Research Study: Machine Learning-Based Statistical Analysis of Global Water Resource Utilization, Economic Factors, and Climate Trends Using R

1. Introduction

Water is a fundamental natural resource essential for human survival, economic development, and environmental stability. However, increasing water stress driven by population growth, climate change, and economic activities presents a significant global challenge. Understanding the dynamics of water resource utilization, economic factors, and climate trends is crucial for effective policymaking and sustainable resource management. This study aims to analyze global and country-level water resource utilization patterns by integrating classical statistical methods with machine learning techniques. The research will explore key variables such as freshwater withdrawals across sectors, water stress levels, GDP per capita, private investment in water and sanitation, and climate indicators like global temperature and sea level rise. The methodological approach includes exploratory data analysis (EDA), regression analysis, correlation analysis, supervised learning (regression & classification), unsupervised learning (clustering & dimensionality reduction), and time-series forecasting to extract insights and predict future trends. The analysis will be conducted using R.

2. Research Objectives

The main objectives of this study are:
- To analyze global and regional trends in water resource utilization.
- To examine sectoral freshwater withdrawals (agriculture, industry, and domestic use) and their changes over time.
- To predict a country's total freshwater withdrawals and water stress levels.
- To apply clustering techniques to group countries with similar water usage behaviors.
- To assess how economic factors (e.g., GDP per capita, private investment in water and sanitation) and climate variables (e.g., temperature, sea level rise, extreme weather events) influence water resource utilization.
- To develop predictive models for future trends in water availability, withdrawals, and stress levels.

3. Data Description

The dataset used in this study is sourced from international organizations, covering global water resource utilization, economic factors, and climate trends.

Key Variables:

Column Name	Description
Water productivity	GDP in constant prices divided by annual

	total water withdrawal
Water stress level	The ratio between total freshwater withdrawn by all major sectors and total renewable freshwater resources
Total freshwater withdrawals	Annual total freshwater withdrawals (billion cubic meters)
Agricultural water withdrawals	Percentage of freshwater withdrawals used for agriculture
Domestic water withdrawals	Percentage of freshwater withdrawals used for domestic purposes
Industrial water withdrawals	Percentage of freshwater withdrawals used for industry
GDP per capita (PPP)	Gross Domestic Product per capita, in purchasing power parity (current international $)
Private investment in water and sanitation	Investment in water and sanitation with private participation(current US$)
Extreme weather events	Percentage of the population affected by droughts, floods, and extreme temperatures (average from 1990 to 2009)
Global sea level	Satellite sea level observations
Global temperature	Global land-ocean temperature index
Groundwater depletion	Groundwater depletion volume and average depletion rate in the USA across distinct historical periods

Data Sources:
https://data.worldbank.org/indicator/ER.GDP.FWTL.M3.KD
https://data.worldbank.org/indicator/ER.H2O.FWST.ZS
https://data.worldbank.org/indicator/ER.H2O.FWTL.K3
https://data.worldbank.org/indicator/ER.H2O.FWAG.ZS
https://data.worldbank.org/indicator/SP.POP.TOTL
https://data.worldbank.org/indicator/ER.H2O.FWDM.ZS
https://data.worldbank.org/indicator/ER.H2O.FWIN.ZS
https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD
https://data.worldbank.org/indicator/IE.PPI.WATR.CD
https://data.worldbank.org/indicator/EN.CLC.MDAT.ZS
https://climate.nasa.gov/vital-signs/sea-level/?intent=121
https://climate.nasa.gov/vital-signs/global-temperature/?intent=121
Groundwater depletion in the United States (1900-2008) - ScienceBase-Catalog

4. Methodology

This study will employ both classical statistical methods and machine learning techniques to analyze and model water resource utilization trends.

Step 1: Exploratory Data Analysis (EDA)

- Descriptive statistics: Mean, median, standard deviation of key variables.
- Data visualization: Time-series plots, heatmaps, and histograms using `ggplot2`.
- Data preprocessing: Handling missing values and normalizing variables using `dplyr`.

Step 2: Classical Statistical Analysis

(1) Regression Analysis (Predicting Total Freshwater Withdrawals and Water Stress Levels)

- Goal: Understand the relationship between key variables and predict total freshwater withdrawals and water stress levels.
- Methods in R:
  - Linear Regression (`lm()`)
  - Correlation Analysis (`cor()`, `corrplot()`)
  - ANOVA for group comparisons (`aov()`)

(2) Trend and Cyclical Pattern Analysis

- Goal: Since the dataset contains only annual data, traditional seasonal analysis (spring, summer, autumn, winter) is not possible. Instead, this study will analyze long-term trends and periodic (cyclical) fluctuations in water withdrawals and stress levels.
- Methods in R:
  - Year-over-year differences in water withdrawals (`diff()`)
  - Correlation between economic/climate factors and water stress
  - Identifying multi-year cyclic patterns (`stats::stl()`, `forecast::decompose()`)

Step 3: Supervised Learning

Advanced Regression Models (Machine Learning for Prediction)

- Methods in R:
  - Random Forest (`randomForest`)
  - Support Vector Regression (`e1071::svm()`)
  - Time Series Forecasting (`forecast::auto.arima()`, `prophet`)

Classification Analysis (Predicting Water Stress Levels)

- Goal: Predict a country’s geographic region based on its water utilization, economic indicators, and climate factors.
- Methods in R:
  - Decision Trees (`rpart`)
  - Random Forest (`randomForest`)
  - Support Vector Machines (`caret`)
  - Logistic Regression (`glm`)
  - Multinomial Logistic Regression (`nnet::multinom()`)

Step 4: Unsupervised Learning

Clustering Analysis (Grouping Countries Based on Water Usage Patterns)

- Goal: Group countries with similar water resource utilization patterns.
- Methods in R:
  - K-Means Clustering (`kmeans()`)
  - Hierarchical Clustering (`hclust()`)
  - DBSCAN (`dbscan`)

Dimensionality Reduction (PCA)

- Goal: Reduce dataset dimensions while retaining important features.
- Methods in R:
  - Principal Component Analysis (`prcomp()`)
  - Factor Analysis (`factanal()`)

Step 5: Time Series Forecasting

- Goal: Predict future freshwater withdrawals, water stress levels, and economic/climate trends influencing water resources.
- Methods in R:
  - ARIMA (`forecast::auto.arima()`)
  - Facebook Prophet (`prophet`)

5. Expected Outcomes

This study aims to provide the following insights:

- Identification of global water resource utilization trends based on total and sectoral freshwater withdrawals.
- Comparison of water productivity (economic efficiency of water use) between countries.
- Analysis of how economic factors (GDP per capita, private investment) and climate trends (temperature, sea level rise, extreme weather events) influence water stress levels.
- Sectoral analysis to determine which sectors (agriculture, industry, domestic) drive the highest freshwater withdrawals.
- Regional classification of water usage patterns, highlighting geographic variations in water stress and withdrawals.
- Predictions for future freshwater withdrawals and water stress levels, aiding policymakers in sustainable water management.

6. Timeline

Task	**Duration**
Topic Selection & Proposal Writing	1 week
Data Collection, Cleaning & Preprocessing	2 weeks
Exploratory Data Analysis (EDA) & Feature Engineering	2 weeks
Classical Statistical Analysis (Regression, Correlation, Trends)	2 weeks

Machine Learning Models (Supervised & Unsupervised Learning)	3 weeks
Time Series Forecasting & Validation	2 weeks
Interpretation, Insights, and Model Refinement	2 weeks
Report Writing & Presentation Preparation	1 week
Final Review & Adjustments	1 week
Total duration	16 weeks

7. Challenges & Limitations

- Data is missing for certain years and countries.
- External factors (e.g., policy changes, economic shifts, technological advancements) are not included in the dataset but could significantly influence water resource utilization.
- Limitations on forecasting exist due to the complexity of climate change impacts on water availability and stress levels, as well as the nonlinear interactions between economic and environmental factors.
- Groundwater depletion data is only available for the USA, lacks annual observations, and covers inconsistent time periods, limiting its comparability and use in statistical modeling.

8. Conclusion

This project will provide valuable insights into global water resource utilization patterns, their economic and climate-related drivers, and future trends, contributing to efforts in sustainable water management. The results can aid governments, environmental organizations, and industries in optimizing water usage, mitigating water stress, and making data-driven policy decisions for long-term resource sustainability.