LS-SVM based spectral clustering and regression for predicting maintenance of industrial machines

Rocco Langone[1,5], Carlos Alzate[1,4], Bart De Ketelaere[2], Jonas Vlasselaer,[3] Wannes Meert,[3] Johan A. K. Suykens[1,5]

[1]*Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, B-3001 Leuven Belgium*
Email: {rocco.langone,carlos.alzate,johan.suykens}@esat.kuleuven.be

[2]*Faculty of Bioscience Engineering, BIOSYST MeBioS Qualimatrics, KU Leuven, B-3001 Leuven Belgium*
Email: bart.deketelaere@biw.kuleuven.be

[3]*Department of Computer Science, KU Leuven, Celestijnenlaan 200a B-3001 Leuven Belgium*
Email: {jonas.vlasselaer,wannes.meert}@cs.kuleuven.be

[4]*Smarter Cities Technology Center, IBM Research-Ireland*
Email: carlos.alzate@ie.ibm.com

[5] *iMinds Future Health Department*

Abstract

Accurate prediction of forthcoming faults in modern industrial machines plays a key role in reducing production arrest, increasing the safety of plant operations, and optimizing manufacturing costs. The most effective condition monitoring techniques are based on the analysis of historical process data. In this paper we show how Least Squares Support Vector Machines (LS-SVMs) can be used effectively for early fault detection in an online fashion. Although LS-SVMs are existing artificial intelligence methods, in this paper the novelty is represented by their successful application to a complex industrial use case, where other approaches are commonly used in practice. In particular, in the first part we present an unsupervised approach that uses Kernel Spectral Clustering (KSC) on the sensor data coming from a vertical form seal and fill (VFFS) machine, in order to distinguish between normal operating condition and abnormal situations. Basically, we describe how KSC is able to detect in advance the need of maintenance actions in the analysed machine, due the degradation of the sealing jaws. In the second part we illustrate a nonlinear auto-regressive model (NAR), thus a supervised learning technique, in the LS-SVM framework. We show that we succeed in modelling appropriately the degradation process affecting the machine, and we are capable to accurately predict the evolution of

dirt accumulation in the sealing jaws.

Keywords: kernel spectral clustering; LS-SVMs; NAR; time-series prediction; condition monitoring; maintenance; fault detection, machine degradation, artificial intelligence.

1. Introduction

In industrial processes, the detection and analysis of faults ensure product quality and operational safety [26]. Traditionally, three ways to deal with sensory faults have been used [30],[31],[32]: corrective maintenance, preventive maintenance and predictive maintenance. Corrective maintenance is performed only when the machine fails, it is expensive and safety and environmental issues arise. Preventive maintenance [16] is based on periodic replacement of components, which are then utilized in a non-optimal way. Predictive maintenance [5] can be performed in a manual or automatic fashion. In the first case machines are manually checked with expensive monitoring techniques and the components are replaced according to their real status. In the second case a machine's status is automatically inspected and maintenance is planned accordingly. The continuous monitoring of machine parts leads to reliable and accurate lifetime predictions, and maintenance operations can be fully automated and implemented in a cost effective way.

Nowadays, in many industries several process variables like temperature, pressure etc. can be measured. These measurements give an information on the current status of a machine and can be used to predict the faults due to deterioration of key components [23, 10]. As a consequence, an optimal maintenance strategy can be planned.

Condition monitoring using sensor data has been performed for long time by means of basic methods like exponentially weighted moving average, cumulative sum, principal component analysis (PCA) [9], [6]. Only in the past few years engineers in companies started to get convinced to use more advanced techniques for fault detection, like for instance SVMs approaches in semiconductor manufacturing [25, 15]. In this realm, with the aim of contributing to bridge the gap between academia and industry, we propose to use LS-SVMs for predictive maintenance. LS-SVMs [28, 27] are an artificial intelligence technique characterized by an high quality generalization capability [29], the flexibility in the model design and a clear procedure for model selection. Basically, for a given machine we can construct a reliable model of the degradation process to be able to opti-

mize the time of maintenance. For this purpose here we propose two alternative approaches with different associated costs, applied to predictive maintenance in a vertical form seal and fill (VFFS) machine:

- clustering of vibration signals collected by accelerometers. This technique is cheap because of the relatively low cost of the accelerometers. On the other hand, the clustering model does not directly predict machine's deterioration, but it infers the degradation based on vibration data.

- regression applied to thermal camera data. The number of hot area pixels present in the images acquired by the camera represent a direct measure of degradation. A nonlinear autoregressive model is then used to forecast the future trend of machine's deterioration. This procedure is more reliable than the clustering-based methodology because in this case a direct prediction of the degradation is achieved. On the other hand, it is more expensive because of the higher cost of the thermal camera w.r.t. the accelerometers.

In the first part of this work we describe the use of kernel spectral clustering (KSC, [2]) for identifying in advance when the (VFFS) machine enters critical conditions. A similar analysis has been carried out in [12], where only two experiments instead of three were considered and no comparisons with other techniques were performed. In contrast to [11] and [13] where an initial clustering model is updated over time, in this framework we assume stationarity. Thus the algorithm is built off-line and then used online (by means of the out-of-sample extension property) in order to distinguish between normal operating condition and abnormal situations. Then we use the model in an online fashion via the out-of-sample extension property of KSC to recognize these two regimes. In particular, KSC correctly infers the degradation process from the vibration signals registered by the accelerometers placed on the sealing jaws.

In the second part of the paper we utilize a nonlinear auto-regressive model (NAR [17]) to catch the deterioration phenomenon acting on the machine. Like KSC, the model is cast in the LS-SVM framework and, once properly trained, it can be used in an online manner thanks to the out-of-sample extension ability. In this experiment, the machine is equipped with a thermal camera that directly measures the degradation in terms of number of hot area pixels in the acquired images. Then the NAR model is used to predict the evolution of the number of hot area pixels over time.

The remainder of this paper is structured as follows: Section 2 summarizes the KSC model and the related model selection scheme, i.e. BLF (Balanced Line

Fit). Section 3 explains the basics of the LS-SVM regression method and how it can be used to formulate a NAR model. Section 4 describes the data-sets used in the experiments. In Section 5 the simulation results are presented. Regarding KSC, a comparison with k-means and self organizing maps is also depicted. For what concerns the LS-SVM NAR model, we show that it outperforms a linear auto-regressive model (AR) and gives very good predictions up to 10 steps ahead. Section 6 presents a general discussion. Finally, Section 7 concludes the paper.

2. Kernel Spectral Clustering

2.1. Model

Spectral clustering methods use the eigenvectors of the Laplacian to properly group the data-points [7, 33, 21]. KSC is a spectral clustering formulation cast in the LS-SVM primal-dual setting, being the primal problem a constrained optimization problem with a quadratic loss function and equality constraints. As mentioned in the introduction, the KSC method has two main advantages, a systematic model selection scheme for the correct tuning of the parameters and the extension of the clustering model to out-of-sample points. In KSC, a clustering model can be trained on a subset of the data and then applied to the rest of the data in a learning framework. The out-of-sample extension allows then to predict the memberships of a new point thanks to the model learned during the training phase. In this way, once a model of the operation of a machine has been constructed, we can use it in an online fashion to discover when the machine enters critical conditions.

Given a training data set $\mathcal{D} = \{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$, the multi-cluster KSC model [2] is formulated as a weighted kernel PCA problem [19] decomposed in $l = k-1$ binary problems, where k is the number of clusters to find:

$$\min_{w^{(l)},e^{(l)},b_l} \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)} \tag{1}$$

$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N \tag{2}$$

The $e^{(l)} = [e_1^{(l)}, \ldots, e_N^{(l)}]^T$ are the projections of the data points $\{x_i\}_{i=1}^N$ mapped in the feature space along the direction $w^{(l)}$, also called score variables. The optimization problem (1) can then be interpreted as the maximization of the weighted variances $C_l = e^{(l)^T} D^{-1} e^{(l)}$ and the contextual minimization of the squared norm of the vector $w^{(l)}$, $\forall l$. Through the regularization constants $\gamma_l \in \mathbb{R}^+$ we trade-off the model complexity expressed by $w^{(l)}$ with the correct representation of the

training data. The symbol 1_N represents a column vector with the N components equal to 1, $D^{-1} \in \mathbb{R}^{N \times N}$ is the inverse of the degree matrix D, Φ is the $N \times d_h$ feature matrix $\Phi = [\varphi(x_1)^T; \ldots; \varphi(x_N)^T]$ and $\gamma_l \in \mathbb{R}^+$ are regularization constants. The clustering model is formulated as:

$$e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l, i = 1, \ldots, N \tag{3}$$

where $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$ is the mapping to a high-dimensional feature space, b_l are bias terms, with $l = 1, \ldots, k-1$. The projections $e_i^{(l)}$ represent the latent variables of a set of $k-1$ binary clustering indicators given by sign$(e_i^{(l)})$. The set of binary indicators form a code-book $\mathcal{CB} = \{c_p\}_{p=1}^k$, where each code-word is a binary word of length $k-1$ representing a cluster. After constructing the Lagrangian and solving the Karush-Kuhn-Tucker (KKT) conditions for optimality the following dual problem is obtained:

$$D^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)} \tag{4}$$

where Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, D is the graph degree matrix which is diagonal with positive elements $D_{ii} = \sum_j \Omega_{ij}$, M_D is a centering matrix defined as $M_D = I_N - \frac{1}{1_N^T D^{-1} 1_N} 1_N 1_N^T D^{-1}$, the $\alpha^{(l)}$ are dual variables, $\lambda_l = \frac{N}{\gamma_l}$. $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is the kernel function and captures the similarity between the data-points. In all the experiments we use the RBF kernel function defined by $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2 / \sigma^2)$, with σ meaning the bandwidth parameter. The out-of-sample extension to new nodes is done by an Error Correcting Output Codes (ECOC) decoding procedure. In the decoding process the cluster indicators obtained in the validation/test stage are compared with the code-book and the nearest code-word in terms of Hamming distance is selected. The cluster indicators can be obtained by binarizing the score variables for test points in the following way:

$$\text{sign}(e_{\text{test}}^{(l)}) = \text{sign}(\Omega_{\text{test}} \alpha^{(l)} + b_l 1_{\text{Ntest}}) \tag{5}$$

with $l = 1, \ldots, k-1$. Ω_{test} is the $N_{\text{test}} \times N$ kernel matrix evaluated using the test nodes with entries $\Omega_{\text{test,ri}} = K(x_r^{\text{test}}, x_i)$, $r = 1, \ldots, N_{\text{test}}$, $i = 1, \ldots, N$.

2.2. Tuning scheme

The number of clusters in which to group the data and the parameters of the kernel function (if any) have to be selected properly to ensure good performances.

The model selection scheme used in the experiments is Balanced Line Fit (BLF). The BLF criterion reaches its maximum value 1 when the clusters are well separated, and in this ideal situation they are represented as lines in the space of the projections $e^{(l)}$. The analytical formula of BLF is [2]:

$$\text{BLF}(\mathcal{D}^V, k) = \eta \text{linefit}(\mathcal{D}^V, k) + (1 - \eta)\text{balance}(\mathcal{D}^V, k) \tag{6}$$

where $\mathcal{D}^V$ indicates the validation set and k the number of clusters. The linefit index equals 0 when the score variables are distributed spherically and equals 1 when the score variables related to points in the same cluster are collinear. The balance index equals 1 when the clusters have the same size and tends to 0 in extremely unbalanced cases. The parameter η controls the importance given to the linefit with respect to the balance index and takes a value in the range $[0, 1]$. The BLF can be used to select the number of clusters and the kernel tuning parameters in the following way:

1. Define a grid of values for the parameters to select
2. Train the related KSC model
3. Compute the memberships of the validation points by means of the out-of-sample extension
4. For every partition of the validation set calculate the related score in terms of BLF
5. Choose the model with the highest score[1].

An example of the model selection procedure on a synthetic toy data-set is depicted in Figure 1. It can be noticed that the maximum value of the BLF criterion is reached when the points in the projections space have a strong line structure, corresponding to optimal parameters. Although in a real-life problem the cluster structure can be less clear than in this toy example, BLF has shown to be a useful model selection criterion in several applications.

3. LS-SVMs for regression

3.1. Primal-Dual Formulation

As already pointed out previously, LS-SVMs are a kind of SVM where a quadratic loss function is used in the primal objective and equality instead of

[1]Instead of considering only the highest value of BLF, also additional local maxima may be taken into account.

inequality constraints are present. This typically leads to a linear system or an eigenvalue problem at the dual level for classification and regression or (weighted) kernel PCA respectively. The main advantage of LS-SVMs with respect to other artificial intelligence methods like artificial neural networks is that a global optimum can be obtained, since the problem formulation is usually convex [28].

Given training data $\{x_i, y_i\}_{i=1}^N$, with $x_i \in \mathbb{R}^{d_x}$ and $y_i \in \mathbb{R}$, the regression problem in the primal space can be expressed as follows:

$$\min_{w,e,b} \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \tag{7}$$

$$\text{such that } y_i = w^T \varphi(x_i) + b + e_i, i = 1, \ldots, N. \tag{8}$$

The expression $\hat{y} = w^T \varphi(x) + b$ indicates the model in the primal space, and the objective function (7) is in fact a ridge regression cost function. With γ we indicate the regularization parameter which controls the trade-off between the model complexity and the minimization of the training error.

$$\left[\begin{array}{c|c} 0 & 1_{\mathrm{N}}^{\mathrm{T}} \\ \hline 1_{\mathrm{N}}^{\mathrm{T}} & \Omega + \mathrm{I}/\gamma \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline y \end{array}\right] \tag{9}$$

where, as before for KSC, $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_h}$ represents the mapping to a high-dimensional feature space, and $y = [y_1; \ldots; y_N]$, 1_N is a column vector of ones, $\alpha = [\alpha_1; \ldots; \alpha_N]$. The term Ω means the kernel matrix with entries $\Omega_{ij} = \varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$, and with $K : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \rightarrow \mathbb{R}$ we denote the kernel function. By using a radial basis function (RBF) kernel, expressed by $K(x_i, x_j) = \exp(-||x_i - x_j||_2^2/\sigma^2)$, one is able to construct a model of arbitrary complexity. Finally, after solving the previous linear system, the LS-SVM model for function estimation in the dual form becomes:

$$\hat{y} = \sum_{i=1}^{N} \alpha_i K(x, x_i) + b. \tag{10}$$

3.2. NAR model

A nonlinear auto-regressive model (NAR) describes time-varying phenomena specifying that the output variable depends non-linearly on its own previous values [17]:

$$\hat{y}_{k+1} = f([y_k; \ldots; y_{k-p}]) \tag{11}$$

where $\hat{y}_{k+1}$ is the predicted future value based on the previous p values, $f : \mathbb{R}^p \rightarrow \mathbb{R}$ is the nonlinear mapping. The parameter p is also called order of the NAR model and has to be properly tuned. In the LS-SVM framework the NAR model can be described by:

$$\hat{y}_{k+1} = w^T \varphi([y_k; \ldots; y_{k-p}]) + b \tag{12}$$

For multi-step ahead prediction a recursive approach can be used:

$$\hat{y}_{k+1} = f([y_k; \ldots; y_{k-p}])$$
$$\hat{y}_{k+2} = f([\hat{y}_{k+1}; \ldots; y_{k-p+1}])$$
$$\ldots$$
$$\hat{y}_{k+m} = f([\hat{y}_{k+m-1}; \ldots; \hat{y}_{k-p+m}])$$

where m is the number of steps ahead.

3.3. Model selection

In order to build-up a valid LS-SVM NAR model for time-series prediction, we have to choose very carefully the parameters γ, σ and p. The parameters γ and σ are selected using 10-fold cross-validation. The Coupled Simulated Annealing (CSA) is used to minimize the mean square error (MSE) in the cross-validation process. Also the simplex algorithm [20] is used afterwards to further improve the search of the optimum. CSA leads to an improved optimization efficiency because it reduces the sensitivity of the algorithm to the initialization parameters [34]. The optimal order p of the NAR model is found using a grid search approach.

4. Data-sets

The data are collected from a Vertical Form Fill and Seal (VFFS) machine used to fill and seal packages in different industries, but mainly in the food industry. The VFFS machine, illustrated in Figure 2, supplies film from a roll which is formed into a bag over the vertical cylinder. Sealing jaws close the bag at the bottom before it is filled. At the end of the cycle, the bag is sealed and cut off with a knife. From previous experimental studies the dirt accumulation on the sealing jaws was observed to strongly affect the process quality. For this reason, in the experiments described here the jaws were monitored to predict in advance the maintenance actions. Maintenance consists of stopping the machine and cleaning the sealing jaws. A total of three experiments have been performed. In the first

two experiments only accelerometers mounted on the jaws are used in order to indirectly detect the dirt accumulation. In the third experiment the machine is also equipped with a thermal camera, which directly measures the dirt accumulation in terms of number of hot area pixels in the acquired images (with hot we mean that the temperature of the sealing jaws is above a user-defined threshold). This has resulted in three data-sets:

- DS_I: this data-set consists of 771 events and 3 external maintenance actions. An event is related to a particular processed bag and takes place every two seconds (i.e. the sampling frequency is $0.5Hz$). Each event is associated with a 150-dimensional accelerometer signal, that is each signal is a vector of length 150 (see top of Figure 3).

- DS_II: it contains a total of $11\,632$ processed bags and 15 maintenance actions. Here the vibration signals used to monitor the dirt accumulation in the jaws are 190-dimensional time-series (as shown in the bottom part of Figure 3). This is due to a different setting of the data acquisition system for this experiment.

- DS_III: there are $3\,519$ processed bags, each one associated with a 150-dimensional vibration signal, and 11 maintenance actions. For this data-set we are given, beyond the the accelerometer signals, also the thermal camera measurements, as depicted in Figure 4.

The accelerometer signals registered just before the maintenance events are depicted in red in Figures 3 and 4 (top). They usually correspond to signals with high amplitude, which is related to the increased amount of dirt accumulated in the sealing jaws.

5. Experimental results

Here we present the analysis of the data described in the previous Section. The experiments have been conducted in Matlab R2013a, on a CORE i7 desktop PC with 16 GB of RAM memory. We used *KSClab* and *LS-SVMlab* available at `http://www.esat.kuleuven.be/stadius/ADB/software.php`

In the first part (unsupervised learning) we illustrate the performance of KSC and we compare it with k-means [18] and self organizing maps (SOMs [8]). K-means is considered among the most popular data clustering methods for industrial applications [4] and will be briefly described later in Section 5.1.4. SOMs

are an artificial neural network model able to recognize groups of similar input vectors in an unsupervised fashion, thanks to a competitive learning strategy allowing to perform vector quantization. The performance of KSC, k-means and SOMs are evaluated according to a standard internal cluster quality measure, that is the mean silhouette index, and the results are reported in table 1. The mean silhouette criterion represents how well each object lies within its cluster, averaged over every cluster [24].

In the second part (supervised learning) we explain the time-series prediction on the thermal camera data by means of an LS-SVM NAR model. The comparison with a linear model and a zero and first order extrapolation methods is also discussed.

5.1. Unsupervised learning

In this Section it is shown how KSC can be used to perform just-in-time maintenance, not too early to take full advantage of component lifetime but also not too late to avoid catastrophic failures and unplanned downtime. In particular, 2 regimes were identified, where one of them can be interpreted as normal behaviour (low degradation) and the other as critical conditions (high degradation inducing the need of maintenance). Moreover, a probabilistic interpretation of the results is also provided, which better describes the degradation process experienced by the sealing jaws of the packing machine.

We perform clustering on the raw accelerometer signals. However, since the KSC technique is formulated as a weighted kernel PCA model (see Section 2.1), it automatically extracts features from the vibration signals when performing clustering. Another procedure could be related to exploiting the frequency content of the signals by using for instance a wavelet kernel [35], but this approach has not been considered in this work.

5.1.1. Model selection

In order to catch the ongoing deterioration process of the jaws we need to use historical values of sealing quality in our analysis. For this purpose we apply a windowing operation on the data (Figure 5). If we do not apply this preprocessing step, we would perform clustering of only one signal at each time. In this way, the algorithm would detect if the single bag has been sealed correctly or not. On the other hand, when concatenating a certain number of signals, the clustering method is able to distinguish between good working conditions (many good bags in the current window) and faulty state (many bad bags in the current window).

We have a total of 3 parameters to determine: the window size (i.e. the number of signals to concatenate), the number of clusters k and the RBF kernel parameter σ. According to the BLF criterion the optimal window size is 40 and the optimal number of clusters is $k = 2$ for the three data-sets, while σ is data-set dependent.

Regarding data-set DS_I we have used 140 data points for training and the remaining 631 as test set. In the other experiments, we directly applied the model trained on the dataset DS_I on datasets DS_II and DS_III. Here, we exploited the generalization capability of the KSC algorithm and the fact that the experiments are similar to each other.

5.1.2. Hard clustering

In Figure 6 the KSC prediction for the data-set DS_I is shown. We can interpret one of the clusters as normal behaviour or low degradation and the other as maintenance cluster or high degradation[2]. Notice that the KSC model is able to predict some minutes in advance the maintenance actions before they are actually performed by the operator. Concerning the data-set DS_II we can draw the same comments: KSC is very accurate in predicting the worsening of the packing process around the actual maintenance events (see top of Figure 7). Finally, Figure 8 illustrates the results on data-set DS_III. In this case KSC predicts the need of maintenance also in zones where maintenance has not been really performed. As we will see later, surprisingly it would have been more logical to perform maintenance as suggested by KSC and not as actually done by the operator[3].

5.1.3. Soft clustering

In the previous Section we demonstrated the effectiveness of KSC in predicting in advance the maintenance events. Nevertheless the predicted output is binary (it goes suddenly from normal operation to maintenance). On the other hand a soft clustering output is in general preferable, since it can provide more interpretable results [14]. Furthermore, in our case study the discrete output does not give us a continuous indicator of the incoming maintenance actions. To solve this issue we can use the latent variable $e(x)$ instead of the binarized clustering output

[2]Also for $k = 3$ the clustering results are meaningful: the BLF reaches the second highest value and three distinct regimes interpretable as normal behaviour, warning state and maintenance state have been observed. However, in this paper we only consider two clusters, corresponding to the maximum of the BLF model selection criterion.

[3]This conclusion can be corroborated by observing that some signals in proximity of maintenance events (depicted in red in at the top of Figure 4) have rather low amplitude.

sign$(e(x))$ (see Section 2.1).

Furthermore, we can rescale it between 0 and 1 to improve the interpretability. This transformation is based on the structure of the latent variable space. As already mentioned in 2.2, in this space every cluster is ideally represented as a line. The tips of the lines can be considered prototypes of their cluster, since they have more certainty to belong to it because they are further from the decision boundaries [1]. Thus, the Euclidean distance from every point to the cluster prototype can be seen as a confidence measure of the cluster membership. The transformed latent variable is depicted at the bottom of Figure 6 (data-set DS_I), Figure 7 (data-set DS_II) and Figure 8 (data-set DS_III). The value can be considered as a soft membership or "probability" to maintenance. As explained in [3], this soft membership is given as a subjective probability and it indicates the strength of belief in the clustering assignment. In other words, it is not a probability based on statistic analysis of a large number of samples. Also, it is worth mentioning that since the KSC algorithm does not naturally provide a soft output, we had to come up with this post-processing step. However, other clustering methods like Hidden Markov Models (HMMs, [22]) are designed in a probabilistic setting and can automatically supply moderated outputs (in case of HMMs also the probabilities of transition between the different regimes are given).

By looking at the bottom of Figures 6, 7 and 8 we can see how the probability increases as the number of faulty bags in the window increases. The value can decrease since the window can move onto zones with good seals after a period of bad seals. This is probably due to a self-cleaning mechanism. Maintenance is predicted when the probability reaches the value 1.

For data-sets DS_I and DS_II it can be noticed how KSC is able to discover from the vibration signals registered by the accelerometers the dirt accumulation in the jaws that leads to the maintenance actions. Since clustering is an unsupervised technique, this is achieved by not making use of any information on the location of the maintenance actions (like it occurs for classification). For what concerns data-set DS_III, also regions where no maintenance actions have been really performed appear associated with high probability to maintenance.

To understand these unexpected outcome, we can rescale the probability to maintenance in the same range of the the number of hot area pixels in the images captured by the thermal camera. This is illustrated in Figure 9, where we depict in red the number of hot area pixels (measured degradation) and in blue the rescaled soft memberships (modelled degradation). We can recognize similar patterns, meaning that KSC was able to catch the degradation process also in this last experiment. This last observation is quite surprising, since the clustering al-

gorithm is only supposed to divide the data into two main groups, as observed at the top side of Figures 6, 7, 8. The fact that the latent variable $e^{(l)}$ of the clustering model trained on the vibration signals follows the trend of the hot area pixels data (Figure 9) is somehow unexpected. To summarize, in the third experiment probably the operator should have performed the maintenance operations in a different way, being coherent with his behaviour in the first two experiments.

5.1.4. k-means clustering

K-means clustering is among the most popular methods for finding clusters in a set of data points. After choosing the desired number of cluster centers, the k-means procedure iteratively moves the centers to minimize the total within cluster variance. k-means has several drawbacks:

- the results are strongly influenced by the initialization
- the number of clusters should be provided by the user. This can be done by using a trial and error approach: the method is applied with different number of clusters and the partition corresponding to the highest value of some quality criterion is chosen.
- it can discover only spherical boundaries.

Despite these disadvantages, together with subtractive and hierarchical clustering, it is still widely used since it works effectively in many scientific and industrial applications. For this reason, here we present the results of k-means applied on the three data-sets described in Section 4. Before discussing the results, it is worth to mention that, thanks to the model selection scheme of KSC described in Section 2.2, we give optimal parameters to k-means (number of clusters = 2 and window size of concatenated accelerometer signals = 40). Figure 10 visualizes the outcomes of the hard and soft clustering. Similarly to KSC, the soft clustering results are based on the distance between the final centroids and the data-points in the input space (see [3]). Concerning data-sets DS_I and DS_III the results of KSC and k-means are very similar, while in the data-set DS_II analysis k-means performs worse than KSC, indicating the need of maintenance where it was not really performed. In this case k-means would suggest too many maintenance actions, which is not cost-effective.

5.2. Supervised learning

In this paragraph we present the results of applying the NAR model explained in Section 3.2 for predicting the evolution of the hot area pixels in data-set DS_III.

We have used the RBF kernel function $K(u, v) = \exp(-||u - v||_2^2/\sigma^2)$ to build the nonlinear model. The optimal parameters obtained using the 10 fold cross-validation procedure summarized in Section 3.3 are: $\gamma = 1356.7$, $\sigma^2 = 301.6$. Moreover, the optimal model order, tuned using a grid search approach, is $p = 5$, as depicted in Figure 11. After selecting the optimal parameters, we performed multi-step ahead recursive prediction. As can be noticed in Figure 12, the NAR model performed very well up to 10 steps ahead prediction, i.e. about 25 seconds since a bag is processed every $2 - 3$ seconds. Afterwards, the mean absolute error between outputs and actual values starts increasing progressively, even if the predictions remains good up to 1 minute ahead prediction. In Figure 13 we visualize the results related to 10 steps ahead prediction. We can notice how the modelled degradation follows quite well the trend of the true degradation, in contrast with what was observed in Figure 9. This is not surprising because in this case we explicitly constructed a regression model to explain the deterioration, while the latent variable $e^{(l)}$ of the clustering model trained on the vibration signals in principle is not supposed to follow the trend of the hot area pixels data. The performance statistics are summarized in Table 2, where a comparison with two baseline techniques[4] and a linear autoregressive model (AR) is shown.

6. Discussion

In the first part of this work we explained the use of KSC for predictive maintenance. KSC has been chosen instead of classification because of the high unbalance of the data (few maintenance events compared to normal operating condition). After applying a windowing operation on the data in order to catch the deterioration process affecting the sealing jaws, we showed how KSC is able to recognize the presence of at least two working regimes of the VFFS machine, identifiable respectively as normal and critical operating condition. Moreover, we proposed also a soft clustering output that can be interpreted as "probability" to maintenance, and it is more directly related to the degradation phenomenon affecting the jaws (see for example Figure 9). In addition, as mentioned in Section 5.1.3, this probability is not defined in a statistical sense, but it is rather constructed to reflect the reliability level of the mechanical equipment at current time. KSC is also compared with k-means and self organizing maps. While in data-sets DS_I and DS_III all methods give quite similar results, in case of data-set DS_II k-means

[4]The two baseline models correspond to zero order and first order extrapolation methods.

performs worse than KSC and SOMs, since it predicts maintenance in regions where it was not actually performed by the operator.

In the second part a LS-SVM NAR model for predicting the evolution of the dirt accumulation in the jaws has been constructed. The dirt accumulation is directly measured by means of the number of hot area pixels present in the images obtained by the thermal camera. First we selected the parameters of the model using 10-fold cross-validation with coupled simulated annealing for γ and σ^2 and a grid search approach for the optimal model order. Then we trained the NAR model and we tested its ability in predicting the evolution of the number of hot area pixels. The performance is very good up to 10 steps ahead prediction (around 25 seconds), even if the forecast remains good up to 1 minute ahead. Moreover, the NAR outperforms the linear auto-regressive model (AR) and a zero and first order extrapolation methods.

To summarize, both the LS-SVM NAR model and KSC, thanks to their forecasting capabilities, could help to optimize the timing of maintenance actions for the machine under study. In particular, a company who would like to use the LS-SVM techniques described in this work to maximize production capacity has two choices:

- option 1 (the cheapest): install on every packing machine only accelerometers and use the proposed unsupervised learning method, that is KSC, to predict in advance, in an online fashion, when the machine starts entering critical conditions. In this way the operator can perform maintenance at the right time and avoid long stops of the production.

- option 2: install the thermal camera and use the LS-SVM NAR model to predict up to 1 minute ahead the dirt accumulation on the jaws. Since the thermal camera costs more than the accelerometers, this option is more expensive. Nevertheless it is more reliable because it is a direct prediction of the deterioration while KSC, as surprisingly effective as it is completely unsupervised, infers the degradation indirectly from the vibration signals.

7. Conclusions

Predictive maintenance of industrial machines is receiving increasing attention in the last years due to its many advantages, like cost efficiency, reduced ecological impact etc. It is based on constant monitoring the health of machines joined with advanced signal processing and prognostics techniques for optimal maintenance management. In this paper we proposed a least squares support vector

machine (LS-SVM) framework for maintenance strategy optimization based on real-time condition monitoring of a packing machine. We presented an unsupervised approach through kernel spectral clustering (KSC), and a supervised learning method, namely nonlinear auto-regression (NAR). In the first case we used the data collected by accelerometers positioned on the jaws of a Vertical Form Fill and Seal (VFFS) machine, from which the degradation process of the machine conditions has been inferred. For the time-series analysis with the NAR model data acquired by a thermal camera, which directly measures the dirt accumulation in the jaws, are processed. We showed that LS-SVM can successfully assess and predict mechanical conditions based on sensor data, thanks to their ability to model the degradation process. Moreover LS-SVM achieved higher performance than basic methods, which are commonly used in practice to predict the forthcoming faults. Finally, we proposed two options to use LS-SVM to schedule maintenance on the packing machine with different associated costs.

Acknowledgements

EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC AdG A-DATADRIVE-B (290923). This paper reflects only the authors views and the Union is not liable for any use that may be made of the contained information. Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; PhD/Postdoc grants Flemish Government: FWO: projects: G.0377.12 (Structured systems), G.088114N (Tensor based data similarity); PhD/Postdoc grants IWT: projects: SBO POM (100031); PhD/Postdoc grants iMinds Medical Information Technologies SBO 2014 Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, Dynamical systems, control and optimization, 2012-2017).

References

[1] Alzate, C., Suykens, J. A. K., 2010. Highly sparse kernel spectral clustering with predictive out-of-sample extensions. In: Proc. of the 18th European Symposium on Artificial Neural Networks (ESANN 2010). pp. 235–240.

[2] Alzate, C., Suykens, J. A. K., February 2010. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (2), 335–347.

[3] Ben-Israel, A., Iyigun, C., Jun. 2008. Probabilistic d-clustering. J. Classif. 25 (1), 5–26.

[4] Berkhin, P., 2006. A survey of clustering data mining techniques. Grouping Multidimensional Data, 25–71.

[5] Bey-Temsamani, A., Engels, M., Motten, A., Vandenplas, S., Ompusunggu, A. P., 2009. A practical approach to combine data mining and prognostics for improved predictive maintenance. In: Proceedings of the Third International Workshop on Data Mining Case Studies (DMCS). pp. 36–43.

[6] Choi, S. W., Yoo, C. K., Lee, I.-B., 2003. Overall statistical monitoring of static and dynamic patterns. Ind. Eng. Chem. Res. 42, 108 – 117.

[7] Chung, F. R. K., 1997. Spectral Graph Theory. American Mathematical Society.

[8] Kohonen, T., Schroeder, M. R., Huang, T. S. (Eds.), 2001. Self-Organizing Maps, 3rd Edition. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

[9] Kourti, T., MacGregor, J. F., 1995. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems 28 (1), 3 – 21.

[10] Kriegel, H.-P., Kroger, P., Zimek, A., 2010. Outlier detection techniques. 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD).

[11] Langone, R., Agudelo, O. M., De Moor, B., Suykens, J. A. K., September 2014. Incremental kernel spectral clustering for online learning of non-stationary data. Neurocomputing 139, 246–260.

[12] Langone, R., Alzate, C., De Ketelaere, B., Suykens, J. A. K., 2013. Kernel spectral clustering for predicting maintenance of industrial machines. In: IEEE Symposium Series on Computational Intelligence and Data Mining (CIDM) 2013. pp. 39–45.

[13] Langone, R., Alzate, C., Suykens, J. A. K., 2013. Kernel spectral clustering with memory effect. Physica A: Statistical Mechanics and its Applications 392 (10), 2588–2606.

[14] Langone, R., Mall, R., Suykens, J. A. K., 2013. Soft kernel spectral clustering. In: Proc. of the International Joint Conference on Neural Networks (IJCNN 2013). pp. 1028–1035.

[15] Li, T.-S., Huang, C.-L., 2009. Defect spatial pattern recognition using a hybrid somsvm approach in semiconductor manufacturing. Expert Systems with Applications 36 (1), 374 – 385.

[16] Liao, W., Pan, E., Xi, L., Dec. 2010. Preventive maintenance scheduling for repairable system with deterioration. J. Intell. Manuf. 21 (6), 875–884.

[17] Ljung, L. (Ed.), 1999. System identification (2nd ed.): theory for the user. Prentice Hall PTR.

[18] MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: of California Press., U. (Ed.), Fifth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. pp. 281–297.

[19] Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., Rätsch, G., 1999. Kernel pca and de-noising in feature spaces. In: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II. pp. 536–542.

[20] Murty, K. G. (Ed.), 1983. Linear programming. Wiley.

[21] Ng, A. Y., Jordan, M. I., Weiss, Y., 2002. On spectral clustering: Analysis and an algorithm. In: Dietterich, T. G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, pp. 849–856.

[22] Rabiner, L., Juang, B., 1986. An introduction to hidden markov models. IEEE Acoutics, Speech and Signal Processing Magazine 3, 4–16.

[23] Randall, R. B., 2011. Vibration-based condition monitoring industrial, aerospace and automotive applications. Wiley.

[24] Rousseeuw, P., Nov. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (1), 53–65.

[25] Sarmiento, T., Hong, S., May, G., April 2005. Fault detection in reactive ion etching systems using one-class support vector machines. In: Advanced Semiconductor Manufacturing Conference and Workshop, 2005 IEEE/SEMI. pp. 139–142.

[26] Sharma, A., Yadava, G., Deshmukh, S., 2011. A literature review and future perspectives on maintenance optimization. Journal of Quality in Maintenance Engineering 17, 5–25.

[27] Suykens, J. A. K., Alzate, C., Pelckmans, K., 2010. Primal and dual model representations in kernel-based learning. Statistics Surveys 4, 148–183.

[28] Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J., 2002. Least Squares Support Vector Machines. World Scientific, Singapore.

[29] Van Gestel, T., Suykens, J. A., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., de Moor, B., Vandewalle, J., 2004. Benchmarking least squares support vector machine classifiers. Machine Learning 54 (1), 5–32.

[30] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part i: Quantitative model-based methods. Computers and chemical engineering 27 (3), 293–311.

[31] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part ii: Qualitative models and search strategies. Computers and chemical engineering 27 (3), 313–326.

[32] Venkatasubramanian, V., Rengaswamy, R., Kavuri, S., 2003. A review of process fault detection and diagnosis. part iii: Process history based methods. Computers and chemical engineering 27 (3), 327–346.

[33] von Luxburg, U., 2007. A tutorial on spectral clustering. Statistics and Computing 17 (4), 395–416.

[34] Xavier-De-Souza, S., Suykens, J. A. K., Vandewalle, J., Bollé, D., Apr. 2010. Coupled simulated annealing. Trans. Sys. Man Cyber. Part B 40 (2), 320–335.

[35] Zhang, L., Zhou, W., Jiao, L., 2004. Wavelet support vector machine. IEEE Transactions on Systems, Man, and Cybernetics, Part B 34 (1), 34–39.

	DS_I	DS_II	DS_III
KSC	0.29	**0.25**	0.27
k-means	0.28	0.12	0.27
SOMs	0.29	0.21	**0.28**

Table 1: **Cluster quality evaluation.** Mean Silhouette index (see end of section 5.1.4), with best performance in bold. In the case of data-set DS_II the low value of Silhouette indicates that k-means does not succeed in correctly separating the normal behaviour and the maintenance cluster.

	LS-SVM NAR	AR	Baseline 1	Baseline 2
MAE	**227.03**	247.25	371.99	251.34
Percentage error	**1.42%**	4.11%	6.17%	4.18%
R^2	**0.93**	0.92	0.77	0.92

Table 2: **Summary NAR performance on test data.** Best performance in bold. By baseline 1 we mean a zero order extrapolation model and baseline 2 is related to a first order extrapolation method.

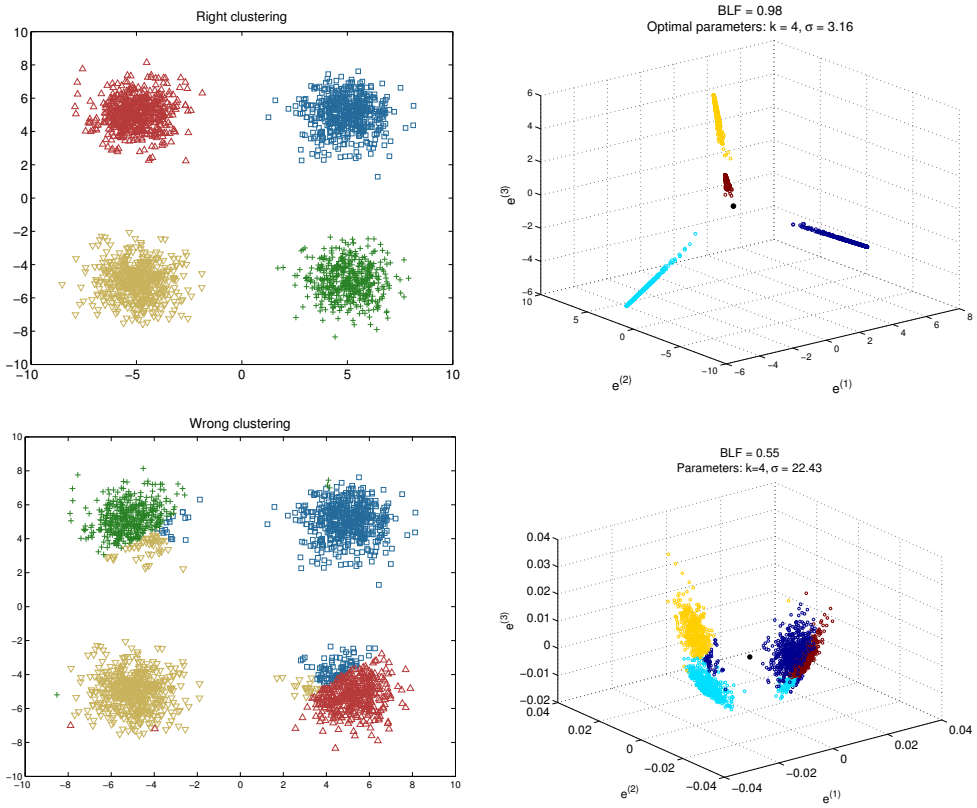

Figure 1: **Model selection illustrative example on synthetic toy data:** for well-chosen kernel parameters, the clusters for out-of-sample data are collinear in the projection space. **Top left:** Correct clustering of 4 Gaussian clouds. **Top right:** Projection space corresponding to an optimal σ value, for which the BLF is maximal. **Bottom left:** Wrong clustering results. **Bottom right:** Projection space for a wrong σ value.

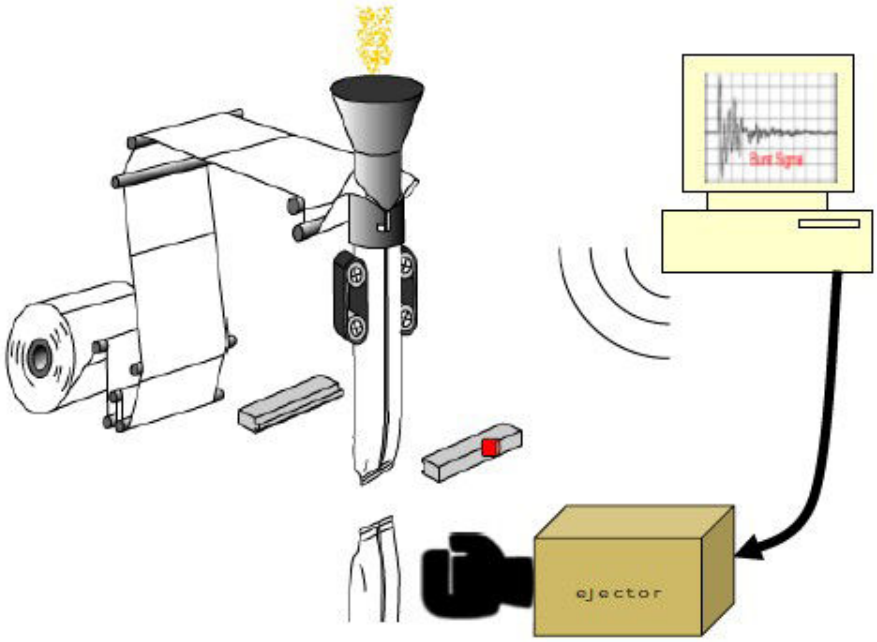

Figure 2: **VFFS machine.** Seal quality monitoring in a packing machine.

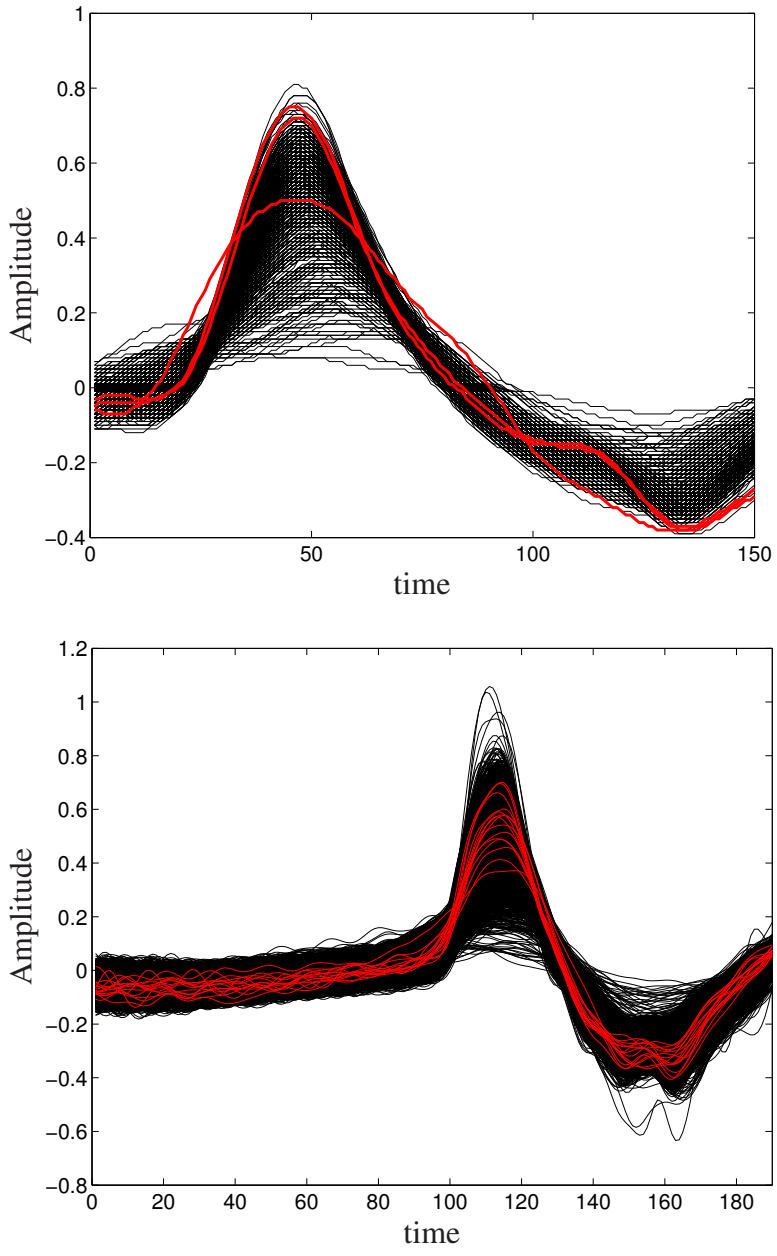

Figure 3: **Data-sets DS_I-DS_II. Top:** Accelerometer signals for the data-set DS_I. **Bottom:** Accelerometer signals for the entire data-set DS_II. The signals corresponding to maintenance actions are depicted in red (best visible in colors).

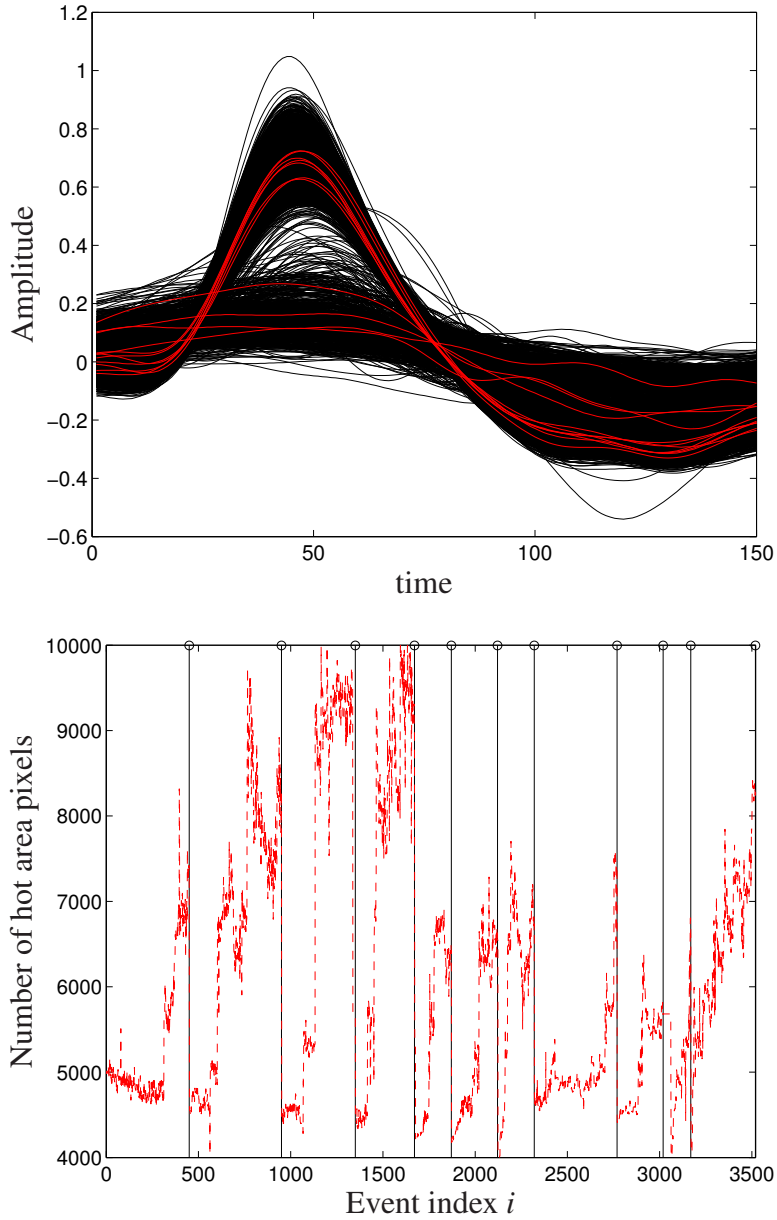

Figure 4: **Data-set DS_III. Top:** Accelerometer signals (signals corresponding to maintenance actions are pictured in red). **Bottom:** thermal camera data (the vertical black lines indicates true maintenance actions).

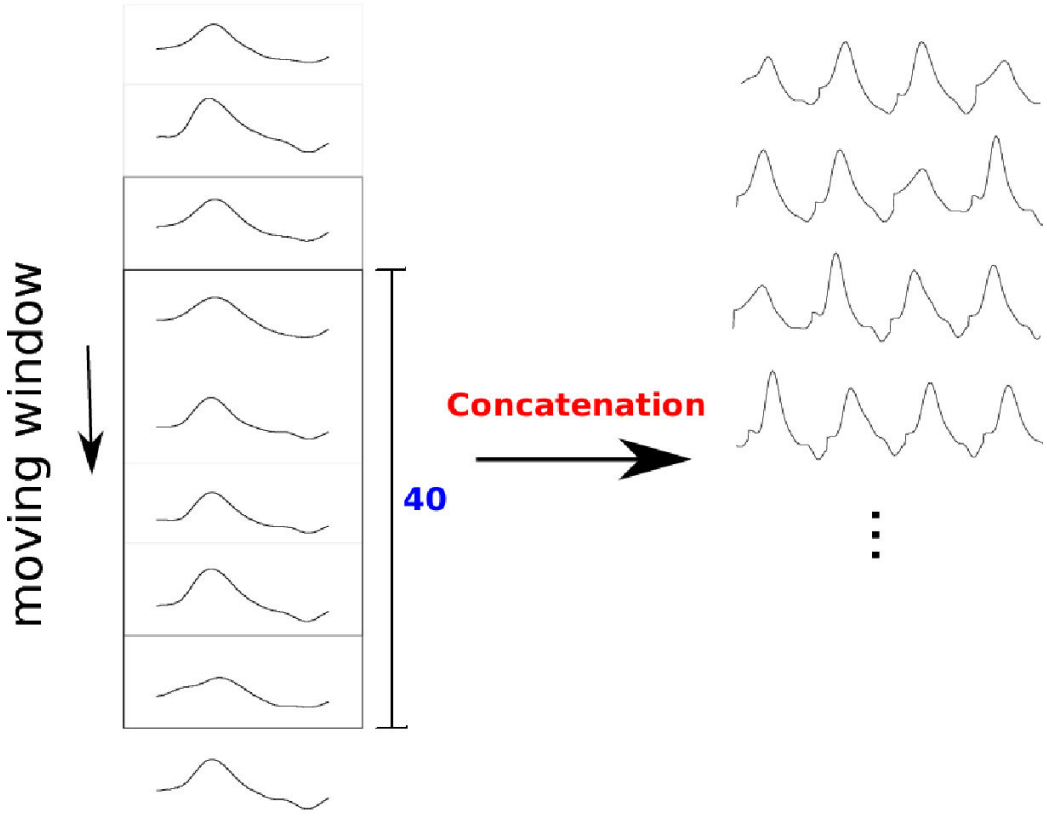

Figure 5: Concatenation of accelerometer signals. After the windowing operation, each data-point is now a time-series of dimension $d = 40 \times 150$ for the first and third data-sets and $d = 40 \times 190$ for data-set DS_II.

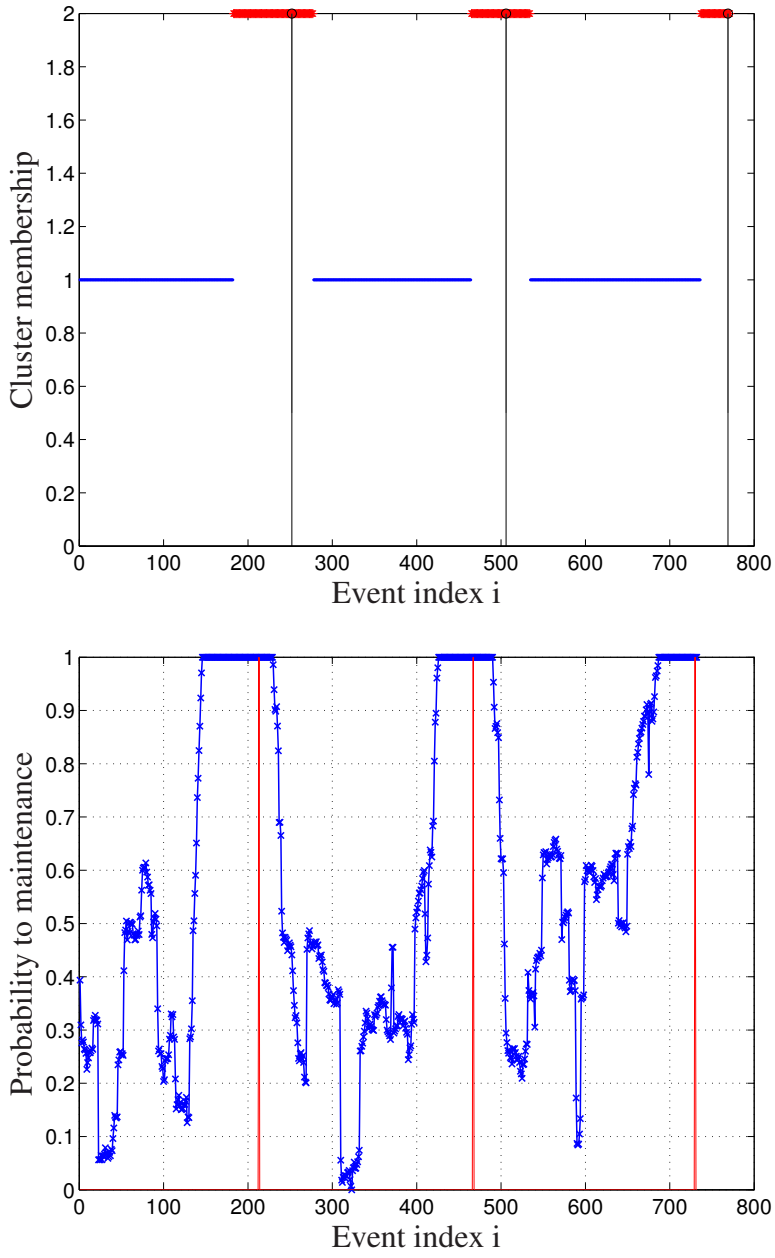

Figure 6: **KSC results data-set DS_I. Top:** Hard clustering results for the whole data-set. Cluster 2 represents predicted maintenance events. The vertical black lines show the true maintenance. **Bottom:** Soft clustering results in terms of probability to maintenance.

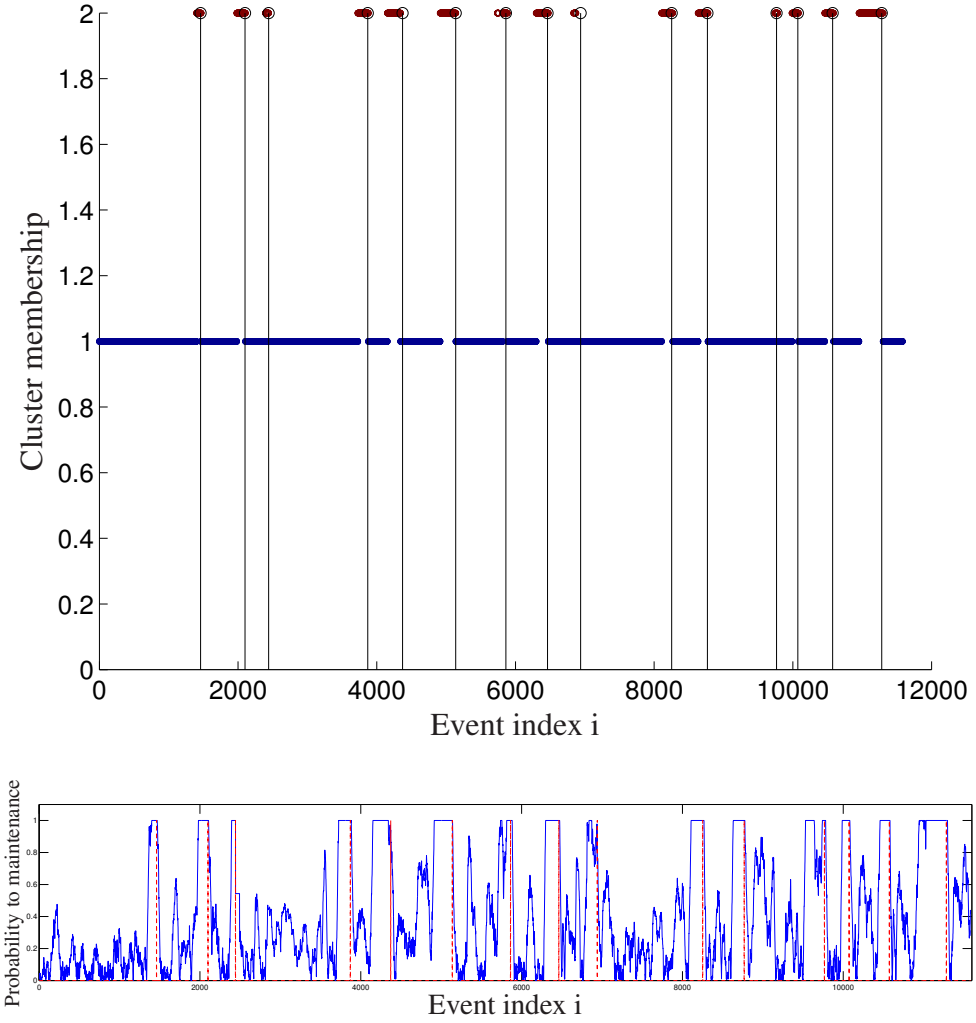

Figure 7: **KSC results data-set DS_II. Top:** Hard clustering, cluster 2 represents predicted maintenance events. **Bottom:** Soft clustering in terms of probability to maintenance.

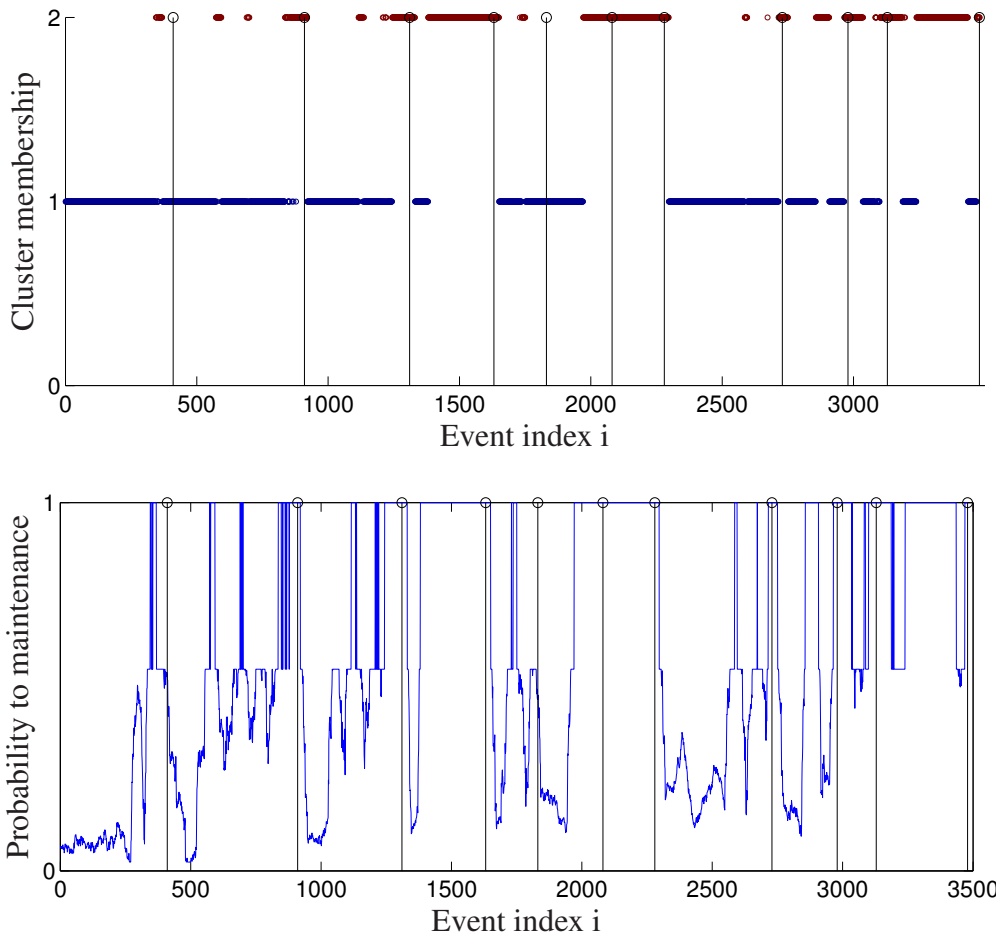

Figure 8: **KSC results data-set DS_III. Top:** Hard clustering, cluster 2 represents predicted maintenance events. **Bottom:** Soft clustering in terms of probability to maintenance.

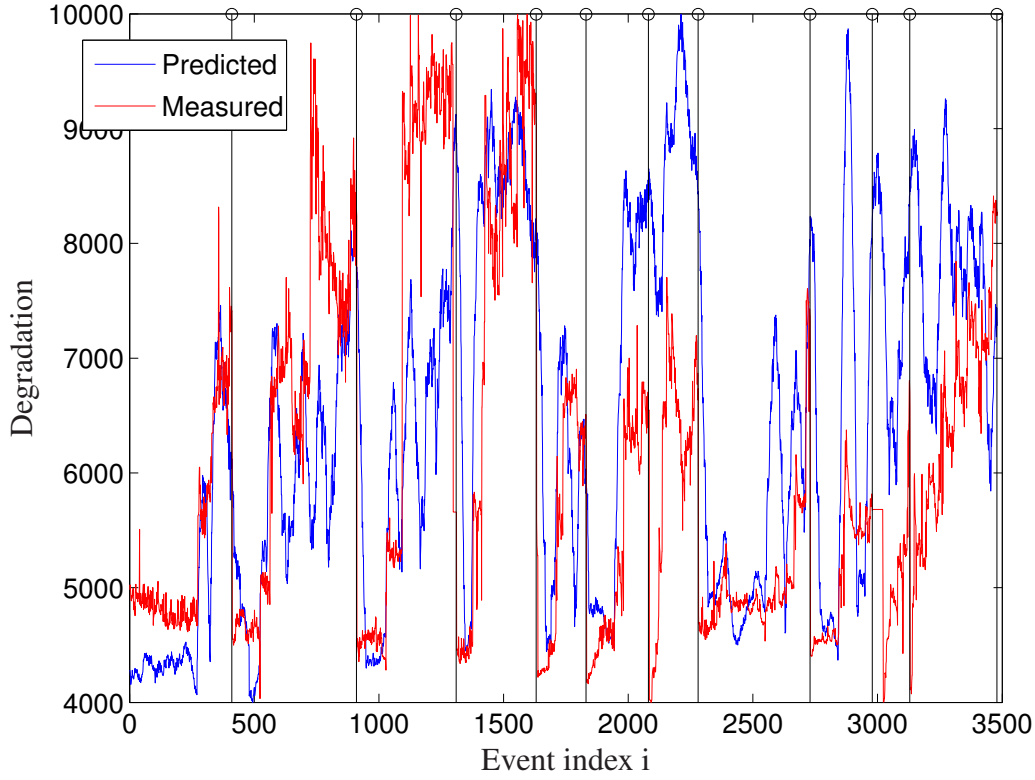

Figure 9: **Degradation dataset DS_III**. Degradation inferred by KSC using as input data the vibration signals (blue) and measured degradation in terms of hot area pixels in the thermal camera images (red).

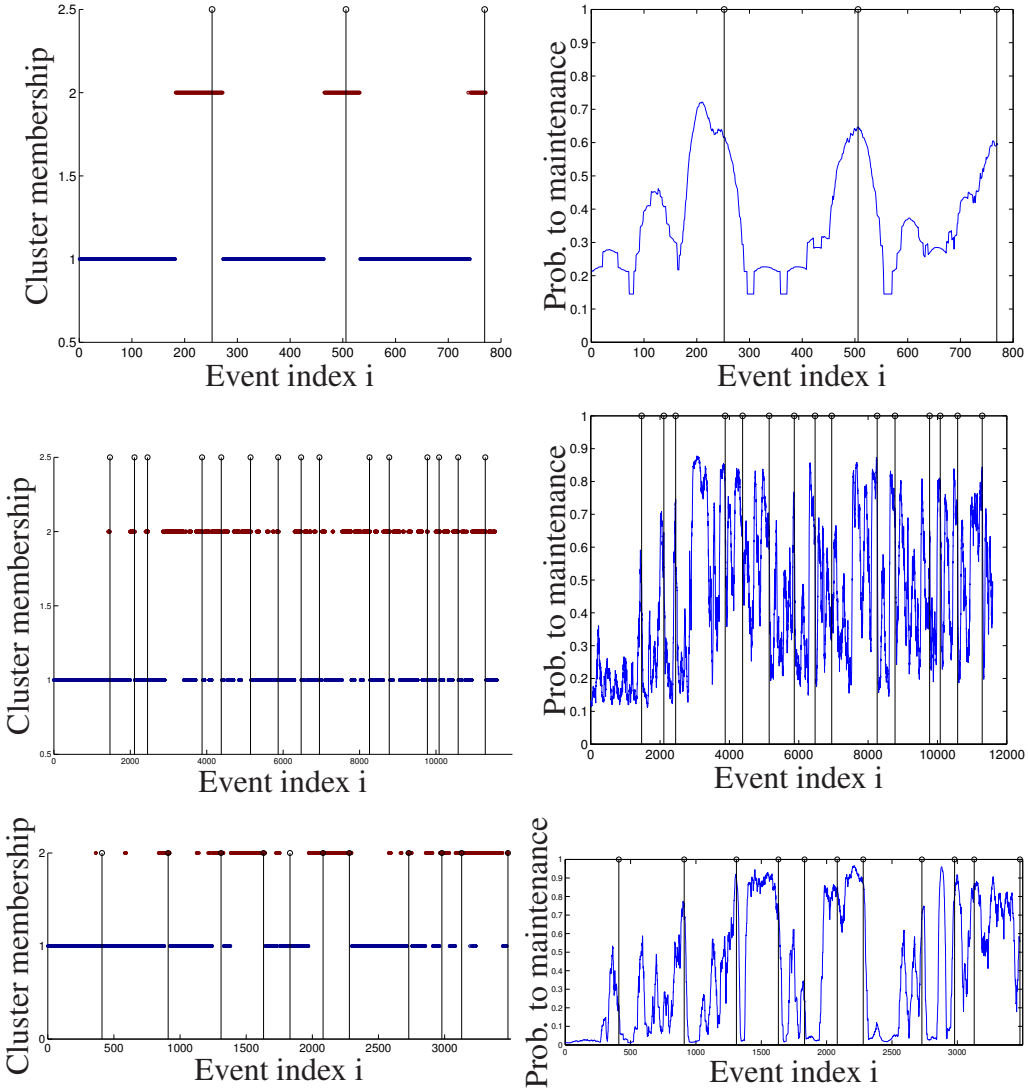

Figure 10: **k-means results.** Cluster 1 symbolizes the normal operating condition, cluster 2 represents predicted maintenance events. The vertical black lines show the true maintenance. **Top**: data-set DS_I. The results are similar to KSC outcomes, even if slightly worse since in the end there is a kind of false alarm (a single prediction of maintenance followed by normal behaviour, before the final maintenance cluster). **Center**: data-set DS_II. In this case k-means performs much worse than KSC, suggesting maintenance in regions not corresponding to actual maintenance events. **Bottom**: data-set DS_III. The results of KSC and k-means are very similar.

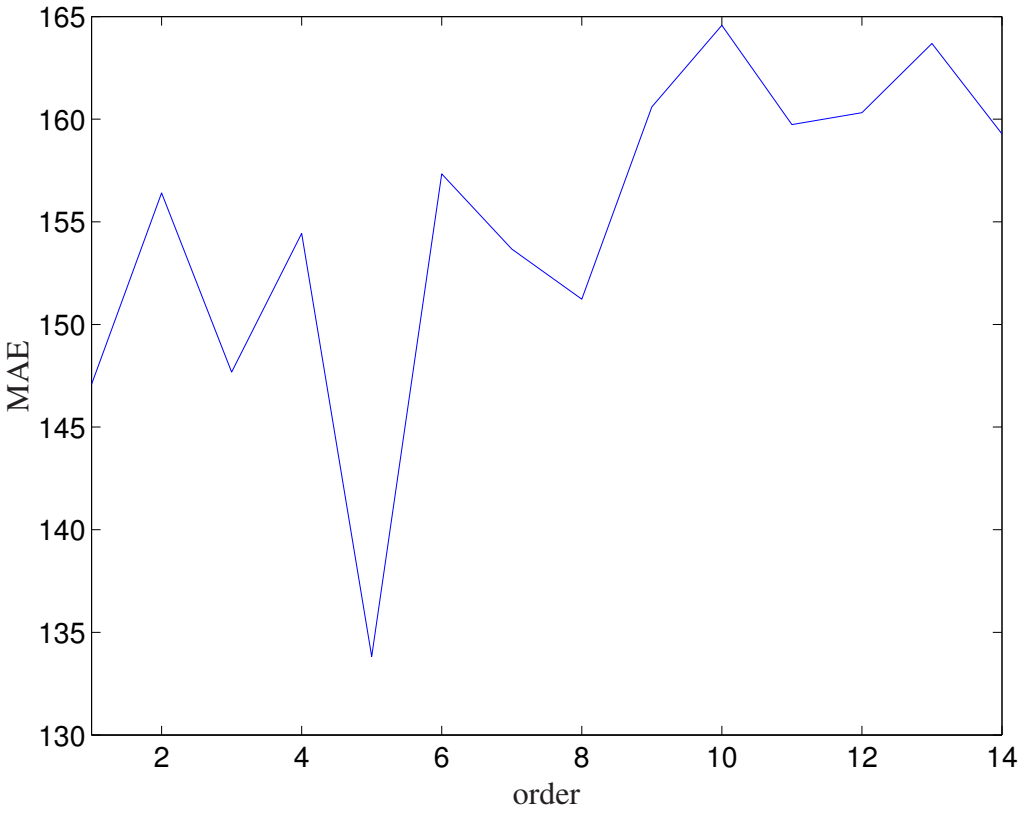

Figure 11: **Tuning order NAR.** Mean absolute error (MAE) between one step ahead prediction and true values with respect to the order of the NAR model. A clear minimum is present at $p = 5$, suggesting to consider this value as the optimal parameter to construct the final NAR model.

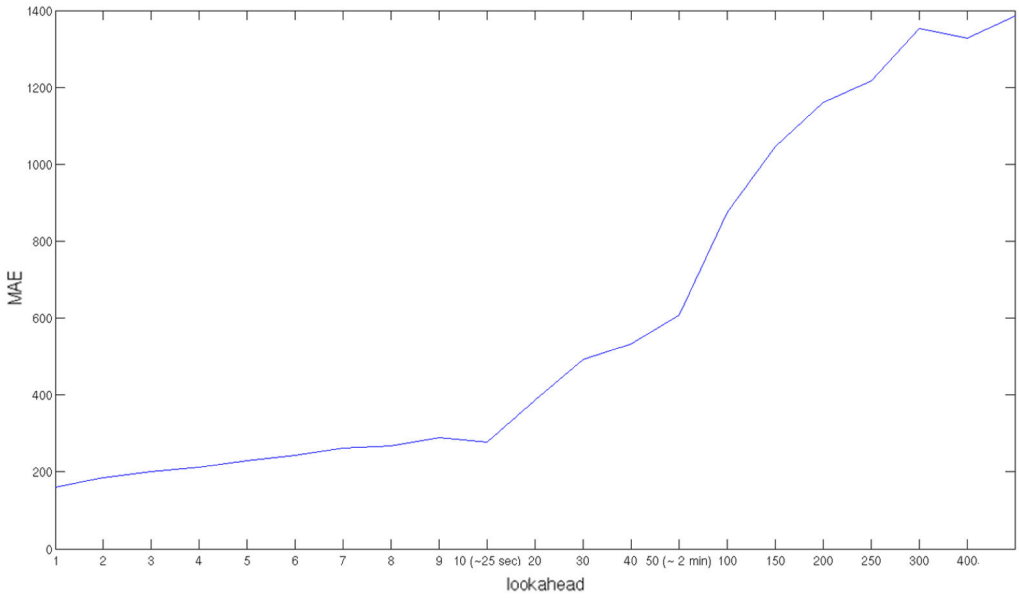

Figure 12: **NAR forecasts versus lookahead.** Prediction performances at the change of the number of steps ahead prediction.

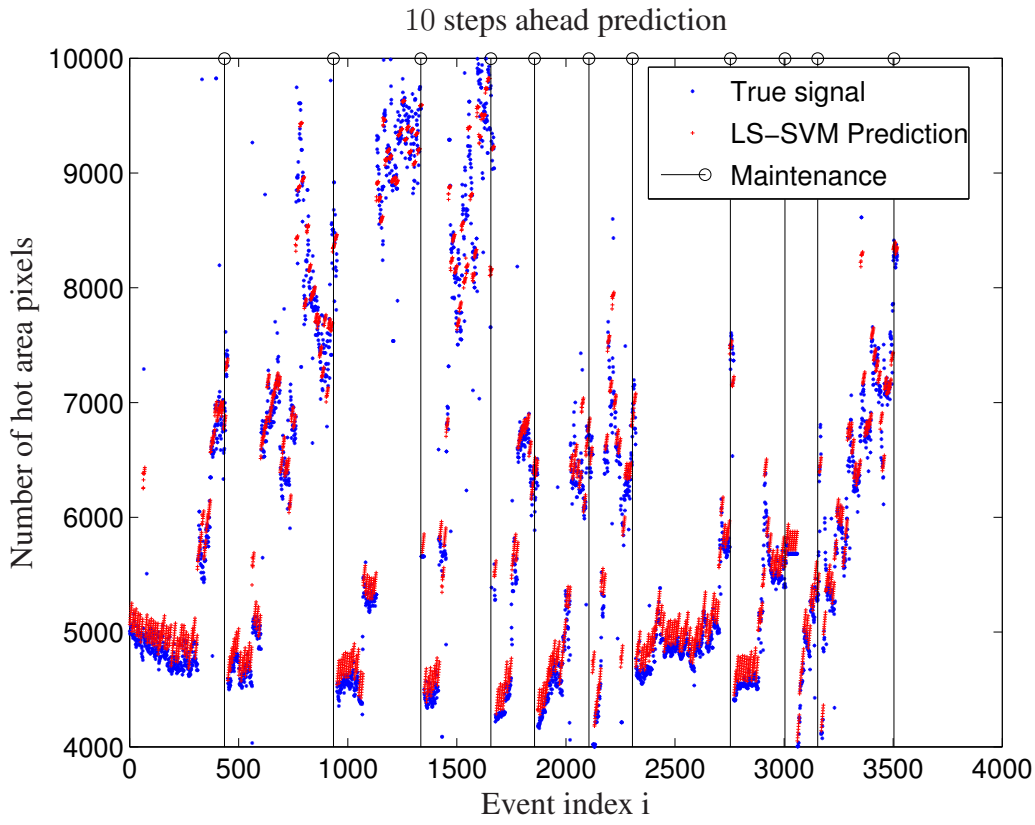

Figure 13: **Prediction NAR.** The NAR model is able to predict very well even 10 steps ahead the future behavior of the thermal camera data based on a window of the 5 past values.