Quantifying the Causes of Human Migration

Austin P. Wright
apwright@gatech.edu
Georgia Institute of Technology
Atlanta, GA

Chitwan Kaudan
ckaudan3@gatech.edu
Georgia Institute of Technology
Atlanta, GA

Arjun Goyal
agoyal305@gatech.edu
Georgia Institute of Technology
Atlanta, GA

Lipi Shah
lshah3@gatech.edu
Georgia Institute of Technology
Atlanta, GA

Frank Whitesell
fwhitesell3@gatech.edu
Georgia Institute of Technology
Atlanta, GA

Austin Himschoot
ahimschoot3@gatech.edu
Georgia Institute of Technology
Atlanta, GA

ABSTRACT

Understanding the causes of global migration has become increasingly important as the effects of climate change as well as global conflicts and economic development begin to take shape. Most current work only analyzes causes in isolation, finding many individually important factors but not quantifying how effects interact with each other. This work aims to build an interpretable and interactive analysis tool using a multiplex network modeling framework together with general additive model regression, in order to understand and visualize the relationships of these causes and effects on international migration.

ACM Reference Format:
Austin P. Wright, Chitwan Kaudan, Arjun Goyal, Lipi Shah, Frank Whitesell, and Austin Himschoot. 2019. Quantifying the Causes of Human Migration. In *Proceedings of Data and Visual Analytics (DVA'19).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

DVA'19, December 2019, Atlanta, GA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As evidenced by publications on this topic, global migration is of interest to governments [17] [25], international agencies, non-profits and businesses, academic institutions, the media, and societies as whole [14]. It is important to note that migration always impacts two communities, the origin and the destination.

Understanding the relative importance and the mutual interactions between the drivers of global migration is of interest for building future migration policies. Existing work studies the effect of climate change on migration independent of the traditional drivers of human migration. This work models climate and traditional drivers simultaneously to understand each variable's relative impact. This work introduces a multiplex network to model relevant features to global migration, and uses a general additive model to perform a regression over those features, predicting migration flows. Previous work in nuclear proliferation has shown that multiplex network analysis is a good framework for quantitative international relations[21], and generalized additive modeling has been shown to be effective in both pure regression performance, and providing useful local and global explanations for its predictions[23].

2 PROBLEM DEFINITION

Our objective is to use network analysis to predict international migration patterns, and understand their causes. Within the network framework, nodes represent countries and weighted edges represent relationships between countries. The goal of the model is to predict migration flows between countries as a dependent variable of several independent data sources. These data sources are an amalgamation of a diverse range of factors that have been found to contribute to global migration flows: political, socioeconomic, demographic, climate, and more. We can then define the regression task as finding a function, f, of the node and edge features for each migration flow pair. Here $\vec{n}$ is vector of node attributes of all layers for a given country in a given year, and $\vec{e}$ is the set of edge attributes of all layers for that year.

$$migration_{source \to target} \approx f\left(\vec{n}_{source}, \vec{n}_{target}, \vec{e}_{source \to target}, \vec{e}_{target \to source}\right)$$

Furthermore by using an explainable model to approximate this f, our goal is to provide useful and interactive visualization tools to explain the specific factors of migration.

3 LITERATURE REVIEW

Migration Analysis

Currently, academic research relies on qualitative methods such as country or region-specific case studies to analyze factor impact on migration. [32] [28] [24] [19] [9] Also, quantitative research that examines international migration is frequently one-dimensional: studies focus on a small number of correlated factors in their analysis.[12][8] Because migration motivations are diverse[14][28], it is important to consider these factors from multiple lenses for a more accurate understanding. Furthermore, individual datasets can be sparse and differing data collection methods can affect study quality.[11]

Multiplex Networks

Using a multiplex network to model an international scenario with time-series data, with countries as nodes and relationships as edges, has been used to predict states of nuclear proliferation[21]. We will build on this by including quantitative node and edge attributes in our model. Another multiplex model [27] examining different metrics for edge weights will be useful in our prediction analysis. Coinciding with these different metrics is the idea of regression between layers [13]. Combining edge weight metrics and layer regression can be used for our model. Each layer will examine a unique factor with new data to improve the robustness. Some of the attributes we have considered are trade [29], conflict [20], climate change impact on conflict [12], environmental hazards [32], and socioeconomic data [28]. Including more attributes will improve the quality of our predictions, and increase the scope of our causal analysis. Moreover, current migration theories can be used to assess results.[30] [14][11][24][32]

Generalized Additive Models

Generalized Additive Models (GAM) are a type of generalized linear model where "the impact of the predictive variables is captured through smooth functions which—depending on the underlying patterns in the data—can be nonlinear".[26] The relationship between each predictor variable and the response are modeled by a non-parametric function simultaneously during model estimation. The final estimated response is a linear combination of each predictor's function. Since we are interested in each explanatory variable's contribution to the predicted migration flow between two countries, a GAM enables a clear visual interpretation of the results [22], both for individual instances and the overall model [23].

4 METHOD

The combination of GAM and multiplex network analysis is a novel approach to migration studies.

Multiplex network analysis allows us to model a large number of complex relationships between countries as a set of nodes and edges. GAM gives our model flexibility and interpretability. We are able to model data which has a nonlinear relationship between the predictors and the response in a way which is both accurate and understandable.

Additionally, the scope of our model, which includes 42 predictors and consists of 2.1 million data points from 1960 to 2014, far exceeds that of previous studies we encountered in our literature survey. Previous approaches have a much narrower scope, often considering only migration within one region during a shorter time-frame. Our model also attempts to quantify the influences on global migration, whereas previous studies tend to be qualitative investigations.

Causes of Migration and Data Collection

Past research suggests that migration is driven by individuals searching for better economic conditions for their households[11]. Trade dependencies can help determine the destination country of a person who decides to migrate.

A country's aggregate demographic and socioeconomic metrics could help determine its bilateral migration flow.[17]. We included a layer that represents population growth and another for the level of education of the general public for each country.

Communities' ability to mitigate environmental risk and chronic abundance of this risk has been hypothesized to affect migration [30]. Threats to land and fresh water resources from climate change also put burdens on individuals to move [32]. We included data on average temperature, natural hazards, water scarcity, and landlocked status to incorporate some of the natural effects of migration.

Political systems and conflict also have an impact on a person's ability and desire to emigrate from, or migrate to, a specific country. We explore these aspects by including datasets of autocracy scores, democracy scores, and conflict between countries around the world.

Selecting Countries and Time-frames. Since the regression analysis is dependent on bilateral migration flow, selecting a standard list of countries across datasets was necessary. The initial dataset[7] used the ISO-3166-1 standard for source and target country abbreviations and numeric codes. We applied this standard across all node and edge attribute datasets for consistency. Furthermore, this dataset constructed 5 year estimates of migration flow from 1960-2014. As a result, we restricted our node and edge attribute datasets to this time-frame. Based on the consistently-available data across all datasets, our analysis considers 194 countries.

Data Imputation. The 5 year estimates of migration flow from each source to target were averaged and applied across the intervening years. In some cases, available data was sparse, especially for countries in the developing world and for the earlier time periods in our research timeframe. Backward filling and, if necessary, forward filling imputation methods were utilized for treatment of missing values in these cases.

Node Attributes

Democracy Score: The Democracy score is a subjective assessment of the country's extent of democracy on a scale of 0 through 10.[1]

Autocracy Score: The Autocracy score similarly assesses the country's extent of autocracy on a scale of 0 through 10. [1]

GDP per Capita: The gross domestic product (GDP) per capita is the sum of gross economic value produced by all residents in the country plus any taxes and not including any subsidies, divided by population. [10]

Average Temperature: The average temperature data for NOAA worldwide weather stations. The data was grouped on country and year to calculate average temperatures.[16]

Average Years in School: Average Years in School is the mean years of schooling completed by the

adult population (25 years or older) at a given time.[3]

Birth rate per 1000: Birth rate per 1000 is the live births occurring during the year, per 1000 population estimated at midyear. [10]

Landlocked Status: The landlocked status data identifies countries bordered only by land. A 1 represents landlocked and 0 otherwise.[5]

Water Scarcity Category: This data categorically represents each country by their fresh water scarcity. There are five different levels from extremely-high to low. Aqueduct combines quantitative and qualitative physical risks and regulatory risks to categorize countries by these levels.[15]

Natural Disasters Risk: A table of worldwide locations and natural hazards that affect them. Key natural disaster words were collected and compared to the CIA description. Words were partitioned based on a common theme. A binary score was given if a country did or did not experience these types of disasters.[4]

Country Population: Total count of all residents in each country, regardless of legal status or citizenship. [10]

Edge Attributes

Trade Dependence: Trade dependence is defined between country i and country j in [21] as:

$$D_{ij} = \frac{Ex_{ij} + Im_{ji}}{GDP_i}$$

where Ex_{ij} is the total value of exports from country i to country j, Im_{ji} is the total value of imports from j to i, and GDP_i is the GDP of i.[10][18]

Conflict: A table of interstate armed conflicts and their corresponding intensities from 1946-2018. A 1 represents a minor conflict (25-999 battle-related deaths in a given year), a 2 represents a war (at least 1,000 battle-related deaths), and a zero represents peace between two states in a given year.[6]

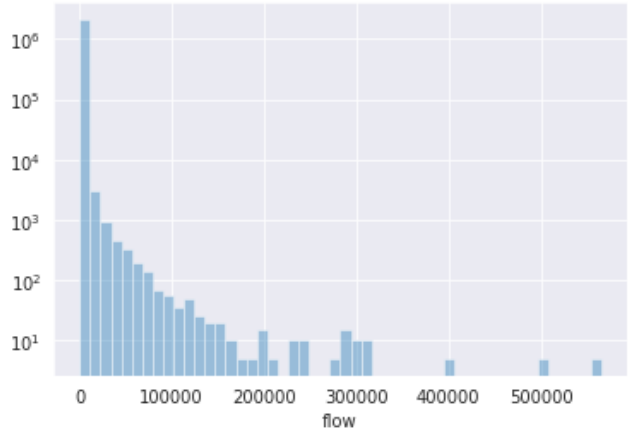

Figure 1: Total Distribution of Migration Flows (log scale)

Contiguous States: A table of contiguous borders between two states. A 1 represents a shared river or land border. 2 to 5 represent contiguity by bodies of water other than rivers in a range of 0-400 miles, in increasing increments. Zero represents no border contiguity between two states.[2]

Bilateral Migration Flow: Estimates of bilateral (country to country) migration flow for 194 countries from 1960-2015 were generated using a combination of demographic and migrant stock datasets.[7]

Model

Generalized additive modelling uses an algorithm called local scoring to create smooth functions $s_i(x_i)$ representing relationships between each feature x_i and the response variable Y. By adding each of these functions together, we can obtain a prediction $E(Y)$ using a link function $g(Y)$ [26]. We used the log link function for our model predictions, since migration flows are exponentially distributed as shown in figure 1.

$$g(E(Y)) = s_0 + s_1(x_1) + \ldots + s_p(x_p)$$

We selected GAM for a few reasons. First, GAM offers a great deal of flexibility. Unlike parametric regression models, we do not have to understand the distribution of our input feature data in order to accurately model it. Nonlinear relationships

Table 1: Features and Encoding

Feature	Encoding	Data Type	Mean	Standard Deviation
Population	Persons	Integer	11,356,500.56	70,714,482.87
Democracy Score	Ordinal $\in [0, 10]$	Integer	5.45	3.90
Autocracy Score	Ordinal $\in [0, 10]$	Integer	2.08	3.02
GDP per Capita	USD	Float	$8911.72	$14965.96
Average Temperature	Degrees Celsius	Float	19.08	7.47
Average Years in School	Years	Float	6.75	3.36
Birth rate per 1000	Annual Birth Rate	Float	25.11	12.69
Land Locked Status	Categorical $\in \{0, 1\}$	Integer	0.23	0.42
Water Disaster Risk Score	Ordinal $\in [1, 5]$	Integer	2.54	1.40
Natural Disasters Risk	Categorical $\in \{0, 1\}^6$	Integer	-	-
Trade Dependence	Relative	Float	$3.48 \cdot 10^{-3}$	$2.81 \cdot 10^{-2}$
Conflict	Ordinal $\in \{0, 2\}$	Integer	$1.89 \cdot 10^{-4}$	$1.87 \cdot 10^{-2}$
Contiguous States	Ordinal $\in \{0, 5\}$	Integer	$5.21 \cdot 10^{-2}$	0.436
Bilateral Migration Flow	Persons (thousands)	Float	160.81	2953.72

between predictors and the response are able to be modeled with GAM, providing powerful state of the art regression performance.[26] Second, because the regression model is additive, the interpretation of one feature's impact on the model prediction does not depend on the surrounding variables. As a result, we can draw conclusions about each predictor's individual effect on the response from its smoothing function. Then, by looking at the summation of all smoothing functions, we can see how each predictor impacts the response on an overall level. We can draw conclusions for the precise impact of each feature as an explanation for a specific prediction through the additive terms, and we can draw insights about the nature of each feature influence globally through the shape of its spline function. This combination of local and global explanations provides a comprehensive tool for understanding the relationships between the independent and dependent variables[23].

Visualization

The visualization was created using R Shiny and D3. Shiny allows the user to host the application locally. Before rendering a figure, the application prompts the user for a year between 1960 and 2014. After selecting a year, a D3 chord diagram is displayed with the top 40 predicted migration flows between selected countries. Individual chords are colored by the country with the larger corresponding outflow, as shown in figure 2 for 1979, with a tooltip over the flow between Canada and the USA. Since the chords are only created for the top 40 flows, migration flows may exist between countries with no displayed chord.

To analyze the factors that affect predicted flows, the user can view a waterfall chart of the log flows by factor for a selected migration. To create this plot, the user has to sequentially select a source country and a target country by their bases on the chord diagram. This selection renders the predicted and actual values on the left side of the application and the waterfall chart below. The first bar represents the intercept and the last is the total predicted flow with other bars ordered by magnitude. Hovering over a bar creates another tooltip which displays the variable and the log value. Selecting another country on the chord diagram or changing the year removes the waterfall chart. This combination of a chord diagram and a waterfall chart

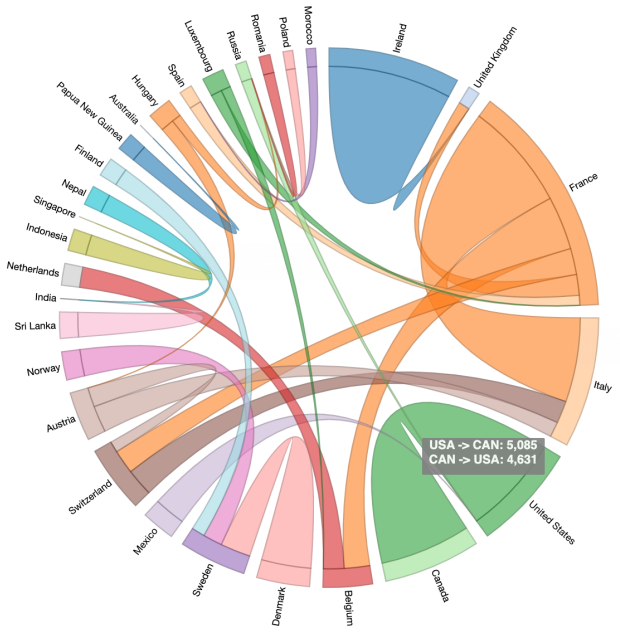

Figure 2: Chord Diagram

gives the user an intuitive representation of the magnitude of the flows between countries and the variables that are predicted to influence the selected migration.

5 EXPERIMENTS/EVALUATION

We hoped to answer the following questions in our investigation which could inform policy makers:

(1) Historically, what factors impact global migration?
(2) In what way do these factors influence global migration?
(3) Is it possible to quantify the impact of these factors?

Quantitative Evaluation

In order to evaluate our model, we calculated an r^2 score of 0.37 which measures the proportion of the variance in the data that the model can explain implying that our model encodes about 37% of the true causes of migration. This statistic quantifies the impact of our considered variables on migration. Our r^2 value is close to the state of the art for migration prediction of 0.42 [31].

Qualitative Evaluation

Local Explanations. Beyond the quantitative metrics for our model, our visualization provides intuitive explanations for individual migration flows. Say for instance, you are interested in examining the migration flow from Canada to the United States in 1979. You would select CAN as the source and USA are the target in the chord diagram in Figure 2. Then you could see the factors of migration from Canada to the USA in 1979 in Figure 3. The resulting estimate shows the effects of each of our

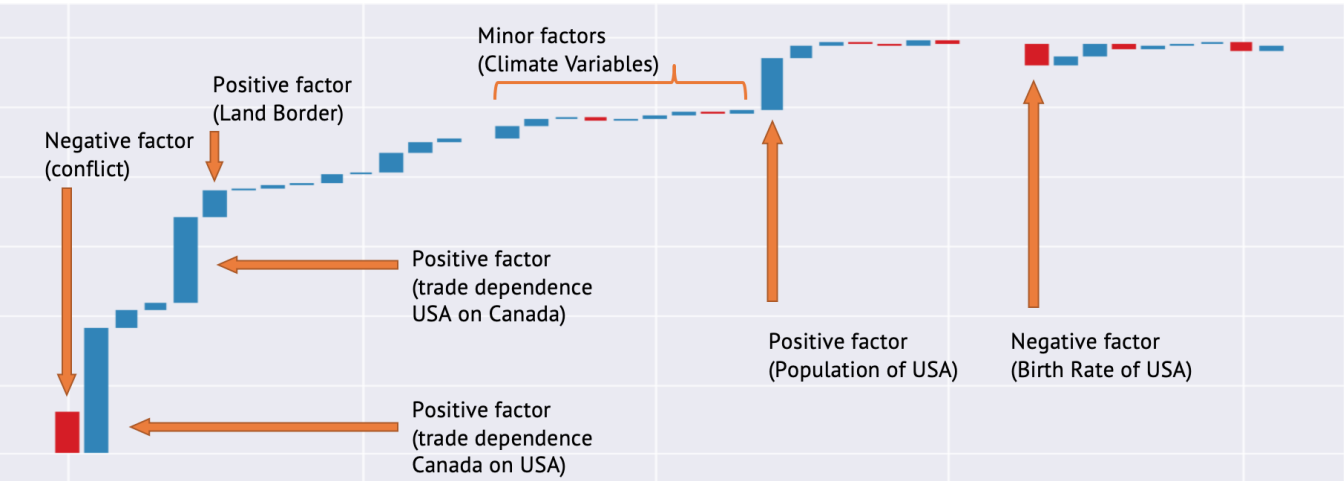

Figure 3: Waterfall Chart of Features of Migration for Canada to the USA in 1979

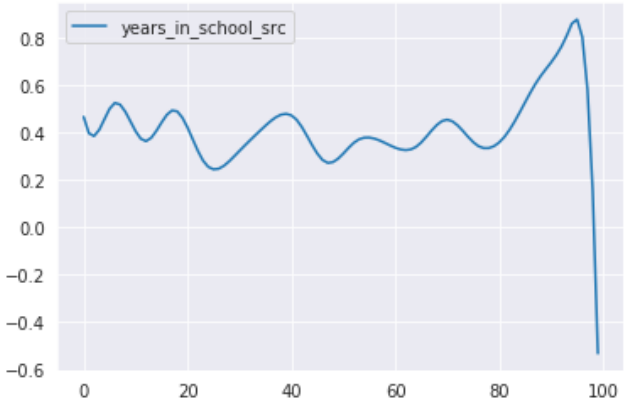

Figure 4: Source Country Average Years in School Impact

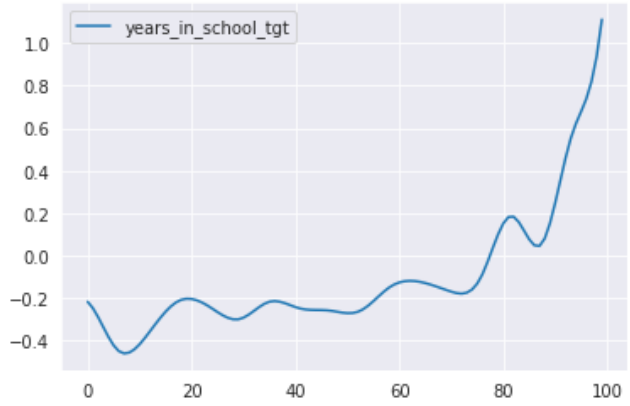

Figure 5: Target Country Average Years in School Impact

variables and how each affects the final estimate of migration. The bulk of the migration in this case can be explained by the strong trade relationship between the US and Canada, while the impact of climate related factors are small. This helps explain that while climate change may have a macro effect on global migration, most bilateral migration flows are still much more dependent on traditional effects like trade dependence.

Global Explanations. To address research question (2), we can turn to the smoothing functions for each predictor variable. For example in figure 4, in the case of the average years in school dataset, we obtained the following smooth function for the data as a source node.

Figure 4 shows that a migration out of a source country is likely to increase as the average years of education in the country also increases, up to a point. This reflects the idea that increased education leads to higher levels of income, and thus greater mobility to move. It likely reaches a peak and then drops because citizens in well-educated countries also likely have higher incomes, and have less incentive to move. Another important note from this chart is that the data points become sparse at the upper end of the range. This is because there are relatively few countries in the world which average the high number of years of education reflected in the table.

It is interesting to compare the Figure 4 to Figure 5, which shows the smooth function for migration into a target country as a function of average years of education. As expected, the higher number of average years of education, the more desirable the country is as a destination for immigrants.

6 CONCLUSIONS AND DISCUSSION

Together, these results seem to indicate that the impact of climate change on international migration may not be as significant as expected. However, it's important to consider that climate change tends to drive internal, not international, migration. Our findings indicate that relationships between countries and economic attributes of respective sources and targets provide considerable influence on predicted flows. Policy makers can use this information to see that climate change impacts will not significantly affect migration to countries. Understanding relationships between countries will provide policy makers and government officials the information they need.

ACKNOWLEDGEMENT

All team members contributed a similar amount of effort.

REFERENCES

[1] [n. d.]. INSCR Data Page. https://www.systemicpeace.org/inscrdata.html.

[2] 2017. Correlates of War Project. Direct Contiguity Data, 1816-2016. Version 3.2. http://www.correlatesofwar.org/data-sets/direct-contiguity

[3] 2018. Mean years of schooling. https://ourworldindata.org/grapher/mean-years-of-schooling-1

[4] 2018. The World Factbook. https://www.cia.gov/library/publications/the-world-factbook/fields/292.html

[5] 2019. Landlocked Countries 2019. http://worldpopulationreview.com/countries/landlocked-countries/

[6] 2019. UCDP Dyadic Dataset version 19.1. https://ucdp.uu.se/downloads/

[7] Guy J. Abel. 2018. Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015. *International Migration Review* 52, 3 (2018), 809–852. https://doi.org/10.1111/imre.12327 arXiv:https://doi.org/10.1111/imre.12327

[8] Fuad Aleskerov, Natalia Meshcheryakova, Anna Rezyapova, and Sergey Shvydun. 2017. Network Analysis of International Migration. In *Models, Algorithms, and Technologies for Network Analysis*, Valery A. Kalyagin, Alexey I. Nikolaev, Panos M. Pardalos, and Oleg A. Prokopyev (Eds.). Springer International Publishing, Cham, 177–185.

[9] Gregory S. Amacher, Wilfrido Cruz, Donald Grebner, and William F. Hyde. 1998. Environmental Motivations for Migration: Population Pressure, Poverty, and Deforestation in the Philippines. *Land Economics* 74, 1 (1998), 92–101. http://www.jstor.org/stable/3147215

[10] The World Bank. [n. d.]. World Bank Open Data. https://data.worldbank.org/

[11] Thomas Bauer and Klaus Zimmermann. 1995. *Modelling International Migration: Economic and Econometric Issues*. 95–115.

[12] Kate Burrows and Patrick Kinney. 2016. Exploring the Climate Change, Migration and Conflict Nexus. *International Journal of Environmental Research and Public Health* 13, 4 (Apr 2016), 443. https://doi.org/10.3390/ijerph13040443

[13] Giona Casiraghi. 2017. Multiplex Network Regression: How do relations drive interactions? arXiv:physics.soc-ph/1702.02048

[14] Francesco Castelli. 2018. Drivers of migration: why do people move? *Journal of Travel Medicine* 25, 1 (07 2018). https://doi.org/10.1093/jtm/tay040 arXiv:http://oup.prod.sis.lan/jtm/article-pdf/25/1/tay040/25811725/tay040.pdf tay040.

[15] Hannah Dormido. 2019. These Countries are the Most at Risk from a Water Crisis. https://www.bloomberg.com/graphics/2019-countries-facing-water-crisis/

[16] National Centers for Environmental Information and Ncei. [n. d.]. Climate Data Online. https://www.ncdc.noaa.gov/cdo-web/

[17] Gary P. Freeman. 1995. Modes of Immigration Politics in Liberal Democratic States. *International Migration Review* 29, 4 (1995), 881–902. https://doi.org/10.1177/019791839502900401 arXiv:https://doi.org/10.1177/019791839502900401 PMID: 12291223.

[18] International Monetary Fund. 2019. Direction of Trade Statistics (DOTS). https://data.imf.org/?sk=9D6028D4-F14A-464C-A2F2-59B2CD424B85

[19] Elizabeth Fussell, Lori M. Hunter, and Clark L. Gray. 2014. Measuring the environmental dimensions of human migration: The demographer's toolkit. *Global Environmental Change* 28 (2014), 182 – 191. https://doi.org/10.1016/j.gloenvcha.2014.07.001

[20] Faten Ghosn, Glenn Palmer, and Stuart A. Bremer. 2004. The MID3 Data Set, 1993–2001: Procedures, Coding Rules, and Description. *Conflict Management and Peace Science* 21, 2 (2004), 133–154. https://doi.org/10.1080/07388940490463861

[21] Bethany L. Goldblum, Andrew W. Reddie, Thomas C. Hickey, James E. Bevins, Sarah Laderman, Nathaniel Mahowald, Austin P. Wright, Elie Katzenson, and Yara Mubarak. 2019. The nuclear network: multiplex network analysis for interconnected systems. *Applied Network Science* 4, 1 (2019), 36. https://doi.org/10.1007/s41109-019-0141-4

[22] Trevor Hastie and Robert Tibshirani. 1986. Generalized Additive Models. *Statist. Sci.* 1, 3 (08 1986), 297–310. https://doi.org/10.1214/ss/1177013604

[23] Fred Hohman, Andrew Head, Rich Caruana, Rob DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

[24] Roel Jennissen. 2007. Causality Chains in the International Migration Systems Approach. *Population Research and Policy Review* 26 (08 2007), 411–436. https://doi.org/10.1007/s11113-007-9039-4

[25] Rey Koslowski. 2014. Selective Migration Policy Models and Changing Realities of Implementation. *International Migration* 52, 3 (2014), 26–39. https://doi.org/10.1111/imig.12136

[26] Kim Larsen. 2015. GAM: The Predictive Modeling Silver Bullet. https://multithreaded.stitchfix.com/blog/2015/07/30/gam/

[27] Shaghayegh Najari, Mostafa Salehi, Vahid Ranjbar, and Mahdi Jalili. 2019. Link Prediction in Multiplex Networks based on Interlayer Similarity. *CoRR* abs/1904.10169 (2019). arXiv:1904.10169 http://arxiv.org/abs/1904.10169

[28] Wim Naudé. 2008. *Conflict, disasters and no jobs: Reasons for international migration from Sub-Saharan Africa.* WIDER Research Paper 2008/85. Helsinki. http://hdl.handle.net/10419/45125

[29] John R. Oneal and Bruce M. Russet. 2002. The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985. *International Studies Quarterly* 41, 2 (12 2002), 267–293. https://doi.org/10.1111/1468-2478.00042 arXiv:http://oup.prod.sis.lan/isq/article-pdf/41/2/267/5154908/41-2-267.pdf

[30] Clionadh Raleigh and Lisa Jordan. 2008. *Assessing the Impact of Climate Change on Migration and Conflict.* Technical Report. World Bank Social Development Department.

[31] Caleb Robinson and Bistra Dilkina. 2017. A Machine Learning Approach to Modeling Human Migration. *CoRR* abs/1711.05462 (2017). arXiv:1711.05462 http://arxiv.org/abs/1711.05462

[32] K. Warner, M. Hamza, A. Oliver-Smith, F. Renaud, and A. Julca. 2010. Climate change, environmental degradation and migration. *Natural Hazards* 55, 3 (01 Dec 2010), 689–715. https://doi.org/10.1007/s11069-009-9419-7