Marine Water Quality Assessment using Support Vector Regression and Neutrosophic control charts

Ishah Maria Mathew[1], O.S. Deepa[2]

[1,2] Department of Mathematics, Amrita School of Physical Sciences, Coimbatore, Amrita Vishwa Vidyapeetham, India

Abstract

In recent years, with the growth of industries in coastal cities, the wastewater including organic and inorganic substances were discharged into seawater, triggering seawater pollution which leads to many social problems. Hence, it is essential to examine the marine environment by considering the key factors like oxygen-related parameters, including Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Dissolved Oxygen (DO). The present study suggests a hybrid statistical process monitoring framework by combining the V-Exponentially Weighted Moving Average (VEWMA) control chart with principal component analysis (PCA) . PCA is considered in this study to extract the dominating linear features from the multivariate marine water quality data. The PCA scores are integrated into NVEWMA monitoring statistics to strengthen the identification of process changes. To evaluate the effectiveness of the proposed method, a real-life marine water quality data is used. The results show that the integration of Machine Learning methods with NVEWMA control chart provides a useful method for keeping an eye on environmental processes and to identify possible shifts.

Keywords: Neutrosophic VEWMA, Machine Learning, PCA, Random Forest, Gradient boosting, SVR

Introduction

A popular Statistical Process Framework (SPC) framework introduced by [9], the EWMA control chart which is known for its ability to identify slight changes in the process parameters. Compared to the traditional control charts, EWMA provides improved sensitivity by giving more weight to current observations while keeping the past observations. The statistical properties and performance of EWMA charts were studied by [12], providing guidance for parameter selection. An adaptive EWMA chart proposed by [1] where the parameters are automatically adjusted to improve detection performance for various mean shifts, thereby increasing flexibility under varying process conditions.

Extended beyond traditional methods, several studies have focused on developing control chart for non-normal process A V-type control chart introduced by [8] follows Maxwell distribution, which is designed for monitoring processes, further a Maxwell-CUSUM control chart by [7] was proposed to effectively detect changes in failure rates in manufacturing systems. These approaches often rely on accurate parameter estimation; in this regard, [11] investigated minimax estimation methods for the Maxwell distribution under different loss functions. In addition to all the prior works a statistical modeling based on the distribution of sum of independent gamma random variable was proposed by [13] which has broad application in process monitoring context.

Due to dimensionality and complexity of present-day industrial situations, Machine Learning approaches have gained popularity in SPC framework. These techniques provide enhanced capacity to handle complicated structures and non-linear interactions. To monitor non-linear process [10] proposed model using SVR. Similarly to demonstrate robustness and predictive ability to identify irregular behavior [6,16] have used ensemble learning techniques like Random Forest and Gradient Boosting. To enhance serially correlated data, [15] proposed a sequential learning framework. Exponentiated exponential distribution was also explained with average sample number [14].

Recent research has increasingly focused on extending SPC techniques to account for uncertainty through neutrosophic statistics. In the present-day indeterminacy [2] have proposed the neutrosophic geometric distribution to model data. To enhance monitoring performance [3] has developed neutrosophic EWMA control chart, similarly, to handle imprecise and uncertain data [4,5] has proposed attribute control chart. Together the studies illustrate a growing potential on neutrosophic approach for improving process monitoring while dealing with uncertain conditions. Mathew Ishah Maria and Deepa[12] explained Neutrosophic EWMA and DEWMA control chart on Exponential and Transformed Exponential Distributions [12]

These studies indicate how crucial it is to combine machine learning methods with traditional SPC tools to improve monitoring efficiency and to evaluate efficient process variation identification in complex scenarios.

These advancements demonstrate how crucial it is to combine machine learning methods with conventional SPC tools in order to improve monitoring efficiency and guarantee efficient process variation identification in complicated contexts.

1.1 Research Gap

Since there are significant developments in Statistical Process Control(SPC) most of the existing control chart such as EWMA and VEWMA control chart are designed for precise data do not address uncertainty and indeterminacy in present real-world observations.

1. Traditional VEWMA control chart cannot handle neutrosophic or interval valued data
2. Instead of monitoring process variability, majority of neutrosophic control charts concentrate on tracking process location
3. Limited research exist on integrating VEWMA control chart with neutrosophic concept.
4. Neutrosophic VEWMA control chart application to real world dataset remains scarce

1.2 Motivation

Traditional control chart provides less reliable monitoring performance in case of environmental and industrial quality assessments involving uncertain and imprecise data.

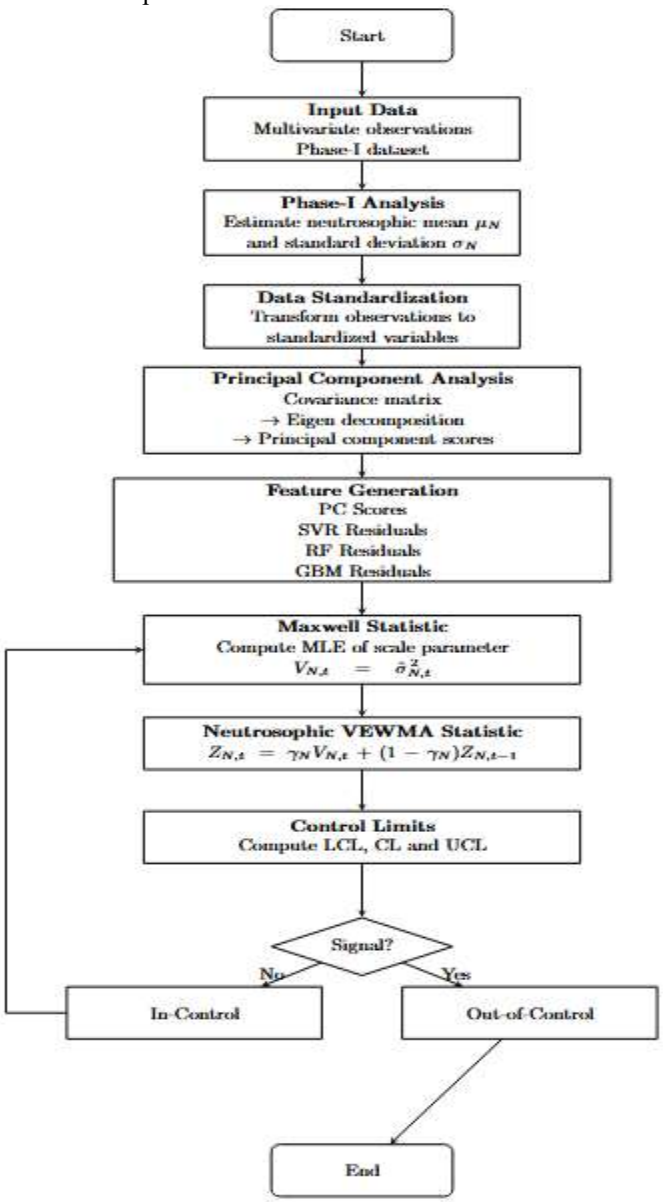


Fig.1: Framework of the proposed PCA based Neutrosophic VEWMA control chart

To address this limitation this study proposes a neutrosophic VEWMA (NVEWMA) control chart that combines neutrosophic framework with VEWMA control chart. The proposed NVEWMA control chart aims to:

1. In cooperating uncertainty and indeterminacy in process observations
2. To monitor changes in process variability
3. Provide a reliable tool for monitoring real world applications involving uncertain data

Figure 1 shows the overall monitoring framework used in this study. It shows the data pre-processing, feature extraction and control chart implementation stages in a systematic way.

2 Literature Review

2.1 Neutrosophic Random Variable

According to Aslam [2], suppose that X_L be a random variable having mean μ and variance σ^2 based on this, we define a neutrosophic random variable $X_N = X_L + X_L I_N; I_N \in [I_L, I_U]$ where $X_L I_N$ be the indeterminate part and $I_N \in [I_L, I_U]$ be the indeterminacy. When $X_L = 0$, the proposed neutrosophic random variable reverts to the classical random variable. Neutrosophic logic is the generalization of fuzzy logic where an indeterminacy component is added additionally. Note that $I_{N^2} = I_N, \ldots, I_{N^n} = I_N, 0.I_N = 0; n \in N$

2.2 Machine Learning Techniques

2.2.1 Principal Component Analysis (PCA):

PCA is a linear transformation method that keeps much of the variability in multivariate data while lowering its dimensionality. Considering the data matrix X with a mean of zero, PCA is obtained by breaking down the covariance matrix $\Sigma = \frac{1}{n} X'X$ into its eigenvalues. The eigenvectors that come out of this process are the principal components, and the eigenvalues that go with them show how much variance each component explains. The leading principal components give a representation in a lower dimensional subspace:

$$Z = XW_k$$

Where W_k has the eigenvectors that go with the k biggest eigenvalues. PCA is often used to cut down on noise, make things easier to see, and as a first step in statistical process control, where it makes it easier to keep an eye on things by projecting correlated variables into uncorrelated directions of greatest variance.

Other machine learning techniques used are:

1. Support Vector Regression (SVR)
2. Boosting
3. Random Forest Regression

3 PROPOSED CONTROL CHART

3.1 Maxwell distribution:

Consider R as a continuous random variable which follows Maxwell distribution with scale parameter σ. The probability density function can be written as:

$$f(r, \sigma) = \sqrt{\frac{2}{\pi}} \frac{r^2}{\sigma^3} e^{-\frac{r^2}{2\sigma^2}}, \quad r > 0, \ \sigma > 0$$

The MLE of σ is defined as $\hat{\sigma} = \sqrt{\frac{1}{3n}\sum_{i=1}^{n} r_i^2}$ (8). Here $V = \hat{\sigma}^2$ is the square of the estimate following gamma distribution with parameters $\frac{3n}{2}$ and $\frac{2\sigma^2}{3n}$. Here σ^2 is the mean and $\frac{2\sigma^4}{3n}$ is the variance (**7**). Considering n as the sample size.

3.2 The proposed Neutrosophic VEWMA control chart:

The null and alternative hypotheses for monitoring the scale parameter σ^2 are given by H_0:The process is in control, i.e., $\sigma_N^2 = \delta\sigma_{N,0}^2$ with $\delta = 1$, and H_1:The process is out of control, i.e., $\sigma_N^2 = \delta\sigma_{N,0}^2$ with $\delta \neq 1$. The purpose of this study is to use an EWMA chart to track a Maxwell process to identify minute process fluctuations. We can use an EWMA chart to track V as it is a significant statistic for the Maxwell distribution (**11**). This charting scheme's plotting statistic is intended to be:

$$Z_{N,i} = \gamma_N V_{N,i} + (1 - \gamma_N) Z_{N,i-1},$$

where $Z_{N,i-1}$ contains past information about the scale parameter and $V_{N,i}$ represents the current value of the maximum likelihood estimator of σ_N^2, (for $i = 1,2, \ldots$). Eq. (1) can also be expressed as

$$Z_{N,i} = \sum_{j=0}^{i-1} \gamma_N (1 - \gamma_N)^j V_{N,i-j} + (1 - \gamma_N)^i Z_{N,0}.$$

In Eq. (2), the weights $\gamma_N(1 - \gamma_N)^j$ drop exponentially with the age of the sample observations. The amount of shift (δ) to be detected and the in-control average run length decide which γ_N should be used for process monitoring (**12**). To determine the ideal value of γ_N, some researchers proposed predicted weighted run-length and iterative least squares methods (**1**), (**9**). As a result, one can set γ_N to any appropriate value, making it larger for large shifts and smaller for little shifts.

$Z_{N,0}$, the starting value for the historical data, is selected to be equal to $\sigma_{N,0}^2$. The average preliminary data can be used to determine the target scale parameter $\sigma_{N,0}^2$ if no information is available. The mean and variance of the VEWMA statistic are σ_N^2 and $\frac{2\sigma_N^4}{3n}$ respectively, since V_N has a gamma distribution with these values. Therefore, the mean and

variance of VEWMA statistic are $E(Z_{N,i}) = \sigma_N^2(1 + I_N)$ and $Var(Z_{N,i}) = \frac{2\sigma_N^4}{3n}\left\{\frac{\gamma_N}{2-\gamma_N}(1-(1-\gamma_N)^{2i})\right\}(1 + I_N)^2$.

TABLE 1: L_N factors to estimate $W_{N,i}$ coefficients.

Parameters		False Alarm rate (α)		
Smoothing constant (γ_N)	Sample size (n)	0.005	0.0027	0.002
[0.25, 0.255]	2	[3.070, 3.170]	[3.395, 3.495]	[3.556, 3.656]
	5	[2.893, 2.993]	[3.161, 3.261]	[3.288, 3.388]
	9	[2.830, 2.930]	[3.070, 3.170]	[3.181, 3.281]
[0.5, 0.55]	2	[3.341, 3.441]	[3.713, 3.813]	[3.900, 4.000]
	5	[3.043, 3.143]	[3.358, 3.458]	[3.510, 3.610]
	9	[2.910, 3.010]	[3.195, 3.295]	[3.328, 3.428]
[0.75, 0.75]	2	[3.500, 3.600]	[3.930, 4.030]	[4.126, 4.226]
	5	[3.160, 3.260]	[3.485, 3.585]	[3.652, 3.752]
	9	[3.001, 3.101]	[3.305, 3.405]	[3.447, 3.547]

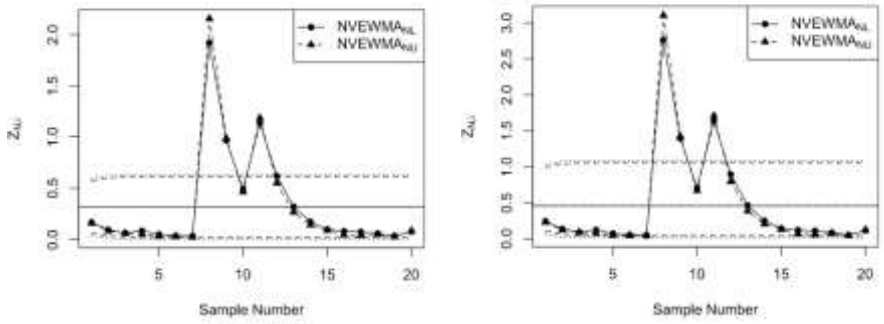
Therefore, the control limits of a VEWMA control chart are given by,

$$\text{LCL} = W_{N,1}\sigma_N^2(1 + I_N),$$
$$\text{CL} = \sigma_N^2(1 + I_N)$$
$$\text{UCL} = W_{N,2}\sigma_N^2(1 + I_N).$$

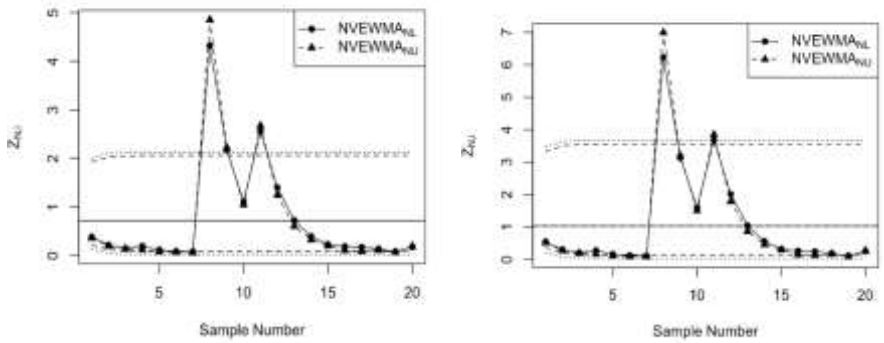
Here, CL represents the central line and $W_{N,1} = \left[1 - L_N\sqrt{\frac{2}{3n} \times \frac{\gamma_N}{2-\gamma_N}(1-(1-\gamma_N)^{2i})}\right]$, $W_{N,2} = \left[1 + L\sqrt{\frac{2}{3n} \times \frac{\gamma_N}{2-\gamma_N}(1-(1-\gamma_N)^{2i})}\right]$. Since $V_{N,i}$ is a gamma random variable and $Z_{N,i}$ is the linear combination of $V_{N,i}$, the distribution of $Z_{N,i}$ is a gamma-series distribution, making it simple to determine the $W_{N,i}$ coefficients for a given L_N (**13**). Consequently, if we achieve the desired fixed false alarm rate (α) for a given sample size, the factor L_N is derived from the quantile of the gamma-series distribution. Since there isn't a closed form equation for the quantile of the gamma-series distribution, we use simulation to get the values, which are listed in Table 1. The quantile of the gamma-series distribution is used to estimate these values. For illustrative purposes, the sample size, false alarm rate, and smoothing constant values were used.

4 REAL LIFE APPLICATION

The application of the proposed control charts is made on real time data. The water quality data of coastal area of 2022 is collected from Odisha (Marine), India (https://cpcb.nic.in/nwmp-data-2022/). The data were recorded for three different measurements such as (Dissolved O2(mg/L), pH and BOD(mg/L) and the average values are considered. Measurements are recorded in the interval having the minimum and maximum values as that of a neutrosophic series. The dataset consists of 20 observations each of sample size $n_N = 4$. The mean and standard deviation estimated from the dataset are $\mu_{N,0} = [5.7,7.0]$ and $\sigma_N = [4.140164797,5.897148707]$.

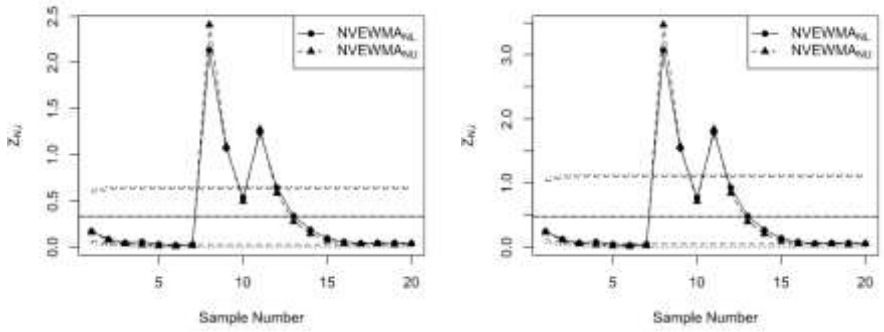


(a) I_N=0 (b) I_N=0.2

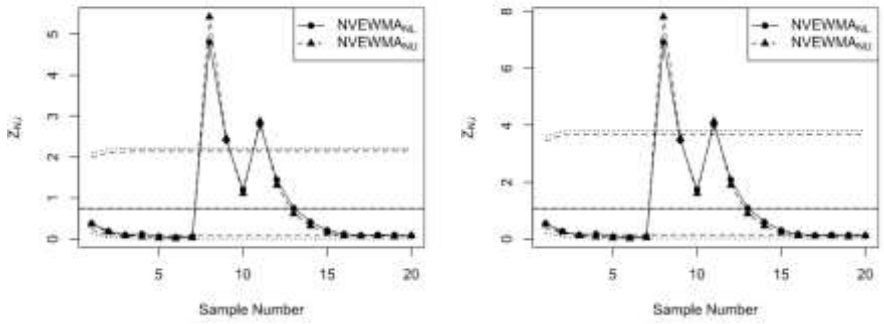


(a) $I_N e$=0.5 (b) I_N=0.8

Fig 1. PCA Neutrosophic VEWMA control charts for various Indeterminacy values for monitoring water quality data of coastal area

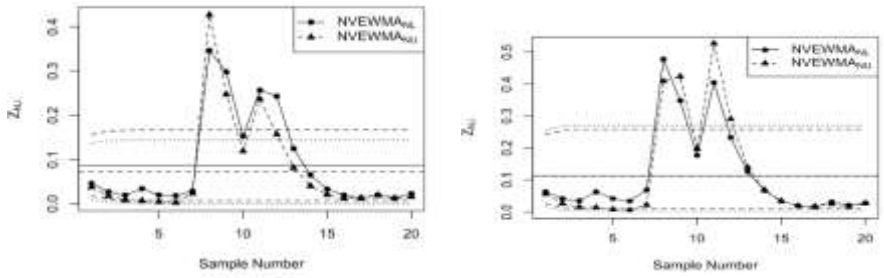


(a) I_N=0 (b) I_N=0.2

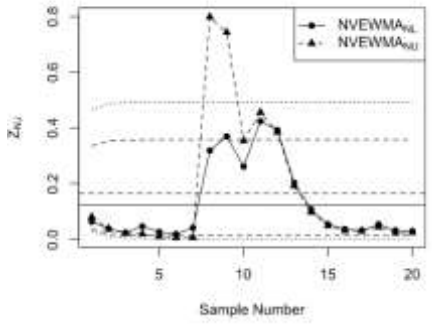


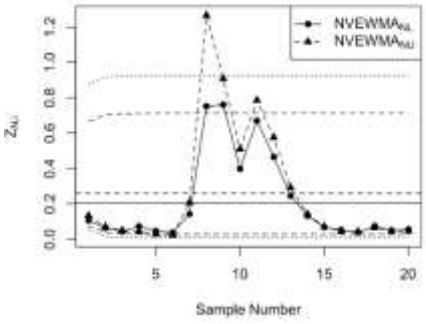
(b) I_N=0.5 (b) I_N=0.8

Fig 2. PCA-SVR Neutrosophic VEWMA control charts for various Indeterminacy values for monitoring water quality data of coastal area



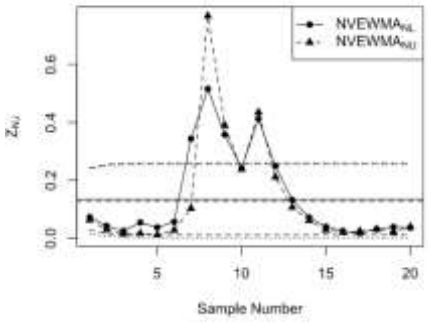
(a) I_N=0 (b) I_N=0.2

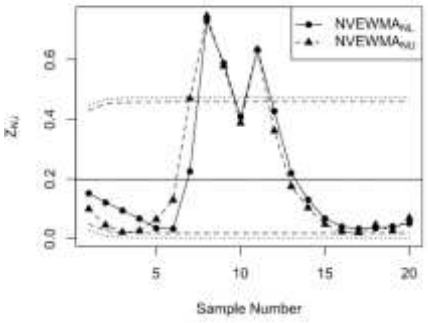




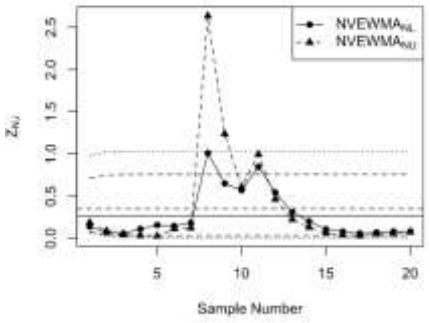
(c) I_N=0.5 (b) I_N=0.8

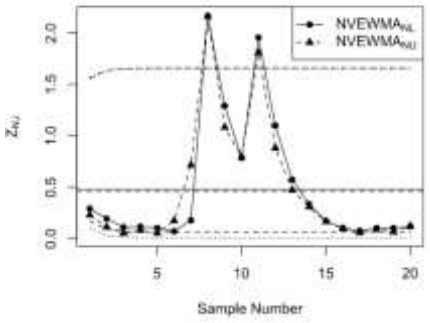
Fig 3. PCA-RF Neutrosophic VEWMA control charts for various Indeterminacy values for monitoring water quality data of coastal area





(a) I_N=0 (b) I_N=0.2





(d) I_N=0.5 (b) I_N=0.8

Fig 4. PCA-GB Neutrosophic VEWMA control charts for various Indeterminacy values for monitoring water quality data of coastal area

To efficiently manage correlation between the process variables, principal component analysis (PCA) was included into the VEWMA control chart. Initially, in Phase-I (First 20 datapoints) mean and standard deviation of the original variables were used to standardize them. The normalized data was then subjected to PCA to get orthogonal principal components via the covariance matrix's eigen decomposition. The main component score vector at time t was calculated as $\boldsymbol{P}_{N,t} = \boldsymbol{Z}_{\boldsymbol{N}_t}\boldsymbol{V}_{\boldsymbol{N}}$, where $\boldsymbol{V}_{\boldsymbol{N}}$ is the matrix of eigenvectors and $\boldsymbol{Z}_{N,t}$ is the standardized observation vector. $V_{N,t} = \frac{1}{3n}\sum_{j=1}^{n} P_{N,tj}^2$, was used as the input to the VEWMA statistics to describe the variability of the PC scores. After that, the VEWMA statistics were updated recursively using the formula $Z_{N,t} = \gamma_N V_{N,t} + (1-\gamma_N)Z_{N,t-1}$,where the smoothing parameter is denoted by γ_N. This integration increases the PCA-VEWMA chart's sensitivity to minor changes while allowing it to track the multivariate process's dominating variation structure. Additionally, by modeling the PC scores and utilizing the generated residual components as inputs to the VEWMA statistic, machine learning

models including Support Vector Regression (SVR), Random Forest (RF), and Gradient Boosting Machines (GBM) were integrated.

The Neutrosophic PCA–VEWMA control charts for different indeterminacy levels $I_N \in 0,0.2,0.5,0.8$ are presented in Figures 2–5. These figures provide a clear illustration of how the proposed monitoring framework performs under increasing levels of uncertainty, along with its hybrid extensions that incorporate machine learning techniques such as SVR, RF, and GBM.

Figure 2, refers to the baseline PCA-NVEWMA model. The control chart is able to successfully capture major deviation in the multivariate marine water quality process. Samples points from 7-10 show a significant change with both the neutrosophic statistics $NVEWMA_{N,L}$ and $NVEWMA_{N,U}$ are above the Upper control limit. Therefore, the model shows an OOC situation which is consistent with the unfavorable environmental condition. As the indeterminacy parameter I_N grows the CL is wider. Thus the graph shows smooth behavior but with a minor delay in detecting smaller variations.

The findings in Figure 3 significantly enhance monitoring performance after integrating SVR. The proposed method shows better sensitivity to smaller process changes while preserving robustness to noise. Compared to the baseline model, the SVR based method generates paths with less variability and better stability in the in-control region. It is also seen that the OOC points are around the 8^{th} sample for all level of indeterminacy values. This indicates that the SVR is able to capture non-linear relationships in data in case of multivariate marine water quality process

Figure 4 represents the PCA-RF-NVEWMA control chart. The Random Forest model is able to handle complex interaction between variables. The results create a strong separation between the in-control and OOC states specifically for large I_N values. There is a peak around the 7-10 sample points which is sharper in comparison with the prior models indicating a better detection strength. Due to the ensemble nature of Random Forest method there is a tiny increase in the variability.

Similarly, the PCA-GB–VEWMA charts in Figure 5 highlight the performance of Gradient Boosting approach in modeling residual pattern. The GB based control chart shows smoother pattern than the Random Forest model but it's very sensitive to smaller shifts. The irregular signals are clearly identified for all levels of uncertainty with better separation on normal and abnormal process behavior. Moreover, the Gradient Boosting model yields a good balance between bias and variance which guarantees consistent and reliable monitoring performance.

Overall, a comparison across Figures 2–5 is made . It is clearly seen that the control limits become wider with the increase of indeterminacy parameter I_N and greater tolerance to variability, which reflects the capability of the neutrosophic framework in dealing with uncertainty. This reduces the risk of false alarms but may slow down detection of minor modifications in case of multivariate marine water quality process. Among the hybrid techniques, PCA-SVR leads to smoother and stable monitoring, PCA-RF produces sharper identification and shifts and PCA-GB has a balance tradeoff between sensitivity and robustness.

5 Conclusion

The proposed study develops neutrosophic VEWMA control chart for monitoring multivariate marine water quality data in the presence of uncertainty while further incorporating PCA and ML models. The developed method combines VEWMA for improving shift detection, neutrosophic concept to handle indeterminacy and PCA for dimensionality reduction. The results show that PCA-NVEWMA control chart identifies process shifts associated with extreme environmental conditions, while with an increase in the indeterminacy, the control limit widens with a slight delay in detecting smaller changes. While incorporating SVR , the model ensures smooth and stable monitoring, RF sharpens detection and GB achieves a balance between sensitivity and robustness. Overall it can be concluded that the proposed approach provides an efficient tool for monitoring complex environmental conditions particularly under uncertain and non-linear conditions, especially for monitoring multivariate marine water quality data.

References

1. Aly, A.A., Hamed, R.M. and Mahmoud, M.A. (2015) 'Optimal design of the adaptive exponentially weighted moving average control chart over a range of mean shifts', Communications in Statistics – Simulation and Computation, 46(2), pp. 890–902.
2. Aslam, M. and Albassam, M. (2024) 'Neutrosophic geometric distribution: Data generation under uncertainty and practical applications', AIMS Mathematics, 9(6), pp. 16436–16452.
3. Aslam, M., Al-Marshadi, A.H. and Khan, N. (2019) 'A new X-bar control chart for using neutrosophic exponentially weighted moving average', Mathematics, 7, p. 957.
4. Aslam, M., Bantan, R.A.R. and Khan, N. (2019) 'Design of a new attribute control chart under neutrosophic statistics', International Journal of Fuzzy Systems, 21, pp. 433–440.
5. Friedman, J.H. (2001) 'Greedy function approximation: A gradient boosting machine', Annals of Statistics, 29(5), pp. 1189–1232.
6. Hossain, M.P., Omar, M.H. and Riaz, M. (2017) 'New V control chart for the Maxwell distribution', Journal of Statistical Computation and Simulation, 87(3), pp. 594–606. doi:10.1080/00949655.2016.1222391.

7. Hossain, M.P., Sanusi, R.A., Omar, M.H. and Riaz, M. (2019) 'On designing Maxwell CUSUM control chart: An efficient way to monitor failure rates in boring processes', Int J Adv Manuf Technol **100**, 1923–1930.
8. Hunter, J.S. (1986) 'The exponentially weighted moving average', Journal of Quality Technology, 18(4), pp. 203–210.
9. Khediri, I., Limam, M. and Weihs, C. (2010) 'Support vector regression control chart for monitoring nonlinear profiles', Computational Statistics & Data Analysis, 54(11), pp. 2769–2781.
10. Li, L. (2016) 'Minimax estimation of the parameter of Maxwell distribution under different loss functions', American Journal of Theoretical and Applied Statistics, 5(4), p. 202.
11. Lucas, J.M. and Saccucci, M.S. (1990) 'Exponentially weighted moving average control schemes: Properties and enhancements', Technometrics, 32(1), pp. 1–12.
12. Mathew Ishah Maria, Deepa, O. S.(2025) 'Neutrosophic EWMA and DEWMA control chart on Exponential and Transformed Exponential Distributions' International Journal of Neutrosophic Science, 26(4), 2025, pp. 184-203.
13. Moschopoulos, P.G. (1985) 'The distribution of the sum of independent gamma random variables', Annals of the Institute of Statistical Mathematics, 37(3), pp. 541–544.
14. O. S. Deepa , (2020) Modified Average Sample Number for Improved Double Sampling Plan Based on Truncated Life Test Using Exponentiated Distributions, Mathematics and Statistics, Vol. 8, No. 5, pp. 542 – 550.
15. Qiu, P. and Xie, X. (2021) 'Transparent sequential learning for statistical process control of serially correlated data', Technometrics, 63(2), pp. 172–184.
16. Sabahno, H. and Amiri, A. (2023) 'Machine learning based control charts using random forest for monitoring process parameters', Mathematics, 11(16), p. 3566.