Multiple factor regression analysis of vegetation changes and land degradation in drylands

Poster SE2. Author: Maria Chzhen, University of Toronto

Background

Desertification is a form of land degradation in drylands. It is measured with the Normalized Difference Vegetation Index (NDVI).

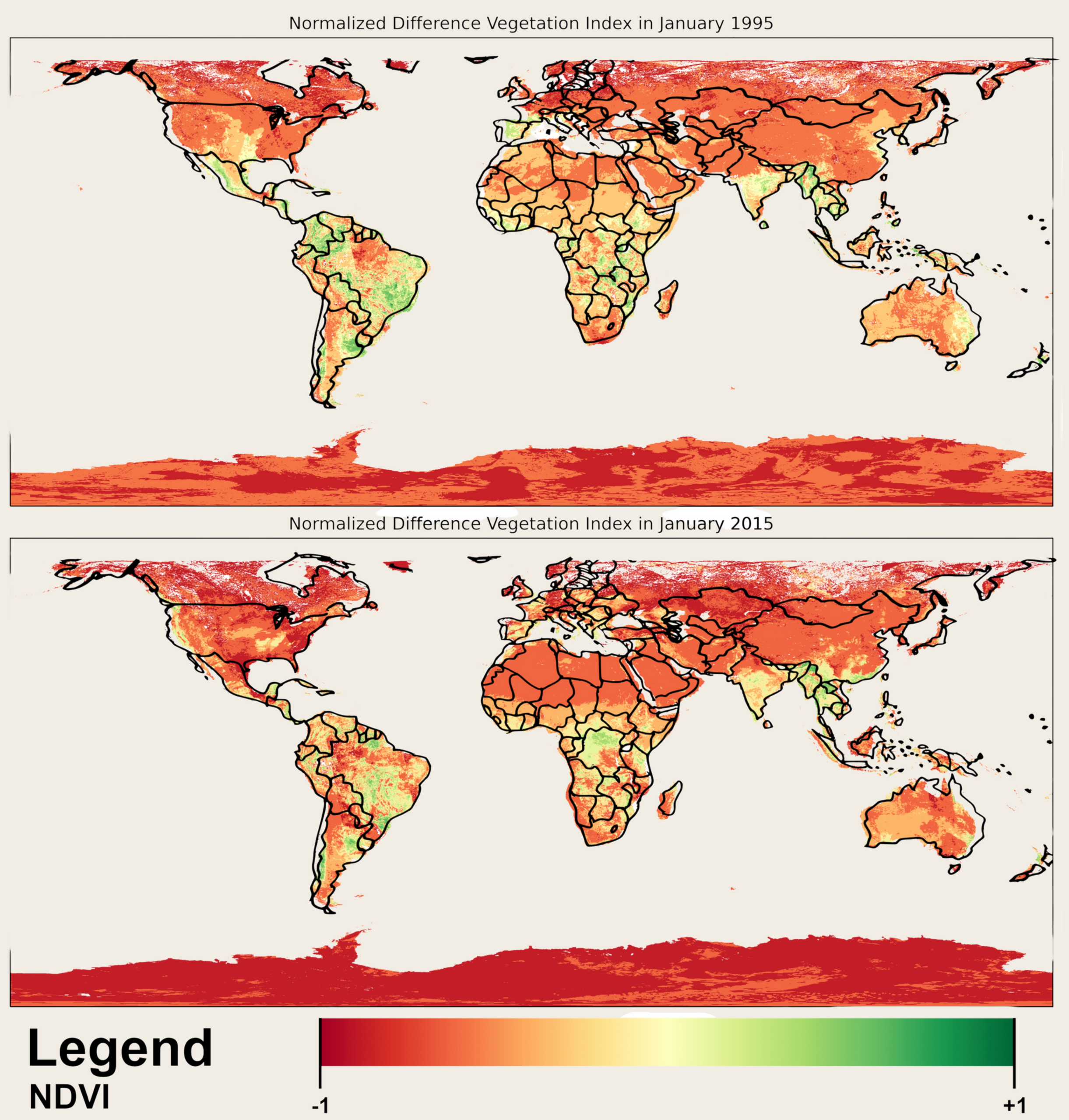

Figure 1: A comparison of NDVI scores between January 1995 and January 2015. The global mean NDVI value has decreased from 0.0875 in 1995 to 0.0750 in 2015 (net decrease of 0.0125). Matplotlib.

Hypothesis

In most regions, the NDVI value will reportedly decrease and each feature will have a varying effect on the vegetation cover in different regions. CO2 emissions will play the greatest role in the mean NDVI values.

Resources

- **netCDF4**. Geoscientific data access.
- **Cartopy, Matplotlib**. Geospatial data processing and visualization.
- **SciPy**. Scientific/technical computing.
- **Scikit-learn**. Machine learning library. Numpy, Pandas.
- **Requests, BeautifulSoup**. Pulling data out of HTML and XML files.
- Advanced Very High Resolution Radiometer (**AVHRR**) database by the National Oceanic and Atmospheric Administration (**NOAA**).
- **TerraClimate**. A database of monthly climate changes.
- **Worldometer, Our World in Data**. Live world statistics.

Methods

1. Geoscientific data (>67GB) was imported using Requests and BeautifulSoup.
2. Data was normalized. Each feature was divided by the maximum value of the feature in a column.

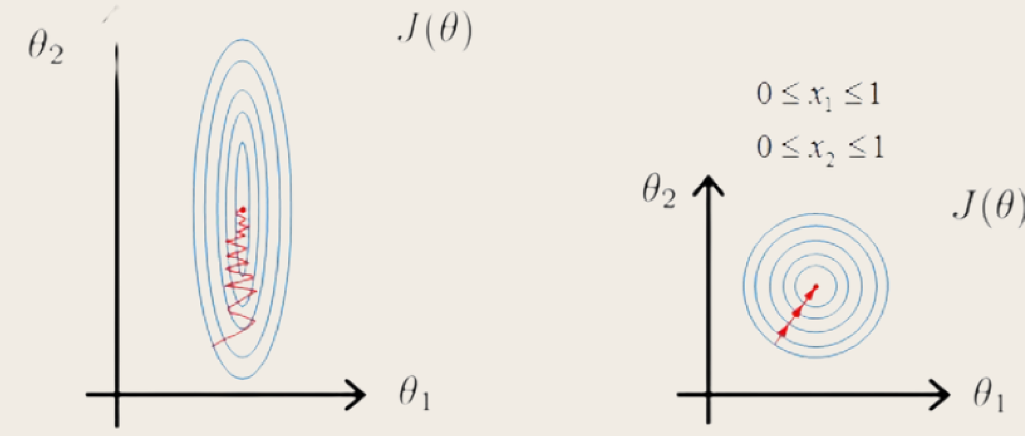

Figure 2: Ease of gradient descent after normalization

3. Multiple factor regression was used to analyze the data.

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \in R^{n+1} \qquad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_n \end{bmatrix} \in R^{n+1}$$

Figure 3: The variables and parameters can be represented in a matrix.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n = \theta^T x$$

Figure 3: Hypothesis of the regression

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\left(\sum_{j=0}^{n} \theta_j x_j^{(i)} \right) - y^{(i)} \right)^2$$

Figure 4: Error function - least-squares.

Evaluating the partial derivative gives the formula for the gradient descent algorithm, which calculates the minimum of a function.

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

Figure 5: Partial derivative of the error function.

4. Linear and asymptotic regression temporal models were generated for each feature and then compared using their R squared values.
5. Prediction intervals were generated for each model.

Purpose

Quantify the effect of various factors, including drought, precipitation, CO2 emissions, forest area, and electrical energy consumption, on the vegetation cover of various regions using multiple factor regression; identify measures that humans can take to increase the vegetation cover.

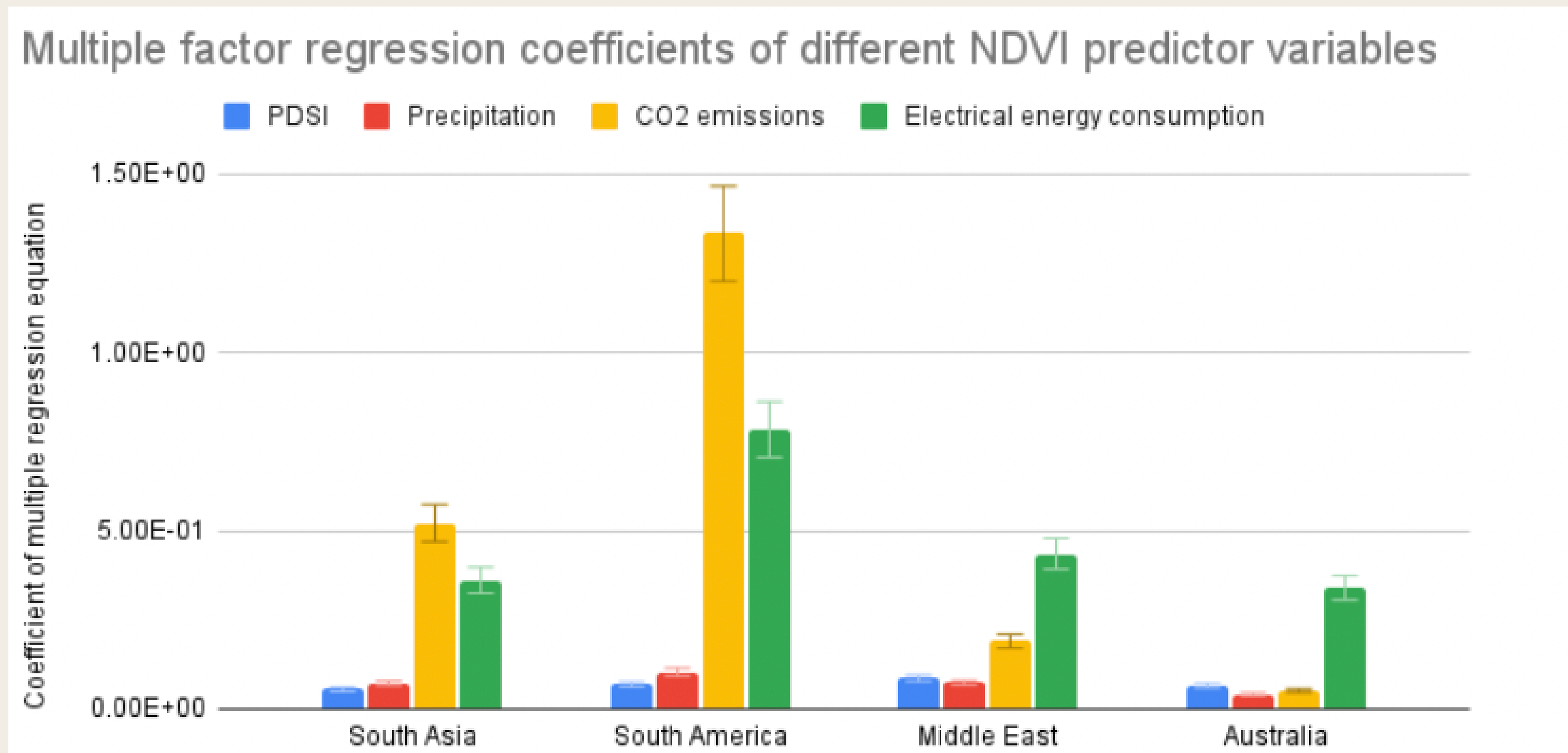

Figure 7: Multiple factor regression coefficients for different NDVI predictor variables.

Results and Discussion

- Identified fundamental causes of decrease in vegetation cover
- Continue to invest in energy alternatives and reducing carbon dioxide emissions
- Pave the way for future prevention of land degradation with identified causes

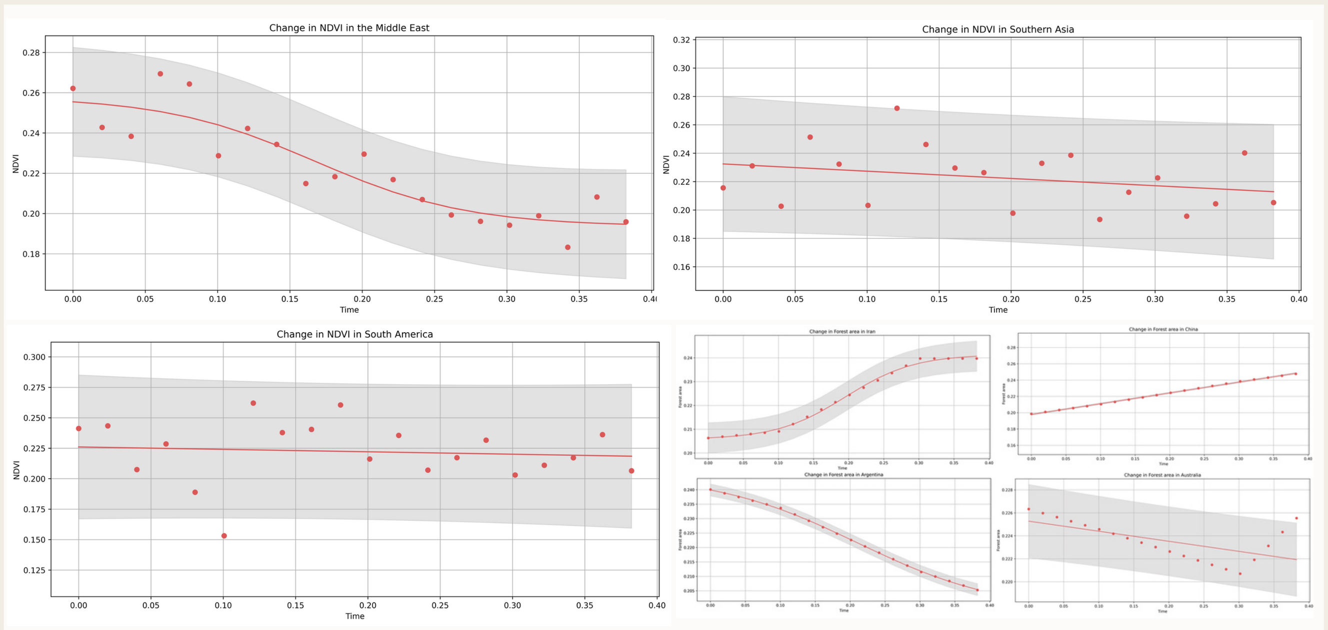

Figure 6: Curve fitting for changes in NDVI and forest area.

- Electrical energy and CO2 emissions both played the most prominent role in NDVI scores.
- Electrical energy consumption was the most significant factor in the Middle East and Australia.
- CO2 emissions were the most significant affecting factor in South America and South Asia.