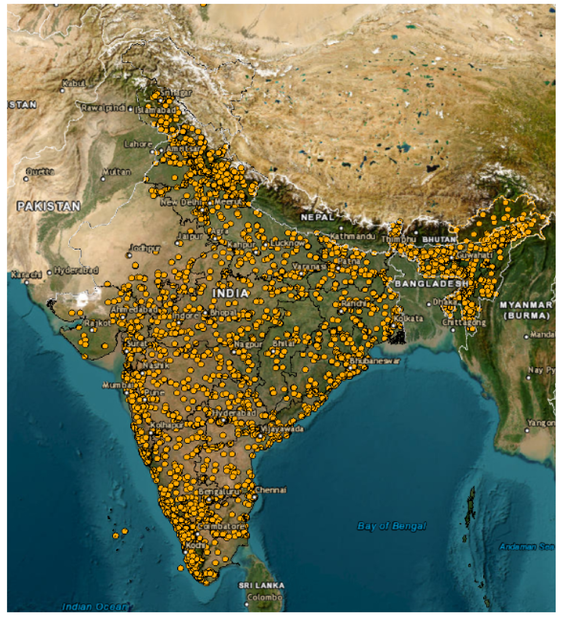
Data Preprocessing and Geospatial Analysis of WQI

2430521 | Shivani Gupta | Mentor: Dr. Shankar Prawesh
Department of Management Sciences

INTRODUCTION

This project involves analyzing mineral concentrations from various monitoring stations across India, using a dataset of over 61,677 entries from more than 30,000 water bodies and 84 physio-chemical parameters. Our objective is to preprocess and clean this data to provide reliable insights into Indian water bodies' mineral concentrations.



DATA DESCRIPTION

The dataset includes 61,667 data points covering key minerals such as calcium, chloride, electrical conductivity, fluoride, bicarbonates, potassium, magnesium, sodium, and pH. Out of 84 parameters, only 18 have more than 45% data availability, with 9 key parameters having over 85% data coverage.

REFERENCE

1. India-WRIS
2. GEMStat - The global water quality database

PROBLEMS IDENTIFIED

Data Sparsity

Many minerals (e.g., silver, aluminum, aldrin, barium, etc.) have extremely sparse data, limiting comprehensive analysis.

Data Consistency

Discrepancies in longitude and latitude values due to new monitoring sites and spelling errors. Lack of uniformity in data recording and entry.

Incomplete Dataset

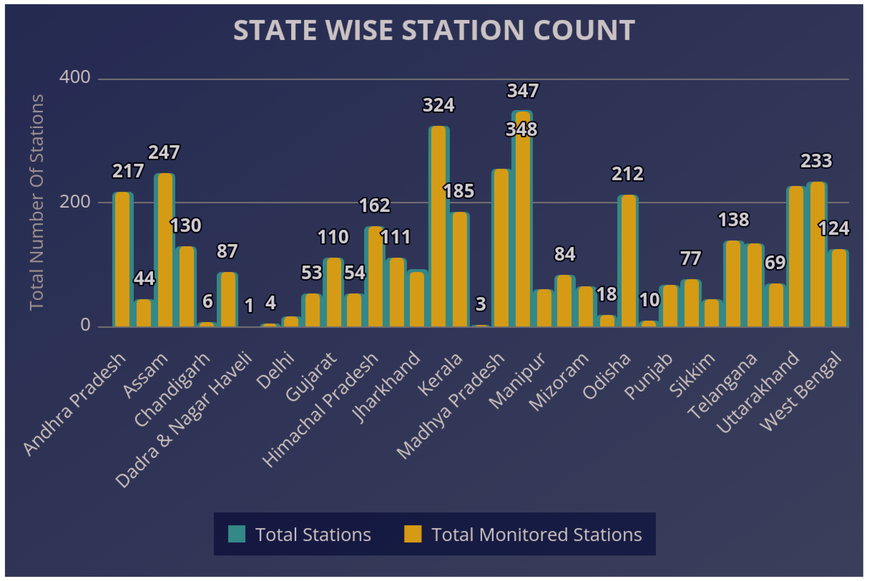
Missing values for important minerals on the India-WRIS site, affecting dataset comprehensiveness.

Manual Corrections Required

Errors identified through mismatches or error messages during the data integration process.

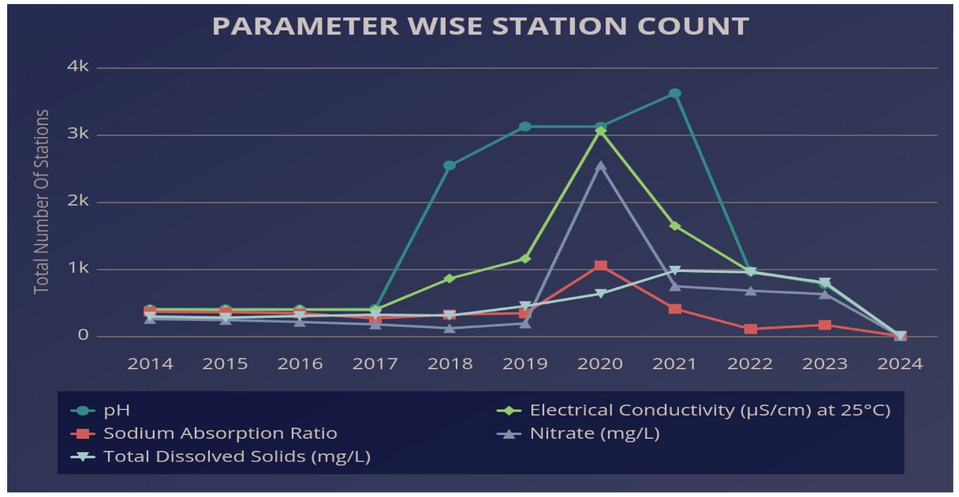
Geographic Gaps

Absence of monitoring sites in various regions of India, leading to incomplete geographic coverage and potential data gaps.



CONCLUSION

Preprocessing and cleaning the dataset is crucial for accurate and reliable analysis. Addressing data sparsity, consistency, completeness, and geographic gaps significantly enhances the quality of research. Implementing these solutions enables more robust and comprehensive studies of mineral concentrations across India.



SOLUTIONS AND METHODOLOGIES

Data Imputation

Employ statistical methods or machine learning techniques to impute missing values. Use multiple imputation techniques to handle uncertainties in imputed values.

Data Standardization

Standardize names and locations of monitoring sites to ensure consistency. Use geocoding and fuzzy matching techniques to correct spelling errors and align old and new site names.

Data Augmentation

Supplement the dataset with additional data from reliable sources to fill in gaps. Use external databases and previous research to cross-validate and augment existing data.

Geospatial Data Integration

Utilize pre-existing datasets containing latitude, longitude, and WRIS IDs up to the year 2018. Apply the Index-Match function in Microsoft Excel to cross-reference and retrieve corresponding geospatial data. Incorporate matched latitude, longitude, and WRIS IDs into the new dataset.

Manual Corrections

Identify errors through mismatches during the Index-Match process. Manually locate and integrate correct WRIS IDs with the corresponding latitude and longitude.

Expanding Monitoring Coverage

Advocate for the establishment of new monitoring sites in underrepresented regions. Use remote sensing and satellite data to estimate water quality in areas lacking direct monitoring.

Times New Roman