



Applying machine learning to social datasets: a study of migration in southwestern Bangladesh using random forests

Kelsea Best¹ · Jonathan Gilligan¹ · Hiba Baroud² · Amanda Carrico³ · Katharine Donato⁴ · Bishawjit Mallick⁵

Received: 5 December 2020 / Accepted: 13 March 2022
© The Author(s) 2022

Abstract

As researchers collect large amounts of data in the social sciences through household surveys, challenges may arise in how best to analyze such datasets, especially where motivating theories are unclear or conflicting. New analytical methods may be necessary to extract information from these datasets. Machine learning techniques are promising methods for identifying patterns in large datasets, but have not yet been widely used to identify important variables in social surveys with many questions. To demonstrate the potential of machine learning to analyze large social datasets, we apply machine learning techniques to the study of migration in Bangladesh. The complexity of migration decisions makes them suitable for analysis with machine learning techniques, which enable pattern identification in large datasets with many covariates. In this paper, we apply random forest methods to analyzing a large survey which captures approximately 2000 variables from approximately 1700 households in southwestern Bangladesh. Our analysis ranked the covariates in the dataset in terms of their predictive power for migration decisions. The results identified the most important covariates, but there exists a tradeoff between predictive ability and interpretability. To address this tradeoff, random forests and other machine learning algorithms may be especially useful in combination with more traditional regression methods. To develop insights into how the important variables identified by the random forest algorithm impact migration, we performed a survival analysis of household time to first migration. With this combined analysis, we found that variables related to wealth and household composition are important predictors of migration. Such multi-methods approaches may help to shed light on factors contributing to migration and non-migration.

Keywords Random forests · Machine learning · Human migration · Bangladesh

Communicated by Robbert Biesbroek and accepted by Topical Collection Chief Editor Christopher Reyer.

This article is part of the Topical Collection on *Environmental Non-Migration: Frameworks, Methods, and Cases*

✉ Kelsea Best
kelsea.b.best@vanderbilt.edu

¹ Earth and Environmental Sciences, Vanderbilt University, 5726 Stevenson Center, Nashville, TN 37240, USA

² Civil and Environmental Engineering, Vanderbilt University, 274 Jacobs Hall, Nashville, TN, USA

³ Environmental Studies, University of Colorado Boulder, Boulder, CO 397 UCB, USA

⁴ Institute for the Study of International Migration, Georgetown University, Harris Building, 3300 Whitehaven Street, N.W., Washington D.C., USA

⁵ Environmental Sciences, Technische Universität Dresden, Dresden, Germany

Introduction

The complexity of processes influencing human migration poses a challenge for researchers who aim to study the interactions between environmental changes and migration (McLeman 2013). Migration in the form of a decision to move or stay is often influenced by a combination of political, social, economic, and environmental drivers, and the dynamics of this combination are unclear and difficult to quantify (Adams and Kay 2019; Black et al. 2011). Environmental migration, which focuses on the influence of environmental conditions and changes on migration, is especially complex, so prediction may be highly uncertain (Gemenne 2011). This uncertainty is further exacerbated by the uncertainty related to future climate and socioeconomic scenarios (Hugo 2011) and the difficulty of predicting human decision-making (Subrahmanian and Kumar 2017). Yet, accurate modeling and prediction of environmental migration

are critical for informing future policy and adaptation strategies, especially as the impacts of climate change continue to increase (Stern 2006; Piguet 2022; Hugo 1996; Biermann and Boas 2010; Black et al. 2011; Boas et al. 2019; Ahsan et al. 2011).

Questions remain about how to best model environmental migration and how to obtain appropriate and accurate data to test these models (Neumann and Hilderink 2015). Current work studying environmental migration uses a wide range of methods and models from strictly conceptual models (Perch-Nielsen et al. 2008; Renaud et al. 2011), to logistic regression (Koubi et al. 2016), multi-variate regression (Hino et al. 2017), and other forms of regression modeling (Henry et al. 2003, 2004), and agent-based models (Cai and Oppenheimer 2013; Hassani-Mahmooei and Parris 2012; Kniveton et al. 2011; Silveira et al. 2006; Smith 2014; Klabunde et al. 2015; Thober et al. 2018). Identifying appropriate data sources is an additional challenge to studies of environmental migration, and there is no agreement about what data are best (Tejero et al. 2020). For example, Fussell et al. (2014) advocate for using a combination of population censuses, surveys, and multi-level modeling. Recently, Lu et al. (2016) utilized mobile phone data from more than six million anonymous phone users in Bangladesh to track movement across short time scales. Household surveys have been a common source of data for migration research (Bilsborrow and Henry 2012), and some researchers claim survey data are the most appropriate level for obtaining information about the causes of migration (Neumann and Hilderink 2015).

Several reviews of existing methods and challenges call for the exploration of new methods that can improve prediction and better address nonlinearities in environmental migration (Neumann and Hilderink 2015; Obokata et al. 2014; Piguet 2010). As Obokata et al. (2014) suggest, existing quantitative methods of studying environmental migration often simplify complex variables and limit the number of variables studied (Obokata et al. 2014). The emergent theory of voluntary non-migration, or the decision to remain in place, further complicates the conceptual understanding of how environmental stress may increase or dampen migration (Adams 2016; Mallick and Schanze 2020). Because of this complexity, researchers who study migration will often use expert judgement or theory to select which variables to assess. Though this approach can be useful to test theoretically motivated hypotheses and provide insights into how specific drivers might impact migration decisions, it does little to identify which variables might be the most important at driving decisions, especially when considering nonlinear interactions among variables.

As researchers continue to collect large amounts of data with household surveys, challenges may arise in how best to analyze such datasets, especially where motivating theories are unclear or conflicting. To advance the study

of environmental migration and non-migration, especially as large datasets and surveys become more readily available, new methods will need to be employed (Neumann and Hilderink 2015). This work aims to address this need by applying machine learning, specifically random forests, to social survey data for the study of environmental migration in Bangladesh. Random forest is a machine learning approach that has been shown to perform well in environmental and ecological contexts (Cutler et al. 2007; Prasad et al. 2006). However, reviews of methodologies used in studying environmental migration did not mention machine learning techniques (Piguet 2010), and to our knowledge, our application of random forest methods to the topic of environmental migration is novel.

In this work, we present machine learning as a potential tool for social scientists studying environmental migration and non-migration and we describe a case study in which we used, random forests to determine the importance of each covariate in a large dataset for predicting migration outcomes. Though random forest models are able to identify correlates of migration, there exists a tradeoff between high predictive ability and low interpretability. To address this tradeoff, random forests and other complex machine learning algorithms may be especially useful in combination with more traditional, simpler methods. We conduct a survival analysis of household time to first migration using a subset of important variables identified by the random forest algorithm, which provides deeper insight into how important variables impact migration. This multi-methods approach of random forest models and survival analysis provides a data-driven method for identifying and further investigating key variables that impact migration from social datasets.

Machine learning

Machine learning, broadly, refers to a variety of methods that enable a computer or “machine” to automatically recognize patterns in data and use these patterns to build and refine a statistical model of the data without being explicitly programmed to do so and without theoretical or phenomenological preconceptions about the causal mechanisms that gave rise to the data. Machine learning methods are often categorized as supervised or unsupervised. Supervised methods are used to predict one or more specified dependent variables. Unsupervised methods are used to identify patterns in the data (Jordan and Mitchell 2015). To give examples from common statistical methods, regression analyses are supervised methods and exploratory factor analyses are unsupervised methods. In order to guard against overfitting, machine learning models are trained using a subset of the complete data, known as the training set, while the remaining data, known as the holdout or testing set, is withheld and

used for validating the model's performance after the model is fully trained.

Machine learning techniques can outperform standard regression analysis in predictive ability, especially when studying complex social problems (Hindman 2015). Recently, there has been discussion of broadly incorporating machine learning into the social sciences, especially in the place of traditional regression analysis (Hindman 2015; Mason et al. 2014). However, some machine learning algorithms can be very difficult to interpret due to their complexity and this complexity makes it difficult to assess how well a machine learning model is likely to apply outside the specific context in which the data was gathered (Buolawmini and Gebru 2018). While a traditional regression results in coefficients that can be easily interpreted, a more complex machine learning model may be “black box,” making it difficult to draw insights from the model. As the complexity of the model increases, interpretability may decrease, representing a tradeoff between model performance and interpretability (Fig. 1).

Where the predictive power of the model is a priority, complex machine learning algorithms may perform very well. Yet, they are a less appropriate tool for theory development or testing specific hypotheses. The greater predictive power that complex models often possess may arise from models reflecting details of the context in which the data set being analyzed was collected and the models may not transfer as well to other contexts as simpler or theory-driven models would. When the complexity of a model impedes interpretation, it can be difficult to draw on theory or other domain knowledge of the context to evaluate the applicability of a machine learning model to different contexts. Therefore, it is especially important for researchers to carefully consider the goals of their research when selecting a machine learning algorithm, as there is no one size fits all approach.

Nevertheless, machine learning can complement more traditional theory-driven approaches and may have advantages, especially where theory is unclear. Machine learning should be incorporated into social scientists' toolkits for

studying migration because of its ability to identify patterns in complex datasets. We demonstrate one such case study where machine learning—specifically random forest models—are useful in identifying salient variables in a large, complex social survey dataset from Bangladesh.

Case study: migration in Bangladesh

Bangladeshi context

Bangladesh is a country located on the floodplain of the Ganges–Brahmaputra–Jamuna Delta, one of the largest river deltas in the world (Passalacqua et al. 2013). Bangladesh faces environmental vulnerabilities such as flooding and waterlogging, cyclones, and rapid river erosion and accretion (Dewan et al. 2007; Hallegatte 2013; Higgins et al. 2014; Islam and Sado 2000; McGranahan et al. 2007). Bangladesh is also considered one of the most vulnerable countries to climate change (Black et al. 2008; Walsham 2010). Future climate change is expected to create additional environmental stress and uncertainty in the future (Ackerly et al. 2015; Auerbach et al. 2015; Benneyworth et al. 2016; Brammer 2014; Nicholls et al. 2007, 2008; Tessler et al. 2015; Xu et al. 2009).

In Bangladesh, migration is common as a method of livelihood diversification and adaptation (Alam et al. 2017; Bryan et al. 2014; Amrith 2013; Black et al. 2005; Martin et al. 2014). Because of the combined complexity of the human and natural systems, it is unclear how patterns of migration are influenced by environmental change and may be influenced in the future. To begin to address these uncertainties, environmental migration has been widely studied in Bangladesh (Afsar 2003; Ahsan et al. 2011; Bell et al. 2021; Call et al. 2017; Carrico and Donato 2019; Chen and Mueller 2018; Donato et al. 2016; Gray and Mueller 2012; Islam 2017; Joarder and Miller 2013). Some studies in Bangladesh have focused on the impacts of extreme weather events such as cyclones or floods on migration (Kartiki 2011; Gray and Mueller 2012; Lu et al. 2016; Mallick and Vogt 2014). For

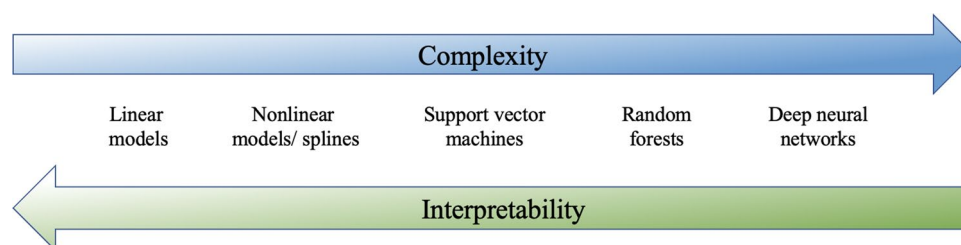


Fig. 1 Schematic demonstrating the tradeoff between complexity and interpretability of common machine learning algorithms. For example, ensemble tree-based methods such as random forests are highly

complex and sometimes challenging to interpret. Researchers should consider where a method falls on this continuum along with specific research goals when selecting an appropriate algorithm

example, Mallick and Vogt assess “disaster-induced population displacement” in the context of the 2009 cyclone Aila in Bangladesh (Mallick and Vogt 2014). They found that male household members tended to migrate towards cities to access livelihood opportunities after the cessation of emergency aid (Mallick and Vogt 2014). In contrast, Gray and Mueller (2012) found that flood events did not influence migration and that crop loss did, but in a complex manner: if crop loss affected only a small number of households in a community, out-migration would decline, whereas if crop loss affected many households, out-migration would rise, with higher status and more affluent households more likely to migrate than lower status and less affluent ones.

Other research has considered slower onset environmental change such as salinity encroachment, temperature change, and changes in precipitation (Call et al. 2017; Carrico and Donato 2019; Chen and Mueller 2018; Perch-Nielsen et al. 2008). Call et al. studied the impacts of temperature, precipitation, and flooding on temporary migration in a non-coastal area, Matlab, Bangladesh (2017). Their work showed that temporary migration declines immediately after a flood, but quickly recovers, while high temperatures consistently increase temporary migration, and precipitation has a strongly nonlinear effect on migration rates (Call et al. 2017). This work supports other research that has indicated that environmental stress could decrease migration and limit the effectiveness of migration as an adaptation strategy (Adger et al. 2015; Gray and Mueller 2012; Bennett et al. 2011). In a more recent study, Carrico and Donato (2019) find that prolonged dry periods, warm periods, and increases in precipitation in Bangladesh may increase migration, especially for households with agricultural livelihoods.

Even within the literature on environmental migration in Bangladesh, there is disagreement in terms of the potential of migration to be a positive adaptation strategy to environmental stress. Though temporary migration is common in Bangladeshi communities, some authors have asserted that permanent migration due to environmental stress may be a last resort for households whose environment becomes inhospitable, potentially suggesting that voluntary non-migration may be influencing such communities’ decision-making (Penning-Rowsell et al. 2013).

Data

Household survey data used in this analysis was collected in the southwest region of Bangladesh by the Bangladesh Environment and Migration Survey (BEMS) in 2014. This survey contains migration, employment, and livelihood histories on more than 3000 individuals affiliated with 1695 households. The data represents 1695 randomly sampled households in nine sites in Bangladesh, which were surveyed in 2014. The survey specifically asks for histories of migration within

Bangladesh, to India, and to any other country (Donato et al. 2016). Here, we focus only on each household’s reported migrations internal to Bangladesh. The original dataset consists of 1695 observations of 1997 distinct variables.

The survey asks respondents to recall the total number of migrations that any member of the household has made, without attributing underlying motivation. This provides the total number of migration trips per household, normalized by total person-years. Person-years were calculated for each member of the household, beginning at age 11, which is the age that many Bangladeshis begin migrating for livelihood opportunities, until 2014 when the survey was collected (Donato et al. 2016). Our analysis takes as its dependent variable this number of trips per person-years, which may be interpreted as the annual probability of making a migration. This is represented as a continuous variable at the household level.

Random forest models

Random forest models are an ensemble method of decision trees and represent a subset of machine learning known as tree-based methods. Tree-based methods, including random forests, can be used for the classification of discrete outcome variables, or regression of continuous variables. They are especially powerful tools when there are strong nonlinearities or interactions between variables in the data.

Random forests models work by fitting many decision trees, where each tree uses a random subset of the predictor variables at each split in its decision tree. The final prediction is then calculated by averaging across the outputs of all of the individual decision trees (Hastie et al. 2009, Ch. 15). This allows random forest models to achieve high predictive accuracy without overfitting (James et al. 2013). One strength of random forest models, especially over other “black box” statistical models, is their ability to assess variable importance and account for complex, nonlinear interactions between variables. Random forest models are also able to use combinations of categorical, ordinal, and continuously valued variables as inputs without requiring dummy variables or scaled data. This makes them especially appealing tools for analyzing large social surveys and studying complex challenges such as migration. However, it can be difficult to interpret a random forest model. The ensemble of trees, each with a different subset of predictor variables, makes it impractical, if not impossible, to establish a clear or descriptive relationship between independent and dependent variables. Thus, while these models are powerful, they are very much black boxes in comparison to the ways we can understand and interpret regression or single-tree models. Random forest models were chosen for this analysis because we found in previous work that they outperformed linear

regressions and support vector machines in predicting the migration outcome for this data set (Best et al. 2020).

Random forests allow us to rank variables by their importance (i.e., their contribution to the overall model performance) (Hastie et al. 2009). For regression random forest models, importance is calculated by node impurity, which is a calculation of how much a split in the decision trees by a specific variable can decrease variance in the outcome (Hastie et al. 2009). Variable importance cannot be meaningfully compared across different datasets or different models, but is useful for comparing the significance of variables within a specific model trained on a specific dataset.

We fit 10 models to the survey data. For each model, we divided the data, randomly assigning 80% of the household responses to a training set and the remaining 20% to a testing data set. We used the *randomForest* package in R to fit a complete random forest regression model to each of the 10 training data sets and evaluated its out-of-sample predictive performance on the corresponding testing data set (Cutler et al. 2018). The regression models predicted the continuous outcome variable of total internal migration trips per household normalized by person-years (Donato et al. 2016). For each of the 10 models, the parameter for the number of variables randomly sampled at each split was tuned by minimizing the out-of-sample error using the *tuneRF* function in the *randomForest* package. After tuning, 10 complete models were fitted using the optimum tuning parameters. Variable importance was ranked and averaged across the 10 complete models, each model's predictive performance was assessed using its testing data set (the 20% of data not used for training). A full explanation of the random forest methods and results, including model performance metrics, can be found in Best et al. 2020.

To further validate the ranked variable importance from our random forest models beyond Best et al. (2020), we divided the complete survey dataset into five groups, each consisting of 20% of the data, and conducted a fivefold cross-validation where each fold chose a different one of the five subsets of data to use as the validation set for a random forest model fit to the other four subsets. We then compared the predictive performance of a model using all the variables and another random forest model using just the top 15 variables identified from the training set when fit to the holdout validation set. We found that, across the five models in this fivefold cross-validation, the top 10 variables of importance were consistent, and there was a slight movement in the bottom five variables between models. We found that models using all the variables had a mean RMSE of 2.65, while the models using just the identified top 15 variables had an average RMSE of 2.67. This is consistent with our understanding that the top 15 variables are robust and account for almost all of the model performance across different subsets of the data.

Survival analysis

Survival analysis is a technique used to study the occurrence of a discrete event where the time until the event matters (Harrell 2015). The response variable in survival models is time until the event, usually referred to as failure time, survival time, or event time. Survival analysis has been widely used in biomedical research to describe times to a disease event (Bull and Spiegelhalter 1997; Crowley and Hu 1977; Prentice et al. 1981), failure or recovery times in engineering systems (Ansell and Philipps 1997; Barker and Baroud 2014), and binary events in demography and the social sciences, including the timing of a woman's first child (Teachman 1983) and when people make a first migrant trip (Donato et al. 1992). Survival analysis also allows for some responses to be incomplete, meaning that the event of interest has not occurred within the observed time. Such responses are censored, and responses for which the event of interest did occur within the study time are uncensored.

For the survival analysis, time in person-years from age 11 to first internal migration by the head of the household was used. This generated a discrete-time person-year file that followed the male head of the household. The age of 11 was chosen as the starting point because this is the age at which many Bangladeshi males begin engaging in paid work. For each year from age 11 to the date of the survey, each male head of household received a 1 if they did complete a trip and a 0 if they did not complete a trip. In this way, the individual migration data was divided into censored and uncensored data for a survival model, as some heads have not completed their first migration by the time of the data collection. Only a small minority of heads of household had ever migrated, so 17.3% of the data was uncensored and the remaining 82.7% was censored.

We used Cox proportional hazards models to estimate the survival and hazards function corresponding to the probability of internal migration and to assess the relative effects of the different covariates (Ansell and Philipps 1997; Harrell 2015). The Cox model is a semi-parametric proportional hazards model, but the regression portion of the model is parametric and assumes that covariates are linearly related to the log of the hazard. This approach is ideal when data is not easily fit to a distribution and when the form of the true hazard function is complex. It is also a useful approach when the key question of concern is how covariates impact the hazard, rather than the shape of the hazard itself (Harrell 2015).

However, Cox models, like most proportional hazards models, can only represent monotonic relationships between covariates and hazard, whereas tree-based models, such as random forests, can represent arbitrarily complex nonlinear and non-monotonic relationships. Thus, if a covariate identified by the random forest models has a non-monotonic

relationship to migration, a Cox model will perform poorly with that covariate.

Results

Salient variables from random forest models

The fitted random forest models provided a rank order of variable importance, which were averaged across all 10 models. The results of the variable importance assessment from the random forest model of the survey data have been highlighted in previous work (Best et al. 2020) and are given in Table 1. The 15 most important variables are presented in order of descending model importance. The range of variable importance rank across the 10 variables is available in Supplementary Materials (Figure S1).

Survival analysis

Univariate Cox proportional hazards models were fit for each of the salient variables in Table 1 identified by the random forest models. For each univariate model, the estimated value of the coefficient “Beta” and the estimated hazard ratio (HR) and 95% confidence interval boundaries are presented (Table 2). In Table 2, we also present the concordance statistic, which is a measure of predictive ability for survival analysis which measures the proportion of pairs of observations in which predictions and outcomes agree (Harrell et al. 1996). While concordance is a common method of measuring predictive ability in survival analysis, we also present the R^2 value and the p -value as commonly employed and widely understood measures of model performance (Table 2).

The hazard ratio describes how a covariate impacts the hazard (whether it has a positive or negative effect) (Harrell 2015). The hazard ratio for a covariate is calculated by computing the ratio of the hazard for that covariate over the baseline hazard. Therefore, a hazard ratio of 1 indicates that the covariate has no effect on the hazard. A hazard ratio less than 1 means that the covariate reduces the hazard of an event, and a hazard ratio greater than 1 means that the covariate increases the hazard from the baseline.

While we do not employ an arbitrary p -value significance threshold, the variables “Latitude,” “Longitude,” and “Who owns water source” have large p -values which are greater than 0.2 and orders of magnitude greater than the p -values for all other variables. This led us to conclude that these are not useful predictors, so we excluded them from the analysis going forward. Furthermore, the uncertainties in regression coefficients for the variables related to the most recent cyclone, female toilet, and water source were very large, with 95% confidence intervals that include the hazard ratio of 1. This means that we cannot be confident that these variables affect the survival function, so we excluded them from the continued analysis.

Next, a series of nested Cox proportional hazards models were developed with the remaining variables by starting with a univariate model and systematically adding an additional significant covariate to the model (Table 3). The hazard ratios for the covariates of the complete model are given in Fig. 2.

Discussion

Random forests applied to migration

Variables of importance were identified using random forest models to identify patterns in the data (Table 1, Best et al.

Table 1 Variables of importance identified by random forest model of migration and original survey questions

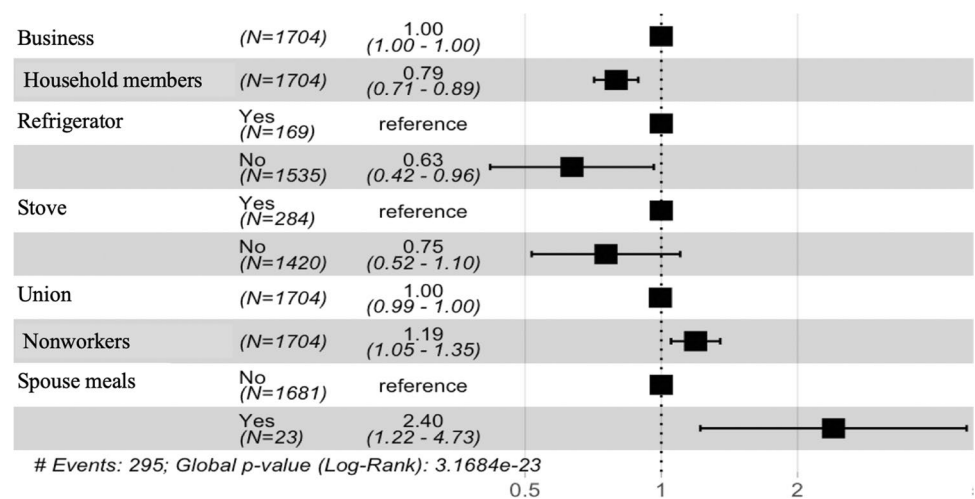
Variable name	Survey question
Latitude	Household latitude
Water Sources: Who owns?	Who owns the primary water source?
Business: year started	What is the year that your business was started?
Longitude	Household longitude
Household: total number of members	How many household members are living in the home?
Cyclone: source of water	What was your principle source of water during the last cyclone?
Cyclone: Source of food	What was your principle source of food during the last cyclone?
Kerosene/gas cooker	Do you own a gas or kerosene cooker?
Where would you go for future cyclone	Where would you go if there was a future cyclone?
Refrigerator	Do you own a refrigerator?
Non-workers	What is the total number of non-workers in the household?
Union	Local government administrative division
Primary water source	What is the household's primary water source?
Toilet-female	What kind of toilet facility do female household members use?
Spouse prepared meals consumed	Has household consumed prepared meals? If yes, who? Spouse

Table 2 Results of univariate Cox proportional hazards models with each salient variable identified by the random forest models. For each univariate model, the fitted coefficient Beta is presented, along with the hazard ratio (HR) and 95% confidence intervals for HR, the generalized R^2 , concordance statistic, and p -value

Variable	Beta	HR (95% CI for HR)	R^2	Concordance	P-value
Latitude	-0.0012	0.999 (0.994–1)	0.000181	0.514	0.653
Business: year started	-0.000406	1 (0.999–1)	0.0260	0.559	1.10e-10
Longitude	-0.000651	0.999 (0.995–1)	7.14e-05	0.434	0.750
Total members in household	-0.129	0.879 (0.831–0.93)	0.0133	0.592	7.45e-06
Kerosene/ gas cooker	-1.02	0.360 (0.274–0.472)	0.0256	0.557	1.94e-13
Refrigerator	-0.801	0.449 (0.351–0.574)	0.0211	0.558	1.71e-19
Cyclone: Source of water	0.902	2.22 (0.223–27.2)	0.0164	0.555	4.14e-05
Cyclone: Source of Food	0.232	-0.242 (0.874–1.82)	0.0160	0.539	9.43e-05
Where would you go for future cyclone	-13.5	0.146 (0.853–1.57)	0.0111	0.527	0.000221
Who owns water source	-0.357	0.0357 (0.378–1.3)	0.00405	0.537	0.292
Spouse prepared meals consumed	1.14	3.11 (1.6–6.05)	0.00474	0.513	0.000801
Union	-0.0098	0.990 (0.986–0.994)	0.0131	0.558	3.38e-06
Non-workers	-0.0909	0.913 (0.859–0.971)	0.00532	0.561	0.00356
Toilet- female	0.226	0.646 (0.934–1.68)	0.0141	0.562	0.000154
Primary water source	0.4581	1.581 (0.9928–2.518)	0.00889	0.515	0.0988

Table 3 Nested Cox proportional hazards models of increasing complexity, generalized R^2 , and concordance

Model	R^2	Concordance
Business	0.026	0.559
Business + household members	0.047	0.637
Business + household members + refrigerator	0.057	0.637
Business + household members + refrigerator + stove	0.06	0.642
Business + household members + refrigerator + stove + union	0.063	0.647
Business + household members + refrigerator + stove + union + Non-workers	0.066	0.657
Business + household members + refrigerator + stove + union + non-workers + spouse prepared meals	0.069	0.657

Fig. 2 Hazard ratios for the final Cox proportional hazards model. A hazard ratio greater than 1 (to the right of the dashed line) indicates that the variable increases mobility, while a hazard ratio less than 1 (to the left of the dashed line) indicates that the variable decreases mobility

2020). No researcher judgement or selection of variables from the large survey dataset was required. This work demonstrates that random forest models can help researchers identify salient variables from large social surveys when studying migration. This is especially useful when dealing with large, complex datasets from social surveys, where it can be challenging to decide which variables are worthwhile for further investigation. In this work, random forest models were able to identify the most important predictors of migration from an original set of approximately 2000 total predictors. Thus, the random forest served as a method of variable reduction which allowed us to conduct our regression analysis with fewer variables and more degrees of freedom.

Variable impact on migration

While random forest models can tell researchers which variables are the most important for predicting the migration outcome, they do little to provide insight into how specific variables impact migration. To dig deeper into the variables identified by the random forest models, survival analysis was implemented, which further illuminates how salient variables related to location, livelihood, and family structure might impact a household's risk of internal migration in coastal Bangladeshi communities. The univariate Cox proportional hazards models outlined in Table 2 demonstrate that the number of members in a household, the year a business is started, whether or not the household owns a refrigerator, and whether or not the household owns a gas cooker were significant. Latitude, longitude, and variables related to the most recent cyclone did not contribute significantly to the hazard function or reflected too much uncertainty to be reliable.

It is especially surprising that latitude and longitude were not significant covariates given that they were the first and fourth most important variables identified by the previous work using random forest algorithms. It was thought that latitude especially would be significant, because there is a clear gradient of increasing soil salinity from north to south in Bangladesh, and previous studies have suggested that soil salinity is important for driving migration in Bangladesh (Chen and Mueller 2018). It is possible that the random forest algorithm is able to identify nonlinear and non-monotonic patterns in the latitude and longitude data, whereas the Cox proportional hazards model assumes monotonicity in the baseline hazard function and multiplicative effects of the predictors on the hazard. For example, the random forest algorithm would be able to identify geographic clusters of migration and the Cox proportional hazards model would not.

The best performing one of the nested models was the complete model with *year business was started*, *refrigerator ownership*, *gas cooker ownership*, *total members in the household*, *union*, and *prepared meal consumption by household head*. This final model had a generalized R^2 value of 0.069 and a concordance of 0.657 (Table 3). It is possible that this value of R^2 is so low because, again, the covariates are unlikely to follow a simple multiplicative relationship assumed by the Cox proportional hazards model.

Despite the low value of R^2 and concordance, the multivariate Cox proportional hazards model is useful in beginning to understand how these variables influence the underlying risk of migrating. The values of hazard ratios shown in Fig. 2 quantify these impacts. The hazard ratios to the left of the dotted line in the figure show the variables have a negative impact on the overall risk of migration. This means that these variables decrease the underlying hazard. These variables include *total household members*, *not owning a refrigerator*, and *not owning a gas or kerosene cooker*. Hazard ratios that fall to the right of the dotted line in Fig. 2 show variables that have a positive impact on migration, meaning they increase the underlying hazard of migration. These variables are the *number of non-workers in the household* and *prepared meal consumption by the spouse of the household head*.

These results show that the *total number of members of a household* has a negative impact on migration, while *number of non-workers in the household* seems to increase migration by the household head. It is possible that this is in part due to the importance of remittances that migratory members of a household can send home to support their families (Massey 1990). A household with a higher number of non-workers to support may be more dependent on remittances from a migratory head of household. However, it seems that larger households may also create an anchoring effect that keeps the head of household from migrating, perhaps because migrating from the household, even temporarily, would leave the household more vulnerable and economically stressed. Such results could also indicate that having additional household members increases attachment to place and voluntary non-migration (Adams 2016; Mallick and Schanze 2020). This suggests that household size has a complex and possibly non-monotonic effect on probabilities of migration which reflects household livelihood capacity as well as vulnerability. This also supports existing literature that suggest that migration decisions may be primarily made at the household level (Massey et al. 1993).

Implications for migration / non-migration research

As several researchers have noted, the field of environmental migration has been growing over time, as have the methods employed (Piguet 2022). Despite advancements in the field,

there remain important and unanswered questions related to how environmental or climatic change interact with mobility and immobility (i.e., migration versus non-migration) decisions (Mallick and Schanze 2020). In addition, how do migration and non-migration decisions vary across individuals and households and across contexts? Just as the drivers of environmental migration are acknowledged to be complex and interconnected, the drivers of environmental non-migration (both voluntary and involuntary) must be similarly studied in detail by the field (Mallick and Schanze 2020).

This work, which combines survey data, random forest algorithms, and survival analysis to investigate migration in rural Bangladeshi communities, has several important implications for migration and non-migration research. First, we provide specific insights into drivers of migration and non-migration in Bangladesh. We show that indicators of lower economic resources (*not owning a refrigerator* and *not owning a gas or kerosene cooker*) work to reduce mobility, suggesting that much of the non-mobility in our study location may be involuntary and driven by a household's inability to afford to move. Similarly, *number of non-workers in the household* increases mobility, which supports the idea that much mobility in the area is primarily motivated by the desire to seek livelihood opportunities outside of the origin community (Bernzen et al. 2019; Biswas et al. 2019).

More broadly, by using multiple machine learning methods in combination, we provide an example of how survey data can be used to provide insights into (non-)migration when the relevant underlying theory is unknown or unclear. Mallick and Schanze propose that migration and non-migration may be considered on a spectrum of aspirations and capabilities (2020). However, how these aspirations and capabilities may be operationalized in data remains unclear. The methods demonstrated here can be used to identify important variables from existing datasets and then quantitatively show how those variables amplify or dampen mobility. These methods may be applied to different datasets and contexts and would yield context-specific insights into which factors influence (non-)migration.

Conclusion

Machine learning methods can be useful tools for researchers to study environmental migration when theory is not clearly established, as is the case with the emergent theory of voluntary non-migration. Though the specific machine learning algorithm used will vary based on research objectives and data used, this work applies random forest models to a household survey of migration in Bangladesh in order to identify salient variables. An important downside to random forest models is that despite quantifying variable importance, they do not provide insights into how the

individual predictors relate to the outcome variable (e.g., does increasing the predictor variable increase or decrease the outcome variable?). Therefore, where theory testing or development is the goal, complex machine learning algorithms such as random forest models may not be useful in isolation. Instead, researchers may use machine learning to direct additional analysis using more traditional regression analysis or, as in this case, survival analysis. This multi-methods analysis provides insights into migration dynamics, but it does not begin to accurately quantify migration risks. Assumptions of linearity in the survival analysis contribute to low predictive power, further demonstrating the strengths and weaknesses of different algorithms and methods.

Future work should continue to develop modeling methods that are able to capture the complex relationship between the many factors that contribute to migration or non-migration decisions. In this process of improving methods, it is likely that no one method will be a clear solution to existing challenges, but methods that draw from the best available computer science methods will likely be important (Neumann and Hilderink 2015; Obokata et al. 2014). In this way, researchers should remain open to investigating new techniques that may be useful, such as more complex machine learning algorithms.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10113-022-01915-1>.

Funding This work was supported by the National Science Foundation Coupled Human-Natural Systems Grant No. 1716909.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Ackerly BA, Anam MM, Gilligan J (2015) Environment, political economies and livelihood change. In B. Mallick & B. Etzold (Eds.), *Environment, migration and adaptation: evidence and politics of climate change in Bangladesh*. Retrieved from <http://eprints.qut.edu.au/84192/>
- Adams H (2016) Why populations persist: mobility, place attachment and climate change. *Population and Environment* 37(4):429–448. <https://doi.org/10.1007/s11111-015-0246-3>

- Adams H, Kay S (2019) Migration as a human affair: Integrating individual stress thresholds into quantitative models of climate migration. *Environ Sci Policy* 93:129–138. <https://doi.org/10.1016/j.envsci.2018.10.015>
- Adger WN, Arnell NW, Black R, Dercon S, Geddes A et al (2015) Focus on environmental risks and migration: causes and consequences. *Environ Res Lett* 10(6):060201. <https://doi.org/10.1088/1748-9326/10/6/060201>
- Afsar R (2003) Internal migration and the development nexus: the case of Bangladesh. *Regional Conference on Migration, Development and Pro-Poor Policy Choices in Asia*, 22–24
- Ahsan R, Karuppannan S, Kellett J (2011) Climate migration and urban planning system: a study of Bangladesh. *Environmental Justice* 4(3):163–170. <https://doi.org/10.1089/env.2011.0005>
- Alam GMM, Alam K, Mushtaq S (2017) Climate change perceptions and local adaptation strategies of hazard-prone rural households in Bangladesh. *Clim Risk Manag* 17:52–63. <https://doi.org/10.1016/j.crm.2017.06.006>
- Amrith SS (2013) *Crossing the Bay of Bengal: the furies of nature and the fortunes of migrants*. Cambridge, Mass. London: Harvard University Press
- Ansell JJ, Philipps MJ (1997) Practical aspects of modelling of repairable systems data using proportional hazards models. *Reliab Eng Syst Saf* 58(2):165–171. [https://doi.org/10.1016/S0951-8320\(97\)00026-4](https://doi.org/10.1016/S0951-8320(97)00026-4)
- Auerbach LW, Goodbred SL Jr, Mondal DR, Wilson CA, Ahmed KR et al (2015) Flood risk of natural and embanked landscapes on the Ganges–Brahmaputra tidal delta plain. *Nat Climate Chang* 5(2):153–157. <https://doi.org/10.1038/nclimate2472>
- Barker K, Baroud H (2014) Proportional hazards models of infrastructure system recovery. *Reliab Eng Syst Saf* 124:201–206. <https://doi.org/10.1016/j.res.2013.12.004>
- Bell AR, Wrathall DJ, Mueller V, Chen J, Oppenheimer M et al (2021) Migration towards Bangladesh coastlines projected to increase with sea-level rise through 2100. *Environ Res Lett*. <https://doi.org/10.1088/1748-9326/abdc5b>
- Bennett G, Thomas SM, Beddington JR (2011) Migration as adaptation. *Nature* 478:447–449. <https://doi.org/10.1038/478477a>
- Benneyworth L, Gilligan J, Ayers JC, Goodbred S, George G et al (2016) Drinking water insecurity: water quality and access in coastal south-western Bangladesh. *Int J Environ Health Res* 26(5–6):508–524. <https://doi.org/10.1080/09603123.2016.1194383>
- Bernzen A, Jenkins JC, Braun B (2019) Climate change-induced migration in coastal Bangladesh? A critical assessment of migration drivers in rural households under economic and environmental stress. *Geosciences* 9(1):51. <https://doi.org/10.3390/geosciences9010051>
- Best KB, Gilligan JM, Baroud H, Carrico AR, Donato KM et al (2020) Random forest analysis of two household surveys can identify important predictors of migration in Bangladesh. *Journal of Computational Social Science*. <https://doi.org/10.1007/s42001-020-00066-9>
- Biermann F, Boas I (2010) Preparing for a warmer world: towards a global governance system to protect climate refugees (Vol. 10). <https://doi.org/10.1162/glep.2010.10.1.60>
- Bilsborrow RE, Henry SJF (2012) The use of survey data to study migration–environment relationships in developing countries: alternative approaches to data collection. *Popul Environ* 34(1):113–141. <https://doi.org/10.1007/s11111-012-0177-1>
- Biswas RK, Kabir E, Khan H (2019) Causes of urban migration in Bangladesh: evidence from the urban health survey. *Popul Res Policy Rev*. <https://doi.org/10.1007/s11113-019-09532-3>
- Black R, Adger WN, Arnell NW, Dercon S, Geddes A et al (2011) The effect of environmental change on human migration. *Glob Environ Chang* 21:S3–S11. <https://doi.org/10.1016/j.gloenvcha.2011.10.001>
- Black R, Natali C, Skinner J (2005) *Migration and inequality*. World Bank Washington, DC
- Black R, Kniveton D, Skeldon R, Coppard D, Murata A et al (2008) *Demographics and climate change: future trends and their policy implications for migration*. Development Research Centre on Migration, Globalisation and Poverty. Brighton: University of Sussex
- Boas I, Farbotko C, Adams H, Sterly H, Bush S et al (2019) Climate migration myths. *Nat Clim Change* 9(12):901–903. <https://doi.org/10.1038/s41558-019-0633-3>
- Brammer H (2014) Bangladesh's dynamic coastal regions and sea-level rise. *Clim Risk Manag* 1:51–62. <https://doi.org/10.1016/j.crm.2013.10.001>
- Bryan G, Chowdhury S, Mobarak AM (2014) Underinvestment in a profitable technology: the case of seasonal migration in Bangladesh. *Econometrica* 82(5):1671–1748. <https://doi.org/10.3982/ECTA10489>
- Bull K, Spiegelhalter DJ (1997) Survival analysis in observational studies. *Stat Med* 16(9):1041–1074
- Buolawmini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Machine Learning Res* 81:77–91. http://proceedings.mlr.press/v81/buolawmini18a.html?mod=article_inline
- Cai R, Oppenheimer M (2013) An agent-based model of climate-induced agricultural labor migration. 2013 Annual Meeting, August, 4–6
- Call MA, Gray C, Yunus M, Emch M (2017) Disruption, not displacement: Environmental variability and temporary migration in Bangladesh. *Glob Environ Chang* 46:157–165. <https://doi.org/10.1016/j.gloenvcha.2017.08.008>
- Carrico AR, Donato K (2019) Extreme weather and migration: evidence from Bangladesh. *Popul Environ*. <https://doi.org/10.1007/s11111-019-00322-9>
- Chen J, Mueller V (2018) Coastal climate change, soil salinity and human migration in Bangladesh. *Nat Clim Chang*. <https://doi.org/10.1038/s41558-018-0313-8>
- Crowley J, Hu M (1977) Covariance analysis of heart transplant survival data. *J Am Stat Assoc* 72(357):27–36. <https://doi.org/10.1080/01621459.1977.10479903>
- Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT et al (2007) Random forests for classification in ecology. *Ecology* 88(11):2783–2792. <https://doi.org/10.1890/07-0539.1>
- Cutler F. original by L. B. and A., & Wiener, R. port by A. L. and M. (2018) *RandomForest: Breiman and Cutler's random forests for classification and regression* (Version 4.6–14). Retrieved from <https://CRAN.R-project.org/package=randomForest>
- Dewan AM, Islam MM, Kumamoto T, Nishigaki M (2007) Evaluating flood hazard for land-use planning in Greater Dhaka of Bangladesh using remote sensing and GIS techniques. *Water Resour Manage* 21(9):1601–1612. <https://doi.org/10.1007/s11269-006-9116-1>
- Donato KM, Durand J, Massey DS (1992) Changing conditions in the US labor market. *Population Res Policy Rev* 11(2):93–115. <https://doi.org/10.1007/BF00125533>
- Donato KM, Carrico AR, Sisk B, Piya B (2016) Different but the same: how legal status affects international migration from Bangladesh. *Ann Am Acad Pol Soc Sci* 666(1):203–218. <https://doi.org/10.1177/0002716216650843>
- Fussell E, Hunter LM, Gray CL (2014) Measuring the environmental dimensions of human migration: the demographer's toolkit. *Glob Environ Chang* 28:182–191. <https://doi.org/10.1016/j.gloenvcha.2014.07.001>

- Gemenne F (2011) Why the numbers don't add up: a review of estimates and predictions of people displaced by environmental changes. *Glob Environ Chang* 21:S41–S49. <https://doi.org/10.1016/j.gloenvcha.2011.09.005>
- Gray CL, Mueller V (2012) Natural disasters and population mobility in Bangladesh. *Proc Natl Acad Sci* 109(16):6000–6005. <https://doi.org/10.1073/pnas.1115944109>
- Harrell FE (2015). Regression Modeling Strategies. <https://doi.org/10.1007/978-3-319-19425-7>
- Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960229\)15:4%3c361::AID-SIM168%3e3.0.CO;2-4](https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3c361::AID-SIM168%3e3.0.CO;2-4)
- Hassani-Mahmooei B, Parris BW (2012) Climate change and internal migration patterns in Bangladesh: an agent-based model. *Environ Dev Econ* 17(6):763–780. <https://doi.org/10.1017/S1355770X12000290>
- Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
- Henry S, Boyle P, Lambin EF (2003) Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. *Appl Geogr* 23(2–3):115–136. <https://doi.org/10.1016/j.apgeog.2002.08.001>
- Henry S, Schoumaker B, Beauchemin C (2004) The impact of rainfall on the first out-migration: a multi-level event-history analysis in Burkina Faso. *Popul Environ* 25(5):423–460. <https://doi.org/10.1023/B:POEN.0000036928.17696.e8>
- Higgins SA, Overeem I, Steckler MS, Syvitski JPM, Seeber L et al (2014) InSAR measurements of compaction and subsidence in the Ganges-Brahmaputra Delta, Bangladesh. *J Geophys Res Earth Surf* 119(8):1768–1781. <https://doi.org/10.1002/2014JF003117>
- Hindman M (2015) Building better models: prediction, replication, and machine learning in the social sciences. *Ann Am Acad Pol Soc Sci* 659(1):48–62. <https://doi.org/10.1177/0002716215570279>
- Hino M, Field CB, Mach KJ (2017) Managed retreat as a response to natural hazard risk. *Nat Clim Chang* 7(5):364–370. <https://doi.org/10.1038/nclimate3252>
- Hugo G (1996) Environmental concerns and international migration. *Int Migr Rev* 30(1):105–131. <https://doi.org/10.2307/2547462>
- Hugo G (2011) Future demographic change and its interactions with migration and climate change. *Glob Environ Chang* 21:S21–S33. <https://doi.org/10.1016/j.gloenvcha.2011.09.008>
- Islam MR (2017) Climate change, natural disasters and socioeconomic livelihood vulnerabilities: migration decision among the Char Land People in Bangladesh. *Soc Indic Res*. <https://doi.org/10.1007/s11205-017-1563-y>
- Islam MM, Sado K (2000) Development of flood hazard maps of Bangladesh using NOAA-AVHRR images with GIS. *Hydrol Sci J* 45(3):337–355. <https://doi.org/10.1080/02626660009492334>
- James G, Witten D, Hastie T, Tibshirani R (eds) (2013) An introduction to statistical learning: with applications in R. Springer, New York
- Joarder MAM, Miller PW (2013) Factors affecting whether environmental migration is temporary or permanent: evidence from Bangladesh. *Glob Environ Chang* 23(6):1511–1524. <https://doi.org/10.1016/j.gloenvcha.2013.07.026>
- Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. *Science* 349(6245):255–260. <https://doi.org/10.1126/science.aaa8415>
- Kartiki K (2011) Climate change and migration: a case study from rural Bangladesh. *Gend Dev* 19(1):23–38. <https://doi.org/10.1080/13552074.2011.554017>
- Klabunde A, Zinn S, Leuchter M, Willekens F (2015) An agent-based decision model of migration, embedded in the life course -model description in ODD+D format (Working Paper No. 2015– 002) (p. 32). Rostock, Germany: Max Planck Institute for Demographic Research. Retrieved from <https://www.demogr.mpg.de/papers/working/wp-2015-002.pdf>
- Kniveton D, Smith C, Wood S (2011) Agent-based model simulations of future changes in migration flows for Burkina Faso. *Glob Environ Chang* 21:S34–S40. <https://doi.org/10.1016/j.gloenvcha.2011.09.006>
- Koubi V, Spilker G, Schaffer L, Bernauer T (2016) Environmental stressors and migration: evidence from Vietnam. *World Dev* 79:197–210. <https://doi.org/10.1016/j.worlddev.2015.11.016>
- Lu X, Wrathall DJ, Sundsøy PR, Nadiruzzaman Md, Wetter E et al (2016) Unveiling hidden migration and mobility patterns in climate stressed regions: a longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environ Chang* 38:1–7. <https://doi.org/10.1016/j.gloenvcha.2016.02.002>
- Mallick B, Schanze J (2020) Trapped or voluntary? Non-Migration despite Climate Risks. *Sustainability* 12(11):4718. <https://doi.org/10.3390/su12114718>
- Mallick B, Vogt J (2014) Population displacement after cyclone and its consequences: empirical evidence from coastal Bangladesh. *Nat Hazards* 73(2):191–212. <https://doi.org/10.1007/s11069-013-0803-y>
- Martin M, Billah M, Siddiqui T, Abrar C, Black R et al (2014) Climate-related migration in rural Bangladesh: a behavioural model. *Popul Environ* 36(1):85–110. <https://doi.org/10.1007/s11111-014-0207-2>
- Mason W, Vaughan JW, Wallach H (2014) Computational social science and social computing. *Mach Learn* 95(3):257–260. <https://doi.org/10.1007/s10994-013-5426-8>
- Massey DS (1990) Social structure, household strategies, and the cumulative causation of migration. *Popul Index* 56(1):3–26. <https://doi.org/10.2307/3644186>
- Massey DS, Arango J, Hugo G, Kouaouci A, Pellegrino A et al (1993) Theories of international migration: a review and appraisal. *Popul Dev Rev* 19(3):431–466. <https://doi.org/10.2307/2938462>
- McGranahan G, Balk D, Anderson B (2007) The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environ Urban* 19(1):17–37. <https://doi.org/10.1177/0956247807076960>
- McLeman R (2013) Developments in modelling of climate change-related migration. *Clim Change* 117(3):599–611. <https://doi.org/10.1007/s10584-012-0578-2>
- Neumann K, Hilderink H (2015) Opportunities and challenges for investigating the environment-migration Nexus. *Hum Ecol* 43(2):309–322. <https://doi.org/10.1007/s10745-015-9733-5>
- Nicholls RJ, Wong PP, Burkett V, Woodroffe CD, Hay J (2008) Climate change and coastal vulnerability assessment: scenarios for integrated assessment. *Sustain Sci* 3(1):89–102. <https://doi.org/10.1007/s11625-008-0050-4>
- Nicholls RJ, Wong PP, Burkett V, Codignotto J, Hay J et al (2007) Coastal systems and low-lying areas
- Obokata R, Veronis L, McLeman R (2014) Empirical research on international environmental migration: a systematic review. *Popul Environ* 36(1):111–135. <https://doi.org/10.1007/s11111-014-0210-7>
- Passalacqua P, Lanzoni S, Paola C, Rinaldo A (2013) Geomorphic signatures of deltaic processes and vegetation: the Ganges-Brahmaputra-Jamuna case study. *J Geophys Res Earth Surf* 118(3):1838–1849. <https://doi.org/10.1002/jgrf.20128>
- Penning-Rowsell EC, Sultana P, Thompson PM (2013) The 'last resort'? Population movement in response to climate-related hazards in Bangladesh. *Environ Sci Policy* 27:S44–S59. <https://doi.org/10.1016/j.envsci.2012.03.009>
- Perch-Nielsen SL, Bättig MB, Imboden D (2008) Exploring the link between climate change and migration. *Climatic Change* 91(3–4):375–393. <https://doi.org/10.1007/s10584-008-9416-y>

- Piguet E (2010) Linking climate change, environmental degradation, and migration: a methodological overview. *Wiley Interdisciplinary Reviews: Climate Change* 1(4):517–524. <https://doi.org/10.1002/wcc.54>
- Piguet E (2022) Linking climate change, environmental degradation, and migration: An update after 10 years. *Wires Clim Change* 13(1):e746. <https://doi.org/10.1002/wcc.746>
- Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems* 9(2):181–199. <https://doi.org/10.1007/s10021-005-0054-1>
- Prentice RL, Williams BJ, Peterson AV (1981) On the regression analysis of multivariate failure time data. *Biometrika* 68(2):373–379. <https://doi.org/10.1093/biomet/68.2.373>
- Renaud FG, Dun O, Warner K, Bogardi J (2011) A decision framework for environmentally induced migration: framework for environmentally induced migration. *Int Migr* 49:e5–e29. <https://doi.org/10.1111/j.1468-2435.2010.00678.x>
- Silveira JJ, Espindola AL, Penna TJP (2006) An agent-based model to rural-urban migration analysis. *Physica A* 364:445–456. <https://doi.org/10.1016/j.physa.2005.08.055>
- Smith CD (2014) Modelling migration futures: development and testing of the Rainfalls Agent-Based Migration Model – Tanzania. *Climate Dev* 6(1):77–91. <https://doi.org/10.1080/17565529.2013.872593>
- Stern N (2006) the Price of Change. *IAEA Bull* 48(2):25
- Subrahmanian VS, Kumar S (2017) Predicting human behavior: The next frontiers. *Science* 355(6324):489–489. <https://doi.org/10.1126/science.aam7032>
- Teachman JD (1983) Analyzing social processes: Life tables and proportional hazards models. *Soc Sci Res* 12(3):263–301. [https://doi.org/10.1016/0049-089X\(83\)90015-7](https://doi.org/10.1016/0049-089X(83)90015-7)
- Tejero DG, Guadagno L, Nicoletti A (2020) Human mobility and the environment: challenges for data collection and policymaking. <https://www.semanticscholar.org/paper/Human-mobility-and-the-environment-%3A-Challenges-for-Tejero-Guadagno/8060883bcf4213b3d177c5c7be2af2616649a690>. Accessed 30 Aug 2021
- Tessler ZD, Vörösmarty CJ, Grossberg M, Gladkova I, Aizenman H et al (2015) Profiling risk and sustainability in coastal deltas of the world. *Science* 349(6248):638–643. <https://doi.org/10.1126/science.aab3574>
- Thober J, Schwarz N, Hermans K (2018) Agent-based modeling of environment-migration linkages: a review. *Ecol Soc* 23:<https://doi.org/10.5751/ES-10200-230241>
- Walsham M (2010) Assessing the evidence: environment, climate change and migration in Bangladesh. International Organization for Migration
- Xu J, Grumbine RE, Shrestha A, Eriksson M, Yang X et al (2009) The Melting Himalayas: Cascading Effects of Climate Change on Water, Biodiversity, and Livelihoods. *Conserv Biol* 23(3):520–530. <https://doi.org/10.1111/j.1523-1739.2009.01237.x>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.