Nowhere to Go

Why do Some Civil Wars Generate More Refugees than Others?

Oguzhan Turkoglu[*] Thomas Chadefaux[*]

July 13, 2017

Abstract

Civil wars greatly vary in the number of refugees they generate, ranging from zero to over six millions in a given conflict. Work on this variation has largely focused on 'push' factors—deleterious attributes of the home country that lead to refugee flows, such as violence and repression. Yet, few have studied the importance of 'pull' factors—attractive features of the potential host countries. Here we show in particular the importance of the expected quality of life in possible destinations. Using data on civil wars from 1951 to 2008, we find that the proximity of democratic and wealthy potential hosts accounts for much of the variation in the number of refugees. In fact, we show using out-of-sample validation that these 'pull' factors account for almost as much of the predictive power as a combination of all the main variables previously identified in the literature.

Keywords: civil war, refugees, geography, spatial regression, network, connectivity

Word Count ≈ 4,943

[*] Department of Political Science, Trinity College Dublin, 2 College Green, Dublin 2, Ireland. Email: turkoglo@tcd.ie and thomas.chadefaux@tcd.ie

Civil wars are the main cause of refugee flows. By 2016, for example, the Syrian conflict alone had generated more than five million refugees. These flows affect not only the refugees themselves, but also impose a strain on the economic, social and political life of their host countries. The Syrian refugee crisis, for example, is currently high on the European Union political agenda and has caused significant tensions between member states.

Yet, civil conflicts greatly vary in the number of refugees they generate, ranging from zero to more than six millions for Afghanistan in 1990. Unfortunately, little is known about what accounts for this variation, despite the importance for international actors and hosting countries of anticipating population movements. Previous work on refugee flows has mostly focused on the characteristics of the country at war, with a particular attention to 'push' factors—deleterious attributes of the home country that lead to refugee flows such as violence and repression.

While important, we find however that push factors explain only some of the variation in the number of refugees. Here, we argue instead that the options available to refugees are key in their choice to leave or stay. Refugees need to find an appealing host, and to be able to get there. In other words, geography and neighbors matter. In fact, we show that the most important factor in the decision to leave is the availability of suitable host countries in nearby proximity—'Pull' factors. We show in particular that the economic and political attractiveness of surrounding countries is key to refugees' decision to leave. Using data from 1951 to 2008, we find that measures of GDP per capita and regime type in neighboring countries explain much of the variation in refugee numbers. Both in-sample regressions and out-of-sample predictions corroborate the key role of these 'pull' factors in explaining the variation in refugee numbers.

We first review existing work on refugee flows and propose hypotheses related to the role of geography and suitable hosts. We then present our empirical strategy and data, after which we report on our results using both in- and out-of-sample validation.

Push and Pull Explanations of Refugee Flows

Studies on refugee flows have typically focused on single case studies analyzing the impact of economic conditions (Osborne 1980, Stanley 1987), ethnic relations (Newland 1993), genocide (Midlarsky 2005, Uzonyi 2014) or conflict (Ibez & Velsquez 2009). The level of analysis ranges from the subnational (Czaika & Kis-Katos 2009) to the national (Adhikari 2012, Adhikari 2013) and regional levels (Zolberg, Shurke & Aguayo 1989, Iqbal 2007, Neumayer 2005). The few quantitative analysis available often suffer from methodological shortcomings. Apodaca (1998), for example, analyzes the main causes of forced migration at the monadic level but only considering countries that do generate forced migration—ignoring those that do not, i.e., omitting the zeros. This leads to a biased sample with questionable inferences. Similar biases apply to studies at the dyadic level (e.g., Moore & Shellman 2007), which only include countries that generate a refugee flow within a given year.[1] Adopting this approach to our data would lead us to discard 30% of our observations. Others similarly limit their analysis to cases involving a high number of refugees (e.g., Wood 1994).

More generally, the literature on the causes of refugee migration has mostly focused on 'push' factors—deleterious attributes of the home country that lead to refugee flows. Most emphasize the effect that violence and repression have on people's decision to leave their country (Weiner 1978). Interstate wars (Moore & Shellman 2004, Melander & Oberg 2006, Schmeidl 1997), dissident violence (Davenport, Moore & Poe 2003), but mostly civil wars (Weiner 1996) and genocide (Davenport, Moore & Poe 2003, Moore & Shellman 2004, Schmeidl 1997, Melander & Oberg 2006) are the main culprits. The role of regime type is

[1] For example, Romania generated refugees in 1970, so all possible Romania dyads are included in the dataset. Yet of these, only Turkey and Greece hosted about 50 refugees each, whereas other countries hosted none. Therefore, other than Turkey and Greece, all countries are coded as 0. Contrast this with 1969, when Romania did not generate any refugees and hence none of the dyads including Romania appear in the data. This is problematic for a number of reasons. First, it is inflating the number of zeros by including irrelevant dyads such as Romania-Burundi. Second, even the very small number of refugees generated by Romania in 1970 lead to the addition of $N-1$ observations to the data (for $N-1$ Romania-dyads). But 1969, with 0 refugees, creates no observation. Yet the absence of refugees is itself valuable information, as it may reflect the absence of valuable opportunities to leave. To understand the impact of this operationalization, we replicated Moore and Shellman's study, but this time coding countries that do not generate any refugee flow as 0 (as opposed to missing). As a result of this change, we find that the coefficient associated with the host's regime type changes sign, and many others are strongly affected.

also emphasized: democratic states generate fewer refugees than autocratic ones (Melander & Oberg 2006, Melander & Oberg 2007, Moore & Shellman 2004), though support for that hypothesis is mixed (Davenport, Moore & Poe 2003).[2] The role of socio-economic variables such as economic underdevelopment or population pressures has also been examined, though with mixed conclusions (e.g., Melander & Oberg (2006), Melander & Oberg (2007), Moore & Shellman (2004)).

However, 'pull' factors—attractive features of the potential host countries—have largely been ignored, especially in terms of their effect on the decision to leave. Existing studies on pull factors instead generally focus on why some countries host more refugees than others (Neumayer 2005, Moore & Shellman 2007). Although these studies offer important insights into refugee hosting at the dyadic level, they fail to provide explanations about the role of pull factors in the generation of refugees. While they explain why some countries host refugees from a given country rather than another, they fail to grasp why some people do not seek refuge and hence, why some countries generate a higher number of refugees than others. In other words, they tend to focus on where refugees go, as opposed to how many refugees are generated in the first place.

Explaining the Decision to Leave

Leaving one's homeland and settling in a foreign country is typically dangerous and costly. Other than the economic, social and cultural aspect of adjustment to the host country, the physical journey itself involves important risks in terms of safety and economic well-being. Refugees can therefore first be expected to favor destinations that are geographically close to their home country, as this facilitates migration.[3]

Second, we expect refugees to prefer democratic destinations over autocratic ones. Au-

[2] Some scholars also analyze the effect of regime collapse, change in polity score and regime transition. While regime collapse and change in polity are positively correlated with the number of refugees, the effect of regime transition is unclear (Davenport, Moore & Poe 2003, Melander & Oberg 2006, Melander & Oberg 2007, Moore & Shellman 2004). Furthermore, some studies examine human rights violations instead of regime type, but with mixed results (Schmeidl 1997).

[3] For example, 92 % of nearly 5.1 million Syrian refugees in 2017 went to Turkey, Jordan and Lebanon.

thoritarian regimes tend to be repressive, whereas democracies tend to respect fundamental human rights and to follow the rule of law. As a result, we expect refugees to target democracies because they are less likely to be persecuted on the basis of their race, religion, nationality, membership of particular social group or political opinion. We therefore expect democracies to act as a pull factor, though this effect will be mitigated by distance—while the United States may an appealing destination, its distance makes it unreachable for many. We therefore include the effect of other countries' regime as a measure of their polity score weighted by their distance from the home country. More specifically, we add a variable $\mathbf{w}_i$Polity, where $\mathbf{w}_i$ denotes a vector of spatial weights for country i, which we discuss in more detail below.

Hypothesis 1 *Increases in the polity score of states surrounding source country i increase the number of refugees in a civil war.*

Finally, we expect refugees to be find wealthier countries more attractive, as they tend to offer better economic opportunities and living conditions. Surrounding countries with a high GDP per capita—also weighted by their distance as $\mathbf{w}_i$GDPPC—are therefore expected to increase the number of refugees.

Hypothesis 2 *Increases in the GDP per capita of states surrounding source country i increase the number of refugees in a civil war.*

Of course, other factors also affect the number of refugees. First, refugees may not always have the ability to choose their destination. Wars may be so severe that people may want to escape at any cost, regardless of the regime or development of their destination. However, most civil wars are not this severe. More than 90% of civil war years saw less than 10,000 battle-related deaths, and more than 99% less than 37,500. In these situations, people may choose to stay in their home country if the experience in the surrounding countries is one of similar violence and repression, and therefore offers little improvement over their current situation at home.

Second, the regulatory environment, in particular, can be a strong brake on refugee patterns. Because of its policy not to admit refugees, Saudi Arabia, for example, was a destination for only a small number of the Syrian refugees, despite the Kingdom's high standards of living. On the contrary, Germany was a magnet because of its welcoming political and legal environment. In other words, pull factors can also be affected by external constraints which can be hard to measure and quantify. In that sense the present study has limits and could be improved with better data. However, our results—both in and out of sample—show strong evidence that pull factors greatly matter and are important to incorporate.

Model and Data

Our model is defined as:

$$\text{Refugees}_{i,t} = \mathbf{x}'_{i,t}\boldsymbol{\beta} + \boldsymbol{\rho}\mathbf{w}'_{i,t}\mathbf{h}_{it} + u_i + \varepsilon_{it}, \tag{1}$$

where $\text{Refugees}_{i,t}$ denotes the total number of refugees originating from a given country-conflict i and year t ($t \in [1951, 2008]$).[4] $\mathbf{x}_{i,t}$ is a vector of K control variables for country-conflict i and year t; $\mathbf{w}_{i,t}$ is a vector of spatial weights for each of the N countries of the world—i.e., for each country i, we calculate the output of a distance function to each of the other countries in the world (more on this below). $\mathbf{h}_{it}$ is an $N \times M$ matrix of M attributes of possible host countries (e.g., GDP per capita). u_i are country-level fixed effects and ε_{it} are residuals at the country-conflict-year level. $\boldsymbol{\beta}$ is vector of K coefficients to be estimated, and $\boldsymbol{\rho}$ is a vector of M spatial coefficients to be estimated (one for each of the M variables

[4] Some studies instead use as dependent variable the *flow* of refugees or forced migration by calculating the change in the stock from one year to the next—usually truncating negative values at zero (Schmeidl 1997, Moore & Shellman 2004, Melander & Oberg 2006, Melander & Oberg 2007, Melander, Oberg & Hall 2009), i.e. Refugee/Forced migration flow = $\max(\text{refugees}_t - \text{refugees}_{t-1}, 0)$. Others use the net stock of forced migration, subtracting hosted refugees from 'exported' ones (Davenport, Moore, and Poe 2003): Net stock of forced migration = (Total number of refugees and IDPs generated by country i) - (Total number refugees hosted by country i). However, using the flow of refugees as the dependent variable is problematic, as it omits refugees who have chosen not to return to their country of origin yet. In addition, the main data sources (the UNHCR and the United States Committee for Refugees and Immigrants) do not keep a record of flows but only of the total number of refugees, so that refugee flow calculations are controversial. We explain in more detail the problems associated with using the flow of refugees rather than the stock in appendix A. Here, we follow Moore & Shellman (2004) and use the UNHCR database.

associated with neighboring countries).[5]

Because our dependent variable is an observed count of refugees, it only takes nonnegative integer values. As such, ordinary least squares regression is inappropriate and generalized linear model of the Poisson family should be preferred. Here we used the negative binomial regression, which relaxes the Poisson model's restrictive assumption that the variance be equal to the mean. This is appropriate here because our dependent variable is over-dispersed (i.e., its variance is greater than its mean—see table 1). However, our results are robust to alternative estimation methods, including OLS and zero-inflated negative binomial regression (see section on robustness checks below).

To measure the number of refugees, we follow Moore & Shellman (2004) and Uzonyi (2015) and use the definition of the Office of the United Nations High Commissioner for Refugees' (UNHCR) definition.[6] For conflict, we use the definition of UCDP/PRIO's Armed Conflict dataset as 'a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths' (Gleditsch 2002).

Our main independent variables ($\mathbf{h}_{i,t}$) are based on the attractiveness of neighboring countries: their GDP per capita is used as a measure of wealth and economic prospects (data from Gleditsch (2002)); and their polity score, as a measure of the attractiveness of their political regime (data from Polity IV's Polity2 variable (Marshall, Gurr & Harff 2016)). Each of these variables is weighted by a function of the distance between the host and source country. More specifically, we first created a connectivity matrix $\mathbf{W}$ which records for each

[5] We did not include a lagged dependent variable in our main specifications (table 2) for two main reasons: first, the inclusion of a lagged dependent variable implies the loss of the first year of data from every conflict (about 16% of our observations). More problematically, it removes the most interesting—and difficult—observations to explain: the number of refugees in the first year of conflict, and hence loses some of our ability to distinguish between models. Just like forecasting the onset of conflict is much more difficult than its incidence, correctly predicting the first year of refugees without any past reference is much more challenging and discriminating than forecasting subsequent years. Regardless, we show below that our results hold even with the inclusion of the lagged dependent variable.

[6] UNHCR defines refugees as people who are 'unable or unwilling to return to their country of origin due to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion.' (Introductory note to the text of the Convention and Protocol Relating to the Status of Refugees, Office of the UNHCR, 2010).

pair of country $\{i, j\}$ and time t the inverse of the logged minimum distance between the boundaries of the source i and the host country j.[7] This represents the idea that remote countries are less attractive, but that the marginal cost associated with long distances is diminishing.[8] We weigh each country in the same way and compile a weighted sum of their GDP per capita and Polity to obtain for each country a weighted measure of their neighbors' polity and GDP per capita, $\mathbf{w_i}$Polity and $\mathbf{w_i}$GDPPC.

Control variables ($\mathbf{x}_{it}$) include: the number of neighboring countries within 500 km of the border;[9] the year in war (starting at one); the number of battle-related deaths (from PRIO's Battle Deaths Dataset (Lacina & Gleditsch 2005));[10] the source country's democracy level (Polity2), GDP per capita (Gleditsch 2002), and population (Gleditsch & Ward 1999); dummy variables for: an ongoing interstate war (UCDP/PRIO Armed Conflict Dataset, Pettersson & Wallensteen (2015)); the occurrence of a genocide in that country-year (Political Instability Task Force, Goldstone, Bates, Epstein, Gurr, Lustik, Marshall, Ulfeder & Woodward (2010), Marshall, Gurr & Harff (2016)); internationalization—i.e., whether a secondary party has intervened in the conflict from (UCDP/PRIO); whether the conflict is driven by territorial or government incompatibility (UCDP/PRIO); whether the country is contemporaneously involved in an interstate conflict with at least 25 battle deaths ('Interstate War', from UCDP/PRIO's Armed Conflict Dataset). Finally, we added country-level fixed effects (regional effects make little difference). Summary of descriptive statistics are reported in table 1.

[7] More specifically: $w_{i,j,t} = \frac{1}{\ln(\text{distance}_{ij,t})}$. $w_{i,j,t}$ is coded as 0 for countries that share a border. Data on distances (in km) was obtained from Weidmann, Kuse & Gleditsch (2010).

[8] We found that other specifications of $\mathbf{w}_i$ had far less predictive power, in line with our theoretical expectation that the marginal effect of distance is decreasing. Using raw distance or only countries within a certain radius, for example, resulted in worse out-of-sample forecasts than our choice of the log.

[9] We use the distance from border to border, as it is probably the most relevant for refugees, rather than the distance between capitals. The 500km threshold simply follows Weidmann, Kuse & Gleditsch (2010), but for robustness purposes, we also varied the threshold from 0 to 900km and found that the results do not change qualitatively.

[10] Using either the 'low', 'high' or 'best' estimate for battle deaths makes no qualitative difference to our results.

Table 1: Summary statistics.

	mean	sd	min	max
Refugees	144251.6	472005.9	0	6339095
$\mathbf{w}_i$Polity	0.36	48.6	-101.6	122.8
$\mathbf{w}_i$GDPPC	13.7	8.91	0.85	55.2
Polity	-0.19	6.41	-10	10
GDPPC (log)	7.10	1.22	4.09	10.7
Battledeaths (log)	7.44	1.63	3.22	12.4
Population (log)	16.8	1.40	12.9	20.9
Year in Conflict	8.84	9.21	1	48
N neighbors	8.23	4.17	0	37
Territory incompatibility	0.48	0.50	0	1
Internationalization	0.13	0.34	0	1
Interstate	0.063	0.24	0	1
Genocide	0.16	0.37	0	1

Results

We report our results in two ways. First, we estimate the model above on our sample data and report our inferences below. Second, we performed a number of out-of-sample cross-validation procedures and show the significant contribution of our independent variable to the performance of out-of-sample forecasts.

In-sample

Table 2 reports on the in-sample results of various specifications of the number of refugees originating from a given country-conflict-year between 1951 and 2008. Standard errors are always clustered by country to account for non-independent panel observations, and country-fixed effects are also included to control for average differences across countries in possible unobserved predictors.

In line with our hypotheses, we find strong support for the role of pull factors. The location of the source country and characteristics of its neighbors play an important role in the number of refugees generated by a particular conflict. In particular, an increase in

Table 2: Negative binomial regression of the yearly number of refugees in Civil Wars, 1951–2008. Similar results are obtained using zero-inflated negative binomial regression, OLS, regional dummies.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
								Full model
				$\mathbf{w}_i$GDPPC	$\mathbf{w}_i$Polity	Full model	Full model	($\mathbf{w}_i$GDPPC +
	Base	Literature	geo only	only	only	($\mathbf{w}_i$GDPPC)	($\mathbf{w}_i$Polity)	$\mathbf{w}_i$Polity)
Battledeaths (log)	0.079**	0.163*	0.143	0.236*	0.168**	0.213**	0.105	0.165**
	(0.021)	(0.081)	(0.083)	(0.100)	(0.008)	(0.067)	(0.056)	(0.060)
Year in Conflict	0.067**	0.067	0.076	0.049**	0.046	0.052	0.069	0.058
	(0.016)	(0.048)	(0.047)	(0.001)	(0.047)	(0.039)	(0.038)	(0.032)
Year in Conflict ^{2}	-0.002*	-0.002	-0.002	-0.001**	-0.001	-0.001	-0.002	-0.002
	(0.001)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
Territory incompatibility		-0.515	-0.710			-0.467	-0.296	-0.345
		(0.357)	(0.399)			(0.311)	(0.263)	(0.253)
Internationalization		0.648*	0.637*			0.411	0.542*	0.459
		(0.287)	(0.278)			(0.263)	(0.240)	(0.251)
Polity		-0.072**	-0.057*			-0.044*	-0.069**	-0.057*
		(0.021)	(0.027)			(0.022)	(0.026)	(0.025)
GDPPC (log)		0.880**	0.787**			0.347	0.468**	0.367*
		(0.154)	(0.218)			(0.179)	(0.167)	(0.176)
Interstate		-0.291	-0.256			0.165	0.325	0.295
		(0.289)	(0.277)			(0.250)	(0.282)	(0.314)
Genocide		0.140	0.303			0.470	0.675	0.620*
		(0.456)	(0.453)			(0.339)	(0.348)	(0.309)
Population (log)		0.233	0.205			0.001	-0.013	-0.036
		(0.144)	(0.175)			(0.121)	(0.103)	(0.103)
N neighbors			0.082			-0.027	0.061	0.015
			(0.048)			(0.050)	(0.045)	(0.047)
$\mathbf{w}_i$GDPPC				0.106**		0.102**		0.054*
				(0.008)		(0.018)		(0.022)
$\mathbf{w}_i$Polity					0.022**		0.019**	0.011*
					(0.000)		(0.003)	(0.005)
Constant	-2.764**	-13.143**	-12.514**	-5.000**	-3.082**	-6.913**	-6.212**	-5.828**
	(0.038)	(2.523)	(3.347)	(0.659)	(0.101)	(2.410)	(2.168)	(1.841)
Observations	1094	1094	1094	1094	1094	1094	1094	1094
BIC	18483.642	17985.291	17936.804	17641.247	17860.476	17561.771	17563.607	17525.275

Standard errors clustered by country in parenthesis. Each model includes country fixed effects (not reported)

* $p < 0.05$, ** $p < 0.01$

either GDP per capita or polity weighted by distance leads to an increase in the expected number of refugees. In other words, civil wars in countries surrounded by developed and democratic countries generate more refugees than in those surrounded by underdeveloped and autocratic states.

The importance of the 'pull' variables $\mathbf{w_i}$GDPPC and $\mathbf{w_i}$Polity is further supported by the large improvement in model fit, as evidenced by the reduction in BIC (i.e., improvement) in models that include either of the pull variables (or both: models 6–8 in table 2). In fact, this reduction in the BIC score is nearly as large as the one obtained by adding all of the variables identified in the literature to the most basic model, which only includes information about the year in conflict and the number of battle-related deaths. Loosely, then, our variables contribute as much to the fit as all the existing literature's variables combined.

Among control variables, we find that the internationalization of a civil war, battle deaths, polity score and GDP per capita all significantly affect the number of refugees, as expected. Territory incompatibility (ethnic civil wars), genocide, population and interstate war, however, have no explanatory power over the number of refugees.

Out-of-sample

Beyond statistical inference, out-of-sample performance is another critical measure of a model's value (Ward, Greenhill & Bakke 2010, Chadefaux 2017*a*). It reinforces the causal claim and helps overcome the overfitting problem (Beck, King & Zeng 2000). We therefore estimated our model on a subset of the data (the 'learning' set), and tested its performance on out-of-sample data (the 'testing' set). In particular, we cross-validated our results using the 'leave-one-out' method, by which coefficients are estimated on all conflicts with the exception of one, and used to estimate the number of refugees for the one conflict left out. This process is repeated for all $M = 148$ conflicts in our sample, yielding 1,391 forecasts (each conflict may have more than one year to predict). We then calculate the forecasting error as the absolute value of the difference between each of these forecasts and the observed value, i.e., $e_i = |\hat{y}_i - y_i|$, and for each model take the median of these errors to obtain the Median

Absolute Error (MAE).[11] A large MAE indicates that the model tends to produce forecasts that are far from the actual number of refugees observed.

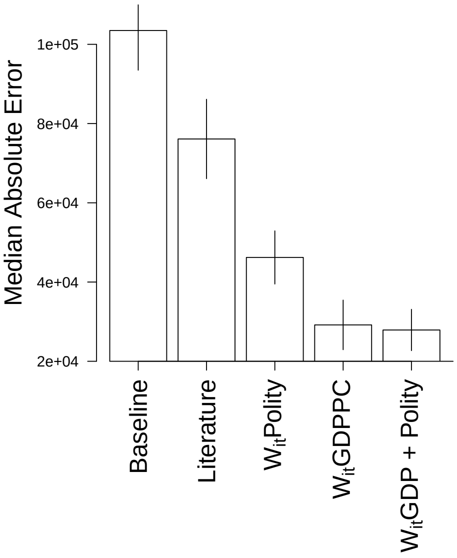


Figure 1: Median Absolute Errors for out-of-sample forecasts of the number of refugees in a given country-year (corresponding to models 1, 2, 6, 7 and 8 in table 2). Standard errors were obtained by bootstrapping.

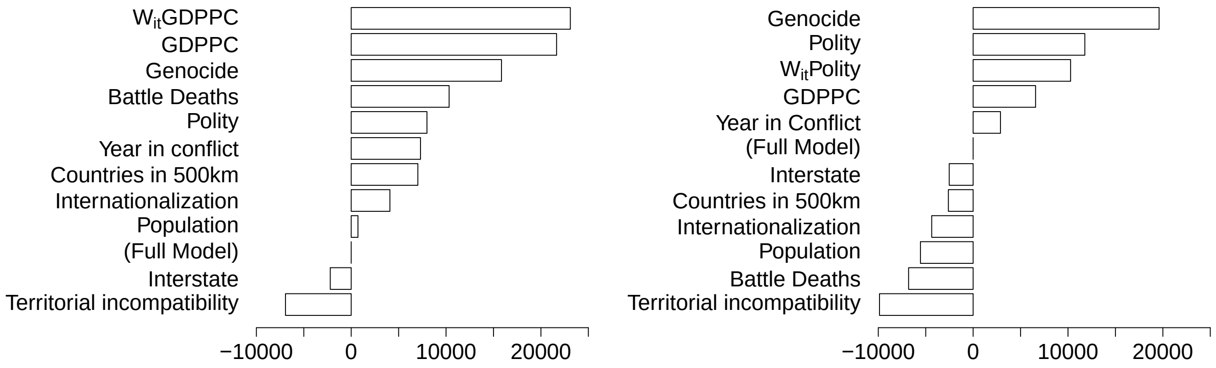
The results displayed in figure 1 strongly corroborate the in-sample findings. In line with the lower BICs of models 6–8 (table 2), we find that these models' median absolute error is substantially and significantly lower than the error of the model derived from the existing literature.[12] In fact, we find that the improvement gained by adding $\mathbf{w}_i$GDPPC alone is much larger than the one provided by all of the literature's variables combined over the baseline model.

To further demonstrate the importance of pull factors, we conducted the same 'leave-one-out' analysis as above, but this time estimating the performance of a model from which one single variable was removed. This gives us a sense of the contribution of each variable to the out-of-sample forecasting performance, and hence of the importance of that variable in

[11] We obtain similar results using the squared difference, but with results that are less easily interpretable.

[12] Paired Mann-Whitney test for the full model compared to the 'literature' model: $U = 797,900$, $p < 0.01$. Similar results apply for the $\mathbf{w}_{it}$Polity model only or the $\mathbf{w}_{it}$GDPPC model only (see also Chadefaux 2014).

the model, both for explanatory and forecasting purposes.[13] The results in figure 2 confirm



(a) Model 6 (table 2)

(b) Model 7 (table 2)

Figure 2: Out of sample predictive power. The out-of-sample Median Absolute Error (MAE) of models (6) and (7) was recomputed after the removal of one of their variables at a time. Large positive values associated with a given variable imply that the model yields larger errors without that variable—i.e., that variable was essential to the model. Negative values imply that removing that variable actually improved the model's predictive power. Removing $\mathbf{w}_{it}$GDPPC from model 6, for example, increases the MAE (i.e., the median of $|\hat{y}_i - y_i|$) by more than 20,000.

the importance of pull factors identified in the in-sample analysis. In particular, $\mathbf{w_i}$GDPPC turns out to be the most important contributor to reducing forecasting errors. $\mathbf{w_i}$Polity also plays an important role, even though Genocide and Polity outperform it. Adding pull variables to existing models in the literature thus reduces the typical forecasting error by more than 20,000 refugees—more than any other variable identified by the literature.

Robustness Checks

To ensure the robustness of our results, we first tested the effect of alternate variable operationalizations and model specifications. First, including a lagged dependent variable did not substantially affect any of our main inferences regarding $\mathbf{w}_i$GDPPC and $\mathbf{w}_i$Polity. Regional dummies and various estimates of battle deaths (Low/High/Best) also had little substantial effect on our results. Clustering standard errors by civil war instead of by country, as well

[13] See also Ward, Greenhill & Bakke (2010) for another application of this strategy.

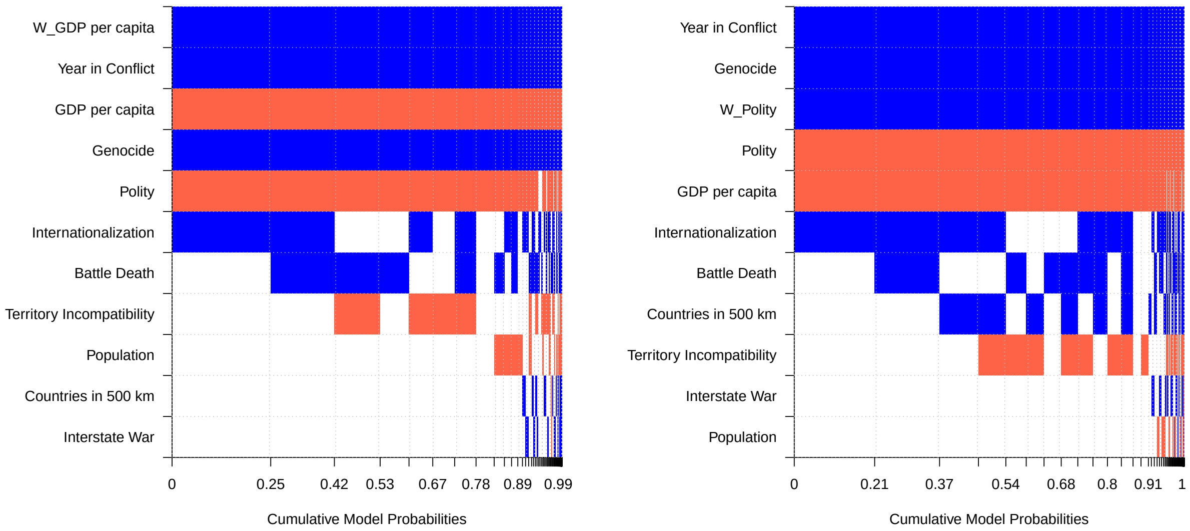
as with panel-correlated standard errors (PCSE) also had little effect.[14] Finally, we also obtained similar results using a zero-inflated negative binomial regression or an ordinary least square.

More generally, one common difficulty of regression models is model uncertainty. In particular, which variables should be included in the model? There may well be a different subset of variables that better fits the data. While table 2 reports on a number of specifications, it is possible that we are missing a better model. Bayesian Model Averaging (BMA) addresses this problem by estimating many combinations of the independent variables (Hoeting, Madigan, Raftery & Volinsky 1999, Chadefaux 2017*b*). With K variables, this implies the estimation of up to 2^K models. For each model M_j, $j = 1, \ldots, 2^K$, a prior $P(M_j)$ is specified and the data X can be used to derive a posterior $P(M_j|X)$ using Bayes' theorem (see Hoeting et al. 1999). The posterior mass associated with each model then gives us a sense of which models are best, and the frequency with which a given variable is part of these successful models tells us about its usefulness and contribution to a wide set of models. In other words, we should have more confidence about the effect of a variable if that variable tends to be included in a large number of 'good' models (i.e., those with a high Posterior Model Probability—PMP).

The results of the Bayesian model averaging are summarized in figure 3. The plot displays the Posterior Model Probability for all $2^{11} = 2048$ models estimated.[15] In the left plot, for example, the best model, with 25% posterior model probability (PMP), includes six variables (i.e., $\mathbf{w}_i$GDPPC, year in conflict, GPPC, Genocide, polity and Internationalization). The second best model adds Battle deaths for a PMP of 17%; the third best removes internationalization but adds territory incompatibility; and so on. Regardless of the details of each model, the most important variables are those that appear in the largest number of model specifications. We note that both $\mathbf{w}_i$GDPPC and $\mathbf{w}_i$Polity appear in all models, and with the same sign, which suggests that our results are not contingent on a specific subset

[14] Loosely, PCSE assumes that the observations are independent across time but not across space; in contrast, clustered standard errors assume that they are independent in space but not in time.

[15] Uniform priors were used, though similar results obtain using fixed ($K/2 = 5.5$) or random priors. We also find similar results using model (8) in table 2.



(a) $\mathbf{w}_i$ GDP per capita (model 6 in table 2)

(b) $\mathbf{w}_i$ Polity (model 7 in table 2)

Figure 3: Bayesian model averaging. Blue (red, white) cells represent positive (negative, zero) coefficients.

of variables.

Conclusion

This study has set out to analyze why some civil wars generate more refugees than others. We found that the quality of life in surrounding countries is critical to the decision of potential refugees to leave. In particular, a higher number of democratic and developed countries in the region increases the number of people who flee their country.

This study is also, to the best of our knowledge, the first to apply out-of-sample forecasting to the analysis of refugees. This matters because variables that are statistically significant need not in fact have much predictive power (Ward, Greenhill & Bakke 2010). Out-of-sample forecasting allowed us to assess the relative importance of each variable beyond its p-value, and to avoid overfitting in-sample data. Thus, out-of-sample forecasts showed that GDP per capita and Polity weighted by distance increase our capacity to predict the number of refugees generated by a particular conflict almost as much as the variables identified by the existing literature. This suggests the importance of hitherto neglected pull factors.

A limitation of the present study is the absence of information about the legal and policy framework of potential hosts. This is unfortunate, as the absence of countries willing to accept refugees will have a clear negative effect on their numbers. Unfortunately, data on these legal frameworks and policies is lacking at the moment, but we hope that the present study's encouraging out-of-sample results will show the importance of incorporating more pull factors in studies on refugees, and hence of collecting these data and extending the present results.

References

Adhikari, Prakash. 2012. "The Plight of the Forgotten Ones: Civil War and Forced Migration." *International Studies Quarterly* 56(3):590–606.

Adhikari, Prakash. 2013. "Conflict-Induced Displacement, Understanding the Causes of Flight." *American Journal of Political Science* 57(1):82–9.

Apodaca, Clair. 1998. "Human Rights Abuses: Precursor to Refugee Flight?" *Journal of Refugee Studies* 11(1):80–93.

Beck, Nathaniel, Gary King & Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94(1):21–35.

Chadefaux, Thomas. 2014. "Early Warning Signals for War in the News." *Journal of Peace Research* 51(1):5–18.

Chadefaux, Thomas. 2017*a*. "Conflict Forecasting and Its Limits." *Data Science* 1(1):1–11.

Chadefaux, Thomas. 2017*b*. "Market Anticipations of Conflict Onsets." *Journal of Peace Research* 54(2):313–327.

Czaika, Mathias & Krisztina Kis-Katos. 2009. "Civil Conflict and Displacement: Village-Level of Determinants of Forced Migration in Aceh." *Journal of Peace Research* 46(3):399–418.

Davenport, Christian A., Will H. Moore & Steven C. Poe. 2003. "Sometimes You Just Have to Leave: Domestic Threats and Forced Migration, 1964-1989." *International Interactions* 29(1):27–55.

Gleditsch, Kristian S. 2002. "Expanded Trade and GDP Data." *Journal of Conflict Resolution* 46(5):712–24.

Gleditsch, Kristian S. & Michael D. Ward. 1999. "A Revised List of the Independent States since the Congress of Vienna." *International Interactions* 25(4):393–413.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted R. Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfeder & Mark Woodward. 2010. "A Global Model for Forecasting Political Instability." *American Journal of Political Science* 54(1):190–208.

Hoeting, Jennifer A, David Madigan, Adrian E Raftery & Chris T Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical science* pp. 382–401.

Ibez, Ana Mara & Andrea Velsquez. 2009. "Identifying Victims of Civil Conflict: An Evaluation of Forced Displaced Households in Colombia." *Journal of Peace Research* 46(3):431–51.

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

Lacina, Bethany & Nils P. Gleditsch. 2005. "Monitoring Trends in Global Combat: A New Dataset of Battle Deaths." *European Journal of Population* 21(2-3):145–66.

Marshall, Monty G., Ted R. Gurr & Barbara Harff. 2016. "PITF: Internal Wars and Failures of Governance, 1955-2015 Political Instability Task Force." `http://www.systemicpeace.org/inscrdata.html`.

Melander, Erik & Magnus Oberg. 2006. "Time to Go? Duration Dependence in Forced Migration." *International Interactions* 32(2):129–52.

Melander, Erik & Magnus Oberg. 2007. "The Threat of Violence and Forced Migration: Geographical Scope Trumps Intensity of Fighting." *Civil Wars* 9(2):156–73.

Melander, Erik, Magnus Oberg & Jonathan Hall. 2009. "Are New Wars More Atrocious? Battle Severity, Civilians Killed and Forced Migration Before and After the End of the Cold War." *European Journal of International Relations* 15(3):505–36.

Midlarsky, Manus I. 2005. "The Demographics of Genocide: Refugees and Territorial Loss in the Mass Murder of European Jewry." *Journal of Peace Research* 42(4):375–91.

Moore, Will H. & Stephen M. Shellman. 2004. "Fear of Persecution: Forced Migration, 1952-1995." *Journal of Conflict Resolution* 48(5):723–45.

Moore, Will H. & Stephen M. Shellman. 2007. "Whither Will They Go? A Global Study of Refugees Destinations, 1965-1995." *International Studies Quarterly* 51(4):811–34.

Neumayer, Eric. 2005. "Bogus Refugees? The Determinants of Asylum Migration to Western Europe." *International Studies Quarterly* 49(3):389–409.

Newland, Kathleen. 1993. "Ethnic Conflict and Refugees." *Survival* 35(1):81–101.

Osborne, Milton. 1980. "The Indochinese Refugees: Causes and Effects." *International Affairs* 56(1):37–53.

Pettersson, Therese & Peter Wallensteen. 2015. "Armed Conflicts, 1946-2014." *Journal of Peace Research* 52(4):536–50.

Schmeidl, Susanne. 1997. "Exploring the Causes of Forced Migration: A Pooled Time-Series Analysis, 1971-1990." *Social Science Quarterly* 78(2):284–308.

Stanley, William D. 1987. "Economic Migrants or Refugee from Violence? A Time Series Analysis of El Salvadoran Migration to the United States." *Latin American Research Review* 22(1):132–55.

Uzonyi, Gary. 2014. "Unpacking the Effects of Genocide and Politicide on Forced Migration." *Conflict Management and Peace Science* 31(3):225–43.

Uzonyi, Gary. 2015. "Refugee flows and state contributions to post-Cold War UN peacekeeping missions." *Journal of Peace Research* 52(6):743–757.

Ward, Michael D, Brian D Greenhill & Kristin M Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflicts." *Journal of Peace Research* 47(4):363–375.

Weidmann, Nils B., Doreen Kuse & Kristian S. Gleditsch. 2010. "The Geography of the International System: The CShapes Dataset." *International Interactions* 36(1):86–106.

Weiner, Myron. 1978. *Sons of the Soil: Migration and Ethnic Conflict in India.* Princeton: Princeton University Press.

Weiner, Myron. 1996. “Bad Neighbors, Bad Neighborhoods: An Inquiry into the Causes of Refugee Flows.” *International Security* 21(1):5–42.

Wood, William B. 1994. “Forced Migration: Local Conflicts and International Dilemmas.” *Annals of the Association of American Geographers* 84(4):607–34.

Zolberg, Aristide R., Astri Shurke & Sergio Aguayo. 1989. *Escape from Violence: Conflict and Refugee Crisis in the Developing World.* Oxford: Oxford University Press.

Appendix

A Flow vs. Stock

We explain here in greater detail our choice to use the total number of refugees (stock) rather than the flow as our dependent variable. Consider for example the case of Rwanda and Afghanistan. Table A.1 reports their total number of refugees (‘stock’), and the difference (‘flow’) between consecutive years (table A.1). Note that in 1992, the flow of Rwandan refugees is only 3,496, but this is ignoring the more than 400,000 refugees who remain out of the country. Their choice not to return is in itself significant, and the 434,736 refugees who made that choice should not be removed from the data. The choice to use the flow would also imply that the situation is more dire in 1990 than it is in 1995 because there are more refugees leaving the country in 1990 than in 1995. But this would be ignoring the fact that 1.8 million refugees remain out of the country in 1995—far more than the 361,322 in 1990.

Table A.1: Stock and Flow of Refugees in Rwanda, 1990–97.

Rwanda			Afghanistan		
Year	Stock	Flow	Year	Stock	Flow
1990	361,322	41,821	1987	5,511,740	417,457
1991	431,240	69,918	1988	5,622,982	111,242
1992	434,736	3,496	1989	5,643,989	21,007
1993	450,462	15,726	1990	6,339,095	695,106
1994	2,257,573	1,807,111	1991	6,306,301	-32,794
1995	1,819,366	-438,207			
1996	469,136	-1,350,230			
1997	68,003	-401,133			

Moreover, the flow may be affected by the number of people left in the country. In the extreme, all people who can flee may already have done so, so that the flow would stop. This does not mean that there are fewer refugees and hence does not match our intuitive understanding of the problem. Consider for example the case of Afghanistan in 1990. By then, more than a third of the country’s population (6.3 million out of a total remaining population of 12 million) had left the country. According to the ‘flow’ model, the number of refugees in 1991 was zero, since the flow was negative. But this completely ignores the 6 million refugees still out of the country.

Finally, the main data sources (the UNHCR and the United States Committee for Refugees and Immigrants) do not keep a record of flows but only of the total number of refugees. As a result, refugee flow calculations are controversial. Using the difference in stock from one year to the next is problematic. For example, it is theoretically possible that in 1991 in Afghanistan, 6,339,095 refugees returned home and 6,306,301 other people left the country.