German J Pharm Biomaterials 2025; 4(4): 26-32
https://doi.org/10.5530/gjpb.2025.4.12

GJPB

Short Communication

Modeling microbial density trend in pharmaceutical water with irregular intervals using CART regression

Mostafa Essam Ahmed Mostafa Eissa

Department of Pharmaceutical SPC, Pharmaceutical and Medicinal Research Facility, Cairo, Egypt.

*Correspondence: mostafaessameissa@yahoo.com

Received: 14 May 2025; Revised: 21 July 2025; Accepted: 27 August 2025

Abstract

Monitoring and controlling the microbiological quality of water in the pharmaceutical and biopharmaceutical industries is paramount to ensure the safety and quality of intermediate and final medicinal products. An integral task of the quality system is to trend, interpret, and investigate this crucial inspection characteristic. Analyzing time series data with irregular sampling intervals presents a significant challenge, particularly when standard methods like ARIMA, which assume fixed-frequency observations, are desired but achieving consistent sample spacing is unattainable. This study investigated the applicability of Classification and Regression Tree (CART) regression as a practical alternative, using Minitab® Statistical Software, to model microbial density collected at irregular intervals. Three response variables were studied with respect to Elapsed Time: a sequential counter, cumulative untransformed microbial density, and cumulative log-transformed microbial density. CART could not model the sequential counter but succeeded with both cumulative variables. The log-transformed cumulative model achieved a test R-squared of 97.88% and a Mean Absolute Percent Error (MAPE) of 0.1610%. The cumulative untransformed model also performed well, with a test R-squared of 95.40% and MAPE of 0.5477 %. The transformed data yielded slightly better results. These findings show that CART regression with Elapsed Time robustly models cumulative trends in irregularly sampled data and is a valuable alternative when fixed-frequency model assumptions cannot be met.

Keywords: RMSE; CART regression; Microbial density; MAPE; Pharmaceutical water; MAD

Introduction

The provision of high-quality purified water is an indispensable component of healthcare products within healthcare facilities [1,2]. Purified water is widely used in critical applications such as mixing and compounding medications, sterilization processes, and preparation procedures [3,4]. Even low levels of microbial contamination in this water can compromise drug safety and lead to serious quality-related issues, particularly in vulnerable and immunocompromised individuals [4,5]. Such infections can result in prolonged hospital stays, increased healthcare costs, significant morbidity, and even mortality [5,6]. Consequently, stringent microbiological control and continuous monitoring of purified water systems are paramount to ensure the safety of medicinal products and to maintain compliance with regulatory standards [6,7]. Thus, the ability to comprehend and analyze microbial water trends is critical to proactive measures before catastrophic events occur. Time series analysis – as one of the important approaches - is fundamental to understanding processes that evolve over time, including environmental monitoring data such as microbial density [8]. A prevalent assumption in many traditional time series models, including AutoRegressive Integrated Moving Average (ARIMA) models, is that observations occur at fixed, equally spaced intervals [9]. However, real-world data collection often results in irregular sampling intervals due to logistical constraints, resource availability, or event-driven monitoring [10]. Applying standard time series modeling methods to such data typically requires resampling or interpolation, which can be complex and may introduce artifacts or information loss, particularly without access to specialized programming resources [11-13]. This study explores an

alternative modeling strategy using Classification and Regression Tree (CART) regression to analyze microbial density data recorded at irregular time points, leveraging readily available statistical software (Minitab® Statistical Software).

Materials and Methods

Microbiological bioburden analysis results for the newly installed water purification small station were extracted from the database center of the healthcare facility in Giza governorate, Egypt (n=68), covering January 2023 to May 2024 [14]. The testing frequency was estimated to be approximately days, with a standard deviation of 7.3 ± 2.26 days. The dataset comprised 68 chronologically arranged microbial population density measurements, recorded at irregular dates [15,16] and obtained after sampling and testing using standard microbiological quality control procedures as provided in previous works. From this, "Elapsed Time" was calculated as the cumulative days from the first observation. Three different forms of the data were considered as potential response variables for modeling (the numbering will be used to refer to these three types of datasets in subsequent texts):

- **Raw data (simple counter):** A sequential counter (0, 1, 2, ...) representing raw microbiological count data as Colony Forming Unit (CFU)/100 ml in the order of observation.
- **Cumulative untransformed data:** The cumulative sum of the previous original microbial density measurements.
- **Cumulative log-transformed data:** The cumulative sum of microbial density after applying logarithmic transformation to the base ten.

CART regression was employed using "Elapsed Time" as the sole predictor variable. CART is a non-parametric technique that recursively partitions the data space based on predictor variables to predict a continuous response [12]. This method can model nonlinear relationships and does not require the predictor variable to be equally spaced. All analyses were performed using Minitab® statistical software, utilizing 10-fold cross-validation to assess model performance on unseen data. Model fit was evaluated using R-squared, Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Mean Absolute Deviation (MAD), and Mean Absolute Percent Error (MAPE) [13]. Residual plots and fits vs. actual plots were examined for diagnostic purposes.

Results and Discussion

Initial attempts to apply CART regression using the **Raw Data (Simple Counter)** as the response variable against "Elapsed Time" resulted in a model failure, indicated by the error message "* ERROR * The optimal tree is the root node. No analysis was performed." This outcome suggested that the simple sequential nature of the counter variable did not present a structure that CART could effectively exploit through splits on Elapsed Time to reduce prediction error [9,12]. Subsequently, CART regression was applied using the cumulative data as the response variables [17,18]. Figure 1 shows the time series plot for the three types of data in separate panels for clarity due to different y-axis scales (CFU) versus fixed elapsed time in days.

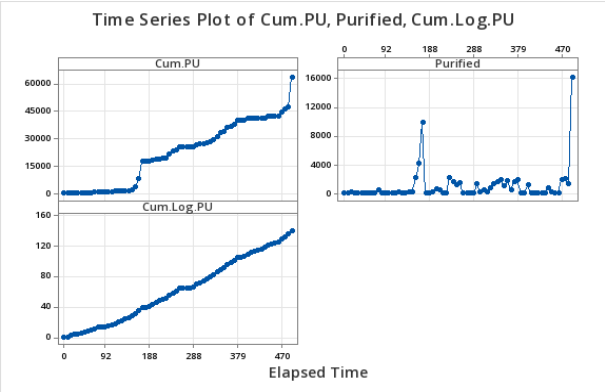

Figure 1. Time series graph showing three levels of data processing (1, 2, 3) for microbiological density in purified water showing that the last phase showed greatest level of linearity.

When using the Cumulative Untransformed Data as the response variable, the CART model fit the data well. The optimal tree had 7 terminal nodes with a minimum of 3 (Figure 2). The model demonstrated high accuracy, achieving a training R-squared of 98.32%, a training RMSE of 2202.68, a training MSE of 4.852e+06, a training MAD of 1270.46, a training MAPE of 0.5233%, and a test R-squared of 95.40%.

Absolute error metrics on the test set were substantial (Test RMSE: 3645.55, Test MSE: 1.329e+07, Test MAD: 1877.56), reflecting the large scale of the cumulative untransformed data. However, the MAPE on the test set was a low 0.5477%, indicating high relative accuracy. The Scatterplot of Response Fits vs Actual Values (Figure 3) showed that the fitted values closely tracked the actual values, with characteristic horizontal segments in the training data. The Boxplot of Residuals (Figure 4) showed residuals centered around zero with a large absolute spread and visible outliers. The MSE vs Terminal Node plot (Figure 5) and Residual Plot by Terminal Node (Figure 6) confirmed varying absolute error levels across the segments defined by the tree in Figure 8. Figure 7 visually summarizes the outcome of a CART regression analysis applied to the cumulative purified water count data.

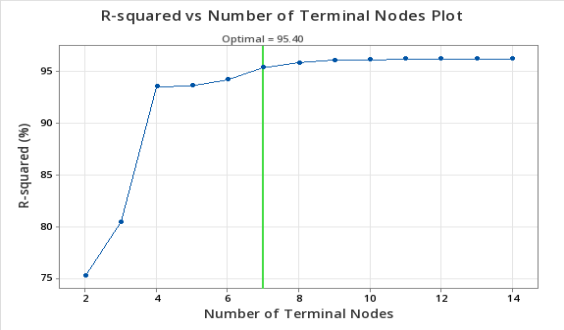

Figure 2. Test R-squared versus number of terminal nodes for the CART regression model (cumulative untransformed data).

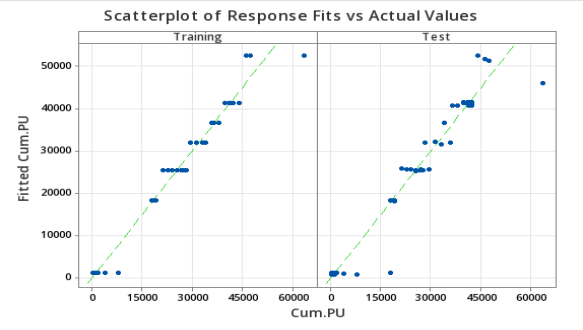

Figure 3. Scatterplot of fitted versus actual values for the CART regression model (cumulative untransformed data).

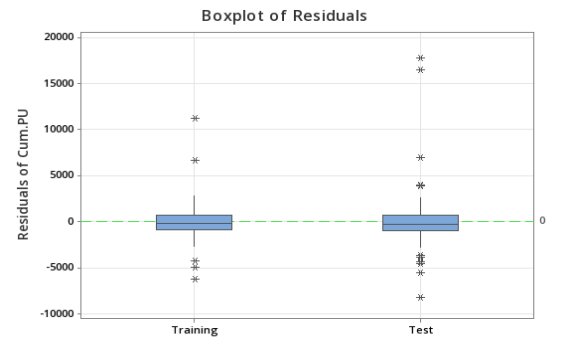

Figure 4. Boxplot of residuals for the CART regression model (cumulative untransformed data).

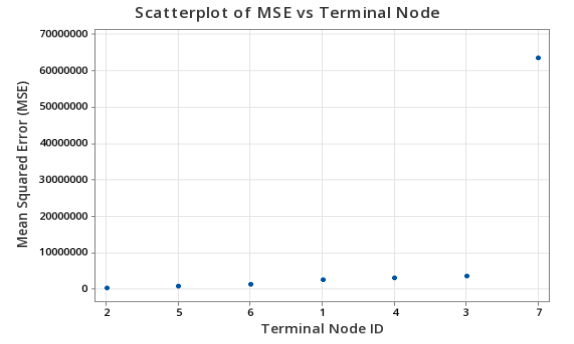

Figure 5. Scatterplot of mean squared error versus Terminal Node ID for the CART Regression Model (Cumulative Untransformed Data).

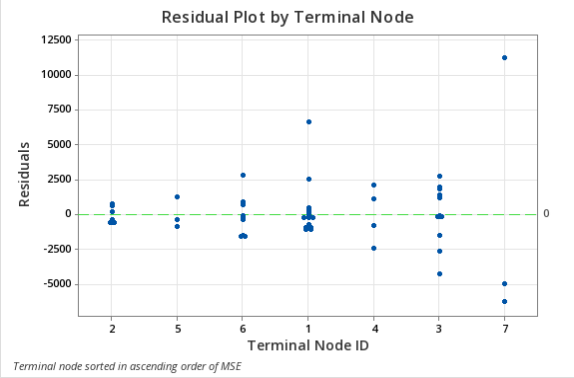

Figure 6. Residual Plot by terminal node for the CART regression model (cumulative untransformed data).

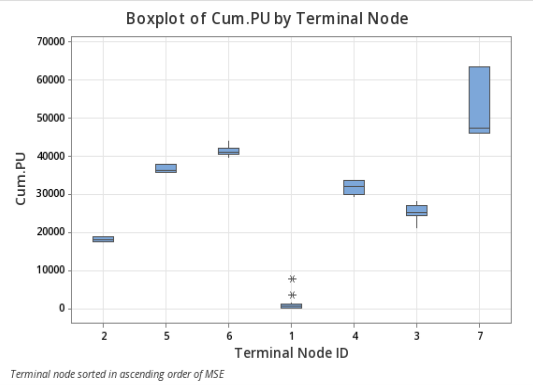

Figure 7. Boxplot of actual values by terminal node for the CART Regression Model (Cumulative Untransformed Data).

When using the Cumulative Log-Transformed Data as the response, the CART model also fit the data, yielding an optimal tree with 9 terminal nodes (Figure 9). This model exhibited slightly better performance metrics. The training R-squared was 98.90%, and the test R-squared was 97.88%. Absolute error metrics were significantly smaller (Test RMSE: 6.18, Test MSE: 38.13, Test MAD: 5.34), commensurate with the compressed scale of the log-transformed response. Crucially, the MAPE on the

test set was very low, 0.1610%. The Scatterplot of Response Fits vs Actual Values (Figure 10) showed an excellent visual fit with points tightly clustered around the diagonal line. The Boxplot of Residuals (Figure 11) showed residuals centered around zero with a smaller absolute spread and no apparent outliers compared to the untransformed model. The MSE vs Terminal Node plot (Figure 12) and the Residual Plot by Terminal Node (Figure 13) indicated varying absolute errors across segments, but on a much smaller scale than in the non-transformed data. Relative Variable Importance confirmed Elapsed Time as the sole predictor (100% importance). The Boxplot of Cum.Log.PU by Terminal Node (Figure 14) showed the distribution of actual values within each segment. The Optimal Tree Diagram (Figure 15) visually represented the splits based on Elapsed Time and the predicted mean log-transformed cumulative values within each terminal node.

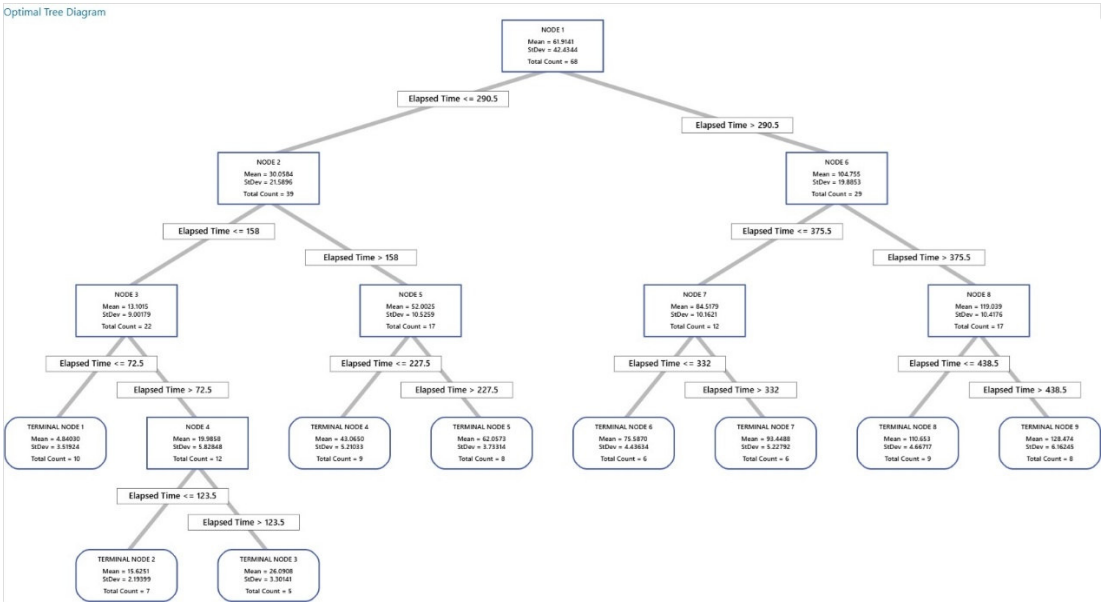

Figure 8. Optimal Tree Diagram for the CART Regression Model (Cumulative Untransformed Data).

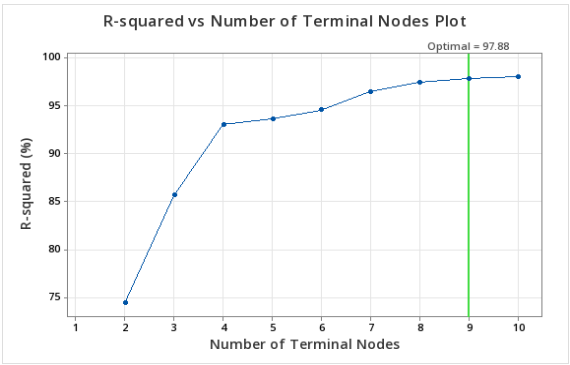

Figure 9. Test R-squared versus number of terminal nodes for the CART regression model (cumulative log-transformed data).

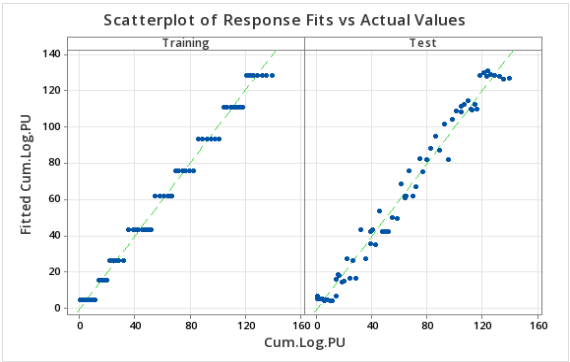

Figure 10. Scatterplot of fitted versus actual values for the CART regression model (cumulative log-transformed data).

The results demonstrate that while CART regression was not applicable to the simple sequential counter data, likely due to the lack of a predictable structure with elapsed time amenable to piecewise constant modeling, it proved highly effective for modeling both the cumulative untransformed and cumulative log-transformed microbial density [12,19-22]. This highlights a crucial point: the success of a modeling technique is dependent not only on the method itself but also on the form of the data being modeled and its inherent relationship with the predictors.

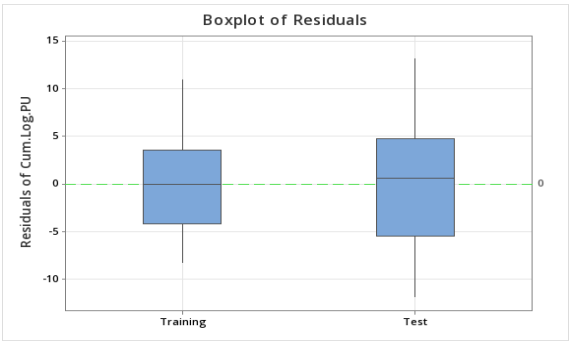

Figure 11. Boxplot of residuals for the CART regression model (cumulative log-transformed data).

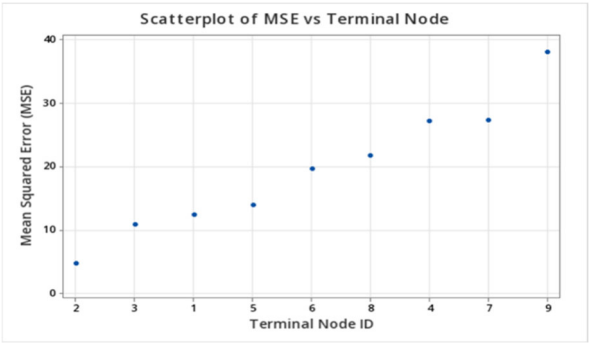

Figure 12. Scatterplot of mean squared error versus terminal node ID for the CART regression model (cumulative log-transformed data).

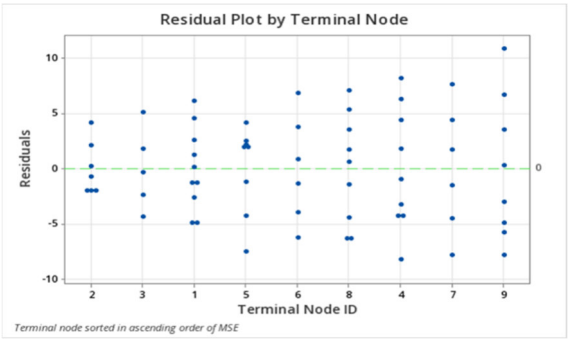

Figure 13. Residual plot by terminal node for the CART regression model (cumulative log-transformed data).

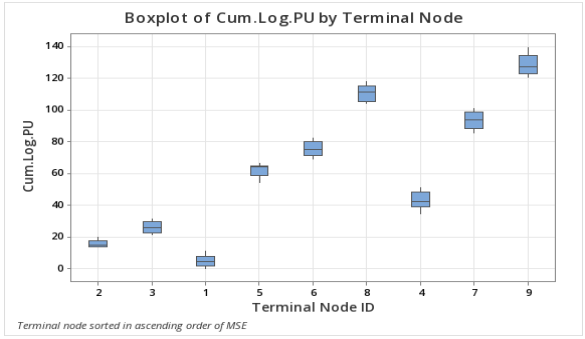

Figure 14. Boxplot of actual values by terminal node for the CART Regression Model (Cumulative Log-Transformed Data).

The primary obstacle posed by the irregular sampling intervals for standard ARIMA was effectively bypassed by using "Elapsed Time" as a continuous predictor in the CART regression framework. This approach allowed the direct use of irregularly spaced observations without the need for complex, potentially impractical data transformation steps within the available software.

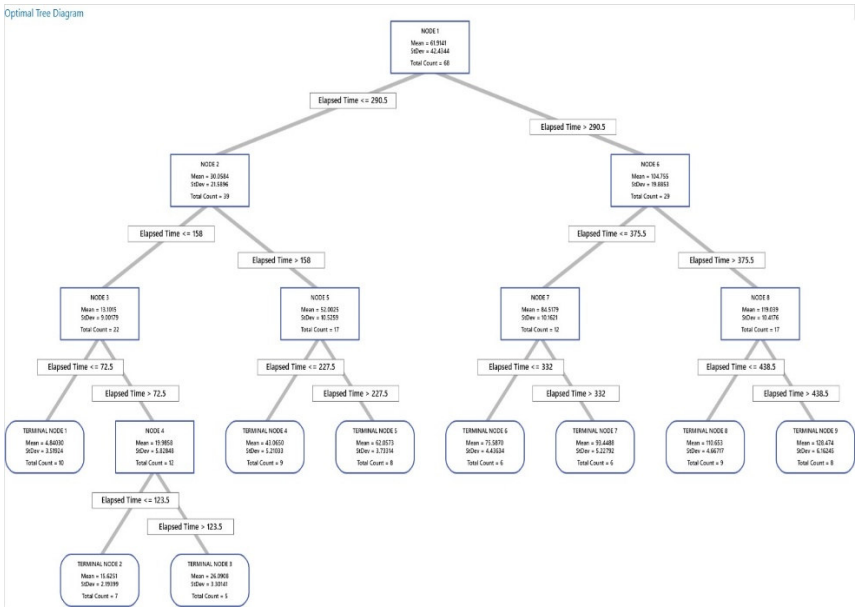

Figure 15. Optimal tree diagram for the CART regression model (cumulative log-transformed data).

Comparing the two successful models, the CART regression predicting the cumulative log-transformed data ("Cum.Log.PU") exhibited slightly superior test R-squared and significantly lower MAPE [21,22]. This suggests that the log transformation likely helped stabilize the variance or linearize

the relationship between cumulative density and elapsed time, making the pattern more readily captured by the CART algorithm's splitting process, as evident in Figure 1 [23-25]. While the untransformed model provided predictions in the original units, which can be more intuitively interpretable, the log-transformed model offered higher relative predictive accuracy.

The interpretation of the CART models provides valuable insights into the relationship between microbial density accumulation and elapsed time. The tree structures (Figures 8 and 15) reveal specific thresholds in Elapsed Time where the predicted cumulative density changes, effectively approximating the non-linear trend observed in the data (Figure 1) [12,26]. The residual analysis and fits vs. actual plots confirmed the high accuracy and lack of major systematic errors within the segments (Figures 2-7 and 9-14) defined by the trees (Figures 8 and 15) [27,28]. The MSE vs Terminal Node plots were further detailed, where the model's absolute errors were higher or lower across the time segments [18]. Thresholds like those in optimal tree diagrams align with USP <1231> action limits for microbial contamination, enabling proactive system sanitization [6]. Moreover, future work could integrate CART with Internet of Things (IoT) sensors for real-time alerts when elapsed time approaches critical thresholds. This study provides a practical, robust solution for modeling cumulative trends in irregularly sampled data using readily available commercial statistical software. While other time-series methods for irregular data (e.g., state-space models) could be explored, their implementation often requires specialized programming environments [9]. Recent advances in machine learning, such as Random Forests and Generalized Additive Models (GAMs), have also addressed challenges with irregular data, but their complexity often limits adoption in the pharmaceutical industry [23].

Conclusion

CART regression, using elapsed time as a predictor, provides a highly effective and practical solution for modeling cumulative trends with irregularly sampled microbial density trends. While CART failed on the raw sequential counter data, it successfully and accurately modeled both cumulative untransformed and cumulative log-transformed microbial density. The log-transformed cumulative data yielded a model with superior accuracy in overall predictive performance. The log-transformed cumulative data yielded a model with superior accuracy in overall predictive performance. This method provides a practical solution for pharmaceutical water monitoring, circumventing the limitations of traditional time-series approaches. Future work could explore hybrid models or additional predictors to further enhance predictive capability. Hybrid AI/ML frameworks that combine CART's interpretability with deep learning's pattern recognition could further enhance predictive accuracy.

Declaration of interest

The author declare no conflict of interest.

Financial support

This work has not received any funds from national and international agencies.

References

1. Eissa ME. Microbiological quality of purified water assessment using two different trending approaches: A case study. Sumerianz Journal of Scientific Research. 2018;1(3):75-9.
2. Eissa M. Evaluation of microbiological purified water trend using two types of control chart. Eur Pharm Rev. 2018;23(5):36-8.
3. World Health Organization. Annex 2: WHO good manufacturing practices: water for pharmaceutical use. In: WHO Expert Committee on Specifications for Pharmaceutical Preparations. Forty-sixth report. Geneva: World Health Organization; 2012. p. 77-126. (WHO Technical Report Series, No. 970).
4. EISSA M. Enhancing microbiological stability in municipal water distribution: A descriptive statistical analysis for public health assurance. Journal of Biometry Studies. 2024;4(1):11-30.
5. Ferraz MP. Antimicrobial Resistance: The Impact from and on Society According to One Health Approach. Societies. 2024;14(9):187.

6. United States Pharmacopeial Convention. Chapter Water for Pharmaceutical Purposes. In: United States Pharmacopeia and National Formulary (USP-NF). Rockville (MD): United States Pharmacopeial Convention; 2024.
7. Sia CH, Hong SD, Teo J, Chan R, Lai W, Chan. Contamination Trends & Proposed Solutions. Pharm Eng. 2023 Mar-Apr;43(2):50-61. Available from: https://ispe.org/pharmaceutical-engineering/march-april-2023/contamination-trends-proposed-solutions (Accessed: May 14, 2025).
8. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken (NJ): John Wiley & Sons Inc; 2015.
9. Shumway RH, Stoffer DS. Time Series Analysis and Its Applications: With R Examples. 4th ed. New York (NY): Springer; 2017.
10. Broersen PM, Bos R. Estimating time-series models from irregularly spaced data. IEEE transactions on instrumentation and measurement. 2006;55(4):1124-31.
11. Rehfeld K, Marwan N, Heitzig J, Kurths J. Comparison of correlation analysis techniques for irregularly sampled time series. Clim Past. 2011;7(5):1545-68.
12. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. New York (NY): Chapman and Hall/CRC; 2017.
13. Minitab LLC. Minitab® Statistical Software [computer program]. Version 21. State College (PA): Minitab LLC; 2022.
14. Eissa ME. Determination of the microbiological quality of feed city water to pharmaceutical facility: distribution study and statistical analysis. Athens J Sci. 2017;4(2):143-60.
15. Eissa ME, Rashed ER, Eissa DE. Case of preferential selection of attribute over variable control charts in trend analysis of microbiological count in water. Acta Nat Sci. 2023 Feb 14;4(1):1-9.
16. Eissa M, Rashed E, Eissa D. 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. 2022;5(2):165-71.
17. Minitab. Tree diagram for CART® Regression - Support. [cited 2025 May 14]. Available from: https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/predictive-analytics/how-to/cart-regression/interpret-the-results/all-statistics-and-graphs/tree-diagram/
18. Minitab. Interpret the key results for CART® Regression - Support. [cited 2025 May 14]. Available from: https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/predictive-analytics/how-to/cart-regression/interpret-the-results/key-results/
19. Lewis RJ. An introduction to classification and regression tree (CART) analysis. Proc (Bayl Univ Med Cent). 2000 Oct;13(4):415-6.
20. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell. 2005 Jun 1;27(2):83-5.
21. Atkinson AC. Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford University Press; 1985.
22. Cho Y, Molinaro AM, Hu C, Strawderman RL. Regression trees and ensembles for cumulative incidence functions. Int J Biostat. 2022;18(2):397-419.
23. Feng X, Shi X, Ho HC, Xu Y, Gao J, Li Y, et al. Regression trees and ensembles for cumulative incidence functions. BMC Med Res Methodol. 2022 Sep 21;22(1):244.
24. John Lu ZQ. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. J R Stat Soc: Series A (Statistics in Society). 2010 Jan 4;173(3):693–4.
25. Ruppert D. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. J Am Stat Assoc. 2004 Jun;99(466):567–7.
26. De'Ath G, Fabricius KE. Classification and regression trees: a powerful tool for ecological analysis. Ecology. 2000 Sep;81(11):3178-3189.
27. Draper NR, Smith H. Applied Regression Analysis. 3rd ed. Wiley-Interscience; 1998.
28. Kutner MH, Nachtsheim CJ, Neter J. Applied Linear Regression Models. 4th ed. McGraw-Hill/Irwin; 2004.

How to cite this article:
Mostafa Eissa MEA. Modeling microbial density trend in pharmaceutical water with irregular intervals using CART regression. German J Pharm Biomaterials. 2025;4(4):26-32.