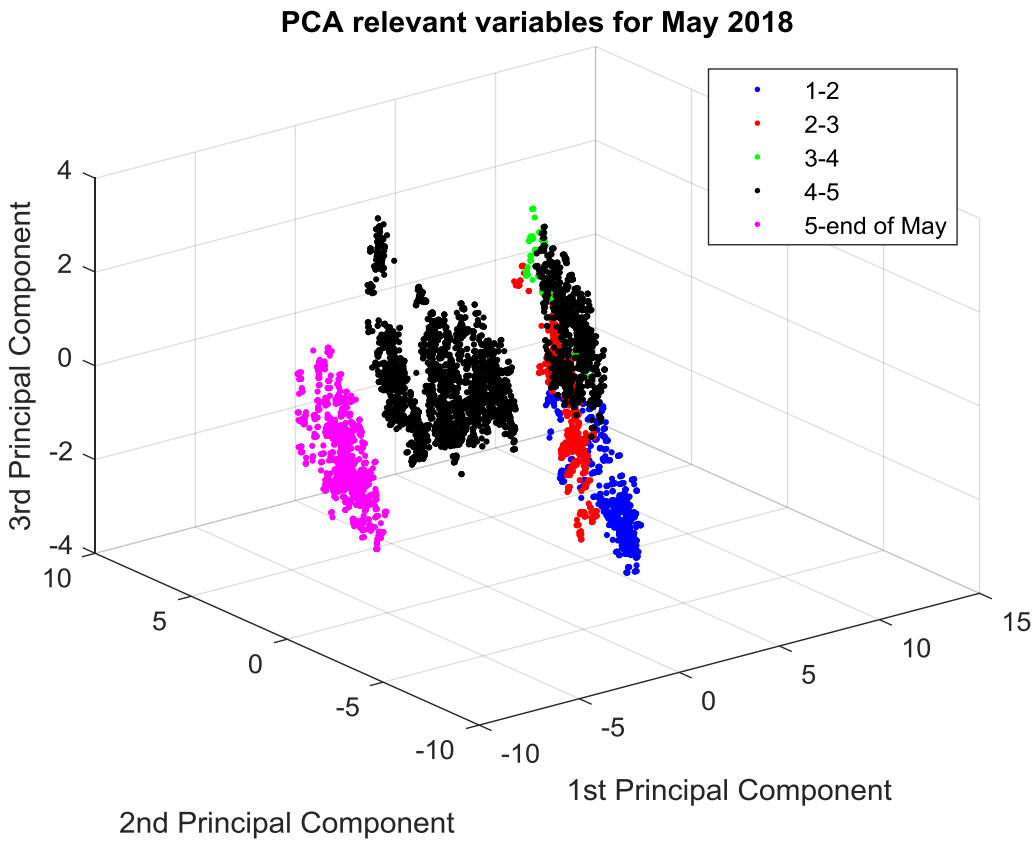


Figure 22. Plot of the first three PCs for the large set of parameters in May 2018

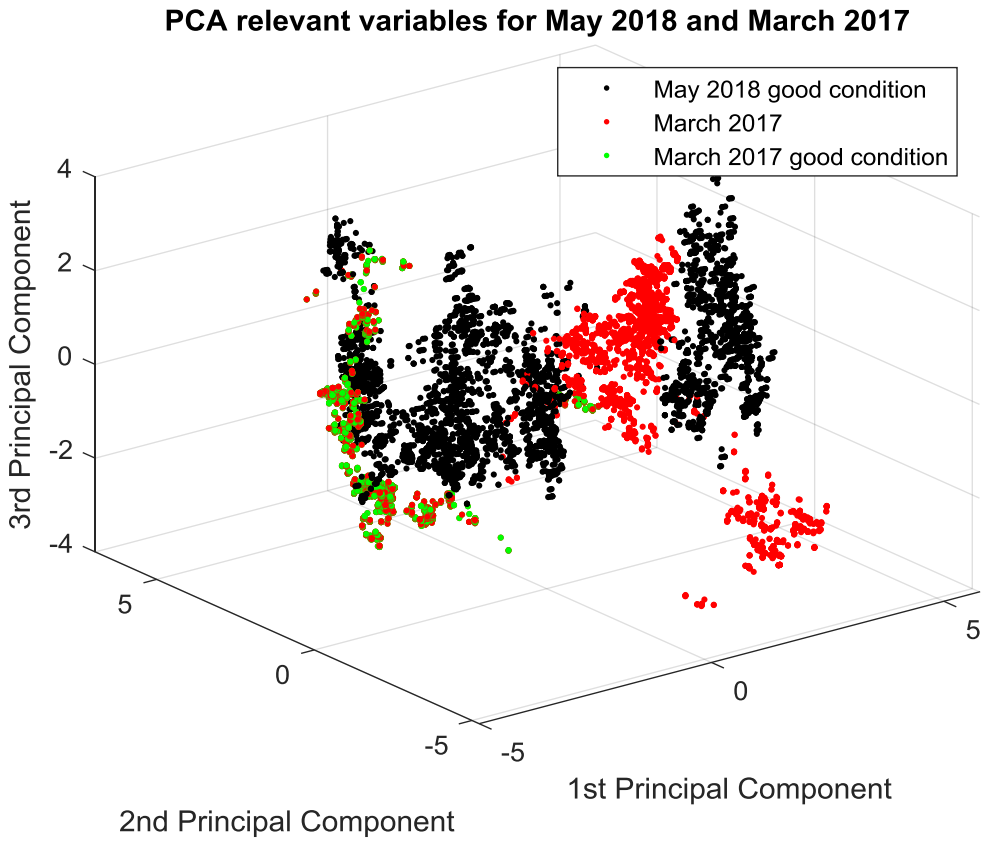


Figure 23. 3D comparison between May 2018 and March 2017

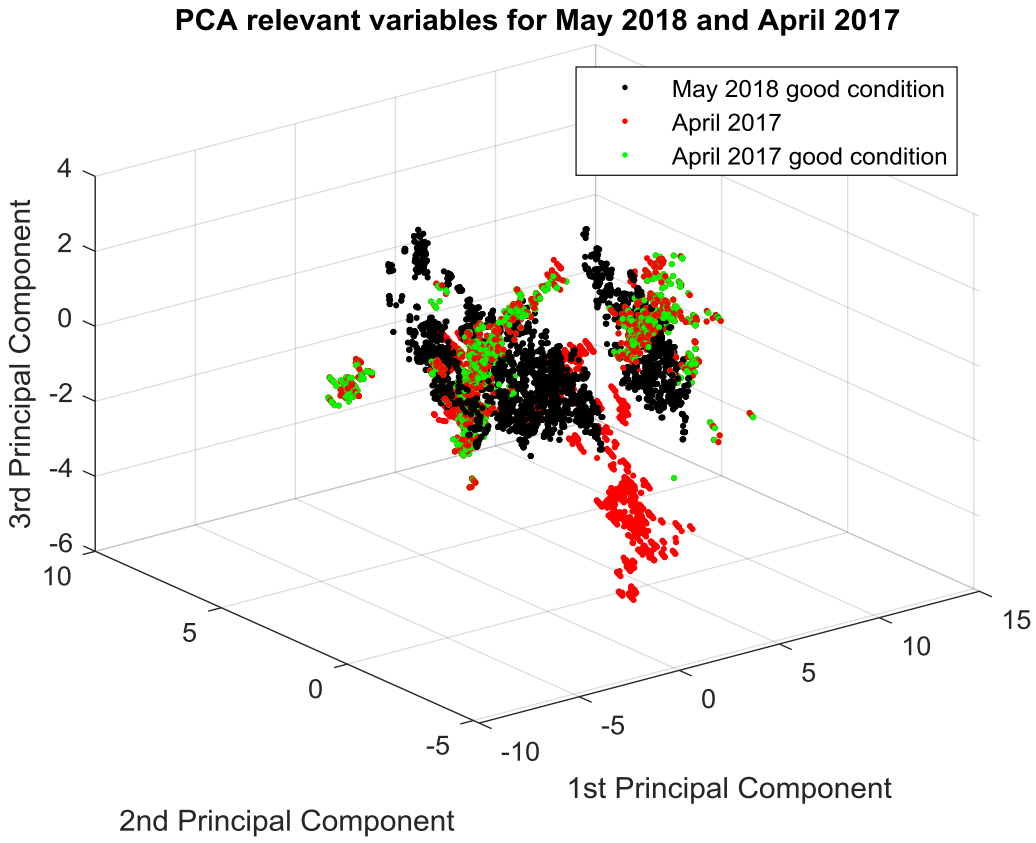


Figure 24. 3D comparison between May 2018 and April 2017

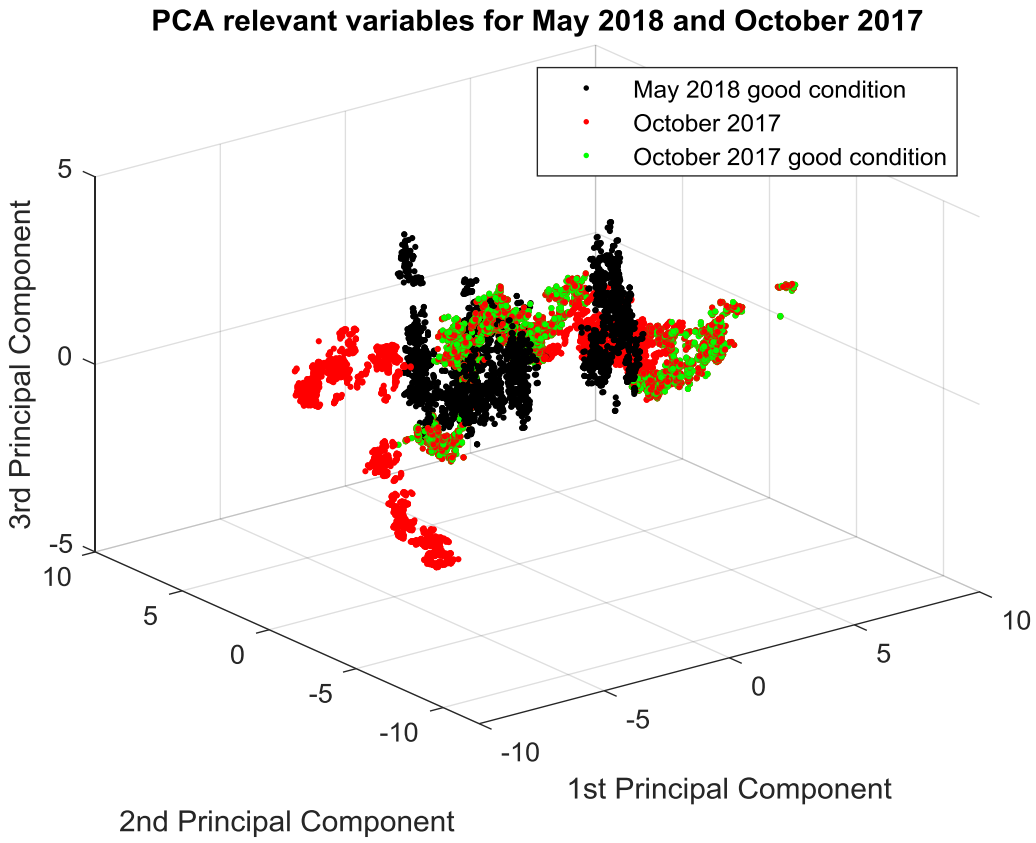


Figure 25. 3D comparison between May 2018 and October 2017

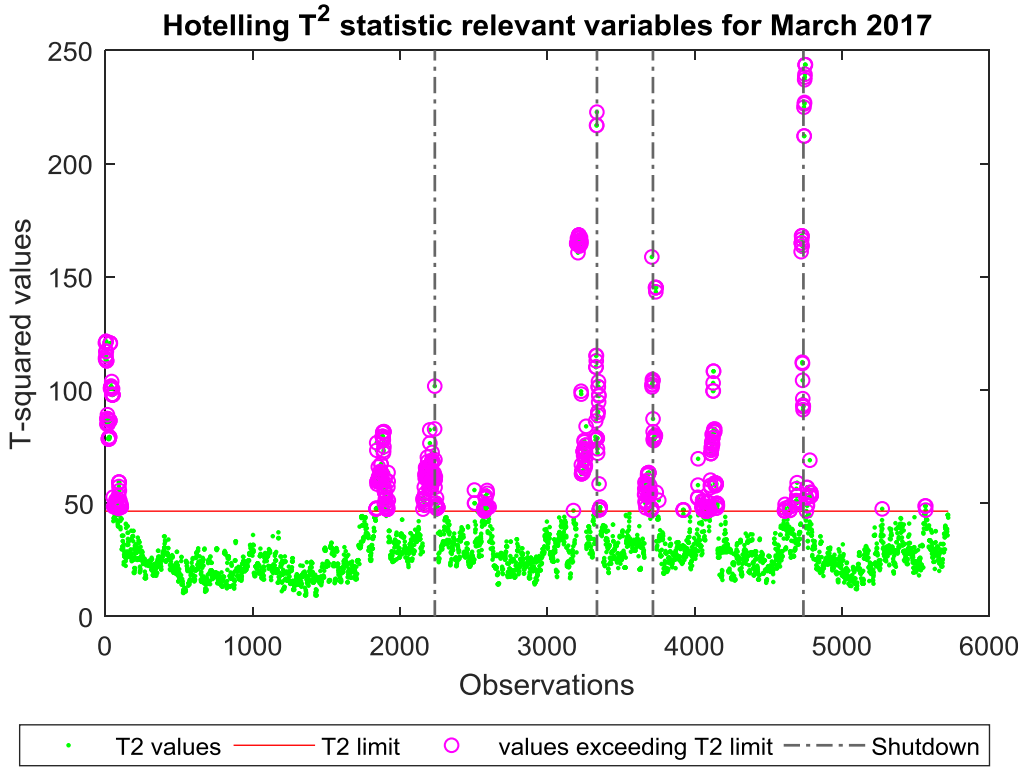


Figure 26. Hotelling T^2 statistic for the large set of parameters in March 2017

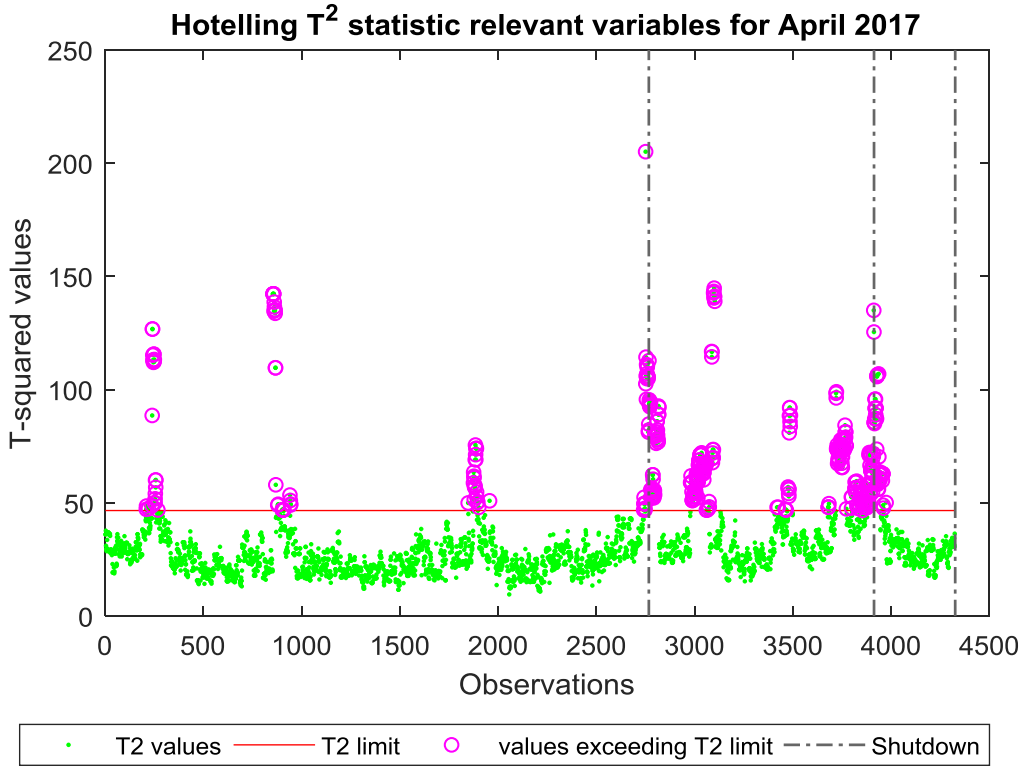


Figure 27. Hotelling T^2 statistic for the large set of parameters in April 2017

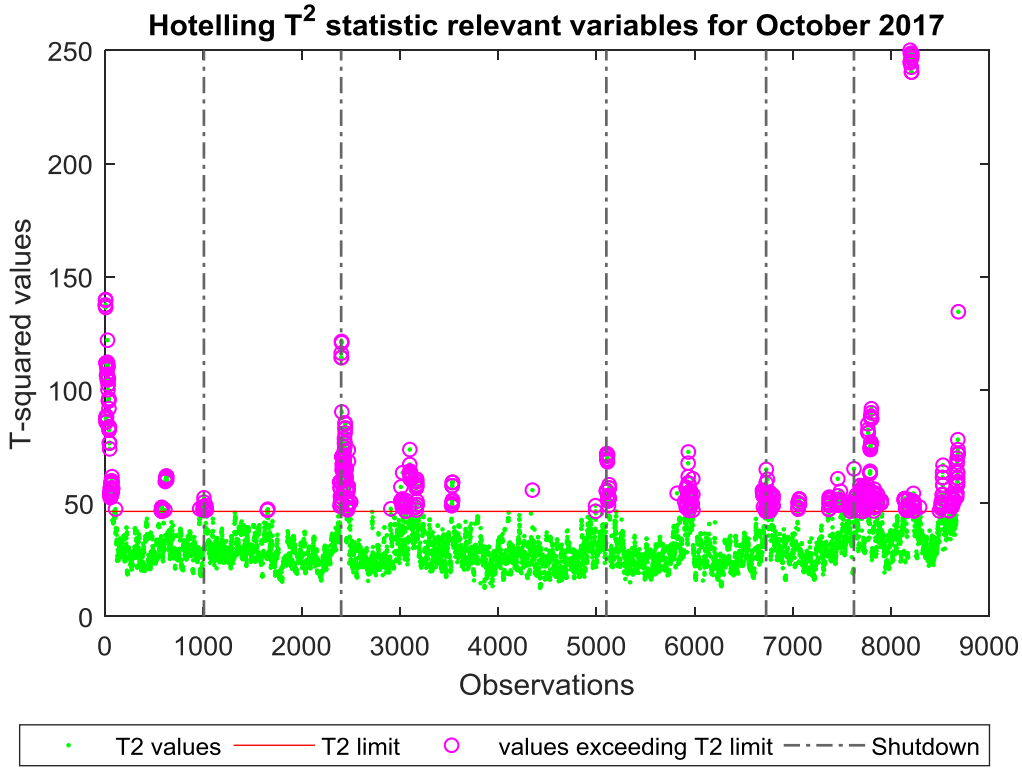


Figure 28. Hotelling T^2 statistic for the large set of parameters in October 2017

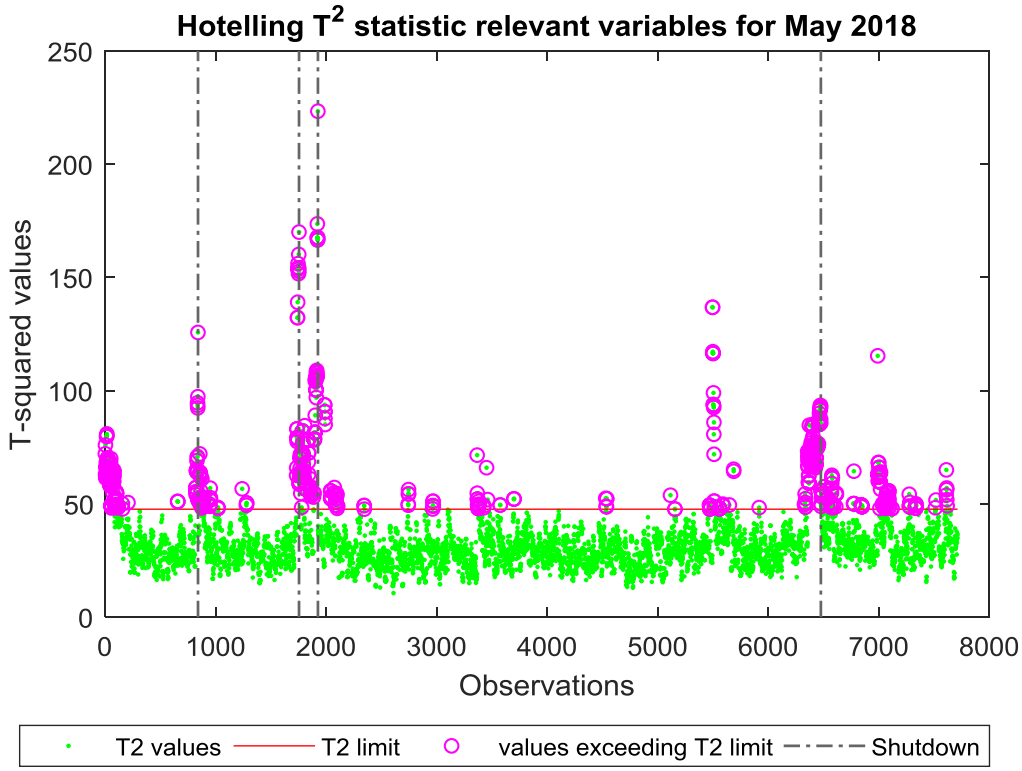


Figure 29. Hotelling T^2 statistic for the large set of parameters in May 2018

Type of retrieved information	2017			2018
	March	April	October	May
Detected incipient failures	4	2	3	3
Missed warning signals	0	1	2	1
False alarms	1	0	1	1

Table 18. Final results inferred from Hotelling T^2 charts

The first row of Table 18 keeps a careful tally of the number of impending failures which are correctly identified through Hotelling T^2 control charts: this is one expected result of the current study. The second row contains the number of cases in which incipient failures are not detected; these unwelcome outcomes are the worst situation, since a component breaks down without warning and it is urgent for the maintenance crew to put the machine back to work. The described scenario threatens the success of predictive maintenance. False alarms (last row of Table 18) imply the misclassification of healthy components as a component that is close to a failure. These alerts can cause trouble and they may lead to unnecessary maintenance costs.

Chapter 3 – Conclusions

Before drawing conclusions from the results, some considerations should be mentioned. The goal of this thesis was the evaluation of the effectiveness of PCA and Hotelling T^2 statistic on monitoring health condition of a complex equipment, which operates in a cogeneration plant. This objective entailed the creation of a sophisticated MATLAB code, to pre-process a massive amount of data (more than 20 millions of numbers). After the intricate but necessary rationalization and validation of the available database, the focus was shifted to the application of Principal Component Analysis. A multivariate analysis technique like PCA helps in handling data with complicated correlation structure. Moreover, PCA is a dimension reduction method and thereby is fit for purpose in the present research. It did not give remarkable results with the two-dimensional approach, whereas it highlighted discernible engine's problems in the three-dimensional approach. As showed in Figure 22, magenta dots represent points of engine regular working conditions. The cloud of black dots that occupy the central part of the graph corresponds to a loss of engine performance, due to fouling phenomenon at turbocharger level. The cloud of dots at the right side of the plot can be connected to imminent failures of fuel injection systems. PCA was followed by Hotelling T^2 statistic, which is suitable for establishing whether or not the system failures were foreseeable events. As reported in the last part of Chapter 2, in the considered period (4 months), 12 out of 16 incipient failures would have been detected. Even if prognostics, like any other prediction techniques, cannot have a 100% success rate in forecasting failures, a model tuning process is recommended. Nevertheless, one can imagine a real-time monitoring system; the opportunity to inform a human operator about an incipient failure, together with the possibility of associating that forthcoming failure with a position in the PCA 3D space, would be a very interesting support in maintenance decision making process. After a model refinement, components residual useful life estimation can be carried out. The transition from diagnosis to prognosis requires classification and regression models, which are used for forecasting time-series data. Both the methods belong to the class of supervised learning techniques and they are effective when the failure reference model is constructed.

Finally, the following additional comments on the specific case study seem legitimate. There is margin for improvement in data collection. For example, some files containing data were poorly or incorrectly formatted and this situation should be avoided. The number of long sequences of missing data should be minimized. Furthermore, the goal of turning large volumes of data into actionable information was pursued without having fuel oil consumption for March, April and October. Obtained results cannot be profitably used in the immediate future also because chemical composition and physical characteristics of fuel oil were unknown. Since the combustion process is influenced by the type of fuel, a data sheet with defined reference ranges for the fuel physical-chemical characteristics would be strongly advised. This statement is especially valid when esterified fuel oil are used, like in the considered cogeneration plant.

Annex A

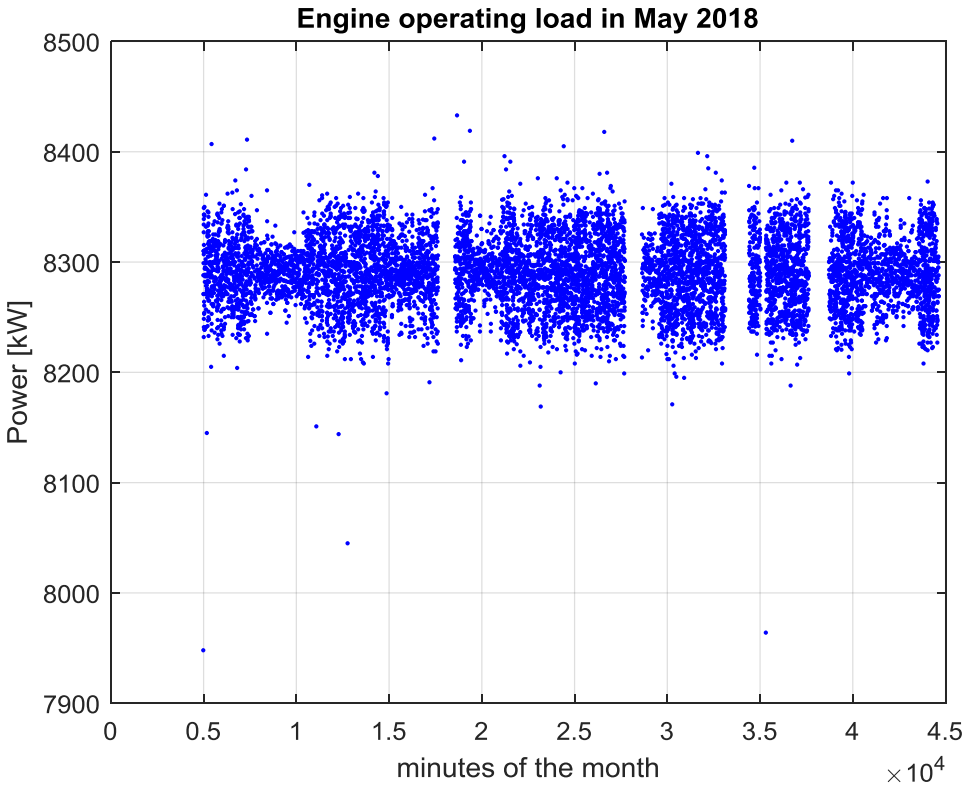


Figure A-1. Trend of engine operating load

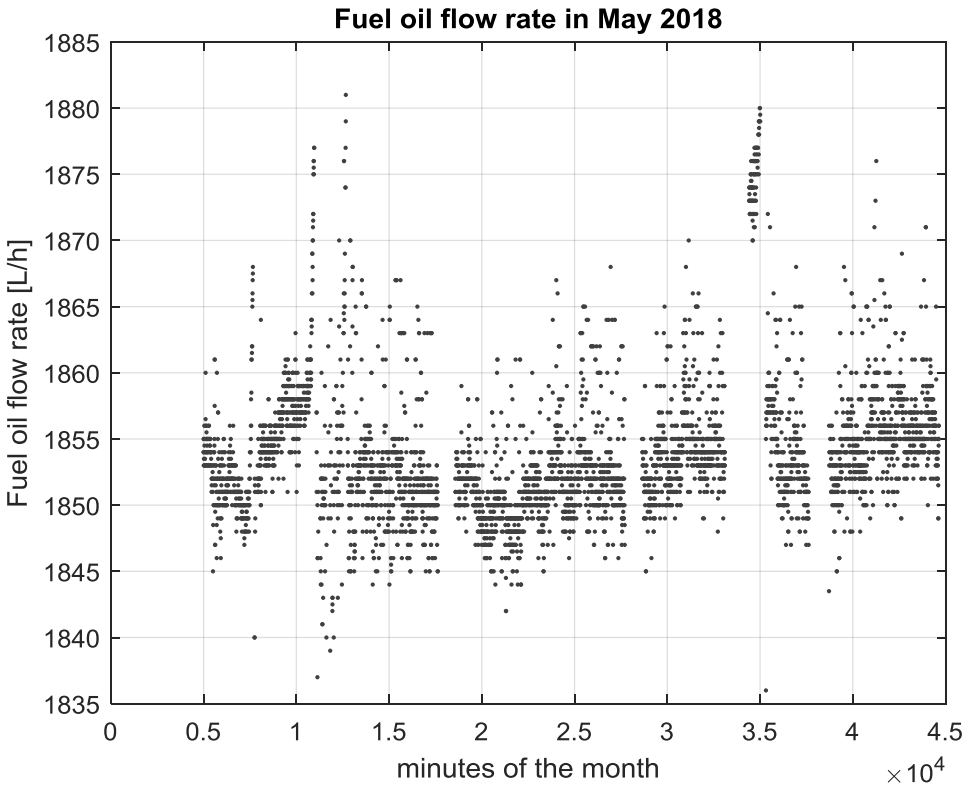


Figure A-2. Trend of fuel oil flow rate

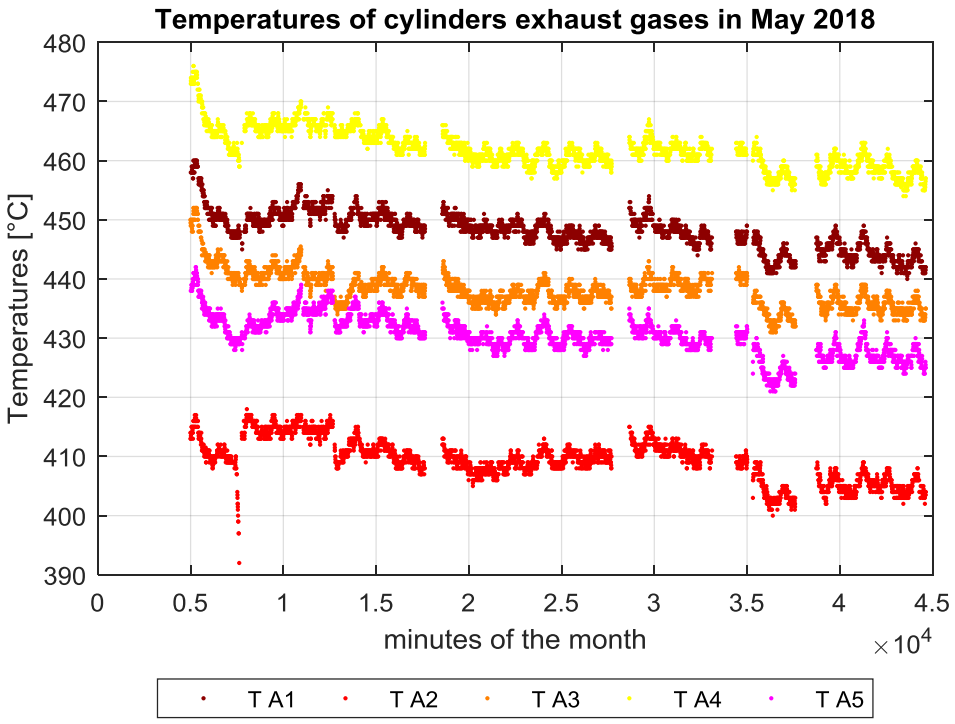


Figure A-3. Trend of temperatures of cylinders A1, A2, A3, A4 and A5 exhaust gases

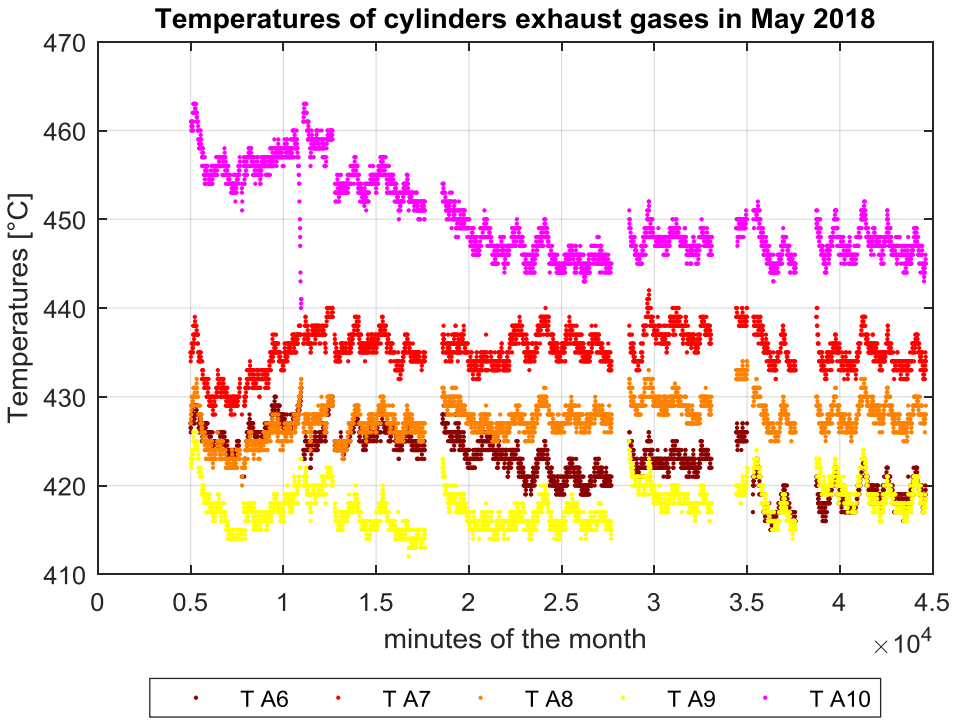


Figure A-4. Trend of temperatures of cylinders A6, A7, A8, A9 and A10 exhaust gases

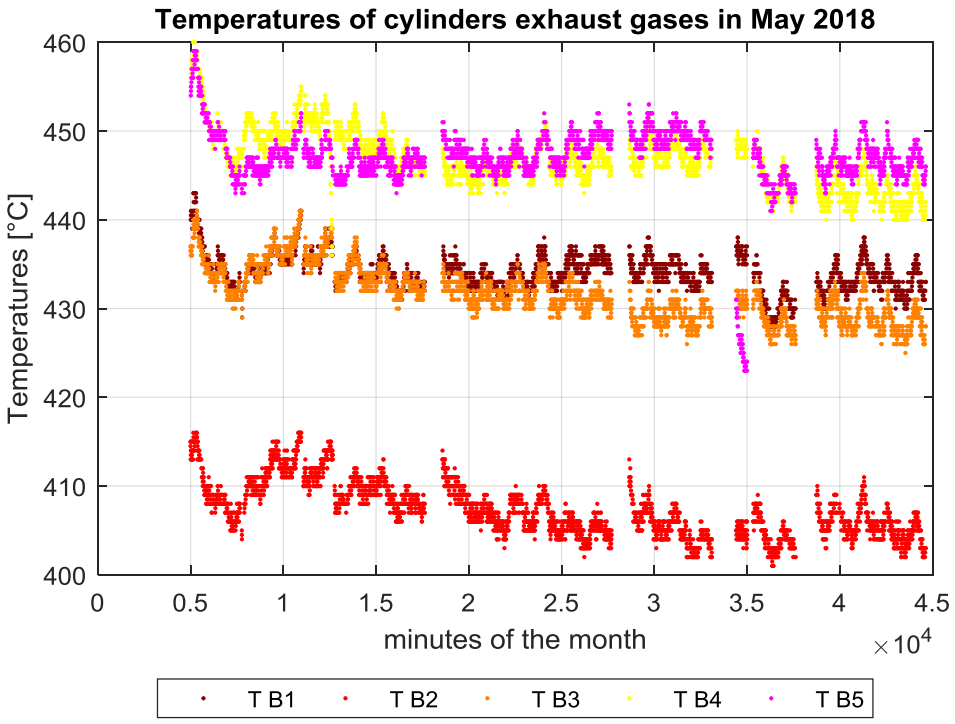


Figure A-5. Trend of temperatures of cylinders B1, B2, B3, B4 and B5 exhaust gases

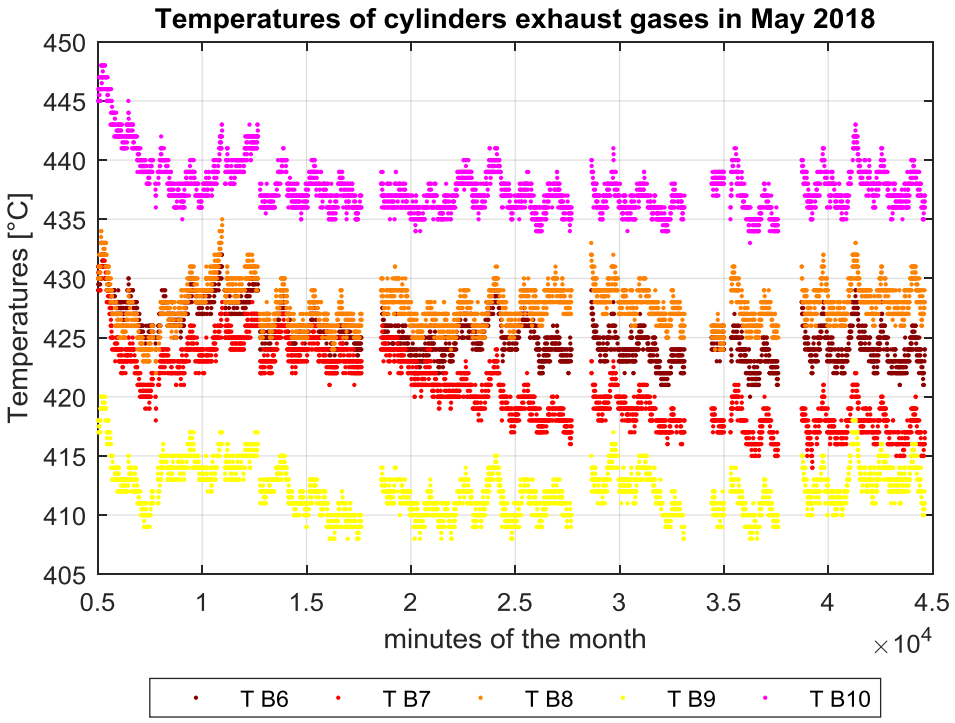


Figure A-6. Trend of temperatures of cylinders B6, B7, B8, B9 and B10 exhaust gases

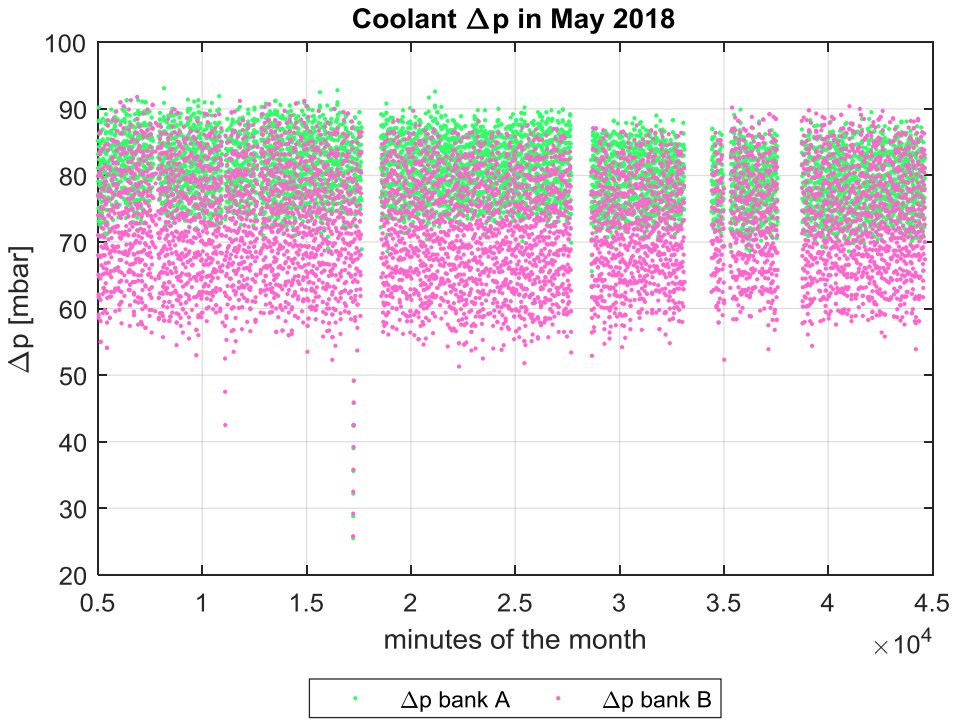


Figure A-7. Trend of coolant Δp in Charge-Air Cooler

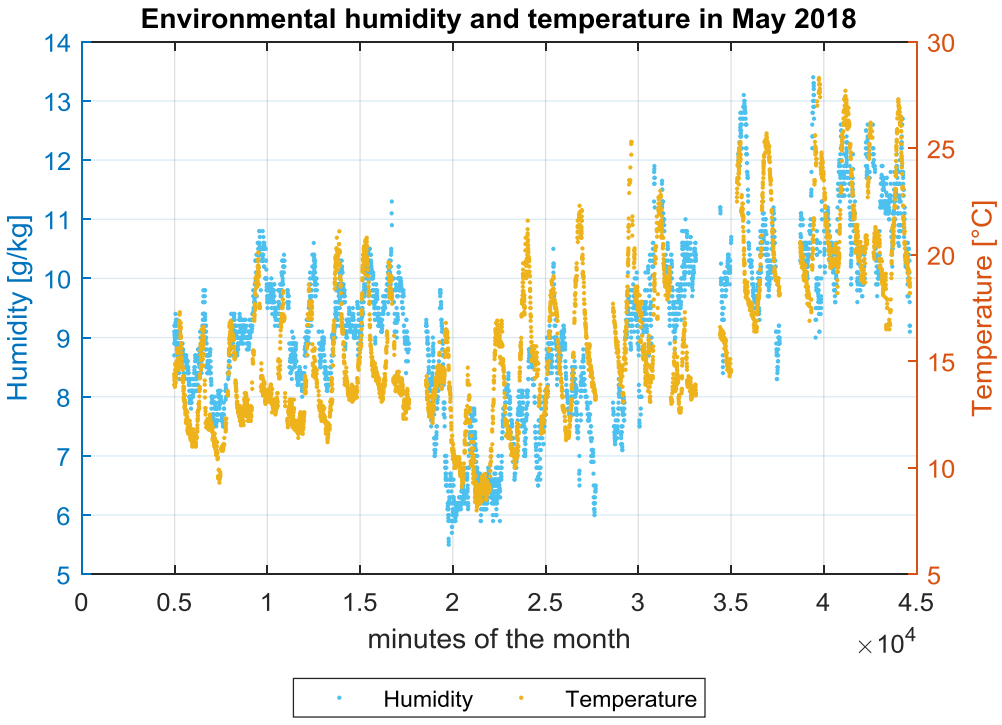


Figure A-8. Trend of environmental humidity and temperature

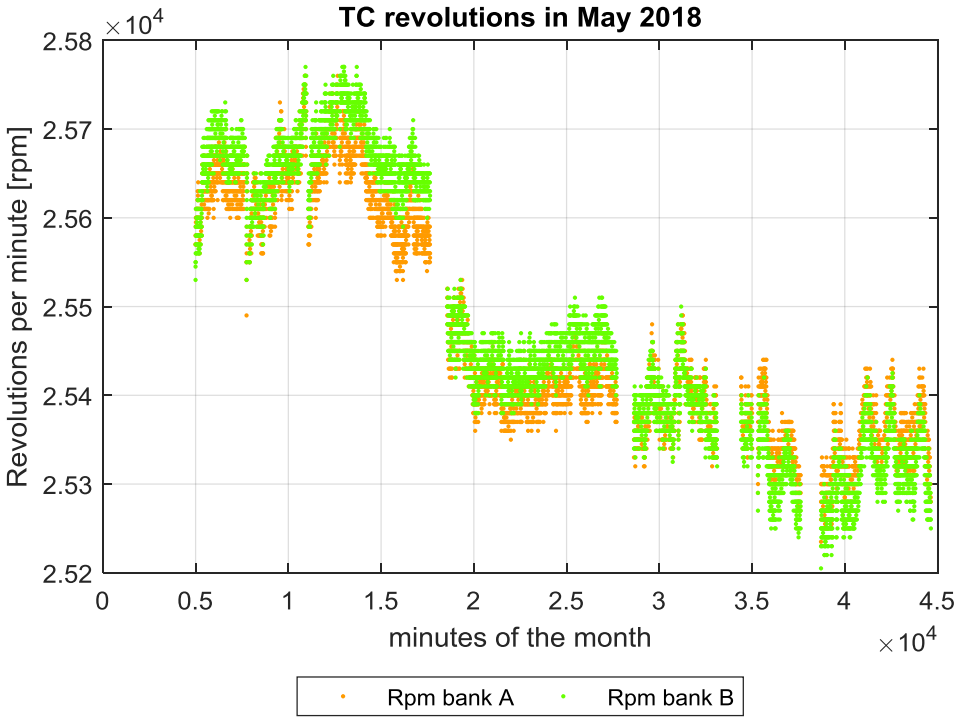


Figure A-9. Trend of turbocharger revolutions per minute

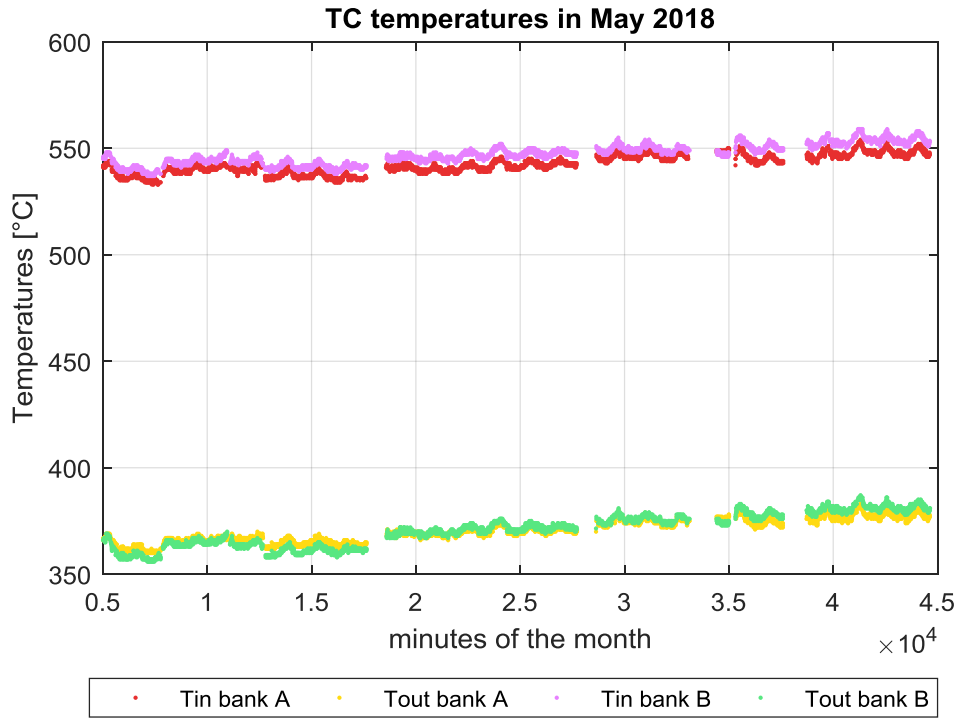


Figure A-10. Trend of turbocharger temperatures

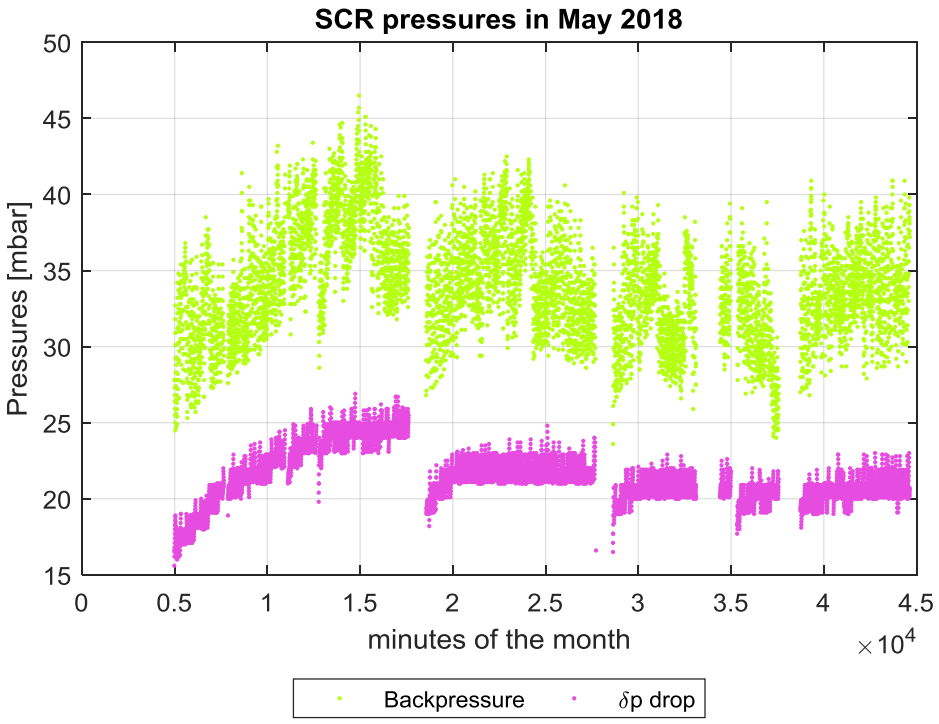


Figure A-11. Trend of SCR system pressures

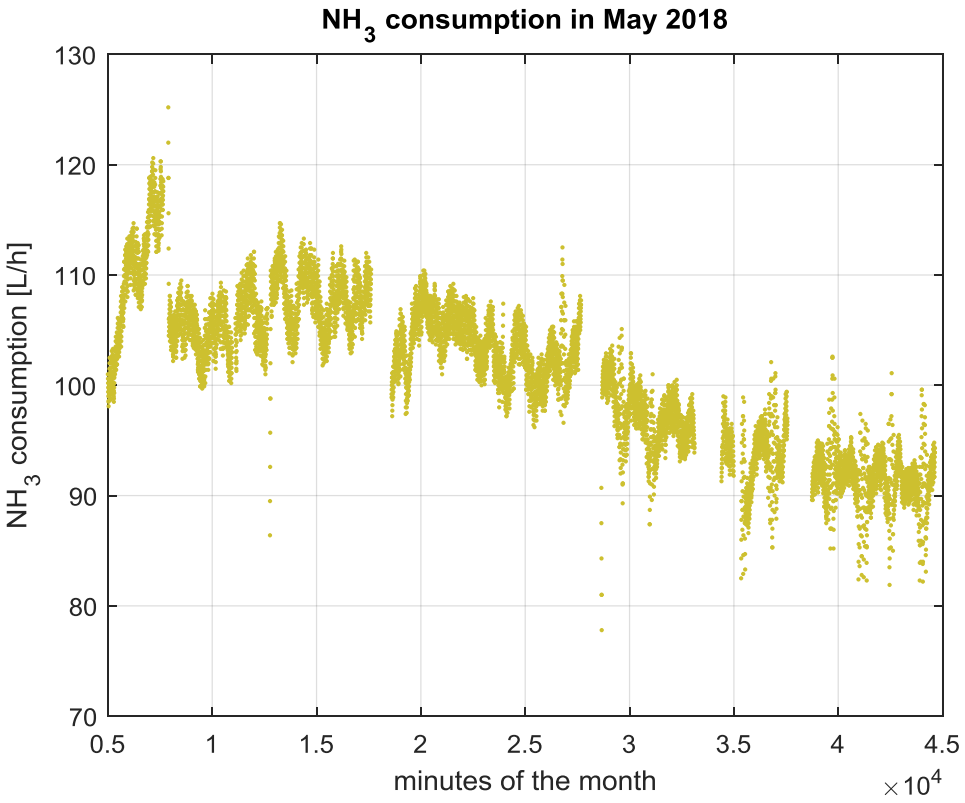


Figure A-12. Trend of NH_3 consumption

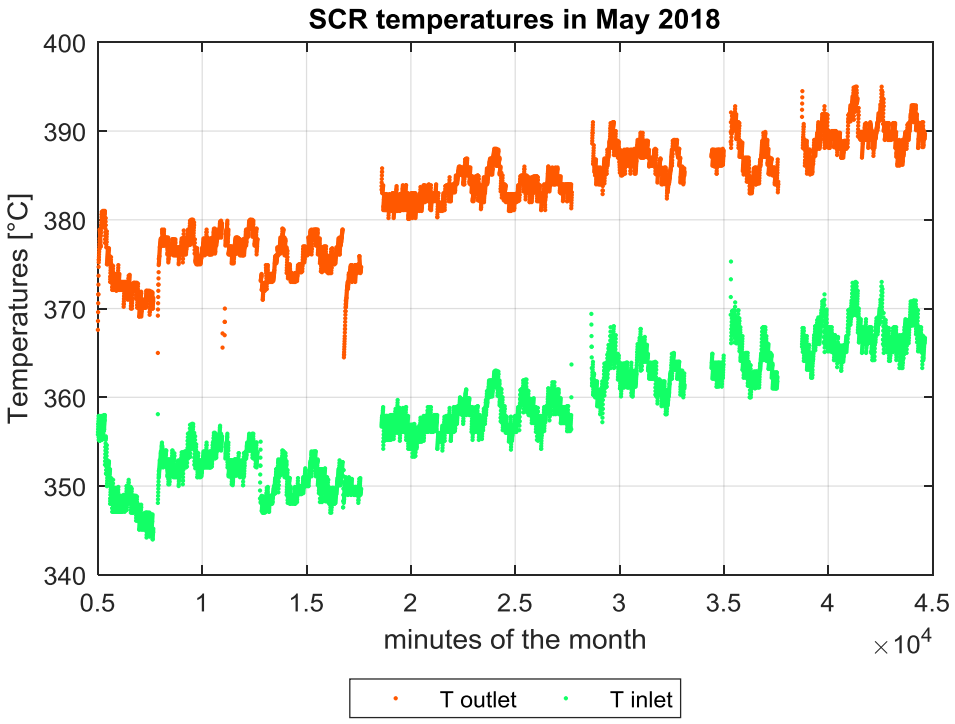


Figure A-13. Trend of temperatures of SCR system exhaust fumes

Annex B

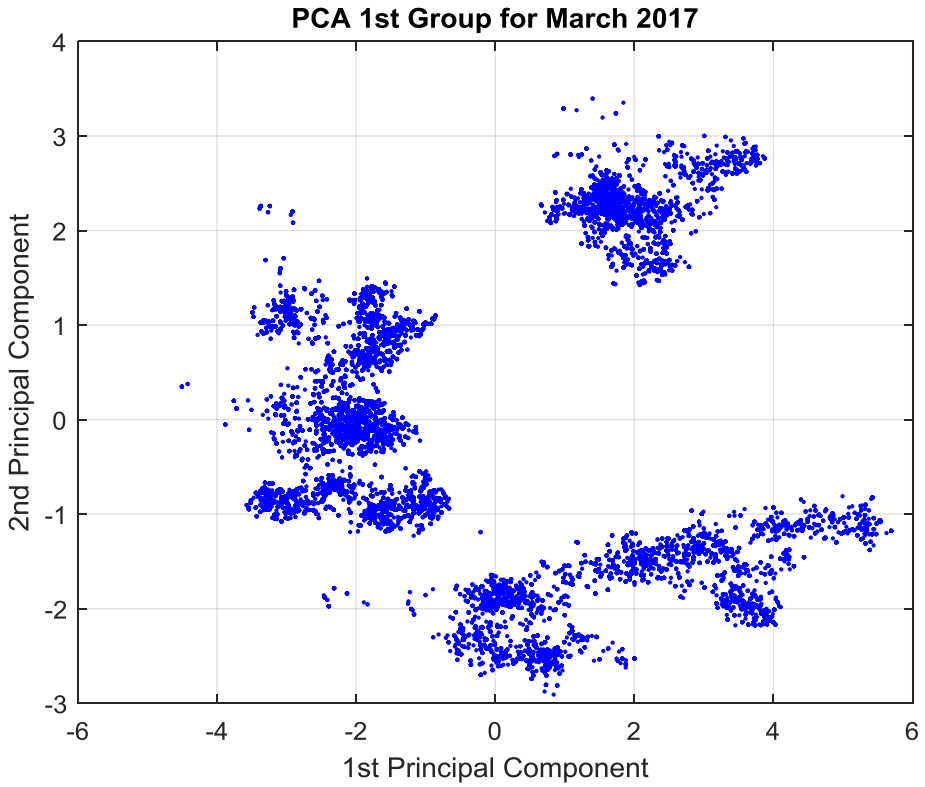


Figure B-1. Plot of the first two PCs for the first set of parameters in March 2017

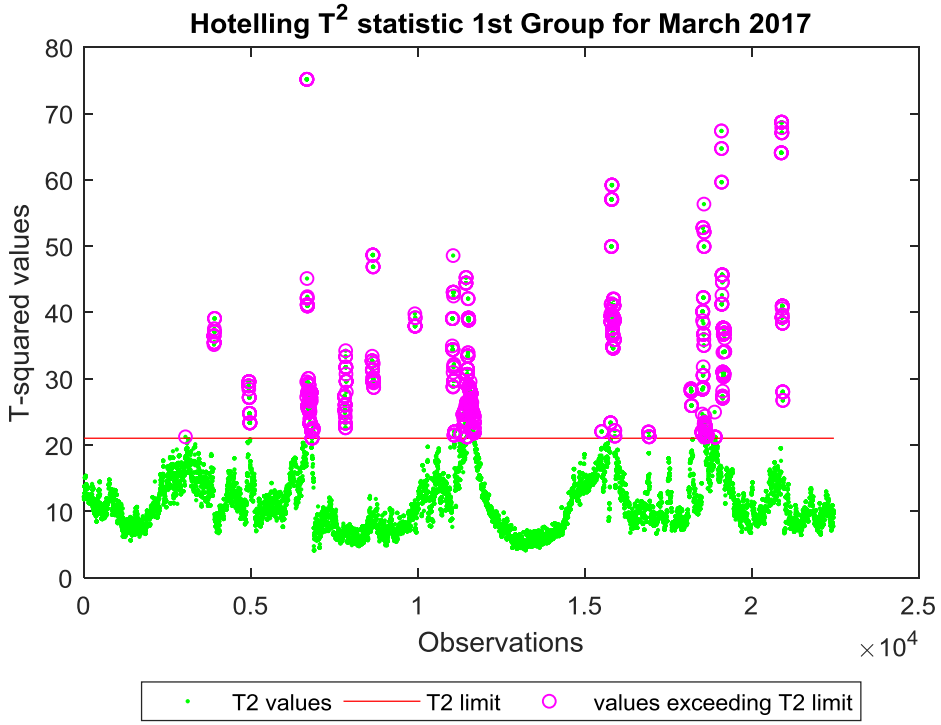


Figure B-2. Hotelling T^2 statistic for the first group of data in March 2017

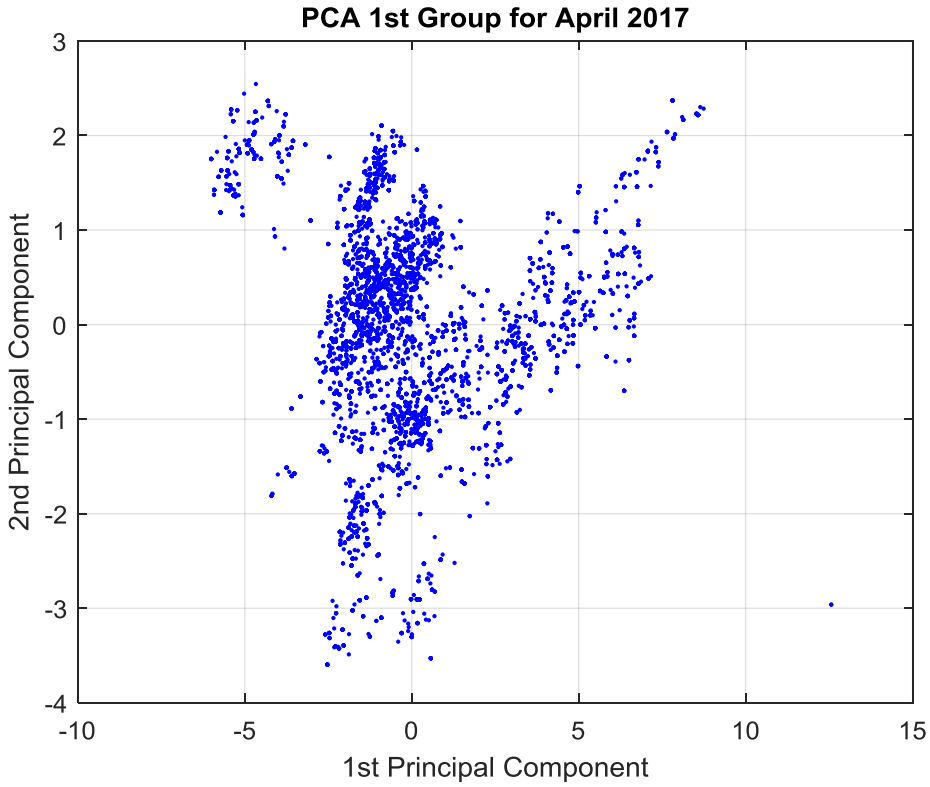


Figure B-3. Plot of the first two PCs for the first set of parameters in April 2017

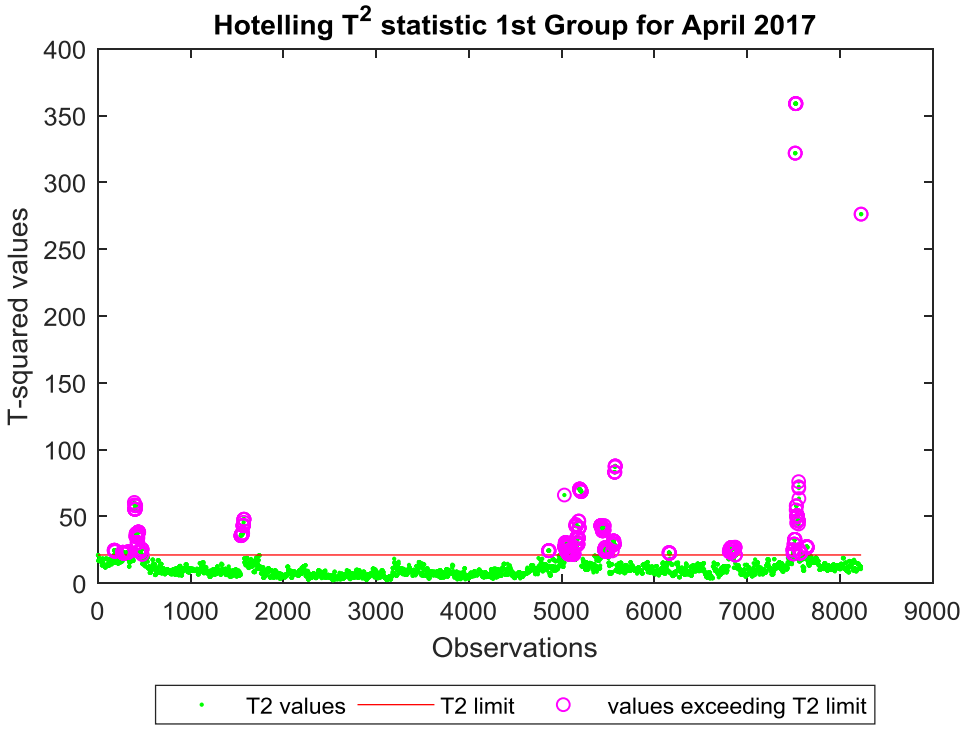


Figure B-4. Hotelling T^2 statistic for the first group of data in April 2017

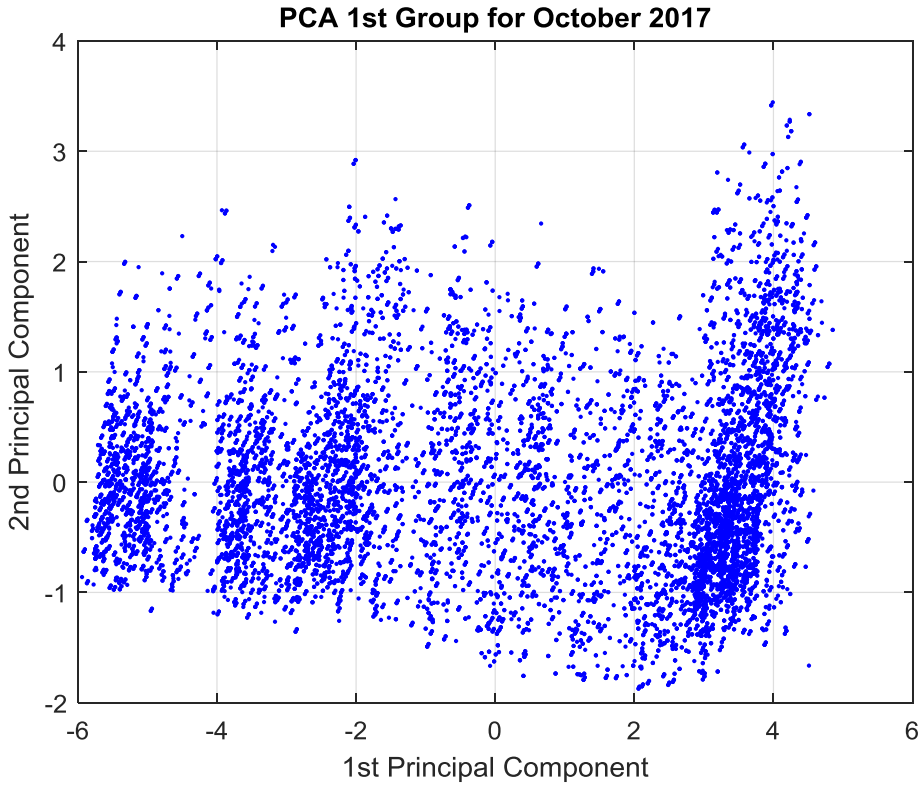


Figure B-5. Plot of the first two PCs for the first set of parameters in October 2017

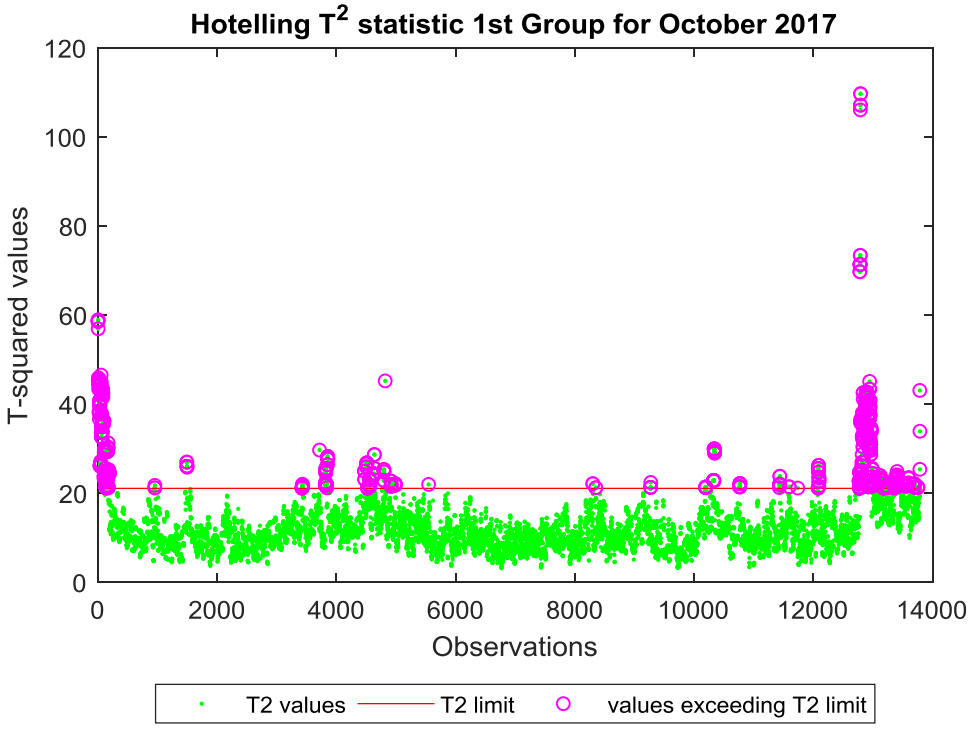


Figure B-6. Hotelling T^2 statistic for the first group of data in October 2017

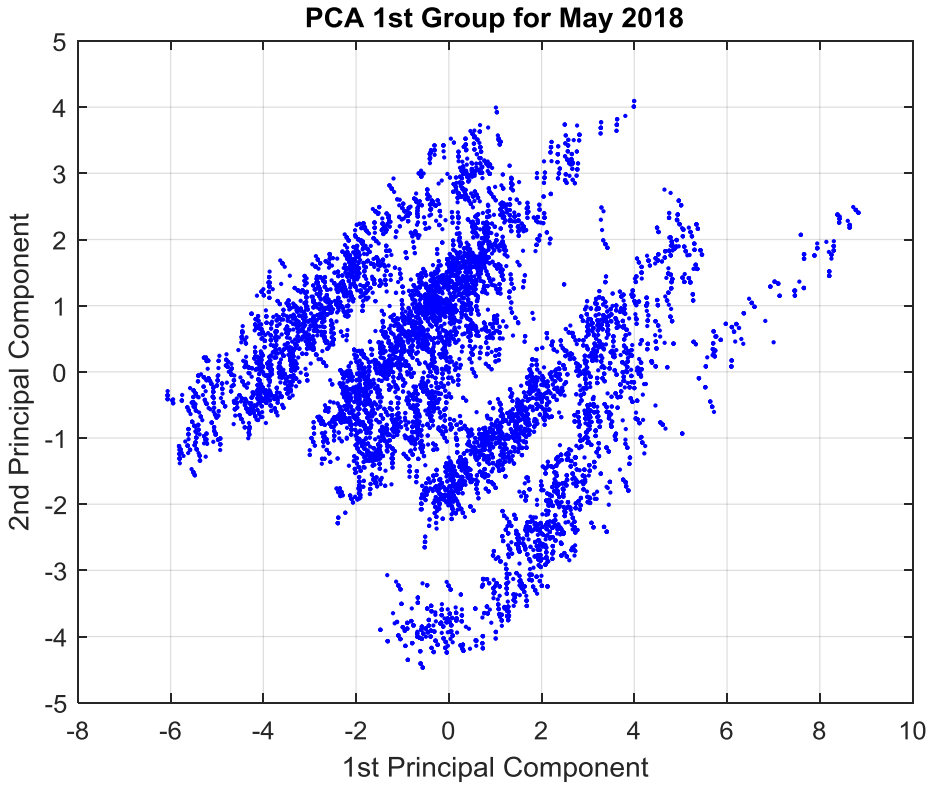


Figure B-7. Plot of the first two PCs for the first set of parameters in May 2018

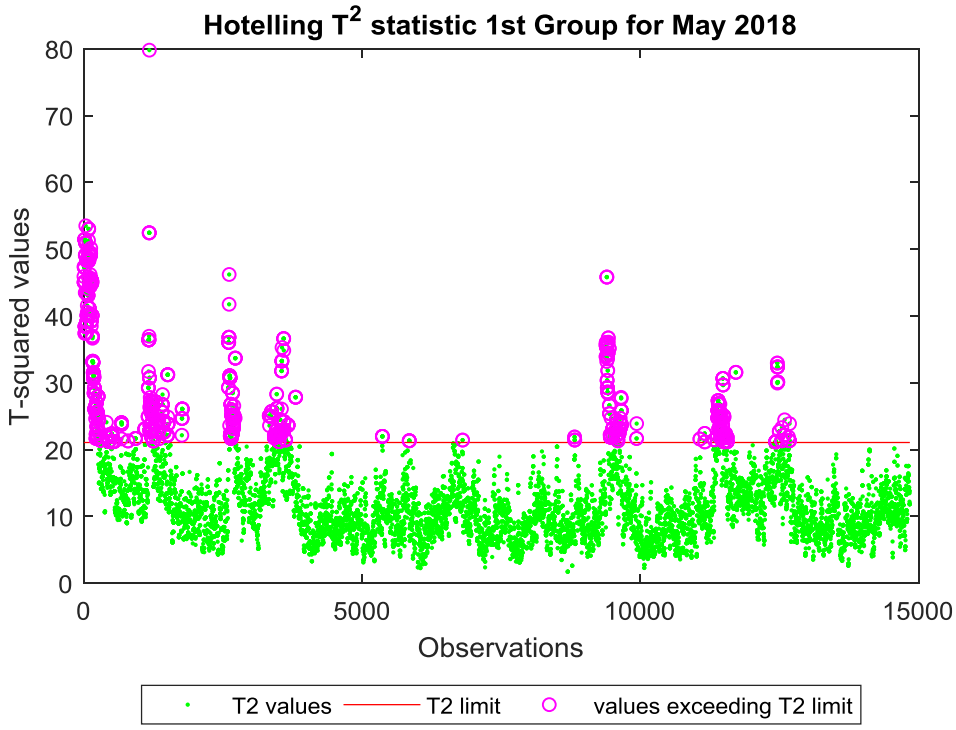


Figure B-8. Hotelling T^2 statistic for the first group of data in May 2018

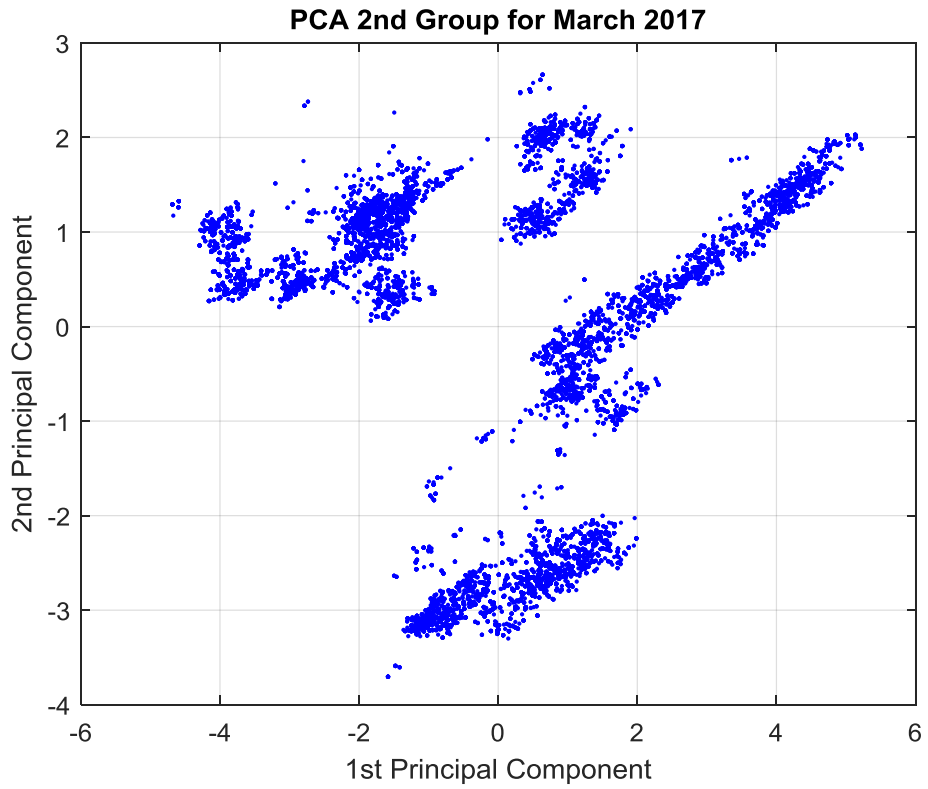


Figure B-9. Plot of the first two PCs for the second set of parameters in March 2017

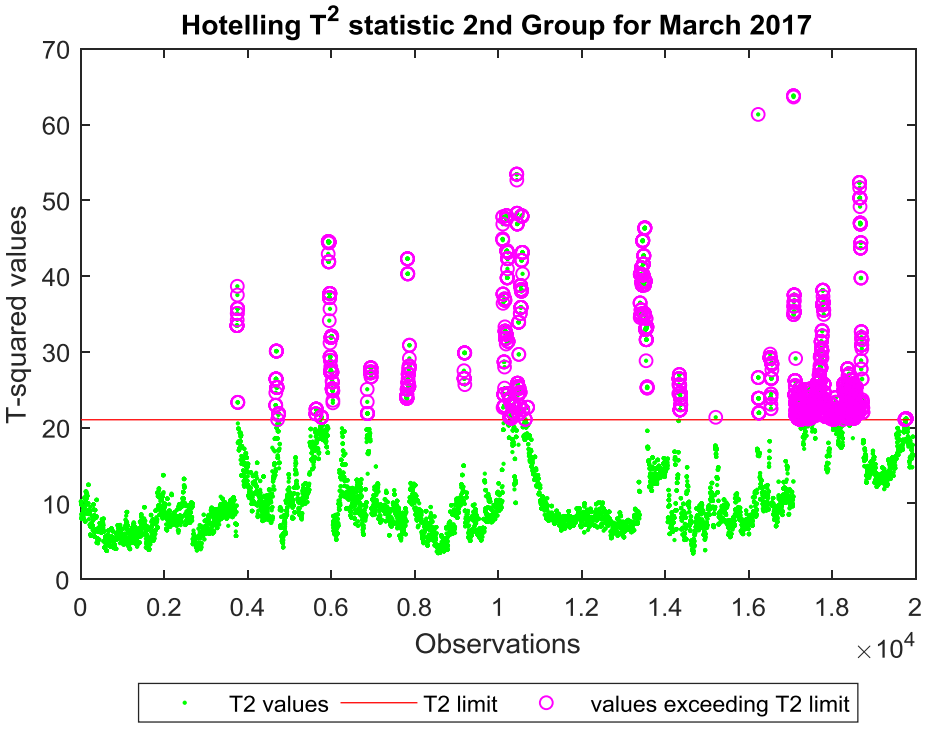


Figure B-10. Hotelling T^2 statistic for the second group of data in March 2017

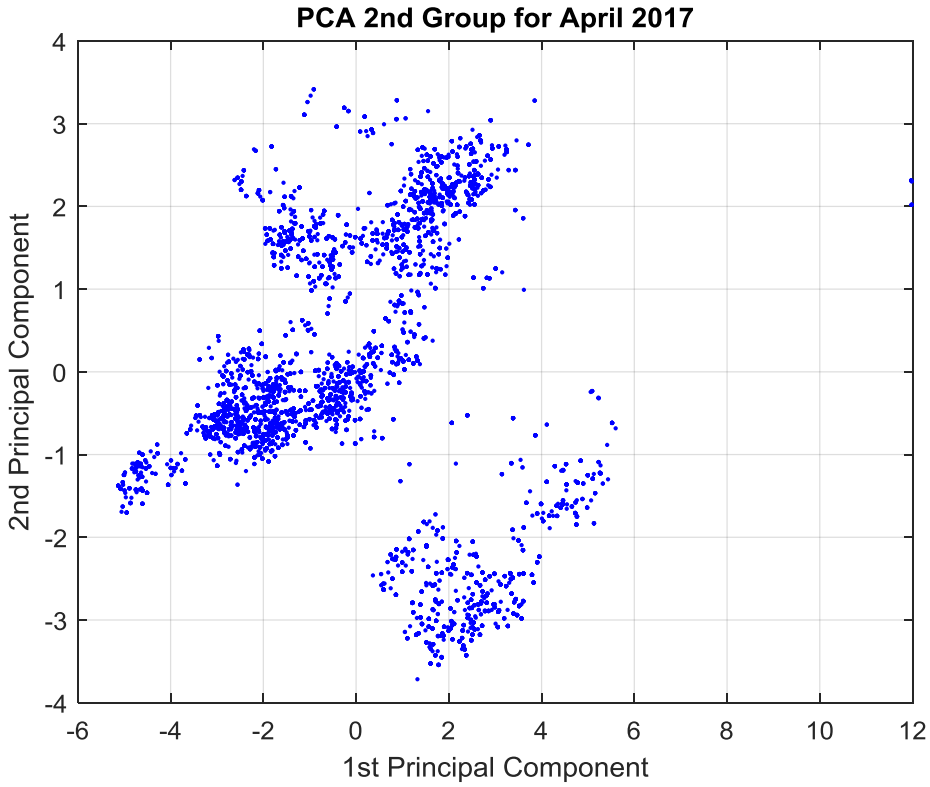


Figure B-11. Plot of the first two PCs for the second set of parameters in April 2017

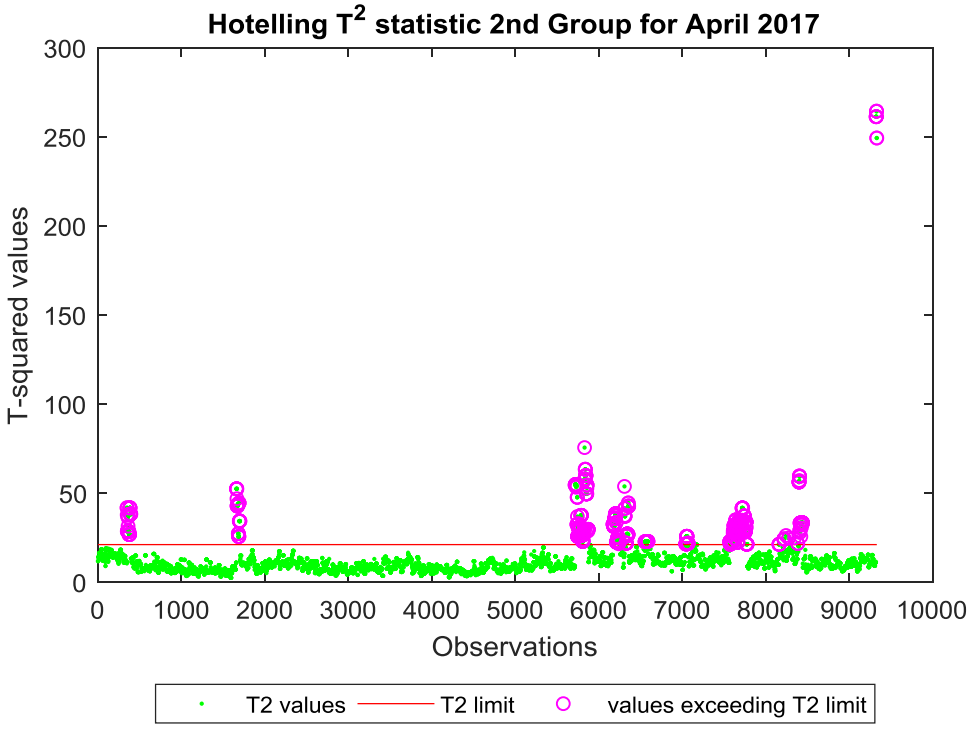


Figure B-12. Hotelling T^2 statistic for the second group of data in April 2017

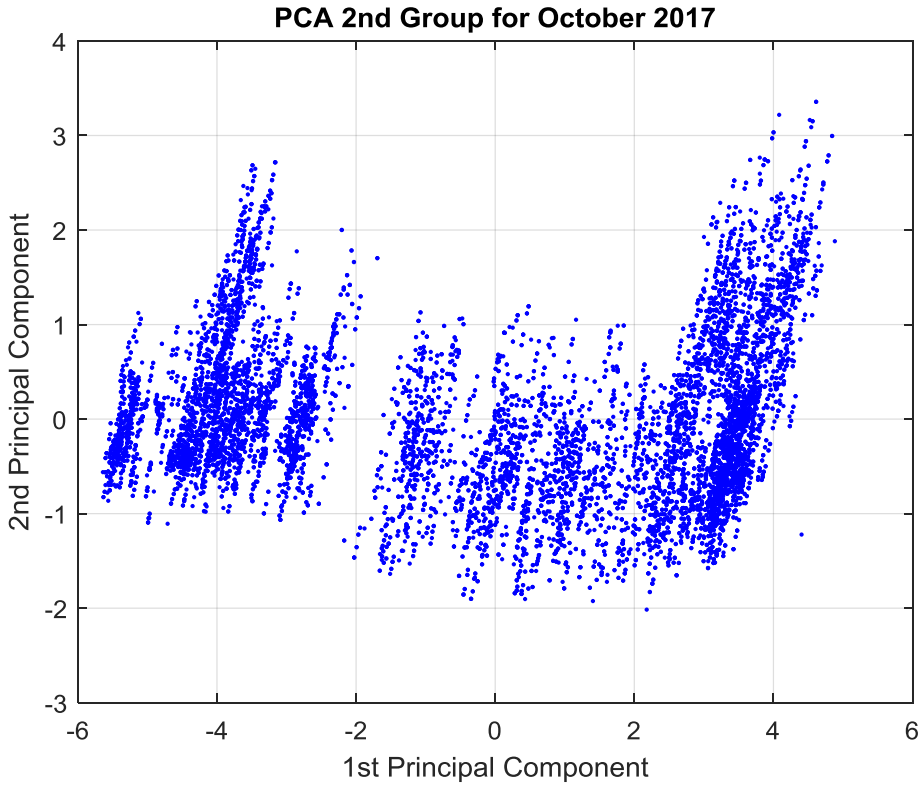


Figure B-13. Plot of the first two PCs for the second set of parameters in October 2017

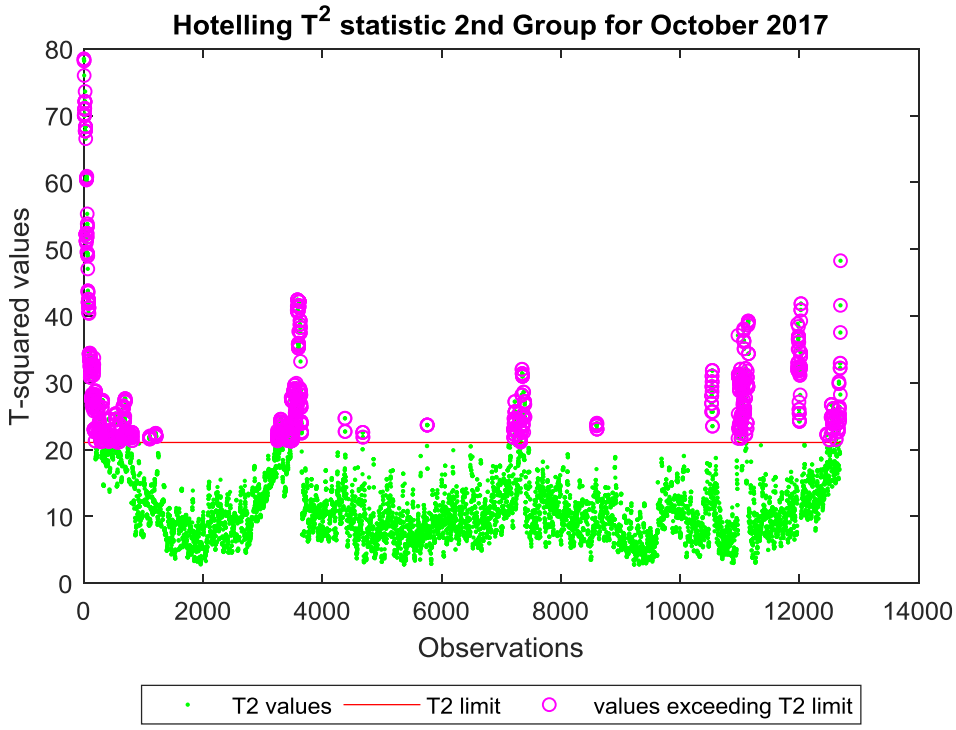


Figure B-14. Hotelling T^2 statistic for the second group of data in October 2017

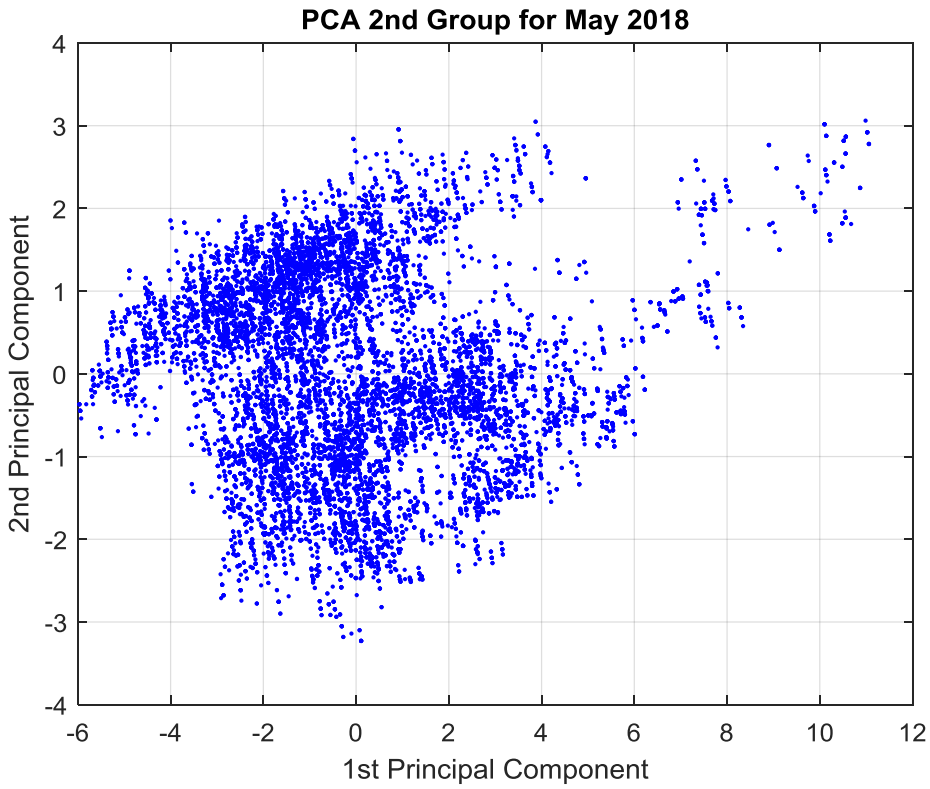


Figure B-15. Plot of the first two PCs for the second set of parameters in May 2018

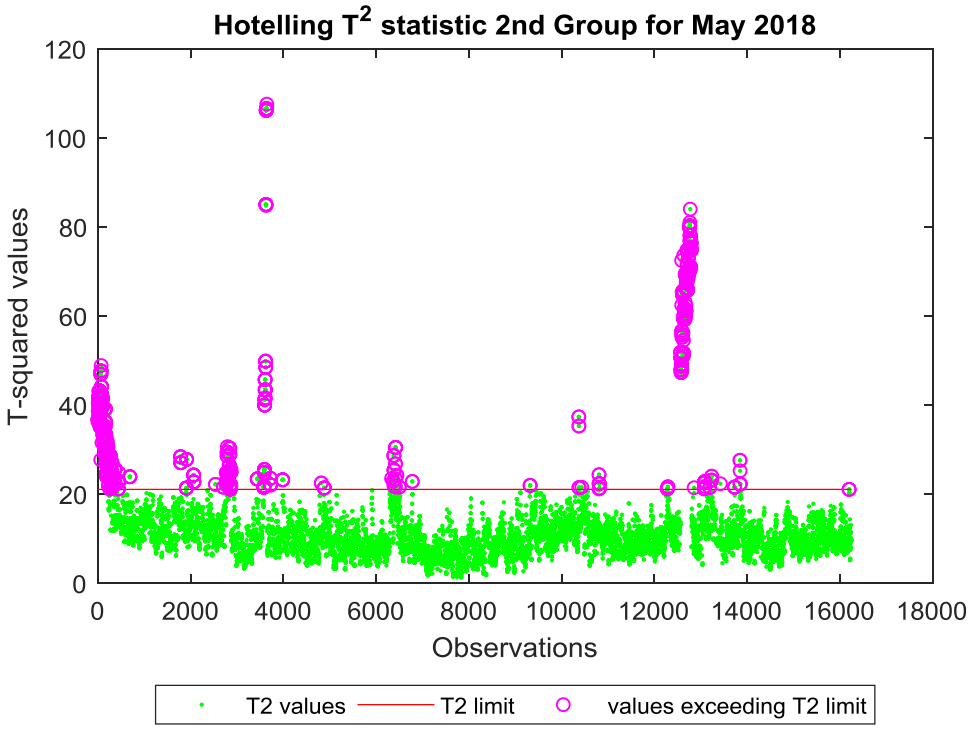


Figure B-16. Hotelling T^2 statistic for the second group of data in May 2018

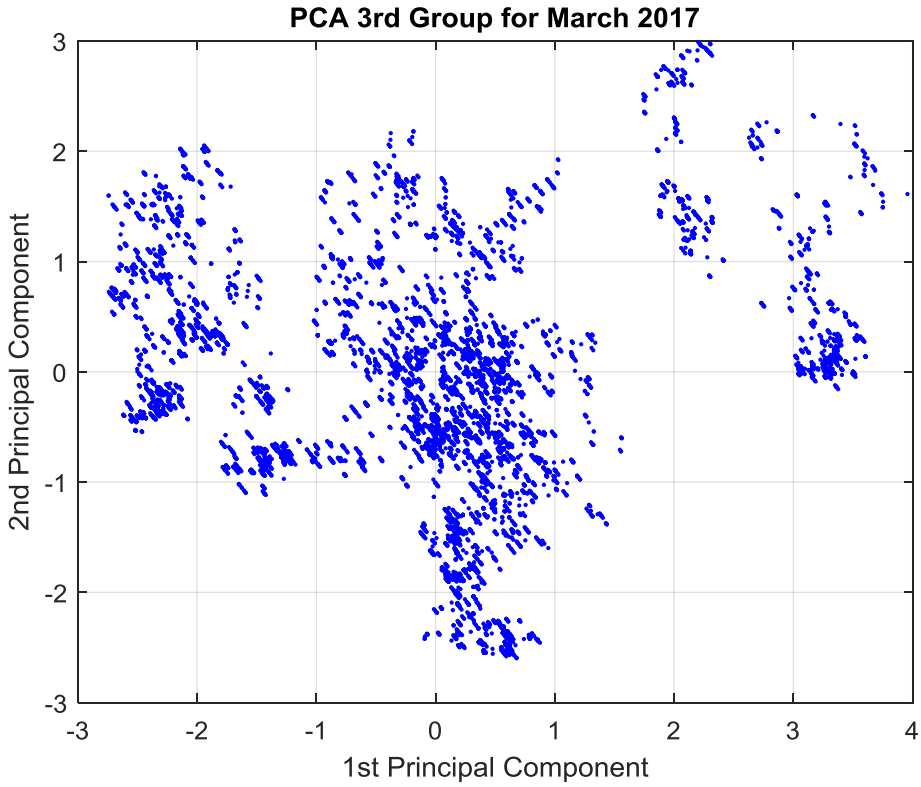


Figure B-17. Plot of the first two PCs for the third set of parameters in March 2017

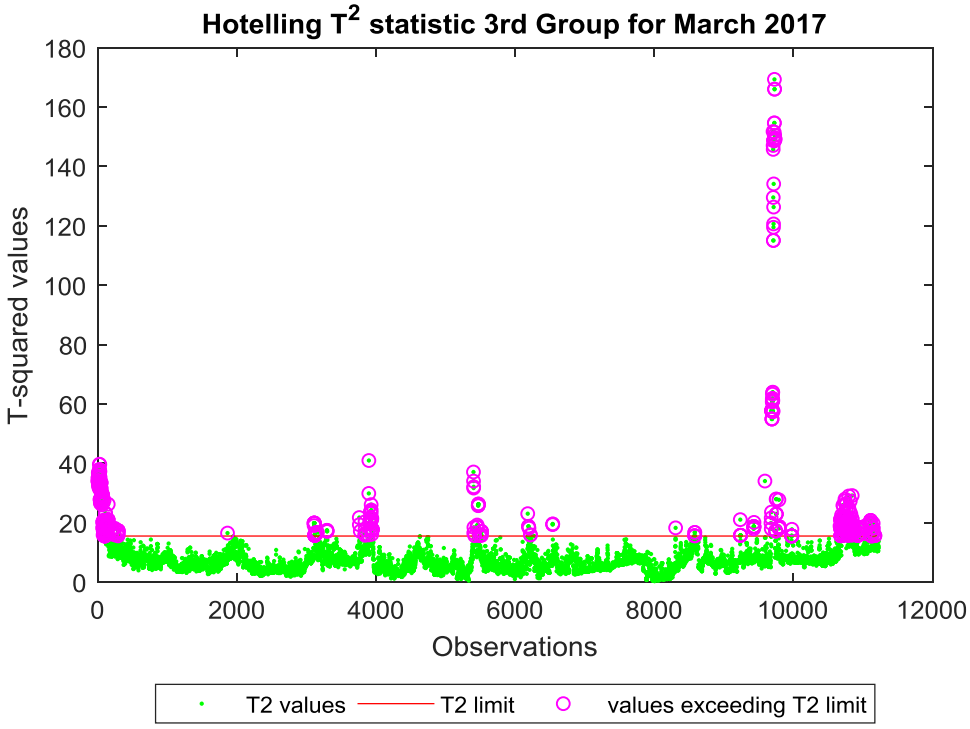


Figure B-18. Hotelling T^2 statistic for the third group of data in March 2017

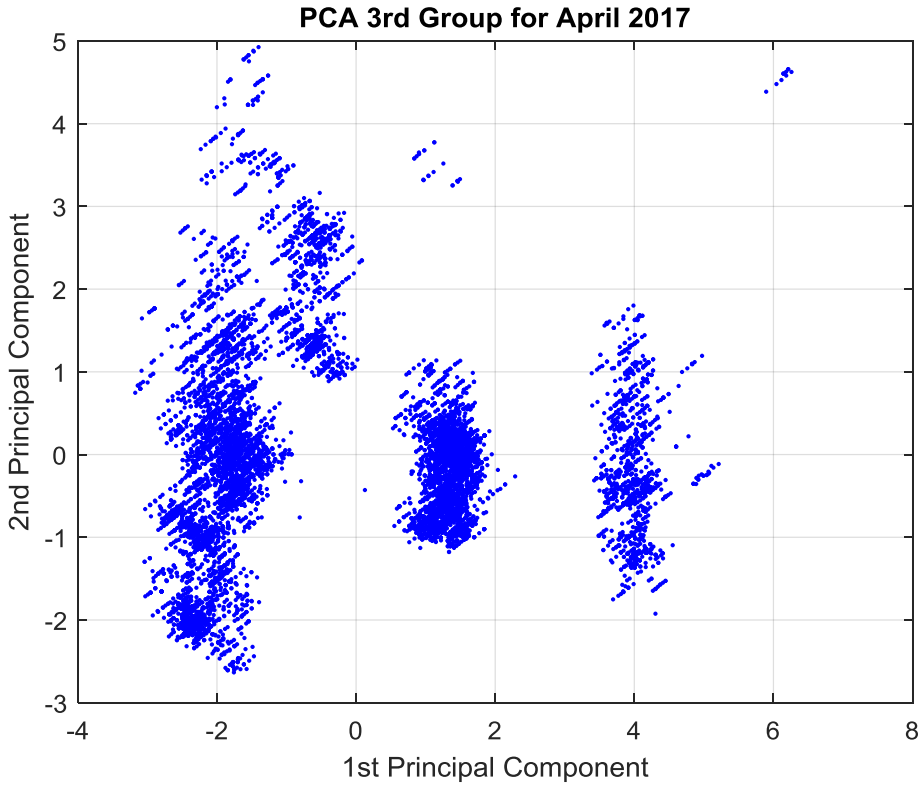


Figure B-19. Plot of the first two PCs for the third set of parameters in April 2017

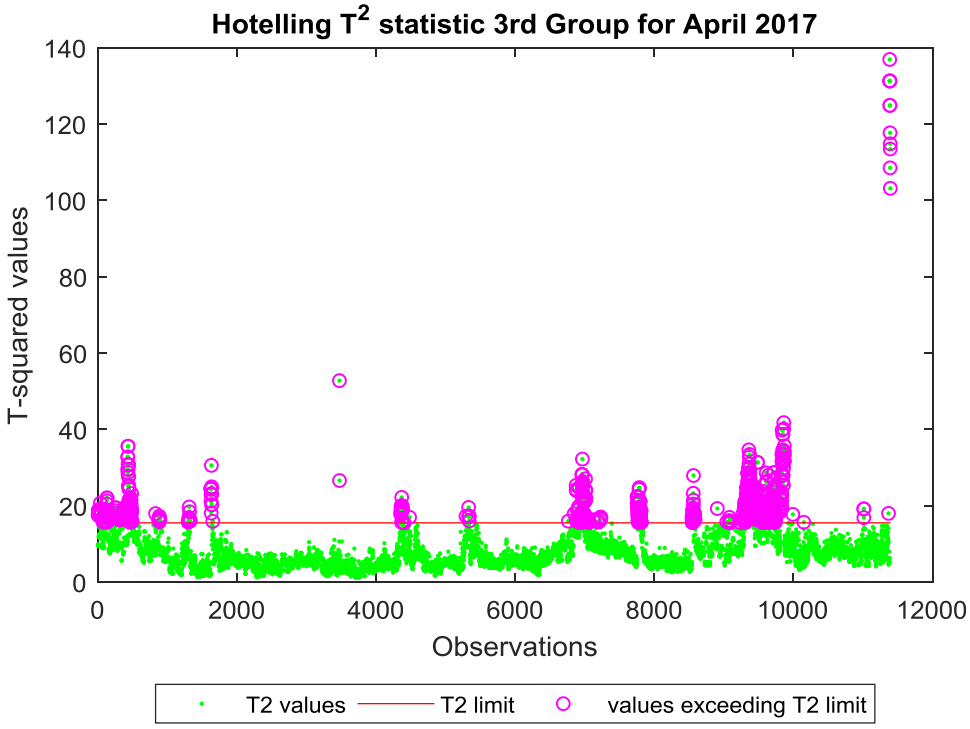


Figure B-20. Hotelling T^2 statistic for the third group of data in April 2017

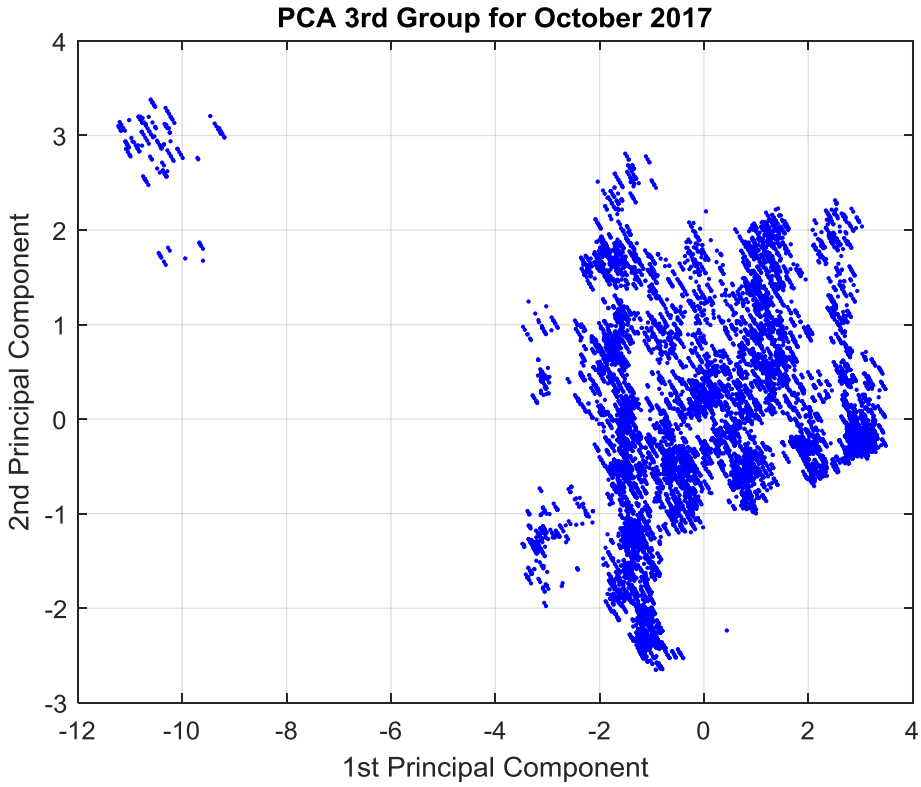


Figure B-21. Plot of the first two PCs for the third set of parameters in October 2017

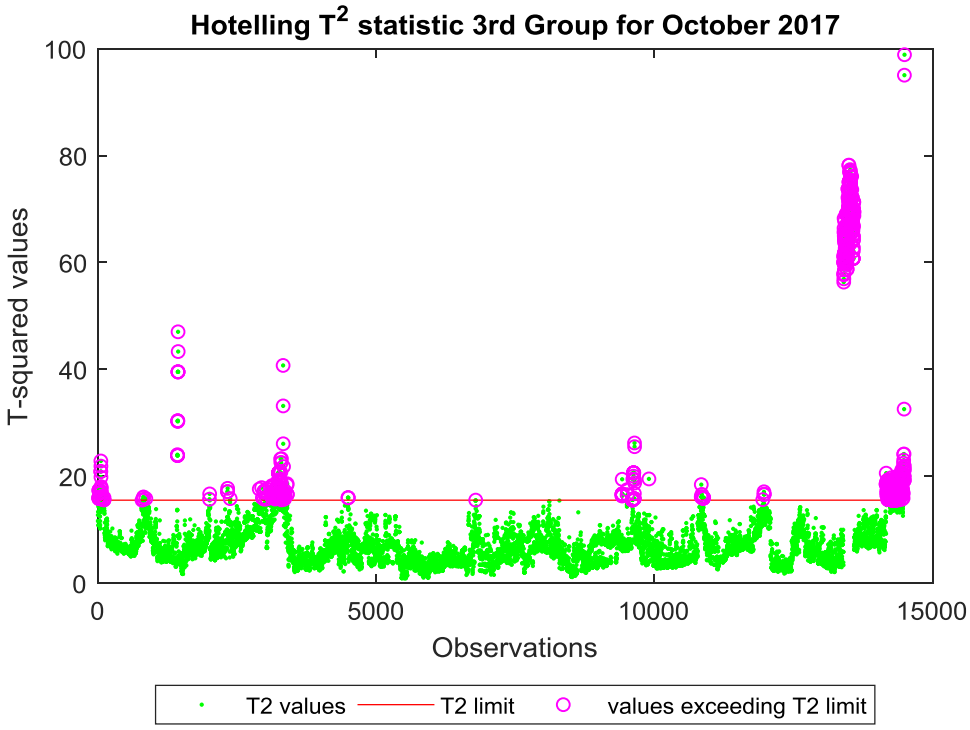


Figure B-22. Hotelling T^2 statistic for the third group of data in October 2017

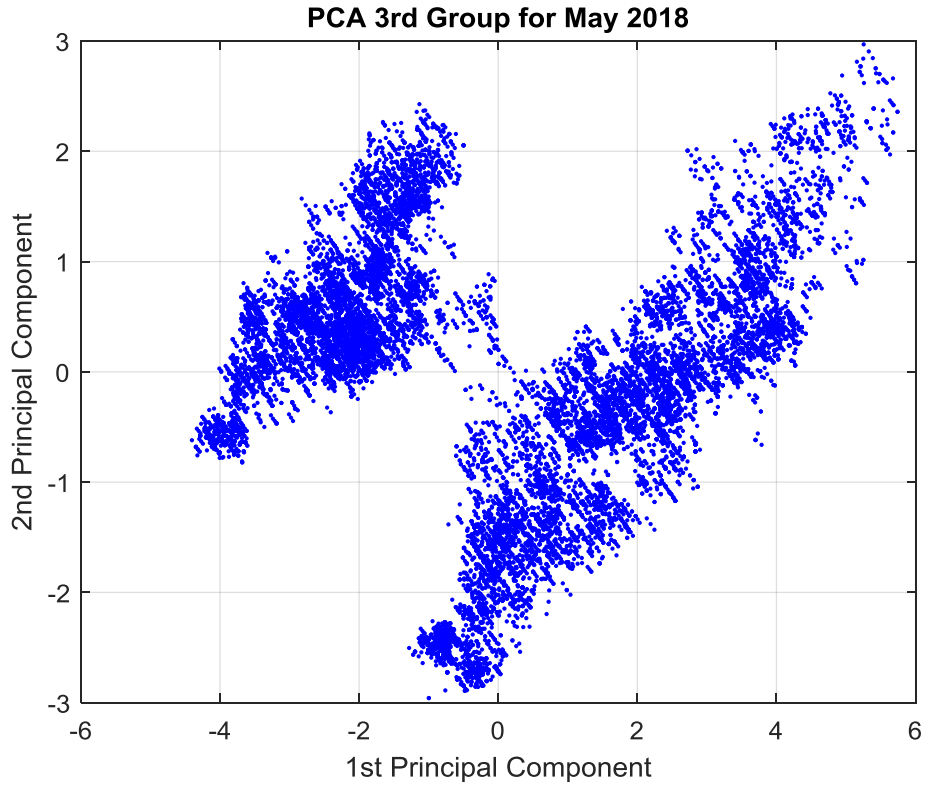


Figure B-23. Plot of the first two PCs for the third set of parameters in May 2018

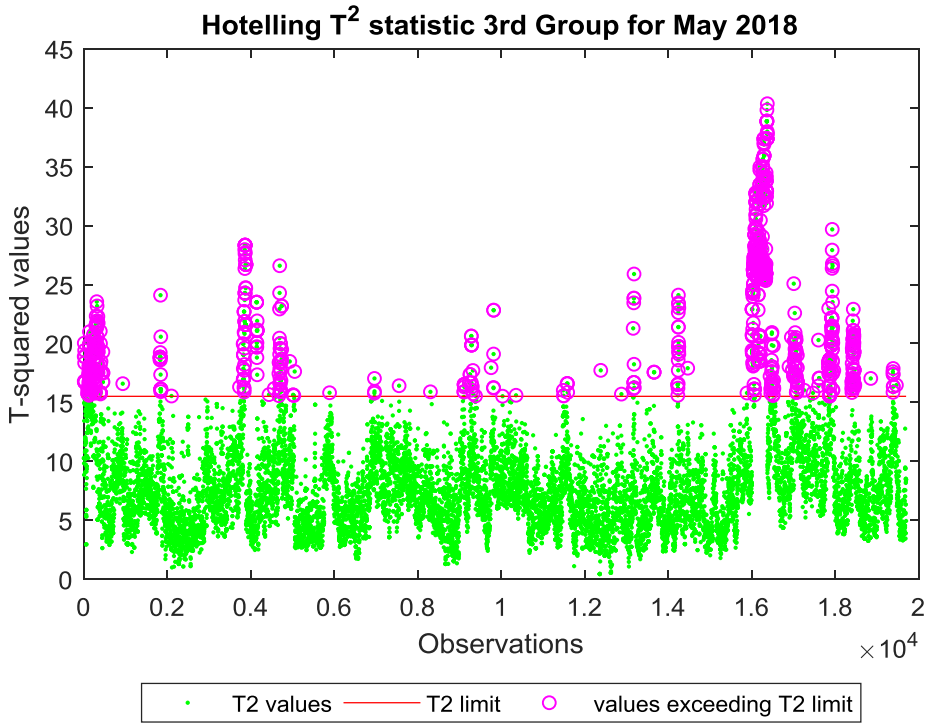


Figure B-24. Hotelling T^2 statistic for the third group of data in May 2018

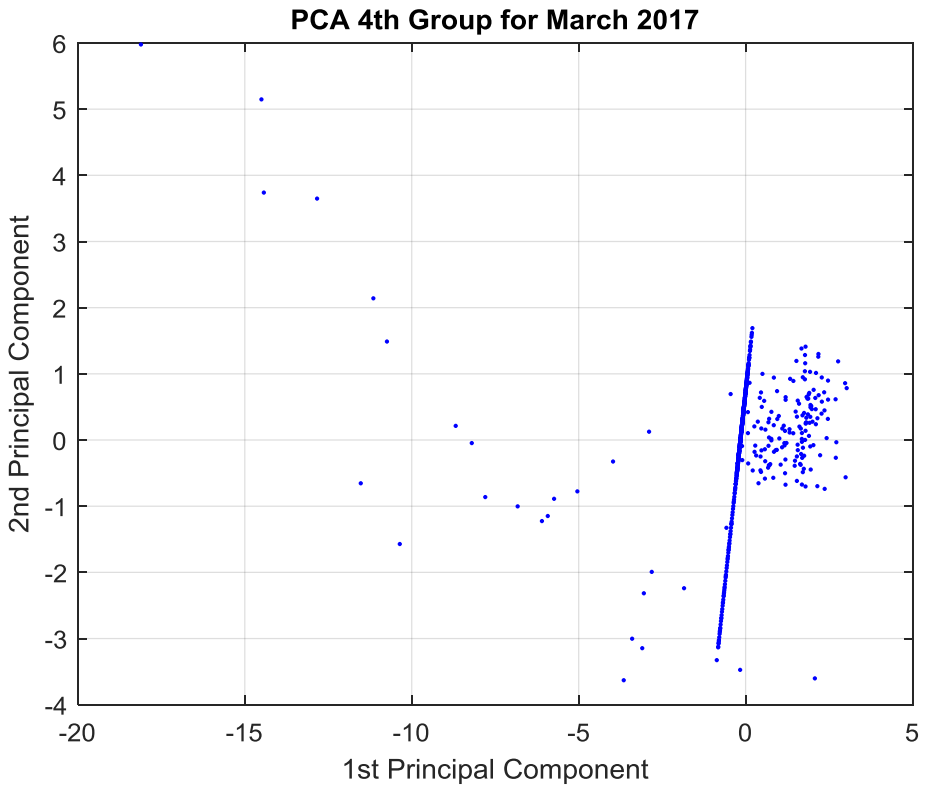


Figure B-25. Plot of the first two PCs for the fourth set of parameters in March 2017

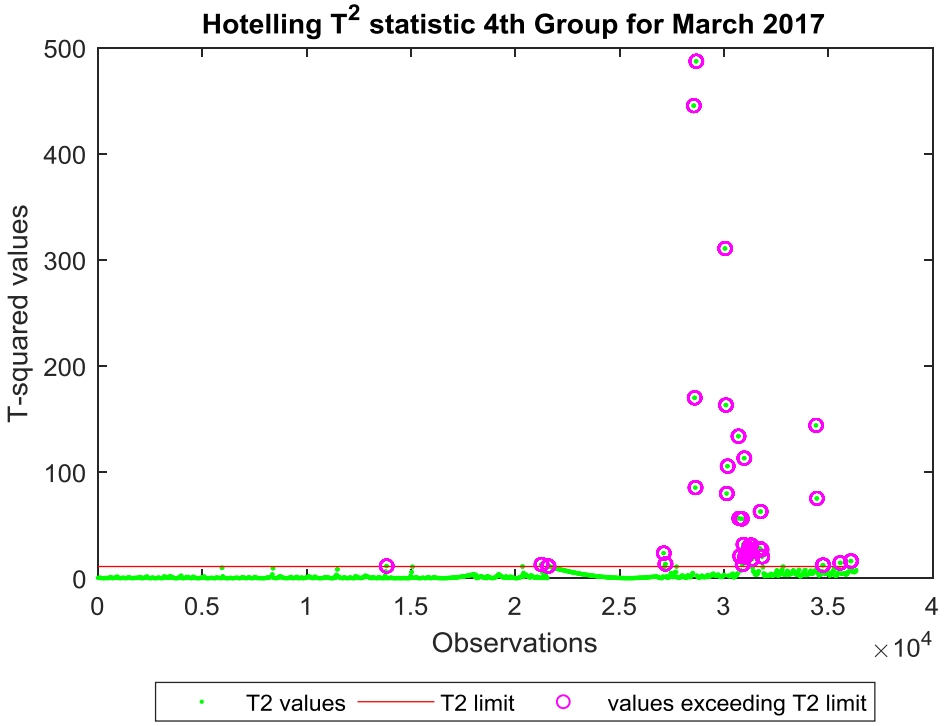


Figure B-26. Hotelling T^2 statistic for the fourth group of data in March 2017

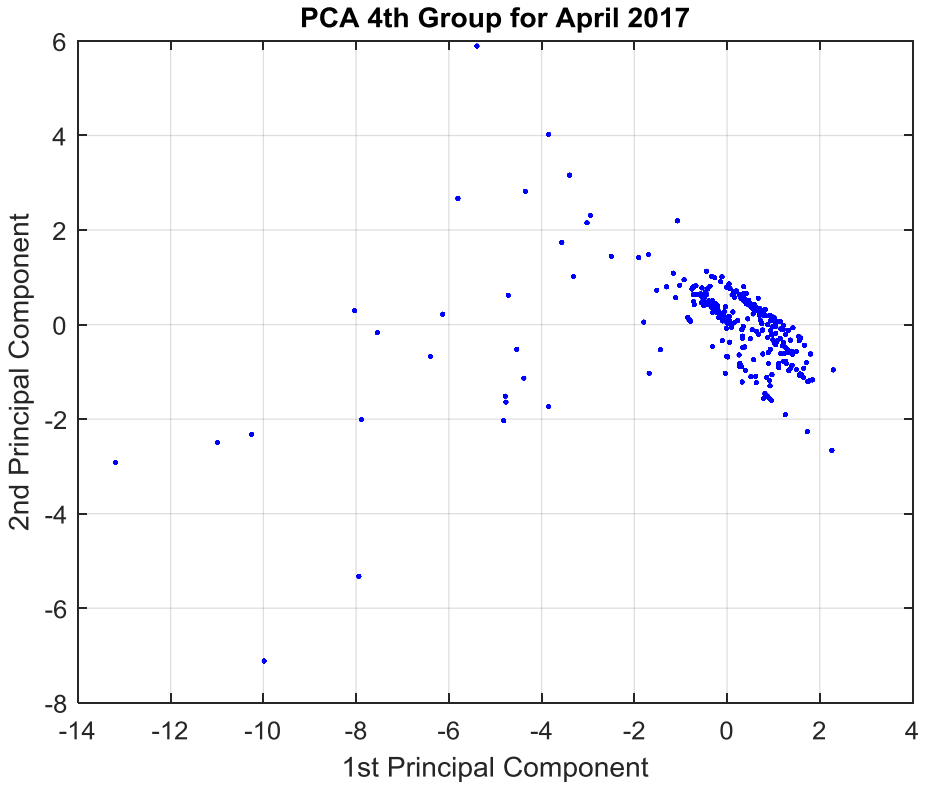


Figure B-27. Plot of the first two PCs for the fourth set of parameters in April 2017

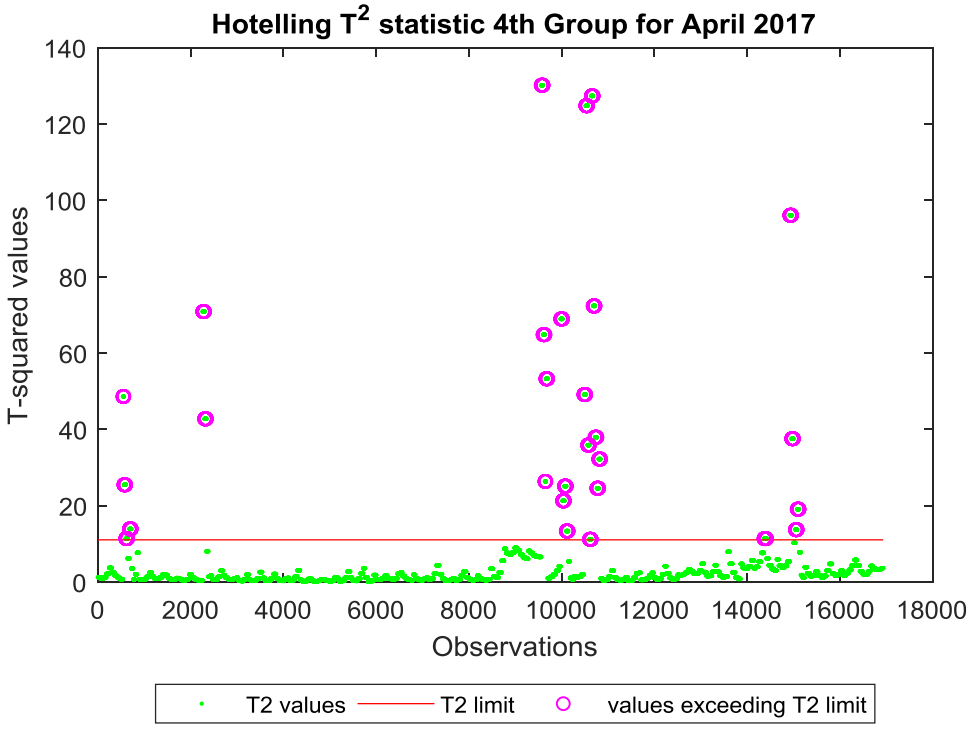


Figure B-28. Hotelling T^2 statistic for the fourth group of data in April 2017

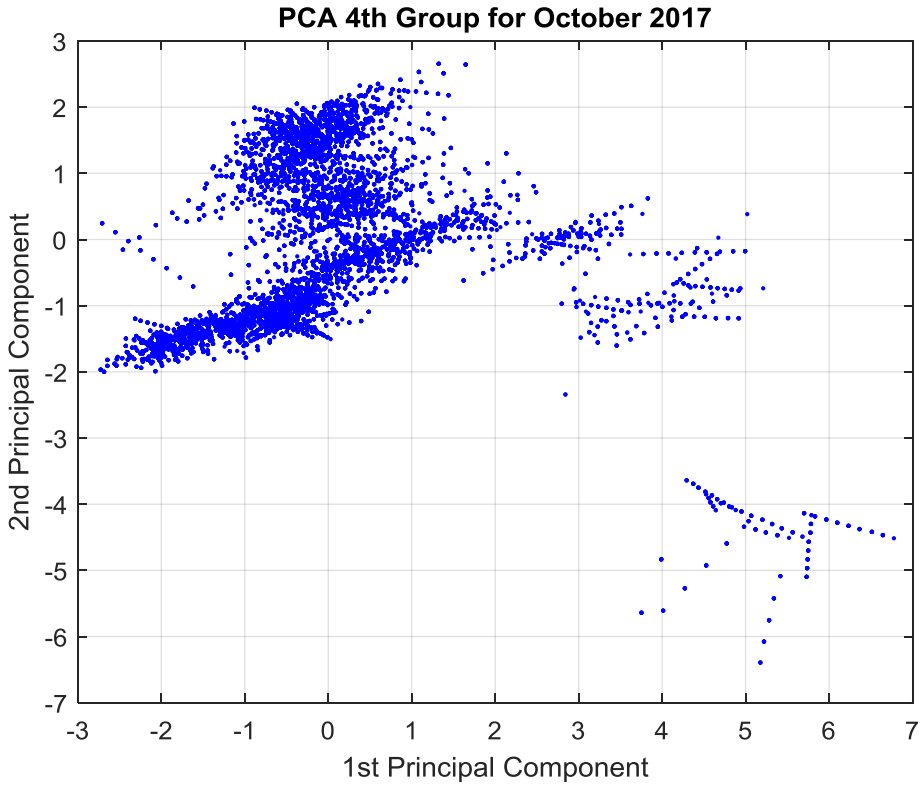


Figure B-29. Plot of the first two PCs for the fourth set of parameters in October 2017

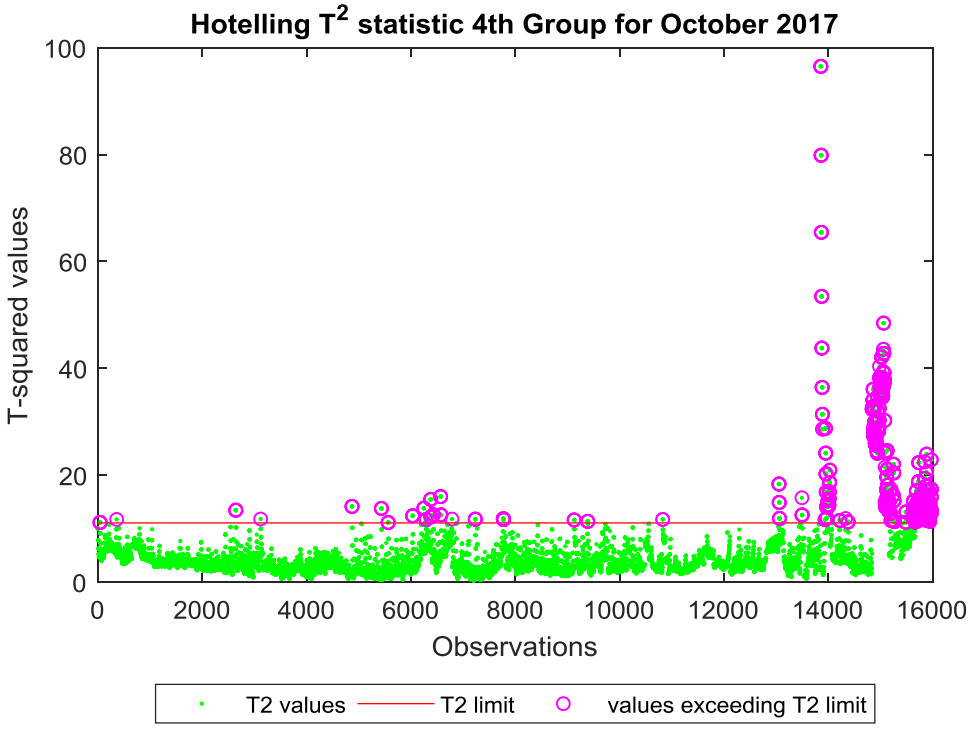


Figure B-30. Hotelling T^2 statistic for the fourth group of data in October 2017

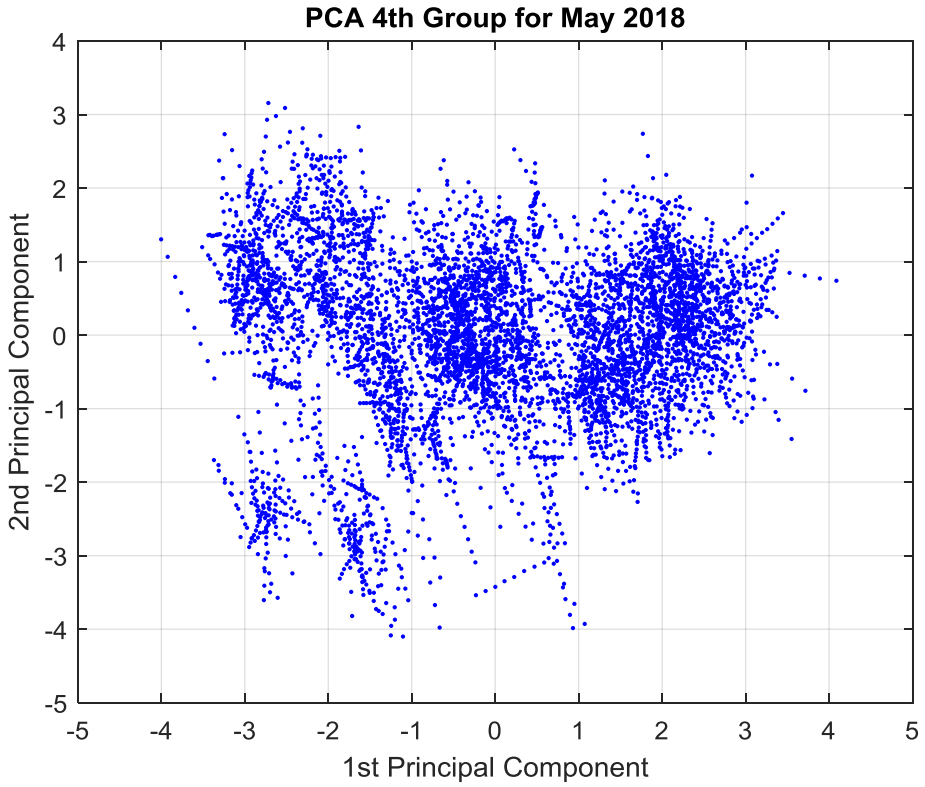


Figure B-31. Plot of the first two PCs for the fourth set of parameters in May 2018

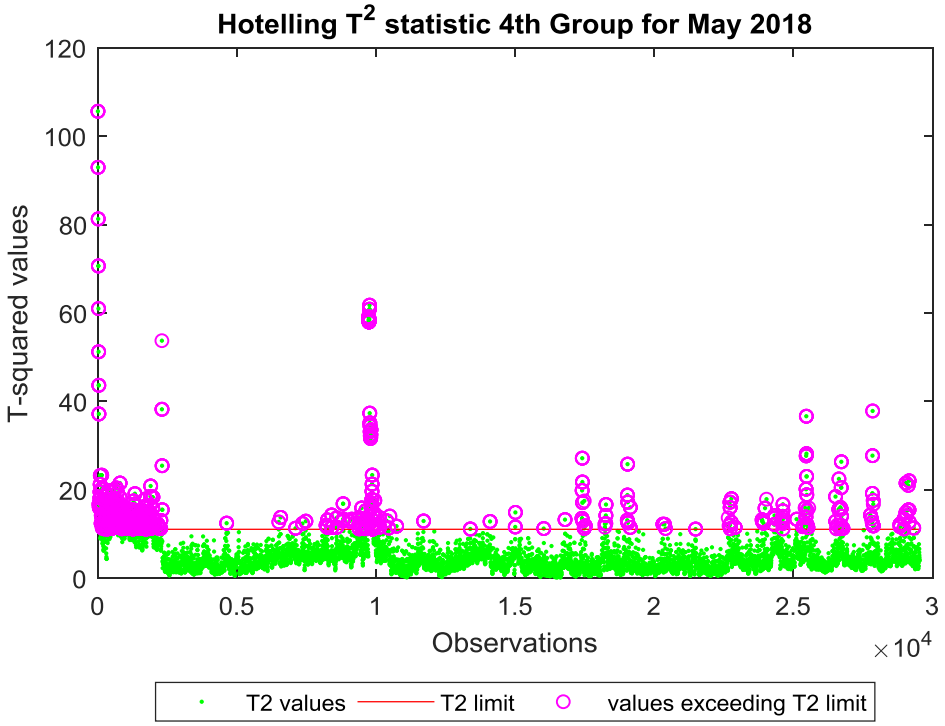


Figure B-32. Hotelling T^2 statistic for the fourth group of data in May 2018

Bibliography

[1] J. Yan, Machinery Prognostics and Prognosis Oriented Maintenance Management, John Wiley & Sons, Incorporated, 2015.

[2] A. Heng, S. Zhang, A. C. C. Tan and J. Mathew, "Rotating machinery prognostics: State of the art, challenges and opportunities," *Mechanical Systems and Signal Processing,* no. 23, pp. 724-739, 2009.

[3] M. Tahan, E. Tsoutsanis, M. Muhammad and Z. A. A. Karim, "Performance-based health monitoring, diagnostics and prognostics for condition-based maintenance of gas turbines: A review," *Applied Energy,* no. 198, pp. 122-144, 2017.

[4] J. Lee, F. Wu, W. Zhao, M. Ghaffari, L. Liao and D. Siegel, "Prognostics and health management design for rotary machinery systems-Reviews, methodology and applications," *Mechanical Systems and Signal Processing,* no. 42, pp. 314-334, 2014.

[5] A. E. Brom, I. N. Omelchenko and O. V. Belova, "Lifecycle costs for energy equipment FMECA for gas turbine," *Procedia Engineering,* no. 152, pp. 177-181, 2016.

[6] A. Carpignano, Risk Analysis, Lecture notes, 2008.

[7] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.

[8] S. M. Ross, Introduction to probability and statistics for engineers and scientists, Academic Press, 2009.

[9] P. Erto, Probabilità e statistica per le scienze e l'ingegneria, McGraw-Hill, 2008.

[10] D. C. Montgomery, Introduction to Statistical Quality Control, John Wiley & Sons, Incorporated, 2009.

[11] O. E. Dragomir, R. Gouriveau, F. Dragomir, E. Minca and N. Zerhouni, "Review of prognostic problem in condition-based maintenance," in *European Control Conference*, Budapest, Hungary, 2009.

[12] J. E. Jackson, A User's Guide To Principal Components, John Wiley & Sons, Inc., 1991.

[13] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.

[14] R. Penha and J. W. Hines, "Using principal component analysis modeling to monitor temperature sensors in a nuclear research reactor," in *Proceedings of the 2001 Maintenance and Reliability Conference (MARCON 2001)*, Knoxville, TN, USA, 2001.

[15] ISO 17359, Condition monitoring and diagnostics of machines - General guidelines,

2011.

[16] [Online]. Available: https://www.wartsila.com. [Accessed 27 May 2019].

[17] [Online]. Available: https://new.abb.com. [Accessed 30 May 2019].