NBER WORKING PAPER SERIES

INSURGENCY AND SMALL WARS: ESTIMATION OF UNOBSERVED COALITION STRUCTURES

Francesco Trebbi
Eric Weese

Working Paper 21202
http://www.nber.org/papers/w21202

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2015

University of British Columbia, Vancouver School of Economics, Canadian Institute For Advanced Research and NBER, francesco.trebbi@ubc.ca; Yale University, eric.weese@yale.edu respectively.
The authors would like to thank Ethan Bueno de Mesquita, James Fearon, Camilo Garcia-Jimeno, Carlos Sanchez-Martinez, and seminar participants at Stanford, Berkeley, Rochester, and UQAM for useful comments and discussion and the researchers at the Princeton University Empirical Studies of Conflict Project for generously sharing their incident data online. Nathan Canen provided excellent research assistance. We are grateful to the Social Science and Humanities Research Council for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Insurgency and Small Wars: Estimation of Unobserved Coalition Structures
Francesco Trebbi and Eric Weese
NBER Working Paper No. 21202
May 2015
JEL No. O1,P48

ABSTRACT

Insurgency and guerrilla warfare impose enormous socioeconomic costs and often persist for decades. This paper studies the detection of unobserved coalitions of insurgent groups in conflict areas, and their main socioeconomic determinants. We present a novel methodology based on daily geocoded incident-level data on insurgent attacks, and provide an application in the context of the Afghan conflict during the 2004-2009 period. We show statistically that the Afghani Taliban are not an umbrella coalition, but rather a highly unified group, and that their span of control has grown substantially beyond ethnic Pashtun areas post-2007.

Francesco Trebbi
University of British Columbia
1873 East Mall
Vancouver, BC, V6T1Z1
Canada
and CIFAR
and also NBER
ftrebbi@mail.ubc.ca

Eric Weese
Yale University
Department of Economics
Box 208269
New Haven, CT 06520-8269
eric.weese@yale.edu

1 Introduction

Insurgency is typically defined as armed rebellion against a central authority.[1] It is one of the most opaque forms of armed conflict, as intertwining connections with the population blur the lines between combatants and civilians [Kilcullen, 2009]. The relative strength and even the identity of potential negotiating counterparts are unclear, and in the words of Fearon [2008] "*there are no clear front lines.*" In the post-World War II era, insurgency and guerrilla conflict have been enormously costly in socioeconomic terms, and they rank among the most detrimental and perduring forms of internal conflict and political violence [O'Neill, 1990].[2] Our paper offers a novel contribution to the empirical analysis of these asymmetric and irregular wars.

This paper's focus is the insurgency in Afghanistan. For the U.S. side alone, operations cost the lives of more than $1,800$ troops between 2001 and 2011, and led to more than $444 billion in military expenses.[3] Statistics for Afghan citizens are less certain, but the adverse effects are obvious. Soon into the operation, the U.S. military acknowledged through a drastic adjustment in tactics that the Afghan conflict differed substantially from previous large scale military operations.

Fighting against an alliance between Afghan Taliban insurgents and al-Qaeda, front lines were uncertain and the unity of the adversary doubtful. Experts disagreed about whether the Taliban were a unified fighting organization, or rather an umbrella coalition of heterogeneous forces. Some were skeptical regarding the degree of control Taliban leader Mullah Mohammed Omar exerted over the powerful Haqqani faction and the Dadullah network.[4] Similarly, the

[1] According to O'Neill [1990] "*Insurgency may be defined as a struggle between a nonruling group and the ruling authorities in which the nonruling group consciously uses political resources (e.g., organizational expertise, propaganda, and demonstrations) and violence to destroy, reformulate, or sustain the basis of one or more aspects of politics.*"

[2] For a recent and exhaustive review see Blattman and Miguel [2010].

[3] Table 1 provides a summary of the US Afghan counterinsurgency timeline.

[4] For example, the UN report [2013] stated that

> *Despite what passes for a zonal command structure across Afghanistan, the Taliban have shown themselves unwilling or unable to monopolize anti-State violence. The persistent presence and autonomy of the Haqqani Network and the manner in which other, non-Taliban, groupings like the Lashkar-e-Tayyiba are operating in Afghanistan raises questions about the true extent of the influence exerted by the Taliban leadership.*

Brahimi [2010] reports a statement by Ashraf Ghani, current Afghan president, in a lecture for the Miliband Programme at LSE indicating *"The Taliban are not a unified force - they are not the SPLA in Sudan or the*

Hizb-i Islami faction was considered by many a separate entity from the Taliban proper.[5] On the other hand, Dorronsoro [2009] offers the following in an insightful qualitative essay:[6]

> *The Taliban are often described as an umbrella movement comprising loosely connected groups that are essentially local and unorganized. On the contrary, this report's analysis of the structure and strategy of the insurgency reveals a resilient adversary, engaged in strategic planning and coordinated action.*

This disagreement is unsettling: understanding the extent of territorial control and population support of insurgent groups is essential for military operations. Furthermore, knowledge of the internal organization and cohesion of rebel groups can be used to prevent selective violence by insurgents, and ultimately helps with the reconstruction of areas affected by the conflict.[7]

This paper shows how data regarding simultaneous violent events can be used to estimate the number of different insurgent groups active and their territories, features that are typically unobservable to the econometrician/analyst. We make use of the fact that insurgent groups with the ability to launch joint attacks generally appear to do so. Our analysis thus relies on the conclusions of the existing literature regarding the incentives for organized groups to launch this sort of attack. Deloughery [2013] provides a recent review of this literature, and presents systematic evidence of the advantages of simultaneous attacks for terrorist organizations in terms of media coverage and appeal in the recruitment of new fighters – incentives that operate within insurgent organizations as well.[8] We take these

Maoists in Nepal" while Giustozzi [2009] states that "*The Taliban themselves are not fully united and the insurgency is not limited to the Taliban."*

[5]Fotini and Semple [2009] state explicitly that "*the Taliban is not a unified or monolithic movement*" and Thruelsen [2010] that "*the movement should not be seen as a unified hierarchical actor that can be dealt with as part of a generic approach covering the whole of Afghanistan.*"See also Giustozzi [2007].

[6]In contrast, the Pakistani Taliban are described in the same essay as an umbrella organization that is clearly non-unitary.

[7]One example of the importance of understanding insurgent group structures for post-conflict negotiations comes from Colombia. The recent appearance of the Bandas Criminales Emergentes (BACRIM) in lieu of the AUC paramilitary combatants has been a central issue in the work of Colombia's Reconciliation Commission in deploying resources and rebuilding state institutions and control at the local level.

[8]From a western perspective, the 9/11 attacks in the United States are the most obvious example of the salience of such simultaneous violence, but the phenomenon is widespread. For example, in southern Thailand insurgent movements have adopted similar tacticts: "On April 28, 2004 groups of militants gathered at mosques in Yala, Pattani, and Songkhla provinces before conducting simultaneous attacks on security checkpoints, police stations and army bases" [Fernandes, 2008]. The Indian Mujahideen, responsible for

incentives as given, and assume that an organization with the capability of launching such attacks will choose to do so. The simultaneous attacks that are observed can thus be used to analyze the underlying structure of the insurgent groups present.

The basic assumption in our model is that attacks on the same day in different locations are either the result of random chance, or represent an insurgent group with a presence in both locations. A simple structural model of attacks is used to distinguish between "random chance" and organized group behaviour. After estimating the number of different guerrilla groups and their territorial extent, we assess the main empirical determinants of insurgent presence, and also produce an analysis of shifts in insurgent presence over time.

The paper thus addresses three broad questions: two methodological and one empirical. First, when faced with multiple violent incidents in multiple regions, how can one decide whether the simultaneous incidents observed are isolated idiosyncratic events, as opposed to organized attacks by coalitions of assailants? Second, how can one identify from incident data alone how many distinct insurgent groups (if any) are attacking? Third, what are the socioeconomic determinants driving the diffusion and segmentation of the rebels within a specific region and across regions?

The estimation method consists of modelling a country experiencing an insurgency as a set of points at which violent incidents can occur in each period. For the application to Afghanistan, each point in this set represents the centroid of an administrative district, and each period is one day. Our main working assumption is that attacks on the same day in two different districts will occur with greater-than-random frequency if the same insurgent group is operating in both. Using a variety of assumptions regarding what the "reference" cross-district covariance in attacks would be in the case where there were no organized groups, we calculate which sets of districts are more correlated than would be expected. In general, these districts will be ones that have repeatedly experienced simultaneous attacks. We then use this information to estimate the set of districts in which each guerrilla group operates.

the 2008 Mumbai attacks, typically carry out simultaneous attacks [Subrahmanian et al., 2013]. Kurdish nationalists and the Tamil Tigers are known to have adopted simultaneous attacks as a strategy. In Africa, Boko Haram in northern Nigeria has carried out coordinated attacks on multiple targets such as churches, and Anderson [1974] describes coordinated attacks in Portuguese colonies. Simultaneous attacks and suicides have been a trademark of international jihadist organizations and of al-Qaeda in particular, making our approach particularly well-suited to the Afghan insurgency case.

We present estimators that allow for a single district to be contested by many guerrilla groups or by none. The estimators provide the number of guerrilla groups operating, the geographic area of each of these, and the intensity of each group's activity in each district. In order to provide these estimates, the data generating process used to model the occurrence of violent incidents is somewhat stylized. There is thus some loss of generality, but parsimony is necessary given the lack of detailed data on the internal organization and planning strategy of insurgents and counterinsurgency forces.

The main empirical results of the paper are as follows. We conclude that insurgent activity in Afghanistan is best represented by a single organized group, rather than several independent groups, and that the extent of this group is largely determined by ethnic boundaries. This result is robust to employing only within-month covariance in attacks (i.e. ignoring between-month variation), to constraining the analysis to districts with a number of incidents above specific thresholds, and to limiting the analysis to incidents explicitly claimed by the Taliban. As a verification of our methodology, we conduct an analysis of the Pakistani Taliban (Tehrik-i-Taliban Pakistan, or TTP), using data from Pakistan. The TTP are completely separate from their Afghan counterparts and are unanimously considered an umbrella coalition of diverse violent actors.[9] We show that in the case of the TTP, our methodology finds multiple groups, and is thus consistent with the qualitative literature.

We then consider changes in the extent of the Afghan Taliban between two time periods: 2004-2007 compared to 2008-2009. We find that insurgents spread largely to districts adjacent to those where they were already present: this is sometimes described as an "oil spot" strategy.[10] We also find that there has been penetration by the insurgents into areas traditionally occupied by non-Pashtun ethnic groups. We finish by discussing several case studies outside of Afghanistan where application of the methodologies we present could be helpful in assessing the economics of post-conflict reconstruction and power sharing with former insurgent groups from Latin America, Asia, and Africa.

An increasing amount of attention has been devoted within the fields of development

[9]Dorronsoro [2009] discusses how *"The Pakistani Taliban have different structures, different leaders, and a different social base* [relative to the Afghan Taliban -AN]. *They are, in fact, an umbrella movement comprising loosely connected groups."*

[10]See Krepinevich [2005] for the relevance of this approach in Iraq by U.S.-led coalition forces.

economics and political economy to the study of armed conflict within countries, in particular civil wars and insurgency. Economists have been interested in the analysis of violence and conflict at least as far back as Schelling [1960] and Tullock [1974], with Hirshleifer [1991, 1995a, 1995b, 2001] and Grossman [1991, 2002] offering more recent theoretical contributions. Outside of economics the interest is even greater: political scientists have dedicated to the study of conflict a large part of their work in the field of international relations.

Political science and economics have provided some of the most recent and novel insights in the study of insurgency.[11] As underlined by Blattman and Miguel [2010], a remarkable characteristic of this recent wave of research has been a strong empirical inclination and an increasing attention to micro-level (typically incident-level) information. The use of precisely geocoded micro data in this research is a departure from more established "macro" empirical approaches, which were based on country level information or aggregate conflict information.[12]

This paper is one in the new "micro" style, with a specific emphasis on the analysis of insurgency and small wars. We do not address conventional warfare. Currently, much less is known about non-conventional warfare (and its consequences on civilian populations) than is known about wars between regular armed forces. Economic and statistical evidence on the role of anti-government guerrilla activities is still sparse, even though such activities cause substantial damage worldwide and appear from a quantitative perspective to be the predominant form conflict in civil wars since 1945 [Fearon, 2008; Ghobarah et al., 2003]. Insurgents' strategies are generally not well understood, and neither are the subtleties of their interactions with the noncombatant population [Gutierrez-Sanin, 2008; Kilcullen, 2009]. A particular incentive for further study is that insurgent activity is also often linked to terrorist activities, and thus there is a connection with the growing literature on the economics of terrorism [Bueno de Mesquita and Dickson, 2007; Benmelech, Berrebi, Klor, 2012].

The remainder of this paper is organized as follows. Section 2 develops our methodology for the estimation of coalition structures among insurgent groups. We describe our data in

[11] These include Berman [2009], Berman et al. [2011], Condra et al. [2010], Blair et al. [2012], Condra and Shapiro [2012], Cullen and Wedmnan [2013], and Bueno de Mesquita [2013].

[12] Notable instances of the "macro" approach include Fearon and Laitin [2003], Boix [2008], Collier and Hoeffler [2004], and Collier and Rohner [2008], among many others.

Section 3, particularly the WITS incident-level data for Afghanistan and Pakistan, which have been generously made publicly available by the Empirical Studies Of Conflict (ESOC) project. The analysis of the determinants of insurgent group presence is developed in Section 4. Section 5 presents several case studies focused on the economic importance of understanding insurgent organization in conflict and post-conflict environments in developing countries. Section 6 concludes.

2 Econometric Setup

The objective is to determine whether insurgent activity in Afghanistan features only a single organized group, or several, and what the extent of these groups are. To allow for the possibility that there are no organized groups present in a given location even though attacks occur, the model will include the possibility of random attacks from unorganized local actors. The number of organized groups that best matches the observed data can then be estimated.

Let locations be indexed by i, and let there be a total of N locations at which attacks occur. For the application to the Afghan data, locations will be taken to be administrative districts. Violent occurrences in i can be of two types: random or organized by insurgent groups. That is, we suppose that observed attacks may be initiated either by unorganized local militants or by members of an organized group which operates on that territory. Let ℓ_i be the number of unorganized local militants in district i. Let organized insurgent groups be indexed by j, and let J be the total number of such organized groups. Let α_{ij} be the number of members in district i belonging to organized group j.

Time is discrete and indexed by t. In the Afghan data, the time periods used will be days. This relatively high frequency attack data is useful because it reduces the number of attacks that will be simultaneous simply by random chance.

In each time period, the probability that a unorganized local militant launches an attack is η, which does not change across time. The decision by unorganized militants to attack is independent of the decision of anyone else (unorganized militant or group member). The expected number of attacks by local militants in district i at time t is thus $\eta\ell_i$, and the variance within district i is $\eta(1-\eta)\ell_i$. The covariance in these attacks between two districts

i and i' is zero: the attack decisions are made independently, and the probability of an attack is constant (this assumption is relaxed in what follows).

In contrast to unorganized militants, members of an organized group are more likely to attack on some particular days than on others. Let ϵ_{jt} be the probability that a member of group j will attack at time t. This probability is the same for all members of group j, and whether any given member attacks is independent of other attack decisions after conditioning on the attack probability ϵ_{jt}. Across time, the covariance of attacks between two members of the same group is thus $\text{Var}(\epsilon_j)$. We assume that this variance is constant across groups, and will refer to it as σ^2. Assume that for any other group j', ϵ_{jt} is uncorrelated with $\epsilon_{j't}$. Thus, the covariance of attacks between two members of different groups is zero.

Consider the members of group j. If there are α_{ij} members in district i and $\alpha_{i'j}$ members in district i', then the covariance in attacks over time between these two districts, due to the presence of members of group j, is $\alpha_{ij}\alpha_{i'j}\sigma^2$. Summing over members of all groups, the covariance in attacks between districts i and i' will be $\sum_j \alpha_{ij}\alpha_{i'j}\sigma^2$. Now consider the variance-covariance matrix Γ for attacks, where the entry in row i and column i' gives the covariance in attacks across time for these two districts. This matrix can be decomposed as $\Gamma = \Gamma_D + \Gamma_H$, where Γ_D is a diagonal matrix and Γ_H is a "hollow" matrix (main diagonal zero) with the form

$$\Gamma_H = \sigma^2 \begin{bmatrix} 0 & \sum_j \alpha_{1j}\alpha_{2j} & & \\ \sum_j \alpha_{2j}\alpha_{1j} & 0 & & \\ \dots & & & \\ \sum_j \alpha_{ij}\alpha_{1j} & & \dots & \sum_j \alpha_{ij}\alpha_{i'j} \\ \dots & & & \end{bmatrix} \tag{1}$$

This decomposition is considered because the diagonal entries of the covariance matrix do not provide useful information regarding the group membership of districts: diagonal entries are a sum of variance from unorganized militants and variance from organized groups, and there is no intuitive way to distinguish between these two components.[13] Thus, for estimating

[13]The diagonal entries of Γ do not in general have a useful form. For example, even in the very simple case where there is only one group and ϵ_1 is uniformly distributed on $[0, b]$, the ith diagonal entry would be a non-trivial nonlinear expression $\frac{b^2}{12}(\frac{6}{b} + (\alpha_{i1} - 4))\alpha_{i1} + \ell_i\eta(1-\eta)$. The covariance matrix is thus used throughout rather than the correlation matrix. A simpler form for the diagonal entries could be obtained by using a mixture Poisson approximation, such as the Poisson-Gamma used in Ashford and Hunt [1974].

group structure, only the off-diagonal entries of the covariance matrix will be used. As a normalization, we set $\sigma^2 = 1$.

Let $\gamma_{ii'} = \sum_j \alpha_{ij}\alpha_{i'j}$ denote the off-diagonal entry on row i and column i' of Γ_H. Let $\bar{\gamma}_{ii'}$ be the corresponding entry of the covariance matrix in the observed sample. Estimation will be based on $\bar{\Gamma}_H$, the sample covariance matrix ignoring the diagonal entries.

The model just presented is clearly a stylized model of the attack behavior of insurgent groups, and the covariance structure imposed is not without loss of generality. A particularly strong assumption made in the model is that the members of an insurgent group do not move between districts: a given group j has a certain membership α_{ij} in district i, and those members will either be encouraged to attack in a given period (a high ϵ_{jt}), or not (low ϵ_{jt}). A very different model would be one in which members of an insurgent group are mobile, and in any given period have the choice of attacking in one of many districts. This latter model implies that organized groups should lead to negative covariances $\gamma_{ii'}$, as insurgent group members who attack in district i could not also be attacking in district i' in the same period. In contrast, the model presented above suggests $\gamma_{ii'}$ should be positive if the same insurgent group j has members in both i and i', as attacks in both i and i' will be higher in periods when ϵ_{jt} is high and lower in periods when ϵ_{jt} is low. In the case of the Afghan data, the observed covariances γ are systematically positive.[14] The qualitative research of Deloughery [2013] and others, as discussed in Section 1, also suggests that a model without substantial substitution in attacks across districts appears most appropriate.

The desired estimates are $\hat{J}$, giving the total number of organized insurgent groups, and an $\hat{\alpha}_{ij}$ for each district i and group j, giving the number of insurgent members of that group operating in that district. The set of estimates $\{\hat{\alpha}_{ij}\}$ thus has $N\hat{J}$ elements. J is an integer, and estimation strategies for this sort of parameter do not typically yield confidence intervals of the sort that would be obtained for a continuous parameter. While the setup described above does not appear to correspond exactly to any discussed previously in the literature, it is close to problems addressed by spectral clustering and non-negative matrix factorization,

A variety of these distributions are discussed in Karlis and Xekalaki [2005]. None of the options available, however, appear to simplify the diagonal entries enough to be useful from an empirical perspective.

[14] Permutation tests of the sort discussed later indicate that the mean covariance is positive at any reasonable confidence level.

and these two approaches can be used to produce estimates $\hat{J}$ and $\{\hat{\alpha}_{ij}\}$.

An approach based on spectral clustering will be presented first, while non-negative matrix factorization techniques will be considered in Section 2.2. These two approaches are largely complementary and rely on different assumptions. This provides a form of cross-validation for our results, which is important given the novelty of these methodologies within the field of political economy. It is difficult to determine the properties of the estimator based on spectral clustering, and the statistical tests available appear to have low power. The approach based on non-negative matrix factorization, however, includes permutation tests that occur as a natural part of the estimation procedure. Unfortunately, the threshold approach used produces only point estimates and not accompanying confidence intervals. Confidence intervals for $\{\hat{\alpha}_{ij}\}$ and $\hat{J}$ are thus not reported below. We are, however, able to reject important null hypotheses at reasonable levels of significance.

2.1 A Spectral Clustering Approach

In graph theory, spectral clustering is a technique used to partition nodes of a graph into clusters. A full review of the methodology and some of its applications in statistics and computer science is available in Luxburg [2007]. Traditional clustering algorithms such as k-means are known to perform poorly when used directly on a highly dimensional matrix such as Γ_H. Spectral clustering, on the other hand, is well suited for this sort of data because dimensionality reduction occurs naturally as a part of the algorithm.

Estimation via spectral clustering requires an additional assumption different from those needed for the technique of Section 2.2: specifically, for our spectral clustering estimator it is necessary to assume that the various insurgent groups present do not have overlapping territories. That is, there is at most one organized group j present in any given district i. Based on this assumption, reordering the districts i allows Γ_H to be written as a block-diagonal matrix:

$$\Gamma_H = \begin{bmatrix} \Gamma_H^1 & 0 & & \\ 0 & \Gamma_H^j & & \\ \dots & & & \\ 0 & & \dots & \Gamma_H^J \end{bmatrix} \tag{2}$$

where there are a total of J organized groups, and each block Γ_H^j has the form given in Equation 1. Here Γ_H corresponds to the adjacency matrix for a weighted undirected graph, with the weights given by the off-diagonal matrix entries $\gamma_{ii'}$ being determined by the degree of group presence in each district.

To perform spectral clustering, a technique following Shi and Malik [2000] will be used.[15] This technique is based on a "graph Laplacian" matrix, which is constructed from the adjacency matrix Γ_H: the graph Laplacian has off-diagonal entries equal to the negative of those of the adjacency matrix, and diagonal entries such that all rows and columns sum to zero. The approach is based on examining the eigenvalues of the graph Laplacian. The number of zero eigenvalues of the graph Laplacian matrix will correspond to the number of connected components of the weighted undirected graph described by the adjacency matrix Γ_H. This is J, the number of blocks of Γ_H.

The intuition for this result is relatively straightforward. Setting the diagonal entries so that rows and columns sum to zero ensures that the rows (and columns) of the graph Laplacian corresponding to each Γ_H^j block are linearly dependent. Γ_H^j is full rank: each row of Γ_H^j is a vector of positive numbers with a single 0 on the diagonal. The transformation to the graph Laplacian L reduces the rank of each block by one. Thus, the reduction in rank of the overall graph Laplacian, relative to the initial Γ_H will be equal to the number of blocks of L, which is also the number of blocks of Γ_H and the number of organized groups J. This is equivalent to the number of zero eigenvalues because this is the dimension of the null-space of L.

Let D be a diagonal matrix with entries such that the rows of $L = D - \Gamma_H$ sum to zero. If L were known, the the number of organized groups would be equal to the number of zero eigenvalues of L. However, the data available gives the sample covariances $\bar{\gamma}_{ii'}$ rather than the true $\gamma_{ii'}$, and thus $\bar{\Gamma}_H$ is observed instead of Γ_H. A simple modification of Shi and Malik [2000] is thus used: use $\bar{\Gamma}_H$ to construct $\bar{L}$, and then count the "zero" eigenvalues of $\bar{L}$. Clustering is thus feasible, because it is based on statistics from the observed sample, and the estimator using $\bar{L}$ is a consistent estimator for the clusters that would be obtained using

[15] Luxburg [2007] provides a summary of this method.

the true graph Laplacian L. Further details are provided in Appendix A.[16]

In a finite sample, the eigenvalues calculated from $\bar{L}$ are subject to finite sample variation. In particular, random variation will result in positive $\bar{\gamma}_{ii'}$ entries in some cases where the true $\gamma_{ii'}$ is zero, and negative $\bar{\gamma}_{ii'}$ entries in some cases where the true $\gamma_{ii'}$ is positive. This random variation will tend to increase the rank of the $\bar{L}$ relative to L. This problem is particularly severe for districts i for which there are few attacks: the data provides little information on the group structure in these districts, and if one object of interest is J, the total number of groups, the inclusion of these particularly noisy districts could result in a substantial amount of additional noise in the estimate $\hat{J}$.

A first step to dealing with this problem is to exclude districts with very few attacks from estimation: thus, for the analysis of the Afghan data, the spectral clustering approach will use data only for those districts in which there were 3 or more attacks.[17] This approach does not fully solve the underlying issue, however. Eigenvalues that would be zero asymptotically will not be zero in a finite sample, because some of the entries that are zero in Γ_H will be positive in the observed $\bar{\Gamma}_H$. When using a covariance matrix that includes this finite sample variation, it is thus necessary to account for the fact that eigenvalues that are zero in the population may not be zero in the sample.

The literature on spectral clustering provides a variety of methods to determine how reliable an estimate can be obtained by examining eigenvalues. We check the reliability of our estimated $\hat{J}$ by considering "eigengaps" similar to those used by Ng, Jordan, and Weiss [2002]. This method is based on matrix perturbation theory, and was originally intended for the case where the true Laplacian L was observed directly. When used with a noisy matrix,

[16] An estimate of the number of organized groups present, $\hat{J}$, can also be obtained based only on which matrix entries are zero and which are non-zero. In this case, the sample covariance matrix used would be

$$\tilde{\Gamma}_H = \begin{bmatrix} \tilde{\Gamma}_H^1 & 0 & & \\ 0 & \tilde{\Gamma}_H^j & & \\ ... & & & \\ 0 & & ... & \tilde{\Gamma}_H^J \end{bmatrix} \tag{3}$$

where each $\tilde{\Gamma}_H^j$ matrix has zeros on the diagonal, and ones in all off-diagonal entries. $\tilde{\Gamma}_H$ thus has the form of an adjacency matrix for an undirected graph: districts correspond to the nodes of this graph, and there is an edge present between districts i and i' if the same organized group is active in both districts. One advantage of the this binary classification is that it emphasizes the relationship between spectral clustering and graph theory.

[17] Other cutoffs yield similar results.

the method still provides an heuristic indication of the reliability of the estimated $\hat{J}$.

Begin by sorting the eigenvalues λ of L in increasing order, such that λ_1 is the smallest and λ_N the largest.[18] The difference $\lambda_{k+1} - \lambda_k$ is defined the kth eigengap. Ng, Jordan, and Weiss [2002] argue that a large eigengap indicates that perturbation of the eigenvectors of L would not change the clusters produced by spectral clustering. Luxburg [2007] thus suggests that the right choice for $\hat{J}$ is a number such that λ_k is "small" for $k \leq \hat{J}$, and the $\hat{J}$th eigengap is large. The intuition here is that if there truly are $\hat{J}$ eigenvalues that are zero, then these appear to be non-zero in the finite sample only due to random variation. In contrast, the $\hat{J}+1$th and larger eigenvalues would be strictly positive even if the true L were used. An examination of the $\hat{J}$th eigengap thus provides a heuristic test of whether the choice of $\hat{J}$ was reliable, or whether small changes due to random variation might result in a different number of zero eigenvalues. The underlying difficulty here is determining what exactly constitutes a "zero" eigenvalue, when there is finite sample variation. A large eigengap provides some confirmation that an appropriate definition of "zero" has been chosen.

After calculating an estimate $\hat{J}$ for the number of organized groups, and checking via the eigengap approach whether this estimate appears to be reliable, a remaining problem is to determine which insurgent group is present in which district. This problem is quite close to a classical k-means problem, where k is now known.[19] There are thus numerous possibilities for determining which insurgent group is active in a given district, including approaches based on eigenvectors, such as are mentioned in Ng, Jordan, and Weiss [2002]. The empirical results below will show that $\hat{J} = 1$, and we thus do not discuss further how to deal with the case where $\hat{J} > 1$, other than noting that many standard methods are available.

While it is relatively straightforward to obtain a consistent estimate for J, the total number of organized insurgent groups, a consistent estimate for $\{\alpha_{ij}\}$ is more challenging. The main difficulty here is that the spectral clustering literature generally assumes that the true graph Laplacian $L = D - \Gamma_H$ is observed, whereas the data provides only $\bar{L}$, a graph

[18] Here we continue to consider only districts that have a certain minimum number of attacks, but for simplicity the notation assumes that no districts are excluded on this basis and thus there are still N districts, and N eigenvalues.

[19] For a standard reference here, see Hastie, Tibshirani, and Friedman [2009].

Laplacian that includes noise due to random variation in the attack data. However, in the case where the number of districts, N, is large, there is a computationally trivial approximate estimator for α_{ij}.

Specifically, suppose that each organized group that is present has members in a large number of districts, and that no single district has a particularly large α_{ij}. Let I_j be the set of districts that have members of organized group J. Then, since an assumption of the spectral clustering model was that the organized groups do not overlap, an estimate of α_{ij} for $i \in I_j$ can be produced via the following approximation, using $\bar{\Gamma}_H^j$, the relevant block of $\bar{\Gamma}_H$. Specifically, note that a sum across the row of Γ_H^j corresponding to district i is $\sum_{i' \neq i} \alpha_{ij}\alpha_{i'j}$. If there are a large number of districts with members of j, then it is reasonable to use the approximation

$$\begin{aligned} \sum_{i' \neq i} \alpha_{ij}\alpha_{i'j} &\simeq \sum_{i'} \alpha_{ij}\alpha_{i'j} \\ &= \alpha_{ij} \sum_{i'} \alpha_{i'j} \\ &= \alpha_{ij} a_j \end{aligned} \tag{4}$$

where $a_j = \sum_{i'} \alpha_{i'j}$ is the same for any choice of district i within I_j. The sums of the rows of each block Γ_H^j thus give the relative prevalence of organized group members in each district in I_j. This approximation is particularly interesting in the case where there is only one group: in this case the sums of the rows of Γ_H give the relative of prevalence of group members across districts in the whole country. This approximate estimator becomes increasingly correct as the number of districts that each organized group has members in grows. While it would be possible to use non-linear programming or other techniques to develop an estimator with more desirable properties, the approximate estimator has at least two advantages. First, the estimator has an intuitive interpretation: Γ_H is a covariance matrix, and the sum across the off-diagonal entries of a row of Γ_H thus gives an indication (in a heuristic sense) of how closely linked attacks in a given district are with attacks in other districts. Second, if in the data a given district i experiences only a small number of attacks, then the off-diagonal entries $\bar{\gamma}_{ii'}$ will be relatively small for that district, and thus i will not introduce substantial noise into estimates $\hat{\alpha}_{i'j}$ for other districts i'. Developing an unbiased estimator that also

possesses such properties appears to be a non-trivial undertaking.

To summarize, in spectral clustering the specific estimator used is the following. A graph Laplacian $\bar{L}$ is calculated based on the observed attack covariance matrix $\bar{\Gamma}_H$. The eigenvalues of $\bar{L}$ are examined, considering only those districts that have a certain minimum number of attacks (three in the Afghan data actually used). The estimate $\hat{J}$ for the number of organized groups is equal to the number of $\bar{L}$'s eigenvalues that are "zero". Eigengaps are then examined to determine how reliable this estimate appears to be.

Two potential problems with this approach based on spectral clustering can be addressed using an alternate technique. First, the actual group structure may be overlapping, with multiple groups present in a single district. Second, hypothesis tests are difficult to perform: the distribution of eigenvalues resulting from random variation in finite samples is not obvious, and the existing literature mostly assumes that the observations to be clustered are observed without noise. These issues can be addressed using an approach based on non-negative matrix factorization.

2.2 A Non-Negative Matrix Factorization Approach

Begin by supposing that the number of organized groups J is known, and consider an estimator that chooses $\hat{\alpha}_{ij}$ for each district i and group j to satisfy, to the extent possible, the set of restrictions

$$\bar{\gamma}_{ii'} = \sum_j \hat{\alpha}_{ij}\hat{\alpha}_{i'j}$$

If there are N districts, there are $N(N-1)/2$ restrictions: one for each off-diagonal element in one half of the symmetric covariance matrix. If there are J groups, there are $N \times J$ parameters to be estimated: one $\hat{\alpha}_{ij}$ for each district i and group j.[20] A necessary condition for identification is thus that $(N-1)/2 \geq J$.[21] In the data used the number of districts is large relative to plausible numbers of groups, and thus this inequality holds strictly and a

[20] Ignoring the diagonal entries of $\bar{\Gamma}$ means that the non-negative matrix factorization problem considered in this paper is not the same as that considered in Ding, He, and Simon [2005], where the authors show an equivalence between NNMF and spectral clustering.

[21] The identities of the groups are never identified: the predicted elements of the covariance matrix are identical if $\hat{\alpha}_{ij}$ and $\hat{\alpha}_{ij'}$ are interchanged for all districts. However, labeling groups becomes possible employing very basic additional information. For instance, our group 1 is obviously the Taliban and any activity in the Uzbek areas could be possibly associated with the Islamic Movement of Uzbekistan insurgent faction.

penalty function is required. The estimator used for $\{\alpha_{ij}\}$ will be:

$$\underset{\hat{\alpha}_{ij}\geq 0}{\operatorname{argmin}} ||\bar{\Gamma}_H - \hat{\Gamma}_H||^2$$

where the off-diagonal entry of $\hat{\Gamma}_H$ in row i and column i' is $\sum_j \hat{\alpha}_{ij}\hat{\alpha}_{i'j}$, and the diagonal entries are all zero.

From a numerical perspective, the easiest norm to use is the element-wise norm. With this norm, the estimator can also be expressed as

$$\underset{\hat{\alpha}_{ij}\geq 0}{\operatorname{argmin}} \sum_i \sum_{i'\neq i} \left(\bar{\gamma}_{ii'} - \sum_j \hat{\alpha}_{ij}\hat{\alpha}_{i'j} \right)^2 \tag{5}$$

The major difficulty with implementing this estimator is that N is large. Thus, even when considering only a small number of groups J, the number of parameters that must be estimated is large. Recent optimization algorithms such as Birgin, Martinez, and Raudan [2000] appear to be computationally feasible so long as there are only about one thousand variables.[22] Thus, with $N \simeq 250$, a direct approach based on method of moments is feasible so long as $J \leq 5$. This will turn out to be the case in the data used, and would also likely be the case for many other data sets of interest.

The above assumed that J was known, but this is of course not the case. A heuristic technique from the clustering literature will again be applied to deal with this problem. Tibshirani, Walther, and Hastie [2001] propose the "gap statistic" as a means of determining the number of clusters to use with a clustering algorithm. Following Mohajer, Englmeier, and Schmid [2010], this can be expressed as

$$\text{Gap}(k) = E^*[W_k] - W_k \tag{6}$$

Here W_k is the variation that is not explained by the k clusters: for this paper, this is taken to be the squared residuals in Equation 5. E^* is the expectation taken with respect to a "reference distribution" chosen to correspond to no cluster structure. This distribution

[22] A very different approach would be to attempt to use the fact that the set of completely positive matrices is convex. Unfortunately, there is no barrier function available for optimization over this set. Vasiloglou, Gray, and Anderson [2009] present some options for various relaxation-based approaches. The "brute force" approach used in this paper, however, appears to yield much better results for the type of data considered: relaxations would presumably perform better if the data were of much higher dimension.

is generated via Monte Carlo permutations of the actually observed attacks, and different choices of the sort of permutation used allow for robustness checks with respect to some of the assumptions made in the structural model. Good [2005] provides an accessible introduction to permutation tests.

First, suppose that the structural model presented above is correct. In this case, the distribution of the number of attacks by disorganized militants in district i is the same for all periods, with expected value $\eta\ell_i$. Thus, under the null hypothesis that there is no group structure, the observed attack data is weakly exchangeable: within a given district, permuting the time indices does not change the joint distribution of the attacks.[23] The total number of such permutations is huge, and thus rather than perform calculations using the entire set we consider only a random subset of these permutations. By construction, the permutated data exhibits no group structure: all the off-diagonal entries of the sample covariance matrix will be zero asymptotically. To construct the desired reference distribution, we treat each of these permutations as if it were the observed data, and estimate a group structure for each of $J = \{1, 2, 3, 4, 5\}$. We then calculate the residual variation not explained by this estimated group structure. The average of this residual variation gives $E^*[W_k]$ for $k = \{1, 2, 3, 4, 5\}$. This is the amount of residual variation we would expect to result from our estimator, if the null hypothesis were true and there was no group structure.

Now, suppose that the structural model assumed is not exactly correct, and there is some cross-time variation in the expected number of attacks by disorganized militants within a district. Specifically, suppose that the probability that a disorganized militant launches an attack is not a constant η, but rather varies across months. The expected number of attacks on a given day in month m is then $\eta_{im}\ell_i$, and will differ by month. In this case, the observed attack data is still weakly exchangeable, but only within a given district *and* a given month. We can thus still construct a reference distribution, provided that observations are permuted only within each month for each district. In this case, the covariance matrices may not have all off-diagonal entries zero asymptotically: it could be that η_{im} and $\eta_{i'm}$ are positively correlated, for example. In this case, the gap statistic in Equation 6 will be positive if a

[23]The intuition here can be provided by an example. Suppose there are three periods. If there is no group structure, then the probability of observing $\{x_1, x_2, x_3\}$ in a given district must be equal to the probability of observing $\{x_1, x_3, x_2\}$, because the number of attacks is i.i.d. across time within a given district.

group structure with k groups, when applied to the actual data, leaves less residual variance W_k than would be expected if the data were the outcome of disorganized militants attacking with different probabilities in different months.

Finally, suppose that the expected number of attacks by disorganized militants varies at the daily level, rather than the monthly level. The general case, with $\eta_{it}\ell_i$ attacks expected in district i at time t, is so general that it does not appear to allow for any permutations. However, suppose that the number of expected attacks is instead $\eta_t\ell_i$, where η_t now does not differ across districts.[24] This might be the case, for example, if there were particular days that, for whatever reason, generated large amounts of random violence. In this case, observations are "approximately" weakly exchangeable via the following sort of permutation, inspired by Good [2002]. Find a pair of districts i and i', and a pair of times t and t', such that the following two conditions hold: there were the same number of attacks x in district i at time t and in district i' at time t', and there were the same number of attacks x' in district i at time t' and in district i' at time t. Permute the data by swapping x and x' in these four entries.[25] These permutations are attractive from an intuitive perspective, as they retain not only the same number of total attacks in each district, but also the same number of total attacks on each day. In the Afghan data, there are relatively few attacks on any given day and thus an enormous number of possible permutations of this sort. A random sample of these permutations is used. The gap statistic in this case describes the degree to which adding organized groups better explains the observed covariance matrix, compared to random attack covariance matrices of the sort that would be generated by per-day variation in random violence.

[24] This gives the disorganized militants the same structure an additional organized group. The test against the null hypothesis in this case is thus related to whether there is an organized group present that is active in some districts but not others. Under the null hypothesis, the off-diagonal entries of the sample covariance matrix should be directly proportional to the total number of attacks in the districts in question.

[25] To see why this weak exchangeability holds "approximately", note that the distribution of attacks is binomial. Approximate the binomial with a Poisson distribution with expectation $\eta_t\ell_i$. Then for observations of the type just described

$$\begin{aligned}\Pr(x|\eta_t\ell_i)\Pr(x'|\eta_{t'}\ell_i)\Pr(x'|\eta_t\ell_{i'})\Pr(x|\eta_{t'}\ell_{i'}) &= \frac{(\eta_t\ell_i)^x}{x!}e^{-\eta_t\ell_i}\frac{(\eta_{t'}\ell_i)^{x'}}{x'!}e^{-\eta_{t'}\ell_i}\frac{(\eta_t\ell_{i'})^{x'}}{x'!}e^{-\eta_t\ell_{i'}}\frac{(\eta_{t'}\ell_{i'})^x}{x!}e^{-\eta_{t'}\ell_{i'}} \\ &= \Pr(x'|\eta_t\ell_i)\Pr(x|\eta_{t'}\ell_i)\Pr(x|\eta_t\ell_{i'})\Pr(x'|\eta_{t'}\ell_{i'})\end{aligned}$$

by rearranging terms. The canonical reference for multivariate permutations appears to be Pesarin [2001], although this specific type of permutation is not described.

In Equation 6, Gap(k) quantifies an intuitive definition of the fit of a $k-$cluster structure to the observed data: the fit is "good" only to the extent that it is *better* than the fit to randomly generated data with no group structure at all. Following Tibshirani, Walther, and Hastie [2001], the estimated number of clusters $\hat{J}$ is selected to be the smallest k such that:

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1} \tag{7}$$

where s_{k+1} is the estimated standard error for the objective function, obtained by randomly drawing a large number of covariance matrices from the reference distribution, and then calculating W_{k+1} for each of these matrices. The intuition for this technique is that adding an additional cluster will always improve the fit to the observed covariance matrix, and thus an appropriate estimate $\hat{J}$ must balance this against the risk of overfitting the data. This is done by comparing the improvement in fit in the actual data to the case with randomly generated data that is known not to have any group structure. Consistency of this estimator is discussed in Appendix B.

2.3 Robustness: potentially changing district environments

Both the spectral clustering approach and the non-negative matrix factorization approach just described assume that the covariance in attacks by group members across districts remains the same even across long periods of time. In the observed data, however, it could be the case that in earlier years certain districts are the focus of many attacks, while in later years activity shifts to other districts. These sorts of long term changes can be accounted for by considering only the covariance in attacks across districts within shorter time windows.

Let $\bar{\Gamma}_{Hm}$ be calculated the same as $\bar{\Gamma}_H$ from Equation 1, but using only daily attack data from month m. As the number of days of data used to calculate $\bar{\Gamma}_{Hm}$ does not increase asymptotically for any given month m, estimation based on a single $\bar{\Gamma}_{Hm}$ would be inconsistent. Aggregating across months, however, results in a consistent estimator that is robust to changes in attack probabilities between districts at the month level.

Specifically, assume that the probability of an attack in district i in month m, either from unorganized militants or an organized group, now changes with a parameter ζ_{im}. That is, the probability of an attack from a unorganized militant is now $\zeta_{im}\eta$, and the probability

of an attack from member of organized group j is now $\zeta_{im}\epsilon_{jt}$. Let $D(\cdot)$ indicate a diagonal matrix with the given entries on the diagonal. If ζ were known, the standardized matrix $\tilde{\Gamma}_{Hm} = D(\frac{1}{\zeta_m})\Gamma_{Hm}D(\frac{1}{\zeta_m})$ could be summed to create $\tilde{\Gamma}_H = D(\sum_m \zeta_m)\tilde{\Gamma}_{Hm}D(\sum_m \zeta_m)$. $\tilde{\Gamma}_H$ could then be used to estimate $\{\alpha_{ij}\}$. In reality, ζ is unobserved; however, dividing by the observed number of attacks creates a feasible estimator, with α identified up to scale. This approach can be used with both estimation based on spectral clustering and that based on non-negative matrix factorization.

3 Data

Afghanistan is covered by the Empirical Studies Of Conflict project (ESOC) at Princeton University, which "*identifies, compiles, and analyzes micro-level conflict data and information on insurgency, civil war, and other sources of politically motivated violence worldwide.*"[26] The ESOC data currently reports a location, date, and type for violent incidents from the beginning of 2003 to the end of 2009. This data is based on the Worldwide Incidents Tracking System (WITS), a declassified U.S. government military database.[27] The following two examples illustrate the typical form of incident descriptions:

> "*On 27 March 2005, in Laghman, Afghanistan, assailants fired rockets at the Governor House, killing four Afghan soldiers and causing minor damage. The Taliban claimed responsibility for the attack.*"
>
> "*On 19 February 2006, in Nangarhar, Afghanistan, a suicide bomber detonated an improvised explosive device (IED) prematurely near a road used by government and military personnel, causing no injuries or damage. No group claimed responsibility.*"

The violent incidents cataloged in the ESOC data are episodes of violence initiated by insurgents, or acts of random violence. The data does not include violence directly connected to military counterinsurgency operations, such as for instance a U.S. military attack on a Taliban safe house or the bombing of a fortified compound.

[26] See https://esoc.princeton.edu/

[27] "Worldwide Incidents Tracking System." National Counterterrorism Center (wits.nctc.gov).

According to the data, there are some days where as many as 64 different districts are affected by simultaneous insurgent attacks. However, there are also 123 districts with no reported incidents over the entire 2004-2009 time period. It is apparent to even the most casual observer that attacks are concentrated in certain areas of the country.

The location reported for an attack in ESOC is given as latitude and longitude coordinates. This would seem to suggest that attacks could be analyzed as some sort of spatial point process. Closer inspection, however, reveals that the latitude and longitude coordinates reported are not those of the actual location of the attack, but rather the coordinates of a prominent nearby geographic feature. Sometimes this is a city or village, but for the vast majority of incidents the location given is that of the centroid of the district in which the incident occurred. In Afghanistan, the "district" is the lowest-level political unit and the unit of geographic location in our model. We also note that a few districts have been split in recent years: this paper uses 2005 administrative boundaries, which specify 398 districts. The ESOC data effectively provides panel data at the district-day level, with $N = 398$ and $T = 2082$. WITS data are also collected for Pakistan and available from ESOC. We make use of this additional data for falsification exercises.

Additional geographic information reported in ESOC includes the location of roads, rivers, and settlements. We aggregate this data to the district level in order to use it jointly with the district-level attack data. ESOC does not report information on the distribution of ethnicity in Afghanistan. For geographic data on ethnicities, we thus use the Soviet Atlas Narodov Mira data. The version used is the "Geo-referencing of ethnic groups" (GREG) data set of the Swiss Federal Institute of Technology Zurich.[28]

In Figure 1 we report the ethnic distribution map by district, and Figure 2 shows the main Afghan highway. Figure 3 gives the attacks observed in the data, aggregated by district. Without further analysis, it is clear that the data confirm two well known qualitative features regarding insurgent attacks: they are more likely to occur in Pashtun areas, and there is a particular concentration on the ring road highway running south from the capital, Kabul. An analysis by the methods developed above, however, reveals some additional patterns that are not immediately obvious from an inspection of the raw data.

[28] http://www.icr.ethz.ch/data/other/greg

As we discuss further below, pooling WITS attacks over the entire 2004-2009 period masks some changes in the distribution of violent activity in the country over time. Evidence of the deterioration of the security environment is reported in Figures 4 and 5, which show the distribution of incidents by district for the years 2004-2007 and 2008-2009 respectively, in per capita terms. To provide some context for the reader in interpreting the maps, Table 1 provides a summary of the US Afghan counterinsurgency timeline produced by the Council of Foreign Relations. Table 2 includes summary statistics for total incidents, ethnic fragmentation, roads, rivers, and settlements by district.

4 Results

Figure 6 shows the eigenvalues obtained by using the spectral clustering approach described in Section 2.1 on the Afghan data. There is only one zero eigenvalue, with the following eigenvalues being substantially larger. Thus, the appropriate estimate for the number of organized insurgent groups is $\hat{J} = 1$. Figure 7 shows the eigengaps for these eigenvalues. The first eigengap is the largest by a substantial margin, suggesting that small random perturbations would not likely change the estimated number of groups. Figure 8 shows the presence of organized group members based on the approximation given in Equation 4. A disadvantage of the estimation strategies used in this paper is that they only provide information about the relative prevalence of each organized group across districts. The units reported in Figure 8 thus do not have an interpretation in levels: 0 corresponds to no attacks being attributable to organized group members, but the numeric scale of the legend is arbitrary, and it is not possible to interpret the results in terms of "fraction of attacks due to organized groups" without additional assumptions.[29]

The approximation in Equation 4 reveals the latent geographic distribution of the Taliban and, with this, we can investigate the geographic spread of the insurgency. Table 3 shows regression results based on this approximation including a set of ethnic and geographic controls, as well as province fixed effects. Most of the estimated coefficients, which should be read as correlates of Taliban control over each specific district, are not surprising. Ethnicities

[29]The numbers reported in the legend are the number of attacks per million people the organized group would have been responsible for if $\sigma^2 = 1$, but this choice is arbitrary.

other than Pashtun (the omitted ethnicity) are substantially less likely to be associated with organized group activity.[30] These include the Hazara, a Shia group hostile to the (Sunni) Taliban, as well as the Tajik and Uzbek communities. These latter groups have historically found themselves in conflict with the Taliban, and were participants in the Northern Alliance.

With respect to geographic characteristics of districts, there is more group activity in districts with more roads, particularly the ring road artery connecting Kabul with other provincial capitals. As in Figure 3, which shows raw total attacks, Figure 8 shows graphically that organized attacks are concentrated in Pashtun-majority areas, and also near the main highway passing through Kabul and other cities. A feature that is apparent in Figure 8, however, that does not show up clearly in the raw attack data of Figure 3 is that there appears to be a substantial organized insurgency operating near the highway north of Kabul, as well as the highway running south from it. This area is not as heavily populated by Pashtuns, and perhaps because of this, the number of total attacks is not as high. The attack covariance matrix, however, reveals that the attacks that did occur appear to exhibit substantial coordination.

The main results from analysis via spectral clustering are thus that insurgent attacks in Afghanistan are best represented as the work of a single organized group (plus "unorganized" local militants) and that this single insurgent group is active both to the north of Kabul and to the south. The eigengap analysis suggests that the conclusion regarding the number of groups would not change under small perturbations of the data.

Although the analysis using spectral clustering makes a strong case for a unitary insurgent actor, there could still be concerns that any analysis based on the off-diagonal entries of the sample covariance matrix is fundamentally misguided, because random variation will overwhelm any signal from actual coordinated attacks. Table 4 addresses this concern by using total attacks in a district as the left hand side variable: the results are similar to those in Table 3.[31]

We now present the results based on the non-negative matrix factorization technique of

[30] Ethnic variables are based on share of settlements in districts covered by GREG ethnic boundaries.

[31] As long as the number of attacks from "disorganized" local militants is not too high, one would expect the analysis of total attacks in Table 4 to give similar results to those in Table 3: the random attacks are simply a form of measurement error.

Section 2.2. As this method involves a comparison with permuted data that by construction has no group structure, results come with some indication of their statistical robustness.[32] Tables 5 - 7 show the results of the "gap statistic" type procedure based on Inequality 7, following Tibshirani, Walther and Hastie [2001]. The tables differ in the "reference distribution" used for comparison with the actual clustering results. In Table 5 we permute the attacks across time within each district (thus holding the total number of attacks in each district constant). In Table 6 we permute the attack time series for each district within each given month (hence preserving low frequency trends in attacks within districts). In Table 7 we permute the attack time series across districts and time holding constant the total number of attacks within a district and the number of attacks in a day across districts (hence preserving the marginal distributions of attacks across days and across districts). These three tables correspond to the three types of permutations discussed in Section 2.2.

Each of Tables 5, 6 and 7 are organized identically. The first set of four columns reports results for Afghanistan, and the second set for Pakistan. Columns marked I and II use the attack covariance matrix as discussed in Section 2.2, while the within-month covariance model of Section 2.3 is estimated in Columns marked III and IV. Columns marked I and III use exactly the data used for the analysis by spectral clustering, where districts with fewer than 3 attacks were excluded. Columns marked II and IV use data from all districts, but with a penalty function that weights each $\gamma_{ii'}$ entry proportionally to the total number of attacks in districts i and i'. This weighting is ad hoc, but accounts for the fact that estimates of insurgent prevalence for districts with very few attacks will be very noisy, because little information is available.[33]

For illustration, consider the estimates reported in Column I for Afghanistan of Table 5. Under the null hypothesis of zero organized groups of insurgents, the model is without degrees of freedom and hence the variation left unexplained in both the actual data and the permuted (reference distribution) data is 1 (i.e. all of it), leaving $\text{Gap}(0) = 0$ in the third row

[32]More specifically, there is an obvious permutation test of the null hypothesis that there is no group structure, and the null is rejected at the 95% level. A formal test of $J = 1$ against $J = 2$ (and so forth) appears more complicated, and in this case the results shown have the heuristic interpretation that is common in the clustering literature.

[33]As is often the case, weighting does not affect the consistency of the estimator. Here weights are used in order to ensure reasonable performance with the sample actually observed.

(marked "A"). Allowing for one group of insurgents in the data leads to N free parameters (an $\hat{\alpha}_{i1}$ for each district i), and leaves 88.3% of the variation in the data unexplained in the randomly permuted data; there is a better fit in the actual data, with only 60.0% of the variation left unexplained. This produces Gap(1) = 28.4% (marked "B"). The gap statistic (B-A) is thus 28.4%, which is higher than 5.5%, the estimated standard error for the objective function obtained by randomly drawing a large number of covariance matrices from the reference distribution and calculating W_1 for each of these matrices. This suggests 1 or more groups are present.

Continuing down the rows of Table 5, we proceed to consider the possibility of two organized groups of insurgents. This model has $2N$ free parameters and leaves 82% of the variation unexplained in the permuted data. Again, the fit is better in the actual data, with only 53.4% unexplained variation. This produces Gap(2) = 28.6% (marked "C"). The gap statistic (C-B) is thus 0.2%, which is now lower than 4.5%, the estimated standard error for W_2 in the reference distribution. This satisfies Inequality 7, and we thus conclude that there are not 2 (or more) groups. That is, the conclusion from this column is that the Taliban operate as a unified organization (i.e. $\hat{J} = 1$), as previously suggested by spectral clustering. The same conclusion is obtained in Columns II and IV in this Table and by Columns I-IV in both 6 and 7, a large set of alternative specifications all pointing in the same direction.[34] In addition, the method outlined in Section 2.3 confirms that the claim that the data is best represented by only one organized insurgent group is not due to long-term trends in attacks that are the same across districts, but is indeed due to coordinated attacks at a day-by-day frequency.

The "Pakistan" columns replace the Afghan attack data with comparable data from WITS covering Pakistan. Tables 5, 6 and 7 show that the Pakistani results differ markedly from those presented in the Afghan Taliban case. Whereas adding a group structure to the attacks is able to explain a statistically significant fraction of the Afghan attacks, as compared to random attacks, the attacks in Pakistan do not appear to match this sort of clustered structure as well. Typically Columns I and II point to zero groups being present,

[34]Only Column III of Table 5 shows a very thin case for $\hat{J} = 2$. This is one out of 12 specifications, and may be the result of random variation.

while discordantly Columns III and IV point to three or more groups. Overall, the unified insurgent structure we recover for the Afghan case appears not to be present in Pakistan. This accords with the qualitative analysis in Dorronsoro [2009].

Figure 9 shows the estimated prevalence of organized insurgents under the non-negative matrix factorization approach. The general pattern appears to agree with the qualitative description of insurgent activity just given for the spectral clustering method and shown in Figure 8. The estimates from the non-negative matrix factorization method appear to make it slightly clearer that the majority of organized insurgent activity is on the ring highway passing through Kabul, and that this activity extends to the north as well as the south of Kabul, possibly with the goal of isolating it. Estimates of the prevalence of the organized group can also be produced using the method in Section 2.3. As the estimates in this case are effectively based only on variation within months, estimates appear slightly noisier. Figure 10 shows these estimates. The tendency towards organized insurgent activity along the main highway can still be seen, although it is not as clear as in Figures 8 and 9.

4.1 Changes in group structure across time

The econometric model outlined so far assumes that the extent and prevalence of the organized insurgent group remains constant across time. This section considers how we can relax this strong assumption and, in the process, reveal novel information on the organization and strategy of the insurgents.

A formal model that allows for this structure to change over time appears challenging to develop. An informal analysis of potential changes can be conducted, however, by splitting the data. Specifically, we create an "early" data set, including only attacks in 2004-2007, and a "late" data set, including only attacks in 2008-2009. The total daily number of attacks is substantially higher in the later period compared to the earlier one, as already discussed for Figures 4 and 5. Estimates of the prevalence of organized insurgents from the earlier data can be compared to estimates from the later data, yielding a description of how the structure and location of insurgent groups has changed over time.[35]

[35]The informal nature of this analysis is due to the fact that the cut point of January 1, 2008, was chosen based on qualitative information: the econometric model is not one of structural breaks.

A first important finding here is that the unified Taliban organization is detectable in both the early and the complete samples, with one group of insurgents. The coordinated and unified behavior of the Taliban is not a feature developing over time, rather it is present from the onset.[36]

Concerning the territorial control of the Taliban, Figure 11 shows an estimate of the number of attacks due to organized insurgent groups in the earlier period, while Figure 12 shows this for the later period. The colors of the figures are aligned so that the same color indicates the same number of attacks per capita per year, although the "early" and "late" data have a different number of months. Comparing the two pictures shows unambiguously how much gain in territorial control and coordination have been characteristic of the Taliban offensive since 2008.

With respect to the distribution of attacks across districts, Figure 11 shows a lower frequency of attacks overall, and most districts that do see a high frequency of attacks are near the main highway to the south and west of Kabul. Figure 12 shows a higher frequency of attacks, and also shows districts in the north with high frequencies of attacks. One example of this is the highway north of Kabul, where now appear to be a number of districts with high frequencies of attacks. This claim is difficult to test statistically, because of the small number of districts in question.

A statistical analysis of changes in the distribution of attacks does reveal some patterns that are statistically significant and of relevance for current efforts in the management of the conflict. Table 8 investigates the correlates of insurgent group control in each district in the early and late periods by stacking the set of districts and employing interactions with POST dummy for the 2008-09 period. Control by the insurgents is measured through the sum of off-diagonal entries of the covariance matrix of attacks for district i according to the approximation in Equation 4. The Table reports both OLS in Columns I-IV and a Generalized Linear Model (Poisson distribution, allowing for overdispersion) in Columns V-VIII.

[36]This also shows that our results are not due purely to the August 2009 presidential election, when there were many attacks on and around election day. While there is substantial evidence that many of these attacks were in fact coordinated by the Taliban, it would be worrisome if the results presented thus far changed drastically when the attacks around the 2009 elections were excluded.

As in Table 3, the results in Table 8 show a clear relationship between ethnicity and simultaneous attacks. Table 8 reports coefficients for seven ethnic group dummy variables (indicating the largest ethnic group in the district), with Pashtun as the omitted dummy variable. For interpreting the Table, recall that the Afghan Taliban are traditionally ethnically Pashtun and follow Sunni Islam. Their historical opposition to Uzbeks, Hazara, and Tajiks – the main ethnic minorities in Afghanistan – is well documented. It is therefore unsurprising that our measure of Taliban activity in Afghan districts is negatively correlated with dummy variables indicating non-Pashtun ethnicity.

What is more surprising is that the coefficient in the POST interaction with the ethnicity variables is generally positive and of magnitude between 25 and 70 percent of the main effect. Consider for instance the case of the Uzbeks in Column II: the main effect is a statistically significant -2.11, indicating a much lower penetration of the Taliban in Uzbek areas in 2004-2007. In the 2008-09 period Uzbek districts are still less likely to experience Taliban activity, but now the coefficient falls by more than half, to -0.76 $(= -2.11 + 1.35)$. The distinction between Uzbek and Pashtun districts is thus decreased in the later period. In Column IV, where we exploit within-province variation, the distinction between Uzbek and Pashtun districts disappears completely or even reverses $(0.3 = -1.05 + 1.35)$. Although less statistically precise, Tajik and Hazara areas appear to display a similar pattern: districts with non-Pashtun ethnicities exhibit relatively greater activity in the later period, indicating a substantial penetration of the Taliban into areas previously outside their reach.

For added robustness, Table 9 employs as a dependent variable the pairwise off-diagonal entries of the covariance matrix of attacks, essentially carrying out the analysis of the insurgents' change in strategy at the district-pair level as opposed to district level. While the direction of the findings in Table 8 in terms of Taliban penetration in non-Pashtun dominated areas is confirmed, the statistical precision of our estimates is much more pronounced in Table 9.

Finally, as a check on the district-level analysis, Table 10 considers total attacks as the dependent variable. Results are consistent with those of Table 8.

4.1.1 An Oil Spot Strategy

Krepinevich [2005] discusses a state of the art counterinsurgency doctrine where control is expanded gradually across space. This is sometimes referred to an "oil spot" strategy, as the area controlled expands like an oil stain.[37] Our methodology allows us to ask a reverse question, regarding the strategy of insurgents: over time, how do attacks expand across space? Do the Taliban appear in completely new and disconnected areas, or do they launch attacks in districts that are adjacent to those that they were operating in previously? We conclude that the Taliban follow an "oil spot" insurgency strategy based on gradual expansion.

We begin by calculating, for each district, the estimated number of attacks by organized groups in adjacent districts, following Equation 4. The results of this calculation are shown in Figure 13: the districts where this variable is zero are shown in blue.[38] All but one of these blue districts are also not estimated to have any organized attacks in the later period, as can be seen by comparing Figure 12 with Figure 13. In particular, there were no attacks in the central part of Afghanistan in the early period, or much of the northeast, and these areas similarly do not have any attacks in the later period. On the other hand, districts immediately adjacent to estimated early Taliban strongholds appear prone to insurgent expansion.

Table 11 shows that this qualitative pattern is statistically significant. The basic specification used here is

$$\begin{aligned}\text{ATTACKS\_LATE}_i &= \beta_0 + \beta_1 \text{ATTACKS\_EARLY}_i \\ &+\beta_2 1(\text{ATTACKS\_EARLY\_ADJACENT}_i = 0) + \epsilon_i\end{aligned}$$

where ATTACKS_LATE is the number of attacks estimated to be due to organized insurgents in the later period, and ATTACKS_EARLY this number for the earlier period. ATTACKS_EARLY_ADJACENT is the average number of attacks in geographically adjacent districts. This last variable is used only as indicator variable: are there an estimated positive number of attacks attributed to organized groups in adjacent districts?[39] Columns I-III of

[37] This strategy appears to date back to the 19th century, as part of French colonial doctrine. Potiron de Boisfleury [2010] provides a detailed historical account.

[38] As in previous figures showing estimates of organized group activity, the units for "number of attacks" shown in the legend here are arbitrary, and thus only relative comparisons can be made.

[39] The dummy recoding is used because there is a long-standing problem in the analysis of spatial data

Table 11 show that districts where there was no insurgent group activity in the early period are less likely to experience insurgent group activity in the later period, and that this result is robust to a variety of controls, including province fixed effects.

Based on the definition of organized group attacks in Section 2, there should never be a negative number of attacks attributed to organized group members. Columns IV-VI of Table 11 thus present the same analysis using a Poisson GLM model, in order to take this non-negativity into account. An additional advantage of the Poisson model is that districts with few attacks are (correctly) treated as having higher variance relative to mean.[40] The results in Columns IV-VI confirm that there is very little organized insurgent activity in the late period in districts that did not border a district with such activity in the early period. The large coefficient on the ATTACKS_EARLY_ADJACENT indicator variable is due to the fact that the data exhibits "almost" complete separation: if there were zero districts rather than one that saw organized insurgent activity in the late period without any adjacent activity in the early period, the estimated coefficient here would be negative infinity, and it would not be possible to calculate standard errors by standard methods.[41]

5 Insurgency Organization and Economic Recovery

This section briefly discusses case studies chosen to highlight the economic importance of understanding insurgent organization in conflict and post-conflict environments. We focus on two different episodes: Iraq, and Syria.

Insurgent groups owe their success to their deep ties with noncombatant populations. By impeding reconstruction efforts, they can fuel popular dissatisfaction with central authorities, thereby maintaining a steady flow of recruits and ensuring logistic assistance for their agents. Insurgencies thus have a particular incentive to delay aggregate economic recovery.

regarding how to use this type of "adjacent observations" data, and there does not appear to be a satisfactory solution in this case.

[40] Weighted least squares could also be used here, but the Poisson model is natural as the underlying attack data is positive integers. The estimated number of attacks attributed to organized group members are non-integer, but this does not cause a problem for generalized linear models of the sort used.

[41] As an additional test, Table 12 repeats the regressions in Columns I-VI of Table 11 without the ATTACKS_EARLY variable. The estimated coefficient on the ATTACKS_EARLY_ADJACENT indicator variable is still negative (and large in the case of Columns IV-VI), although no longer statistically significant when province fixed effects are included. Table 12 shows that the results in Table 11 are not due purely to statistical relationships within the districts that did have attacks in 2004-2007.

In Iraq, insurgents disrupted the electricity grid and seized control of oil resources. Henderson [2005] describes the loop that linked insecurity and economic stagnation:

> *Inability to provide security had a profound impact on Iraq's economic recovery. In turn, inability to provide recovery had a profound impact on Iraq's security. Reconstruction delays fed into Iraqi feelings of resentment and despair, which fueled insurgency and crime, thereby worsening the security climate.*

The connection of the study of insurgency with economic development comes from this tight link between insurgent strategies and the failure of reconstruction efforts. Understanding the exact nature of the Iraqi insurgency early on in the conflict could have proven crucial in breaking the vicious cycle that Henderson [2005] observes.[42]

Uncertainty about the organization of the insurgency in post-2003 Iraq took several forms. First, there was disagreement regarding the extent to which attacks represented an insurgency at all.[43] There was also confusion regarding its magnitude: as late as the fall of 2004, the U.S. military still attributed 80 percent of attacks to random and not political violence. Finally, there was debate about the organization of the insurgency, once it was clear that one existed.[44] Further complexity in the Iraqi case stemmed from signs of evolution over time: *"the insurgency was now organized regionally, and that evidence pointed to some planning across regional boundaries".*[45]

The difficulty, and the importance, of understanding the structure of insurgencies is not limited to Iraq. Consider recent Western efforts in Syria:

> *Sixteen months into the uprising in Syria, the United States is struggling to de-*

[42]Henderson is critical of the strategy actually used: *"as violence worsened, the response of coalition officials in charge of reconstruction was not to find a way to fight it more effectively. Instead, their response was to withdraw into the heavily protected world of the Green Zone."*

[43]Eisenstadt and White [2005] write that *"In the summer of 2003, Secretary of Defense Donald Rumsfeld and General John Abizaid (head of U.S. Central Command) publicly disagreed about whether the violence in the Sunni Triangle was the final act of former regime "dead-enders" or an incipient insurgency against the emerging political order*". There was a similar disagreement in 2005 between Vice President Richard Cheney and General Abizaid.

[44]The New York Times quotes senior U.S. intelligence sources stating that "*It's not just one group of insurgents rallying under one cause. It's multiple groups with different causes loosely tied together by the threads of anti-U.S. sentiment, some sort of Iraqi nationalism, Muslim-Arab unity or greed*". The lack of familiarity with this type of enemy appeared evident: "*What makes it more difficult is that you're dealing with an insurgency without a single face".*

[45]http://www.nytimes.com/2004/10/22/international/middleeast/22insurgents.html?pagewanted=2&_r=0

velop a clear understanding of opposition forces inside the country, according to U.S. officials who said that intelligence gaps have impeded efforts to support the ouster of Syrian President Bashar al-Assad.[46]

Beginning with a series of pro-democracy protests in 2011, the situation in Syria quickly escalated into a full-blown civil war that has cost 200,000 lives and displaced at least 4 million Syrian citizens. Lack of understanding of the structure of the insurgency in Syria has been one of the strongest deterrents to military and humanitarian involvement of Western powers in this conflict.

Western countries were willing to lend support and provide prompt international aid to moderate Sunni organizations, but the difficulty lay in identifying these rebels. The impossibility of separating the secular moderates from the religious extremists among the Sunni opponents of the Alawite-led government resulted in international paralysis. This led to further economic and social deterioration, radicalization, and escalation of the conflict. Syria is now a nearly failed state, fought over by Assad loyalists, the Islamic State, and the al-Qaeda affiliated Nusra front. Numerous attempts at a political solution by the Arab League and the United Nations have failed.

6 Conclusions

This paper focuses on the empirical analysis of insurgency, with an application to post-2001 Afghanistan. Often the only type of data available concerning the amount and geographical diffusion of insurgent activity comes from incident-level data on insurgent attacks. However limited such information might be, recent important advances in the analysis of the economics of conflict and reconstruction in war zones have been possible thanks to this data.[47]

This paper shows how incident-level data contains information that can be used to estimate the structure and geographic span of influence of insurgent groups. We present a set of econometric methods to detect unobserved insurgent coalition structures, based on

[46] http://www.washingtonpost.com/world/national-security/in-syria-conflict-us-struggles-to-fill-intelligence-gaps/2012/07/23/gJQAW8DG5W_story.html

[47] Berman, Shapiro, and Felter [2011] is one recent example.

co-occurrences of violent incidents across districts over time. If incidents in two districts occur simultaneously more than would be expected by random chance, then this suggests that these districts share an organized insurgent movement, one capable of cross-district coordination. We then carry out an analysis of the spread and frequency of attacks. Specific geographic and historical characteristics, in particular highways and the ethnic composition of the local population, predict insurgent presence and growth.

Progress in understanding insurgency is key in furthering our knowledge of the determinants and consequences of political violence in developing countries. Although much of the analysis in this paper is necessarily context-dependent, it is informative nonetheless for regional stabilization and local development goals [Drozdova, 2012]. From a methodological perspective, our contributions have a more general appeal.

Figure 1: Ethnicities of Afghanistan

Figure 2: Afghan Ring Road

(via Wall Street Journal)

Figure 3: Total attacks per capita

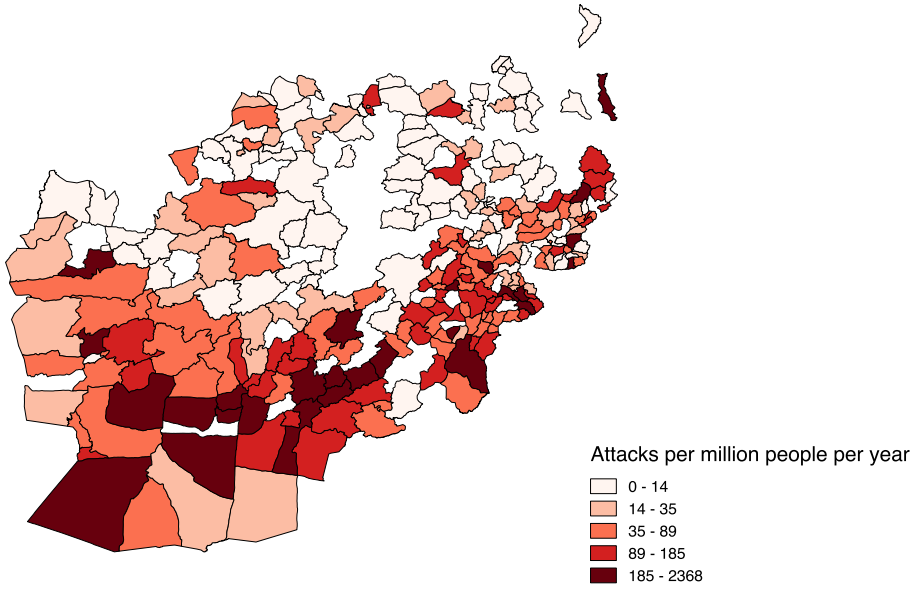

Figure 4: Attacks per capita 2004-2007

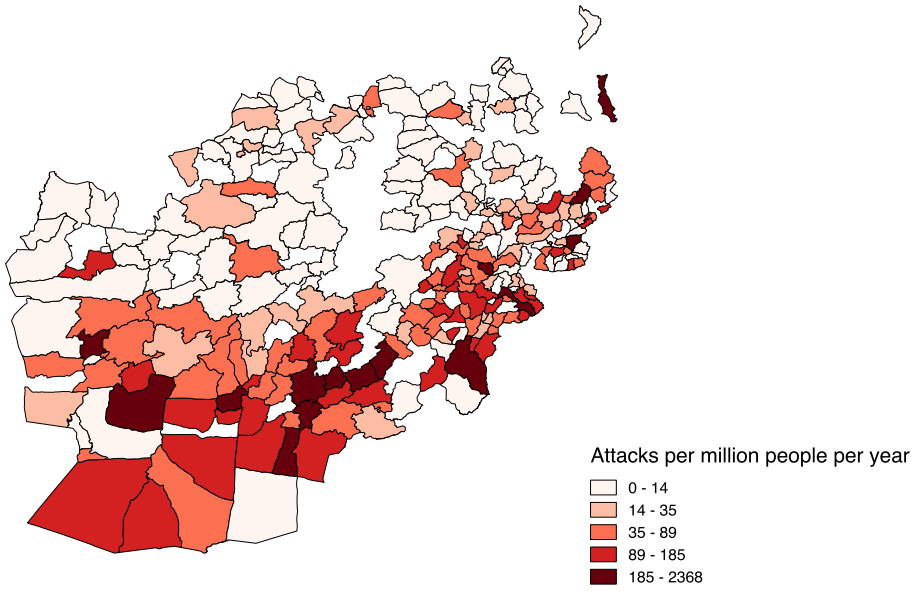

Figure 5: Attacks per capita 2008-2009

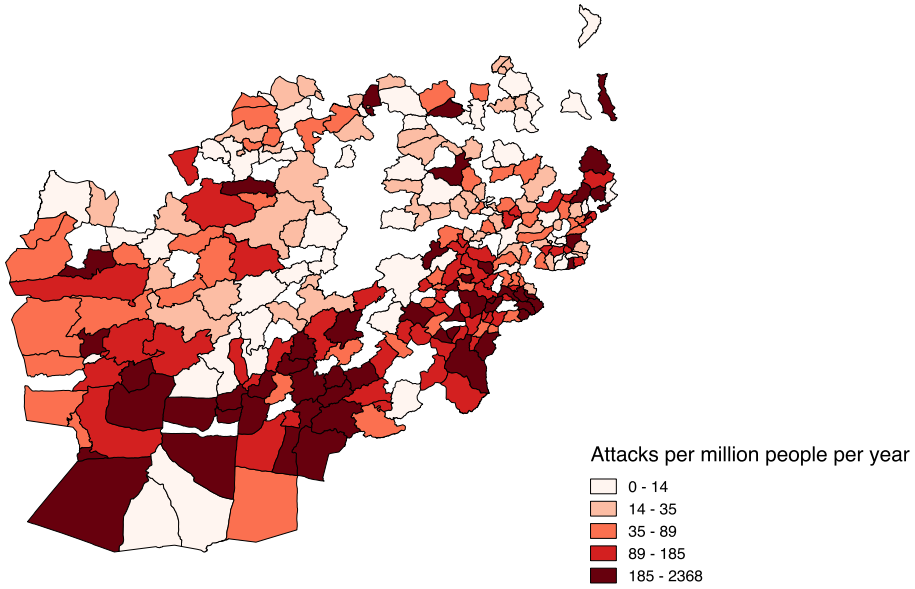

Figure 6: (Sorted) Eigenvalues for Spectral Clustering

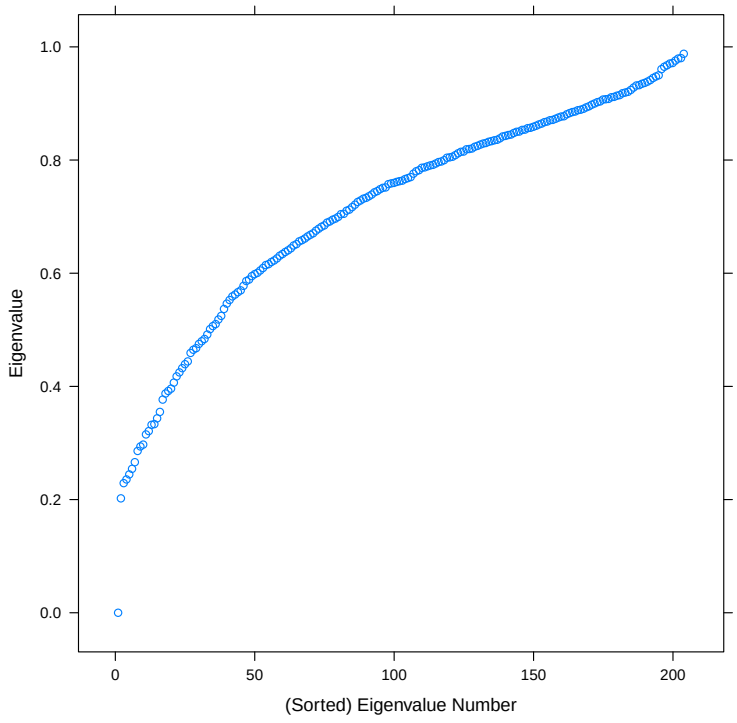

Figure 7: Eigengaps for Spectral Clustering

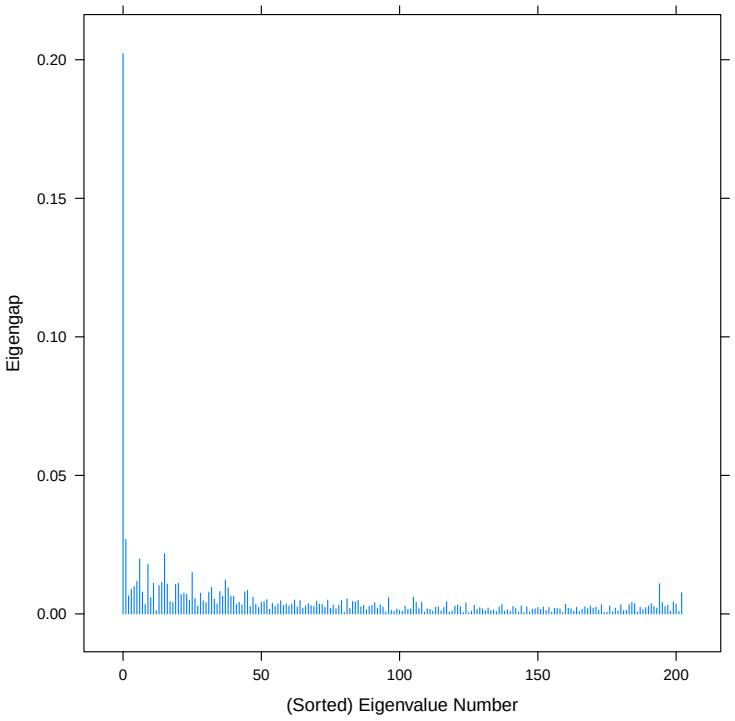

Figure 8: Organized group members: Spectral clustering (Equation 3)

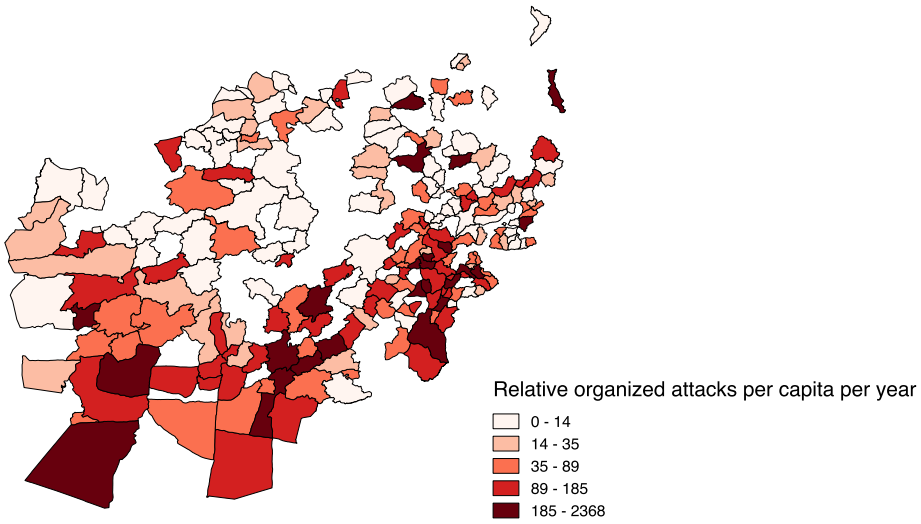

Figure 9: Organized group members: NNMF method (Section 3.2)

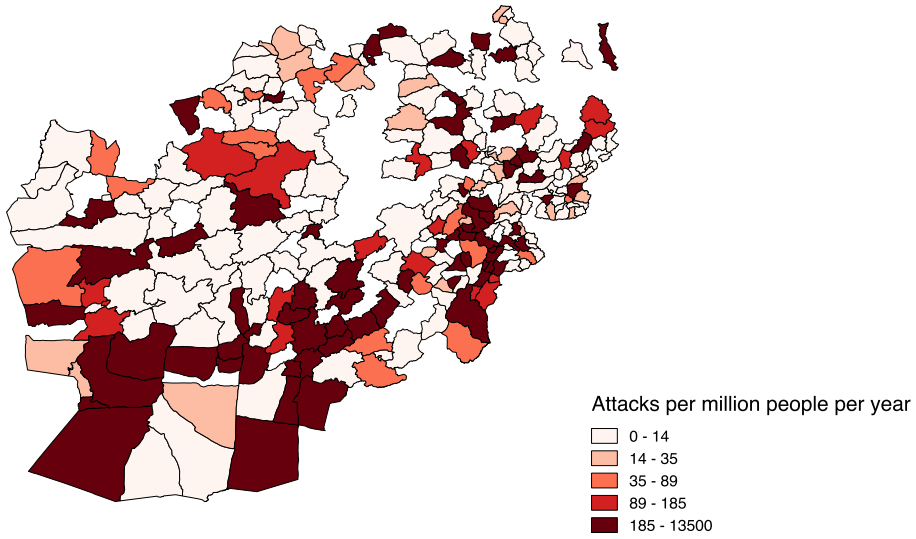

Figure 10: Organized group members: NNMF method (Section 3.3)

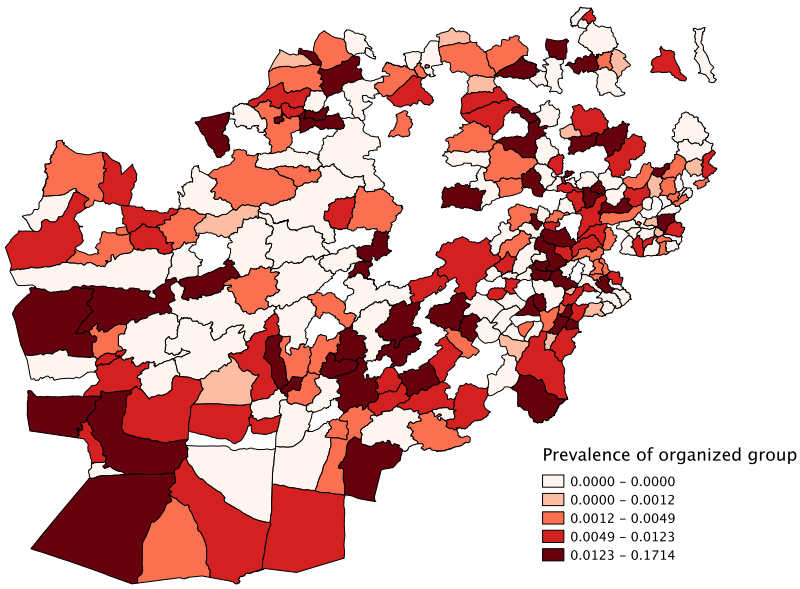

Figure 11: Organized group members: Spectral clustering (2004-2007)

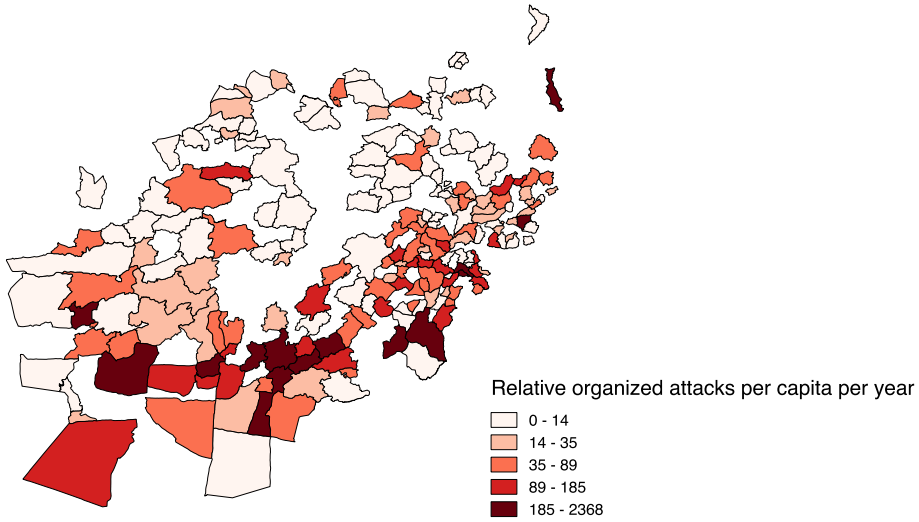

Figure 12: Organized group members: Spectral clustering (2008-2009)

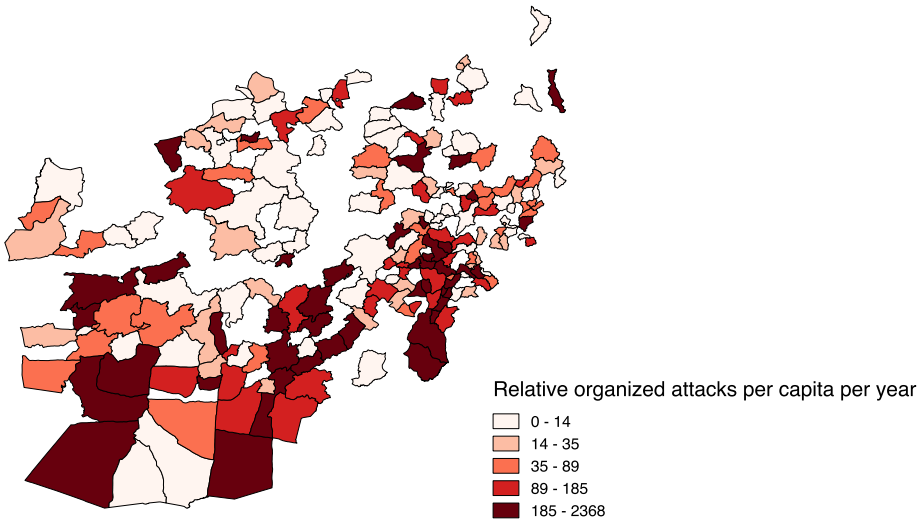

Figure 13: (Estimated) attacks by organized group members (2004-2007, average over adjacent districts)

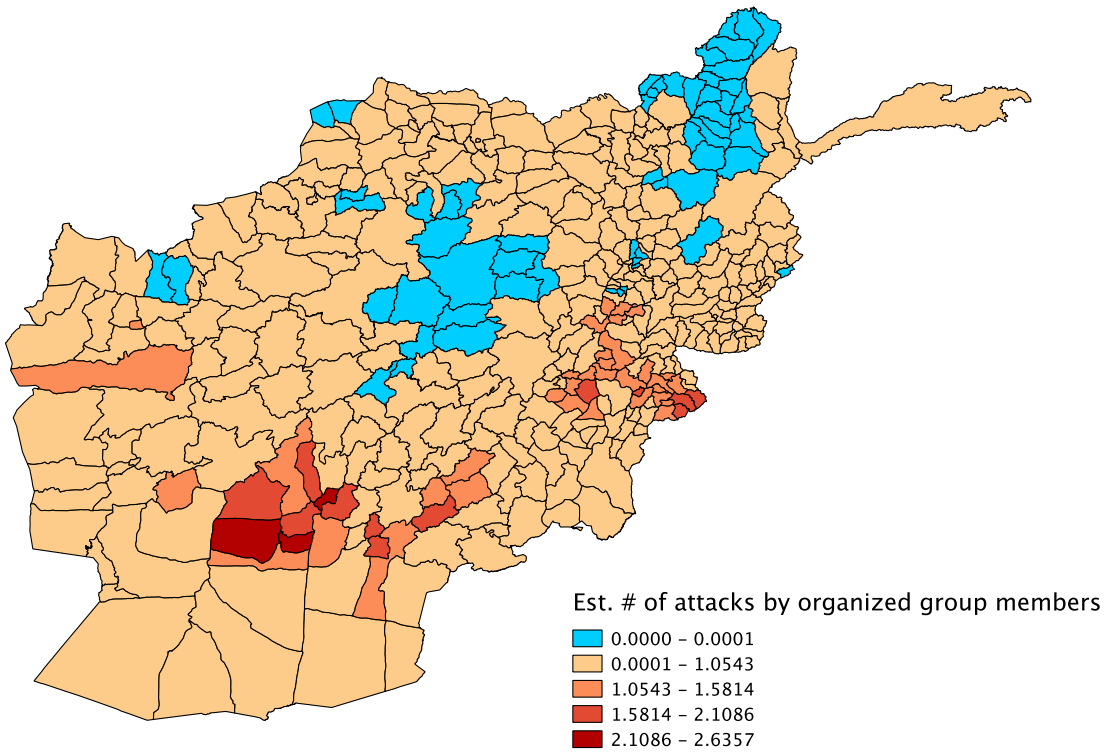

Table 1: Afghanistan timeline 2001-2011

18-Sep-01	President George W. Bush signs into law a joint resolution authorizing the use of force against those responsible for attacking the United States on 9/11.
7-Oct-01	The U.S. military, with British support, begins a bombing campaign against Taliban
Nov-01	The Taliban regime unravels rapidly after its loss at Mazar-e-Sharif on November 9th
Dec-01	Osama bin Laden escapes from Tora Bora
5-Dec-01	Hamid Karzai is installed as interim administration head after the Bonn Agreement
9-Dec-01	The Taliban surrender Kandahar, their regime collapses.
17-Apr-02	U.S. Congress appropriates over $38 billion in humanitarian and reconstruction assistance to Afghanistan from 2001 to 2009.
1-May-03	U.S. Secretary of Defense Donald Rumsfeld declares an end to "major combat."
8-Aug-03	NATO assumes control of international security forces (ISAF) in Afghanistan
Jan-04	Afghan Constitution is approved.
9-Oct-04	Hamid Karzai is popularly elected as president.
29-Oct-04	Osama bin Laden releases a videotaped message three weeks after the country's presidential election.
18-Sep-05	Legislative elections in Afghanistan for the Wolesi Jirga (Council of People) and the Meshrano Jirga (Council of Elders)
Jul-06	Violence increases across the country, including suicide attacks.
Nov-06	U.S. Secretary of Defense Robert Gates criticizes NATO countries in late 2007 for not sending more soldiers.
22-Aug-08	Afghan civilian casualties mount. Gen. Stanley A. McChrystal orders an overhaul of U.S. air strike procedures.
17-Feb-09	New U.S. president Barack Obama announces plans to send seventeen thousand more troops to Afghanistan. Reinforcements focus on countering a "resurgent" Taliban and stemming the flow of foreign fighters over the Afghan-Pakistan border in the south.
27-Mar-09	New American strategy focused on disrupting Taliban safe havens in Pakistan
11-May-09	Secretary of Defense Robert Gates replaces the top U.S. commander in Afghanistan, Gen. David D. McKiernan, with counterinsurgency and special operations guru Gen. Stanley A. McChrystal.
Jul-09	U.S. Marines launch a major offensive in southern Afghanistan (Helmand Province), representing a major test for the U.S. military's new counterinsurgency strategy.
Nov-09	Hamid Karzai is popularly re-elected as president.
1-Dec-09	President Obama announces a major escalation of the U.S. mission, an Afghan surge.
23-Jun-10	Gen. Stanley McChrystal is relieved of his post as commander of U.S. forces in Afghanistan
1-May-11	Osama bin Laden killed in Pakistan
Jun-11	President Obama outlines a plan to withdraw troops according to NATO plans of complete withdrawn by 2014
7-Oct-11	10 years of counterinsurgency war. 1,800 U.S. troop casualties and $444 billion in spending

Source: Council on Foreign Relations
http://www.cfr.org/afghanistan/us-war-afghanistan/p20018

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
PASHTUN	396	0.516	0.439	0.000	1.000
UZBEK	396	0.123	0.285	0.000	1.000
BALOCH	396	0.015	0.099	0.000	1.000
HAZARA	396	0.097	0.257	0.000	1.000
TAJIK	396	0.219	0.357	0.000	1.000
PAMIR.TAJIK	396	0.013	0.094	0.000	1.000
ORMURI	396	0.005	0.050	0.000	0.731
NURISTANI	396	0.012	0.084	0.000	0.846
POPULATION	398	58.673	150.129	1.841	2,882.164
AREA	398	1.948	2.624	0.032	25.128
LIGHT	398	0.051	0.192	0.000	2.000
LATITUDE	398	34.580	1.724	29.889	38.225
LONGITUDE	398	67.796	2.607	61.156	73.349
ROADS	398	1.063	1.212	0	6
RIVERS	398	0.798	1.687	0.000	13.598

The first eight variables indicate the shares of ethnicities in each district. PASHTUN also includes Pashai, Tirahi, Afghan Arabs, and Persians. UZBEK also includes Turkmens and Kirghis. BALOCH also includes Brahui. HAZARA includes Mongols, in addition to Hazaraberberi and Hazaradehizainat. TAJIK also includes Jamshidis, Taimanis, Firozkohis, Teymurs. ORMURI includes Parachi. There are two districts for which ethnic information is not available.

POPULATION is in thousands of people. AREA is in thousands of square km. LIGHT is a index of nightime light emissions. LATITUDE and LONGITUDE are in degrees. ROADS is the number of major roads in the district. RIVERS is the total length of rivers in the district.

Table 3: Dep. variable is sum of off-diagonal entries of cov. matrix for a given district i

	I	II	III	IV	V	VI	VII	VIII
(Intercept)	2.57*	1.86*	0.18	−2.28*	3.83*	3.93*	1.87*	−0.36
	(0.19)	(0.75)	(0.62)	(0.96)	(0.12)	(0.54)	(0.38)	(0.83)
UZBEK	−0.56	−0.20	−1.06*	−0.08	−1.17*	−0.98	−1.57*	−0.98
	(0.38)	(0.73)	(0.39)	(0.71)	(0.36)	(0.55)	(0.41)	(0.69)
BALOCH	−1.78	−2.65	−1.93	−1.55	−2.24*	−2.80*	−2.19*	−1.57
	(1.49)	(1.42)	(1.47)	(1.34)	(0.93)	(1.03)	(1.07)	(1.09)
HAZARA	−1.46*	−2.27*	−2.14*	−2.44*	−0.75	−0.71	−1.17	−0.74
	(0.38)	(0.54)	(0.40)	(0.73)	(0.61)	(0.61)	(0.66)	(0.65)
TAJIK	−0.89*	−0.19	−1.33*	−0.35	−0.44	−0.26	−0.81*	−0.26
	(0.38)	(0.78)	(0.37)	(0.62)	(0.29)	(0.81)	(0.28)	(0.47)
PAMIR.TAJIK	0.97*	3.77*	1.95*	4.49*	−0.35*	3.66*	0.20	4.28*
	(0.22)	(0.81)	(0.44)	(0.74)	(0.13)	(0.88)	(0.32)	(0.61)
ORMURI	1.64	−0.28	1.05	−1.64*	0.61	−0.18	0.24	−1.55*
	(0.87)	(0.44)	(0.62)	(0.69)	(0.75)	(0.28)	(0.54)	(0.70)
NURISTANI	−1.45	0.51	−0.94	0.91	−3.02*	−0.76*	−1.85	−0.43
	(1.21)	(0.32)	(1.37)	(2.05)	(1.31)	(0.19)	(1.13)	(1.06)
logPOP			0.52*	0.73*			0.43*	0.59*
			(0.17)	(0.19)			(0.09)	(0.13)
logAREA			0.38*	0.18			0.28*	0.19
			(0.13)	(0.16)			(0.08)	(0.12)
logROADS			0.55*	0.59*			0.43*	0.60*
			(0.23)	(0.26)			(0.17)	(0.17)
logRIVERS			−0.22*	−0.13			−0.03	−0.01
			(0.09)	(0.13)			(0.07)	(0.09)
PROV		Y		Y		Y		Y
N	262	262	262	262	262	262	262	262
R^2	0.06	0.24	0.18	0.35				
adj. R^2	0.04	0.10	0.15	0.22				
Resid. sd	1.93	1.86	1.82	1.74				

Columns I - IV use OLS with dependent variable log transformed

Columns V - VIII use GLM/Poisson allowing for overdispersion

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Table 4: Dependent variable is total attacks for district i

	I	II	III	IV	V	VI	VII	VIII
(Intercept)	2.58*	1.96*	0.23	−1.97*	3.37*	3.98*	1.02*	−0.97
	(0.11)	(0.82)	(0.32)	(0.71)	(0.14)	(0.61)	(0.34)	(0.77)
UZBEK	−1.65*	−1.36*	−2.04*	−1.30*	−2.17*	−1.74*	−2.70*	−2.15*
	(0.24)	(0.47)	(0.22)	(0.44)	(0.29)	(0.51)	(0.33)	(0.50)
BALOCH	−2.02*	−3.03*	−1.59*	−1.50*	−2.38*	−3.29*	−1.93*	−1.71*
	(0.54)	(0.43)	(0.49)	(0.48)	(0.44)	(0.38)	(0.42)	(0.41)
HAZARA	−1.71*	−1.72*	−2.21*	−1.73*	−1.72*	−1.31*	−2.12*	−1.18*
	(0.26)	(0.38)	(0.28)	(0.33)	(0.43)	(0.53)	(0.43)	(0.51)
TAJIK	−1.12*	−0.52	−1.58*	−0.66	−0.82*	−0.05	−1.22*	−0.41
	(0.24)	(0.55)	(0.21)	(0.38)	(0.40)	(0.91)	(0.38)	(0.49)
PAMIR.TAJIK	0.16	2.01*	0.72*	2.36*	−0.64*	2.45*	0.02	2.80*
	(0.13)	(0.57)	(0.28)	(0.45)	(0.14)	(0.93)	(0.33)	(0.60)
ORMURI	0.85	0.61	0.10	−0.82	−0.00	0.36	−0.59*	−1.04*
	(0.51)	(0.54)	(0.24)	(0.50)	(0.28)	(0.37)	(0.20)	(0.36)
NURISTANI	−1.27*	−1.85*	−0.34	−1.29	−2.45*	−2.82*	−0.92	−2.89
	(0.50)	(0.38)	(0.60)	(1.08)	(0.56)	(0.40)	(0.52)	(1.57)
logPOP			0.60*	0.69*			0.49*	0.63*
			(0.08)	(0.10)			(0.08)	(0.11)
logAREA			0.19*	−0.04			0.24*	0.10
			(0.07)	(0.09)			(0.07)	(0.08)
logROADS			0.37*	0.57*			0.59*	0.77*
			(0.16)	(0.15)			(0.15)	(0.14)
logRIVERS			−0.03	0.03			−0.03	−0.01
			(0.06)	(0.07)			(0.08)	(0.07)
PROV		Y		Y		Y		Y
N	262	262	262	262	262	262	262	262

Columns I - IV use OLS with dependent variable log transformed

Columns V - VIII use GLM/Poisson allowing for overdispersion

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Table 5: Non-negative matrix factorization ("full shuffle" reference distribution)

			Afghanistan				Pakistan			
			I	II	III	IV	I	II	III	IV
0 grps	rnd shuffled data (mean)		1	1	1	1	1	1	1	1
	actual data	-	1	1	1	1	1	1	1	1
	gap	A	0	0	0	0	0	0	0	0
1 grp	rnd shuffled data (mean)		0.883	0.950	0.962	0.972	0.682	0.918	0.947	0.972
	actual data	-	0.600	0.721	0.921	0.883	0.773	0.655	0.903	0.889
	gap	B	0.284	0.229	0.042	0.089	-0.091	0.263	0.043	0.084
	gap statistic (B minus A)		0.284	0.229	0.042	0.089	-0.091	0.263	0.043	0.084
	rnd shuffled data (std. dev.)		0.055	0.018	0.011	0.008	0.133	0.040	0.008	0.008
2 grps	rnd shuffled data (mean)		0.820	0.913	0.937	0.951	0.537	0.861	0.903	0.951
	actual data	-	0.534	0.668	0.884	0.863	0.631	0.610	0.845	0.829
	gap	C	0.286	0.245	0.053	0.088	-0.094	0.251	0.058	0.122
	gap statistic (C minus B)		0.002	0.016	0.011	-0.001	-0.003	-0.012	0.015	0.038
	rnd shuffled data (std. dev.)		0.045	0.023	0.010	0.011	0.130	0.049	0.012	0.011
3 grps	rnd shuffled data (mean)		0.787	0.887	0.910	0.935	0.433	0.817	0.864	0.935
	actual data	-	0.493	0.633	0.858	0.842	0.501	0.580	0.785	0.783
	gap	D	0.294	0.254	0.052	0.093	-0.068	0.237	0.079	0.152
	gap statistic (D minus C)		0.009	0.009	-0.001	0.005	0.026	-0.015	0.021	0.030
	rnd shuffled data (std. dev.)		0.070	0.031	0.012	0.012	0.126	0.053	0.015	0.012
4 grps	rnd shuffled data (mean)		0.845	0.904	0.921	0.921	0.366	0.780	0.828	0.921
	actual data	-	0.458	0.603	0.836	0.825	0.419	0.543	0.750	0.750
	gap	E	0.387	0.301	0.085	0.096	-0.053	0.237	0.078	0.172
	gap statistic (E minus D)		0.093	0.047	0.033	0.003	0.015	0.001	-0.001	0.019
	rnd shuffled data (std. dev.)		0.073	0.038	0.032	0.013	0.114	0.055	0.016	0.013
5 grps	rnd shuffled data (mean)		0.880	0.918	0.956	0.908	0.315	0.749	0.796	0.908
	actual data	-	0.427	0.576	0.816	0.809	0.352	0.514	0.721	0.738
	gap	F	0.453	0.343	0.140	0.099	-0.037	0.235	0.076	0.170
	gap statistic (F minus E)		0.066	0.042	0.054	0.003	0.016	-0.002	-0.002	-0.002
	rnd shuffled data (std. dev.)		0.098	0.042	0.028	0.015	0.103	0.055	0.017	0.015

Columns I-II use the model in Section 2.2; III-IV use the model from Section 2.3.
Columns I and III consider only districts with more than three attacks.
Columns II and IV use all districts, but weight districts by the number of attacks.

Table 6: Non-negative matrix factorization ("monthly shuffle" reference distribution)

			Afghanistan				Pakistan			
			I	II	III	IV	I	II	III	IV
0 grps	rnd shuffled data (mean)		1	1	1	1	1	1	1	1
	actual data	-	1	1	1	1	1	1	1	1
	gap	A	0	0	0	0	0	0	0	0
1 grp	rnd shuffled data (mean)		0.880	0.939	0.958	0.968	0.736	0.667	0.943	0.956
	actual data	-	0.600	0.722	0.921	0.883	0.773	0.655	0.903	0.889
	gap	B	0.281	0.217	0.037	0.085	-0.037	0.012	0.039	0.067
	gap statistic (B minus A)		0.281	0.217	0.037	0.085	-0.037	0.012	0.039	0.067
	rnd shuffled data (std. dev.)		0.049	0.015	0.014	0.013	0.109	0.022	0.010	0.009
2 grps	rnd shuffled data (mean)		0.800	0.886	0.928	0.944	0.597	0.611	0.897	0.925
	actual data	-	0.534	0.669	0.884	0.863	0.631	0.603	0.845	0.829
	gap	C	0.266	0.217	0.044	0.081	-0.034	0.008	0.052	0.096
	gap statistic (C minus B)		-0.014	-0.001	0.007	-0.004	0.002	-0.004	0.013	0.029
	rnd shuffled data (std. dev.)		0.047	0.026	0.018	0.015	0.122	0.022	0.014	0.012
3 grps	rnd shuffled data (mean)		0.732	0.854	0.905	0.928	0.504	0.579	0.856	0.901
	actual data	-	0.493	0.634	0.858	0.842	0.501	0.572	0.785	0.782
	gap	D	0.239	0.220	0.046	0.086	0.003	0.006	0.071	0.120
	gap statistic (D minus C)		-0.028	0.003	0.003	0.005	0.038	-0.001	0.019	0.024
	rnd shuffled data (std. dev.)		0.065	0.030	0.018	0.015	0.120	0.024	0.017	0.014
4 grps	rnd shuffled data (mean)		0.680	0.813	0.884	0.917	0.434	0.551	0.820	0.881
	actual data	-	0.458	0.604	0.836	0.825	0.419	0.540	0.750	0.761
	gap	E	0.222	0.209	0.048	0.091	0.015	0.011	0.070	0.120
	gap statistic (E minus D)		-0.016	-0.011	0.002	0.005	0.012	0.004	-0.002	0.000
	rnd shuffled data (std. dev.)		0.051	0.032	0.0181	0.013	0.112	0.024	0.019	0.015
5 grps	rnd shuffled data (mean)		0.673	0.794	0.864	0.901	0.379	0.529	0.786	0.861
	actual data	-	0.427	0.577	0.816	0.809	0.353	0.519	0.720	0.733
	gap	F	0.246	0.217	0.048	0.091	0.026	0.010	0.066	0.129
	gap statistic (F minus E)		0.024	0.009	0.000	0.000	0.011	-0.001	-0.003	0.009
	rnd shuffled data (std. dev.)		0.096	0.033	0.018	0.015	0.104	0.024	0.020	0.017

Columns I-II use the model in Section 2.2; III-IV use the model from Section 2.3.
Columns I and III consider only districts with more than three attacks.
Columns II and IV use all districts, but weight districts by the number of attacks.

Table 7: Non-negative matrix factorization ("constant marginals" reference distribution)

			Afghanistan				Pakistan			
			I	II	III	IV	I	II	III	IV
0 grps	rnd shuffled data (mean)		1	1	1	1	1	1	1	1
	actual data	-	1	1	1	1	1	1	1	1
	gap	A	0	0	0	0	0	0	0	0
1 grp	rnd shuffled data (mean)		0.894	0.952	0.956	0.960	0.713	0.891	0.843	0.810
	actual data	-	0.600	0.722	0.921	0.883	0.773	0.654	0.903	0.888
	gap	B	0.294	0.230	0.035	0.077	-0.060	0.237	-0.060	-0.079
	gap statistic (B minus A)		0.294	0.230	0.035	0.077	-0.060	0.237	-0.060	-0.079
	rnd shuffled data (std. dev.)		0.040	0.018	0.011	0.009	0.106	0.020	0.022	0.033
2 grps	rnd shuffled data (mean)		0.821	0.919	0.922	0.934	0.560	0.823	0.785	0.760
	actual data	-	0.534	0.669	0.884	0.852	0.631	0.605	0.845	0.829
	gap	C	0.287	0.250	0.038	0.082	-0.071	0.217	-0.059	-0.069
	gap statistic (C minus B)		-0.008	0.020	0.003	0.005	-0.011	-0.019	0.001	0.009
	rnd shuffled data (std. dev.)		0.051	0.025	0.014	0.012	0.117	0.031	0.023	0.035
3 grps	rnd shuffled data (mean)		0.764	0.892	0.894	0.912	0.463	0.778	0.739	0.723
	actual data	-	0.493	0.634	0.858	0.831	0.501	0.576	0.785	0.780
	gap	D	0.271	0.258	0.035	0.082	-0.038	0.202	-0.046	-0.058
	gap statistic (D minus C)		-0.016	0.008	-0.003	0.000	0.033	-0.015	0.013	0.012
	rnd shuffled data (std. dev.)		0.053	0.028	0.016	0.013	0.119	0.035	0.024	0.036
4 grps	rnd shuffled data (mean)		0.716	0.868	0.869	0.894	0.396	0.741	0.700	0.692
	actual data	-	0.458	0.604	0.836	0.815	0.419	0.543	0.750	0.756
	gap	E	0.258	0.264	0.033	0.080	-0.023	0.198	-0.050	-0.064
	gap statistic (E minus D)		-0.013	0.006	-0.002	-0.002	0.015	-0.004	-0.005	-0.006
	rnd shuffled data (std. dev.)		0.055	0.030	0.017	0.013	0.115	0.036	0.024	0.037
5 grps	rnd shuffled data (mean)		0.674	0.847	0.846	0.879	0.347	0.710	0.664	0.668
	actual data	-	0.427	0.577	0.816	0.798	0.353	0.518	0.720	0.737
	gap	F	0.247	0.270	0.030	0.080	-0.006	0.192	-0.056	-0.069
	gap statistic (F minus E)		-0.011	0.006	-0.003	0.001	0.017	-0.006	-0.005	-0.005
	rnd shuffled data (std. dev.)		0.055	0.031	0.017	0.013	0.108	0.036	0.024	0.037

Columns I-II use the model in Section 2.2; III-IV use the model from Section 2.3.
Columns I and III consider only districts with more than three attacks.
Columns II and IV use all districts, but weight districts by the number of attacks.

Table 8: Dep. variable is sum of off-diagonal entries of cov. matrix for a given district i

	I	II	III	IV	V	VI	VII	VIII
(Intercept)	−1.08*	−0.96*	−1.23*	−3.23*	0.63	0.75*	0.61	−1.69
	(0.40)	(0.40)	(0.50)	(0.71)	(0.36)	(0.35)	(0.38)	(0.87)
POST	0.31*	0.09	0.62	0.62	0.58*	0.40*	0.67	0.76
	(0.15)	(0.20)	(0.79)	(0.73)	(0.15)	(0.17)	(0.62)	(0.63)
UZBEK	−1.44*	−2.11*	−2.11*	−1.05*	−1.54*	−3.28*	−3.34*	−2.44*
	(0.27)	(0.31)	(0.30)	(0.50)	(0.40)	(0.65)	(0.67)	(0.76)
BALOCH	−1.02	−1.54*	−1.05	−0.93	−1.90*	−2.45*	−2.23*	−1.56*
	(0.89)	(0.63)	(0.60)	(0.71)	(0.84)	(0.68)	(0.68)	(0.77)
HAZARA	−1.82*	−1.98*	−1.92*	−1.73*	−1.05	−2.20*	−2.18*	−1.85*
	(0.32)	(0.36)	(0.38)	(0.49)	(0.60)	(0.44)	(0.45)	(0.50)
TAJIK	−1.39*	−1.62*	−1.68*	−0.66	−0.84*	−1.00*	−1.03*	−0.57
	(0.23)	(0.30)	(0.29)	(0.47)	(0.25)	(0.38)	(0.38)	(0.48)
PAMIR.TAJIK	1.92*	2.03*	1.88*	3.60*	0.34	0.64*	0.60	4.43*
	(0.35)	(0.31)	(0.38)	(0.63)	(0.40)	(0.29)	(0.35)	(0.93)
ORMURI	−0.12	0.82*	0.60	−1.66*	0.17	−0.26	−0.36	−2.37*
	(1.24)	(0.32)	(0.36)	(0.75)	(0.57)	(0.20)	(0.19)	(0.82)
NURISTANI	−0.53	0.11	0.47	1.49	−2.35	−1.11	−0.77	0.51
	(0.76)	(1.00)	(0.95)	(1.22)	(1.23)	(0.97)	(0.91)	(0.82)
logPOP	0.60*	0.60*	0.70*	0.85*	0.44*	0.44*	0.45*	0.64*
	(0.11)	(0.11)	(0.13)	(0.14)	(0.09)	(0.08)	(0.09)	(0.14)
logAREA	0.25*	0.25*	0.14	0.03	0.26*	0.26*	0.24*	0.16
	(0.08)	(0.08)	(0.09)	(0.12)	(0.07)	(0.07)	(0.08)	(0.12)
logROADS	0.44*	0.44*	0.46*	0.55*	0.39*	0.39*	0.58*	0.71*
	(0.16)	(0.16)	(0.20)	(0.21)	(0.15)	(0.16)	(0.20)	(0.23)
logRIVERS	−0.08	−0.08	0.01	0.10	−0.03	−0.03	0.00	0.02
	(0.07)	(0.07)	(0.08)	(0.09)	(0.07)	(0.06)	(0.08)	(0.10)
POST:UZBEK		1.35*	1.35*	1.35*		2.22*	2.29*	2.09*
		(0.51)	(0.52)	(0.52)		(0.77)	(0.80)	(0.65)
POST:BALOCH		1.04	0.06	0.06		0.80	0.43	0.34
		(1.48)	(1.48)	(1.45)		(1.23)	(1.27)	(1.30)
POST:HAZARA		0.32	0.19	0.19		1.53	1.50	1.55*
		(0.60)	(0.65)	(0.53)		(0.80)	(0.85)	(0.70)
POST:TAJIK		0.46	0.57	0.57		0.26	0.31	0.36
		(0.45)	(0.46)	(0.41)		(0.51)	(0.50)	(0.46)
POST:PAMIR.TAJIK		−0.21	0.09	0.09		−0.59*	−0.53	−0.55
		(0.23)	(0.60)	(0.50)		(0.18)	(0.54)	(0.52)
POST:ORMURI		−1.88	−1.43	−1.43		0.64	0.81	0.99
		(1.97)	(1.95)	(1.52)		(0.86)	(0.91)	(0.77)
POST:NURISTANI		−1.28	−2.01	−2.01		−3.11	−3.74	−2.86
		(1.60)	(1.65)	(1.48)		(2.68)	(2.81)	(1.99)
POST:logPOP			−0.20	−0.20			−0.02	−0.05
			(0.21)	(0.20)			(0.15)	(0.15)
POST:logAREA			0.23	0.23			0.03	0.04
			(0.15)	(0.14)			(0.13)	(0.14)
POST:logROADS			−0.03	−0.03			−0.29	−0.26
			(0.33)	(0.29)			(0.30)	(0.29)
POST:logRIVERS			−0.18	−0.18			−0.04	−0.04
			(0.13)	(0.12)			(0.12)	(0.12)
N	524	524	524	524	524	524	524	524

Columns I - IV use OLS with dependent variable log transformed. Column IV has province fixed effects.
Columns V - VIII use GLM/Poisson allowing for overdispersion. Column VIII has province fixed effects.

Table 9: Dependent variable is off diagonal covariance matrix entry $i\ i'$

	I	II	III	IV
POST	0.234*	0.905*	0.909*	0.491
	(0.032)	(0.173)	(0.173)	(0.467)
UZBEK	−2.431*	−2.433*	−1.472*	−1.703*
	(0.229)	(0.229)	(0.334)	(0.352)
BALOCH	−0.936	−0.713	−0.421	−0.756
	(0.773)	(0.775)	(0.701)	(0.718)
HAZARA	−2.490*	−2.476*	−2.310*	−1.741*
	(0.269)	(0.269)	(0.325)	(0.331)
TAJIK	−1.454*	−1.525*	−0.538*	−0.604*
	(0.166)	(0.167)	(0.238)	(0.244)
PAMIR.TAJIK	1.254	1.237	3.627*	2.548*
	(0.832)	(0.834)	(0.909)	(0.932)
ORMURI	−0.182	−0.278	−1.866*	−1.058
	(0.739)	(0.739)	(0.746)	(0.754)
NURISTANI	−0.104	0.114	−0.404	−1.065
	(0.719)	(0.719)	(0.970)	(0.992)
logPOP	0.479*	0.530*	0.638*	0.633*
	(0.081)	(0.082)	(0.080)	(0.082)
logAREA	0.194*	0.167*	−0.004	0.019
	(0.052)	(0.053)	(0.064)	(0.066)
logROADS	0.368*	0.428*	0.540*	0.518*
	(0.111)	(0.113)	(0.106)	(0.107)
logRIVERS	−0.079	−0.047	0.018	0.072
	(0.049)	(0.051)	(0.056)	(0.057)
POST:UZBEK	1.364*	1.364*	1.365*	1.637*
	(0.122)	(0.123)	(0.123)	(0.190)
POST:BALOCH	−0.180	−0.529	−0.530	0.027
	(0.475)	(0.479)	(0.479)	(0.501)
POST:HAZARA	1.020*	0.992*	0.994*	0.252
	(0.117)	(0.117)	(0.118)	(0.167)
POST:TAJIK	0.382*	0.483*	0.485*	0.561*
	(0.057)	(0.060)	(0.060)	(0.098)
POST:PAMIR.TAJIK	−0.487	−0.471	−0.470	1.606*
	(0.256)	(0.265)	(0.265)	(0.761)
POST:ORMURI	0.500*	0.650*	0.651*	−0.533*
	(0.197)	(0.199)	(0.199)	(0.250)
POST:NURISTANI	0.076	−0.265	−0.266	0.858*
	(0.302)	(0.305)	(0.305)	(0.433)
POST:logPOP		−0.081*	−0.082*	−0.071*
		(0.022)	(0.022)	(0.032)
POST:logAREA		0.042*	0.042*	0.001
		(0.018)	(0.018)	(0.027)
POST:logROADS		−0.095*	−0.095*	−0.053
		(0.037)	(0.037)	(0.042)
POST:logRIVERS		−0.047*	−0.047*	−0.120*
		(0.018)	(0.018)	(0.025)
Constant	−6.889*	−7.303*	−11.288*	−11.044*
	(0.590)	(0.601)	(1.055)	(1.085)
N	68,382	68,382	68,382	68,382

GLMM/Poisson allowing for overdispersion, with random effects at district level

Overdispersion modelled via random effects at observation level. Column IV has province fixed effects.

Table 10: Dependent variable is total attacks in district i

	I	II	III	IV	V	VI	VII	VIII
(Intercept)	0.67*	0.26	−0.09	−1.82*	1.02*	0.45	0.31	−1.77*
	(0.28)	(0.22)	(0.28)	(0.44)	(0.34)	(0.28)	(0.37)	(0.66)
UZBEK	−1.74*	−1.85*	−1.85*	−1.24*	−2.70*	−3.27*	−3.23*	−2.64*
	(0.18)	(0.16)	(0.17)	(0.29)	(0.33)	(0.43)	(0.44)	(0.45)
BALOCH	−1.25*	−1.07	−0.88	−0.83	−1.93*	−1.50*	−1.36	−1.08
	(0.34)	(0.60)	(0.57)	(0.59)	(0.42)	(0.72)	(0.73)	(0.72)
HAZARA	−1.84*	−1.58*	−1.58*	−1.13*	−2.12*	−2.15*	−2.11*	−1.18*
	(0.23)	(0.23)	(0.23)	(0.23)	(0.43)	(0.39)	(0.40)	(0.37)
TAJIK	−1.34*	−1.42*	−1.45*	−0.70*	−1.22*	−1.40*	−1.40*	−0.61
	(0.18)	(0.19)	(0.19)	(0.26)	(0.38)	(0.44)	(0.45)	(0.39)
PAMIR.TAJIK	0.54*	0.69*	0.75*	1.84*	0.02	0.17	0.21	3.04*
	(0.24)	(0.18)	(0.25)	(0.39)	(0.33)	(0.27)	(0.37)	(0.60)
ORMURI	0.05	−0.19	−0.29	−1.04*	−0.59*	−1.00*	−1.01*	−1.47*
	(0.22)	(0.20)	(0.23)	(0.35)	(0.20)	(0.20)	(0.23)	(0.30)
NURISTANI	−0.31	−0.62	−0.45	−1.38*	−0.92	−1.27	−1.14	−3.16*
	(0.46)	(0.50)	(0.52)	(0.66)	(0.52)	(0.67)	(0.68)	(1.21)
logPOP	0.52*	0.47*	0.59*	0.67*	0.49*	0.49*	0.53*	0.68*
	(0.07)	(0.06)	(0.07)	(0.07)	(0.08)	(0.06)	(0.09)	(0.11)
logAREA	0.16*	0.15*	0.12*	−0.04	0.24*	0.24*	0.22*	0.08
	(0.06)	(0.04)	(0.06)	(0.07)	(0.07)	(0.05)	(0.08)	(0.09)
logROADS	0.36*	0.36*	0.27*	0.42*	0.59*	0.59*	0.58*	0.76*
	(0.13)	(0.09)	(0.13)	(0.11)	(0.15)	(0.12)	(0.17)	(0.15)
logRIVERS	−0.02	−0.02	0.01	0.07	−0.03	−0.03	−0.02	−0.00
	(0.05)	(0.04)	(0.06)	(0.06)	(0.08)	(0.06)	(0.09)	(0.06)
POST		−0.12	0.60	0.60		−0.25	0.08	0.21
		(0.13)	(0.41)	(0.40)		(0.16)	(0.49)	(0.41)
POST:UZBEK		0.57*	0.56*	0.56*		1.05*	0.99	0.90*
		(0.22)	(0.24)	(0.23)		(0.52)	(0.53)	(0.34)
POST:BALOCH		−0.35	−0.74	−0.74		−1.39	−1.69	−1.87
		(0.83)	(0.80)	(0.82)		(1.39)	(1.39)	(1.50)
POST:HAZARA		−0.12	−0.13	−0.13		0.08	−0.01	0.01
		(0.29)	(0.31)	(0.23)		(0.64)	(0.64)	(0.58)
POST:TAJIK		0.40	0.46	0.46*		0.39	0.37	0.40
		(0.26)	(0.26)	(0.22)		(0.56)	(0.56)	(0.32)
POST:PAMIR.TAJIK		−0.47*	−0.59	−0.59*		−0.41*	−0.51	−0.57
		(0.13)	(0.34)	(0.29)		(0.16)	(0.49)	(0.34)
POST:ORMURI		0.53*	0.74*	0.74*		0.79*	0.83*	0.84*
		(0.24)	(0.32)	(0.23)		(0.25)	(0.30)	(0.19)
POST:NURISTANI		0.69	0.35	0.35		0.72	0.42	0.51
		(0.58)	(0.62)	(0.59)		(0.77)	(0.85)	(1.09)
POST:logPOP			−0.24*	−0.24*			−0.09	−0.11
			(0.11)	(0.11)			(0.12)	(0.10)
POST:logAREA			0.05	0.05			0.05	0.05
			(0.08)	(0.07)			(0.10)	(0.09)
POST:logROADS			0.17	0.17			0.01	0.01
			(0.18)	(0.15)			(0.23)	(0.19)
POST:logRIVERS			−0.06	−0.06			−0.02	−0.01
			(0.07)	(0.06)			(0.12)	(0.07)
N	262	524	524	524	262	524	524	524

Columns I - IV use OLS with dependent variable log transformed. Column IV has province fixed effects.
Columns V - VIII use GLM/Poisson allowing for overdispersion. Column VIII has province fixed effects.
Robust standard errors in parentheses

Table 11: Estimated Organized Attacks, 2008-2009

	I	II	III	IV	V	VI
(Intercept)	-1.31^*	-1.23	-11.78	0.48^*	-4.11	-8.59
	(0.09)	(2.71)	(12.30)	(0.14)	(4.98)	(21.03)
I(ATTACKS_EARLY_ADJACENT == 0)	-0.95^*	-0.69^*	-0.49^*	-4.71^*	-4.39^*	-3.67^*
	(0.10)	(0.12)	(0.18)	(1.02)	(1.04)	(1.11)
ATTACKS_EARLY	0.78^*	0.68^*	0.69^*	0.34^*	0.28^*	0.34^*
	(0.11)	(0.11)	(0.11)	(0.07)	(0.08)	(0.13)
logPOP		0.27^*	0.24		0.54^*	0.61^*
		(0.10)	(0.13)		(0.17)	(0.23)
logAREA		0.08	0.06		0.22	0.10
		(0.07)	(0.10)		(0.14)	(0.17)
LIGHTS		-0.71^*	-0.63^*		-1.94	-1.04
		(0.27)	(0.29)		(1.18)	(1.30)
LATITUDE		-0.17^*	0.01		-0.18	-0.27
		(0.05)	(0.17)		(0.09)	(0.41)
LONGITUDE		0.05	0.10		0.08	0.14
		(0.03)	(0.17)		(0.05)	(0.29)
Provice FE	N	N	Y	N	N	Y
N	398	398	398	398	398	398
R^2	0.36	0.39	0.46			
adj. R^2	0.35	0.38	0.40			
Resid. sd	1.41	1.38	1.36			

Columns I-III use OLS with log(ATTACKS+0.1) as dependent variable

Columns IV-VI use Poisson regression with ATTACKS as dependent variable

ATTACKS_EARLY is (estimated) number of organized attacks in 2004 - 2007.

ATTACKS_EARLY_ADJACENT is (est.) # of organized attacks per capita in adj. districts in 2004-2007.

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Table 12: Estimated Organized Attacks, 2008-2009 (no attacks in 2004-2007)

	I	II	III	IV	V	VI
(Intercept)	-1.93^*	2.60	20.39	-0.93^*	14.67	106.11^*
	(0.07)	(2.38)	(12.69)	(0.43)	(7.82)	(42.20)
I(ATTACKS_EARLY_ADJACENT == 0)	-0.34^*	-0.15	-0.13	-3.29^*	-2.61^*	-1.78
	(0.08)	(0.08)	(0.12)	(1.10)	(1.15)	(1.26)
logPOP		0.10	0.02		0.62	0.02
		(0.08)	(0.09)		(0.38)	(0.52)
logAREA		0.01	0.02		-0.32	0.44
		(0.05)	(0.06)		(0.40)	(0.58)
LIGHTS		-0.29^*	-0.12		-7.43	-1.52
		(0.13)	(0.15)		(7.62)	(3.00)
LATITUDE		-0.10^*	-0.10		-0.46^*	-1.50
		(0.04)	(0.09)		(0.19)	(0.82)
LONGITUDE		-0.03	-0.28		-0.10	-0.81
		(0.03)	(0.17)		(0.11)	(0.51)
Provice FE	N	N	Y	N	N	Y
N	235	235	235	235	235	235
R^2	0.03	0.09	0.50			
adj. R^2	0.02	0.06	0.40			
Resid. sd	0.87	0.85	0.69			

Sample is districts with zero (estimated) number of organized attacks in 2004-2007.

Columns I-III use OLS with log(ATTACKS+0.1) as dependent variable

Columns IV-VI use Poisson regression with ATTACKS as dependent variable

ATTACKS_EARLY is (estimated) number of organized attacks in 2004 - 2007.

ATTACKS_EARLY_ADJACENT is (est.) # of organized attacks per capita in adj. districts in 2004-2007.

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

A Spectral Clustering Consistency

Each off-diagonal $\bar{\gamma}_{ii'}$ entry will converges to $\gamma_{ii'}$ as the number of time periods grows, and the $\bar{\Gamma}_H$ matrix will converge to Γ_H. Thus, $\bar{L}$ will converge to L. Asymptotically, the correct number of the sample eigenvalues of $\bar{L}$ will approach zero. From a theoretical perspective, a test statistic similar to that given in Yao, Zheng, and Bai [2015] could be used to determine the number of zero eigenvalues. This test statistic appears to have originated from Anderson [1963], and a simplified version appears to be appropriate in this case: the eigenvalues that are converging to zero are doing so at a $\sqrt{T}$ rate, and thus for the K smallest eigenvalues, the test statistic $\sqrt{T}\sum_{k=1}^{K}\lambda_k$ or $T\sum_{k=1}^{K}\lambda_k^2$ could be used.[48] However, the asymptotic distribution of these test statistics is not clear, and it is also not obvious that a subsampling bootstrap approach would yield the correct distribution either. Simulations suggest that here are certain cases where the correct number of groups will only be obtained with high probability when a very large number of time periods are observed. Specifically, consider the case where α_{ij} is positive but very close to zero for some i and j. That is, there are members of group j in district i, but there are very few of them. In this case $\gamma_{ii'}$ will be very close to zero for all the other i' that contain members of group j. It is thus difficult to distinguish between i containing its own separate group, and i being a part of group j. Given the difficulty of a formal test, heuristic methods are used.

The estimate $\hat{J}$ corresponds to an eigenvalue such that λ_k is "small" for all $k \leq \hat{J}$. The presence of high eigengaps on the right hand side of Figure 7 is not relevant for the eigengap procedure, as eigenvalues preceding the gaps on the right hand side of Figure 7 are "large". In particular, Luxburg [2007] suggests that the cutoff between "small" and "large" should not be larger than the minimum degree in the graph, and this is trivially met by $\hat{J} = 1$ but would be violated by any much larger estimate. Although the "eigengap" approach is intended to be heuristic rather than formal, it is possible to compare the first eigengap to simulated data where there is no group structure. Compared to data where the attacks in each district have been reassigned to a random date, the first eigengap shown in Figure 7 is larger, and this difference is statistically significant at the 95% level.

[48] The asymptotic argument is made with a fixed number of districts, N, and a growing number of time periods, T.

B NNMF Consistency

$\bar{\Gamma}_H$ will converge to Γ_H with an asymptotically normal distribution, by the Cramer-Wold device and the fact that the underlying distribution of attacks has finite fourth moments. Let $W_k = ||\hat{\Gamma}_H^k - \bar{\Gamma}_H||$, where $\hat{\Gamma}_H^k$ is the estimated covariance matrix for the model with k groups. When $k = J$, $\hat{\Gamma}_H^k$ will converge to Γ_H, and thus W_J will converge to zero. The estimated $\hat{\alpha}$ that produce $\hat{\Gamma}_H$ will be a consistent estimator for the true α so long as the standard GMM assumptions are satisfied. As is usually the case, however, the GMM identification condition is challenging to prove. Huang, Sidiropoulos, and Swami [2014] discuss uniqueness of symmetric non-negative factorizations at some length. They conclude that while there are no obvious necessary conditions to check for uniqueness, simulations reveal that multiplicity of solutions does not appear to be a problem unless the correct factorization is extremely dense: factorizations with 80% non-zero entries are still reconstructed successfully. The Γ_H matrices considered in this paper would generally be expected to have a relatively sparse factorization, so long as insurgent groups have geographic territories. One concern might be that diagonal entries has been zeroed out in Γ_H, and disregarding these entries would increase the probability of factorizations being non-unique. There is no evidence of problems with non-uniqueness, however in the results reported in Tables 5 to 7.

Additional groups will not worsen the model fit, and thus W_{J+1} will also converge to zero. For values $k < J$, W_k will converge to a positive value, so long as $\alpha_{ik'} > 0$ for at least two districts i and $k' > k$. The main difficulty is thus in selecting a threshold such that asymptotically $k = J$ will be selected instead of $k = J+1$ or $K < J$. Convergence of W_J and W_{J+1} is at the standard $\sqrt{T}$ rate, and thus any threshold that also shrinks at this rate will lead to an inconsistent estimator: this includes any the rule of thumb "one standard error" rule from Tibshirani, Walther and Hastie [2001], as the errors in the random model with no group structure will also shrink at $\sqrt{T}$ rate. The solution would be to use a threshold that shrinks to zero, but at a rate slower than $\sqrt{T}$. The probability of an incorrect selection of $k = J + 1$ or higher number of groups would then decrease to zero asymptotically, and the probability of $k < J$ being selected would similarly decrease. The asymptotic argument is theoretical, in the sense that only one data set is actually available: the "one standard error"

rule is used with it, and a hypothetical larger data set would call for a more stringent rule.

C Estimation using monthly covariance matrices

Suppose that attack probabilities are relatively small. Then the number of attacks by unorganized militants can be approximated using a Poisson($\zeta_{im}\eta\ell_i$) distribution instead of using the actual Binomial($\zeta_{im}\eta, \ell_i$) distribution. Similarly, the distribution of attacks by members of an organized group can be approximated with Poisson($\zeta_{im}\epsilon_{tj}\alpha_{ij}$) in place of Binomial($\zeta_{im}\epsilon_{tj}, \alpha_{ij}$).

Now, suppose that there are a total of x_{im} attacks in district i. Conditional on there being a total of x_{im} attacks, the distribution of these attacks across days is given by a Multinomial(x_{im}, p_i) distribution, where p_i is a probability vector with elements of the form

$$p_{it} = \frac{\eta\ell_i + \sum_j \epsilon_{tj}\alpha_{ij}}{\sum_{t'} (\eta\ell_i + \sum_j \epsilon_{t'j}\alpha_{ij})}$$

If in some other district i' there were $x_{i'm}$ attacks, then the covariance of daily attacks has the useful form

$$\begin{aligned}
\text{Cov}(x_{im\cdot}, x_{i'm\cdot}) &= x_{im}x_{i'm}\sum_t p_{it}p_{i't} - \frac{x_{im}}{T}\cdot\frac{x_{i'm}}{T} \\
&= x_{im}x_{i'm}(\sum_t p_{it}p_{i't} - \frac{1}{T}\cdot\frac{1}{T}) \\
\frac{\text{Cov}(x_{im\cdot}, x_{i'm\cdot})}{x_{im}x_{i'm}} &= \text{SCov}(p_{it}, p_{i't})
\end{aligned}$$

where $\text{SCov}(p_{it}, p_{i't})$ gives the sample covariance for a given draw of ϵ. The first line of the above holds because each attack decision is independent given both the total number of attacks and the realization of ϵ. If the ϵ are constructed such that $\sum_{t'} \epsilon_{t'j} = 1$, then the denominator in the expression above for p_{it} will simplify such that

$$\text{SCov}(p_{it}, p_{i't}) = \frac{\sum_j \alpha_{ij}\alpha_{i'j}\sigma_j^2}{(T\eta\ell_i + \sum_j \alpha_{ij})(T\eta\ell_{i'} + \sum_j \alpha_{i'j})}$$

If the distribution of ϵ conditional on the number of attacks is the same as the unconditional distribution of ϵ, then the above will hold because the number of attacks is a sufficient statistic (if the ϵ are independent of the number of attacks?). The $T\eta\ell_i + \sum_j \alpha_{ij}$ term can

be taken to be the "average" number of attacks, which implies that $\tilde{\alpha}_{ij} = \frac{\alpha_{ij}}{T\eta\ell_i + \sum_j \alpha_{ij}}$ is the fraction of attacks in district i that group j will be responsible for. Then

$$\text{Cov}(p_{it}, p_{i't}) = \sum_j \tilde{\alpha}_{ij}\tilde{\alpha}_{i'j}\sigma_j^2$$

Here $\tilde{\alpha}$ and σ^2 are not separately identified. If the normalization $\sigma_j^2 = 1$ is used, then the estimated $\tilde{\alpha}$ describe relative degrees to which groups are more or less responsible for attacks, across districts.

REFERENCES

[1] **Anderson, Carl A.** (1974) "Portuguese Africa: A Brief History of United Nations Involvement" *Denver Journal of International Law & Policy* 133

[2] **Anderson, T.W.** (1963) "Asymptotic Theory for Principal Component Analysis" *Annals of Mathematical Statistics* 122-148.

[3] **Ashford, J.R. and R.G. Hunt** (1973) "The Distribution of Doctor-Patient Contacts in the National Health Service" *Journal of the Royal Statistical Society Series A* 137 (3), 347-383.

[4] **Benmelech, Efraim, Claude Berrebi, and Esteban F. Klor. (**2012). "Economic Conditions and the Quality of Suicide Terrorism." *The Journal of Politics* 74 (1): 113–128.

[5] **Berman, Eli** (2009). *Radical, Religious and Violent: The New Economics of Terrorism.* MIT Press.

[6] **Berman, Eli, Joseph H. Felter, Jacob N. Shapiro,** (2011) Can Hearts and Minds Be Bought? The Economics of Counterinsurgency in Iraq. *Journal of Political Economy* Vol. 119, No. 4: 766-819

[7] **Birtle, Andrew J.** (2008). "Persuasion and Coercion in Counterinsurgency Warfare." *Military Review* (July-August): 45-53.

[8] **Blair, Graeme, C. Christine Fair, Neil Malhotra, Jacob N. Shapiro** (2012) "Poverty and Support for Militant Politics: Evidence from Pakistan". *American Journal of Political Science* 57(1): 30-48

[9] **Blattman, Christopher and Edward Miguel** (2010) "Civil War" *Journal of Economic Literature* 2010, 48:1, 3–57

[10] **Boix, Carles** (2008) Civil Wars and Guerrilla Warfare in the Contemporary World. Toward a Joint Theory of Motivations and Opportunities. In Stathis Kalyvas, Ian Shapiro and Tarek Masoud, ed., *Order, Conflict and Violence*. Cambridge University Press. Chapter 8, pages 197-218.

[11] **Bueno de Mesquita, Ethan.** (2013). "Rebel Tactics." *Journal of Political Economy* 121 (2): 323–357

[12] **Bueno de Mesquita, Ethan, and Eric S. Dickson.** (2007). "The Propaganda of the Deed: Terrorism, Counterterrorism, and Mobilization." *American Journal of Political Science* 51 (2): 364–381.

[13] **Callen, Michael, Nils B. Weidmann** (2013) Violence and Election Fraud: Evidence from Afghanistan. *British Journal of Political Science* 43(1): 53-75

[14] **Collier, Paul, and Anke Hoeffler** (2004). "Greed and Grievance in Civil War." *Oxford Economic Papers*, 56, 563-595.

[15] **Collier, P. and Rohner, D.** (2008), Democracy, Development, and Conflict. *Journal of the European Economic Association*, 6: 531–540.

[16] **Condra, Luke Joseph H. Felter, Radha Iyengar, Jacob N. Shapiro,** (2010) The Effect of Civilian Casualties in Afghanistan and Iraq. N*BER Working Paper* 16152.

[17] **Condra, Luke N., Jacob N. Shapiro,** (2012) Who Takes the Blame? The Strategic Effects of Collateral Damage. A*merican Journal of Political Science* Vol. 56, No. 1: 167-187.

[18] **Deloughery Kathleen** (2013) Simultaneous Attacks by Terrorist Organisations. *Perspectives on Terrorism*, 7(6): 79-90.

[19] **Ding, C., He, X., and Simon, H.** (2005) On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering *Proceedings of the Fifth SIAM International Conference on Data Mining*, 606-610.

[20] **Dorronsoro, Gilles** (2009) The Taliban's Winning Strategy in Afghanistan. *Carnegie Endowment for International Peace Paper.*

[21] **Dorronsoro, Gilles** (2012) Waiting for the Taliban in Afghanistan. *Carnegie Endowment for International Peace Paper.*

[22] **Drozdova, Katya,** (2012). Divide and COIN: Evaluating Strategies for Stabilizing Afghanistan and the Region APSA 2012 Annual Meeting Paper.

[23] **Eisenstadt, Michael and Jeffrey White** (2005) "Assessing Iraq's Sunni Arab Insurgency" *The Washington Institute for Near East Policy Policy Focus* No.50.

[24] **Fearon, James D.** (2007) Iraq's Civil War. *Foreign Affairs* 86(2):2-16.

[25] **Fearon, James** (2008) "Economic development, insurgency, and civil war" in *Institutions and Economic Performance*, ed. Elhanan Helpman, Harvard University Press

[26] **Fearon, James D. and David D. Laitin.** (2003) Ethnicity, Insurgency, and Civil War. *American Political Science Review* 97(1):75-90.

[27] **Fernandes, Clinton** (2008) *Hot Spot: Asia and Oceania.* ABC-CLIO

[28] **Fotini, Christia, Semple, Michael** (2009) "Flipping the Taliban- How to Win in Afaghanistan" *Foreign Affairs,* 88, 34-45

[29] **Ghobarah, Hazem Adam, Paul Huth and Bruce Russett.** (2003) Civil Wars Kill and Maim People Long After the Shooting Stops." *American Political Science Review* 97(2):189-202.

[30] **Giustozzi, Antonio.** *Koran, Kalashnikov and Laptop: The Neo-Taliban Insurgency in Afghanistan*, Hurst & Company, London, 2007.

[31] **Giustozzi, Antonio (**2009). "The Pygmy who turned into a Giant: The Afghan Taliban in 2009", LSE mimeo.

[32] **Good, Phillip.** (2002) Extensions of the Concept of Exchangeability and their Applications. *Journal of Modern Applied Statistical Methods.* 1(2) 243-247.

[33] **Good, Phillip.** (2005) *Permutation, Parametric, and Bootstrap Tests of Hypotheses.* New York: Springer.

[34] **Grossman, Herschel I.** (1991) A General Equilibrium Model of Insurrections. *American Economic Review* 81(4):912-21.

[35] **Grossman, Herschel I.** (2002) Make Us a King: Anarchy, Predation, and the State. *European Journal of Political Economy* 18:31-46.

[36] **Gutierrez-Sanin, Francisco**. (2008) Telling the Difference: Guerrillas and Paramilitaries in the Colombian War. *Politics and Society* 36(1):3-34.

[37] **Hastie, T., Tibshirani, R., and Friedman, J.** (2001). *The elements of statistical learning.* New York: Springer.

[38] **Henderson, Anne.** (2005) The Coalition Provisional Authority's Experience: with Economic Reconstruction in Iraq: Lessons Identified. *USIP Special Report* No. 138. http://www.usip.org/files/resources/sr138.pdf

[39] **Hirshleifer, Jack** (1991) The Technology of Conflict as an Economic Activity. *American Economic Review*, Vol. 81, No. 2, pp. 130-134

[40] **Hirshleifer, Jack** (1995a) Anarchy and Its Breakdown. *Journal of Political Economy* 103(1):26-52.

[41] **Hirshleifer, Jack** (1995b) Theorizing about conflict. *Handbook of defense economics*, Elsevier.

[42] **Hirshleifer, Jack** (2001) *The dark side of the force: Economic foundations of conflict theory.* Cambridge University Press.

[43] **Hovil, Lucy and Eric Werker.** (2005) Portrait of a Failed Rebellion: An Account of Rational, Sub-Optimal Violence in Western Uganda. *Rationality and Society* 17(1):5-34.

[44] **Huang, K., Sidiropoulos, N., and Swami, A.** (2014) Non-Negative Matrix Factorization Revisited: Uniqueness and Algorithm for Symmetric Decomposition. *IEEE Transactions on Signal Processing* 62(1):211-224.

[45] **Humphreys, Macartan.** (2005) Natural Resources, Conflict, and Conflict Resolution: Uncovering the Mechanisms. *Journal of Conflict Resolution* 49(4):508-537.

[46] **Karlis, Dimitris and Evdokia Xekalaki.** (2005) Mixed Poisson Distributions *International Statistical Review* 73(1):35-58.

[47] **Kilcullen, David** (2009) *The accidental guerrilla: Fighting small wars in the midst of a big one* Oxford University Press.

[48] **Krepinevich, Andrew** (2005) "How to Win in Iraq". *Foreign Affairs*, September/October.

Kriegel, H.-P.; Kröger, P., Zimek, A. (2009). "Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering". *ACM Transactions on Knowledge Discovery from Data*. 3 (1): 1–58.

[49] **Leites, Nathan and Charles Wolf.** (1970). Rebellion and Authority. Chicago, IL: Markham.

[50] **Luxburg, Ulrike von** (2007) "A tutorial on spectral clustering" *Statistics and Computing* Volume 17, Issue 4, pp 395-416

[51] **Luxburg, Ulrike von**, **Mikhail Belkin AND Olivier Bousquet,** (2008) "Consistency of Spectral Clustering" *The Annals of Statistics*, Vol. 36, No. 2, pp 555–586

[52] **Mohajer, M., Englmeier, K., and Schmid, V.** (2010). "A comparison of Gap statistic definitions with and with-out logarithm function". *Technical Report*. Department of Statistics, University of Munich. 096.

[53] **Ng, A. Y., Jordan, M., &Weiss, Y.** (2002). On spectral clustering: Analysis and an algorithm. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems*, 14. Cambridge, MA: MIT Press.

[54] **O'Loughlin, John, Frank Witmer, and Andrew Linke** (2010a) "The Afghanistan-Pakistan Wars 2008–2009: Micro-geographies, Conflict Diffusion, and Clusters of Violence" *Eurasian Geography and Economics*, 51 No.4, pp.437-71.

[55] **O'Loughlin, John, Frank Witmer, Andrew Linke, and Nancy Thorwardson.** (2010b) "Peering into the Fog of War: The Geography of the WikiLeaks Afghanistan War Logs 2004-2009" *Eurasian Geography and Economics*, 51 No.4, pp.472-95.

[56] **O'Neill, Bard** (1990) *Insurgency and Terrorism, Inside Modern Revolutionary Warfare*, Dulles, VA.: Brassey's Inc.

[57] **Pesarin, Fortunato** (2001) *Multivariate Permutation Tests*, New York: Wiley.

[58] **Potiron de Boisfleury, Gregoire** (2010) *The origins of Marshal Lyauteys pacification doctrine in Morocco from 1912 to 1925*, Master's Thesis, US Army Command and General Staff College.

[59] **Schelling, Thomas C.** (1960) *The Strategy of Conflict.* Cambridge: Harvard University Press.

[60] **Shi, J. and Malik, J.** (2000). "Normalized cuts and image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8), 888 – 905.

[61] **Subrahmanian, V. S., Aaron Mannes, Animesh Roul, R. K. Raghavan** (2013) *Indian Mujahideen: Computational Analysis and Public Policy*, Springer.

[62] **Thruelsen, Peter Dahl** (2010) "The Taliban in southern Afghanistan: a localised insurgency with a local objective" *Small Wars & Insurgencies*, Volume 21, Issue 2, pp.259-276

[63] **Tibshirani, R., Walther, G., and Hastie, T.** (2001) "Estimating the number of clusters in a data set via the gap statistic". *J. R. Statist. Soc. B*, 63, Part 2, 411-423.

[64] **Tullock, Gordon** (1974) *The Social Dilemma*, Blacksburg: Center for the Study of Public Choice, VPISU Press.

[65] **United Nations** (2013) *Third report of the Analytical Support and Sanctions Monitoring Team, submitted pursuant to resolution 2082 (2012) concerning the Taliban and other associated individuals and entities constituting a threat to the peace, stability and security of Afghanistan.* S/2013/656

[66] **N. Vasiloglou, A. Gray, and D. Anderson** (2009) Non-Negative Matrix Factorization, Convexity and Isometry". *Proc. SIAM Data Mining Conf.*, 673-684.

[67] **Yao, J., Zheng, S., and Bai, Z.** (2015) *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press.