**Revisiting forced migration: A machine learning perspective (EJPE, 2021)**

Summary: Machine learning to predict forced migration

Data: Panel data from 45 African countries for the years 1997–2017

Method: (ML) Random Forests:

* Useful for capturing nonlinear relationships and interactions between variables
* Helps us in identifying variables most important for predicting migration using several specifications, instead of the “effects” of those variables
* Explicitly allows for bias by making the model very rich, works best with a large number of predictors
* Doesn’t give us a parametric model but tells us which variables could be included in one
* No assumptions about the data generating process required to ensure consistency as both predictions and outcomes are observed (training/testing data)
* Builds multiple decision trees on a bootstrapped sample of observations
* Each decision tree is optimized to find observations that are homogeneous with respect to the outcome variable given a set of predictors
* The random forest is the average of the resulting trees
* A random sample of predictors is taken at each split to avoid averaging highly correlated trees
* 500 trees with 30–50% of variables sampled for each split of the trees
* Author uses LASSO regression model using random forest estimates (also useful for when multicollinearity in high-dimensional data) to select the 10 most important variables
* Advantage: Larger reduction in variance, more reliable than a single tree
* Disadvantage: Not easily interpretable (like regressions), allow you to rank variables by importance

**How to model the weather-migration link: a machine-learning approach to variable selection in the Mexico-U.S. context (Journal of Ethnic and Migration Studies, 2023)**

Tradeoff between model complexity and generalizability

Summary: Machine learning to predict migration using weather variables

Data: Mexican Migration Project (MMP) (130,000 individuals, 1980-2016)

Method: Random Forests:

* Classification of individuals as migrants or non-migrants using 36 weather variables
* More complex models (more splits in variable space) may not generalize well to the validation or test data, so they use a regularizer to decide depth of trees
* Regularizing less means better in-sample performance but potentially worse out of sample performance (overfitting)
* Fixed ratio of 1:10 between migrants and non-migrants in test and training set
* Penalize false negatives (classified as “no migration” but person migrated) more than false positives using weights
* Each model is estimated using k-fold cross-validation (with k=5), where different model choices estimated on k-1 folds are then evaluated out of sample on the remaining fold
* Goal again is to identify the most important predictors

**A Machine Learning Approach to Modeling Human Migration (Comp Sci)**

* Estimate a function , which takes the features of zone and , as well as the joint features between them, as input, and outputs the estimated number of migrants that travel from to .
* Two models: “Extreme” gradient boosting regression (XGBoost), Artificial neural network model (ANN)
* XGBoost: Based on gradient boosting trees, gives a ranking of relative feature importance
* XGBoost hyperparameter tuning for maximum tree depth, number of estimators, and learning rate
* ANN hyperparameter tuning for loss function, number of layers, layer width, number of training epochs, and training mini-batch size
* Migration data is heavily zero-inflated, where in any given year, most pairs of zones do not have any migrants traveling between them, i.e., for most pairs
* Undersample “negative” samples between pairs of zones for which there are no observed migrations to address this, offset it by including a hyperparameter
* ANN with extended features performs the best (more features are good)