Point process modelling of the Afghan War Diary

Andrew Zammit-Mangion[a,b], Michael Dewar[c], Visakan Kadirkamanathan[d], and Guido Sanguinetti[a,e,1]

[a]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom; [b]University/British Heart Foundation Centre for Cardiovascular Science, Queen's Medical Research Institute, Edinburgh EH16 4TJ, United Kingdom; [c]Department of Applied Physics and Applied Mathematics, Columbia University, New York, NY 10027; [d]Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, United Kingdom; and [e]SynthSys—Systems and Synethic Biology, University of Edinburgh, Edinburgh EH9 3JD, United Kingdom

Edited by Stephen E. Fienberg, Carnegie Mellon University, Pittsburgh, PA, and approved June 8, 2012 (received for review February 25, 2012)

Modern conflicts are characterized by an ever increasing use of information and sensing technology, resulting in vast amounts of high resolution data. Modelling and prediction of conflict, however, remain challenging tasks due to the heterogeneous and dynamic nature of the data typically available. Here we propose the use of dynamic spatiotemporal modelling tools for the identification of complex underlying processes in conflict, such as diffusion, relocation, heterogeneous escalation, and volatility. Using ideas from statistics, signal processing, and ecology, we provide a predictive framework able to assimilate data and give confidence estimates on the predictions. We demonstrate our methods on the WikiLeaks Afghan War Diary. Our results show that the approach allows deeper insights into conflict dynamics and allows a strikingly statistically accurate forward prediction of armed opposition group activity in 2010, based solely on data from previous years.

conflict prediction | point processes | variational Bayes

The last decade has witnessed a tremendous increase in the availability of data relating to conflicts. For example, the collection of media reports in the 'Armed Conflict Location and Event Dataset' (1) provides a small scale but highly curated record of conflict events. More prominently, the release of confidential documents by the WikiLeaks whistleblower website in July 2010 has provided for the first time a large scale (but uncurated) description of the current Afghan conflict. However, most analyses of these and similar data sources do not go beyond visualization and descriptive statistical methods (2–5), for good reasons: first, conflict data is highly heterogeneous and often poorly annotated. For example, the WikiLeaks Afghan War Diary (AWD) data used in this study (Dataset S1) consists of event entries as diverse as elaborate preplanned military activity and spontaneous stop-and-search events. Any plausible attempt to model this data will need to be statistical in nature in order to handle the high levels of noise. Second, it is very difficult to define simple mechanisms that would allow the bottom-up construction of a plausible model.

Here, we develop statistical dynamical modelling methodologies to provide a predictive framework that may be used in policy making. We show that the temporal and spatial dependencies (6, 7) as well as diffusion and advection effects (8, 9) inherent in conflict data make it suitable for the use of a broad class of models, widely employed in ecology and epidemiology, in order to describe the dynamics of disaggregated data. We then develop tools based on ideas from point process statistics (10) to constrain the models. The approach enables us to leverage powerful techniques from point process filtering theory and spatiotemporal statistics (11–14) to carry out inference of the underlying system's dynamics and to predict the future behavior of the system.

We test the performance of our methods on the AWD, a WikiLeaks release which contains over 75,000 military logs by the USA military, describing events which occurred between the beginning of 2004 and the end of 2009 and providing a high temporal and spatial resolution description of the Afghan war in that period. We show that our approach allows deeper insights in the conflict dynamics than simple descriptive methods by providing a spatially resolved map of the growth and volatility of the conflict. Most remarkably, we show that a model trained on the AWD can predict with surprising statistical accuracy the progression of the conflict in 2010; i.e., a year after the end of the AWD data. We conclude the paper by discussing the importance and potential of statistical modelling of conflict data, as well as offering some consideration as to its wide applicability.

Statistical Methods

Spatial Point Processes and the Stochastic Integro-Difference Equation (SIDE). Conflict data typically consists of a set of incidents labeled through spatiotemporal coordinates which, when visualized as event markers, are highly spatiotemporally correlated, generally clustered and representative of some underlying structure. In this regard, these data sets are very similar to others encountered in a variety of fields, such as epidemiology (15) and agricultural sciences (16). Poisson point processes provide a convenient and frequently used mathematical framework to model event-based data; in this framework, the probability of observing a certain number of events within a region of interest $\mathcal{O}$ is given by a Poisson distribution whose mean is the integral over $\mathcal{O}$ of an intensity function $\lambda(\boldsymbol{s}), \boldsymbol{s} \in \mathcal{O}$. In order to accommodate phenomena such as event clustering, the intensity itself is often modeled as a random function, giving rise to *doubly stochastic* or *Cox* processes. A popular class of Cox processes, which will also be considered here, is the log-Gaussian Cox process (LGCP) where the logarithm of the event intensity is assumed to be a Gaussian process (GP). We recall that a GP is wholly defined by (*i*) a mean function $\mu(\boldsymbol{s})$ describing a global trend and (*ii*) a covariance function $k(\boldsymbol{s}, \boldsymbol{r})$ indicating how the field at distinct points in space ($\boldsymbol{s}$ and $\boldsymbol{r}$) covary (17).

Because conflict data is often logged in a discrete-time format (e.g., the day of an event as opposed to the precise time), we will consider a discrete-time series of continuous-space LGCPs. Formally, let $k \in \mathcal{K}$, $\mathcal{K} = \{1, 2, \ldots, K\}$ denote a discrete-time index set and $\{z_k(\boldsymbol{s})\}$, $z_k(\boldsymbol{s}) \sim \mathcal{GP}(\mu_k(\boldsymbol{s}), \sigma_k^2 \psi_k(\boldsymbol{s}, \boldsymbol{r}))$, a set of temporally correlated spatial GPs, each with mean $\mu_k(\boldsymbol{s})$ and covariance function $\sigma_k^2 \psi_k(\boldsymbol{s}, \boldsymbol{r})$. For each k, we then define the point process intensity function as $\lambda_k(\boldsymbol{s}) = \exp(z_k(\boldsymbol{s}))$. Frequently, the mean function of $z_k(\boldsymbol{s})$, $k \in \mathcal{K}$, can be related to explanatory variables, such as population density, which help to reduce prediction uncertainty. We hence let $\boldsymbol{d}(\boldsymbol{s})$ be a vector of spatially referenced covariates and $\boldsymbol{b}^T$ the corresponding regression parameters; the LGCP at time k then has intensity $\lambda_k(\boldsymbol{s}) = \exp(\boldsymbol{b}^T \boldsymbol{d}(\boldsymbol{s}) + z_k(\boldsymbol{s}))$.

Naturally, the key question is how to specify the temporal dynamics of the intensity functions through $z_k(\boldsymbol{s})$; we need a sufficiently flexible modelling approach to incorporate the complexity of conflict dynamics. One such representation is the stochastic

Author contributions: A.Z.M., V.K., and G.S. designed research; A.Z.M. and G.S. performed research; A.Z.M., M.D., and G.S. analyzed data; and A.Z.M. and G.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: G.Sanguinetti@ed.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1203177109/-/DCSupplemental.

integro-difference equation (SIDE), a model originally introduced in ecology (18) which has rapidly gained popularity in spatiotemporal statistics (19). The SIDE relates the spatiotemporal dependent variable $z_k(\boldsymbol{s})$ to $z_{k+1}(\boldsymbol{s})$ through the following integral equation

$$z_{k+1}(\boldsymbol{s}) = \int_{\mathcal{O}} k_I(\boldsymbol{s},\boldsymbol{r}) f_1(z_k(\boldsymbol{r}))\mathrm{d}\boldsymbol{r} + e_k(\boldsymbol{s}), \quad [1]$$

where $k_I(\boldsymbol{s},\boldsymbol{r})$ is the mixing kernel in the integral and $e_k(\boldsymbol{s})$ is an added disturbance, modeled as a Gaussian field with mean $\mu_Q(\boldsymbol{s})$ and covariance function $k_Q(\boldsymbol{s},\boldsymbol{r})$, $e_k(\boldsymbol{s}) \sim \mathcal{GP}(\mu_Q(\boldsymbol{s}), k_Q(\boldsymbol{s},\boldsymbol{r}))$, and $\mathcal{O}$ is the spatial domain under investigation. The nonlinear mapping $f_1(\cdot)$ distorts the field in the *sedentary stage*; in this work we will employ the identity map $f_1(z_k(\boldsymbol{r})) = z_k(\boldsymbol{r})$, an assumption usually adopted in the absence of a priori knowledge (20). The SIDE is, in its original form, a very flexible modelling tool, capable of representing a number of dynamic effects such as diffusion and dispersal (or both simultaneously) even under considerably restrictive conditions (19). Although the AWD will suggest the use of only a special case of SIDE, the two-pronged methodological approach we present here to estimate unknown components is in principle applicable to the more general case.

Nonparametric Analysis. We start by studying the correlation between the conflict events within the same and across subsequent time frames. We are interested in the probabilities of finding a conflict event at $\boldsymbol{r}$ given that an event has occurred at $\boldsymbol{s}$ within the same time frame k or at the previous time frame $k-1$. In point process statistics these are quantified through the pair auto-correlation function (PACF) $g_{k,k}(\boldsymbol{s},\boldsymbol{r})$, and what we term the pair cross-correlation function (PCCF) $g_{k,k+1}(\boldsymbol{s},\boldsymbol{r})$ defined as

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \frac{\lambda^{(2)}_{k,k}(\boldsymbol{s},\boldsymbol{r})}{\lambda^{(1)}_{k}(\boldsymbol{s})\lambda^{(1)}_{k}(\boldsymbol{r})}, \quad [2]$$

$$g_{k,k+1}(\boldsymbol{s},\boldsymbol{r}) = \frac{\lambda^{(2)}_{k,k+1}(\boldsymbol{s},\boldsymbol{r})}{\lambda^{(1)}_{k}(\boldsymbol{s})\lambda^{(1)}_{k+1}(\boldsymbol{r})}, \quad [3]$$

where $\lambda^{(1)}_k(\boldsymbol{s}) = \mathbb{E}[\lambda_k(\boldsymbol{s})]$ and $\lambda^{(2)}_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \mathbb{E}[\lambda_k(\boldsymbol{s})\lambda_k(\boldsymbol{r})]$ are real and positive and $\mathbb{E}[\cdot]$ denotes the expectation operator.

The PACF may be used to determine qualitative characteristics of the conflict; for instance if $g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = 1$, then no spatial pattern can be extracted from the data; $g_{k,k}(\boldsymbol{s},\boldsymbol{r}) > 1$ and $g_{k,k}(\boldsymbol{s},\boldsymbol{r}) < 1$ can be used to indicate conflict aggregation and repulsion respectively. The PACF can also be used as a preprocessing tool for dimensionality reduction. Direct use of the PACF and PCCF for nonparametric field estimation is also possible (*SI Text*) but our preliminary investigation showed that this is only a reliable proposition for homogeneous datasets with a very large number of events (*SI Text*).

Dimensionality Reduction and Bayesian Inference. In order to develop an inferential approach for SIDE driven LGCPs, we adopt a basis function representation of the spatiotemporal field, which we will then truncate at a level which enables sufficient accuracy (21). This representation, frequently employed in spatiotemporal modelling [e.g., process convolution models (22, 23)], in turn facilitates the implementation of computationally efficient inference algorithms.

The choice of basis functions is a problem that deserves attention; as far as we are aware, there are no standard solutions for LGCPs. We propose here a general approach to selecting basis functions based on the nonparametric estimation of the PACF. Specifically, we capitalize on (*i*) a fundamental lemma of LGCPs

$$g_{k,k}(\boldsymbol{s},\boldsymbol{r}) = \exp(\sigma_k^2 \psi_k(\boldsymbol{s},\boldsymbol{r})), \quad [4]$$

which states that the log PACF is proportional to the field auto-correlation function and (*ii*) the auto-correlation theorem (24) which states that the Fourier transform of the auto-correlation function is the spectrum of the signal. Hence, a relationship between the frequency content of the point process and the PACF is found, which in turn may be used to select a set of sufficiently representative basis functions, much on the lines of refs. 21 and 25. We then obtain a decomposition of the kernel, the mean disturbance and the field as

$$z_k(\boldsymbol{s}) = \boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{x}_k, \quad [5]$$

$$\mu_Q(\boldsymbol{s}) = \boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{\vartheta}, \quad [6]$$

$$k_I(\boldsymbol{s},\boldsymbol{r}) = \boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{\Sigma}_I \boldsymbol{\phi}(\boldsymbol{r}), \quad [7]$$

$$k_Q(\boldsymbol{s},\boldsymbol{r}) = \boldsymbol{\phi}(\boldsymbol{s})^T \boldsymbol{\Sigma}_Q \boldsymbol{\phi}(\boldsymbol{r}), \quad [8]$$

where $\boldsymbol{\phi}(\boldsymbol{s}) \in \mathbb{R}^n$ is the vector of basis functions, $\boldsymbol{x}_k \in \mathbb{R}^n$ and $\boldsymbol{\vartheta} \in \mathbb{R}^n$ are weights which reconstruct the spatiotemporal field and the disturbance mean respectively and where $\boldsymbol{\Sigma}_I \in \mathbb{R}^{n\times n}$ and $\boldsymbol{\Sigma}_Q \in \mathbb{R}^{n\times n}$ reconstruct the kernel covariance function and the disturbance covariance function respectively.

It can be shown (*SI Text*) that under this decomposition, the SIDE of Eq. **1** can be represented in the compact form

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}(\boldsymbol{\Sigma}_I)\boldsymbol{x}_k + \boldsymbol{w}_k(\boldsymbol{\vartheta},\boldsymbol{\Sigma}_Q), \quad [9]$$

where $\boldsymbol{A}(\boldsymbol{\Sigma}_I) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{w}_k \in \mathbb{R}^n$ is a Gaussian colored noise term with mean $\mathbb{E}[\boldsymbol{w}_k] = \boldsymbol{\vartheta}$ and covariance $\mathrm{cov}[\boldsymbol{w}_k] = \boldsymbol{\Sigma}_Q$. Eq. **9** is a standard linear dynamical system where both the states $\mathcal{X}_K = \boldsymbol{x}_{0:K} = \{\boldsymbol{x}_k\}_{k=0}^K$ and the unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{\vartheta}, \boldsymbol{\Sigma}_I, \boldsymbol{\Sigma}_Q^{-1}\}$ need to be estimated from the data $\mathcal{Y}_K = \{\boldsymbol{y}_k\}_{k=1}^K$ where we define each $\boldsymbol{y}_k$ to be the set of coordinates of the logged events at the kth time point.

For inference, we make use of the likelihood function

$$p(\boldsymbol{y}_k|\lambda_k(\boldsymbol{s})) = \prod_{\boldsymbol{s}_j\in\boldsymbol{y}_k} \lambda_k(\boldsymbol{s}_j) \exp\left(-\int_{\mathcal{O}} \lambda_k(\boldsymbol{s})\mathrm{d}\boldsymbol{s}\right), \quad [10]$$

and approximate each $\lambda_k(\boldsymbol{s})$ using the same basis representation:

$$\lambda_k(\boldsymbol{s}) = \exp(\boldsymbol{b}^T\boldsymbol{d}(\boldsymbol{s}) + z_k(\boldsymbol{s})) \approx \exp(\boldsymbol{b}^T\boldsymbol{d}(\boldsymbol{s}) + \boldsymbol{\phi}(\boldsymbol{s})^T\boldsymbol{x}_k). \quad [11]$$

We proceed with a computationally efficient variational Bayes (VB) method by approximating the full posterior distribution

$$\begin{aligned} p(\mathcal{X}_K, \boldsymbol{\theta}, \boldsymbol{b}|\mathcal{Y}_K) &= p(\mathcal{X}_K, \boldsymbol{\vartheta}, \boldsymbol{\Sigma}_I, \boldsymbol{\Sigma}_Q^{-1}, \boldsymbol{b}|\mathcal{Y}_K) \\ &\approx \tilde{p}(\mathcal{X}_K)\tilde{p}(\boldsymbol{\vartheta})\tilde{p}(\boldsymbol{\Sigma}_I)\tilde{p}(\boldsymbol{\Sigma}_Q^{-1})\tilde{p}(\boldsymbol{b}), \end{aligned} \quad [12]$$

where $\tilde{p}(\cdot)$ are the variational marginals (26, 27).

The variational marginals are able to reveal important properties of the conflict progression; $\mathcal{X}_K$ is used to reconstruct the spatiotemporal field at every time point, $\boldsymbol{\vartheta}$ reveals the spatially varying escalation in conflict, $\boldsymbol{\Sigma}_I$ the extent of any spatial dynamics, if any, and $\boldsymbol{\Sigma}_Q$ the *volatility* of the conflict which can either be localized or dependent on events happening at remote geographical locations. The number of unknown parameters in the reduced model scales as $\mathcal{O}(n^2)$, where n is the number of basis functions retained. However, as we will see later, nonparametric

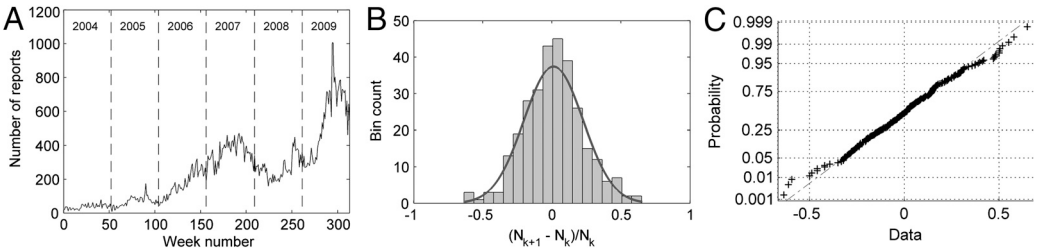

Fig. 1. Temporal analysis of the AWD. (*A*) Weekly number of activity reports in Afghanistan between January 2004 and December 2009 (bin size = 1 w). (*B*) Distribution of weekly fractional increments in report count in the AWD where N_k denotes the number of report counts at week k. (*C*) Corresponding normality probability plot. Fourteen points (4.5% of data) were marked as clear outliers as a result of low report count and not used in this analysis.

data analysis can suggest further simplifications which can considerably lower the complexity of the model.

The Afghan War Diary

On July 25 , 2010, WikiLeaks publicly made available a compendium of US military war logs in Afghanistan dating between 2004 and 2009. The so-called Afghan War Diary contains a detailed insider's description of the military machinery of the world's largest power; it consists of roughly 77,000 logs and entries detail the time and position of an event, which could be anything from a stop-and-search episode to a gunfight. The dataset is considered a reliable description of the Afghan war and systematic verification efforts carried out by several organizations such as the New York Times* have found little reason to dispute its authenticity. *SI Text* reports some of our own tests which show significant correlations between the logged event rate in the AWD and that in other datasets. In what follows we adopt the spatiotemporal point process approach to infer a model from the data in the AWD and use it to analyze the heterogeneous growth (through $\boldsymbol{\vartheta}$) and volatility (through $\boldsymbol{\Sigma}_Q$) of the conflict in Afghanistan and also to predict violence of armed opposition groups in 2010, a year after the end of the WikiLeaks dataset.

We start with a nonparametric analysis (*SI Text*) of the data by splitting the data into weekly intervals ($\Delta_t = 1$ week) and looking at the temporally averaged PACF and PCCF fitted to Gaussian radial basis functions. It is found that, on average, the log PACF is nearly identical to the log PCCF and that a nonparametric estimate of a homogeneous kernel $k_I(||\boldsymbol{s}-\boldsymbol{r}||)$, computed with the direct inverse filter, is very narrow in relation to the extent of the spatial correlations in the field (*SI Text*). This observation suggests that $k_I(\cdot)$ in the SIDE may be safely approximated to $\gamma(\boldsymbol{s})\delta(\boldsymbol{s}-\boldsymbol{r})$, corresponding to negligible spatial interactions across adjacent time frames. Note that if $e_k(\boldsymbol{s})$ is restricted to be homogeneous and $\gamma(\boldsymbol{s}) = \gamma$, the spatiotemporal covariance function is separable, a common assumption in several fields such as epidemiology (15). However, given the data characteristics, we chose to maintain the spatial heterogeneity in $e_k(\boldsymbol{s})$. We also set $\gamma(\boldsymbol{s}) = 1$ as we found no evidence of mean reversion both at a national and a provincial level; additionally, we found that a spatially dependent $\gamma(\boldsymbol{s})$ did not contribute to increased prediction accuracy.

The resulting formulation is validated by studying the temporal dynamics of the AWD (Fig. 1*A*). A quantitative analysis reveals that the fractional increments of the event incidence nationwide are normally distributed (with a one-tailed Shapiro Wilk's test and a Levene's test with $\alpha = 0.1$, $n = 312$ w. See also Fig. 1 *B* and *C*)†. This statistic characterizes systems following a geometric Brownian motion given by

$$\mathrm{d}\lambda(\boldsymbol{s},t) = \widetilde{R}(\boldsymbol{s})\lambda(\boldsymbol{s},t)\mathrm{d}t + \lambda(\boldsymbol{s},t)\mathrm{d}W(\boldsymbol{s},t), \qquad [13]$$

where the increment $\mathrm{d}W(\boldsymbol{s},t)$ is a Gaussian process with zero mean and covariance function $k_Q(\boldsymbol{s},\boldsymbol{r})\mathrm{d}t$ and $\widetilde{R}(\boldsymbol{s})$ is a spatially varying percentage drift. Applying Ito's Lemma (28) to $\ln\lambda(\boldsymbol{s},t)$ and noting that the continuous-time intensity $\ln\lambda(\boldsymbol{s},t) = \boldsymbol{b}^T\boldsymbol{d}(\boldsymbol{s}) + z(\boldsymbol{s},t)$, we obtain the following form for $z(\boldsymbol{s},t)$:

$$\mathrm{d}z(\boldsymbol{s},t) = R(\boldsymbol{s})\mathrm{d}t + \mathrm{d}W(\boldsymbol{s},t), \qquad [14]$$

where $R(\boldsymbol{s}) = \widetilde{R}(\boldsymbol{s}) - \frac{1}{2}\sigma(\boldsymbol{s})^2$ is a heterogeneous temporally independent spatial growth rate and $\sigma(\boldsymbol{s})^2$ is the variance field. Applying an explicit Euler discretization scheme to Eq. **14**, one obtains the model $z_{k+1}(\boldsymbol{s}) = z_k(\boldsymbol{s}) + e_k(\boldsymbol{s})$ where $e_k(\boldsymbol{s})$ has mean $\mu_Q(\boldsymbol{s}) = R(\boldsymbol{s})\Delta_t$ and covariance function $k_Q(\boldsymbol{s},\boldsymbol{r})\Delta_t$. This model is, as expected, the SIDE with the delta-Dirac kernel.

The field is next decomposed and Eqs. **6** and **8** are applied to finally obtain the random walk model occasionally employed in spatiotemporal studies (29)

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{w}_k(\boldsymbol{\vartheta},\boldsymbol{\Sigma}_Q). \qquad [15]$$

For basis function selection we employed the aforementioned frequency-based approach (see *SI Text* for complete details). Finally, we chose population density and the distance to the nearest major city as covariates (see *SI Text* for details on how this choice was made). Inference was carried out using the VB algorithm described above. Full derivatons, algorithmic details, and configuration parameters (priors and stopping conditions), as well as indicative run times, are given in *SI Text* respectively whilst a detailed simulation study showing the identifiability of the model under flat priors and a comparison with kernel-based estimators (30), is given in *SI Text*.

Results

Conflict Intensity and Regression Parameters. State inference leads to broad conclusions to where and how the conflict intensity has increased, decreased or shifted in time. We show the posterior mean intensity at regular intervals in *SI Text* and also in Movie S1 together with the underlying AWD events at a weekly resolution. The progression of the intensity captures important geographical features of the war scenario. Regions of high intensity in 2009 include Sangin in northern Helmand (see *SI Text* for a provincial map), one of the most dangerous places in Afghanistan, notorious for thousands of improvised explosive devices and frequent suicide bombings (2). Other regions, such as Kabul, Nangarhar, and Paktya provinces, on the other hand have witnessed high activity throughout the six-year interval. Also very apparent is the emergence in later years of a high intensity ring starting from Kabul extending southwards towards Kandahar, up through Herat, through Balkh and back to Kabul. This roughly elliptical shape corresponds to the country's 'ring road', commonly targeted by insurgent activity and placement of improvised explosive devices (2). We note that a representative spatiotem-

*http://www.nytimes.com/2010/07/26/world/26editors-note.html?_r=1

†The Levene's test failed to reject the null hypothesis of constant variance for the years 2006 to 2009 but not when including 2004 and 2005. The reason for rejection when including the earlier two years can be safely attributed to relatively low report count, arising in noisy quantities when computing the fractional increments.

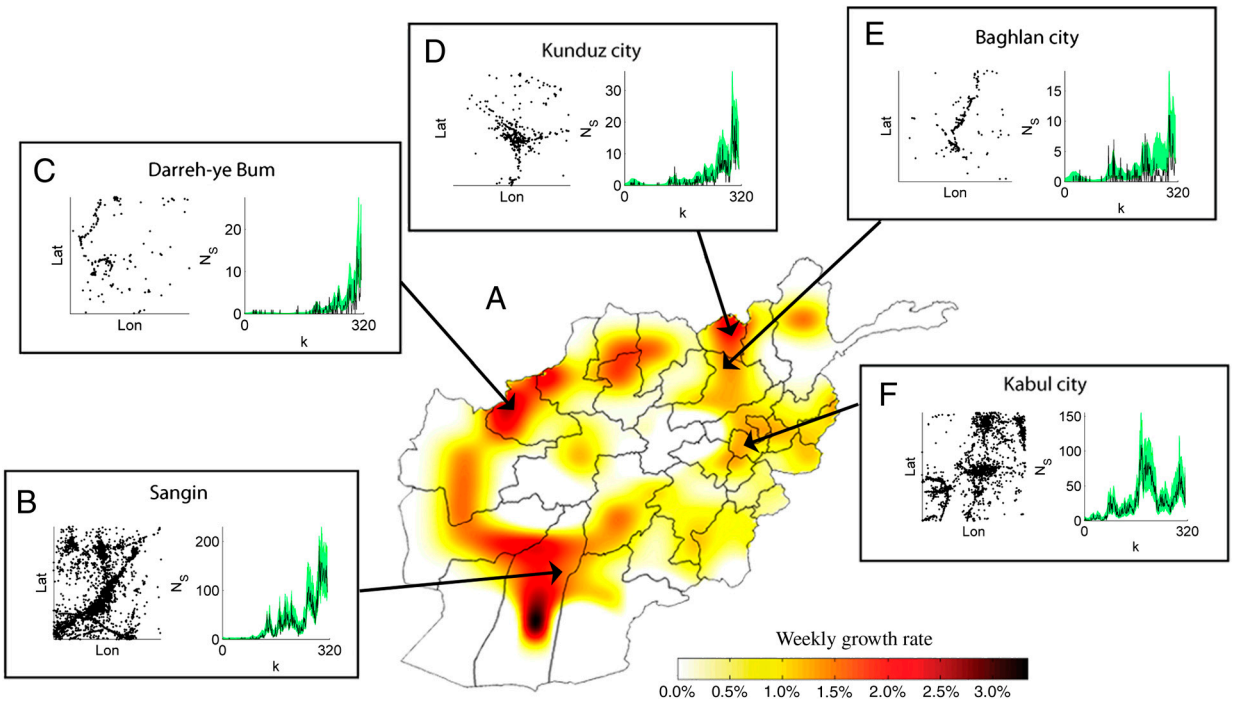

Fig. 2. AWD activity growth in Afghanistan. (*A*) Posterior mean fractional increase in logs per week in the AWD between 2004 and 2009. Only regions with positive overall growth are shown. (*B–F* left) Spatial map of all events occurring in a square of side 100 km centered on the city under study. (*B–F* right) Number of weekly events N_k in these regions (-) together with the estimated 90% confidence intervals (green shading).

poral intensity map may also be obtained with the use of standard nonparametric kernel estimators (30), seen in Movie S2.

The regression parameters corresponding to population density and distance to the closest major city were estimated to be 1.97×10^{-4} $\pm 6.2 \times 10^{-6}$ (2σ) and $-0.037 \pm 2.1 \times 10^{-4}$ respectively. This result reflects the fact that a vast majority of logs in the AWD, as with typical conflict datasets, are present in urban and highly populated areas (7).

Conflict Escalation and Volatility. A major advantage of the adopted model-based approach is the ability to establish quantitative conclusions on aspects of the conflict scenario other than the intensity. For instance, in the AWD we have modeled the spatially varying escalation of conflict in Afghanistan between 2004 and 2009 through $\boldsymbol{\vartheta}$ (Fig. 2) and the volatility of the conflict progression in the same period through the diagonal elements of $\boldsymbol{\Sigma}_Q$ (Fig. 3).

Escalation (or deescalation) may be used to distinguish between event hot spots and growth hot spots. This feature is, in itself, a major advantage over conflict clustering analysis which cannot discern whether a cluster was a one-off, or a sign of a deteriorating situation. In the AWD it is very evident, for instance, that while some of the high growth areas such as Helmand also had an overall high count of events, this was not the general case; for example, Sar-e Pul and Balkh in the north and the Badghis province in the west all had witnessed a modest number of total event count but are seen to have had a significant overall growth in activity throughout the years.

The volatility/predictability of the conflict is also of considerable interest. In our case, a small diagonal value in $\boldsymbol{\Sigma}_Q$ indicates that based on the data so far the future intensity may be predicted with reasonable accuracy. On the other hand, a large value is a sign of considerable volatility; little can be said about the future. Such inferences are vital for decision purposes—simply stated it might prove a better option to admit a large uncertainty about the future, than to base a policy decision on a highly uncertain prediction. Consider for instance the high volatility on the eastern part of Farah province in western Afghanistan (see *SI Text*). A subsequent analysis of the video shows spurious clusters emerging in April 2005 and towards the end of 2006, an indication that the conflict dynamics in this part of Afghanistan are relatively hard to predict; even more so than in Sangin which had seen a drastic, but relatively smooth, increase in events in the latter years.

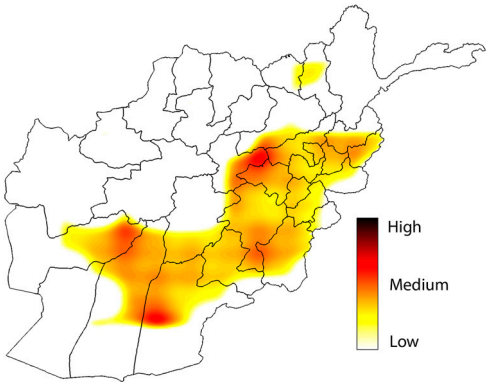

Fig. 3. Volatility in conflict events between 2004 and 2009 in the WikiLeaks AWD. Only regions with a high volatility ($\sigma^2 > 0.055$) are shown.

Prediction. The key advantage of dynamic point process modelling is the ability to make statistical predictions of the system's behavior for decision making. To illustrate this feature we considered the frequency of incidents by armed opposition groups (AOG) and predicted it in 2010, a year after the termination of the WikiLeaks dataset. AOG activity on a provincial scale was obtained from the Afghanistan NGO Safety Office (ANSO) safety reports[‡]. Prediction was carried out by (*i*) sampling a trajectory z_k through $\tilde{p}(\mathcal{X})$ in 2009, (*ii*) forward simulating each trajectory for 52 weeks (2010) using the generative model with the parameters ϑ, $\boldsymbol{\Sigma}_Q$ and $\boldsymbol{b}$ set to $\mathbb{E}_{\tilde{p}(\vartheta)}[\boldsymbol{\vartheta}]$, $(\mathbb{E}_{\Sigma_Q^{-1}}[\boldsymbol{\Sigma}_Q^{-1}])^{-1}$ and $\mathbb{E}_{\tilde{p}(\boldsymbol{b})}[\boldsymbol{b}]$ respectively,

[‡]Reports are freely available from the official ANSO website http://www.afgnso.org.

(*iii*) integrating the interpolated sample over each *i*th province to give $\hat{z}_{k,i}$, (*iv*) finding the corresponding intensity $\hat{\lambda}_{k,i}$, (*v*) averaging the intensity over 52 w invervals to obtain $\hat{\lambda}_{2009,i}$ and $\hat{\lambda}_{2010,i}$, (*vi*) generating two samples $N_{i,2009}$ and $N_{i,2010}$ from Poisson random variables with intensity $\hat{\lambda}_{2009,i}$ and $\hat{\lambda}_{2010,i}$, (*vii*) predicting a provincial AOG count in 2010, $\mathrm{AOG}_{i,2010}$, from that in 2009, $\mathrm{AOG}_{i,2009}$, through the formula

$$\mathrm{AOG}_{i,2010} = \frac{N_{i,2010}}{N_{i,2009}} \mathrm{AOG}_{i,2009}, \qquad [16]$$

and (*viii*) repeating (*i*)–(*vii*) for $N = 2,000$ times. Note that although Eq. **16** is a very simple predictor, one which assumes a linear relationship (without offset), it reflects the fact that the frequency of the logs in the WikiLeaks dataset is significantly correlated with the saliency of AOG initiated attacks in Afghanistan, particularly in 2009 (*SI Text*).

As seen from Fig. 4 *A* and *B*, the prediction medians from the model match closely the observed values. In Baghlan, for instance, AOG activity rose by 120% (17.3% using log counts) from 100 incidents in 2009 to 222 in 2010; the model predicted a median 2010 increase of 128% (17.9%) to a count of 228. Badakhshan saw a −19% (−5.5%) growth in 2010; our model predicted a median of −23% (−7.0%) growth. Further, a correlation test between the predicted medians and actual incident count for all 32 provinces gave a Pearson's correlation coefficient of 0.81 on a linear scale and 0.89 under a log transform (Fig. 4*B*), showing strong support for prediction capability.

Despite this, for some provinces (such as Badghis), the median remains substantially offset from the true value. The disparities are, however, consistent with the predictive distributions. From Fig. 4*A* it is seen that counts in 62.5% of the predicted provinces lie between the lower and upper quartiles and more importantly all of them lie within the 99% confidence intervals. The same holds for the predicted change in AOG activity in 2010, the distributions of which are given in *SI Text*. Even here, the model is seen to be well tuned and supply confidence intervals which consistently capture the true activity growth (Fig. 4*C*).

Thus, although the true count is not always close to the pointwise median predictions, we see that the predictions are accurate *in a statistical sense*; i.e., the predicted and observed distribution of AOG growth across provinces match closely. Further, the above results were obtained merely from the AWD up to 2009 and did not include any knowledge of events in 2010 such as military plans or deployments/withdrawal of troops. Incorporation of domain knowledge to reduce the predictive variance would, in principle, be straightforward in our model through manipulation of the prior distributions or inclusion of further relevant exogeneous inputs.

Discussion and Conclusions

Our results demonstrate that statistical spatiotemporal modelling can be an extremely valuable tool in the analysis of conflict. The analysis of the AWD data shows that data modelling can yield insights that cannot be achieved by simple visualization or by the use of descriptive statistics. This claim is borne out by the availability of a spatially resolved map of the growth of conflict intensity, as well as the volatility/predictability of the conflict. Furthermore, the availability of statistical confidence intervals associated with all model predictions is an important feature of our modelling framework and a potentially crucial feature for decision making.

The most striking result of our analysis is the ability to accurately predict (in a statistical sense) conflict dynamics for a whole year after the end of the AWD data on which the model was trained. While we do not have a simple mechanism underlying our model, the fact that a latent Gaussian model can produce predictions of this quality cannot be by chance. Intuitively, we believe that the type of conflict we are modelling may be the main reason

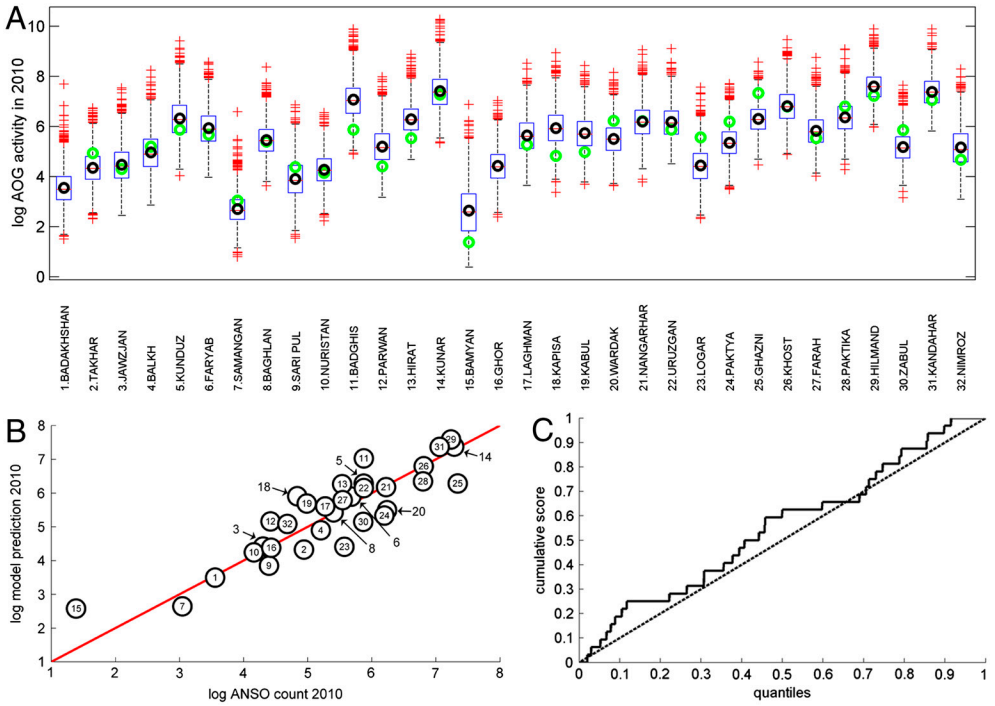

Fig. 4. Prediction of AOG growth in 2010. (*A*) Box-and-whisker plots of the predicted log AOG activity in 2010 using 2000 MC runs. For each province, the box marks the first and third quartiles; the median (red line), mean (black circle), and true reported count (green circle) are also given. The whiskers extend to the furthest MC points that are within 1.5 times the interquartile range (≈99% coverage) and the outliers are plotted individually (red cross). (*B*) Comparison between the median log model prediction and log AOG count in 2010 where the mark number corresponds to the province number denoted in (*A*). (–) Ideal prediction. (*C*) Cumulative distribution of growth prediction on a province-by-province basis. The graph shows correct tuning of the model, with approximately *x*% of provinces lying within the *x*th percentile of the predictive distribution. (–) True cumulative score. (dotted line) Ideal cumulative score.

why our method works. The Afghan conflict is characterized by insurgent movements and qualifies as a case of irregular warfare where activity is only loosely dependent and actioned by a myriad of disparate groups. Some averaging effects may be leading to the Gaussian behavior of the conflict's intensity, which in turn may be exploited for modelling purposes.

Naturally, as with all modelling techniques, our approach comes also with limitations, as well as benefits. From the technical point of view, reliable parameter estimation in point processes requires a sufficiently large number of events within the region of interest. While it is difficult to put a precise figure to this number, we found that parameter estimation in provinces with fewer event counts than a few dozens a year was extremely difficult. Another limitation may be the suitability of the modelling approach to generic conflict scenarios. Our approach appears to be more suitable for fragmented scenarios such as Afghanistan rather than conventional wars between well organized armies. Finally, we have assumed temporal-invariance of the parameters. Sequential implementations allowing continuous estimation of slowly varying governing parameters are in principle straightforward (11) and offer an attractive way forward to the study of conflict.

In conclusion, the analysis presented in this paper has been made possible by the development of statistical methodologies to handle large scale spatiotemporal datasets. Given the increased availability of such datasets from remote sensing or social networking sources, we envisage that methods such as those used here will become increasingly useful in a number of disciplines.

ACKNOWLEDGMENTS. This work was supported in part by the Pattern Analysis, Statistical Modelling, and Computational Learning 2 (PASCAL) FP7 Network of Excellence, and by a studentship from the University of Sheffield to A.Z.-M. G.S. is funded by the Scottish Government through the SICSA initiative. V.K. is part-funded by the EPSRC platform grant EP/H00453X/1.

1. Raleigh C, Linke A, Hegre H, Karlsen J (2010) Introducing ACLED: an armed conflict location and event dataset. *J Peace Res* 47:651–660.
2. O'Loughlin J, Witmer FDW, Linke AM, Thorwardson N (2010) Peering into the fog of war: the geography of the WikiLeaks Afghanistan war logs, 2004–2009. *Eurasian Geogr Econ* 51:472–495.
3. O'Loughlin J, Witmer FDW, Linke AM (2010) The Afghanistan-Pakistan wars, 2008–2009: micro-geographies, conflict diffusion, and clusters of violence. *Eurasian Geogr Econ* 51:437–471.
4. Gleditsch KS, Weidmann NB (2012) Richardson in the information age: GIS and spatial data in international studies. *Annu Rev Polit Sci* 15:461–481.
5. Bohannon J (2011) Counting the dead in Afghanistan. *Science* 331:1256–1260.
6. Haushofer J, Biletzki A, Kanwisher N (2010) Both sides retaliate in the Israeli–Palestinian conflict. *Proc Natl Acad Sci USA* 107:17927–17932.
7. Weidmann NB, Ward MD (2010) Predicting conflict in space and time. *J Conflict Resolut* 54:883–901.
8. Schutte S, Weidmann NB (2011) Diffusion patterns of violence in civil wars. *Polit Geogr* 30:143–152.
9. Zhukov YM (2012) Roads and the diffusion of insurgent violence: the logistics of conflict in Russia's North Caucasus. *Polit Geogr* 31:144–156.
10. Moeller J, Waagepetersen R (2004) *Statistical Inference and Simulation for Spatial Point Processes* (CRC Press, Boca Raton).
11. Zammit Mangion A, Yuan K, Kadirkamanathan V, Niranjan M, Sanguinetti G (2011) Online variational inference for state-space models with point-process observations. *Neural Comput* 23:1967–1999.
12. Zammit Mangion A, Sanguinetti G, Kadirkamanathan V (2012) Variational estimation in spatiotemporal systems from continuous and point-process observations. *IEEE T Signal Process* 60:3449–3459.
13. Wikle CK, Holan SH (2011) Polynomial nonlinear spatio-temporal integro-difference equation models. *J Time Ser Anal* 32:339–350.
14. Cressie NAC, Wikle CK (2011) *Statistics for Spatio-temporal Data* (Wiley, New Jersey).
15. Diggle P, Rowlingson B, Su T (2005) Point process methodology for online spatio-temporal disease surveillance. *Environmetrics* 16:423–434.
16. Brix A, Moeller J (2001) Space-time multi type log Gaussian Cox processes with a view to modelling weeds. *Scandinavian Journal of Statistics* 28:471–488.
17. Rasmussen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning* (The MIT Press, Cambridge, MA).
18. Kot M, Lewis MA, van den Driessche P (1996) Dispersal data and the spread of invading organisms. *Ecology* 77:2027–2042.
19. Wikle CK (2002) A kernel-based spectral model for non-Gaussian spatiotemporal processes. *Stat Model* 2:299–314.
20. Dewar M, Scerri K, Kadirkamanathan V (2009) Data-driven spatiotemporal modeling using the integro-difference equation. *IEEE T Signal Process* 57:83–91.
21. Scerri K, Dewar M, Kadirkamanathan V (2009) Estimation and model selection for an IDE-based spatio-temporal model. *IEEE T Signal Process* 57:482–492.
22. Rodrigues A, Diggle P (2010) A class of convolution-based models for spatiotemporal processes with non-separable covariance structure. *Scandinavian Journal of Statistics* 37:553–567.
23. Higdon D (1998) A process convolution approach to modelling temperatures in the North Atlantic ocean (with Discussion). *Environ Ecol Stat* 5:173–190.
24. Bracewell R (2000) *in The Fourier Transform & its Applications* (McGraw-Hill, Singapore), 3rd Ed, p 122.
25. Freestone DR, et al. (2011) A data-driven framework for neural field modeling. *NeuroImage* 56:1043–1058.
26. Beal MJ (2003) Variational Algorithms for Approximate Bayesian Inference. PhD thesis (Gatsby Computational Neuroscience Unit, University College London, United Kingdom).
27. Smidl V, Quinn A (2005) *The Variational Bayes Method in Signal Processing* (Springer Verlag, New York).
28. Jazwinski AH (1970) *Stochastic Processes and Filtering Theory* (Academic Press, London).
29. Stroud JR, Mueller P, Sanso B (2001) Dynamic models for spatiotemporal data. *J R Stat Soc B Stat Met* 63:673–689.
30. Diggle P (1985) A kernel method for smoothing point process data. *Appl Stat* 34:138–147.