2026 International Nuclear Atlantic Conference – INAC 2026
Escola de Guerra Naval, Rio de Janeiro, RJ, August 24 - 28, 2026

ISBN: 978-65-272-2398-6

Statistical Analysis of Reactor Variables for Sensor and System Diagnostics through Deep Generative and Recurring Neural Modelling

F. E. Wu[1] and T. N. Conti[1]

[1] fred.wu@alumni.usp.br, [1] tnconti@ipen.br,
Instituto de Pesquisas Energéticas e Nucleares (IPEN / CNEN - SP) Av. Professor Lineu Prestes 2242 05508-000 São Paulo, SP, Brazil

1. Introduction

The use of Artificial Intelligence (AI) models in nuclear reactors offers promising prospects. Among them, systems trained for early diagnostics of accident conditions [1, 2], analytical redundancy for sensors [3, 4], and core monitoring [5]. Another venue of interest is the use of Deep Generative models to produce synthetic data to be used for training of such systems.

This paper presents a preliminary statistical analysis of the variables monitored at the experimental nuclear reactor Institute of Atomic Energy-Reactor One (IEA-R1), of the Nuclear and Energy Institute (IPEN). The goal was to map out correlations and reduce the number of variables to an optimal set that can account for as much of the information in the system as possible.

The data used derives from the reactor's Data Acquisition System (SAD). The system records simultaneous readings and calculated values, at a rate of two per minute, of 69 variables [6], comprising power level, rod positions, water flow and conductivity, radiation and temperature values.

2. Methodology

The general process consisted of applying different metrics to find correlations and remove statistically redundant variables. The main metric used was linear correlation, through Pearson's *r* [7]. Principal Component Analysis (PCA) [8] was performed for both the original and reduced set of variables. The analysis was performed as follows, starting by pre-processing the data:

- Selection of data of an operating cycle without abnormal occurrences or sensor faults;
- Delimitation of an appropriate interval of the operating cycle for the analysis;
- Arrange the data in a Measurement x Variable matrix;
- Calculation of Median Absolute Deviation (MAD) [9] and Coefficient of Variation (CV) [10];
- Removal of variables with MAD or CV near zero;
- Normalization through z-score [11];
- Removal of rows of data containing outliers (z-score higher than 6);
- Renormalization through z-score without outliers.

The removal of variables with low MAD or CV values was necessary because such small variance hinders the analysis. The data also contained a small number of spurious measurements. The small amount meant they could be removed without jeopardizing the analysis. Procedures continued with the steps listed below:

- Calculation of Pearson's *r* for all pairs of variables, building a square Correlation Matrix;

- Estimation of Signal-to-Noise Ratio (SNR);
- List subsets of variables with high correlation (higher than 0.9) using the Correlation Matrix;
- Removal of all variables of each subset, with exception of the one with highest SNR or standard deviation;
- Implementation of PCA for both sets of variables.

All data handling was performed using the free and open-source Scilab software.

3. Results and Discussion

Of the 69 variables recorded by SAD, 57 are sensor measurements, while the others are calculated values derived from the sensors themselves. Four temperature sensors had unreliably high values (e.g. over 240 °C). The reactor's regime in the chosen cycle was of continuous operation for three days, at 4.5 MW. The analysis comprised the interval starting when the reactor reached the operating power level up until start of shutdown. Furthermore, the reactor has two sets of heat exchangers and cooling towers, but only one or the other is turned on in each cycle, so waterflow measured in both sets were added up for simplicity. The resulting waterflow is numerically equal to the readings of another sensor, unless a leak occurs. That excludes one of the two sensors for the purpose of this analysis.

MAD and CV screening revealed two sensors with very low values, leaving 49 variables in total for the analysis. Outlier removal eliminated less than 2% of the total number of measurements, with a total of 6739 rows remaining in the data matrix.

Calculation of Pearson's *r* revealed that several temperature and radiation variables contained high values of correlation, which is expected as they are all affected by the same physical phenomena. Three of the four rod position sensors had high correlation with each other. This comes from the operating procedures adopted, where three rods move in steps further higher as the operation goes on, while one rod moves continuously. The indication that only two rod positions are necessary, however, is deceptive. While most of the movements follow the same direction, the few movements of one rod in the opposite direction are important from a physical standpoint, even if they carry small significance in terms of linear correlation.

The removal of temperature and radiation variables decreased the set to 28 variables. The Correlation Circles and Eigenvalues Graphs of the two PCAs performed are presented In Figures 1 and 2, below.

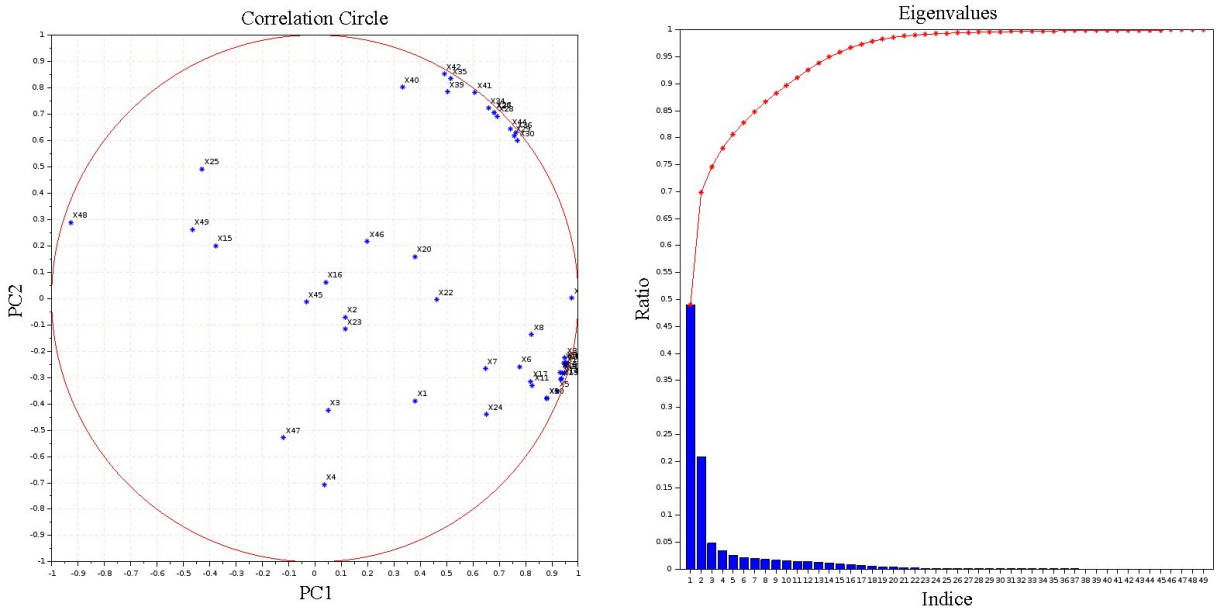

Figure 1: Correlation Circle and Eigenvalues Graph of the pre-Pearson PCA, performed over the 49 variable set.

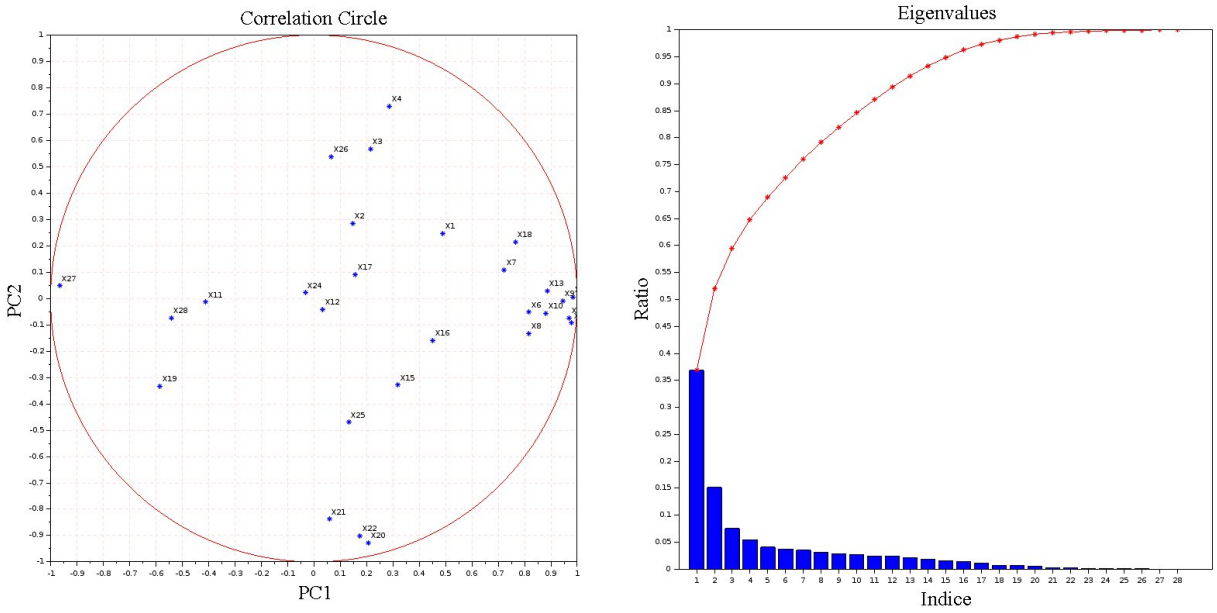

Figure 2: Correlation Circle and Eigenvalues Graph of the post-Pearson PCA, performed over the 28 variable set.

At first glance, Figures 1 and 2 seem to indicate the pre-Pearson PCA achieved better results, with the first two eigenvectors, or Principal Components (PC) accounting for much more of the variance. Furthermore, to surpass 90% of the variance, 11 PCs are required per the pre-Pearson PCA, contrasted with 13 for the post-Pearson PCA. On the other hand, the pre-Pearson Correlation Circle of Figure 1 shows clusters of variables, indicating close relation, whereas the variables in the post-Pearson Correlation Circle of Figure 2 are more spread out. Comparing the first five PCs of each PCA further highlights the differences, listed in Table 1, below.

Table 1: Main features of the first five PCs of pre-Pearson and post-Pearson PCAs.

Principal Component	Pre-Pearson PCA		Post-Pearson PCA	
	Ratio	Dominant Variable Types	Ratio	Dominant Variable Types
PC1	0.49	Temperature, Radiation	0.37	Power Level, Temperature, Radiation
PC2	0.21	Temperature	0.15	Power Level, Conductivity
PC3	0.05	Power Level, Radiation	0.08	Reactor Period, Rod Position, Radiation
PC4	0.03	Power Level, Radiation	0.05	Conductivity, Radiation, Waterflow
PC5	0.03	Waterflow	0.04	Waterflow

The dominant variables in each pre-Pearson PC listed in Table 1 make it clear temperature and radiation variables account for much of the set's variance, obfuscating other physical phenomena. Meanwhile, the dominant variables in the post-Pearson PCA offer much more interpretable and meaningful results. Each PC can be distinguished as a different subset of phenomena.

4. Conclusions

The statistical analysis through Pearson's *r* and PCA successfully mapped out correlations and produced a set of meaningful variables that accounts for much of the reactor's variance and information. A different approach through pure PCA corroborates the results. Further statistical calculations, such as Mutual Information, and the use of different data intervals and cycles can help refine the analysis. The findings form a baseline reference for optimal settings and expected results when designing deep generative and recurring neural models for sensor simulation, reactor state diagnostics and synthetic data generation.

Acknowledgements

The authors would like to thank the IEA-R1 operators for providing data and support for the study, and the Postgraduate Department of the Nuclear and Energy Institute (IPEN) for their assistance to our participation in this event.

References

[1] GOMES, Carla Regina; MEDEIROS, Jose Antonio Carlos Canedo. Neural network of Gaussian radial basis functions applied to the problem of identification of nuclear accidents in a PWR nuclear power plant. **Annals of Nuclear Energy**, v. 77, , p. 285-293, 2015.

[2] LEE, Gyumin; LEE, Seung Jun; LEE, Changyong. A convolutional neural network model for abnormality diagnosis in a nuclear power plant. **Applied Soft Computing**, v. 99, , p. 106874, 2021.

[3] MESSAI, Adnane; MELLIT, Adel; ABDELLANI, Idir; PAVAN, Alessandro M. On-line fault detection of a fuel rod temperature measurement sensor in a nuclear reactor core using ANNs. **Progress in Nuclear Energy**, v. 79, , p. 8-21, 2015.

[4] MANDAL, Shyamapada et. al. Sensor fault detection in Nuclear Power Plant using statistical methods. **Nuclear Engineering and Design**, v. 324, , p. 103-110, 2017.

[5] SAEED, Asim; RASHID, Atif. Development of core monitoring system for a nuclear power plant using artificial neural network technique. **Annals of Nuclear Energy**, v. 144, , p. 107513, 2020.

[6] RICCI FILHO, Walter; SURKOV, Vadim. **Sistema de Aquisição de Dados do reator IEA-R1 - SAD - Lista de Variáveis**. 2007. Not published.

[7] VOTHIHONG, Phuong et al. **Python: End-to-end Data Analysis**. Birmingham [UK]: Packt Publishing Ltd, 2017.

[8] GEWERS, Felipe L. et al. Principal component analysis: A natural approach to data exploration. **ACM Computing Surveys (CSUR)**, v. 54, n. 4, p. 1-34, 2021. Available at: https://dl.acm.org/doi/abs/10.1145/3447755. Date of access: 2026-04-08.

[9] JONES, Alan R. **Probability, Statistics and Other Frightening Stuff**. [S.l.]: Routledge, 2018. p. 73. ISBN 978-1-351-66138-6.

[10] EVERITT, Brian. **The Cambridge Dictionary of Statistics**. [S.l.]: Cambridge University Press, 1998. ISBN 978-0-521-59346-5.

[11] KREYSZIG, E. Advanced Engineering Mathematics. 4. ed. [S.l.]: Wiley, 1979. p. 880 eq. 5. ISBN 0-471-02140-7.