Methods of Advanced Data Engineering

Project Report

Dushyant Supare

23160892

Summer Semester 2024

1. Introduction

The objective of this report is to investigate the relationship between climate change, that is indicated by the surface temperature and the displacement of people due to disasters in the South Asian region of the world from 2008 to 2013.

Climate change is a well-established reality. As we all know, it causes several changes in the world, the most of which are negative. This research will investigate the link between climate change at the nation level in South Asia and the number of internally displaced people inside a country because of it. The key impact point will be determining which nation was more affected and when it happened, between 2008 and 2013. This is required for the tackling of the underlying causes.

2. Data Sources

There are two datasets in this report.

Datasource 1: Kaggle

Metadata URL: https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

Data Type: CSV

This dataset was uploaded to Kaggle by Berkley earth. This includes the information about surface temperatures. There are many files in this but the current focus in on country level so surface temperatures according to countries are selected. The temperatures are recorded since 1750 up to 2013.

Licence: CC BY-NC-SA 4.0

CC BY-NC-SA 4.0 is a form of Creative Commons licence. This licence enables people to Share (copy and redistribute the work in any media or format) and Adapt (remix, alter, and build upon the material), as long as they acknowledge the original creator, use the material for non-commercial purposes, and licence any new creations under the same conditions.

Datasource 2: Our World In data

Metadata URL: https://ourworldindata.org/grapher/internally-displaced-persons-from-disasters?tab=csv

Data Type: CSV

This dataset contains statistics on the number of persons displaced internally. This information is country-specific and contains the overall number of individuals displaced every year by country. This information is relatively easy to understand.

Licence: CC BY-NC-SA 4.0

Data Pipeline

The data pipeline has been implemented using Python. Where the Kaggle API was used for the first data source, a zip file was obtained, which was then unzipped and placed in a specified folder (data folder). Because a downloadable CSV URL was not accessible for the second data source, it was obtained via Selenium automation and saved to the project folder. The fundamental transformations and data cleansing were completed.

3. Analysis

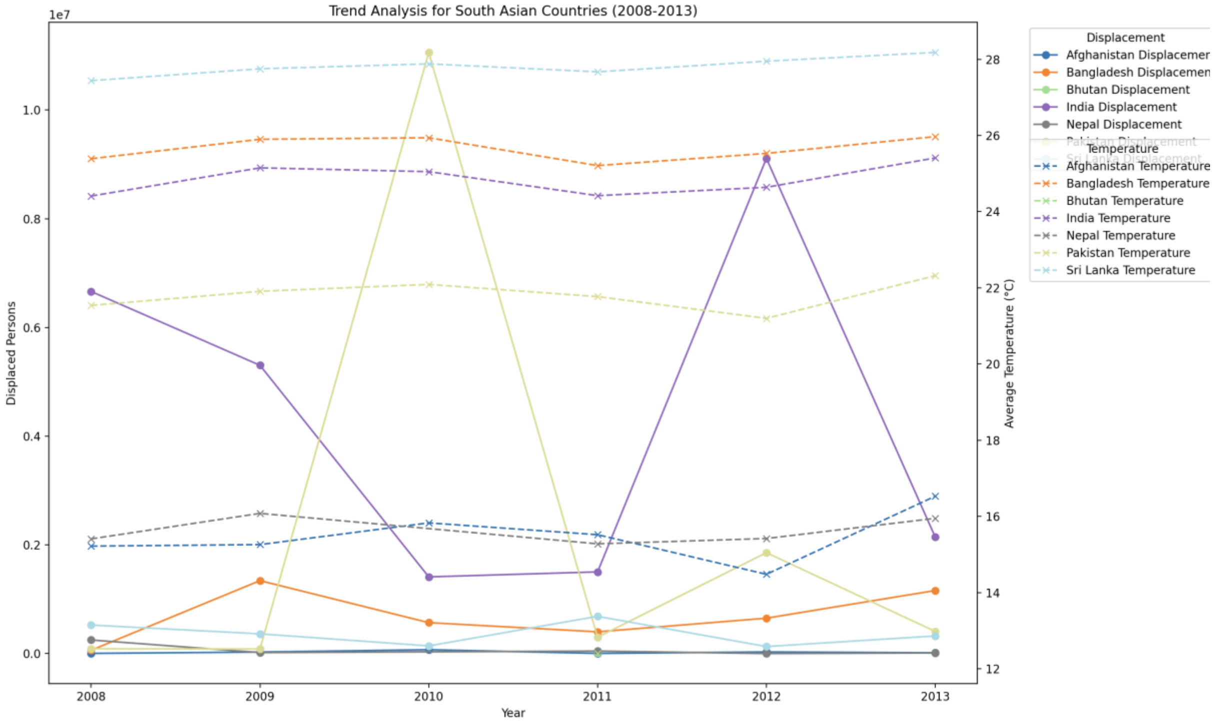
As it was easy to work with the directly fetched CSV's, due to easy accessibility and availability of python libraries to perform better analysis through data frames, the CSV's were used in the further analysis of the data.

To perform the analysis, it was important to join the datasets based on Country and Year. As the Temperature data in dataset 1 was on monthly basis, it was combined and taken average of, so now every country has only one entry of data tuple for each year. As another dataset already has the displacement numbers on yearly basis, this uniformity achievement was necessary.

The datasets were combined, and further analysis was performed, firstly the south Asian countries were defined which includes 'Afghanistan', 'Bangladesh', 'Bhutan', 'India', 'Maldives', 'Nepal', 'Pakistan', 'Sri Lanka'. As the analysis was needed for the years 2008-2013 only that data was selected for further analysis.

After all these initial configurations were done, two charts were plotted, a chart comprising trend analysis and another chart portraying correlation analysis between Displacement and the average temperatures.

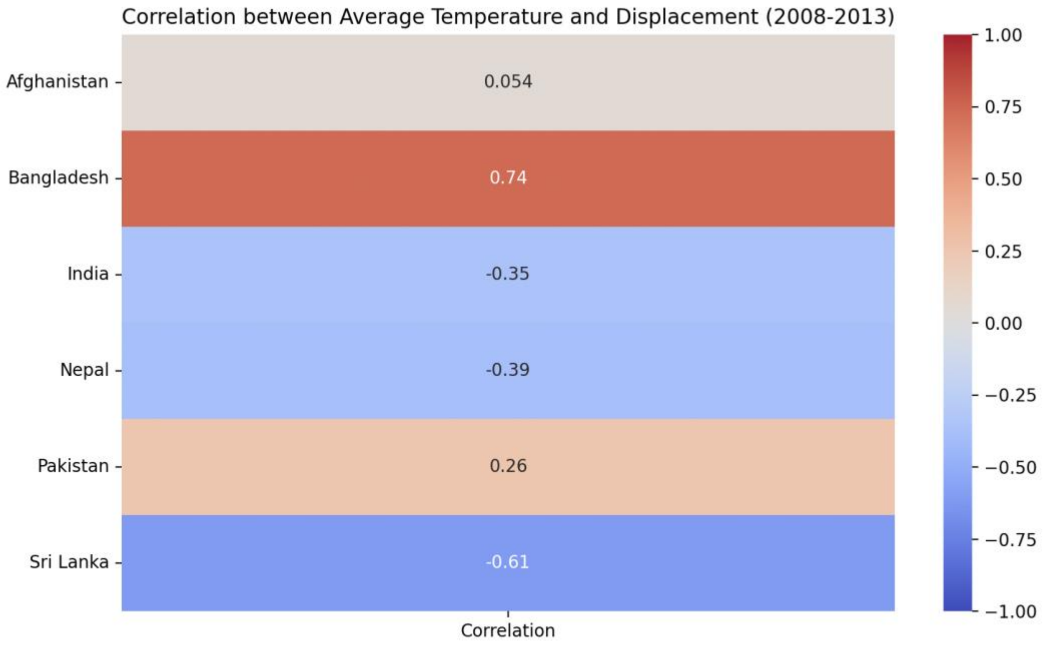
Trend Analysis



This trend analysis was done on the resultant data where the yearly displacement count and yearly average temperature was plotted. It is evident from the picture that there is a steady growth in yearly annual average temperatures, signifying the climate change, from 2008-2010 with a slight drop in 2011 for some countries and 2012 for other, but it has again started increasing in 2013.

When it comes to displacement figures the yearly count is variable, this is an indication that climate change cannot be the only factor in internal displacement of people, there can be more factors. For India and Pakistan, the trend is hugely variable as there is a drastic increase of displaced people in Pakistan in 2010 and in India in 2012.

Correlation Analysis



The correlation analysis gives a clearer insight. It is evident from the correlation analysis that out of the 6 countries, 3 has negative and 1 has a near zero correlation coefficient, that means there is a no direct relation between the surface temperature and total number of displaced people. Only Bangladesh stands out with a correlation analysis of 0.74 signifying the correlation of these two factors in the context of this country.

Overall, it can be said that climate change is not the only cause for displacement of people internally.

4. Conclusion

The study shows that four of the six South Asian nations investigated had a negative or weak association between average temperature and displacement. This shows that temperature fluctuations do not directly correlate with disaster-related displacement in these nations.

The weak or negative association in most nations shows that variables other than average temperature variations may have a greater impact on disaster-related relocation. These characteristics may include socioeconomic status, disaster preparation, and local environmental circumstances. This suggests that a one-size-fits-all strategy for climate adaptation and crisis management may be ineffective. Instead, localized measures that consider each country's specific environment and weaknesses are required.

This study's limitations include the relatively short time period of the data and the aggregate of displacement estimates. The poor connections discovered might be attributed to the intricacy of the processes driving displacement, which cannot be reflected by temperature data alone.

Future study might expand the analysis over longer time periods and include more detailed data, such as additional environmental elements, socioeconomic characteristics, and particular types of catastrophes. Furthermore, qualitative research might give deeper insights into the local environment and aid in understanding the causes of displacement in each nation.