A Clustering Approach for Mining Reliability Big Data for Asset Management

Francesco Cannarile[1,2], Michele Compare[1,2], Francesco Di Maio[1], Enrico Zio[1,2,3]

[1]*Energy Department, Politecnico di Milano, Via la Masa 34, 20156 Milano, Italy*

[2] *Aramis Srl, Via pergolesi 5, Milano, Italy*

[3] *Chair on System Science and the Energetic Challenge, Fondation EDF, Centrale Paris and Supelec, Paris, France*

ABSTRACT

Big data from very large fleets of assets challenge the asset management, as the number of maintenance strategies to optimize and administrate may become very large. To address this issue, we exploit a clustering approach that identifies a small number of sets of assets with similar reliability behaviors. This enables addressing the maintenance strategy optimization issue once for all the assets belonging to the same cluster and, thus, introduces a strong simplification in the asset management. However, the clustering approach may lead to additional maintenance costs, due to the loss of refinement in the cluster reliability model. For this, we propose a cost model to support asset managers in trading-off the simplification brought by the cluster-based approach against the related extra-costs. The proposed approach is applied to a real case study concerning a set of more than 30000 switch point machines.

KEY WORDS: Big Data; Spectral Clustering; Preventive Maintenance.

1. INTRODUCTION

Managing large assets with numerous (e.g., millions) assemblies is nowadays supported by sophisticated Enterprise Resource Planning (ERP) systems, which allow collecting and storing many and diverse data about the asset lives such as their failures and maintenance times, operating and ambient conditions, etc. On the one hand, the increasing capabilities of the ERP systems offer opportunities for new developments in reliability and maintenance engineering, as they provide a sound basis that enables the application of stronger statistical methods supporting more informed predictions and, thus, operations of the asset behaviors [1], [2]. On the other hand, the tracking of millions of assets challenges the asset management, because the full exploitation of the available (big) data requires the maintenance departments to perform expensive analyses and because of the need to handle large (petabytes) databases of possibly unstructured heterogeneous data [3]. In this work, we focus on the first issue, only, while not addressing the computer science perspective, for which the interested reader can refer to [4]. For big data analysis, innovative approaches for data management and mining are required to the benefit of process optimization and decision making [3] in different areas of application, which include computing, telecommunications, mobile services, manufacturing, process industries and railway [3], [5]. For example, the latter railway sector today handles a huge quantity of information from different types of data sources (e.g., unstructured text, signaling or train data streams) that could be used to improve the understanding of risk factors involved in operation [6], [7], and to optimize the assets maintenance [3].

For example, the approach to maintain the assets of the Italian railway system is based on the definition of segmented populations of components and systems in relation to some technical information (such as rail type, switch point machine model, etc.) and/or geographic localization, and on the optimization of the maintenance for every population. Typically, the number of populations turns out to be quite large, with consequent difficulties in managing all different strategies and the related administrative activities. This becomes even more complicated if we consider, that in principle, the maintenance strategies have to be periodically updated to give due account to newly collected data, which may reveal changes in the components hazard rate values.

It is, then, that a sound methodological framework is needed to allow the exploitation of the available data with manageable efforts for the maintenance engineering department. In [3], an exhaustive review of integrated maintenance processes in the railway sector is provided, with emphasis on the importance of these processes and the need of computer-based maintenance systems for the management of big data from multiple sources. In [12], a Support Vector Machine (SVM) framework is proposed for tackling fault detection of the braking system in a high speed train from highly unbalanced data. In [8], a method to discover from data temporal association rules leading to rare events requiring immediate maintenance actions is proposed. In [9], an optimized OnLine Support Vector Regression (OL-SVR) for condition-based maintenance is proposed for streaming analysis of big data in the context of rail transportation systems. In [10], a clustering algorithm for grouping segmented asset populations based on their reliability distributions is proposed, with the goal that clusters of assets with similar failure behavior will be subject to the same (optimal) maintenance strategy. Although the approach proposed in [10] strongly reduces the number of strategies to be handled, it does not give full account to the drawback of the simplification introduced by forcing the different reliability distributions in a cluster to be approximated by one representative of the entire cluster.

In this respect, the present work proposes a method for evaluating whether the simplification brought by the clustering may lead to maintenance extra-costs. This is fundamental for the asset decision maker, who has to trade-off the possible economic loss against the expected savings coming from the management simplification.

The remainder of the paper is organized as follows: Section 2 briefly recalls the methodology to cluster assets based on their reliability distributions. In Section 3, we detail the cost model. Section 4 presents the case study concerning assets of the Italian railways. Finally, in Section 5 some conclusions are drawn. Notice that although the case study discussed in this work is derived from a real industrial application, the data shown have been opportunely re-scaled and modified to respect the non-disclosure agreement with the industrial partner.

2. CLUSTERING

In this Section, we briefly recall the clustering approach developed in [10], whose objective is to group a large number A of assets into C^*clusters, $C^* \ll A$, based on their reliability behaviors. This way, the asset manager can reduce the number of strategies to implement and trace from A to C^*.

We assume that the asset manager can decide among two maintenance strategies: Periodic Maintenance (PM, i.e., assets are preventively maintained at some predetermined periodic times or repaired at failure, whichever comes first) or Corrective Maintenance (CM, i.e., assets are operated until failure) [11].

A perfect maintenance action is performed upon asset failure or preventive task. Then, the asset can be considered “As Good As New” (AGAN) after any maintenance intervention [12].

We suppose that for every asset $a \in \{1, \dots, A\}$ the following pieces of information are available:

1. the values of K control variables $(X_1, \dots, X_K)$ containing technical information about the asset (e.g., its location, the type of railway line, etc.). These are arranged into vectors $\boldsymbol{x}_a = (x_a^1, \dots, x_a^K), a = 1, \dots, A$.
2. A collection of n_a independent field observations registered into the vector $\boldsymbol{D}_a = (\boldsymbol{y}_a, \boldsymbol{\delta}_a)$, where $\boldsymbol{y}_a = \left(y_a^1, \dots, y_a^{n_a}\right)$, is a time variable, $y_a^b \in \mathbb{R}_0^+, b = 1, \dots, n_a$, whereas $\boldsymbol{\delta}_a = \left(\delta_a^1, \dots, \delta_a^{n_a}\right), \delta_a^b \in \{0,1\}, b = 1, \dots, n_a$ is a censoring indicator variable: y_a^b is a current failure time if δ_a^b is equal to 1, or a right-censored observation if δ_a^b is equal to 0, $b \in \{1, \dots, n_a\}$.

2.1. CLUSTERING METHODOLOGY SNAPSHOT

The methodology proposed in [8] is based on the following steps:

1. Based on the knowledge of experts and on considerations pertaining to the organization of the maintenance engineering department of the industrial partner, identify the subset $\left(\tilde{X}_1, \dots, \tilde{X}_{\tilde{K}}\right)$ of decision variables $(X_1, \dots, X_K)$, $\widetilde{K} < K$, which allows partitioning the assets in $N < A$ segmented populations S_i, corresponding to different combinations of decision variables $\left(\tilde{X}_1, \dots, \tilde{X}_{\tilde{K}}\right)$. Thus, each population S_i is associated to the failure dataset $\boldsymbol{O}_i = (\boldsymbol{D}_{i_1}, \dots, \boldsymbol{D}_{i_{n_i}})$, where $\left\{i_1, \dots, i_{n_i}\right\} \subseteq \{1, \dots, A\}$ are the indexes identifying the assets belonging to population S_i, $i = 1 \dots, N$. Notice that the identification of these populations differs from clustering analysis, which, indeed, aims at grouping these populations based on their reliability distributions.

In particular, the reliability distribution of population S_i is assumed to be a Weibull distribution of scale parameter α_i and shape parameter $\beta_i, i = 1, .., N$. The probability density function is given by

$$f_i(y|\alpha_i, \beta_i) = \frac{\beta_i}{\alpha_i}\left(\frac{y}{\alpha_i}\right)^{\beta_i - 1} \quad y > 0, \alpha_i > 0, \beta_i > 0 \tag{1}$$

whereas the corresponding reliability function and hazard rates are, respectively:

$$R_i(y|\alpha_i, \beta_i) = e^{-\left(\frac{y}{\alpha_i}\right)^{\beta_i}} \quad y > 0, \alpha_i > 0, \beta_i > 0 \tag{2}$$

and

$$h_i(y|\alpha_i, \beta_i) = \frac{\beta_i}{\alpha_i}\left(\frac{y}{\alpha_i}\right)^{\beta_i} \quad y > 0, \alpha_i > 0, \beta_i > 0 \tag{3}$$

It is worth mentioning that the results of clustering depend on the subset of decision variables chosen by the experts at this first step, because different partitioning solutions of the assets lead to different reliability behaviors, and, therefore, different similarity values to be fed to the clustering algorithm. Then, a solid and rational selection of the decision variables by the experts is fundamental for setting the maintenance strategies.

2. Apply the Maximum Likelihood Estimation (MLE) technique to each population *i* to estimate the parameters (α_i, β_i) of Eqs. (1), (2) and (3). For this, in this work we assume that there exists at least one observed failure time into the failure dataset $\boldsymbol{O}_i$ (i.e., not subject to any kind of censoring) so that MLE exists for every $i = 1 \dots, N$. In many real applications, the existence of the MLE may not be guaranteed: in these cases, one can use either Bayesian statistical inference techniques [13],

or work on the $\widetilde{K}$ covariates to reduce the number of possible populations, each one consisting of a larger number of assets.

3. Quantify the similarity between all pairs of statistical populations from the reliability perspective. This task is achieved quantifying how similar the reliability distributions are (either Eq. (1) or Eq. (2)) of each pair of statistical populations i and j, where $i, j = 1, \dots, N$. The Symmetric Kullback-Leibler Dissimilarity (SKLD) is here adopted to compute the similarity w_{ij} between densities f_i and f_j representative of the reliability distributions of statistical populations i and j, respectively. This point of the methodology is detailed in Appendix A.
4. The similarity matrix W, whose entries are given by similarities w_{ij}, is given in input to the Spectral Clustering Algorithm (SCA). This point of the methodology is detailed in Appendix B.
5. Infer the best number of clusters C^* quantifying a compromise between the silhouette [14] and Davies-Bauldin coefficients [15].
6. Create the failure time datasets $\boldsymbol{O}_p = (\boldsymbol{O}_{p_1}, \dots, \boldsymbol{O}_{p_{n_p}})$, where $\{p_1, \dots, p_{n_p}\}$ are the indexes referring to the statistical population assigned to cluster $p \in \{1, \dots, C^*\}$. Again, each cluster is assumed to be Weibull distributed with scale parameter α_p and shape parameter β_p.
7. Apply the MLE technique to estimate parameters (α_p, β_p) for each cluster.

3. ASSET MANAGEMENT

The occurrence of failures of assets belonging to the population S_i can be modeled by a Renewal Process (RP) with Renewal Function (RF) [12], $i = 1, \dots N$:

$$H_i(y|\alpha_i, \beta_i) = 1 - R_i(y|\alpha_i, \beta_i) + \int_0^y H_i(y - z|\alpha_i, \beta_i) f_i(y|\alpha_i, \beta_i)\, dz \tag{4}$$

where $H_i(y|\alpha_i, \beta_i)$ indicates the expected number of replacements up to time y for an asset belonging to S_i.

In case of Weibull distribution, the RF in Eq. (4) cannot be analytically solved; thus, we use the following approximation [16]:

$$H_i(y|\alpha_i, \beta_i) \approx \frac{y}{\mathbb{E}\{Y|\alpha_i, \beta_i\}} + \frac{\mathbb{E}\{Y^2|\alpha_i, \beta_i\}}{2\mathbb{E}\{Y|\alpha_i, \beta_i\}^2} - 1 \tag{5}$$

where, $\mathbb{E}\{Y|\alpha_i, \beta_i\}$ and $\mathbb{E}\{Y^2|\alpha_i, \beta_i\}$ are the first moments of a Weibull distribution with scale and shape parameters α_i and β_i, respectively.

We assume that only assets having an Increasing Failure Rate (IFR) (i.e., $\beta_i > 1$) undergo a PM strategy [12], otherwise they are repaired or replaced upon failure.

In case of PM, assets are periodically inspected with period τ_i and re-set into an AGAN state. This entails that probability distribution function $f_i(y|\alpha_i, \beta_i)$ is a τ_i −periodic function.

If we assume that the asset repair time is negligible with respect to its mean time between failures, then the number K_i of PM actions carried out over the mission time T_{miss} is the largest integer smaller then T_{miss}/τ_i, whereas the expected number of replacements up to the time of the last PM maintenance action is

$$\begin{aligned} H_i(\tau_i K_i|\alpha_i, \beta_i) &= 1 - R_i(K_i\tau_i|\alpha_i, \beta_i) + \int_0^{K_i\tau_i} H_i(K_i\tau_i - z|\alpha_i, \beta_i) f_i(y|\alpha_i, \beta_i)\, dz = \\ &= K_i\left(1 - R_i(\tau_i|\alpha_i, \beta_i) + \int_0^{\tau_i} H_i(\tau_i - z|\alpha_i, \beta_i) f_i(y|\alpha_i, \beta_i) dz\right) = K_i H_i(\tau_i|\alpha_i, \beta_i) \end{aligned} \tag{6}$$

where, the second equality in Eq. (6) follows from the fact that $f_i(y|\alpha_i, \beta_i)$ is a τ_i −periodic function under a PM strategy.

3.1 MODELING COSTS AND OPTIMAL PM STRATEGY

Let C_{CM} and C_{PM} be the costs of performing single CM and PM actions, respectively, which we assume being not dependent on the i^{th} population. We also assume that $C_{PM} \leq C_{CM}$, to take into account that although the PM and CM actions yield the same effect (i.e., restoring the asset to the AGAN state), nonetheless the CM actions are not pre-organized and usually entail extra-costs related to the larger downtimes. With no loss of generality, we set $C_{PM} = \chi C_{CM}$, $\chi \in (0,1)$.

Under these assumptions, the expected maintenance cost for an asset belonging to the i^{th} population over the mission time T_{miss} in case of CM and PM policies are, respectively:

$$C_i^{CM} = C_i^{CM}(\alpha_i, \beta_i) = C_{CM} H_i(T_{miss}|\alpha_i, \beta_i) \tag{7}$$

$$C_i^{PM} = C_i^{PM}(\alpha_{i,}\beta_i, \chi, T_{miss}, \tau_i) = K_i H_i(\tau_i|\alpha_i, \beta_i) C_{CM} + H_i(T_{miss} - \tau_i K_i|\alpha_i, \beta_i) C_{CM} + K_i C_{PM} \tag{8}$$

where $H_i(T_{miss} - \tau_i K_i|\alpha_i, \beta_i)$ represents the expected number of replacements between the last PM action undertaken at time $\tau_i K_i$ and the mission time T_{miss}.

The optimal period τ_i^* of the PM strategy can be found by minimizing the expected cost per asset (Eq. (8)):

$$\tau_i^* = \tau_i^*(\alpha_i, \beta_i, \chi, T_{miss}) = argmin_{\tau_i \in (0, T_{miss}]} C_i^{PM}(\alpha_i, \beta_i, \chi, T_{miss}, \tau_i) \tag{9}$$

The number of PM actions and the minimum expected cost in Eq. (8) under the optimal strategy are referred to as K_i^* and $C_i^{PM^*}$, respectively. When $\tau_i^* = T_{miss}$, then $C_i^{PM^*} = C_i^{CM}$ (i.e., only CM actions need to be performed).

3.2 TOTAL EXPECTED COST

In this Subsection, we present the general assumptions to estimate the expected maintenance costs for the population-driven and cluster-driven approaches:

A. *Population-driven approach*: the PM strategy is managed at population level, based on its reliability distribution $R_i(y|\alpha_i, \beta_i)$, $i = 1, .., N$. In this setting, if $\widetilde{N}$ is the number of segmented populations with $\beta_i > 1$ $(\widetilde{N} \leq N)$, then $\widetilde{N}$ different PM policies need to be optimized and managed. The total expected cost is:

$$C_{TOT}^{POP} = \sum_{i:\beta_i > 1} n_i C_i^{PM^*} + \sum_{i:\beta_i \leq 1} n_i C_i^{CM} \tag{10}$$

where the first term sums over all populations with $\beta_i > 1$ (i.e., those for which, an optimal PM maintenance can be scheduled finding the optimal period τ_i^* with Eq. (9)), whereas the second term sums over those populations with $\beta_i < 1$ (i.e., those for which only CM actions can be undertaken), $i \in \{1, \dots, N\}$.

B. *Cluster-driven approach*: the PM strategy is managed at cluster level, based on the cluster reliability distribution $R_p(y|\alpha_p, \beta_p), p = 1, .., C^*$. In this case, only the assets belonging to the $\tilde{C} \leq C^*$ clusters with $\beta_p > 1$ will undergo a PM policy. Then, $\tilde{C}$ different PM policies have to be optimized and scheduled, being in general $\tilde{C} \ll \widetilde{N}$. This yields a strong simplification for the maintenance management process. To estimate the maintenance costs, we observe that the entire asset population can be partitioned in the following four mutually exclusive and exhaustive subsets:

I. Assets belonging to a population with $\beta_i > 1$ and assigned to cluster p with $\beta_p > 1$, whose expected cost is assumed to be

$$C_{i,p}^{I} = C_{i,p}^{I}(\alpha_i, \beta_i, \chi, T_{miss}, \alpha_p, \beta_p, \tau_p^*) = [K_p^* H_i(\tau_p^* | \alpha_i, \beta_i) + H_i(T_{miss} - \tau_p^* K_p^* | \alpha_i, \beta_i)] C_{CM} + K_p^* C_{PM} \quad (11)$$

where τ_p^* is the optimal time interval between successive PMs which minimizes the expected cost $C_p^{PM}(\alpha_p, \beta_p, \chi, T_{miss}, \tau_p)$ in Eq. (7). Notice that Eq. (11) derives from Eq. (8), in which the optimal period of maintenance actions is defined by the reliability function of cluster p. Obviously, in general $\tau_p^* \neq \tau_i^*$; then, assets belonging to population i are required to follow a periodic PM strategy which is not optimal for themselves. This entails that $C_{i,p}^{I} \geq C_i^{PM^*}$, where the equality holds if and only if $\tau_p^* = \tau_i^*$, i.e., $\alpha_i = \alpha_p$ and $\beta_i = \beta_p$.

II. Assets belonging to a population with $\beta_i \leq 1$ and assigned to cluster p, $\beta_p > 1$, whose expected cost is assumed to be

$$C_{i,p}^{II} = C_{i,p}^{II}(\alpha_i, \beta_i, \chi, T_{miss}, \alpha_p, \beta_p, \tau_p^*) = K_p^* H_i(\tau_p^* | \alpha_i, \beta_i) + H_i(T_{miss} - \tau_p^* K_p^* | \alpha_i, \beta_i) + K_p^* C_{PM} \quad (12)$$

In fact, in this case no PM action would be required for the asset, which is forced to follow the optimal PM strategy of cluster p. Notice that $C_{i,p}^{II} > C_i^{CM}$, as PM actions imply a cost which is not counterbalanced by any benefit in preventing failures.

III. Assets belonging to population with $\beta_i > 1$, which are assigned to cluster p, $\beta_p \leq 1$, whose expected cost is assumed to be

$$C_{i,p}^{III} = C_{CM} H(T_{miss} | \alpha_i, \beta_i) \geq C_i^{PM^*} \quad (13)$$

In this case, although an optimal PM could be scheduled for the individual assets, however, they are not preventively maintained since $\beta_p \leq 1$. Notice also that the equality holds if and only if $\tau_i^* = T_{miss}$ (i.e., only CM actions are performed), whereby $C_i^{PM^*} = C_i^{CM}$.

IV. Assets belonging to a population with $\beta_i \leq 1$, and assigned to cluster p, $\beta_p \leq 1$. The expected cost per asset is obviously $C_{i,p}^{IV} = C_i^{CM}$, like for the asset managed under the population-driven approach.

Based on the considerations above, the total expected cost can be computed as:

$$C_{TOT}^{CLU} = \sum_{i \in Subset\ I} n_i C_{i,p}^{I} + \sum_{i \in Subset\ II} n_i C_{i,p}^{II} + \sum_{i \in Subset\ III} n_i C_{i,p}^{III} + \sum_{i \in Subset\ IV} n_i C_{i,p}^{IV} \tag{14}$$

Finally, we consider the following three cost items:

- C^{EC} is the difference between the expected cost in Eq. (14) and that in Eq. (10), i.e., the maintenance extra-cost due to the application of the cluster-driven approach:

$$\begin{gathered} C^{EC} = C_{TOT}^{CLU} - C_{TOT}^{POP} = \\ = \sum_{i \in Subset\ I} n_i\left(C_{i,p}^{I} - C_i^{PM^*}\right) + \sum_{i \in Subset\ II} n_i(C_{i,p}^{II} - C_i^{CM}) + \sum_{i \in Subset\ III} n_i(C_{i,p}^{III} - C_i^{PM^*}) + \sum_{i \in Subset\ IV} n_i(C_{i,p}^{IV} - C_i^{CM}) = \\ = \sum_{i \in Subset\ I} n_i\left(C_{i,p}^{I} - C_i^{PM^*}\right) + \sum_{i \in Subset\ II} n_i(C_{i,p}^{II} - C_i^{CM}) + \sum_{i \in Subset\ III} n_i(C_{i,p}^{III} - C_i^{PM^*}) \end{gathered} \tag{15}$$

Notice that the quantity C^{EC} is always non-negative since, $C_{i,p}^{I} \geq C_i^{PM^*}$, $C_{i,p}^{II} > C_i^{CM}$, $C_{i,p}^{III} \geq C_i^{PM^*}$ and $C_{i,p}^{IV} = C_i^{CM}$.

- $C_{ORG}^{POP} = C_{ORG}(\widetilde{N})$ and $C_{ORG}^{CLU} = C_{ORG}(\tilde{C})$ are the planning expected costs resulting from scheduling $\widetilde{N}$ different PM policies under the population-driven approach and $\tilde{C}$ different PM policies under the simplification brought by the cluster-driven maintenance approach, respectively.

The asset manager will opt for the cluster-driven maintenance approach if the following inequality holds

$$C^{EC} < C_{ORG}^{POP} - C_{ORG}^{CLU} \tag{16}$$

i.e., if the extra-cost due to clustering is balanced by the cost reduction in scheduling $\tilde{C}$ different PM strategies instead of $\widetilde{N}$, being in general $C_{ORG}\left(\widetilde{N}\right) > C_{ORG}\left(\tilde{C}\right)$."

4 CASE STUDY

The available dataset consists of millions of different assets for which the values of many control variables are provided in the form of heterogeneous unstructured data. With the objective of developing a methodological approach for extracting information from all available data for optimizing asset management, in this work, we consider a subset of the original dataset without loss of generality. The subset of considered data consists of $A = 32385$ different assets for which the values of $K = 12$ decision variables $(X_1, \dots, X_K)$ are provided (for confidentiality, details are not given here). Among these decision variables a subset of $\widetilde{K} = 5$ decision variables $(\widetilde{X}_1, \dots, \widetilde{X}_{\widetilde{K}})$ has been selected by experts (Step 1, Subsection 2.1). Based on their values, $N = 374$ populations of assets have been identified, with corresponding failure time datasets $\boldsymbol{O}_i, i \in \{1, \dots, 374\}$. Then, the estimates of the Weibull parameters (α_i, β_i) for all 374 populations have been obtained by resorting to MLE method (Step 2, Subsection 2.1). In Figure 1(a), the estimated values of the scale parameters (abscissas, in logarithmic scale) and shape parameters (ordinates) are shown. Notice that a different choice of the decision variables provided by the experts leads to a different partitioning of the assets, with different reliability behaviors. For example, if we select a subset of $\widetilde{K} = 4$ decision variables among the 5 selected from the experts, the cardinality of populations of assets reduces to $N = 91$, with different reliability behaviors, as shown in Figure 1(b).

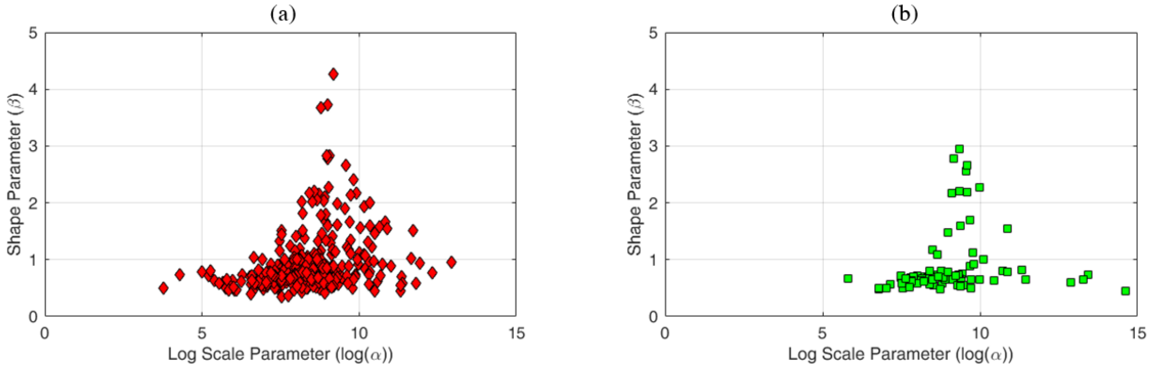

Figure 1: Estimated Weibull parameters for each statistical population when $\widetilde{K} = 5$ (left) and $\widetilde{K} = 4$ (right).”

The similarity matrix W has been obtained by computing the similarity measure w_{ij} of Eq (A.3) between all possible 374 pairs of statistical populations (Step3, Subsection 2.1). To assess the most appropriate number of clusters, we have resorted to the silhouette and Davies-Bauldin coefficients (Step 4, Subsection 2.1). Figures 2 and 3, respectively, show the values of these coefficients in correspondence to the number of clusters varying from 2 to 10.

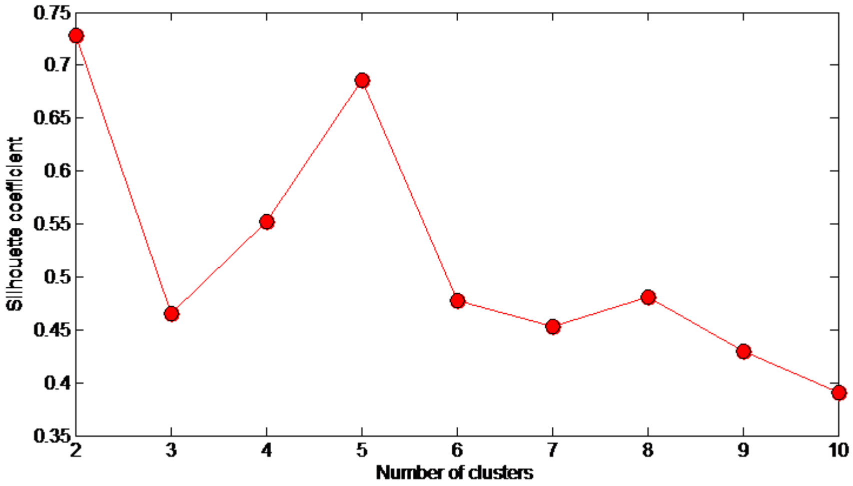

Figure 2: Silhouette coefficient increasing the number of clusters from 2 to 10.

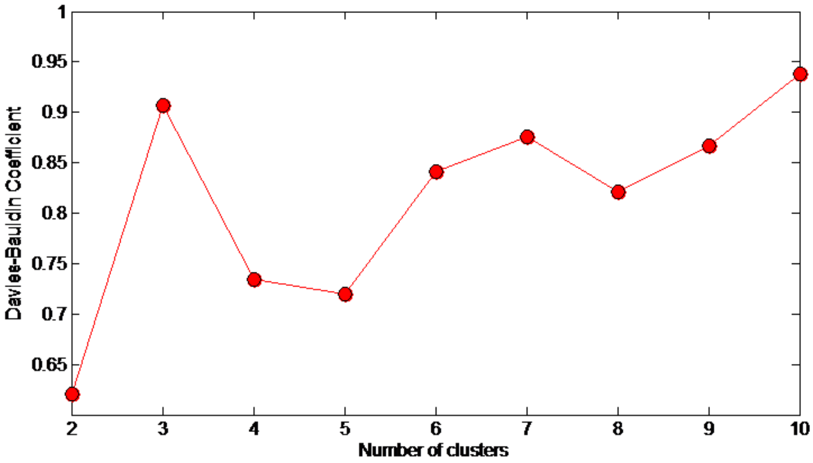

Figure 3: Davies-Bauldin coefficient increasing the number of clusters from 2 to 10.

Let us analyze the best solutions, $C^* = 2$ and $C^* = 5$, as seen in both Figures 3 and 4. Figures 4(a) and 4(b) show, for each statistical population, in abscissa the log scale parameter ($\log(\alpha)$) and in ordinate the corresponding value of the shape parameter (β) when $C^* = 2$ and $C^* = 5$, respectively. From these Figures, it emerges that the cluster 2 and 5 (represented by different markers) divide the semi-plane ($\log(\alpha)$, β) in 2 and 5 pairwise disjoint regions, respectively, and, therefore, we can conclude that these 2 and 5 clusters really identify different reliability behaviours, respectively.

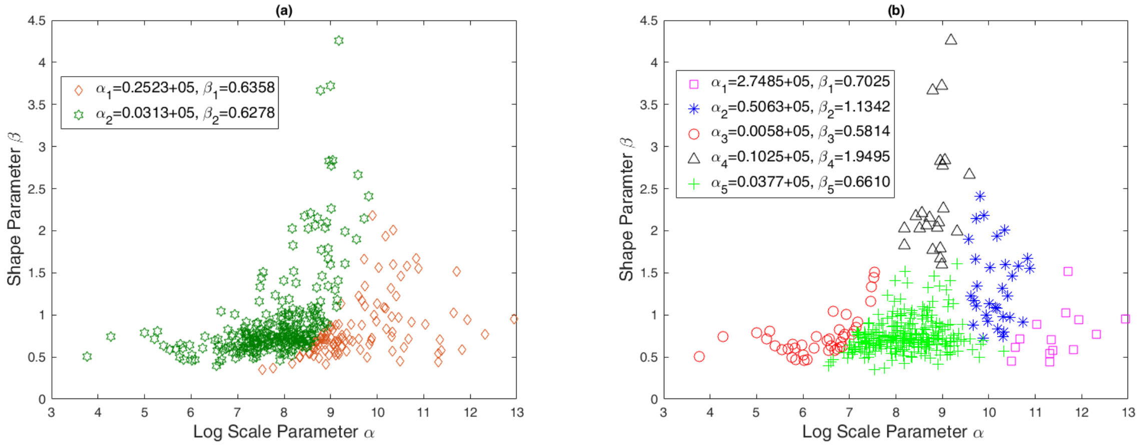

Figure 4: Values of the (log) scale parameters (abscissa axis) and shape parameters (ordinate axis) when $C^* = 2$ (left) and $C^* = 5$ (right), respectively.

Once these clusters are identified, we can estimate the representative reliability distribution of all assets belonging to the same clusters. Again assume that the reliability behaviour of each cluster is described by a Weibull probability distribution, by reason of the flexibility of this distribution [10]. In Tables 1 and 2, the MLE values of the scale parameters and shape parameters are reported for each cluster (Steps 6 and 7, Subsection 2.1) when $C^* = 2$ and $C^* = 5$, respectively. From these, one can conclude that:

- *Case 1:* $C^* = 2$
  1. There are two clusters (indicated by hexograms and diamonds in the Figure) for which the estimated values of the shape parameters are not significantly different to each other, whereas the estimated values of the scale parameters are very different. For these clusters, the hazard rate turns out to be a decreasing function of time and, thus, only CM actions would be considered.
- *Case 2:* $C^* = 5$
  1. There are three clusters (indicated by circles, crosses and squares in the Figure) for which the estimated values of the shape parameters are similar to each other ($\beta < 1$), whereas the estimated values of the scale parameters are very different. For these clusters, the hazard rate is a decreasing function of time and, thus, only CM actions would be considered.
  2. There are two clusters (stars and triangles in the Figure) with shape parameters assuming values larger than one, and with very different values of the estimated scale parameters. For these clusters, the failure rate is an increasing function of time and, thus, an optimal PM maintenance strategy can be scheduled.

Scale Parameter α_p	Shape Parameter β_p	Cluster marker
0.2523+05	0.6358	hexogram
0.0313e+05	0.6278	diamond

Table1: MLEs of scale and shape parameters of each cluster ($\boldsymbol{C^* = 2}$)

Scale Parameter α_p	Shape Parameter β_p	Cluster marker
0.0058 e+05	0.5814	circle
0.5063e+05	1.1342	star
0.0377e+05	0.6610	cross
0.1025e+05	1.9495	triangle
2.7485e+05	0.7025	square

Table2: MLEs of scale and shape parameters of each cluster ($\boldsymbol{C^* = 5}$).

In Figure 5, the reliability functions relative to the $C^* = 5$ identified clusters are shown. This information enables the scheduling of only 5 maintenance strategies, which are applied to all the assets belonging to the corresponding clusters.

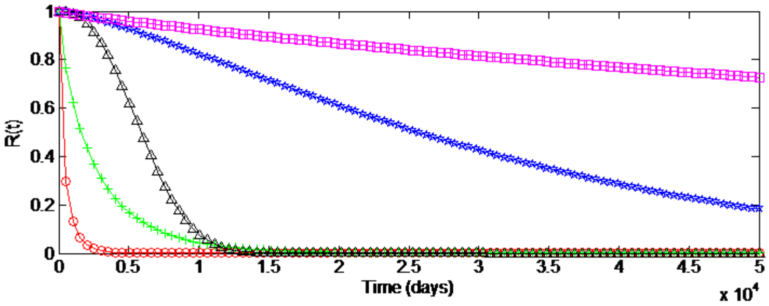

Figure 5: Reliability function $\boldsymbol{R(t)}$ relative to the 5 clusters (different markers correspond to the different clusters)

4.1. COST ANALYSIS: results

In this Subsection, we limit our analysis to case $C^* = 5$. This choice is due to the fact that the case $C^* = 2$ corresponds to a particular case when only CM actions are taken.

We compare the total expected cost under the population-driven approach with that of the cluster-driven approach. For simplicity, the cost of a CM action, C_{CM}, is set equal to 1 (in arbitrary unit), whereas factor χ is assumes 20 evenly spaced values between 0.05 and 1. The mission time $T_{miss} =$ 40 years.

If we approach maintenance under the population-driven approach, there are $\widetilde{N}$ =83 populations with shape parameter β_i>1. Accordingly, we need to find 83 optimal policies by using Eq. (9).

For example, Figure 6 shows the outcomes of the optimization of the maintenance period of the segmented population S_{339}, whose reliability distribution has scale parameter $\alpha_{339} = 4.5817e + 03$ (days) and shape parameter β_{339} =2.1735. From this Figure, we can see that the larger the value of χ, the larger is the optimal maintenance interval τ_{339}^*. When χ is larger than 0.35, it is no longer convenient scheduling a periodic PM strategy, as $\tau_{339}^* \geq 40$.

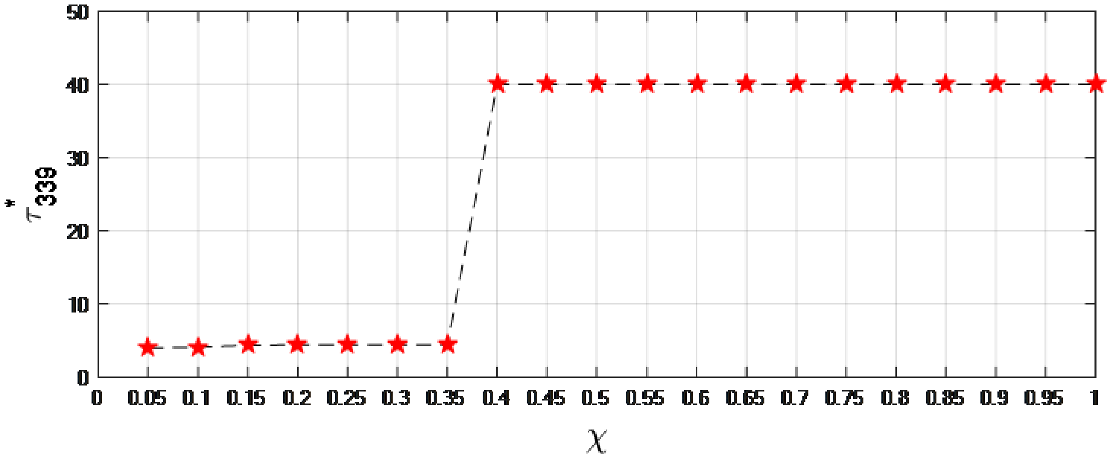

Figure 6: Optimal period τ_{399}^*, corresponding to different values of the discount factor χ for the 339^{th}

The sudden transition from $\tau_{399}^* = 5$ to $\tau_{399}^* = 40$ at $0.35 < \chi < 0.40$ can be justified by looking at Figure 7, which shows the cost function $C_{339}^{PM} = C_{339}^{PM}(\alpha_{339,}\beta_{339,}\chi, 40, \tau_{339})$ in Eq. (8) in logarithmic scale for all considered values of factor χ: the period $\tau_{339} = 40$ is always a local minimum point regardless of the value χ except when $\chi = 0.05$, and it becomes a global minimum point for $\chi \geq 0.40$. This is due to the fact that when $\chi > 0.05$ scheduling a PM action at time $\tau_{399} = 39$ is less convenient that not to undertake any PM action.

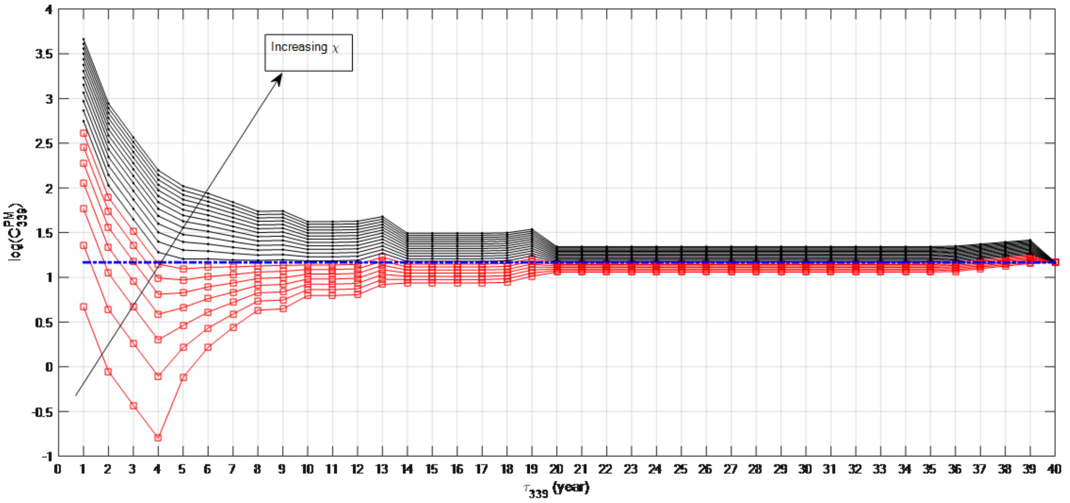

Figure 7: Cost function $C_{339}^{PM} = C_{339}^{PM}(\alpha_{339,}\beta_{339,}\chi, 40, \tau_{339})$ in Eq. (16) in logarithmic scale for all considered values of factor χ (square and point markers correspond to $\chi < 0.35$ and $\chi \geq 0.35$, respectively, the dash line correspond to the expected cost when no PM action are undertaken , i.e., $\tau_{399} \geq 40$, which does not depend on factor χ.

Figure 8 shows the expected cost for every asset belonging to population S_{339} as a function of the discount factor χ, provided that the PM actions are performed at the corresponding optimal periods $\tau_{399}^{*}(\chi)$ shown in Figure 6. According to the results of Figures 6 and 7 the PM strategy is no longer convenient once χ reaches 0.4 and the maintenance cost equal that of the CM strategy.

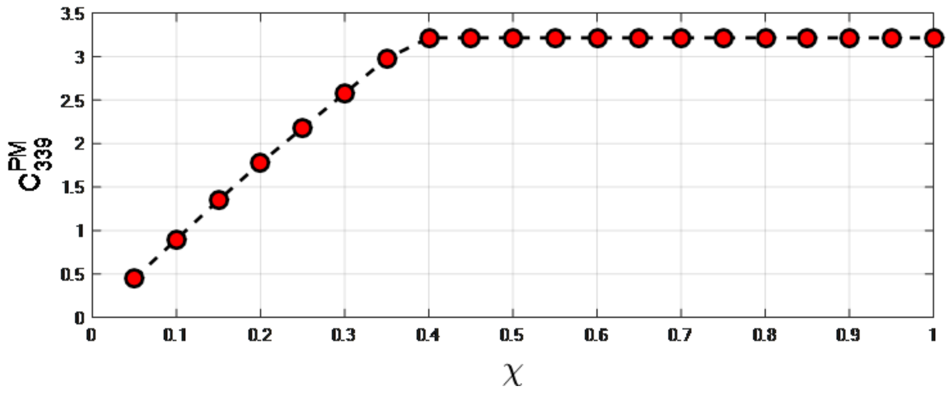

Figure 8: Expected cost per asset under optimal strategy as function of the discount factor χ for the 339^{th} population

With respect to the cluster-driven approach, Figure 9 shows for cluster 2, the optimal period τ_2^* vs χ. From this, it clearly emerges that PM strategy is never convenient, whichever the value of χ.

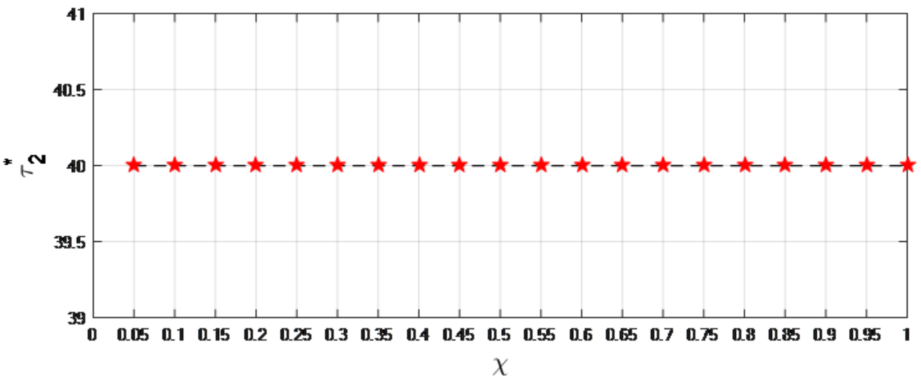

Figure 9: Optimal periods τ_2^* for different values of the discount factor χ for clusters 2

Figure 10 shows the same results related to cluster 4. In this case, an optimal PM strategy can be found until χ is smaller than 0.40. From this value on, the difference between CM and PM does not allow to justify the scheduling of a PM strategy.

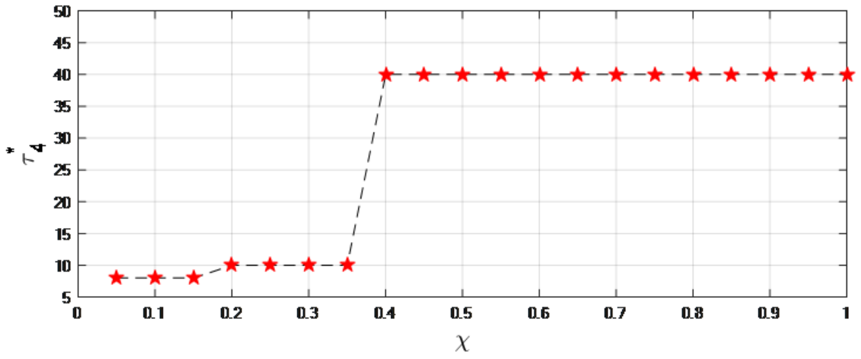

Figure 10: Optimal periods τ_4^* for different values of the discount factor χ for clusters 4

4.2. COST ANALYSIS: discussion

To fairly compare the population and cluster-driven, Tables 3 and Table 4, respectively, report the total numbers $\widetilde{N}$ and $\tilde{C}$ of PM policies that these approaches require to optimize and trace: whenever an optimal PM strategy exists, the number of optimal periodic PM strategy under cluster-driven approach is always smaller than that under the population-driven approach. This difference is even more pronounced for small values of χ; for example, when χ=0.05 one has to schedule only one maintenance action versus the 66 required when the maintenance approach is managed under the population-driven approach.

χ	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
$\widetilde{N}$	66	58	49	43	34	26	21	8	3	0	0	0	0	0	0	0	0	0	0	0

Table 3: Total number of optimal periodic PM strategy $\widetilde{N}$ to be scheduled under the population-driven approach.

χ	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
$\tilde{C}$	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: Total number of optimal periodic PM strategy $\widetilde{C}$ to be scheduled under the cluster-driven approach

The total expected cost versus χ for the two maintenance management approaches are shown in Figure 11. Notice that according to the arguments discussed in Subsection 3.2, the total expected cost in case of cluster-driven maintenance management approach is larger than that obtained with the population-driven approach.

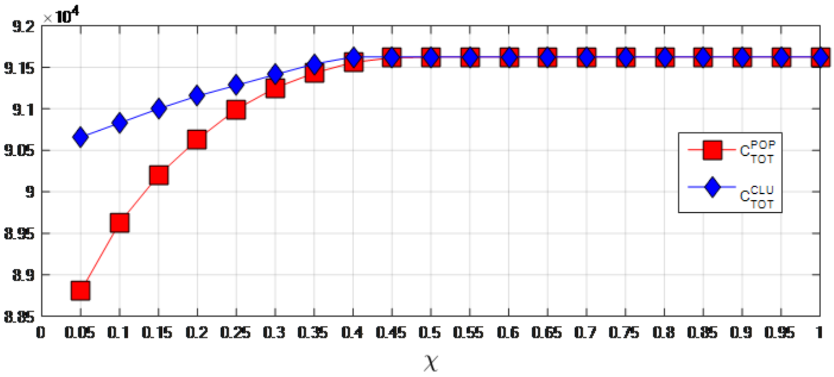

Figure 11: Total expected cost in case of cluster-driven maintenance management approach and segmented population-driven approach (square and diamond, respectively).

Figure 12 shows the percentage value of the total expected cost difference between the cluster and population-driven approaches.

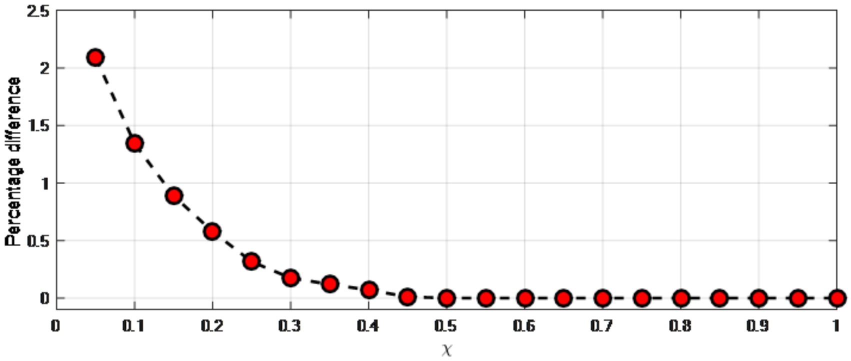

Figure 12: Percentage total expected cost difference between the cluster-driven maintenance management approach and segmented population-driven approach.

Suppose that $\chi = 0.05$. From Table 3 and 4 one has that the total number of optimal periodic PM strategy $\widetilde{N}$ and $\tilde{C}$ to be scheduled under the population and cluster-driven approaches are 66 and 1, respectively; the corresponding scheduling costs are $C_{ORG}^{POP} = C_{ORG}(66)$ and $C_{ORG}^{CLU} = C_{ORG}(1)$, respectively. The total expected cost difference C^{EC} in Eq. (15) is 1815 (in arbitrary unit) which to corresponds a percentage total cost expected difference of 2.05%. If the simplification brought by the cluster-driven maintenance approach is such that condition in Eq. (16) is satisfied: that is, if the difference between planning costs $C_{ORG}(66)$ and $C_{ORG}(1)$ is larger than 1815, i.e., is larger than 0.0205 times the total expected cost under the segmented population-driven approach, then, the asset manager should opt for the cluster-driven maintenance approach otherwise for the segmented population-driven approach.

5 CONCLUSIONS

In this work, we have proposed a cost model to support the asset decision maker in quantifying the possible maintenance extra-costs due to scheduling PM strategies under the cluster-driven maintenance approach proposed in [10] for managing the maintenance of a very large fleet of assets. Our cost model has been applied to a real case study concerning assets of the railway system, with more than 30000 assets which have been grouped in 5 clusters. We found that the optimal PM strategies under the population and cluster-driven approaches are 66 and 1, respectively. Cost analysis shows that the simplification brought by the cluster-driven approach is justified when the expected planning costs resulting from scheduling 1 PM strategy instead of 66 outweigh the extra-costs due to the assets following a maintenance strategies that are optimal at cluster level but not at population level.

7 REFERENCES

[1] Meeker, W.Q. & Hong, Y. 2014. Reliability meets big data: Opportunities and challenges. *Quality Engineering* 26 (1): 102-116.

[2] Zio, E. 2016. Some Challenges and Opportunities in Reliability Engineering. *IEEE Transactions on Reliability,* Article in press.

[3] Nappi, R. 2014. Integrated maintenance: analysis and perspective of innovation in railway sector. *arXiv: 1404.7560.*

[4] Chen, H., Chiang, R.H.L., Storey, V.C. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, 36 (4), pp. 1165-1188.

[5] Thaduri, A., Galar, D., Kumar, U. 2015. Railway assets: A potential domain for big data analytics. *Procedia Computer Science,* 53 (1), pp. 457-467.

[6] Attoh-Okine, N. 2014. Big data challenges in railway engineering. *2014 IEEE International Conference on Big Data*, pp. 7–9.

[7] Figures-Esteban, M., Hughes, O., Van Gulijk, C. 2015. The role of data visualization in railway Big Data Risk Analysis. *Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference,* ESREL 2015, pp. 2877-2882.

[8] Sammpouri, W., Come, E., Oukhellou, L., Aknin, P., Fonllados, C. 2013. Floating train data systems for preventive mainetenance: A data mining approach. *Proceedings of 2013 International Conference on Industrial Engineering and Systems Management, IEEE- IESM 2013,* art. No. 6761372.

[9] Fumeo, E., Oneto, L., Anguita, D. 2015. Condition based maintenance in railway transportation systems based on big data streaming analysis. *Procedia Computer Science,* 53 (1), pp., 437-446.

[10] Cannarile, F., Compare, M., Di Maio, F., Zio, E. 2015. Handling reliability big data: A similarity-based approach for clustering a large fleet of assets (2015) *Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference*, ESREL 2015, pp. 891-896.

[11] Zio, E., Compare, M. Evaluating maintenance policies by quantitative modeling and analysis (2013) Reliability Engineering and System Safety, 109, pp. 53-65.

[12] Barlow RE, Proschan F. Mathematical theory of reliability. New York: John Wiley & Sons; 1965.

[13] Cannarile, F., Compare, M., Mattafirri, S., Carlevaro, F., Zio, E. 2015. Comparison of Weibayes and Markov Chain Monte Carlo methods for the reliability analysis of turbine nozzle components with right censored data only. *Safety and Reliability of Complex Engineered Systems - Proceedings of the 25th European Safety and Reliability Conference,* ESREL 2015, pp. 1937-1944.

[14] Rousseeuw, P. J. 1987. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics* 20: 53–65.

[15] Davies, D. & Bouldin, D. 1979. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1(2): 224–227.

[16] Maghsoodloo, S., Helvaci, D. 2014. Renewal and renewal-intensity functions with minimal repair. *Journal of Quality and Reliability Engineering* art. no. 857437 .

[17] Kullback, S. & Leibler, R.A.. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22 (1): 79–86.

[18] Gower, J. 1985. Measures of similarity, dissimilarity, and distance. *Encyclopedia of Statistical Sciences* 5: 397–405. New York: Wiley.

[19] Gower, J. 1986. Metric and Euclidean properties of dissimilarity coefficients. *Journal of classification* 3:5–48.

[20] Bauckhage, C. 2013. Computing the Kullback-Leibler Divergence between two Weibull Distributions. *arXiv:1310.3713 [cs.IT]*.

[21] Von Luxburg, U. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing* 17(4): 395-416.

[22] Baraldi, P., Di Maio, F., Zio, E. 2013. Unsupervised Clustering for Fault Diagnosis in Nuclear Power Plant Components. *International Journal of Computational Intelligence Systems,*6 (4): 764-777.

[23] Hartigan, J. A. (1975). Clustering Algorithms. New York: Wiley.

APPENDIX

A: SIMILARITY BETWEEN PROBABILITY DISTRIBUTIONS

Let Ω denote the sample space, $\mathcal{F}$ a σ −algebra on Ω, and P the set of probability measures on the measurable space $(\Omega, \mathcal{F})$. Let μ_i and μ be two elements of P , and f_i and f_j their corresponding probability density functions with respect to a dominating measure ρ. Then, the Kullback-Leibler Divergence (KLD) of μ_i from μ_j denoted with $d_{KL}(\mu_i||\mu_j)$, is defined as [17]:

$$d_{KL}(\mu_i||\mu_j) = d_{KL}(f_i||f_j) = \int_\Omega f_i \log\frac{f_i}{f_j} d\rho \qquad \text{(A1)}$$

Note that, in general, $d_{KL}(\mu_i||\mu_j) \neq d_{KL}(\mu_j||\mu_i)$.

Otherwise, if we define the SKLD between μ_i and μ_j as:

$$d_{KL}^{sym}(\mu_i, \mu_j) = \frac{1}{2}(d_{KL}(\mu_i||\mu_j) + d_{KL}(\mu_j||\mu_i)) \qquad \text{(A2)}$$

then d_{KL}^{sym} is a dissimilarity measure (being symmetric). We can, therefore, define the similarity corresponding to the SKLD [18], [19] as in Eq. (A3):

$$w_{ij} = \frac{1}{1 + d_{KL}^{sym}} \qquad \text{(A3)}$$

This measure is used to compute the similarity between the reliability distributions f_i and f_j of two different populations of assets. In this respect, notice that the densities f_i and f_j in Eqs. (A1) and (A2) are assumed Weibull distributions in our case study. This makes the computation of w_{ij} not straightforward. Nonetheless, we can exploit the results provided in [20] to efficiently compute the KLD divergence in Eq. (1) between two Weibull densities $f_i(y|\alpha_i, \beta_i)$ and $f_j(y|\alpha_j, \beta_j)$ as in Eq. (A4):

$$d_{KL}(f_i\|f_j) = \log\left(\frac{\beta_i}{\alpha_i^{\beta_i}}\right) - \log\left(\frac{\beta_j}{\alpha_j^{\beta_j}}\right) - (\beta_i - \beta_j)\left(\log(\alpha_i) - \frac{\gamma}{\beta_i}\right) + \left(\frac{\alpha_i}{\alpha_j}\right)^{\beta_j} \Gamma\left(\frac{\beta_j}{\beta_i}\right)^{\beta_j} - 1 \quad \text{(A4)}$$

where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant, whereas Γ is the gamma function defined as follows:

$$\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt \quad z \geq 0 \qquad \text{(A5)}$$

B: SPECTRAL CLUSTERING

Consider the similarity matrix $W$of size (N, N), whose generic element w_{ij} represents the similarity between the statistical populations S_i and S_j. W is symmetric and its the diagonal elements w_{ii} are set to 1.

From matrix W, a similarity graph $G = (V, E)$ is constructed, where each vertex v_i represents the i^{th} population and the weight associated to the edge e_{ij} connecting the two vertices i and j is the similarity value w_{ij} [21]. The spectral clustering algorithm proceeds as follows [22]:

Step1: normalized Graph Laplacian Matrix

Compute:

- The degree matrix D which is a diagonal matrix with diagonal entries $d_1, \ldots, d_N$ defined by:

$$d_i = \sum_{j=1}^{N} w_{ij} \quad i = 1, \ldots, N \qquad \text{(B1)}$$

- The normalized graph Laplacian matrix

$$L_{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \qquad \text{(B2)}$$

where $L = D - W$, and I is the identity matrix of size (N, N).

Step2: feature extraction

The relevant information on the structure of the matrix W is obtained by considering the eigenvectors $u_1, \ldots, u_C$ associated to the C smallest eigenvalues $\lambda_1, \ldots, \lambda_C$ of its laplacian matrix L_{sym}, where C is the desired number of clusters. The square matrix W is transformed into a reduced matrix U of size (N, C), in which the C columns of U are the eigenvectors $u_1, \dots, u_C$. Thus, the i^{th} object is captured in the C-dimensional vector u_i corresponding to the i^{th} row of the matrix U A matrix T is formed from U by normalizing its row [21]:

$$t_{ic} = \frac{u_{ic}}{(\sum_{c=1}^{C} u_{ik}^2)^{\frac{1}{2}}} \quad i = 1, \dots, N, c = 1, \dots, C \tag{B3}$$

It has been shown that this change of representation enhances the cluster properties in the data, so that clusters can be more easily identified [20].

Step3: Unsupervised clustering

We use the K-means [23] algorithm to get C clusters. Details on this clustering method can be found in [10].