Insurgency and Small Wars: Estimation of Unobserved Coalition Structures

Francesco Trebbi and Eric Weese*

April 2015

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 politico-economic determinants. Through the use of detailed geocoded incident-level data available from the United States Worldwide Incidents Tracking System (WITS) we present a novel methodology to the study of the economics of insurgency and provide an application in the context of the Afghan conflict during the 2005-2009 period. We prove statistically that the Afghani Taliban are not an umbrella coalition, rather a highly unified group and show how their span of control grew substantially post-2007 beyond ethnic Pashtun areas.

*University of British Columbia, Vancouver School of Economics, CIFAR and NBER, francesco.trebbi@ubc.ca; Yale University, eric.weese@yale.edu respectively. The authors would like to thank Ethan Bueno de Mesquita 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.

1 Introduction

Insurgency is typically defined as armed rebellion against a centralized or national authority[1]. Among the various forms of armed conflict, insurgency is possibly one of the most opaque, as intertwining connections with the population blur the lines between combatants and civilians (Kilcullen, 2009). In insurgencies and guerrilla warfare the relative strength and even the identity of potential negotiating counterparties are unsure, and in the words of Fearon (2008) "*there are no clear front lines*". Nonetheless, insurgency and guerrilla conflict have exerted enormous socioeconomic costs and in the post-World War II era they rank among the most detrimental and perduring forms of internal conflict[2] and of political violence (O'Neill, 1990). Our paper offers a novel contribution to the empirical analysis of these asymmetric irregular wars.

As a way of more precise motivation, consider the U.S.-led counterinsurgency operation in Afghanistan[3]. Over the 2001-2011 period and on the U.S. side alone, this operation cost the lives of more than $1,800$ troops and more than $444 billion in military expenses. Uncounted scores of Afghan citizens were also extremely adversely affected. Soon into the operation, the U.S. military acknowledged through drastic adjustment in tactics that the Afghan armed conflict presented complex differences from previous large scale military operations. Fighting an established alliance between the Afghan Taliban insurgents and the al-Qaeda terrorist organization, front lines appeared blurred and even the identity of a unified operating adversary seemed doubtful. There was (and still is) open disagreement among experts on whether the Taliban were (or currently are) a unified fighting organization rather than an umbrella coalition of heterogeneous forces. For instance, the extent of the organizational control of Taliban leader Mullah Mohammed Omar over the powerful Haqqani faction and the Dadullah network is frequent subject of discussion[4]. Similarly, the Hizb-i Islami

[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] Afghanistan is also going to be the object of the paper's main application. Table 1 includes for ease of reference a summary of the US Afghan counterinsurgency timelined produced by the Council of Foreign Relations.

[4] Note for instance the UN report (2013) stating that *"Despite what passes for a zonal command structure*

faction is considered by many a separate entity from the Taliban proper[5]. In an insightful, yet qualitative essay Dorronsoro (2009) discusses precisely how "*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.*"[6] Ultimately, this is unsettling: Understanding the extent of territorial control and population support of insurgent groups are essential not just to military operations, but for our understanding of the internal organization of rebel groups, assessing their cohesion, preventing selective violence by insurgents, and ultimately inform any effort of reconstruction of areas affected by conflict.

This paper shows how, by focusing on specific events of armed violence, one can recover the number and extent of different insurgent groups in activity on a specific territory, features of the data that are typically unobservable to the econometrician/analyst and essentially latent to the conflict. This objective is achieved through inference from the co-occurrence of violent events over time across different areas, under the working assumption that only an insurgent group with foothold in two separate areas can jointly carry out attacks in both areas simultaneously and on a repeated basis. Once the number of different guerrilla groups and their territorial extent has been statistically ascertained, we assess its main empirical determinants and its economic and social consequences for specific areas and for the non-combatant population. We also produce an analysis of shifts in insurgent presence over time.

The paper aims at addressing four broad questions: 1. When faced with multiple violent incidents in multiple regions, is it possible to identify whether specific incidents are isolated idiosyncratic events as opposed to organized attacks by coalitions of assailants? 2. Is it

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."

[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 stark contrast, the Pakistani Taliban are indicated in the same essay as a clearly non-unitary umbrella organization.

possible to identify from incident data alone how many separate and different insurgent groups (if any) are attacking? 3. What are the economic determinants driving the diffusion and segmentation of the rebels within a specific region and across regions? 4. What are the consequences of insurgent control on civilian population wellbeing?

To summarize the estimation methodology in more detail, we can think of a territory of conflict as a grid of points over which at any moment in time violent incidents can happen. A point on the grid represents the stylized district (its centroid). The data generating process underlying the incident occurrence is stylized. This is not without loss of generality, but lacking detailed data on the internal organization and planning strategy of insurgents and counterinsurgency forces, one has to remain parsimonious. Our main working assumption is that attacks in two different districts can occur in the same day if an insurgent group is able to operate in both and finds it worthwhile to carry out such simultaneous missions[7]. Deloughery (2013) reviews extant studies and presents systematic evidence of 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]. With some fairly general assumptions on the time covariance across districts, we detect which sets of districts tend to systematically co-move over time (i.e. repeatedly experience attacks in the same days) and, from this, infer the set of districts in which each guerrilla group operates. The estimators we present are flexible and allow for the same district to be under dispute by many guerrilla groups or by none. The outcome is an estimated number of guerrilla groups and, for each group, a measure of their geographic span of control, and intensity of presence in a district.

The main empirical results of the paper are as follows. We conclude that insurgent

[7]Alternatively, simultaneous incidents may simply happen occasionally by chance, due to random unorganized violence. If it is not worthwhile to carry out simultaneous attacks for a multi-district insurgent coalition, the estimator is bound to reject any coalition.

[8]The tragic events of 9/11 in the United States are also testament to the salience of such simultaneous attacks. In fact, simultaneous attacks and suicides have been a trademark of international jiadist organizations and of al-Qaeda in particular. This makes the approach more suited to the Afghan insurgency case. Other examples abound. In southern Thailand insurgent movements 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, p.258) . The Indian Mujahideen, responsible for the 2008 Mumbai attacks, typically carry out simultaneous attacks (Subrahmanian et al., 2013, ch. 6 on Simultaneous and timed attacks). Kurdish independentists and the Tamil Tigers are also known to have adopted simultaneous attacks.

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. We then consider changes in the extent of this group 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 (following a specific "oil spot" strategy). We also find that there may have been some increase in the support of non-Pashtun ethnic groups for the insurgency; however, this result is somewhat dependent on the econometric specification used.

An increasing amount of attention as been devoted within the fields of development economics and political economy to the study of internal armed conflicts within countries, including prominently civil wars and insurgency. Indeed, economists have long been interested in the analysis of violence and conflict, at least dating back to the theoretical work of Schelling (1960), Tullock (1974), Hirshleifer, (1991,1995a,b, 2001) and Grossman (1991, 2002). Even more emphatically, political scientists have dedicated to the study of conflict a large part of their work in the field of international relations.

Precisely from political science and economics some of the most recent and novel insights in the study of insurgency have emerged (Berman, 2009; Berman et al., 2011; Condra et al., 2010; Blair et al., 2012; Condra and Shapiro, 2012; Cullen and Wedmnan, 2013; Bueno de Mesquita, 2013). As underlined by Blatman and Migueal (2010), most remarkable in this most recent wave of research have been a strong empirical inclination and an increasing attention to micro-level (typically incident-level) information. The use of precisely geocoded micro data has been a point of departure relative to more established "macro" empirical approaches, based on country level information or aggregate conflict information (for notable instances, see Fearon and Laitin (2003), Boix (2008), Collier and Hoeffler (2004), Collier and Rohner (2008), among the many).

This paper follows in these footsteps with a specific emphasis on the analysis of insurgency and small wars. Indeed, we do not address conventional warfare. In the way of motivation for this specific choice, much less is known about non-conventional warfare (and its consequences on civilian populations) than what is known about wars among nations. Economic or statistical evidence on the role of anti-government guerrilla activities is still

sparse, even as such activities cause much damage worldwide and appear quantitatively to be the predominant form conflict in civil wars since 1945 (Fearon, 2008, Ghobarah et al., 2003). Insurgents' strategies are not generally well understood nor the subtleties of their interactions with noncombatant population (Gutierrez-Sanin, 2008; Kilcullen, 2009). Finally, insurgent activity is also often linked to terrorist activities and its economic study directly connects with a similarly growing literature on the economics of terrorism (Bueno de Mesquita and Dickson, 2007; Benmelech, Berrebi, Klor, 2012).

The paper is organized as follows. Section 2 develops our methodology for the estimation of coalition structures among insurgent groups. We describe our data, particularly the Empirical Studies of Conflict Project incident-level Afghan data, which have been generously made publicly available in Section 3. The analysis of the determinants of the insurgent coalition and its consequences is developed in Section 4. Section 5 concludes.

2 Econometric Model

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 based on our econometric model.

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 Afghan districts. In the base model presented below (which will be based on spectral clustering), it is assumed that there is at most one organized group in any given district. A modified model (based on non-negative matrix factorization) will also be presented that can accommodate several organized insurgent groups in a given district.

Let organized insurgent groups be indexed by j, and let J be the total number of such organized groups. Both our approaches allow for there to be no organized groups in a given district, and also allow for the possibility that there is no observed inter-district structure in the attack data.

Specifically, suppose that observed attacks may be initiated either by unorganized local militants, or by local members of an organized group. Let ℓ_i be the number of unorganized local militants in district i. Let α_{ij} be the number of members in district i of organized group j. Initially it will be assumed that for any given i, $\alpha_{ij} > 0$ for at most one j, but this will be relaxed below.

Let time be discrete and indexed by t. In the Afghan data, the time periods used will be days. 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 across time 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.

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 σ^2, which will indicate the variance of ϵ. 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 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, for 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 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 obvious way to distinguish between these two elements.[9] Thus, for estimating group structure, only the off-diagonal entries of the covariance matrix will be used. As a normalization, 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 sample 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 behaviour of insurgent groups. The model is not without loss of generality, but lacking detailed data on the internal organization and planning strategy of insurgents and counterinsurgency forces, a parsimonious model seems most appropriate. 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 (for example a draw of 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

[9]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 $[0, b]$, then 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.

covariances γ are generally positive, and qualitative research also suggests that a model in which there is not substantial substitution in attacks across districts is most appropriate.[10]

2.1 Base model

Estimates $\hat{\alpha}_{ij}$ are desired: these give the estimated number of members in district i of group j. An object of particular interest, however, is also J, the number of organized insurgent groups. J is an integer, and estimation strategies for this sort of parameter do not typically yield confidence intervals of the sort that would be typical for a continuous parameter. While the model described above does not appear to correspond exactly to any discussed previously in the literature, it is close enough to problems addressed by spectral clustering and non-negative matrix factorization. These two approaches can be used to produce estimates $\hat{J}$ and $\hat{\alpha}$.

An approach based on spectral clustering will be used as the base estimation strategy, while other techniques will be considered in following subsections. It is difficult to determine the properties of the estimator based on spectral clustering, but there is a substantial statistics literature related to the later approach based on non-negative matrix factorization. This literature relies heavily on bootstrap simulations, but produces only point estimates and not accompanying confidence intervals. Confidence intervals for $\hat{\alpha}$ and $\hat{J}$ are thus not reported below, but the bootstrap simulations suggest that some important null hypotheses can be rejected.

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 application 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, but spectral clustering is well suited for this sort of data structures through the addition of an explicit dimensionality reduction phase in its design. Estimation via

[10]There is substantial evidence of the strategic role of simultaneous and timed attacks initiated by insurgents (Deloughery, 2013). The evidence reported in the literature on international jihadist movements and the tactics used by insurgents in Afghanistan supports our assumption. So does evidence from insurgencies across Asia and Africa (Subrahmanian et al. 2013; Fernandes, 2008; Anderson, 1974). In addition, attack data for Afghanistan is available at a daily frequency. As will be shown by later bootstrap simulations, the large number of time periods available makes it possible to reject at reasonable confidence levels the possibility that the observed structure of attacks is due purely to random variation.

spectral clustering requires an additional assumption that is relaxed below: specifically, 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 present in any given district j.

Based on this assumption, reordering of the districts i allows Γ_H to be written as a block-diagonal matrix:

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

where there are a total of J organized groups, and each Γ_H^j has the form given in Equation 1. The value of individual matrix entries $\gamma_{ii'}$ is essential for estimating the degree of group presence in each district. Here Γ_H corresponds to the adjacency matrix for a weighted undirected graph.[11]

To perform spectral clustering, a technique following following Shi and Malik [2000] will be used.[12] This technique is based on a "graph Laplacian" matrix, which is constructed from the adjacency matrix: 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.

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

[11] 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.

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

Laplacian corresponding to each Γ_H^j block are linearly dependent. Γ_H^j is full rank: each row of Γ^j is a vector of 1s with a single 0 in the diagonal. The transformation to the graph Laplacian reduces the rank of each block by one. Thus, the reduction in rank of the overall graph Laplacian, relative to the intial Γ_H will be equal to the number of blocks, which is the number of organized groups. This is equivalent to the number of zero eigenvalues because this is the dimension of the nullspace.

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 examine the eigenvalues of this matrix. Clustering is thus feasible, because it is based on statistics from the observed sample. estimator using $\bar{L}$ is a consistent estimator for the clusters that would be obtained using the true graph Laplacian L. Further details are provided in Appendix A.

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 where the true $\gamma_{ii'}$ is positive. 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 of 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 and we experimented with several different cutoffs. 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 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, 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.[13] The difference $\lambda_{k+1} - \lambda_k$ is 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 are 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.[14] 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.

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

[14] For a standard reference here, see Hastie, Tibshirani, and Friedman (2009).

While it is relatively straightforward to obtain a consistent estimate for J, the total number of organized insurgent groups, a consistent estimate for α_{ij} is more challenging. The main difficulty here is that 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 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. 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 in 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. 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. 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 approximation 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 do not introduce substantial noise into estimates $\hat{\alpha}_{i'j}$ for other districts i'. Developing an unbiased estimator that also posessessuch properties appears to be a non-trivial undertaking.

To summarize, for the base model the specific estimator used is the following. A graph Laplacian $\bar{L}$ is calculated based on observed attack data. 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 these 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 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 Non-Negative Matrix Factorization Model

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 NJ parameters to be estimated: one $\hat{\alpha}_{ij}$ for each district i and group j.[15] A necessary condition for

[15]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.

identification is thus that $(N-1)/2 \geq J$.[16] In the data used the number of districts is large relative to plausible numbers of groups, and thus this inequality holds strictly and a penalty function is required. The estimator used for the α_{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

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

The major difficulty with implementing this estimator is that N is large. Thus, even when considering only a small number of groups, 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.[17] 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 datasets 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

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

[16] 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.

[17] 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.

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 is generated by randomly rearranging the time indices of the observations for each district. As this is done independently for each district, the result is data with a covariance matrix that is solely the result of random variation.

The estimated number of clusters $\hat{J}$ is selected to be the smallest k such that

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

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. 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 technique just described assumes 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 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. Aggregrating 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 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 α. In reality, ζ is unobserved; however, dividing by the observed number of attacks creates a feasible estimators, with α identified up to scale. This approach can be used with both with 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), which "*identifies, compiles, and analyzes micro-level conflict data and information on insurgency, civil war, and other sources of politically motivated violence worldwide.*"[18] The ESOC data currently reports a location, date, and type for violent incidents from the beginning of 2003 to the end 2009. This data is based on the Worldwide Incidents Tracking System (WITS), a U.S. government military database[19]. 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 dataset are episodes of violence initiated by insurgents, or acts of random violence. The data does not include violence directly connected to a military counterinsurgency operation, such as for instance a U.S. military attack on a Taliban safe house.

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

[19] "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. 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 $T = 2082$.

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 with the district-level attack data. ESOC does not report information on the distribution of ethnicities 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) dataset of the Swiss Federal Institute of Technology Zurich[20].

In Figure 1 we report the ethnic distribution map by district and in Figures 3 and 4 the incident distribution map by district for the years 2004-2007 and 2008-2009 respectively (in per capita terms). Table 1 includes for ease of reference a summary of the US Afghan counterinsurgency timelined produced by the Council of Foreign Relations. Table 2 includes summary statistics for total incidents, ethnic fragmentation, roads, rivers, and settlements by district. Figure 2 shows the attacks observed in the data, by district in per capita terms. 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.

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

An analysis by the methods developed above, however, reveals some additional patterns that are not immediately obvious from an inspection of the raw data.

4 Results

Figure 5 shows the eigenvalues obtained by using the spectral clustering approach described above 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 6 shows the eigengaps for these eigenvalues. The first eigengap is the largest by a substantial margin, suggesting that small random perturbations would not change the estimated number of groups. Figure 7 shows the presence of organized group members based on the approximation given in Equation 4. The units reported in the figure are normalized so that 0 corresponds to no attacks being attributable to organized group members, and 1 corresponds (approximately) to all attacks being attributable to members of an organized group.[21] Table 3 shows regression results based on the approximation given in Equation 4. Most of these are unsurprising: ethnicities other than Pashtun (the omitted ethnicity) are generally less likely to be associated with organized group activity, while there is more group activity in districts with more roads.

As in Figure 2, which shows raw attacks, Figure 7 shows that organized attacks are concentrated in Pashtun areas, and also near the main highway passing through Kabul and other cities. A feature that is apparent in Figure 7, however, that does not show up clearly in the raw attack data of Figure 2 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.[22] This area is not as heavily populated by Pashtuns, and perhaps because of this the number of total attacks is not as high. An investigation of the attack data via spectral clustering, however, reveals that the attacks that did occur appear to exhibit substantial coordination.

The major 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"

[21] Values less than 0 or greater than 1 can be obtained because of finite sample variation. A small number of these occur in Figure 7, and are reflected in the legend.

[22] Table 5 shows regression results based on total attack data, using the same specifications as Table 3. The results here are similar.

local militants), rather than multiple groups, and that this insurgent group is active both to the north of Kabul as well as to the south. The eigengap analysis suggests that the conclusion regarding the number of groups would not change under small perturbations of the data; however, this check is heuristic in nature. An analysis is thus performed using technique based on non-negative matrix factorization outlined in Sections 3.2 and 3.3, as this method involves bootstrap simulations.

Table 12 shows the results of the Tibshirani, Walther and Hastie (2001) gap statistic procedure, using the estimation approach based on non-negative matrix factorization. Columns I and III use exactly the data used for the analysis by spectral clustering, where districts with fewer than 3 attacks were excluded. Columns II and IV use data from all districts, but with a penalty function that weights the penalty for each $\gamma_{ii'}$ entry proportional 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.[23] The "Pakistan" column replaces the Afghan attack data with roughly comparable data from Pakistan: the results here differ markedly from those presented in the first four columns. 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 structure.

Figure 8 shows the estimated prevalence of organized insurgents. As both the numerator and the denominator here are subject to random variation, there is substantial noise due to finite sample variation; however, the general pattern appears to agree with the qualitative description of insurgent activity just given for the spectral clustering method, and shown in Figure 7. The estimates from the non-negative matrix factorization method appear to make it slightly clearly 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.

Estimates of the prevalence of the organized group can also be produced using the method in Section 3.3. As the estimates in this case effectively control for variation across months,

[23] 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.

estimates appear slightly noisier.[24] Figure 9 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 7 and 8. The main benefit of using the method outlined in Section 3.3 appeared in Table 5: it 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.

One potential problem with the above analysis is that the observed results are due to an external event that causes widespread protests and many attacks. In the context of the model, this would appear to suggest a set of coordinated attacks, when in reality the attacks were by independent actors who were merely responding to the same event. The most obvious case of this in the Afghan data is the 2009 elections, which led to 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 when the attacks around the time of the 2009 elections were excluded. A reanalysis of the data stopping at July 2009 (before the August 2009 elections) however, gives the same results for the gap statistic analysis shown in Table 5: the observed pattern of attacks is best represented by one coordinated group. Estimates of the prevalence of organized insurgents are qualitatively similar to Figures 7 and 8.

4.1 Changes in group structure across time

The econometric model outlined in Section 3 assumes that the extent and prevalence of the organized insurgent group remains constant across time. 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 dividing the data into two. Create an "early" data set, including only attacks in 2004-2007, and a "late" data set, including only attacks in 2008-2009. Estimates of the prevalence of organized insurgents from the ealier data can then be compared to estimates from the later data, yielding a description of how the location of insurgent groups has changed over time.[25] The average number of attacks

[24] Calculation of correct standard errors for these estimates appears challenging.

[25] 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.

per day is substantially higher in the later period compared to the earlier one: Figure 10 shows the total number of attacks estimated to be due to organized insurgent groups in the earlier period, while Figure 11 shows this for the later period. The colors of the figures are aligned so that the same colour indicates the same number of attacks per capita per year, although the "early" and "late" data have a different number of months.

An interesting feature of these figures is that attacks in the later period mainly occur in districts in which there were attacks in the earlier period, or districts adjoining districts where there were attacks. For example, there were no attacks in the central part of Afghanistan, or much of the northeast, and these areas similarly do not have any attacks in the later period. Figure 12 shows in blue the districts where the indicator variable is coded to be zero. All but one of these are not estimated to have any organized attacks due to insurgent group members in the later period: this can be seen by comparing Figure 11 with Figure 12.

Table 8 shows that this qualitative pattern is statistically significant. The basic specification used here is

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

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?[26] Columns I-III of Table 8 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 econometric model presented in Section 3, there should never be a negative number of attacks attributed to organized group members. Columns IV-VI present the same analysis using a Poisson model, in order to take this into account. An additional advantage of the Poisson model is that districts with few attacks are (correctly) treated as having

[26]This is because there is a long-standing problem in the analysis of spatial data regarding how to use this type of "adjacent observations" data, and there does not appear to be a satisfactory solutin in this case.

higher variance.[27] 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 "almost" exhibits 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.[28]

With respect to the distribution of attacks across districts, Figure 10 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 11 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 there are 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. Table 4 shows that areas with non-Pashtun ethnicities appeared to exhibit relatively greater activitiy in the later period, although this is only statistically significant in the case of the Uzbeks.[29]

[27]This could also have been obtained using weighted least squares of some sort, but the Poisson model is natural here, 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.

[28]As an additional test, Table 9 repeats the regressions in Columns I-VI of Table 8 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.

[29]Table 7 shows similar results using a specification where each entry in the covariance matrix is treated as its own observation, rather than summing across rows: some additional coefficients are statistically significant with this specification. Table 6 shows that roughly similar results are obtained using data based on total attacks.

5 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, that is instances of attacks led by either insurgency or counterinsurgency participants (often deadly and highly damaging). 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 (see Berman, Shapiro, and Felter, 2011 for an example).

This paper shows how incident-level data contains useful information on the coalition structure and the geographic span of influence of insurgent groups and how this can be assessed systematically. We present a class of econometric methods useful for detecting unobserved insurgent coalition structures employing the "excess covariance" in terms of violent incident co-occurrences across regions and over time. Intuitively, if incidents in districts i and i' tend to co-occur simultaneously more than what would be predicted by random chance, then most likely areas i and i' share an insurgent movement capable of spatial coordination across i and i'. The paper also carries out an economic analysis of the spread and frequency of attacks. Specific historical and topological constraints like ethnic composition of the local population, terrain ruggedness, local resources, or ease of access to safe havens (e.g. vicinity to the Pakistani border in the South) explain a considerable amount of insurgent diffusion and its dynamics over time.

Progress in understanding insurgency appears key in furthering our knowledge of the determinants and consequences of political violence in developing countries. In this sense, much of the analysis in this paper is necessarily context-dependent, but informative nonetheless for regional stabilization and local development goals (Drozdova, 2012). On the other hand, our methodological contributions have a more general appeal.

Figure 1: Ethnicities of Afghanistan

Figure 2: Total attacks per capita

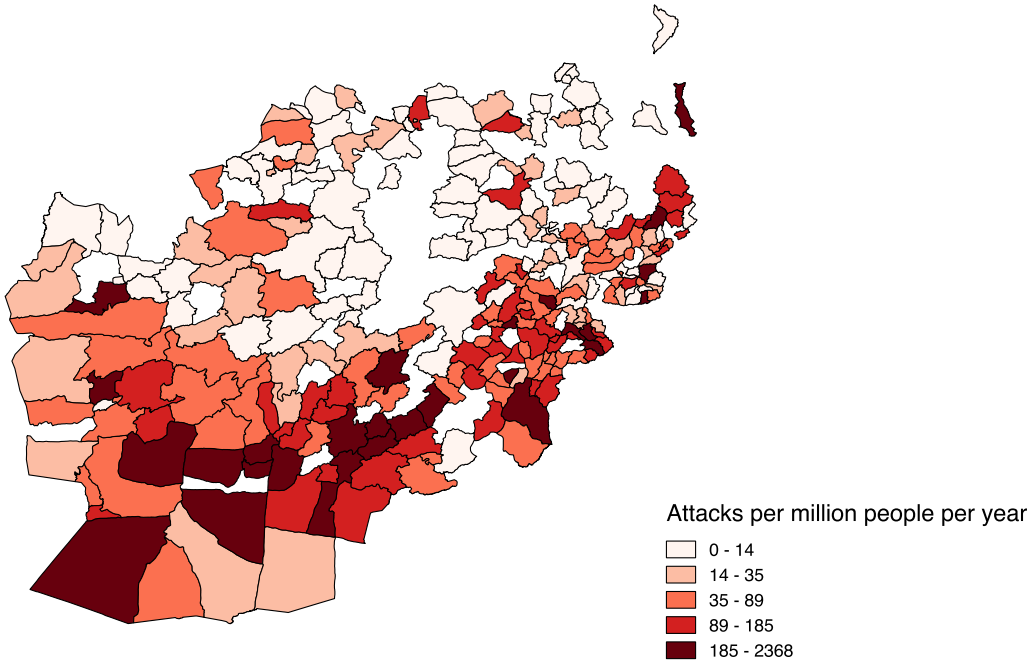

Figure 3: Attacks per capita 2004-2007

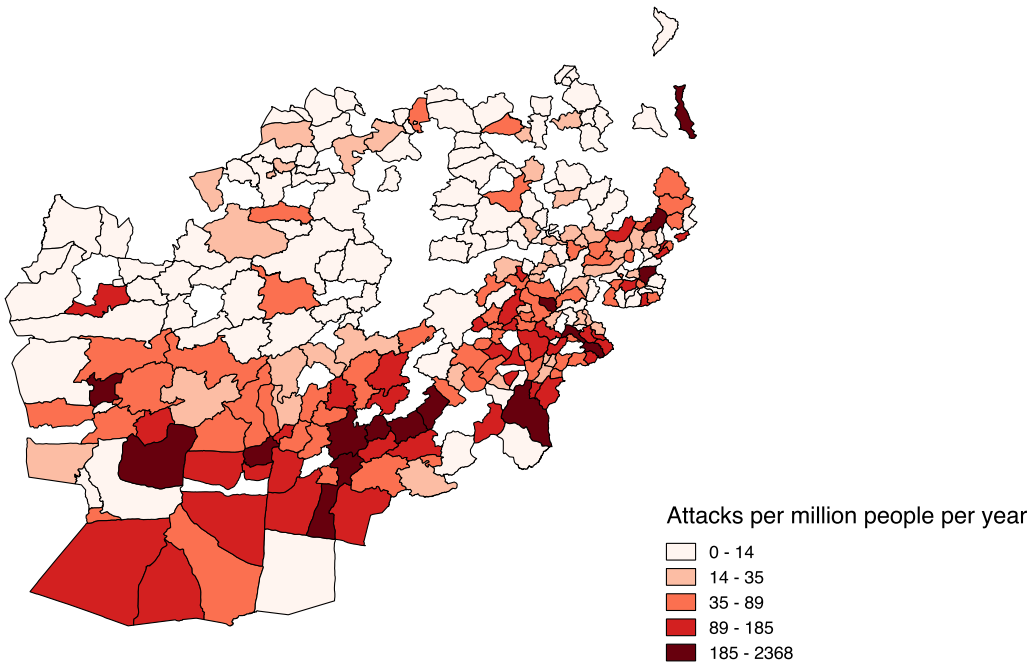

Figure 4: Attacks per capita 2008-2009

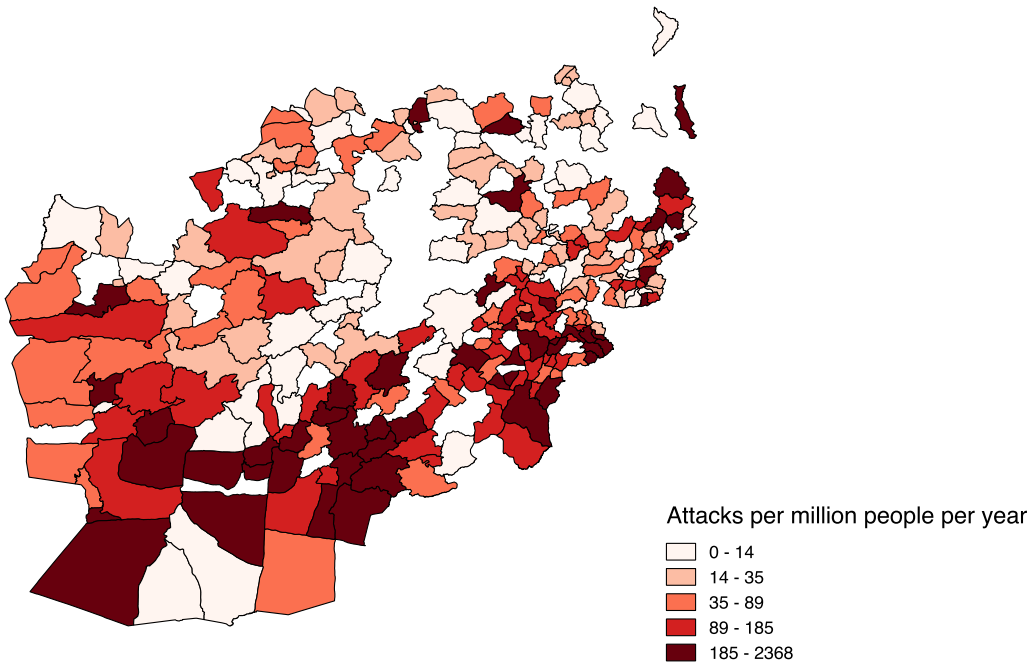

Figure 5: (Sorted) Eigenvalues for Spectral Clustering

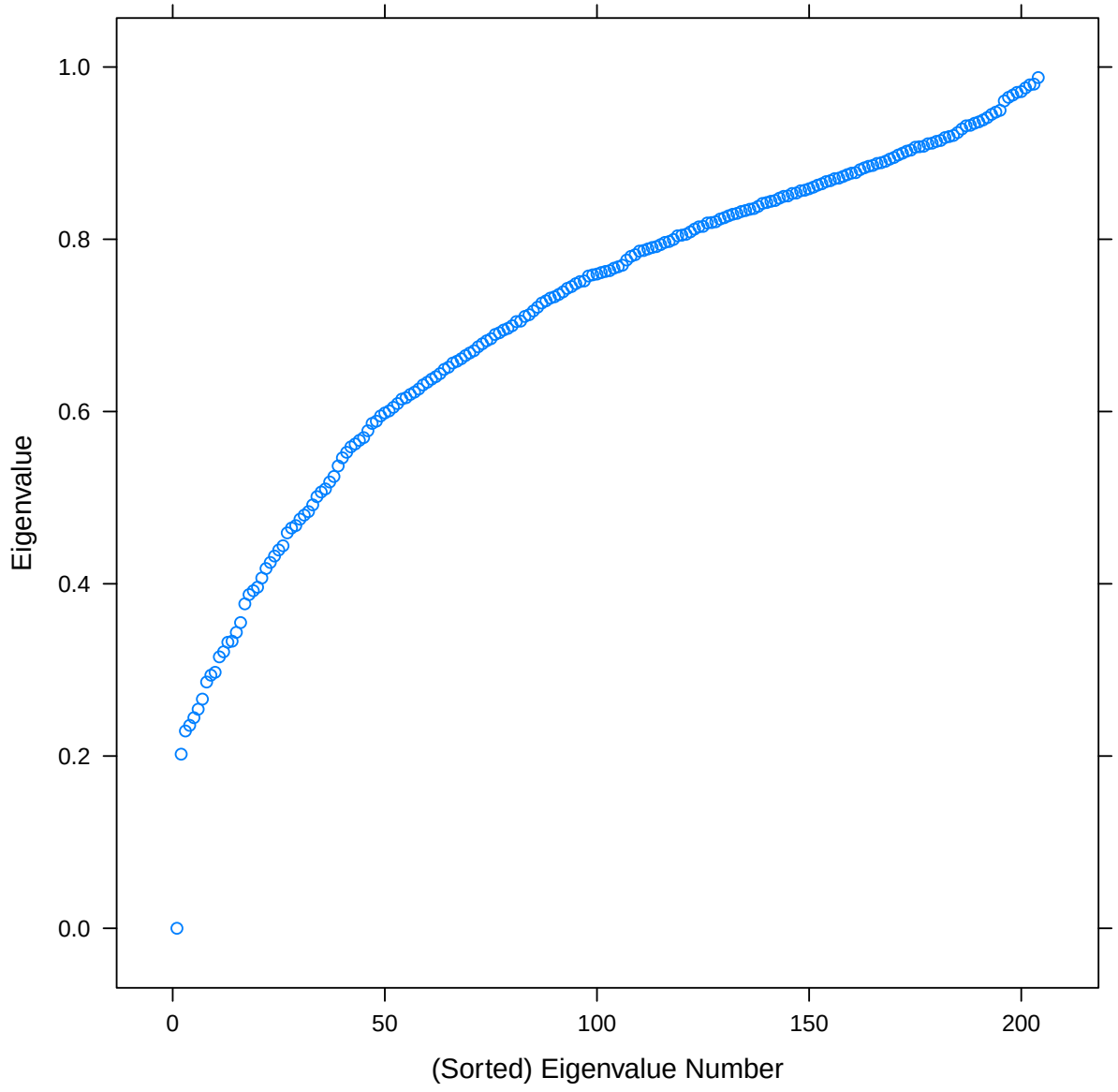

Figure 6: Eigengaps for Spectral Clustering

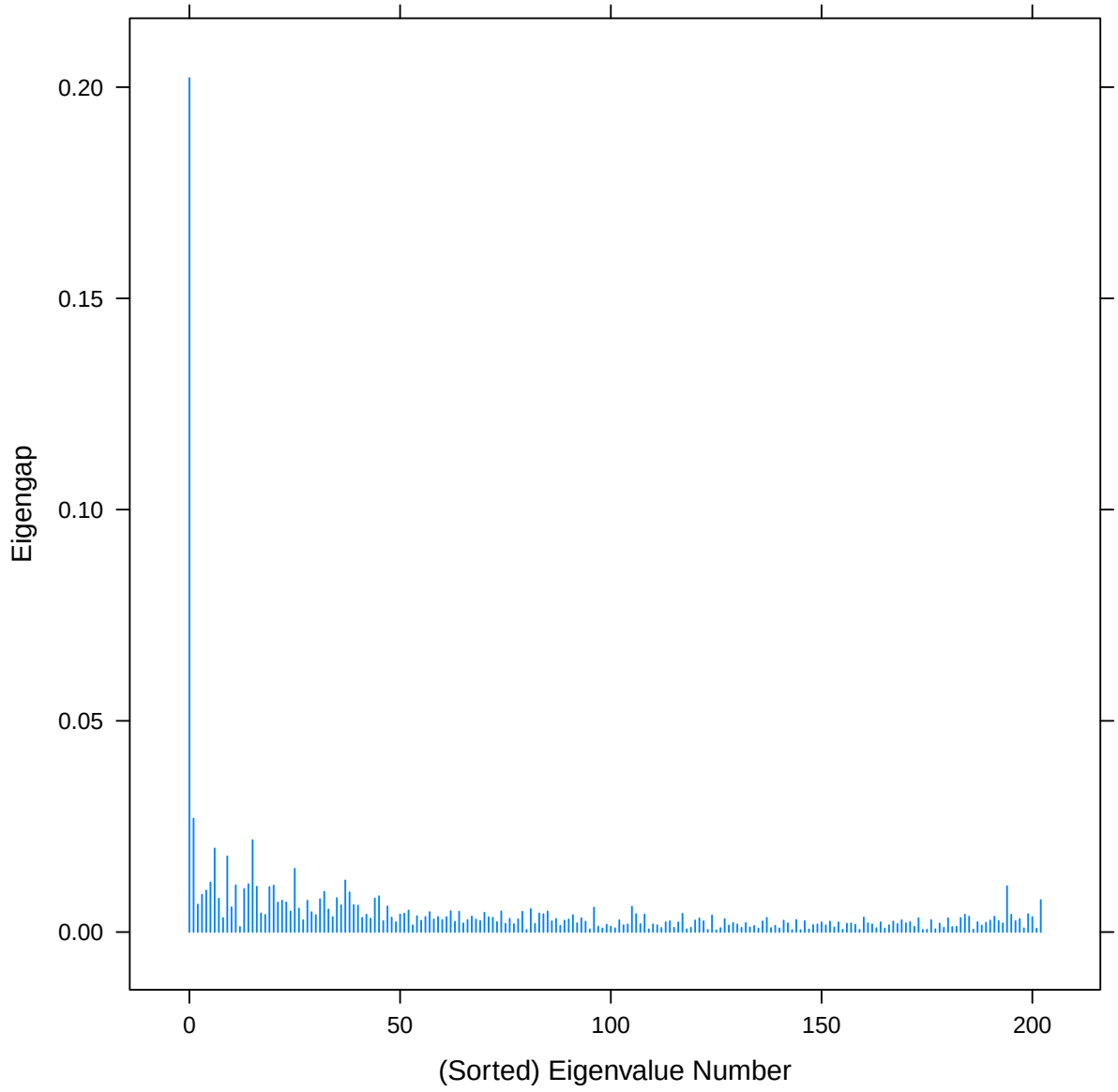

Figure 7: Organized group members: Spectral clustering (Equation 3)

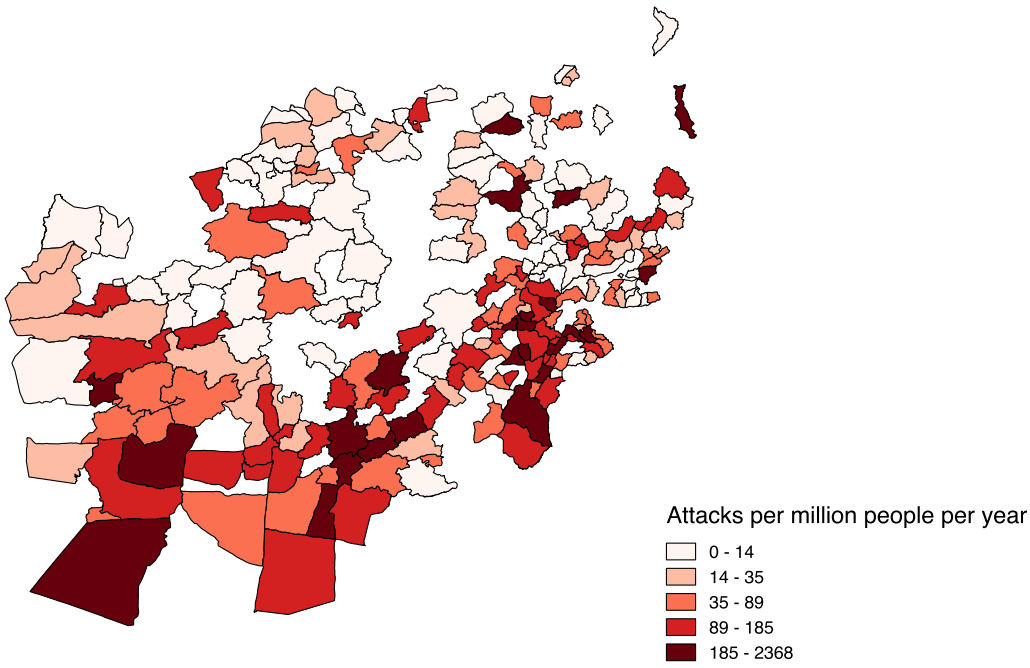

Figure 8: Organized group members: NNMF method (Section 3.2)

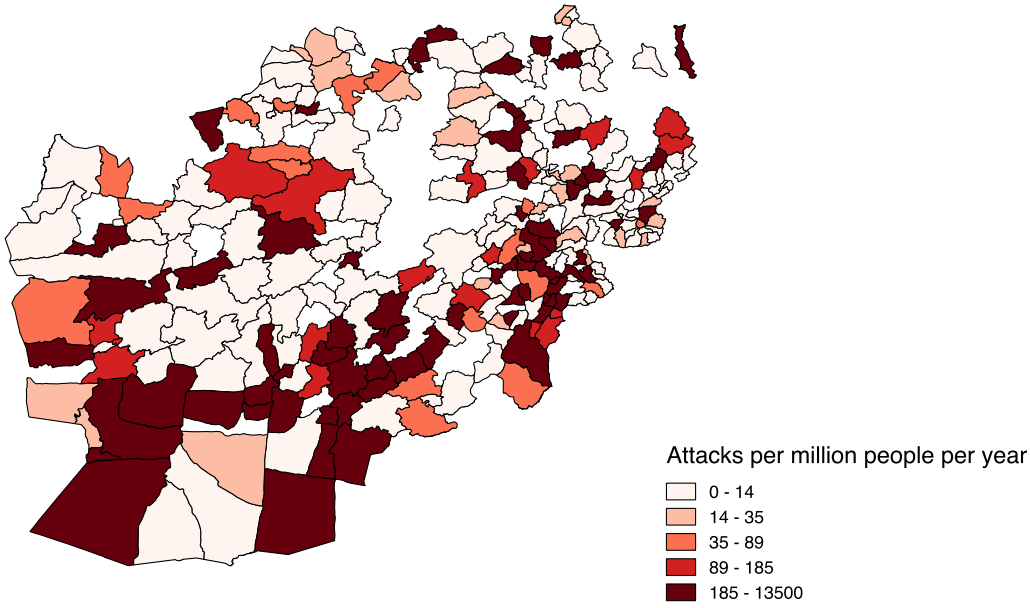

Figure 9: Organized group members: NNMF method (Section 3.3)

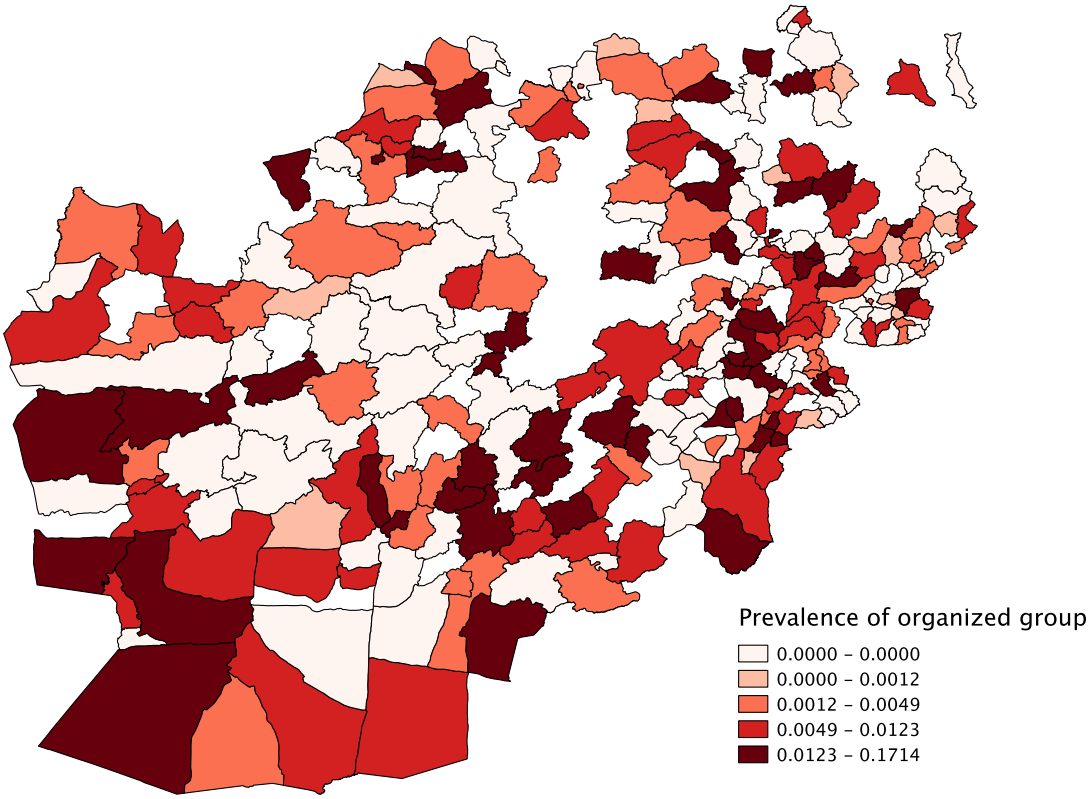

Figure 10: Organized group members: Spectral clustering (2004-2007)

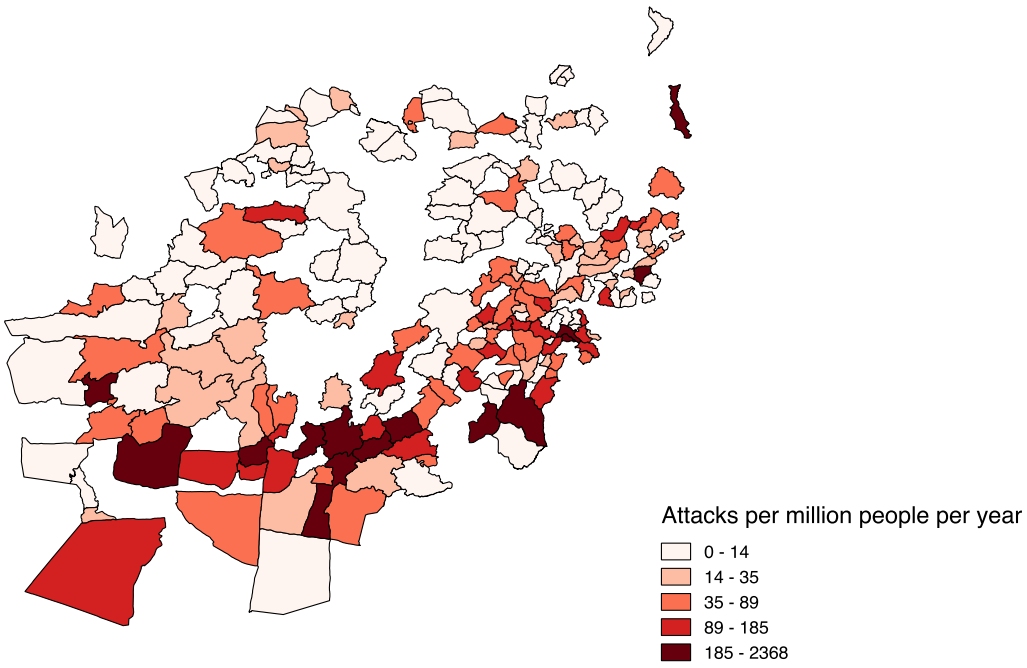

Figure 11: Organized group members: Spectral clustering (2008-2009)

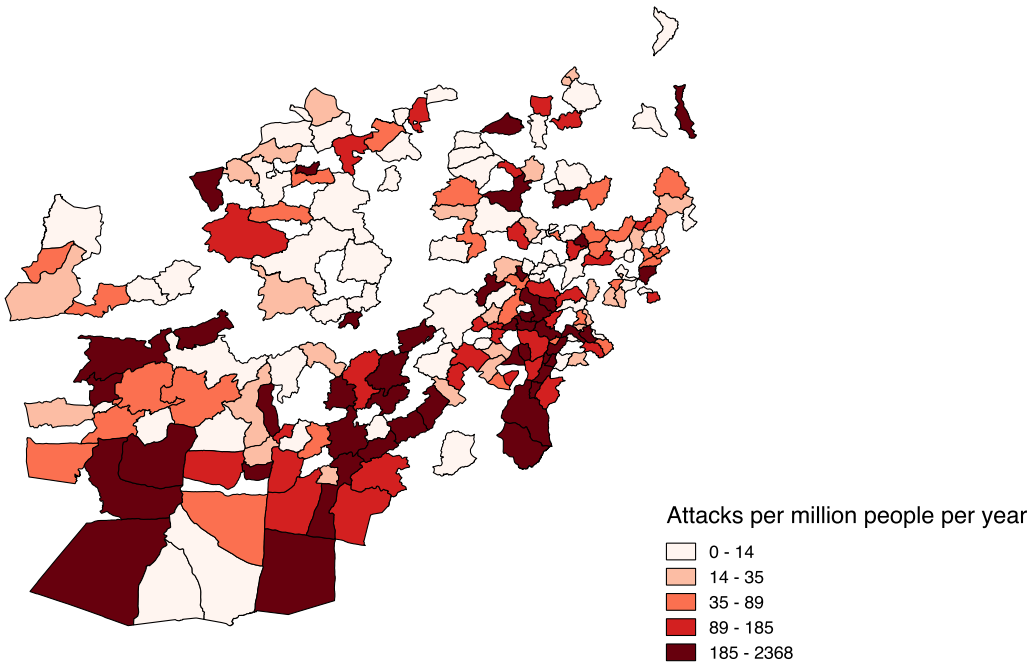

Figure 12: (Estimated) attacks by organized group members (2004-2007, average over adjacent districts)

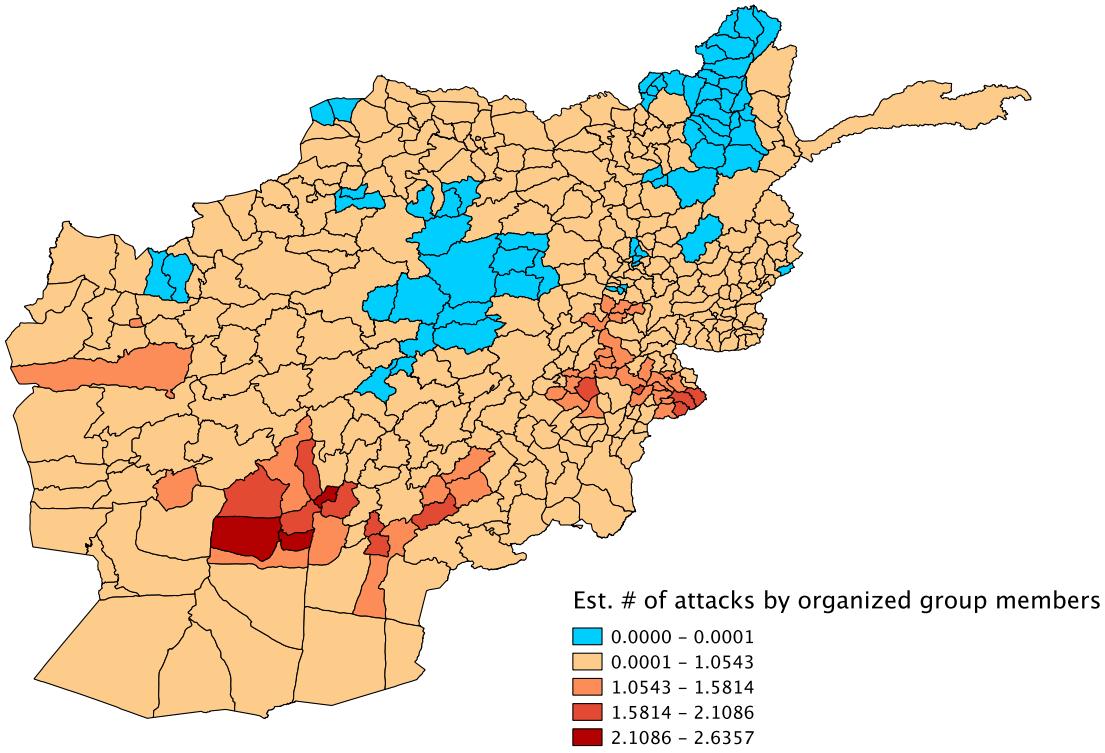

Table 1: Afghanistan timeline 2001-2011

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
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: 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 5: 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
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 6: 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 7: 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 8: 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(adj.attacks.pc == 0)TRUE	-0.95^*	-0.69^*	-0.49^*	-4.71^*	-4.39^*	-3.67^*
	(0.10)	(0.12)	(0.18)	(1.02)	(1.04)	(1.11)
mod.pre	0.78^*	0.68^*	0.69^*	0.34^*	0.28^*	0.34^*
	(0.11)	(0.11)	(0.11)	(0.07)	(0.08)	(0.13)
log(population08)		0.27^*	0.24		0.54^*	0.61^*
		(0.10)	(0.13)		(0.17)	(0.23)
log(areadeg)		0.08	0.06		0.22	0.10
		(0.07)	(0.10)		(0.14)	(0.17)
mean_2000		-0.71^*	-0.63^*		-1.94	-1.04
		(0.27)	(0.29)		(1.18)	(1.30)
centroidlat		-0.17^*	0.01		-0.18	-0.27
		(0.05)	(0.17)		(0.09)	(0.41)
centroidlon		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 is (estimated) number of organized attacks in Jan 2008 - July 2009.

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

adj.attacks is (estimated) number of organized attacks per capita in adjacent districts in 2004-2007.

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Table 9: 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(adj.attacks.pc == 0)TRUE	−0.34*	−0.15	−0.13	−3.29*	−2.61*	−1.78
	(0.08)	(0.08)	(0.12)	(1.10)	(1.15)	(1.26)
log(population08)		0.10	0.02		0.62	0.02
		(0.08)	(0.09)		(0.38)	(0.52)
log(areadeg)		0.01	0.02		−0.32	0.44
		(0.05)	(0.06)		(0.40)	(0.58)
mean_2000		−0.29*	−0.12		−7.43	−1.52
		(0.13)	(0.15)		(7.62)	(3.00)
centroidlat		−0.10*	−0.10		−0.46*	−1.50
		(0.04)	(0.09)		(0.19)	(0.82)
centroidlon		−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 is (estimated) number of organized attacks in Jan 2008 - July 2009.

adj.attacks is (estimated) number of organized attacks per capita in adjacent districts in 2004-2007.

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Table 10: 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 2.1; III-IV use the model from Section 2.2.
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 11: 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 2.1; III-IV use the model from Section 2.2.
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 12: 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 2.1; III-IV use the model from Section 2.2.
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.

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.[30] 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 6 is not relevant for the eigengap procedure, as eigenvalues preceding the gaps on the right hand side of Figure 6 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 6 is larger, and this difference is statistically significant at the 95% level.

[30] 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 10 to 12.

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 ansymptotically, 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 avaible: 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 (and groups?).

REFERENCES

[1] **Anderson, Carl A.** (1974) "Portugese 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] **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.

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

[5] **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

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

[7] **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

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

[9] **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.

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

[11] **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.

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

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

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

[15] **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.

[16] **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.

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

[18] **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.

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

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

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

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

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

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

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

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

[27] **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.

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

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

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

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

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

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

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

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

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

[37] **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.

[38] **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.

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

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

[41] **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.

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

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

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

[45] **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.

[46] **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.

[47] **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.

[48] **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.

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

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

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

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

[53] **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

[54] **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.

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

[56] **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

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

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