State detection and RUL prediction of industrial plant components in the absence of fault data. Comparison between multivariate control charts and one-class SVM: a case study.

Navicelli A.*, De Carlo F.*, Tucci M.*

** Department of Industrial Engineering, University of Florence, Via di S. Marta, 3 50139 - Florence - Italy*
(andrea.navicelli@unifi.it, filippo.decarlo@unifi.it)

Abstract: Predictive maintenance for critical components' monitoring in industrial plants has aroused the interest of many researchers in recent years. The typical phenomenology of industrial plants' failures shows degradation of performance before the occurrence of the failure event; therefore, predictive maintenance is the most suitable technique to intercept them. To implement prognostics is necessary to have a lot of data on system behavior in both nominal and degraded conditions up to the failure event. With this information, it is possible to build a suitable prognostic model using a mathematical-statistical or machine learning technique. The advent of the fourth industrial revolution favored the collection of real-time assets' data. The low failure rate that characterizes most critical assets of industrial plants, result in a lot of nominal conditions' and an absence of degraded conditions' data, hampering the implementation of prognostic. In this article, have been developed, validated, and compared on a case study two prognostic techniques using only nominal condition data. The first one is based on the multivariate control charts (Hotelling); the second one uses the one-class Support Vector Machine model. Both techniques, combined with an Autoregressive Integrated Moving Average time series analysis model, allow the real-time prediction of the anomalous operating condition of the monitored asset. Since we don't have any fault data acquired on field, both the prognostic models developed can predict significant deviations from nominal operating conditions due to the degradation phenomena, but they can't characterize the failure mode that will arise until the failure occurs for the first time. The two models were applied to a case study to verify their robustness in predicting deviations from the nominal operating conditions of a multistage compressor caused by surge phenomenon.

Keywords: Predictive maintenance, state detection, RUL prediction, multivariate control charts, Support Vector Machine

1.Introduction

In such a competitive industrial context, it is necessary to control both production (De Carlo et al., 2014) and maintenance performance. with regard to maintenance performance, predictive maintenance is a maintenance philosophy that is part of preventive maintenance techniques together with periodic and condition-based maintenance (Ahmad and Kamaruddin, 2012). It is based on real-time monitoring of the operating conditions of industrial plant components and process systems to estimate their Remaining Useful Life (RUL). Once this parameter is known, maintenance engineering can easily optimize and schedule the maintenance interventions necessary to restore the performance of the machines, with the consequent reduction of corrective interventions, responsible for most of the total maintenance costs (Löfsten, 1999). The overall maintenance costs are defined as the cost for the ordinary interventions, planned and organized in advance, and for the extraordinary interventions, caused by unforeseen breakdowns, which have a greater economic impact on the company due to the longer period of out of service of the machinery (Komonen, 2002). Another obvious direct benefit of a predictive maintenance system is the increased availability of machinery. (Si et al., 2011) The implementation of a predictive maintenance system is also the most appropriate technique for 89% of failure modes, compared to periodic maintenance, which is appropriate in the remaining 11%. (Hashemian, 2010) So the interest of researchers in predictive maintenance has increased exponentially since 1952 when the first article was published, as shown in Figure 1.

Many studies have sought to develop maintenance support systems (De Carlo et al., 2013) and the last frontier is the development of a predictive maintenance system that requires data of the analyzed asset during its evolution towards failure and an advanced internet of things (IoT) infrastructure (Kanawaday and Sane, 2017). The high reliability of most of the critical plant components, combined with the strong use of periodic maintenance, result in a low number of failure data that makes it

apparently impossible to estimate the RUL of the industrial critical components.

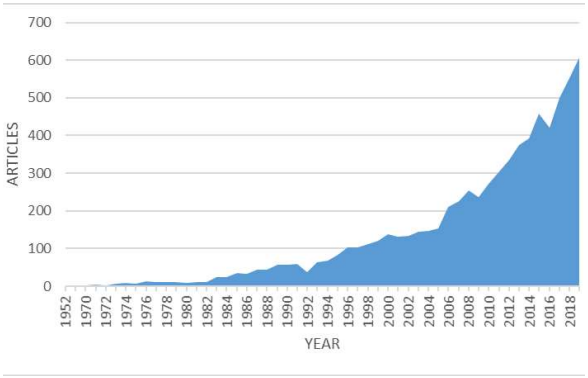

Figure 1: Number of published articles per year (Keywords: Predictive Maintenance, Limits: Articles, Source: Scopus.com).

In this case, the literature proposes, state detection techniques based on Statistical Process Control (SPC), which simply warns in case of anomalous behavior of the asset (Alwan and Roberts, 1988; Bersimis et al., 2007; MacGregor and Kourti, 1995; Mason and Young, 2002; Oakland, 2007). In this context, this study aims to build a RUL prediction methodology that can be used in the absence of failure data. The idea is to combine a state detection model with a time series analysis to estimate the RUL. For this purpose, two different state detection methods based on the SPC principles have been built: the first one is based on multivariate control charts (Hotelling charts) and the second one instead is based on the Machine Learning model called One-class Support Vector Machine (SVM). Both techniques combined the selected control variables, that are related to the failure mode under analysis, and gives as output a single variable proxy of the component performance. Applying the ARIMA (AutoRegressive Integrated Mouving Average) time series analysis technique to the output of the state detection models, it is possible to predict a possible excessive drop in component performance and assimilate it to a failure event. ARIMA model is widely used and considered one of the most comprehensive mathematical techniques of time series analysis also for industrial application. (Chen et al., 2009; Pai and Lin, 2005; Zhang, 2003) The two constructed RUL prediction models were then compared by applying them to a case study.

In the next chapter, the methodology developed, and the main mathematical techniques used for its development will be presented. The chapter "Case study" presents its application to the surge failure mode of the multistage centrifugal compressor of a geothermal power plant, and finally, the methodology and future developments will be discussed in the last 2 sections.

2 Methodology

In this chapter, it is presented first all the mathematical techniques used for the development and comparison of the prognostic models. In the last section has instead presented the logic used to build and compare them.

2.1 Multivariate control charts

The multivariate control charts, introduced in 1947 by Harold Hotelling, make it possible to aggregate information about certain process variables on a diagram using the so-called T^2 statistics. The T^2 control charts are based on the normal multivariate distribution and the Mahalanobis distance, i.e. the distance of the set of variables acquired from the average of the Gaussian multivariate distribution fitted on the training data set. The values of the mean vector and covariance matrix of the population are estimated from the historical data set of the selected control variables. The T^2 Hotelling statistic is defined as follows:

are:

$$x_1, \dots, x_n \quad (1)$$

column vectors of real numbers where each column represents the history of a selected control variable and:

$$\bar{x} = (x_1 + \cdots + x_n)/n \quad (2)$$

Their averages. Is:

$$W = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'/(n-1) \quad (3)$$

The matrix of their variances and μ are the column vector of the estimated population averages, you have:

$$T^2 = (\bar{x} - \mu)' W^{-1} (\bar{x} - \mu) \quad (4)$$

(Bersimis et al., 2007; Ilin and Raiko, 2010).

The T^2 statistic is the single proxy control variable of the performance of the analyzed component used to detect significant performance variations and to be analyzed using the time series analysis model to predict when performance will fall below the acceptability threshold.

Since T^2 is the distance of the set of variables from the average of the Gaussian multivariate built on the historical data of full operation condition, an increase of the T^2 parameter corresponds to a decrease in the performance of the component.

2.2 One-class SVM

The unsupervised one-class learning or SVM aims to separate data from the origin in n-dimensional space, with n equal to the number of control variables. It is an algorithm used to detect abnormal values. The algorithm is trained on the historical data set of the selected control variables related to the full capacity period of the analyzed component to minimize the double expression:

$$0{,}5 \sum_{jk} \propto_j \propto_k G(x_j, x_k) \quad (5)$$

Compared to $\propto_1, \dots, \propto_n$ liable to:

$$\sum \alpha_j = n\nu \quad (6)$$

With $0 \leq \propto_j \leq 1$ for each $j = 1, \dots, n$. The value of $G(x_j, x_k)$ is an element (j, k) of the Gram matrix.

A small value of ν leads to fewer support vectors and, therefore, to a smooth and raw decision-making boundary. A great value of ν leads to a greater number of support vectors and, therefore, to a sinuous and flexible decision-making contour. The optimal value of ν should be large

enough to capture the complexity of the data and small enough to avoid overtraining. Also,

$0 < \nu \leq 1$.

(Schölkopf et al., 2001)

Once trained on the historical data of the control variables, the model assigns to each set of new data the likelihood value of belonging to the training population. The likelihood value is therefore the single proxy control variable of the performance of the controlled component to be used to detect significant changes in performance and to be analysed using the time series analysis model to predict when performance will fall below the acceptability threshold.

The higher is the likelihood value of the new set of data the closer they are to the data used for training the SVM model. As the training is done on the data related to the full capacity conditions of the machinery, the higher the likelihood value and the better the performance of the component will be.

2.3 ARIMA

The ARIMA (AutoRegressive Integrated Moving Average) method is a model that, using 3 parameters (p,d,q), can model the trend of a time series and predict what will happen in the following temporal instants. The 3 parameters indicate:

- p: order of the self-regressive model;
- d: degree of the first differential;
- q: moving average model order.

With $p,d,q \in \mathbb{N}$.

The determination of the best parameters to be used for the construction of the ARIMA model is fundamental to obtain a good forecast.

The model used in the case study performs an automatic optimization of the three parameters by minimizing the Aikaike Information Criterion (AIC). This parameter has a value equal to:

$$aic = -2(logL) + 2(numParam) \quad (7)$$

With:

$$numParam = p + d + q \quad (8)$$

This criterion prefers ARIMA models with high loglikelihood compared to the data used to build the model and penalizes those with a high number of parameters (Box et al., 2015; Pai and Lin, 2005; Zhang, 2003). By setting the desired number of forecasts points, the model returns for each forecast point its expected value and its standard deviation assuming a Gaussian noise around the variable's trend. In the next chapter, the two developed RUL prediction models will be applied to a single case study and the prediction performance of the RUL will be compared.

2.4 RUL prediction models

As described in the previous chapter, the RUL prediction models have been built only on the full capacity data of the component under analysis. The application to a case study of which we have also the failure data, allows the comparison of the two developed models in the RUL forecast. The scheme followed for the development and comparison of RUL prediction models is schematically represented in Figure 2.

The logic developed to estimate the RUL of industrial plant components involves the use of two different classes of statistical mathematical models:

- Statistical process control model (models 1 and 2 in Figure 2);
- Time series analysis model (model 3 in Figure 2).

The first one is a model designed to identify anomalous behaviour of the component under control, while the second one analyses the evolution of its performance over time and forecasts its future trend. By setting a minimum threshold of acceptability of the component performance, the model can provide an estimation of the RUL of the component under analysis.

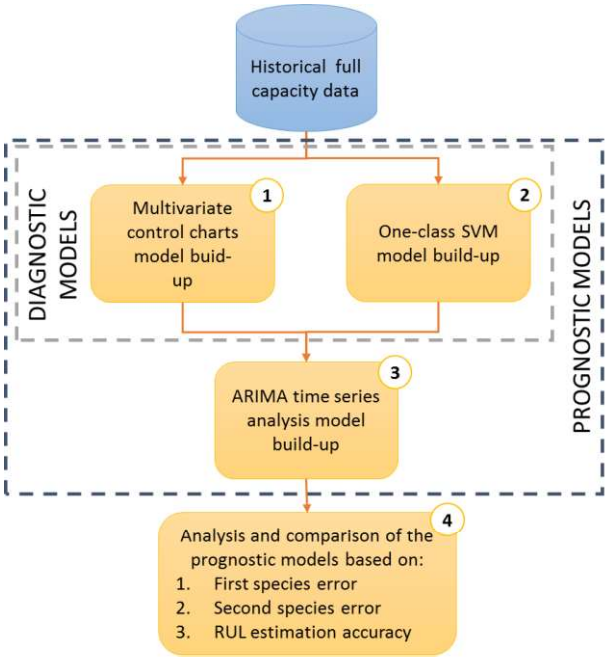

Figure 2: Scheme of development and comparison of RUL prediction **models**

The performance of industrial plants can be derived from a set of process variables that describe their operation. To build a RUL prediction model, it is necessary to acquire in real-time the value of these variables and process them through the SPC model to obtain a single control variable related to the performance of the analysed component. The control variables that can be acquired in real-time to estimate the performance of industrial plant components are typically: vibrations, temperatures, pressures and flow rates. For this purpose, two different multivariate statistical process control models were used, applied to a single case study and compared:

- Multivariate control charts or Hotelling charts;
- One-class SVM;

The first one is a classical statistical SPC model while the second one is a machine learning model. The component's control variable, proxy of the performance, is then analysed by a time series analysis model that predicts its trend over time. The model selected and developed (activity 3 in Figure 2) is an autoregressive integrated moving average model (ARIMA). Once the two RUL prediction models have been built on historical data, the application to the case study allows us to compare their performance. The comparison parameters used are 3:

- First species state detection model error;
- Second species state detection model error;
- Accuracy in estimating the RUL

3 Case study

The methodology developed was applied to a geothermal energy production plant. The plant follows a single flash condensing cycle and can develop power, when fully operational, of about 20 MW. The machinery studied is the compressor for the extraction of incondensable gases. It is 3-stage centrifugal turbomachinery that brings back the incondensable gases, processed together with the steam by the turbine and the condenser, under atmospheric pressure conditions and at a temperature of about 170 °C; conditions that allow it to be effectively treated before their release in the atmosphere.

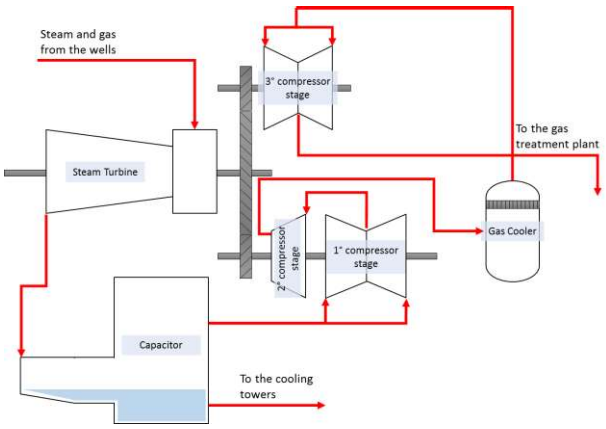

Figure 3: Case study compressor diagram

The analysis of the historical data of the compressor together with the specialized personnel of the case study company, allowed us to select the data related to the full capacity of the machinery for the training of the state detection models and to use the data related to the surge phenomenon to verify that these give a performance value of the compressor significantly lower than those used for the training of the model. The application of the time series analysis model also allows us to evaluate the accuracy of the RUL estimation of the two models. Given the nature of the machinery, all the sensors related to the thermofluid-dynamic conditions (pressure, temperature, and flow rate) of the processed gas at the inlet and outlet of all compression stages have been selected for a total of 16 control variables related to the compressor performance.

The historical period analyzed, corresponding to about 8 days of data acquisition with 1-second sampling, was divided together with the plant personnel into 3 different periods: full capacity, anomaly and surge.

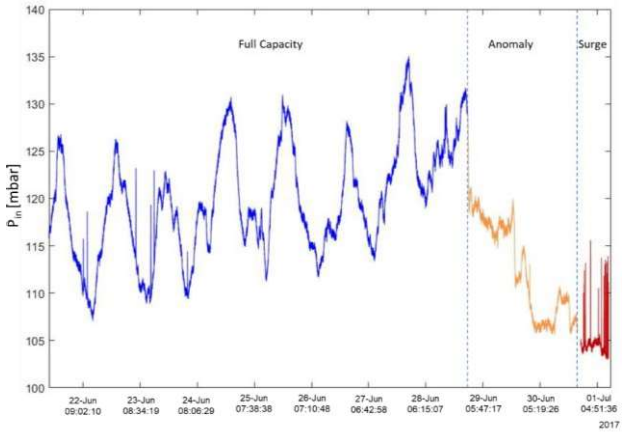

Figure 4: Example of operating conditions periods division

The data for the period of full capacity count 635782 elements for each of the 16 control variables, 166473 for the period of anomaly operation, and 42103 for the surge period. A large amount of data during the regime period used to train the two state detection models ensures their robustness.

3.1 Application of multivariate control cards

On the data has been built the multivariate control charts using the "*pca*" function of the Matlab software. One of the function outputs is the T^2 statistic, control variable related to compressor performance. The trained model was then applied to the anomaly and surge period data. Given the high noise of T^2 statistics, it was decided to make a moving average with a time bucket of 300 seconds to make the ARIMA forecast more stable and the SPC more robust (Kay, 1993). Figure 5 shows the trend of the moving average of the T^2 statistics over time with the associated upper control limit (UCL) of 25,06 with 95% confidence and the lower control limit for the surge period (SCL) of 77,36 with 5% confidence:

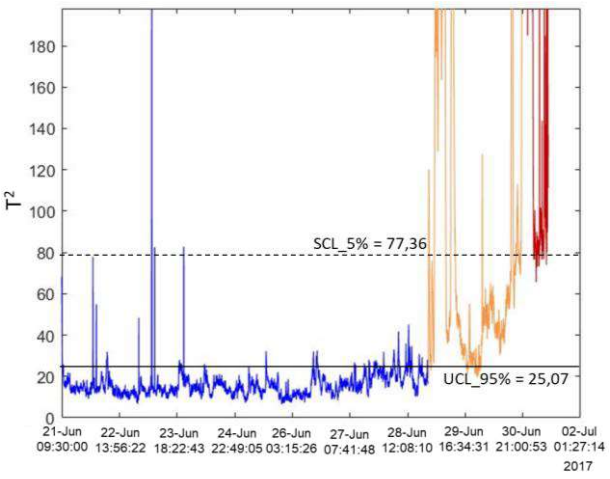

Figure 5: Evolution over time of the moving average of the control variable T^2

It is clear from Figure 5 that the T^2 values for the period marked as surge (in red) are well above the upper control

limit, in particular. the minimum T^2 value in the surge period is 65.61 and the average value is 190.66. The first species error of the model, i.e. the model mistake in reporting surge period, is null. Having chosen an upper control limit of 95%, the error of the second species, i.e. the incorrect reporting of full capacity as surge, is 5%. To obtain also a second species null error, it is necessary to use as a signaling logic of the anomalous operation condition of the compressor, the T^2 statistic above the UCL threshold for a time greater than 2587 seconds. Using this logic, the model can signal an operating anomaly 45,34 hours before the onset of the failure mode under examination.

3.2 Application of the One-class SVM

As in the previous case, the model has been trained on historical data related to the full capacity period, represented in Figure 4 in blue, to train the SVM model, Matlab's "fitcsvm" function with Gaussian kernel and automatic scale parameter optimization was used. Given the high noise of the output control variable of the trained model, it was decided to make a moving average with a time bucket of 300 seconds as in the previous case (Kay, 1993). Figure 6 shows the trend of the moving average of the Likelihood control variable over time with the associated lower control limit (LCL) of 5883 with 95% confidence and upper control limit for the surge period (SCL) of -1.148 with 5% confidence:

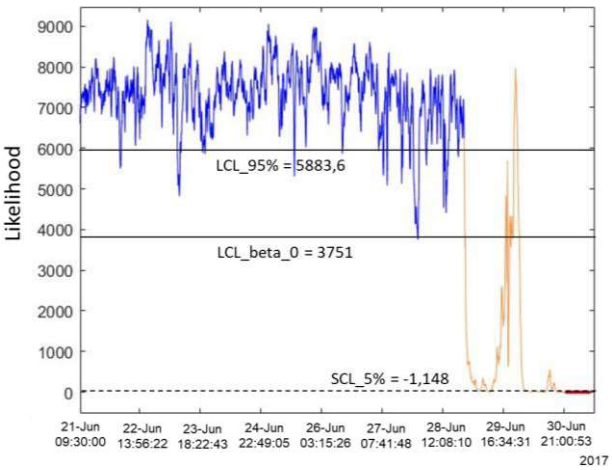

Figure 6: Evolution of the moving average of the Likelihood control variable over time

Also in this case it is evident that the surge data (in red in the figure) are extremely below the LCL control limit; in particular, the maximum value of the control variable in the surge period is -1.06 and has an average value of -1.19. The model, therefore, has a high state detection capability of surge conditions. The first species error on the analysis period is null and have chosen a 95% lower limit, the second species error is 5%. To obtain an error of second species null, it is necessary to use, as a signaling logic of the anomalous operation condition of the compressor, the permanence below the LCL threshold for a time of more than 146732 seconds (40 hours approximately), thus canceling the state detection power of the model. Alternatively, the same result can be achieved by lowering the LCL threshold limit to 3751 (LCL_beta_0). In practice, as shown in Figure 6, LCL_beta_0 is the lower control limit such that the likelihood does not go beyond that limit during the full capacity period (blue in Figure 6). Using this threshold value, the model can signal the compressor surge 45,96 hours before the failure mode occurs.

3.3 Application of the ARIMA model

Once we have verified that both models have a high state detection power for the analyzed case study, we have built a method to evaluate the quality of the compressor RUL estimation by imagining its use in real-time. Figure 7 shows the functioning scheme of the ARIMA algorithm developed and the evaluation of the RUL prediction power of the model:

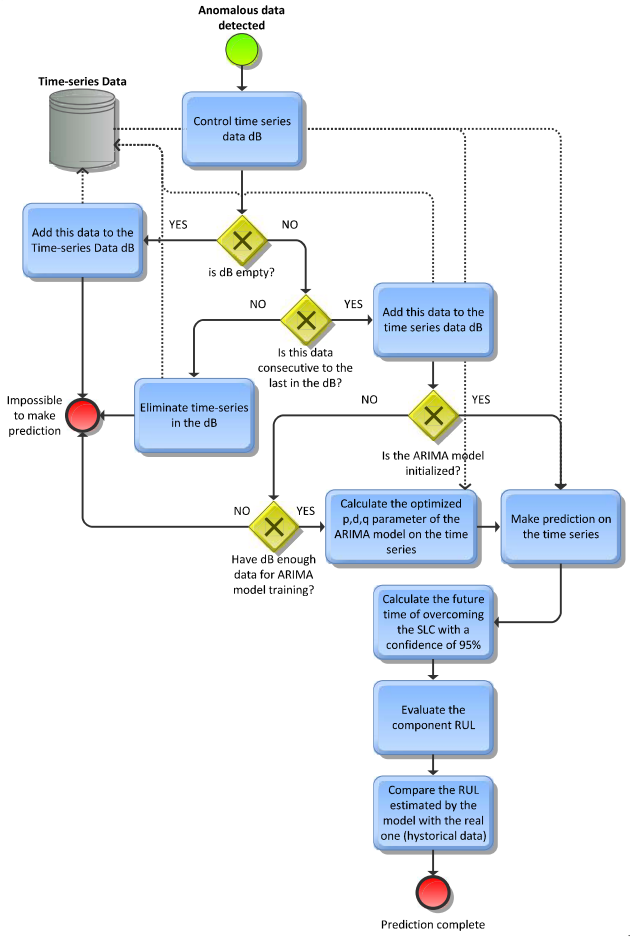

Figure 7: Schema of the ARIMA model in the case study and evaluation of its performances

The ARIMA model was tested on the output data of the two constructed state detection models, i.e. the historical T^2 values for the multivariate control charts and the Likelihood output values of the One-class SVM model, to compare their predictive performance. In the case of the model that uses Hotelling charts, T^2 data starts to be marked as anomalous after they are above the UCL_95% threshold for longer than 2857s. In the case of the one-class state detection model, all data below the LCL_beta_0 threshold are considered abnormal. In both cases, the ARIMA model has been initialized on the first 1000 anomalous values. The model initialization calculates the

parameters that best fit the data through automatic optimization based on AIC minimization. The model output is the expected value and 95% confidence band of the future moment in which the value of the control variable proxy of the compressor performance exceeds the SCL_5% limit threshold. The initialized ARIMA model is first adapted on the time series of which we want the forecast (Time-series Data in Figure 7) using the "estimate" function and then we make the forecast using the "forecast" function on MATLAB software. The strong instability of the T^2 control parameter in the anomaly period makes the ARIMA forecast extremely unstable for all anomalous data until the surge event occurs, making the forecast unreliable. As far as the Loglikelihood output parameter of the One-class SVM model is concerned, this initially shows a very stable downward trend. The forecast with the ARIMA model is also stable and robust from the very first forecasts as shown in Figure 8.

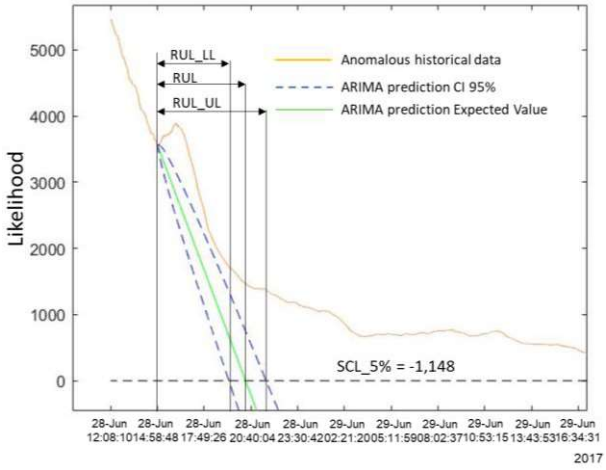

Figure 8: First RUL ARIMA prediction

The ARIMA model estimates an expected RUL of the component of 31,4 minutes and considering a 95% confidence interval on the forecast, a lower limit of 25,5 minutes and an upper limit of 38,8 minutes. Comparing this value with the real RUL calculated between the timestamp of the first RUL prediction with the ARIMA model andthe start of the surge period i.e. the first red data in figure 4, we found that the RUL is underestimated of about 45,68 hours. Subsequently, as shown in Figure 6, the Loglikelihood value becomes very unstable and returns over the LCL_95% control limit, making it impossible to continue forecasting with the ARIMA model.

4 Discussion

The extremely unstable behavior of the compressor performance proxy parameter in both state detection models may be due to the choice of the failure phenomenon under analysis. The surge phenomenon is not a real failure event, but rather a phenomenon that, if prolonged over time, leads to compressor failure, due to the high vibrations it causes. The surge is not caused by a mechanical degradation of the compressor, but by the thermofluid-dynamic conditions of the processed gas and the conditions of the circuit in which it is inserted. Ultimately, depending not only on the performance of the compressor but also on the whole circuit, this phenomenon is very difficult to predict (Gravdahl and Egeland, 2012, 1999). When applying this RUL prediction methodology in the field, the absence of fault condition data means that the SCL_5% control limit cannot be calculated a priori but assumed at some distance from the full capacity population used to train the state detection models. From the single case study analyzed, it appears that for the surge phenomenon, the regime and failure population are extremely distant, especially using the One-Class SVM model. In addition, the same low performance conditions of the monitored machinery may be caused by several failure events. The choice of the correct training control variables for state detection models can partially mitigate this uncertainty. When you have available data related to the spy variables of the failure phenomenon even in bad conditions, you can implement a more robust RUL prediction and also a prognostic model system using a two-class SVM model. This will then be able to recognise the class of full capacity from that of the specific failure mode under analysis without uncertainty. (Navicelli et al., 2019)

5 Conclusions and future developments

Both process control models (multivariate control cards and One-class SVM) applied to the case study show high performance in terms of state detection power of abnormal conditions of the studied compressor. We cannot say the same about the ARIMA model applied to the proxy control variable of the compressor performance output of the two state detection models. In one case, the extremely unstable trend of the compressor performance in the transient between full capacity and surge does not allow to make any RUL forecast using the ARIMA model. In the second case, the forecast is extremely stable in the initial period, but the RUL estimation is strongly underestimated.

The table below provides a summary of the performance of the two RUL prediction models studied.

Table 1: Performance of state detection and RUL prediction **models**

	Hotelling + ARIMA	**One-Class SVM +ARIMA**
First species error	0%	0%
Second species error	0%	0%
Early State Detection	45,34 ore	45,96 ore
RUL estimate	n.d	31,4 minuti
Real RUL - RUL estimate	n.d	45,68 ore

The future steps of this research are the application of the same RUL prediction methodology to different case studies and failure modes to obtain a generalized result.

References

Ahmad, R., Kamaruddin, S., 2012. An overview of time-based and condition-based maintenance in industrial application. Comput. Ind. Eng. 63, 135–149.

Alwan, L.C., Roberts, H.V., 1988. Time-series modeling for statistical process control. J. Bus. Econ. Stat. 6, 87–95.

Bersimis, S., Psarakis, S., Panaretos, J., 2007. Multivariate statistical process control charts: an overview. Qual. Reliab. Eng. Int. 23, 517–543.

Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. John Wiley & Sons.

Chen, P., Pedersen, T., Bak-Jensen, B., Chen, Z., 2009. ARIMA-based time series model of stochastic wind power generation. IEEE Trans. Power Syst. 25, 667–676.

De Carlo, F., Arleo, M.A., Tucci, M., 2014. OEE evaluation of a paced assembly line through different calculation and simulation methods: A case study in the pharmaceutical environment. Int. J. Eng. Bus. Manag. 6, 6–27.

De Carlo, F., Tucci, M., Borgia, O., 2013. Conception of a prototype to validate a maintenance expert system. Int J Eng Technol 5.

Gravdahl, J.T., Egeland, O., 2012. Compressor surge and rotating stall: modeling and control. Springer Science & Business Media.

Gravdahl, J.T., Egeland, O., 1999. Centrifugal compressor surge and speed control. IEEE Trans. Control Syst. Technol. 7, 567–579.

Hashemian, H.M., 2010. State-of-the-art predictive maintenance techniques. IEEE Trans. Instrum. Meas. 60, 226–236.

Ilin, A., Raiko, T., 2010. Practical approaches to principal component analysis in the presence of missing values. J. Mach. Learn. Res. 11, 1957–2000.

Kanawaday, A., Sane, A., 2017. Machine learning for predictive maintenance of industrial machines using IoT sensor data, in: 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, pp. 87–90.

Kay, S.M., 1993. Fundamentals of statistical signal processing. Prentice Hall PTR.

Komonen, K., 2002. A cost model of industrial maintenance for profitability analysis and benchmarking. Int. J. Prod. Econ. 79, 15–31.

Löfsten, H., 1999. Management of industrial maintenance–economic evaluation of maintenance policies. Int. J. Oper. Prod. Manag.

MacGregor, J.F., Kourti, T., 1995. Statistical process control of multivariate processes. Control Eng. Pract. 3, 403–414.

Mason, R.L., Young, J.C., 2002. Multivariate statistical process control with industrial applications. Siam.

Navicelli, A., Vincitorio, M., De Carlo, F., Tucci, M., 2019. Predictive maintenance in industrial plants: real application of Machine Learning models for prognostics. XXIV Summer Sch. "Francesco Turco" – Ind. Syst. Eng. 165–171.

Oakland, J.S., 2007. Statistical process control. Routledge.

Pai, P.-F., Lin, C.-S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33, 497–505.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural Comput. 13, 1443–1471.

Si, X.-S., Wang, W., Hu, C.-H., Zhou, D.-H., 2011. Remaining useful life estimation–a review on the statistical data driven approaches. Eur. J. Oper. Res. 213, 1–14.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.