Forecasting Internally Displaced Population Migration Patterns in Syria and Yemen

Benjamin Q. Huynh, BS; Sanjay Basu, MD, PhD

ABSTRACT

Objectives: Armed conflict has contributed to an unprecedented number of internally displaced persons (IDPs), individuals who are forced out of their homes but remain within their country. IDPs often urgently require shelter, food, and healthcare, yet prediction of when IDPs will migrate to an area remains a major challenge for aid delivery organizations. We sought to develop an IDP migration forecasting framework that could empower humanitarian aid groups to more effectively allocate resources during conflicts.

Methods: We modeled monthly IDP migration between provinces within Syria and within Yemen using data on food prices, fuel prices, wages, location, time, and conflict reports. We compared machine learning methods with baseline persistence methods of forecasting.

Results: We found a machine learning approach that more accurately forecast migration trends than baseline persistence methods. A random forest model outperformed the best persistence model in terms of root mean square error of log migration by 26% and 17% for the Syria and Yemen datasets, respectively.

Conclusions: Integrating diverse data sources into a machine learning model appears to improve IDP migration prediction. Further work should examine whether implementation of such models can enable proactive aid allocation for IDPs in anticipation of forecast arrivals.

Key Words: Syria, Yemen, machine learning, internal displacement, forecasting

Armed conflict has contributed to an alarming rate of migration, with over 68.5 million forcibly displaced people worldwide.[1] Of those displaced, over 40 million are considered internally displaced persons (IDPs), individuals who are forced out of their homes but remain within their country. IDPs often require aid in the form of food, shelter, or health care. However, because much of internal displacement arises from regional instability for which local authorities are either unwilling or unequipped to provide aid, it is rare for IDPs to find support from their governments.[2]

Humanitarian response to IDPs is instead typically provided by international nongovernmental organizations. Such aid groups face many logistical challenges providing support to IDPs within conflict-rife zones, one of which is resource allocation across many possible sites to which IDPs may migrate. Given the unpredictability of conflict zones, it is difficult to anticipate when and where IDPs will arrive, so it is unclear which shelters and camps will reach capacity soonest, and where supplies and workers should be sent. At present, allocations are often ad hoc and delayed. It would, therefore, be valuable to forecast IDP migration so that aid groups can proactively prepare to distribute resources in an anticipatory manner.

Some previous works use simulation modeling[3–5] or linear statistical models to forecast migration.[6,7] These methods are useful for data-scarce scenarios because they do not require much data to implement, but they might not be able to leverage the information found in large and heterogeneous datasets to the extent that machine learning methods can.[8] Additionally, no study to our knowledge has modeled IDP migration, which is more granular and noisy than international migration, despite being more common. Historically, IDP migration data have been delayed, missing, and sometimes anecdotal rather than consistently collected, due to difficulties in obtaining reliable data. Our work builds on previous works in that we use machine learning models on large, heterogeneous datasets (the scale of which has previously been unavailable) for internal displacement migration forecasting (as opposed to international migration forecasting).

In recent years, internal displacement task forces have been established to collect on-site data on IDP migration through surveys, registrations, and site monitoring, all of which are triangulated and verified through multiple sources. Additionally, other public data sources have emerged, providing potential predictors of migration, including market prices for staple goods, wages, and conflict events. By using these data, we are able to provide monthly forecasts of IDP migration across provinces in Syria and Yemen using machine learning.

METHODS

Data

We obtained monthly IDP migration data within Syria and Yemen from publicly available datasets provided by the United Nations Office for the Coordination of Humanitarian Affairs and the International Organization for Migration, respectively. The Syria dataset spans from January 2016 to December 2017; the Yemen dataset spans from January 2014 to September 2017. We determined each observation within our dataset to be each unique grouping of month, origin province, and destination province. The Syria and Yemen datasets thus contain 1505 and 3563 observations, respectively.

We obtained monthly food prices, fuel prices, and wages within provinces of both Syria and Yemen from the World Food Program's global food price database. The dataset contains values for commodities such as cheese, wheat, diesel, and gas. We calculated the median value for commodities for the categories of food, fuel, and wages. The values were originally recorded across distinct districts and marketplaces within provinces, so we calculated the medians of all values across provinces per month to get province-level values for the model. The global food price database is updated monthly and spans from the early 2000s to December 2017. We used within-month median imputation for missing values. In the Syria and Yemen price datasets, 6% and 46% of values were missing, respectively. Wage data were only available for July 2016 onward for Yemen, so we excluded the wage data from the Yemen analysis.

We obtained conflict data from 2 separate sources: The Integrated Conflict Early Warning System (ICEWS) dataset and the Armed Conflict Location and Event Data (ACLED) collection. The ICEWS dataset consists of political events across the globe and is publicly available for data spanning from 1995 to 2016. We took the subset of ICEWS events based on codes that corresponded to armed conflict events. The ACLED dataset consists of global armed conflict event data; the Middle East ACLED data span from 2017 to May 2018. We defined a conflict intensity variable using these data, defined as the number of violent events per month for a given province, scaled to zero mean and unit variance. Scaling was done at the dataset level (separately for ICEWS and ACLED) to account for potential frequency biases in data collection between the 2 datasets.

We finally created a distance metric by taking the coordinates of each province's centroid and calculating the Haversine distance[9] between each province pair, ie, the distance across a sphere of 2 points given their coordinates.

Models

We trained various statistical and machine learning models to predict next month's migrations for each observation. Broadly, we formulate the models as $\hat{y}_{ijk} = f(x)$, where $i = 1, \ldots, N$ origin provinces, $j = 1, \ldots, n_i$ origin-destination pairs, $k = 1, \ldots, n_{ij}$ monthly observations for each origin-destination pair, $\hat{y}$ represents the estimated number of IDP migrations for a given location, and x represents our set of covariates. Details on the specific models can be found in the Supplementary Materials, which are available online.

For each model, our covariates consisted of monthly features derived from the aforementioned datasets. We used monthly data from both the origin and destination for each destination-origin pair to model the "push and pull" factors[10] of migration. Our model covariates for each observation were thus the date, monthly food prices, fuel prices, wages, and conflict intensity from both the origin and destination, as well as the distance between the origin and destination. To account for the fact that marketplace data typically take 3 months to be collected and shared, we used food/fuel prices and wages with a 3-month lag to reflect available covariates for real-time forecasting. For example, we use prices and wages from January 2017 to predict migrations for April 2017. Conflict data are updated monthly, so we used the previous month's conflict intensity metric for each observation. However, we also include a 3-month lagged conflict intensity variable to account for interactions with price data. An autoregressive term, the previous month's IDP migrations for a given origin–destination pair, was included as the final covariate.

For comparison against our machine learning models, we applied the baseline persistence methods of last observation carried forward (LOCF) and historical mean (HM). Both methods are done from within origin–destination pairs. More specifically, HM was calculated as $\hat{y}_{ijk} = (y_{ij1} + \ldots + y_{ijnij})/n_{ij}$ and LOCF was calculated as $\hat{y}_{ijk} = y_{ijk\text{-}1}$.

We evaluated our models by forecasting out-of-sample 1 month ahead using a rolling origin. We started at month 5 for Syria (out of 24) and month 23 for Yemen (out of 44); these are the time points at which all origin and destination provinces became present in the datasets. We use root mean squared error (RMSE), mean absolute error (MAE), and sign accuracy as metrics for evaluation. Sign accuracy is a metric we introduce that measures how well a model can predict whether a given observation will be an increase in migrations from the previous month, or not. Specifically, we measure this as a binary classifier evaluation at each observation: an observation is a 1 for an increase in migrations, and 0 otherwise. We test our methods both on log and absolute units (see Supplementary Materials).

RESULTS

The baseline persistence models we tested, HM and LOCF, were able to capture trends of IDP migration and log-migration within each province with RMSE (RMSE = 10587 and 10661

https://doi.org/10.1017/dmp.2019.73 Published online by Cambridge University Press

TABLE 1

Predictive Performance of Forecasting Methods for Syria (a) and Yemen (b) on Both Migration and Log-Migration

(a) Syria Predictive Performance

Model	RMSE	MAE	R^2	RMSE (log)	MAE (log)	R^2 (log)	Sign Acc.
HM	10587.07	3066.02	0.24	2.15	1.66	0.38	0.63
LOCF	10660.7	2577.37	0.34	2.01	1.44	0.46	0.59
RF	**9576.61**	**2237.73**	**0.45**	**1.49**	**1.14**	**0.59**	**0.70**

(b) Yemen Predictive Performance

Model	RMSE	MAE	R^2	RMSE (log)	MAE (log)	R^2 (log)	Sign Acc.
HM	1332.29	287.78	0.08	2.10	1.75	0.30	0.67
LOCF	1413.30	325.92	0.17	1.48	1.13	0.33	0.60
RF	**1140.01**	**247.05**	**0.21**	**1.23**	**0.98**	**0.39**	**0.74**

Abbreviations: HM, historical mean; LOCF, last observation carried forward; MAE, mean absolute error; RF, random forest; RMSE, root mean squared error; R^2, coefficient of determination.

[a] Boldface type indicates best forecasting performance for a given metric.

for Syria HM and LOCF, 1332 and 1413 for Yemen HM and LOCF) and MAE (MAE = 3066 and 2577 for Syria HM and LOCF, 288 and 326 for Yemen HM and LOCF) values that are moderately low, with poor R^2 ($R^2 = 0.24$ and 0.34 for Syria HM and LOCF, 0.08 and 0.17 for Yemen HM and LOCF) values (Table 1). Because the persistence models relied solely on historical data, they were unable to provide forecasts for regions for which there previously had been no IDP arrivals, instead producing erroneous zero predictions. Additionally, both persistence models performed poorly in terms of sign accuracy (63% and 59% for Syria HM and LOCF, 67% and 60% for Yemen HM and LOCF).

By comparison, machine learning models we trained outperformed HM and LOCF forecasts in terms of RMSE, MAE, and R^2 for both predicting migration and log-migration, as well as sign accuracy (Table 1). Furthermore, the machine learning models were able to make predictions for regions without previous data, avoiding erroneous zero predictions. The machine learning models had similar predictive performances, although the random forest machine learning algorithm in particular appeared slightly better overall than the others across both countries. The random forest specifically outperformed LOCF in terms of RMSE of log-migration by 26% and 17% for the Syria and Yemen datasets, respectively.

DISCUSSION

We found that forecasting IDP migration by integrating diverse data sources into a machine learning model appears to improve IDP migration prediction. If we assume human-level performance to be similar to that of the persistence models, then we would expect machine learning models to improve forecasting by 20-30% in terms of RMSE, potentially facilitating more accurate resource allocation.

We observed that the random forest, our best machine learning model, captured overall trends of IDP migrations for each province but occasionally failed to capture sudden spikes in displacement (Figure 1). Our model obtained a 70% and 74% sign accuracy for Syria and Yemen, respectively (Table 1). These are relatively high values (± 2% sign accuracy compared with other machine learning models we tested), but they also suggest room for improvement in absolute terms of detecting spikes. This is likely because our features are unable to fully characterize when spikes occur. For example, our conflict intensity metric is determined by how many armed conflict events occur in a month, but did not consider the magnitude of the armed conflict event. In comparison, the baseline persistence methods were fundamentally poor at detecting large spikes in displacement because they simply projected past data.

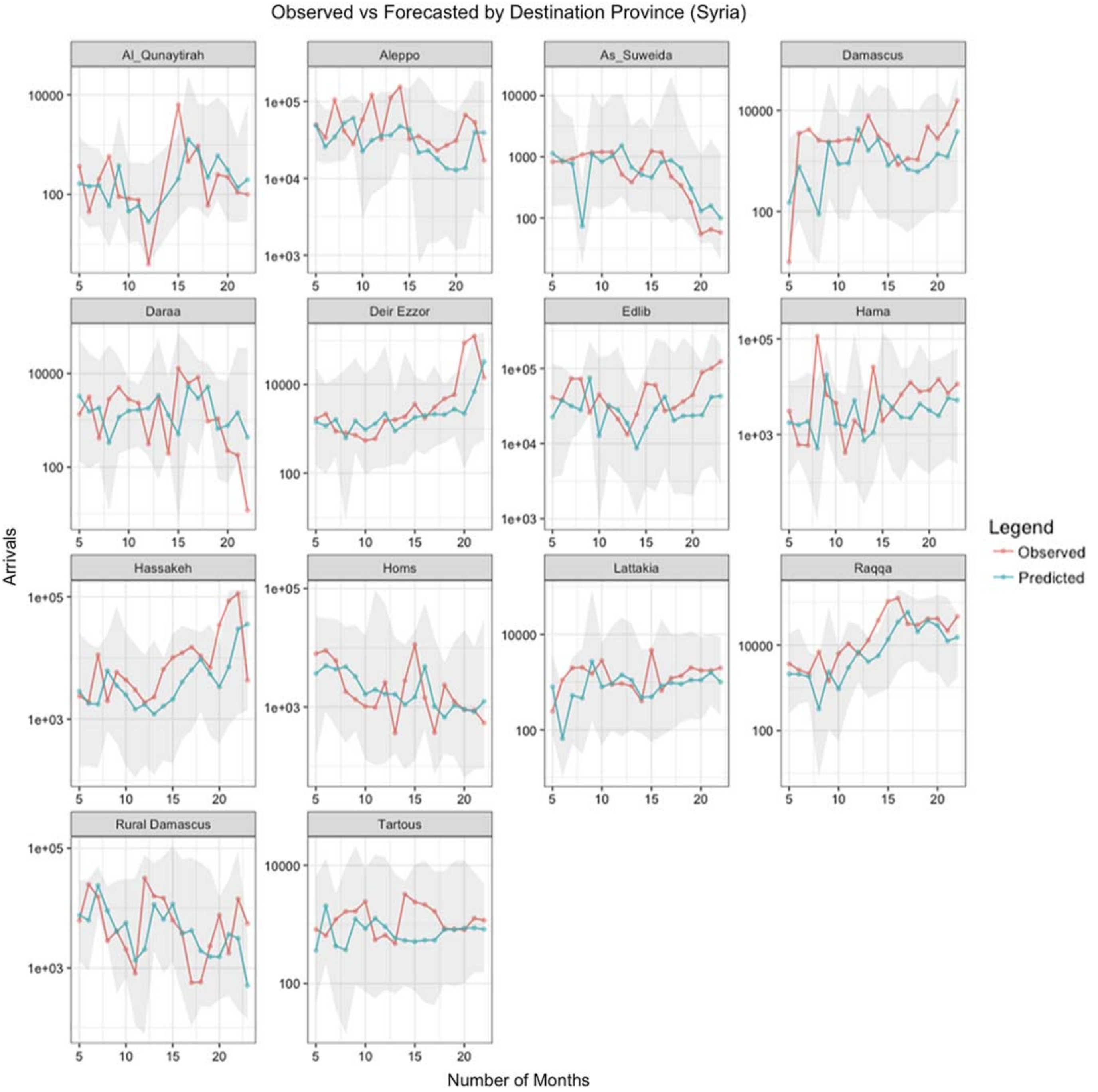
The limitations of our work are largely related to the quality of the available data. There is substantial uncertainty inherent to the datasets we used: the ground truth for IDP migration numbers, conflict events, prices, and wages are all subject to the unreliability of on-site data collection. The IDP migration values also lack potentially valuable disaggregated information, such as age or sex, or more granular information, such as daily migrations (instead of monthly) or subdistrict-level migrations (instead of provinces). There were also substantial amounts of missing data from the price dataset for Yemen that were imputed, possibly explaining the poorer predictive performance of the machine learning models when applied to the data from Yemen.

Future work could involve incorporating new kinds of data into our models. Other approaches could involve obtaining new datasets, such as acquiring annotated satellite imagery, cell phone data, or relevant social media posts and adding them to our models. Additionally, incorporating our models directly into the workflow of aid agencies using their supply chain data would allow for explicit optimization of resource allocation based on forecast IDP migration.

https://doi.org/10.1017/dmp.2019.73 Published online by Cambridge University Press

FIGURE 1

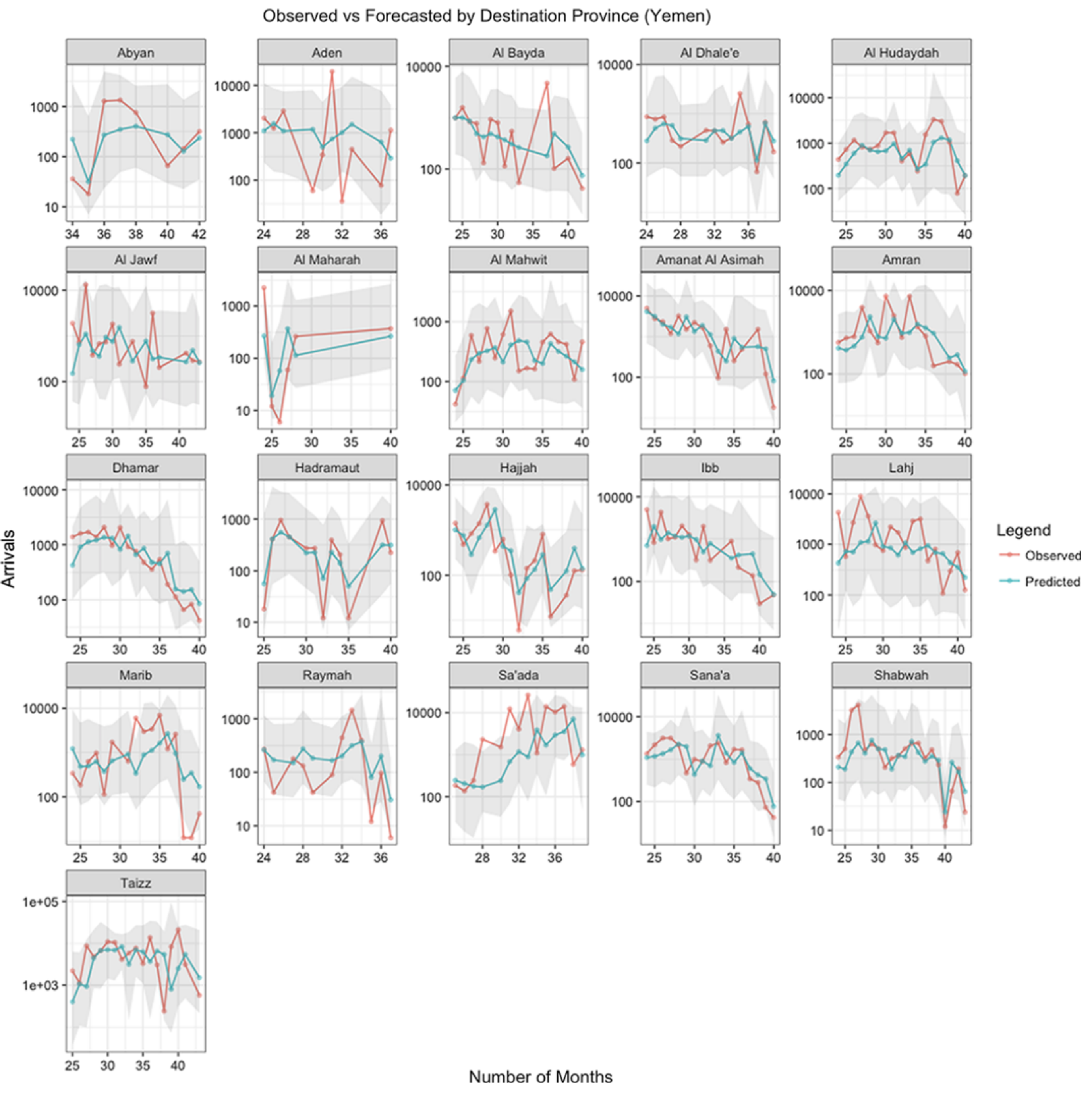
Observed and Forecast Number of IDP Arrivals in Each Province by Month for Syria and Yemen. Forecasts for each month were made from a random forest model trained on data from prior months. Gray shaded regions denote 95% prediction intervals determined by the quantiles of the individual trees for each prediction.



https://doi.org/10.1017/dmp.2019.73 Published online by Cambridge University Press

FIGURE 1

Continued.



https://doi.org/10.1017/dmp.2019.73 Published online by Cambridge University Press

CONCLUSIONS

Forecasting IDP migration patterns appears to be viable using machine learning methods. In the future, more comprehensive data collection (disaggregated information, higher resolution, and less missing data) would likely lead to better predictive performance and thus better decision-making. We hope that our work can help promote discussion on what can be accomplished with IDP data, with the eventual goal of direct operational use by aid agencies.

About the Authors

Stanford University, Department of Medicine, Stanford, California. (Mr Huynh); Harvard Medical School, Center for Primary Care, Cambridge, Massachusetts (Dr Basu); and Imperial College London, School of Public Health, London, England (Dr Basu)

Correspondence and reprint requests to Benjamin Q. Huynh, 615 Crothers Way, Office 211, Stanford, CA 94305, USA (e-mail: benhuynh@stanford.edu)

Ethics Approval and Consent to Participate

Not applicable. No IRB review was required for this study.

Availability of Data and Materials

The datasets analyzed during the current study are available in the following repository: https://github.com/benhuynh/migrationPatterns/tree/master/data. They were derived from publicly available datasets described as follows. The IDP migration data for Yemen can be found at the International Organization for Migration's data repository (https://displacement.iom.int/yemen). The price data and Syria IDP migration data we used can be found at the Humanitarian Data Exchange's data repository (https://data.humdata.org/). The ACLED conflict data can be found at https://www.acleddata.com/. The ICEWS conflict data can be found at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/28075.

Acknowledgment

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE – 1656518.

Conflicts of Interest

The authors declare that they have no competing interests.

Authors' Contributions

Mr. Huynh and and Dr. Basu designed the study and wrote the paper. Mr. Huynh assembled the dataset and performed the analyses. Both authors discussed the results and commented on the manuscript.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1017/dmp.2019.73

REFERENCES

1. Internal Displacement Monitoring Centre. Global report on internal displacement 2018; 2018. http://www.internal-displacement.org/global-report/grid2018/downloads/2018-GRID.pdf. Accessed July 18, 2019.
2. Goodwin-Gill GS, McAdam J. *The Refugee in International Law*. Oxford: Oxford University Press; 2007.
3. Suleimenova D, Bell D, Groen D. A generalized simulation development approach for predicting refugee destinations. *Sci Rep*. 2017;7(1):13377.
4. Kniveton D, Smith C, Wood S. Agent-based model simulations of future changes in migration flows for Burkina Faso. *Global Environ Change*. 2011;21:S34-S40.
5. Vernon-Bido D, Frydenlund E, Padilla JJ, Earnest DC. Durable Solutions and Potential Protraction: The Syrian Refugee Case. In: Proceedings of the 50th Annual Simulation Symposium. ANSS '17. San Diego, CA: Society for Computer Simulation International; 2017. p. 19:1-19:9. http://dl.acm.org/citation.cfm?id=3106388.3106407. Accessed July 18, 2019.
6. Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake. *Proc Natl Acad Sci U S A*. 2012;109(29):11576-11581.
7. Cohen JE, Roig M, Reuman DC, et al. International migration beyond gravity: a statistical model for use in population projections. *Proc Natl Acad Sci U S A*. 2008;105(40):15269-15274.
8. Ahmed MN, Barlacchi G, Braghin S, et al. A multi-scale approach to data-driven mass migration analysis. 2016. http://ceur-ws.org/Vol-1831/paper_4.pdf. Accessed July 18, 2019.
9. Hijmans RJ. Geosphere: Spherical trigonometry; 2017. R package version 1.5-7. https://CRAN.R-project.org/package=geosphere. Accessed July 18, 2019.
10. Schoorl J, Heering L, Esveldt I, et al. Push and pull factors of international migration: a comparative report. 2000. https://www.nidi.nl/shared/content/output/2000/eurostat-2000-theme1-pushpull.pdf. Accessed July 18, 2019.

https://doi.org/10.1017/dmp.2019.73 Published online by Cambridge University Press