Forecasting Internally Displaced Population Migration Patterns in Syria and Yemen

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

Stanford University, Department of Medicine

Supplementary Information

Previous Work The study of migration modeling can be traced to 1885 with Ravenstein's seminal work on the "Laws of Migration."[1] From there, models such as the Gravity Model emerged, which derived migration from distances and populations of neighboring areas.[2–4] Afterwards, quantitative models assessing the risk factors (and later developing early warning systems) for migration were created.[5–10]. Later, models for explicitly forecasting migration were developed.[11–17]

Within the literature of migration forecasting, approaches can be roughly divided into two categories: simulation modeling and statistical modeling. Simulation modeling includes techniques such as agent based modeling and dynamical systems. It works by simulating agents collectively interacting inside an environment with predetermined or derived rules and parameters.[18] Statistical modeling in the context of forecasting refers to predicting the future based on previous sample data, using techniques such as generalized linear models or machine learning models.[19]

Naturally, the two approaches have their contextual advantages and disadvantages. Simulation modeling requires users to predefine or estimate parameters and rules into a system, which may introduce bias. However, simulation modeling for migration works well in data-scarce scenarios, making it potentially helpful for new or emerging migration crises. Statistical modeling encodes fewer assumptions about the context from the user, instead fitting predictions strictly based on past data. Complex statistical models are useful for situations in which large, heterogeneous datasets are available, as they are able to find patterns that a human user may not be able to discover easily.

For our work, we used a statistical model approach (a variety of machine learning models) due to the long-lasting and ongoing nature of the crises in Syria and Yemen. Because these crises have been ongoing for nearly a decade, large datasets based on historical events are available for use, making statistical modeling possible.

Background Since 2011, populations in both Syria and Yemen have experienced severe levels of displacement; there are over 7 million IDPs in Syria (with a total population of 18 million) and 2.5 million in Yemen (total population of 28 million).[20] Despite their similarities in terms of geography and timeframe, each country has its own unique factors that contribute to displacement beyond armed conflict. Syria has had extreme levels of massacres and airstrikes against civilians, as well as frequently shifting territorial control. Yemen faces a famine largely due to a blockade, a massive cholera outbreak, and the threat of tropical cyclones.

Data There exist two kinds of missing data in the IDP migration data: missing observations due to zero counts and missing observations due to some areas being inaccessible to surveyors. Despite the missing data, we opted for a complete case analysis instead of imputing the migration data. Our rationale is that if an area is inaccessible to surveyors, then it is also most likely inaccessible to humanitarian aid, so forecasting movements to those areas is not useful for assisting aid groups. Furthermore, it is unclear which missing observations represent zero migrations and which ones represent unrecorded data, so imputation is not sensible.

We observed that the distribution of IDP migration could be modeled as log-normal for the sake of forecasting large, rare displacement events (Figure 4). Because our focus is on predictive performance and not statistical inference, the loss of effect size interpretability from log-transforming the response variable is not relevant. Furthermore, we empirically find that despite the bias introduced by transforming and untransforming the response variable, doing so provided better predictions than directly modeling IDP migration (Table 3). Thus, all evaluation metrics for statistical models are reported from models trained on log-migration, where predictions are untransformed back into migration. We ran our baseline persistence models both on log-migration and migration separately, so as not to introduce bias from transforming and untransforming.

Methods For our linear mixed effects model, we used a three-level structure with random slopes and intercepts.[21] For notation, we defined $i = 1, ..., N$ origin provinces, $j = 1, ..., n_i$ origin-destination pairs, and $k = 1, ..., n_{ij}$ monthly observations for each origin-destination pair. Our formulation of the model was as follows:

$$y_i = X_i\beta + Z_i v_i + \epsilon_i$$

X_i was the known design matrix for the fixed effects, β was the unknown vector of regression coefficients, Z_i was the known design matrix for the random effects, v_i was the unknown vector of random effects with $v_{ij} \sim N(0, \Sigma_v)$, and ϵ_i was the error term vector with $e_{ijk} \sim N(0, \sigma^2)$.

We trained a support vector regression model[22], a machine learning algorithm that seeks to find a function $f(x)$ that approximates y by minimizing a loss function that ignores errors within a given distance ϵ of the true values. We specified it with a polynomial kernel $K(x, y) = (x^T y + c)^d$. Hyperparameters c and d were selected through five-fold cross-validation on a training set.

We trained a random forest[23], an algorithm that trains a large number of individual decision trees and takes the mean output as the prediction. We tuned the optimal number of variables randomly sampled at each split through five-fold cross-validation. We also trained a mixed-effects random forest[24], with a similar specification to our linear mixed effects model: $y_i = f(X) + Z_i v_i + \epsilon_i$, where $f(X)$ was a standard random forest model.

We also used a tree boosting method, XGBoost[25], which forms an ensemble of regression trees and builds a model in stages during training. The hyperparameters tuned through five-fold cross-validation were maximum tree depth, step size shrinkage, subsample ratio of columns (by tree), and subsample ratio of the training instance.

We trained a multi-layer perceptron (MLP), which is a class of feedforward deep neural networks.[26] Briefly, MLPs consists of layers of nodes, where each node is a neuron with a nonlinear activation function; the resulting network is thus a nonlinear function approximator. We specified our MLP with two hidden layers and rectifiers as activation functions. We selected the number of nodes through five-fold cross-validation.

Results The data from both Syria and Yemen revealed large province-to-province and month-to-month variations in IDP migration, as well as in key covariates we studied for prediction: food prices, fuel prices, and wages. The relative standard deviations of IDP migration were extremely large for both Syria and Yemen (389% and 517% respectively), suggesting high variability in migration across provinces and months (Figure 1, Table 2). The price data also yielded large relative standard deviations: 34%, 52%, and 27%, for food prices, fuel prices, and wages in Syria; 55% and 142% for food and fuel prices in Yemen.

Interpretation of the random forest model yielded sensible results, suggesting our models are finding patterns within the data and not just fitting to noise. The minimal depth levels, a measurement of how much impact a given variable had on the final prediction, appeared plausible for both datasets (Figure 3). The autoregressive term is unsurprisingly the strongest predictor - we expected last month's migrations to be a good estimate of this month's migrations. Distance was the second strongest predictor - most IDPs become displaced within their home province (Figure 4), so we expect shorter distances between the origin and destination provinces to be associated with larger migration numbers. Food prices and conflict intensity are also strong predictors, likely due to famine and severe civil conflict in both countries.

Ethics We recognize there are ethical concerns involved with developing public forecasts within conflict zones. The primary concern is that malicious actors could use our modeling to more effectively target civilians and/or combatants. Because all of the data we use are publicly available, and since our methods do not require any special tools or access to be replicated, we believe it would be irresponsible to avoid disseminating our research - malicious actors could develop similar work without publishing it.

We hope that by publishing an open-source, public use case, our work will facilitate discussions on proper access and use of the available data. In particular, it is important to discuss whether or not malicious actors are sufficiently equipped to perform machine learning with data on internal displacement migration, and whether or not such learning tools are more likely to be used by more organized humanitarian aid agencies. Additionally, as data collection is improved and disaggregated, care should be taken to avoid personally identifiable information within migration datasets.

Figures and Tables

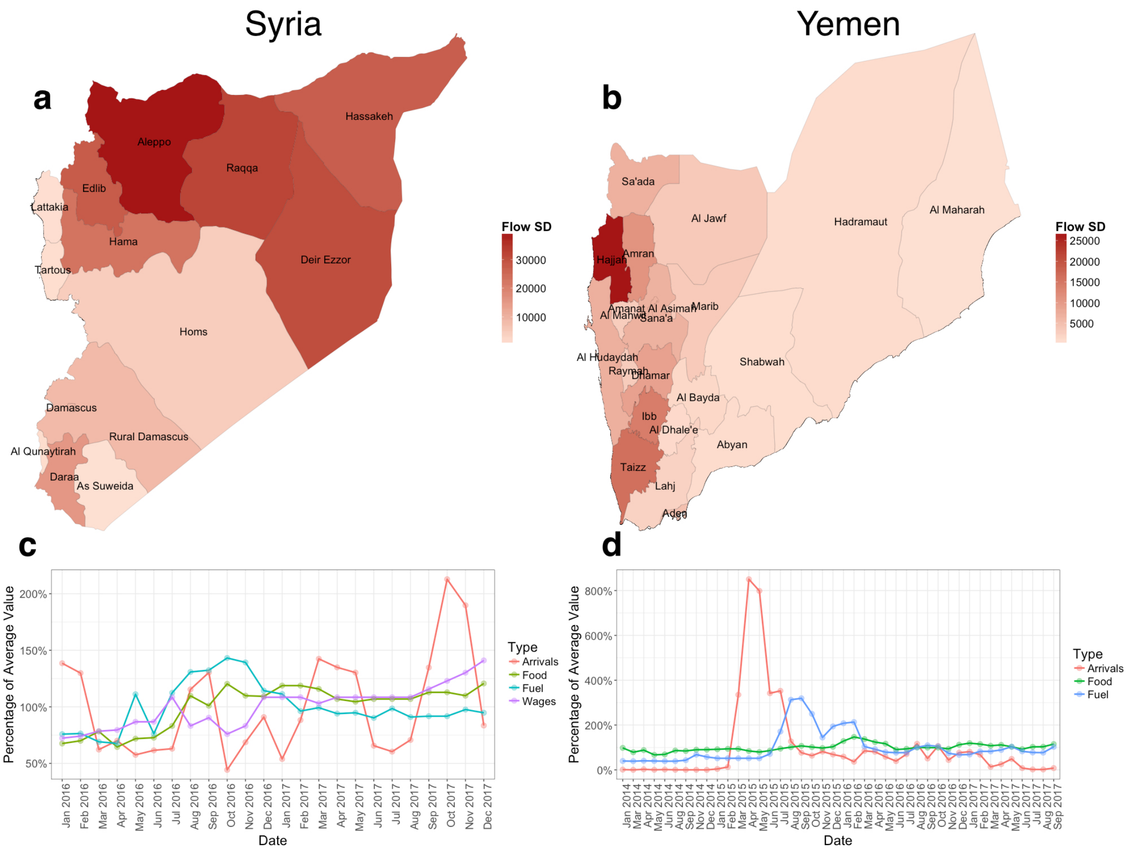

Figure 1: Measurement variability over time for Syria and Yemen. a,b: Provinces of each country color coded by standard deviation of IDP migrations aggregated over time. Darker shades indicate larger variability in IDP migrations for a given province. c,d: Country level statistics on IDP migrations, food prices, fuel prices, and wages over time for Syria and Yemen. Values are presented as percentages of their historical averages. Wage data are unavailable for Yemen.

Table 1: Predictive performance of forecasting methods for Syria (a) and Yemen (b) on both migration and log-migration. HM denotes historical mean, LOCF denotes last observation carried forward, LMM denotes linear mixed effects model, SVM denotes support vector machine, RF denotes random forest, MERF denotes mixed-effects random forest, XGB denotes gradient boosting, and MLP denotes multi-layer perceptron. RMSE is root mean squared error, MAE is mean absolute error, and R^2 is the coefficient of determination.

(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
LMM	10074.47	2370.81	0.31	1.55	1.19	0.56	**0.70**
SVM	10292.38	2383.21	0.26	1.53	1.16	0.57	**0.70**
RF	**9576.61**	**2237.73**	**0.45**	**1.49**	**1.14**	**0.59**	**0.70**
MERF	9627.89	2304.97	0.34	1.53	1.18	0.57	**0.70**
XGB	9760.46	2351.41	0.35	1.59	1.23	0.53	0.68
MLP	10283.04	2378.43	0.35	1.59	1.23	0.53	0.68

(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
LMM	1175.50	276.59	0.17	1.31	1.02	0.37	0.73
SVM	1149.05	254.37	**0.22**	1.37	1.06	0.33	0.74
RF	**1140.01**	**247.05**	0.21	**1.23**	**0.98**	**0.39**	0.74
MERF	1161.15	250.41	0.19	1.25	0.98	0.38	0.75
XGB	1236.94	258.51	0.12	1.27	0.98	0.37	**0.76**
MLP	1588.56	329.39	0.10	1.44	1.09	0.26	0.72

Table 2: Descriptive statistics on IDP arrivals, food prices, wages, fuel prices, and conflict intensity for Syria and Yemen. N denotes the number of observations and SD denotes standard deviation. Wage data are unavailable for Yemen. Units for food/wage/fuel data are in Syrian and Yemeni currency, respectively.

Country	N	Flow Mean	Flow SD	Food Mean	Food SD	Wage Mean	Wage SD	Fuel Mean	Fuel SD	Conflict Mean	Conflict SD
Syria	1505	3098.59	12066.28	474.93	163.42	1383.18	373.97	2054.13	1077.89	0.36	1.33
Yemen	3589	563.54	2912.12	280.04	155.13			977.83	1387.37	0.40	1.17

Table 3: Predictive performance differences between models trained directly on IDP migration and models trained on log-migration (and then transformed back to migrations). Positive values for RMSE and MAE and negative values for R^2 and sign accuracy indicate the model trained directly on migrations performed worse by the given amount.

(a) Syria

Model	RMSE	MAE	R^2	Sign Acc
LMM	2136.71	573.64	-0.18	-0.13
SVM	-325.86	962.05	0.03	-0.10
RF	20.32	1119.93	-0.10	-0.11
MERF	375.20	1112.60	0.01	-0.11
XGB	533.16	1310.00	0.03	-0.09
MLP	1555.76	1516.87	-0.17	-0.06

(b) Yemen

Model	RMSE	MAE	R^2	Sign Acc.
LMM	127.53	12.08	-0.10	-0.16
SVM	785.60	651.41	-0.12	-0.11
RF	-26.89	139.77	0.11	-0.09
MERF	272.47	202.70	0.06	-0.11
XGB	187.72	237.65	-0.01	-0.12
MLP	1464.27	259.35	0.00	-0.06

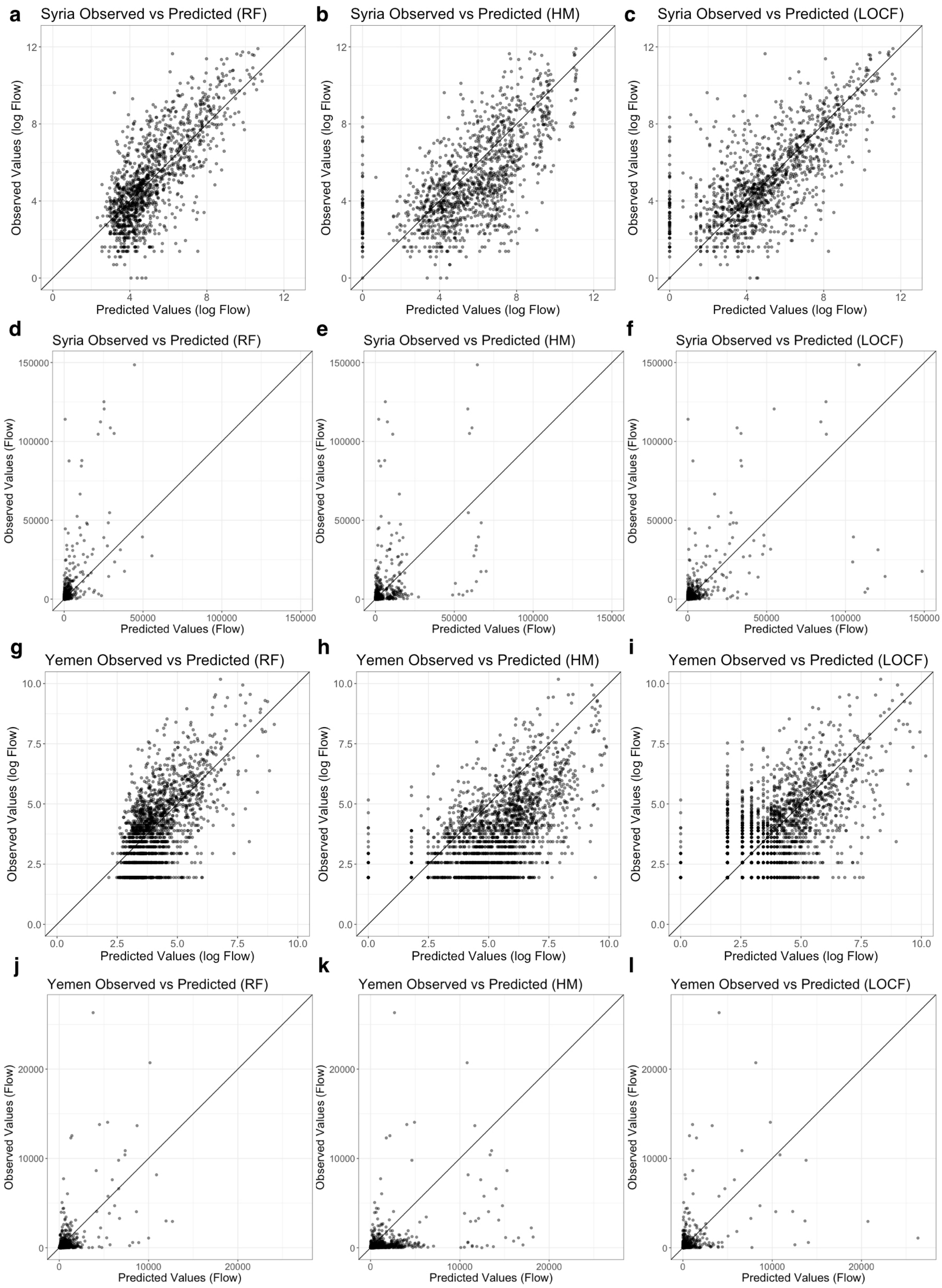

Figure 2: Observed vs predicted values for IDP migration (d-f, j-l) and log-migration (a-c, g-i) aggregated across all available months and provinces. Left column (a,d,g,j) depicts plots from a random forest model (RF), middle column (b,e,h,k) depicts historical mean values (HM), and right column (c,f,i,l) depicts last observations carried forward (LOCF).

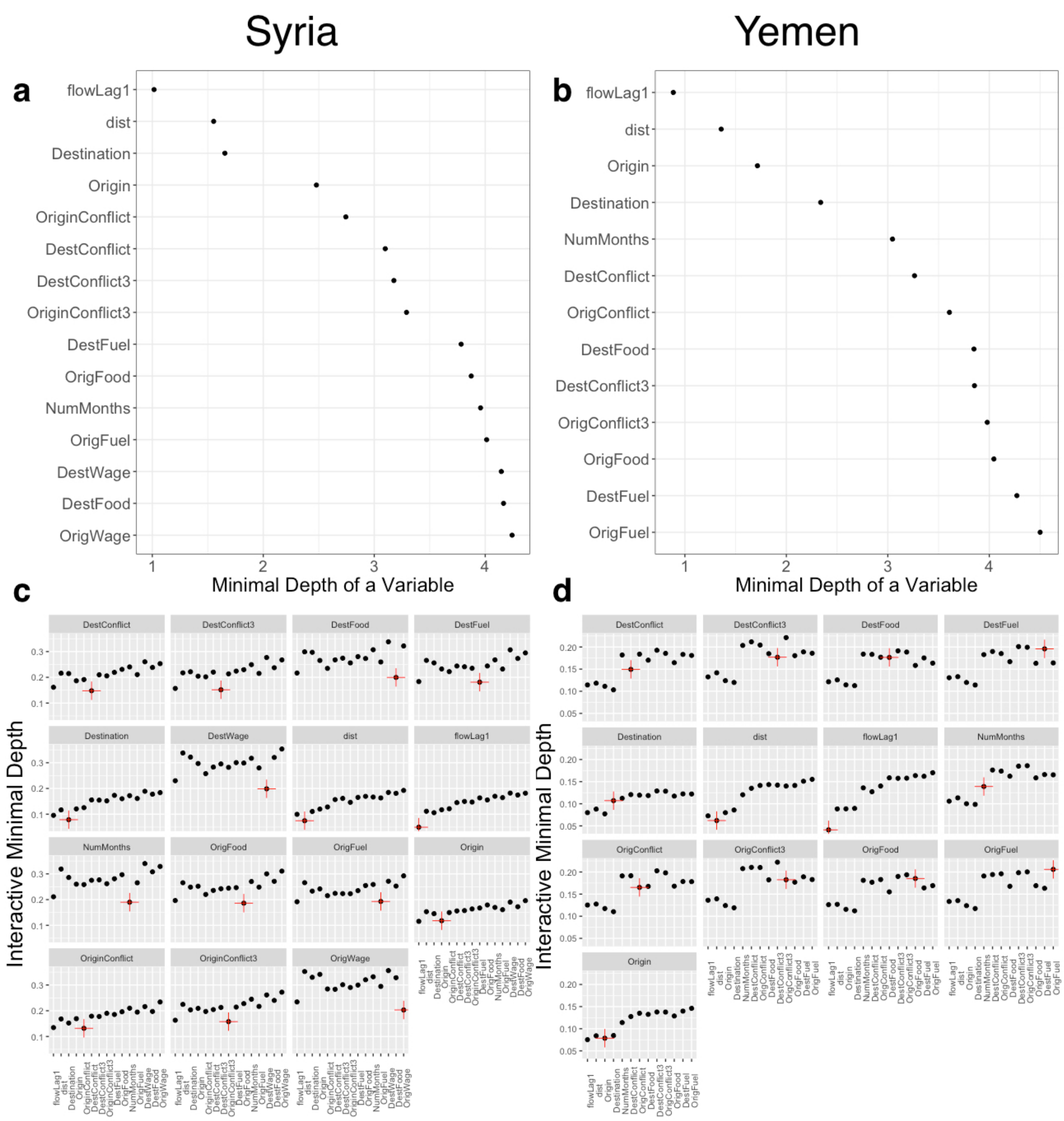

Figure 3: a,b: Random forest minimal depth variables in ranked order for Syria (a) and Yemen (b), with the most important variables at the top. Smaller values of minimal depth indicate a stronger impact on the forest prediction. c,d: Minimal depth variable interactions for Syria (c) and Yemen (d). Red cross indicates the reference variable for each panel. Higher levels of interactivity are indicated by lower levels of minimal depth.

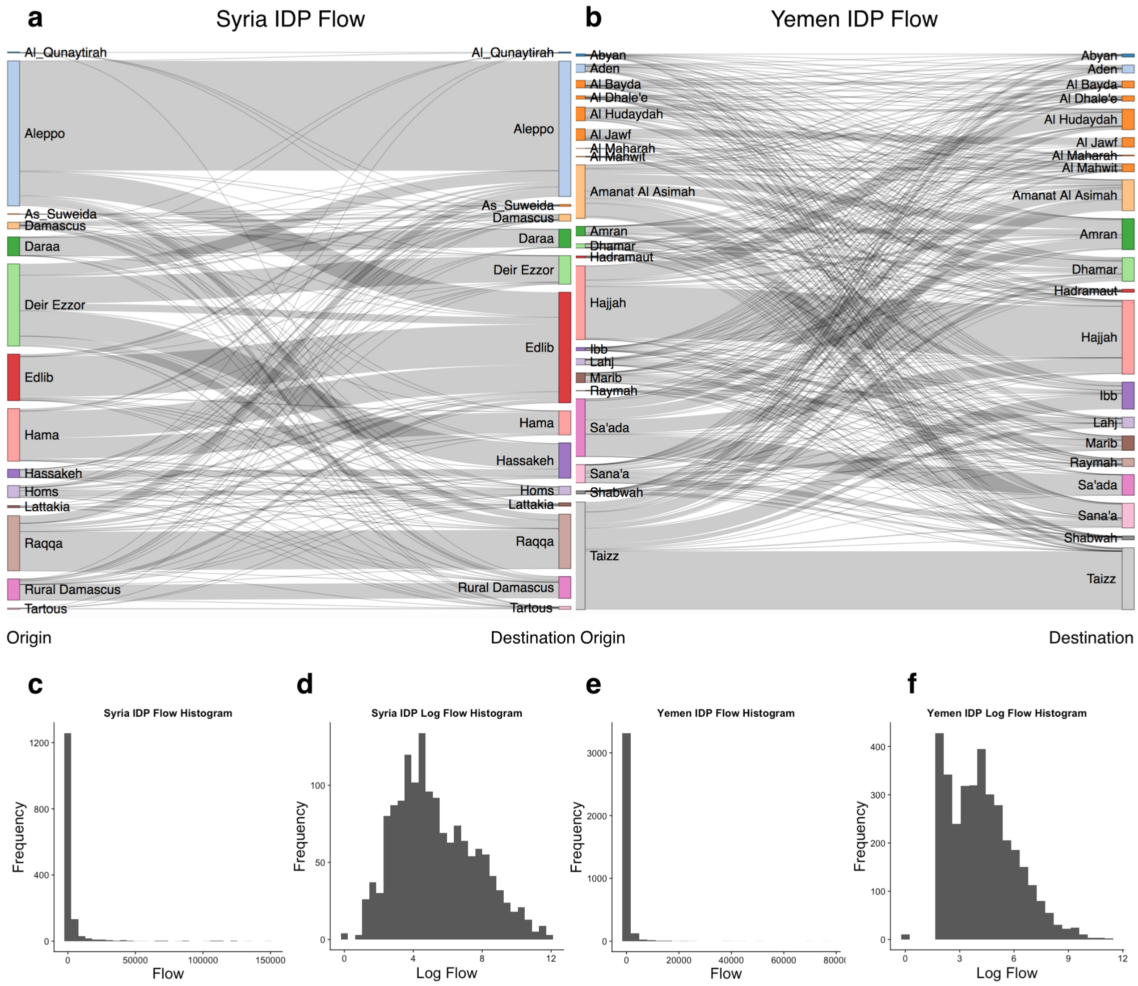

Figure 4: a,b: IDP migration from province to province aggregated over all time periods for Syria (a) and Yemen (b). Each node represents a province. The widths of the bands represent the number of migrations. c-f: Distribution of IDP migration across all time points and provinces for Syria (c,d) and Yemen (e,f). Both log-transformed (d,f) and untransformed IDP migration values (c,e) are shown.

References

1. Ravenstein EG. The laws of migration. Journal of the royal statistical society. 1889;52(2):241–305.

2. Zipf GK. The P 1 P 2/D hypothesis: on the intercity movement of persons. American sociological review. 1946;11(6):677–686.

3. Willekens F. Migration flows: Measurement, analysis and modeling. In: International handbook of migration and population distribution. Springer; 2016. p. 225–241.

4. Stewart JQ. The development of social physics. American Journal of Physics. 1950;18(5):239–253.

5. Shellman SM, Stewart BM. Predicting risk factors associated with forced migration: An early warning model of Haitian flight. Civil Wars. 2007;9(2):174–199.

6. Martineau JS. Red flags: A model for the early warning of refugee outflows. Journal of Immigrant & Refugee Studies. 2010;8(2):135–157.

7. BUNOIU MD, UDROIU I. Spotting trouble in migration flows: An indicator-based early warning model. Bucharest–2016;.

8. Azose JJ, Ševčíková H, Raftery AE. Probabilistic population projections with migration uncertainty. Proceedings of the National Academy of Sciences. 2016;113(23):6460–6465.

9. Moore WH, Shellman SM. Whither will they go? A global study of refugees' destinations, 1965–1995. International Studies Quarterly. 2007;51(4):811–834.

10. Schoorl J, Heering L, Esveldt I, Groenewold G, Van der Erf R. Push and pull factors of international migration: a comparative report. 2000;.

11. Suleimenova D, Bell D, Groen D. A generalized simulation development approach for predicting refugee destinations. Scientific reports. 2017;7(1):13377.

12. Ahmed MN, Barlacchi G, Braghin S, Calabrese F, Ferretti M, Lonij V, et al. A Multi-Scale Approach to Data-Driven Mass Migration Analysis. In: SoGood@ ECML-PKDD; 2016. .

13. Sokolowski JA, Banks CM, Hayes RL. Modeling population displacement in the Syrian city of Aleppo. In: Simulation Conference (WSC), 2014 Winter. IEEE; 2014. p. 252–263.

14. Harrison E. Modeling Syrian Internally Displaced Person Movements: A Case Study of Conflict, Travel, Accessibility, and Resource Availability. 2016;.

15. Lu X, Bengtsson L, Holme P. Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences. 2012;109(29):11576–11581.

16. Cohen JE, Roig M, Reuman DC, GoGwilt C. International migration beyond gravity: A statistical model for use in population projections. Proceedings of the National Academy of Sciences. 2008;105(40):15269–15274.

17. 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, USA: Society for Computer

Simulation International; 2017. p. 19:1–19:9. Available from: http://dl.acm.org/citation.cfm?id=3106388.3106407.

18. Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the National Academy of Sciences. 2002;99(suppl 3):7280–7287.

19. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics New York, NY, USA:; 2001.

20. Centre TIDM. Global Report on Internal Displacement 2018; 2018. http://www.internal-displacement.org/global-report/grid2018/downloads/2018-GRID.pdf.

21. Hedeker D, Gibbons RD. Longitudinal data analysis. vol. 451. John Wiley & Sons; 2006.

22. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995;20(3):273–297.

23. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

24. Hajjem A, Bellavance F, Larocque D. Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation. 2014;84(6):1313–1328.

25. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. New York, NY, USA: ACM; 2016. p. 785–794. Available from: http://doi.acm.org/10.1145/2939672.2939785.

26. Haykin S. Neural Networks: A Comprehensive Foundation. Neural networks. 2004;2(2004):41.