Predictive Machine Learning Models for Exploring Environment and Migration Linkages in Historical Mexico-U.S. Flows

Filiz Garip
Cornell University

Mario Molina
Cornell University

Abstract
Prior research identifies several linkages through which environmental factors – both gradual changes and sudden-onset events – might shape internal and international migration flows. These linkages point to various explanatory variables, and complex interactions among them. Empirical work tests some of these linkages in isolation, but reports results that vary considerably depending on the particular variables or models used. To address this issue, we propose to use machine learning (ML) tools that allow us to include all potential indicators (and all possible interactions) to predict U.S.-bound migration outcomes among 120,000+ individuals in 1980-2017 in the Mexican Migration Project data. These tools rely on data-driven model selection, optimize predictive performance, but often produce 'black-box' results. To overcome this shortcoming, we propose to use the predictions as a starting point, and analyze discrepant communities for which our model offers poor predictions.

Migration as a result of environment and climate stressors has recently gained widespread interest, and dominates the political discourse in many countries. Climate change is projected to accelerate human displacement in the future by increasing the frequency and severity of extreme environmental events, such as droughts, sea level rise, floods, and hurricanes (Homer-Dixon 2010; Gray 2013). The extent to which we can respond to these challenges effectively, as a global community, relies in part on our ability to predict future patterns of mobility.

Researchers have recently started to link weather variation (shorter-run changes in temperature or precipitation) to internal and international migration patterns.[1] But empirical results have remained mixed. Studies have relied on varying specifications of weather events (e.g., using different lags, linear or non-linear functions, interactions with other indicators), making it hard to generalize from the observed patterns. Studies have also considered various mechanisms underlying weather-migration linkages, and in some cases, singled out one mechanism to establish a causal relationship (Feng et al. 2010).

This study takes an alternative approach. Rather than providing yet another description of weather-migration relationship, we seek to test the predictive power of weather indicators (along with other potential factors) for migration outcomes. Inspired by the budding machine-learning applications in the social sciences (Molina and Garip 2019), we use random-forest models to predict migration decisions. Unlike traditional regression approaches, these models do not yield interpretable parameter estimates, but instead optimize predictive performance.

We focus on the historical migration flows from Mexico the United States, and use the largest existing survey (Mexican Migration Project data) that captures 150,000+ individuals' movements within Mexico and to the United States from 1980 to 2017. We combine the surveys with fine-grained gridded temperature and precipitation data, and build a predictive model of individuals' first-migration decision. In addition to weather indicators, our model contains demographic, economic, and social indicators at the individual, household, and community levels. The predictive modeling framework allows us to include varying specifications of these indicators (for example, time-lagged versions, non-linear terms, interactions), and let the data guide the model selection process.

While we care about predictive performance, we do not see it as an end goal, but rather a starting point. That is, once we choose the optimal model, we go back to our data, and try to identify patterns in what we can (and cannot) predict with our model. We inquire, for example, whether our model performs better in rural or urban communities, or across earlier or later time periods, or for different population groups. By doing so, we aim to understand the ability of our current survey measurement efforts to capture the factors relevant to migration decisions for different groups or in different contexts. Our goal is to inform future data collection efforts.

[1] 'Climate' refers to distribution of outcomes over a longer time span. 'Weather' can be thought of as a particular empirical realization from that distribution.

Linkages between migration and the environment

Social scientists are increasingly interested in how weather fluctuations shape various outcomes, including agriculture, labor productivity, health, political stability, and conflict. Empirical findings to date suggest that temperature, precipitation, and extreme weather events have statistically significant effects on a variety of economic and political outcomes (Dell et al. 2013).

Until recently, migration scholars have not considered weather variation as a potential determinant of rural-to-urban or international moves. Instead, our current understanding of migration highlights individual-level motivations and community- or country-level economic, social, and political drivers. As a result, we still know relatively little about the influence of the environmental factors on migration decisions.

Some researchers view environment as the primary factor in migration, and study the so-called 'environmental refugees' who flee droughts (El-Hinnawi 1985), land de-gradation (Kavanagh and Lonergan 1992), or sea level rises (Myers 1993). Other researchers critique the implied separation of environmentally-forced migrants from those responding to economic or political factors (Castles 2002), or point to missing evidence on the linkages between environmental indicators and migration (Baldwin 2017; Black 2001).

Figure 1 offers a synthesis of various mechanisms that researchers posit for linking environmental factors to migration decisions. It reflects the pre-dominant thinking in the literature that environmental factors interact with economic, social, and political processes to shape migration patterns (Black et al. 2011; Hunter et al. 2015; McLeman and Smit 2006). Two concepts – vulnerability and adaptive capacity – capture this general idea. For example, a community that relies on rainfall for agricultural production is more 'vulnerable' to drought-related economic stress, and hence, more prone to sending migrants in response. Similarly, a community with social connections to a particular destination is more likely to turn to migration as an adaptation strategy in the face of climate stress compared to a community with few connections.

It is difficult to empirically identify these linkages altogether. As a result, empirical work has focused on reduced-form results, or considered particular mechanisms in isolation. Munshi (2003), for example, connects low rainfall in Mexican communities to higher out-migration rates to the United States. Feng et al. (2010), for example, focus on crop yields, and show precipitation-related crop declines to be associated with U.S.-bound migration in Mexico. They do not consider any other potential mechanism.

More generally, empirical results on environment-migration link tend to be varied, and often hard to reconcile. For example, Riosmena et al. (2018) report two conflicting findings using Mexican census data. First, lower rainfall increases out-migration in communities of low vulnerability (that is, those that are rich). Second, higher-than-average temperature increases out-migration in communities of high vulnerability (that is, those that are poor). Two similar weather stressors generate two different responses to vulnerability.

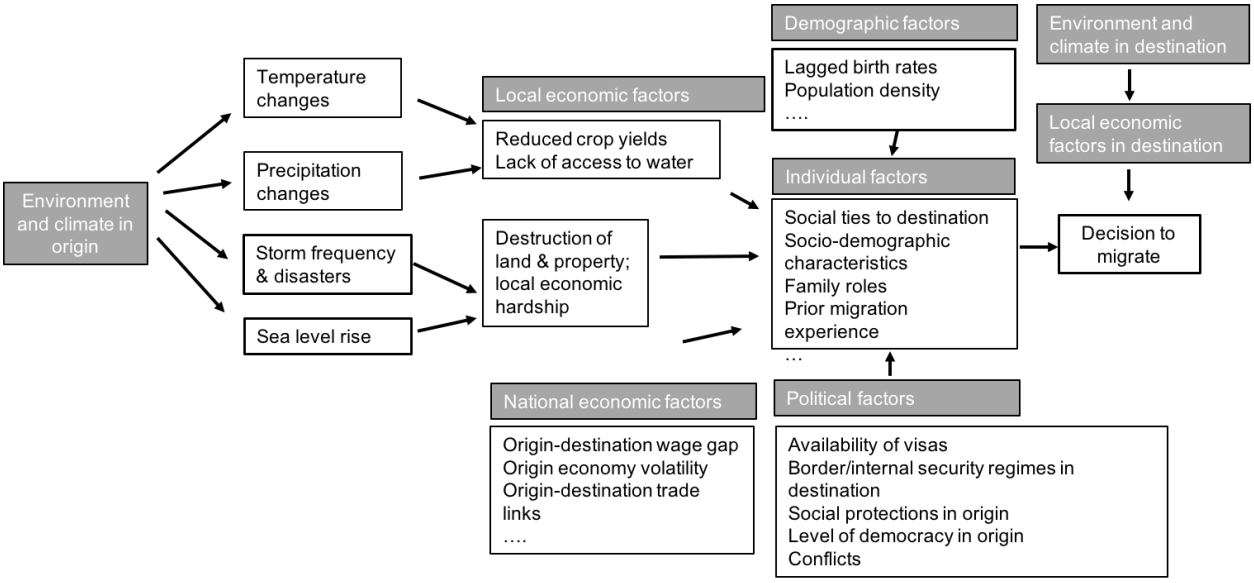


Figure 1 Theoretical framework for environment-migration linkages

Another mixed result is related to the moderating effect of migrant networks (that is, social connections to prior migrants to a given destination). While networks in a community increase the migration response to climate events according to Hunter et al.'s (2013) analysis of the MMP data, they decrease the migration propensity to climate events in Riosmena et al.'s (2018) census-based computations.

The literature, in other words, is full of mixed results, even though the studies often share the authors, and rely on similar data sets (e.g., Nawrotzki et al. 2015; Riosmena et al. 2018; Hunter et al. 2013). This pattern might signal potential model misspecification. That is, small perturbations on the variable definitions (e.g., maximum 5-day precipitation versus number of days with heavy precipitation) seems to alter the results considerably. Similarly, small changes in the sample used (e.g., rural communities versus all communities in the MMP) seems to change the observed associations.

This is a major obstacle to cumulative knowledge, but it is not one that is easy to resolve. Many weather-related indicators are highly correlated; various definitions of measures are equally reasonable (e.g., coldest day temperature versus % of cool days) at the outset; and it is hard to determine which measures are the right ones a-priori.

One particular issue, then, is to account for many potential indicators of weather variations. Another one, noted in the literature, is to flexibility consider nonlinearities in the weather-migration relationship. Basically, we do not know if there are particular 'thresholds' over which weather variation exerts an effect on migration, and how these different thresholds might vary across different settings or different groups of individuals. A third –related- issue is the

potential 'adaptation' or 'intensification' effects. That is, communities/individuals can adapt to weather changes (e.g., by switching to different crops in rural settings), or the effects of weather changes can be felt more over time (e.g., once the reservoir runs dry in irrigated regions).

These issues all point to the uncertainty (i) what the best measures to capture weather variation are, (ii) what time-lags are appropriate to include in our models to detect potential adaptive behaviors, and (iii) what particular specifications (linear, quadratic, step-function, and so on) would capture possible intensification effects.

To address these issues that relate to parameterization as well as heterogeneity in potential effects, we suggest changing the goal of our analysis from the identification of 'effects' of particular indicators, to the prediction of migration outcome where many indicators (as well as their time-lags, and alternative specifications) are flexibly included in the model. Below we describe this approach in greater detail. But first we briefly describe the Mexican setting.

Mexican setting

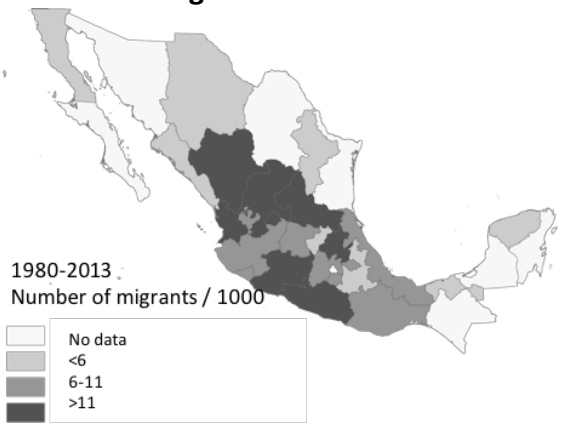


Figure 2. Map of Mexico with 24 states categorized into three groups based on the number of U.S.-bound migrants in 1990-2013 per 1000 residents (Source: Authors' calculations from the MMP data)

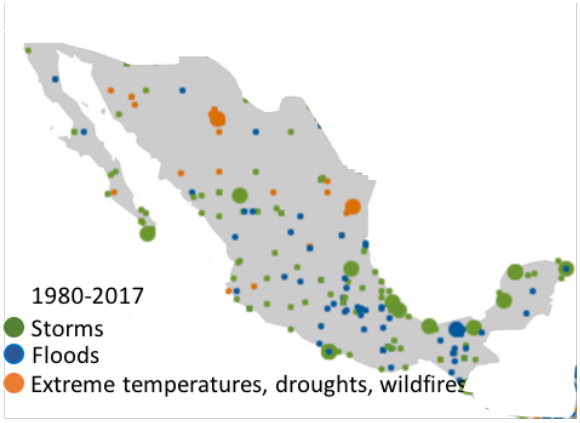


Figure 3. Map of Mexico with select environmental events 1980-2017. Size of the circle indicates the number of events. (Source: NatCatSERVICE [2])

Mexico has been the origin country for the largest international migration in the world. Between 1960 and 2010, an estimated 12 million Mexicans have migrated to the United States (Garip 2016). The country has experienced numerous environmental events including storms, floods, extreme temperatures, droughts and wildfires leading to an estimated 4,864 fatalities between 1980 and 2017 alone (Muenchen RE 2017). Yet, only few studies have linked the patterns of human mobility to environmental events (Gray 2013; Hunter, Murray and Riosmena 2013). Figure 2 illustrates the states with the highest numbers of U.S.-bound migrants, and Figure 3 highlights the regional distribution of select extreme-weather events.

Prior work has attempted to dissect the environment-migration link in the Mexican setting with descriptive regression models (Hunter et al. 2013), and in some cases, with further attempts at causal identification (Feng et al. 2010). In what follows, we propose an empirical strategy that instead focuses on prediction of international moves from Mexico, using slow-onset weather events (extreme temperatures and rainfalls) as predictors.

Machine-learning approach – a brief review

Machine learning is a field at the intersection of statistics and computer science that uses algorithms to extract information and knowledge from data. Its applications increasingly find their way into economics, political science, and sociology. For a recent review of this vast toolbox, see Molina and Garip (2019).

Supervised machine learning (SML) involves searching for functions, f(*X*), that predict an output (*Y*) given an input (*X*). One can consider different classes of functions, such as linear models, decision trees, or neural networks.

Let's take the linear model as a tool for prediction.[2] We have an input vector, *X*, and want to make a prediction on the output, *Y*, denoted as $\hat{Y}$ ('y-hat') with the model

$$\hat{Y} = \mathrm{f}(X) = X^{\mathrm{T}}\beta$$

where X^T is the vector transpose and β ('beta') is the vector of coefficients.

Suppose we use ordinary least squares (OLS) – the most commonly used method in sociology – to estimate the function, f(*X*), from data. We pick the coefficients, β, that minimize the sum of squared residuals

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 \qquad (1)$$

This strategy ensures estimates of β that give the best fit *in sample*, but not necessarily the best predictions *out of sample* (i.e., on new data) (see sidebar titled Classical Statistics versus Machine Learning).

To see that, consider the generalization error of the OLS model, that is, the expected prediction error on new data. This error comprises of two components: bias and variance (Hastie et al. 2009). A model has bias if it produces estimates of the outcome that are consistently wrong in a particular direction (e.g., a clock that is always an hour late). A model has variance if its estimates deviate from the expected values across samples (e.g., a clock that alternates

[2] Uppercase letters, such as *X* or *Y,* denote variable vectors, and lowercase letters refer to observed values (e.g., x_i is the *i*-th value of *X*).

between fast and slow) (Domingos 2015). OLS minimizes in-sample error (equation 1), but it can still have high generalization error if it yields high-variance estimates (Kleinberg et al. 2015).

To minimize generalization error, SML makes a trade-off between bias and variance. That is, unlike OLS, the methods allow for bias in order to reduce variance (Athey and Imbens 2017). For example, an SML technique is to minimize

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda R(f) \qquad (2)$$

that is, in-sample error plus a **regularizer**, $R(f)$, that penalizes functions that create variance (Kleinberg et al. 2015; Mullainathan and Speiss 2017). An important decision is to select λ ('lambda'), which sets the relative 'price' for variance (Kleinberg et al. 2015). In OLS, that price is set to zero. In SML methods, the price is determined using the data (more on that later).

SML techniques seek to achieve an ideal balance between reducing the in-sample and out-of-sample error (aka training and generalization error, respectively). This goal helps avoid two pitfalls of data analysis: underfitting and overfitting. Underfitting occurs when a model fits the data at hand poorly. Take a simple example. An OLS model with only a linear term linking an input (X) to output (Y) offers a poor fit if the true relationship is quadratic. Overfitting occurs when a model fits the data at hand too well, and fails to predict the output for new inputs. Consider an extreme case. An OLS model with N inputs (plus a constant) will perfectly fit N data points, but likely not generalize well to new observations (Belloni et al. 2014).

Underfitting means we miss part of the signal in the data; we remain blind to some of its patterns. Overfitting means we capture not just the signal, but also the noise, that is, the idiosyncratic factors that vary from sample to sample. We hallucinate patterns that are not there (Domingos 2015).

Through regularization, SML effectively searches for functions that are sufficiently complex to fit the underlying signal without fitting the noise. To see that, note that a complex function will typically have low bias but high variance (Hastie et al. 2009). And recall that the regularizer, $R(f)$, penalizes functions that create variance; it often does so by expressing a model's complexity.

Let us consider regression trees, a function class in SML. The method proceeds by partitioning the inputs (X) into separate regions in a tree-like structure, and returning a separate output estimate ($\hat{Y}$) for each region. Say we want to predict whether someone migrates using individual attributes of age and education. A tree might first split into two branches by age (young and old), and then each branch might split into two by education (college degree or not). Each terminal node ('leaf') corresponds to a migration prediction (e.g., 1 for young college graduates). With enough splits in the tree, one can perfectly predict each observation within sample. To prevent overfitting, a typical regularizer controls the tree depth, and thus, makes us

search not for the best fitting tree overall, but the best fitting tree among those of a certain depth (Mullainathan and Speiss 2017).

What sets SML apart from classical statistical estimation, then, are two essential features: regularization, and the data driven-choice of regularization parameters (aka empirical tuning) (Mullainathan and Speiss 2017; Athey and Imbens 2017; Kleinberg et al. 2015). These features allow researchers to consider complex functions and more inputs (polynomial terms, high-order interactions, and, in some cases, more variables than observations) without overfitting the data. This flexibility contrasts sharply with classical statistics, where one typically selects a small number of inputs (*X*), and a simple functional form to relate the inputs to the output (*Y*).

One way SML uses data, therefore, is for model selection, that is, to estimate the performance of alternative models (functions, regularization parameters) to choose the best one. This process requires solving an optimization problem. Another way SML uses data is for model assessment, that is, having settled on a final model, to estimate its generalization (prediction) error on new data (Hastie et al. 2009).

A crucial step in SML is to separate the data used for model selection from the data used for model assessment. In fact, in an idealized set-up, one creates three, not two, separate data sets. Training data is used to fit the model; validation data is put aside to select among different models (or to select among the different parameterizations of the same model), and finally, test (or hold-out) data is kept 'in the vault' to compute the generalization error of the selected model.

Models – Random Forests

We use random forests as an ML tool which average over multiple 'decision trees' and capture nonlinearities and interactions in inputs (Breiman 2001a). We train a random forest to obtain our migration predictions. Random forests are statistical models based on classification and regression trees (CARTs) that capture nonlinearities and interactions between covariates (Breiman 2001a). CARTs build a decision tree that makes partitions of the covariate space based on the minimization of a loss function, whose goal is to provide information on partitions that contain elements as similar as possible. Each leave is constructed so that observations that are more similar to each other fall within the same leaf. In a classification task – i.e. when we have a binary outcome –, the ultimate goal of the tree-like process is to assign a probability to each observation that informs about its class membership (e.g. migrants or non-migrants).

Random forests are bootstrap bags of *k* trees. Each tree is obtained using a random sample of *D* covariates (with replacement). We set *k*=100 and sample D covariates. Since each tree estimates a probability using a subset of the covariates, estimates are slightly biased with respect to a non-random tree. However, random forests average all probabilities obtained from each tree and therefore provide an unbiased estimate for the membership probability.

We start our analysis by splitting our data into training (50% of individuals), validation (25%) and test (25%) data. We use the training data to fit our models, validation data to 'tune' it (that

is, figure out which model parameters give the best predictive accuracy), and finally rely on the test data to report the final accuracy of our predictions. Because migration is a rare-event in the data, the average prediction accuracy is likely to overstate our actual ability to predict migration. (Say, 95 percent of individuals are non-migrants in the data. A model that can predict 'no migration' by default – without any input – and thus seem to have high accuracy.) Therefore, we report prediction accuracy separately for migrants and non-migrants.

Data

We use data from the Mexican Migration Project (MMP) with retrospective life histories of 120,000+ individuals (20,000+ U.S.-bound migrants) between 1980 and 2017. The data provide detailed records of international migration trips in addition to information on individual characteristics, household demographic and economic make-up, and community institutions over time.

The majority of quantitative results on Mexico-U.S. migration are based on data from two surveys: the Mexican National Survey of Population Dynamics (ENADID) and the Mexican Migration Project (MMP). The former is a representative national sample, but contains information on only labor migrants. The latter is from specific Mexican communities, but covers all migrants, including those who have moved to the United States to join family members.

The inclusion of all migrants, not just labor-force participants, makes the MMP data more advantageous to fully understand the Mexico-U.S. stream. These data are not strictly representative of the Mexican population. Yet, prior work found that the MMP data yield an accurate profile of the U.S. migrants in Mexico, and this profile is largely consistent with that observed in the ENADID data (Durand et al. 2001; Zenteno and Massey 1998).

The MMP data come from 161 communities located in major migrant-sending areas in 21 Mexican states. Each community was surveyed once between 1987 and 2017, during December and January, when the U.S. migrants are mostly likely to visit their families in Mexico. In each community, individuals (or informants for absent individuals) from about 200 randomly selected households were asked to provide demographic and economic information and to state the timing of their first and last trip to the United States. Household heads were additionally asked to report the trips in between. These data were supplemented with information from a non-random sample of migrants identified with snowball sampling in the United States (about 10% of the sample).

Because more detailed information is available for household heads, most studies of the MMP have restricted attention to this sub-population. To provide a more representative portrait of migrants, we consider all household members. We focus on the first trip to the United States. Subsequent trips are not considered as they are recorded only for household heads, and also to avoid a complication that has haunted prior work on migration. This complication arises from the fact that many attributes related to migration behavior are also changed by it. Over successive trips, migrants gradually gain more experience, establish stronger ties to destination, and become wealthier. Their attributes change, not as a result of the changing selectivity of the

stream, but due to the changes caused by prior migration trips. Focusing on first-time migrants allows us to observe migrants' attributes independently from this reciprocal relationship.

A concern with the MMP data is the retrospective nature of the information on migrants. Let's take a household surveyed in 1990, where the daughter has migrated to the United States for the first time in 1980. Her attributes, like age and education, were recorded in 1990, but could be projected linearly to 1980. The economic status of her household could be reconstructed using the data on the timing of asset purchases. The characteristics of her community could be traced back using the retrospective community history. All these plausible steps rely on one crucial assumption: that the daughter in question was living in the same household and community in 1980. While this assumption is viable for most cases, the study cannot account for the cases for which it is not.

We combine the MMP data with daily gridded weather data obtained from ORNL DAAC (Oak Ridge National Laboratory Distributed Active Archive Center), one of the NASA Earth Observing System Data and Information System data centers. These data offer fine-grained information (and more complete spatial coverage) by interpolating data obtained from ground stations over a 'grid' (in our case, each grid is 4km x 4km). Gridded data offer a balanced panel (rather than scattered data based on the locations of stations), and thus are commonly preferred by social scientists. Yet, the data require the selection of a particular interpolation scheme (which typically work better for temperature than precipitation measures) (Dell et al. 2013).

We obtain shape files for the 161 communities in the MMP data, and then overlay the weather grids on community boundaries. We aggregate the weather information for each community by taking the average measure across the grids within the community boundary.

We then build a predictive model of taking a first U.S.-bound trip, and include demographic (age, sex), socio-economic (household wealth, education), social-network (number of prior migrants in the household and community) and community-level (share employed in agriculture, urban-rural status, population density) measures. We also include measures of temperature and precipitation. The gridded weather data are available starting in 1980.

We use different measures of temperature and precipitation, but in each measure, we make an attempt to account for differences across regions in what weather conditions are considered 'normal'. Basically, we first the average temperature at the state-level in the 1960-1979 period. Then, for each community-time period, we compute the deviation from that normal (i.e., difference from the 1960-1979 mean divided by the standard deviation in 1960-1979).

Descriptive information

Table 1 shows the mean, minimum, and maximum values for the key indicators in the MMP data. Our data includes nearly 2 million person-years that come from 129,968 unique persons nested in 26,907 households in 161 communities. Because detailed temperature and precipitation data (at the grid level) are available from 1980 onwards, we restrict the survey data to the 1980-2016 period.

Table 1 Descriptive statistics for indicators from the MMP data

Variables	Mean	Min	Max
Age	32.6	15	99
Years of education	7.2	0	28
Value of land owned (in 2010 US$)	3863	0	199754
Number of rooms in properties owned	0.5	0	18
Household owns a business?	0.2	0	1
Share of men in agriculture in community	0.5	0	1
Share of households earning 2 x min wage+	0.3	0	0.8
Share of ever-migrants in community	0.1	0	0.6
N (years)	37		
N (communities)	161		
N (households)	26,907		
N (persons)	129,968		
N (person-years)	1,983,249		

Figure 4 shows the variation in the distribution of migration prevalence (defined as the percent of the population who has ever migrated) across communities in 5-year periods in 1980-2000. Note that the median values in the box plots are not necessarily increasing in time as different communities enter the panel data in different years (depending on the time the survey was conducted in each community). The figure shows a striking variability in migration outcomes across communities, with some communities having a negligible share of their population migrate, while other communities sending a third or more of their members to the United States at least once.

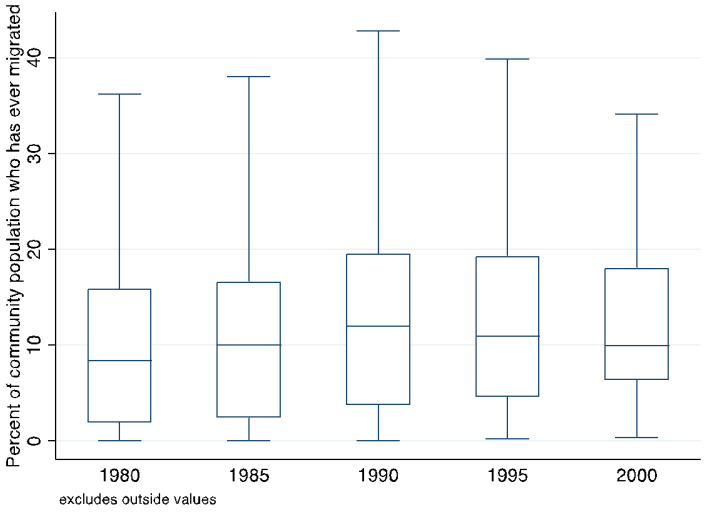


Figure 4. Box plot of percent of community population who has ever migrated to the United States across 161 communities

We also observe quite a bit of weather variation across communities and over time. Figures 5 and 6 respectively show the monthly maximum temperature and average precipitation rates (shown as standardized deviations from the 1960-1979 mean) for 4 selected communities (out of 161 in total). In each figure, the long-dashed line corresponds to 1 standard deviation bounds from the 1960-1979 mean, and the short-dashed line shows the 2 standard deviation bounds.

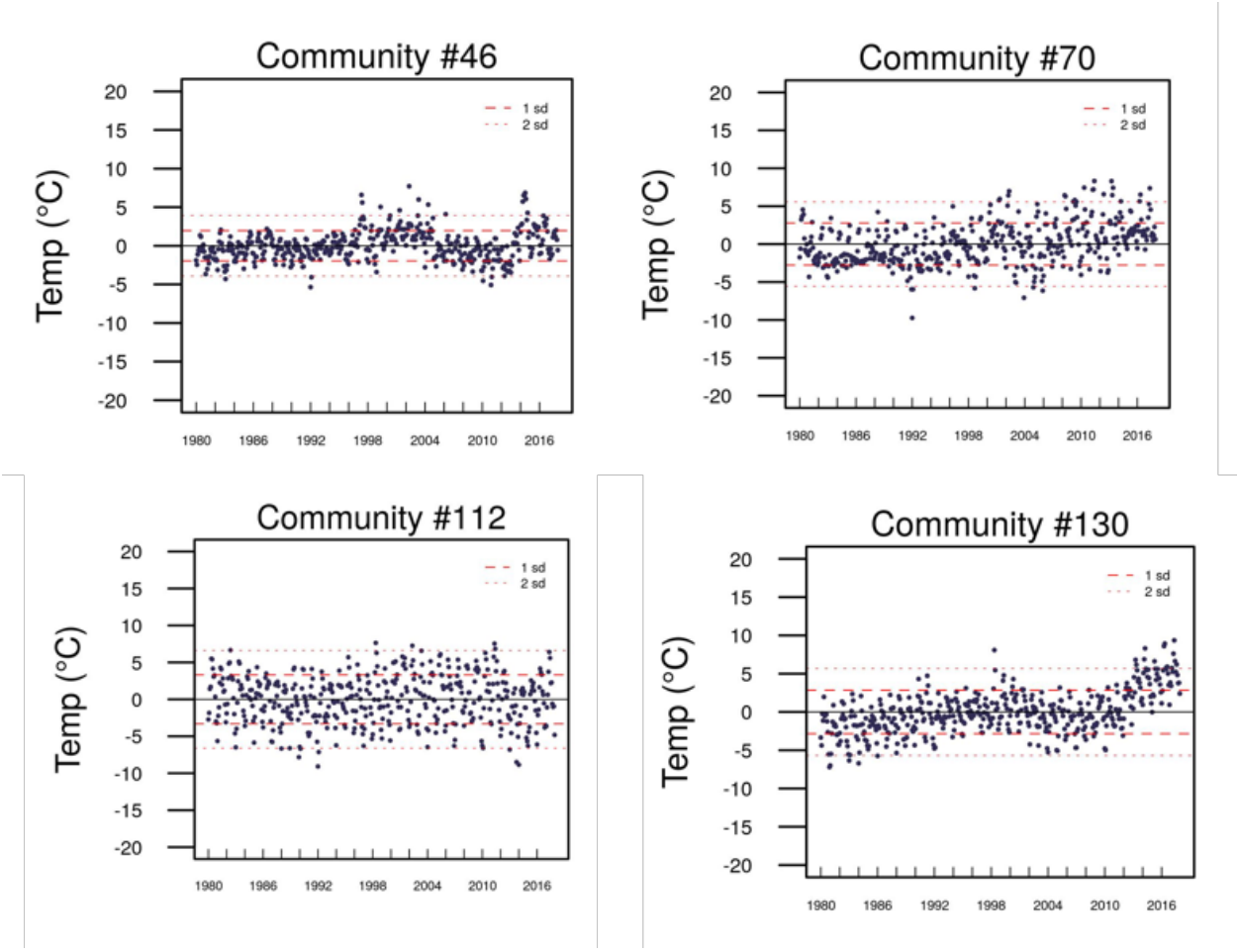


Figure 5. Maximum temperature in each month (1980-2016) for selected four communities. The results are presented as deviations from the historical mean (1960-1979) in the state divided by the historical standard deviation. The red lines show the 1 (long dashes) and 2 (short dashes) standard deviation bounds set by historical observations.

Let's take community #112 (lower-left panel). This community does not steer far off from its 'normal' temperature and rainfall experience from 1980 to 2016, as almost all monthly values in this period fall within the 2-standard deviation bounds of the 1960-1979 values in its state. Community #46 experiences slightly higher maximum temperatures in 2000s, and lower ones in 2010s compared to its earlier average; and it receives occasional excessive rains over the entire

period. Community #130 starts to consistently exceed the bounds on maximum temperature set by its historical average from 2010 onwards, and it simultaneously receives excessive rainfall in some months in the same period.

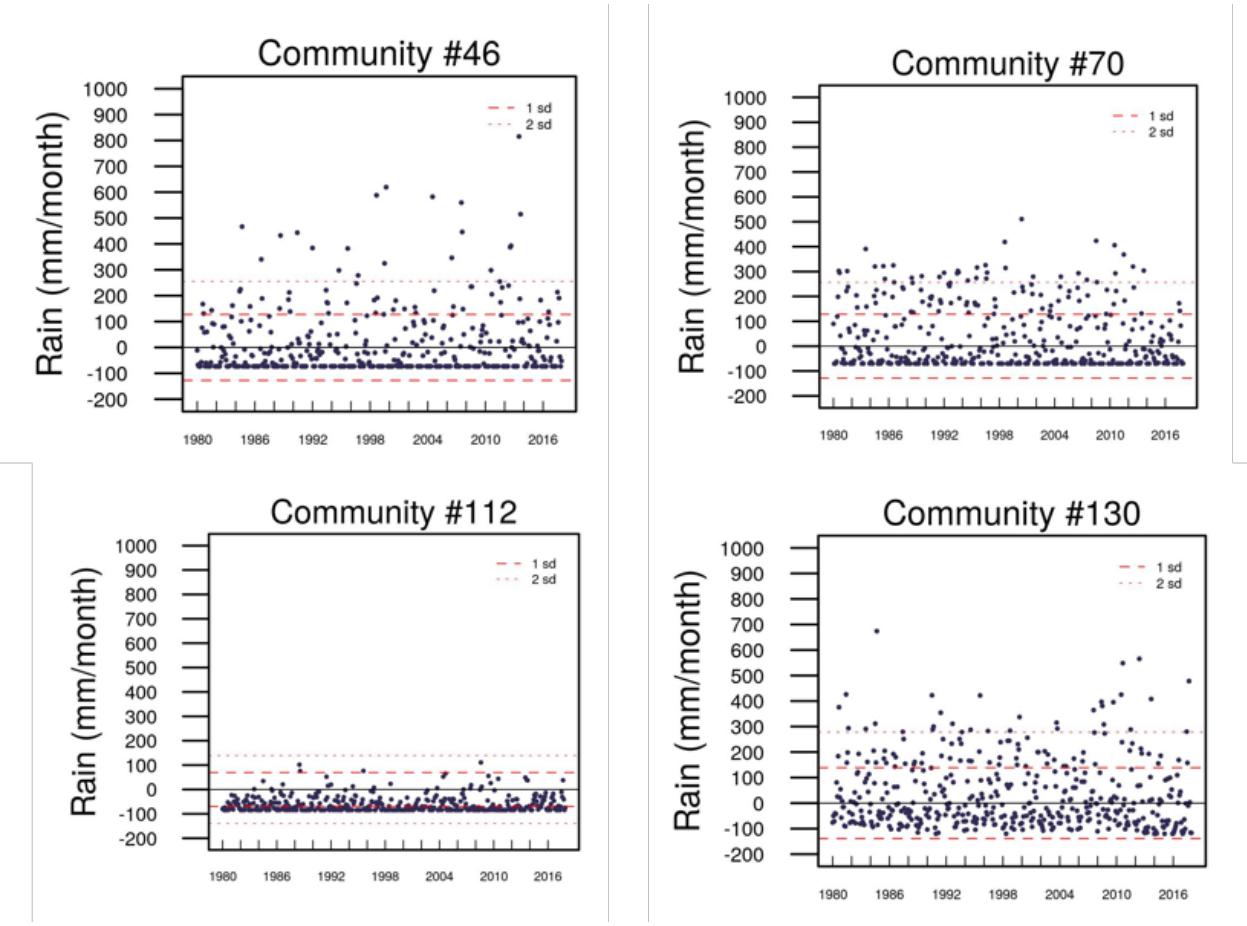


Figure 6. Average rainfall in each month (1980-2016) for selected four communities. The results are presented as deviations from the historical mean (1960-1979) in the state divided by the historical standard deviation. The red lines show the 1 (long dashes) and 2 (short dashes) standard deviation bounds set by historical observations.

Results

We estimate 2 random forest models to predict migration: one without precipitation variables and another one with only 2 precipitation lags. Since our weather data are collected from 1980 onwards and we want to explore the impact of weather shocks on the decision to migrate, we drop out all individuals who have migrated before 1985 and include precipitation lags before 1985. We add two precipitation lags, one in 1980 and 1983.

Since migration is a rare event (i.e. most of the people do not migrate), the overall accuracy for both models is very high (0.92 and 0.93 for each model). This occurs because the models

correctly classified people who decided to not migrate. In fact, the false positive rate in both models is approximately ~0.02, which means that both models rarely classified as non-migrant someone who has actually migrated.

This suggests that we focus instead in the precision and recall of the models. The precision of the model is given by the proportion of values correctly classified as migrants relative to the total number of predicted migrants (which can be either correctly or correctly classified). This will tell us how much of our predictions are actually relevant. The recall of a model is the proportion of values correctly classified as migrants relative to total number of actual migrants (who were either correctly or incorrectly classified). This value tells us how good our predictions are relative to the ground truth.

Figure 7 shows the precision-recall curve for both random forest models, with and without precipitation lags. This curve shows the trade-off between precision and recall, and reveals that high levels of precision come with a high cost of not correctly identifying the actual migrants. Both models seem to be fairly similar in terms of this trade-off, with the random forest with precipitation being slightly better for lower recall levels. In general, this suggests that adding two precipitation lags may not be enough information to correctly classify migrants. In fact, the area under curve (AUC) for both models are 0.32 and 0.33 for models without and with precipitation lags. The AUC ranges between 0 and 1, with a value of zero meaning that our model has terrible performance, misclassifying all values. These AUCs roughly suggests that the performance of the models is not very good.

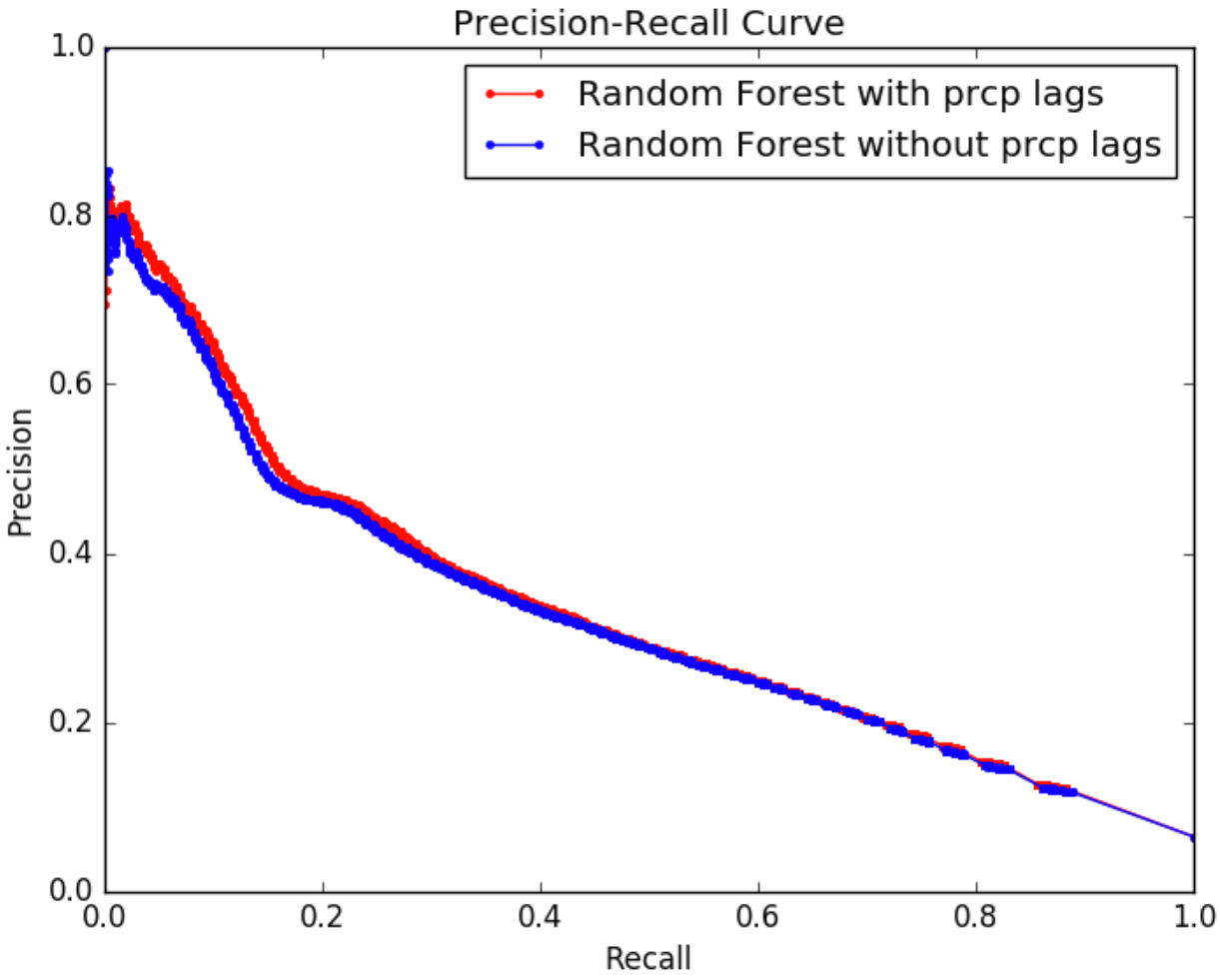
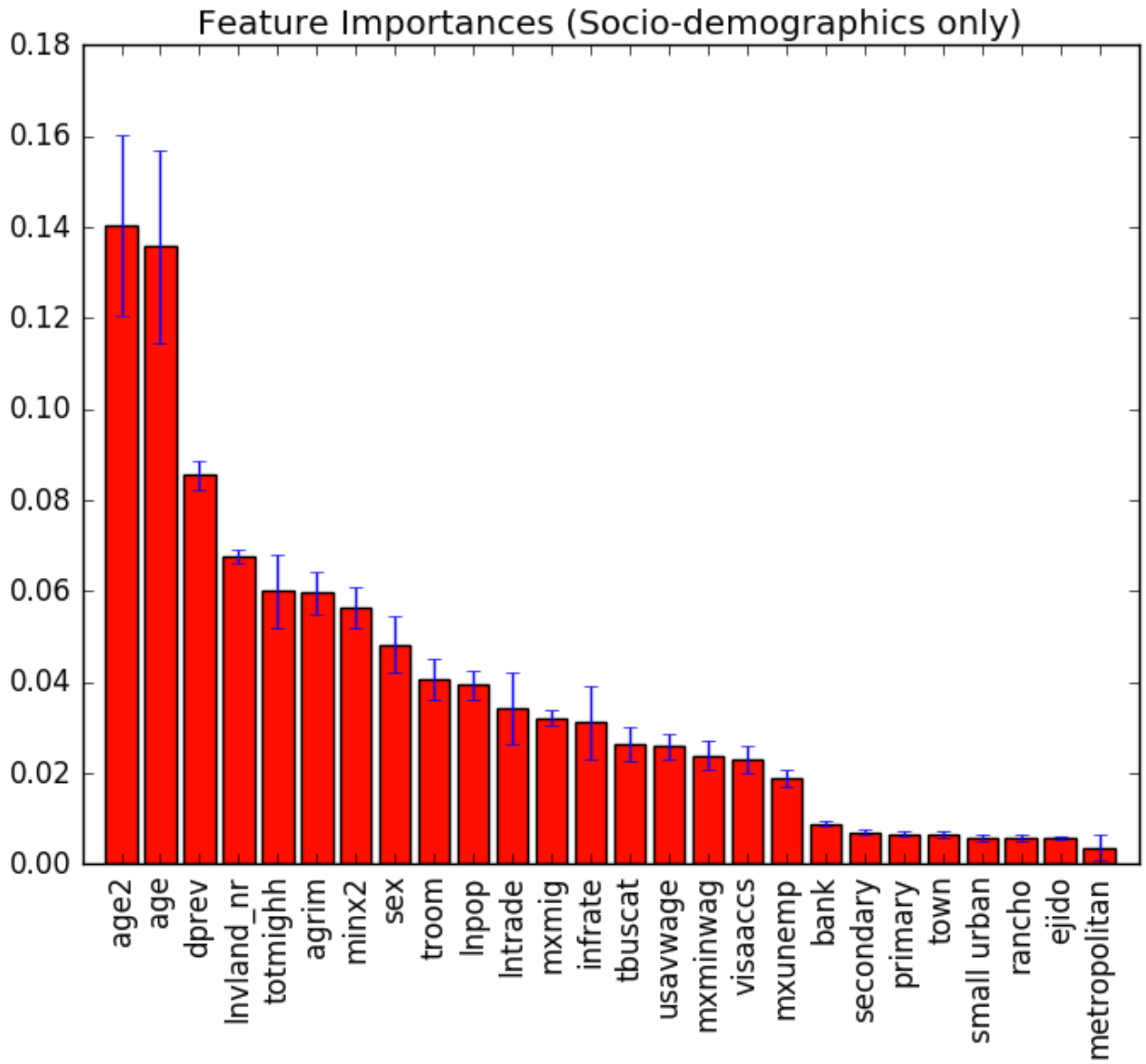
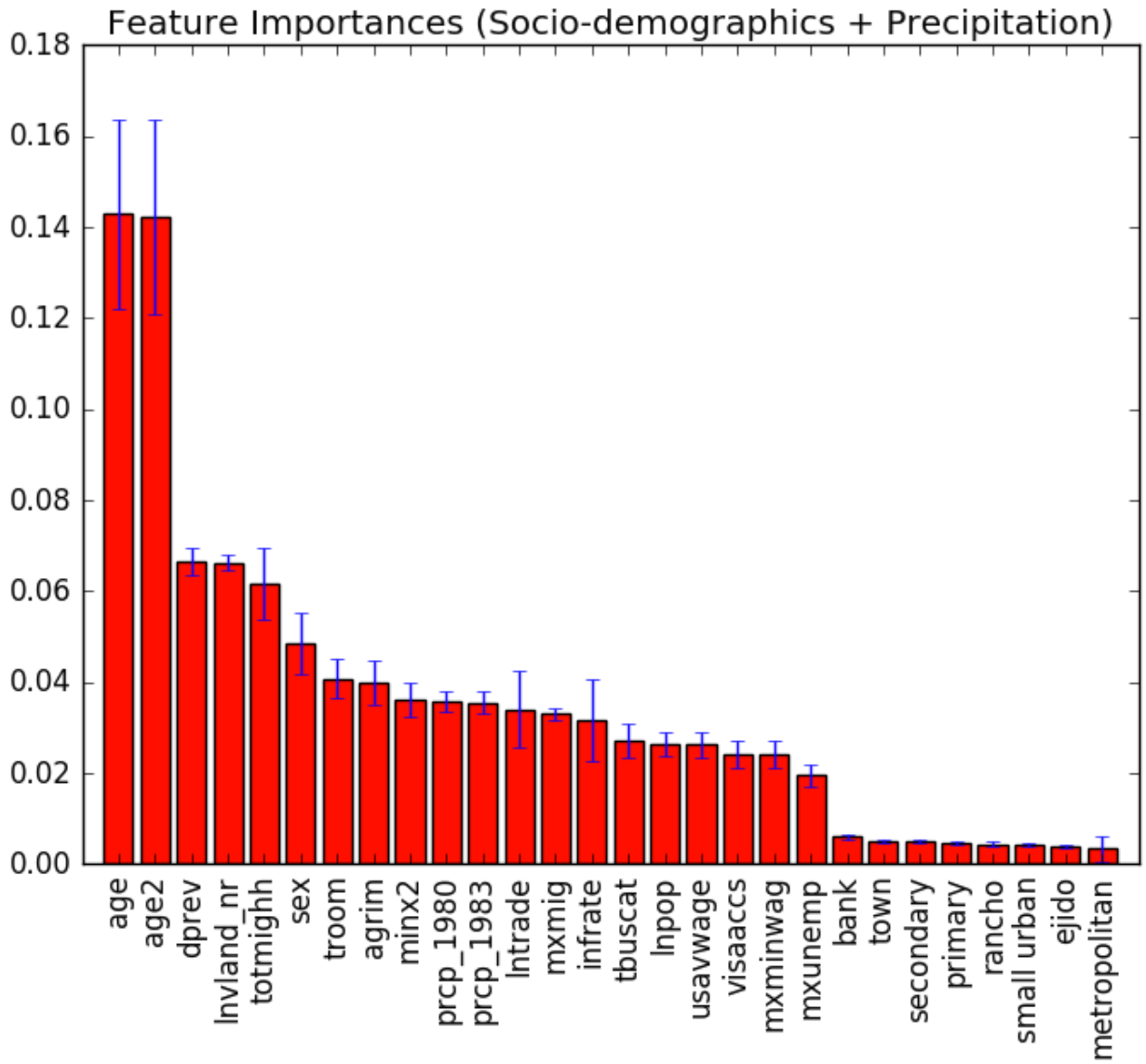


Figure 7. Precision and recall of random forest, with and without precipitation lags.

When we estimate our random forest models, it is also interesting to see how much each covariate contributes to the predictive performance of the models. The following figures show the relative importance of all covariates in predicting migration. Blue bars represent one standard deviation across all estimated trees (k=100). We observe that in a model with only individual and household socio-demographics and community-level covariates, the most important features are age, experience (or age squared), the prevalence of people in the community who have migrated, the value of the land, and the total number of people in the household who have migrated.



When we include two precipitation lags, we observe that the same variables mentioned before remain as the most important for predicting migration. However, our precipitation lags are not completely irrelevant and contribute roughly 4% to the predictive capacity of the model.



Discussion

Our analysis has built a predictive model of migration behavior in the Mexico-U.S. setting, and tested the accuracy of the predictions with out-of-sample data. The results suggest that our model cannot predict migration outcomes with great accuracy. We improve a slight improvement in our model performance with the inclusion of weather variations. There are several potential avenues to go from here, the most straight-forward being to include more weather information that can help us improve our predictions for migration.

Concerns on future climate change, and its potential impact on human mobility, certainly provide a crucial motive for this 'predictive' exercise. But, as social scientists, our interest is also to use these predictions to improve future research. To that end, our analysis will investigate the variability in the model's predictive accuracy across rural-urban communities, over time, and by age, education, and gender groups. By doing so, we will seek to understand what (unmeasured) factors might account for this variation, and design additional analysis (and data collection) to probe further.

We also plan to use simpler decision trees (rather than random forests that average over multiple trees) to increase the interpretability of our findings. These methods are increasingly used by economists, and offer a bridge to causal-inference approach in econometrics/statistics (Athey and Imbens 2017).

And, finally we are collecting additional data that could account for the linkages between weather and migration. We have recently secured detailed data on agricultural productivity in Mexico (annual, for each municipality). These data will allow us to create weather measures that are specific to the crops grown in each region. We are looking to include data on exports which can offer an intervening mechanism between weather and migration in urban regions as suggested in prior work (Jones and Olken 2010).

References

Athey S, Imbens GW. 2017. The State of Applied Econometrics: Causality and Policy Evaluation. *J. Econ. Perspect*. 31:3-32

Baldwin, A., 2017. Climate change, migration, and the crisis of humanism. *Climate change*, 8(3).

Black, R. 2001. *Environmental refugees: myth or reality*? United Nations High Commissioner for Refugees, Geneva, pp 1–19.

Black, R., Adger, W.N., Arnell, N.W., Dercon, S., Geddes, A. and Thomas, D., 2011. The effect of environmental change on human migration. *Global environmental change*, *21*, pp.S3-S11.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Castles, S., 2010. Understanding global migration: A social transformation perspective. *Journal of ethnic and migration studies*, *36*(10), pp.1565-1586.

Dell, M., Jones, B.F. and Olken, B.A., 2014. What do we learn from the weather? The new climate-economy literature. *Journal of Economic Literature*, *52*(3), pp.740-98.

Domingos P. 2015. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books

El-Hinnawi, E. 1985. *Environmental refugees*. United Nations Environment Programme, Nairobi, p 41.

Feng, S., A.B. Krueger, and M. Oppenheimer. 2010. "Linkages among climate change, crop yields and Mexico–US cross-border migration." *PNAS* 107.32: 14257-14262.

Garip, F. 2016. *On the Move: Changing Mechanisms of Mexico-U.S. Migration.* Princeton, NJ: Princeton University Press.

Gray, C., 2013. Human Migration in a Changing Climate. *Global Environmental Politics*, *13*(1), pp.128-132.

Homer-Dixon, T.F., 2010. *Environment, scarcity, and violence*. Princeton University Press.

Hunter, L.M., Murray, S. and Riosmena, F., 2013. Rainfall patterns and US migration from rural Mexico. *International Migration Review*, 47(4), pp.874-909.

Hunter, L.M., Luna, J.K. and Norton, R.M., 2015. Environmental dimensions of migration. *Annual Review of Sociology*, *41*, pp.377-397.

Jones, B. F., and B. A. Olken. 2010. “Climate Shocks and Exports.” *American Economic Review* 100 (2): 454–59.

Kavanagh, B. and Lonergan S. 1992. Environmental degradation, population displacement and global security. *Canadian Global Change Program*, Ottawa, p 55.

Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. 2015. Prediction Policy Problems. *Am. Econ. Rev*. 105:491-495

McLeman, R. and Smit, B., 2006. Migration as an adaptation to climate change. *Climatic change*, *76*(1-2), pp.31-53.

Molina, M. and F. Garip. (forthcoming) Machine learning for sociology. *Annual Reviews of Sociology.*

Muenchen RE. 2017. “Natural catastrophes 2016: Analyses, assessments, positions.” *Topic GEO.*

Mullainathan S, Spiess J. 2017. Machine Learning: An Applied Econometric Approach. J*. Econ. Perspect.* 31:87-106

Munshi, K. 2003. “Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market.” *Quarterly Journal of Economics* 118 (2): 549–99.

Myers, N. 1993. *Ultimate security: the environmental basis of political stability*. Norton, New York, p XI, 308.

Nawrotzki, R.J., Hunter, L.M., Runfola, D.M. and Riosmena, F., 2015. Climate change as a migration driver from rural and urban Mexico. *Environmental Research Letters*, *10*(11), p.114023.

Riosmena, F., Nawrotzki, R. and Hunter, L. 2018. Climate Migration at the Height and End of the Great Mexican Emigration Era. *Population and Development Review*.