Fitting Multivariate Time Series Models: Analyzing Environmental and Chemical Data

Sajad Ahmad Mir(M21MA207)
Himalaya(M21MA204)
Nasrat Jahan (M21MA205)
Raman Reshi(M21MA206)

Abstract

Environmental and chemical datasets often exhibit complex interdependencies and temporal structures, necessitating robust statistical approaches for meaningful analysis. This study focuses on modeling a multivariate time series dataset obtained from the India Meteorological Department (IMD) Pune via its data supply portal for Jodhpur city. The dataset comprises variables such as rainfall, pH, conductivity, sulphate, nitrate, chloride, ammonium, calcium, magnesium, sodium, and potassium. Our primary objective is to uncover the dynamic relationships between these variables and forecast their future behavior. We began our analysis by attempting to fit ARIMA models, but no suitable model was found. Subsequently, we applied a GARCH model to the transformed variable log(1+ph) to capture its volatility structure. Building on this, we employed vector autoregression (VAR) to explore the temporal dependencies and assess the statistical significance of the interactions between the components. This report outlines the methodology, data preprocessing steps, model fitting procedures, and diagnostic checks performed to validate the fitted models. The results provide insights into the intricate interplay between environmental and chemical parameters, with potential applications in environmental monitoring and resource management.

1 Introduction

Time series analysis has become an essential tool in understanding temporal patterns and dependencies in environmental and chemical datasets. Multivariate time series models, in particular, allow for simultaneous analysis of multiple interrelated variables, offering insights into their dynamic relationships. The ability to model and forecast such data is crucial for applications in environmental management, agriculture, and public health.

In this study, we examine a dataset obtained from the India Meteorological Department (IMD) Pune via its data supply portal for Jodhpur city. The dataset consists of eleven variables: rainfall, pH, conductivity, sulphate, nitrate, chloride, ammonium, calcium, magnesium, sodium, and potassium. These variables represent key environmental and chemical parameters, which are not only interdependent but also influenced by temporal dynamics. The data has been recorded over multiple time periods and offers a comprehensive view of the environmental conditions in the region.

Our aim is to identify patterns in the data, evaluate the relationships between the variables, and develop predictive models using multivariate time series approaches. The analysis began with an attempt to fit ARIMA models to capture temporal dependencies in individual variables. However, no suitable ARIMA model was found. To address this, we applied a GARCH model to the transformed variable log(1+ph) to account for its volatility. Building on this, we employed vector autoregression (VAR) and vector autoregressive moving average (VARMA) models to explore the interdependencies and temporal relationships among the variables.

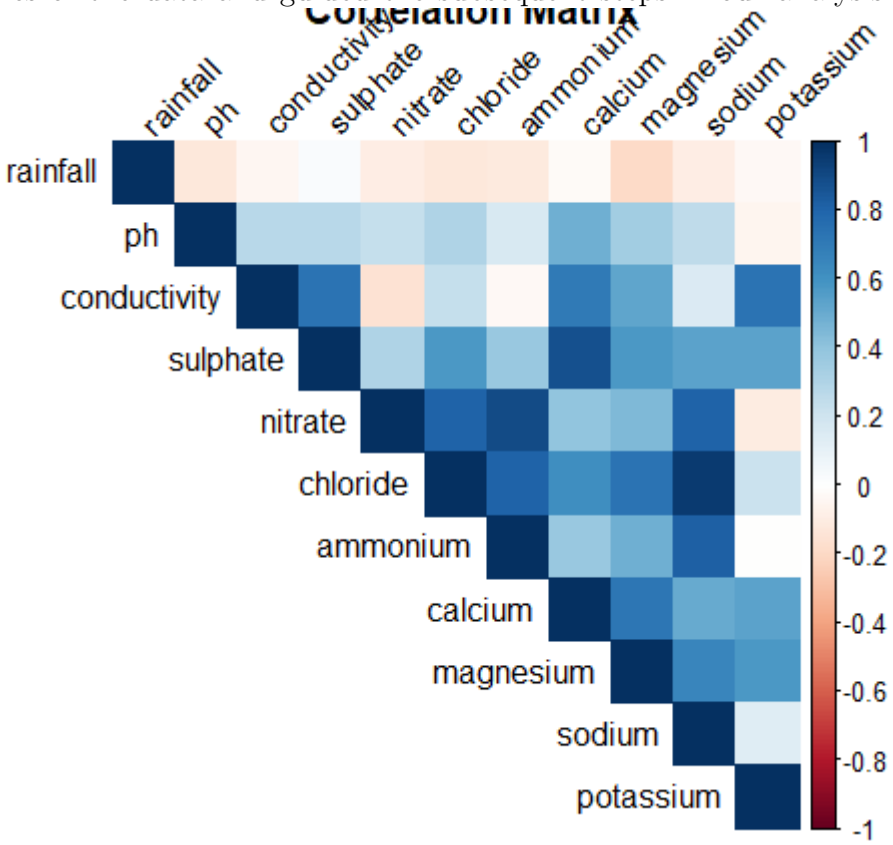
The analysis involves rigorous preprocessing steps undertaken to handle irregularities such as missing values and non-stationarity, followed by the model fitting and evaluation process. By conducting diagnostic checks, we ensure the validity and reliability of the fitted models.

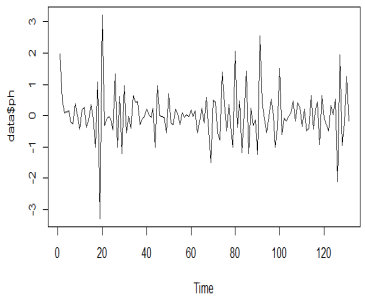
2 Data

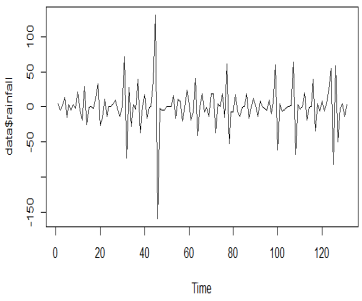
	station	year	month	date	rainfall	ph	conductivity	sulphate	nitrate	chloride	ammonium	calcium	magnesium	sodium	potassium	ph
1	42027	2015	2	2	12	7.55	79.7	3.02	7.27	1.85	0.16	12.36	1.04	0.84	0.45	
2	42027	2015	2	3	4	6.86	18.4	0.26	2.24	0.88	0.14	2.81	0.14	0.08	0.09	
3	42027	2015	2	8	4	7.63	59.8	2.07	4.22	1.70	0.05	9.30	0.63	0.76	0.45	
4	42027	2015	2	16	11	6.83	21.1	1.10	3.16	1.33	0.08	2.87	0.24	0.45	0.18	
5	42027	2015	2	19	21	6.46	13.0	0.39	2.86	3.72	0.07	3.37	0.16	0.12	0.11	
6	42027	2015	2	25	45	4.83	31.9	2.50	7.71	1.59	0.12	2.93	0.34	0.54	0.25	
7	42027	2015	3	3	10	6.86	17.5	0.24	2.95	0.81	0.03	2.75	0.15	0.09	0.11	
8	42027	2015	3	4	1	6.92	53.3	3.03	16.23	1.48	0.07	6.91	0.75	0.39	0.55	
9	42027	2015	3	5	2	5.60	13.9	0.50	4.06	0.64	0.16	1.81	0.12	0.05	0.11	
10	42027	2015	3	9	54	6.45	19.6	1.65	4.53	0.78	0.02	2.84	0.19	0.32	0.12	
11	42027	2015	3	16	37	7.47	70.3	1.93	11.87	1.31	0.02	10.88	0.65	0.46	0.43	
12	42027	2015	3	25	11	7.25	34.3	0.97	3.67	3.05	0.05	6.31	0.42	0.70	0.28	
13	42027	2015	3	29	44	6.96	21.7	0.83	3.47	4.27	0.05	5.04	0.62	0.31	0.20	
14	42027	2015	4	2	27	6.32	15.2	0.35	4.32	0.83	0.16	2.35	0.13	0.11	0.24	

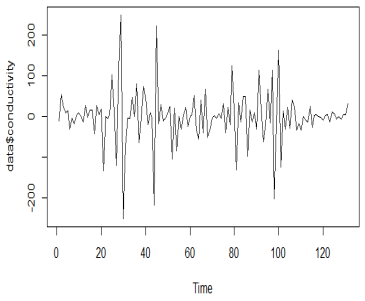
3 Basic Analysis

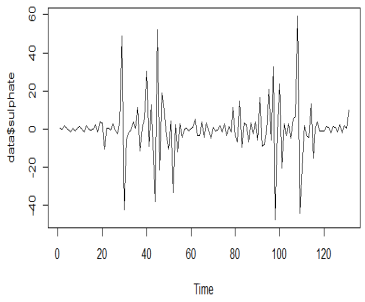
we performed an initial analysis of the data by visualizing it through time-series plots. These plots allowed us to observe the temporal patterns and trends within the data, providing valuable insights into its behavior over time. By examining the time series, we were able to identify key fluctuations, trends, and potential anomalies, which served as the foundation for further analysis. This approach helped in understanding the dynamics of the data and guided the subsequent steps in our analysis process.

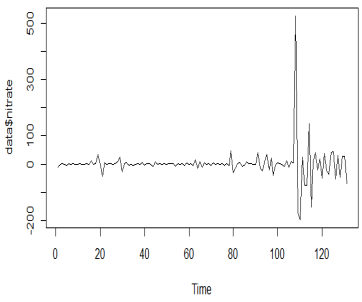


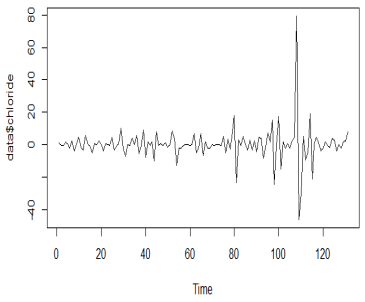

data$ph
Time

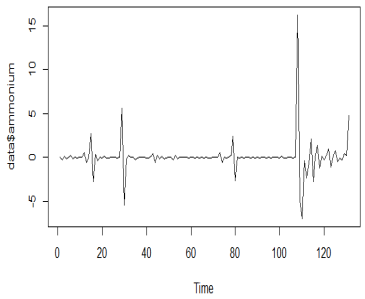

data$rainfall
Time

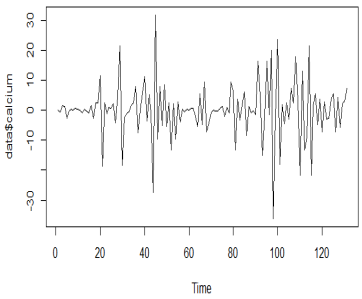

data$conductivity
Time

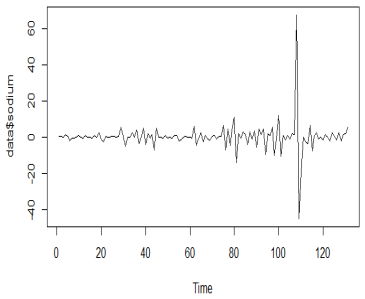

data$sulphate
Time

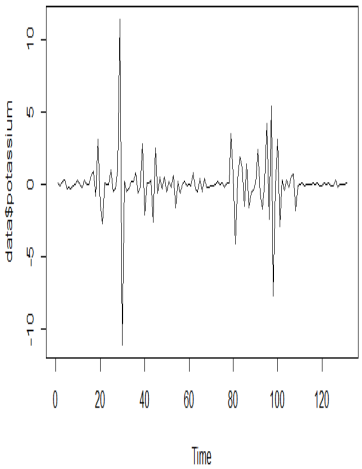

data$nitrate
Time


data$chloride
Time


data$ammonium
Time

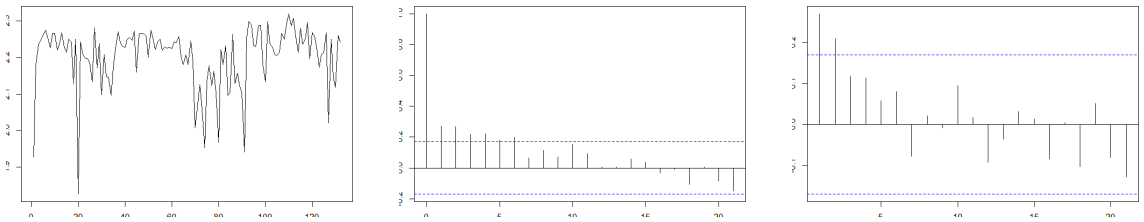

data$calcium
Time


data$sodium
Time


data$potassium
Time

4 GARCH Modelling

We began our analysis by fitting ARIMA models to the data to capture any underlying temporal dependencies. However, after several attempts, we were unable to find a suitable ARIMA model that adequately described the data. As a result, we shifted our focus to the GARCH model, specifically modeling the transformation log(1+ph). This approach provided a more effective way to model the volatility and fluctuations in the data, offering a better fit for the observed patterns.



log(1+ph) **ACF of log(1+ph)** **PACF of log(1+ph)**

Table 1: Figures of log(1+ph) and its ACF and PACF.

4.1 Residual Diagnostics

To evaluate the goodness-of-fit and check for any remaining structure in the residuals, diagnostic tests were conducted on the transformed series log(data$ph + 1). The results are summarized below.

4.1.1 Box-Ljung Test

The Box-Ljung test was performed on the residuals (r^2) of the transformed series to check for autocorrelation. The null hypothesis of the test is that there is no autocorrelation in the residuals. The test results are as follows:

$$\text{Test Statistic: } X^2 = 9.9072, \quad \text{Degrees of Freedom: } df = 1, \quad \text{p-value: } 0.001646.$$

Based on the p-value (< 0.05), we reject the null hypothesis and conclude that there is significant autocorrelation in the residuals.

4.1.2 ARCH LM Test

The ARCH LM test was applied to the residuals (r^2) of the transformed series to check for the presence of autoregressive conditional heteroskedasticity (ARCH) effects. The null hypothesis is that there are no ARCH effects. The test results are as follows:

$$\text{Test Statistic: } \chi^2 = 24.758, \quad \text{Degrees of Freedom: } df = 12, \quad \text{p-value: } 0.01601.$$

Since the p-value is less than 0.05, we reject the null hypothesis, indicating the presence of significant ARCH effects in the residuals.

4.2 Conclusion

The diagnostic tests reveal significant autocorrelation and ARCH effects in the residuals of log(data\$ph + 1), suggesting that the model may require further refinement or alternative specifications.

4.3 Model Description

The GARCH model is fitted to the log-transformed pH data using the following specification:

$$\begin{aligned} y_t &= \mu + \epsilon_t, \\ \epsilon_t &= \sigma_t z_t, \\ \sigma_t^2 &= \omega + \alpha_1 \epsilon_{t-1}^2 + \alpha_2 \epsilon_{t-2}^2 + \beta_1 \sigma_{t-1}^2, \end{aligned}$$

where y_t is the log-transformed pH data, $z_t \sim N(0, 1)$, and the parameters are estimated using maximum likelihood.

4.4 Model Output

The model equation and estimated coefficients are as follows:

Call: `garchFit(formula = garch(2, 1), data = log(data$ph + 1), trace = F)`

Mean and Variance Equation:

$$\text{data} \sim \text{GARCH}(2, 1), \quad [\text{data} = \log(\text{data\$ph} + 1)].$$

Conditional Distribution: Normal distribution.

Estimated Coefficients:

$$\begin{aligned} \mu &= 2.0896, \\ \omega &= 0.00079265, \\ \alpha_1 &= 0.0664, \\ \alpha_2 &= 0.00000001, \\ \beta_1 &= 0.8210. \end{aligned}$$

Error Analysis

The parameter estimates, standard errors, t-values, and significance levels are summarized in Table 2.

4.5 Log Likelihood

The log-likelihood of the fitted model is:

$$\text{Log Likelihood} = 329.7767, \quad \text{normalized: } 1.074191.$$

Table 2: Error Analysis of GARCH(2,1) Model.

Parameter	**Estimate**	**Std. Error**	**t-value**	**Pr($>|t|$)**
μ	2.0896	0.004983	419.333	$< 2 \times 10^{-16}$***
ω	0.00079265	0.0005166	1.534	0.125
α_1	0.0664	0.04679	1.419	0.156
α_2	0.00000001	0.05872	0.000	1.000
β_1	0.8210	0.09434	8.703	$< 2 \times 10^{-16}$***

Significance codes: ***0.001,** 0.01,* 0.05, .0.1.

4.6 Standardized Residuals Tests

The diagnostic checks for standardized residuals are provided in Table 3.

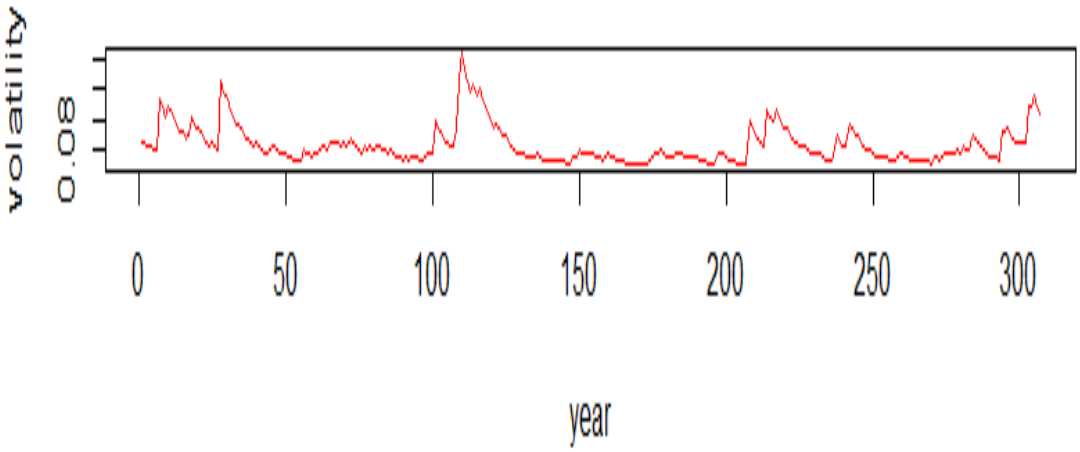
Table 3: Diagnostic Tests for Standardized Residuals.

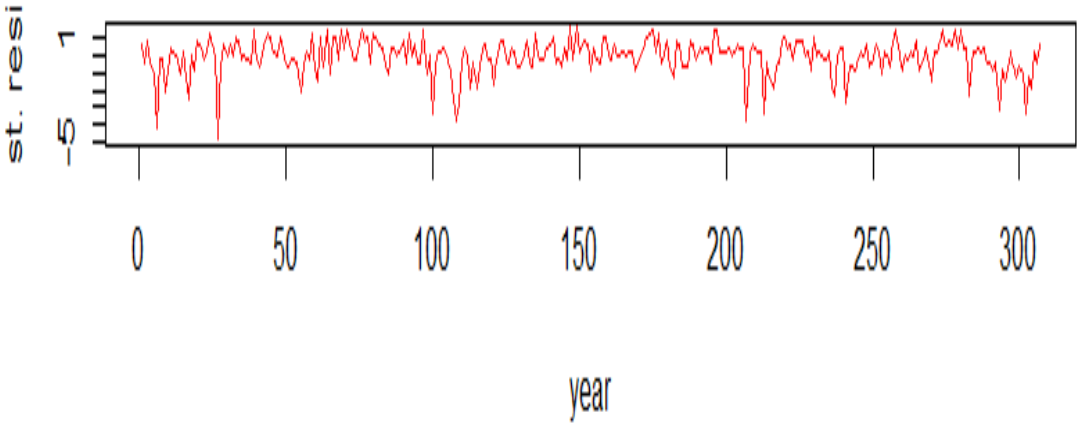
Test	**Statistic**	**p-Value**
Jarque-Bera Test R, χ^2	278.3886	$< 10^{-16}$
Shapiro-Wilk Test R, W	0.9000	2.30×10^{-13}
Ljung-Box Test R, $Q(10)$	45.2332	1.97×10^{-6}
Ljung-Box Test R, $Q(15)$	47.9649	2.58×10^{-5}
Ljung-Box Test R, $Q(20)$	56.8161	2.19×10^{-5}
Ljung-Box Test R^2, $Q(10)$	5.9232	0.822
Ljung-Box Test R^2, $Q(15)$	8.8288	0.886
Ljung-Box Test R^2, $Q(20)$	10.3515	0.961
LM ARCH Test R, TR^2	7.7602	0.804

4.7 Information Criterion Statistics

The information criteria for model selection are as follows:

$$\begin{aligned} \text{AIC} &= -2.1158, \\ \text{BIC} &= -2.0551, \\ \text{SIC} &= -2.1163, \\ \text{HQIC} &= -2.0915. \end{aligned}$$





5 Final VAR Model Results

After applying the GARCH model to log(1+ph), we proceeded to explore the Vector Autoregressive (VAR) model. The VAR model allowed us to capture the interdependencies between multiple time series variables, providing a more comprehensive analysis of the relationships within the data. This approach helped us examine the dynamic interactions between the variables and assess their collective behavior over time. The table below

summarizes the results from the final Vector Autoregression (VAR) model, including the coefficients for the first and second lags (AR(1) and AR(2)) for each variable, along with their corresponding standard errors.

AR(1) Matrix

The AR(1) coefficients matrix is as follows:

$$\mathbf{AR(1)} = \begin{bmatrix} 0.1516 & 3.4544 & -0.1654 & -0.2187 & -0.1184 & -0.6534 & 5.7524 & -0.0727 & 8.4613 & -0.0976 \\ 0.6856 & & & & & & & & & \\ 0.0048 & 0.0649 & 0.0012 & -0.0194 & 0.0002 & 0.0057 & 0.0512 & 0.0338 & -0.0352 & 0.0045 \\ -0.0129 & & & & & & & & & \\ 0.1312 & -4.8985 & 1.1531 & -0.2044 & 0.6859 & 0.8863 & -13.7250 & -3.5664 & -19.0668 & -1.4476 \\ -3.3183 & & & & & & & & & \\ 0.1024 & 0.8773 & 0.0978 & 0.5145 & 0.1377 & 0.2184 & -3.4267 & -0.8161 & -2.1351 & -0.4074 \\ -0.2015 & & & & & & & & & \\ 0.6903 & 4.5272 & 0.0892 & 0.7612 & 0.5673 & 1.7614 & -2.9333 & -3.1043 & 3.8381 & -0.7153 \\ -0.4340 & & & & & & & & & \\ 0.1015 & 0.7773 & 0.0416 & 0.0392 & 0.0619 & 0.4743 & -0.0117 & -0.4704 & 0.0601 & -0.4082 \\ 0.4191 & & & & & & & & & \\ 0.0190 & 0.4190 & 0.0092 & 0.0301 & 0.0223 & 0.0713 & -0.1590 & -0.1590 & 0.0761 & -0.0490 \\ -0.0416 & & & & & & & & & \\ 0.0658 & -0.3858 & 0.0697 & 0.1749 & 0.1131 & 0.0237 & -2.0973 & -0.3956 & -0.7058 & -0.0698 \\ -0.0789 & & & & & & & & & \\ 0.0084 & -0.0742 & 0.0101 & 0.0103 & 0.0106 & 0.0310 & -0.0532 & -0.0664 & 0.0816 & -0.0498 \\ -0.0167 & & & & & & & & & \\ 0.0790 & 0.1754 & 0.0232 & -0.0227 & 0.0354 & 0.1577 & -0.0230 & -0.1846 & 0.1429 & -0.0309 \\ 0.3472 & & & & & & & & & \\ -0.0002 & -0.0756 & 0.0255 & -0.0129 & 0.0271 & -0.0278 & -0.4930 & -0.1208 & -0.1594 & -0.0232 \\ 0.0947 & & & & & & & & & \end{bmatrix}$$

AR(2)-Matrix

$$\mathbf{AR(2)} = \begin{bmatrix} -0.04125 & -0.51380 & 0.01409 & 0.44433 & -0.12600 & -0.5631 & 3.55120 & 0.5173 & -2.0554 & -0.1315 \\ 2.2177 & & & & & & & & & \\ -0.00506 & 0.03261 & 0.00375 & -0.01006 & 0.00455 & -0.0052 & -0.10851 & 0.0172 & -0.0657 & -0.0064 \\ -0.0929 & & & & & & & & & \\ 0.20417 & 3.96806 & -0.13174 & 1.39086 & -0.30119 & -2.5116 & -3.12179 & 0.7216 & 5.7326 & 4.7923 \\ -5.4692 & & & & & & & & & \\ 0.03718 & -0.05538 & -0.05477 & 0.16165 & -0.07139 & -0.9388 & -0.05907 & 0.8499 & -0.3539 & 1.4627 \\ -0.2528 & & & & & & & & & \\ -0.10532 & 8.02827 & -0.14893 & -1.20296 & 0.15182 & -3.6381 & 0.42683 & 3.7576 & -9.9728 & 3.6866 \\ 4.4782 & & & & & & & & & \\ -0.02233 & 0.13124 & -0.04189 & -0.07012 & -0.03445 & -0.5664 & -0.22309 & 0.6741 & -0.9833 & 0.7713 \\ 0.3733 & & & & & & & & & \\ 0.00114 & 0.00404 & -0.00727 & -0.01005 & -0.00276 & -0.0640 & -0.00099 & 0.0881 & -0.1172 & 0.0827 \\ 0.0194 & & & & & & & & & \\ 0.01324 & 0.09584 & -0.02971 & 0.02057 & -0.02910 & -0.6606 & 0.13732 & 0.6754 & 0.1452 & 0.7871 \\ -0.4617 & & & & & & & & & \\ -0.00335 & -0.04861 & -0.00653 & -0.00964 & -0.00254 & -0.0778 & -0.09612 & 0.0618 & 0.1239 & 0.1092 \\ 0.0768 & & & & & & & & & \\ -0.02363 & -0.02779 & -0.03304 & -0.05080 & -0.02540 & -0.3720 & -0.28834 & 0.4230 & -0.4586 & 0.5547 \\ 0.3822 & & & & & & & & & \\ 0.00230 & 0.18762 & -0.01779 & 0.04284 & -0.01277 & -0.1069 & -0.04129 & 0.0378 & 0.1685 & 0.1952 \\ 0.2775 & & & & & & & & & \end{bmatrix}$$

Standard Error

The standard error matrix for the AR(1) coefficients:

$$\text{SE} = \begin{bmatrix} 0.0956 & 4.448 & 0.0897 & 0.4626 & 0.1485 & 0.9053 & 3.791 & 0.9268 & 4.621 & 1.2014 \\ 2.5827 & & & & & & & & & \\ 0.0028 & 0.132 & 0.0027 & 0.0137 & 0.0044 & 0.0269 & 0.113 & 0.0275 & 0.137 & 0.0357 \\ 0.0767 & & & & & & & & & \\ 0.2547 & 11.857 & 0.2391 & 1.2330 & 0.3957 & 2.4130 & 10.105 & 2.4702 & 12.318 & 3.2021 \\ 6.8839 & & & & & & & & & \\ 0.0533 & 2.480 & 0.0500 & 0.2579 & 0.0828 & 0.5048 & 2.114 & 0.5167 & 2.577 & 0.6698 \\ 1.4400 & & & & & & & & & \\ 0.2394 & 11.144 & 0.2248 & 1.1589 & 0.3720 & 2.2680 & 9.498 & 2.3217 & 11.578 & 3.0097 \\ 6.4701 & & & & & & & & & \\ 0.0408 & 1.897 & 0.0383 & 0.1973 & 0.0633 & 0.3862 & 1.617 & 0.3953 & 1.971 & 0.5124 \\ 1.1016 & & & & & & & & & \\ 0.0077 & 0.360 & 0.0073 & 0.0375 & 0.0120 & 0.0733 & 0.307 & 0.0750 & 0.374 & 0.0973 \\ 0.2091 & & & & & & & & & \\ 0.0341 & 1.589 & 0.0320 & 0.1652 & 0.0530 & 0.3234 & 1.354 & 0.3310 & 1.651 & 0.4291 \\ 0.9225 & & & & & & & & & \\ 0.0043 & 0.199 & 0.0040 & 0.0207 & 0.0067 & 0.0406 & 0.170 & 0.0416 & 0.207 & 0.0539 \\ 0.1158 & & & & & & & & & \\ 0.0314 & 1.464 & 0.0295 & 0.1522 & 0.0489 & 0.2979 & 1.248 & 0.3050 & 1.521 & 0.3954 \\ 0.8500 & & & & & & & & & \\ 0.0070 & 0.325 & 0.0066 & 0.0338 & 0.0108 & 0.0661 & 0.277 & 0.0677 & 0.337 & 0.0877 \\ 0.1886 & & & & & & & & & \end{bmatrix}$$

Residual Covariance Matrix

The residual covariance matrix is:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 358.12 & -1.15 & 41.12 & 11.63 & -98.54 & -20.21 & -2.97 & 1.03 & -2.84 & -12.58 \\ -0.57 & & & & & & & & & \\ -1.15 & 0.32 & 9.31 & 1.48 & 2.99 & 1.00 & 0.04 & 1.76 & 0.15 & 0.66 \\ -0.03 & & & & & & & & & \\ 41.12 & 9.31 & 2544.21 & 377.18 & 80.68 & 116.34 & 8.76 & 272.90 & 27.63 & 62.81 \\ 53.76 & & & & & & & & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & & & & & & & & \end{bmatrix}$$

Additional Statistics

$$\det(\text{SSE}) = 94864561636$$
$$\text{AIC} = 28.94238$$
$$\text{BIC} = 34.22752$$
$$\text{HQ} = 31.09002$$

6 Theory Behind Vector Autoregression (VAR)

Vector Autoregression (VAR) is a statistical model used to analyze the interdependencies and dynamic relationships between multiple time series variables. The key idea is that each variable in the system is explained by its own past values and the past values of other variables.

Key Points of VAR

- **Multiple Time Series:** VAR models allow for the analysis of multiple interrelated time series simultaneously, unlike univariate time series models which focus on a single variable.
- **Lag Structure:** The model includes lags (previous time periods) of both the dependent and independent variables, capturing temporal dependencies and understanding how past values influence future outcomes.
- **Model Equations:** Each equation in a VAR model represents one variable in terms of its own past values and the past values of all other variables. For example, the equation for Rainfall may look like:

$$\text{Rainfall}_t = c_1 + \phi_{11}\text{Rainfall}_{t-1} + \phi_{12}\text{pH}_{t-1} + \cdots + \phi_{12}\text{Rainfall}_{t-2} + \epsilon_1$$

  where c_1 is the constant, ϕ are the coefficients, and ϵ_1 is the error term.
- **Model Estimation:** The coefficients of the VAR model are estimated using methods like Ordinary Least Squares (OLS). These coefficients represent the strength and direction of relationships between the variables at different lags.
- **Model Interpretation:** By analyzing the coefficients and their statistical significance, we can assess how past values of each variable affect its own future values and the future values of other variables.

6.1 Strengths and Limitations of VAR

Strengths:

- Useful when there are no strong assumptions about the relationship between the variables.
- Can capture complex interdependencies between multiple time series.

Limitations:

- Requires large amounts of data, especially when dealing with many variables.
- Sensitive to the chosen lag length, making model selection (like AIC or BIC) crucial.

7 Results

- **Influence of Past Lags:** The AR(1) and AR(2) coefficients indicate varying degrees of influence from previous periods on future values. For example, Rainfall shows a strong positive impact from its past values (AR(1) coefficient of 404.097), while other variables like Nitrate show a weaker relationship.
- **Statistical Significance:** Some variables, such as Nitrate and Magnesium, show large standard errors, suggesting that their relationships with past values may not be statistically significant. Further refinement may be necessary to improve accuracy.
- **Dynamic Relationships:** The AR coefficients vary across variables, indicating different time dependencies. For example, Calcium and Conductivity show stronger dependencies on their past values, while Potassium shows weaker relationships.

In conclusion, the VAR model reveals that there are significant temporal dependencies for certain variables, though some relationships are weak or statistically insignificant. This suggests that further model refinement is necessary to improve prediction accuracy.