Using Ordinal Regression Analysis For Probabilistic Ranking And Identification Of Sustainable Carbon Storage Sites In California

Rudraksh Mohapatra

Nov 19, 2024

Stanford | Doerr School of Sustainability | Stanford Center for Carbon Storage

Research objective

- Much of the existing literature on CCS focuses on identifying the geologic and economic potential of sites within a given area, often near emission sources (sink-source matching) or already well explored geologic formation like saline aquifers and unmineable oil fields.

- In existing literature, an aspect of the site selection process remains unaddressed, which is the geographical location of the sites within a larger system.

- This research aims to develop a method of ranking carbon storage sites in California, based on economic, social, geographical and geological factors.

Research Questions

- How can we use machine learning to rank carbon storage sites?
  Ordinal regression gives us the probability of any location being a viable site, which can be used to rank them.

- Which technical input features can be used?
  - Saline Aquifers
  - CO2 Intensity
  - Seismic moment
  - Distance to fault lines
  - Distance to natural gas pipelines
  - Distance to power plants

- Finally, what is the relative importance of each feature?
  It was found that Saline is the most important feature, followed by CO2 Intensity and distance to natural gas pipelines

Methodology

1
- **Data Collection and Processing**
  Data for the input and output features were collected from various publicly available sources, and were processed in a way that made it easy to apply computational methods

2
- **Applying classification models**
  Using the Yggdrasil Decision Forest library, three classification models were applied: CART, Random Forest, and Gradient Boosted Trees. This helped in figuring out the input features used in the final step.

3
- **Applying regression models**
  Next regression analysis was done using the same models. The input features were normalized, and used in creating a ranked map of all locations. This gave us a reference for the final output.

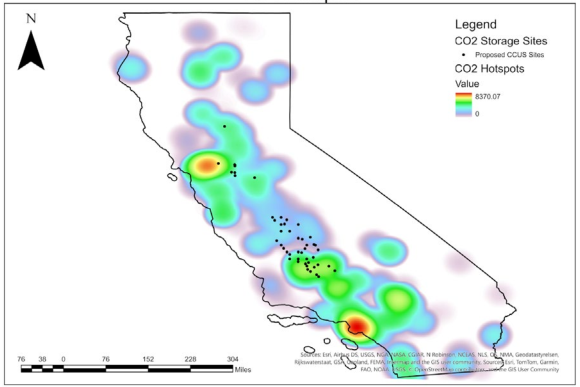
4
- **Ordinal regression**
  Finally, the ordinal regression model was applied using a logit distribution. For the input, equal number of location that were sites, as well as non-sites were used, the result was a probability based ranking of the entirety of California

5
- **Final results**
  Finally, the top sites were selected based on a threshold defined by the probability. In addition, the sites were overlaid with exclusion zones (areas that are densely-populated, protected lands, critical habitats, near faults)

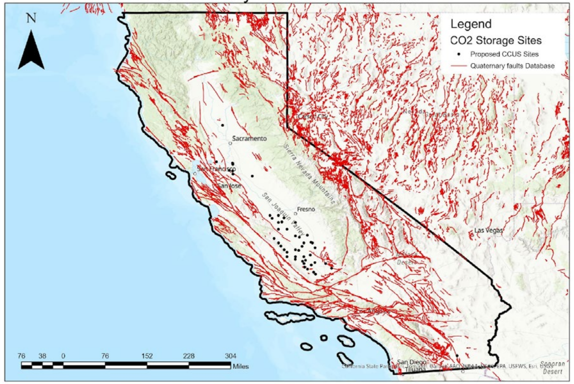
Input Features Used

- After going through the exploratory data analysis step, the following input features were used:
  - Saline Aquifers
  - CO_2 Intensity Hotspots
  - Seismic moment
  - Distance to fault lines
  - Distance to natural gas pipelines
  - Distance to power plants
- Some of the maps used for input features are shown:

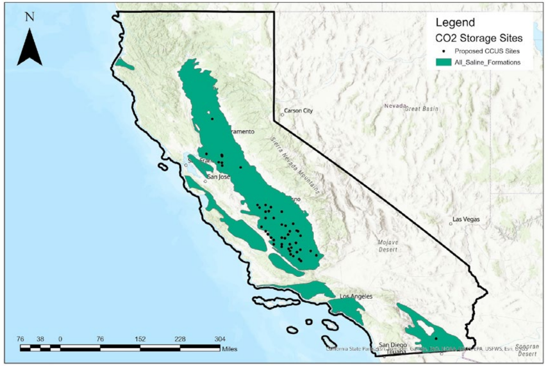
CO2 Hotspots



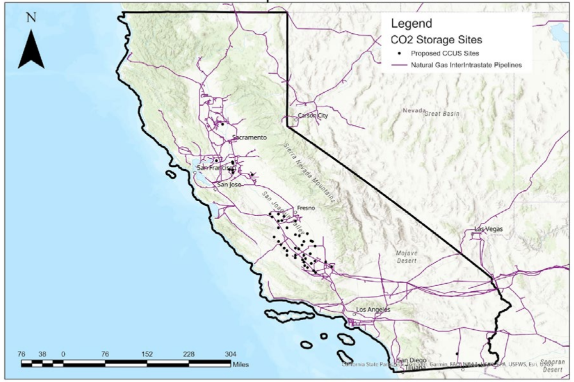
Quaternary Faults in California



Saline Formations in California

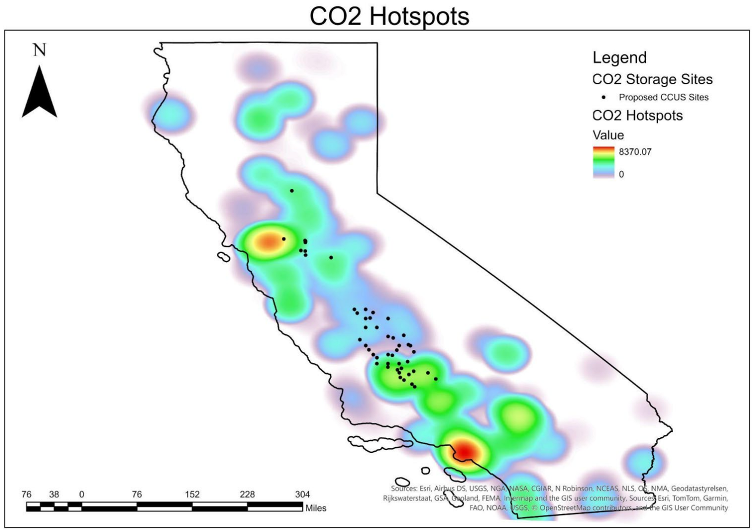


Natural Gas Pipelines in California



Motivation Behind Each Input

- There are two features which require more detailed explanation:
  - The Powerplants input
  - The CO2 Hotspots input
- For these features, we needed a method of accurately capturing both the scale of the powerplants output and co2 emissions, while also taking into account the distance from these sources.
- The Kernel density function in GIS was used for this. This allowed us to use one input instead of two (scale of source + distance from source)



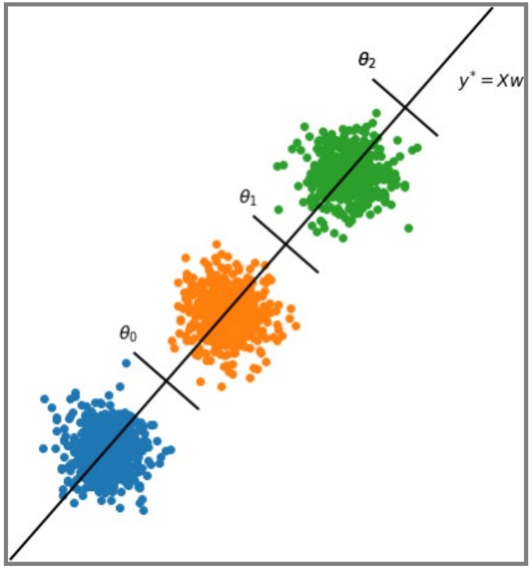
Applying ML Models

- There were three main machine learning techniques explored in the study.
    1. Classification Models
    2. Regression Models
    3. Ordinal Regression Models

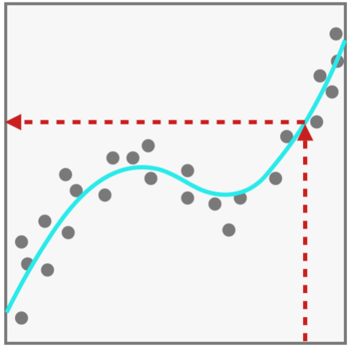
- The classification and regression models served two main purposes. By applying the classification models we could determine which input features were the most important – and only include those for the final model.

- The regression model used a formula we derived to calculate some initial results, which gave us an understanding of how the model's final result should look like.

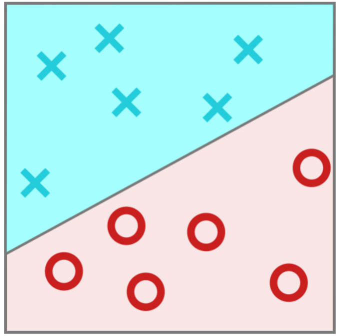
Ordinal regression gives the probability of an observation being a class



Regression predicts a numeric value



Classification Groups observations into "classes"



Ordinal Regression

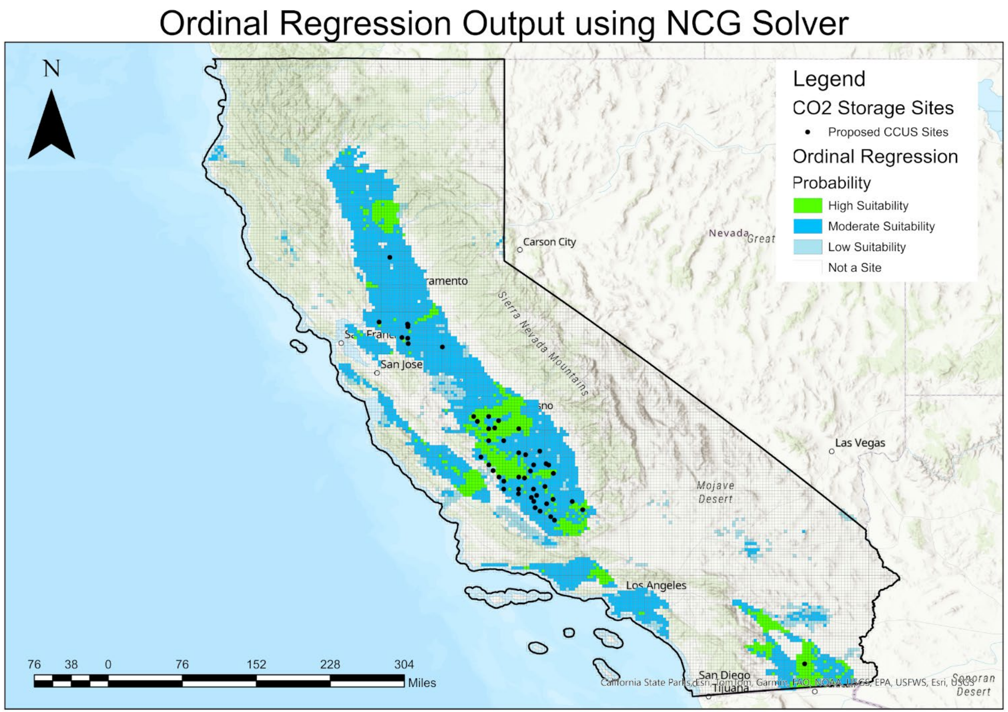
- Ordinal regression gives us the probability of an item to be a certain classification. This probability can be used for a relative ranking of the Carbon Storage sites, as a location with a higher probability is more suited to be a site, than a lower probability one.
- The following probability:

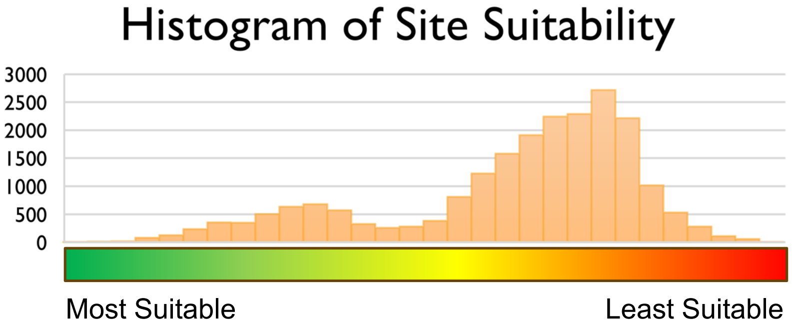
$$log\frac{P(Y \leq j)}{P(Y > j)} = logit(P(Y \leq j)).$$

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \cdots - \eta_p x_p.$$

j =classification x_i = input feature η_i=coefficients for input feature

Final Results





- We see that the model has several areas of interest in the San Joaquin and Sacramento Basins.
- While many of the proposed sites are in suitable locations, their locations can be further improved with insights provided by the model

Research Importance

- Why do we need this model?
  We need a unified model that allows us to screen Carbon Storage sites in a much larger scale. While we may have optimization methods to identify the best sites, being able to do this faster and in a more uniform manner is the need of the hour.

- We already know the technical inputs needed to identify good sites, so what new information is the model providing us?
  So far, we have qualitatively attempted to understand the importance of these models, and in some studies, normalized their values to come up with a generalized method. However, the model gives us an exact mathematical formula that can be applied to any region.

- Why weren't more social features used in the model?
  This was done to ensure that the model could give a mathematical relation to quantifiable variables. The social variables can be changed based on policies of each region, and therefore weren't applied to a more general model like this one.

Conclusions

- **How does the model perform?**
  The model performs pretty accurately, however there is still room for improvement.

  The model gives valuable insights into how well placed the suggested EPA sites are.

- **Limitations of the model?**
  Heavily dependent on the input features used. The model's predictions may vary based on whether the inputs are what humans consider important and what features are actually important.

- **What's next?**
  Currently I am working on improving the model by taking into account social factors like economic development, AQI, and more in order to ensure that communities aren't negatively affected.