The Analysis of Oceanic Oxygen Concentration and Influencing Factors

Saiem Ilyas, Riyaan Ahmed, Trisha Bhatnagar, Yoatam Gebremicael

Problem

The primary objective of this study is to investigate how oceanic temperature, salinity, and geographic location influence oxygen concentration in seawater. Oxygen levels in the ocean play a crucial role in sustaining marine ecosystems, influencing aquatic life, and regulating biochemical cycles. Variations in temperature and salinity significantly affect the solubility and availability of oxygen, while geographic location determines exposure to oceanic currents and environmental changes.

By analyzing the relationships between these variables, we aim to identify patterns and trends that provide a deeper understanding of oxygen concentration fluctuations. This is essential for marine biology, climate studies, and ecosystem sustainability, as it allows scientists to predict hypoxic conditions and take necessary conservation measures.

Impact

The successful execution of this study has the potential to make meaningful contributions to oceanographic research, environmental conservation, and climate change mitigation.

- **Institutional Impact:** This study helps ocean research centers learn why the sea has less oxygen. It offers facts that build future models plus guide regular checks.
- **Ecological Benefits:** The results help choose ways to keep the sea safe, especially for creatures that depend on plenty of oxygen. Fishers use this information to pick better fishing areas and stop catching too many fish in weak spots.
- **Policy & Decision-Making:** Finding areas with little oxygen helps shape rules for guarding the sea moreover controlling the climate. Government groups use these facts to write safety plans and set rules that ease the harms of climate change on the sea's oxygen.

Data

The dataset consists of **504 observations**. It shows oxygen level temperature, salinity as well as location. The data comes from ocean sensors and sample bottles and it gives us complete checks on environmental conditions.

Data Cleaning & Preprocessing:

During analysis, certain **missing values** were identified, particularly in the **micromoles_of_oxygen_per_unit_mass_in_sea_water** variable. These gaps in data required handling through appropriate statistical techniques such as mean imputation or regression-based filling methods to maintain analytical integrity.

Limitations:

The dataset will show some useful facts but it has a few weak points:

- **Additional Variables Needed:** The inclusion of **ocean depth, pressure, and biological activity** would enhance the precision of the analysis.
- **Potential Biases:** The dataset may not account for external environmental factors such as ocean currents, pollution, or seasonal variations, which could affect oxygen distribution.

Methods

Variables

The dataset includes several key variables essential for analysis:

- **best_Oxygen:** Oxygen concentration measured through sampling bottles.
- **micromoles_of_oxygen_per_unit_mass_in_sea_water:** Sensor-measured oxygen concentration.
- **CTDTEMP:** Temperature of seawater.
- **CTDSAL:** Salinity of seawater.
- **Latitude & Longitude:** Geographic coordinates of measurement points.

No new variables were created, but **transformations and imputation** were applied where necessary to manage missing values.

Visualizations

We will try different ways to show plots that reveal links in the data:

- **Pairplot:** To examine correlations between oxygen levels, temperature, and salinity to detect patterns and interdependencies.
- **Time-Series Plot:** To illustrate temporal variations in oxygen levels over time and highlight seasonal or cyclical trends.
- **Scatterplot:** To provide a spatial representation of oxygen concentration across different geographic locations, revealing regional disparities in oxygen levels.

The **time-series plot** will be particularly significant in identifying trends over time, helping to understand long-term fluctuations in oxygen concentration and potential environmental impacts.

Models and Statistical Tools

- **Correlation Matrix:** To find relationships between variables, revealing expected strong negative correlations between oxygen concentration and salinity as well as temperature.
- **Descriptive Statistics:** To provide insights into data distribution, including mean, standard deviation, and range.
- **Linear Regression:** To examine the dependency of oxygen concentration on temperature and salinity.
- **Random Forest Regression:** To provide a non-linear approach to model complex interactions between variables, ensuring robustness in predictions.

Challenges Encountered:

- **Handling Missing Data:** The presence of missing values in the sensor-measured oxygen variable may create limitations in conducting a comparative analysis.
- **High Correlation Between Variables:** A strong correlation between oxygen levels and salinity may require careful interpretation to avoid multicollinearity issues in regression models.

Validation Techniques:

- **Train-Test Split:** We will divide the dataset into training and testing subsets to evaluate model performance.
- **Cross-Validation:** We will implement cross-validation to ensure model robustness and prevent overfitting.

Concerns

Dataset Integrity:

The presence of missing values in the **micromoles_of_oxygen_per_unit_mass_in_sea_water** variable limited the depth of analysis. Although statistical imputation was applied, the accuracy of results could have been improved with complete data.

Potential Drawbacks of Analysis:

- **Correlation vs. Causation:** While significant correlations are expected, external factors such as ocean currents, biological activity, and human-induced changes may also influence oxygen levels but are not accounted for in this dataset.
- **Simplification of Complex Interactions:** Some visualizations may present an oversimplified view of the relationships between variables, requiring more advanced modeling techniques to uncover hidden patterns.