NBER WORKING PAPER SERIES

INSURGENCY AND SMALL WARS: ESTIMATION OF UNOBSERVED COALITION STRUCTURES

Francesco Trebbi
Eric Weese

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

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

The authors would like to thank Eli Berman, Ethan Bueno de Mesquita, James Fearon, Camilo Garcia-Jimeno, Jason Lyall, Carlos Sanchez-Martinez, Jake Shapiro, Drew Shaver, Austin Wright and seminar participants at UCSD, Penn, Chicago Harris, Columbia, Stanford, Berkeley, Rochester, UQAM, Kobe, Tokyo, and Princeton for useful comments and discussion and the researchers at the Princeton University Empirical Studies of Conflict Project for generously sharing their incident data online. Nathan Canen provided excellent research assistance. We are grateful to the Social Science and Humanities Research Council for financial support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Insurgency and Small Wars: Estimation of Unobserved Coalition Structures
Francesco Trebbi and Eric Weese
NBER Working Paper No. 21202
May 2015, Revised May 2016
JEL No. O1,P48

ABSTRACT

Insurgency and guerrilla warfare impose enormous socio-economic costs and often persist for decades. The opacity of such forms of conflict is often an obstacle to effective international humanitarian intervention and development programs. To shed light on the internal organization of otherwise unknown insurgent groups, this paper proposes two methodologies for the detection of unobserved coalitions of militant factions in conflict areas, and studies their main determinants. Our approach is parsimonious and based on daily geocoded incident-level data on insurgent attacks alone. We provide applications to the Afghan conflict during the 2004-2009 period and to Pakistan during the 2008-2011 period, identifying systematically different coalition structures. Further applications are discussed.

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

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

1 Introduction

Among many political sources of welfare loss, few compare in magnitude to military conflict and, in the post World War II period in particular, to the losses ascribed to civil war and violent insurgency [O'Neill, 1990]. Insurgency, defined as armed rebellion against a central authority[1], is also one of the most opaque forms of conflict. Intertwining connections with the population blur the lines between combatants and civilians [Kilcullen, 2009]. The relative strength and even the identity of potential negotiating counterparties are often unclear, and in the words of Fearon [2008] "*there are no clear front lines.*" Such forms of conflict have disproportionately affected poor countries and are gaining central status in the literature on the political economy of development [Blattman and Miguel, 2010; Berman and Matanock 2015]. Our paper offers a contribution to the economic analysis of these irregular wars that is both methodological and empirical.

Theoretically, we propose a set of approaches aimed at estimating the latent internal organization of multiple insurgent groups. Motivating our interest in this particular problem is uncontrived. Take the case of Syria after 2011. Lack of knowledge about the organization of the anti-Assad insurgency and the structure of cross-alliances among militant groups emerged as a crucial obstacle to targeted Western military action and a deterrent to humanitarian relief intervention in the Syrian civil war [Jenkins, 2014]. The political risks of supporting anti-regime rebels (possibly aligned with radical Islamist movements) were deemed too high. In this backdrop, to February 2016, the Syrian conflict alone displaced 10.7 million civilians and caused the death of over $250,000$ individuals according to the United Nations Office for the Coordination of Humanitarian Affairs [United Nations, 2016].

Empirically, we focus on the costly insurgencies of Afghanistan and Pakistan[2]. On the U.S. side alone, Afghan operations cost the lives of more than $1,800$ troops between 2001 and 2011, and led to more than $444 billion in military expenses.[3]

[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] Unfortunately, we are not aware of suitable data for Syria. In Section 5 we further discuss the case cases of Iraq, Syria, and Libya, all instances where our methodologies could be potentially of use.

[3] Soon into the operation, the U.S. military acknowledged through a drastic adjustment in tactics that the Afghan conflict differed substantially from previous large-scale military operations.

Statistics for Afghan civilians appear less certain, but the adverse effects are painfully obvious even to the casual observer. In Pakistan, a perduring effort in placating nationalist, militant, and sectarian-related insurgencies has been accompanied by a heavy toll of 18,583 people killed and 19,356 injured between 2012 and 2015 alone [Pak Institute for Peace Studies, 2016, p.8]. This effort has diverted valuable resources from aid and development assistance programs and public goods provision in the country.

As an more concrete example, consider the case of Afghanistan specifically. The insurgency during 2004-2009 has been approximately described as originating from an alliance between the Afghan Taliban insurgents and al-Qaeda foreign fighters against the Afghan government and the supporting international coalition forces. While this was surely a facet of this early phase, front lines were uncertain and the unity of the insurgents doubtful. Importantly, policymakers disagree about whether the Taliban were a unified fighting organization, or rather an umbrella coalition of heterogeneous forces. Some were skeptical of the degree of control that Taliban leader Mullah Mohammed Omar exerted over the powerful Haqqani faction and the Dadullah network.[4] Similarly, the Hizb-i Islami faction was considered by many a separate entity from the Taliban proper.[5]

Other observers promoted an opposite view. In an insightful qualitative essay Dorronsoro [2009] states: "*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] Evidence

Appendix A includes a reference time line for conflict events in Afghanistan.

[4]For example, the UN report [2013] stated that *"Despite what passes for a zonal command structure across Afghanistan, the Taliban have shown themselves unwilling or unable to monopolize anti-State violence. The persistent presence and autonomy of the Haqqani Network and the manner in which other, non-Taliban, groupings like the Lashkar-e-Tayyiba are operating in Afghanistan raises questions about the true extent of the influence exerted by the Taliban leadership."* Brahimi [2010] reports a statement by Ashraf Ghani, current Afghan president, in a lecture for the Miliband Programme at LSE indicating *"The Taliban are not a unified force - they are not the SPLA in Sudan or the Maoists in Nepal"* while Giustozzi [2009] states that "*The Taliban themselves are not fully united and the insurgency is not limited to the Taliban."*

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

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

in support of this position includes the existence of the *Layha* (a centralized code of conduct for Mujahidin), as well as the strong centralizing tendencies of the *Obedience to the Amir* (a manual endorsed by Mullah Omar).[7]

This disagreement among experts is unsettling. Understanding the extent of territorial control and population support of insurgent groups is essential for military operations. Furthermore, knowledge of the internal organization and cohesion of rebel groups could have been used to prevent selective violence by insurgents, and ultimately help with the relief and reconstruction of areas affected by conflict.[8]

This paper shows how the covariance structure of data on violent events can be used to estimate the number of different active insurgent groups and their territories of influence, features that are typically unobservable to the econometrician. In this particular application we make use of the specific covariance structure emerging from the fact that insurgent groups with the ability to launch simultaneous and geographically separated attacks generally appear to do so. This relies on the conclusions of the existing literature regarding the incentives for organized violent groups to launch coordinated attack in multiple locations. For instance, Deloughery [2013] provides a recent review of this literature and presents systematic evidence of the advantages of simultaneous attacks for terrorist organizations in terms of media coverage and appeal in the recruitment of new fighters –incentives that operate within insurgencies as well.[9]

[7]These are available in English translation as Munir [2011] and Ludhianvi [2015], respectively.

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

[9]From a western perspective, the 9/11 attacks in the United States are the most obvious example of the salience of such simultaneous violence, but the phenomenon is widespread. For example, in southern Thailand insurgent movements have adopted similar tactics: "On April 28, 2004 groups of militants gathered at mosques in Yala, Pattani, and Songkhla provinces before conducting simultaneous attacks on security checkpoints, police stations and army bases" [Fernandes, 2008]. The Indian Mujahideen, responsible for the 2008 Mumbai attacks, typically carry out simultaneous attacks [Subrahmanian et al., 2013]. Kurdish nationalists and the Tamil Tigers are known to have adopted simultaneous attacks as a strategy. In Africa, Boko Haram in northern Nigeria has carried out coordinated attacks on multiple targets such as churches, and Anderson [1974] describes coordinated attacks in Portuguese colonies. Simultaneous attacks and suicides have been a trademark of international jihadist organizations and of al-Qaeda in particular, making our approach well-suited to the Afghan insurgency case. Because the empirical covariance matrix of attacks is observed, these assumptions implying positive covariances driven by co-occurring incidents are readily verifiable and they are in fact supported by the data. See discussion in Section 2.

We take these incentives as given, and assume that an organization with the capability of launching such attacks will choose to do so. The covariance of the attacks that are observed in excess to simple random correlation can thus be employed to analyze the underlying structure of the insurgent groups organization. Attacks across different locations on the same day are assumed either as the result of random chance or to represent an insurgent group with a presence in those locations. A transparent structural model of attacks is used to distinguish between random chance and organized group behavior. After estimating the number of different guerrilla groups and their territorial extent, we assess the main empirical determinants of insurgent presence and produce an analysis of shifts in insurgent presence over time.

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

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

We present estimators that allow for a single district to be contested by multiple guerrilla groups or by a single group, or by none. The estimators provide the number of guerrilla groups operating, the geographic area of each of these, and the intensity of each group's activity in each district. Let us emphasize that the methods we present can accommodate slow-moving trends in violence over time and are robust to aggregate shocks (e.g. weather, seasonality, or U.S. troop movements) that might

affect insurgent activity in many areas simultaneously.

In order to provide our estimates, the data generating process used to model the occurrence of violent incidents is somewhat stylized. The partial loss of generality has to be traded off with the parsimony necessary due to the lack of any detailed data on the internal organization and planning strategies of insurgent forces.

The main empirical results of the paper are as follows. We conclude that for the early period 2004-2009 insurgent activity in Afghanistan is best represented by a single organized group, rather than several independent groups, and that the extent of this group is largely determined by ethnic boundaries. This result is probed by partialling out slow-moving trends and spurious correlation induced by large-scale U.S. military activity, to constraining the analysis to districts with a number of incidents above specific thresholds, and to limiting the analysis to incidents explicitly claimed by the Taliban.

We also conduct an in-depth analysis of the Pakistani insurgency (which includes the Pakistani Taliban, known as Tehrik-i-Taliban Pakistan or TTP), using multiple data sets for the period 2008-2011. Interestingly, the TTP are completely separate from their Afghan counterparts and unanimously considered an umbrella coalition of diverse violent actors.[10] We show that in the case of Pakistan, our methodology detects multiple insurgent groups (four, in fact) and is completely consistent with the extant qualitative literature on insurgency in Pakistan.

We also consider changes in the extent of the Afghan Taliban over two time periods: 2004-2007 compared to 2008-2009. We find that insurgents spread largely to districts adjacent to those where they were already present: this is sometimes described as an "oil spot" strategy.[11] We also find that there has been penetration by the insurgents into areas traditionally occupied by non-Pashtun ethnic groups. We finish by discussing several case studies outside of Afghanistan and Pakistan where application of the methodologies we present could be helpful in assessing the economics of post-conflict reconstruction and power sharing with former insurgent groups from Latin America, Asia, and Africa.

An increasing amount of attention has been devoted within the fields of development economics and political economy to the study of armed conflict within countries,

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

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

in particular civil wars and insurgency. Economists have been interested in the analysis of violence and conflict at least as far back as Schelling [1960] and Tullock [1974], with Hirshleifer [1991, 1995a, 1995b, 2001] and Grossman [1991, 2002] offering more recent theoretical contributions. Outside of Economics the interest has been even greater: Political scientists have dedicated to the study of conflict a substantial part of their work in the field of International Relations.

Political Science and Economics have provided some of the most recent and novel insights in the study of insurgency.[12] As underlined by Blattman and Miguel [2010], a remarkable characteristic of this recent wave of research has been a strong empirical bend and an increasing attention to micro-level (typically incident-level) information. The use of geocoded micro data in this area is a departure from more established "macro" empirical approaches, which were based on country level information or aggregate conflict information.[13]

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

The remainder of this paper is organized as follows. Section 2 develops our methodology for the estimation of coalition structures among insurgent groups. We describe our data in Section 3, particularly the incident-level data for Afghanistan and Pakistan, which have been generously made available by the Empirical Studies Of Conflict (ESOC) project. The analysis of the determinants of insurgent group presence is developed in Section 4. Section 5 presents several case studies focused on the economic importance of understanding insurgent organization in conflict and

[12]These include Berman [2009], Berman et al. [2011], Condra et al. [2010], Blair et al. [2012], Condra and Shapiro [2012], Callen and Weidmann [2013], and Bueno de Mesquita [2013].

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

post-conflict environments in developing countries. Section 6 concludes.

2 Model

We are interested in estimating parameters describing the presence and identity of organized insurgent groups in each of many districts in a country, and also the total number of such insurgent groups active in the country. We do not propose a complete theory of insurgent activity. Instead, we develop a parsimonious econometric model based on assumptions regarding insurgent behavior. This approach is motivated by both lack of data on opaque insurgency organization and by the computational difficulties involved in estimating models of the type considered; these difficulties appear to preclude consideration of all but the simplest structural models.

We first present a model of insurgent behavior that links insurgent presence to entries in the cross-district covariance matrix of insurgent attacks. Next, we discuss how to decompose this covariance matrix into a form useful for estimation. We then describe two estimation approaches based on this model: the first assumes that exactly one insurgent group is present in each district, while the second relaxes this assumption. We conclude by explaining how we can avoid potential bias from long-term trends in attacks across districts. Avoiding bias from weather, seasonal fluctuations, and other such trends is important when applying our approach to actual data, but we discuss these details last in order to simplify the exposition of the model.

2.1 Insurgent Attacks

Let districts be indexed by i, and let there be a total of N districts in which attacks occur. Violent occurrences in i can be of two types: unorganized or organized by an insurgent group. We make a distinction between attacks initiated by unorganized local militants and those initiated by members of an organized group because we also allow for the possibility that there are no organized insurgent groups present in a district even though attacks are observed there.

Let ℓ_i be the number of unorganized local militants in district i. Let organized insurgent groups be indexed by j, and let J be the total number of such organized groups active anywhere in the country. Let α_{ij} be the number of members in district i belonging to organized group j. Time is discrete and indexed by t. In our analysis

below, the time periods used will be days. This relatively high-frequency attack data is useful because it reduces the number of attacks that are simultaneous simply by random chance.

In each time period, the probability that an unorganized local militant launches an attack is η, which does not change across time (this assumption is relaxed in Section 2.5). The decision by unorganized militants to attack is independent of the decision of anyone else (unorganized militant or group member). The expected number of attacks by local militants in district i at time t is thus $\eta\ell_i$, and the variance within district i is $\eta(1-\eta)\ell_i$. The covariance in these attacks between two districts i and i' is zero: the attack decisions are made independently, and the probability of an attack is constant.

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

Consider the members of group j. If there are α_{ij} members in district i and $\alpha_{i'j}$ members in district i', then the covariance in attacks over time between these two districts, due to the presence of members of group j, is $\alpha_{ij}\alpha_{i'j}\sigma^2$. Summing over members of all groups, the covariance in attacks between districts i and i' will be $\sum_j \alpha_{ij}\alpha_{i'j}\sigma^2$.

The model just presented is clearly a stylized model of the attack behavior of insurgent groups, and the covariance structure imposed is not without loss of generality. A particularly strong assumption made in the model is that the members of an insurgent group do not move between districts: a given group j has a certain membership α_{ij} in district i, and those members will either be encouraged to attack in a given period (a high ϵ_{jt}), or not (low ϵ_{jt}).

A very different model would be one in which members of an insurgent group

are mobile, and in any given period have the choice of attacking in one of many districts. This latter model implies that organized groups should lead to negative covariances between districts, 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 that this covariance 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 data used, the observed covariances are systematically positive, validating our main assumption.[14] The qualitative research of Deloughery [2013] and others, as discussed in Section 1, also suggests that a model without substantial substitution in attacks across districts appears most appropriate.

2.2 Covariance Decomposition

Let Γ be the covariance matrix for attacks discussed in Section 2.1, where the entry in row i and column i' gives the covariance in attacks across time for these two districts. Analysis will be based on this matrix, and others created from it.[15]

The covariance matrix Γ can be decomposed as

$$\Gamma = \Gamma_D + \Gamma_L, \tag{1}$$

where Γ_D is a diagonal matrix and Γ_L is a low rank matrix of the form

$$\Gamma_L = \sigma^2 \begin{bmatrix} \sum_j \alpha_{1j}\alpha_{1j} & \sum_j \alpha_{1j}\alpha_{2j} & & & \\ \sum_j \alpha_{2j}\alpha_{1j} & \sum_j \alpha_{2j}\alpha_{2j} & & & \\ \quad ... & & \sum_j \alpha_{ij}\alpha_{ij} & & \\ \sum_j \alpha_{ij}\alpha_{1j} & & & ... & \sum_j \alpha_{ij}\alpha_{i'j} \\ \quad ... & & & & \end{bmatrix} \tag{2}$$

This decomposition is considered because the diagonal entries of the covariance matrix are a sum of variance from unorganized militants and variance from organized

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

[15] This specification is similar to the "random flock" of Bottegal and Picci [2015], except that an attack either happens or does not in each case, and this is observed, rather than the underlying probability being observed.

groups, and only the latter is of interest.[16] As a normalization, we set $\sigma^2 = 1$.

Let $\gamma_{ii'} = \sum_j \alpha_{ij}\alpha_{i'j}$ denote the off-diagonal entry on row i and column i' of Γ_L. Let $\bar{\gamma}_{ii'}$ be the corresponding entry of the covariance matrix in the observed sample. Unfortunately, no empirical counterpart to Γ_L is observed, and thus one will have to be created by modifying the diagonal of the observed covariance matrix $\bar{\Gamma}$.

To create a $\hat{\Gamma}_L$ from $\bar{\Gamma}$, a diagonal matrix $\hat{\Gamma}_D$ will be subtracted from the latter to produce the former. An intuitive method for doing this is "trace minimization", discussed at least as early as Ledermann [1940]. First, note that $\bar{\Gamma}$ is a (sample) covariance matrix, and is thus positive semi-definite. $\hat{\Gamma}_L$ should also correspond to a covariance matrix, and thus should also be positive semi-definite. Consider the optimization problem

$$
\begin{aligned}
(3) \qquad & \min_{\hat{\Gamma}_D} \operatorname{Tr}(\hat{\Gamma}_L) \\
& \text{s.t. } \hat{\Gamma}_L = \bar{\Gamma} - \hat{\Gamma}_D, \quad \hat{\Gamma}_D \text{ diagonal}, \\
& \qquad \hat{\Gamma}_D \succ 0, \hat{\Gamma}_L \succ 0
\end{aligned}
$$

Here Tr() denotes the sum of diagonal entries of a matrix, and $\succ 0$ indicates positive semi-definiteness. The intuition for trace minimization is that the "extra" variance present in the diagonal entries of Γ has the form of a full rank matrix, and thus in order to recover a low rank matrix such as Γ_L, as much of this as possible needs to be removed.[17]

If $N = 200$, the the semi-definite program corresponding to (3) involves $200 \times 199 = 39,800$ constraints: each off-diagonal entry $\bar{\gamma}_{ii'}$ in the positive semi-definite matrix $\bar{\Gamma}$ must be equal to the corresponding entry in $\hat{\Gamma}_L$. Problems of this size, however, are feasible using modern semidefinite programming algorithms. We thus compute $\hat{\Gamma}_L$ using (3), and will use it as the basis for producing an estimate of

[16] The diagonal entries of Γ do not in general have a useful form. For example, even in the very simple case where there is only one group and ϵ_1 is uniformly distributed on $[0, b]$, the ith diagonal entry would be a non-trivial nonlinear expression $\frac{b^2}{12}(\frac{6}{b} + (\alpha_{i1} - 4))\alpha_{i1} + \ell_i\eta(1-\eta)$. A simpler form for the diagonal entries could be obtained by using a mixture Poisson approximation, such as the Poisson-Gamma used in Ashford and Hunt [1973]. A variety of these distributions are discussed in Karlis and Xekalaki [2005]. None of the options available, however, appear to simplify the diagonal entries enough to be directly useful from an empirical perspective.

[17] Saunderson et al. [2012] show that the intuition of Ledermann and others was correct in general. Specifically, the positive semi-definite matrix Γ_L can be recovered given Γ so long as it is sufficiently "incoherent", and this property is satisfied by most low rank matrices. Details are provided in Appendix B.

insurgent group presence in the next two subsections.

2.3 Non-overlapping Insurgent Groups

We desire both an estimate $\hat{J}$, the total number of organized insurgent groups, as well as an estimate $\hat{\alpha}_{ij}$ for each district i and group j, giving the number of insurgent members of the group operating in that district. The set of estimates $\{\hat{\alpha}_{ij}\}$ will have a total of $N \times \hat{J}$ elements. It turns out to be easiest to first produce the $\{\hat{\alpha}_{ij}\}$ estimates for each value of $J \in \{1, ..., J_{\max}\}$, and then choose a $\hat{J}$ based on examining this set of estimates.[18] We will thus begin by assuming that J is known, and consider how to compute estimates $\{\hat{\alpha}_{ij}\}$ given J. After this, we will then consider how to choose $\hat{J}$.

An approach based on standard clustering techniques will be presented in this subsection, as well as one based on eigenvalues. Matrix factorization will be considered in Section 2.4. The approach based on clustering and that based on matrix factorization are largely complementary in that they rely on different assumptions. This provides a form of cross-validation for our results, which is important given the novelty of these methodologies within the field of political economy.

Estimation via standard clustering techniques requires an additional assumption different from those that will be needed for the techniques of Section 2.4: specifically, it is necessary to assume that the various insurgent groups present do not have overlapping territories. That is, there one organized group j present in any given district i.[19] Based on this assumption, reordering the districts i allows Γ_L to be written as a block-diagonal matrix:

$$\Gamma_L = \begin{bmatrix} \Gamma_L^1 & 0 & & \\ 0 & \Gamma_L^j & & \\ ... & & & \\ 0 & & ... & \Gamma_L^J \end{bmatrix} \tag{4}$$

where there are a total of J organized groups, and each block Γ_L^j has the form given in Equation 2. To produce estimates $\{\hat{\alpha}_{ij}\}$ we will first determine which organized

[18] The exact choice of $J_{\max}$ is not important.

[19] This is a direct consequence of the standard assumption that factors should be orthogonal, combined with the fact that insurgent prevalence α must be non-negative.

group is present in each district, and then afterwards we will determine the strength of this group in the district.

To determine which organized group is present in each district, we will follow a modified k-means type approach[20]. Begin by constructing a scaled version of Γ_L:

$$\Gamma_L^{\text{cor}} = D(\sum_j \alpha_{\cdot j}\alpha_{\cdot j})^{-1/2}\Gamma_L D(\sum_j \alpha_{\cdot j}\alpha_{\cdot j})^{-1/2},$$

where $D()$ indicates a diagonal matrix with the specified vector on the diagonal. This process is occasionally referred to as "sphering" and it often improves the quality of the clustering. By assumption, for each district i, $\alpha_{ij} = 0$ for all but one group j, and thus Γ_L^{cor} is constructed by dividing row i and column i of Γ_L by the value of α_{ij} for the single group j that is present in i.

The "cor" superscript is used because Γ_L^{cor} is positive semi-definite with all diagonal entries equal to one, and thus has the form of a correlation matrix. However, observe that each off-diagonal entry $\gamma_{ii'}$ has now been divided by $\alpha_{ij}\alpha_{i'j}$ if the same insurgent group is present in districts i and i'. Thus, in exactly the same way as (4), after suitable rearrangement Γ_L^{cor} is a block diagonal matrix with entries consisting only of zeros and ones:

$$\Gamma_L^{\text{cor}} = \begin{bmatrix} 1_{N_1} & 0 & & \\ 0 & 1_{N_j} & & \\ ... & & & \\ 0 & & ... & 1_{N_J} \end{bmatrix} \tag{5}$$

where 1_{N_j} is an N_j by N_j matrix consisting entirely of ones, and corresponding to the N_j districts that have group j present in them.[21]

Running a k-means type clustering algorithm on Γ_L^{cor} would be trivial, but only the finite sample version is available. Let $\hat{\Gamma}_L^{\text{cor}}$ be the correlation matrix associated with the finite sample covariance matrix $\hat{\Gamma}_L$.[22] This $\hat{\Gamma}_L^{\text{cor}}$ will have off-diagonal entries

[20]In many clustering approaches transformations or scaling of the original data is common. An example is spectral clustering, a methodology we also employed in the working paper version of the paper. We discuss this approach, and reasons why our current approach may be preferable, in Appendix C. Luxburg [2007] provides a general discussion of the technique.

[21]Note that this step relies on the assumption that there is an organized insurgent group present in every district: if there were a district with no organized presence, creating the correlation matrix would involve division by zero.

[22]The correlation matrix $\hat{\Gamma}_L^{\text{cor}}$ is readily obtained by imposing diagonal elements equal to 1 and

that are neither zero nor one. For many k-means algorithms, a distance matrix rather than a correlation matrix is needed. Such a distance matrix can easily be constructed using cosine distances: $1 - \gamma^{\text{cor}}_{ii'}$ is the cosine distance between i and i', where $\gamma^{\text{cor}}_{ii'}$ is the off-diagonal entry of Γ^{cor}_L corresponding to districts i and i'.[23] The cosine distance between two districts with the same group present will be zero asymptotically, while it will be one when the districts have different groups present. Given J, any reasonable clustering algorithm should thus be able to recover which insurgent group is present in which district, given enough data.[24]

Once districts have been clustered into groups, estimates for $\{\alpha_{ij}\}$ can be obtained. Begin by supposing that each organized group that is present has members in a large number of districts, and that no single district has a particularly large α_{ij}. Let I_j be the set of districts that have members of organized group j. Then, since an assumption of the 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 $\bar{\Gamma}^j$, the relevant block of the original $\bar{\Gamma}$.[25]

Specifically, note that a sum across the off-diagonal entries of a row of $\bar{\Gamma}$ 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

$$\begin{aligned} \sum_{i' \neq i} \alpha_{ij}\alpha_{i'j} &\simeq \sum_{i'} \alpha_{ij}\alpha_{i'j} \\ &= \alpha_{ij} \sum_{i'} \alpha_{i'j} \\ &= \alpha_{ij} a_j \end{aligned} \tag{6}$$

appropriately rescaling rows and columns of the covariance matrix $\hat{\Gamma}_L$ by the square root of the corresponding diagonal entry of $\hat{\Gamma}_L$.

[23] The construction of a distance matrix is trivial because any correlation matrix is also an inter-point angle matrix, and these angles can be used directly to construct a cosine distance matrix.

[24] A weighted clustering approach appears to be called for, because a district with very low α_{ij} for the group j that is present will have very noisy off-diagonal entries. We do not explore optimal weights, instead using ad-hoc weights corresponding to the square root of the diagonal entries of $\hat{\Gamma}_L$. Krishna and Narasimha [1999] provide a weighted k-means algorithm, based on genetic optimization: we use the Hornik, Feinerer, Kober, and Buchta [2012] implementation of this algorithm. Using unweighted clustering instead does not change any of the results discussed below substantially.

[25] A potential alternative approach to the one presented here would be to use the diagonal entries of $\hat{\Gamma}_L$ to produce estimates of $\{\alpha_{ij}\}$. However, this matrix is itself the output of a semi-definite program based on $\bar{\Gamma}$. The approach presented below has the advantage of using the off-diagonal entries of $\bar{\Gamma}$ directly.

where $a_j = \sum_{i'} \alpha_{i'j}$ is the same for any choice of district i within I_j. The row sums of the off-diagonal entries of each block of $\bar{\Gamma}^j$ thus give the relative prevalence of organized group members in each district in I_j.[26]

We now consider how to produce an estimate $\hat{J}$ of the number of insurgent groups present. Our approach is based on a modification of the "gap statistic" of Tibshirani, Walther, and Hastie [2001].[27] The intuition behind this technique is that adding an additional cluster increases the number of degrees of freedom in the model, and thus an appropriate estimate $\hat{J}$ must somehow counterbalance this in order to avoid overfitting the data. This is done by comparing the performance of the model with the actual data to a case with randomly generated data that is known not to have any group structure. In general, computational difficulties loom large in the clustering literature and asymptotic behaviour is given only limited consideration. Below, we present a simple computationally feasible estimator for $\hat{J}$. In Appendix D we discuss modifications that would be necessary for the consistency of this estimator.

In general, the gap statistic approach makes use of the value

$$\text{Gap}(k) = E^*[\log(W_k)] - \log(W_k). \tag{7}$$

Here W_k is some measure of variation left unexplained by the k clusters, and E^* is the expectation taken with respect to a "reference distribution" chosen to correspond to no cluster structure. In Equation 7, $\text{Gap}(k)$ quantifies an intuitive definition of the quality of a k group clustering of the observed data: the clustering is "good" only to the extent that it is *better* than what would occur with randomly generated data that had no group structure by construction. Following Tibshirani, Walther, and Hastie [2001], the estimated number of clusters $\hat{J}$ is then selected to be the smallest k such

[26] While it would be possible to use non-linear programming or other techniques to develop an estimator with more desirable properties, the approximate estimator has at least two advantages. First, the estimator has an intuitive interpretation: $\bar{\Gamma}$ is a covariance matrix, and the sum across the off-diagonal entries of a row of $\bar{\Gamma}$ thus gives an indication (in a heuristic sense) of how closely linked attacks in a given district are with attacks in other districts. Second, if in the data a given district i experiences only a small number of attacks, then the off-diagonal entries $\bar{\gamma}_{ii'}$ will be relatively small for that district, and thus i will not introduce substantial noise into estimates $\hat{\alpha}_{i'j}$ for other districts i'. Developing an unbiased estimator that also possesses such properties appears to be a non-trivial undertaking.

[27] An alternative approach would be to compare $\hat{\Gamma}_L^{\text{cor}}$ to the correlation matrix that would be predicted from the k means results, and choose the number of groups that minimized the distance between these two matrices. We do not pursue this approach because it is not clear what sort of distance function might be appropriate.

that

$$\text{Gap}(k) \geq \text{Gap}(k+1) - s_{k+1}, \tag{8}$$

where s_{k+1} is an estimated standard error of $\log(W_k)$ in the case where there is no cluster structure. $\text{Gap}(k)$ describes how much better the estimated cluster structure of k groups looks, compared to how it would be expected to look if there were in reality no cluster structure. If $\text{Gap}(k) \geq \text{Gap}(k+1)$, then this means that adding the $k+1$th cluster did not improve the clustering any more than would be expected with random unclustered data.

In the standard Tibshirani, Walther, and Hastie [2001] setup, if clustering were performed using Γ^{cor}, W_k would be the residual sum of squares or "within sum of squares": the variation in Γ^{cor} that is not explained by the group structure. However, this standard approach runs into problems with the calculation of $E^*[\log(W_k)]$ because a reference distribution for $\hat{\Gamma}_L^{\text{cor}}$ needs to be calculated.[28] This calculation appears to be extremely complicated, because the answer depends on the finite sample behavior of $\hat{\Gamma}_L^{\text{cor}}$, which is not well understood. We avoid this problem by modifying the standard Tibshirani, Walther, and Hastie [2001] approach, and use a W_k defined with respect to a set of auxiliary covariates Z, rather than Γ^{cor}.[29]

To see why this simplifies the problem, note that in the model the only source of correlation in insurgent attacks across districts is through ϵ. In particular, our model assumes that if the same insurgent group is present in both districts i and i', the correlation in attacks between districts will not depend on the relationship between any other covariates of i and i': for example, it does not matter whether i is geographically close to i', or geographically distant. We will now add one additional assumption. Suppose that the districts where a given insurgent group is present are less dispersed in terms of these auxiliary covariates Z than a set of randomly chosen districts. For simplicity, we will focus specifically on geography, that is $Z_i = (\text{lat}_i, \text{long}_i)$, but our approach is potentially more general.

Let W_J describe the geographic dispersion of the insurgent group territories when

[28] Specifically, we would need to know how much adding a $J+1$th group should improve model fit, if there are only actually J groups in the data.

[29] We employed the standard Tibshirani, Walther, and Hastie approach described here in a previous version of this paper. The results for Afghanistan were identical as those presented below. However, the results for Pakistan were not as easy to interpret. All results are available on request.

there are J groups, according to the following formula:

$$W_J = \sum_{j=1}^{J} \frac{1}{2N_j} \sum_{i,i' \in I_j} d_{i,i'}^2 \tag{9}$$

where $d_{i,i'}$ is the geographic (Haversine) distance between districts i and i'. As before, I_j is the set of districts where insurgent group J is present, and N_j is the cardinality of this set. W_J follows the "within sum of squares" formula from the analysis of variance literature. The intuition here is that, at numbers of groups beyond the true number of groups J, the additional groups will be based on finite sample noise, which is by assumption uncorrelated with geography. Thus, the additional groups should not be correlated with geography, and thus values of W_J should not decrease any faster than would be expected in the case where groups were randomly assigned.[30] The "random assignment" case, needed for determining $E^*[\log(W_k)]$, can be generated via Monte Carlo permutations of the group structure: randomly reassign geographic coordinates to each of the districts, thereby forcing group membership to be unrelated to geographic location.

One does not have to exclusively rely on the gap statistic to estimate the number of groups J. In fact, this parameter is also recoverable using an entirely different approach, one based on the spectral properties of Γ_L. Notice that each Γ_L^j in (4) has rank 1, implying the rank of Γ_L is J.[31] The rank of Γ_L can be then consistently estimated by applying the intuition of Ahn and Horenstein [2013], using the eigenvalues of Γ_L. Ahn and Horenstein's "eigenratio" approach proceeds as follows. Suppose that we were interested in estimating the rank of Γ_L. Let $\hat{\lambda}_k$ be the $k-$th largest eigenvalue of $\hat{\Gamma}_L$. Asymptotically, the first J of these eigenvalues will be positive and bounded away from zero, while the remaining $N - J$ will go to zero. Ahn and Horenstein consider the "eigenratio"

$$\mathrm{ER}_k = \hat{\lambda}_k / \hat{\lambda}_{k+1}. \tag{10}$$

Asymptotically, ER_k will converge to some positive value c_k for $k < J$. However, it

[30] There may still be a decrease, but it should not be rapid. Eventually W must fall, because $W_N = 0$, with each group containing only one district.

[31] This is because the vectors $\alpha_{\cdot j}$ and $\alpha_{\cdot j'}$ describing insurgent group presence are orthogonal for $j \neq j'$.

will diverge to infinity for $k = J$, as the denominator becomes increasingly close to zero while the numerator remains bounded away from zero. A simple estimate for $\hat{J}$ can then be obtained by choosing the $\hat{J}$ that gives the highest value for $\text{ER}_{\hat{J}}$.[32]

To summarize, in this subsection the specific estimator used is the following. We begin with $\hat{\Gamma}_L$, which approximates a low rank block-diagonal matrix where each block corresponds to the districts where a given organized group is present. $\hat{\Gamma}_L$ is then normalized to obtain the correlation matrix $\hat{\Gamma}_L^{\text{cor}}$, which provides cosine distances that are used for k-means clustering, which determines which group is present in each district. The number of groups to use in this clustering is determined by the gain in fit (defined as log within sum of squares based on a set of covariates Z) associated with various choices for the number of groups, or alternatively by eigenratio methods.

2.4 Potentially Overlapping Insurgent Groups

We now relax the assumption that insurgent groups do not overlap. As in the approach discussed above, we will begin by assuming that J is known, and estimate $\{\alpha_{ij}\}$. We then produce an estimate $\hat{J}$ based on a comparison of these estimates for different values of J. The estimates for $\{\alpha_{ij}\}$ will be based on non-negative matrix factorization, and the $\hat{J}$ estimate will be based on a modification of the eigenratio approach discussed above.

We first construct an estimator for the $\{\alpha_{ij}\}$, given an assumed number of groups J. Consider choosing $\hat{\alpha}_{ij}$ for each district i and group j to satisfy, to the extent possible, the set of restrictions

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

where $\hat{\gamma}_{ii'}$ is the relevant entry in $\hat{\Gamma}_L$, estimated in (3), above. 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, plus the diagonal elements. If there are J groups, there are $N \times J$ parameters to be estimated: one $\hat{\alpha}_{ij}$ for each district i and group j. A necessary condition for identification is thus that $(N+1)/2 \geq J$.

In the data used the number of districts is large relative to plausible numbers of

[32] Ahn and Horenstein [2013] require that there be some exogenous maximum number of possible factors, $J_{\max}$. We follow this, and use $J_{\max} = 80$ for this paper. Simulations provided in Appendix E show why this $J_{\max}$ is necessary.

groups, and thus this inequality holds strictly and a penalty function is required. An obvious estimator for $\{\alpha_{ij}\}$ would then be the squared Frobenius norm

$$\underset{\hat{\alpha}_{ij} \geq 0}{\operatorname{argmin}} \sum_{i} \sum_{i'} \left(\hat{\gamma}_{ii'} - \sum_{j} \hat{\alpha}_{ij} \hat{\alpha}_{i'j} \right)^2 . \tag{11}$$

Unfortunately, solving this optimization problem directly by searching the space of $\{\alpha_{ij}\}$ is challenging because the problem as stated is non-convex in $\{\alpha_{ij}\}$. A variety of algorithms have been proposed for solving this problem. We will use the "Procrustes rotation" algorithm of Huang, Sidiropoulos, and Swami [2014]. This algorithm does not attempt to minimize (11), but instead solves a related optimization problem based on a spectral decomposition of $\hat{\Gamma}_L$. Huang and Sidiropoulos [2014] show that this algorithm is effective at solving (11), despite the fact that this objective is not used as part of the algorithm.[33]

We now consider how to produce an estimate $\hat{J}$. In Section 2.3, the rank of Γ_L was J, because the vectors $\alpha_{\cdot j}$ and $\alpha_{\cdot j'}$ describing insurgent group presence would be orthogonal for $j \neq j'$. This is no longer true if groups have the potential to overlap. Instead of the rank of Γ_L, we thus base our estimate $\hat{J}$ on the completely positive rank of Γ_L: that is, the rank of A, where $\Gamma_L = AA^{\mathrm{T}}$, and all entries of A are non-negative. Without further assumptions this decomposition is not identified: for example,

$$\Gamma_L = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{12}$$

could be decomposed either into

$$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{13}$$

or

$$A = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}. \tag{14}$$

[33] A previous version of this paper optimized (11) directly, using the algorithm of Birgin, Martinez, and Raydan [2000]. The (qualitatively identical) results from this direct approach are available upon request. Huang, Sidiropoulos, and Swami [2014] is orders of magnitude faster, converging in seconds or minutes rather than hours or days.

Huang and Sidiropoulos [2014] summarize assumptions under which the non-negative factorization of Γ_L becomes unique for practical purposes:[34] each factor must have at least $J-1$ non-zero entries, and the non-zero entries of one factor must not be a subset of the non-zero entries of any other factor. In the example above, the second assumption is violated by (14). We will assume that the Huang and Sidiropoulos [2014] assumptions are satisfied. Thus, faced with the covariance matrix in (12), we would conclude that $\hat{J}=1$ based on the factorization in (13).

If Γ_L were known, the number of organized groups could thus be calculated immediately by producing a non-negative factorization of Γ_L. However, only the finite sample $\hat{\Gamma}_L$ is actually available, and in general this matrix will not have a non-negative factorization due to finite sample variation.

To address this problem, we will use a modification of the "eigenratio" approach. In fact, the intuition behind the Ahn and Horenstein [2013] approach appears very general. Consider a rank k approximation to an $N \times N$ matrix. The first k eigenvectors can be used to create such an approximation. How much better is a rank $k+1$ approximation? If the $k+1$th eigenvalue is very small relative to the kth eigenvalue, then considering a rank $k+1$ matrix instead of a rank k matrix does not improve the approximation very much, and ER_k will thus be very high. We can apply this general intuition to the case of the group structure of $\hat{\Gamma}_L$. Let A_k be an approximate non-negative factorization of $\hat{\Gamma}_L$ with k factors. How much better would A_{k+1} be as an approximation to $\hat{\Gamma}_L$? Asymptotically, if Γ_L was produced by k groups, the improvement will be zero.

A ratio equivalent to Ahn and Horenstein's "eigenratio" can then be expressed as

$$\text{NNR}_k = \frac{||\hat{\Gamma}_L - A_k A_k^{\mathrm{T}}||_F^2 - ||\hat{\Gamma}_L - A_{k-1} A_{k-1}^{\mathrm{T}}||_F^2}{||\hat{\Gamma}_L - A_{k+1} A_{k+1}^{\mathrm{T}}||_F^2 - ||\hat{\Gamma}_L - A_k A_k^{\mathrm{T}}||_F^2} \tag{15}$$

where $||||_F$ is the Frobenius norm. The intuition for NNR_k is exactly that of the eigenratio approach: if Γ_L has a completely positive rank of k, then the $k+1$th factor should not help explain Γ_L, and thus NNR_k should diverge to infinity. In contrast, values of NNR_k for $k < J$ will converge to finite values.[35]

The ER estimator has a finite sample tendency to estimate $\hat{J}=1$, because the

[34] Conditions theoretically guaranteeing the uniqueness of the factorization are more complicated: see the references in Huang and Sidiropoulos [2014].

[35] Ding, He, and Simon [2005] describe from a more general perspective the similarities between NNMF and spectral clustering.

eigenvalues of random matrices are generally distributed so that the first few eigenvalues are spaced further apart than most of the remaining eigenvalues.[36] This effect has been noted previously by Ferson and Kim [2012], and Guo-Fitoussi and Darne [2014] perform an extensive simulation-based analysis.[37]

The attack datasets that we consider in this paper, however, are noisier than the data generally used by researchers studying factor models in macro or finance. We are thus particulary interested in the finite sample properties of the estimator when the signal-to-noise ratio is very low. In Appendix E, we provide figures illustrating the behaviour of the eigenratio estimator as the signal vanishes. The simulations that we perform are effectively identical to those conducted in Guo-Fitoussi and Darne [2014], as well as the original monte carlo exercises of Ahn and Horenstein [2013]. To the best of our knowledge, however, the figures that we produce have not previously appeared in the literature: this includes Appendix Figure E.3d, showing the distribution of the eigenratio estimator under the null hypothesis that there is no group structure.

The finite sample behaviour of the eigenratio estimator is shared by estimators using NNR. It will thus be important to check whether the values of NNR obtained might have arisen by random chance from data with no actual group structure. Consider the value of $\max_{k<J_{\max}} \text{NNR}_k$. We wish to compare this test statistic to its distribution under the assumption that there is no actual group structure, obtaining appropriate p-values.[38]

To do so, consider a "reference distribution" where there are no organized groups. Randomly generate attack data based on this distribution, calculate an equivalent to $\hat{\Gamma}_L$ based on this randomly generated data, calculate a value for NNR_k based on this matrix, and then repeat this process 100 times. We consider three different reference distributions: specifics are provided in Appendix F.

[36] Classic references here include the Wigner [1955] semi-circular distribution and the Marchenko-Pastur [1967] distribution.

[37] As Mirza and Storjohann [2014] point out, the effect is visible in the original Ahn and Horenstein [2013] simulations.

[38] Other hypothesis tests are difficult to perform: the distribution of eigenvalues resulting from random variation in finite samples is not obvious. We thus do not report confidence intervals for $\hat{J}$. For similar reasons, we also do not report confidence intervals for $\{\hat{\alpha}_{ij}\}$ below.

2.5 Robustness: potentially changing district environments

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

Let $\bar{\Gamma}_m$ be calculated the same as $\bar{\Gamma}$ from Equation 2, but using only daily attack data from month m. As the number of days of data used to calculate $\bar{\Gamma}_m$ does not increase asymptotically for any given month m, estimation based on a single $\bar{\Gamma}_m$ would be inconsistent. Aggregating across months, however, results in a consistent estimator that is robust to changes in attack probabilities between districts at monthly frequency.

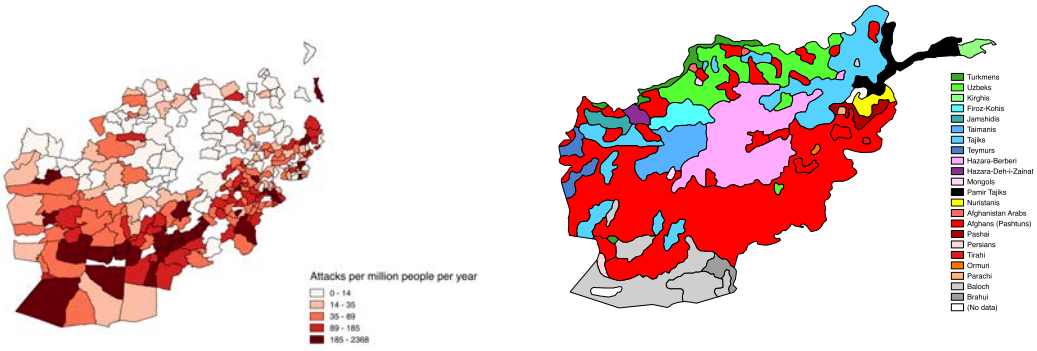
Specifically, assume that the probability of an attack in district i in month m, either from unorganized militants or an organized group, now changes with a parameter ζ_{im}. That is, the probability of an attack from a unorganized militant is now $\zeta_{im}\eta$, and the probability of an attack from member of organized group j is now $\zeta_{im}\epsilon_{jt}$. Let $D(\cdot)$ indicate a diagonal matrix with the given entries on the diagonal. If ζ were known, the standardized matrix $\tilde{\Gamma}_m = D(\frac{1}{\zeta_m})\Gamma_m D(\frac{1}{\zeta_m})$ could be summed to create $\tilde{\Gamma} = D(\sum_m \zeta_m)\tilde{\Gamma}_m D(\sum_m \zeta_m)$. $\tilde{\Gamma}$ could then be used to estimate $\{\alpha_{ij}\}$. In reality, ζ is unobserved; however, dividing by the observed number of attacks creates a feasible estimator, with α identified up to scale. This approach can be employed with both estimation based on clustering and that based on non-negative matrix factorization. Further details are provided in Appendix G.

3 Data

Both Afghanistan and Pakistan were covered by the Worldwide Incidents Tracking System, a discontinued U.S. government database [Wigle 2010].[39] Data is available for location, date, and type of violent incidents from the beginning of 2003 to the end

[39]The data remains accessible online courtesy of the Empirical Studies Of Conflict (ESOC) project at Princeton University.

Figure 1: Afghanistan data



(a) Total attacks per capita

(b) Ethnic Groups

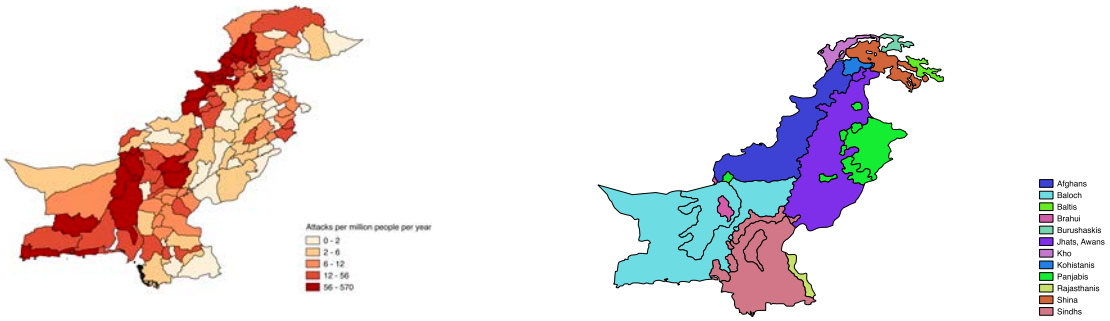
of 2009.[40]

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

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

[40]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.*" And "*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.*"

Figure 2: Pakistan data



(a) Total attacks per capita

(b) Ethnic Groups

According to the data, there are some days where as many as 64 different districts in Afghanistan 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.

For Pakistan, the BFRS dataset [Mesquita et al. 2015] is also available. This is similar to WITS, in that it provides daily data on violent incidents, including geographic information. BFRS data is available until 2011, and over the WITS time frame of 2004-2009, BFRS contains approximately twice as many incidents as WITS. Because of the greater number of attacks recorded, we prefer the BFRS data to the WITS data, We make use only of BFRS data from mid-2008 onwards, because qualitative evidence suggests that the structure of insurgent groups during this 3.5 year period was relatively stable.

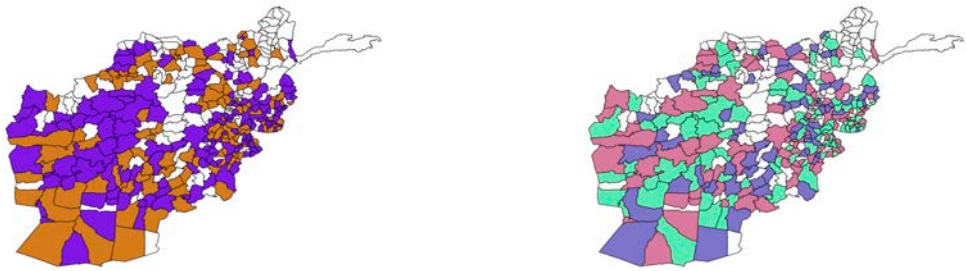
Additional geographic information available for Afghanistan includes the location of roads, rivers, and settlements.[41] We aggregate this data to the district level in order to use it jointly with the district-level attack data. For geographic data on ethnicities, we use the Soviet Atlas Narodov Mira data.[42]

In Figure 1 we show the pattern of attacks by district in Afghanistan and the distribution of ethnicities. The concentration of attacks in the ethnic Pashtun areas

[41] This data was also obtained from ESOC.

[42] The version used is the "Geo-referencing of ethnic groups" (GREG) data set of Weidmann et al. [2010].

Figure 3: Afghanistan groups via spherical k-means



(a) 2 Groups

(b) 3 Groups

is evident. In Figure 2 we report the same information for Pakistan. This data forms the basis for most of the analysis that will be performed in the following section.

We will also perform some additional follow-up analysis regarding Afghanistan. To provide context for this, and help the reader in interpreting maps, we provide additional maps and tables in Appendix A.

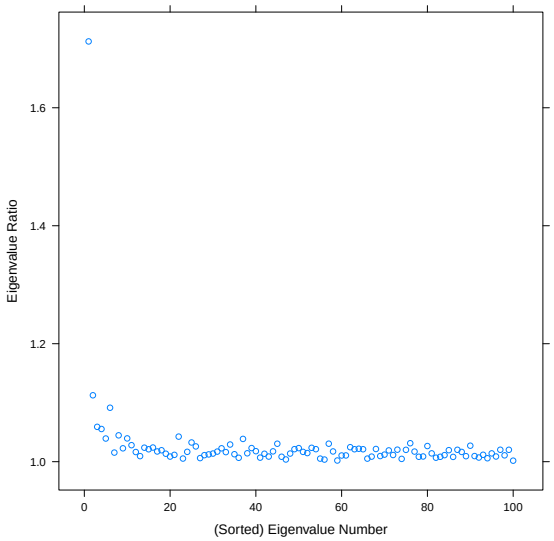
Appendix Figure A.1 shows the main Afghan highways. Even without formal analysis, it is clear that the data confirm two well known qualitative features regarding insurgent attacks in Afghanistan: they are more likely to occur in Pashtun areas, and there is a particular concentration on the ring road highway running south from the capital, Kabul.

Appendix Table A.1 provides a summary of the US Afghan counterinsurgency timeline produced by the Council of Foreign Relations. Evidence of the deterioration of the security environment in Afghanistan is reported in Appendix Figure A.2, which shows the distribution of incidents by district for the 2004-2007 and 2008-2009 periods, in per capita terms.

4 Results

We will first analyze the attack data from Afghanistan, and then consider the case of Pakistan. In both cases, we begin with the spherical k-means clustering and "gap statistic" approach outlined Section 2.3, and then proceed to the non-negative

Figure 4: Eigenratios, Afghanistan



matrix factorization approach of Section 2.4. For all these analyses, we will use attack covariance matrices calculated using only within-month variation, as described in Section 2.5, unless specifically noted otherwise. This is because it is important to avoid contamination by long-term trends in the conflict, as well as seasonal variation. We end this section by further examining the Afghanistan attack data, looking for changes between the earlier part of the 2004-2009 period and the latter.

4.1 Afghanistan

We begin by analyzing the attack data from Afghanistan. Figure 3 shows clustering based on spherical k-means, as outlined in Section 2.3.[43] Qualitatively, the clusters shown in the figure here appear indistinguishable from random noise. We test this using the gap statistic presented in Equation 7. Column I of Table 1 shows the results of this analysis for Afghanistan. We will use this column to provide some further detailed exposition of this approach.

Under the hypothesis of one organized groups of insurgents, the model is without degrees of freedom (all districts must be associated with that group) and hence the geographic variation left unexplained in both the actual data and the permuted

[43]These figures are calculated based on a "within month" covariance matrix, as described in Section 2.5. Results do not change with other approaches.

Table 1: Gap Statistic

			Afghanistan I	Pakistan II
1 group	Randomly shuffled data (mean)		16.765	16.929
	Actual data	-	16.765	16.929
	Gap	A	0.000	0.000
2 groups	Randomly shuffled data (mean)		16.760	16.921
	Actual data	-	16.765	16.897
	Gap	B	-0.005	0.024
	Gap statistic (B minus A)		-0.005	0.024
	Randomly shuffled data (std. dev.)		0.006	0.011
3 groups	Randomly shuffled data (mean)		16.755	16.912
	Actual data	-	16.757	16.736
	Gap	C	-0.002	0.176
	Gap statistic (C minus B)		0.003	0.152
	Randomly shuffled data (std. dev.)		0.008	0.015
4 groups	Randomly shuffled data (mean)		16.749	16.903
	Actual data	-	16.755	16.622
	Gap	D	-0.005	0.281
	Gap statistic (D minus C)		-0.004	0.105
	Randomly shuffled data (std. dev.)		0.011	0.019
5 groups	Randomly shuffled data (mean)		16.745	16.894
	Actual data	-	16.738	16.758
	Gap	E	0.007	0.136
	Gap statistic (E minus D)		0.012	-0.145
	Randomly shuffled data (std. dev.)		0.011	0.021

Each column computes the gap statistic as described in Section 2.3, based on a within-month covariance matrix as described in Section 2.5. Columns differ in the underlying attack data used:
Column I uses the full Afghanistan WITS dataset.
Column II uses the Pakistan BFRS dataset for May 2008 - October 2011.

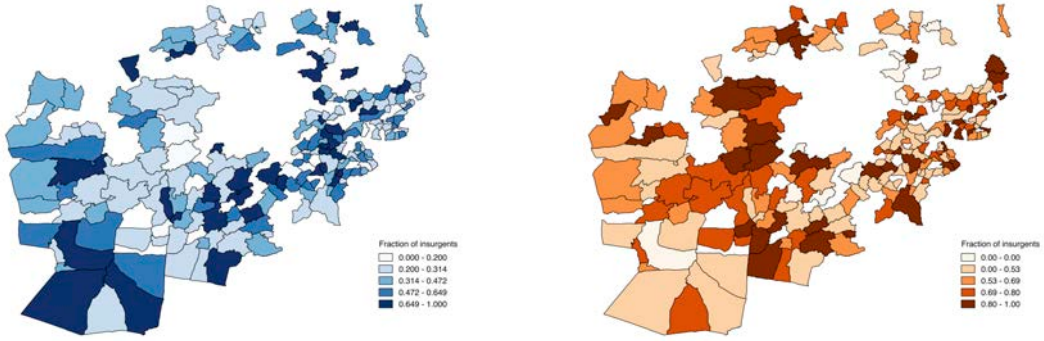
(reference distribution) data is the same (all of it), leaving $\text{Gap}(0) = 0$ in the third row (marked "A"). Allowing for two groups of insurgents in the data leads to free parameters (which districts are associated with each group). However, it turns out that there is actually slightly less geographic variation left unexplained in the randomly permuted data compared to the actual data (16.760 vs. 16.765). This produces $\text{Gap}(2) = -0.005$ (marked "B"). The gap statistic (B-A) is thus -0.005, which, since it is negative, is definitely lower than 0.006, the estimated standard deviation for unexplained geographic variation in the case where districts are randomly assigned to insurgent groups. Thus, $\hat{J} = 1$ satisfies Inequality 8, and we conclude that there is only one insurgent group present in Afghanistan. The eigenratio approach of Ahn and Horenstein [2013] produces an identical result, as illustrated in Figure 4.

The group membership shown in Figure 3 involves a discrete partition of districts into insurgent groups. Some districts, however, might have many insurgents, while other districts might have few. Furthermore, there might be districts where more than one insurgent group is active. The model presented in Section 2.4 allows for these possibilities. An additional advantage of this model is that it provides a test against that null hypothesis that $J = 0$, and all attacks are the result of disorganized local actors. In contrast, the model used in Section 2.3 assumes that there is exactly one organized group present in each district, and thus this model cannot be used to test the hypothesis that there are actually no groups.

We thus move on to the alternate model of Section 2.4. A non-negative factorization of the attack covariance matrix, following Equation 11, turns out to give very similar results to those just discussed above. In contrast to the previous approach, multiple insurgent groups may now be active in any single district, and thus it is no longer possible to easily display the estimated insurgent group structure on a single map. Instead, we produce one map for each group. Figure 5 provides a visualization of this factorization in the case where $J = 2$. Again, there is no discernible pattern to the estimated insurgent groups.

We are particularly interested, however, in whether we can reject the null hypothesis that $J = 0$, and there are no organized insurgent groups present at all. We begin by calculating the ratio described in Equation 15, and choose $\hat{J}$ so as to maximize this ratio. We then consider the distribution that this ratio would have if there were actually no organized groups. To do this, we use the permutation approach described in Section 2.4 and Appendix F.

Figure 5: Afghanistan, 2 groups via NNMF



(a) Group 1

(b) Group 2

Table 2: Estimated number of groups via NNR, Afghanistan

	Not by Month		By Month		
	I	II	III	IV	V
Afghanistan (WITS, Jan 2004 - Sept 2009, weighted)	4	4	1	1	1
(p value, vs. no group structure)	0.57	0.41	0.01	0.01	0.03
Afghanistan (WITS, Jan 2004 - Sept 2009, unweighted)	1	1	1	1	1
	0.02	0.02	0.02	0.03	0.06

Each row presents two estimates of $\hat{J}$, the number of groups present. Columns I and II show the first estimate, described in Section 2.4. Columns III through V show the second estimate, based on the within-month covariance matrix as described in Section 2.5.
In each column, the p values presented are a test of the null hypothesis that there is no group structure. Other tests (e.g. $J = 1$ vs. $J = 2$) appear difficult to construct.
Columns I and III compute p values by comparing to a reference distribution where the time of the attacks within each district has been permuted. See Appendix F for a description of this and other reference distributions.
Column IV is the same as Column III, but the time of attacks is permuted only within each month.
Columns II and V consider only permutations that keep constant the total number of attacks in each district and on each day.

Table 2 shows the results of this analysis for the Afghanistan data. There are four estimates of J provided. Beginning with the first two columns of the first row, $\hat{J} = 4$ in the case where districts are weighted proportional to the number of attacks in the district. Continuing to the next three columns of the first row, $\hat{J} = 1$ if, in addition to this weighting, the covariance matrix is calculated considering only within-month variation in attacks using the approach described in Section 2.5. The second row provides estimates without weighting districts, and results in $\hat{J} = 1$ regardless of whether the approach in Section 2.5 is used or not.

Below each $\hat{J}$ estimate a p value is shown, corresponding to a test of the null hypothesis that in fact there is no group structure, with $J = 0$. We see that in general the null is rejected at the 95% level. The exception is the case where our estimate was $\hat{J} = 4$: with this specification, the model appears to have low power. This analysis supports the results obtained in Table 1, in that there appears to be one organized group of insurgents, rather than more than one. Furthermore, the observed NNR_1 values, calculated according to Equation 15, appear to be more extreme than would be the case if there were no organized groups at all. Table 2 shows that the observed data appears to be inconsistent with $J = 0$, a conclusion that we were not above to draw from the gap statistic results shown in Table 1.

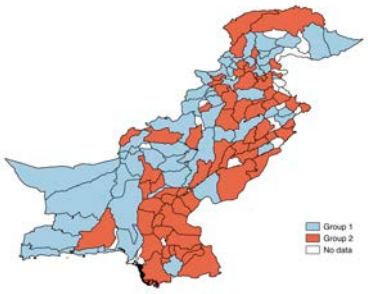
4.2 Pakistan

We now consider the Pakistan attack data. We are particularly interested in the period from mid-2008 until the fall of 2011, as this was a period when the structure of terrorist groups was relatively stable. We begin by clustering Pakistani districts into groups for each $J \in \{1, 2, 3, 4, 5\}$. Our assumption is that a clustering that is due to random noise will not have any particular structure in terms of covariates Z, whereas a cluster that represents an actual insurgent group will feature districts that are more similar in terms of covariates.

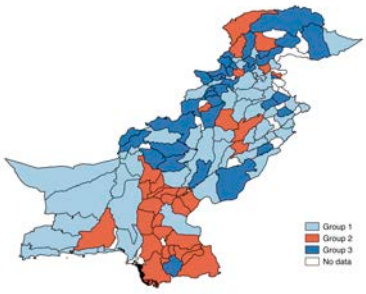
Clusterings based on the Pakistan attack data are shown in Figure 6. Unlike the results for Afghanistan shown in Figure 3, our clusterings for Pakistan, computed on the basis of the attack covariance matrix, result in groups that appear to be clustered geographically. For a more formal analysis, we consider the gap statistic results in Table 1.

Column II of Table 1 reports results for Pakistan for the period of interest. Unlike

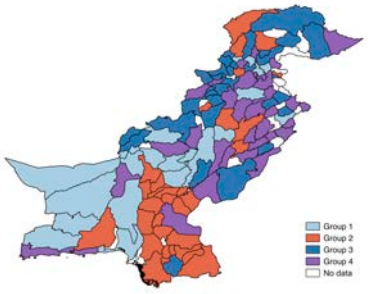
Figure 6: Pakistan groups via spherical k-means



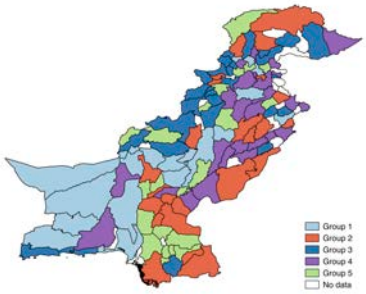
(a) 2 Groups



(b) 3 Groups



(c) 4 Groups

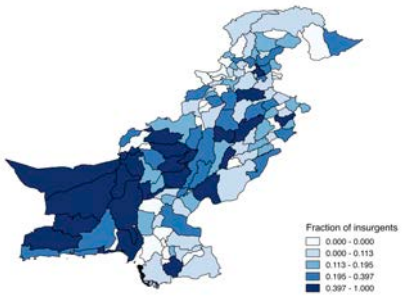


(d) 5 Groups

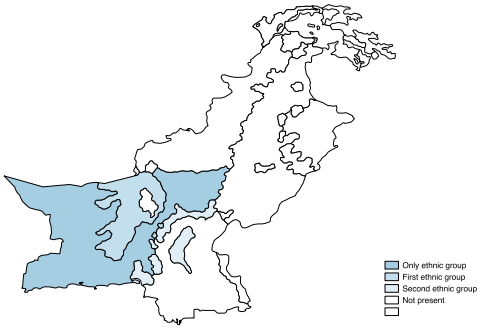
the case with Afghanistan in Column I, a comparison of $J = 1$ with $J = 2$ shows that a clustering with two groups based on the attack covariance matrix is substantially more clustered geographically (with a log within sum of squares of 16.897) than would be expected by random chance (an average log within sum of squares of 16.921). This difference (0.024) is large relative to the standard deviation of clustering based on random noise (0.011), and thus Inequality 8 is not satisfied for $J = 1$.

We thus continue down the rows of Table 1, and proceed to consider the possibility of three organized groups of insurgents. Here again, a clustering based on the actual attack covariance matrix is more geographically clustered than one would expect by random chance. This produces Gap(3) = 0.176 (marked "C"). The gap statistic (C-B) is thus 0.152, which is again higher than 0.015, the estimated standard deviation for $\log(W_3)$. The same situation arises with four groups (where $0.105 > 0.019$). In

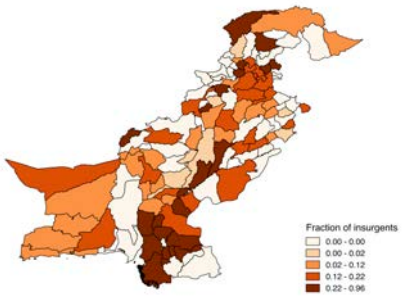
Figure 7: Pakistan, 4 groups via NNMF (left column) and ethnicities (right column)



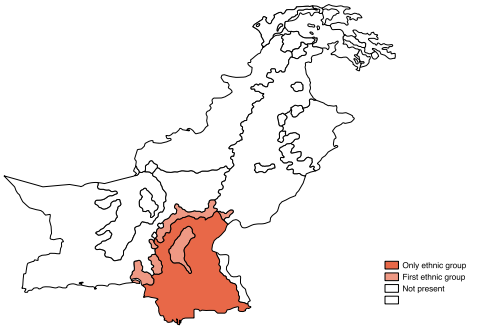
(a) Group 1



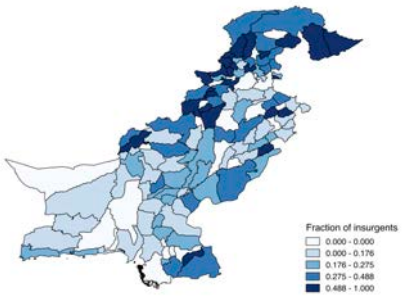
(b) Balochis



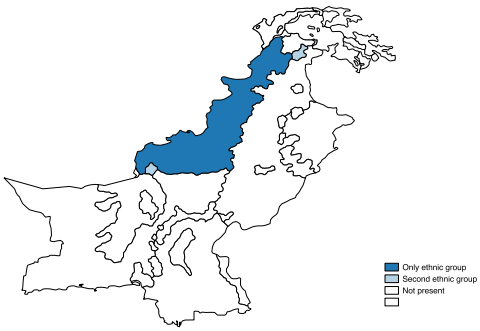
(c) Group 2



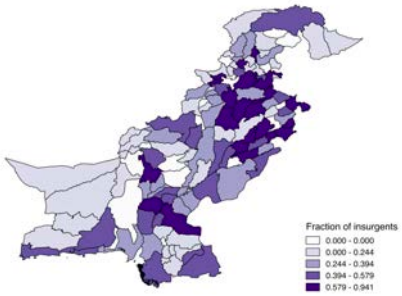
(d) Sindhis



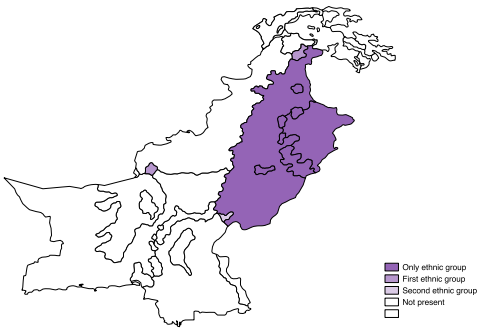
(e) Group 3



(f) Afghans



(g) Group 4



(h) Panjabis, Jhats, Awans

Table 3: Ethnic composition of groups shown in Figure 6c

	Group 1	Group 2	Group 3	Group 4
Baloch	0.62	0.25	0.00	0.12
	(0.09)	(0.09)	(0.10)	(0.10)
Sindhs	0.04	0.87	0.04	0.04
	(0.07)	(0.07)	(0.08)	(0.09)
Afghans	0.12	0.08	0.58	0.23
	(0.07)	(0.07)	(0.08)	(0.08)
Panjabis, Jhats, Awans	0.16	0.12	0.23	0.49
	(0.05)	(0.05)	(0.06)	(0.06)
Other	0.00	0.29	0.57	0.14
	(0.13)	(0.13)	(0.15)	(0.16)
N	115	115	115	115

Each column corresponds to a single regression without intercept.
The dependent variable is a dummy variable indicating whether a given district was clustered into the specified group number in the clustering shown in Figure 6c. Districts shown as white in the figure ("no data") are dropped: the remaining 115 districts are used in the regression.
The independent variables are a set of dummy variables, indicating whether the specified ethnicity was listed as the first ethnicity at the centroid of a given district.
Each row should sum to 1 because each coefficient in the table is a conditional mean giving the fraction of districts of the specified ethnicity that were clustered into the specified group, and the clustering in Figure 6c assigns each district to one group. Rows may not sum exactly to 1 because of rounding.
Standard errors in parentheses.

contrast, adding a fifth group results in a clustering that is not as geographically clustered as when there were only four groups. We thus conclude that $J = 4$ in the case of Pakistan, at least in the 2008-2011 period. The unified insurgent structure ($J = 1$) that we recover for the Afghan case thus appears not to be present in Pakistan. This accords with the qualitative analysis in Dorronsoro [2009].

We now analyze the Pakistan data using the non-negative matrix factorization approach of Section 2.4. As with the Afghan data, we obtain a result generally in line with that obtained using the gap statistic approach just discussed. The left-hand column of Figure 7 shows a non-negative matrix factorization of the attack covariance matrix for Pakistan, using four factors. The result here is very close to that shown in Figure 6c. Furthermore, both of these figures show what appears to be a close

relationship between the estimated group structure and the arrangement of ethnic groups in Pakistan. The relevant breakdown of these ethnic groups is shown in the right-hand column of Figure 7.

A qualitative comparison of the left and right columns of Figure 7 shows that there is one insurgent group present in Balochistan, another in the area populated by Sindhs, a third in the area populated by "Afghans" (i.e. Pashtuns), and a fourth in the Punjabi areas of Pakistan. The northernmost areas of Pakistan, with numerous smaller ethnicities, appear to be associated most closely with the "Afghans".[44] The major ethnic divisions of Pakistan can thus be reproduced successfully using only the covariance matrix of data on insurgent attacks.

Tables 3 and 4 show the relationship between estimated insurgent groups and ethnic groups in a quantitative fashion. These tables are constructed to show the distribution of ethnicities across the estimated groups: each row corresponds to an ethnicity, and each row sums to 100% (plus or minus rounding error). The rows have been ordered so that the diagonal entries correspond to the qualitative relationship between groups and ethnicities just discussed; this is the same ordering of rows that is used in Figure 7.

We now consider an "eigenratio" type analysis of the Pakistan attack data. In the case of the Afghanistan data, analysis based on the gap statistic approach with Table 1 resulted in an estimate of $\hat{J} = 1$, but it was then necessary to use the results shown in Table 2 to show that the null hypothesis of $J = 0$ could be rejected. In contrast, with the Pakistan data, Table 1 gives $\hat{J} = 4$. This result would be extreme under a null hypothesis of $J = 0$, and thus it is not as important to seek alternate confirmation that there is indeed a group structure in the data.[45] This turns out to be fortunate, as the "eigenratio" type analysis shown in Table 5 is inconclusive in the case of the Pakistan data.[46] Very few entries are statistically significant at

[44] Adding a fifth group does not result in these "other" ethnicities being clustered into their own separate group: see Figure 6d. This may be because these areas consist of many small ethnic groups, and there is not a sufficient number of attacks for these smaller groups to be estimated correctly.

[45] In order for Table 1 to lead to an estimate of $\hat{J} = 4$ it must be that for each of two groups, three groups, and four groups, the improvement in geographic clustering is at least one standard deviation better than would be expected if there were no group structure. A result of $\hat{J} = 4$ is thus already very extreme under the null that $J = 0$.

[46] We join a number of other researchers in discarding a low $\hat{J}$ estimate based on eigenratios in favour of other evidence: both Henzel and Rengel [2014] and Alquist and Coibion [2014] discard $\hat{J} = 1$ in favour of two factors, and Bleany et al. [2012] discards $\hat{J} = 1$ or $\hat{J} = 3$ in favour of four or more factors, while Rezitis [2015] discards $\hat{J} = 2$ in favour of five factors.

Table 4: Ethnic composition of groups shown in Figure 7

	Group 1	Group 2	Group 3	Group 4
Baloch	0.58	0.07	0.12	0.23
	(0.05)	(0.04)	(0.06)	(0.06)
Sindhs	0.14	0.35	0.19	0.32
	(0.04)	(0.03)	(0.05)	(0.05)
Afghans	0.15	0.07	0.48	0.30
	(0.04)	(0.03)	(0.04)	(0.05)
Panjabis, Jhats, Awans	0.22	0.11	0.23	0.44
	(0.03)	(0.03)	(0.03)	(0.04)
Other	0.05	0.11	0.61	0.24
	(0.07)	(0.06)	(0.08)	(0.09)
N	115	115	115	115

Each column corresponds to a single regression without intercept, concerning group $j \in \{1, 2, 3, 4\}$.
The dependent variable is $\hat{\alpha}_{ij} / \sum_{j' \in \{1,2,3,4\}} \hat{\alpha}_{ij'}$. This is the fraction of organized insurgents present in a district that are from group j. This data is displayed in the left column of Figure 7, and it is available for the same 115 districts that were analyzed in Table 3.
The independent variables are a set of dummy variables, indicating whether the specified ethnicity was listed as the first ethnicity at the centroid of a given district.
Each row should sum to 1 (up to rounding) by the same argument as in Table 3: each coefficient in the table is a conditional mean for the ethnicity in question, every district is coded as one ethnicity, and the group shares must sum to one.
Standard errors in parentheses.

Table 5: Estimated number of groups via NNR, Pakistan

	Not by Month		By Month		
	I	II	III	IV	V
Pakistan (BFRS, May 2008 - Oct 2011, weighted)	1	1	1	1	1
	0.63	0.63	0.16	0.28	0.55
Pakistan (BFRS, May 2008 - Oct 2011, unweighted)	2	2	16	16	16
	0.73	0.68	0.00	0.01	0.03

Notes: same as Table 2, except with Pakistan data.

the 95% level, and those that are appear to be computational artifacts of some sort, giving very high estimates for $\hat{J}$. We confirm this by rerunning the analysis using the original eigenratio from Equation 10.[47]

4.3 Further Analysis: Afghanistan

Based on Tables 1 and 2, we have estimated $\hat{J} = 1$ for Afghanistan. It is thus not relevant to compute factorizations such as those shown in Figure 7 for Pakistan, because we believe that there is only group. Instead, we will first consider our estimates for how this organized insurgent group is distributed across Afghanistan. We will then extend our analysis by considering potential changes to the prevalence of the members of this organized insurgent group across time. We will use a number of covariates as part of this analysis: Table 6 includes summary statistics for total incidents, ethnic fragmentation, roads, rivers, and settlements by district for Afghanistan.

The approximation in Equation 6 can be used to reveal the latent geographic distribution of the Taliban and, with this, we can investigate the geographic spread of the insurgency. Table 7 shows regression results based on this approximation including a set of ethnic and geographic controls, as well as province fixed effects. The dependent variable used in these regressions is displayed in Appendix Figure A.3.[48]

The coefficient estimates reported in Table 7 should be read as correlates of Taliban presence in each district. Most of these results are not surprising. Ethnicities other than Pashtun (the omitted ethnicity) are substantially less likely to be associated with organized group activity.[49] These include the Hazara, a Shia group hostile to the (Sunni) Taliban, as well as the Tajik and Uzbek communities. These latter groups

[47] This is theoretically less appealing, because it is associated with eigenvectors that could imply a negative presence of insurgent groups in certain districts. However, it is numerically less challenging to compute. Most of these results are statistically insignificant, and are available upon request.

[48] Appendix Figure A.4 shows the estimated prevalence of the single group of organized insurgents under the non-negative matrix factorization approach. A disadvantage of the estimation strategies used in this paper is that they only provide information about the relative prevalence of each organized group across districts. The units reported in Appendix Figure A.3 thus do not have an interpretation in levels: 0 corresponds to no attacks being attributable to organized group members, but the numeric scale of the legend is arbitrary, and it is not possible to interpret the results in terms of "fraction of attacks due to organized groups" without additional assumptions. The numbers reported in the legend are the number of attacks per million people the organized group would have been responsible for if $\sigma^2 = 1$, but this choice is arbitrary.

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

Table 6: 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.

have historically found themselves in conflict with the Taliban, and were participants in the Northern Alliance.

With respect to other characteristics of districts, there is more group activity in districts with more roads, particularly the ring road artery connecting Kabul with other provincial capitals. As in Figure 1a, which shows raw total attacks, Appendix Figure A.3a shows graphically that organized attacks are concentrated in Pashtun-majority areas, and also near the main highway passing through Kabul and other cities.[50] A feature that is apparent in Appendix Figure A.3a, however, that does not show up clearly in the raw attack data of Figure 1a is that there appears to be a substantial organized insurgency operating near the highway north of Kabul, as well as the highway running south from it. This area is not as heavily populated by Pashtuns, and perhaps because of this, the number of total attacks is not as high. The attack covariance matrix, however, reveals that the attacks that did occur appear to exhibit substantial coordination.

4.3.1 Changes in group structure across time

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

A formal model that allows for this structure to change over time appears challenging to develop. An informal analysis of potential changes can be conducted, however, by appropriately splitting the data. Specifically, we create an "early" data set, including only attacks in 2004-2007, and a "late" data set, including only attacks in 2008-2009.[51] The total daily number of attacks is substantially higher in the later period compared to the earlier one.[52] Estimates of the prevalence of organized insur-

[50]Estimates based on non-negative matrix factorization, shown in Appendix Figure A.4a, appear to make it slightly clearer that the majority of organized insurgent activity is on the ring highway passing through Kabul, and that this activity extends to the north as well as the south of Kabul, possibly with the goal of isolating it. Estimates of the prevalence of the organized group can be produced using the method in Section 2.5. As the estimates in this case are effectively based only on variation within months, estimates appear slightly noisier. These estimates are shown in Appendix Figure A.4b. The tendency towards organized insurgent activity along the main highway can still be seen, although it is no longer as clear.

[51]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.

[52]The distribution of these attacks is shown in Appendix Figure A.2. The unified Taliban orga-

Table 7: Dep. variable is sum of off-diagonal entries of cov. matrix for district i

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

Columns I - IV use OLS with dependent variable log transformed

Columns V - VIII use GLM/Poisson allowing for overdispersion

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

gents from the earlier data can be compared to estimates from the later data, yielding a description of how the structure and location of insurgent groups has changed over time.

Appendix Figure A.3b shows an estimate of the number of attacks due to organized insurgent groups in the earlier period, while Appendix Figure A.3c shows this for the later period. Comparing the two figures shows how much gain in territorial control and coordination have been characteristic of the Taliban offensive since 2008.[53] Appendix Figure A.3b 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. Appendix Figure A.3c shows a higher frequency of attacks, and also shows districts in the north with high frequencies of attacks. One example of this is the highway north of Kabul, where now a number of districts with high frequencies of attacks appears.

A statistical analysis of changes in the distribution of attacks does reveal some patterns that are statistically significant and of relevance for current efforts in the management of the conflict. Appendix Table A.3 investigates the correlates of insurgent group control in each district in the early and late periods by stacking the set of districts and employing interactions with POST dummy for the 2008-09 period. Control by the insurgents is measured through the sum of off-diagonal entries of the covariance matrix of attacks for district i according to the approximation in Equation 6.

As in Table 7, the results in Appendix Table A.3 show a clear relationship between ethnicity and simultaneous attacks. Appendix Table A.3 reports coefficients for seven ethnic group dummy variables (indicating the largest ethnic group in the district), with Pashtun as the omitted dummy variable.[54] The coefficient in the POST inter-

nization is detectable in both the early and the complete samples, with one group of insurgents. The coordinated and unified behavior of the Taliban is not a feature developing over time, rather it is present from the onset. This shows that our results are not due purely to the August 2009 presidential election, when there were many attacks on and around election day. While there is substantial evidence that many of these attacks were in fact coordinated by the Taliban, it would be worrisome if the results presented thus far changed drastically when the attacks around the 2009 elections were excluded. These results are available on request.

[53]The colors of the figures are aligned so that the same color indicates the same number of attacks per capita per year, although the "early" and "late" data have a different number of months.

[54]For interpreting the table, recall that the Afghan Taliban are traditionally ethnically Pashtun and follow Sunni Islam. Their historical opposition to Uzbeks, Hazara, and Tajiks – the main ethnic minorities in Afghanistan – is well documented. It is therefore not surprising that our measure of Taliban activity in Afghan districts is negatively correlated with dummy variables indicating

action with the ethnicity variables is generally positive and of magnitude between 25 and 70 percent of the main effect: districts with non-Pashtun ethnicities exhibit relatively greater activity in the later period, indicating a substantial penetration of the Taliban into areas previously outside their reach.[55]

4.3.2 An Oil-Spot Strategy

Krepinevich [2005] discusses a state of the art counterinsurgency doctrine where control is expanded gradually across space. This is sometimes referred to an "oil spot" strategy, as the area controlled expands contiguously like an oil stain.[56] Our methodology allows us to ask a reverse question, regarding the strategy of insurgents: over time, how do attacks expand across space? Do the Taliban appear in completely new and disconnected areas, or do they launch attacks in districts that are adjacent to those that they were operating in previously? We conclude that the Taliban follow an "oil spot" insurgency strategy based on gradual expansion.

We begin by calculating, for each district, the estimated number of attacks by organized groups in adjacent districts, following Equation 6. The results of this calculation are shown in Appendix Figure A.5: the districts where this variable is zero are shown in blue.[57] All but one of these blue districts are also not estimated to have any organized attacks in the later period, as can be seen by comparing Appendix Figure A.3c with Appendix Figure A.5. In particular, there were no attacks in the central part of Afghanistan in the early period, or much of the northeast, and these areas similarly do not have any attacks in the later period. On the other hand, districts immediately adjacent to estimated early Taliban strongholds appear prone to insurgent expansion.

---

non-Pashtun ethnicity.

[55]Consider for instance the case of the Uzbeks in Column II: the main effect is a statistically significant -2.11, indicating a much lower penetration of the Taliban in Uzbek areas in 2004-2007. In the 2008-09 period Uzbek districts are still less likely to experience Taliban activity, but now the coefficient falls by more than half, to -0.76 ($= -2.11 + 1.35$). The distinction between Uzbek and Pashtun districts is thus decreased in the later period. In Column IV, where we exploit within-province variation, the distinction between Uzbek and Pashtun districts disappears completely or even reverses ($0.3 = -1.05 + 1.35$). Although less statistically precise, Tajik and Hazara areas appear to display a similar pattern.

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

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

Appendix Table A.6 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?[58] Columns I-III of Appendix Table A.6 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.[59]

5 Insurgency Organization & Economic Recovery

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

[58]The dummy recoding is used 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 solution in this case.

[59]Based on the definition of organized group attacks in Section 2, there should never be a negative number of attacks attributed to organized group members. Columns IV-VI of Appendix Table A.6 thus present the same analysis using a Poisson GLM model, in order to take this non-negativity into account. 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. An additional advantage of the Poisson model is that districts with few attacks are (correctly) treated as having higher variance relative to mean (weighted least squares could also be used here, but the Poisson model is natural as the underlying attack data is positive integers). The results in Columns IV-VI confirm that there is very little organized insurgent activity in the late period in districts that did not border a district with such activity in the early period. The large coefficient on the ATTACKS_EARLY_ADJACENT indicator variable is due to the fact that the data exhibits "almost" complete separation: if there were zero districts rather than one that saw organized insurgent activity in the late period without any adjacent activity in the early period, the estimated coefficient here would be negative infinity, and it would not be possible to calculate standard errors by standard methods.

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

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

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

The connection of the study of insurgency with economic development comes from this tight link between insurgent strategies and the failure of reconstruction efforts. Understanding the exact nature of the Iraqi insurgency early on in the conflict could have proven crucial in breaking the vicious cycle that Henderson [2005] observes.[60]

Uncertainty about the organization of the insurgency in post-2003 Iraq took several forms. First, there was disagreement regarding the extent to which attacks represented an insurgency at all.[61] There was also confusion regarding its magnitude: as late as the fall of 2004, the U.S. military still attributed 80 percent of attacks to random and not to political violence. Finally, there was heated debate about the organization of the insurgency, once it was clear that one existed.[62] Further complexity in the Iraqi case stemmed from signs of evolution over time, as the New York Times

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

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

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

reported: *"the insurgency was now organized regionally, and that evidence pointed to some planning across regional boundaries".*[63]

The difficulty, and the importance, of understanding the structure of insurgencies is not limited to Iraq. Consider recent Western efforts in Syria: *"Sixteen months into the uprising in Syria, the United States is struggling to develop a clear understanding of opposition forces inside the country, according to U.S. officials who said that intelligence gaps have impeded efforts to support the ouster of Syrian President Bashar al-Assad."*[64]

Beginning with a series of pro-democracy protests in 2011, the situation in Syria quickly escalated into a full-blown civil war that has cost $250,000$ lives and displaced almost 11 million Syrian citizens to the beginning of 2016. In the backdrop of a ethnically and religiously divided population, this conflict quickly displayed a high degree of complexity in the heterogeneity of parties involved [Smith, 2012], including the Syrian state army loyal to Bashar al-Assad, Sunni Syrian rebels, the Islamic State, Jabhat al-Nusra, Kurdish forces, and Hezbollah. Lack of understanding of the structure of the insurgency in Syria has been one of the strongest deterrents to military and humanitarian involvement of Western powers in this conflict [Jenkins, 2014] and slowed down relief efforts.

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

Another relevant case is Libya post-Colonel Gaddafi. This event would require in itself a fully accurate discussion, but as above for Iraq and Syria, we try to provide a basic picture from the perspective of the analysis of multi-group conflicts. After 2011 and the violent overthrowing of the Gaddafi regime, Libya gradually de-

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

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

scended into full-blown factional violence with Islamic State factions jockeying for control of oil rich areas together with two main armed groups: the Tobruk government (elected democratically but in a deeply unstable political environment) and the Muslim Brotherhood-supported General National Congress. To further complicate the picture, other ethnic-based groups, like the Touareg, have also laid claim to certain parts of the former Libyan state. Repeated failures to achieve stable Unity governments and substantial ambiguity in the set of alliances struck among the various groups have severely hindered the pacification response led by the United Nations in the region. While the United Nations and the European Union have been holding off decisive intervention, the east/west divide in the country has been increasingly exacerbating.

6 Conclusions

This paper focuses on the empirical analysis of insurgency, with applications to Afghanistan and Pakistan. Often the only type of data available concerning the level and geographical diffusion of insurgent activity comes from incident-level data on insurgent attacks. However limited such information might be, recent advances in the analysis of the economics of conflict and reconstruction in war zones have been possible thanks to this data.[65]

This paper shows how incident-level data contains information that can be employed to estimate the structure and geographic span of influence of insurgent groups. We present a set of methods to detect unobserved insurgent coalition structures, based on co-occurrences of violent incidents across districts over time. If incidents in two districts occur simultaneously more than would be expected by random chance, then this suggests that these districts share an organized insurgent movement, one capable of cross-district coordination. We then carry out an analysis of the spread and frequency of attacks.

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

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

REFERENCES

[1] **Ahn, S.; Horenstein, A.** (2013) "Eigenvalue Ratio Test for the Number of Factors. "*Econometrica*. 81(3): 1203-1227.

[2] **Alquist, R.; Coibion, O.** (2014) "Commodity-Price Comovement and Global Economic Activity ". NBER Working Paper 20003. March 2014.

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

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

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

[6] **Baurle, G.** (2013) "Structural Dynamic Factor Analysis Using Prior Information From Macroeconomic Theory " *Journal of Business & Economic Statistics*. 31(2):136-150.

[7] **Bleaney, M.; Mizen, P.; Veleanu, V.**2012). "Bond Spreads as Predictors of Economic Activity in Eight European Economies ." University of Nottingham, Centre for Finance, Credit and Macroeconomics (CFCM) Discussion Paper. December 2011.

[8] **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.

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

[10] **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

[11] **Berman, Eli, Aila Matanock.** (2015). "The Empiricists' Insurgency." *Annual Review of Political Science*, Vol 18: 443-464.

[12] **Birgin, E.; Martinez, J.M.; Raydan, M.** (2000). "Nonmonotone Spectral Projected Gradient Methods on Convex Sets." *SIAM J. Optim.*. 10(4): 1196–1211.

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

[14] **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

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

[16] **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.

[17] **Bottegal, G.; Picci, G.** (2015). "Modeling Complex Systems by Generalized Factor Analysis." *IEEE Transactions on Automatic Control* 60(3) : 759–774.

[18] **Brahimi, A.** (2010). "The Taliban's Evolving Ideology." *Working Paper*. LSE Global Governance. WP 02/2010.

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

[20] **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.

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

[22] **Choi, W.G.; Kang, T.; Kim, G.Y.; Lee, B.** (2014) "Global Liquidity Transmission to Emerging Market Economies, and Their Policy Responses" . SSRN Scholarly Paper 2580627. December 2014.

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

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

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

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

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

[28] **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.

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

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

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

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

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

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

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

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

[37] **Ferson, W.; Kim, M.** (2012) "The factor structure of mutual fund flows. " *International Journal of Portfolio Analysis and Management,* 1(2), 112-143.

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

[39] **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.

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

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

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

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

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

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

[46] **Guo-Fitoussi, L.; Darne, O.**. (2014) "A Comparison of the Finite Sample Properties of Selection Rules of Factor Numbers in Large Datasets". HAL Working Paper hal-00962247. March 2014.

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

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

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

[50] **Henzel, S.; Rengel, M.** (2014) Dimensions of Macroeconomic Uncertainty: A Common Factor Analysis. . SSRN Scholarly Paper 2507743.

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

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

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

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

[55] **Hornik, K.; Feinerer, I.; Kober, M.; Buchta, C.** (2012). Spherical k-Means Clustering. *Journal of Statistical Software*. 50(10).

[56] **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.

[57] **Huang, K., Sidiropoulos, N.** (2014) "Putting nonnegative matrix factorization to the test: a tutorial derivation of pertinent cramer-rao bounds and performance benchmarking." *IEEE Signal Processing Magazine*. 31(3):76–86.

[58] **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.

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

[60] **Jenkins, Brian Michael** (2014) The Dynamics of Syria's Civil War, *RAND Perspectives no. 115.*

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

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

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

[64] **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.

[65] **Krishna, K.; Narasimha, M.** (1999). "Clustering High Dimensional Data: A Survey on Subspace Clustering, Pattern-based Clustering, and Correlation Clustering". *Trans. Sys. Man Cyber. Part B.* 29 (3): 433–439.

[66] **Ledermann, W.** (1940). "On a Problem concerning Matrices with Variable Diagonal Elements." *Proceedings of the Royal Society of Edinburgh.* 60(1). 1–17.

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

[68] **Ludhianvi, M.R.** (2015). *Obedience to the Amir: An early text on the Afghan Taliban Movement.* Trans. Y. Mitha and M. Semple. Berlin: First Draft Publishing.

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

[70] **Marchenko, V.A.; Pastur, L.A.** (1967). "Distribution of Eigenvalues for Some Sets of Random Matrices". *Matematicheskii Sbornik.* 72 (114): 507–536

[71] **Mirza, H.; Storjohann, L.** (2014). "Making Weak Instrument Sets Stronger: Factor-Based Estimation of Inflation Dynamics and a Monetary Policy Rule." *Journal of Money, Credit and Banking*. 46(4): 643-664.

[72] **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.

[73] **Munir, M.** (2011). "The Layha for the Mujahideen: an analysis of the code of conduct for the Taliban fighters under Islamic law ". *International Review of the Red Cross*. Vol. 93 No. 881.

[74] **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.

[75] **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.

[76] **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.

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

[78] **Pak Institute for Peace Studies** (2016) *Pakistan Security Report 2015*.

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

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

[81] **Rezitis, A.N.** (2015). "Empirical Analysis of Agricultural Commodity Prices, Crude Oil Prices and US Dollar Exchange Rates Using Panel Data Econometric Methods." SSRN Scholarly Paper 2631534. July 2015.

[82] **Saunderson, J.; Chandrasekaran, V.; Parrilo, P.; Willsky, A.** (2012). "Diagonal and Low-Rank Matrix Decompositions, Correlation Matrices, and Ellipsoid Fitting." *SIAM Journal on Matrix Analysis and Applications.* 33(4): 1395–1416.

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

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

[85] **Smith, B.** (2012). "Syria: No End in Sight?" *House of Commons Library.* Research Paper 12/48.

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

[87] **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

[88] **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.

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

[90] **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

[91] **United Nations** (2016) "Humanitarian Response Plan - Syrian Arab Republic" United Nations Office for the Coordination of Humanitarian Affairs.

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

[93] **Weidmann, N.; Rod, J.K.; Cederman, L.E.** (2010) "Representing ethnic groups in space: A new datset." *Journal of Peace Research*, 47(4), 491–499.

[94] **Wigner, E.P.** (1955) "Characteristic Vectors of Bordered Matrices With Infinite Dimensions". *The Annals of Mathematics*, 62(3), 548–564.

[95] **Wu, Y.; Moon, H.R.; Deng, Y.** (2011) "Factor Analysis on US Housing Price Indexes." USC Lusk Center Working Paper.

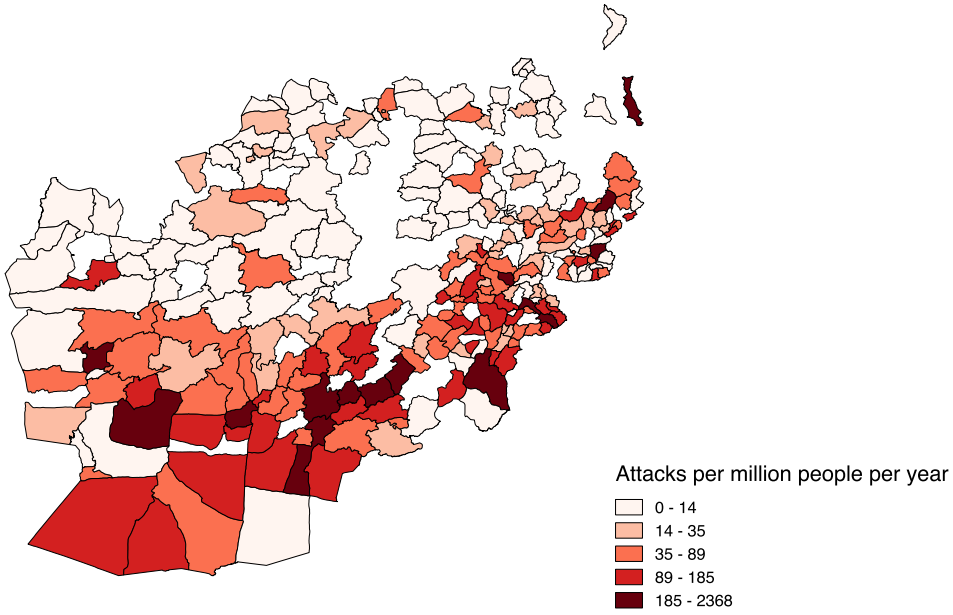
[96] **Yao, J., Zheng, S., and Bai, Z.** (2015) *Large Sample Covariance Matrices and High-Dimensional Data Analysis*, Cambridge University Press.

A Appendix Figures and Tables – Not For Publication

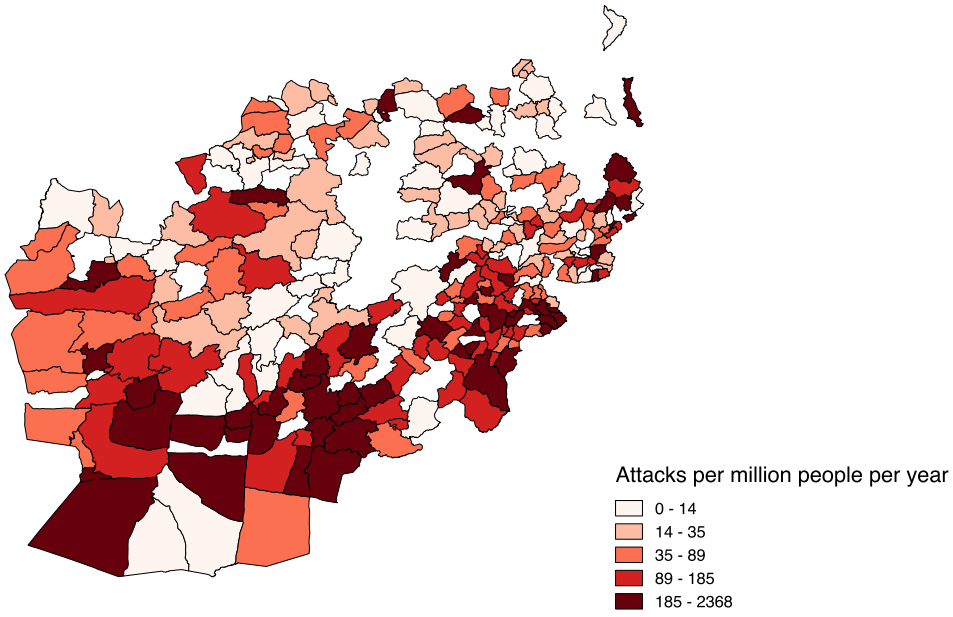
Appendix Figure A.1: Afghan Ring Road

(via Wall Street Journal)

Appendix Figure A.2: Afghanistan attacks by time period

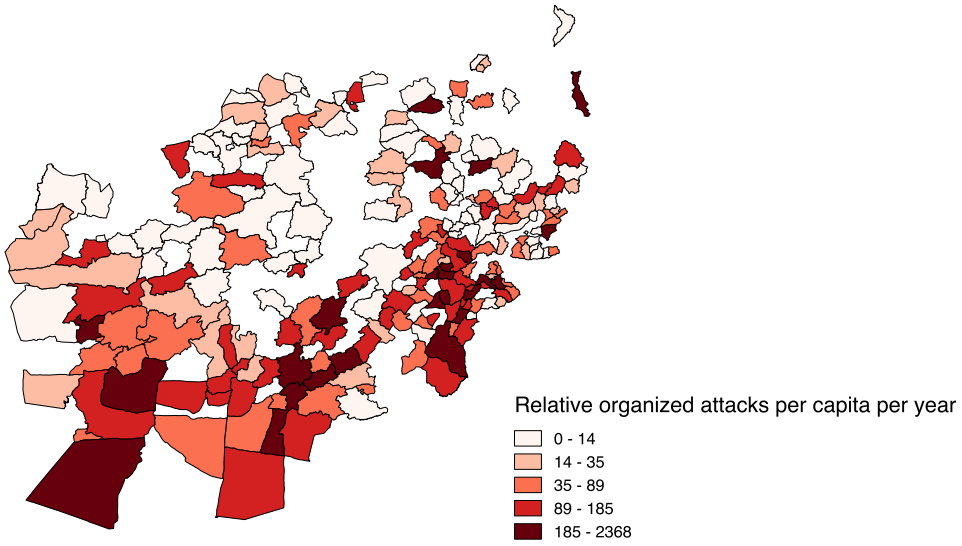


(a) Attacks per capita 2004-2007

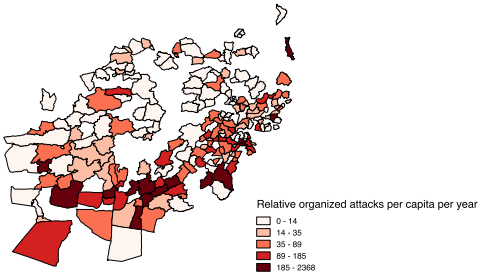


(b) Attacks per capita 2008-2009

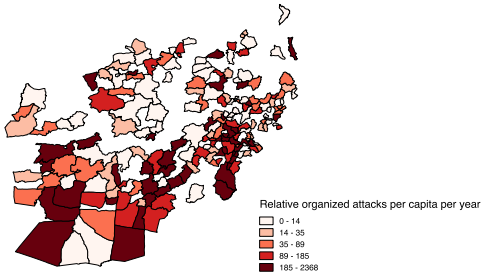
Appendix Figure A.3: Organized group members (Equation 6 approximation)



(a) full dataset (2004-2009)

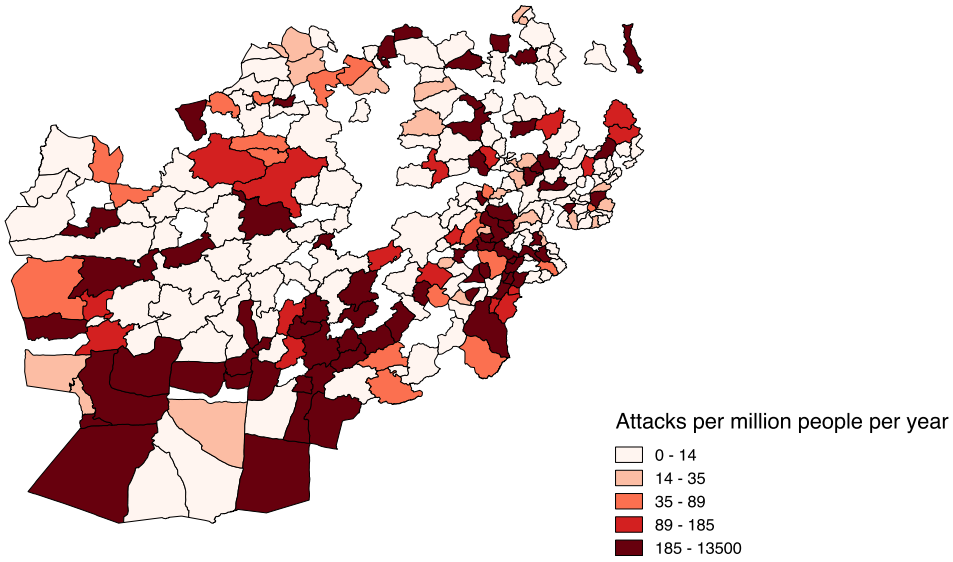


(b) 2004-2007 period only

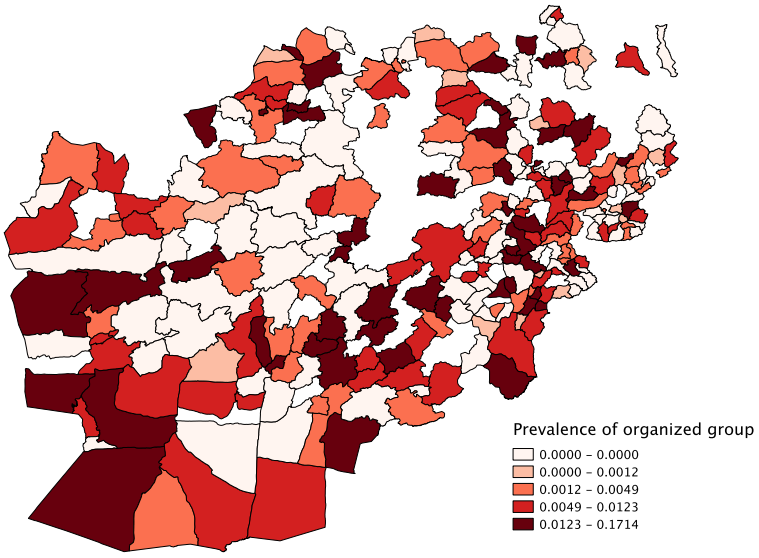


(c) 2008-2009 period only

Appendix Figure A.4: Organized group members: NNMF

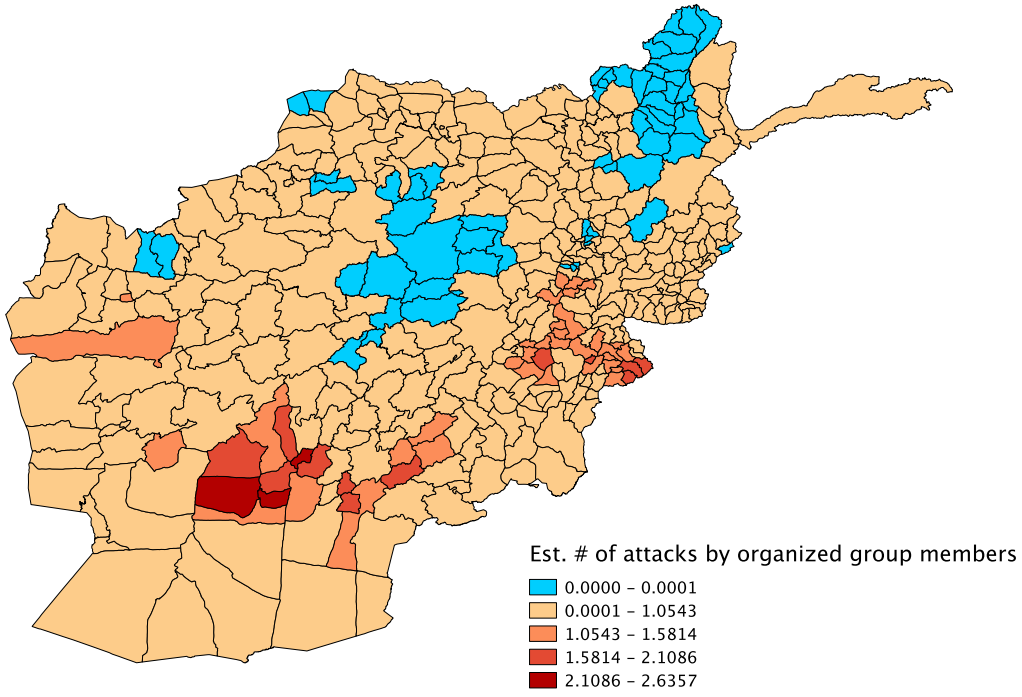


(a) "Not by month"



(b) "By month" covariance matrix (Section 2.5)

Appendix Figure A.5: (Estimated) attacks by organized group members (2004-2007, average over adjacent districts)



Appendix Table A.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

Appendix Table A.2: Dependent variable is total attacks for district i

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

Columns I - IV use OLS with dependent variable log transformed

Columns V - VIII use GLM/Poisson allowing for overdispersion

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Appendix Table A.3: Dep. var. is sum of off-diag. entries of cov. mat. for 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			60−0.18	−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.

Appendix Table A.4: 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

Appendix Table A.5: 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			62 −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.

Appendix Table A.6: Estimated Organized Attacks, 2008-2009

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

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

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

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

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

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

Appendix Table A.7: Estimated Organized Attacks, '08-'09 (no attacks in '04-'07)

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

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

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

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

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

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

Robust standard errors in parentheses

* indicates significance at $p < 0.05$

B Recoverability of Low-Rank Matrix

We are interested in the conditions under which the $\hat{\Gamma}_L$ resulting from (3) will be a consistent estimator for Γ_L. It is clear that there are some matrices Γ_L for which the proposed method will be inconsistent:

Example 1. *Suppose that there are three districts, and two groups. Group memberships are* $\alpha_{\cdot 1} = (1, 0, \delta)$ *and* $\alpha_{\cdot 2} = (0, 1, \delta)$*, and thus*

$$\Gamma_L = \begin{bmatrix} 1 & 0 & \delta \\ 0 & 1 & \delta \\ \delta & \delta & 2\delta^2 \end{bmatrix}$$

for some small value δ*. Suppose that there are disorganized insurgents such that* $\Gamma_D = I_3$*. The minimum trace heuristic of (3), will then give an estimate*

$$\hat{\Gamma}_L = \begin{bmatrix} \delta & 0 & \delta \\ 0 & \delta & \delta \\ \delta & \delta & 2\delta \end{bmatrix}$$

which has lower trace than the true Γ_L *so long as* δ *is small.*

It is thus important to provide conditions for the matrix Γ_L such that the proposed method gives a consistent estimator. Saunderson et al. [2012] give such a characterization. First, Saunderson et al. [2012] define a subspace $\mathcal{U}$ as realizable if, for any Γ_L having column space $\mathcal{U}$, and any Γ_D, the minimum trace factorization algorithm of (3) applied to $\Gamma = \Gamma_D + \Gamma_L$ returns $\hat{\Gamma}_L = \Gamma_L$. Next, they define the "coherence" $\mu(\mathcal{U})$ of a subspace $\mathcal{U}$ of $\mathbb{R}^n$ as

$$\mu(\mathcal{U}) = \max_{i \in \{1,2,\ldots n\}} ||P_{\mathcal{U}} e_i|| \tag{16}$$

where e_i are the standard basis vectors, and $P_{\mathcal{U}}$ is the orthogonal projection matrix onto $\mathcal{U}$. They then provide the following sufficient condition:

Theorem 2 (Saunderson et al. 2012)**.** *If* $\mathcal{U}$ *is a subspace of* $\mathbb{R}^n$ *and* $\mu(\mathcal{U}) < 1/2$*, then* $\mathcal{U}$ *is realizable.*

From an intuitive perspective, this restriction on coherence is equivalent to nothing in the column space of Γ_L being too close to the standard basis vectors. In the context of estimating insurgent groups, the standard basis vectors represent groups that are only present in one district. It makes sense that groups of this sort will result in the procedure in (3) being inconsistent: a group that is only present in one district is indistinguishable from disorganized insurgents, as they both only appear in the diagonal entries of the covariance matrix.

Saunderson et al. [2012] also provide a further result, regarding the "realizability of random subspaces". They argue that "most" subspaces of dimension less than $n/2$ are realizable. The intuition here appears to be that a random subspace of low dimension is unlikely to include anything close to a standard basis vector. In general, then, if the number of groups is small relative to the number of districts, the heuristic given in (3) will provide a consistent estimator for the group structure. Cases where the estimator will not be consistent are those where one of the groups is overwhelmingly located in a single district.

C Spectral Clustering Estimator

Spectral clustering is based on the "graph Laplacian" matrix

$$L = D - \Gamma_L \tag{17}$$

where D is a diagonal matrix with entries equal to the row sums of Γ_L. The graph Laplacian thus 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 graph Laplacian L has a rank of $N-J$, and thus has J zero eigenvalues.[66] Spectral clustering focusses on the number of zero eigenvalues for the associated graph Laplacian matrix

[66]The number of zero eigenvalues of the graph Laplacian matrix corresponds to the number of connected components of the weighted undirected graph described by the adjacency matrix Γ_L. This is J, the number of blocks of Γ_L.

The intuition for this result is relatively straightforward. Each Γ_L^j block has rank one. The corresponding block of the diagonal matrix D has full rank. Setting the entries in this diagonal matrix so that rows and columns of the graph Laplacian L sum to zero ensures that the rows (and columns) of L corresponding to each Γ_L^j block are linearly dependent. The Γ_L^j block that was subtracted, however, is only rank one, and thus the null space of the resulting block of L must be rank one. This is true for every block in L, and thus the null space of L has dimension J. This will also be the number of zero eigenvalues of L.

L, whereas the method used in the main text produces an estimate $\hat{J}$ of the number of insurgent groups by examining (in a very broad sense) the rank of Γ_L.

If Γ_L were known, the number of organized groups could be calculated immediately, and it would equal both the rank of Γ_L and 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 a noisy $\hat{\Gamma}_L$ must be used instead of the true Γ_L. The simplest option for actually implementing a spectral clustering approach is to use a modification of Shi and Malik [2000]: use $\bar{\Gamma}$ to construct $\bar{L}$, and then count the "zero" eigenvalues of $\bar{L}$.

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

A similar problem affects the approach presented in the main text, which is based on using the the largest eigenvalues (or other components) of $\hat{\Gamma}_L$. Finite sample variation will also affect these eigenvalues. The question thus arises whether it is better to use $\hat{\Gamma}_L$ directly, or instead use the corresponding graph Laplacian matrix L. Direct use of $\hat{\Gamma}_L$ requires confidence that the trace minimization algorithm in (3) will work well in finite samples, while use of L avoids this issue because the diagonal entries in question are subtracted away and thus are irrelevant. On the other hand, using L requires labelling some eigenvalues as "zero" eigenvalues, despite the fact that due to random noise all eigenvalues will probably be non-zero.[67] A particular concern here is that the eigenvalues in question are the smallest out of N eigenvalues. Monte Carlo exercises (available upon request) suggest that the approach based on using $\hat{\Gamma}_L$ directly has better finite sample performance. We thus use this approach in our analysis, as described in the main text. Below, we briefly discuss how the alternative

[67]Eigenvalues that would be zero asymptotically will not be zero in a finite sample, because some of the entries that are zero in Γ_L will be positive in the calculated $\hat{\Gamma}_L$. 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.

approach (based on the smallest eigenvalues of the graph Laplacian) might be applied.

A heuristic method is available based on "eigengaps" similar to those used by Ng, Jordan, and Weiss [2002]. Sort the eigenvalues λ of L in increasing order, such that λ_1 is the smallest and λ_N the largest.[68] The difference $\lambda_{k+1} - \lambda_k$ is defined the kth eigengap. Ng, Jordan, and Weiss [2002] argue that a large eigengap indicates that perturbation of the eigenvectors of L would not change the clusters produced by spectral clustering. Luxburg [2007] thus suggests that the right choice for $\hat{J}$ is a number such that λ_k is "small" for $k \leq \hat{J}$, and the $\hat{J}$th eigengap is large.[69] The intuition here is that if there truly are $\hat{J}$ eigenvalues that are zero, then these appear to be non-zero in the finite sample only due to random variation. In contrast, the $\hat{J}+1$th and larger eigenvalues would be strictly positive even if the true L were used. An examination of the $\hat{J}$th eigengap thus provides a heuristic test of whether the choice of $\hat{J}$ was reliable, or whether small changes due to random variation might result in a different number of zero eigenvalues.

Using this approach, the estimated $\hat{J}$ corresponds to an eigenvalue such that λ_k is "small" for all $k \leq \hat{J}$. The presence of high eigengaps for very high values of k is not relevant for the eigengap procedure, so long as $J_{\max}$ is lower than these values. Luxburg [2007] suggests that the cutoff between "small" and "large" should not be larger than the minimum degree in the graph. 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 in the actual Afghanistan attack data is larger, and this difference is statistically significant at the 95% level.

More formal tests could also be constructed. Each off-diagonal $\bar{\gamma}_{ii'}$ entry will converges to $\gamma_{ii'}$ as the number of time periods grows, and the $\bar{\Gamma}_L$ matrix will converge to Γ_L. Thus, $\bar{L}$ will converge to L. Asymptotically, the correct number of the sample

[68] A first step to dealing with the problem of finite sample is to exclude districts with very few attacks from estimation: for the analysis of the Afghan data, we used data only for those districts in which there were 3 or more attacks (other cutoffs yielded similar results). This approach does not fully solve the underlying issue, however. For simplicity the notation here assumes that no districts are excluded on this basis and thus there are still N districts, and N eigenvalues.

[69] The underlying difficulty here is determining what exactly constitutes a "zero" eigenvalue, when there is finite sample variation. The presence of a large eigengap would thus provide some confirmation that an appropriate definition of "zero" has been chosen.

eigenvalues of $\bar{L}$ will approach zero. Thus, 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.[70]

Unfortunately, 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 there 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. This suggests that a formal test following this approach might be difficult to implement.

D NNMF Consistency

Conditions under which $\hat{\Gamma}_L$ will converge to Γ_L have been discussed in Appendix B. We now consider conditions under which a non-negative matrix factorization of Γ_L will recover the $\{\alpha_{ij}\}$ group structure. It is clear that the index numbering of the groups cannot be recovered, because Γ_L is invariant to relabelling of groups. The index numbering of groups is irrelevant throughout our analysis, however, and thus we are only concerned with whether the group structure can be recovered up to a reindexing.

The question of recoverability can be broken down into two parts. First, is the non-negative factorization of Γ_L unique? Second, what is required for the approach for estimating the number of groups, $\hat{J}$ to be consistent? We consider these questions in turn.

Huang, Sidiropoulos, and Swami [2014] discuss uniqueness of symmetric non-negative factorizations at some length. They conclude that while there are no obvious

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

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_L matrices considered in this paper would generally be expected to have a relatively sparse factorization.

We now consider the approach for generating an estimate $\hat{J}$ for the number of groups. Suppose that the true number of groups is J. If clustering is performed with $J+1$ groups, the arrangement of these groups will be based on finite sample variation in the attack covariance matrix. In expectation, $\text{Gap}(J) \geq \text{Gap}(J+1)$, but in any given finite sample this may not be true. The method of estimating $\hat{J}$ given in the main text, then, is not consistent as written, because regardless of the sample size, it will sometimes be the case that $\text{Gap}(J) < \text{Gap}(J+1)$, and changes to the "rule of thumb" s_{J+1} cannot rectify the situation.

In general, the "gap statistic" literature focuses on clustering the sample actually available, and does not emphasize consistency of the proposed estimators. In the case considered in this paper, consistency would require two modifications to basic model presented. First, N would have to grow at some slow but non-zero rate: the easiest way of handling this would be via infill asymptotics, where the country in question were divided into progressively smaller districts. Second, the clustering method employed should be hierarchical. This is because if number of groups is below the true value of J, multiple groups will be clustered together. The way this occurs will be determined by finite sample variation. It could be that at $J/2$, geographically close groups are placed together, while at $J/2+1$, geographically distant groups are placed together. Using a hierarchical clustering method would prevent this possibility. We do not use such a method because implementation would be computationally more challenging, and the problem just discussed does not appear to occur in the data we use.[71]

[71] There are of course many other potential approaches. For example, at large sample sizes, one could progressively revert to the "standard" Tibshirani, Walther, and Hastie [2001] setup, where the gap statistic is calculated with respect to the same Γ_L distance matrix that is used to perform the clustering. In this case, 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 would then be 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} would be 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

E Eigenratio type estimators: simulations

To better understand the finite sample properties of eigenratio type estimators, we conduct a series of simulations. For simplicity, we do not use a model with discrete attacks, as presented in Section 2, but instead use a more standard model with normally distributed random variables. Let there be $J = 4$ groups, $N = 100$ districts, and $T = 2000$ days. Let there be exactly one group in each district, with $\alpha_{i1} \sim \text{Uniform}(0, 1)$ i.i.d. for $i \in \{1, ..., 25\}$, and no other group present in those districts. In the same fashion, only Group 2 is present in districts 26-50, only Group 3 in districts 51-75, and only Group 4 in 76-100.

Our simplified model of attacks is that in each period t for each group j, an i.i.d. draw $\epsilon_{tj} \sim N(0, \sigma^2)$ is made. The number of attacks is then given by

$$x_{it} = \sum_j \alpha_{ij}\epsilon_{tj} + u_{it} \tag{18}$$

where $u_{it} \sim N(0, 1)$, i.i.d.

We then consider eigenvalues associated with the (N by N) covariance matrix of attacks. We perform 100000 simulations for each of $\sigma^2 = 1$, $\sigma^2 = 0.1$, $\sigma^2 = 0.05$, and $\sigma^2 = 0$, generating a total of 400000 simulated sample covariance matrices.[72]

Figures E.1 - E.3 graphically display the results of these simulations. Figure E.1 shows the eigenvalues of the covariance matrix. We see that the group structure is immediately apparent at $\sigma^2 = 1$, still clear at $\sigma^2 = 0.1$, but somewhat unclear at $\sigma^2 = 0.05$. There is no group structure with $\sigma^2 = 0$, and thus Figure E.1d shows the distribution of eigenvalues under $J = 0$.

Figure E.2 shows eigenratios, with the leftmost eigenratio being the ratio between the largest (i.e. leftmost) and second-largest eigenvalues, and so forth. Here, on average the largest eigenratio clearly corresponds to $J = 4$ when σ^2 is large, but this is no longer the case with $\sigma^2 = 0.05$. Figure E.2d shows that the distribution of eigenvalues when $J = 0$ leads to a somewhat peculiar distribution of eigenratios: the

a rate slower than $\sqrt{T}$. The probability of an incorrect selection of $k = J + 1$ or higher number of groups would then decrease to zero asymptotically, and the probability of $k < J$ being selected would similarly decrease.

[72]Note that in the main text, the choice of $\sigma^2 = 1$ is a normalization, because the $\{\alpha_{ij}\}$ are unknown, and a decrease in the choice of σ^2 would simply result in higher $\hat{\alpha}$ estimates. In contrast, in the simulations in this appendix, the distribution of the $\{\alpha_{ij}\}$ are given, and thus choosing a different value σ^2 changes the signal to noise ratio for the attack covariance matrix.

first few and last few eigenratios are much larger than the others. Figure E.2d thus illustrates why it is important to have some maximum number number of possible groups, $J_{\max}$. The eigenratios associated with the very smallest eigenvalues (towards the right hand side of Figure E.1d) become quite large. With $N = 100$, and no $J_{\max}$, choosing $\hat{J}$ based on the largest of all the eigenratios would lead to many $\hat{J}$ estimates of 99 groups. However, as noted in Ahn and Horenstein [2013], any intermediate choice of $J_{\max}$ is unlikely to affect the results.

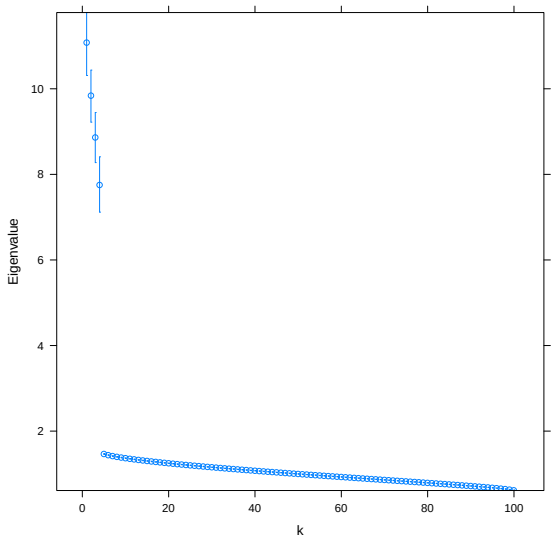
Figure E.3 shows the distribution of estimates $\hat{J}$ with $J_{\max} = 50$. Figures E.3a and E.3b show that the eigenratio approach works very well when the signal to noise ratio in the covariance matrix is relatively high. Figure E.3c, however, shows that with a noisier covariance matrix, the estimated values for $\hat{J}$ tend to be too low. Figure E.3d shows the distribution of estimates of $\hat{J}$ when there is no group structure.

In both of Figures E.3c and E.3d, $\hat{J} = 1$ is the modal estimate. Figure E.3d shows that the median estimated $\hat{J}$ is below the true value $J = 4$ (the mean is above, but this is less apparent from the figure). However, Figure E.3d shows the case with no group structure at all, and thus would not change regardless of the true value of J. The bias of the estimator thus cannot be signed: this is a natural result of J and $\hat{J}$ both being integers bounded between 0 and 50. Bias correction appears to be non-trivial.

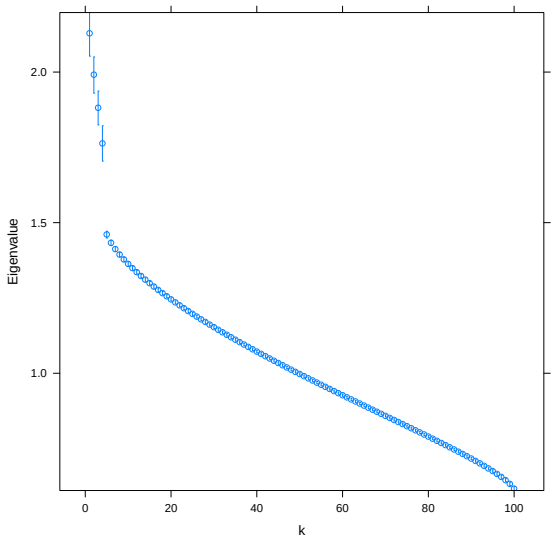
Figure E.3c provides a possible explanation for why estimates of $\hat{J} = 1$ appear so frequently in Table 5. The finite sample properties of eigenratio type estimators are such that there is a tendency to estimate low values of $\hat{J}$ in cases where the covariance matrix is noisy. This is due to the distribution of eigenvalues resulting from the noise, as shown in Figure E.1d. The evidence provided in Table 5 should thus mainly be taken as an indication that the null hypothesis of no group structure should be rejected. Figure E.3c shows how estimates $\hat{J} = 1$ occur frequently when there is actually a group structure with $J > 1$.[73]

[73] In the empirical literature, "low" estimates for the number of factors (compared to other methods) are obtained by Choi et al. [2014] and Wu et al. [2011]. Figures E.3b and E.3c appear in line with results reported (using actual data) in the supplement to Baurle [2013].

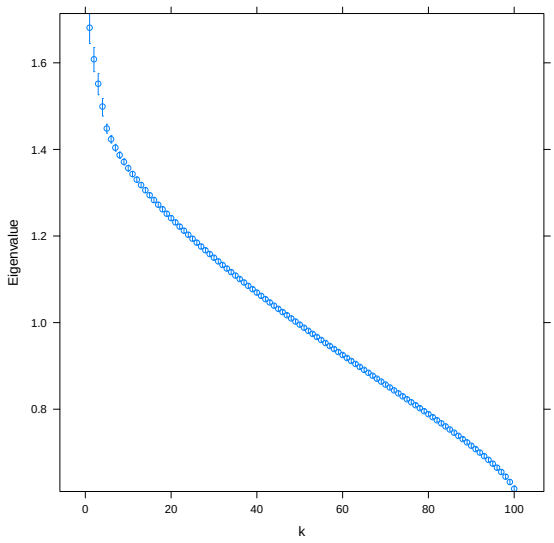
Appendix Figure E.1: Eigenvalues



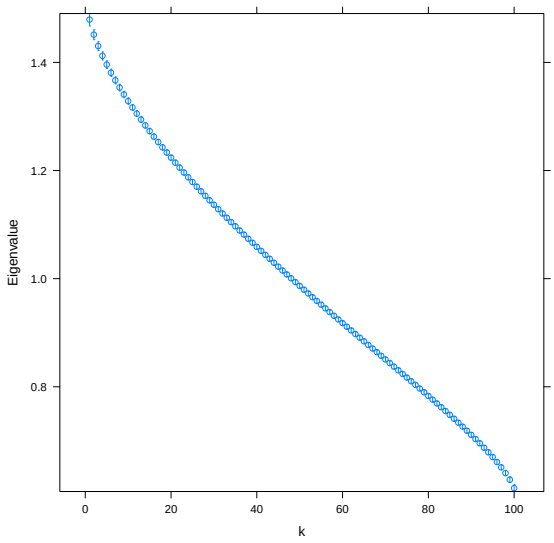
(a) $\sigma^2 = 1$



(b) $\sigma^2 = 0.1$



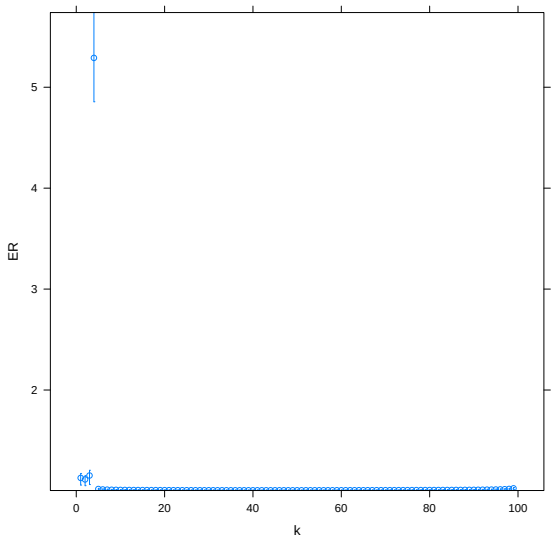
(c) $\sigma^2 = 0.05$



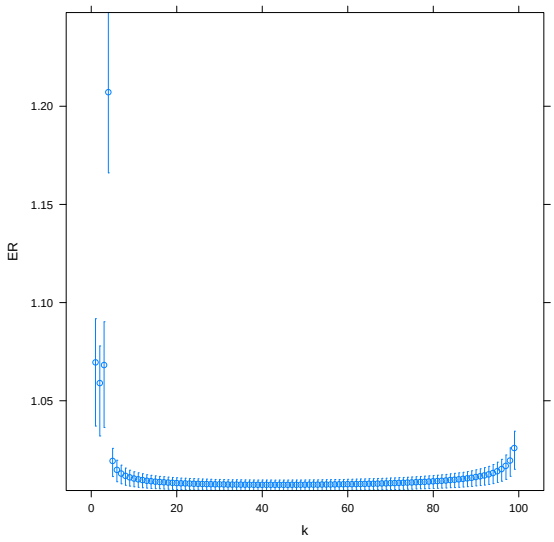
(d) $\sigma^2 = 0$

Points indicate means over 100000 simulations. Bars show interquartile range.

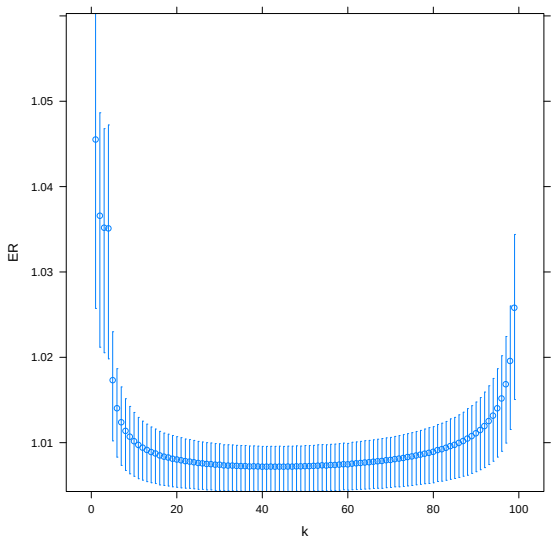
Appendix Figure E.2: Eigenratios



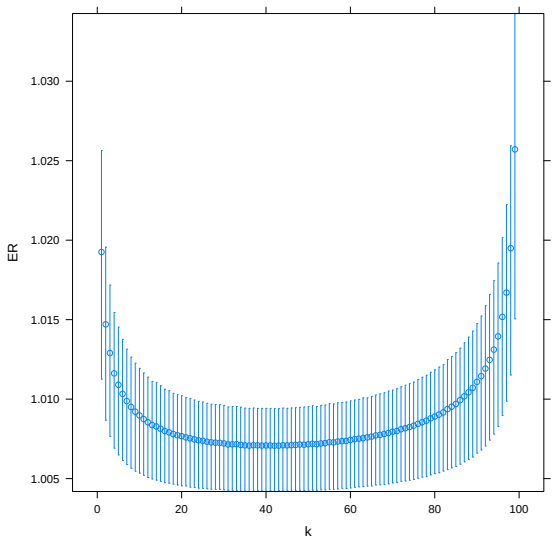
(a) $\sigma^2 = 1$



(b) $\sigma^2 = 0.1$



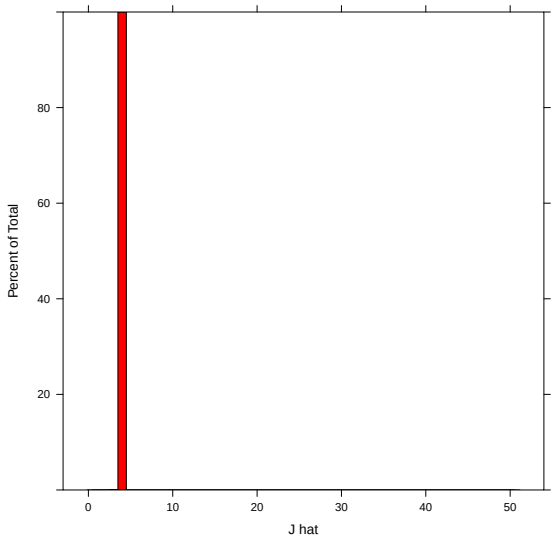
(c) $\sigma^2 = 0.05$



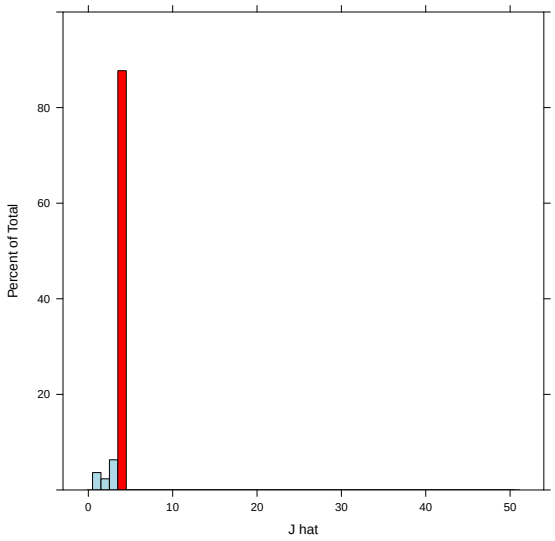
(d) $\sigma^2 = 0$

Points indicate means over 100000 simulations. Bars show interquartile range.

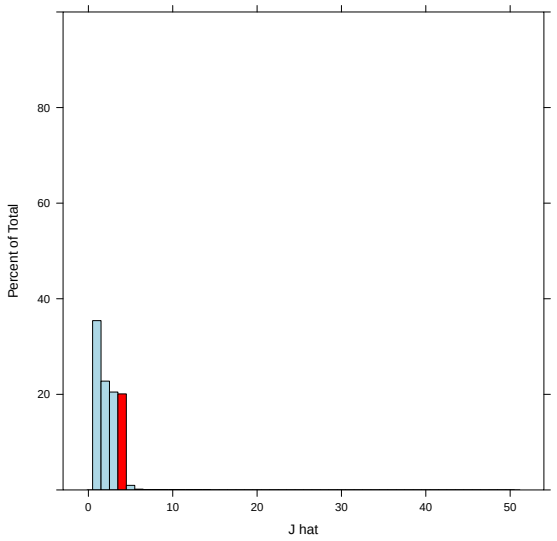
Appendix Figure E.3: Estimated number of groups ($\hat{J}$)



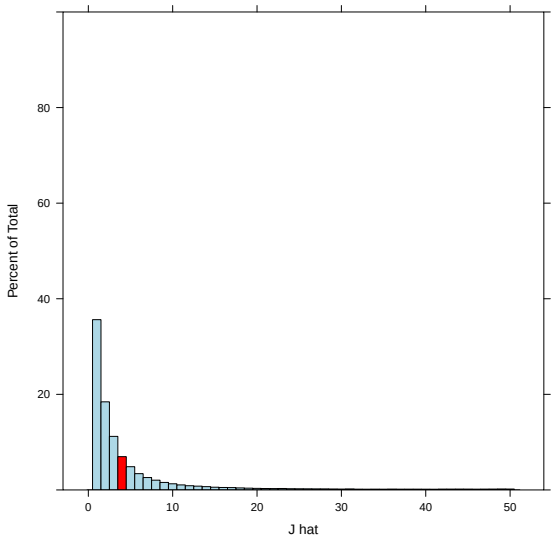
(a) $\sigma^2 = 1$



(b) $\sigma^2 = 0.1$



(c) $\sigma^2 = 0.05$



(d) $\sigma^2 = 0$

Histograms of estimated number of groups, over 100000 simulations. True value $J = 4$ shown in red.

F Reference Distributions

We consider three different "reference distributions". First, suppose that the structural model presented in Section 2.1 is correct. In this case, the distribution of the number of attacks by disorganized militants in district i is the same for all periods, with expected value $\eta\ell_i$. Thus, under the null hypothesis that there is no group structure, the observed attack data is weakly exchangeable: within a given district, permuting the time indices does not change the joint distribution of the attacks.[74] The total number of such permutations is huge, and thus rather than perform calculations using the entire set we consider only a random subset of these permutations. By construction, the permuted data exhibits no group structure: all the off-diagonal entries of the sample covariance matrix will be zero asymptotically. To construct the desired reference distribution, we treat each of these permutations as if it were the observed data.

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

Finally, suppose that the expected number of attacks by disorganized militants varies at the daily level, rather than the monthly level. The general case, with $\eta_{it}\ell_i$ attacks expected in district i at time t, is so general that it does not appear to allow for any permutations. However, suppose that the number of expected attacks is instead $\eta_t\ell_i$, where η_t now does not differ across districts.[75] This might be the case,

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

[75] This gives the disorganized militants the same structure an additional organized group. The test against the null hypothesis in this case is thus related to whether there is an organized group present

for example, if there were particular days that, for whatever reason, generated large amounts of random violence. In this case, observations are "approximately" weakly exchangeable via the following sort of permutation, inspired by Good [2002]. Find a pair of districts i and i', and a pair of times t and t', such that the following two conditions hold: there were the same number of attacks x in district i at time t and in district i' at time t', and there were the same number of attacks x' in district i at time t' and in district i' at time t. Permute the data by swapping x and x' in these four entries.[76] These permutations are attractive from an intuitive perspective, as they retain not only the same number of total attacks in each district, but also the same number of total attacks on each day. In the Afghan data, there are relatively few attacks on any given day and thus an enormous number of possible permutations of this sort. A random sample of these permutations is used.

G 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

that is active in some districts but not others. Under the null hypothesis, the off-diagonal entries of the sample covariance matrix should be directly proportional to the total number of attacks in the districts in question.

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

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

by rearranging terms. The canonical reference for multivariate permutations appears to be Pesarin [2001], although this specific type of permutation is not described. Good [2005] provides an accessible introduction to permutation tests.

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

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

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.