SCIENTIFIC REPORTS

OPEN

Correlations and forecast of death tolls in the Syrian conflict

Kazuki Fujita[1], Shigeru Shinomoto[1] & Luis E. C. Rocha[2,3]

Received: 14 June 2017
Accepted: 5 November 2017
Published online: 16 November 2017

The Syrian armed conflict has been ongoing since 2011 and has already caused thousands of deaths. The analysis of death tolls helps to understand the dynamics of the conflict and to better allocate resources and aid to the affected areas. In this article, we use information on the daily number of deaths to study temporal and spatial correlations in the data, and exploit this information to forecast events of deaths. We found that the number of violent deaths per day in Syria varies more widely than that in England in which non-violent deaths dominate. We have identified strong positive auto-correlations in Syrian cities and non-trivial cross-correlations across some of them. The results indicate synchronization in the number of deaths at different times and locations, suggesting respectively that local attacks are followed by more attacks at subsequent days and that coordinated attacks may also take place across different locations. Thus the analysis of high temporal resolution data across multiple cities makes it possible to infer attack strategies, warn potential occurrence of future events, and hopefully avoid further deaths.

The outbreak of the current Syrian armed conflict occurred in March 2011 as a consequence of protests demanding democratic reforms and the end of the current government. These protests quickly escalated and within weeks were widespread in key cities all over Syria, eventually giving rise to groups against and in favor of the government[1]. Since then, the Syrian conflict has been marked by a large number of deaths of both civilians and military personnel. Although a matter of debate[2], estimates claim that 470,000 people have been killed and at least 3 million refuged or migrated to foreign countries by the end of 2015[3].

The dynamics of wars is complex and involves interdependent cultural, ethnic, political and economic variables that are difficult to model and predict[4–6]. Modeling efforts have been employed to forecast the outbreak of conflicts[6–11] and to understand their dynamics[5,12–14]. There is also much interest on estimating the number of casualties and death tolls. Such information helps to allocate resources, estimate the magnitude of the conflict, develop war strategies from the military and political points of view[15], and to quantify the burden of the war on health systems (needed for example to deliver humanitarian aid) and on the society[16–18]. Reliable data on death tolls are difficult to obtain and different methods exist to improve data collection during and after the conflict[17]. Higher resolution temporal data sets (at daily and weekly resolution) have however became increasingly available in recent years, allowing researchers to employ advanced methods of time-series analysis to make predictions on death tolls[19–22], extreme massacres[23] and to study the dynamics of conflicts[24–26].

In this article, we study the daily time series of death tolls in the current Syrian conflict and look for the possibility of detecting signs of war-related tragic events based on temporal correlation within individual cities and spatial correlation across different cities. We compare these results to the statistics of a benchmark country, England, which is a representative country not undergoing domestic armed conflicts, in which the daily number of deaths at different cities is expected to have no direct causal relation. We exploit these correlations to improve the forecast of violent deaths in Syria, information that can be used to better allocate resources and aid to affected regions.

We firstly investigate auto-correlations in the number of deaths in individual cities and cross-correlations across cities, and find that these correlations are significant and positive in Syria, suggesting a possible coordination of events in multiple cities. Secondly, we perform simulations by assimilating models to the characteristics of the real data such as slow non-stationary fluctuations or rapid daily correlations within each city, and examine the extent to which the observed temporal and inter-city correlations are explained by these apparent characteristics. Thirdly, we carry out the Granger causality analysis[27,28] to see if there are statistical causal relations across

[1]Department of Physics, Kyoto University, Kyoto, 606-8502, Japan. [2]Department of Sociology, Stockholm University, Universitetsvägen 10B, Stockholm, S-10691, Sweden. [3]Department of Public Health Sciences, Karolinska Institutet, 18A Tomtebodavägen, Stockholm, S-17177, Sweden. Correspondence and requests for materials should be addressed to L.E.C.R. (email: luis.rocha@ki.se)

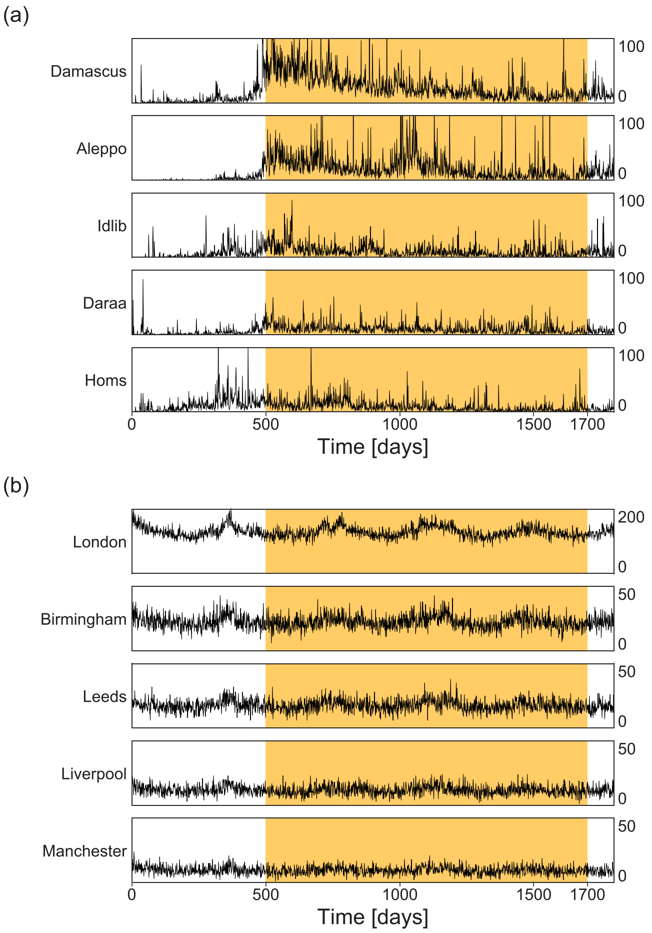


Figure 1. Daily time series of the number of deaths. The figure shows the number of deaths per day (y-axis) in different cities of (**a**) Syria and (**b**) England. The highlighted interval shows the time period of 1200 days for which we have performed the correlation and the Granger causality analyses. For the prediction analysis, the initial 600 days (500 to 1099 in Fig. 1(a)) is used to fix the parameters of the prediction models, and the remaining 600 days (from 1100 to 1699) is used to examine the predictability.

different cities. Finally, we attempt to predict the number of deaths; given a set of data for each country, we fix the parameters of prediction models using the initial part of the time series and then apply the models to the rest to validate if the models give a reasonable prediction on the future. The predictability depends not on the model but essentially on whether there is statistical causal relation between cities in each data set. We find for Syrian data that the vector auto-regression (VAR)[29] model that takes account of inter-city correlations outperforms the auto-regression (AR)[29] model that uses only single time series of individual cities separately. This indicates that the information of war-related events occurring in some cities may be used for warning of potential occurrence of future events in other cities.

Results

Death tolls.

The data used in this study come from The Violations Documentation Center (VDC) in Syria (www.vdc-sy.info). All data analyzed during this study are included in the Supplementary Information files. The VDC has been collecting information on death tolls in the Syrian conflict since June 2011 and retrospectively since the outbreak of the conflict in March 18, 2011. We aggregate data at the province level, adding together adults, children, civilians and military personnel to get a unified number of deaths per day. Figure 1(a) shows the time-series of death tolls after the outbreak, as to the top 5 provinces with most casualties, i.e. Damascus (including the suburbs), Aleppo, Idlib, Daraa and Homs. We focus our analysis in the times after the shock, starting at 500 days from the outbreak of the conflict, and during the following 1200 days.

The office for National Statistics (www.ons.gov.uk) provides the daily number of deaths in England, that we use as a reference, from 2010 to 2014. We select 5 large cities in England: the greater London area (hereafter referred simply as London), Birmingham, Leeds, Liverpool and Manchester. We also aggregate adults and children in this case. Figure 1(b) shows a seasonal pattern in which deaths are more common during winter in all studied cities. We match the starting dates in the two datasets.

Figure 1 suggests large variation in the number of deaths in the case of Syria but not in the case of England. To study these fluctuations, we build a histogram of the number of deaths per day and analyze different parametric models to the data (Fig. 2). We have modeled three versions of the normal distribution of the transformed

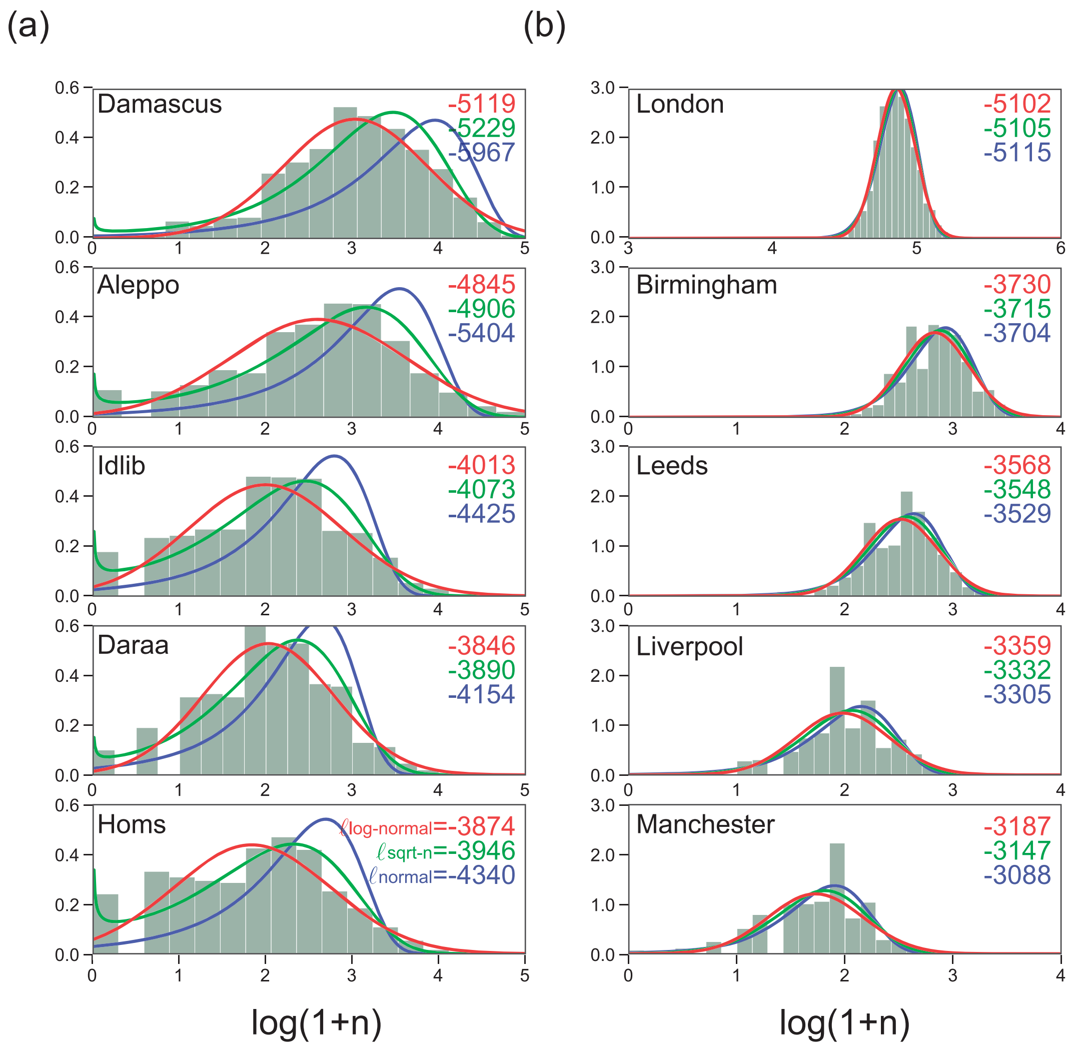


Figure 2. Distribution of number of deaths per day. (**a**) Syrian and (**b**) English cities. Distributions of $x = \log(1 + n)$ are plotted. The normal distributions of variables $x = n$, $x = \sqrt{n}$ and $x = \log(1 + n)$ fitted to each data are displayed on top of the histograms, with their goodness-of-fit represented in terms of the log-likelihood $\ell = \log L$. The bin-size of the histograms is determined by Scott's rule that provides the best fitting for a compact set of data, so to minimize the mean square error[35].

variables $x = n$, $x = \sqrt{n}$ and $x = \log(1 + n)$, where n is the original number of deaths per day. The model goodness-of-fit was corroborated by estimating the log-likelihood $\ell \equiv \log L$ of each model to the data. Our analysis indicates that the log-normal distribution (i.e. normal with $x = \log(n + 1)$) is the best model (highest ℓ) among the three alternatives for the Syrian data but all models give similar results for the English data.

The log-normal is a non-symmetric right skewed distribution in n, meaning that values much larger than the mean, i.e. many deaths in a single day, are relatively common in comparison to the standard normal distribution. Although alternative models could be tried, the log-normal distribution is appropriate and a standard choice when the underlying process consists of multiplicative random variables[30]. This is a reasonable assumption in our case given that the effects of shooting or bombing are non-linear and an attack may trigger multiplicative attacks locally or at different locations. Gomez-Lievano and collaborators[31] also suggest that a log-normal distribution is an appropriate model of violent deaths (homicides) in the context of Brazil, Colombia and Mexico. In the literature, different mechanisms have been proposed to generate log-normal distribution of events[32]. Lewis and collaborators[21], for example, recently suggested that a self-exciting point process may explain up to 50% of the violent deaths in the case of Iraq between the years 2003 and 2007. It remains an open question to determine the specific mechanisms driving violent deaths in our context and further analysis is necessary.

The Syrian data only counts violent deaths. The English data counts both violent and non-violent deaths. A more detailed model for England could thus combine log-normal and normal distributions to account simultaneously to both violent and non-violent deaths. However, the majority (more than 90%) of deaths in England are non-violent (e.g. non-communicable diseases–www.ons.gov.uk) and a sum of independent Bernoulli random variables given rise to a normal distribution is a reasonable model. Though most English cities exhibited slightly larger likelihood for the normal distribution, Syrian data in question exhibited absolutely larger likelihood for the log-normal. Accordingly, from now on, we analyze all the time series (including English data) using the transformed variables $x = \log(1 + n)$. Note that the results are essentially unchanged even if we consider $x = n$ for modeling the English data.

Correlation Analysis. The correlation analysis reveals the direction and strength of the relationship between two time-series, or of the same time-series at different times[29]. In our context, we will analyze the temporal structure of the daily deaths in individual cities (i.e. auto-correlation) and the spatial correlation across cities in the same country (i.e. cross-correlation). The cross-correlation of the daily deaths between cities i and j is given by

$$\phi_{ij}(t) = \frac{\frac{1}{T-t}\sum_{s=1}^{T-t}\left(x_i(s+t) - \overline{x}_i)(x_j(s) - \overline{x}_j\right)}{\frac{1}{T}\left((\sum_{s=1}^{T}(x_i(s) - \overline{x}_i)^2)(\sum_{s=1}^{T}(x_j(s) - \overline{x}_j)^2)\right)^{\frac{1}{2}}} \tag{1}$$

where t is the time difference measured in the unit of day, $T = 1200$ in our case, and $\overline{x}_i \equiv \sum_{t=1}^{T} x_i(t)/T$. The auto-correlation in city i is thus given by $\phi_{ii}(t)$. The value of $\phi_{ij}(t)$ varies between -1 (negatively correlated) and $+1$ (positively correlated) at each time t, with $\phi_{ij}(t) = 0$ indicating no correlation.

The correlation functions $\{\phi_{ij}(t)\}$ computed for the time series of Syria and England are displayed in Fig. 3(a) and (b), respectively. The diagonal elements represent auto-correlations in individual cities, i.e. $\{\phi_{ii}(t)\}$. For Syria, the prominent positive auto-correlation lasting for a week, particularly in Damascus ($\phi_{11}(t) \sim 0.6$) and Aleppo ($0.4 \lesssim \phi_{22}(t) \lesssim 0.5$) but also in Homs ($0.3 \lesssim \phi_{55}(t) \lesssim 0.4$), indicates that high (low) death tolls in one day are followed by high (low) death tolls on the next day in the same city. The correlation analysis does not necessarily mean causality but the results suggest that individual war-related events may have caused a number of deaths in the subsequent days or that major attacks (and thus deaths) trigger a series of new attacks with further deaths. Similarly, peaceful days are followed by further peaceful days. Note that auto-correlation is also present in London to some extent ($\phi_{11}(t) \sim 0.4$), where war-related conflicts are inexistent, but we shall see that this is mostly due to seasonal variation.

The spatial, inter-city or cross-correlation is represented as the off-diagonal elements in Fig. 3. We observe positive cross-correlation across Syrian cities, in particular from Damascus to Aleppo ($\phi_{21}(t) > 0.3$, from Damascus to Idlib ($\phi_{31}(t) \sim 0.3$), and from Damascus to Homs ($\phi_{51}(t) > 0.3$). These correlations are not as strong as the auto-correlations but their positive direction and strength suggest that high (small) death tolls in Damascus are followed by high (small) death tolls in Aleppo, Idlib and Homs (Fig. 3(a)). Similarly, significant cross-correlation is observed from Homs to Aleppo ($\phi_{25}(t) > 0.3$), from Aleppo, Idlib and Homs to Damascus, respectively $\phi_{12}(t) > 0.3, \phi_{13}(t) \sim 0.3$ and $\phi_{15}(t) > 0.3$, and from Aleppo to Homs ($\phi_{52}(t) > 0.3$). Note that some cross-correlation is also present in a few cases in England, though at a relatively low intensity (Fig. 3(b)).

Figure 1 indicates that non-stationary fluctuations of long timescale are taking place commonly across different cities in each country; the death tolls in Syria have been slowly decreasing on average (Fig. 1(a)), while those in England exhibit seasonal modulation (Fig. 1(b)). To examine the extent to which the observed temporal and spatial correlations measured in the real data are explained by these slow fluctuations, we create three assimilated time series null models for each city and repeat the correlation analysis for each case. Each model reproduces different features of the time series, with increasing complexity.

Model 0: Stationary uncorrelated time series. We first generate a series of independent Gaussian random values $\xi_i(t)$, given the mean ($\overline{x}_i$) and variance ($\overline{x}_i^2 - \overline{x}_i^2$) of each city i (see Table 1):

$$x_i^{(0)}(t) = \xi_i(t). \tag{2}$$

Model 1: Non-stationary uncorrelated time series. We modulate the stationary time series of Model 0 according to the slow modulation observed in each city, which can be obtained by smoothing the original data with the Gaussian kernel with the standard deviation of $k = 10$ days,

$$\Delta\tilde{x}_i(t) = \sum_{s=1}^{T} x_i(s)\frac{1}{\sqrt{2\pi}\,k}e^{-\frac{(t-s)^2}{2k^2}} - \overline{x}_i. \tag{3}$$

The smoothed modulation is added to the time series of Model 0 as

$$x_i^{(1)}(t) = x_i^{(0)}(t) + \Delta\tilde{x}_i(t). \tag{4}$$

This addition in terms of the logarithmic coefficient $x = \log(1 + n)$ corresponds to multiplying the modulation to the original number of deaths.

Model 2: Non-stationary correlated time series. The strong temporal correlation $\phi_{ii}(t)$ lasting for a few days in Syrian cities may be reproduced by adding memory h_i to the random variable $y_i(t)$, such that

$$y_i(t) = ((1 - h_i)\xi_i(t) + h_i y_i(t - 1). \tag{5}$$

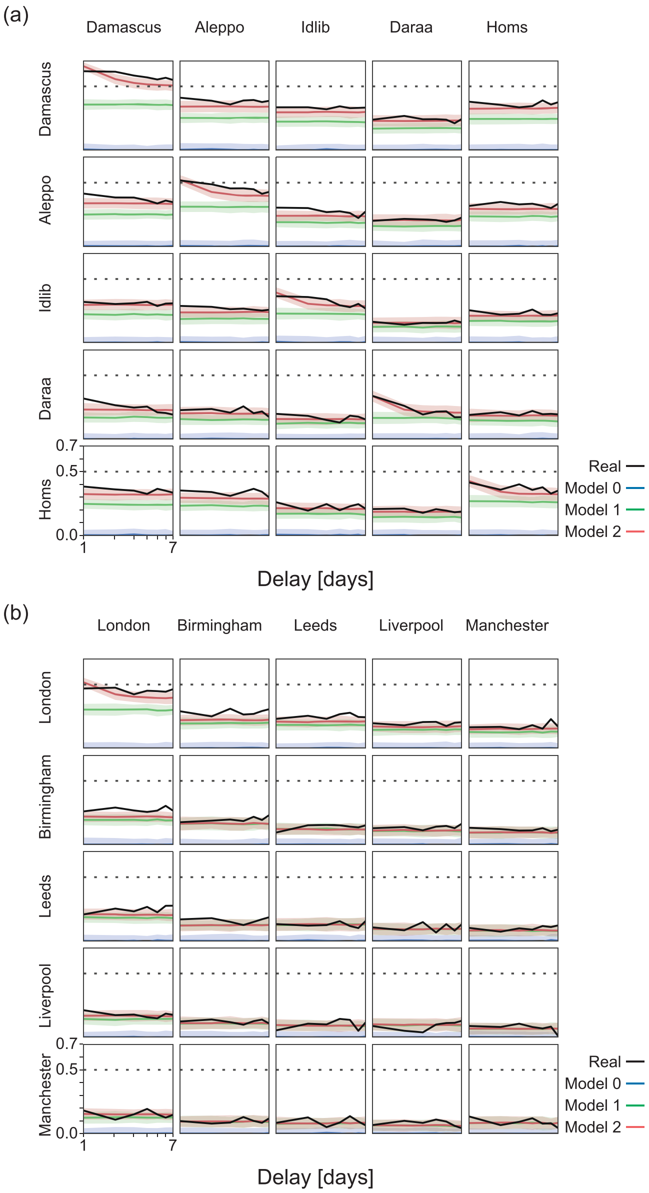


Figure 3. Correlation analysis. Temporal and spatial correlation represented by the auto- and cross-correlations across cities, $\{\phi_{ij}(t)\}$ (y-axis) in (**a**) Syria and (**b**) England. The real data is show by black curves. Model 0 (blue curves) represents a stationary time series of independent Gaussian random numbers, given the mean and variance of each city. Model 1 (green curves) takes account of slow non-stationary modulation in each city. Model 2 (red curves) takes account of daily correlation. The memory parameters were chosen as $h_{\text{Damascus}} = 0.3$, $h_{\text{Aleppo}} = 0.2$, $h_{\text{Idlib}} = 0.15$, $h_{\text{Daraa}} = 0.15$, $h_{\text{Homs}} = 0.15$, $h_{\text{London}} = 0.2$, $h_{\text{Birmingham}} = 0$, $h_{\text{Leeds}} = 0$, $h_{\text{Liverpool}} = 0$, $h_{\text{Manchester}} = 0$. The shaded area represents the 90% interval (obtained by repeating simulations) of the stochastic deviation of models 0, 1, and 2. The delay t in the x-axis is represented in the logarithmic scale. Dotted lines are used as eye-guides and indicate $\{\phi_{ij}(t)\} = 0.5$.

We might add higher order memory terms if needed. A stationary correlated time series may be constructed by iterating this equation. By adding the slow fluctuation $\Delta\tilde{x}_i(t)$ (Eq. 3) to the stationary time series $y_i(t)$, we obtain a non-stationary correlated time series:

$$x_i^{(2)}(t) = y_i(t) + \Delta\tilde{x}_i(t). \tag{6}$$

Figure 3 summarizes the results of the correlation analysis (Eq. 1) applied to Models 0, 1, and 2 in comparison to the real data. As expected, the uncorrelated stationary time series (Model 0) did not exhibit significant correlation. Model 1 that adopted the slow modulation has reproduced the most part of the slow correlations in English data, but has not succeeded in reproducing the strong auto-correlation and cross-correlation in Syrian data. Model 2 was able to reproduce the strong auto-correlation of the Syrian data by suitably accommodating the memory parameter h_i (Fig. 3). Nevertheless, the strong correlations across real Syrian cities (and weaker correla-

Syrian Cities	Mean	Variance
Damascus	3.03	0.83
Aleppo	2.60	1.01
Idlib	1.99	0.89
Daraa	2.03	0.75
Homs	1.84	0.91
English Cities	**Mean**	**Variance**
London	4.86	0.13
Birmingham	3.12	0.23
Leeds	2.89	0.26
Liverpool	2.48	0.32
Manchester	2.3	0.33

Table 1. Mean ($\overline{x}_i$) and variance ($\overline{x}_i^2 - \overline{x}_i^2$) of the number of deaths in Syrian and English cities. The number of deaths is measured as $x = \log(1 + n)$.

tions across real English cities) were not fully reproduced with Model 2. These results suggest that observed inter-city correlations are genuine and these violent death tolls are correlated between some cities. For example, an increase (decrease) of deaths in Damascus is accompanied by an increase (decrease) of deaths in Aleppo, Daraa and Homs, or an increase (decrease) in Aleppo is followed by an increase (decrease) in Homs. The weak cross-correlations observed in some English cities may reflect the known effect of synchronization of deaths due to seasonality with peaks during winter months[33]. The large number of deaths and seasonality may also explain the larger auto-correlations observed in London but not in other English cities.

Altogether, these results suggest that there may be some coordination in the attacks (and consequently deaths) at the different Syrian cities, or in other words, attacks may spread to different locations. Although this effect is not as strong, it is significantly higher than one would expect if no violent conflict was going on across the country, given that cross-correlation is close to zero between cities in England. In the next section, we will look more carefully to these inter-city correlations and exploit this information to make better forecast of death tolls.

Granger causality. We try to detect statistical causal relation between the different cities in both countries, using a standard pairwise-conditional Granger causality (GC) analysis[27,28]. The GC analysis indicates if the evolution of the number of deaths at city *i* is better predicted by combining information from the past deaths in both cities *i* and *j* than if using only the past information of deaths in city *i*. This is an extension of the previous analysis because it measures the improvement on forecast in the presence of information from other variables.

The naive GC analysis indicated many inter-city correlations (indicated by low *p*-values) not only in Syria (Fig. 4(a)) but also in England (Fig. 4(b)), although the direct causal relations between English cities are expected to be absent. The GC analysis is known to be vulnerable to non-stationary fluctuations and likely to suggest spurious correlations[34]. By applying the same GC analysis to the assimilated data (Model 2 described in the previous section), we also obtained similar correlations (Fig. 4(c) and (d)). The result suggests that the seasonal modulation in the number of deaths has induced the observed causality in the English cities.

The influence of slow non-stationary fluctuations (such as seasonality) to the GC analysis may be mitigated by analyzing the temporal difference[29],

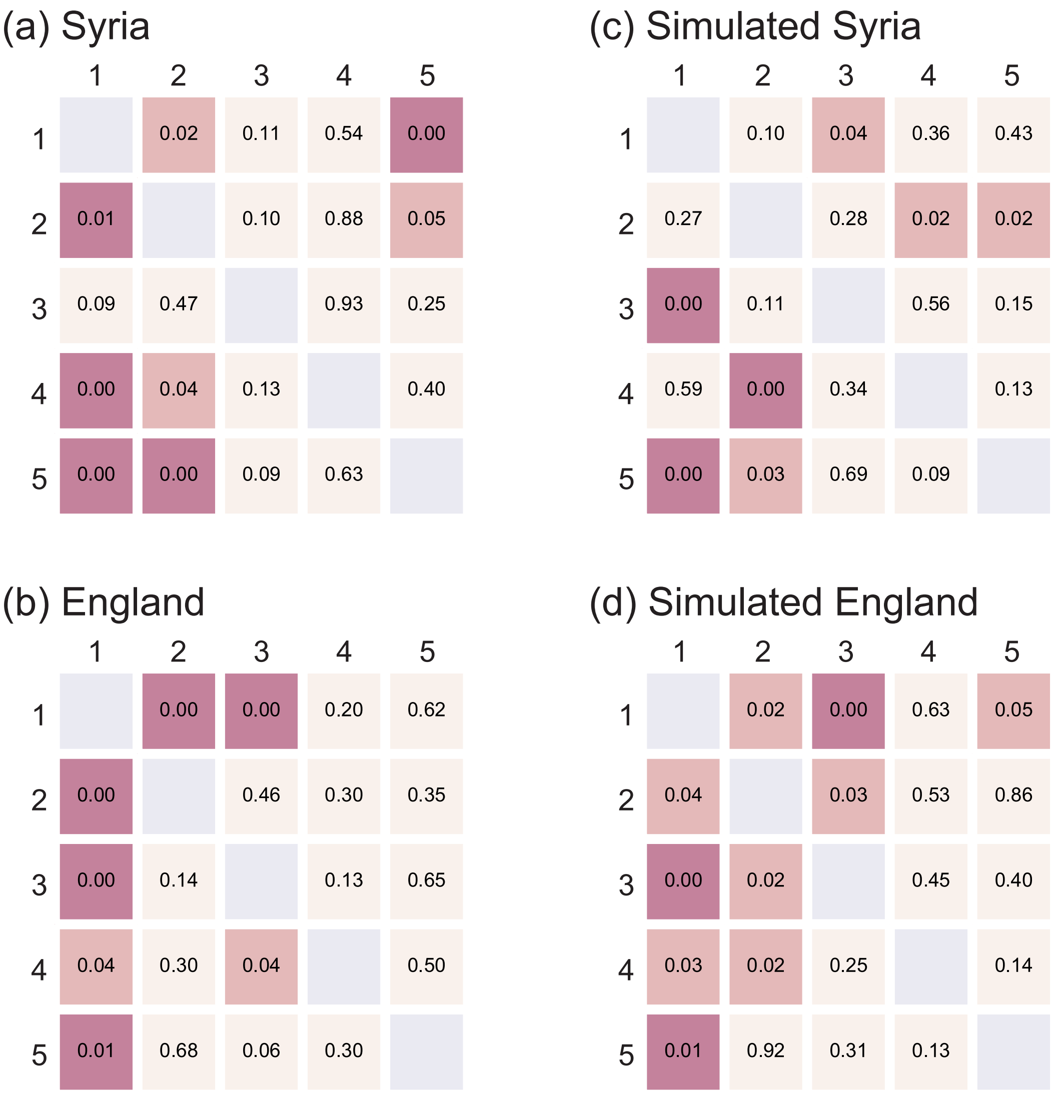
$$z_i(t) \equiv x_i(t) - x_i(t-1). \tag{7}$$

In our case, this operation succeeded in removing inter-city correlations from not only the simulated Model 2 (Fig. 5(c) and (d)), but also from real English data (Fig. 5(b)). However, the real Syrian data is left with directed inter-city correlations from Damascus to Aleppo and Daraa, from Aleppo to Damascus, and from Idlib to Homs (low *p*-values in Fig. 5(a)). Thus the GC analysis indicates that death tolls are dependent across these Syrian cities but not across English or on simulated cities, in which the original causal relations were an effect of the slow fluctuations.

We note that "slow" and "fast" fluctuations should be treated with care. It is easier for the GC or other analysis methods to detect rapid changes in the number of events (in our case, death tolls). For example, the GC analysis may interpret "causal" if a deadly infectious disease propagates from city to city in a few days, even if there is no physical causal influence between cities. This is unavoidable, because the analysis method simply count the death tolls without inspecting the other information. A slower propagation of an infectious deadly disease such as the plague in the 14th century might also be detected using GC if the changes in the death tolls were large.

Forecast. We attempt to predict the death tolls in Syria and England using the auto-regression (AR) model (Eq. 8) and the vector auto-regression (VAR) model (Eq. 9)[29]. We use the initial 600 days of the time series (500 to 1099 in Fig. 1(a)) to fix the model parameters, and apply the models to the remaining 600 days (from 1100 to 1699) to see if they give efficient prediction on the future number of deaths in each country.

The AR model only uses information of a single time series to forecast values of the corresponding time series, as given by



(a) Syria

	1	2	3	4	5
1		0.02	0.11	0.54	0.00
2	0.01		0.10	0.88	0.05
3	0.09	0.47		0.93	0.25
4	0.00	0.04	0.13		0.40
5	0.00	0.00	0.09	0.63	

(c) Simulated Syria

	1	2	3	4	5
1		0.10	0.04	0.36	0.43
2	0.27		0.28	0.02	0.02
3	0.00	0.11		0.56	0.15
4	0.59	0.00	0.34		0.13
5	0.00	0.03	0.69	0.09	

(b) England

	1	2	3	4	5
1		0.00	0.00	0.20	0.62
2	0.00		0.46	0.30	0.35
3	0.00	0.14		0.13	0.65
4	0.04	0.30	0.04		0.50
5	0.01	0.68	0.06	0.30	

(d) Simulated England

	1	2	3	4	5
1		0.02	0.00	0.63	0.05
2	0.04		0.03	0.53	0.86
3	0.00	0.02		0.45	0.40
4	0.03	0.02	0.25		0.14
5	0.01	0.92	0.31	0.13	

Figure 4. Granger causality analysis applied to the original data. The *p*-values of Granger causality for real (**a**) Syrian and (**b**) English cities, and for the simulated (**c**) Syrian and (**d**) English cities given by Model 2. For Syrian data: 1: Damascus; 2: Aleppo; 3: Idlib; 4: Daraa; 5: Homs. For English data: 1: London; 2: Birmingham; 3: Leeds; 4: Liverpool; 5: Manchester. Low *p*-values indicate statistical causality.

$$\hat{x}_i(t) = c + \sum_{s=1}^{m} a_i(s)x_i(t-s) + \varepsilon(t), \tag{8}$$

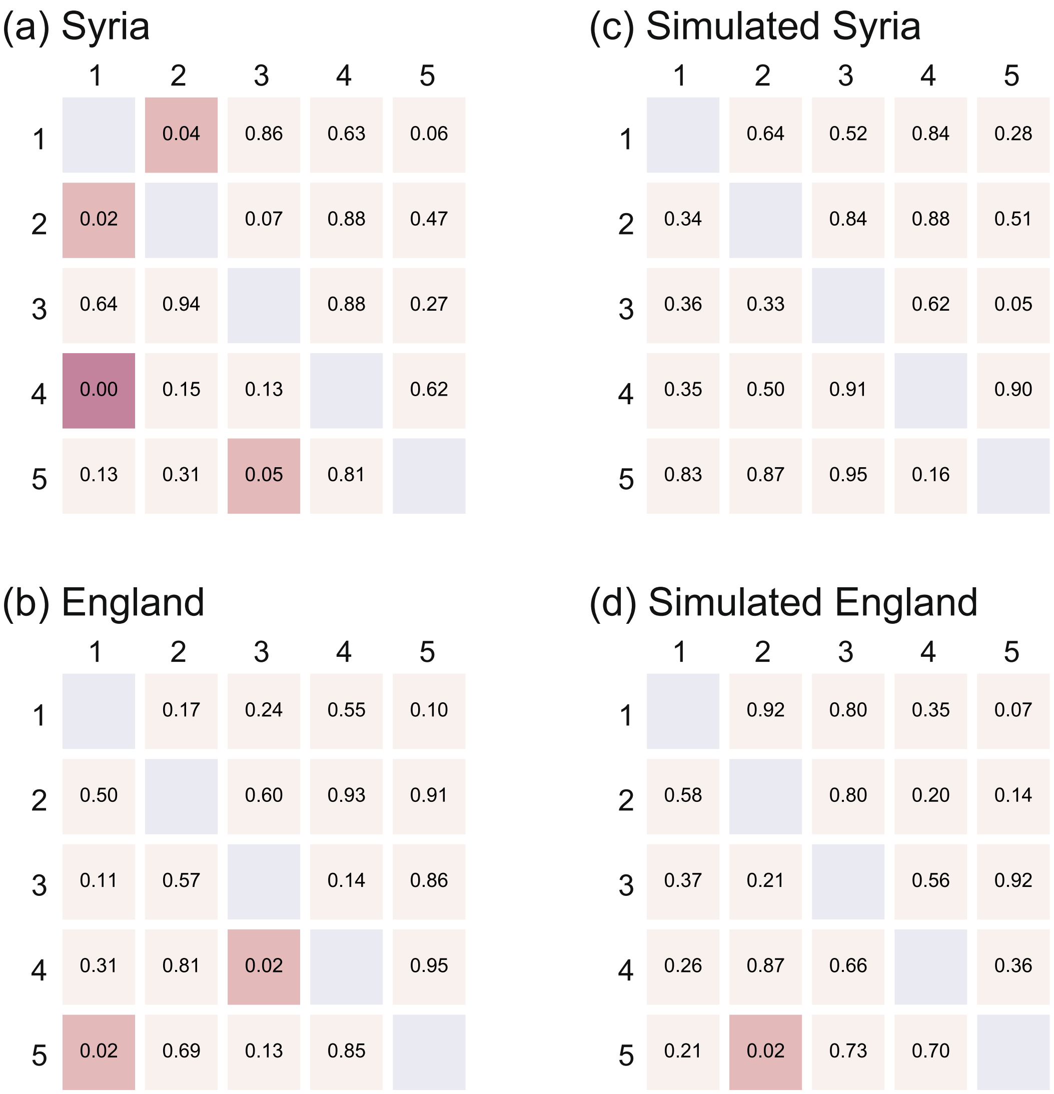
where $\hat{x}_i(t)$ represents the predicted value, $\{a_i(s)\}_{s=1,\cdots,m}$ are parameters determined with the temporal auto-correlation $\{\phi_{ii}(s)\}_i$ (Eq. 1) computed for the initial 600 days of the time series, and m is the order of regression. We take $m = 5$ for the current analysis.

The VAR model uses the time series of all cities simultaneously to forecast values of all cities at once, as given by

$$\hat{x}(t) = c + \sum_{s=1}^{m} \mathbf{A}(s)x(t-s) + \varepsilon(t), \tag{9}$$

where $\mathbf{x}(t)$ represents a vector comprising of five cities $(x_1(t), x_2(t), \cdots, x_5(t))^t$, and $\mathbf{A}(s)$ is a matrix whose elements are determined with the temporal correlations $\{\phi_{ij}(s)\}_{ij}$ (Eq. 1).

If there are significant correlations between cities, there is room for the VAR model to make the better forecast in comparison to the AR model that only uses information of a single time series (i.e. single city). As a reference, these two models are compared with two simpler models: (i) predicting the future deaths simply with the number of deaths of the preceding date (Preceding), and (ii) predicting deaths with a single fixed value given by averaging over the number of deaths during the past first half of the time series (Average).



(a) Syria

	1	2	3	4	5
1		0.04	0.86	0.63	0.06
2	0.02		0.07	0.88	0.47
3	0.64	0.94		0.88	0.27
4	0.00	0.15	0.13		0.62
5	0.13	0.31	0.05	0.81	

(c) Simulated Syria

	1	2	3	4	5
1		0.64	0.52	0.84	0.28
2	0.34		0.84	0.88	0.51
3	0.36	0.33		0.62	0.05
4	0.35	0.50	0.91		0.90
5	0.83	0.87	0.95	0.16	

(b) England

	1	2	3	4	5
1		0.17	0.24	0.55	0.10
2	0.50		0.60	0.93	0.91
3	0.11	0.57		0.14	0.86
4	0.31	0.81	0.02		0.95
5	0.02	0.69	0.13	0.85	

(d) Simulated England

	1	2	3	4	5
1		0.92	0.80	0.35	0.07
2	0.58		0.80	0.20	0.14
3	0.37	0.21		0.56	0.92
4	0.26	0.87	0.66		0.36
5	0.21	0.02	0.73	0.70	

Figure 5. Granger causality analysis applied to the time difference data $z_i(t) \equiv x_i(t) - x_i(t-1)$. The p-values of Granger causality for real (**a**) Syrian and (**b**) English cities, and for simulated (**c**) Syrian and (**d**) English cities given by Model 2. For Syrian data: 1: Damascus; 2: Aleppo; 3: Idlib; 4: Daraa; 5: Homs. For English data: 1: London; 2: Birmingham; 3: Leeds; 4: Liverpool; 5: Manchester. Low p-values indicate statistical causality.

Table 2 demonstrates the performance of these four models for the Syrian and English cities in terms of the average prediction error $MSE \equiv \sum_{t=T+1}^{2T}(\hat{x}_i(t) - x_i(t))^2/T$ over the testing period of $T = 600$ days, where the forecast provided by each model is given by $\hat{x}_i(t)$ and the true value is given by $x_i(t) = \log(1 + n_i(t))$. The AR and VAR models generally perform much better than the other two methods. Among the two regression models, VAR outperforms AR model for the Syrian data. The increase in the performance for the prediction errors $\Delta \equiv (MSE_{\text{AR}} - MSE_{\text{VAR}})/MSE_{\text{AR}}$ in Table 2 is large for three cities in Syria, in contrast to the negligible improvement in the English cities (less than 2%).

The comparison of prediction models in Table 2 is made assuming the log-normal distribution, or by setting $x = \log(1 + n)$. We firstly compared different kinds of distribution such as the normal distributions of variables $x = n, x = \sqrt{n}$ and $x = \log(1 + n)$ in terms of the likelihood values. Here we also compare them in terms of the performance of predicting the number of deaths. Figure 6(a) and (b) demonstrate the predictions made for the Syrian city of Daraa and the English city of Liverpool using the VAR of $x = n, x = \sqrt{n}$ and $x = \log(1 + n)$. It is observed that $x = \log(1 + n)$ is superior to $x = \sqrt{n}$ and $x = n$ for the data of Daraa, while three variables make no significant difference for the data of Liverpool.

To better visualize the impact of these errors in the forecast of death tolls, we measure the total prediction error (E) in the unit of the number of dead people. Considering all five cities of each country together, we have

$$E = \sum_{i=1}^{5}\sum_{t=1}^{T}|n_i(t+T) - \hat{n}_i(t+T)|, \tag{10}$$

	Preceding	Average	AR	VAR	Δ(%)
Damascus	0.717	0.552	0.500	0.478	+4.4
Aleppo	1.383	1.064	1.016	1.006	+1.0
Idlib	1.295	0.823	0.787	0.791	−0.5
Daraa	0.901	0.638	0.635	0.570	+10.2
Homs	1.173	0.711	0.831	0.732	+11.9
London	0.017	0.017	0.011	0.011	0
Birmingham	0.088	0.059	0.052	0.051	+1.9
Leeds	0.110	0.066	0.061	0.060	+1.6
Liverpool	0.187	0.101	0.101	0.100	+1.0
Manchester	0.194	0.113	0.108	0.108	0

Table 2. Forecast error for Syrian and English cities. The table shows the mean square error (*MSE*) between the forecast provided by each model $\hat{x}_i(t)$ and the true value $x_i(t)$. The column Δ stands for the improvement of the prediction errors of the VAR model in respect to the AR model.

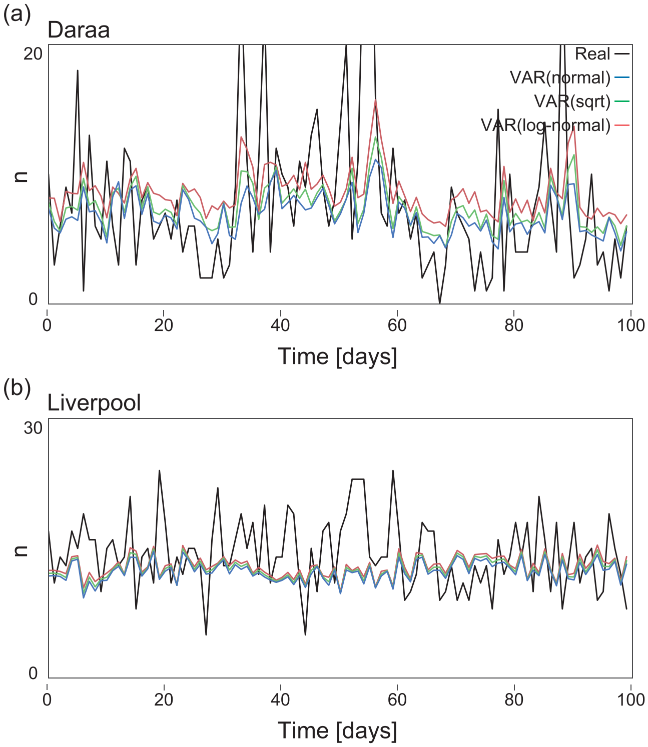


Figure 6. Prediction of the number of deaths. The plots show the number of deaths (n) from 1100 to 1199 days (i.e. the initial 100 days of the latter half of the time series) which is used for evaluating the predictability power of the models. (**a**) Syrian city of Daara and (**b**) English city of Liverpool. Predictions made by VAR of $x = n$, $x = \sqrt{n}$ and $x = \log(1 + n)$ are plotted on top of the real time series.

where $T = 600$ days, $n_i(t)$ is the number of deaths of city i at day t, and $\hat{n}_i(t)$ is the predicted value given by $\hat{n}_i(t) = x_i(t)$ if we use the transformed variable $x_i(t) = n_i(t)$, $\hat{n}_i(t) = \hat{x}_i(t)^2$ if we use $x_i(t) = \sqrt{n_i(t)}$ and $\hat{n}_i(t) = e^{\hat{x}_i(t)} - 1$ if we use $x_i(t) = \log(1 + n_i(t))$. Figure 7 shows that the VAR model using the transformed variable $x_i(t) = \log(1 + n_i(t))$ gives the best prediction of future deaths. In the best case, the difference in the total prediction error between AR and VAR models corresponds to 451 people in the period of 600 days. In fact, for all transformed variables, the VAR model gives better prediction than the AR model, that is, a difference of 1010 deaths for $x_i(t) = \sqrt{n_i(t)}$ and 2337 deaths for $x_i(t) = n_i(t)$. These differences between VAR and AR are substantially smaller in the case of England; the cumulative difference sums up to 98 deaths for $x_i(t) = \log(1 + n_i(t))$. For the remaining, 115 deaths for $x_i(t) = \sqrt{n_i(t)}$ and 111 for $x_i(t) = n_i(t)$.

These results indicate a significant improvement in the prediction power of death tolls in Syria when using information from multiple cities and the log-transformed variables. In comparison, the second best model (AR

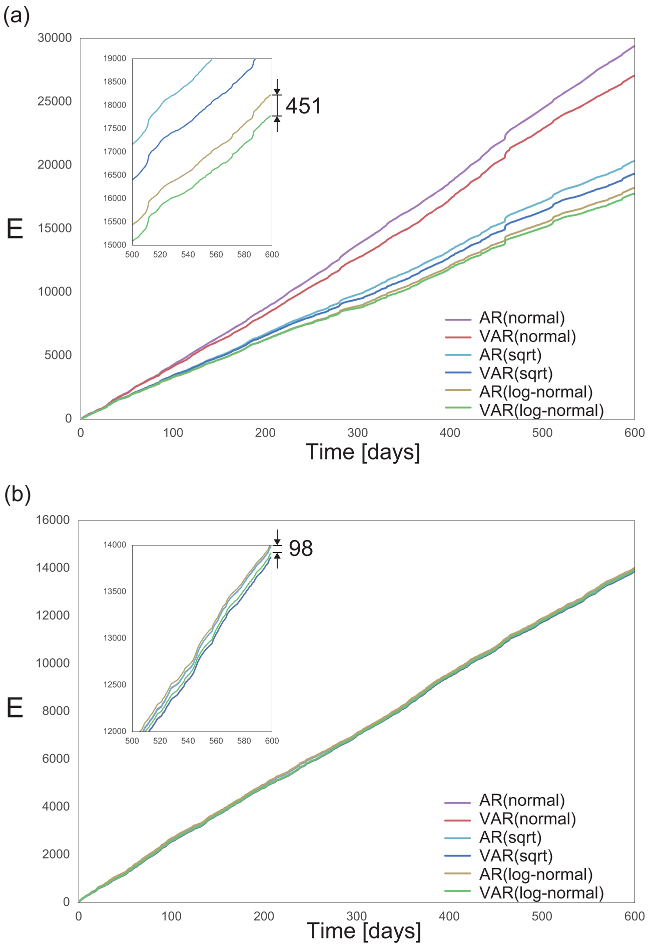


Figure 7. Prediction of the number of deaths. The plots show the cumulative difference *E* (Eq. 10) between the predicted and the actual number of deaths from 1100 to 1699 days for the AR and VAR models, using three different transformed variables for (**a**) Syria and (**b**) England.

with log-transformed variables) predicts about 0.75 more deaths per day. While this might sound a small difference, it accumulates to about 23 people within a month. From the analytical point of view, these results emphasize that the cross-correlations in the death tolls at different cities in Syria are significantly higher than those in the English cities, in which all tested models give relatively similar results.

Discussions

Armed conflicts typically cause significant life losses on all sides. Estimates of death tolls are important in order to quantify the magnitude of the war, encourage peacemaking and to allocate resources and humanitarian aid. The availability of high-resolution spatiotemporal data on the number of deaths allows researchers to analyze correlations between different cities at different times and to identify trends that could possibly be used to reduce future causalities.

In this paper, we use daily information on the number of deaths in a given city to study spatial and temporal correlations of death tolls in the current Syrian conflict and compared the results with the daily number of deaths in English cities that are not undergoing any domestic conflict. We have explored different models to remove potential virtual correlations in the empirical data, as was the case in English cities, mainly due to seasonality. Our analysis showed that significant positive auto-correlation exist in Syrian cities, meaning that days with a high (low) number of deaths in a particular city were followed by days with many (few) deaths in the same city, possibly reflecting a sequence of attacks within short periods triggered by a single attack. Similarly, we have also observed significant cross-correlation (i.e. spatial correlation) between some cities in Syria. This means that deaths in one city were accompanied by deaths at another city, for example, from Damascus to Aleppo and vice-versa, from Damascus to Daraa, and from Idlib to Homs. Given the available data, we cannot infer mechanistic causality in deaths between different cities but the cross-correlations and Granger causality analysis suggest that events are not completely random but coordinated attacks at multiple locations possibly take place. At the same time, correlations are typically not super strong and are not observed between all cities too. Such results are useful since one can exploit them to develop a warning system monitoring the sequence of events taking place at different days and cities aiming to better understand attack strategies and avoid further deaths. For example, since the number of deaths increases in both Damascus and Aleppo, attacks in one city may trigger warnings to the other.

We have also explored the possibilities to forecast death tolls at different cities. Our analysis have shown that due to the significant cross-correlations, improved forecast is obtained for Syria if using information from all cities simultaneously in a vector auto-regression model in comparison to single cities in independent auto-regression models. Contrastingly, the difference is very small (typically less than 2%) for the case of England. The important conclusion here is that death tolls during the conflict in Syria can be better predicted if information from multiple locations and times are simultaneously available. Such prediction could be used to organize the allocation of resources and aid during the conflict.

We observe that daily death tolls in Syria can be well-described by log-normal distributions which is in contrast to English cities, in which death tolls can be described equally well by the three distributions examined. This means that death events in Syria are not uniform and days with a much larger number of deaths than the average are expected during the conflict. This is consistent with the previous studies suggesting that the distribution of the number of violent deaths may be described by log-normal distributions given the multiplicative nature of violent events.

We should keep in mind that the numbers of casualties in Syria have been very large and the counting of deaths is a very difficult task. Accordingly the counts might have been accompanied with errors that could have affected the correlations between specific cities and the overall forecast exercise. Future work should focus on the analysis of different datasets to determine the intensity of the correlations and thus the possibilities for forecast on other contexts. Furthermore, it remains an open question to determine the mechanisms driving the positive correlations between the death tolls in different cities and in the same city at different times.

References

1. Lynch, M. *The political science of Syria's war*. POMEPS Briefings 22 (2013).
2. Taylor, A. *The Syrian war's death toll is absolutely staggering. But no one can agree on the number*. The Washington Post (2016).
3. Syrian Centre for Policy Research. *Confronting fragmentation! Impact of Syrian crisis report*. 1–66 (2015).
4. Hegre, H., Karlsen, J., Nygard, H. M., Strand, H. & Urdal, H. Predicting Armed Conflict. *Inter. Stud. Q.* **57**, 250–270 (2012).
5. Kress, M. Modeling Armed Conflict. *Science* **336**, 865–869 (2012).
6. Cederman, L.-E. & Weidmann, N. B. Predicting armed conflict: Time to adjust our expectations? *Science* **355**, 474–476 (2017).
7. Lim, M., Metzler, R. & Bar-Yam, Y. Global pattern formation and ethnic/cultural violence. *Science* **317**, 1540–1544 (2007).
8. Brandt, P. T., Freeman, J. R. & Schrodt, P. A. Real time, time series forecasting of inter- and intra-state political conflict. *Conflict Manag. Peace Sci.* **28**, 41–64 (2011).
9. Chadefaux, T. Early warning signals for war in the news. *J. Peace Res.* **51**, 5–18 (2014).
10. Helbing, D. *et al*. Saving human lives: What complexity science and information systems can contribute. *J. Stat. Phys.* **158**, 735–781 (2015).
11. Parens, R. & Bar-Yam, Y. Step by step to peace in Syria. *arXiv:1602.06835* (2016).
12. Morgenstern, A. P. *et al*. Modeling political conflict, violence, and wars: A survey. *Am. J. Phys.* **81**, 805–814 (2013).
13. Stauffer, D. A biased review of sociophysics. *J. Stat. Phys.* **151**, 9–20 (2013).
14. Johnson, N. F., Restrepo, E. M. & Johnson, D. E. *Modeling Human Conflict and Terrorism Across Geographic Scales*, pages 209–233. (Springer International Publishing, Cham 2015).
15. The Economist. *And now, the war forecast*. (Technology Quarterly, Q3 2005).
16. Hicks, M. H.-R., Dardagan, H., Serdan, G. G., Bagnall, P. M., Sloboda, J. A. & Spagat, M. Violent deaths of Iraqi civilians: Analysis by perpetrator, weapon, time, and location. *PLoS Med.* **8**, e1000415 (2011).
17. Seybolt, T. B., Aronson, J. D. & Fischhoff, B. *Counting Civilian Casualties*. (Oxford University Press, New York 2013).
18. Guha-Sapir, D. *et al*. Civilian deaths from weapons used in the Syrian conflict. *BMJ* **351**, 1–5 (2015).
19. Rusch, T., Hofmarcher, P., Hatzinger, R. & Hornik, K. Modeling mortality rates in the Wikileaks Afghanistan war logs. *Institute for Statistics and Mathematics, WU Wirtschaftsunversit Wien,* Vienna, Technical Report 112 (2012).
20. Zammit-Mangion, A., Dewar, M., Kadirkamanathan, V. & Sanguinetti, G. Point process modelling of the Afghan war diary. *Proc. Nat. Acad. Sci. USA* **109**, 12414–12419 (2012).
21. Lewis, E., Mohler, G., Brantingham, P. J. & Bertozzi, A. L. Self-exciting point process models of civilian deaths in Iraq. *Security J.* **25**, 244–264 (2012).
22. White, G., Ruggeri, F. & Porter, M. D. Endogenous and exogenous effects in contagion and diffusion models of terrorist activity. *arXiv:1612.02527* (2016).
23. Scharpf, A., Schneider, G., Nöh, A. & Clauset, A. Forecasting the risk of extreme massacres in Syria. *Eur. Rev. Int. Stud.* **2**, S50–S68 (2014).
24. Jaeger, D. A. & Paserman, M. D. The cycle of violence? An empirical analysis of fatalities in the Palestinian - Israeli conflict. *Am. Econ. Rev.* **98**, 1591–1604 (2008).
25. Haushofer, J., Biletzki, A. & Kanwisher, N. Both sides retaliate in the Israeli-Palestinian conflict. *Proc. Nat. Acad. Sci. USA* **107**, 17927–17932 (2010).
26. Kurzman, C. & Hasnain, A. When forecasts fail: Unpredictability in Israeli-Palestinian interaction. *Sociol. Sci.* **1**, 239–259 (2014).
27. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
28. Honerkamp, J. *Statistical Physics: An advanced approach with applications*. (Springer-Verlag, Berlin 1998).
29. Hamilton, J. D. *Time Series Analysis*. (Princeton University Press, Princeton 1994).
30. Telser, L. G. The Lognormal Distribution Am. *J. Agricul. Econom* **41** (1959).
31. Gomez-Lievano, A., Youn, H. & Bettencourt, L. M. A. The Statistics of Urban Scaling and Their Connection to Zipfs Law. *PLoS ONE* **7**, e40393 (2012).
32. Crow, E. L. & Shimizu, K. *Lognormal Distributions: Theory and Applications*. (Marcel Dekker, New York 1987).
33. Johnson, H. & Griffiths, C. Estimating excess winter mortality in England and Wales. *Health Stat. Q.* **20**, 19–24 (2003).
34. Granger, C. W. J. Some recent development in a concept of causality. *J. Econom.* **39**, 199–211 (1988).
35. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. (John Wiley, New York 1992).

Acknowledgements

The authors thank Ryota Kobayashi for fruitful discussions. SS and LECR were supported by the Bilateral Joint Research Project between JSPS, Japan, and FRS-FNRS, Belgium. SS was supported by Grants-in-Aid for Scientific Research from MEXT Japan (26280007 and 17H06028) and by CREST.

Author Contributions

S.S. and L.R. conceived the project, K.F. performed the analysis, L.R. and S.S wrote the main text, K.F. prepared figures. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-15945-x.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.