Migrant Stock Nowcasting Proposal

Jordan Klein, Shreyas Gandlur, Dorothy Zhao

12/1/2020

Problem

Goal is to see if social media data can improve demography.

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 UN publishes estimates of global international migrant stock biennially. However, there is a great degree of variance between different countries in the quality of, methods they use, and frequency with which they collect the data used to generate these estimates. Simultaneously, social media platforms, in particular Facebook, are ever more ubiquitous, even in parts of the world without high quality migration data. We aim to nowcast current UN migrant stock estimates by training models on a sample of countries and measuring their out-of-sample predictive accuracy to determine:

1. Whether models using Facebook expat data have superior predictive accuracy compared to a baseline model using previous UN migrant stock estimates.
2. How the predictive accuracy of a combined model using both Facebook expat data and previous UN migrant stock estimates compares to models using either of these data sources alone.
3. Does adjusting for Facebook's sampling bias with respect to age and sex-specific rates of penetration improve predictive accuracy compared to a naive model using Facebook expat data that does not account for these biases?
4. How do the countries we include in our sample with respect to the quality of their migration data influence our predictive accuracy? How do our models' predictive accuracy compare when trained and tested on random samples of countries, when trained on developed countries with higher quality data and tested on developing countries with poorer quality data, or trained on developing countries with poorer quality data and tested on developed countries with higher quality data? How does their predictive accuracy compare when we limit our training and test samples to just countries with higher quality data or just countries with lower quality data?
5. Do certain types of predictive models have superior accuracy?

Related Work

Using Facebook data to predict migrant stock in the United States

In 2017, Zagheni et al used Facebook's advertising platform to predict the number of foreign born individuals and their countries of origin in each of the 50 United States. Using American Community Survey data as their "ground truth", they tried 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 said states. They made their predictions using linear regression, training both 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 find their calibrated model to have superior predictive accuracy to their naive model, they acknowledge several key limits to prediction inherent in their approach, namely biases beyond age and country of origin that they did not adjust for, the lack of transparency in how Facebook classifies users into categories like expats, the limits to using only linear regression as a statistical method for approaching this problem, and most importantly, the fact that national migration statistics are not truly a “ground truth” measurement but are themselves only estimates of the “ground truth”.

While all US states measure migrant stock using the ACS, we extend Zagheni et al’s approach by predicting migrant stock on a global scale across countries with highly heterogeneous methods of estimating migrant population and differentiate between countries with superior and lower quality migration data. We also evaluate the performance of prediction models using Facebook data against an autoregressive baseline to determine whether it meaningfully improves migrant stock prediction in the first place. We also use different modeling methods besides just linear regression. However, compared to Zagheni et al our data are limited as the UN does not further break down migrant stock age group and sex estimastes by country of origin and we therefore cannot adjust our models for differential age and sex distributions of Facebook use by country of origin.

Comparing predictive performance using digital trace data to a baseline

In their 2010 paper, Goel et al. predict various consumer behavior and health outcomes using search volume. Using linear regression, they compare the predictive performance of models using just search data to baseline autoregressive models using previous values of the outcome of interest and to combined models using both search data and autoregressive terms. This allows them to systematically evaluate the added value of digital trace data in various prediction problems.

We build on this approach by using it to study the outcome of migrant stock and evaluating whether adjusting digital trace data for sampling bias improves its value-added for prediction. In sum, we synthesize Zagheni et al. and Goel et al’s approaches, extending them by using modeling methods beyond just linear regression, exploring the best approaches to predicting migrant stock across countries with varying data quality, studying the role of this varying quality of data in predictive accuracy, and examining implications for limits to prediction.

Implementation

Dataset Collection

Facebook data

We will start by collecting our training data using the Facebook for Business Marketing API. Using the Marketing API, we are able to query Facebook’s estimate for the number of expats living in a specific country. Furthermore, we are also able to collect more specific demographic information about these expats, such as age, sex, and country of origin in addition to demographic information about Facebook users in that country as a whole. A key limitatino of this data is that only current data can be queried, not retrospective data.

UN migrant stock data

The UN estimates the migrant stock of each country in the world biennially in aggregate, by age and sex, and by country of origin and sex. They use the best available data from each country, either current or older population registers, censuses, or surveys and use estimation to fill in gaps. For countries with missing data, they use regional imputation. Country of birth is used for defining migrants whenever available, with country of citizenship used when it is not. In countries that don’t include refugees or asylum seekers in their official migration statistics, refugee/asylum seeker estimates are added to total migrant stock estimates. The UN has migrant stock estimates for 2019, 2017, and 2015, and plans to publish estimates in the future for 2021.

Making Predictions

Models

We will use the following models to try to nowcast UN migrant stock estimates. Presented here are their linear regression equations:

1. Autoregressive baseline

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

   Where **foreign_born** is total migrant stock, subscript C indicates country, and $t = 2019$.

2. Facebook naive

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

   Where **FB_expats** is the number of Facebook expats.

3. Combined with Facebook naive

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

4. Facebook age-sex corrected

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

   Where $\mathbf{FB\_penetration}^z_C = \frac{\mathbf{FB\_users}^z_C}{\mathbf{Total\_pop}^z_C}$, the superscript z indicates age-sex group and I^z is an indicator variable for each age-sex group.

5. Combined with Facebook age-sex corrected

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

Countries' level of development/data quality

The UN migrant stock estimates divides countries into "more developed regions" where migration data quality is generally higher, and "less developed regions" where migration data quality is generally lower. We will compare the predictive when using the following approaches for constructing our training and test sets:

1. Randomly sample from all countries for training and test sets.
2. Train on a sample of more developed countries, test on a sample of less developed countries.
3. Train on a sample of less developed countries, test on a sample of more developed countries.
4. Randomly sample from only more developed countries for training and test sets.
5. Randomly sample from only less developed countries for training and test sets.

Prediction modeling methods

We will use the following methods for generating prediction models:

1. Linear regression
2. Machine learning methods that will be considered

    Random Forest

    AdaBoost

    XGBoost

Measuring predictive accuracy

We will measure the predictive accuracy of our models using mean absolute percentage error (MAPE) from UN migrant stock estimates in the countries in our test sets.

Limitations

There are several limitations to our approach:

1. The data we are using as our "ground truth", UN migrant stock estimates, is not actually the "ground truth" of what we are trying to predict but is only an estimate of the "ground truth" itself. Both the facts that the outcome we are predicting is only an estimate of the "ground truth" and that its deviation from the actual "ground truth" highly varies between different countries imposes hard limits to prediction on this problem. Though we attempt to quanity the limits to prediction imposed by different countries having different degrees of deviation from the "ground truth" by comparing predictive accuracy between countries with better and poorer quality data, quantifying the limits to prediction imposed by our outcome of interest only being an estimate of the "ground truth" is beyond the scope of our exercise, but may be the grounds for future published work on this topic.
2. We note that we are predicting 2019 UN migrant stock data using Facebook data from 2020 in this exercise; in effect we are not nowcasting but rather predicting the past. In order to conduct a proper exercise in nowcasting we will have to wait until the 2021 UN migrant stock data is published for and nowcast this data using Facebook data from 2021. We will preregister and pull facebook data from July 1 2021
3. We don’t have the country of origin of migrants broken down by age and sex and therefore cannot adjust our models for differential age and sex distributions of Facebook use by country of origin.
4. The way Facebook classifies expats is proprietary.
5. UN doesn’t publish methodology for how migrant stock is estimated when they publish their estimates.

Possible extensions

Although we adjust our models by age and sex-specific rates of Facebook penetration in migrants’ destination countries, the representativeness of Facebook data of migrants is influenced by its penetration in their origin countries as well as their destination countries. One possible extension of our work could be to examine whether adjusting for Facebook penetration rates in migrants’ origin countries improves our models’ predictive accuracy. We could implement this by calculating a weighted average of Facebook penetration in migrants’ origin countries in each destination country and include this as an additional predictor.

References

- Gil-Clavel, Sofia, and Emilio Zagheni. “Demographic Differentials in Facebook Usage around the World.” Proceedings of the International AAAI Conference on Web and Social Media 13 (July 6, 2019): 647–50.
- Goel, Sharad, Jake M. Hofman, Sébastien Lahaie, David M. Pennock, and Duncan J. Watts. “Predicting Consumer Behavior with Web Search.” Proceedings of the National Academy of Sciences 107, no. 41 (October 12, 2010): 17486–90. https://doi.org/10.1073/pnas.1005962107.
- Marketing API. Facebook. https://developers.facebook.com/docs/marketing-apis/.
- Statistical Commission. “Report on the Forty-Sixth Session.” United Nations Economic and Social Council, March 3, 2015. https://unstats.un.org/unsd/statcom/46th-session/documents/statcom-2015-46th-report-E.pdf.
- United Nations Population Division | Department of Economic and Social Affairs. “International Migrant Stock 2019,” August 2019. https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates1
- Zagheni, Emilio, Ingmar Weber, and Krishna Gummadi. “Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants.” Population and Development Review 43, no. 4 (2017): 721–34. https://doi.org/10.1111/padr.12102.