Exploring the Relationship Between Water Salinity and Temperature in the California Current

Yiheng Yao
University of Michigan
yyiheng@umich.edu

December 9, 2024

1 Introduction

The California Cooperative Oceanic Fisheries Investigations (CalCOFI) dataset is one of the most comprehensive oceanographic time series, providing valuable insights into the marine ecosystem of the California Current. Salinity and temperature are crucial parameters that influence marine life, nutrient cycles, and ocean circulation. Understanding the relationship between these variables can provide insights into broader environmental processes such as climate change and ocean health.

This research aims to explore the relationship between water salinity and temperature in the California Current. Specifically, we investigate whether salinity, combined with depth, can accurately predict water temperature. To achieve this, several modeling approaches, ranging from linear to non-linear models, are employed and compared in terms of accuracy and interpretability.

2 Methodology

Data from the CalCOFI dataset, including temperature, salinity, and depth variables, were used for this study. The dataset was first preprocessed to handle missing values, and exploratory analysis was conducted to understand basic trends and relationships. To model the relationship between temperature and salinity, we used the following approaches:

- **Basic Linear Regression**: This model serves as a baseline for assessing the linear relationship between salinity and temperature. The simplicity of this method makes it an effective starting point, providing a straightforward interpretation of how salinity influences temperature.
- **Multiple Linear Regression**: In addition to salinity, depth was included as a predictor to understand its impact on temperature. This method helps to assess how multiple factors interact to influence temperature, providing a more comprehensive understanding compared to simple linear regression.

- **Polynomial Regression**: To capture non-linear relationships, polynomial regression with degrees of up to four was used. Polynomial regression is useful when the relationship between predictors and response is more complex, allowing us to fit curves that better capture non-linearities in the data. Cross-validation was applied to evaluate model performance.

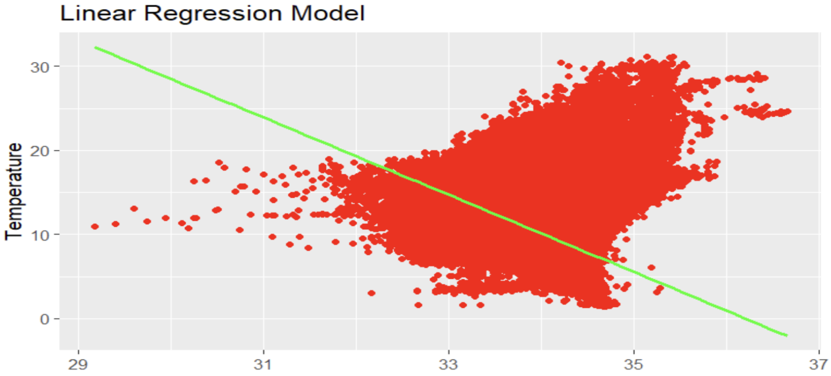
- **Decision Tree Regression**: Decision trees provided a flexible way to handle non-linearities without requiring feature scaling. Decision trees are highly interpretable and can capture complex interactions between variables. Cross-validation was applied to evaluate model performance.

- **Random Forest Regression**: Random forests, an ensemble learning method, were used to improve accuracy and provide insight into feature importance. Random forests combine multiple decision trees to reduce overfitting and increase prediction accuracy. Cross-validation was applied to evaluate model performance.

By combining these methods, I will gain a more reliable assessment of the model's true performance, ensuring that findings are both reproducible and applicable to real-world scenarios. This balanced approach enhances the robustness and validity of results. Metrics such as Root Mean Squared Error (RMSE) was used to assess the models' effectiveness.

3 Results

The **basic linear regression** model provided a simple and interpretable relationship between salinity and temperature, with an adjusted R^2 value of 0.2507 and an RMSE of 3.653, suggesting a bad linear correlation. From figure1, it seems that we need more methods to improve.

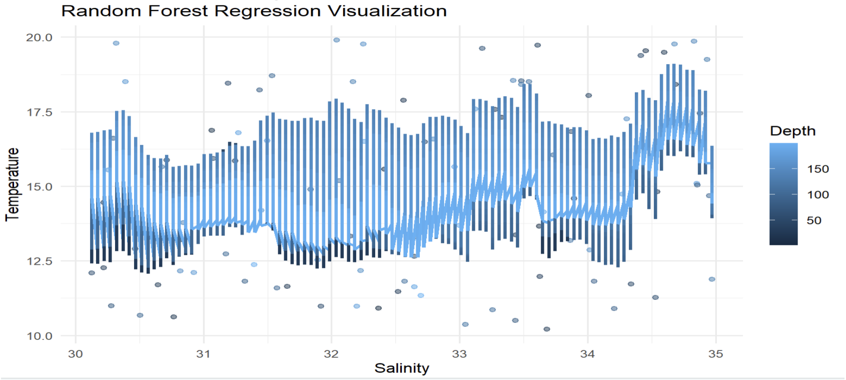


Adding **depth** as an additional predictor in the **multiple linear regression** model improved the adjusted R^2 value to 0.475 and reduced the RMSE to 3.064, indicating that depth plays a significant role in explaining temperature variation.

The **polynomial regression** model changes the RMSE to 3.068, capturing some of the non-linear patterns present in the data. However, there was a risk of overfitting at higher polynomial degrees, which was mitigated using cross-validation.

The **decision tree** model, while highly interpretable, resulted in overfitting, especially when the tree was allowed to grow without constraints. The cross-validated RMSE for the decision tree model was 3.124. Despite this, it provided valuable insights into the decision rules, such as the importance of certain depth ranges in predicting temperature.

The **random forest** model performed best among all approaches, achieving the lowest cross-validated RMSE of 3.053. Random forest also provided insights into **feature importance**, revealing that depth had a slightly higher influence on temperature than salinity. This model's robustness can be attributed to its ensemble nature, which reduces variance and improves generalization.



4 Discussion

The results demonstrate that while **linear models** provide a good baseline and are easy to interpret, they may not fully capture the complex relationship between salinity, depth, and temperature. **Polynomial regression** added flexibility to the model, helping to address non-linearities, but required careful handling to avoid overfitting. **Tree-based models**, particularly the **random forest**, proved to be the most effective in terms of predictive accuracy.

The random forest model's ability to capture complex interactions makes it an ideal choice for modeling the relationship between oceanographic parameters in dynamic environments like the California Current. However, the model's complexity comes at the cost of interpretability, making it challenging to extract simple, actionable insights about the influence of individual predictors.

5 Conclusion

In this study, we explored the relationship between salinity, depth, and temperature in the California Current using a variety of modeling techniques. Our findings indicate that both salinity and depth are important predictors of temperature, with depth showing a slightly greater influence. **Random forest** emerged as the best-performing model, while simpler models such as **linear regression** offered ease of interpretation but at the cost of accuracy.

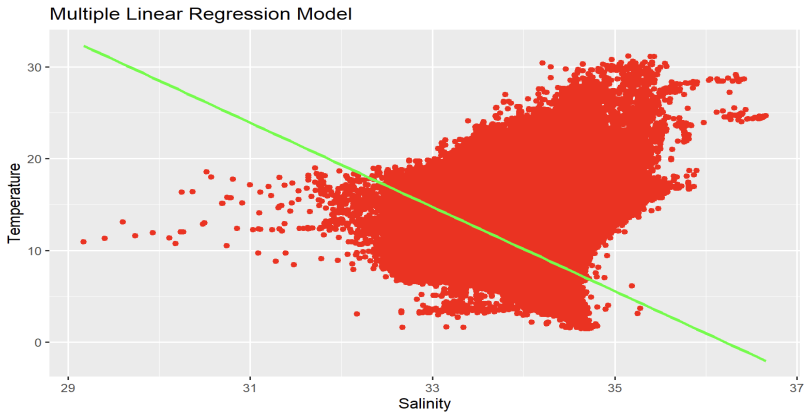
Future work could include exploring other features from the CalCOFI dataset, such as nutrient concentrations or seasonal variations, to further improve temperature predictions. Additionally, considering time-series models could provide deeper insights into the temporal dynamics of oceanographic changes.

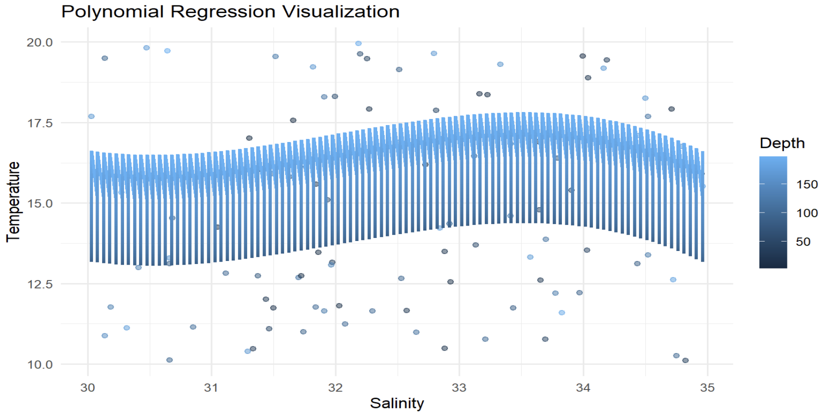
6 Link To Github

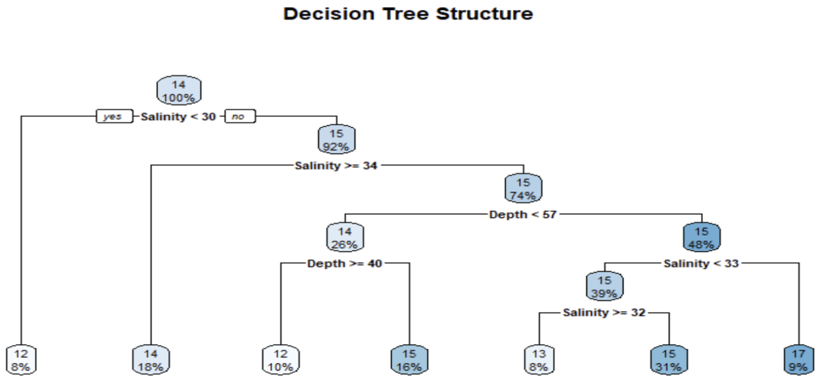
https://github.com/yyhken/stats-506/tree/main/506

7 Appendix

Model Evaluation Metrics


Multiple Linear Regression Model


Polynomial Regression Visualization


Decision Tree Structure

Reference

García-Reyes, M., Koval, G., & Vazquez-Cuervo, J. (2024). Characterizing the California Current System through Sea Surface Temperature and Salinity. *Remote Sensing*, *16*(8), 1311. https://doi.org/10.3390/rs16081311