Climate Refugees in Brazil: A Case Study

Connor Hall, Lorenzo Salgado, Rebecca Wolf

Abstract

The aim of this research is to produce a proof of concept in predicting when a large number of people are going to be displaced in a certain state at a certain time in Brazil. Mean temperature, cumulative precipitation, GDP and population were the indicator variables used. A Random Forest Classifier and a Logistic Regression were used to predict these instances. The Random Forest Classifier performed relatively well, while the Logistic Regression was little better than a coin flip. Our team met many obstacles in this project, resulting from a dearth of data availability at the state level and poor data reliability since displaced people are difficult to track and count. This paper serves as a starting point that further research efforts can build off of and a jumping off point for what doesn't work.
Key words: Climate Displaced People, Brazil, Random Forest, Logistic Regression, SMOTE

Introduction

Climate change is acknowledged as a serious threat across the world, and it is creating a new group of refugees and displaced people, climate refugees. Existing literature cites the link between climate change and conflict as a result of increased migration [1]. Links also exist in declines in agricultural productivity that lead to food shortages, water scarcity, and competition for gas and oil resources [2,6]. Previous literature has emphasized the prevalence between conflict and migration while explaining the difficulties of linking climate change to migration [6].

Climate refugees have an uncertain legal status. According to the 1951 UN Refugee convention, the definition of refugees does not include those displaced due to environmental factors [2,8]. Some refugees are not logged in data sets until obtaining official refugee status, which can take many months, up to several years [2]. This brings up the problem of validity in using data over a certain temporal frame, as well as the degree of 'completeness' of the UN Refugee Agency's (UNHCR) refugee data and the Internal Displacement Monitoring Centre's (IDMC) internally displaced peoples data [7,8]. These bureaucratic factors can negatively impact knowledge about the number and location of refugees around the world.

Climate change and its effects disproportionally affect less-developed countries (LDCs) compared to more-developed countries [4,5]. Beyond the obvious effects of extreme weather events, the limited resources and investments in preventative infrastructure, as well as dependence on agriculture for sustenance, makes the issue of climate migration in and from LDCs interesting to explore. [5]

This paper investigates the phenomena of climate refugees in Brazil. Analysis is done at the level of the federative units of Brazil, which is a developing nation with great extremes in wealth, geography, and industries from state to state. These disparities are best exemplified by the differences in amount of rainfall between different regions of Brazil - with the Amazon receiving 2000-3000 mm of precipitation compared to the northeast region where drought is common. [3]

Methods

The response variables data were from the IDMC, and went from 2008 to 2019. However, before 2012 all of the dates listed were January first of the year, likely a placeholder date. Some data also lacked a location listed, or identified a region or multiple states that the disaster took place in. The number of displacements recorded for those data points were not properly sorted. As a short fix, the data were divided into the number of states involved in the disaster.

Thresholds were set for large scale displacements of people. Initially 0.1% of the state's population was the threshold for whether a displacement event was of interest. However, it was decided that the amount of people displaced was of primary interest for disaster planning, in terms of resource and relocation. The threshold was then set at 3,000 people displaced. This threshold was chosen as it is half of a standard deviation above the average event displacement. In the final response vector, rows with less than 3000 people displaced were set to zero, and those with greater than 300 were set to 1.

Weather data was acquired from the Brazilian National institute of Meteorology (INMET) [10]. Three weather stations per state were randomly selected, unless the state had less than three. In the case of Rondônia, where zero weather stations that provided the desired data existed, three stations were selected from surrounding the state. INMET offers a lot of weather related monthly reported data. Average temperature, temperature high, temperature low, cumulative rainfall, and number of days with rainfall were the available data. Only average temperature and cumulative rainfall were usable as the other three features were 50-90% nan's. In order to create more data points, each year was broken up into 4 quarters; January- March; April-June; July-Sept; Oct-Dec.

Another predictor variable used was GDP per capita and population from the Brazilian Institute of Geography and Statistics (IBGE) [11]. This data was in yearly resolution so the same value for GDP and population for all 4 quarters in a year for each state were used. This allowed us to continue using 4 quarters per year. The GDP data came as total GDP. This was normalized with the population of the state, resulting in GDP per capita.

The response data came in the form of a list of events that occurred in certain starters starting on certain dates. Each event needed to be placed with their corresponding box in terms of year, quarter and state. This was done by generating a day that was in the exact middle of each quarter and compared the reported starting displacement date to all the middle dates. The date closest to the displacement date corresponded to the box that event went into. The last step of

matching the correct state was much simpler. With 27 states, 7 years and 4 quarters per year this gave us 756 samples. 122 of them had displaced people. 31 of them being “large” events, as defined above.

With the predictor matrix and response vector set, the analysis began. A Random Forest model was chosen because of its applicability to the decision making process of organizations that assist refugees. Given certain conditions, this amount of climate refugees could be expected. It is highly interpretable and accessible, which is useful when presenting tools to organizations to use in disaster situations. Secondly the data logically can be broken up into chunks, like a lot of rain and a little rain, and very hot versus not as hot. This suggests that a Random Forest might work well. Using 10 fold cross validation, a Random Forest was trained and tested with a max depth parameter ranging from 2 to 60 with 100 estimators each. Inside each fold, SMOTE was performed on the training data to even out the number of samples in each category. The held-out validation set predictions were joined together such that 756 predictions resulted - all of which had not been seen by the model when training. These predictions compared the max depth hyperparameter to determine the best balanced accuracy. An ROC curve was created to observe the true positive rate vs. false positive rate. With the best max depth hyper parameter (max depth of 4), a confusion matrix was produced to learn more about how our model was predicting each class. See results section for these figures.

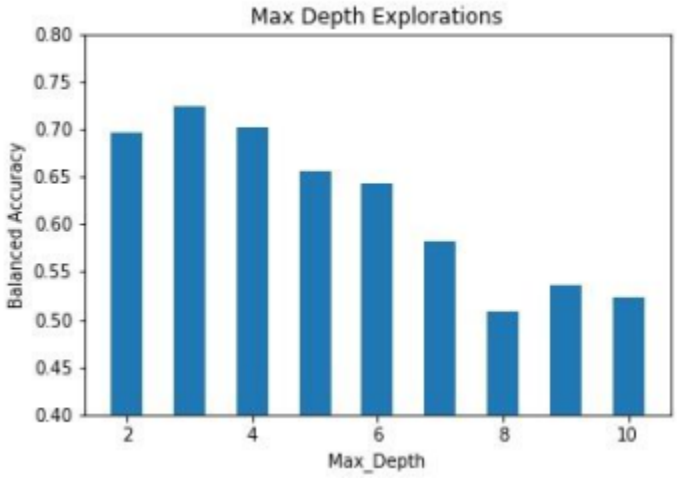

Figure 1: Using Cross Validation to choose the hyperparameter max depth based on Balanced Accuracy.

It was hypothesized that a Logistic Regression would work best for this data. Intuitively there wouldn’t be a complex relationship between the predictor variables and the response. If there is more rain, it’s more likely to flood and displace people. There is a similar relationship between hot weather and droughts as well as GDP and having the resources to stay put. Because of these intuitive relationships between the predictors and the response variables, a simple Logistic Regression is a logical model to apply. A range of C penalties were tested using 10 fold cross validation. SMOTE was applied within each fold to balance out the training data only. The

best C penalty was used to produce a ROC Curve as well as a Threshold to Balanced Accuracy plot. See results for figures.

Results

Note on reasoning for using balanced accuracy: Balanced accuracy is the main performance metric used throughout this study. This was strategically chosen because the displacement data is highly imbalanced. With 31 samples being positive (more than 3000 displaced people) and 725 samples being negative (less than 3000 displaced people), if a model chose negative everytime it would result in an accuracy rate of 95% because the model would have gotten all of the negative class samples correct. However balanced accuracy is defined as the average accuracy of each class weighted equally. So the balanced accuracy of a model choosing all negative would be 50%. 100% for the negative class and 0% for the positive class. This shows a clearer better picture of how well the model is performing on the test data.

Random Forest

The result for the Random Forest Classifier with a max depth of 4 and 100 estimators was a balanced accuracy of 72.3%. However a closer look at the predictions is necessary to understand how well the model is doing. Table 1 shows the confusion matrix of all the predictions compared to the true data. The model correctly identified 22 out of the 31 large events (True Positive Rate of 70.9%) with only a 25.9% false positive rate. This suggests that there is a signal here and that it is better than a coin flip.

		True Class	
		Positive	Negative
Predicted Class	Positive	22	188
	Negative	9	527

Table 1: Confusion matrix showing Random Forest model performance

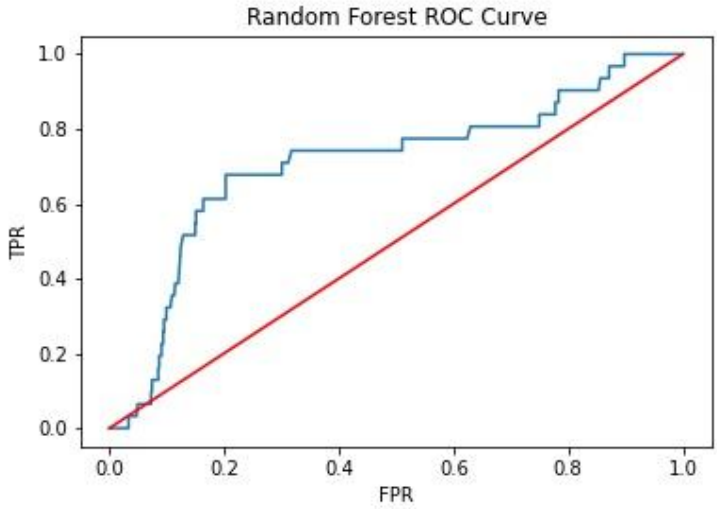

Figure 2: Blue Curve represents the Random Forest ROC curve. Red Line represents random what the curve would look like if it was just random chance.

A ROC Curve can be interpreted as displaying the tradeoff between the sensitivity and the specificity. Classifiers that display ROC Curves that reach further toward the top left are considered better classifiers. One advantage to using a ROC curve in our instance is that it does depend on class distribution which is crucial when working with highly imbalanced classes. The area under the ROC curve, or AUC, is also indicative of how accurate the classifier is. The AUC of our Random Forest Classifier was 0.708. Random chance would give an AUC of .5 and a perfect classifier would give you an AUC of 1. So .708 is a moderately good classifier.

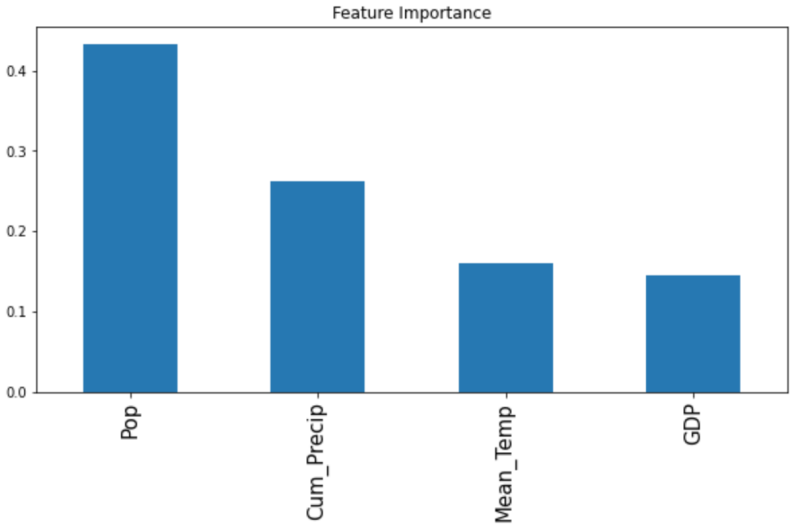

Figure 3: Feature importance chart displays the prominent input features from Random Forests modeling

Feature importance is a useful attribute of Random Forests. It allows the model builder to learn which predictor variables are the most useful in predicting the event. Figure 3 displays a bar graph of how important each feature is in predicting the outcomes. Feature importance is calculated by using a certain criterion - in this case the Gini impurity was chosen. The Gini impurity is minimized when only objects of the same class are in groups, and larger when classes are mixed within groups. The feature importance is calculated by averaging how much the gini impurity decreases every time the model uses that feature to split the data. If splitting the data using that feature doesn't decrease the Gini impurity by very much, then the feature importance of that feature will be low, and vice versa. The results indicate the most important feature is the state's population. This is logical because if there are more people in a given state then it is more likely that a lot of them might be displaced. If population were the only important feature, then that would indicate that our other features have no predictive value. However the second most important feature is cumulative rainfall, which makes intuitive sense since flooding is the cause of many of the displacements every year in Brazil.

Logistic Regression

The Logistic Regression did not yield as promising results. As in the Random Forest, SMOTE was used to even out class samples. The predictions obtained from the Logistic regression all predicted a very low probability that any of the samples were positive, predicting all samples to be negative with a probability of 90% or higher. The threshold for predicting displaced people had to be lowered down to 0.09 before any predictions were made positively. Figure 4 below illustrates how the balanced accuracy changes as the probability threshold for picking a positive prediction fluctuates. A threshold of .065 gave us our maximum BAC or .519. This being only slightly better than a coin flip isn't a promising result.

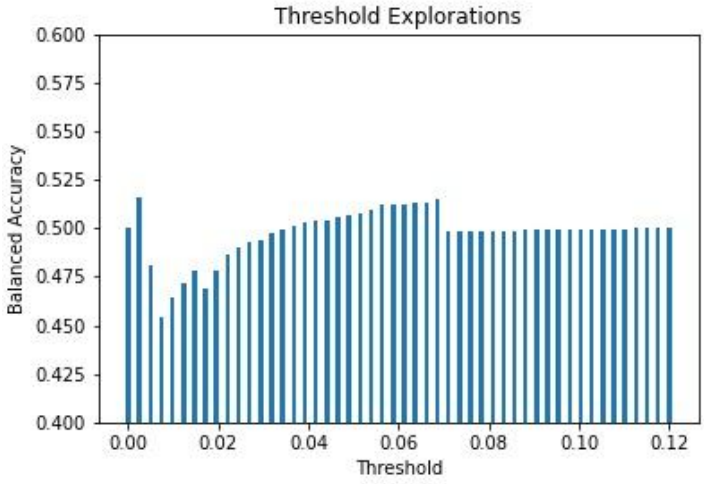

Figure 4: Logistic Regression Threshold for Decision Making vs. BAC to determine an appropriate threshold for identifying extreme weather events.

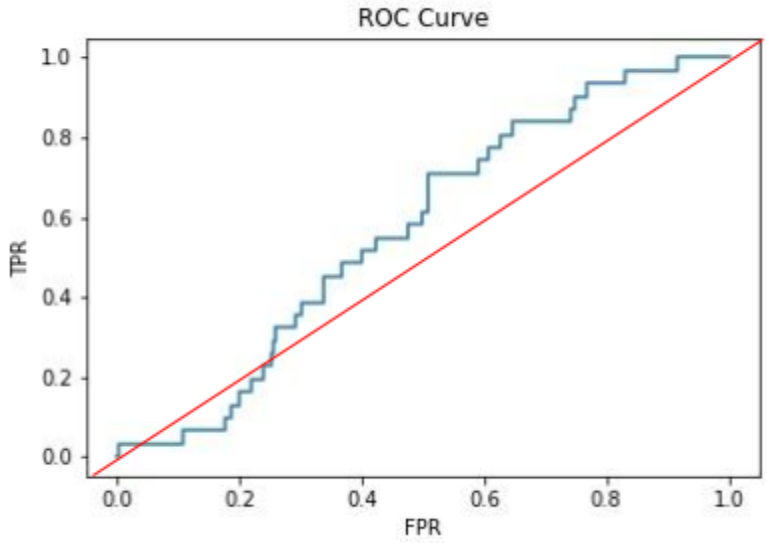

Figure 5: ROC Curve showing the performance of our Logistic Regression model in classification selection.

Figure 5 shows the ROC curve for the Logistic Regression model. This is another way to demonstrate what happens when the threshold for predicting positive is fluctuated up and down. As stated above a better classifier reaches out toward the top left of the plot and a poor one appears more diagonal. The classifier is very near the red random chance line throughout.

Discussion

In this study, our team used publicly available meteorological, geographic, and demographic data published by the Brazilian federal government and its 27 federative units (26 states and 1 federal district), made available for download through the 2011 Brazilian Data Protection Law. The data were collected with the goal to uncover macroscopic displacement patterns of people in response to extreme weather events and local economic factors such as GDP. An initial iteration of this project used data from the United Nations Refugee Agency (UNHCR) [8]. However, the UNHCR data only had displacements per year at the national level. This did not have a high enough temporal or spatial granularity to give the analysis of interest.

Model improvements were necessary after initial model iterations proved to be no better than the null model. These initial iterations included Logistic Regression, Decision Trees, and Random Forests. The methods previously selected did not properly account for non-linear or time-series data. Using Gradient Boosting for time-series data helped in slightly reducing bias and somewhat improving model performance. Additionally, crop yield data for the nation of

Brazil was not appropriate to use when trying to predict internal displacements. A different direction was taken to collect state-level data and to account for geographic variability.

The main changes included obtaining data at the state level such as quarterly cumulative precipitation, average temperature, GDP per capita and population. State level predictor variables are useful in granting greater granularity to our data, both in temporal terms and geographic terms. This results in a tradeoff that data doesn't extend far beyond 2013, bringing up concerns of the validity of the application of model predictions in other years. It also resulted in another tradeoff in that very limited data is available at the state level from Brazil. This cut the number of predictor variables down from twelve to four. Overall, results seem to indicate an ability to predict the occurrence of displaced peoples following a major weather event. This information can be helpful for NGO and government organizations to send aid to displaced peoples in a more timely manner.

Conclusion

The literature review suggests that experts by and large agree that the link between climate change and migration exists. The climate is changing, and the frequency and magnitude of natural disasters is increasing. This is going to create an unprecedented humanitarian crisis, due to the creation of climate refugees. Understanding the factors that relate to the creation of climate refugees, and being able to predict how many and where they are, will help with the movement of aid and the organization of relief efforts. Further work is needed in this area of study. These are questions and future possible projects the team suggests.

1. Can an aggregated rating for crop yield at the state level from year to year be formulated?
    a. In our first iteration of this project country wide crop production of soybeans, rice and wheat data were used. These variables seemed to be helpful in predicting displaced people at the country level. However, making accurate predictions proved difficult when using Brazil's crop production data at the state level since crops vary from state to state.
    b. Aggregation ratings can be useful in comparing crop production between states and between years.
2. Would Lead-lag analysis lead to better model predictions?
    a. A lead-lag analysis helps in correlating the nature of weather pattern effects on migration, since migration is best explained through a delay after a weather event. Typically, drought conditions persist long enough to affect agricultural production before people are motivated enough to migrate outside of their current location [4]. Drought affects the degree of soil degradation that occurs which leads to short term labor migration [15]. As a result, a lead-lag estimation has an enhanced ability to analyse extreme weather events effects on internal displacements.

3. Are Brazil's data collection methods adequate? Do they meet a standard for free and open data?
    a. The Sunlight Foundation, an international open data watchdog, defines open data principles which include: completeness, primacy, timeliness, accessibility, machine readability, nondiscriminatory, nonproprietary, license-free, permanence, and low usage costs.
    b. According to Brito et al. 2015, Brazil fails in terms of completeness, timeliness, nondiscrimination, and being license-free [16]. These add barriers to developers who note the lack of data centralization and publishing standards. Trends during the pandemic [14] show possible future concerns for the future of open data in Brazil[12,13]. Existing wealth disparities between wealthy and poorer states result in irregular data collection that only further irregular data collection.

4. Since anomalies in weather patterns are the relevant parameters, can thresholds relative to each state be set that signify extreme weather events ?
    a. The necessity to consider thresholds for every state lies is exemplified best in the difference in precipitation between tropical states, like Amazon, and drier northeast states, like the Caatinga region [3].

References

[1] - Reuveny, R. Climate change-induced migration and violent conflict. Political Geogr. **2007**, 26, 656–673.

[2] - Gleick, P.H. The implications of global climatic changes for international security. *Climatic Change* **15,** 309–325 (1989). https://doi.org/10.1007/BF00138857

[3] - Hudson, Rex A., ed. (1998). Brazil : a country study. https://www.loc.gov/item/97036500/

[4] - IPCC Fifth Assessment Report: Synthesis Report. Available online: http://www.ipcc.ch/report/ar5/syr/

[5] - Migration and Global Environmental Change (2011) Final Project Report The Government Office for Science, London

[6] - Burrows K, Kinney PL. Exploring the climate change, migration and conflict nexus. Int J Environ Res Public Health. 2016;13(4):443.

[7] - "IDMC | Global Report on Internal Displacement 2018." Accessed December 10, 2020. https://www.internal-displacement.org/global-report/grid2018/.

[8] - UNHCR. "'Refugees' and 'Migrants' – Frequently Asked Questions (FAQs)," March 16, 2016. https://www.unhcr.org/news/latest/2016/3/56e95c676/refugees-migrants-frequently-asked-questions-faqs.html.

[9] - Cortes, Amber. "The Refugees You Don't See — How Climate Change Fuels Forced Migration." Global Washington, June 19, 2019. https://globalwa.org/issue-brief/acknowledging-climate-refugees-on-world-refugee-day/.

[10] - Instituto Nacional de Meteorologia. "Instituto Nacional de Meteorologia - INMET." Accessed December 10, 2020. https://portal.inmet.gov.br/dadoshistoricos

[11] - "Downloads | IBGE." Accessed December 10, 2020. https://www.ibge.gov.br/en/statistics/downloads-statistics.html.

[12] - Sarkis, Marcelo. "Access to Public Information in Brazil: What Will Change With Law No. 12.527/2011? — Right2Info.Org," May 16, 2012. https://www.right2info.org/recent/access-to-public-information-in-brazil-what-will-change-with-

law-no.-12.527-2011.

[13] - Serbin, Kenneth P. “The Ghosts of Brazil’s Military Dictatorship,” February 13, 2019. https://www.foreignaffairs.com/articles/brazil/2019-01-01/ghosts-brazils-military-dictatorship.

[14] - Committee to Protect Journalists. “Brazil Restricts Access to Government Information amid COVID-19 Emergency,” March 26, 2020. https://cpj.org/2020/03/brazil-restricts-access-to-government-information/.

[15] - Gray, C. Soil quality and human migration in Kenya and Uganda. Glob. Environ. Chang. **2011**, 21, 421–430.

[16] - Brito, Kellyton & Silva Costa, Marcos & Garcia, Vinicius & Meira, Silvio. (2015). Is Brazilian Open Government Data Actually Open Data?. International Journal of E-Planning Research. 4. 57-73. 10.4018/ijepr.2015040104.

Appendix

The code for this study can be found at: https://colab.research.google.com/drive/1f-f23RQgVbtnHD88W1u7ZZIR6W4I3xmm?usp=sharing

The data for this study can be found at: https://github.com/alyssa-rose/DSS_Final_Project