THE DEMOGRAPHIC DETERMINANTS OF CONFLICT

Neil G. Bennett
CUNY Institute for Demographic Research

Bochen Cao
The University of Pennsylvania

Paper submitted for consideration for presentation at the 2013 meetings of the Population Association of America

Introduction

Armed conflict has long been a subject of inquiry among policy makers and scholars. A large volume of research has focused particularly on the political and economic factors that have given rise to such conflict. The great majority of demographic literature on the subject, however, has examined conflicts' consequences. One critical demographic consequence, a surge in mortality, may seem rather obvious, but often the quantification of worsening mortality rates is not all together straightforward. Ramifications with respect to other demographic phenomena, such as marriage and fertility, are also often difficult to assess (see, e.g., Heuveline and Poch 2007).

In contrast to the sizeable literature that explores the demographic consequences of conflict, the corresponding number of studies of the demographic determinants of conflict is relatively paltry. The proposed research aims to fill gaps in this literature. Ultimately, our objective is to be able to identify "hot spots" around the world and determine which countries are at relatively high risk of experiencing armed conflict at some time in the next several years. The hope is that, upon identification, other nations may be in a position to provide positive interventions that would cool down these areas and reduce the likelihood of conflict.

By no means is demography destiny. That is, the presence of certain demographic factors will not necessarily ensure that a country will devolve into armed conflict. That said, however, we explore the extent to which any of a variety of demographic factors, both in and of themselves, and in interaction with other sorts of factors – most notably, social, economic, political, and environmental – will heighten tension beyond the necessary threshold to result in conflict.

Literature Review

Past researchers have tested a large number of hypotheses that relate conflict to a variety of political, economic, social, environmental, and demographic risk factors. A widely accepted argument is that countries/regions with greater economic well-being are less likely to experience civil conflict. Economic well-being is measured in many different ways. Collier and Hoeffler (1998) use per capita income as one independent variable and argue that rebels will initiate a civil war if the perceived economic benefits outweigh the costs of rebellion. In a later paper, Collier and Hoeffler (2004) extend their model by grouping the risk factors of civil war into two categories, opportunity and grievance.

Many economic variables serve as a proxy for opportunity, such as GDP per capita, GDP per capita growth, and the fraction of primary commodity exports in GDP. Besides these economic variables, they also incorporate male secondary schooling to proxy opportunity. Grievance is proxied by ethnic or religious fractionalization, polarization, level of democracy, ethnic dominance, and inequality. In addition to the opportunity and the grievance model, Collier and Hoeffler also test these two sets of risk factors jointly and find that opportunity variables better explain the onset of civil war. In general, lower per capita income, lower male education, and decline in GDP growth, are significantly associated with increased risk of civil war. The proportion of primary commodity exports in GDP is significantly associated with the risk of civil war, but its effect is nonlinear, where the risk of civil war peaks when primary commodity exports compose 33 percent of the GDP. Most proxies for grievances, however, are insignificant; only dominance of one ethnic group would increase the likelihood of civil war. In contrast, Sambanis (2001) shows that the causes for ethnic civil wars are predominantly due to political grievances rather than the lack of economic opportunity.

Many other scholars have tested the effects of ethno-linguistic fractionalization on the incidence of civil war. Ellingsen (2000), for example, shows that a country has a substantially higher risk of civil conflict if the largest ethnic, religious, or linguistic group is less than 80 percent of the total population. Societies with several minority groups, semi-democracies (i.e., those neither extremely democratic (with a polity score close to 10) nor extremely autocratic (with a polity score close to -10) and lower per capita energy consumption are also more likely to experience domestic conflict.

Based on the Collier-Hoeffler (1998) model, de Soysa (2002) tests the effects of renewable and sub-soil assets, ethno-linguistic fractionalization, and percentage of Muslims and Christians in the total population on the onset of civil war during the post-Cold War era. Instead of allowing the variables to vary with time, de Soysa uses GDP per capita adjusted by purchasing power parity (PPP), trade openness, population size and density, and a democracy score and its squared term for one year only, 1989. The five-year average income growth rate between 1985 and 1989 is included as well. Controlling for these variables, de Soysa concludes that having abundant mineral wealth, being ethnically homogeneous, and being semi-democratic contribute to an increased incidence of civil war.

Hegre and Sambanis (2006) find that in addition to large population size, low income levels, and low rates of economic growth, recent political instability and inconsistently democratic institutions, small military establishments, rough terrain, and war-prone and undemocratic neighbors were robust factors that associated positively with the risk of domestic conflict.

Henderson and Singer investigate political, economic, and cultural factors among 90 post-colonial states of Africa, Asia, and the Middle East. Their findings show that semi-democracy has the greatest impact on the probability of civil war and further supports

previous findings that greater economic development reduces the likelihood. Also, militarized and Asian post-colonial states are more likely to experience civil war. Sambanis (2001) concludes that countries bordering with neighboring countries at war are significantly more likely to experience ethnic civil war. The country's own and their neighbors' levels of democracy were both significantly and negatively associated with conflicts.

Urdal has examined many demographic risk factors, the youth bulge in particular, in a series of studies. The youth bulge is measured as the proportion of 15 to 24 year-olds in the total adult population (15 years and above). In his modeling, Urdal (2004) also includes infant mortality rates, regime type, average annual change in GDP per capita over the five-year period prior to a given year, total population size, political dependency status, Communist state dissolution, brevity of peace, as well as previous conflict history. He finds that youth bulges are positively correlated with the risk of armed conflict. The effect of youth bulges is magnified following periods of negative economic growth. Youth bulges were also more strongly associated with conflict during the cold war period rather than later periods. Urdal (2005) includes population density and an environmental variable – cropland per capita, which is a single non-time-variant measure based on 1994-2001 land and population figures – to test neo-Malthusian concerns. He concludes that there is not a significant relationship between population growth and risk of internal armed conflict.

Urdal (2006) tested the effect of demographic youth bulges alone and in interaction with demographic dividend, economic growth, expansion in higher education, level of democracy, and urbanization on domestic armed conflict for the period 1950-2000 and terrorism and rioting for the years 1984-1995. He concludes that relatively large youth cohorts are associated with both armed conflict and terrorism/riots. However, its effect is not significantly magnified in countries with low levels of development and democracy. The

interactions of youth bulges with economic decline and expansion in higher education appear to increase the risk of terrorism. Presumably, well-educated young men are most frustrated about the lack of economic opportunity in the developing world and are the group most vulnerable to terrorist agitation.

Data

The data set we compile includes key variables related to our various hypotheses, drawn from a range of international data sets.

In this paper, our outcome of interest is the onset of domestic conflict. Data on domestic armed conflict are drawn from the Uppsala/PRIO data set (Gleditsch et al., 2002). A conflict is defined 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." The starting date and the ending date for each conflict are recorded and are then used to determine the duration of a conflict and duration between conflicts.

These start and end dates become crucial in determining whether a domestic conflict is new or a restart of an existing conflict that was inactive for some years. As Urdal (2006) suggests, we use two years as the threshold for distinguishing between the two possibilities. If a conflict is followed by at least two years of inactivity, then we define the subsequent conflict as a new onset.

The duration variable helps us to account for the time dependency in cross-sectional time-series data such as ours. An underlying assumption of logistic regression is independence across observations. Our data clearly violate this (Beck et al., 1998). As previous research has noted, countries that have experienced conflict in the past are more

prone to experience new incidents of domestic conflict. We adopt the same approach as Hegre et al. (2001) and Urdal (2006) to control for temporal dependence, by introducing a variable that is used to measure the reduction in the likelihood of domestic conflict after consecutive years of peace. The underlying assumption is that the longer a country is at peace, the less likely a new conflict will occur. Specifically, we model this reduction in risk as $exp\{(\text{-years in peace})/x\}$ (Hegre et al. 2001, Urdal 2006), where the value of x determines the rate of reduction of the effect from a previous conflict. The smaller the value of x, the faster the reduction. We assume the risk of occurrence of new domestic conflict reduces to half (0.5) three years after the last armed conflict onset, therefore the value of x is set to 4.328. (We have set the half-life to many other values and the results of the analysis are not affected significantly.) Therefore, the value of this new variable is one immediately after the onset of a conflict and declines towards zero over time. Since the observation for the data starts in 1950, we have no information regarding domestic conflict before that time. We must assign the value of the new variable arbitrarily to zero, thus assuming no country has experienced any conflict at a time not long before 1950. This may introduce bias since some countries actually did experience conflict in 1940s. However, since these events are not very common among the large number of observations we have, the bias can be ignored. Because the countries that have an ongoing conflict in a certain year may be more prone to experience new onsets of conflict in that and those new conflicts could be highly dependent on the ongoing conflicts, the coefficients tend to be overestimated. As a result, we omit the country-years for which there are ongoing conflicts.

Existing literature addressing the association between demographic factors and conflict has viewed the demographic transition and its relevant consequences as major determinants of conflict. In general, countries in the late stages of the demographic transition

tend to have a lower risk of conflict than those still in the earlier or middle stages, which are characterized by a high rate of infant mortality, a large proportion of youth, and rapid urban population growth. These features can exacerbate existing political instability and vulnerability to conflict. Other factors found to undermine political stability are the decline of natural resources on a per capita basis, changes in ethnic or religious composition, and the dramatic mortality increase due to HIV/AIDS (Cincotta et al., 2003).

Most of our demographic variables are drawn from the United Nations' *World Population Prospects* (UN, 2008). Since the United Nations data, spanning 1950-2008, are available for only every five years, the data for years in between are interpolated. Variables employed in this study are the following:

Population growth rate: Average exponential rate of growth of the population over a given period. As civil conflicts are associated with greed or grievance (Collier and Hoeffler, 2000), the competition over natural resources may worsen in some areas of the world given the unprecedentedly rapid population growth over the past century.

Infant mortality rate: Probability that a baby will die before his or her first birthday. It is expressed as deaths per 1,000 births. The infant mortality rate is usually employed in the literature as a proxy of development, where a lower infant mortality rate is more likely to be associated with a higher level of development. (To reduce the skewness of this variable and account for its diminishing effect, we employ a log-transformation.)

Population density: Population density is measured by population per square kilometer. Presumably, countries with high population density have a greater risk of domestic conflict than those with lower population density because of greater competition over resources in the former than in the latter. In addition, high population density, of course, results in a greater likelihood of encounter between people and may thus heighten the risk of conflict.

Population composition: Population by five-year age groups. Population composition incorporates the concept of youth bulge. On the one hand, a youth bulge could be of benefit to the welfare of a nation if the young adults are adequately educated and employed. Higher employment rates of youth would produce a "demographic bonus" by providing more financial support to children and the elderly, who are commonly not employed. This would create a vibrant economy in the country and in turn lower the risk of occurrence of domestic conflict. On the other hand, in countries where a large proportion of young adults are unemployed and they tend to lack education, greater tension may caused by frustrated youth who are likely to participate in rebellion and increase the chance of political instability, thereby rendering the country susceptible to domestic conflicts. Previous research shows that "the greater the unemployment among the educated youth, the greater are the propensities for dissatisfactions, instability and violence" (Choucri, 1974). Specifically, in the Arab world, youth bulges contribute considerably in explaining current political instability (Urdal, 2004).

The proportion of youth and the dependency ratios in a population may be calculated in a variety of ways. The dependency ratio is defined as the ratio of the number of people under age 15 and over age 65 to the number aged 15 through 64. Different definitions of "youth," such as people aged 15 through 24 years, 15 through 29 years, or 15 through 34 years are used to test the impact of the so-called "youth bulge." The various measures do not yield significantly different estimates, so we choose to use the conventional definition of youth bulge, namely those aged 15 through 24 years.

Urbanization:

Economic and educational opportunities and a more modern life style are attracting increasing numbers of people, especially youth, into cities. This trend is primarily seen in the developing world. On the one hand, this unprecedented urbanization in the developing world can create

a more vibrant economic environment. On the other, it also may lead to an oversaturated job market and exacerbated competition over current resources. Existing civil services and infrastructure as well as city budgets may no longer be adequate in the presence of an increasing number of in-migrants, especially when the most economically disadvantaged urban population continues to grow rapidly (National Intelligence Council, 2000; Nichiporuk, 2000). Characterized by disparate ethnic, religious, and regional subgroups and a large number of young students and workers, urban areas are often the center of social protest and labor unrest, as well as ethnic and religious conflicts (Gizewski & Homer-Dixon, 1995; Renner, 2000). The risk of domestic conflict rises dramatically when discontent rises among young politicized students and the unemployed and when groups find the need to compete rather than cooperate with each other.

Variables related to urbanization, between 1950 and 2008, are drawn from the United Nations' *World Urbanization Prospects* (UN, 2009). As was the case for data from *World Population Prospects*, data for in between years are interpolated. We use the urbanization rate in our models, which is defined here as the growth rate of the proportion of urban population in the entire population. We use a five-year period of this growth rate to account for likely sudden changes before or after conflicts.

Rural/Urban annual growth rate: Like rates of population growth, these rates are calculated on an exponential basis.

Natural Resources:

Some countries with scarce natural resources are experiencing especially heated competition as a result of the rapid population growth over the past century. Some developing countries are facing declining levels of arable land and renewable fresh water per capita because of other factors such as land degradation and failed government policies (Cincotta et al., 2003).

Insufficient sources of arable land and fresh water could impose constraints on food production, especially in many developing countries that have less open policies towards international trade. However, because of technology and trade, the scarcities of natural resources can be mediated to some extent. Hence, the association between natural resource scarcity and the risk of conflict may be weaker than other relationships we observe. Variables related to natural resources, for 1962 through 2007, are drawn from the AQUASTAT database (FAO, 2007). Data for in between years are here, too, interpolated. In conjunction with population data, these variables are then used to compute per capita values.

Arable land: This is operationalized as the percentage of all land under temporary crops, temporary meadows for mowing or pasture, land under market and kitchen gardens, and land temporarily fallow (less than five years).

Total internal renewable water: Long-term average annual flow of rivers and recharge of aquifers generated from endogenous precipitation.

Total actual external renewable water: The proportion of the country's annual renewable water resources that are not generated in the country.

Political Data: Variables related to political issues are collected from the Polity IV data set (Marshall and Jaggers, 2000). The key variable is the type of political regime, which assumes a value from -10 (most autocratic) to 10 (most democratic). Previous studies have found that highly autocratic regimes (more negative polity scores) and highly democratic regimes (more positive polity scores) are less likely to experience conflicts (Collier and Hoeffler, 1998; Hegre et al., 2001; Urdal, 2006). We first created several categories (dummies) for these scores to determine the functional form of the polity scores. We employ a quadratic form, as that represents the relationship well.

Ethnic/Religious Data: The percentages of the largest and second largest ethnic and religious groups in the total population are collected from the Ethnic Witches' Brew data set, covering the years 1945 through 1994 (Ellingsen, 2000). The growth rates of each individual group and the difference of growth rates for the largest and second largest ethnic and religious groups are then calculated. We hypothesize that if there is a difference in the growth rates of different ethnic/religious groups, tensions between these groups will emerge. Due to data availability, we calculate only the growth rates of the largest group and all other groups combined, and create a dummy variable that indicates if the largest group is growing faster than the others.

Economic Data: Most variables related to economic status, between 1950 and 2007, are drawn from the Penn World Tables (Heston et al., 2002), such as GDP per capita and GDP adjusted for purchasing power parities (PPP). Their logarithms are used in the analysis. In addition to these variables, we are able to measure countries' economic growth over time.

Trade openness: The sum of imports and exports as a fraction of gross domestic product. This variable is often used in the literature on growth and development as a proxy for globalization. It is predicated on the notion that trade leads to growth and development and in turn may reduce domestic tension and violence (Burgoon, 2006). Trade openness could also be a proxy of a country's attitude to open markets. We hypothesize that the greater tolerance a country's policy towards open markets, the less likely domestic conflicts will occur.

Education data: We focus on male secondary school enrollment as this is the group from which most rebels are recruited (Collier & Hoeffler, 2004). The data are drawn from the Global Education Database (USAID,2011) and are available from 1970 to 2010. Data before 1970 and missing data are extrapolated and imputed.

Geographic proximity: We hypothesize that a domestic conflict could diffuse beyond a nation's borders. A country whose neighboring lands have experienced domestic conflict recently may be more likely to have one itself. We retrieve the geographic coordinates of all countries and compute a distance matrix containing the distance between any two countries. For each country-year, we determine which other countries experienced the onset of domestic conflict within a certain period (12 months) and then find the nearest country/region among these and adopt that value of the distance. The choice of the length of the period is arbitrary. We assume the diffusion effect of domestic conflicts in neighboring countries to be minimal after two years and tested many values under two years, and the estimates are not affected substantially.

We merge all of the information mentioned above to form an extensive data set, covering 217 countries and regions over the period 1950 through 2008 containing 10,783 country-years.

Multiple Imputation

Bias that may be introduced by missing data is one of the major issues in studies of armed conflict. Most previous research has adopted a listwise deletion approach to address missing data. In cross-sectional research that involves a large number of countries/regions over a long period of time such as ours, data are usually difficult to collect for every country at every point of time for every variable due to a variety of reasons. There tend to be many observations in the data – that is, for a country in a given year – have at least one variable with a missing value. One method for handling missing data that is widely used by researchers and incorporated in statistical software is listwise deletion. However, it is an inefficient procedure in that it discards every observation with missing data and therefore

much potentially usable information is dropped even though the data are not missing. This causes larger standard errors, and hence weaker power in testing hypotheses. This approach also constrains the selection of regression models. The more variables we want to include in a model, the fewer complete observations will be available. In addition, whether data are missing depends on the dependent variable, listwise deletion may lead to biased estimates of regression coefficients (Allison, 2001). Since data are more likely to be missing for the years immediately after an armed conflict, we must be especially cautious implementing this approach.

As an alternative, many studies use dummy variables to address missing data. For example, Urdal (2006) created two dummy variables for inclusion in his regression to indicate respectively whether GDP growth data and regime data were missing. As a result, one can judge a missing value's "effect" on the probability of conflict, relative to the "effect" of non-missing values for a given independent variable. Although the estimates of standard errors are accurate in general (Jones, 1996), one cannot understand the content of that variable and thus cannot properly interpret the results.

In order to maximize our use of the available information and at the same time minimize potential biases, we use multiple imputation to impute missing values. As indicated by its name, multiple imputation fills in each missing value with several imputed values, producing several complete data sets with the same observed data but different imputed data. Since we retain all the available information in the original data set and minimize the uncertainty by imputing the data set multiple times, the estimates we get are more robust. We use the Markov Chain Monte Carlo (MCMC) algorithm to accomplish this task.

The multiple imputation procedure can be divided into three steps: First, we generate several data sets by randomly drawing different imputed values each trial randomly.

Generally, approximately five trials of imputations are sufficient to produce robust and reasonably efficient results (Allison, 2001; Honaker & King, 2010). We generate ten imputed data sets to assure the results are reliable and to get sufficiently stable p-values as well as confidence intervals (Allison, 2001).

Second, we apply conventional regression methods to each of the imputed data sets. Last, we combine the results from this second step and follow Rubin's (1987) rules to compute the final point estimates and standard errors. For the point estimates, we simply take the average of the estimates over all imputed data sets. For the standard errors, we first compute the "within" variance by taking the mean of the squared standard errors across all imputed data sets; then we compute the "between" variance, which is the sample variance of the point estimates of the parameters across all imputed data sets. Last, we take the square root of the sum of the within and between variances, after applying a correction factor to the latter. This is represented by:

$$\sqrt{\frac{1}{M}\sum_{k=1}^{M} s_k^2 + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{k=1}^{M}(b_k - \bar{b})^2}$$

where M is the number of the imputed data sets, s_k is the standard error of the kth data set, and b_k is the point estimate of the parameter in the kth data set.

Model

We use logistic regression to investigate the determinants of domestic armed conflict, since our dependent variable for the onset of domestic conflict is a dichotomous outcome, where '1' indicates the country had experienced a conflict in a given year, and '0' indicates that it had not.

However, we do not use the onset of domestic conflict as our independent variable. Domestic conflicts are rare events. Among all the country-years in our data set, a total of 246 domestic conflicts occurred during the entire time period, comprising 2.28 percent of all the observations. Using logistic models to predict the probabilities of rare events, despite having a sample as large as ours, will lead to underestimation of both the regression coefficients and estimated event probabilities (King and Zeng, 2001). Although this problem is often ignored, we use the software package Relogit (Rare Event Logistic Regression), written by King and his colleagues, to correct these biases. We choose, however, not to present the results here because the main purpose of our models is to predict future incidents of domestic conflicts rather than to infer the causal mechanisms. Testing the model using Relogit, we obtain R-square values that are very low. Another criterion for testing the predictive ability of the model, the area under ROC curves (_____, ____), which we discuss below, is also not quite satisfactory for this model.

Rather than focus upon the onset of domestic conflict in a single year, we examine the probability that an onset of domestic conflict will occur during the next three-year, or five-year, period. We construct a dichotomous variable for which a country or region is attributed a value of one if it had experienced the onset of conflict in the next three years, and zero otherwise. We do the same in constructing a five-year variable. A drawback of relying on such multi-year variables is that we lose several observations referring to the last several years before 2008, the last year for which we have data. Specifically, the last year for which we can conduct out-of-sample forecasting is 2005 using the three-year conflict variable and 2003 for the corresponding five-year variable.

ROC Curves

For logistic model, a cutoff value must be set beforehand to determine how the thousands of continuous values of the estimated probabilities to be classified into only two groups with value 0 and 1. Conventionally, the cutoff point is set at 0.5. In the context of domestic conflict, if the estimated probability for a certain country-year is greater than 0.5, then it is predicted that the country/region is likely to experience an onset of domestic conflict in the next three- or five-year period. If the estimated probability is otherwise lower than or equal to 0.5, then the country is predicted not to experience conflict in that time period. In fact, almost none of the previous studies has predicted the experience of domestic conflict with a probability greater than 0.5. Therefore, we need to chose a cutoff that would maximize the rate of countries/regions are predicted to have an onset of conflict during a certain time period and they actually had one (true positive), and minimize the rate of countries/regions are predicted to not to have an onset of conflict during a certain time period but they actually had one (false negative). We are more concerned about true positive and false negative than true negative and false positive because they are relatively more important for policy makers in making decisions. Maximizing true positive could direct the resources and attentions to be located to the countries/regions at real risks. On the other hand, a false negative may cause under-preparation and end up with undesirable domestic conflicts.

Receiver operating characteristic (ROC) curves could help us to find the cutoff that optimizes the predictions of the model. ROC curves provide a way to summarize the accuracy of predictions visually and comprehensively. ROC curves plot the true positive rate and one minus the false negative rate at different values of the cutoff. Sometimes it is better to have a quantitative summary measure of the ROC curves than just looking at the curves graphically. The area under the ROC curve (AUC) is most commonly used for this purpose.

Typically, when comparing two or models, the larger the AUC, the higher accuracy of the predictions.

In contrast to Urdal's (2004, 2006), we argue that male youth are more violent than female youth in general and are more responsible for the initiation of armed domestic conflicts. Therefore we use the proportion of males aged 15 to 24 years old in the male adult population, where the male adult population is composed by males aged 15 years old and above. As the data show, for 82.66% of the country-years, the proportions of male youth in male adult population are greater than the proportions of both-sexes youth in the entire adult population.

Results

According to past research, a large youth cohorts is one of the most essential demographic risk factors that would increase the probability of the occurrence of domestic conflict. We begin our modeling process with by estimating some parsimonious models with youth bulge variables as well as some other demographic, political, and economic risk factors included. The coefficients estimated in Model 1 are consistent with theory and past findings in general. Higher infant mortality rates, higher proportions of youth in the population, larger population size, slower GDP growth, as well as a smaller value of the square of polity 2 score are associated with higher risks of the occurrence of new domestic conflict onset in the next five-year period. Countries/regions with high infant mortality rates may also perform poorly in many other respects, and hence increase the risk of potential internal conflicts. The positive coefficient of the log of infant mortality rate suggests this hypothesis is plausible. The statistically insignificant coefficient of polity 2 score in conjunction with the statistically significant negative coefficient for its square term confirm the widely accepted argument that

the political regime type has a inverted-U shape relationship with domestic conflict, with the most democratic and most autocratic regimes being less likely to experience domestic conflict than the semi-democratic regimes. Also, countries/regions with larger population sizes are more difficult for governments to manage and therefore may suffer greater internal tension that will eventually lead to armed conflict, as shown by the positive coefficient of population size variable. As we surmised, the coefficient of the variable that accounts for temporal dependency on a previous outbreak of conflicts is positive, meaning the effect of an existing domestic conflict on the next onset reduces over time.

The directions of the effects of the variables discussed above remain virtually unchanged for all the models, and the magnitude of these effects vary only marginally. In Model 2, we included the interaction of the proportion of males aged 15-24 and mean GDP growth rate in five years. The coefficients suggest that greater economic growth is associated with a lower risk of experiencing the onset of domestic conflict when interacted with the presence of a large male youth cohort. The impact of the size of male youth cohorts becomes larger, while the effect of GDP growth becomes positive but statistically insignificant, in contrast to the result shown in Model 1.

Education is viewed as creating opportunities, such that higher levels of educational attainment may enable youth to find better paying jobs and, in turn, increase the opportunity cost for those youth to engage in rebellious activity. Since most rebels are recruited from youth who are of secondary or tertiary school age, as Thyne (2006) has suggested, we test this in Model 3, including tertiary education enrollment. Enrollment is log-transformed to reduce the skewness of the distribution. The result suggests the effect of tertiary education enrollment is not statistically significant. We add the interaction of the youth bulge variable and this education variable in a model not shown, and the coefficient of interaction term is

virtually zero and not significant. We also test secondary school enrollment, and tertiary school enrollment for only males, none of these appear to have significant impact on civil conflict onset in the next five-year period.

We then test the association between a country/region's openness toward international trade and the probability of incidence of new domestic conflicts. Model 4 shows that the larger the fraction of both imports and exports in GDP, the lower the risk of a country experiencing any domestic conflict onset in the next five years. Compared to Model 1, trade openness would mitigate the effect caused by large population size by more than 0.1, while having essentially no effect on other variables. In model 5, we include the interaction between the male youth bulge and trade openness. The magnitude of the coefficient of trade openness is almost three times that in Model 4, while the sign of this variable remains unchanged. The coefficient of the interaction term is positive and of borderline significance (p=0.066). These results suggest that trade openness alone would have reduced the risk of domestic conflicts even more, but if a country is relatively more open towards international trade and has a large proportion of youth in the population at the same time, the positive effect of openness to international trade on maintaining domestic peace would be slightly reduced by this interaction.

Model 6 investigates the association between urbanization and risk of new episodes of domestic conflict in the next five years. The urbanization rate is negatively associated with the risk of domestic conflict onset. Adding the mean urbanization rate over a five-year period in the model barely affects the estimate of any variable except increasing the estimate of the log infant mortality rate by about 0.08 and making the linear term of the regime type variable significant and larger in its magnitude comparing to Model 1. This suggests urbanization may

lower both the infant mortality rate and the risk of new onsets of domestic conflict, as well as increase the degree of democracy in a country.

Model *7* shows how religious composition is associated with the likelihood of domestic conflicts. We compare the growth rate of the largest religious group and the remaining groups as a whole in each country and region. This variable is classified into three categories, namely the largest religious group grows faster than the other groups, the largest religious group grows as fast as the rest groups as a whole, and the largest group grows more slowly. We used the second category as the reference group. Compared to all religious groups growing at the same rate, the coefficients in Model *7* indicates that a country will be subject to lower risk when the largest religious group grows faster. Since the largest religious group in most cases is a majority of the population (true for over 95 percent of the country-years), faster growth rate will make the difference in proportion in population between this group and the rest even larger and hence result in a more homogeneous population. In contrast, if the largest religious group grows more slowly, then the difference in proportion in population will decline and as a result the population would become more diversified. As de Soysa (2002) notes, the organizational costs of engaging in religious competition will be larger if the population is highly diversified, and the opportunity costs for fighting against the largest religious group will also be high. Therefore, the countries/regions in which the largest religious group grows faster will be less likely to experience domestic conflicts. In a model not shown, we include the interaction between the youth bulge variable and this religious measure; the interaction term turns out to be statistically insignificant and the significant effect of the religious variable is washed out. We also attempt to test a similar model for ethnic composition, but find it is not significantly associated with the risk of experiencing

conflicts in the next five-year period, although the coefficient suggest if the largest ethnic group grows faster the risk will be lower.

We expect the abundance of natural resources would increase the likelihood of remaining in peace. In Model 8, we add a variable that measures the log of percentage of arable land in a country/region. The point estimate indicates that the greater the percentage of arable land, the less likely a new domestic conflict would occur in the next five years, however, this association is statistically insignificant. We then add the interaction between the male youth bulge variable and the variable for arable land. The coefficient of the arable land variable now contradicts our hypothesis, in that it is positively associated with the likelihood of an occurrence of civil conflict in the next five years. However, the interaction term has a significant negative effect on the dependent variable, suggesting the abundance of arable land will mitigate the risks of civil conflict brought about by large male youth cohort. We suspect the positive coefficient for the log of percentage of arable land is a result of its correlation with urbanization. When we include the urban growth variable in Model 10, the coefficient of the arable land variable declines slightly, but becomes insignificant.

We take one step further and use the interaction between the urban growth rate, substituting for the logged percentage of arable land, and the size of male youth cohort in Model 11. As expected in our earlier hypothesis, a higher urban growth rate is associated with a lower risk of civil conflict onset in the next five-year period. However, the interaction of the size of male youth cohort and the urban growth rate is positively associated with the risk of civil conflict onset. This indicates that although urbanization itself would make countries/regions less likely to experience internal upheaval due to the boosted economic and educational opportunities produced during the urbanization process, the youth cohort would wash this desirable effect away by depleting urban resources, especially when the size of

youth cohort is relatively large. In addition, the coefficient of the arable land variable becomes negative but insignificant. Once the effects of urbanization are controlled, then, the higher the proportion of the land in a country is arable, the less likely this country would experience civil conflict, although its association is statistically weak. We test other specifications for natural resources variables, such as total arable land, arable land per capita, the growth rate of arable land, total renewable water, renewable water per capita, and the growth rate of total renewable water. None of these variables is associated significantly with the risk of experiencing civil conflict in the next five years.

In Model 12, we included all the significant variables. The coefficient estimates show that large size of population, higher infant mortality rates and larger male youth cohort are associated with a greater risk of experiencing civil conflict onset in the next five-year period, while a larger square term of the regime type variable, faster GDP growth, extended trade openness, higher urban growth rate, and a greater growth rate of the largest religious group relative to the remaining religious groups as a whole are associated with lower risks.

Out-of-Sample Forecasting

Although we have made considerable progress towards understanding the determinants of domestic conflict, we now seek to predict the onset of conflict. We adapt the technique of out-of-sample forecasting to judge our methodology's ability to predict conflict.

We split the observed sample into two parts. Data from 1950-2000 are used as training data to fit our statistical models, while 2001-2008 data are used for out-of-sample testing. This reduces our sample size since we are now pretending we know nothing about the incidents of conflict that occurred between 2001 and 2008 and therefore the dependent variable for the last several years of the training data have to be set to missing. However, the

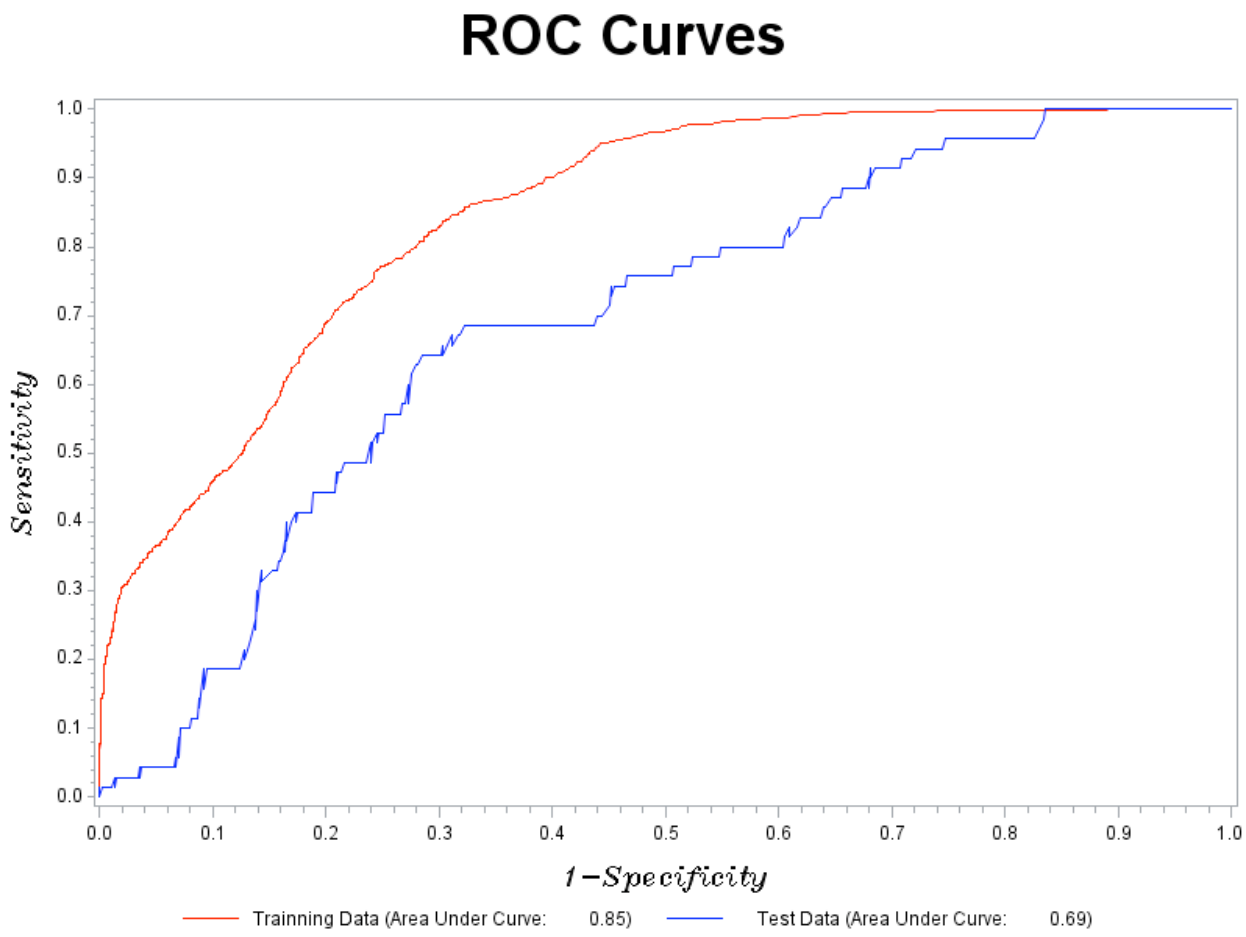
number of observations dropped due to this reason is relatively small and should not be of concern, given we have recovered more data using multiple imputation. Thus, applying our best-fitting model to the earlier data (i.e., 1950-2000), we forecast incidents of conflict for the latter time period (2001-2008). We then compare these predictions with that which actually occurred and determine the accuracy of our forecasting methodology. In the event that we are indeed successful, then we will estimate near-term probabilities of conflict for countries around the world.

We produce the ROC curves, with which we can assess the accuracy of the prediction. Since the independent variable, the predicted probability of domestic conflict onset in the next three or five years, is continuous, while the real event is binary, we will need a threshold to determine whether the event occurs or not based on the predicted probability. The ROC curves provide visual representation of the assessment of accuracy at different threshold levels. The proportion of true positive (sensitivity) is shown on the y-axis, and the proportion of false positive (1-specificity) is shown on the x-axis. The ideal model would have as high a proportion of true positive and as small a proportion of false positives as possible. In other words, we would expect the point that represents the ideal model to be located on the upper left corner of the plane.

As the graph shows, if, for example, we set the threshold at 0.50, which is the default for most previous studies that use logistic models, we would capture only about 25 percent of the cases that would actually experience conflict in the next five years for the training data. We would incorrectly forecast peace for the remaining 75 percent of cases that, in reality, would experience conflict during the five-year period. On the other hand, adopting this threshold correctly predicts peace among more than 98 percent of the cases that would not experience any conflicts in the next five years.

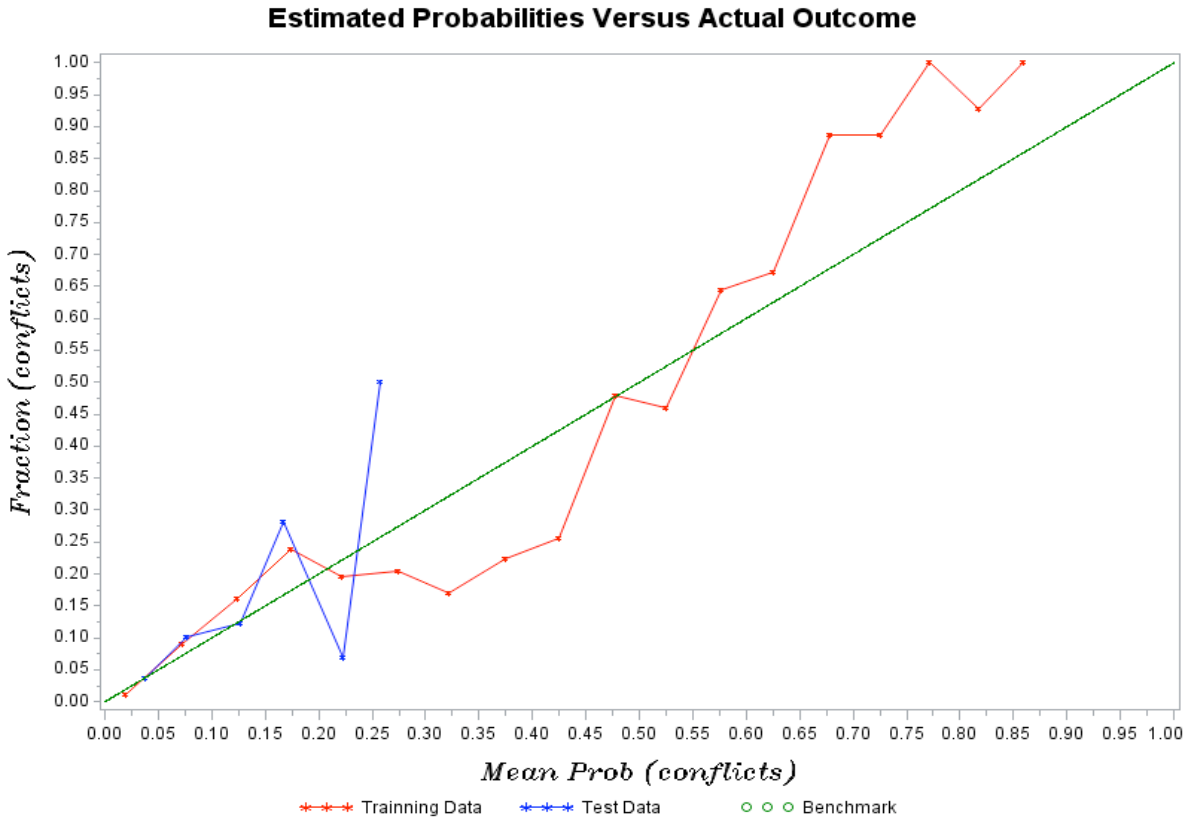
The 0.50 threshold would render a worse prediction for the test data. At this threshold level, the out-of-sample test would capture none of the cases that will actually experience conflict in the next five years and predict all the cases that would not experience conflict in the next five years correctly. The consequences of setting the threshold at 0.50 level would be very costly given the results from the out-of-sample test, since in this scenario we would predict that no country would experience domestic conflict in the next five years. However, if we set the threshold at a much lower level, say 0.08, we would successfully capture about 87 percent of the cases that will actually experience conflicts in the next five years correctly, but at the same time also incorrectly classify 34 percent of those that will actually not experience conflicts in the next five years to experience conflicts for the in-sample test with training data. For the out-of-sample test with test data, we would correctly classify 80 percent of the cases that eventually would experience conflict in the next five years, but also incorrectly predict 55 percent of the cases that would experience only peace in the next five years to experience conflicts. Although we may be able to improve the predictive power of our model by altering the threshold level, we are not likely to produce predictions with the test data as accurately as with the training data given the area under curve (AUC) is much less for the former than the latter (0.69 vs. 0.85).

Setting the threshold at 0.08 is certainly not perfect in that we are still not able to capture all the high-risk countries and would misclassify more of the low-risk countries as high-risk countries. However, one can argue that it is better to set the threshold too low than too high, since such abundance of caution would lead perhaps to unnecessary spending to prevent conflict in places that would not experience domestic conflict in the absence of such spending. In contrast, a lack of caution and necessary intervention might well result in what could have been avoidable death and injury, as well as economic and social deterioration.



An alternative means of testing the predictive power of the models, as King and Zeng (2001) demonstrate, is to plot the observed fraction of conflicts in each predicted probability interval against the mean predicted probability in the same interval. The two quantities would be very close if the model has good predictive power. An ideal model would produce such plots coinciding with the main diagonal. In other words, if the probability of having domestic conflict is predicted to be 0.5 in the next five years, then 50 percent of the countries/regions with such a prediction would actually have conflict in that time period. According to the graph, for the in-sample test, the model tends to underestimate the true likelihood of domestic conflict when the estimated probability is high and tends to overestimate when the estimated

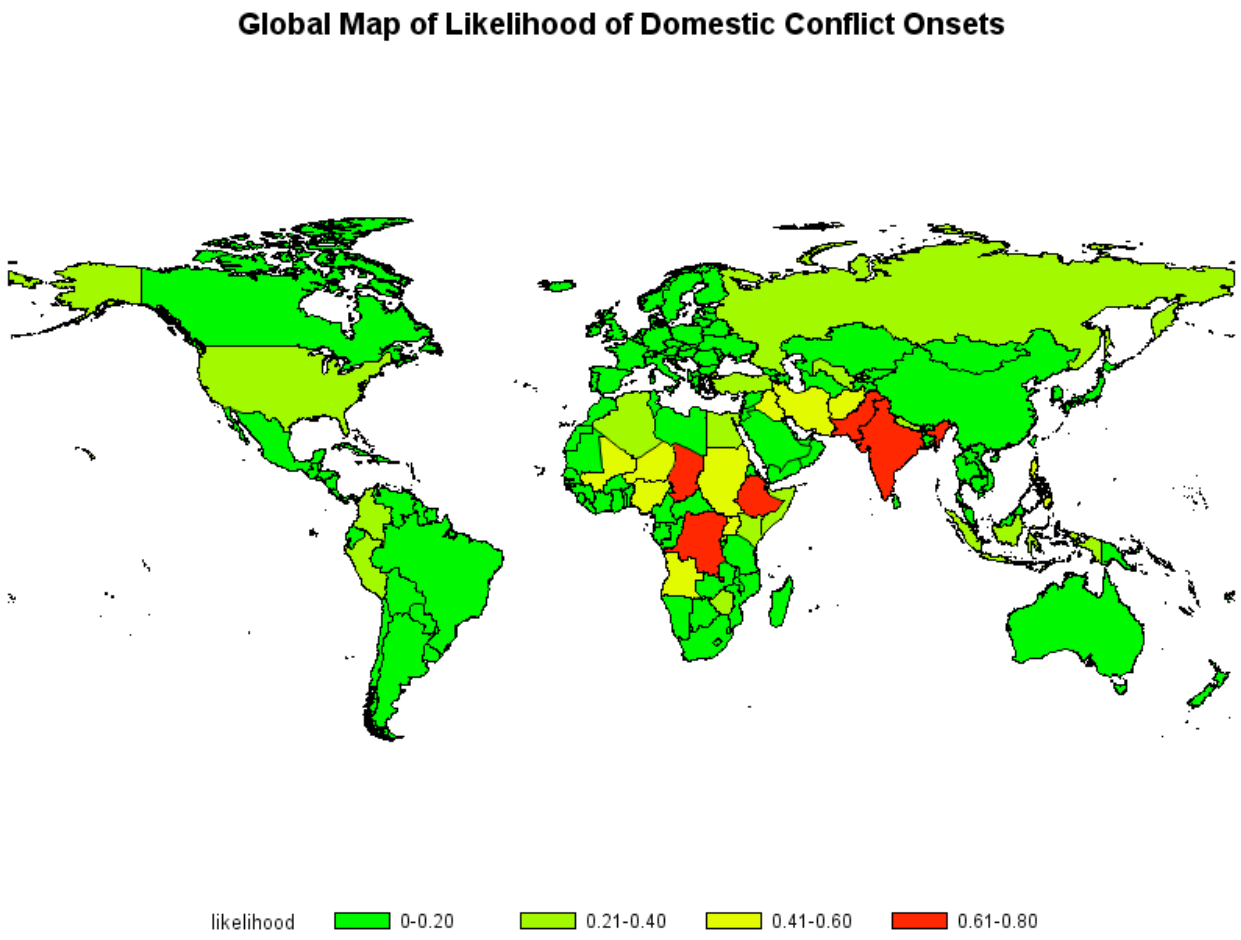
probability is low. One may note that the line is still relatively close to the diagonal line. In addition, since we are using threshold to classify the binary outcomes, the biases in estimated probability will not affect the predicted outcome if we set the threshold at level lower than 0.50, although we might be over cautious especially when the model already overestimate the probability when the predicted probability is below 0.50. For the out-of-sample test, the line does not reach the upper right corner as the one for in-sample test does, meaning the predictive power is weaker in the out-of-sample test than the in-sample test. Also the model predicts the occurrence of domestic conflict in the next five years better when the predicted probability is relatively low.



Using the available data at 2008, we estimate the probability of occurrence of domestic conflict onset in the next five years using the coefficients obtained from the regression with

the training data. In addition, we create a global map of "hot-spots" determined from the estimated probabilities.

{MORE TO COME}



Appendix: Methodological Concerns

An underlying assumption of logistic regression is that the independent variables are linearly related to the dependent variable. However, this assumption is often violated among actual data, as the effect of an explanatory variable is seldom linear or monotonic. Therefore, logistic regression may not always be valid to provide either an acceptable goodness-of-fit or

accurate out-of-sample forecasting given a complex data structure. Such is the case here, in which domestic conflict is driven by highly nonlinear processes and complex interactions, as noted by a number of political scientists (e.g., Beck, King, and Zeng, 2000). The difficulty of choosing an appropriate functional form in these circumstances to model the relationship between dependent and independent variables may cause model misspecification.

We attempt to use nonlinear models to explore the determinants of conflict. The models we choose are generalized additive models and neural network models. There are some issues preventing us to present those models in this paper. Generalized additive models are non-parametric and contain a smoothing function for each independent variable. They provide the assessment of significance of each independent variable using a chi-square test that compares the deviation between the full model and the model without the variable (Hastie and Tibshirani, 1990). It is complex, however, for us to test the out-of-sample forecasting power of this model. Additionally, as we incorporate more variables into the model, it is less likely that the model will converge. Given that we have ten multiple imputed data sets, the problem is compounded. To some extent, neural network models are generalized logistic models with additional levels of hierarchy, if the link functions between different levels of neurons are properly specified. Hence, to some degree, they can be considered parametric models. Neural network models are very flexible and able to approximate any form of data with relatively few variables (Ripley, 1996). This feature of neural network models raises several issues. First, they tend to overfit data since they are good at picking up the idiosyncrasies of data (Beck, King, and Zeng, 2001). Even if we split the data for out-of-sample forecast to validate the models, this approach fits the training data much better than the testing data, given the training data are of larger size. Second, since the neural network models require fewer variables than logistic models do, sometimes they

outperform logistic models with many fewer variables. This renders the comparison between the two families of models difficult because neural network models are more complex and it is not easy to interpret the effects of variables in neural network models in comparison with those in the logistic models.

References

Allison, Paul D. 2001. *Missing Data* (1st ed.). Thousand Oaks: Sage Publications, Inc. ISBN 978-0761916727.

Beck, Nathaniel, Jonathan Katz, and Richard Tucker. 1998. "Taking time seriously: Time-series-cross-section analysis with a binary dependent variable." *American Journal of Political Science* 42(2): 1260-1288.

Beck, Nathaniel, Gary King, and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." *American Political Science Review* 94: 21-36.

Besancon, Marie L. 2005. "Relative Resources: Inequality in Ethnic Wars, Revolutions, and Genocides." *Journal of Peace Research* 42(4): 393-415.

Buhang, Halvard and Jan Ketil Rod. 2006. "Local determinants of African Civil Wars 1970-2001." *Political Geography* 25(3): 315-35.

Burgoon, Brian. 2006. "OnWelfare and Terror." *Journal of Conflict Resolution* 50(2): 176–203.

Choucri, Nazli. 2002. "Migration and Security: Some Key Linkages." *Journal of International Affairs* 56(1): 97-125.

Cincotta, Richard P. 2008-2009. "Half a Chance: Youth Bulges and Transitions to Liberal Democracy." *ECSP Report* 13: 10-18.

Cincotta, Richard P., Robert Engelman, and Daniele Anastasion. 2003. *The Security Demographic: Population and Civil Conflict after the Cold War*. Washington, DC: Population Action International.

Collier, Paul and Anke Hoeffler. 1998. "On Economic Causes of Civil War." *Oxford Economic Papers* 50(4): 563-573.

Collier, Paul and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 50(4): 563-73.

de Soysa, Indra. 2002. "Ecoviolence: Shrinking Pie or Honey Pot?" *Global Environmental Politics* 2(4): 1-36.

de Soysa, Indra. 2002. "Paradise is a Bazaar? Greed, Creed, and Governance in Civil War. 1989-99." *Journal of Peace Research* 39: 395-416.

Ellingsen, Tanja, and Wenche Iren Hauge. 1998. 'Beyond Environmental Scarcity: Causal Pathways to Conflict', *Journal of Peace Research* 35(3): 299–317.

Ellingsen, Tanja. 2000. "Colorful Community or Ethnic Witches' Brew? Multiethnicity and Domestic Conflict during and after the Cold War." *Journal of Conflict Resolution* 44: 228-49.

Fearon, James D. and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97: 75-90.

Gizewski P, and T. Homer-Dixon. "Urban Growth and Violence: Will the Future Resemble the Past?" Project on Environment, Population and Security. Washington,DC: American Association for the Advancement of Sciences, 1995.

Gleditsch, Nils Petter, Petter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Ha Vard Strand. 2002. Armed Conflict 1946–2001: A New Dataset. *Journal of Peace Research* 39:615–637.

Goldstone, Jack A. 2002. "Population and Security: How Demographic Change Can Lead to Violent Conflict." *Journal of International Affairs* 56(1): 3-23.

Hastie, T.J. and R.J. Tibshirani. 1990. *Generalized Additive Models*. New York: Chapman and Hall.

Hauge, Wenche and Tanja Ellingsen. 2001. "Causal Pathways to Conflict," pp. 36-57, in Paul F. Diehl and Nils Petter Gleditsch, eds., *Environmental Conflict*. Boulder, CO: Westview.

Hegre, Havard and Nicholas Sambanis. 2006. "Sensitivity Analysis of Empirical Results on Civil War Onset." *Journal of Conflict Resolution* 50: 508-35.

Henderson, Errol A. 2000. "When States Implode: The Correlates of Africa's Civil Wars 1950-92." *Studies in Comparative International Development* 35(2): 28-47.

Henderson, Errol A. and J. David Singer. 2000. "Civil War in the Post-Colonial World, 1946-92." *Journal of Peace Research* 37: 275-99.

Heuveline, Patrick and Bunnak Poch. 2007. "The Phoenix Population: Demographic Crisis and Rebound in Cambodia." *Demography* 44(2): 405-426.

Honaker, James and Gary King. 2010. "What to Do About Missing Values in Time Series Cross-Section Data." *American Journal of Political Science* 54: 561-581.

Jones, M.P. 1996. "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association* 91: 222–230.

Kahl, Colin. 2002. "Demographic Change, Natural Resources and Violence: The Current Debate." *Journal of International Affairs* 56(1): 257-282.

King, Gary, and Langche Zeng. 2001. "Explaining Rare Events in International Relations." *International Organization* 55: 693-715.

King, Gary and Langche Zeng. "Improving Forecasts of State Failure." *World Politics,* 53 (July 2001):623-58.

Mesquida, Christian G. and Neil I. Wiener. 1996. "Human Collective Aggression: A Behavioral Ecology Perspective." *Ethology and Sociobiology* 17: 247-62.

Nichiporuk B. "The Security Dynamics of Demographic Factors," MR-1088 WFHF/RF/DLPF/A. Santa Monica: RAND,2000.

Østby, Gudrun, & Urdal, H. (2010). *Education and civil conflict: A review of the quantitative, empirical literature* (p. 26).

Raleigh, Clionadh and Henrik Urdal. 2007. "Climate Change, Environmental Degradation, and Armed Conflict." *Political Geography* 26: 674-94.

Renner M. Environmental and Social Stress Factors, Governance, and Small Arms Availability: The Potential for Conflict in Urban Areas, in: *Urbanization, Population, Environment and Security* (Rosan C, Ruble BA, Tulchin JS, eds), pp. 51–72. Washington, DC: Woodrow Wilson International Center for Scholars, 2000.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Rubin, Donald B. 1994. "Missing Data, Imputation, and the Bootstrap: Comment." *Journal of the American Statistical Association* 89(426): 475–78.

Sambanis, Nicholas. 2001. "Do Ethnic and Non-Ethnic Civil Wars Have the Same Causes? A Theoretical and Empirical Inquiry (Part 1)." *Journal of Conflict Resolution* 45: 259-82.

Sambanis, Nicholas. 2002. "A Review of Recent Advances and Future Directions in the Quantitative Literature on Civil War." *Defence and Peace Economics* 13: 215-43.

Staveteig, Sarah. 2005. "The Young and the Restless: Population Age Structure and Civil War." *ECSP Report* 11: 12-19.

Theisen, Ole Magnus. 2006. *Other pathways to conflict? Environmental Scarcities and Domestic Conflict*. Paper presented at the 47th Annual Convention of the International Studies Association, March 22-25, San Diego, CA.

Toft, Monica D. 2007. "Population Shifts and Civil War." *International Interactions* 33: 243-69.

Urdal, Henrik. 2004. "The Devil in the Demographics: The Effects of Youth Bulges on Domestic Armed Conflict 1950-2000." Social Demographic Papers 14. Washington, DC: Conflict Prevention and Reconstruction Unit, World Bank.

Urdal, Henrik. 2005. "People vs. Malthus: Population Pressure, Environmental Degradation, and Armed Conflict Revisited." *Journal of Peace Research* 42(4): 417-434.

Urdal, Henrik. 2006. "A Clash of Generations? Youth Bulges and Political Violence." *International Studies Quarterly* 50: 607-29.

Urdal, Henrik. 2008. "Population, Resources, and Political Violence: A Subnational Study of India, 1956-2002." *Journal of Conflict Resolution* 53(4): 590-617.

USAID. 2011. Global Education Database, http://ged.eads.usaidallnet.gov/data/.

Variables	Model1	Model2	Model3	Model 4	Model 5	Model 6	Model 7a	Model 7b	Model 8	Model 9	Model 10	Model 11	Model 12	Bivariate Models		
Intercept	-9.666 ***	-9.915 ***	-9.403 ***	-10.082 ***	-12.262 ***	-9.777 ***	-9.547 ***	-9.477 ***	-9.458 ***	-10.901 ***	-10.943 ***	-9.001 ***	-10.021 ***	Coeff	R-Square	N
log Population	0.45 ***	0.448 ***	0.348 ***	0.341 ***	0.339 ***	0.446 ***	0.443 ***	0.438 ***	0.396 ***	0.394 ***	0.389 ***	0.394 ***	0.339 ***	0.437 ***	0.13783	7640
log IMR	0.228 **	0.232 ***	0.769 ***	0.29 ***	0.291 ***	0.305 ***	0.24 ***	0.246 ***	0.549 ***	0.572 ***	0.628 ***	0.666 ***	0.354 ***	0.584 ***	0.04674	7640
Polity2	-0.016	-0.015	-0.01	-0.013	-0.013	-0.021 *	-0.015	-0.015	-0.003	-0.003	-0.007	-0.008	-0.017 *	-0.033 ***	-	7640
Polity Sq.	-0.003 *	-0.002 *	-0.002	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.002 *	-0.005 ***	0.02283	7640
%male 15-24	0.092 ***	0.1 ***	0.046 **	0.087 ***	0.153 ***	0.091 ***	0.091 ***	0.09 ***	0.066 ***	0.105 ***	0.104 ***	0.042 *	0.086 ***	0.079 ***	0.03636	7640
Mean GDP Growth in 5 years	-0.012 ***	0.009	-0.008 *	-0.009 ***	-0.01 ***	-0.011 ***	-0.011 ***	-0.011 ***	-0.011 *	-0.011 **	-0.011 **	-0.01 **	-0.009 ***	-0.017 ***	0.04538	7640
%male 15-24*Mean GDP Growth in 5 years		-0.001 *														
Log Tertiary School Enrolment			0.068											-0.236 ***	0.02626	4281
trade openness				-0.16 ***	-0.427 **								-0.148 ***	-0.324 ***	0.1117	7640
%male 15-24 * trade openness					0.008											
Mean Urbanization Rate						-0.091 **					-0.068 *	-0.64 *	-0.067 *	0.017	0.00022	7640
%male 15-24* Mean Urban. Rate in 5-year												0.016 *				
Largest Reli. Group Grows Faster (vs. Equal or Slower)							-0.403 **							-0.673 ***	0.01381	7640
Largest-Reli. Group Grow>0 (vs. Equal)								-0.456 ***					-0.336 *	-0.851 ***	-	
Largest-Reli. Group Grow<0 (vs. Equal)								-0.187					-0.058	-0.606 ***	0.02574	7640
Log Perct of Arable land									-0.032	0.538 *	0.53	-0.035		0.153 ***	0.0109	5641
%male 15-24*Log Perct of Arable land										-0.016 *	-0.016 *					
risk decay	0.564 ***	0.564 ***	0.804 ***	0.571 ***	0.571 ***	0.558 ***	0.569 ***	0.573 ***	0.74 ***	0.74 ***	0.734 ***	0.729 ***	0.572 ***	0.739 ***	0.139	7640
N	7640	7640	4281	7640	7640	7640	7640	7640	5641	5641	5641	5641	7640			
Adj R-Square	0.29914	0.30105	0.37794	0.31365	0.31453	0.30154	0.30233	0.30305	0.35142	0.35291	0.35424	0.35413	0.31697			