Islamic Azad University
Shiraz Branch

Journal of Building Information Modeling
Summer 2025. Vol 1. Issue 1

https://sanad.iau.ir/journal/bim

Original Research Paper

A Data-Driven Framework for Operational Management of Pumping Stations Using Statistical Methods in Water Transmission Infrastructure

Farhad Pazhuheian*: Industrial engineering department, Iran university of science and technology, Tehran, Iran
Alireza Farahbakhsh: Senior manager in systems and methods, Persian Gulf Water Transfer Operation and Supply company, Bandar Abbas, Iran
Ali Liaghat: Department of Civil engineering, Islamic Azad university of Shiraz, Shiraz, Iran

ARTICLE INFO

Received:2025/03/12
Accepted:2025/06/15
PP: 43-54

Use your device to scan and read the article online

Keywords: *Pumping Station, Failure Count, Electricity Consumption, Spatial Regression, Generalized Additive Model*

Abstract

This research presents a data-driven framework that integrates spatial modeling techniques and nonlinear methods to analyze the performance of pumping stations in water transmission infrastructure. The study highlights the importance of considering spatial and temporal factors to enhance operational reliability and optimize resource allocation.Modeling the number of failures against electrical energy consumption in pumping stations enables better maintenance planning. By analyzing the relationship between energy use and failures, patterns can be identified to predict potential breakdowns and schedule preventive maintenance more effectively. This approach helps reduce unexpected downtime, lower costs, and improve system efficiency and equipment lifespan.By combining advanced statistical methods, such as spatial regression and generalized additive model, the study develops a comprehensive tool for predicting pumping station performance. A case study at the Pumping Station in Iran demonstrates how these techniques can help analyze the of pumping stations in water transmission.

Citation: Pazhuheian, F., Pazhuheian, A,R., & Liaghat, A. (2025). **A Data-Driven Framework for Operational Management of Pumping Stations Using Statistical Methods in Water Transmission Infrastructure**. *Journal of Building Information Modeling*, 1(1), 43-54.

* Corresponding author: Farhad Pazhuheian, Email: farhad.pzh@gmail.com

INTRODUCTION

The effective operation of pumping stations is crucial for ensuring the efficient and reliable functioning of water distribution networks. These stations are vital for maintaining the continuous movement of water across extensive distances, yet they face various operational challenges, including failures and variations in energy usage. Successfully predicting and managing these challenges is key to enhancing system efficiency, reducing downtime, and lowering operational expenses (Dadar *et al.*, 2021)

Pumping stations are often equipped with complex machinery and control systems that, when exposed to wear and tear, may lead to unexpected failures. Additionally, fluctuations in energy consumption due to changes in demand and operational conditions can further increase operational costs and reduce efficiency in water transmission systems (Luna et al., 2019).

These challenges not only impact the performance and reliability of the infrastructure but also result in higher operational costs and diminished service quality (Ikramov et al., 2020). Effective management of pumping stations requires continuous monitoring and predictive maintenance strategies to avoid costly breakdowns and extend the lifespan of equipment. To address these challenges, there is an increasing demand for data-driven approaches that can monitor, predict, and optimize the performance of these stations in real time (Yates et al., 2001). Spatial statistics and spatio-temporal modeling provide powerful tools for analyzing the complex relationships between various operational factors, such as equipment failure rates, energy consumption patterns, and environmental influences. These techniques have been widely applied in infrastructure management to enhance decision-making processes and improve system reliability (Blokus-Dziula et al., 2023). By integrating these advanced techniques, predictive models can be developed to identify potential issues before they arise, optimize energy usage, and ultimately improve the overall operational efficiency of pumping stations.

There is a notable correlation between the frequency of mechanical failures in pumping stations and their energy usage. Failures such as impeller wear, bearing degradation, and shaft misalignment tend to reduce hydraulic efficiency by increasing internal resistance. As a result, more electrical energy is required to maintain the same level of output, which leads to higher operational expenses and accelerated equipment aging.

A case study conducted at the Gigiri Pumping Station in Kenya highlighted how different operational configurations impact overall efficiency. The study showed that operating a single pump (Pump No. 4) yielded a significantly higher efficiency of 74%, compared to just 34% when Pumps 1 and 2 were run together. These differences were attributed to variations in maintenance status and mechanical condition (Tiony, 2013).

These findings underscore the importance of preventive maintenance, precise equipment selection, and real-time monitoring in minimizing breakdowns and optimizing energy consumption. Utilizing smart diagnostics and predictive tools allows for early detection of faults, which helps ensure reliable performance while minimizing unnecessary power usage.

Recent advancements in data analytics and spatio-temporal modeling have opened up new possibilities for improving the management of infrastructure systems. However, many conventional approaches still fail to account for the intricate spatial and temporal relationships that influence the performance of these systems. To bridge this gap, combining spatial clustering, spatial regression, and spatio-temporal modeling offers a promising strategy to improve the management and forecasting of pumping station operations (Kofinas et al, 2020) .

This research aims to develop and implement a data-driven analytical framework that integrates these techniques to analyze and predict the performance of pumping stations within a water transmission infrastructure. The study particularly focuses on exploring the connection between energy consumption and failure frequency at these stations, employing spatial regression to forecast future performance trends.

The key contributions of this paper include:

1. Presenting a new method that integrates spatial clustering, regression, and

generalized additive modeling to enhance the management of pumping stations.
2. Demonstrating how data-driven methods can be leveraged to gain a deeper understanding of the spatial and temporal behavior of system performance.
3. Offering a case study based on the water transmission pipeline, providing insights into the practical application of the proposed framework .

Through the integration of these advanced analytical techniques, this study aims to deliver a more precise and holistic understanding of the operational dynamics of pumping stations, ultimately aiding in more informed decision-making and improving infrastructure management.

Literature Review

The efficient management of pumping stations, which are vital components of water transmission infrastructure, has been a subject of significant research in the field of infrastructure management. Various studies have focused on understanding and improving the operational performance of pumping stations by employing advanced analytical and statistical techniques. This literature review explores the key methodologies used in analyzing infrastructure performance, with a specific focus on spatial regression, nonlinear modeling, and data-driven methods**.**

Spatial modelling in Infrastructure Management

Spatial regression techniques have been widely applied to model and analyze spatial dependencies in infrastructure systems. Anselin (1988) introduced the concept of spatial econometrics, emphasizing the importance of spatial autocorrelation in model specification. Later, researchers such as Getis and Ord (1992) extended these concepts by applying spatial regression models to infrastructure systems to account for spatial dependencies in variables like system failures, maintenance costs, and energy consumption. These models are particularly useful in identifying spatial patterns and correlations that may not be evident using traditional regression methods**.**

For instance, Bao & Chen (2017) used spatial econometrics to model water distribution systems and found significant spatial dependencies in the failure rates of different components **.**

As infrastructure systems are influenced by both spatial and temporal factors, spatio-temporal modelling has become an essential tool for analyzing dynamic performance over time. Cressie (2015) laid the foundation for spatio-temporal statistics by developing models that simultaneously account for both spatial and temporal correlations in environmental data. Du et al. (2023) further advanced these models by applying Gaussian Processes (GP) to infrastructure performance, showing that spatio-temporal models can significantly improve the prediction accuracy of system failures and energy consumption**.**

In the context of pumping stations, Qiu et al. (2024) demonstrated the application of Gaussian Process Regression (GPR) for predicting the energy consumption of pumping stations by integrating both spatial and temporal variables. This approach has shown considerable promise in enhancing the predictive capabilities of infrastructure management systems by accounting for the intricate relationships between time, location, and system performance**.**

Clustering Techniques in Infrastructure Analysis

Clustering methods, particularly K-means clustering, have been widely used to group infrastructure units based on their operational characteristics. Jain (2010) discussed the importance of clustering in identifying patterns in large datasets, which can aid in segmenting infrastructure systems into more manageable units for optimization purposes. In the context of pumping stations, clustering has been used to group stations with similar failure rates or energy consumption patterns, thereby facilitating targeted management strategies**.**

For example, Alyu et al. (2023) used clustering algorithms to identify groups of pumping stations with similar failure characteristics, enabling better resource allocation and maintenance scheduling. Similarly, Huo et al. (2020) applied clustering to water distribution networks to optimize energy usage and reduce operational costs**.**

Applications in Water Transmission Systems

The application of spatial and spatio-temporal analysis methods in water transmission systems

has gained traction in recent years. Christodoulou et al. (2012) applied spatial analysis to optimize the maintenance schedules of water pumps in a large distribution system, showing that spatially aware models can lead to more efficient operations. Mutambik (2024) employed spatio-temporal models to predict pipe failure rates in a water supply system, providing a framework for proactive maintenance and resource management**.**

Baerton et al. (2020) investigated the environmental factors influencing pipe failure in clean water networks using Generalized Additive Models (GAMs). GAMs were applied to model and analyze the effects of variables such as temperature, pressure, humidity, and soil quality on pipe failures. The study emphasized the importance of considering environmental impacts in the design and maintenance of water distribution systems, offering valuable insights for predicting future failures and improving the management of water systems**.**

Gaps and Contribution of This Study

While previous research has explored individual aspects of infrastructure performance using spatial regression, clustering, and nonlinear modelling (GAMs), few studies have integrated these methods into a unified framework for managing multiple performance indicators, such as failure count and energy consumption, across the same infrastructure system. This study aims to fill this gap by providing a data-driven analytical framework that integrates spatial clustering, spatial regression, and generalized additive model to predict and manage the performance of pumping stations in a water transmission system. By doing so, this study seeks to enhance the predictive accuracy and operational efficiency of pumping stations, contributing to more effective infrastructure management practices.

Research Methodology

This section describes the analytical methods used to investigate the relationship between electricity consumption and system failures in pumping stations. A combination of spatial analysis, time-series visualization, and nonlinear modeling techniques was employed to capture the complex interdependencies across both space and time.

Moran's I Test for Spatial Autocorrelation

To confirm the existence of spatial autocorrelation in the data, Moran's I statistic was calculated. This global spatial autocorrelation measure was applied separately to the variables of electricity consumption and failure count, using the same spatial weight matrix W as in the regression models.

Moran's I values closer to +1 indicate positive spatial autocorrelation (similar values cluster together), while values closer to –1 suggest negative spatial autocorrelation (dissimilar values are adjacent). The test helped determine whether the observed data exhibit significant spatial patterns, justifying the application of spatial regression (Moran, 1950; Cliff & Ord, 1981).

Cumulative Time-Series Analysis

For temporal trend analysis, cumulative plots of both electricity consumption and failure counts were generated for each pumping station over the course of March 2024 to March 2025. These visualizations enabled identification of long-term patterns, seasonal effects, and sudden surges in failures or electricity use.

Additionally, peak analysis was conducted to isolate time intervals with unusually high values, which may correspond to periods of stress or inefficiency in system operations (Chatfield, 2004).

Clustering of Pumping Stations

To identify groups of stations with similar operational characteristics, unsupervised clustering techniques were applied. In particular, K-means clustering was used to partition the pumping stations into distinct groups based on their electricity consumption and failure counts.

This step aimed to: Discover hidden patterns in station behavior, Facilitate targeted interventions, Improve maintenance strategies and resource allocation.

Spatial mapping of the resulting clusters also revealed potential regional performance trends or infrastructure disparities (Jain, 2010; MacQueen, 1967).

Spatial Regression Analysis

To analyze the spatial dependency between electricity consumption and failure count in

pumping stations, spatial regression models were utilized. The primary objective was to assess whether electricity consumption at a station is associated with failure occurrences in the same station or its neighboring stations.
Specifically, the Spatial Lag Model (SLM) was implemented. The SLM incorporates the influence of neighboring units through a spatially lagged dependent variable and is expressed as follows:

$$Y = \rho WY + X\beta + \varepsilon$$

Where:

Y is the dependent variable (failure count)

X is the matrix of explanatory variables (e.g., electricity consumption)

W is the spatial weight matrix representing spatial relationships among stations

ρ is the spatial autoregressive coefficient

ε is the error term.

The inclusion of the term WY allows the model to account for spatial spillover effects, where failures in one station may be influenced by conditions in nearby stations (Anselin, 1988; Elhorst, 2014).
To complement the analysis and correct for potential spatial autocorrelation in the residuals, the Spatial Error Model (SEM) was also applied. The SEM is suitable when unobserved spatial effects influence the dependent variable indirectly through the error term (Elhorst, 2014).

Generalized Additive Model

In the final stage of the analysis, a Generalized Additive Model (GAM) was employed to capture potential non-linear relationships between the number of failures and electricity consumption. The model was fitted using a Poisson distribution with a log link function, suitable for count data. A smooth term was applied to the electricity consumption variable to allow for flexible, data-driven estimation of its effect, while the categorical effects of month and station id were included as parametric terms. This approach allowed the potential nonlinear influence of energy use on failure counts to be identified without imposing a strict functional form. General form of the GAM is expressed as follow :

$$g(E[Y]) = \beta_0 + \sum_{j=1}^{p} f_j(X_j)$$

Where:

Y is the response variable

$E[Y]$ is the expected value of the response

$g(.)$ is the link function

$\beta_0$is the intercept

$f_j(.)$ is Smooth functions estimated from the data (e.g., splines), allowing for nonlinear relationships

P is the number of predictors

One of the advantages of GAM is its ability to model complex, non-linear effects of predictors, such as electricity consumption, while also accounting for other factors like month and station id (Wood, 2017).

A real word case study

This section presents the case study used in the present research, focusing on a water transmission line located in Kerman Province, Iran. The transmission line includes four pumping stations situated in the southeastern region of the province, responsible for the transportation and pumping of water.
A real-world dataset was collected, including the number of failures at each pumping station and the corresponding electricity consumption over a 12-month period. The geographic coordinates (longitude and latitude) of each station were used as spatial axes for geostatistical analysis.
"The data collection period spans from March 2024 to March 2025. All spatial and statistical analyses in this study were conducted using specialized R packages, including sf, ggplot2, sp, spdep, spatialreg and mgcv.

Research findings

Moran's I statistic was initially employed to detect potential spatial autocorrelation in the distribution of failure counts and electricity usage across stations. The results of the Moran's I test for the failure count and electricity consumption have been presented in Table 1 and Table 2, respectively.

Table 1: Moran's I statistic for failure count

Moran I statistic standard deviate = 5.3308		p-value = 0.7161
Moran I statistic	Expectation	Variance
-0.064384141	-0.021276596	0.005692039

The result of the Moran's I test suggests that there is no significant spatial pattern or clustering in the failure count data, meaning that failures are randomly distributed across the stations and do not exhibit a clear spatial correlation. Since the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that there is no significant spatial autocorrelation in the failure count data. In other words, the number of failures does not show any significant spatial clustering or pattern.

Table 2: Moran's I statistic for electricity consumption

Moran I statistic standard deviate = -0.57137		p-value = 4.889e-08
Moran I statistic	Expectation	Variance
0.0005835877	-0.021276596	0.005835877

Since the p-value is very small (much smaller than 0.05), we reject the null hypothesis and conclude that there is significant spatial autocorrelation in the electricity consumption data. This means that electricity consumption values exhibit a spatial pattern, suggesting that stations located closer to each other tend to have similar levels of electricity consumption.

A cumulative plot for failure count and electricity consumption over time was generated to investigate the temporal trends of these variables. The plot, shown in Fig 1, displays the cumulative sum of failures and electricity consumption for each station, revealing underlying patterns and trends across the period.

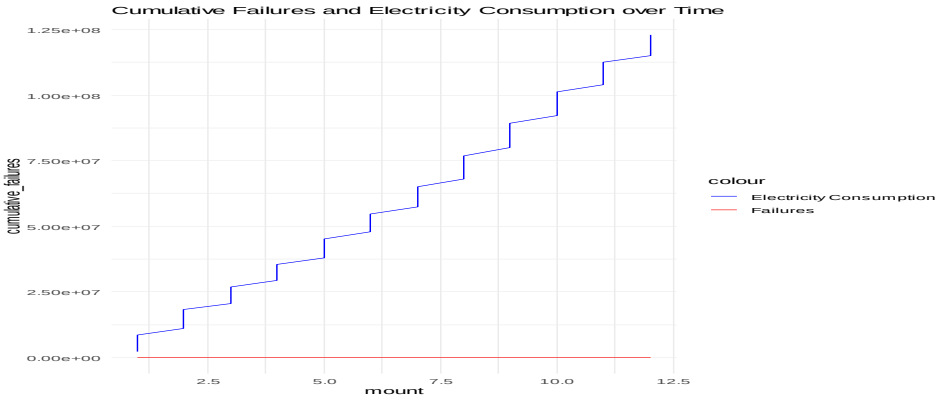

Fig 1: Cumulative failure count and electricity consumption over time

The cumulative plot in Fig 1 shows the temporal trends of failure count and electricity consumption over the period of one year. The horizontal axis represents the months of the year, while the vertical axis shows the cumulative failures. The horizontal line, located at the bottom of the plot, represents the failure count. This line remains relatively constant, showing that failures occur at specific times throughout the year but do not exhibit a rapid increase or decrease over time. The blue stepped line represents electricity consumption. This line increases progressively, showing the cumulative electricity consumption over the months. The stepped nature of the line indicates that electricity consumption increases in increments, reflecting usage over time.

K-means clustering was applied to partition the pumping stations into distinct groups based on their electricity consumption and failure counts. This technique helps to categorize stations that exhibit similar characteristics in terms of energy usage and failure frequency, enabling a better understanding of operational patterns and performance. The result has been shown in Fig 2.

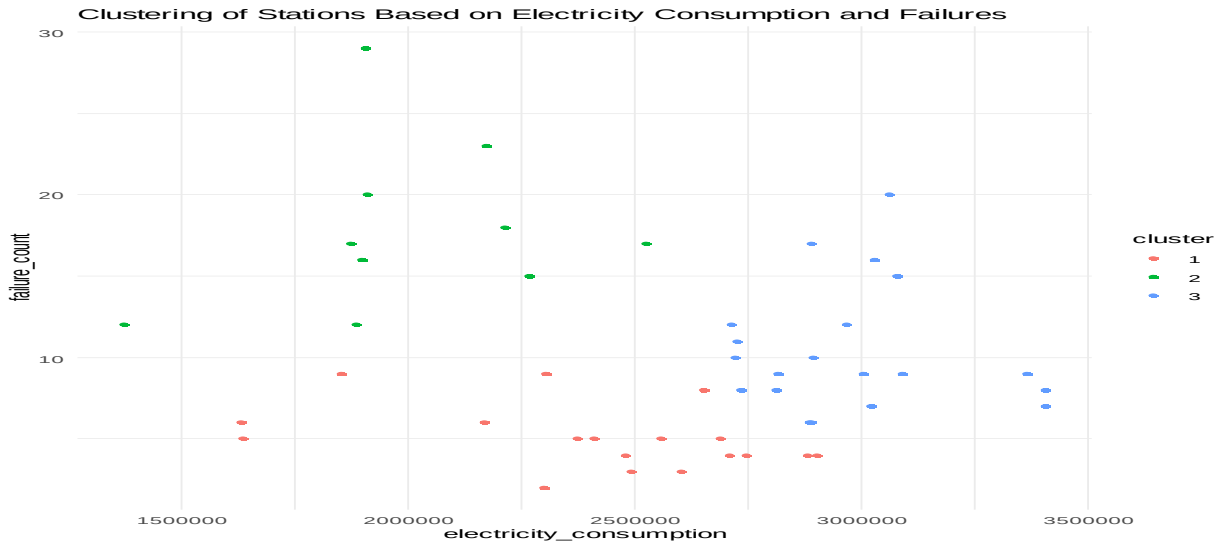

Fig 2: Clustering of stations based on electricity consumption and failure count

The clustering clearly shows that the stations with similar characteristics, in terms of failure count and electricity consumption, are grouped together. The points within each cluster are close to one another, indicating that these stations have similar patterns of performance and energy usage.

- The blue cluster might represent stations with high failure counts and relatively higher electricity consumption.
- The green cluster could indicate stations with moderate failure counts and moderate energy consumption
- The red cluster may correspond to stations with low failure counts and lower electricity consumption

To investigate the relationship between electricity consumption and failure count at the four pumping stations, a simple linear regression model was applied, and the results are presented in Table 3.

Table 3: Simple linear regression analysis

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.674e+01	6.380e+00	2.624	0.0118
electricity_consumption	-2.637e-06	1.978e-06	-1.333	0.1892
station_id	-4.609e-03	8.549e-01	-0.005	0.9957
Multiple R-squared	0.04702	Adjusted R-squared	0.004668	

As observed from the results of the model, the p-value for electricity consumption is greater than 0.05, indicating that the model is not statistically significant. Therefore, changes in electricity consumption does not have a significant impact on the number of failures. Although the model does not show a significant impact of energy consumption on failure count, it is important to consider other factors or use more advanced models to capture potential effects.

Also, for the stations, the p-value is greater than 0.05, indicating that the number of failures is not significantly different between the stations according to the model. In other words, the model does not distinguish significant variations in failure counts based on the station id.

The values of Multiple R-squared and Adjusted R-squared indicate that the model does not effectively captures the variations in the number of failures. These values suggest that the model does not explains a significant portion of the variability in failure counts, reflecting its adequacy in modeling the underlying patterns. The intuitive interpretation of the linear regression model is presented in Fig 3.

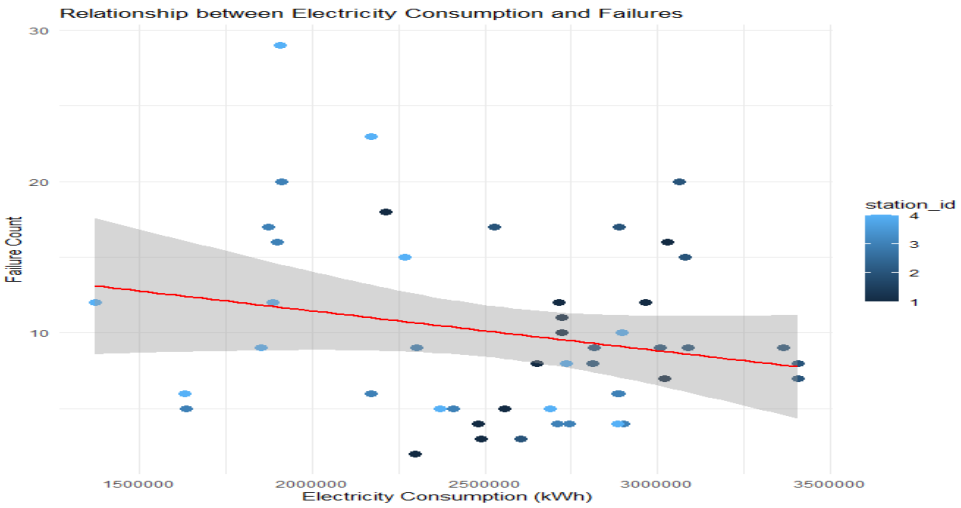

Fig3: Spatial relationship between electricity consumption and failure count

Although the values Multiple R-squared and Adjusted R-squared do not confirm the goodness of fit of the model, the Figure1 suggests that there is an inverse relationship between electricity consumption and the number of failures. Generally, with an increase in electricity consumption, the number of failures decreases. This downward trend is more pronounced at stations 1, 2, and 3, while it is less evident at station 4. In other words, the reduction in failure count with increasing electricity consumption is stronger at stations 1 to 3, while this relationship is less clear at station 4. This could indicate differences in performance or specific characteristics of the stations that should be considered in further analysis.

Due to the inadequacy of the linear regression model in explaining the variations in failure count with respect to electricity consumption, spatial regression has been applied to further investigate the relationship, and the results are presented in Table 4:

Table 4: Spatial Lag model analysis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.1540e+01	4.6631e+00	4.6192	3.852e-06
electricity consumption	-3.7555e-07	1.6707e-06	-0.2248	0.8221431
factor(month)2	-1.0480e+01	3.0181e+00	-3.4723	0.0005161
factor(month)3	-1.2836e+01	3.0121e+00	-4.2616	2.030e-05
factor(month)4	-1.5619e+01	3.0559e+00	-5.1109	3.206e-07
factor(month)5	-1.0977e+01	3.0296e+00	-3.6232	0.0002910
factor(month)6	-6.4960e+00	3.0163e+00	-2.1536	0.0312713
factor(month)7	-8.4140e+00	3.0885e+00	-2.7243	0.0064437
factor(month)8	-1.4270e+01	3.3094e+00	-4.3121	1.617e-05
factor(month)9	-1.1240e+01	3.3698e+00	-3.3355	0.0008515
factor(month)10	-1.3003e+01	3.3445e+00	-3.8878	0.0001012
factor(month)11	-1.3836e+01	3.1951e+00	-4.3303	1.489e-05
factor(month)12	-1.4158e+01	3.0951e+00	-4.5744	4.776e-06
Rho	0.0308	LR test value	0.03283	
z-value	0.13011	p-value	0.89648	

The results of fitting the Spatial Lag Model (SAR) to the data indicate that the spatial lag parameter ($\rho = 0.0308$, $p = 0.896$) is not statistically significant. This suggests that incorporating spatial dependence does not substantially improve the model's explanatory power. Furthermore, electricity consumption (p-value= 0.822) does not have a statistically significant effect on failure counts, confirming that changes in energy consumption are not a significant driver of failure events in this context.

To evaluate the adequacy of the model, a scatter plot of residuals versus fitted values has been used, and the results are presented in Fig 4.

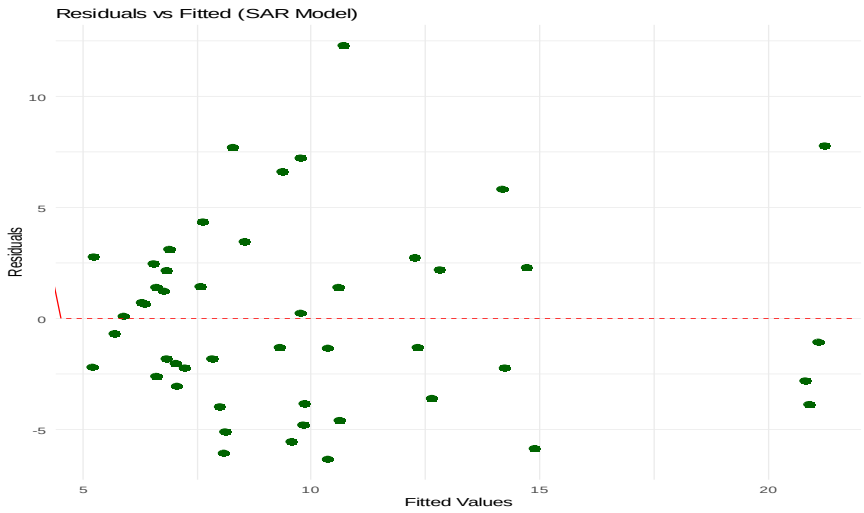

Fig 4: Residuals versus Fitted values

As visually observed in Fig 4, the residuals appear randomly scattered around the fitted values without forming an S-shaped pattern along the diagonal, indicating that the spatial lag model is appropriately fitted to the data."

Despite the adequate fit of the spatial regression model, it failed to reveal a significant relationship between electricity consumption and the number of failures across pumping stations. Therefore, to investigate the effect of electricity consumption on the number of failures, a generalized additive model (GAM) has been applied, and the results are presented in Table 5.

Table 5: Generalized additive model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.949449	0.153235	19.248	< 2e-16
factor(month)2	-0.599345	0.211560	-2.833	0.004612
factor(month)3	-0.876623	0.208450	-4.205	2.61e-05
factor(month)4	-0.440336	0.327143	-1.346	0.178302
factor(month)5	0.000000	0.000000	NaN	NaN
factor(month)6	0.164635	0.269455	0.611	0.541204
factor(month)7	0.719540	0.263956	2.726	0.006411
factor(month)8	0.000000	0.000000	NaN	NaN
factor(month)9	0.301532	0.264218	1.141	0.253777
factor(month)10	0.000000	0.000000	NaN	NaN
factor(month)11	-0.002296	0.298742	-0.008	0.993867
factor(month)12	0.156700	0.292498	0.536	0.592146
factor(station_id)2	-0.627595	0.274741	-2.284	0.022353
factor(station_id)3	-1.143090	0.278886	-4.099	4.15e-05
factor(station_id)4	-1.021784	0.281624	-3.628	0.000285
Electricity consumption	edf	Ref . df	Chi . sq	p-value
	6.756	7.694	24.8	0.00148

To better capture the potential nonlinear relationship between electricity consumption and the number of failures, a Generalized Additive Model (GAM) with a Poisson distribution and a log link function has been applied. In this model, a smooth term was used for electricity consumption, while month and station id were included as categorical covariates.

According to the results, the smooth term (electricity consumption) was found to be statistically significant (p-value = 0.00148), indicating that a nonlinear relationship between electricity consumption and failure count is present. This suggests that a linear model would not have been sufficient to capture this pattern accurately.

Additionally, several levels of the month and station id variables were shown to have significant effects, indicating their influence on failure count. Approximately 67% of the deviance was explained, and the adjusted R-squared was reported as 0.469, suggesting that

the model was reasonably well-fitted to the data.

These results imply that nonlinear effects and spatial-temporal variation should be accounted for when modeling failure counts in pump stations.

The non-linear relationship between electricity consumption and failure count has been depicted in Fig 5 based on the smooth term of the GAM:

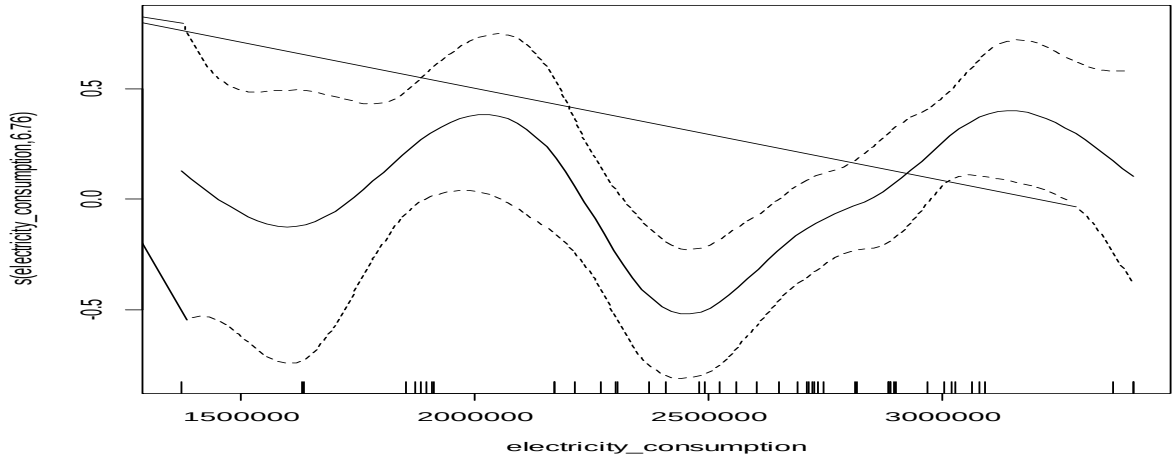

Fig 5: Smooth plot of the GAM

In Fig 5, the smooth term corresponding to electricity consumption has been plotted. The x-axis represents electricity consumption levels, while the y-axis displays the estimated partial effect on failure count. The figure clearly reveals a sinusoidal pattern in the relationship between electricity consumption and failure count.

Results

In conclusion, this research successfully integrates spatial modeling techniques and nonlinear method to analyze the performance of pumping stations in a water transmission infrastructure. The study underscores the importance of considering spatial and temporal factors in infrastructure management to enhance operational reliability and optimize resource allocation. By combining advanced analytical techniques, the study offers a robust framework for improving the management of pumping stations. Further research is needed to refine predictive models and explore the underlying causes of station-specific variations in performance, particularly at stations where performance patterns do not align with general trends.

In summary, this research offers a comprehensive approach to infrastructure management by integrating spatial statistics and predictive modeling techniques. The findings contribute to a deeper understanding of the factors influencing the performance of pumping stations, with the potential to guide future improvements in the management of water transmission systems and beyond.

Lastly, the application of this framework to other types of infrastructure systems beyond water transmission, such as wastewater treatment plants or energy distribution networks, could further validate its versatility and impact. By expanding the scope of analysis to include various systems, researchers and practitioners alike can leverage these advanced analytical techniques to enhance decision-making, optimize system operations, and ensure long-term sustainability.

Future research should aim to uncover these local factors more comprehensively. Additionally, refining the predictive models through machine learning techniques could provide more accurate forecasts and improve the overall predictive power of the framework

.

References

Abadia, R., Rocamora, C., Ruiz, A., & Puerto, H. (2008). Energy efficiency in irrigation distribution networks I: theory. *Biosystems engineering*, *101*(1), 21-27.

Aliyu, R., Mokhtar, A. A., & Hussin, H. (2023). A Study on Comparison of Classification

Algorithms for Pump Failure Prediction. *Journal of Advanced Industrial Technology and Application*, *4*(2), 48-65.

Anselin, L. (1988). *Spatial econometrics: methods and models* (Vol. 4). Springer Science & Business Media.

Bao, C., & Chen, X. (2017). Spatial econometric analysis on influencing factors of water consumption efficiency in urbanizing China. *Journal of Geographical Sciences*, *27*, 1450-1462.

Blokus-Dziula, A., Dziula, P., Kamedulski, B., & Michalak, P. (2023). Operation and maintenance cost of water management systems: Analysis and optimization. *Water*, *15*(17), 3053.

Barton, N. A., Farewell, T. S., & Hallett, S. H. (2020). Using generalized additive models to investigate the environmental effects on pipe failure in clean water networks. *Npj Clean Water*, *3*(1), 31.

Chatfield, C., & Xing, H. (2019). *The analysis of time series: an introduction with R*. Chapman and hall/CRC.

Christodoulou, S., Gagatsis, A., Agathokleous, A., Xanthos, S., & Kranioti, S. (2012). Urban water distribution network asset management using spatio-temporal analysis of pipe-failure data. *Proc. ICCCBE*, 27-29.

Cliff, A. D., & Ord, J. K. (1981). Spatial processes: models & applications. *(No Title)*.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Dadar, S., Đurin, B., Alamatian, E., & Plantak, L. (2021). Impact of the pumping regime on electricity cost savings in urban water supply system. *Water*, *13*(9), 1141.

Du, B., Ye, J., Zhu, H., Sun, L., & Du, Y. (2023). Intelligent monitoring system based on spatio–temporal data for underground space infrastructure. *Engineering*, *25*, 194-203.

Elhorst, J. P. (2014). *Spatial econometrics: from cross-sectional data to spatial panels* (Vol. 479, p. 480). Heidelberg: Springer.

Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189-206.

Hou, C. C., Wang, H. F., Li, C., & Xu, Q. (2020). From metal–organic frameworks to single/dual-atom and cluster metal catalysts for energy applications. *Energy & Environmental Science*, *13*(6), 1658-1693.

Ikramov, N., Majidov, T., Kan, E., & Mukhammadjonov, A. (2020, July). Monitoring system for electricity consumption at pumping stations. In *IOP Conference Series: Materials Science and Engineering* (Vol. 883, No. 1, p. 012101). IOP Publishing.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters*, *31*(8), 651-666.

Kofinas, D., Ulanczyk, R., & Laspidou, C. S. (2020). Simulation of a water distribution network with key performance indicators for spatio-temporal analysis and operation of highly stressed water infrastructure. *Water*, *12*(4), 1149.

Liu, X., Liu, H., Wan, Z., Chen, T., & Tian, K. (2015). Application and study of internet of things used in rural water conservancy project. *Journal of Computational Methods in Science and Engineering*, *15*(3), 477-488.

Luna, T., Ribau, J., Figueiredo, D., & Alves, R. (2019). Improving energy efficiency in water supply systems with pump scheduling optimization. *Journal of cleaner production*, *213*, 342-356.

MacQueen, J. (1967, January). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (Vol. 5, pp. 281-298). University of California press.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17-23.

Mutambik, I. (2024). Assessing Urban Vulnerability to Emergencies: A Spatiotemporal Approach Using K-Means Clustering. *Land*, *13*(11), 1744.

Qiu, P., Yan, J., Xu, H., & Yu, Y. (2024). Health status assessment of pump station units based on spatio-temporal fusion and uncertainty information. *Scientific Reports*, *14*(1), 24096.

Shi, F. (2018). *Data-driven predictive analytics for water infrastructure condition assessment and management* (Doctoral dissertation, University of British Columbia).

Tiony, C. K. (2013). *An energy assessment of the water pumping Systems at the Gigiri pumping station* (Doctoral dissertation, University of Nairobi).

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Yates, M. A., & Weybourne, I. (2001). Improving the energy efficiency of pumping systems. *Journal of Water Supply: Research and Technology—AQUA*, *50*(2), 101-111.