Impact of Weather Factors on Migration Intention using Machine Learning Algorithms

John O.R. Aoga[1*], Juhee Bae[2*], Stefanija Veljanoska[3], Siegfried Nijssen[4] and Pierre Schaus[4]

[1*]Ecole Doctorale Science Pour Ingenieur, Université d'Abomey-Calavi, Abomey-Calavi, Benin.
[2*]School of Informatics, University of Skövde, Skövde, Sweden.
[3*] Université de Rennes 1, CNRS/CREM-UMR621, Rennes, France.
[4*]ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, 1348, Belgium.

*Corresponding author(s). E-mail(s): johnaoga@gmail.com; juhee.bae@his.se;
Contributing authors: stefanija.veljanoska@univ-rennes1.fr; siegfried.nijssen@uclouvain.be; pierre.schaus@uclouvain.be;

Abstract

A growing attention in the empirical literature has been paid on the incidence of climate shocks and change on migration decisions. Previous literature leads to different results and uses a multitude of traditional empirical approaches. This paper proposes a tree-based Machine Learning (ML) approach to analyze the role of the weather shocks toward an individual's intention to migrate in the six agriculture-dependent-economy countries such as Burkina Faso, Ivory Coast, Mali, Mauritania, Niger, and Senegal. We performed several tree-based algorithms (e.g., XGB, Random Forest) using the train-validation-test workflow to build robust and noise-resistant approaches. Then we determine the important features showing in which direction they influence the migration intention. This ML based estimation accounts for features such as weather shocks captured by the Standardized Precipitation-Evapotranspiration Index (SPEI) for different timescales and various socioeconomic features/covariates. We find that (i) the weather features improve the prediction performance, although socioeconomic characteristics have more influence on migration intentions, (ii) a country-specific model is necessary, and (iii) the international move is influenced more by the longer timescales of SPEIs while general move (which includes internal move) by that of shorter timescales.

Keywords: Migration, Weather shocks, Machine learning, Tree-based algorithms

1 Introduction

The climate is changing and its implications for human mobility are at the core of the scientific and political agenda. The profound relationship between migration and the environment is not an unknown phenomenon, but the emergence and acceleration of climate change introduce more complexity to this relationship. The literature that brings together migration and climate change has increased significantly in the past ten years [5, 6, 13, 26]. This literature benefited primarily from greater availability and quality of climate and mobility indicators. Their main goal is to study the extent to which climate events initiated or even forced individuals to move. Although the objective of the research seems straightforward, the findings do not reach a consensus.

The heterogeneity of the results is due to the use of different measurements, different methodological approaches, and different contexts [5]. First, the findings differ in terms of (i) the direction of impact, whether climate acts as a pull or a push factor for migration[1], (ii) the strength of the relationship, and (iii) that this relationship is conditional on other features. Second, the different methodological approaches and ways of measuring climate shocks and migration could explain this divergence from the existent evidence. Third, the findings are context-specific. For example, existing evidence shows that the climate-migration nexus is common in developing societies, with the rain-fed agricultural sector that occupies a vital place in the overall economy.

The primary goal of this article is to bring new insights to this literature by adopting Machine Learning (ML) techniques and a multitude of climate and mobility measurements. Following Bertoli et al [7], we focus on the West African region, namely Burkina Faso, Ivory Coast, Mali, Mauritania, Niger, and Senegal. We rely on data from Gallup World Poll (GWP) surveys [21] and the high-resolution guided dataset from the Climatic Research Unit of the University of East Anglia [22] to construct the climate indicators.

[1] Black et al [8] distinguishes migration, displacement, and immobility. Beine and Jeusette [5] refers to the 'trapped population'.

Unlike traditional methods used by social scientists that specify the relationships between variables, machine learning algorithms are emerging technologies that can learn data without any explicit specification of relationships. Thus, one benefit is the smaller manual function operations compared to the methods used in econometric studies, reducing the possible bias introduced by the expertise of the modeler. Instead, ML models are based on the dataset and can uncover more complex relationships or patterns between more variables during the learning phase. Therefore, ML models are considered more as a "black-box" approach. This higher *capacity* comes with two major drawbacks. First, there is a risk of overfitting the model to the data. Therefore, the ML methodology generally splits the observation dataset into the so-called *training set* and *test set*. The *training set* is used to fit the model, and the *test set* to evaluate the performance of the learned models. The second disadvantage of ML techniques is that the models they generate are less understandable. Linear models can be easily understood by examining the weights of the coefficients for each variable, but some machine learning models may contain thousands or even millions of parameters combined with a complex mathematical or logical formula to make each decision. Although some approaches have been developed to help users interpret ML models, in general, the behavior of ML models is less interpretable than linear methods.

Among the large variety of machine learning approaches, in this work, we have chosen to use (ensemble) tree-based classifiers (decision tree (DT), random forest (RF), and eXtreme Gradient Boosting (XGB)) for the following reasons. The reproducibility of our result/approach is important, and those methods are available today in most of the off-the-shelf data science tools or library. For nearly two decades, these methods have been highly successful in tackling a variety of machine learning problems. They are still the go-to approach for machine learning competitions. In addition, they do not require a large amount of data to achieve good results and the cost of training is relatively low.

Our contributions are as follows.

- We approach the migration-climate nexus using tree-based methods and demonstrate through this paper the interest to use ML.
- We provide evidence on how climate influences migration intentions.
- We propose an ML workflow to the social science community on how to use machine learning techniques.

Section 2 introduces the problem and lists our questions based on our motivation to use machine learning methods to predict migration intentions. Next, Section 3 (and Appendix A) describes our methodology and provides an overview of the machine learning (ML) approaches (e.g., decision tree (DT), random forest (RF), and eXtreme Gradient Boosting (XGB)) we used in this paper. In addition, Section 4 describes the dataset in detail. Then, Section 5 describes our experiments and answers to the research questions, followed by Section 6 which elaborates on our findings and discussions. We finish our paper in Section 7.

2 Conceptual framework

2.1 Formalizing the climate-migration relation

The idea that the weather can affect economic outcomes such as economic growth, agricultural production, migration, among others, is not a recent idea [15]. Establishing such a relationship is a long-standing challenge. It is challenging to separate the effect of climate from other influences (nonclimatic) which are potentially correlated with it.

The climate-migration nexus is typically represented as identifying an unknown functional relationship, f:

$$y_{irkt} = f(C_{rkt}, X_{irkt}). \tag{1}$$

This function links climate variables (C) available for a set of regions r in a country k in a year t and other potentially explanatory variables (X) available at the individual level i to migration outcomes (y).

In this paper, y_{irkt} represents the *willingness to emigrate*, locally or abroad. This is a binary variable that is set to 1 if an individual intends to relocate or to 0 if not. . It is worth noting that this variable captures the possibility of migration or a plan to migrate, but not an official migration. Generally, migration intention is used as an indicator of a possible future migration since reliable data on actual migration is not available [33]. Tjaden et al [33] validates the usefulness of migration intentions data, especially when actual migration flow information is not available.

C_{rkt} represents the weather shocks based on the multiscalar drought index computed from the weather data called *SPEI (Standardized Precipitation-Evapotranspiration Index)* provided at the regional level [35]. SPEI is an improved drought index that accounts for atmospheric water conditions that are affected by temperature, wind, and humidity. SPEI normalizes anomalies in the accumulated climatic water balance (a difference between precipitation and potential evapotranspiration) to measure the severity of the drought. To compare between locations and climates, the log-logistic probability distribution is used for normalization, as suggested by Vicente-Serrano et al [35]. For more details on SPEI, we refer to Section 4.2.1. Other variables considered in the model represent demographic (e.g., age, sex) and socioeconomic variables (e.g., income) related to a sample of individuals living in a country.

A *probit* model is one of the traditional approaches used to estimate Equation (1). Its form is:

$$P(y_{irkt} = 1 \mid C_{rkt}, X_{irkt}) = \Phi(\alpha + \beta C_{rkt} + \gamma X_{irkt}) \tag{2}$$

where β and γ are the parameters that characterize the contribution and the role of the regional climate (C_{rkt}) and individual covariates (X_{irkt}) for individual migration intentions. $\Phi(\cdot)$ is the cumulative standard normal distribution function.

To understand the climate determinants of migration intentions, Bertoli et al [7] uses a *logit* approach to estimate the migration intention decision of an individual i based on the utility they would gain from different migration options or staying at their place of origin. This utility is similar to the one expressed in Equation (2). It is determined by the regional climate and the demographic and socioeconomic characteristics of the other individual, as well as time and regional dummies. The latter controls the possible seasonal effects in the stated intentions to migrate, the time-varying country-level determinants of these intentions, and the time-invariant spatial heterogeneity in the intentions to move. The study comprises two stages of analysis. In their first stage, Bertoli et al [7] performs more than 300,000 regressions to select the weather factors that influence migration intentions on several samples. The study uses these selected variables in the second stage to estimate their parameters and the direction of influence toward migration intentions.

2.2 Benefits of Decision trees over linear models

Finding the best linear model can become very complex as the number of input variables (covariates) grows. In [7], the authors reported more than 300,000 regressions performed resulting from some hand-crafted terms that are added to the model (Equation (2)) by multiplying and/or taking the logarithm of several variables. Additionally, comparing each regression model as done in [7] without a training/test set approach might cause a bias toward more complex models that may perform more poorly on unseen data. We carried out two comparative studies using the dataset described in [7].

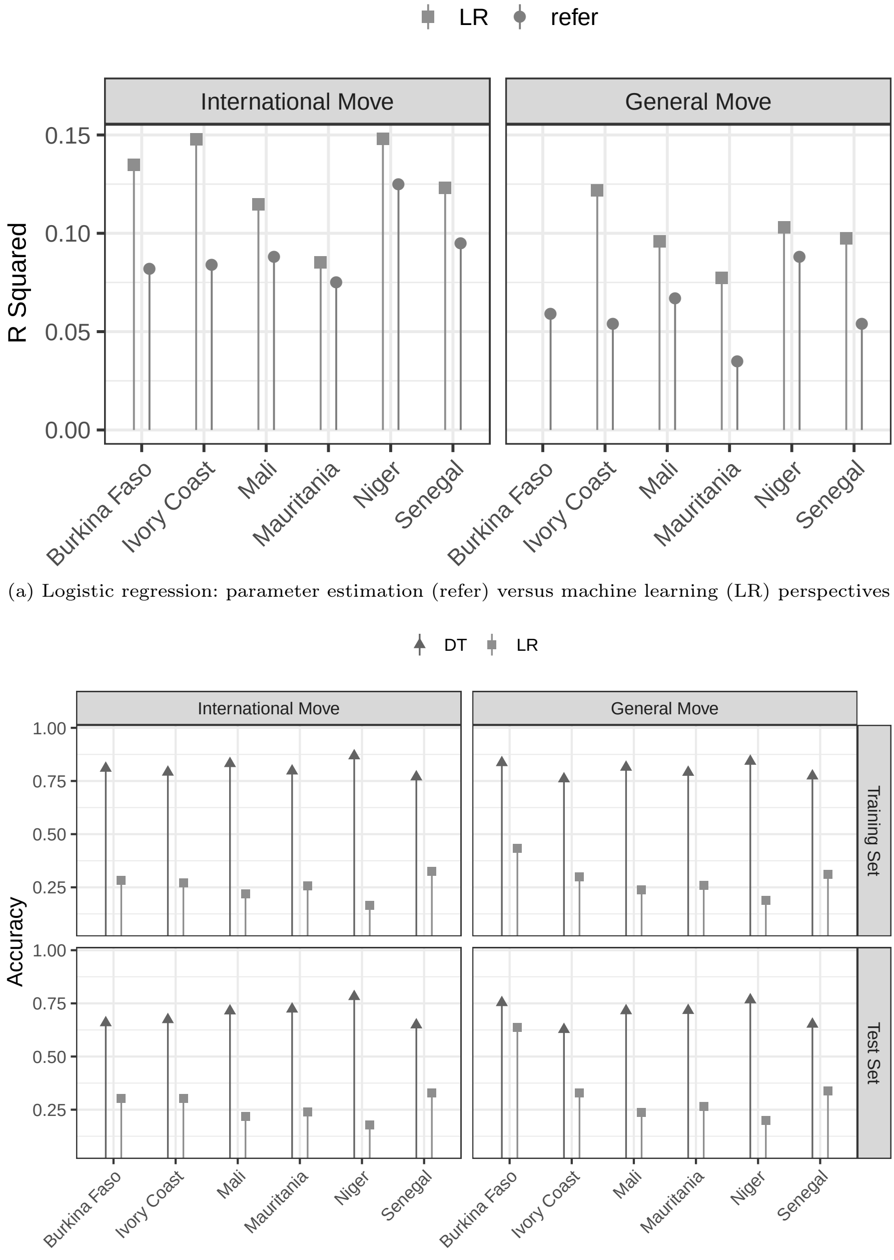
1. Comparing the predictive power between two different ways of running logistic regressions: *several regressions* (as described in Bertoli et al [7]) versus *single regression* (ML's way) to examine their differences[2].
2. Comparing the predictive power of a *linear* model (logistic regression) versus a *non-linear* model (decision trees) to examine the nature of the problem we have.

Figure 1(a) compares the *predictive power* of both regression models using R-squared measurement.[3] ML's logistic regression (LR) outperforms the regressions from Bertoli et al [7] (refer) in terms of the predictive power. Hence, machine learning (ML) yields higher predictive power and more interesting results.

However, it should be noted that a low R-squared measure does not imply that the model performs worse. Models with very low R-squared can fit the

[2] The major difference between the logistic regression from the ML approach and the regressions used in Bertoli et al [7] is that the logistic regression from the ML approach runs a *single regression*, including all features or covariates, while in [7], there are multiple runs of regressions (i.e., a run for each feature).

[3] R-squared can be computed using the *McFadden's* R^2 formula [25]. Bertoli et al [7] use R-squared measure implemented in *STATA* [31]: $1 - L_M/L_0$, where L_M is the log-likelihood of the model and L_0 is the log-likelihood of a *null-model*. A null-model is a model where we learn only from the target attribute with no predictor.



(a) Logistic regression: parameter estimation (refer) versus machine learning (LR) perspectives

(b) Logistic regression (LR) versus decision tree (DT) in machine learning

Fig. 1: Comparing models' predictive power for each country using: **(a)** R^2 measures of logistic estimation provided in Bertoli et al [7] (refer) and ML's logistic regression results (LR), and **(b)** Accuracy measures of ML's logistic regression results (LR) and decision trees (DT). In (a), the logistic regression model (LR) of ML for Burkina Faso (general move) is empty because it does not converge

data very well according to *goodness of fit* tests. This meets the goal of parameter estimation, whereas, in ML, the performance on unknown dataset/instances is more critical.

In a second step, the model is trained on a part of the dataset (i.e., the training set). The other part (i.e., hold-out sample or test set) is used to measure the model's predictive power of the unknown observations. There are a number of metrics to measure the predictive power of a model. Section 3.4 provides more details of the metrics.

The accuracy value is between 0 and 1, the same as the R-squared (the higher the values, the better). Figure 1(b) shows the accuracy of a logistic regression model (LR) and a decision tree (DT). DT outperforms LR on both the training and test sets. However, accuracy is not a reliable measure when class distributions show severe skewness [9]. However, it is common to face this imbalanced distribution with the real dataset (refer to the counts in Figure 4). Positive migration intentions (minority class) are more important to our analysis, but the accuracy shows a lesser impact on this minor representation, as depicted in Figure 1(b). This is why we show alternative metrics for an imbalanced classification problem in Sections 3.4 and 5. Overall, tree-based approaches better capture the nonlinearity of our problem.

Typically, the learning workflow includes (i) selecting a suitable ML method for the problem taking into account several criteria: the quality of the data (e.g., reducing noise), the linearity of the problem, and the interpretability of the outcomes. Subsequently, (ii) we optimize the configuration of the methods to improve the overall predictive power of the models. Finally, (iii) we interpret the outcomes of the models to gain insight into the problem.

In this paper, we use tree-based ML methods to address the issues (e.g., scalability, nonlinearity, and determining driving factors) that have been raised in prior research when exploring the connection between migration and climate. We closely follow the construction of the data of Bertoli et al [7] to investigate this connection. The distinction is that the current study involves actual SPEI values without any transformation over longer time frames and uses tree-based methods to predict the migration intention and capture the nonlinear relationships of the input variables.

2.3 Toward ML approaches

To understand the links between climate, individual characteristics, and migration intentions in a more flexible methodological manner, we propose using machine learning (ML) algorithms. It makes it possible to predict the migration intention with a larger dataset and find influencing features from an existing explainable method (Section 3.5).

We first find an ML approach that shows robust prediction performance (Q1). Furthermore, with a large dataset including the weather shock and individual characteristics from a survey on migration intentions of six different countries, we statistically compare the prediction performance to find the impact of weather (Q2, Q5), individual characteristics (Q3, Q4), and countries

(Q3, Q4). The weather shock involves various SPEI timescales and different lengths of the lags before the interview date (over 4 years from the interview date) (Q5).

- Q1: Which tree-based ML algorithm(s) performs better, that is, with a higher score?
- Q2: Does the weather (i.e., the drought) influence the moving intentions?
- Q3: Can we generalize a model for the six countries or need a country-specific model?
- Q4: Which features have an impact on moving intentions?
- Q5: Does the SPEI index or monthly lags of the weather influence the moving intentions, and if so, which SPEI(s) or lag(s) matter?

For the sake of simplicity, in the remainder of the paper, we will use X to denote the dataset used by the learning models without distinguishing between climate and control variables unless necessary, i.e., $y = f(X)$ instead of $f(C, X)$.

3 Methodological approach

This section focuses on the key concepts of the four methodological stages of our study: (i) data preparation, (ii) model implementation, (iii) model's performance evaluation, and (iv) the interpretation of the model outputs.[4]

3.1 Terminology

We first review the terminology used by social scientists and machine learners. In this section, we establish a link between the naming of concepts in social science and those used in ML (refer to Table 5 in the Appendix). In regression, the model is estimated, whereas, in ML, the model is trained [3]. The sample (in-sample) used to estimate the parameters of a model is called the *training set.* ML also uses a *test set* (or a hold-out sample) that is a distinct dataset separated from the training set. Through the learning process, these two types of samples make sure that the model is robust and noise-resistant.

The *R-squared* is a goodness-of-fit measure that is used in regression models, while *accuracy* is used in classification models. R-squared is a statistical measure that represents the proportion of variance in the dependent variable predictable from the independent variable, also known as the coefficient of determination. When R squared is 1, it shows that the regression prediction model perfectly fits the data. Accuracy is the fraction of predictions that the model measures correctly. It reflects the ability of a model to predict (or classify) classes of unknown vectors, since accuracy is generally measured on the dataset that was not used to optimize the model. Strictly, the metrics are not comparable one to one.

In ML, *features* (variables, columns) refer to regressors, predictors, or covariates. Each row is called an *example, instance, or observation.* Our

[4] For more detailed information, see Appendix A and Provost and Fawcett [28].

approach is a *supervised learning* approach since both the predictors/features X_i and the output y_i are observed. Another way to categorize a problem is regression and classification. When the output is numerical and continuous, it is called a *regression*. However, when the output is categorical or binary, it is a *classification* problem. Based on the research problem and the dataset we have, we solve a *supervised learning classification* problem. In the remainder of the paper, we mainly use ML-based terminologies. Appendix A describes more information.

3.2 Data preprocessing

It is critical to perform data preprocessing before running a model since it removes inconsistencies, missing data, and possible scale/type-related problems. Many ML algorithms only support numerical variables, often for the sake of implementation efficiency. Given a dataset with many categorical variables (e.g., survey questions with yes/no answers), we convert the categorical variables into numerical variables using the *one-hot encoding* method[5]. *Discretization* is also typically used to avoid an over-sensitivity of floating numbers, which we used for SPEI drought index values. Section 4 details how we build and prepare our dataset.

3.3 Model implementation: Tree-based approaches

In this paper, we focus specifically on tree-based methods because these methods are well suited to classification problems and automatically capture nonlinearity [27]. As a result, tree-based algorithms are increasingly being used in applied sciences [2, 3].

The decision tree method consists of approximating the learning function f using decision trees. Figure 2 is an example of a decision tree from our experiment, which is straightforward and highly interpretable. However, in practice, they can be inaccurate [23]. Therefore, several other tree-based methods have been proposed. Random Forest (RF) [10] and eXtreme Gradient Boosting (XGB) [14] are widely used methods. For both RF and XGB, the basic idea is to combine several decision trees to make a prediction. The predictions achieved with multiple trees can then be more accurate, generalizing the data appropriately.

However, obtaining a high-performance and accurate model is not trivial. It involves tuning the model parameters, for which we used the *Bayesian Hyperparameter Optimization* (BHO) [30] to select the appropriate tuning parameter. Appendix A.2 includes more information on the terminology.

3.4 Performance evaluation

In supervised learning, models are evaluated by making one-on-one comparisons between the predicted outcome ($\hat{y}$) and the real outcome (y). From this

[5] A dummy variable that represents categorical data.

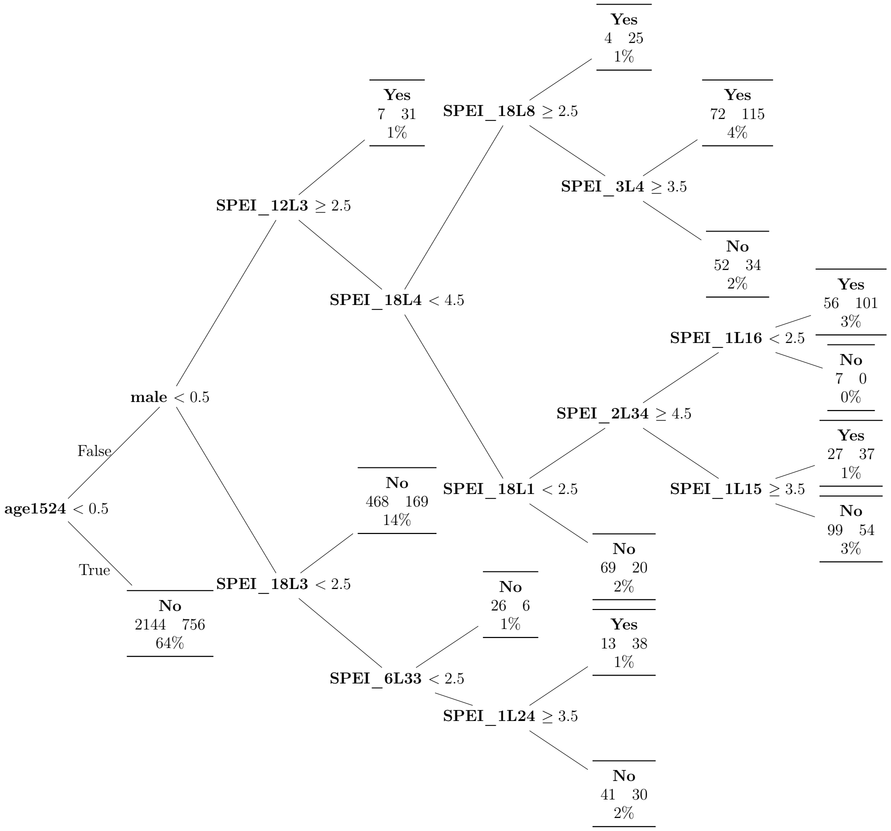


Fig. 2: A decision tree with the features involved in the international moving intention of Burkina Faso

comparison, in ML, several metrics are used to evaluate a model. This is a benefit of ML over parameter estimation, which typically relies on the assumptions from the data generation process to ensure consistency [27].

In this paper, we measure the accuracy, precision, recall, and the Area Under the ROC (Receiver operating Characteristics) curve (AUC) [19, 32]. *Accuracy* is a ratio of correctly predicted observations to total observations. It is an intuitive performance measure, but only when the dataset is symmetric with a balance between false positive and false negative. It measures the total number of predictions that a model makes and gets it correct. But it should be used carefully since, for example, if a model shows high accuracy in an environment where most people do not have a disease, the model has a high tendency to falsely predict someone who has a disease. This is why other metrics are considered simultaneously. The precision represents the ratio of correctly predicted positive observations to the total predicted positive observations. It evaluates how precise a model performs in predicting positive observations and is useful when there are many false positives (e.g., email spam). *Recall* is the ratio of correctly predicted positive observations to all actual true observations. It evaluates how many actual positives are correctly identified and is useful when there are many false negatives (e.g., fraud detection). However, having a high accuracy (or recall, precision) of a model does not necessarily mean that it is good. It is crucial to use an appropriate metric for different problems and to overview all the metrics. The AUC represents the overall performance of a model regardless of any classification threshold, for example, 0.5 to separate positive/Yes (> 0.5) and negative/No (≤ 0.5) predictions. These metrics have values between 0 and 1 (the higher, the better performance). Appendix A.3 and Figure 16 include more information on the terminologies.

3.5 Output interpretation: Feature importance and Partial Dependence Plots (PDP)

The features X used to estimate f in equation $y = f(X)$ are rarely equally relevant. Typically, only a small subset of features is relevant. Therefore, after training the model, the *Relative Feature Importance* (RFI) method is used to determine the most relevant ones. It consists of computing the contribution of each feature to the prediction [11].

RFI has become widespread and is thereby used for other ML methods. To understand in which direction these important features influence the outcome y, *Partial Dependency Plots* illustrates the impact [23, Chap. 14]. It is a *marginal average* of f that describes the effect of a chosen set of features S on f. The most convenient way to compute the partial dependency of a feature X_i, which contains k distinct values ($\{x_{i1}, x_{i2}, \cdots, x_{ik}\}$), is to compute the prediction when $X_i = x_{ij}$ with $j \in [1, k]$. Appendix A.4 includes more details.

4 Data preparation

In this section, we describe the data sources used in this study and its preprocessing. The dataset comprises individual survey data on migration intentions (Section 4.1, Figure 3) based on the Gallup World Poll (GWP) [21] and data on weather shocks based on SPEI (Standardized Precipitation Evapotranspiration Index) [35] of the six Western African countries between 2009 and 2015 (Section 4.2, Figure 5). The two datasets are joined by the months interviewed and the region identifiers[6] based on a finer geographical identifier that corresponds to the location of an interviewee (i.e., regional administrative units).

4.1 Gallup World Poll (GWP) data

We used GWP data to study the influences on the likelihood that people want to move or stay in their country of residence. GWP surveys have interviewed citizens in 160 countries since 2005, covering both urban and rural

[6] GADM: the Database of Global Administrative Areas

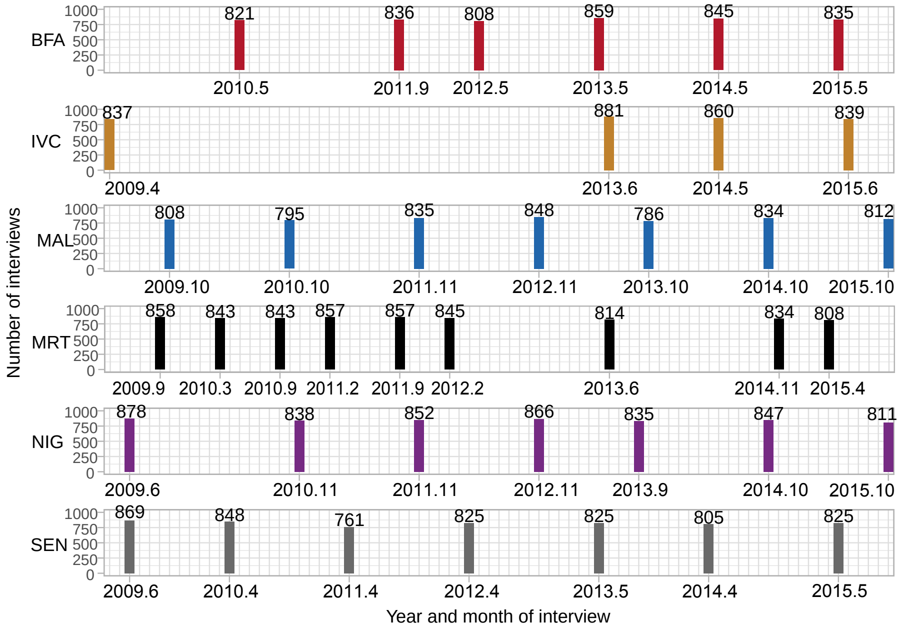


Fig. 3: GWP interview timeline and the number of interviews which are provided in Bertoli et al [7]. BFA: Burkina Faso, IVC: Ivory Coast, MAL: Mali, MRT: Mauritania, NIG: Niger, and SEN: Senegal

areas. These surveys measure the attitudes and behaviors of a random sample of approximately 1,000 individuals in each survey round. Our dataset includes migration responses and other characteristics of interviewees aged from 15 to 49 years from six western African countries (Burkina Faso, Ivory Coast, Mali, Mauritania, Niger, and Senegal) between 2009 and 2015 (Figures 3 and 4).

4.1.1 Migration intentions

Two questions have been identified from the GWP survey and cited by Bertoli et al [7], which are related to migration intentions.

- *Q1:* In the next 12 months, are you likely or unlikely to move away from the city or area where you live? (general move)
- *Q2:* Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country? (international move)

Q1, which we call the general move, involves migration that includes internal and international moving intentions with a decision period of 12 months. Contrary to *Q1*, *Q2* only involves international migration intentions, excluding a time frame. It should be noted that these questions capture the willingness to emigrate, and one should expect that not all potential migration would realize a move.

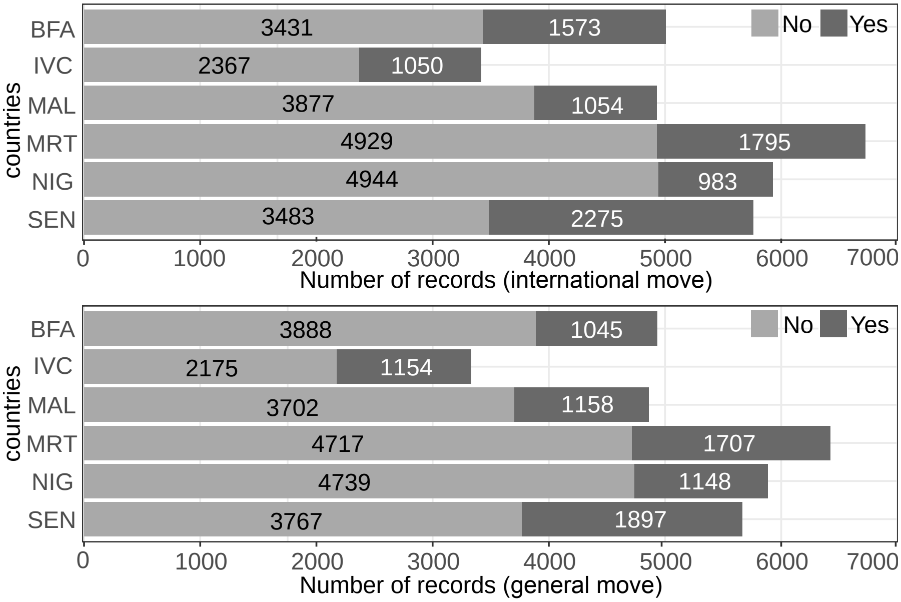


Fig. 4: The number of records in the entire dataset of six countries for international and general move intentions. Refer to Figure 3 for country codes

From these two questions arise the two target variables that we aim to explain in our study: general move in *Q1* and international move in *Q2*. Figure 4 shows the number of records with positive and negative responses for each country towards the two types of moving intentions.

4.1.2 Individual characteristics

Following the empirical approach of Bertoli et al [7], Table 1 summarizes the control variables used in the ML approach, such as country of origin, age, the gender of an individual, and when the interview took place (e.g., month, year). Furthermore, 'urban' attribute shows whether a person lives in an urban or rural area; 'hskill' attribute includes if one is highly educated (i.e., has completed four years of education beyond high school and/or received a 4-year college degree or not); 'hhsize' attribute accounts for the number of household members who are older than five years; 'mabr' attribute includes whether one has family members or relatives living abroad and who can provide assistance if needed. We include 'lnhhincpc' attribute, which is the natural logarithm of self-reported household income per capita in dollars. This attribute is not included in Bertoli et al [7] due to the side effects it might cause in their identification strategy. One of the side effects is that it reduces the sample size, especially since this income question was not asked in all countries. Furthermore, we cannot overlook the bias introduced because it is a self-reported measurement and there may be a potential correlation between income and weather shocks [12]. This is related to one of the limitations of the traditional empirical approaches discussed in Section 2. We include the income variable in our study because it offers an alternative to the explicit selection of variables that is done through machine learning techniques. We do not make assumptions, but we build the model to select the important variables while being noise-tolerant.

The preprocessing of the data consists mainly of either binarizing or one-hot encoding certain variables. The binary variables 'gender', 'mabr', 'hskill', and 'urban' are not involved in this operation. The categorical variables 'origin' and 'year' [7] are preprocessed by one-hot encoding. The numerical variables 'age', 'hhsize', and 'lnhhincpc' are binarized. The age variable is binarized to 15-24 (age1524), 25-34 (age2534), and 35-49 (age35plus) [7].

We binarized the variables 'hhsize' and 'lnhhincpc' into four classes based on a process that tests several subdivisions of the continuous values. It measures the correlation of each class with the dependent variables [17] [8]. The more correlated subdivisions of continuous values are grouped together. For example, variables 'hhsize' 3 and 4 (i.e., interviewees who had three and four residents in a house) are grouped as one class, 'hhsize 3-4', since they show a high correlation with the dependent variables.

The final GWP dataset, illustrated in Table 1, consists of six countries variables (origin), seven-year variables (the interview held between 2009 and 2015),[9] four variables of household size, four variables of self-reported household income per capita, three variables of age, gender, living environment (urban or rural), connections abroad ('mabr'), and the individual's education level ('hskill') variables.

4.2 Weather shocks data

The results of Bertoli et al [7] show that an identified period of shocks, the intensity of the shocks, and the treatment of the (local) crop-growing or crop-planting seasons have impacts on the migration intentions (general and international). This section describes the weather shock information we used.

4.2.1 SPEI

The economic activity of the region that we focus on is highly dependent on the agricultural sector. In the absence of irrigation infrastructure, weather, and in particular, water availability directly influences agricultural production. In such a context, livelihoods are indirectly affected by the climatic condition.

[7] There are several ways to configure the year variable: (i) use the integer value for each year, (ii) subtract each year by the minimum year to have relatively smaller numbers starting with 0, and (iii) treat integer as a categorical variable and perform one-hot encoding. Here, we use the last approach.

[8] We used the R package *correlationfunnel* which is fast and offers visualizations to facilitate this work.

[9] The interviews are conducted in different months for different countries and the month of interview may be different for each year (Figure 3).

Table 1 The number of samples of the binarized and discretized dataset with one-hot encoding for the international and the general move

Type		Feature	1-hot encoding	international	general
GWP (X)	GWP dataset only	origin	Burkina Faso	5,004 (16%)	4,933 (16%)
			Ivory Coast	3,417 (10%)	3,329 (10%)
			Mali	4,931 (16%)	4,860 (16%)
			Mauritania	6,724 (21%)	6,424 (21%)
			Niger	5,927 (19%)	5,887 (19%)
			Senegal	5,758 (18%)	5,664 (18%)
		year	2009	4,143 (13%)	4,037 (13%)
			2010	3,975 (12%)	3,944 (13%)
			2011	4,834 (15%)	4,748 (15%)
			2012	4,019 (13%)	3,951 (13%)
			2013	4,835 (15%)	4,765 (15%)
			2014	5,025 (16%)	4,925 (16%)
			2015	4,930 (16%)	4,727 (15%)
		hhsize	inf.-3	10,905 (34%)	10,663 (34%)
			3-4	5,767 (18%)	5,637 (18%)
			4-6	8,933 (28%)	8,745 (28%)
			6-inf.	6,156 (19%)	6,052 (19%)
		lnhhincpc	inf.-5.605	7,961 (25%)	7,834 (25%)
			5.605-6.446	7,903 (25%)	7,732 (25%)
			6.446-7.231	7,935 (25%)	7,780 (25%)
			7.231-inf.	7,962 (25%)	7,751 (25%)
		age	age1524	11,493 (36%)	11,239 (36%)
			age2534	10,686 (34%)	10,462 (34%)
			age35plus	9,582 (30%)	9,396 (30%)
		gender	male	16,937 (53%)	16,593 (53%)
			female	14,824 (47%)	14,504 (47%)
		urban	urban	7,491 (24%)	7,295 (23%)
			rural	24,270 (76%)	23,802 (77%)
		mabr	yes	14,879 (47%)	14,573 (47%)
			no	16,882 (53%)	16,524 (53%)
		hskill	yes	886 (3%)	862 (3%)
			no	30,875 (97%)	30,235 (97%)
SPEI (C)	SPEI dataset	SPEI timescales	1,2,3,6,12,18,24		
		lags	lag0 - 48		
ALL		GWP + SPEI			

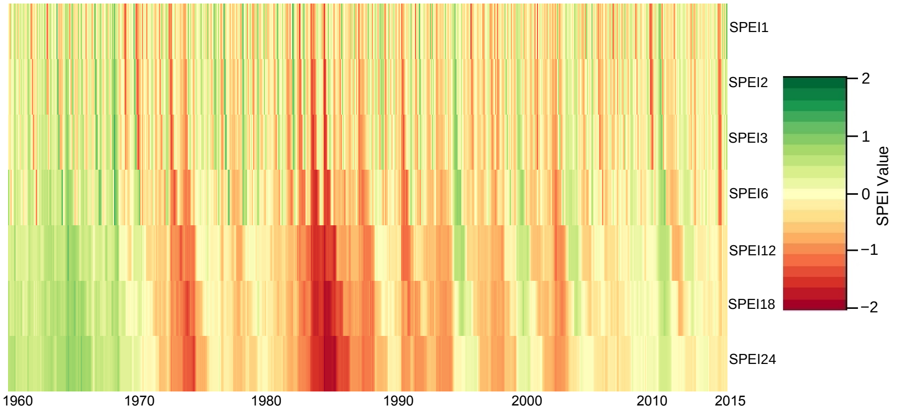


Fig. 5: SPEI timescales for 55 years in six western African countries

One strategy to deal with chronic weather variability, especially when other economic opportunities are limited, is to move.

One of the statistical challenges in the literature studying the impact of climate variability on various economic outcomes, including migration, is how to measure it so that it is comparable over time, space, and several relevant climatic factors. At earlier stages, the literature has focused on precipitation. However, the impact of climate on agricultural yields depends on factors such as temperature and the ability of soil to retain water. Moreover, the emerging global warming issue emphasizes the importance of capturing the impact of temperature. These are assembled by potential evapotranspiration (PET), which in turn depends on temperature, latitude, sunshine exposure, and wind speed.

SPEI is calculated by fitting the time series of differences in precipitation and PET (i.e., climatic water balance) to a probability distribution. This process enables the differences to be expressed as standard normal scores with zero mean and unit variance. It is standardized using a *Log-logistic distribution function* that is found to be the most suitable distribution for SPEI [34]. These standardized units are comparable on different SPEI timescales. The index value below 0 is characterized as drought. These calculated monthly SPEI values are collected at different time scales for each subregion in the six western African countries. Figure 5 illustrates these SPEI values averaged over the six countries between 1960 and 2015 to identify the dry and wet conditions in this area. It explicitly shows moderate conditions before 1970 in green, while increasing drought shocks trend with SPEI at 24 timescales in the early 1970s and between 1980s and 1990 in red.

SPEI outperforms other indices in predicting agricultural yields [34], especially as an index incorporating the effect of temperature, which can assess the effects of emerging global warming. In fact, SPI (Standard Precipitation Index) and SPEI are similar in the way of calculating the index, but SPEI overcomes the limitation of SPI by comparing the water and the evapotranspiration in the atmosphere. Unfortunately, SPI considers only precipitation. Both SPI and SPEI are multiscaled indices [10] that can identify the multitemporal nature of droughts. Another advantage of the SPEI is that it considers the onset, length, and intensity of a climatic event, rather than only the absolute value of precipitation and temperature (Figure 11). Moreover, it is comparable over time and space thanks to its fixed mean and standard deviation.[11]

4.2.2 SPEI and Lags

The SPI, and therefore the SPEI, was originally designed to quantify the precipitation deficit for multiple timescales. These timescales reflect the incidence of drought on the availability of water sources. The climatology community has defined three main types of drought: (i) meteorological drought, (ii) hydrological drought, and (iii) agricultural drought, which differ in intensity, duration, and spatial coverage [18]. The meteorological drought captures the extent to which soil moisture conditions react to precipitation anomalies in the short run, whereas surface and groundwater reservoirs are subject to the longer-term precipitation anomalies as captured by hydrological droughts. Agricultural drought occurs when crop production is affected by precipitation anomalies. To some extent, the meteorological drought is the mildest scenario, whereas the hydrological drought is the most severe scenario of drought occurrence, with the agricultural drought being in between. In this sense, a 1 or 2 months SPEI measure can show the presence and level of meteorological drought; from 1 month to 6 months SPEI for agricultural drought, and from 6 months to 24 months SPEI or more for hydrological droughts [36].

Our key challenges are to understand which measure of climate conditions matters for migration and which time-spans (or timescales) need to be considered. We built the weather shocks dataset with seven SPEI timescales (i.e., 1, 2, 3, 6, 12, 18, 24). We collect the SPEI values based on each subregion and an interview month. We gather the lags of each SPEI timescale of the past four years from the interview month, making a total of 49 lags (lag0–lag48) [12]. To understand how the SPEI drought index affects migration intention (positive or negative), continuous SPEI values are discretized into *seven equal bins*. For a feature X_i, the binning step is therefore (3):

$$\text{bin step}(X_i) = \frac{max(X_i) - min(X_i)}{7} \tag{3}$$

where i is the SPEI timescale. This discretization allows limiting the sensitivity of our models to the high variability of SPEI values. Figure 13 in the Appendix provides an example of discretization.

[10] SPEI at 3 months timescale for May 2015 is a function of the sum of the climatic water balance of March, April, and May 2015.

[11] By construction, SPEI has a zero mean and a standard deviation of unity.

[12] To get a SPEI at 12 months timescale with lag 6 for an individual interviewed in May 2015, the SPEI value is the SPEI12 value 6 months ago in November 2014.

5 Results

In this section, we report the results by applying machine learning algorithms to the constructed dataset combining the migration intention data (GWP) and the weather shock data (SPEI) (see Section 4).

5.1 Protocols

We ran all experiments on a computing environment with Intel Core i5 64-bit processor (2.7GHz) and 16GB of RAM running MacOS. We ran tests on the entire dataset (ALL) and then on the feature groups described in Table 1. To reduce the risk of overfitting, we used the 10-fold cross-validation workflow. For each of these datasets and each dependent variable (general or international migration intention), we conducted experiments separately using the R implementation of the tree-based ML algorithms: DT: Decision trees [29], RF: Random Forest [10], and XGB: eXtreme Gradient Boosting [14]. Table 2 shows the optimal parameters (Section 3.3) used for these algorithms. Our findings follow in the order of questions asked in Section 2.3.

Table 2 Optimal parameters for each algorithm

DT	RF	XGB
cost complexity = 10^{-5} tree depth = 30 number of nodes = 20	mtry[a] = 5 number of trees = 1080	mtry[a] = 3 number of trees = 761

[a] *mtry* represents the number of possible splits at each node.

5.2 *Q1.* Performance comparison of tree-based ML algorithms

We compare the results of tree-based methods over the test set with DT, RF, and XGB in Figure 6 based on AUC (see Section 3.4). The AUC measure is used to compare the models since it characterizes the overall performance of the classifiers. We find that XGB outperforms other algorithms for both dependent variables. Thus, the results described afterward are all from the XGB algorithm.

5.3 *Q2.* Influence of weather towards moving intentions

To answer the impact of weather towards migration intentions, we conducted separate experiments on the individual survey dataset (i.e., GWP) and the entire dataset with individual survey and weather dataset of the six countries (i.e., ALL = GWP + Weather) using the XGB algorithm. The *t-test* comparisons of precision and AUC measures on the test set reveal that the

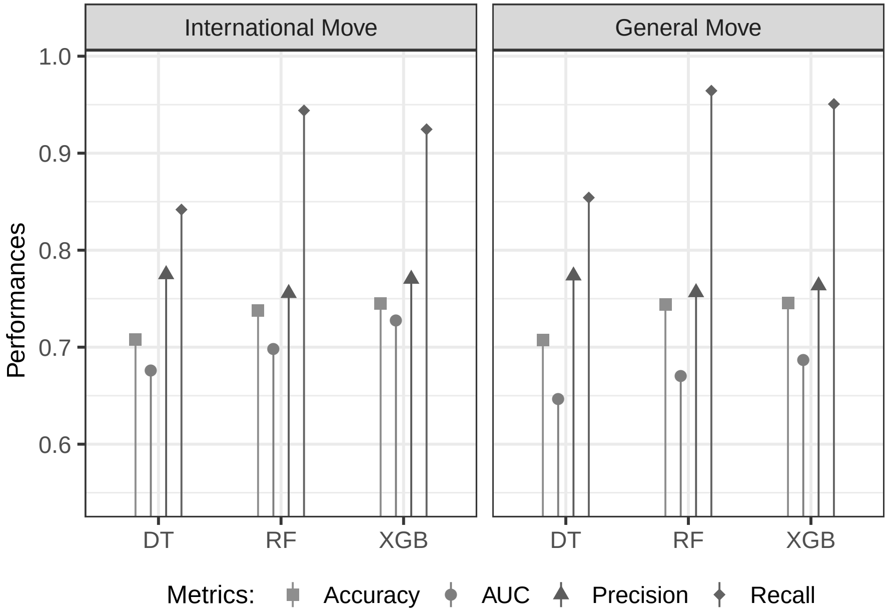


Fig. 6: Performance of tree-based algorithms of the test set in the entire dataset (i.e., GWP and Weather data) for all six countries

Table 3 The t-test of GWP compared to ALL (GWP + weather) using the XGB algorithm on test sets

Metric	move	mean.ALL	mean.GWP	t.value	p.value	comparison
Precision	International	0.77	0.76(0.7567)	4.00	0.00	ALL > GWP
	General	0.76	0.75(0.7462)	4.76	0.00	ALL > GWP
AUC	International	0.73	0.71	5.51	0.00	ALL > GWP
	General	0.69	0.66	7.70	0.00	ALL > GWP
Accuracy	International	0.75	0.74(0.7416)	1.95	0.08	ALL ˜ GWP
	General	0.75	0.74(0.7384)	1.93	0.08	ALL ˜ GWP
Recall	International	0.92	0.95(0.9487)	-10.29	0.00	ALL < GWP
	General	0.95	0.98(0.9793)	-14.25	0.00	ALL < GWP

classifier on the entire dataset (i.e., ALL includes the weather information) significantly outperforms the GWP's classifier for both international and general move dependent variables (Table 3). Thus, including weather data improves the prediction power for precision and AUC.

The t-test comparison in terms of accuracy shows that the models are not different between the individual survey (i.e., GWP) and the entire datasets (i.e., ALL = GWP + Weather). For recall, the GWP's model outperforms the ALL's model, although in general, ALL's model outperforms GWP's model

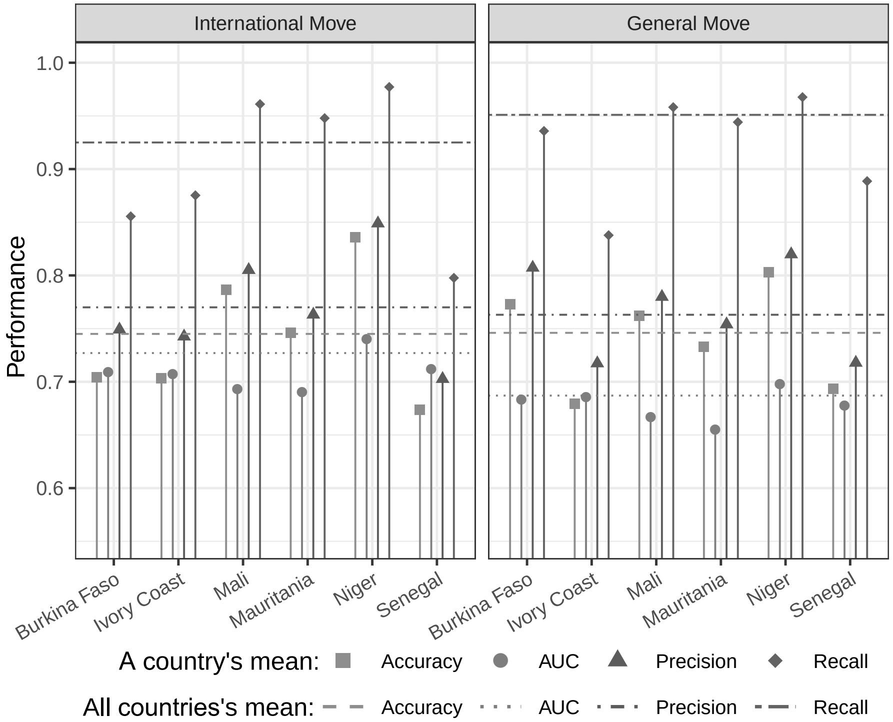


Fig. 7: Comparison between the average performance of six-countries' model (dotted lines) and performances of each country model (shape) on the test set

(which is noisier). This would be due to the presence of several countries that do not have the same weather conditions. We investigate this in the following section.

5.4 *Q3.* A general model or a country-specific model

To decide whether we can achieve a general prediction model for all six countries, we further investigate the prediction performances of six countries (general model) and those of each country (country-specific model) on international and general migration intentions.

Figure 7 shows and compares multiple measures using XGB of each country compared to the mean value with all-countries for the GWP + weather dataset (ALL). The horizontal dashed line for each metric (accuracy, AUC, precision, and recall) represents the average performance of the general model trained with all-countries dataset. Each shape-defined vertical line represents the country-specific performance for each metric. For example, considering precision ▲, country-specific models for Mali and Niger outperform the general model using the entire dataset for the prediction of international moves and the same with Burkina Faso, Mali, and Niger for the prediction of general moves.

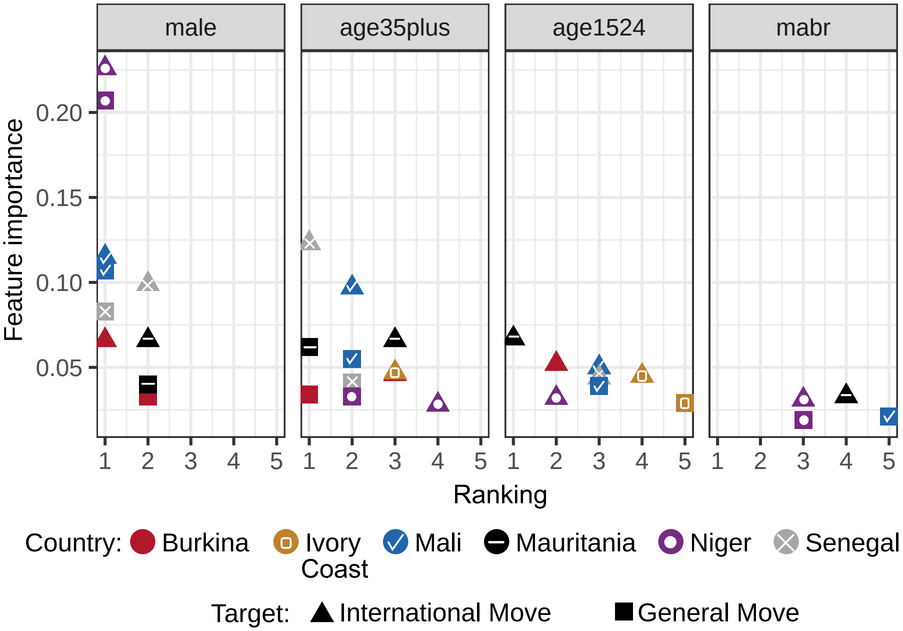


Fig. 8: GWP features with higher feature importance on each target

Hence, generalizing moving intentions with one model can mislead, since some countries do not fit into a general model. Moreover, the performance can be different for different targets in a country. For example, the precision performance of the general move intention of Burkina Faso's one-country model is higher than the all-countries model; however, it shows lower performance with the international move intention.

This guides us to what is critical to investigate in the model for each country when analyzing the relationship between climate and migration.

5.5 *Q4.* Important features

For five out of six countries, we find that gender and age show a stronger influence on both the international and general moving intentions.[13] In Figure 8, gender feature ('male') is considered as the first important feature for three countries, respectively on the international and general moves intentions. Overall, men have higher intentions to move than women, while people between 35 and 49 are likely to stay at their current residence. Although we cannot infer the exact reasons, we assume that men have higher spatial mobility to search for a job and a better economic status. In addition, older adults over 35 years are less inclined to move, as stated in previous studies [16].

[13] We find that the results are similar with permutation feature importance. Refer to Figure 19 in the Appendix.

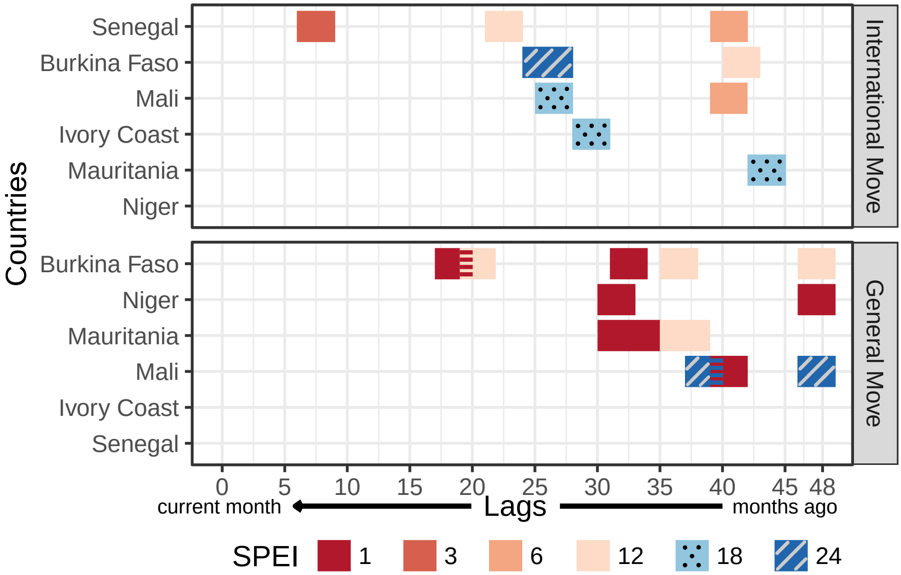


Fig. 9: Important SPEI timescales and periods of lags. The order of countries is based on the emergence of important lag periods

In addition, the younger group, aged between 15 and 24, is more likely to engage in international migration. Furthermore, having a one-distance connection abroad ('mabr') shows a positive impact on migration decisions. SPEI features are the most important features in Ivory Coast, followed by age characteristics. However, these findings cannot be generalized due to the limited number of surveys performed (refer to Table 1). Interestingly, Niger and Mali are countries that show higher accuracy, precision, and recall measures in Q3, both of which have feature importance values higher than 0.1 on both international and general move intentions (Figure 8). The next section shows the influence of weather features with slightly lower feature importance values. The countries for which the weather features show a higher value than the average importance values for each country are Ivory Coast and Mauritania. This finding is relevant for both types of movement, international and general moves. With Senegal, weather features show a higher effect than the average importance for the international move intention.

5.6 *Q5.* Influence of SPEI and monthly lags towards moving intentions

Figure 9 shows the most prominent SPEI timescales and lag periods, for at least three consecutive months, based on the feature importance values greater than the average. Each row (y-axis) represents a country, and each column (x-axis) shows the lags. The reddish periods represent SPEIs comprising shorter timescales (i.e., 1, 2, 3, 6, 12) while bluish representation for longer ones (i.e., 18, 24)[14]. We find that more reddish plots are more visible in the general move intention and more bluish ones for the international moving intention. This means international move has more influence from the longer timescales of SPEI as it may involve a more extended period of time to make permanent decisions[15], while general move which includes internal move, is more influenced by shorter timescales of SPEIs. Especially, Burkina Faso and Mauritania show shorter SPEIs affecting the general moving intentions and Mauritania's international migration intentions that are affected by longer SPEIs. With lags, besides some periods in Senegal and Burkina Faso, we find that lags over 24 months are more likely to affect migration intentions. The interpretation of such a result would be that potential migrants tend to move internally as a result of meteorological/agricultural droughts occurring, whereas international migration is a response to more severe, hydrological droughts. This indicates that the severity of the climate conditions determines the degree to which migration plans would be drastic. As severe weather conditions are highly correlated over time and space, migration is from a longer-distance type (e.g., international migration) in order to be protected from these conditions.

It is challenging to draw global patterns since the results are country- and migration-type- specific, thus it is crucial that further research considers the heterogeneity of such a country and migration type. As mentioned in the previous section regarding the important features, demographic characteristics are the essential drivers for both international and general migration intentions (again considering heterogeneity among countries and migration type). SPEI drought index is important to a lesser extent but seems to explain a part of the migration probability; however, the explanations are different across countries and migration decisions. In Appendix B, Figure 20 shows the feature importance in size for each SPEI and the lag combination for each country. It includes more details of the distribution of each SPEI index and each lag and the overall results of SPEI timescales and lag combinations.

To understand the mechanisms behind the results included in Figure 9, we further examine (i) the way climate shock events are captured by the different SPEI indicators influencing migration and (ii) why these lags appear to be important. It is important to note that the results differ among the different countries for the different international and general migration intentions. For example, our findings are explainable for Burkina Faso where the impact of important SPEI is V-shaped (Figure 10), compared to other countries. This indicates that highly negative (e.g., severe drought) and positive values (e.g., severe flooding) of SPEI increase the probability of moving both internationally and globally, whereas values closer to zero reduce it. We observe this V-shaped impact mainly for the months that fall in the cropping season (Figure 11).[16] The cropping season in Burkina Faso that concerns its main

[14] Longer timescales ($\geq$18 months) referred to https://climatedataguide.ucar.edu/climate-data/standardized-precipitation-evapotranspiration-index-spei.

[15] The international move's question (*Q2*) actually asks people if they want to move permanently to another country.

[16] The economic activity in the countries we consider in this article highly depends on the agricultural sector. Knowing that the irrigation infrastructure is lacking and that these sectors are

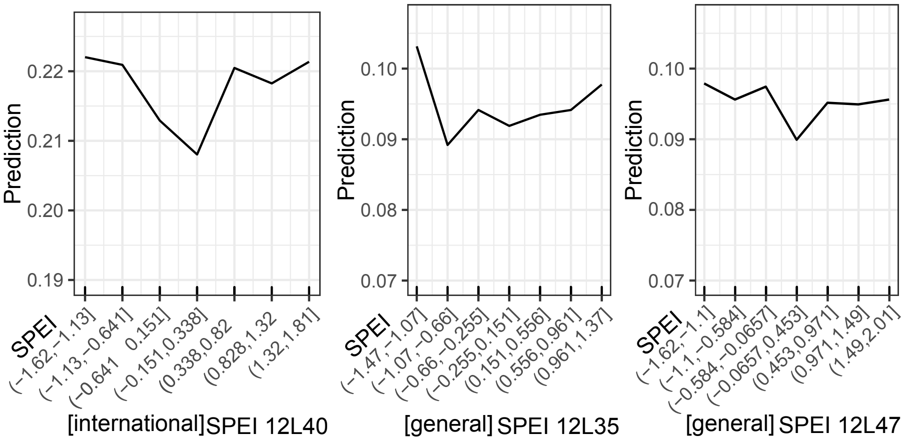


Fig. 10: Partial dependence plot (PDP) of selected SPEI timescale and lag combinations with V-shape in Burkina Faso. Severe drought and flooding are indicators of international and general migration intentions

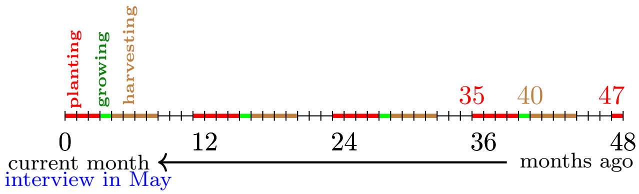


Fig. 11: Cropping seasons of sorghum, maize, and millet in Burkina Faso, including the planting growing and harvesting periods over the four years preceding the GWP interview mainly held in May (Figure 3). Months 35, 40, and 47 show the significant lags presented in Figure 10

crops, sorghum, maize, and millet, begins in April and ends with the harvest period in December. The lack or surplus of rainfall in the Aprils of three and four years, corresponding to lags 35 and 47, before the interview increases the intentions of moving generally and internationally. It is natural to expect that individuals do not immediately react to current events, but rather consider past weather events that occur in periods significant for economic activity.

6 Discussion

Data quality review. Any data analysis process depends on the quality of the data itself. In this study, we have combined several data sources to build

mainly rainfed, the weather conditions contribute greatly to agricultural production and income generation.

our own dataset. From the GWP data, there is a bias introduced by the inaccuracy of the questionnaire, which cannot be overlooked. From the weather data, various kinds of indicators can measure shocks. In this study, we used SPEI values that represent an aggregation of several indicators. However, one can also benefit from using raw indicators such as temperature and precipitation. These indicators can bring a more disaggregated viewpoint to the study. Finally, the equal-bin discretization has the advantage of restricting the sensitivity of our models with varying SPEI values. However, this has the disadvantage of giving the same importance to different ranges of shocks.

Methodology review. Concerning the methodology used, the use of machine learning (ML) brings a new perspective in applied science dealing with several issues such as large datasets, nonlinearity, and multicollinearity. A clear distinction is necessary between the prediction results obtained by these approaches and the causal inference allowed in parameter estimation. To circumvent the limitation of ML approaches to provide causal inference, we have used feature importance and partial dependence metrics to interpret the results obtained. However, more effort is needed in ML to build alternative methods integrating causal inference, as advocated by several authors in applied science [1, 2, 3, 4]. In the plethora of ML methods, in this study, we have focused on tree-based methods because they fit our study. However, future work may investigate other methods such as neural networks and unsupervised learning approaches to improve performance.

Results review. The results show that the weather feature adds more prediction power than only using the GWP dataset based on the XGB algorithm with higher AUC measures. In general, a longer time horizon of SPEIs (e.g., 18, 24 months) drives more international moving intention while shorter timescale SPEIs (e.g., 1, 12 months) affected the internal or general move. It is reasonable since it is likely that international migration decisions take a longer time to collect and assess information over some time. As the social science literature illustrates, we also find that it is vital to investigate based on each country, since there are unique features with different intensity affecting the migration intention. Moreover, men have a higher tendency to move, whereas the age group of 35-49 shows a higher tendency to stay [5]. The younger group (age 15-24) shows a higher tendency to move internationally, as found in the literature [24].

Unfortunately, there were a few consecutive periods of longer than 5 months with the SPEI's feature importance values above the average. Most of the periods found were 3 consecutive months followed by two occurrences of 4 consecutive months and one occurrence of 5 consecutive months. Overall, it is difficult to draw any global patterns that conform to the inconclusive findings from the previous literature.

7 Conclusion

In summary, weather features influence the prediction performance on migration intentions. We examine three tree-based machine learning algorithms and derive results with the better performing XGB algorithm using the GWP dataset and the weather shock dataset based on various SPEI timescales and lags. The weather indicators show a positive impact on prediction performance that is significantly higher than without the weather shock dataset. Furthermore, the longer timescales of SPEI (e.g., 18, 24 months) drive more international migration intention, while shorter timescales of SPEI (e.g., 1, 12 months) affected the general move intention. Yet, it is not easy to generalize global weather patterns for the six countries we investigated. Moreover, country-specific models are necessary due to distinct features with different effects on migration intentions. Overall, among the individual characteristics, gender, age, and networks abroad are revealed as important features. For example, *male* shows higher intention of migration, while the 15-24 age group shows higher intention of international migration.

There can be further improvements with a different way of preprocessing the features, involving more raw indicators besides the SPEI values, using neural networks, and integrating causal inference with ML methods.

Acknowledgments. This research is supported by the ARC Convention on “New approaches to understanding and modeling global migration trends” (convention 18/23-091). This project has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 754412. We thank Professor Frédéric Docquier for the reviews, suggestions, and earlier discussions of the project. The authors thank Dr. Shari De Baets for her feedback and comments. Open access funding provided by University of Skövde.

Declarations

On behalf of all authors, the corresponding author states that there is no conflict of interest. Gallup dataset is a paid dataset that we cannot make available publicly based on its copyrights.

Appendix A Machine learning approaches

A.1 Data preprocessing

We use the sample dataset in Table 4 as an illustrative example. This dataset has four features: age, household size (‘hhsize’), having human network abroad (‘mabr’), and the intensity of the drought (‘drought’); and one target attribute representing the migration intention (‘move’).

The first step is data preprocessing. It allows cleaning up the data by handling missing data and scale/type-related problems. A scale-related issue occurs when variables are displayed in different scales, for example, year (e.g.,

Table 4 A sample dataset with individual characteristics, drought index, and migration intention. 'hhsize': household size. 'mabr': human network abroad. (Note: The table is for an explanation purpose, not the dataset we used.)

instance	age	hhsize	mabr	drought	move
1	young	large	yes	harsh	Yes
2	young	large	no	harsh	Yes
3	middle	large	yes	harsh	Yes
4	old	medium	yes	harsh	No
5	old	small	yes	soft	No
6	old	small	no	soft	Yes
7	middle	small	no	soft	No
8	young	medium	yes	harsh	Yes
9	young	small	yes	soft	No
10	old	medium	yes	soft	No
11	young	medium	no	soft	No
12	middle	medium	no	harsh	No
13	middle	large	yes	soft	No
14	old	medium	no	harsh	Yes

$[2010, 2016]$) and age (e.g., $[0, 100]$). This problem can cause bias on ML models' output and implementation inefficiency.

There are two types of variables, numerical and categorical variables, that may need preprocessing. The categorical variables contain labels instead of numerical values. Many ML algorithms only support numerical variables, often for the sake of implementation efficiency. Hence, it is recommended to convert these variables into numerical variables using *one-hot encoding*.

Definition (One-hot encoding). *It consists of creating new binary variables for the unique labels in the categorical variable.*

It is well known that it produces bias to the model output when using the numerical variable inputs with different scales. We overcome this problem by *binarizing* these numerical variables.

Definition (Data binarization). *It comprises transforming a numerical variable into several binary variables. The binarization workflow is in two steps: (i) split the numerical variable into intervals and create a categorical variable by labeling each range. Then, (ii) use the one-hot encoding method to create the binary variables.*

Example. *Figures 12 and 13 show examples of one-hot encoding and binarization for the age and SPEI12 variables.*

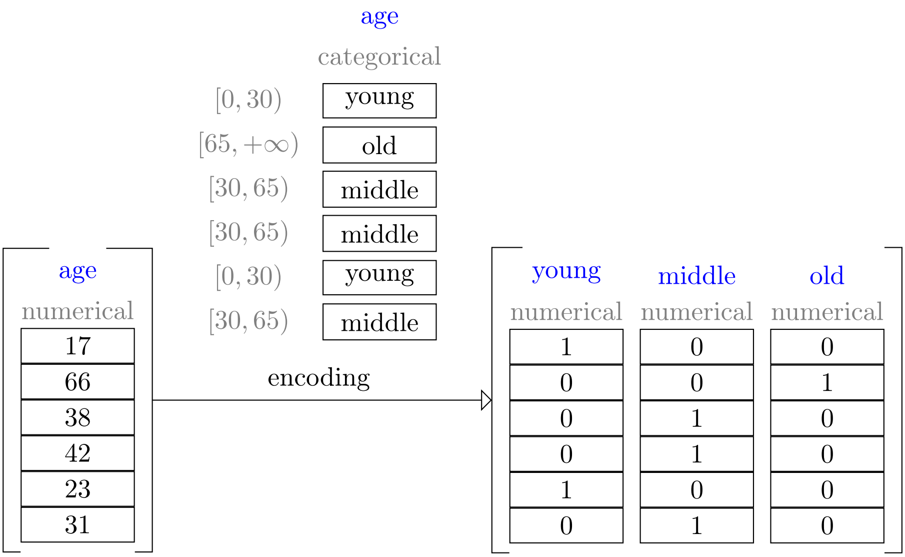


Fig. 12: Example of one-hot encoding and binarization of the variable age

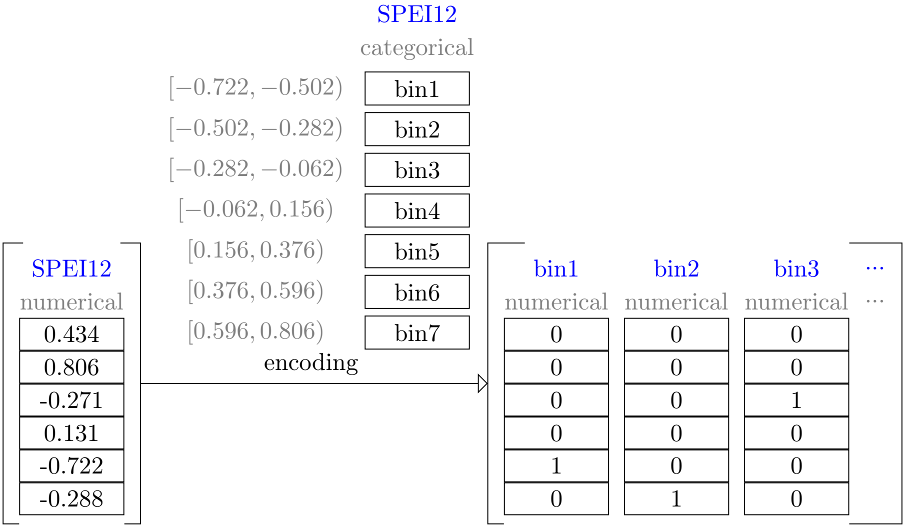


Fig. 13: Example of one-hot-encoding and binarization of the 12 months timescale of SPEI variable

Generalization is an essential concept in ML. It refers to the ability of a method to classify unknown examples to the model correctly. For this, the dataset is split for training and testing the model in the data preprocessing step.

Definition (Training set and test set)**.** *The training set is a part of the dataset used to train the model and the test set is the hold-out part of the dataset to test the model. Typically, 60 to 90% of the database is assigned as a training set while the rest as a test set.*

To have a noise-free and robust model that generalizes well, the training and test sets are extracted iteratively from the dataset. This resampling procedure is called the *cross-validation* process.

Definition (Cross-Validation)**.** *The cross-validation process consists of randomly splitting the dataset into K fairly equal samples $S_1, S_2, \cdots, S_K$. Based on these samples, K folds are created, each containing training and testing sets. At the ith fold, the samples $S_1, S_2, \cdots, S_K$, excluding S_i, are merged to a training set and sample S_i is used as a testing set.*

Example. *Figure 14 shows an example of the second fold.*

A.2 Tree-based approaches

Decision tree method approximates the learning function f using decision trees.

Definition. *A decision tree represents a set of conditions that satisfies the classification of instances. Paths from the root to the leaf represent classification rules.*

Example. *Figure 15 is an example of a decision tree using the sample dataset.*

Decision tree algorithms classify instances from the root to the leaves by providing a classification for each instance to the leaves. Each node represents a test on the features, and each branch corresponds to a potential value of a feature.

Example. *In the tree in Figure 15, age is the root node. This node has three branches (young, middle, and old) representing the age values. The first leaf on at the leftmost of the tree represents all instances where individuals are young, and the drought is harsh, where people have a moving intention (move).*

1	2	3	4	5	6	7	8	9	10
train	test	train	train	train	train	train	train	train	train

Fig. 14: 10-folds cross validation

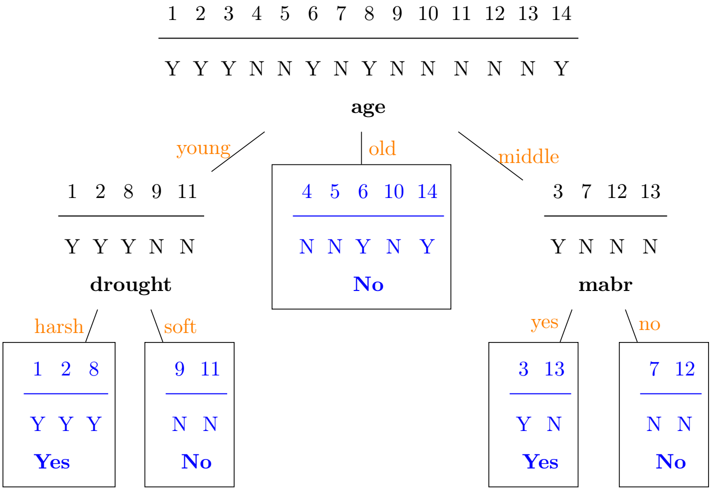


Fig. 15: Example of a decision tree trained with the sample dataset in Table 4. The numbers from 1 to 14 are instance numbers from Table 4. Capitalized Y and N represent the moving intention for each instance

A decision tree is built by selecting the variable at each node that gives the best data split. This split is based on the measure of the impurity rate (obtained by calculating, for example, the entropy or the Gini index) of each variable. The best variable is the one with the lowest impurity rate. Typically, this measure favors splits that allow having the dominant or (strongly) discriminative label over the target attribute.

It is possible to represent a decision tree as a linear function [27]. This is closer to the way that social scientists represent a model. To do so, we represent each leaf of the tree as a variable (feature) of the linear model. This variable is the product of decisions from the root to the leaf. This model thus contains as many variables as there are leaves in the tree. These variables show how decision trees take into account the nonlinearity of the problem automatically.

Example. *Let* $L_1, L_2, \cdots, L_5$ *be the variables of the linear model. These variables represent the leaves of the tree in Figure 15 (from left to right of the tree). The leftmost leaf variable* L_1 *is equal to* $L_1 = 1_{age\ =\ young \wedge drought\ =\ harsh}$*. The variables* L_3 *and* L_5 *are equal to* $L_3 = 1_{age\ =\ old}$*, and* $L_5 = 1_{age\ =\ middle \wedge mabr\ =\ no}$*. Accordingly, the outcome (y) follows:*

$$y = f(L) = \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_5 + \epsilon \tag{A1}$$

As in the example, building and using decision trees (DT) are straightforward and explainable. However, in practice, they might be inaccurate [23]. Thus, several other tree-based methods have been proposed. Random Forest

(RF) [10] and eXtreme Gradient Boosting (XGB) [14] methods are well known and widely used.

Definition (Random Forest)**.** *Random forest consists of several decision trees that operate together as an ensemble. This ensemble of trees is called a forest. Each tree classifies an instance in the forest, and the class label of this instance is decided by a majority vote. Each tree is built on a randomly selected (with replacement) sample of the dataset and a random number of features.*

Example. *With the DT example in Figure 15, instance 1 from our sample example is classified as the class label number (i.e., the individual with instance number 1 has an intention to move). With RF that contains five trees, we classify this instance with each tree and take the majority-class label. Assuming that we have these predictions* {*Yes, Yes, Yes, No, No*}*, RF classifies this instance as Yes.*

Random forest considers the predictions of each tree to have the same weight. By contrast, XGB does not make this assumption, thus, dynamically assigns a certain weight to each tree and instance. At each step of the forest construction, a new tree is added to address the errors made by the existing trees.

By constructing decision trees, one may wonder how deep it needs to go to achieve a better classifier. For a forest, how many trees does it need and how many features must be selected? Basically, in ML, these parameters are determined dynamically by trying several sets of parameters. This process is called parameter tuning. In this paper, we used *Bayesian Hyperparameter Optimization* (BHO) [30].

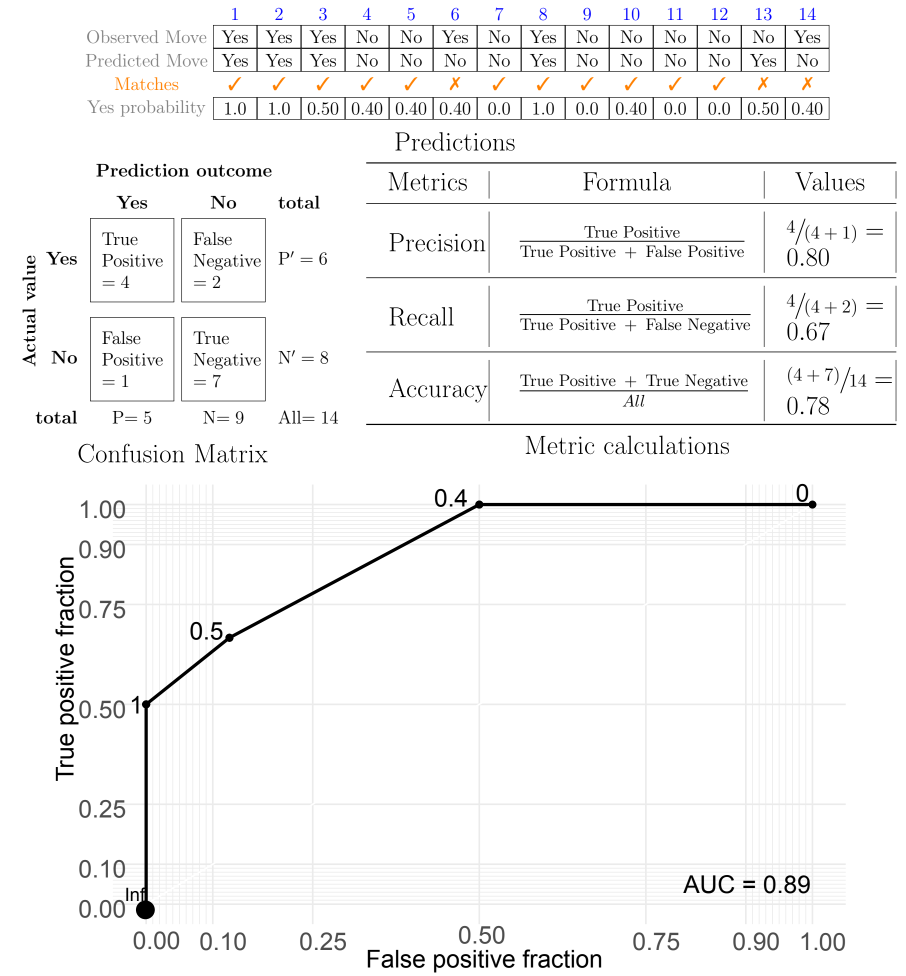
Definition (Bayesian Hyperparameter Optimization)**.** *It consists of testing the models on several parameters and associating each set with a probability to obtain the best performance. A Bayesian model (i.e., probability model) is then used to select the most promising parameters.*

A.3 Performance evaluation

In supervised learning, models are evaluated by making one-on-one comparisons between the predicted outcome ($\hat{y}$) and the real outcome (y). This is a benefit of ML over parameter estimation, where the estimation is usually based on the assumptions made from the data-generating process to ensure consistency [27].

For comparison, in ML, we typically build a confusion matrix.

Definition (confusion matrix)**.** *A confusion matrix is a matrix that compares the predicted values to the ground-truth. It contains four values, namely true positive (actual observation 'Yes' and predicted 'Yes'), false positive (actual*



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Observed Move	Yes	Yes	Yes	No	No	Yes	No	Yes	No	No	No	No	No	Yes
Predicted Move	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	No	Yes	No
Matches	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗
Yes probability	1.0	1.0	0.50	0.40	0.40	0.40	0.0	1.0	0.0	0.40	0.0	0.0	0.50	0.40

Predictions

Prediction outcome

Actual value		**Yes**	**No**	**total**
	Yes	True Positive = 4	False Negative = 2	$P' = 6$
	No	False Positive = 1	True Negative = 7	$N' = 8$
	total	P= 5	N= 9	All= 14

Confusion Matrix

Metrics	Formula	Values
Precision	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$	${}^{4}\!/_{(4+1)} =$ 0.80
Recall	$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$	${}^{4}\!/_{(4+2)} =$ 0.67
Accuracy	$\frac{\text{True Positive} + \text{True Negative}}{All}$	${}^{(4+7)}\!/_{14} =$ 0.78

Metric calculations

Fig. 16: Model performance evaluation with Precision, Recall, Accuracy, and AUC based on the confusion matrix values

observation 'No' but predicted 'Yes', false alarm), true negative (actual observation 'No' and predicted 'No'), and false negative values (actual observation 'Yes' but predicted 'No').

Example. *Figure 16 shows the predicted move intention using the decision tree (DT) and the confusion matrix comparing these predictions to the observed (actual) move intention.*

Based on the confusion matrix, various performance metrics can be computed. The common ones are *accuracy*, *precision*, and *recall*.

Definition (Accuracy - Precision - Recall). *The accuracy is a ratio of correctly predicted observations to the total number of observations. It is an intuitive measure, but only when false positive and false negatives are not too different. Instead, precision shows the ratio of correctly predicted positive observations to the total predicted positive observations, while recall is the ratio of correctly predicted positive observations to all accurate (or true) observations. The formulas are available in Figure 16 with the confusion matrix. These measurements have values between 0 and 1 (the higher, the better performance).*

Predicted class labels typically involve a user-defined threshold (e.g., 0.5). By convention, the probability lesser or equal to the threshold is considered as a 'No' and otherwise a 'Yes'. Differently defined threshold leads to different predictions. The Area under the ROC (Receiver operating Characteristics) curve (AUC) [32, 19], another model performance metric, is used to evaluate the performance regardless of any classification threshold.

Definition (ROC and AUC). *A ROC curve, a two-dimensional graph, is generated by plotting the false-positive fraction (x-axis) against the true-positive fraction (y-axis) of a model for each possible threshold value. The ROC curve shows how well a model classifies binary outcomes. The AUC (Area under the curve), as its name implies, is the area under the ROC curve. Typically, it is computed when a single value is needed to summarize a model's performance to undertake comparisons. The AUC value is also between 0 and 1 (the higher, the better performance).*

Example. *Figure 16 illustrates the ROC curve and the AUC of a decision tree (DT). The AUC of this classifier is 0.89 (i.e. classifier performs well).*

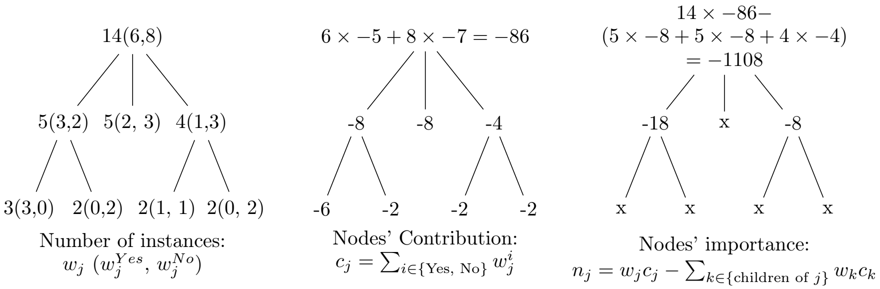
In this paper, we mainly use AUC and precision to determine which method to focus on.

A.4 Output interpretation: Feature importance and Partial Dependence Plots (PDP)

The features X used to estimate f in the equation $f(X) = y$ are rarely equally relevant. Typically, only a small subset of the features is relevant. Hence, after training the model, the *Relative Feature Importance (RFI)* method is used to determine the relevant features. RFI was introduced by Breiman et al [11] for tree-based learning methods.

Definition (RFI). *RFI consists of, (i) for each internal node of a tree T, compute the contribution of each feature to the prediction, (ii) then sum its contributions for each feature, and (iii) arrange the features accordingly.*

To calculate the importance I_j of the feature j (at node j) in a decision tree (A2), five elements are needed: the numbers of 'Yes' (w_j^{Yes}) and 'No'



Number of instances: w_j (w_j^{Yes}, w_j^{No})

Nodes' Contribution: $c_j = \sum_{i \in \{\text{Yes, No}\}} w_j^i$

Nodes' importance: $n_j = w_j c_j - \sum_{k \in \{\text{children of } j\}} w_k c_k$

Fig. 17: The five elements needed to compute the feature importance in DT in Figure 15

(w_j^{No}) instances, the total number of instances ($w_j = w_j^{Yes} + w_j^{No}$) at node j, the contribution of j ($c_j = \sum_{i \in \{\text{Yes, No}\}} w_j^i$), and the importance of node j ($n_j = w_j c_j - \sum_{k \in \{\text{children of } j\}} w_k c_k$).

$$I_j = \frac{n_j}{\sum_{i \in \{\text{all feature nodes}\}} n_i} \tag{A2}$$

Example. *Figure 17 shows how we compute the five elements needed to compute the importance of the feature age, which results in 0.977 using* (A2)*:*

$$I_{age} = \frac{n_{age}}{n_{age} + n_{drought} + n_{mabr}} = \frac{-1108}{-1108 - 18 - 8} = 0.977 \tag{A3}$$

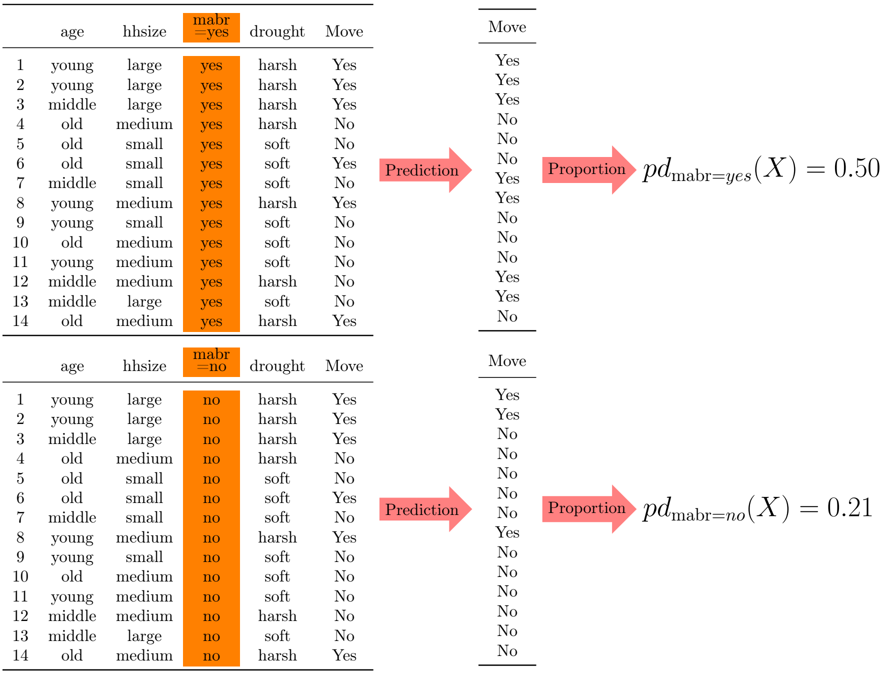
In a single decision tree, it is clear that the most important feature is the feature at the root node. In a forest, (A2) is generalized as follows:

$$RI_j = \frac{\sum_{t \in \{\text{forest}\}} n_j^t}{\sum_{t \in \{\text{forest}\}} \sum_{i \in \{\text{all feature nodes of } t\}} n_i^t} \tag{A4}$$

RFI has become widespread and is used for other ML methods. In order to understand how these important features influence the outcome y, one uses the *Partial Dependency Plots* [23, Chap. 14].

Definition (Partial Dependence)**.** *Assume the features $X = X_1, X_2, \cdots, X_p$, indexed by $P = \{1, 2, \cdots, p\}$. Let S and its complement R be subsets of P, i.e., $S, R \subset P \wedge S \cup R = P \wedge S \cap R = \emptyset$. Assuming that $f(X) = f(X_S, X_R)$, the partial dependence of $f(X)$ on the features X_S is,*

$$PD_S(X_S) = E_{X_R} f(X_S, X_R) \approx \frac{1}{N} \sum_{i=1}^{N} f(X_S, x_{iR}) \tag{A5}$$



	age	hhsize	mabr =yes	drought	Move
1	young	large	yes	harsh	Yes
2	young	large	yes	harsh	Yes
3	middle	large	yes	harsh	Yes
4	old	medium	yes	harsh	No
5	old	small	yes	soft	No
6	old	small	yes	soft	Yes
7	middle	small	yes	soft	No
8	young	medium	yes	harsh	Yes
9	young	small	yes	soft	No
10	old	medium	yes	soft	No
11	young	medium	yes	soft	No
12	middle	medium	yes	harsh	No
13	middle	large	yes	soft	No
14	old	medium	yes	harsh	Yes

Move
Yes
Yes
Yes
No
No
No
Yes
Yes
No
No
No
Yes
Yes
No

	age	hhsize	mabr =no	drought	Move
1	young	large	no	harsh	Yes
2	young	large	no	harsh	Yes
3	middle	large	no	harsh	Yes
4	old	medium	no	harsh	No
5	old	small	no	soft	No
6	old	small	no	soft	Yes
7	middle	small	no	soft	No
8	young	medium	no	harsh	Yes
9	young	small	no	soft	No
10	old	medium	no	soft	No
11	young	medium	no	soft	No
12	middle	medium	no	harsh	No
13	middle	large	no	soft	No
14	old	medium	no	harsh	Yes

Move
Yes
Yes
No
No
No
No
No
Yes
No
No
No
No
No
No

Prediction
Proportion
$pd_{\text{mabr}=yes}(X) = 0.50$
Prediction
Proportion
$pd_{\text{mabr}=no}(X) = 0.21$

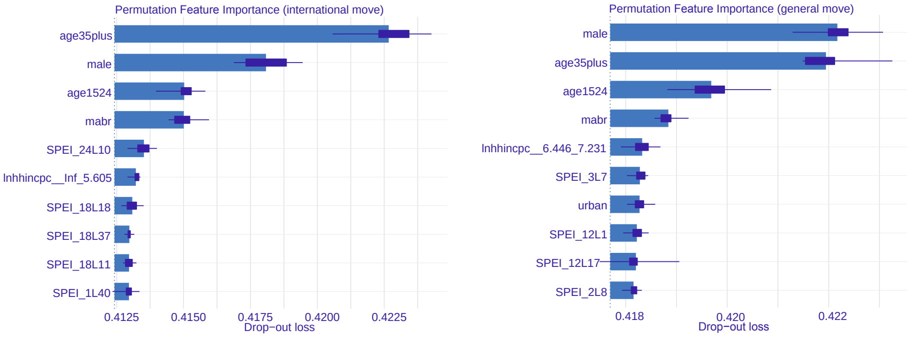
Fig. 18: Partial Dependence computation for the 'mabr' feature (connections abroad)

This is a marginal average of f describing the effect of a chosen set of features S on f. It is approximated as the average over the N instances in the training set (X) of the prediction of each instance (x_{iR}) occurring in the complementary set X_R.

The computation of (A5) requires a pass over the data for each set of joint values of X_S. This can be computationally intensive, and therefore, the partial dependency is usually not calculated with more than three features. Fortunately, partial dependence with only one feature is often informative enough, and it simplifies the calculation with a discrete feature. In practice, for a discrete feature with two class labels 'yes' and 'no', we only compute $PD_S(X_S = yes)$ and $PD_S(X_S = no)$.

Example. *Figure 18 shows how we compute the partial dependence in DT (Figure 15) on a feature 'mabr' (human network abroad).*

From the different values used to calculate the partial dependence, we can draw a chart with the tested values in x-axis against the partial dependence output in y-axis. The plot's role is to show in which direction (towards label



(a) international move

(b) general move

Fig. 19: Male and age appear as top features based on the permutation feature importance

'Yes' or 'No') each feature value drives the outcome y. The plot visualizes the effect of a feature related to the average effects of other features.

Appendix B Additional figures

Figure 19 shows male and age as top influencing features according to the permutation feature importance measures, similar to the results from the Relative Feature Importance (RFI) method. We also observe international move is more affected by longer SPEIs (e.g., 18, 24) while general move is affected by shorter SPEIs (e.g., 2, 3, 12) which aligns with previous findings. Darker box plot shows the uncertainty from the permutations. Permutation feature importance measures the increase of a model's prediction error after a certain feature's value is permuted. The permutation breaks the relationship between the feature and the true outcome. A feature is considered 'important' if the change of a feature value increases the model error since it means that the model relies on that feature for prediction. Fisher et al [20] proposed 'model reliance' measures and a model-agnostic permutation feature importance algorithm.

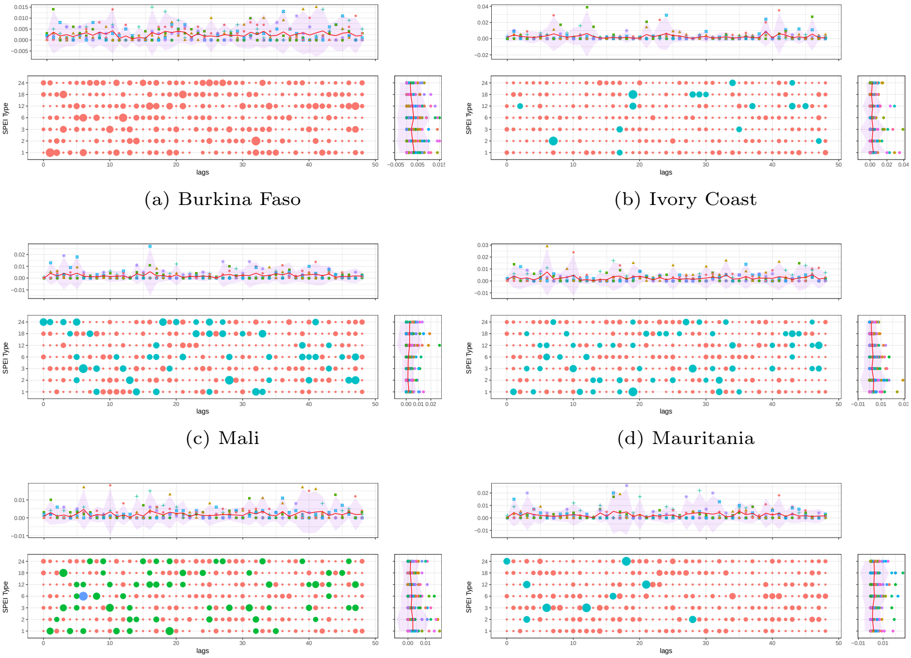
Figure 20 shows the feature importance distributions of the six countries targeting international move over the seven SPEI timescales (i.e., 1, 2, 3, 6, 12, 18, 24) and 49 lags (i.e., 0-48).

Appendix C Terminology comparison

Table 5 compares the common terminology used in social sciences and machine learning.

Appendix D GWP questions

Table 6 describes the World Poll questions used to measure the opinions of the interviewees.



(a) Burkina Faso

(b) Ivory Coast

(c) Mali

(d) Mauritania

(e) Niger

(f) Senegal

Fig. 20: Feature importance (dot size) based on different SPEI timescales and lags with the distribution of those by each lag (top) and each SPEI (right)

Table 5 The mapping of the terminology used in social science and machine learning

Social science	Machine Learning
independent variable, covariate, control variable	variable, feature, attribute, column
observation	observation, row, example, instance
output, dependent variable, outcome	output, dependent variable, target attribute
subsample	training set
subsample	test set

Table 6 GWP questions

Feature	Description
age	Please tell me your age.
hhsize	Including yourself, how many people who are residents of this country, age 15 or over, currently live in this household?
hskill	Education Category
Inhhincpc	Annual household income in International Dollars
mabr	Do you currently have family members or relatives living permanently in other countries, or not? (including countries of the former Soviet Union)
male	Gender
urban	Do you live in . . . ?
year	The year Gallup survey was performed.

References

[1] Athey S (2015) Machine learning and causal inference for policy evaluation. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 5–6, https://doi.org/10.1145/2783258.2785466

[2] Athey S (2018) The Impact of Machine Learning on Economics, University of Chicago Press, pp 507–547. https://doi.org/10.7208/chicago/9780226613475.001.0001

[3] Athey S, Imbens GW (2019) Machine learning methods that economists should know about. Annual Review of Economics 11(1):685–725. https://doi.org/10.1146/annurev-economics-080217-053433

[4] Athey S, Tibshirani J, Wager S, et al (2019) Generalized random forests. The Annals of Statistics 47(2):1148–1178. https://doi.org/10.1214/18-AOS1709

[5] Beine M, Jeusette L (2018) A meta-analysis of the literature on climate change and migration. Journal of Demographic Economics pp 1–52. https://doi.org/10.1017/dem.2019.22

[6] Berlemann M, Steinhardt MF (2017) Climate change, natural disasters, and migration—a survey of the empirical evidence. CESifo Economic Studies 63(4):353–385. https://doi.org/10.1093/cesifo/ifx019

[7] Bertoli S, Docquier F, Rapoport H, et al (2021) Weather shocks and migration intentions in Western Africa: insights from a multilevel analysis. Journal of Economic Geography URL https://academic.oup.com/joeg/advance-article-pdf/doi/10.1093/jeg/lbab043/41299221/lbab043.pdf

[8] Black R, Arnell NW, Adger WN, et al (2013) Migration, immobility and displacement outcomes following extreme events. Environmental Science & Policy 27:32–43. https://doi.org/10.1016/j.envsci.2012.09.001

[9] Branco P, Torgo L, Ribeiro RP (2016) A survey of predictive modeling on imbalanced domains. ACM Comput Surv 49(2). https://doi.org/10.1145/2907070

[10] Breiman L (2001) Random forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324

[11] Breiman L, Friedman J, Stone C, et al (1984) Classification and Regression Trees. The Wadsworth and Brooks-Cole statistics-probability series, Taylor & Francis, https://doi.org/10.1201/9781315139470

[12] Cattaneo C, Peri G (2016) The migration response to increasing temperatures. Journal of Development Economics 122:127–146. https://doi.org/10.1016/j.jdeveco.2016.05.004

[13] Cattaneo C, Beine M, Fröhlich CJ, et al (2019) Human migration in the era of climate change. Review of Environmental Economics and Policy 13(2):189–206. https://doi.org/10.1093/reep/rez008

[14] Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: Krishnapuram B, Shah M, Smola AJ, et al (eds) Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. ACM, pp 785–794, https://doi.org/10.1145/2939672.2939785

[15] Dell M, Jones BF, Olken BA (2014) What do we learn from the weather? the new climate-economy literature. Journal of Economic Literature 52(3):740–98. https://doi.org/10.1257/jel.52.3.740

[16] Djamba YK (2003) Gender differences in motivations and intentions to move: Ethiopia and south africa compared. Genus 59(2):93–111

[17] Duan L, Street WN, Liu Y, et al (2014) Selecting the right correlation measure for binary data. ACM Transactions on Knowledge Discovery from Data (TKDD) 9(2):1–28. https://doi.org/10.1145/2637484

[18] Eslamian S (2014) Handbook of engineering hydrology: environmental hydrology and water management. CRC press, https://doi.org/10.1201/b16766

[19] Fawcett T (2006) An introduction to roc analysis. Pattern recognition letters 27(8):861–874. https://doi.org//10.1016/j.patrec.2005.10.010

[20] Fisher A, Rudin C, Dominici F (2018) Model class reliance: Variable importance measures for any machine learning model class, from the rashomon perspective. arXiv preprint arXiv:180101489 68

[21] Gallup (2015) Worldwide research methodology and codebook

[22] Harris I, Osborn TJ, Jones P, et al (2020) Version 4 of the cru ts monthly high-resolution gridded multivariate climate dataset. Scientific data 7(1):1–18. https://doi.org/10.1038/s41597-020-0453-3

[23] Hastie T, Tibshirani R, Friedman JH (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer Series in Statistics, Springer, https://doi.org/10.1007/978-0-387-84858-7

[24] Mayda A (2010) International migration: a panel data analysis of the determinants of bilateral flows. J Popul Econ 23:1249–1274. https://doi.org/10.1007/s00148-009-0251-x

[25] McFadden D, et al (1973) Conditional logit analysis of qualitative choice behavior. Institute of Urban and Regional Development, University of California

[26] Millock K (2015) Migration and environment. Annu Rev Resour Econ 7(1):35–60. https://doi.org/10.1146/annurev-resource-100814-125031

[27] Mullainathan S, Spiess J (2017) Machine learning: An applied econometric approach. Journal of Economic Perspectives 31(2):87–106. https://doi.org/10.1257/jep.31.2.87

[28] Provost F, Fawcett T (2013) Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly Media, Inc.

[29] Quinlan JR (1986) Induction of decision trees. Machine Learning 1:81–106. https://doi.org/10.1007/BF00116251

[30] Snoek J, Rippel O, Swersky K, et al (2015) Scalable bayesian optimization using deep neural networks. In: Bach FR, Blei DM (eds) Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, JMLR Workshop and Conference Proceedings, vol 37. JMLR.org, pp 2171–2180, URL https://dl.acm.org/doi/10.5555/3045118.3045349

[31] StataCorp L, et al (2007) Stata data analysis and statistical software. Special Edition Release 10:733

[32] Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240(4857):1285–1293. https://doi.org/10.1126/science.3287615

[33] Tjaden J, Auer D, Laczko F (2019) Linking migration intentions with flows: Evidence and potential use. International Migration 57(1):36–57. https://doi.org/10.1111/imig.12502

[34] Vicente-Serrano SM, Beguería S, López-Moreno JI, et al (2010) A new global 0.5 gridded dataset (1901–2006) of a multiscalar drought index: comparison with current drought index datasets based on the palmer drought severity index. Journal of Hydrometeorology 11(4):1033–1043. https://doi.org/10.1175/2010JHM1224.1

[35] Vicente-Serrano SM, Beguería S, López-Moreno JI (2010) A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. Journal of Climate 23(7):1696–1718. https://

doi.org/10.1175/2009JCLI2909.1

[36] Wilhite DA, Svoboda MD (2000) Drought early warning systems in the context of drought preparedness and mitigation. Early warning systems for drought preparedness and drought management (Geneva: World Meteorological Organization) pp 1–21