Warszawa, 19.09.2024 r.

Dr hab. Inż. Szymon Gontarz, prof. PW
Instytut Pojazdów
Wydział Samochodów i Maszyn Roboczych
Politechniki Warszawskiej
Ul. Narbutta 84
02-524 Warszawa

REVIEW OF DOCTORAL DISSERTATION

mgr inż. Hamid SHIRI

„Modelling and analysis of long-term historical data of time-varying complex systems in the presence of impulsive noise for condition monitoring"

Basis for the study: letter RDND08/164/2024, dated 17/07/2024, of the Chairman of the Discipline Council of Environmental, Mining and Energy Engineering, Prof. Dr. Hab. Eng. Robert Król.

1. General characteristics of the dissertation and assessment of the topic and purpose of the work.

The society of the whole world can be successfully called information society. It is precisely 'information' that has become a desired commodity because its proper use can translate into real profits. Additionally, the rapid development of information technologies and other technologies that facilitate obtaining information has caused us to drown in data. This also applies to the area of technology where the main provider of information are sensors, which have become an inseparable element of devices in the human environment or are directly used by them, e.g. in research work. Their capabilities in measuring various physical quantities, possibilities of application in extreme conditions or acquisition with excessive parameters create an unprecedented potential for use in technical diagnostics. However, in order to take advantage of this opportunity, appropriate knowledge is necessary, which the Author developed during his scientific work and which he presented in the discussed dissertation.

The doctoral student set the goal of his work study of knowledge-based data-driven approach for machine health assessment and prognostics, focusing on long-term health index data in the presence of non-Gaussian noise.

The proposed topics and methods are attractive for several reasons. Referring to the introduction and Technological Advancements in Sensor Data Collection: With advancements in sensor technology and Industrial Internet of Things (IIoT), there is an exponential increase in the amount of data collected from industrial machinery.

Effective models like those proposed in the thesis help in making sense of these vast datasets, making it an up-to-date area of study.

The paper concerns the current diagnostic strategy Condition-Based Maintenance (CBM) and Predictive Maintenance (PdM). The industrial shift from reactive to predictive maintenance is critical to reducing operational costs, improving equipment lifespan, and avoiding unplanned downtime. The study's focus on using long-term data for machine health assessment and prognostics aligns with the growing adoption of these strategies in various industries, such as manufacturing, aerospace, mining, and energy (especially wind turbines and aviation as mentioned in the thesis).

Additionally, it refers to complexity of Time-Varying Systems. Modern industrial systems and machinery are highly complex, with varying operational conditions. The study's emphasis on non-stationary and non-Gaussian noise is crucial because real-world systems rarely exhibit ideal conditions (stationarity or normal distribution of noise). The increasing use of data-driven methods to monitor and predict machine health is particularly challenged by such complexities, which the dissertation aims to address through advanced modeling techniques.

In his considerations he assumes the existence of Impulsive Noise and proposes Robust Algorithms: Addressing non-Gaussian, heavy-tailed, and impulsive noise is a contemporary issue in signal processing and data analysis. Robust methods that can handle such noise are essential for accurately predicting machine failure and assessing system health, making this a cutting-edge topic in industrial and academic research.

In conclusion, the dissertation's focus on long-term data modeling, especially in the presence of complex noise, is timely and contributes to the ongoing development of smarter and more efficient predictive maintenance systems. The challenges it addresses are highly relevant to the modern industrial landscape. The doctoral student should therefore be praised for setting an ambitious goal, accompanied by an innovative approach, for conducting interesting analyses of scientific research and their interpretation, but also for largely independently preparing the simulation and analytical work workshop.

Taking into account the current state of knowledge in this area, it should be recognized that the objectives of the dissertation are formulated correctly, the topic of the work is current and important, and the issues raised are of great scientific, cognitive and application importance. The ambitious goal and scope of the doctoral dissertation testify to the broad spectrum of scientific interests and skills of the Doctoral Student.

2. Assessment of the substantive content.

Chapter 1 of the thesis provides an insightful introduction to the topic. It starts by discussing the advancements in technology that have enabled the acquisition, transmission, and storage of large amounts of long-term historical data, which is crucial for condition-based maintenance (CBM). The chapter emphasizes the growing reliance on sensor-rich systems, such as SCADA and CMS, in the industrial sector. The availability of such data is viewed as highly valuable for research into CBM, as it allows for predictive insights into equipment health and failure prevention.

The introduction further outlines the historical evolution of maintenance strategies, transitioning from corrective and preventive maintenance to the more advanced CBM approaches. The shift highlights the industry's move away from reactive models towards data-driven strategies that can predict machinery failure based on condition data, enabling more precise and timely interventions. This evolution is crucial in modern industrial systems that seek to reduce unexpected downtime and optimize maintenance schedules .

The chapter also introduces the OSA-CBM architecture, detailing its seven layers, from data acquisition to prognostics and decision support, making it a comprehensive framework for CBM. These foundational elements are critical to understanding the structure of the dissertation's subsequent chapters, which focus on advanced methods for processing and analyzing long-term data. And here the Author gives the benefits of extracting signal features instead of using the raw signal. This is understandable, but it should be emphasized that there is a risk of omitting valuable diagnostic information. There are few reasons of this result: Simplification of Complex Data, Feature Selection Bias and finally Non-Gaussian and Non-Stationary Signals: In systems where the signal exhibits non-Gaussian noise or non-stationary behavior (as emphasized in this dissertation), relying on traditional feature extraction methods that assume Gaussian or stationary signals might lead to significant information loss. This is especially true when dealing with impulsive noise or heavy-tailed distributions, as these characteristics may not be adequately captured by conventional features like mean or variance. Therefor the key is to ensure that the feature extraction process is robust and tailored to the specific signal characteristics and industrial application before proceed to long term condition monitoring data. Additionally I am not able to find the reason of showing figure 1.3 and 1.4.

In summary, Chapter 1 sets the stage for the dissertation by highlighting the importance of long-term data and outlining the transition to predictive maintenance strategies that are central to the research.

The purpose and scope of the work were described very clearly and factually in chapter 2. It is divided into several key areas, including the formulation of the research problem, objectives, and current research trends in the field. It highlights the challenges associated with machine health assessments, particularly the issues caused by non-stationary behavior and non-Gaussian noise in time-varying complex systems. These factors pose significant difficulties in accurately modeling, analyzing, and predicting machine degradation using conventional methods. Unfortunately, the part of the problem formulation describing non-Gaussian noise introduces some misunderstanding that should be clarified. When the health index is not designed specifically to handle non-Gaussian noise and non-stationary behavior so such behavior will be visible in long-term HI data observation. It will be valid for example if the HI is Inner race energy like in the figure 2.2. But the Author state in 2.7 Limitation of this research that: "health index is the best indicator for describing the degradation process". So, there is some unknown about HI. But when the raw source signal has only Gaussian noise, non-Gaussian noise can still appear in the long-term historical data of the health index due to several factors that arise during the signal processing and data analysis stages. These factors can distort the Gaussian characteristics of the original signal, leading to non-Gaussian noise in the health index. In summary, even when the raw source signal has Gaussian noise, the processes of feature extraction, data aggregation, non-linear transformations, and time-varying system dynamics can introduce non-Gaussian noise into the health index. These factors alter the statistical properties of the noise, leading to heavy-tailed, impulsive, or skewed noise characteristics in the long-term data. The work demonstrates how to handle a health index that already exhibits these challenging characteristics—non-Gaussian noise and non-stationary behavior but not identify the source.

Anyway the chapter establishes the need for novel frameworks capable of handling such complexities in machine health monitoring, particularly in the context of long-term historical data. It articulates the purpose of the research, which is to develop robust models that can deal with non-Gaussian noise and non-linear behavior in the data, thus improving the reliability of condition-based maintenance (CBM) systems. The problem formulation emphasizes how traditional approaches, which assume Gaussian noise, are inadequate for real-world applications where noise is often heavy-tailed and non-Gaussian.

Moreover, the state of the art section covers various degradation modeling approaches, categorizing them into physics-based, data-driven, and hybrid models. It acknowledges the limitations of each, with a particular focus on the shortcomings of models that fail to account for the unpredictability of noise and complex system behavior. By addressing these gaps, the research sets itself apart as an innovative contribution to the field of CBM.

Chapter 3 of the thesis, titled "Experimental Data," describes the various datasets used in the research to validate the proposed models and methodologies. This chapter is essential as it provides real-world context which will validate the theory of dissertation. The datasets discussed include:

FEMTO Dataset: This dataset consists of health monitoring data from bearings, including vibration data collected over time. It is frequently used in CBM research due to its extensive data points and relevance to industrial applications. The dataset is analyzed for degradation patterns, serving as a key test for the proposed noise-handling algorithms.

IMS Dataset: Another bearing dataset, the IMS dataset, includes vibration data from machines operating under different load conditions. It provides a diverse range of data to test the robustness of models in varying operational contexts.

Wind Turbine Dataset: This dataset focuses on the long-term monitoring of wind turbine health. It includes data that capture the non-Gaussian, heavy-tailed noise often found in such systems. Wind turbines operate in unpredictable environments, making this dataset ideal for testing the proposed frameworks.

Chapter 3 also highlights the challenges associated with working with long-term data, such as the presence of noise, outliers, and non-stationary behavior. The synthetic degradation model introduced in this chapter complements the real datasets, allowing the author to control certain parameters for more precise validation of the proposed methods. Nevertheless degradation model is designed to simulate the degradation process describe by HI in three distinct regimes representing different stages of machine degradation. According to this, model is structured with deterministic (trend) component and referring to the topic of dissertation there is also random components. The model's random part is created using an autoregressive process with different distributions Gaussian distribution or an α-stable distribution. I am wonder how the author choose the right distributions, and its parameters values?

Chapter 4 of the thesis, titled "Long-term health index data modeling and identification," focuses on establishing a robust framework for analyzing and modeling long-term health index (HI) data. The chapter introduces a methodology for separating deterministic and random components in complex, non-stationary data, which is particularly relevant for predictive maintenance applications. Traditional approaches often struggle with non-Gaussian, heavy-tailed noise that can be present in real-world data, leading to inefficiencies. The methodology was developed to address the challenges associated with non-Gaussian noise and non-homogeneous characteristics in long-term degradation data. The Author's framework incorporates algorithms for both Gaussian and non-Gaussian data and introduces a novel approach for separating deterministic and random components, standardizing time-varying data, and applying specialized algorithms for non-Gaussian, heavy-tailed noise. This framework is applied to real datasets, such as the FEMTO and wind turbine data, demonstrating the robustness and effectiveness of the proposed methodology. The author explicitly highlights the novelty of this approach, contributing to a field where handling non-Gaussian noise in long-term data has been a persistent challenge. Furthermore, the contributions of this thesis, including this framework, have been published in multiple peer-reviewed journals.

The framework for modeling long-term health index data presented in Chapter 4 was validated through a detailed simulated data analysis. Specifically, the framework was applied to signals modeled under both Gaussian and non-Gaussian scenarios, with the random component modeled as an AR(1) process. The simulation involved generating 100 datasets for four distribution types: one Gaussian and three non-Gaussian (α-stable) distributions with different values of α (1.95, 1.9, and 1.85). Each phase of the methodology, including trend, scale, and autoregressive model identification, was tested using both classical and robust techniques. The simulated data were analyzed across four rows in the results, presenting a comparison of trend and scale identification, AR model parameters, and distribution tail analysis for both Gaussian and non-Gaussian signals. The key metric for validating the approach was the root mean square error (RMSE) for trend and scale identification, alongside boxplots for random components. The robust algorithms were found to be significantly more effective than the classical approaches for handling non-Gaussian heavy-tailed signals, with lower RMSE values and better alignment of estimated parameters with theoretical expectations.

While presented approach is novel and effective, does present some weaknesses, especially when validated against real data. One key limitation is its dependence on prior knowledge of the degradation process and the assumption of an appropriate health index (HI) for predicting Remaining Useful Life (RUL). these dependencies can make the framework less adaptable to new or complex systems where prior knowledge of the degradation process is limited, or where the HI is difficult to define or extract. The thesis does not delve into the methodology of constructing the HI, which could significantly influence the framework's performance. Additionally, the framework shows some constraints when handling datasets with different noise characteristics. For example, in the FEMTO dataset, which exhibited near-Gaussian noise, the impact of the non-Gaussian robust approach was limited. In contrast, the wind turbine dataset displayed stronger non-Gaussian characteristics, revealing some challenges in handling evolving scale and seasonality in the random component. the framework's difficulty in adapting to varying noise characteristics across different datasets limits its generalizability and can result in poor performance when applied to complex, real-world scenarios where noise patterns are not uniform or stationary. Another weakness is the model's assumption that the random component's distribution will always fit within a t location-scale distribution, which may not apply universally across all datasets. However, this assumption might not capture other types of noise patterns that could appear in diverse datasets. For example, some datasets might exhibit different non-Gaussian characteristics, such as skewness, kurtosis, or multimodal distributions, which the t location-scale model may not adequately represent. Therefore, while the t location-scale distribution is robust for certain applications, it might not provide the flexibility required for generalization across all datasets, limiting the framework's adaptability and potentially leading to inaccuracies in diverse industrial environments.

Chapter 5 focuses on segmentation of long-term historical health index data by deterministic approach using various regression models and also involve probabilistic approaches. Those plays a crucial role in dividing the degradation process into identifiable stages. The key challenge addressed here is separating the deterministic components of the degradation (e.g., constant, linear, or exponential trends) from the random noise that could be non-Gaussian in nature. In 5.1 author proposes several regression models (Ordinary Least Squares, Iterative Reweighted Least Squares, Least Absolute Error) and incorporates Student's t-distribution for robust estimation, accounting for extreme values or heavy-tailed distributions that might affect HI data. In chapter 5.2 delves into probabilistic segmentation of long-term health index (HI) data, utilizing probabilistic methods like the Kalman Filter (KF) and the Maximum Correntropy Kalman Filter (MCKF) to handle non-Gaussian noise and time-varying data. These methods provide a more robust framework for predicting the degradation stages and Remaining Useful Life (RUL) by estimating the probability of system states based on observed data. The use of Kalman filters allows the model to adapt over time, making it particularly useful for dynamic systems where noise characteristics and system behavior evolve. In conclusion, as ono could supposed, the deterministic approach worked well with simulated data, especially under Gaussian conditions. However, in datasets with more complex noise, robust methods were essential for accurate segmentation so what was the objective of the Authors showing the results for deterministic approach? Moreover, in which is fully correct, only the probabilistic segmentation results from Chapter 5 are used in Chapter 6 for RUL estimation.

In Chapter 6, the focus shifts to the prediction of Remaining Useful Life (RUL) based on the health index (HI) data, with particular emphasis on handling non-Gaussian noise in the data. The methodology proposed by the Author revolves around a probabilistic approach using a state-space degradation model that dynamically estimates the parameters of the degradation process. The degradation model parameters (such as the coefficients in the exponential model, equation 6.1) are estimated in real time with methods like the Maximum Correntropy Extended Kalman Filter (MCEKF), Extended Kalman Filter (EKF), and Unscented Kalman Filter (UKF). These methods are used to dynamically update the degradation model based on new data and to predict RUL as the system enters the critical degradation phase. One key original achievement presented in this chapter is the development of a robust probabilistic framework that accounts for non-Gaussian noise. This method significantly improves the accuracy of RUL predictions, as demonstrated with various real-world datasets (such as FEMTO and wind turbine datasets) – chapter 6.3. The simulation results (chapter 6.2) reveal that, in the presence of non-Gaussian noise, the MCEKF significantly outperforms both the EKF and UKF, achieving more accurate RUL predictions. The study shows that the MCEKF's robustness against non-Gaussian noise leads to reduced Mean Absolute Error (MAE), especially when data is limited. Early

MCEKF results, appears worse at first compared to the EKF and UKF. Does the change of MCEKF's kernel could improve the situation? In simulation data analysis it seems to be assumed and constant.

MCEKF has been previously known as a robust filtering technique, particularly for systems with non-Gaussian noise. However, what is innovative in this thesis is its application to machine health monitoring and prognosis, particularly within the context of real-time RUL estimation using long-term historical health index data. The methodology for estimating Remaining Useful Life (RUL) is built around the critical stage, using a state-space degradation model that applies the Maximum Correntropy Extended Kalman Filter (MCEKF). The results of Remaining Useful Life (RUL) estimation in Chapter 6 are indeed closely tied to the assumed stage of model described in equations 6.1–6.4. This model summarizes the degradation process with an exponential structure that is continuously updated in real-time using the parameters a, b, and c, which represent the health index's (HI) evolution over time. The filters such as MCEKF, EKF, and UKF estimate these coefficients dynamically to generate RUL predictions. However, while the model is crucial, other factors also play a significant role in determining the accuracy of RUL predictions. One major factor is the accuracy of the First Predicting Time (FPT) detection, which marks when degradation behavior is first observed. It would be nice to hear something more about it during the public defense: how this point is important and how heavily impact the RUL estimation?

Last chapter presents the conclusions drawn from the research and highlights its major contributions. Chapter provides a synthetic overview of the methodologies and their applications, it could be perceived as somewhat general depending on how detailed or reflective the conclusions could be addressing the broader implications of the work, the limitations, and the practical applications of the methodologies. Anyway it seems that Author realize practical perspective on the proposed approach and general direction of research development but "we know our strength when we know our limitation". It would be valuable to clearly define the limitation and doubts. However, the author does acknowledge some limitations throughout the thesis.

3. Evaluation of the editorial side of the dissertation

A detailed assessment of the editorial side of dissertation would focus on several key aspects, such as structure, clarity, consistency, formatting, and overall presentation.

Structure and Organization: the dissertation is well-structured, with clear chapters that logically build upon one another. The flow from the introduction, literature review, methodology, and results sections leads smoothly to the conclusion. The use of consistent headings and subheadings makes it easier for the reader to follow the content. For example, in few chapter one can find clear division on results which refers to simulation and real datasets. Each chapter begins with a good introduction and typically ends with a summary or conclusion. However, some chapters, such as the conclusion in Chapter 7, might benefit from more depth in certain areas like limitations and broader implications.

Clarity of Writing: The technical descriptions are generally clear, with the use of precise terminology, particularly in sections discussing methods like MCEKF, EKF, and UKF. However, complex technical terms and acronyms are introduced (with a few exceptions) and used frequently, which could benefit from additional clarification or glossary inclusion for readers unfamiliar with the subject matter. The dissertation maintains consistency in terminology and referencing key methodologies and datasets, which is critical for such technical works.

Formatting and Presentation: The document appears to follow a consistent formatting style throughout, with clear distinctions between titles, body text, and references. This is essential for academic professionalism. The use of equations is effective in conveying the technical methods, but many figures, in general are just too small to be easy analyze. Additionally, in Chapter 3 in the description of the Wind turbine dataset we will not find data described in metric units and there is no reference in the text to Figure 3.6.

Grammar and Style: The grammar is generally good, with few errors, but there are areas where sentence structure could be simplified for easier comprehension, particularly in complex technical explanations. The tone is appropriately formal, which is essential for a dissertation. The language is scientific and objective, suitable for the intended academic audience.

References and Citations: The dissertation makes consistent use of rich in-text citations, which is essential for referencing prior work and providing credibility to the research. However, it is hard to point out what where the state of art based conclusions which influenced the dissertation.

Summarizing, the dissertation is well-organized and technically sound, with strong editorial presentation overall. Some minor improvements in clarity and depth of conclusions would further enhance its readability and academic impact. The document maintains a high standard of academic professionalism.

4. Final conclusions

To sum up the entire presented dissertation, it is necessary to emphasize the importance of the problem posed in the work and the methods used to analyze it and the proposed innovative solutions. I also state that:

- The scientific problem that the PhD student undertook to solve was selected and formulated correctly.

- The main objective of the work has been achieved. The implementation of all the specific objectives of the doctoral thesis constitutes an original solution to the scientific problem in the scope of the scientific discipline under consideration: Environmental, Mining and Energy Engineering.
- To solve the defined problem, the PhD student skillfully used contemporary scientific and technical achievements in the field of mechanics and machine construction, operation, statistics as well as numerical methods.
- While working on the thesis, the PhD student demonstrated independence in conducting research as well as the ability to work in a research team.
- The results of the dissertation extend the knowledge on the possibilities of modeling and analysis of long-term historical data of time-varying complex systems in the presence of impulsive noise for condition monitoring.

The reviewed work has therefore original authorial features of novelty as well as significant utilitarian values. However, the comments presented in the review do not diminish the fundamental substantive value of the work, because they primarily concern the imperfections of the description of the methodology and the presentation of the results. The author's achievements in the dissertation expand existing knowledge in several key areas related to condition-based maintenance (CBM), machine health monitoring, and Remaining Useful Life (RUL) prediction, particularly under non-Gaussian noise conditions. Here are the primary areas of expansion:

- Robustness in Non-Gaussian Noise Handling: The author contributes significantly by expanding the application of the Maximum Correntropy Extended Kalman Filter (MCEKF) to RUL prediction, which addresses a notable gap in the handling of non-Gaussian noise in machinery data. Prior to this work, traditional filters such as the Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF) were commonly applied, but they relied heavily on the assumption of Gaussian noise. The MCEKF introduces robustness against outliers and heavy-tailed noise distributions, enhancing the accuracy of predictions in real-world scenarios where data is less predictable.
- Improved RUL Prediction Accuracy: The author's work builds upon existing machine health monitoring techniques by improving RUL prediction through more accurate modeling of degradation processes. Specifically, by introducing methodologies that can better capture complex, non-linear degradation behaviors and by refining how First Predicting Time (FPT) is detected, the research expands on conventional data-driven approaches and provides more reliable outcomes in diverse industrial settings.
- Segmentation of Long-term Health Index (HI) Data: The author also expands existing knowledge in the segmentation of long-term HI data, offering new methodologies for separating different degradation stages in machinery. This aspect, which combines historical and real-time data, addresses challenges in

analyzing non-stationary time series data and enhances the ability to predict system failures at an earlier stage.

- Hybrid Models for Prognostics: While traditional models either focused on physics-based or data-driven methods, the dissertation bridges these by offering hybrid approaches that integrate deterministic degradation models with probabilistic methods like MCEKF. This combination improves the precision of RUL estimates and offers a new perspective in the prognostics and health management (PHM) field.
- These achievements not only improve existing models but also push forward the practical implementation of RUL predictions in industries like mining, aviation, and wind energy, where non-Gaussian noise and complex operating conditions present major challenges.

I rate these work achievements very highly and believe that the work meets the requirements provided for in Article 187 paragraphs 1 and 2 of the Act of 20 July 2018.: Prawo o szkolnictwie wyższym i nauce (Dz. U. z 2023 r., poz. 742). This allows me to formulate a motion to allow the Author of the work to publicly defend it. Additionally, due to the broad approach to the issue, the brilliant solution to a non-trivial problem and the significant application potential of the technical solutions developed as part of the doctoral dissertation, I request a possible distinction of the doctoral dissertation by the Council of the Scientific Discipline of Environmental, Mining and Energy Engineering. In conclusion, it should be emphasized that the research were supported by European Commission: Marie Sklodowska Curie programme through the ETN MOIRA project (GA 955681), which indicates the broad impact of the work. Finally I agree with the thought quoted at the beginning: "Prediction is very difficult, especially if it's about the future", but MSc. Hamid Shiri showed that these difficulties can be overcome.

Z poważaniem,

dr hab. inż. Szymon Gontarz, prof. ucz.