NOWCASTING MIGRANT STOCK USING FACEBOOK DATA

Shreyas Gandlur
sgandlur

Jordan Klein
jdklein

Dora Zhao
dorothyz

December 8, 2020

ABSTRACT

Previous work has demonstrated ways in which social media data, specifically Facebook data, can be applied to the problem of predicting international migrant stock. In this paper, we evaluate the utility of using said social media data for predicting international migration by comparing a simple autoregressive baseline with a standalone Facebook model and a combined model that uses both the autoregressive and Facebook data. In addition, we analyze how different machine learning models can perform on this task relative to a log-transformed linear regression model to, perhaps surprisingly, find that in most cases the linear regression model has the highest predictive accuracy. Finally, we conclude by examining the predictive accuracy when varying the quality of the ground-truth data.

1 Introduction

1.1 Background

Estimating international migrant stock, or the number of people living in countries other than those of their birth, is of great importance to demographers and policymakers. To meet these needs, the United Nations Population Division publishes estimates of global international migrant stock biennially [1]. However, there is a great degree of variance between countries in quality of, methods used for, and frequency with which data is collected to generate these estimates.

Simultaneously, digital technologies, such as social media platforms, are ever more ubiquitous, even in parts of the world without high quality migration data. The data these technologies provide has been used by demographers in recent years to try to fill in some of these gaps in studying migration [2, 3, 4, 5, 6]. In particular, data from Facebook (the most widely used social media platform in the world [7]) has been used to estimate both migrant stock and migrant flow (i.e., using changes in migrant stock to measure the rate at which people move from one country to another). Prior work has computed such estimates both in the United States [8, 9] and internationally [10]. In these studies, Facebook is often presented as a "biased census" whose degree of bias varies by country, age, sex, and other demographic characteristics.

1.2 Previous work

In 2017, Zagheni et al. [9] used Facebook's advertising platform to predict the number of foreign born individuals, along with countries of origin, in each of the 50 U.S. states. Using American Community Survey (ACS) data [11] as their "ground truth," they created linear regression models to predict the number of immigrants by country of origin in each state, using the number of Facebook users who were expats from these countries in each state as predictors. They trained two models: a naive model that did not account for biases by age and country of origin and a calibrated model adjusting for age and country of origin bias. While they found that their calibrated model had superior predictive accuracy over their naive model, they acknowledged several key limits inherent in their approach: biases (other than age and country of origin) that they did not adjust for; lack of transparency in how Facebook classifies users as expats and other categories; only using linear regression as a statistical method for approaching this problem; and, most importantly, that ACS data is not truly a "ground truth," but is itself only an estimate of the "ground truth."

In 2019, Spyratos et al. [10] extended this approach to try to measure international migrant flows. They leveraged changes over time in the number of Facebook users classified as expats in 55 countries to perform these estimates. They adjusted for Facebook's sampling bias by age, sex, and country of current and previous residence using migration statistics from official sources: ACS data [11] for the US; Eurostat data [12], consisting of migration data based on European Union population registers, for EU countries; and United Nations [1] and OECD data [13], consisting of generally higher quality migration data from more developed countries and lower quality data from less developed countries, for other countries. These sources are then used to check accuracy of adjusted estimates. While ACS and Eurostat data disaggregates migrants by age, sex, and country of origin, UN data does not and OECD data only does for 2010/2011. Consequently, Spyratos et al. obtain estimates of migrant stock by country of origin that are relatively close to official estimates for the US and EU countries. On the other hand, these estimates are less accurate for countries outside the US and EU, especially estimates for developing countries with lower quality data. While Spyratos et al. conclude that there is potential in using social media-derived migration data to generate timely migration estimates and predictions, they caution that this data should always be validated using official migration statistics and not be regarded as a replacement for such statistics.

While Zagheni [9] and Spyratos et al. [10] provide compelling cases for the use of social media data in predicting migration, they do not analyze the added value social media data provides over simpler models. Goel et al. [14] present an approach for systematically evaluating the added value of digital trace data across a variety of prediction problems. They evaluate the utility of using search data to make predictions regarding consumer behavior (e.g. box office performance, video game sales) and health outcomes (e.g. flu). Using linear regression, they compare the predictive performance of three models: those just using search data, baseline autoregressive models using previous values of the outcome of interest, and combined models using both search data and autoregressive terms. Goel et al. found that search data alone had greater predictive power than a simple autoregressive model in only two of the five tasks, and that the combined search and autoregressive models perform the best overall.

We synthesize these previous approaches to explore the limits in using social media data for predicting migrant stock on a global scale, across countries with highly varying methods for generating migrant population estimates and with varying quality of migration data. We compare the performance of prediction models using Facebook data to autoregressive baselines and analyze whether adjusting Facebook data for sampling bias improves such data's added value in prediction. Furthermore, we go beyond linear regression models to make predictions, evaluating whether machine learning methods, the use of which is still relatively nascent in the field of demography [15, 16], improve performance in demographic prediction problems.

1.3 Research aims

We aim to evaluate the potential for and the limits to using social media data in predicting migration. Towards this end, we present a methodology for attempting to nowcast, or predict present migrant stock populations in countries worldwide, by training models on a sample of countries and measuring their out-of-sample predictive accuracy, using UN migrant stock estimates as our "ground truth." We first explore whether leveraging Facebook data provides added value in improving the accuracy of these predictions, evaluating whether models using Facebook data have superior predictive accuracy compared to baseline models using previous UN migrant stock estimates. We then compare the predictive accuracy of combined models using both Facebook data and previous UN migrant stock estimates to models using either of these data sources alone. Third, we evaluate which machine learning methods have the best predictive accuracy for this task and consider whether correcting for Facebook's sampling bias with respect to age and sex-specific rates of penetration improves its predictive accuracy. Finally, we explore how the countries we include in our sample with respect to the quality of their migration data influences our models' predictive accuracy, comparing models trained and tested on random samples of countries without respect to migration data quality, models trained and tested on samples restricted to just countries with higher quality data, and models trained and tested on samples restricted to countries with lower quality data.

2 Methods

2.1 Data Collection

2.1.1 Facebook

Using the methods introduced in Zagheni et al. [9], we collect our training data using the Facebook for Business Marketing API [17]– a freely accessible resource for anyone with a Facebook account. Using the Marketing API, we query Facebook's reach estimates for individuals labeled as expats (or the behavior "Expats (All)" in the API) living in a specific country. As described by Facebook, this returns an estimate of all "people living outside their home country." In total, Facebook provides reach estimates for 246 countries or territories; however, we use data from only 192 countries in prediction, after only considering countries that were present in both the Facebook and United Nations (UN) datasets.

We also remove countries for which the total population was less than the minimum reach estimate of 1000 provided by the API.

Furthermore, we collect more fine-grained estimates for expats based on characteristics, namely age and sex. Our age groups are bucketed into five year intervals, replicating the age-group breakdowns found in the UN data. Given the limitations put in place by the Facebook API, our youngest age group starts at 15 (e.g. 15-19 age group), and the oldest age group we capture is 65+. We also collect demographic information about Facebook users in each country as a whole.

When computing Facebook penetration rates (equation 7), or the ratio of Facebook users in a country to that country's population, there were instances in which the Facebook penetration exceeded 1. We note that penetration rates greater than one occurred mostly for the younger age categories (e.g. 35 and under) in which Facebook usage is generally higher. For our study, we clip the penetration rates to a maximum value of one.

In total, we identify three limitations to using Facebook reach estimate data. First, Facebook's algorithm for determining the expat status of individuals is proprietary. While Facebook's researchers have published articles that can be used to infer how this category is calculated [18], we treat their algorithm as a black box. Second, Facebook provides coarse reach estimates. They are rounded to the nearest 1000, with a minimum estimate of 1000. Finally, a key limitation is that only current data, not retrospective data, can be queried.

2.1.2 United Nations Population Division

The United Nations Population Division of the Department of Economic and Social Affairs produces the foremost global migrant stock estimates on a biennial basis. They estimate the migrant stock of every country in the world in aggregate, by age and sex, and by country of origin and sex, but not by all three [1]. They use the best and most recent data available from each country's national statistical office, including population registers, censuses, and/or nationally representative surveys, with standard demographic estimation methods used to fill in any gaps [19]. Regional imputation is used for countries with missing data.

The UN Population Division notes how each country defines and measures who qualifies as an international migrant, whether based on country of birth, citizenship, or both, and whether refugees and asylum seekers are included in official migrant counts [19]. Defining migrants based on country of citizenship suffers from key shortcomings, namely the potential for misclassifying individuals born to non-citizen parents in countries without birthright citizenship and individuals who naturalize in their country of residence as migrants and non-migrants, respectively. Consequently, country of birth is used whenever available, with country of citizenship only used in countries that collect no data on country of birth. Inclusion of refugees and asylum seekers in official migrant counts is preferred, but in countries that do not include them, refugee/asylum seeker estimates are added to total migrant stock estimates. The UN further classifies countries into "more developed regions" where migration data quality is generally higher, including the US and EU which publish annual high quality migration statistics through the ACS and Eurostat, respectively, and "less developed regions" where migration data quality is generally lower, including countries in which UN estimates rely on much sparser data sources and/or imputation. The UN has most recently published migrant stock estimates for 2019, 2017, and 2015, and plan to publish estimates for 2021 in the future. They also publish annual estimates of total population for every country in the world by age group and sex, from which we use their 2020 estimates. The UN does not publish an accompanying full methodology on how these estimates are made.

2.2 Making Predictions

We define $\mathbf{foreign\_born}_{C,t}$ to represent the total migrant stock over age 15, in a country C, in year t, as estimated by the UN [1]. We endeavor to nowcast $\mathbf{foreign\_born}_{C,t}$, for each country C and at $t = 2019$, using the following set of 6 models. We present each model below in terms of a linear regression equation. The first three models do not correct for any biases:

1. **Autoregressive baseline**

$$\mathbf{foreign\_born}_{C,t} = \beta_0 + \beta_1 \mathbf{foreign\_born}_{C,t-2} + \beta_2 \mathbf{foreign\_born}_{C,t-4} + \epsilon_C \tag{1}$$

   This is a baseline model of UN migrant stock estimates in 2019 as a function of migrant stock estimates in 2017 and 2015 against which models using Facebook data are compared.

2. **Facebook naive**

$$\mathbf{foreign\_born}_{C,t} = \beta_0 + \beta_1 \mathbf{FB\_expats}_C + \epsilon_C \tag{2}$$

   where $\mathbf{FB\_expats}$ is the number of Facebook expats.

   This is a simple model of migrant stock using just Facebook expat data.

3. **Autoregressive baseline and Facebook naive combined**

$$\mathbf{foreign\_born}_{C,t} = \beta_0 + \beta_1 \mathbf{FB\_expats}_C + \beta_2 \mathbf{foreign\_born}_{C,t-2} + \beta_3 \mathbf{foreign\_born}_{C,t-4} + \epsilon_C \quad (3)$$

   This model combines Facebook expat data ($\mathbf{FB\_expats}_C$) with autoregressive terms: migrant stock estimates in 2017 ($\mathbf{foreign\_born}_{C,t-2}$) and 2015 ($\mathbf{foreign\_born}_{C,t-4}$).

In the next three models, we correct for sampling bias by age and sex. We model the migrant stock in country C, in year t, in age-sex group z, with the outcome $\mathbf{foreign\_born}^z_{C,t}$. There are 22 different categories for z: the eleven age intervals $\{[15, 20) \cup [20, 25) \cup \cdots \cup [60, 65) \cup [65, \infty)\}$ for both males and females. The total migrant stock over age 15 in each country, $\mathbf{foreign\_born}_{C,t}$, is then simply the following equation:

$$\mathbf{foreign\_born}_{C,t} = \sum_{z=1}^{22} \mathbf{foreign\_born}^z_{C,t} \quad (4)$$

We now provide the models for the age-sex adjusted outcomes, $\mathbf{foreign\_born}^z_{C,t}$, as follows:

4. **Autoregressive baseline, age-sex adjusted**

$$\mathbf{foreign\_born}^z_{C,t} = \beta_0 + \beta_1 I^z + \beta_2 \mathbf{foreign\_born}^z_{C,t-2} + \beta_3 \mathbf{foreign\_born}^z_{C,t-4} + \epsilon^z_C \quad (5)$$

   where I^z is an indicator variable for age-sex group z.

   This is the baseline autoregressive model of migrant stock in 2019 as a function of migrant stock in 2017 and 2015, against which we compare the age-sex corrected models that use Facebook data.

5. **Facebook, age-sex adjusted**

$$\mathbf{foreign\_born}^z_{C,t} = \beta_0 + \beta_1 \frac{\mathbf{FB\_expats}^z_C}{\mathbf{FB\_penetration}^z_C} + \beta_2 I^z + \epsilon^z_C \quad (6)$$

   where $\mathbf{FB\_penetration}^z_C$ is Facebook penetration for age-sex group z and defined as

$$\mathbf{FB\_penetration}^z_C = \frac{\mathbf{Facebook\_users}^z_C}{\mathbf{Total\_population}^z_C} \quad (7)$$

   This model represents migrant stock as a function of Facebook expats in each age-sex group, adjusted by Facebook penetration in that group.

6. **Autoregressive with Facebook combined, age-sex adjusted**

$$\mathbf{foreign\_born}^z_{C,t} = \beta_0 + \beta_1 \frac{\mathbf{FB\_expats}^z_C}{\mathbf{FB\_penetration}^z_C} + \beta_2 I^z + \beta_3 \mathbf{foreign\_born}^z_{C,t-2} + \beta_4 \mathbf{foreign\_born}^z_{C,t-4} + \epsilon^z_C \quad (8)$$

   This model combines age-sex corrected Facebook expat data and autoregressive terms.

We note that since our outcome, $\mathbf{foreign\_born}_{C,t}$, is migrant stock in 2019 and the Facebook data we use are from 2020, we are in effect predicting the past rather than nowcasting. If it were available to us, it would be more appropriate to use Facebook data from mid-2019 to nowcast 2019 migrant stock.

2.3 Modeling Techniques

We use linear regression as our initial prediction method, employing it to generate predictions on each model given in Section 2.2. We log-transform all continuous predictors and the outcome for each model. Logarithmic transforms are generally a good approach when dealing with heteroscedastic data, as non-linear transforms may help linearize relationships. Additionally, prior work suggests that transformations are useful in the context of migrant stock data [9].

We employ two machine learning methods: random forests and Extreme Gradient Boosting (XGBoost) [20]. Random forests are a widely available method that have, in other domains, shown increased predictive power over traditional statistical regression techniques [21] [22]. XGBoost is a boosting algorithm that has achieved state-of-the-art results in other domains [20]. Given the relatively small size of our dataset, it is not appropriate to use deep learning methods for this problem. Thus, XGBoost is a good representation of the best current machine learning method for the task of predicting migrant stock.

We initially tuned hyperparameters for these models by performing a randomized search. Ultimately, we use a maximum depth of 10 for both random forest and XGBoost models. We use 100 estimators, or number of trees, for random forest and 300 estimators for XGBoost. In addition, we specify a learning rate of 0.01 and a gamma of 1 for XGBoost. We found that increasing estimators or modifying other hyperparameters resulted in minimal improvement in predictivity, with sharply increased training time.

For the machine learning methods, we take a similar approach as we did with linear regression, by training incrementally on different sets of predictors as specified in Section 2.2. This approach allows us to systematically evaluate how using Facebook data affects the predictive power of our models. We keep the hyperparameters constant across the six different models as we manipulate the predictors chosen.

2.4 Sampling Countries

The UN's division of countries into "more developed regions" where migration data quality is generally higher, and "less developed regions" where migration data quality is generally lower [19]. We compare the predictive abilities of our models using the following splits for constructing our training and test sets:

1. Randomly sample from all countries for training and test sets.
2. Randomly sample from only more developed countries for training and test sets.
3. Randomly sample from only less developed countries for training and test sets.

2.5 Measuring Predictive Accuracy

We measure the predictive accuracy of our models using mean absolute percentage error (MAPE). MAPE provides an interpretable evaluation metric useful for relative comparison, which is desirable for our study. MAPE suffers from weaknesses in datasets with small or zero values; this is because errors on small values will skew MAPE or cause divide-by-zero errors. This is not an issue in our study – we are dealing with large population numbers, and the smallest value in our outcome data is 1881.

Additionally, prior work in migrant stock data prediction [9] has used MAPE as an evaluation metric. An added benefit of using MAPE is that it allows for comparison across studies.

k-fold cross validation is widely used to assess predictive accuracy of statistical and machine learning models. We use 5-fold cross validation to measure predictive accuracy of our models. We randomly partition the set of countries into five subsets. Data from one subset of countries is then used as a test set, with data from the other four subsets of countries used for training. Models are trained on the training set and MAPEs are computed on the test set. This procedure was repeated with each subset of countries used as the test set. After these five iterations were complete, MAPEs were averaged to compute the final errors. Note that, for a given split, the same subsets were used for each prediction model and method. This allows for comparisons across these, as the same data was used for training and testing them.

3 Results

We first evaluate the effect that adding Facebook data has on migrant stock prediction performance, comparing the performances of the baseline (equation 1), the Facebook naive (equation 2), and the combined models (equation 3), heretofore referred to as our non-adjusted models, on random samples of all countries. As seen in Figure 1, models using only Facebook data have significantly higher MAPEs than the autoregressive baseline, regardless of the modeling method used. Adding Facebook data to the autoregressive model barely improves predictive performance for random forests only (from a MAPE of 19.90% to 19.85%), while yielding slightly a slightly worse predictive performance for both XGBoost and the linear regression, by about 2%.

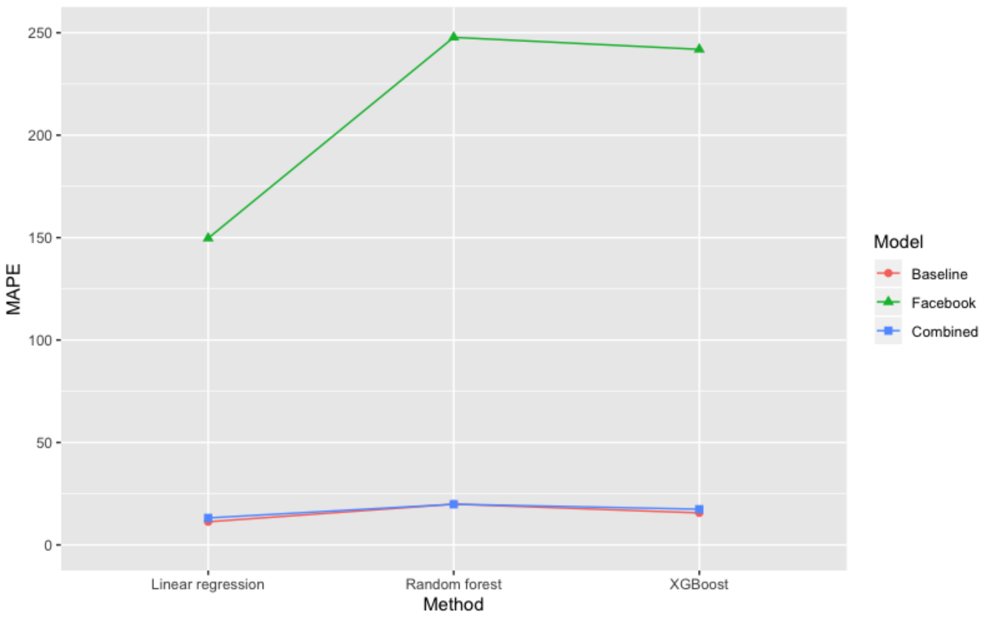

Figure 1: MAPEs for out-of-sample prediction of 2019 UN migrant stock estimates using linear regression, random forest, and XGBoost, on a sample of all countries without adjusting for age and sex. The models compared are autoregressive baseline (equation 1), Facebook naive (equation 2), and autoregressive baseline and Facebook naive combined (equation 3). Lower MAPEs indicate superior predictive performance.

Turning to our age-sex adjusted models, we observe that models using Facebook data only perform worse when age-sex adjusted (equation 6) than when not (equation 2), regardless of the modeling method used (Fig. 2). Models that use Facebook data and autoregressive terms however improve when age-sex adjusted (Fig. 3). We can also observe that when adjusted by age and sex, combined models have superior performance to baseline models that do not use Facebook data, specifically when using linear regression or XGBoost. However, the single best performing model trained and tested on a random sample of countries is the linear regression baseline model.

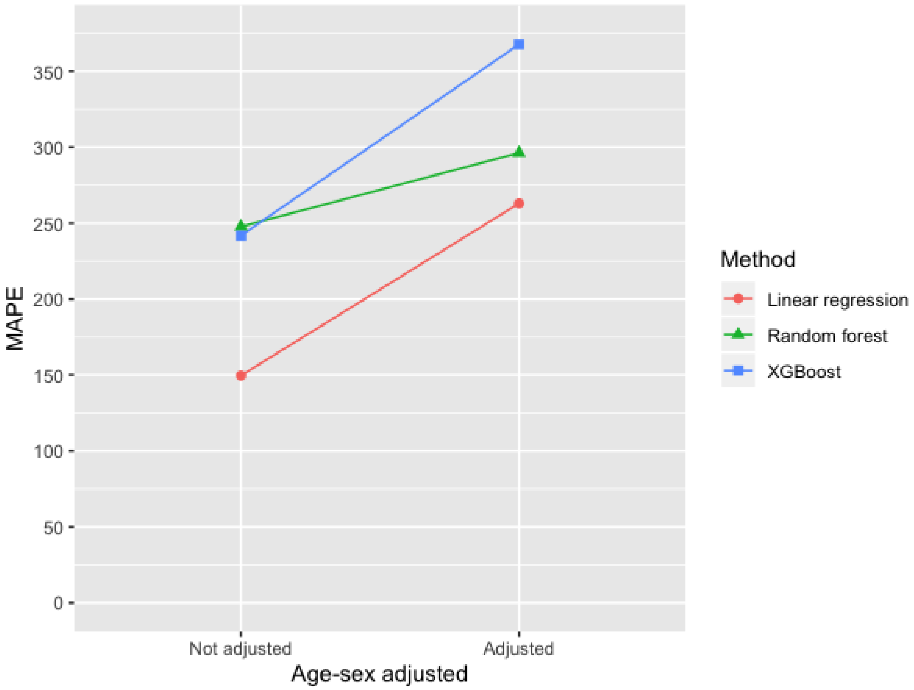

Figure 2: MAPEs for out-of-sample prediction of 2019 UN migrant stock estimates on a sample of all countries with a naive model of Facebook data (equation 2) and a model of age and sex-adjusted Facebook data (equation 6).

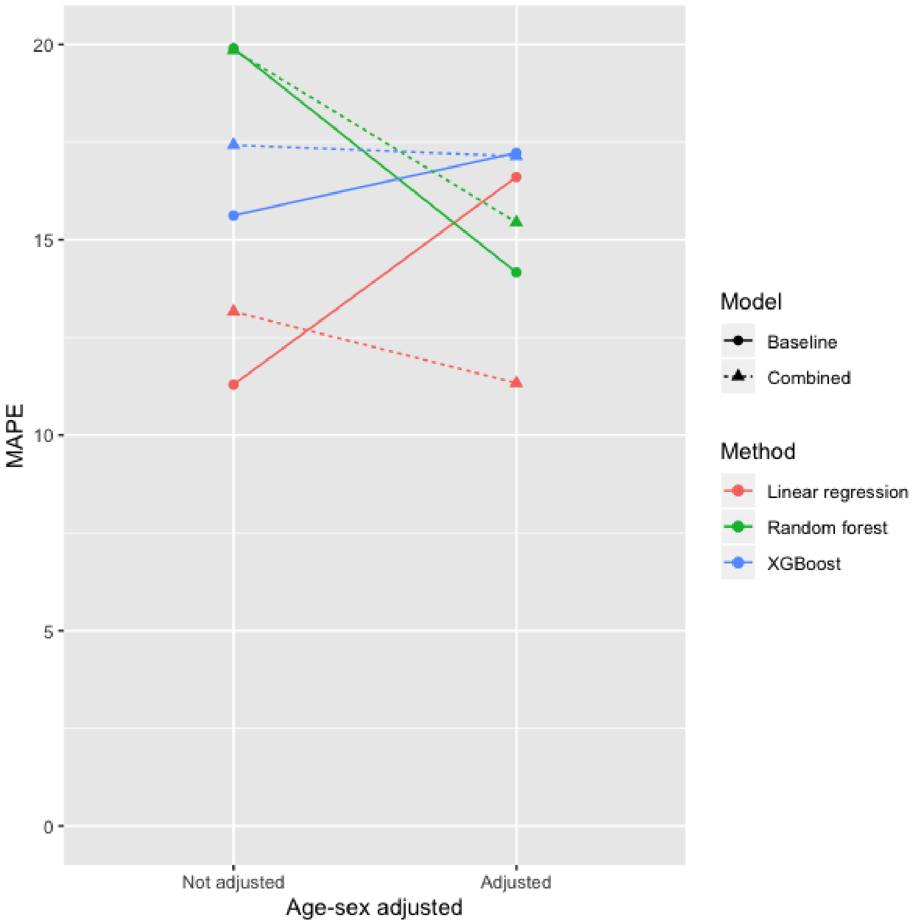

Figure 3: MAPEs for out-of-sample migrant stock prediction on a sample of all countries of autoregressive and combined autoregressive-Facebook models, age-sex adjusted (equations 5, 8) and not age-sex adjusted (equations 1, 3).

When comparing the performance of our three modeling techniques, we observe that linear regression performs best overall across every prediction model, including both age-sex adjusted and non-adjusted models (Figs. 1, 2, 3). XGBoost performs better than random forests on non-adjusted models, while on age-sex adjusted models random forests is superior.

Specifically for the linear regression model, the combined baseline and social media data (equation 8) improves the performance as compared to just the autoregressive baseline (equation 5) for all three data splits. Again, looking at the performance on the random split, we see that the MAPE improves from 32.1% for the autoregressive baseline to 19.3% with the combined data (see Figure 3). However, for the random forest, training on the combined age-sex adjusted data actually leads to higher MAPE than on just the baseline.

We can also look at the performance of the models that include only the total migrant populations for 2015 and 2017 (i.e. non-adjusted) versus those that include age-sex subgroup populations (i.e. adjusted). Interestingly, we found that the adjusted models do not outperform the non-adjusted models. This is particularly pronounced when we compared the models that only use the Facebook data (e.g. Facebook naive equation 2 and Facebook, age-sex adjusted equation 5). As seen in Figure 2, all three models actually end up performing worse when using the age-sex adjusted data as opposed to the non-adjusted. We have a similar result when looking at the autoregressive models, with the exception of random forest which does improve with the adjusted age-sex data. However, if we consider the combined data, we find that all three regression techniques using the adjusted data outperform their non-adjusted counterparts, although the difference is slight for XGBoost.

When analyzing our results separately by models trained and tested on random samples of all countries, samples restricted to more developed countries, and samples restricted to less developed countries, we find that the linear regression performs better than the machine-learning methods over all country splits for the non-adjusted models. As seen in Figure 4, at its best, the linear regression method is able to achieve a MAPE of 7.11% when using the combined autoregressive and Facebook data (equation 3) to predict the migrant stock of highly developed countries. For the age-sex adjusted models, we find that random forests generally performs better when only using the autoregressive model. However, when we train on either only the normalized Facebook data or the combined data, linear regression again outperforms or performs in-line with the machine learning methods (Fig. 4, Appendix Fig. 1).

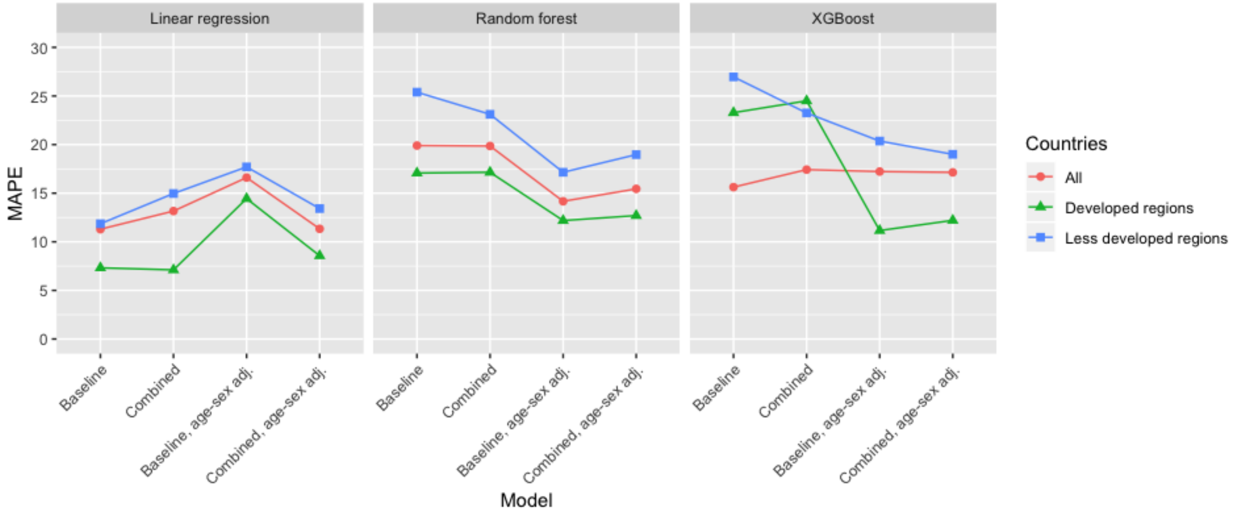

Figure 4: MAPEs for out-of-sample migrant stock prediction on samples of all countries, countries in "developed regions", and countries in "less developed regions" using autoregressive (equation 1), combined autoregressive-Facebook (equation 3), autoregressive age-sex-adjusted (equation 5), and combined autoregressive-Facebook age-sex-adjusted (equation 8) models.

Finally, when looking more closely at the effects of country development level and migration data quality on our models' predictive accuracy, we observe that training and testing on high development countries leads to the lowest MAPEs, training and testing on less developed countries yields the highest MAPEs, and training and testing on all countries yields MAPEs between these two extremes (Fig. 4). This is in line with our expectations that more developed countries have higher quality migration data that deviate less from the "ground truth". The only exceptions to this finding are for the baseline and combined models of XGBoost which actually have lower error rates when training across all countries. Furthermore, as seen in Figure 4 across all models and regression techniques, we get a higher accuracy when we randomly sample from all countries compared to when we sample from only less developed regions. In general, we have the lowest predictive accuracy when training and testing on the less developed regions.

4 Discussion

As we anticipated, Facebook expat data used alone did not produce a benefit in prediction, as seen in Figure 1. Across the different models and data splits, though, we notice that the combined data does not consistently improve the predictive accuracy of our models either. In the case of linear regression models, age-sex adjusted combined data provided significant benefit compared to the age-sex adjusted baseline, but even these predictions were comparable to the non-adjusted baseline. Furthermore, as seen in Figure 4, looking at combined data did not produce much of an effect in our other machine learning models.

We generally found, as expected, that correcting Facebook data for age-sex biases does have a positive impact on prediction. Looking at Figure 4, when comparing models trained on combined data to models training on age-sex adjusted combined data, adjusting Facebook data for age-sex outcomes almost uniformly had a positive impact on prediction. However, this does not mean that training on age-sex adjusted combined data is uniformly the best strategy. In the case of linear regression, combined age-sex adjusted data was comparable to the non-adjusted original baseline, whereas in the case of machine learning models, combined age-sex adjusted data was comparable to our age-sex autoregression baseline.

We also found that models training and testing on data from "developed regions" performed better than those testing on the entire dataset or those testing on data only from "less developed regions." We do not have enough information to conclude the exact causal mechanism for this: this could be a result of higher quality data, greater correlation between migrant populations in "developed regions" making prediction easier, or some other mechanism altogether.

In addition, contrary to our initial expectations, linear regression models generally perform comparably to or better than the machine learning methods we chose. Moreover, the MAPEs that our linear regression models were able to achieve were generally quite low (consistently less than 20% for baseline and combined models), indicating that the method has a rather high predictive accuracy. Given that we compare the linear regression to XGBoost, which has achieved state-of-the-art results in many other domains, it does not appear to be caused by the machine learning models we chose. While the values of the MAPEs could differ with further hyperparameter tuning (though we find this unlikely, as MAPEs did not change significantly with other hyperparameters) , we believe that our general claim that machine learning models do not provide better predictive accuracy over linear regression, for the predictors we used, holds. This perhaps suggests that international migrant stock data might genuinely be linear once log-transformed, in which case we would expect linear regression to outperform machine learning methods.

Our results, overall, draw skepticism upon the predictive ability of Facebook data for migrant stock prediction. What we found does not support the conclusion that Facebook data increases predictive ability beyond that of a simple autoregressive baseline. Nonetheless, Facebook data may still have some usefulness, given its timeliness and availability; for example, we did not study the predictive ability of Facebook data to measure migrant *flows* rather than total migrant populations. It is possible that Facebook data is useful in that setting, since it would be better placed, e.g., to react to changes in flows because of local or global shocks. Future work should continue

5 Conclusion

In this study, we address three main questions related to predicting present migrant stock on an international basis. First, we address the question of whether using Facebook data provides additional predictive power for the task at hand. To accomplish this, we collect Facebook reach estimates using the Marketing API for expat populations in different countries around the world. We then compare the MAPEs of three different models: a simple autoregressive baseline (equations 1 and 5), a combined Facebook data and autoregressive model (equations 3 and 8), and a Facebook-only model (equations 2 and 6). We find that Facebook data, generally, did not increase predictive ability beyond a simple autoregressive baseline. However, we do not conclude that Facebook data categorically lacks utility and future work should continue exploring its use in predicting migrant *flows*, instead of just migrant *stock*.

We also compare three different regression techniques to better understand whether machine learning methods can provide higher predictive accuracy for the migrant prediction task as compared to a linear regression. In total, we use a log-transformed linear regression as a baseline in addition to random forest and XGBoost. We find that the machine learning models do not provide improvement, and somewhat unexpectedly, generally perform worse than linear regression models.

Last, we explore the effects of training data quality on predictive accuracy. We address this question by introducing three data splits: one that includes all of the countries and then two others that are split into "developed" and "less developed" based on the UN migrant stock classification. The development level is used as a proxy for the data quality. Here, we find that the predictive accuracy is highest when training and predicting on "developed" countries. We are unable to conclude whether this is caused by data quality specifically, as this could have been caused by other correlations as well. However, this suggests future exploration would be useful.

As an extension, we are planning to evaluate our prediction methods on the UN's release of international migrant stock data in 2021. At the moment, we have prepared two models. The first is akin to the autoregressive baseline (equation 1) except we now use $t = 2021$ and are subsequently using 2019 and 2017 migrant totals rather than 2017 and 2015. The second model is the autoregressive baseline, age-sex adjusted model (equation 5). Again, we are setting $t = 2021$ and updating the migrant subgroup populations as well. We will include the Facebook-only and combined models we proposed as well but will pull the real-time Facebook reach estimates in 2021. Finally, we will preregister our models and methodology ahead of the UN's data release.

6 Appendix

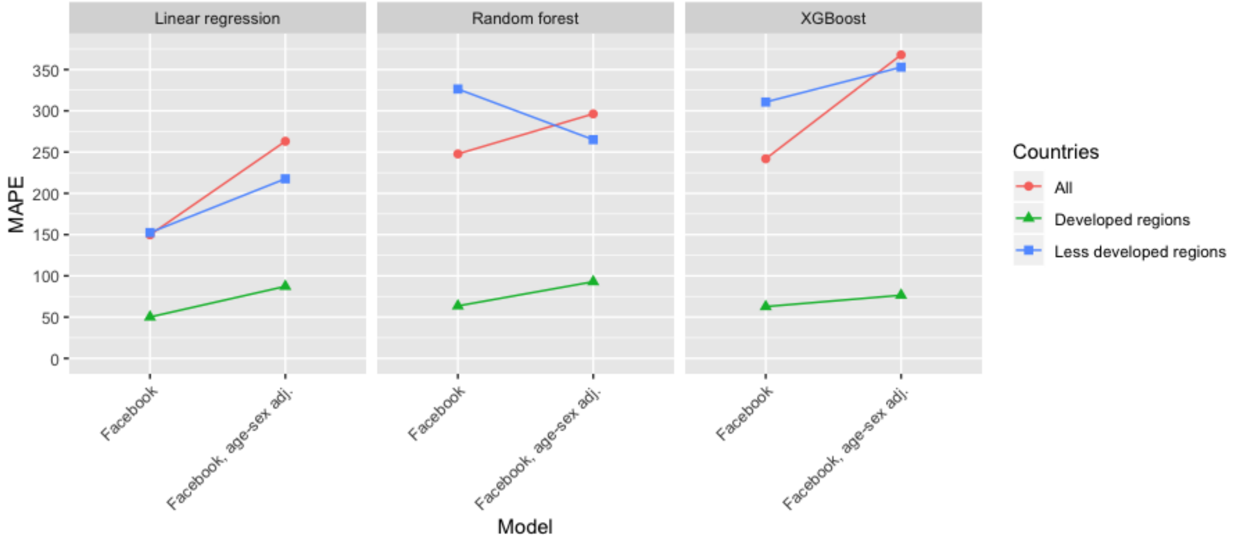

Figure 1: MAPEs for out-of-sample migrant stock prediction using Facebook data adjusted by age and sex (equation 6) and Facebook data not adjusted by age and sex (equation 2). Color-coded by country data quality and faceted by modeling technique.

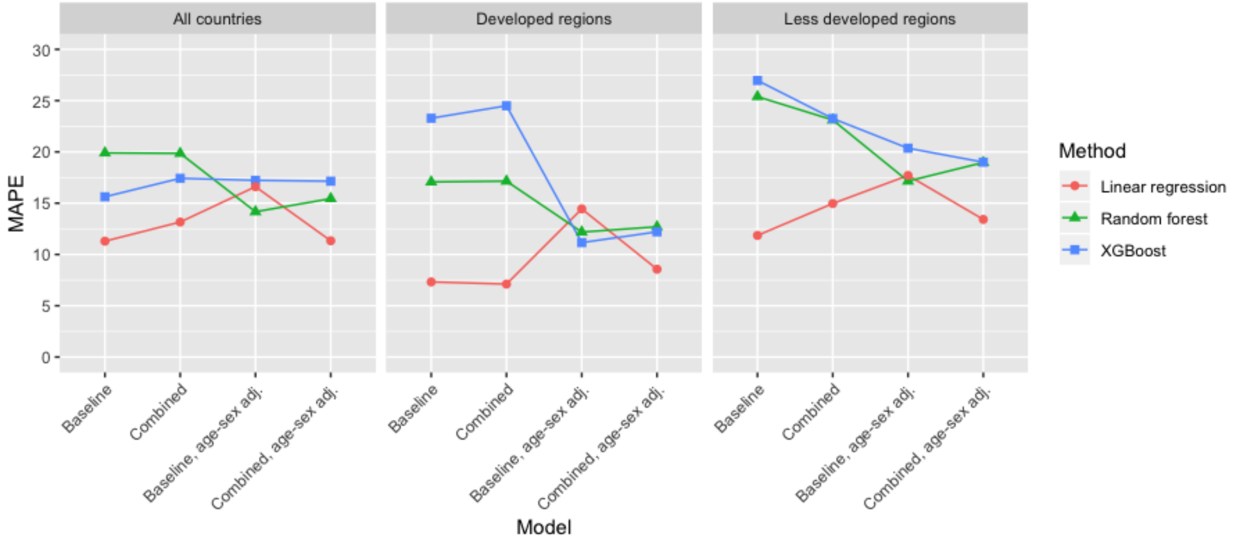

Figure 2: MAPEs for out-of-sample migrant stock prediction using autoregressive (equation 1), combined autoregressive-Facebook (equation 3), autoregressive age-sex-adjusted (equation 5), and combined autoregressive-Facebook age-sex-adjusted (equation 8) models, color-coded by modeling technique and faceted by country data quality, rather than color-coded by country data quality and faceted by modeling technique (Fig. 4).

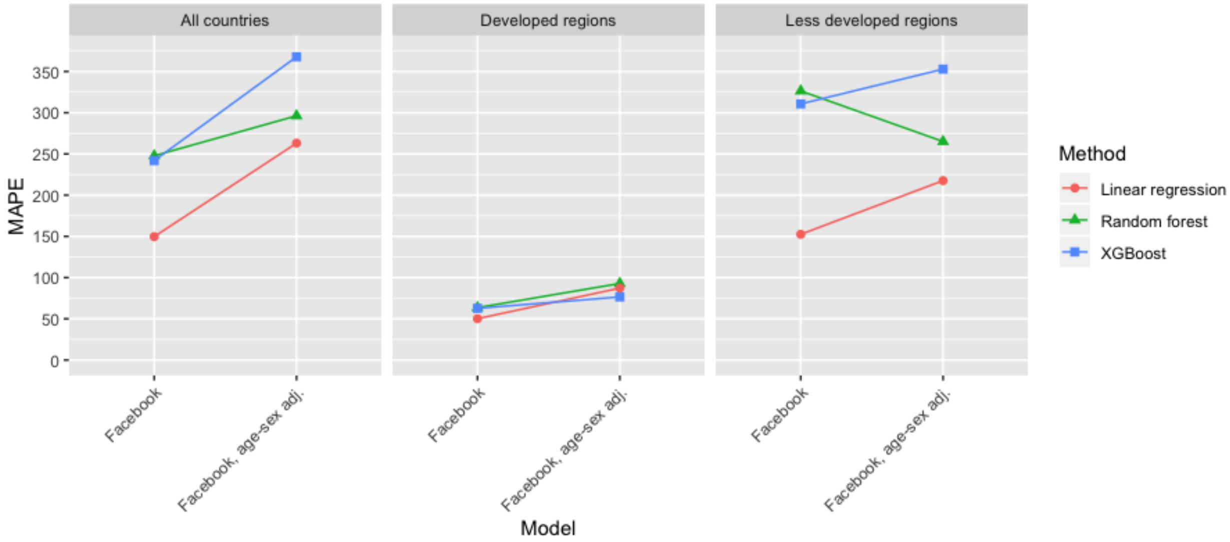

Figure 3: MAPEs for out-of-sample migrant stock prediction using Facebook data adjusted by age and sex (equation 6) and Facebook data not adjusted by age and sex (equation 2). Color-coded by modeling technique and faceted by country data quality.

7 Code

Our code can be found on Github in the following repository: Migrant Stock Nowcasting.

References

[1] U. N. P. D. |. D. of Economic and Social Affairs, “International migrant stock 2019,” 2019.

[2] G. Chi, F. Lin, G. Chi, and J. Blumenstock, “A general approach to detecting migration events in digital trace data,” *PLOS ONE*, vol. 15, p. e0239408, Oct. 2020.

[3] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State, “Inferring international and internal migration patterns from Twitter data,” in *Proceedings of the 23rd International Conference on World Wide Web - WWW ’14 Companion*, (Seoul, Korea), pp. 439–444, ACM Press, 2014.

[4] B. State, I. Weber, and E. Zagheni, “Studying inter-national mobility through IP geolocation,” in *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM ’13, (New York, NY, USA), pp. 265–274, Association for Computing Machinery, Feb. 2013.

[5] E. Zagheni and I. Weber, “You are where you e-mail: using e-mail data to estimate international migration rates,” in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci ’12, (New York, NY, USA), pp. 348–351, Association for Computing Machinery, June 2012.

[6] E. Zagheni and I. Weber, “Demographic research with non-representative internet data,” *International Journal of Manpower*, vol. 36, pp. 13–25, Jan. 2015.

[7] M. Iqbal, “Facebook Revenue and Usage Statistics (2020),” Oct. 2020.

[8] M. Alexander, K. Polimis, and E. Zagheni, “The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data,” *Population and Development Review*, vol. 45, no. 3, pp. 617–630, 2019.

[9] E. Zagheni, I. Weber, and K. Gummadi, “Leveraging facebook’s advertising platform to monitor stocks of migrants,” *Population and Development Review*, pp. 721–734, 2017.

[10] S. Spyratos, M. Vespe, F. Natale, I. Weber, E. Zagheni, and M. Rango, “Quantifying international human mobility patterns using Facebook Network data,” *PLOS ONE*, vol. 14, p. e0224134, Oct. 2019.

[11] U. C. Bureau, “American Community Survey (ACS),” Dec. 2020.

[12] Eurostat, “Migration and migrant population statistics - Statistics Explained,” May 2020.

[13] OECD, “OECD International Migration Database,” 2020.

[14] S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts, “Predicting consumer behavior with web search,” *Proceedings of the National academy of sciences*, vol. 107, no. 41, pp. 17486–17490, 2010.

[15] I. Tingzon, A. Orden, K. T. Go, S. Sy, V. Sekara, I. Weber, M. Fatehkia, M. García-Herranz, and D. Kim, “MAPPING POVERTY IN THE PHILIPPINES USING MACHINE LEARNING, SATELLITE IMAGERY, AND CROWD-SOURCED GEOSPATIAL INFORMATION,” *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-4/W19, pp. 425–431, Dec. 2019.

[16] W. Luo, T. Nguyen, M. Nichols, T. Tran, S. Rana, S. Gupta, D. Phung, S. Venkatesh, and S. Allender, “Is Demography Destiny? Application of Machine Learning Techniques to Accurately Predict Population Health Outcomes from a Minimal Demographic Dataset,” *PLOS ONE*, vol. 10, p. e0125602, May 2015.

[17] Facebook, “Facebook Marketing API.” `https://developers.facebook.com/docs/marketing-apis/`, November 2020. Version 9.0.

[18] A. Herdağdelen, B. State, L. Adamic, and W. Mason, “The social ties of immigrant communities in the united states,” in *Proceedings of the 8th ACM Conference on Web Science*, WebSci ’16, (New York, NY, USA), p. 78–84, Association for Computing Machinery, 2016.

[19] U. N. P. D. | . D. of Economic and Social Affairs, “INTERNATIONAL MIGRANT STOCK 2019,” Aug. 2019.

[20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[21] D. Muchlinski, D. Siroky, J. He, and M. Kocher, “Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data,” *Political Analysis*, pp. 87–103, 2016.

[22] R. Couronné, P. Probst, and A.-L. Boulesteix, “Random forest versus logistic regression: a large-scale benchmark experiment,” *BMC bioinformatics*, vol. 19, no. 1, p. 270, 2018.