Forecasting Internal Displacement Trends with Agent-based Models

Tyler Amos

6 June 2018

Abstract

Accurately forecasting forced migration trends allows humanitarian agencies to more efficiently position resources and respond to emergencies. In this paper, I present an agent-based model built on the FLEE environment, using the case study of Iraq (2017-01 to 2018-04). Results indicate the FLEE environment has potential for accurately predicting internal displacement flows in protracted displacement scenarios.

Can agent-based simulation accurately forecast the volume and geographic distribution of internal displacement? Building on recent work by Suleimenova et al. (2017), I use the FLEE agent-based modelling environment to study internal displacement in Iraq from January 2017 through to April 2018. I make four main contributions to the forced migration literature: i) I apply simulation methods, which are relatively rare in forced migration studies; ii) I develop granular, and therefore actionable, forecasts of internal displacement; iii) I validate the potential of the FLEE environment for simulating forced displacement scenarios; and iv) I establish benchmarks for the FLEE environment's parameters by optimization. The paper proceeds in three parts. First, I review a common formal model of migration, simulation methods in general, and agent-based models in particular. Second, I summarize the data and FLEE environment. Third, I discuss results and conclusions, from which I cautiously conclude agent-based models have the potential to accurately forecast internal displacement trends.

Formal Models of Migration: The Gravity Model

The most prominent of formal migration models is the gravity model. First used to explain economic migration (See Ravenstein 1885; in Edwards 2009), it is frequently employed in econometric studies of trade, with some applications to forced migration. (Iqbal 2007; in Edwards 2009) In this model "objects" are drawn to one or other locations by their "mass". The attractive power (mass) of a location is determined by some characteristic, usually population. (Edwards 2009) This attraction is then limited by distance, which is assumed to have a negative relationship with attraction; closer locations are preferred to those more distant.

(1)

$$I_{i,j} = \frac{f(R_i, A_j)}{f(D_{i,j})}$$

> Where I is the interaction between locations i and j, determined by R_i, repelling forces at location i and A_j, attraction at location j. $D_{i,j}$ is the distance between locations i and j. (Edwards 2009 pp 21)[1]

For its robustness across a number of applications, long history, and appealing simplicity, the gravity model's shortcomings are numerous. For one, it overemphasizes macro trends. (Edwards 2009 pp 21) Simini et al. (2012) identify a number of further issues: i) the wide latitude available in determining the cost function $f(D_{i,j})$; ii) poor predictive performance in certain applications; iii) an over-reliance on population, and; iv) a number of free parameters.[2]

The gravity model is drawn from economic studies of commuting and more routine forms of migration. This leads it to rely on certain assumptions commonly found in rational choice models. (Edwards 2009 pp 20-22) These models assume an individual chooses to move from one location to another on the basis of some calculus - weighing the benefits and costs of staying or leaving. (Edwards 2009 pp 16)

The rationality assumption can be a useful approximation of human decision-making in many environments. Indeed the gravity model performs well and has been successfully applied to forced migration phenomena. In forced migration studies, however, the objects of analysis are regularly coerced and operate in environments with poor information flow. As such, the rationality assumption may be a theoretical weakness. (Edwards 2009 pp 16)

To relax this assumption, some mixed models use elements of the above model with new approaches from areas such as network theory. In one informal variant, the attractiveness of a location is determined by its extant migrant population. Social ties between source and location are strengthened by successive migrant arrivals. (Lindstrom and Ramírez 2010; Garip and Asad 2016) More formal alternatives combine heuristics based on formal models with network or graph models. (Ahmed et al. 2016; Suleimenova, Bell, and Groen 2017) Edwards (2009) explores ways to account for a relaxed rationality assumption in detail, presenting an agent-based model on a generic lattice which combines elements of macro models (e.g., gravity) and observations from research at the micro level like the limitations of the rational choice assumption (e.g., informational assymmetry). Most recently, Suleimanova et al. (2017) preserve the rationality assumption but instead develop a network based on real locations and events with a set of rules based on the gravity model (1).

[1] See Simini et al. (2012) for a concise description of both the gravity and radiation models.

[2] Alternatives exist, such as the radiation, intervening opportunity or random utility models. (Simini et al. 2012)

Simulation in Forced Migration Studies

Simulation studies by researchers such as Suleimenova et al. (2017) have a number of interesting applications in forced migration studies. The principal benefits are fourfold: i) the ability to imitate experiments, which would be neither ethical nor practical (Hartmann 1996 pp 2-10; in Edwards 2009); ii) the ability to explore the internal dynamics of a phenomenon without comprehensive or highly reliable data (Ibid); iii) apart from time, little to no cost for the researcher, and; iv) no need for specialized knowledge beyond intermediate programming.

Simulation also implies certain drawbacks. Most germane to this discussion are those of generalizability and realism.[3] Simulations are questionably generalizable, as they are only valid under the specific scenarios (i.e., combinations of parameters and assumptions) used by the researchers. Furthermore, simulations are only approximations of reality, capturing but a fraction of the true complexity in social phenomena. In a purely theoretical study, this second objection is of less concern, but when a more applied orientation is desired, this shortcoming raises substantial questions about the real-world implications of simulation studies. Maldonado and Greenland (1997) propose results from simulations should be interpreted similar to clinical studies in medicine, with: i) great caution; ii) requirement for corroboration, and; iii) reference to real-world data.

Researchers using simulation in forced migration studies account for these challenges by employing complex, multi-layered models (Edwards 2009), or basing their simulations on real parameters (e.g., locations and distances between those locations) of a specific instance of the phenomena of interest. (Suleimenova, Bell, and Groen 2017) The challenge of demonstrating real-world validity can also be addressed by structuring a simulation to produce predictions which can then be compared to real-world data.[4]

Reasonable approximations of displacement via simulation are achieved through a number of methods. One common method is agent-based simulation. Agent-based simulations require the modeller to specify all rules by which individuals make decisions, which are then explicitly coded into the situation as decision rules for the agents. In a displacement simulation, agents (a household or individual) move across a virtual space and interact with elements of the simulation according to some set of rules.[5]

Many of the critiques of simulation generally, such as those from Maldonado and Greenland (1997) apply to agent-based models in particular. The rules by which agents “live” may be unreasonable simplifications or require bold assumptions which, in the worst case, limit simulation results’ generalizability to just other simulations. This is notable given one of the reasons for choosing simulation methods is to manipulate virtual ecosystems to enhance our understanding of

[3] See Maldonado and Greenland (1997) pp 454-455 for a more extensive discussion.

[4] One example of this strategy is found in Suleimenova, Bell, and Groen (2017), and is the strategy I adopt here.

[5] See Edwards (2009) for an accessible explanation of agent-based models.

reality.

Formal models borrowed from economics (gravity), and simulation methods like agent-based models have potential to produce valuable insights in forced migration studies. To date, a substantial amount of empirical work in this space focuses on causal drivers of displacement. Researchers investigate how significant events, regimes, or geographical scope and intensity of conflict effect long-term movement patterns. (Schon 2015; Melander and Öberg 2007; Iqbal 2007) These studies produce insightful results about broad displacement trends. However for applications like humanitarian response, more granular results about volume and geographic distribution of displaced people in-country are valuable. There is thus an unmet need for actionable insights at a relatively granular scale.

This study seeks to address this gap through the application of the model and methods described above. Specifically, using an agent-based simulation, I develop governorate-level forecasts for internal displacement volume in Iraq based on data from United Nations agencies and non-profit monitoring groups.

Data

The principal data sources for this analysis are: (i) records of the volume and geographic distribution of internally displaced people collected by the International Organization for Migration (IOM); (ii) records of violent incidents from the Armed Conflict Location and Event Database (ACLED), and; (iii) spatial data for populated locations from the United Nations Office for the Coordination of Humanitarian Affairs (UNOCHA).[6]

IOM Displacement Tracking Matrix (DTM)

IOM conducts regular surveys of the location and number of displaced households in Iraq. Surveys are conducted approximately every two weeks as part of IOM's assessment system.[7] For this simulation, the IDP Master Lists for rounds 84 (November 29, 2017) through 91 (April 30, 2018) were used. On average, each round of the survey has 2,436 cases.

IDP Master Lists have consistent formats across rounds, which motivated their selection over more granular round-specific reports. Each Master List provides updated figures of the number of households at each reported location, as well as adding new locations as they are surveyed. Master Lists were downloaded as MS Excel documents and converted to CSV format.

Armed Conflict Location and Event Database (ACLED)

ACLED is an initiative which catalogues incidents of violence across a number of countries. ACLED data are frequently used by researchers studying conflict/crisis. For this simulation, ACLED data for Iraq were accessed from the Humanitarian Data Exchange Portal's live update link as a CSV. (Exchange, n.d.)

[6] Links to download all data files are available in the Appendix.

[7] See Migration (n.d.) for details of the survey methodology.

For each event, ACLED records the approximate location, date and time, estimated fatalities, a short description, as well as other features. This simulation uses the approximate location and date in the simulation environment. Events with no estimated fatalities were removed from the data under the assumption they are not indicative of a level of conflict sufficient to provoke new displacements. In the complete dataset there are 6,354 cases, 3,730 of which involve at least one fatality. Of these, 704 fit into the specified time period and were used in the simulation to "switch on" locations' conflict status in the appropriate round.

Populated Locations in Iraq

Spatial data on the location of 23,991 populated places in Iraq was collected from UNOCHA via the HumData portal in ShapeFile format. (Coordination of Humanitarian Affairs, n.d.) This dataset was chosen specifically as it is compatible with the Displacement Tracking Matrix and is derived from IOM's internal placename database. It provides the names and locations of not only official settlements, but neighbourhoods and other unofficial locations, allowing for regional centroids to be weighted by density of settlements, not area.

Source	Usage	Start	End
IOM	IDP Location	2017-11-29	2018-04-30
IOM	IDP Population	2017-22-29	2018-04-30
ACLED	Location Type	2017-01-01	2018-04-28
UNOCHA	Network Node Location	NA	NA

Data Joining and Aggregation[8]

To align the three data sources in the desired format, a series of aggregation and spatial join operations were performed. The spatial geometries of IDP, event, and populated places were transformed into representative polygons using the GeoPandas convex hull functionality after being aggregated to the level of one of Iraq's 18 first-level sub-national administrative regions (governorate/province). On these transformed datasets, two sets of spatial joins were used. First, the populated locations were joined with the observed IDP populations data. Population numbers were summed by region to establish the initial and future populations of each location. Second, the populated locations, aggregated to the regional level and represented geometrically as a convex hull, were joined with ACLED records involving at least one fatality. Population location geometries were then transformed into centroids for use as node locations. Lastly, the distance between these centroids was calculated using the GeoPandas `distance` function.

These joins and aggregation produced four datasets: (i) population centres represented as a point, with starting IDP populations (nodes); (ii) final observed

[8] Algorithms used to perform these operations are available in the Appendix.

IDP populations by population centre; (iii) conflict locations and their date of observation, and; (iv) distance between all population centres (edges).

Missingness

The validation period used in this iteration of the simulation uses observed values from round 91 (late April 2018) to calculate error rates. Round 91 of the Displacement Tracking Matrix does not contain IDP totals for all governorates. The error rate used is calculated based on those governorates for which there are observed values. In future iterations, alternative error calculations will be explored.

Models and Methods

The FLEE Agent-based Modelling Environment

FLEE is a purpose-built Agent-based Modelling (ABM) environment for simulating the flow of people. (Suleimenova, Bell, and Groen 2017) The initial development of the environment has focused on modelling forced displacement, specifically refugee movements.[9] In FLEE, agents traverse a network where each node represents a town, camp, or conflict. Agents follow a series of rules in order to determine where they will travel where conflict and distant locations are less likely to be selected, and non-conflict and proximate locations are more likely.

In the simulation environment, each agent represents a household (family). At each step of the ecosystem, in this case a 2-week period, agents navigate the ecosystem according to a set of rules inspired by the gravity model of migration. In short, under the gravity model the relative attractiveness of a location is a function of the population size of the destination location and the distance to that location. In this simulation, the population size is the number of internally displaced people, and the distance is the euclidean distance between points as calculated by the GeoPandas `distance` function. A fixed number of agents (100) are added to locations at random once per step. Agents at those locations then decide to stay or move based on the population of their current location and the distance to other locations, as in the gravity model. In this simulation, there were seven possible parameters which could be adjusted.

Parameters Varied In the Simulation

Name	Description
`CampWeight`	The factor by which camps 'attract' agents.
`ConflictWeight`	The factor by which conflicts 'repel' agents.
`MinMoveSpeed`	Minimum distance an agent covers in one step.
`MaxMoveSpeed`	Maximum distance an agent covers in one step.

[9] The principal distinction between refugees and internally displaced people is refugees have crossed an international border while internally displaced peope have not.

Name	Description
`ConflictMoveChance`	Default probability for leaving a conflict zone.
`CampMoveChance`	Default probability for leaving a camp.
`DefaultMoveChance`	Default probability for leaving any location.

Parameters and Constructs

Name	Construct (Gravity Analogy)
`CampWeight`	The attractive power of camps (mass)
`ConflictWeight`	The repellent factor of conflicts (mass)
`ConflictMoveChance`	Agents' decision to leave a conflict.
`CampMoveChance`	Agents' decision to leave a camp.
`DefaultMoveChance`	Agents' decision to leave any given location

Apart from these parameters, the simulation was set so that agents introduced added to existing populations (`TakeRefugeesFromPopulation = False`), camp weights were dynamically calculated based on the agent population in the camp at each step (`UseDynamicCampWeights = True`), agent awareness was limited to their location (`AwarenessLevel = 1`), IDP mode was enabled (`UseIDPMode = True`), and agents did not accumulate knowledge about the network over time (`UseDynamicAwareness = False`).

The ecosystem was initialized with the processed list of locations (i).[10] The distances between nodes were used to create links between locations (iv). Throughout the simulation, if a location was included in the list of conflict locations for that step, the location was changed to a 'conflict' zone, and the weights agents use to implement their decision function were re-calculated.

Error Metric and Objective Function

To evaluate the appropriateness of the simulation parameters, an error function was calculated from the difference in forecasted proportions of displaced people in each governorate and the true observed proportions of IDPs in each governorate. As in Suleimenova et al. (2017), this error metric was the Mean Absolute Scaled Error (MASE).

$$MASE = \frac{1}{N_L} \sum_{l_i}^{L} \frac{1}{T} \sum_{t=1|l_i}^{T} \left(\frac{|r_t - f_t|}{\frac{|r_t - r_{t-1}|}{T-1}} \right)$$

[10] See "Data Joining and Aggregation" for the corresponding values associated with these numerals.

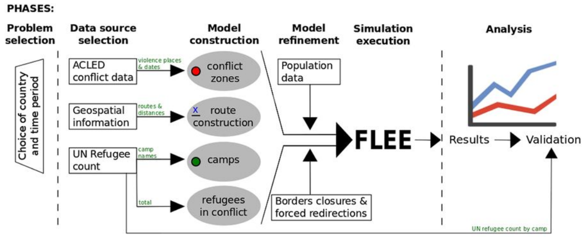

Figure 1: Illustration of original FLEE simulation workflow. Source: Suleimenova, Bell, and Groen (2017)

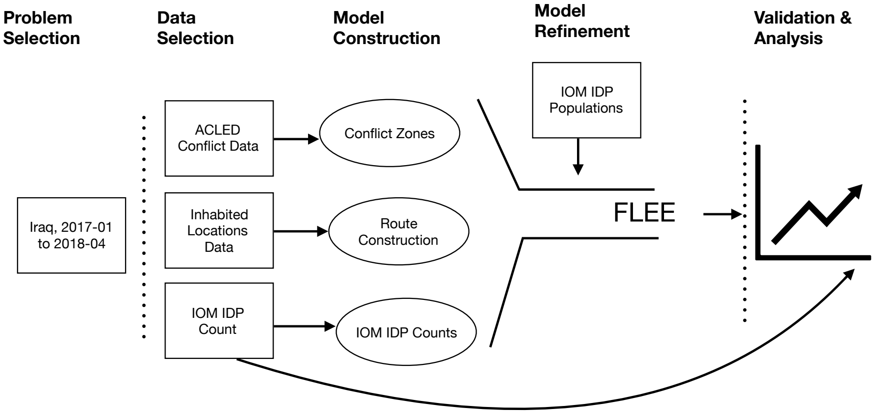

Figure 2: Illustration of modified FLEE workflow.

Where N_L is the number of locations, l_i is a given location in L, T is all time periods, r_t is the observed values of a given location at time t, f_t is the forecasted values of a given location at time t. A MASE of 10% should be interpreted as a 10% improvement in forecasting accuracy for time t as compared to simply using time $t-1$ as the forecast for time t.

Results

Algorithmic optimization of the simulation parameters was done using three different methods. The first, the basin hopper algorithm,[11] produced a minimum MASE of 58%. A simple optimizer using the Broyden-Fletcher-Goldfarb-Shanno algorithm also produced a minimum MASE of 59%. Heuristic optimization, based on the values used in Suleimenova et al. (2017) produced similar results, with a minumum of 57%. A brute force algorithm was also applied, but quickly became computationally intractable.

The relative consistency of results across optimizers supports assertions made in Suleimenova et al. (2017) that the FLEE environment has weak sensitivity to most simulation parameters.[12]

Method	Minimum MASE
Basinhopper	0.58
BFGS	0.59
Heuristic	0.57

Limitations

In the interest of reducing computational complexity, regional geographies were simplified to simple centroids, calculated using the algorithms noted above. However, in a country such as Iraq, where the majority of populated locations are not uniformly distributed[13] this simplification may involve substantial information loss. The threshold for a location to be considered a "conflict" location was set quite low in this simulation (one fatality, as reported by ACLED). This may have resulted in a liberal categorization of locations as "conflict".

Conclusion

These results indicate the FLEE environment, originally developed for simulating initial refugee displacement, appears to perform well in simulating protracted internal displacement scenarios. Moreover, it confirms the conclusions of

[11] See https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.basinhopping.html

[12] See the Appendix for the specific paramter values for each method.

[13] See Appendix, "Populated Locations in Iraq".

Suleimenova et al (2017) that the FLEE environment generalizes beyond just one country.

The significance of these results is twofold. First, the extension of the environment to a non-African country, Iraq, suggests FLEE is robust to changes in geography. Second, the application of FLEE to a protracted displacement scenario, as opposed to the inital displacement scenario presented in Suleimenova et al. (2017), suggests FLEE is also robust to changes in displacement dynamics.

Future research could validate these conclusions with a variety of error metrics and against some more robust alternatives such as random walks. For example, there is some indication that random walk simulations may predict displacement trends with a high degree of accuracy. (Migration 2017)[14] More sophisticated methods for calculating distances between points could also be applied. Lastly, modifications to the decision function used by agents in FLEE could be made, such as using a radiation model-based implementation rather than the current gravity model-inspired method.

Appendix

Data Sources

IOM Displacement Tracking Matrix: http://iraqdtm.iom.int

Populated Places in Iraq: https://data.humdata.org/dataset/settlements-villages-towns-cities

Iraq Administrative Boundaries: https://data.humdata.org/dataset/iraq-admin-level-1-boundaries[15]

ACLED Event Data: https://www.acleddata.com/data/

ACLED Data is also available with live updates from: www.data.humdata.org

Algorithms

Aggregation of Populated Locations

```
generate spatial geometries of all populated locations
dissolve geometries by Governorate
convert geometries to convex hulls
spatial join with second dataset (if needed)
```

Creation of new geometries

[14] This claim is yet to be tested in a peer-reviewed publication.

[15] This data was used for visualization purposes only.

```
load non-spatial data with requisite spatial features (lat, lon)
for each lat, lon pair
    create geometry
assign new geometries as a feature of the dataset
```

Calculation of edge lengths

```
load dissolved geometries of Populated Locations
convert geometries to centroids (a single point)
for each row
    calculate the distance between the row's geometry and all other geometries
    convert distances to an integer value
    for each trip
        if the start and end are equal
            move to the next trip
        if the trip has not yet been seen
            add the trip to a master trip list
            write the start, end, and distance to a file
```

Identification of Conflict Zones

```
load ACLED conflict data
remove cases with no fatalities
create new geometries (above)
sort by event date
convert event dates to integers
remove cases with event dates before the lower date bound
remove cases with event dates after the upper date bound
assign round number, equivalent to step in simulation
write location names and steps to file
```

Creation of test dataset

```
aggregate population locations
load final observed dataset
create geometries for final observed dataset
spatial join observed counts with population locations
sum observed counts by governorate
write to file
```

Links to Code

Code used to produce the simulation environment, clean, and represent data are available at: https://github.com/tamos/MACS30200proj.

Figures

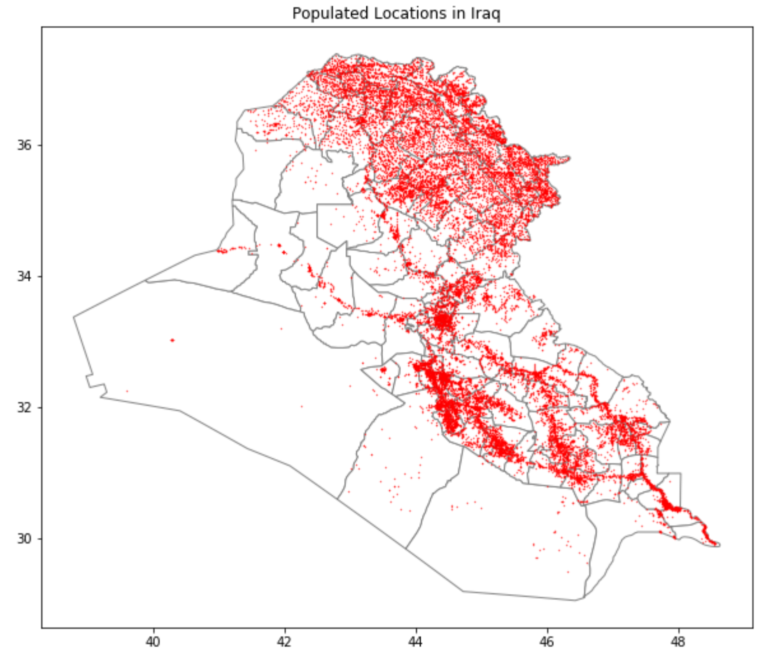

Figure 3: Iraq Populated Places with level 1 administrative boundaries. Source: UNOCHA

Optimized Parameter Values

Parameter	Basinhopper	BFGS	Heuristic
`CampWeight`	1.45814991	0.9	1.0
`ConflictWeight`	0.62591585	0.24	0.25
`MinMoveSpeed`	0.35781887	0.5	0.5
`MaxMoveSpeed`	5.00099374	4.9	5.0
`ConflictMoveChance`	1.1834612	0.9	1.0
`CampMoveChance`	0.31270562	0.09	0.1
`DefaultMoveChance`	0.17024767		0.29

References

Ahmed, Mohammed N, Gianni Barlacchi, Stefano Braghin, Francesco Calabrese, Michele Ferretti, Vincent Lonij, Rahul Nair, Rana Novack, Jurij Paraszczak, and Andeep S Toor. 2016. "A Multi-Scale Approach to Data-Driven Mass Migration Analysis." In *SoGood@ Ecml-Pkdd*.

Coordination of Humanitarian Affairs, United Nations Office for. n.d. "Iraq - Settlements (villages, towns, cities)." https://data.humdata.org/dataset/settlements-villages-towns-cities.

Edwards, Scott. 2009. *The Chaos of Forced Migration: A Means of Modeling Complexity for Humanitarian Ends.*

Exchange, Humanitarian Data. n.d. "Iraq - Conflict Data." https://data.humdata.org/dataset/acled-data-for-iraq.

Garip, Filiz, and Asad L. Asad. 2016. "Network Effects in Mexico–U.s. Migration: Disentangling the Underlying Social Mechanisms." *American Behavioral Scientist* 60 (10):1168–93. https://doi.org/10.1177/0002764216643131.

Hartmann, Stephan. 1996. "The World as a Process." In *Modelling and Simulation in the Social Sciences from the Philosophy of Science Point of View*, 77–100. Springer.

Iqbal, Zaryab. 2007. "The Geo-Politics of Forced Migration in Africa, 1992–2001." *Conflict Management and Peace Science* 24 (2). Taylor & Francis:105–19.

Lindstrom, David P., and Adriana López Ramírez. 2010. "Pioneers and Followers: Migrant Selectivity and the Development of U.s. Migration Streams in Latin America." *The ANNALS of the American Academy of Political and Social Science* 630 (1):53–77. https://doi.org/10.1177/0002716210368103.

Maldonado, George, and Sander Greenland. 1997. "The Importance of Critically Interpreting Simulation Studies." *Epidemiology* 8 (4). Lippincott Williams & Wilkins:453–56. http://www.jstor.org/stable/3702591.

Melander, Erik, and Magnus Öberg. 2007. "The Threat of Violence and Forced Migration: Geographical Scope Trumps Intensity of Fighting." *Civil Wars* 9 (2). Routledge:156–73. https://doi.org/10.1080/13698240701207310.

Migration, International Organization for. 2017. "Prepositioning relief: Using a random walk model to predict the distribution of IDPs in Nigeria." https://displacement.iom.int/content/prepositioning-relief-using-random-walk-model-predict-distribution-idps-nigeria.

———. n.d. "Methodology." http://iraqdtm.iom.int/Methodology.aspx.

Ravenstein, E. G. 1885. "The Laws of Migration." *Journal of the Statistical Society of London* 48 (2). [Royal Statistical Society, Wiley]:167–235. http://www.jstor.org/stable/2979181.

Schon, Justin. 2015. "Focus on the Forest, Not the Trees: A Changepoint Model of Forced Displacement." *Journal of Refugee Studies* 28 (April).

Simini, Filippo, Marta C González, Amos Maritan, and Albert-László Barabási. 2012. "A Universal Model for Mobility and Migration Patterns." *Nature* 484 (7392). Nature Publishing Group:96.

Suleimenova, Diana, David Bell, and Derek Groen. 2017. "A Generalized Simulation Development Approach for Predicting Refugee Destinations." *Scientific Reports* 7 (1). Nature Publishing Group:13377.