

Statistical Research: Water Resource Analysis with R

01.05.2025

Seunghyun Kim
Niran Raj Pradhan
Heejung Heo

Table of Contents

Research Objectives

Data Overview & Preprocessing Steps

1. **Data Description**
 2. **Data Preprocessing**
 - 2.1 R Libraries Used
 - 2.2 Missing Data Assessment and Handling
 - 2.3 Data Overview After Handling Missing Data
 - 2.4 Additional Feature Derivation
 3. **Summary**
-

Supervised Learning: Predicting Population Growth Categories

1. **Introduction**
 2. **R Libraries Used**
 3. **Labels and Features**
 - 3.1 Label Construction
 - 3.2 Feature Selection and Summary Statistics
 4. **Classification Methodology**
 - 4.1 Candidate Model Selection
 - 4.2 Nested Cross-Validation Procedure
 - 4.3 Evaluation Metric Selection
 5. **Model Evaluation and Comparison**
 - 5.1 Baseline Model Introduction
 - 5.2 Model Performance Summary and Visualization
 - 5.3 Pairwise Model Comparison via Statistical Testing
 6. **GLMNET Interpretation and Error Analysis**
 - 6.1 Variable Importance, Coefficients, and Feature Effects
 - Variable Importance Rankings
 - Coefficient Interpretation
 - Partial Dependence Analysis
 - 6.2 Country-Level Misclassification Analysis
 7. **Reflections and Possible Applications**
 8. **Conclusion**
-

Unsupervised Learning: PCA and Clustering Analysis

1. **Introduction**
 2. **R Libraries Used**
 3. **Clustering Methodology**
 - 3.1 Feature Selection & Normalization
 - 3.2 Model Selection
 - 3.3 Principal Component Analysis
 - 3.4 Grid Search
 4. **Results and Interpretation**
 - 4.1 Clustering Configuration Selection
 - 4.2 Evaluation and Visualization
 - 4.3 Cluster Profiles
 5. **Discussion on Policy Implications**
 6. **Conclusion**
 7. **Conclusion**
-

Time Series Forecasting: Freshwater Withdrawals Prediction

1. **Introduction**
 2. **R Libraries Used**
 3. **Variable and Country Selection Rationale**
 - 3.1 Target Variable: Freshwater Withdrawals
 - 3.2 Target Country: Germany
 4. **Forecasting Framework**
 - 4.1 Arima Forecasting
 - 4.2 Prophet Forecasting
 5. **Forecasting Results and Evaluation**
 - 5.1 Arima Forecasting
 - 5.2 Prophet Forecasting
 - 5.3 Model Performance Evaluation and Comparison
 6. **Interpretation of Forecast Results and Conditional Policy Recommendations**
 - 6.1 Potential Drivers Behind the Forecasted Withdrawal Trend
 - 6.2. Scenario-Based Policy Recommendations
 7. **Conclusion**
-

Data Sources, References, and R Scripts

- Data Sources
 - References
 - R Scripts
-

Appendix

- Model Explanation
-

Research Objectives

Water is a fundamental resource essential for human survival, economic development, and environmental stability. Yet growing water stress concerns—driven by population dynamics, climate change, and economic activity—pose a major global challenge. A deep understanding of water-use dynamics, and their links to economic and climate factors, is therefore critical for informed policymaking and sustainable management.

This study analyzes global and country-level water-use patterns by combining classical statistical techniques with machine learning workflows in R. Water-related variables were examined through a structured pipeline comprising exploratory data analysis, supervised learning(binary classification), unsupervised learning(PCA & clustering), and time-series forecasting. A major emphasis was placed on rigorous validation and evaluation, such as constructing a reliable analytical workflow, optimizing model performance, critically reviewing assumptions and limitations, and examining generalizability of the models—to ensure that all findings are both reliable and applicable.

