Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration

Lisa Singh, Laila Wahedi, Yanchen Wang, Yifang Wei, Christo Kirov, Susan Martin, Katharine Donato, Yaguang Liu, Kornraphop Kawintiranon*
Georgetown University
Washington, DC, USA
{lisa.singh,law98,yw516,yw255,ck746,martinsf,kmd285,yl947,kk1155}@georgetown.edu

ABSTRACT

Worldwide displacement due to war and conflict is at all-time high. Unfortunately, determining if, when, and where people will move is a complex problem. This paper proposes integrating both publicly available organic data from social media and newspapers with more traditional indicators of forced migration to determine when and where people will move. We combine movement and organic variables with spatial and temporal variation within different Bayesian models and show the viability of our method using a case study involving displacement in Iraq. Our analysis shows that incorporating open-source generated conversation and event variables maintains or improves predictive accuracy over traditional variables alone. This work is an important step toward understanding how to leverage organic big data for societal–scale problems.

CCS CONCEPTS

• **Information systems → Data mining**; • **Human-centered computing → Social engineering (social sciences)**;

KEYWORDS

forced migration; Bayesian; open source data; text mining

ACM Reference Format:
Lisa Singh, Laila Wahedi, Yanchen Wang, Yifang Wei, Christo Kirov, Susan Martin, Katharine Donato, Yaguang Liu, Kornraphop Kawintiranon. 2019. Blending Noisy Social Media Signals with Traditional Movement Variables to Predict Forced Migration. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3292500.3330774

1 MOTIVATION

Worldwide displacement due to war and conflict is at all-time high. UNHCR [18] reports that one in every 113 people is a refugee, internally displaced, or seeking asylum. Currently, more than 3 million persons are still displaced in Iraq, and more than 6 million persons are displaced in Syria [12]. As millions have become displaced in the Middle East and elsewhere, many countries were surprised by the volume of asylum seekers or refugees at, or near, their borders. No early warning system existed to warn governments when and how many people would be on the move. If some warning occurred, government and international aid community responses might have more effectively addressed the causes of displacement or provided safer alternatives to flight before millions left their homelands and many risked their lives crossing the Mediterranean.

Unfortunately, determining if, when, and where people will move is a complex problem. The literature on the causes of migration suggests that the scale, duration, and geographical location of violence or some other type of contextual pressure together with other social factors are major determinants of displacement [16]. Even though a number of theoretical frameworks of movement exist, getting data for these variables is difficult, if not impossible because of the difficultly of collecting data in conflict zones. To meet this challenge, this paper proposes integrating both publicly available organic data from social media and newspaper outlets with more traditional migration data to capture changing conditions in a region. More specifically, we combine variables extracted from text (indirect indicators of variables that cannot be captured easily) with more traditional movement variables (variables that are collected by government agencies and NGOs). We believe this form of *data blending* is crucial to tackle large-scale societal problems for which large data gaps exist. We then combine all these variables (with spatial and temporal variation) within different Bayesian models to predict when and where people will migrate. Our central case study is Iraq because of the magnitude of displacement, the difficulty associated with on the ground data collection, and the scale of resources being mobilized to respond to the humanitarian crisis.

Our contributions to the field are as follows: 1) we use organic data sources to extract these meaningful indirect indicators of movement and show their value in a case study related to Iraq; 2) we present experiments comparing different modeling approaches that integrate direct and indirect indicators and show that we can predict movement between 2015 and 2017; and 3) we make our generated social media variables, and interactive comparative visual analytics available to the migration community [39].

2 RELATED LITERATURE

2.1 Current Models of Migration

Prior studies have attempted to build mathematical models of migration in response to conflict [13, 29, 30, 36, 42], natural disaster

*Singh, Wahedi, and Wang contributed equally to this research.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330774

[23], climate change [34], economic underdevelopment [28], and human rights violations [35]. These models are designed to test theory with limited data samples, but temporally fine-grained, localized data are difficult and expensive to collect. For example, death counts [25] and destruction of property [1] are rarely collected at the local level and, if available, tend to be annual national measures. Moreover, the reliability of such data is even more limited [14], especially in conflict zones.

This lack of available micro-level direct indicators means that studies examining micro-drivers tend to be based on survey research after a conflict has ended [1], to be qualitative, or to have a narrower scope such as those on specific conflict dynamics [3] or environmental shocks (e.g. [9]). Broader studies examining the interaction between multiple drivers, on the other hand, tend to be limited to annual-country level data [13, 32]. So far, there has been little integration between micro-level drivers and broader system-level approaches [38]. What is needed is a way to fill the data gaps, and to incorporate variables from multiple domains and levels of analysis, from micro-drivers to macro-drivers, from factors ranging from conflict to environmental to economic.

2.2 Big Data Use for Social Causes

Data scientists and social scientists have begun to use social media data, and more generally, big data to tackle similar problems in other areas [19, 26]. Here we highlight a few examples. Blumenstock [6] used data from mobile phone networks to estimate shifts in national population distributions in Rwanda. Following the 2010 Haitian earthquake and cholera outbreak, Bengtsson et al. [4] estimated population mobility shifts using the daily position of SIM cards tracked by mobile phone towers. Deville and colleagues [15] monitored seasonal population changes in France and Portugal. LinkedIn data has been used to investigate labor market migration in parts of the U.S. [41]. These data have also been shown to be a good source of data for monitoring the diffusion of epidemics and the effectiveness of public health measures [20]. The United States Geological Survey shows an approach for tracking earthquakes using Twitter data [17]. Finally, EMBERS is a system for monitoring civil unrest that combines social media and other publicly available data. It predicts when, where and why protests occur [33]. EMBERS is the closest work to our work since it also uses social media data for its task. The methods we use and the variables we construct differ for predicting movement, partially, because online conversations about movement itself is not as readily available as discussions about civil unrest. This is one important reason why we need to capture indirect indicators of movement.

3 METHODOLOGY OVERVIEW

Figure 1 shows our high level approach for tackling this complex issue. We begin by working with migration researchers to understand the broad factors people consider when deciding whether or not to migrate (Step 1). This resulted in a factor model containing sixteen push and pull factors, including national and regional macro-level factors (economic, political, environmental), intervening meso-level factors (relief, weather, infrastructure), and individual household micro-level factors (household demographics, capital) [27]. Each factor can be operationalized as a set of variables generated from

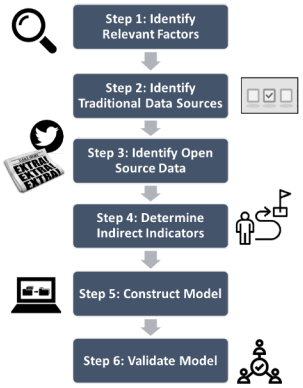


Figure 1: High Level Methodology

different data sources. For example, a traditional environment variable may be *amount of precipitation*, while *death count* may be a political/social variable. Therefore, our next step is to identify data sources for relevant variables (Step 2). Unfortunately, data collection and survey research are dangerous and challenging in conflict zones and regions with high poverty or political instability.

To fill this gap, we identify alternative sources of data (social media and newspaper) that can serve as indirect indicators of movement for variables that cannot reasonably be collected directly (Step 3). There are a number of strengths associated with using these data. 1) These data are timely. In places where people use social media, they use it before, during, and after conflict. 2) These data are not controlled: anyone with Internet access may communicate via social media and become an information provider. 3) In certain conflict regions, social media data may be the only detailed data to which we have access.

Of course, social media data sources also have potential problems. These data are noisy, and can have biases. Samples of such data are not typically representative of an entire population, since not everyone uses social media and levels of penetration vary in different locations. Even with these limitations, when events occur (e.g., a factory has reopened in Baghdad) or a situation is deteriorating in a region (e.g. a bridge has been blown out in Ninewa governorate), online conversations do emerge and these discussions, even if they are not specifically about migration, capture indirect indicators of one or more migration variables. Step 4 focuses on determining which conversation topics/signals are reasonable indirect indicators for specific conflict situations. For example, is conversation about violence in Iraq a reasonable proxy for death counts - does a relationship between the two exist? This step helps us map our indirect signals to the variables that have been identified as theoretically important or as proxies for gaps in the data. This in turn makes our modeling approach useful for both prediction, and understanding relationships among drivers of movement.

Next, we consider different Bayesian models for generating movement predictions (Step 5). Most modeling of migration decisions lacks a formal mathematical framework. Although this is unsurprising given the complexity of migration, we are interested in comparing different models to better understand the types of models that are most informative. Because we have data at different

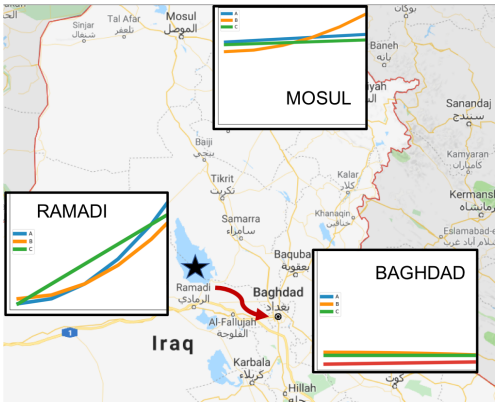


Figure 2: Example of the Changing Dynamics of Factors

spatial scales and temporal resolutions, we need models that can handle this variation effectively. For this reason we focus on hierarchical models. Finally, we must validate our results using available ground truth data or through manual validation (Step 6).

4 CAPTURING USEFUL BIG DATA SIGNALS

Social media and newspaper data give us an opportunity to learn what people are talking about, what events are occurring, and how people's perceptions are changing about a situation. How do we use these readily available data sources to identify meaningful forced-migration related signals? Our approach requires an understanding of the importance and the changing dynamics of each migration factor/driver in a particular location. Figure 2 is an illustration of what we are trying to capture. Suppose we are monitoring movement variables in three locations in Iraq - Mosul, Ramadi and Baghdad. Let's also suppose that larger values of the shown variables are an indication of a worsening situation. Our figure shows people in Ramadi (see star) dealing with worsening conditions (increases in all variables values). If they are considering moving north to Mosul, they will face bad conditions (high variable values). If they move south to Baghdad, the situation is better (constant, low variable values). In this section, we show an example case explaining how we can capture some of these spatio-temporal movement variables using open source data sets.

4.1 Events, Buzz, & Perception

While we will continue to augment our list of signals/variables captured from open-source data, the dynamics we consider in this paper are drawn from events, buzz, and perception. Events of interest are those that occur in a particular location and are associated with one or more factors/topics identified by social scientists as important for predicting movement. Tracking the frequency of these events allows us to compute a time series containing the number of events related to forced migration factors by day, week, or month. Buzz represents the amount of interest in a topic [21]. Topics are constructed by migration experts to reflect drivers of migration. Buzz about a topic may be high, low, increasing, or decreasing. We are interested in the variation of buzz strength of a topic over time. Finally, we are interested in understanding people's perceptions about relevant direct and indirect indicators, e.g. wages, schools, etc. Perceptions can be measured in different ways. Three that are important in the context of migration are tone (sentiment), stance (position - for or against), and emotion.

Over the last two years, much work has focused on extracting these types of signals from newspaper and public social media data, including detecting relevant dynamic topics and events from newspapers [2, 11, 21, 31, 44, 45]. We take advantage of this previous work to extract topic buzz, sentiment (as a proxy for perception), and event volume as an initial set of social media and newspaper related signals. We translate these signals into longitudinal variables that can be monitored daily, weekly, or monthly depending on the level of granularity of the data and the model. For example, we can compute the buzz about a topic by determining the number of posts associated with that topic in a particular location of interest, or the event volume by counting the number of events identified in newspapers/social media that map to different factors, e.g. political violence events. Because we are interested in understanding both when and where people will move, we need to capture signals for both source and destination locations. While newspaper articles generally contain the locations that the articles refer to, social media posts are less consistent. There are three possible ways to capture location in social media data: based on geolocated posts, based on the location of the user posting, or based on mentions of locations in the post text. In this work, we focus on geolocated tags and mentions of locations in posts (given the scale of our data).

4.2 Identifying Indirect Indicators – Example

While there is insufficient space to go through the process associated with identifying each social media/newspaper variable that can be used to represent a movement variable, we describe the process using a small case study - using social media data for understanding the relationship between displacement in Iraq and chatter and tone on Twitter about the Islamic State (ISIS). We focus on six signals from Twitter and their relationship to different local and regional dynamics. The data in this case example have been collected using the Twitter Streaming Application Programming Interface (API) since late 2014. During this time, we have collect tweets containing the hashtag #ISIS, and #ISIS in Arabic. We have over 45 million tweets with these hashtags over a 3 year period. After preprocessing, and governorate location identification, we focus in on 2.6 million tweets that were mapped to a governate in Iraq. To keep this simple, we construct two types of variables, volume and sentiment.[1] The longitudinal tweet variables were aggregated into one-month periods with a given governorate as their primary location. The variables were the following: total volume of tweets, total count of tweets classified as positive, and total count of tweets classified as negative, for both English and Arabic. This resulted in three Arabic and three English variables. Because we are interested in these variables as leading indicators about displacement, we lag all the variables by one month.

We compared these social media signals against a common traditional variable: monthly conflict-related deaths, curated by Iraq-BodyCount.com. We know from the literature that conflict-related

[1] Our analysis presented in Section 6 is much more detailed, looking at variables at both the factor and location level. Here we show the variables across each location. Recall, the goal of this section is to show that these signals are identifiable and reasonable as indirect indicators for some variables that can be hard to capture.

violence is a major driver of movement [25]. Because our movement variable, described in more detail below, captures people once they have already moved and checked in, we lag the death count by one month, using it as a leading indicator of movement. Much of the displacement in Iraq from 2014-2017 was due to ISIS related violence. Here we ask the following – to what extent do changes in the level of discussion of ISIS capture the variation in death counts. Understanding this relationship will allow us to study conflict dynamics in other regions where death counts are not available.

Our outcome variable of interest is the number of families fleeing a governorate in a particular month. We use the Displacement Tracking Matrix from the International Organization of Migration (IOM) to determine the number who have moved. Finally, for this example we use a Bayesian negative binomial regression with random intercepts for month and governorate.[2]

Figure 3 is a coefficient plot showing the substantive effect of each signal with 95% credible intervals. From each of seven random-effects negative binomial models, we present three coefficient estimates to demonstrate consistency and model convergence. We report substantive effects, which are calculated from the exponent of a negative binomial regression coefficient. They represent the percent change in the outcome variable expected with a 1-unit change in the explanatory variable. We begin with the effect of conflict-related deaths in a given location and month on displacement from that location in the following month. As shown in the final row of Figure 3, one additional death per thousand population is associated with an 11.1 percent increase in displacement. Given the literature on forced migration, it makes sense that death is a strong indicator of movement.

Figure 3 also reveals a relationship between each of our constructed Twitter variables and movement. The most significant relationship is the negative sentiment of the Arabic #ISIS tweets mentioning a location in a given month and the number of people who flee from that location in the following month. An additional hundred negative sentiment tweets in a month is associated with 10.4 percent more families displaced the following month. This is consistent with previous literature that suggests the more public opposition to a group, the more movement when the group takes control of the territory [3]. The Arabic signal is understandably stronger, and more reliable, given Arabic is the spoken language in Iraq and neighboring countries. Overall, tone is more reliable a signal than volume in Arabic. Tweet sentiment is a much weaker signal in English, where the vast majority of tweets are negative. Moreover, when death is included in the same model as these Twitter variables, the effect of the Twitter variables is reduced almost to 0, which suggests that the signals are capturing some of the same variation as death. These findings suggest that Arabic sentiment and tweet volume are indirect indicators of movement associated with conflict factors. Section 6 further explains how different signals affect movements.

This case was one example of step 4 in our methodology. By examining the effect of a social media signal in a model that controls for other structural variation (see Section 5), we isolate the effect of variation in that variable. We then examine the spatial and temporal correlation between that variable and the theoretical variable of interest. Finally, we use them together to plot movement. If the variation that had been explained by the theoretical variable is explained by the social media signal (e.g. if the coefficient estimate drops) then we know we have captured that signal, and may be able to use it as an indirect indicator of a theoretically important factor in order to better study that factor. In this example, there is a clear, indirect relationship between the Arabic Twitter signals and movement, and Arabic sentiment appears to capture much of the variation explained by death counts. Given the difficulty of collecting real-time, accurate death counts in conflict zones, uncovering this relationship should help us to study the relationship between conflict and migration in more detail than previously possible.

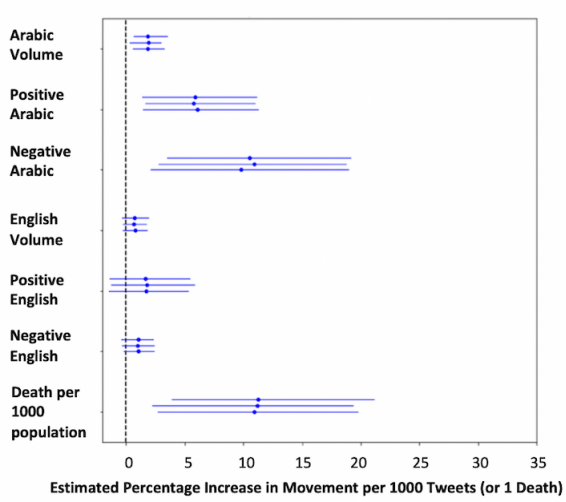


Figure 3: Twitter Signal Effects with 95% Credible Intervals

5 PREDICTION USING A HIERARCHICAL BAYESIAN APPROACH

Our prediction task is to estimate the number of families moving from one location to another. We consider movements as origin and destination pairs,and refer to this model as a *dyadic model*.

While there are many modeling options for these prediction tasks, we use a hierarchical Bayesian approach for three reasons. 1) Our data have temporal (daily, weekly, monthly, annual, static), and spatial (district, governorate, country) variation. Hierarchical models allow us to incorporate data at multiple levels of analysis, without using repeated observations. 2) Bayesian estimation of mixed hierarchical models make specifying complex dependence structures between these levels of data explicit. Instead of giving each governorate an independent fixed effect as in traditional panel analysis, we specify a distribution that relates these effects to one another. This is important for generating an accurate representation of uncertainty in predicted outcomes. 3) Predictions produced by Bayesian models are posterior probability densities rather than point predictions. This means that rather than telling you the number of families that will flee, the model gives you the probability that any possible number of families will flee. This can be used to give aid workers a credible range of the amount of displacement to expect, or to tell them the probability that there will be more

[2]In our full empirical analysis, we consider a number of different models. To show the viability of our methodology for this example, we show only one model.

than n families displaced in the following week. Having explicit posterior probabilities will help them to better manage risk in the distribution of limited aid resources.

Because we are predicting counts, we will compare the following three count models: the Negative Binomial, zero-inflated Poisson, and zero-inflated Negative Binomial. The Negative Binomial is effective for capturing over-dispersion of the outcome variable by not forcing the mean and variance to be the same, like the Poisson. Because large-scale movement does not tend to happen all over the country at the same time, we have a large number of zeros and want to consider models that account for this.

While we do not have space to show the equations associated with all the models, a generalized representation of the model is shown below. The count of families is distributed (Z) as either negative binomial, zero-inflated Poisson, or zero-inflated negative binomial, where λ is the expected count, α represents the dispersion parameter for the negative binomial family, and ψ represents the proportion of zeros for the zero-inflated family. β is the coefficient for covariates at the highest time resolution for origin-destination pairs. Examples might include differences in social media signals between locations. $\hat{\eta}$ are random intercepts for each origin-time, and are parameterized to incorporate η coefficients for temporal signals within each origin location, such as the sentiment of tweets in each governorate. ζ and $\hat{\zeta}$ are the corresponding parameters for destination temporal district signals. Finally, $\hat{c}$ and $\hat{d}$ are the random intercepts for destination districts, with c and d as a vector of coefficients for static location variables such as population.

Generalized Dyad Model:

$$P(Y) \sim Z(\boldsymbol{\lambda}, \alpha, \psi)$$
$$\lambda_{s,r,t} = exp(\boldsymbol{\beta X_{s,r,t}} + \hat{\eta}_{s,t} + \hat{\zeta}_{r,t} + \hat{c}_s + \hat{d}_r)$$
$$\hat{\eta} \sim N(\mu_\eta + \boldsymbol{\eta S_T}, \sigma_\eta)$$
$$\hat{\zeta} \sim N(\mu_\zeta + \boldsymbol{\zeta R_T}, \sigma_\zeta)$$
$$\hat{c} \sim N(\mu_c + \boldsymbol{cS}, \sigma_c)$$
$$\hat{d} \sim N(\mu_d + \boldsymbol{dR}, \sigma_d)$$

We want to ensure that our model allows us to examine the effect of a particular signal in a particular place at a particular time. This means that we need to consider both spatial and temporal dependencies. In all models, we include spatial random intercepts for the governorate of the origin and the destination district. We also test a variant model that uses a conditional auto-regression function (CAR) to account for the distance between locations, encoding the expectation that places near one another, at similar points in time, will be more similar than those places that are farther apart. (CAR models replace the μ terms in $\hat{c}$ with a CAR term.)

We account for temporal dependence using both time polynomials [10], as well as random time intercepts. In the end, we have a highly parameterized model with few degrees of freedom. This design is important because there are many known confounding factors that vary spatially and over time for which we do not have data. For example, forced and economic migration exist on a spectrum and unemployment rates drive both the decision to move, and the destination [24]. The random intercepts absorb the spatial and temporal variation caused by unemployment and other unobserved confounds. This allows us to draw generalizable inferences about the value of these social media signals as indirect indicators of factors driving forced migration.

The structure of the model also allows us to incorporate time and path dependence. One of the most important indicators of movement is past movement, both because of the presence of the drivers of migration, as well as the fact that it is easier to take a well-trodden path. By incorporating location and dyad random intercepts, the model accounts for past movement. In the model, governorates such as Ninewa with levels of movement have a higher random intercept than governorates with less movement, and popular destinations are also accounted for, as well as popular origin-destination pairs.

6 EMPIRICAL EVALUATION

In this section, we evaluate different predictive models and show the strength of using indirect variables from big data/open-source data sets. We begin by explaining our data sets and our ground truth. We then discuss the traditional and the big data variables used in this analysis. Finally, we present a comparative analysis that highlights the strengths and weaknesses of our movement models.

6.1 Data Sets

In an ideal scenario, during times of crisis we would ask a large sample of people in randomly selected households in origin communities about whether they have moved or plan to move. We would also capture detailed information about how they feel about different factors or drivers of forced migration, and capture why they prefer certain locations over others. During such periods, however, people do not have time to stop and give interviews and researchers are unable to access areas where violence and persecution are occurring. This means that we do not have an optimal data set containing values for all the traditional variables associated with movement. This is one reason the problem is so challenging to address. Instead, we have traditional variables that are readily available in these situations – variables that NGOs and agencies make available or that we can construct from more general data that are available online. Given this, we will compare three different groupings of variables in our Hierarchical Bayesian Models - traditional variables only (TRADITIONAL), big data variables only (BIG DATA), and a combination of the two (BLENDED).

In order to validate the accuracy of our prediction results, we construct a ground truth data set using the Displacement Tracking Matrix (DTM) generated by the International Organization for Migration (IOM). The DTM contains bi-weekly reports on internally displaced persons (IDPs) within Iraq. These data indicate the number of families residing near a reporting location during this period, and the governorate from which they originated. The IOM data are likely to capture most IDPs in Iraq because families must check in each reporting period in order to receive benefits.[4]

6.1.1 Traditional Variables. Table 1 shows a sample of the types of traditional variables we have available for our analysis between

[4]These reports do not measure movement directly. We aggregate these reporting locations to the district level. We then estimate the number of families fleeing from each governorate, to each district, by subtracting the number families present in a district who had originated in that governorate in the previous time period from the number residing in the location in the current time period. We zero any negative numbers to focus on flight instead of returns. There are 101 districts for reporting locations and 8 governorates for origins, with a total of 808 pairs of locations.

Variable	Data Source	Movement Factor	Driver Type
Death counts - weekly	Iraq Body Count (IBC) [3]	Physical Insecurity	Threat
Historically Shia region	Empirical Studies on Conflict	Demographic	Macro
Agricultural region	PRIO grid	Environmental	Macro
Population	Empirical Studies on Conflict	Demographic	Macro
Distance	Google API	Practicality	Intervening
Number of Schools	Google API	Infrastructure	Intervening
Precipitation Amount	NOAA API	Weather	Intervening

Table 1: Traditional Variables of Movement In Our Analysis

2015 and 2017. Along with the variable, the table lists the source of the variable, the movement factor the variable is associated with, and the type of factor – macro, micro, intervening (meso), or threat/risk. For space reasons, we show only one variable per source.

We have a range of macro and intervening variables, and one threat variable, death counts. IBC provides weekly death counts by governorate, as well as a death incident database. We used this incident-level data, combined with a location ontology, to estimate daily death counts by district. For this conflict area, this variable is an important indicator of movement. In migration context, previous movement is also a significant factor predicting movements, but we do not include it in our list of variables because our time dependent Bayesian model already incorporates past movement by design. Much of the terrain of Iraq is inhospitable, meaning that direct distance may not correspond to travel distance. We used the Google API to calculate both travel time between locations, as well as miles by road. We use the number of hospitals, clinics and medical facilities; and the number of schools within each district to understand the infrastructure in different locations. We use measures of population, population density, and indicators of whether a region was historically Sunni, Shia, or mixed [7]. Finally, we use indicators of whether a region is on a border, is a dessert, urban, or farmland, and the luminosity of the region in satellite images to understand more about urbanization levels [43].

6.1.2 Big Data Variables. After social scientists on the team began developing a unifying theoretical model of movement drivers [27], we began a multi-year process of working with experts to collect keywords related to factors within the unifying model in both English and Arabic. These keywords and phrases contain both general, e.g. *militia*, and Iraqi specific vocabulary, e.g. ISIS. We also further divided these into sub-categories relating to specific variables of interest derived from a comprehensive literature review.

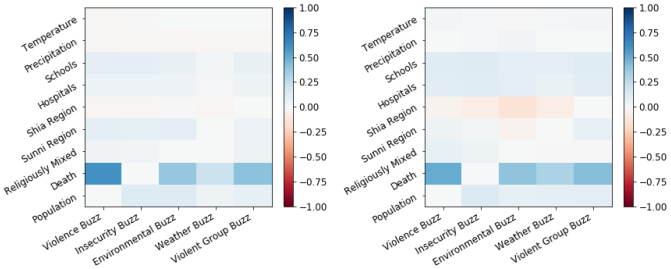
Experts simultaneously identified Twitter hashtags, regional newspaper Twitter accounts, and the most relevant keywords – all of which were used to collect data from the Twitter API. For this analysis we use over 1.3 Billion tweets in both English and Arabic. There were a number of preprocessing steps, but the most significant was identifying the location of each tweet. Iraq is divided into administrative territories called governorates. We searched the text of each tweet for location keywords.[5] Any time a tweet mentioned either a governorate, or a location within that governorate, be it a district, city, or town, that tweet was assigned the most detailed level location mentioned. When multiple locations were mentioned, the tweet was assigned all the locations mentioned. If there was no location information, the tweet was removed from the sample. After completing this preprocessing, 31 million English and 41 million Arabic tweets were used to compute buzz and sentiment variables at the district and governorate levels. Our team also has access to an unstructured archive of over 700 million articles [40]. We use a subset of more than 1.4 million English-language articles that either contained the name of a location in Iraq or were from a news source in Iraq. The buzz variables generated from these newspaper data are at the country spatial granularity.

Finally, our last type of open-source data is events. We extracted significant events from Wikipedia, and curated political events from the Integrated Crisis Early Warning System [8]. We segment them into categories of interest for different factors and construct variables by computing the frequency of categories of events in specific locations. In total, we constructed over 400 buzz, sentiment, and event variables spanning all of the factors identified by social scientists as important drivers of movement. To understand the impact of language, we consider English and Arabic separately and together.

6.2 Comparison of Different Variables

Because many of our big data variables are intended to capture gaps in the available data, we cannot fully assess the extent to which they capture variation in the direct factors influencing migration. However, we can get a sense of the efficacy of the approach by comparing those variables that are intended to capture factors related to our traditional variables. Some of these variables are more closely related than others. For example, violent events, death, and violence buzz all relate to a similar component of physical insecurity and threat. Others are only partially related. For example, derogatory slurs about certain ethnic groups are related to indicators of whether that ethnic group has historically lived in a region, but the former also captures tensions over time, rather than mere presence. Finally, others measure components of the same factor without being directly related, such as buzz about transportation infrastructure and the number of schools. Figure 4 show correlations between some of our Arabic (Figure 4a) and English (Figure 4b) big data generated variables (x-axis) and related traditional variables (y-axis). The more blue the cell, the more positive the correlation, with redder hues representing negative correlations. Those that are more conceptually related tend to be more closely correlated, e.g. death and violence buzz, but there are exceptions, weather-related buzz and precipitation. Looking at this correlation matrix can sometimes be a straightforward way to identify variables that

[5] We used a location ontology for Iraq generated by fallingrain.com. This ontology was supplemented with location names that we frequently encountered. Locations that were particularly ambiguous were manually removed from the ontology. The location ontology was used to localize deaths, events, newspaper, and social media signals.



(a) Arabic **(b) English**

Figure 4: Correlation Maxtrix - Shows example correlations between big data and traditional variables

are reasonable indirect indicators for situations when we can only use big data variables to predict movements.

Correlation alone does not tell us enough about how different big data variables can be used to predict movements and their performance compared to traditional variables. To resolve this, we examine how the variation in movement explained by important social media variables relates to the variation explained by available traditional variables. This followed the methodology in the case study presented in Section 4. We run each model with traditional variables alone (TRADITIONAL), with big data variables alone (BIG DATA), and with both traditional and big variables (BLENDED). We use the model with traditional variables as a benchmark to evaluate performance of using big data alone and use the model with both traditional and big variables to see if we can make better predications with blended variables.

6.3 Model Evaluation

Our data range is from January 2015 to December 2017. We split out data into training and validation sets. We set the training set to be data between January 2015 and February 2017 and the validation set between March 2017 and December 2017. All figures and tables in this section are obtained using the validation set. All data are aggregated into bi-weekly periods because our movement data are bi-weekly and all other data are either daily or weekly. We lag all traditional and big data variables by two weeks, using them as leading indicators to predict current movements.

We begin by identifying the underlying probability distribution that best describes our data. One benefit of Bayesian modeling is that rather than producing point predictions, it produces posterior predictive distributions. For example, given a set of covariates, the model might predict 5 families will flee with a probability of 0.25, 6 will flee with a probability of 0.4, and 7 will flee with a probability of 0.25, with the remaining 0.1 probability density in the distribution tail. These distributions can be used to generate credible ranges of possible outcomes, in this example that 5-7 families will flee. These distributions were estimated by taking 5000 draws from the model parameter posterior distributions, which produces 5000 estimated datasets. In order to improve convergence, we used Metropolis sampling for initialization and scaling, followed by a more efficient No U-Turn Sampler to sample the covariate space smoothly [22]. Sampling was conducted using the PyMC3 in Python [37]. Because the parameters in complex hierarchical models can be difficult to estimate, we used an asymmetrical sampling procedure [5].

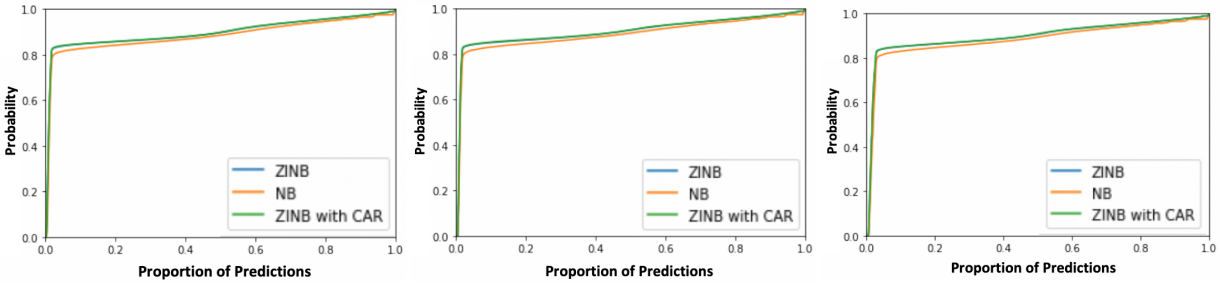
Here, we present goodness of fit results from negative binomial, zero inflated negative binomial, and zero inflated negative binomial with the addition of a CAR term. While we also tested the zero inflated Poisson model with and without the addition of a CAR term, it did not converge. This is not surprising because of the over-dispersion in movement data. Traditional fit tests rely on a single prediction. Instead of asking how close a single predicted value is to the observed outcome, we calculate the probability that each observed outcome was drawn from the posterior predictive distribution. Continuing our example, if the true observed value was 3 families, that observation would receive a lower probability score than if it was 5. The highest probability score would be if the value was 6.

Figure 5 shows this idea plotted as a CDF: the x axis represents the proportion of observations with a probability score of at least the value on the y axis. The closer the area under the curve is to 1, the higher the probability of the model overall. If a value of .1 on the x axis corresponds to a value of .8, that means that only 10% of observed outcomes had below a .8 probability of being drawn from our model distribution. In the model, the zero-inflated negative binomial model performs best across our data combinations. This is likely because most origin-destination pairs do not have movement in most weeks: new refugee families do not move to every district every week. This means that there are many zeros in the model. Models with CAR terms often performs slightly better than those without. In all further figures, we therefore use the zero-inflated negative binomial distribution with the CAR term.

Focusing only on the best model for the dyadic models, we plot all the different variable combinations on the same plot (Figure 6). Traditional variables perform better than big data variables, but the big data variables perform very well. Moreover, blending those traditional variables to which we have access with variables constructed from social media improves performance across nearly all conditions. These figures include both the Arabic and English data, but using Arabic or English alone performs similarly.

Within the migration context, it is important to predict destinations of mass movements so humanitarian organizations can position humanitarian aid in locations with greater need. Using our dyadic model we compute the precision, recall and F1 scores for estimating large movements to destination districts (see Table 2). Based on a sensitivity analysis, we consider a large movement to be a one that is at least the 75th percentile of overall movement in that district. We see that the traditional and the Arabic blended models perform the best, and the Big Data only model is comparable.

Finally, we analyze the coefficients for our social media signals. Figure 7 shows the estimated coefficients for our English and Arabic Twitter variables that serve as indirect indicators to some of the major factors in our theoretical model. Figure 8 shows the estimated coefficients for the death and events variables. The dyadic method models the decision to move from one location to a particular destination. The model was estimated with signals for both the origin and destination, but because the model is oriented around the choice of destination, we present only the coefficients for the destination coordinates.



(a) TRADITIONAL (b) BLENDED (c) BIG DATA

Figure 5: CDF of probability scores. The x axis represents the proportion of observations with a probability score of at least the value on the y axis. The closer the area under the curve is to 1, the higher the probability of the model overall.

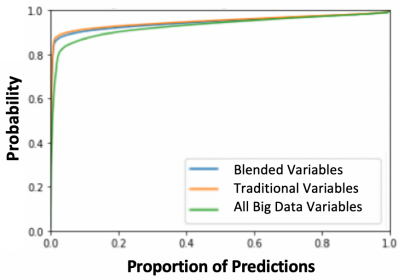


Figure 6: ZINB model comparison of different variable sets

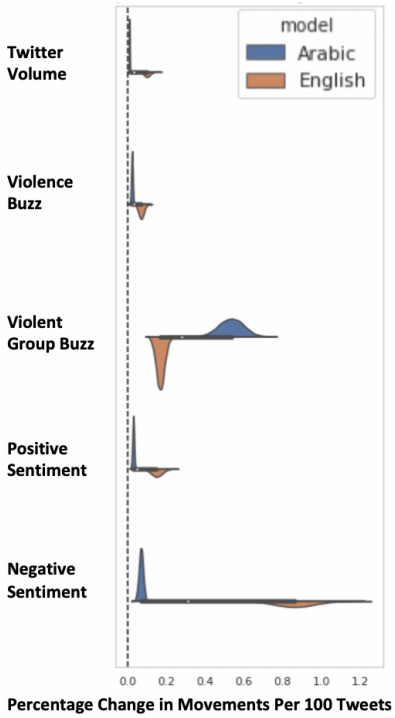


Figure 7: ZINB - Subset of Big Data Variables

One thing that stands out is that different signals have varying levels of predictive power for movement. Among traditional variables, death is the strongest signal for movement. However, death is not easy to estimate in real time. To show the significance of death in the traditional variables, we run our model using all of the traditional variables except death count. Table 2 shows the prediction scores for the traditional dyadic model without death count. We see that without death, the model is unable to capture fluctuations in movements, thereby leading to poor results (e.g. F1=0.33).

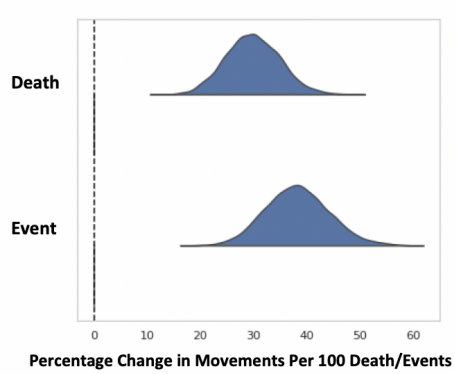


Figure 8: ZINB- Death & Event Counts

Among the big data variables, events are the strongest signal. This is consistent with movement studies conducted using traditional interview and survey variables. Among social media variables, violent group buzz is the strongest signal. It is also highly correlated to death, suggesting that this buzz signal may be capturing similar dynamics to that of death. Finally, English signals differ from Arabic signals, highlighting the importance of capturing signals in the local language. For example, tweet sentiment in Arabic is a more effective signal for movement than tweet sentiment in English. On the whole, while each individual social media signal is noisy, these figures demonstrate how they capture different salient dynamics relevant to predicting movement.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we presented a systematic approach to constructing big data variables for a traditional humanitarian problem, and blending those variables with traditional movement indicators. This study is an important step toward taking advantage of the value big data has to offer in studying age old problems. By working with social scientists to construct dynamic conversation, perception, and event variables and systematically evaluating them, we

Score	Traditional	Traditional - No Death	Arabic Big Data	Arabic Blended	English Big Data	English Blended	All Big Data
Precision	0.74	0.29	0.72	0.76	0.69	0.74	**0.77**
Recall	**0.85**	0.38	0.79	0.82	0.78	0.82	0.80
F1	**0.79**	0.33	0.75	**0.79**	0.73	0.78	0.78

Table 2: Model Prediction of Large Movements to Destination Districts

have laid the groundwork toward more nuanced study of forced migration. Moreover, we have shown the value of incorporating big data variables in predicting forced migration in the context of Iraq. Our indirect indicators and exploratory tools are available to social scientists studying forced migration via a data portal we have created [39]. As we help generate more of these signals and learn to customize and calibrate them for different regions of the world, we hope that they will become the foundation for an early warning system that will help UNHCR and other NGOs direct support and aid more efficiently.

ACKNOWLEDGEMENTS

This work was supported in part by the Massive Data Institute (MDI), the Institute for the Study of International Migration (ISIM), the MacArthur Foundation, the National Science Foundation (NSF), and the Canadian Social Science and Humanities Research Council (SSHRC). We would also like to thank the rest of the Georgetown computer science and migration team for their insight.

REFERENCES

[1] P. Adhikari. 2013. Conflict-Induced Displacement, Understanding the Causes of Flight. *American Journal of Political Science* 57, 1 (2013), 82–89.
[2] A. Agrawal and A. An. 2016. Selective Co-occurrences for Word-Emotion Association. In *International Conference on Computational Linguistics (COLING)*.
[3] L. Balcells and A. Steele. 2016. Warfare, political identities, and displacement in Spain and Colombia. *Political Geography* 51 (2016), 15–29.
[4] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, and J. Von Schreeb. 2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. 8 (2011).
[5] M. Betancourt. 2017. Diagnosing biased inference with divergences. http://mc-stan.org/documentation/case-studies/divergences_and_bias.html. (2017).
[6] J. Blumenstock. 2012. Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda. *Information Technology for Development* 18 (2012), 107–125.
[7] J. Borokowski and Z. Bulutgil. 2012. Ethnicity Study : Ethnic Composition at District Level. (2012). https://esoc.princeton.edu/files/ethnicity-study-ethnic-composition-district-level The Empirical Studies of Conflict (ESOC).
[8] E. Boschee, J. Lautenschlager, S. O'Brien, S. Shellman, and et al. 2015. ICEWS Coded Event Data. (2015).
[9] M. Bylander. 2015. Depending on the sky: Environmental distress, migration, and coping in rural Cambodia. *International Migration* 53, 5 (2015), 135–147.
[10] D. Carter and C. Signorino. 2010. Back to the future: Modeling time dependence in binary data. *Political Analysis* 18, 3 (2010), 271–292.
[11] R. Churchill, L. Singh, and C. Kirov. 2018. A Temporal Topic Model for Noisy Mediums. In *Pasific Asian Conference on Knowledge Discovery and Data Mining*.
[12] Norwegian Refugee Council and Chemin de Balexert. 2018. Internal Displacement Monitoring Centre. *Guidelines on Profiling Internally Displaced Persons* (2018).
[13] C. Davenport, W. Moore, and S. Poe. 2003. Sometimes you just have to leave: Domestic threats and forced migration, 1964-1989. *International Interactions* 29, 1 (2003), 27–55.
[14] C. Davenport and R. Talibova. 2019. Counting Battle Deaths: Shaping Outcomes in Civil Wars. (2019). Unpublished Manuscript.
[15] P. Deville, C. Linard, S. Martin, M. Gilbert, Forrest Stevens, Andrea Gaughan, Vincent Blondel, and Andrew Tatem. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences(PNAS)* 111 (2014), 15888–15893.
[16] K. Donato and E. Ferris. 2017. A Dynamic View of Forced Migration: Revisiting the Push-Pull Framework. *TransAtlantic Council on Migration* (2017).
[17] P. Earle, D. Bowden, and M. Guy. 2012. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics* 54 (2012). Issue 6.
[18] United Nations High Commissioner for Refugees. 2016. UNHCR Global Trends: Forced Displacement in 2015. (2016). Technical Report.
[19] I. Foster, R. Ghani, R. Jarmin, F. Kreuter, and J. Lane. 2016. *Big data and social science: A practical guide to methods and tools.* CRC Press.
[20] E. Frias-Martinez, G. Williamson, and V. Frias-Martinez. An agent-based model of epidemic spread using human mobility and social network information. In *IEEE International Conference on Social Computing* (2011).
[21] J. Hockett, Y. Liu, Y. Wei, L. Singh, and N. Schneider. Detecting and using buzz from newspapers to understand patterns of movement. In *IEEE International Conference on Big Data* (2018).
[22] M. Hoffman and A. Gelman. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* 15, 1 (2014), 1593–1623.
[23] L. Hunter. 2005. Migration and environmental hazards. *Population and environment* 26, 4 (2005), 273–302.
[24] D. Karemera, V. Oguledo, and B. Davis. 2000. A gravity model analysis of international migration to North America. *Applied Economics* 32, 13 (2000), 1745–1755.
[25] N. Lozano-Gracia, G. Piras, A. M. Ibáñez, and G. Hewings. 2010. The journey to safety: conflict-driven migration flows in Colombia. *International Regional Science Review* 33, 2 (2010), 157–180.
[26] S. Martin and L. Singh. 2018. Data Analytics and Displacement: Using Big Data to Forecast Mass Movements of People. (2018).
[27] S. Martin, L. Singh, A. Taylor, and Wahedi L. 2019. What drives forced migration?: A dynamic factor analysis. (2019). Working Paper.
[28] D. Massey. 2015. A missing element in migration theories. *international Journal of Migration Studies* (2015), 279.
[29] E. Melander and M. Öberg. 2006. Time to go? Duration dependence in forced migration. *International Interactions* 32, 2 (2006), 129–152.
[30] E. Melander and M. Öberg. 2007. The threat of violence and forced migration: Geographical scope trumps intensity of fighting. *Civil Wars* 9, 2 (2007), 156–173.
[31] S. Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 174–179.
[32] W. Moore and S. Shellman. 2004. Fear of persecution: Forced migration, 1952-1995. *Journal of Conflict Resolution* 48, 5 (2004), 723–745.
[33] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, and et al. 2014. 'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.
[34] R. Reuveny. 2007. Climate change-induced migration and violent conflict. *Political Geography* 26, 6 (2007), 656–673.
[35] J. Rubin and W. Moore. 2007. Risk factors for forced migrant flight. *Conflict Management and Peace Science* 24, 2 (2007), 85–104.
[36] I. Salehyan and K. Gleditsch. 2006. Refugees and the spread of civil war. *International Organization* 60, 2 (2006), 335–366.
[37] J. Salvatier, T. Wiecki, and C. Fonnesbeck. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2 (2016), e55.
[38] J. Schon. 2015. Focus on the forest, not the trees: A changepoint model of forced displacement. *Journal of Refugee Studies* 28, 4 (2015), 437–467.
[39] L. Singh and DataLab. Forced Migration Interactive Visualizations of Big Data Variables. http://forcedmigration.cs.georgetown.edu. (????).
[40] L. Singh and R. Pemmaraju. 2017. EOS: A multilingual text archive of international newspaper & blog articles. In *IEEE International Conference on Big Data*.
[41] B. State, M. Rodriguez, D. Helbing, and E. Zagheni. 2014. Migration of professionals to the US: evidence from linkedin data. In *International Conference on Social Informatics*. Springer.
[42] J. Tarver. 1961. Predicting migration. *Social Forces* 39, 3 (1961), 207–213.
[43] A. Tollefsen, H. Strand, and H. Buhaug. 2012. PRIO-GRID: A unified spatial data structure. *Journal of Peace Research* 49, 2 (2012), 363–374.
[44] X. Wang and A. McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*.
[45] Y. Wei, L. Singh, D. Buttler, and B. Gallagher. 2018. Using semantic graphs to detect overlapping target events and story lines from newspaper articles. *International Journal of Data Science and Analytics* 5, 1 (2018), 41–60.