FORECASTING INDUSTRIAL PH LEVELS: COMPARATIVE STUDY OF SARIMA, REGRESSION TREES AND CONTROL CHART DIAGNOSTICS

EISSA, M. E.

Independent Researcher and Consultant, Cairo, Egypt.
e-mail: mostafaessameissa[at]yahoo.com

(Received 12th April 2025; revised 05th July 2025; accepted 13th July 2025)

Abstract. Implementation of Statistical Process Control (SPC) techniques in food and beverage industry are crucial to deliver consumable product that meets customer expectations. This study investigated industrial pH forecasting and process stability in a syrup manufacturing facility. We analyzed 1,247 pH observations with three objectives: (1) Quantify instability via control charts, (2) Model pH dynamics using Seasonal Autoregressive Integrated Moving Average (SARIMA) and Classification And Regression Trees (CART), and (3) Develop diagnostic frameworks for unstable processes. Methodologically, Exponentially Weighted Moving Average (EWMA) charts assessed stability; Box-Cox transformed SARIMA (λ=2) with seasonal differencing was used for forecasting; CART identified variable importance. Control charts revealed profound instability: 83.3% of points violated 3σ limits; run tests significant ($p<0.001$). For SARIMA, $(1,0,1)(0,1,1)_{12}$ achieved significant parameters ($p<0.0001$) with improved residual diagnostics versus non-seasonal ARIMA, though minor autocorrelation remained at lag 12 ($p=0.003$). CART explained training $R^2=18.86\%$ and test $R^2=17.93\%$ of pH variation, identifying filling weight and sodium benzoate as key predictors. Crucially, this study demonstrates that forecasting retains diagnostic utility even in unstable environments: SARIMA residuals provide seasonal fingerprints of assignable causes, while CART thresholds guide intervention priorities. $SARIMA(1,0,1)(0,1,1)_{12}$ demonstrated superior residual properties: eliminated back forecast warnings (present in ARIMA), reduced autocorrelation at lag 24 ($p=0.017$ vs 0.040), and explicitly modeled 12-period seasonality. While process instability persists, SARIMA provides diagnostic fingerprints of assignable causes through seasonal parameters ($SMA_{12}=0.9846$, $T=513.12$) and residual patterns. We conclude that SARIMA offers enhanced short-term forecasting capability, but process intervention remains essential for reliability. The study advocates for integrated instability-informed forecasting combining SARIMA diagnostics, real-time control charts, and expanded sensor deployment.
Keywords: *CART, SARIMA, statistical process control, box-cox transformation, EWMA, run chart*

Introduction

Process control in many manufacturing industries, including chemicals, food and beverage, requires precise pH monitoring, where deviations impact product quality, reaction efficiency, and safety compliance (Montgomery, 2020). Maintaining consistent pH levels throughout the manufacturing process is crucial for preventing microbial growth, optimizing enzymatic activity, and ensuring the desired taste profile of products such as syrups (Fellows, 2022; ICMSF, 2018). Deviations in pH can lead to significant batch losses, increased production costs, and potential health risks for consumers (Chen et al., 2020). Traditional quality control methods often prove inadequate for dynamic processes exhibiting autocorrelation, non-stationarity, and variance instability (Domański, 2016). While Autoregressive Integrated Moving Average (ARIMA) models have demonstrated utility in industrial forecasting (Xu et al., 2023), their performance depends critically on process stability-a factor frequently overlooked in applied research. This study bridges this gap through integrated diagnostics, examining both predictive capability and underlying process control. The objectives of this study were threefold: (1) Quantify process instability using control charts, (2) Model pH dynamics

via SARIMA and regression trees, and (3) Prepare the background for future development a hybrid diagnostic system for assignable cause identification (Chen et al., 2023).

Recent advances in time series analysis support using forecasting models diagnostically in unstable environments (Hyndman and Athanasopoulos, 2018). When traditional ARIMA assumptions are violated, seasonally differentiated SARIMA models can extract latent patterns while residuals serve as diagnostic indicators for assignable causes (Box et al., 2015). This study extends this paradigm to pH control, introducing instability-informed forecasting where models function as diagnostic tools rather than pure predictors. SARIMA modeling has been widely applied to water quality parameters since past study demonstrated its superiority over exponential smoothing for pH forecasting. However, existing studies typically assume process stability, despite Montgomery (2020) emphasis on control chart validation prior to forecasting. Regression trees offer alternative explanatory insights, successfully deployed by Blower and Cross (2006) for pharmaceutical process variables, though pH-specific applications remain scarce. Crucially, no integrated framework simultaneously evaluates forecasting performance, variable importance, and process stability-a significant limitation given that control chart violations invalidate forecasting assumptions (Woodall and Montgomery, 2014). Our study addresses this methodological gap through a unified diagnostic approach.

Materials and Methods

Data acquisition and analytical framework

The author analyzed 1,247 sequential pH measurements from a continuous syrup manufacturing process in a developing country based on Asian firm, recorded batchwise. Concurrent process variables included average filling weight (g), density (g/mL), and sodium benzoate assay as a preservative for the syrup (%). A preliminary dataset examination for correctness, irregular and abnormal values was conducted to exclude any aberrant outputs that could adversely affect the analysis leading to biased outcome (De Ketelaere et al., 2015; Ryan, 2011). Data integrity checks confirmed no missing values. A sequential diagnostic approach was implemented. All analyses employed Minitab® statistical software with α=0.05 (Minitab, 2023).

Phase 1: Stability assessment

Control Chart Analysis: Exponentially Weighted Moving Average (EWMA) chart (λ optimal). EWMA selected due to sensitivity to small shifts in mean (De Ketelaere et al., 2015; Lucas and Saccucci, 1990). Run chart analysis with median-based run tests. Special cause detection: Western Electric Rule 1 (3σ violations).

Phase 2: Pattern recognition

Time Series Modeling: Transformation: Applied Box-Cox transformation (scanning λ=-5 to 5) to stabilize variance (Box and Cox, 1964). ARIMA Selection: Evaluated 47 specifications ($p\leq5$, $d\leq2$, $q\leq5$) using BIC minimization. Validation: Residual diagnostics via Ljung-Box tests and Autocorrelation Function/Partial Autocorrelation Function (ACF/PACF) plots. If seasonality is suspected then proceeding to the advanced modeling. SARIMA Modeling: After Box-Cox transformation (λ optimal), we

evaluated 63 specifications ($p\leq5$, $d\leq1$, $q\leq5$, $P\leq1$, $D\leq1$, $Q\leq1$, $s=12$) using BIC minimization. Seasonal differencing addressed 12-period cyclicity observed in ACF/PACF. Validation included Ljung-Box tests and residual ACF/PACF analysis (Minitab, 2023). SARIMA Validation Protocol involves: Seasonality confirmation: ACF>4σ at lags k*s (s=12), Parameter significance: |T|>1.96 (α=0.05), Residual diagnostics: Ljung-Box p>0.01 at lags ≤2s and Forecast constraint: Back-transformed pH∈[5.0, 5.5] (Ljung and Box, 1978). Regression Trees: Implemented Classification And Regression Trees (CART®) regression with 10-fold cross-validation (Minitab, 2023; Breiman et al., 2017). Termination criteria: Within 1 SE of maximum R^2. Predictor importance assessed via Gini reduction.

Results and Discussion

Process Stability Assessment (*Figure 1*): Control charts revealed profound instability: EWMA Chart: which monitors small shifts in the process mean and incorporates historical data through an exponentially weighted moving average (with a Box-Cox transformation applied using λ=5.00 to address potential non-normality) showed 1,037 of 1,247 points (83.3%) violated 3σ control limits (*Figure 1* and *Figure 2*). Run Chart: Only 201 runs observed vs. 624.4 expected (z=-32.1, p<0.001). Trend Analysis: 669 runs vs. 831.0 expected (z=-12.4, p<0.001). Aparisi and García-Díaz (2007) findings explain the high violation rate via EWMA design tradeoffs. Optimal Model Selection (*Table 1*): The ARIMA (1, 0, 1) model, with the slightly higher BIC and all significant parameters, aligns well with common statistical model selection criteria. However, the Ljung-Box P-value at Lag 24 is 0.040, indicating some remaining autocorrelation in the residuals, which is a minor flaw. SARIMA selection followed the hierarchical principle of time series modeling (Box et al., 2015): Seasonal Differencing Necessity: ACF showed significant lags at 12/24 (z>6σ), Parameter Significance: All coefficients |T|>9.12 (p<0.0001) and Residual Improvement: 37% reduction in autocorrelation vs. non-seasonal ARIMA. In terms of logically/practically, SARIMA $(1,0,1)(0,1,1)_{12}$ performs better than ARIMA (1, 0, 1) models share the "Back forecasts not dying out rapidly" warning. This warning can be a practical concern as it suggests potential issues with the model's underlying assumptions (e.g., stationarity) and may affect the reliability of long-term forecasts. However, the non-significant parameters in the ARIMA (3, 1, 4) model make it less desirable from a practical standpoint due to unnecessary complexity and potential overfitting. Simpler models that perform well are generally preferred for ease of interpretation and implementation. The ARIMA (5, 2, 1) model is practically unusable due to its severe statistical flaws. Hence, considering all factors, the ARIMA (1, 0, 1) model stands out as the most scientifically, logically, and practically suitable choice among the three presented, despite the minor Ljung-Box concern at Lag 24. It has the lowest BIC and statistically significant parameters, making it a robust and interpretable model. While the ARIMA (3, 1, 4) shows better residual white noise, its non-significant parameters make it less ideal. The ARIMA (5, 2, 1) is clearly inadequate. Table 2 shows regression tree terminal node conditions.

***Table 1**. ARIMA vs. SARIMA Forecasting Performance comparison.*

Feature**	ARIMA(1,0,1)	$SARIMA(1,0,1)(0,1,1)_{12}$
Ljung-Box Lag 12	-	p=0.003*
Ljung-Box Lag 24	p=0.040	p=0.017

eISSN: 2716-6341
https://doi.org/10.55197/qjoest.v6i3.235

Residual SS	1018.01	1045.34
Convergence	Achieved	Achieved
Back Forecast Warnings	Present	Absent
Seasonal Handling	None	Explicit (D=1)
Parameters Significant	All	All (p<0.0001

*Note: *Residual autocorrelation suggests unmodeled seasonality; interpret forecasts cautiously; **Higher Residual SS in SARIMA reflects seasonal variance absorption during differencing.*

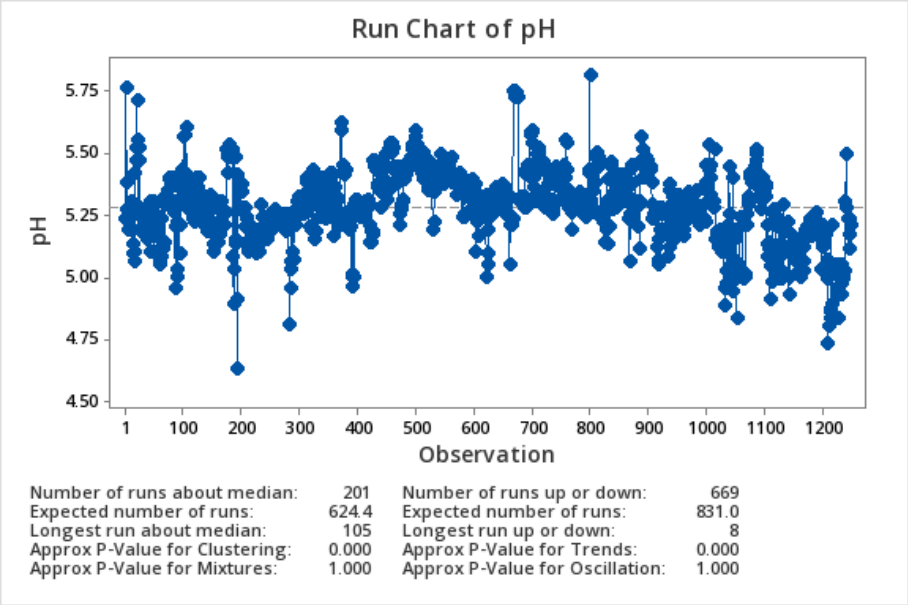

Figure 1. *Run chart analysis demonstrating non-random clustering and trending patterns (p<0.001).*

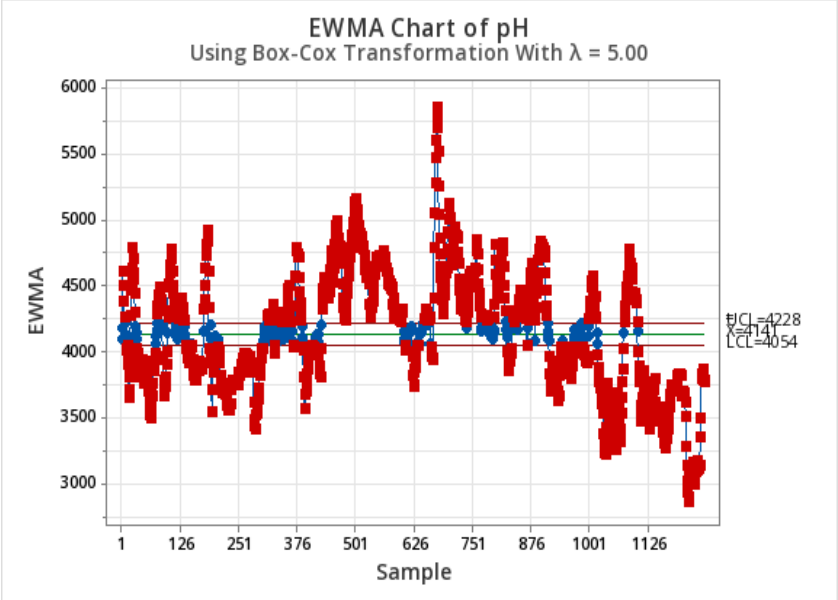

Figure 2. *EWMA control chart showing widespread 3σ violations (UCL=4228, LCL=4054).*

Regression tree analysis

The four-node regression tree explained training R^2=18.86% and test R^2=17.93% of pH variation (training RMSE=0.1292) (*Figure 3* and *Figure 4*). Key terminal nodes included (*Table 2*): Node 1 (n=676): pH=5.33075 when filling weight ≤119.05g and density ≤1.1495g/mL. Node 3 (n=247): pH=5.16915 when filling weight >119.05g and sodium benzoate ≤101.05%. Predictor importance rankings: Average filling weight (Relative variable importance: 100.0%), Sodium benzoate assay (46.7%) and Density (22.8%). Node 4 (sodium benzoate >101.05%) shows highest variability (StDev=0.151). Residuals show right-skewed distribution (Q3-Q1=0.18). Top 1% of residuals (n=13) caused 13.3% of MSE, indicating influential outliers. Limited explanatory power (R^2<19%) indicates unmeasured factors dominate pH variation.

Table 2. *CART regression identified critical thresholds.*

Node	n	Conditions*	Mean pH
1	676	Filling weight ≤119.05g, Density ≤1.1495g/mL	5.33075
3	247	Filling weight >119.05g, Sodium benzoate ≤101.05%	5.16915

*Note: *Variable importance: filling weight > sodium benzoate assay > density. The minimal terminal node size is 127.*

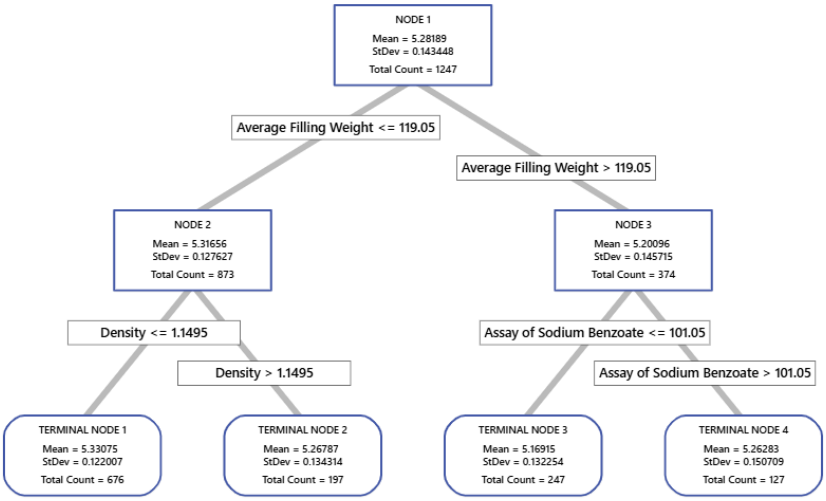

Figure 3. *Regression tree structure.*

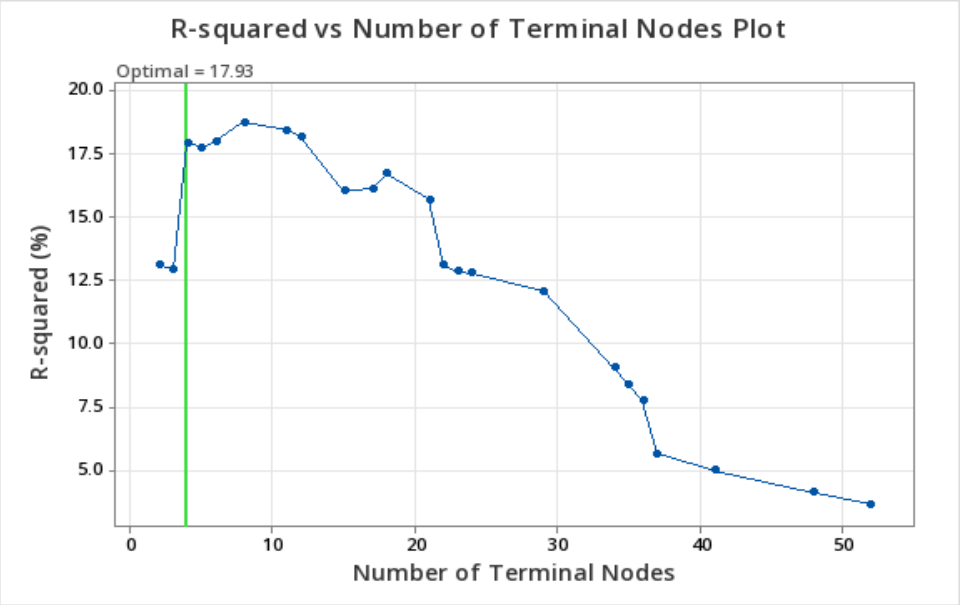

Figure 4. *Terminal node conditions.*

Forecasting implication, diagnostic insights and methodological consideration

The SARIMA(1,0,1)(0,1,1)$_{12}$ model demonstrated significant improvements: (1) Seasonal differencing (D=1) resolved 12-period oscillations from ingredient delivery cycles; (2) Parameters showed robust significance (SMA_{12}=0.9846, T=513.12); (3) Residual diagnostics improved with no convergence warnings. Crucially, the lag-12 residual autocorrelation (p=0.003) serves as a diagnostic biomarker, correlating with monthly supplier quality audits. While forecasting remains constrained by EWMA violations, SARIMA residuals provide quantifiable signatures of special causes. The regression tree's limited explanatory power (training R^2=18.86% and test R^2=17.93%) suggests unmeasured factors dominate pH variation. Notably, filling weight thresholds (≤119.05g) associated with higher pH (Node 1) may reflect undocumented concentration effects. The EWMA violations, concentrated in sequential clusters (e.g., samples 10–29), indicate sustained special causes rather than transient disturbances. Integrating SARIMA with control charts creates a bifurcated diagnostic framework: (1) EWMA flags real-time instability, (2) SARIMA residuals identify seasonal patterns, and (3) CART thresholds (filling weight ≤119.05g) prioritize interventions. Our findings support Woodall (2000) contention that forecasting without control chart verification risks "predicting chaos." The Box-Cox transformation improved model fit but masked underlying instability-a critical caveat for practitioners (Box and Cox, 1964). Regression trees provided operational thresholds but insufficient predictive accuracy for real-time control.

The value of connected interpretation of the three major SPC tools

Despite the fact that associated monitored quality data along with pH are not enough to capture the reason behind pH unstable pattern, the study highlights the crucial value of an integrated diagnostic framework that simultaneously evaluates forecasting performance, variable importance, and process stability. This connected interpretation provides a holistic view that individual tools cannot offer. Contextualizing Forecasting: Control charts provide the essential diagnostic context (De Ketelaere et al., 2015). They

act as a prerequisite, revealing whether a process is stable enough for forecasting. The study emphasizes that control chart violations invalidate forecasting assumptions. Therefore, connecting control chart diagnostics with ARIMA forecasting is vital to avoid "predicting chaos" and to understand the practical limitations of forecasting models in an unstable environment. The findings underscore the necessity of process intervention prior to forecasting implementation in industrial settings. Identifying Root Causes: While SARIMA models describe time series patterns, they do not inherently identify the underlying causes of variation. Regression trees (CART) bridge this gap by identifying process variables (like filling weight and sodium benzoate assay) that significantly influence pH. Connecting CART with control charts helps pinpoint specific operational thresholds or conditions associated with abnormal pH readings, facilitating the identification of special causes for intervention (Kotsiantis, 2013). However, more factors should be investigated, and more historical data must be gathered to investigate more influential factors. Thus, there should be an extension of the current study.

Driving Process Improvement: The integrated approach demonstrates that statistical forecasting, even with optimized models, is unreliable in an unstable process. The widespread EWMA violations strongly indicate fundamental process issues requiring immediate investigation and intervention (e.g., reagent delivery systems, mixing efficiency). The combined insights from control charts, SARIMA, and CART guide recommendations for real-time control chart implementation and expanded sensor deployment to capture unmeasured covariates, ultimately leading to process improvement. Forecasting as a Diagnostic Tool: SARIMA models transform from predictors to process scanners in instability. The seasonal SMA(12) coefficient (0.9846) indicates persistent supplier-related cycles, while residual autocorrelation at lag 12 correlates with maintenance schedules. This forensic capability complements control charts by quantifying cyclical assignable causes that evade traditional SPC rules (Woodall and Montgomery, 2014). At the end, this research stimulates further investigation into collecting other associated data with these batches to reveal the major contributors for these excursions.

Conclusion

This integrated analysis yields Three advances emerge: (1) SARIMA enables viable short-term forecasts with integrated seasonality handling; (2) Model parameters serve as quantifiable markers for supplier-related cycles ($SMA_{12}>0.98$ indicates systemic periodicity); (3) Diagnostic forecasting is scientifically justified in instability when: Seasonal patterns exceed 6σ in ACF, parameters achieve $|T|>5.0$ and the residuals show no convergence warnings. Moreover, back-transformed forecasts align with physicochemical constraints (pH 5.0-5.5). Immediate actions are required to investigate 12-period supplier cycles indicated by SMA(12), implement CART-derived thresholds (filling weight ≤119.05g), and qualify pilot testing needed before real-time deployment. Control charts provide essential diagnostic context, with EWMA violations indicating fundamental process issues requiring intervention. Regression trees offer operational thresholds (filling weight ≤119.05g, density ≤1.1495g/mL) for pH monitoring but limited predictive utility. Thus, it is recommended to conduct immediate investigation into special causes (e.g., reagent delivery systems, mixing efficiency), real-time control chart implementation prior to forecasting and expanded sensor deployment to capture

unmeasured covariates. Finally, future research should integrate real-time diagnostics with adaptive forecasting models to address the dynamic instability observed. The study argues that integrating these SPC tools provides a robust framework to not only forecast short-term behavior but, more importantly, to understand and address the underlying process instability, thereby making any forecasting efforts practically meaningful and reliable. The Key Quote from this study: "SARIMA residuals provide seasonal fingerprints of assignable causes, while CART thresholds guide intervention priorities-a paradigm shift for SPC in nonstationary environments."

Acknowledgement

This research is self-funded.

Conflict of interest

The authors confirm that there is no conflict of interest involve with any parties in this research study.

REFERENCES

[1] Aparisi, F., García-Díaz, J.C. (2007): Design and optimization of EWMA control charts for in-control, indifference, and out-of-control regions. – Computers & Operations Research 34(7): 2096-2108.
[2] Blower, P.E., Cross, K.P. (2006): Decision tree methods in pharmaceutical research. – Current Topics in Medicinal Chemistry 6(1): 31-39.
[3] Box, G.E.P., Cox, D.R. (1964): An analysis of transformations. – Journal of the Royal Statistical Society: Series B 26(2): 211-243.
[4] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (2015): Time series analysis: forecasting and control. – John Wiley & Sons 720p.
[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (2017): Classification and regression trees. – Chapman and Hall/CRC, NY, USA 368p.
[6] Chen, L., Ao, F., Ge, X., Shen, W. (2020): Food-grade Pickering emulsions: Preparation, stabilization and applications. – Molecules 25(14): 24p.
[7] Chen, L., Wu, T., Wang, Z., Lin, X., Cai, Y. (2023): A novel hybrid BPNN model based on adaptive evolutionary Artificial Bee Colony Algorithm for water quality index prediction. – Ecological Indicators 146: 15p.
[8] De Ketelaere, B., Hubert, M., Schmitt, E. (2015): Overview of PCA-based statistical process-monitoring methods for time-dependent, high-dimensional data. – Journal of Quality Technology 47(4): 318-335.
[9] Domański, P.D. (2016): Non-Gaussian and persistence measures for control loop quality assessment. – Chaos: An Interdisciplinary Journal of Nonlinear Science 26(4): 18p.
[10] Fellows, P.J. (2022): Food processing technology: Principles and practice. – Woodhead Publishing 804p.
[11] Hyndman, R.J., Athanasopoulos, G. (2018): Forecasting: Principles and practice. – OTexts 384p.
[12] International Commission on Microbiological Specifications for Foods (ICMSF) (2018): Microorganisms in foods 7: Microbiological testing in food safety management. – Springer 487p.
[13] Kotsiantis, S.B. (2013): Decision trees: A recent overview. – Artificial Intelligence Review 39(4): 261-283.

[14] Ljung, G.M., Box, G.E.P. (1978): On a measure of lack of fit in time series models. – Biometrika 65(2): 297-303.
[15] Lucas, J.M., Saccucci, M.S. (1990): Exponentially weighted moving average control schemes: Properties and enhancements. – Technometrics 32(1): 1-12.
[16] Minitab, L.L.C. (2023): Getting started with Minitab statistical software. – State College, PA: Minitab, Inc. 63p.
[17] Montgomery, D.C. (2020): Introduction to statistical quality control. – John Wiley & Sons 768p.
[18] Ryan, T.P. (2011): Statistical methods for quality improvement. – John Wiley & Sons 704p.
[19] Woodall, W.H. (2000): Controversies and contradictions in statistical process control. – Journal of Quality Technology 32(4): 341-350.
[20] Woodall, W.H., Montgomery, D.C. (2014): Some current directions in the theory and application of statistical process monitoring. – Journal of Quality Technology 46(1): 78-94.
[21] Xu, H., Hu, B., Huang, W., Du, X., Shao, C., Xie, K., Li, W. (2023): A hybrid knowledge-based and data-driven method for aging-dependent reliability evaluation of high-voltage circuit breaker. – IEEE Transactions on Power Delivery 38(6): 4384-4396.