MODELING MICROBIOLOGICAL COUNTS IN PURIFIED WATER AT A HEALTHCARE FACILITY USING ARIMA

EISSA, M. E.

Biometric & Statistics Department, Pharmaceutical Research Facility, Cairo, Egypt.
e-mail: mostafaessameissa[at]yahoo.com

(Received 27th February 2025; revised 10th June 2025; accepted 19th June 2025)

Abstract. The microbiological quality of purified water is a crucial aspect in the healthcare industry to ensure safety for different applications and uses. Understanding the trend and forecasting would be of prime importance to take proactive control and protective measures before catastrophic excursions might occur leading financial and health casualties. This study analyzes microbial density, a key metric for monitoring water purification system efficacy in healthcare facilities. The objective was to transform irregular, cumulative data into a regular time series and identify the optimal ARIMA model for forecasting to support predictive maintenance and regulatory compliance. Preliminary modeling attempts were conducted using simpler approaches such as linear, exponential and Holt-Winters methods without showing promising outcomes. Descriptive statistics and distribution analysis, including the Johnson Transformation for normality, were performed. ARIMA models with differencing orders d=0, d=1, and d=2 were fitted to the Aggregated cumulative logarithmically transformed data series, with the best model at each order selected based on minimum AICc. Model adequacy was assessed through parameter significance and residual diagnostics (Ljung-Box test). Descriptive statistics showed the aggregated series non-normal ($p<0.005$). ARIMA(5, 0, 4) (d=0) performed poorly (AICc=319.39) with residual autocorrelation. ARIMA(0, 2, 1) (d=2) showed improved fit (AICc=258.98) and white noise residuals ($p>0.5$). The ARIMA(2, 1, 2) model (d=1) was optimal (AICc=256.91), with all significant parameters and white noise residuals ($p>0.3$), effectively addressing non-stationarity. Forecasts from ARIMA(2, 1, 2) predict stable future growth. The ARIMA(2, 1, 2) model with first-order differencing is the most appropriate and robust model for forecasting data trends. Its strong statistical fit and reliable residual properties make it a valuable tool for predictive maintenance, optimizing resources, and enhancing patient safety in healthcare water systems, provided model performance is continuously monitored. Addressing data limitations and processing requires monitoring and exploring alternative models for future improvement.

Keywords: *ARIMA, autocorrelation, bioburden, healthcare facility, Akaike information criterion corrected, purified water*

Introduction

High-quality purified water is crucial for patient care in healthcare facilities and for public health broadly. It is used in vital applications like hemodialysis, medication compounding, sterilization, and surgical procedures. Even low levels of microbial contamination in this water can endanger patient safety, leading to serious healthcare-associated infections, particularly in vulnerable and immunocompromised individuals (Kanamori et al., 2016). Such infections can result in longer hospital stays, increased healthcare costs, significant illness, and even death (Kmentt et al., 2021). Therefore, strict microbiological control and continuous monitoring of purified water systems are essential to ensure patient safety and maintain regulatory compliance (Yetiş et al., 2023). Traditional methods for monitoring the microbiological quality of purified water typically involve routine sampling and laboratory analysis to quantify viable microorganisms, often expressed as Colony Forming Units (CFU) per unit volume. While this provides essential information about the current microbiological status at specific sampling points, it offers a retrospective "snapshot" (Kmentt et al., 2021). This reactive approach has limitations in predicting future trends in microbial populations, which is problematic if bioburden levels approach critical thresholds that could compromise water safety (Zafra-Mejía et al., 2024). Bioburden, defined as the population of viable microorganisms in or on a product or material, is a relevant concept for purified water systems (Dishan et al., 2024). Maintaining low bioburden levels is challenging due to factors inherent in these systems. Microbial ingress can occur through breaches in system integrity, such as leaks or improperly maintained filters (Kanamori et al., 2016). Nutrient availability, even at trace levels from source water or system materials, can support microbial growth. Temperature fluctuations can also influence microbial proliferation, with warmer temperatures generally favoring growth. Furthermore, the formation of biofilms on the internal surfaces of piping and components provides a protective niche for microorganisms, making them more resistant to disinfection strategies (Yetiş et al., 2023).

Purified water systems utilize a combination of treatment technologies to minimize bioburden, including filtration, reverse osmosis, deionization, and disinfection methods like ultraviolet (UV) irradiation or chemical treatment. Despite these measures, the potential for microbial contamination persists, underscoring the need for continuous monitoring and proactive management strategies (Kanamori et al., 2016). Forecasting microbiological counts in purified water systems emerges as a valuable proactive tool. By anticipating potential increases in microbial populations, healthcare facilities can implement timely interventions, such as optimizing preventive maintenance schedules (e.g., filter replacement, system disinfection) before bioburden levels exceed acceptable limits. This can help prevent system downtime, minimize disruptions to patient care, and ultimately enhance patient safety by ensuring the consistent availability of microbiologically acceptable water (Zafra-Mejía et al., 2024). Time series analysis techniques offer a robust framework for modeling and forecasting data collected sequentially over time, such as microbiological monitoring results. Autoregressive Integrated Moving Average (ARIMA) models are a widely used class of time series models capable of capturing various temporal dependencies, including trends, seasonality, and short-term correlations (Zafra-Mejía et al., 2024). Given the potential for temporal patterns and trends in microbiological contamination data, ARIMA modeling presents a promising approach for developing predictive tools for purified water systems in healthcare facilities. This study aimed to explore the effectiveness and applicability of ARIMA modeling for forecasting microbiological counts in purified water within a healthcare facility. By analyzing historical data and applying ARIMA techniques, the study sought to demonstrate the feasibility and potential benefits of using such models for proactive bioburden management and to identify a suitable ARIMA model for this specific application.

Materials and Methods

While the current work aimed to primarily focus on ARIMA modeling, a rapid screening of other approaches will be briefly mentioned to show if they are able to capture any trends successfully or the modeling will be compromised. Various time series forecasting methods exist that could potentially be applied to water quality data, including other statistical models and machine learning techniques (Chidiac et al., 2023). Data Acquisition and Preprocessing: Historical microbiological count data from the Purified Water (PW or PU) system at a healthcare facility were obtained for this study from January 2023 to May 2024 on almost weekly basis from database center of the plant in Giza governorate, Egypt. The dataset comprised observations aggregated to n=72 of microbiological counts-expressed as Colony Forming Unit (CFU)-in 100 mL samples of purified water (PW) collected from a sampling port over a period spanning almost from the installation of treatment unit (Eissa et al., 2022). The data included the date of sampling and the corresponding microbiological count in Colony Forming Units (CFU) per 100 mL. No missing values were present in the dataset. The use of historical water quality data for analysis and forecasting is a common practice in environmental and public health studies (Zafra-Mejía et al., 2024). Prior to modeling, the data underwent several preprocessing steps. Elapsed time in days was calculated from the first sampling date. Given the often-skewed nature of microbiological count data and to stabilize variance, the data was logarithmically transformed (Feldl et al., 2025; Eissa, 2024; Eissa et al., 2023). Cumulative counts were then calculated for the logarithmically transformed data (Cum.Log.PW), representing the sum of transformed counts up to each time point. This cumulative transformed data was used for the subsequent time series modeling. Cumulative sum analysis can be a useful technique in analyzing environmental time series data (Lange and Hauhs, 2025).

Exploratory Data Analysis (EDA): Exploratory data analysis was conducted to understand the characteristics of the microbiological count data. Descriptive statistics were calculated for the raw microbiological counts, including minimum, maximum, mean, median, standard deviation, skewness, and interquartile range (IQR). Visualizations were generated, including a Run Chart, Cumulative Histogram, and Box Plot of the raw data (PW), to assess distribution patterns, central tendency, dispersion, and potential outliers. These EDA techniques are standard practice in analyzing datasets, including environmental time series, to gain initial insights into the data structure and identify potential issues before formal modeling (Lokanan, 2024). ARIMA Model Selection: ARIMA model selection was performed using statistical software (Minitab). The process involved assessing the stationarity of the time series, identifying potential model orders (p, d, q), fitting various candidate ARIMA models, and selecting the best model based on information criteria and residual diagnostics. Stationarity is a key assumption for ARIMA models, and differencing is a common technique used to transform non-stationary time series into stationary ones by removing trends or seasonality (FasterCapital Web Portal, 2025; Xue, 2024). If visual inspection of the time series plot of the cumulative transformed data suggested the presence of a trend then to address the trend and achieve stationarity for ARIMA modeling, a zero, first and second-order differencing were scanned to the cumulative transformed data. This step effectively detects the optimal model parameters to be selected for the time-series data. After screening for the best differencing parameter, various ARIMA models were then considered by varying the orders of the autoregressive (p) and moving average (q) terms (FasterCapital Web Portal, 2025). Candidate models included ARIMA (p,d (optimal),q) for different values of p and q. Each candidate ARIMA model was fitted to the differenced cumulative transformed data. The Akaike Information Criterion corrected for small samples (AICc) was used as the primary metric for model selection. AICc is a widely accepted criterion for selecting among competing statistical models, particularly in time series analysis with limited data points, balancing model fit and complexity (Mazerolle, 2023; Portet, 2020). The model with the minimum AICc value was considered the best fit among the candidate models, balancing goodness of fit with model complexity. ARIMA models have been successfully applied to water quality prediction in various studies (Zafra-Mejía et al., 2024).

Model Diagnostics: The selected ARIMA model underwent diagnostic checking to evaluate its assumptions and adequacy. Residual analysis was performed to assess the properties of the model's residuals (the differences between the observed and fitted values). Ideally, the residuals of a well-fitting ARIMA model should resemble white noise-a series of independent and identically distributed random variables with a mean of zero and constant variance. The Ljung-Box test was employed to statistically assess the presence of autocorrelation in the residuals at various lags (Alsheheri, 2025). High p-values (typically>0.05) for the Ljung-Box test indicate no significant autocorrelation in the residuals, supporting the adequacy of the model. Forecasting: Once the optimal ARIMA model was selected and validated, it was used to generate forecasts for future cumulative transformed microbiological counts. Forecasts were generated for a specified forecasting horizon beyond the historical data period. The software provided point forecasts (the most likely future values) and 95% confidence intervals around these forecasts, reflecting the uncertainty associated with the predictions. Generating confidence intervals around forecasts is crucial for quantifying the uncertainty of predictions in time series analysis (Wang and Xu, 2025). A time series plot illustrating the historical data, the fitted values, and the future forecasts with their confidence limits was generated.

Results and Discussion

Descriptive Analysis of Microbiological Counts: The descriptive statistics of the raw microbiological counts (PW) from the purified water system are presented in the analysis output (*Figure 1*). With 68 observations, the counts ranged from a minimum of 0 to a maximum of 100 CFU/mL. The mean count was 4.476 CFU/mL, while the median was considerably lower at 0.409 CFU/mL. This substantial difference between the mean and median, coupled with a high Pearson skewness of 5.850, confirmed that the distribution of microbiological counts was highly positively skewed, meaning the majority of counts were low, but there were infrequent higher values (Kirkwood and Sterne, 2010). The standard deviation of 13.423 CFU/mL and a coefficient of variation of 2.999 indicated notable variability in the data relative to the mean. The interquartile range (IQR) of 3.805 CFU/mL provided insight into the spread of the middle 50% of the data. The Cumulative Histogram visually demonstrated the distribution of the raw PW data and its close fit to a Negative Binomial distribution with specified parameters (*Figure 2*). This finding is consistent with the nature of count data, particularly in situations where rare events (microbial detection) occur, and supports the understanding of the underlying data generation process (Hilbe, 2011). The Box Plot of PW further illustrates the skewed nature of the data, with a median positioned towards the lower end of the box (*Figure 3*). The presence of an outlier at approximately 100 CFU/mL highlighted an infrequent but significant elevated count, reinforcing the non-normal distribution and the impact of such events on the mean (Lemenkova, 2019).

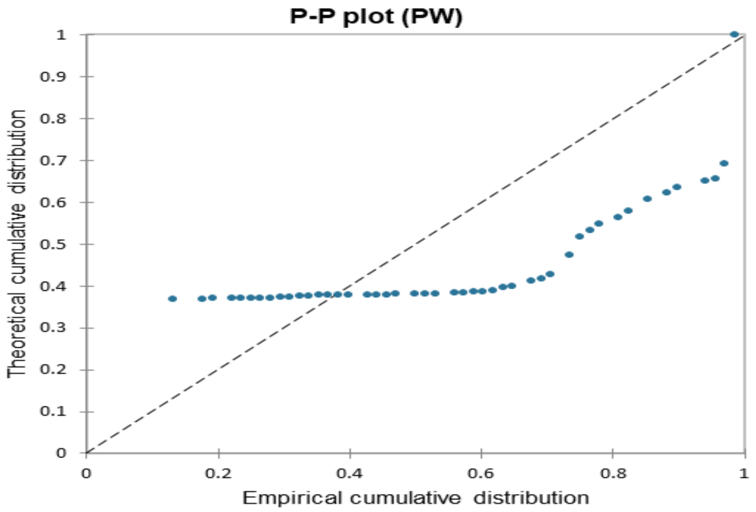

Figure 1. *Descriptive analysis of microbiological water data trend showing non-Gaussian behavior with negative binomial distribution pattern with a rising tendency in the bioburden level with time.*

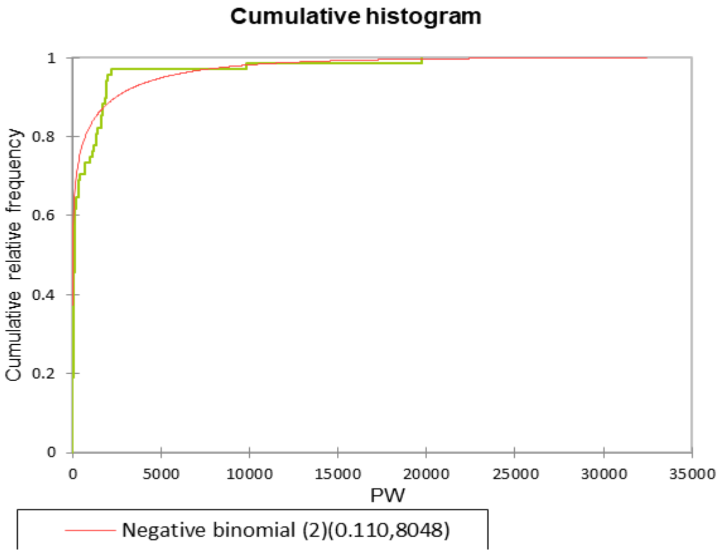

Figure 2. *The cumulative histogram.*

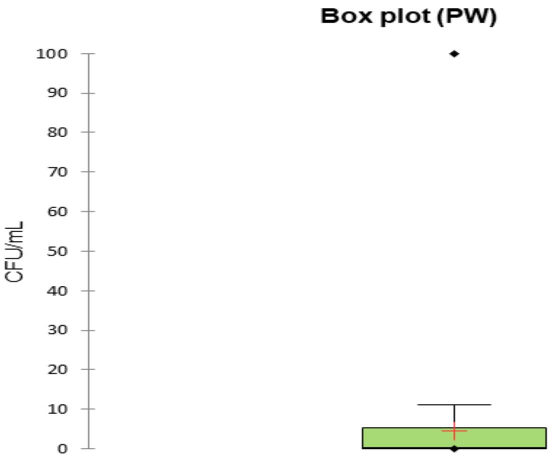

Figure 3. *The box plot of PW.*

Formal distribution identification testing confirms that the data does not follow a normal distribution, with the Anderson-Darling test yielding a p-value less than 0.005. The Anderson-Darling test is a widely used statistical test for assessing if a sample comes from a specified distribution, such as the normal distribution (Razali and Wah, 2011). Challenges were encountered when attempting to fit several common distributions directly, such as Exponential, Lognormal, Weibull, Gamma, and Loglogistic, due to the presence of non-positive values in the data. However, the Johnson Transformation proved highly effective in achieving normality for the data, as evidenced by an Anderson-Darling statistic of 0.131 and a p-value of 0.981 after applying the transformation. The Johnson transformation is a flexible method for transforming data to approximate a normal distribution (Johnson, 1949). The Run Chart of PW (*Figure 4*) provided a visual time series representation of the raw data. While the number of runs about the median (25) and the number of runs up or down (46) were somewhat close to the expected values for a random process, the approximate p-value for clustering (0.007) suggested potential non-random clustering behavior. Conversely, high p-values for mixtures, trends, and oscillation (0.993, 0.615, and 0.385, respectively) indicated no statistically significant evidence of these specific patterns. Run charts and analysis of runs are useful tools in time series analysis for detecting non-random patterns like trends, shifts, or clustering (Anhøj and Olesen, 2014). This preliminary analysis suggested that while a clear linear trend or oscillation was not dominant in the raw counts, there might be underlying patterns of clustering or intermittent excursions from low baseline levels, which is characteristic of bioburden monitoring data in controlled water systems.

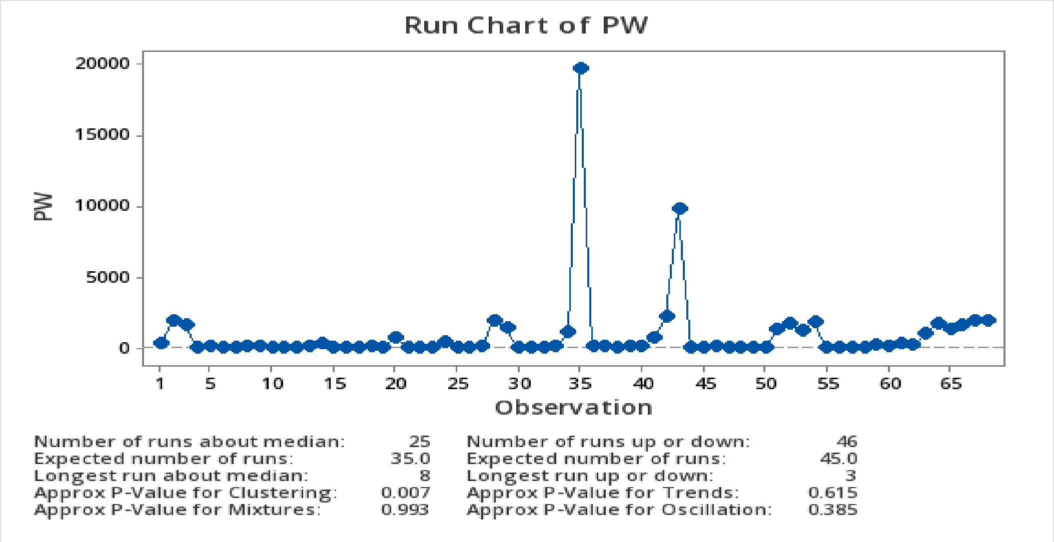

Figure 4. *The bioburden in purified water: It displays the number of runs above the median, the number of runs up or down, and the expected number of runs for each time period.*

Evaluation of Alternative Models: Prior to selecting the ARIMA model, simpler smoothing techniques were considered to understand the data's characteristics and justify the need for a more complex approach. Simple Exponential Smoothing was applied to the cumulative transformed data. This method is suitable for data without a trend or seasonality. The analysis revealed a low R^2 and high error metrics compared to other models. Crucially, the residual analysis showed consistently positive and increasing residuals, indicating a significant systematic under-prediction by the model. This demonstrated that Simple Exponential Smoothing was inappropriate for this data due to its inability to account for the clear upward trend in the cumulative transformed counts (Hyndman et al., 2008). A simple Linear Regression model, using Elapsed Time as the predictor for Cum.Log.UV, was also evaluated. This model captured the overall linear trend effectively, resulting in a high R^2. The model parameters (Intercept and Elapsed Time coefficient) were highly statistically significant. However, the Durbin-Watson statistic provided strong evidence of significant positive autocorrelation in the residuals. The Durbin-Watson statistic is commonly used to detect the presence of autocorrelation in the residuals of a regression analysis (Durbin and Watson, 1992). This indicated that the linear model, while describing the average trend well, failed to capture the time-dependent structure of the variations around that trend. The presence of autocorrelated residuals violates a key assumption of linear regression and compromises the reliability of statistical inferences and forecasts (Chatfield and Xing, 2019). The limitations of both Simple Exponential Smoothing (inability to handle trend) and Linear Regression (autocorrelated residuals despite capturing the trend) underscored the necessity of employing a time series model capable of explicitly addressing both the trend and the temporal dependencies in the data, such as an ARIMA model.

ARIMA Model Selection and Evaluation: Given the trending nature observed in the cumulative transformed data and the potential temporal dependencies suggested by the initial analysis and the limitations of simpler models, ARIMA modeling was a suitable approach. The analysis focused on finding an appropriate model for the cumulative logarithmic transformed data (Cum.Log.PU), which exhibited a clear upward trend. To address the trend and achieve stationarity for ARIMA modeling, a first-order differencing (d=1) was applied. This was chosen after careful analysis and selection between models at d=0 and 2 where the model with d=1 was superior to the others. The model selection process evaluated various ARIMA(p,1,q) models. The criterion for selecting the best model was the minimum AICc value, which balances model fit and complexity. To model the time series behavior of the Aggregated Cum.Log.PU data, ARIMA models were explored with different orders of regular differencing (d=0, d=1, and d=2), with model selection guided by minimizing the AICc criterion. The initial exploration involved fitting ARIMA models with no differencing (d=0). The best model identified in this category based on AICc was ARIMA(5, 0, 4). This model, however, displayed significant issues, including a relatively high AICc of 319.39, indicating poorer performance compared to models with differencing. Furthermore, many of its parameters, such as AR1 (P=0.386) and MA1 (P=0.608), were not statistically significant, suggesting potential overfitting. The residual diagnostics for this model showed statistically significant remaining autocorrelation at lags 12 and 24 (P=0.026 and P=0.015 respectively), indicating that the model did not fully capture the time series dependencies. Forecasts from this model were described as erratic with wide confidence intervals.

Next, ARIMA models with first-order differencing (d=1) were considered. Differencing is typically necessary for time series exhibiting a trend, as observed in the Aggregated Cum.Log.PU data, to achieve stationarity (Herrera-González et al., 2024; Box et al., 2015). The best model with d=1, according to the minimum AICc criterion, was ARIMA(2, 1, 2). This model achieved the lowest AICc among the three primary candidates, at 256.91. All its parameters, including AR1 (P=0.000), AR2 (P=0.000), MA1 (P=0.007), MA2 (P=0.000), and the Constant (P=0.000), were found to be statistically significant, suggesting a well-specified model. The constant term's significance indicates a deterministic trend in the original series. Crucially, the residual diagnostics for the ARIMA(2, 1, 2) model were favorable, with Ljung-Box test P-values well above 0.05 for all tested lags (e.g., Lag 12: 0.317, Lag 24: 0.455), indicating that the residuals are effectively white noise (Ljung and Box, 1978). This confirms the model's success in capturing the underlying time series structure after differencing. Forecasts from this model showed a gradual increase with relatively tight confidence intervals. Finally, ARIMA models with second-order differencing (d=2) were analyzed. The best model in this category was ARIMA(0, 2, 1), with an AICc of 258.98. While this AICc is close to that of the d=1 model, it is slightly higher. The ARIMA(0, 2, 1) model includes a single significant parameter, MA1 (P=0.000). Similar to the d=1 model, the residual diagnostics for the ARIMA(0, 2, 1) model were strong, with Ljung-Box test P-values all significantly above 0.05 (e.g., Lag 12: 0.883, Lag 24: 0.545), confirming the removal of residual autocorrelation. Forecasts from this model projected a steady linear growth.

Based on the comprehensive comparison, the ARIMA(2, 1, 2) with d=1 is the most robust choice for modeling the Aggregated Cum.Log.PU time series. This selection is primarily driven by its lowest AICc, which signifies the best balance between model fit and parsimony. Furthermore, all its parameters are statistically significant, and its residuals conform well to the white noise assumption, as demonstrated by the Ljung-Box test. While the distribution analysis showed that the raw data is not normal and benefits from a Johnson transformation, the ARIMA framework effectively handles the non-stationarity through differencing (d=1), making explicit data transformation unnecessary for the ARIMA modeling itself, provided the residuals are white noise. The forecasts generated by the d=1 model align plausibly with the observed historical trend, showing moderate continued growth with manageable uncertainty (*Figure 5(a), Figure 5(b) and Figure 5(c)*. The d=2 model, while also producing white noise residuals, had a slightly higher AICc and its linear forecast behavior might be less nuanced compared to the d=1 model's projection which is influenced by both AR and MA components of the differenced series. Applying ARIMA on the raw untransformed datasets was excluded because the initial diagnostic outputs were obviously inferior to the logarithmically transformed results. A comparative analysis of ARIMA models was conducted by examining the best-fitting model for each differencing order (d=0, d=1, and d=2) based on the criterion of minimum AICc. The results of this comparison are summarized in *Table 1*.

***Table 1**. Comparison of best ARIMA models at different differencing orders.*

Model*	ARIMA (p,d,q)	AICc	Significant Parameters (p < 0.05)	Residual Diagnostics (Ljung-Box p-value)	Forecast Behavior
d=0	(5,0,4)	319.39	AR3 (p=0.006); others insignificant (e.g., AR1 p=0.386)	p=0.026 (Lag 12) → Residual autocorrelation	Erratic, wide CIs (e.g., 142.47 ± 3.41 at Week 72)
d=1 (Best)	(2,1,2)	256.91	AR1=0.25, AR2=-1.04, MA1=0.29, MA2=-0.97, Constant=3.60	p=0.317 (Lag 12) → White noise	Stable growth (142.83 → 161.40 over 10 weeks)
d=2	(0,2,1)	258.98	MA1=0.95 (p=0.000)	p=0.883 (Lag 12) → White noise	Steeper linear trend (141.67 → 162.30)

*Note: *Key observations from this comparison highlight that the ARIMA(2,1,2) model with d=1 is optimal. It exhibits the lowest AICc (256.91) among the compared models. This model features statistically significant parameters for its AR and MA terms, as well as a significant constant term (3.60, p=0.000), which indicates the presence of a deterministic trend in the original series, aligning plausibly with the cumulative nature of the data. Furthermore, the* *residual diagnostics for the ARIMA(2,1,2) model strongly support the white noise assumption, with Ljung-Box test p-values exceeding 0.3 for tested lags. In contrast, the ARIMA(5,0,4) model (d=0) had a considerably higher AICc (319.39) and failed residual checks due to remaining autocorrelation. The ARIMA(0,2,1) model (d=2), while achieving white noise residuals, had a slightly higher AICc (258.98) compared to the d=1 model. The d=2 model, lacking autoregressive terms, projects a simpler linear trend.*

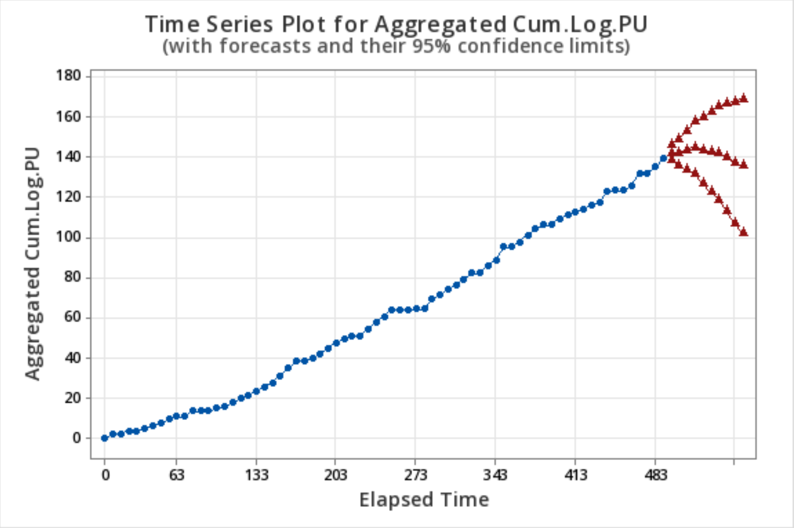

Figure 5(a). *Forecast of microbiological trend in PU at d=0.*

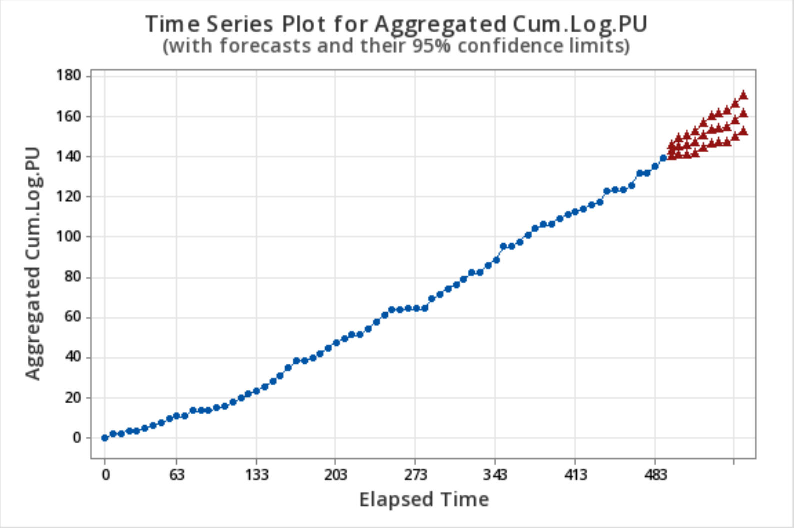

Figure 5(b). *Forecast of microbiological trend in PU at d=1.*

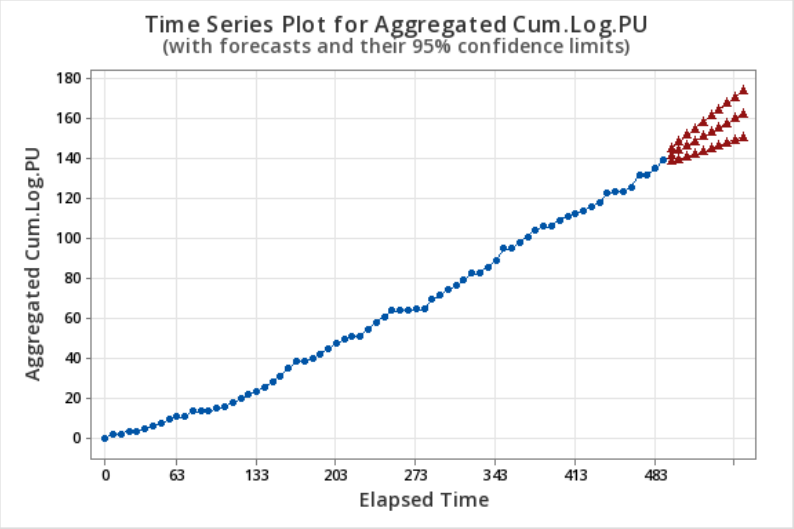

Figure 5(b). *Forecast of microbiological trend in PU at d=2.*

Limitations and Ethical Considerations: This analysis is subject to certain limitations. The models were selected based on a specific set of ARIMA (p, d, q) orders explored; other model structures or seasonal components, if present, were not explicitly investigated and could potentially provide a better fit. The slight deviation of the normality of the residuals, while not invalidating the model for point forecasting, may impact the accuracy of the computed confidence intervals which assume normality (Chatfield and Xing, 2019). The Johnson transformation, while effective for achieving normality, is complex and less directly interpretable in terms of the original scale compared to simpler transformations (Sakia, 1992). The forecasts are based on historical patterns and assume that the underlying process generating the data remains consistent; significant changes in water treatment protocols or environmental factors could affect future accuracy. The application of this analysis in healthcare facility water safety necessitates careful ethical consideration. Relying on model forecasts for critical decisions, such as adjusting testing frequency or scheduling maintenance requires a high degree of confidence in the model's accuracy and robustness. Over-reliance could lead to delayed interventions if the model underestimates risk, potentially compromising patient safety (WHO, 2022). It is ethically imperative to ensure that model performance is regularly monitored and validated against actual measurements. Transparency in communicating model limitations and uncertainties, particularly regarding forecast confidence intervals, is crucial. Furthermore, access to and interpretation of the model's outputs should be managed responsibly to avoid misinformed decisions. The potential to optimize resources must not outweigh the primary ethical responsibility to maintain the highest standards of water quality and patient safety.

Conclusion

The ARIMA(2, 1, 2) model provides a statistically sound and well-fitting representation of the Aggregated Cum.Log.PU time series. Its performance characteristics, including the lowest AICc and favorable residual properties, make it the recommended model for forecasting future values of this series. This model is suitable for generating reliable near-term predictions. It is recommended to adopt the ARIMA(2,1,2) model for future forecasting applications. To ensure ongoing model validity and reliability, residuals should be monitored periodically, perhaps quarterly, to detect any shifts or patterns that might indicate changes in the underlying process, such as modifications in water treatment protocols. While the Johnson-transformed data demonstrated good normality, cross-validating critical secondary analyses using the transformed data can be beneficial if assumptions of normality are paramount for those specific analyses. This integrated approach provides a solid foundation for data-driven decision-making, contributing to the proactive maintenance of water safety in healthcare facilities, with a direct positive impact on patient health and regulatory compliance. Future techniques such as the last observation carried forward for cumulative data treatment could be investigated for transforming irregular cumulative measurements into a regular time series suitable for modeling.

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] Alsheheri, G. (2025): Comparative Analysis of ARIMA and NNAR Models for Time Series Forecasting. – Journal of Applied Mathematics and Physics 13(1): 267-280.

[2] Anhøj, J., Olesen, A.V. (2014): Run charts revisited: a simulation study of run chart rules for detection of non-random variation in health care processes. – PLoS One 9(11): 13p.

[3] Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M. (2015): Time series analysis: forecasting and control. – John Wiley & Sons 720p.

[4] Chatfield, C., Xing, H. (2019): The analysis of time series: an introduction with R. – Chapman and hall/CRC 414p.

[5] Chidiac, S., El Najjar, P., Ouaini, N., El Rayess, Y., El Azzi, D. (2023): A comprehensive review of water quality indices (WQIs): history, models, attempts and perspectives. – Reviews in Environmental Science and Bio/Technology 22(2): 349-395.

[6] Dishan, A., Barel, M., Hizlisoy, S., Arslan, R.S., Hizlisoy, H., Gundog, D.A., Al, S., Gonulalan, Z. (2024): The ARIMA model approach for the biofilm-forming capacity prediction of Listeria monocytogenes recovered from carcasses. – BMC Veterinary Research 20(1): 13p.

[7] Durbin, J., Watson, G.S. (1992): Testing for serial correlation in least squares regression. – In Breakthroughs in Statistics: Methodology and Distribution, New York, NY: Springer New York 22p.

[8] Eissa, M. (2024): Bioburden analysis and microbiological stability of municipal distribution system through examination of transformed total microbial count dataset. – Frontiers in Scientific Research and Technology 8(1): 14p.

[9] Eissa, M., Rashed, E., Eissa, D. (2022): Principal component analysis in long term assessment of total viable plate count of municipal water distribution network system in healthcare facility. – Environmental Research and Technology 5(2): 165-171.

[10] Eissa, M.E., Rashed, E.R., Eissa, D.E. (2023): Case of preferential selection of attribute over variable control charts in trend analysis of microbiological count in water. – Acta Natura et Scientia 4(1): 1-9.

[11] FasterCapital Web Portal (2025): Differencing: Bridging the Gap: The Power of Differencing in Time Series Analysis. – FasterCapital Web Portal 27p.

[12] Feldl, M., Olayo-Alarcon, R., Amstalden, M.K., Zannoni, A., Peschel, S., Sharma, C.M., Brochado, A.R., Müller, C.L. (2025): Statistical end-to-end analysis of large-scale microbial growth data with DGrowthR. – BioRxiv 18p.

[13] Hilbe, J.M. (2011): Negative binomial regression. – Cambridge University Press 553p.

[14] Herrera-González, J. L., Rodríguez-Venegas, R., Legarreta-González, M. A., Robles-Trillo, P. A., De-Santiago-Miramontes, Á., Loya-González, D., & Rodríguez-Martínez, R. (2024). Time series (ARIMA) as a tool to predict the temperature-humidity index in the dairy region of the northern desert of Mexico. – PeerJ 12: 13p.

[15] Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D. (2008): Forecasting with exponential smoothing: the state space approach. – Springer Science & Business Media 360p.

[16] Johnson, N.L. (1949): Systems of Frequency Curves Generated by Methods of Translation. – Biometrika 36(1/2): 149-176.

[17] Kanamori, H., Weber, D.J., Rutala, W.A. (2016): Healthcare outbreaks associated with a water reservoir and infection prevention strategies. – Clinical Infectious Diseases 62(11): 1423-1435.

[18] Kirkwood, B.R., Sterne, J.A. (2010): Essential medical statistics. – John Wiley & Sons 516p.

[19] Kmentt, L., Cronk, R., Tidwell, J.B., Rogers, E. (2021): Water, sanitation, and hygiene (WASH) in healthcare facilities of 14 low-and middle-income countries: to what extent is WASH implemented and what are the ‘drivers’ of improvement in their service levels? – H2Open Journal 4(1): 129-137.

[20] Lemenkova, P. (2019): Generic Mapping Tools and Matplotlib Package of Python for Geospatial Data Analysis in Marine Geology. – International Journal of Environment and Geoinformatics 6(3): 225-237.

[21] Ljung, G.M., Box, G.E.P. (1978): On a Measure of Lack of Fit in Time Series Models. – Biometrika 65(2): 297-303.

[22] Lokanan, M. (2024): Harnessing Exploratory Data Analysis for Robust Financial Fraud Detection and Model Enhancement. – SSRN Electronic Journal 49p.

[23] Mazerolle, M.J. (2023): Package ‘AICcmodavg’: Model Selection and Multimodel Inference Based on (Q)AIC(c). – CRAN 229p.

[24] Lange, H., Hauhs, M. (2025): Complexity Analysis of Environmental Time Series. – Entropy 27(4): 31p.

[25] Portet, S. (2020): A primer on model selection using the Akaike Information Criterion. – Infectious Disease Modelling 5: 111-128.

[26] Razali, N.M., Wah, Y.B. (2011): Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. – Journal of Statistical Modeling and Analytics 2(1): 21-33.

[27] Sakia, R.M. (1992): The Box-Cox transformation technique: a review. – Journal of the Royal Statistical Society: Series D (The Statistician) 41(2): 169-178.

[28] Wang, S., Xu, O. (2025): Confidence interval forecasting model of small watershed flood based on compound recurrent neural networks and Bayesian. – PloS One 20(4): 24p.

[29] World Health Organization (WHO) (2022): Guidelines for drinking-water quality: incorporating the first and second addenda. – World Health Organization 614p.

[30] Xue, K. (2024): Prediction and Analysis of Total Nitrogen in a Sewage Treatment Plant Effluent. – Open Journal of Modelling and Simulation 12(4): 114-129.

[31] Yetiş, Ö., Ali, S., Karia, K., Bassett, P., Wilson, P. (2023): Enhanced monitoring of healthcare shower water in augmented and non-augmented care wards showing persistence of Pseudomonas aeruginosa despite remediation work. – Journal of Medical Microbiology 72(5): 12p.

[32] Zafra-Mejía, C.A., Rondón-Quintana, H.A., Urazán-Bonells, C.F. (2024): ARIMA and TFARIMA Analysis of the Main Water Quality Parameters in the Initial Components of a Megacity’s Drinking Water Supply System. – Hydrology 11(1): 19p.