BRILL

ASIAN JOURNAL OF
SOCIAL SCIENCE 43 (2015) 613–633

Asian Journal
of Social Science
brill.com/ajss

Analysing "Long Data" on Collective Violence in Indonesia*

David A. Meyer
University of California, San Diego

Arthur Stein
University of California, Los Angeles

Abstract

"Long data", i.e., temporal data disaggregated to short time intervals to form a long time series, is a particularly interesting type of "big data". Financial data are often available in this form (e.g., many years of daily stock prices), but until recently long data for other social, and even other economic, processes have been rare. Over the last decade, however, long data have begun to be extracted from (digitized) text, and then used to assess or formulate micro-level and macro-level theories. The UN Support Facility for Indonesian Recovery (UNSFIR) collected a long data set of incidents of collective violence in 14 Indonesian provinces during the 14 year period 1990–2003. In this paper we exploit the "length" of the UNSFIR data by applying several time series analysis methods. These reveal some previously unobserved features of collective violence in Indonesia—including periodic components and long time correlations—with important social/political interpretations and consequences for explanatory model building.

* This work was partially supported by Office of Naval Research grant N00014-10-1-0138, by National Science Foundation grants 1120888 and 1223137, and by Minerva Research Initiative/ARO grant W911NF-12-1-0389. We thank Ashutosh Varshney, and Patrick Barron, Sana Jaffrey, and Blair Palmer at the World Bank for sharing these data, and Risa Toha for working with them to obtain the data. And we thank the organizers and participants in the 2014 Association of Asian Studies Annual Conference Big Data panel organizers and participants: Gerry van Klinken, Jacky Hicks, Allen Carlson, Benjamin Nyblade, Tom Pepinsky, Molly Roberts, Ross Tapsell and Vincent Traag for useful comments.

 | DOI: 10.1163/15685314-04305006

Keywords

Indonesia – violence – seasonal components – autocorrelation – long-range dependence

Introduction

The increasing collection of data by semi-automated and automated tools is producing large data sets in a variety of domains. These include disciplines that use traditional sources of quantitative data such as astronomy (e.g., Sloan Digital Sky Survey), high energy physics (e.g., *Organisation Européenne pour la Recherche Nucléaire*—CERN), genomics (e.g., European Bioinformatics Institute), and economics (e.g., UN System of National Accounts), to give just a few examples. But ubiquitous use of modern information and communication technologies now produces many kinds of digital data about people and their behaviors: online news (e.g., Baidu News, Twitter), web searches (e.g., Yahoo!, Google), commercial web transactions (e.g., Amazon, PayPal), web-mediated social networks (e.g., Facebook, Sina Weibo), anonymized mobile phone records (e.g., Télécom France–Orange, Telecom Italia), etc. Economists (Einav and Levin, 2014), and social scientists more generally (Lazer et al., 2009), have begun to recognize that analysis of such data may generate otherwise unavailable insights.

The promise of big data extends beyond their use to assess precise hypotheses about how the world works. Even if big data does not mean the end of theory, as some have suggested (Anderson, 2008; Cukier and Mayer-Schoenberger, 2013), they do provide vastly richer possibilities for elucidating empirical patterns and covariations. As the statistician Edward Tufte (2006) has summarized, "Empirically observed covariation is a necessary but not sufficient condition for causality. [...] Correlation is not causation but it sure is a hint." (p. 4). Our goal in this paper is to illustrate some of the hints that big data can provide.

There are, of course, formidable challenges involved in dealing with increasingly large datasets, ranging from storage and transmission to efficient computation. Our interest here, however, is in modes of *analysis* of large sets of data: we believe that novel insights into the underlying human phenomena will depend upon imaginative use of these new kinds of (quantitative) data, together with the development of methodologies designed to analyse them. Some previous examples of such innovative analysis follow, with no pretence of completeness.

One of the first, and best known, is Google's use of a large, digital dataset, previously non-existent—namely records of hundreds of billions of search queries over five years—to measure and predict the incidence of influenza in spatially localized populations (Ginsberg et al., 2009). This analysis was initially very successful compared with the US Center for Disease Control's observations, but then deteriorated, bringing to light a set of novel problems with this kind of data (Lazer et al., 2014). One of the most interesting of these is algorithm drift, namely changes over time in the digital processes generating the data.

Telecommunications are another source of immense quantities of data, which can be used to address perhaps more traditional social science questions. Blondel and Lambiotte and collaborators, for example, have studied a dataset consisting of 810 million mobile phone calls among 2.5 million people in Belgium over 6 months (Lambiotte et al., 2008). By clustering the resulting network of people connected by calls they detect language communities in Belgium (Blondel et al., 2008). Similar results have been obtained in Côte d'Ivoire's substantially more complicated linguistic environment (Bucicovschi et al., 2013).

Analysis of information and communication technologies can now be undertaken to address political questions even in less open political systems. King et al. used natural language processing on 11,382,221 posts on Chinese social media sites to classify their content and search for patterns of censorship. Their results contradicted conventional wisdom that the purpose of Chinese censorship is to silence criticism of the government, and showed instead that the primary purpose is to disrupt collective actions (King et al., 2013).

Large (although not, of course as large as in these three examples) sets of data can, and have been, accumulated by more traditional methods and analysed using more familiar methodologies. Jia's recent study of peasant revolts in 267 Chinese prefectures across more than four centuries, compiled by the Editing Committee of China's Military History (1985), comprises over a hundred thousand prefecture-years of data on revolts, weather and agricultural prices. She uses regression methods, not dissimilar to machine learning techniques typically used on big data, to demonstrate a positive effect of drought on incidence of peasant revolts, mitigated by the introduction of sweet potatoes (Jia, 2014).

This last example shares important features with the Indonesian dataset we study in this paper: It includes conflict events with dates (and geographic locations), collected manually, although in principle they could have been collected by at least semi-automated processing of digital texts. Such "long" time series data support certain inferences even without completely specifying the

dynamics. First, they allow assessment of historical evolution, determining whether some historical experience, in our case violence in Indonesia, has undergone a structural change. In general, the analytic methods we develop in this paper are required to demarcate historical periods by temporal discontinuities (i.e., historiography) and also to assess the consequences of policy interventions (i.e., policy analysis). Second, sufficiently long time series data can reveal periodicities as well as exhibit discontinuities, *even in the presence of multiple confounding factors*. Both discontinuity and periodicity are essential elements for understanding the nature and dynamics of historical change.

In the next section we describe the data we analyse in detail. The following sections describe a series of statistical analyses, both relatively standard and less so, and their implications for understanding political violence in Indonesia. We conclude by summarising our results and pointing in directions for further work.

Data on Incidents of Group Violence

Funded by the United Nations Development Programme, a team of 14 people at the United Nations Facility for Indonesian Recovery (UNSFIR) collected data on incidents of group violence in Indonesia that occurred between 1 January 1990 and 31 December 2003 (Varshney et al., 2008). The data were collected manually from provincial level newspapers. Due to the danger of local data collection in parts of Indonesia with active secessionist conflicts, notably Aceh and Papua, the resulting database[1] is not comprehensive, but does cover the 14 provinces of Riau, DKI Jakarta, Central Java, West Java, East Java, Banten, Central Kalimantan, West Kalimantan, South Sulawesi, Central Sulawesi, East Nusatenggara, West Nusatenggara, Maluku, and North Maluku, which had been the locations of the vast majority of fatalities recorded in an earlier data collection effort (Tadjoeddin, 2002).

Using provincial level newspapers as sources for reports of violent events allowed the UNSFIR team to compile a substantially more complete database than had been produced by the earlier effort, which used the national news sources *Kompas* and *Antara* (Tadjoeddin, 2002). There are multiple possible explanations for this, including the lack of national significance of small-scale violent events, and the New Order SARA policy of not reporting on ethnic

1 The database was accessible online at http://www.unsfir.or.id, but that site seems to be defunct now.

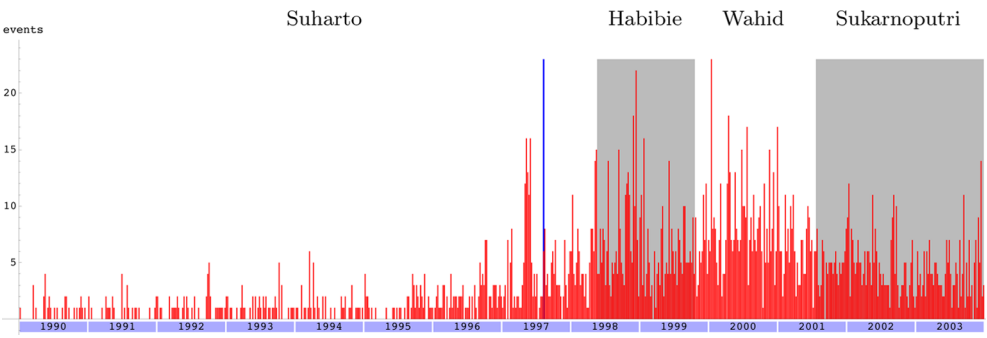

FIGURE 1 *UNSFIR violent event counts for Java*

(*Suku*), religious (*Agama*), racial (*Ras*), and intergroup (*Antar-golongan*) differences, which had less influence on provincial level newspapers (Varshney et al., 2008). Nevertheless, there can be no expectation of completeness for this database. At best the events recorded constitute a representative sample of all events, from which informative statistics can be computed. Very possibly, however, despite the provincial level sources, there are systematic biases, including differences between the New Order and post-New Order periods; in fact Varshney et al. (2008) argue that the level of violence during the New Order estimated from their data may be low due to underreporting.

The database records the day and location of each event, together with the number of fatalities, the weapons used, and a classification of the event as "ethnocommunal", "state versus community", "economic", and "other".[2] There are 3,608 events, which resulted in 10,758 fatalities. The majority of the Indonesian population resides on the island of Java, where the first six provinces listed above are located (the remaining Javan province, the special region of Yogyakarta, was not surveyed as it had been found to be extremely peaceful in the earlier study); 2,323 of the events in the database occurred here. Figure 1 shows the *long*, 5,112 days, time series of these events, aggregated by week, with successive presidential administrations labeled and indicated by alternating white and grey backgrounds. The blue line indicates the date, 14 August, during the Asian currency crisis of 1997, on which the Rupiah was allowed to float freely (Williamson, 2004; Djiwandono, 2005).

The New Order, President Suharto's regime, is often considered to have been a time of relative peace. Bertrand (2004), for example, argues that "The third

2 Other classifications are possible, of course. Van Klinken (2007), for example, distinguishes between large-scale communal violence and localized communal riots.

juncture began with the resignation of Suharto in May 1998. This period of instability exacerbated tensions that had accumulated during the previous thirty years and led to violent outbreaks of several ethnic conflicts." (p. 5). Using comprehensively collected data to address such theories about differences in violence before and after the fall of the New Order was a primary motivation for Varshney et al.'s study (2008). Subsequently these data have been combined with other types of data on economic crises, income, education and demography to build explanatory models for violence in Indonesia (Tadjoeddin and Murshed, 2007; Tadjoeddin et al., 2012; Tadjoeddin 2013).

Our focus is on detailed analysis of the long time series of violent event counts. As we explain in the next two sections, the length of the time series allows us to observe long-time correlations; their presence, rather than correlations that decay over short times, not only suggests novel features to include in explanatory models, but also changes how we assess the significance of differences in mean levels of violence, even without constructing a dynamical model—an important point that we introduce in the next section, and return to in our concluding discussion. As will become evident when we compute the autocorrelation function for the time series, it is long enough that we can decompose it into meaningful periodic components. The significant presence of specific periodicities should also inform future explanatory models.

Mean Levels of Violence

Looking at Figure 1, the level of violence certainly seems higher after the end of the New Order. Let $\hat{\mu}_1$ be the average number of violent events during the New Order (the first 3,063 days of the time series), $\hat{\mu}_2$ be the average number of violent events after the New Order (the remaining 2,049 days), and $\hat{\sigma}_i^2$ be the corresponding variances. In the following table we show these means and variances for various aggregation scales.[3] We also include Welch's t-statistic for assessing the significance of the difference $\hat{\mu}_2 - \hat{\mu}_1$ in mean values,[4] and its number of degrees of freedom ν. In the last column, p is the probability that t would be at least as big as it is, *if the true mean values were the same*.

3 Since the lengths of the two segments of the time series are not evenly divisible by each of the aggregation scales, blocks are measured from the time dividing the two time series, and the resulting two blocks of short length at the beginning and the end of the full time series are weighted proportionally to their lengths.

4 See Appendix A.1 for the relevant statistical formulas.

TABLE 1 *Dependence of violent events statistics on aggregation scale*

days	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	$\hat{\mu}_2$	$\hat{\sigma}_2^2$	t	ν	p
1	0.224	0.379	0.855	1.20	23.69	2917	10^{-113}
7	1.57	5.71	5.98	13.3	18.24	458	10^{-56}
30	6.74	70.1	25.5	97.8	12.96	129	10^{-25}
365	78.1	5744	306	6103	5.61	11	10^{-4}

There are several observations to be made about these calculations. First, the probabilities are extraordinarily small—the first line, for example, tells us that the probability of observing $\hat{\mu}_2$ to be as much bigger than $\hat{\mu}_1$ as it is would be only $p \approx 10^{-113}$ if the true daily mean levels were actually the same. It is hard to imagine how small this probability really is; for all practical purposes it is 0. Thus we would be inclined to conclude that there is *certainly* an increase in the level of violence after the New Order, with the *caveats* implied by the discussion of the data in the previous section.

The second observation to make, however, is that the probability of observing a t value at least as large as we do increases *exponentially* as the aggregation scale increases. While it is still very small at an aggregation scale of 365 days—a probability $p \approx 10^{-4}$ supports as definite a conclusion as we could hope for in empirical work like this—the fact that it has increased so much reveals a problem with any simple assessment of the significance of the difference in the mean level of violence across the two periods. We would, of course, expect the probability to increase because the t-statistic is sensitive to the number of observations and with the fewer observations available when the aggregation scale is larger, a bigger t-score is required to reject the null hypothesis of equal means. But the t-score declines with the variance (see Appendix A.1) and the variances are growing super-linearly (more on this below); thus, the p values are growing more than they would simply as a result of aggregation.

Third, the variances, as just noted, increase disproportionally with the aggregation scale. As the scale increases we expect changes in the average number of events, and the variance in the number of events. As it must, the average number of events increases proportionally to the aggregation scale; for example, $1.57 \approx 7 \times 0.224$. On the other hand, the variance does not increase proportionally to the aggregation scale, as it would *if the events were independent*; instead, we see, for example, $5.71 \gg 2.65 \approx 7 \times 0.379$ when we compare the second with the first row in the table. The disproportionate growth in variance

is an indication that the events are *not* independent and also the reason that the p values grow exponentially.

Fourth, at each aggregation scale $\hat{\sigma}_i^2 > \hat{\mu}_i$, i.e., the data are *over-dispersed*, or clustered, or "bursty", which is visible in Figure 1. If violent events happened independently of previous events their distribution would be uniform, i.e., Poisson, with variance equal to the mean number of events, at any scale. That this is not the case with these data indicates that the events are (positively) correlated, and do not become approximately independent *even at large aggregation scales*. This implies that the most basic of the assumptions underlying Welch's t-test, that the numbers of events per day, X_t, are independent random variables, is violated. Thus the probabilities in the table should not be trusted—they are much too small, i.e., *the difference in levels of violence is less significant than it appears*. To determine how to correct this we must investigate the correlation between X_t and $X_{t'}$; we do so in the next section.

The Autocorrelation Function

Assuming a time series $\{X_t \mid t \in \mathbb{Z}\}$ is *second-order stationary*, so that $\mathsf{E}[X_t] = \mathsf{E}[X_{t+s}]$ and $\mathsf{Cov}[X_t, X_{t'}] = \mathsf{Cov}[X_{t+s}, X_{t'+s}]$, for all $t, t', s \in \mathbb{Z}$, then the *autocorrelation function* $\rho(s)$ is well defined, and the *sample autocorrelation function* $\hat{\rho}(s)$ can be computed, by the formulas in Appendix A.2.

Although we have seen that the condition of constant $\mathsf{E}[X_t]$ seems to fail for the whole UNSFIR time series, being lower during the New Order than afterwards, it does appear to be (approximately) true for each of these two time intervals separately. Figure 2 shows the sample autocorrelation function for each subtimeseries, assuming second-order stationarity.

Recall that for a time series of independent events, the autocorrelation $\rho(s)$ is 1 when the lag $s = 0$, and vanishes for all lags $s \neq 0$. Notice that both of the sample autocorrelation functions shown in Figure 2 decay very slowly, staying positive out to lags of almost three years for the New Order time series, and out to lags of almost one year even for the much noisier post-New Order time series. This strongly suggests *long-range dependencies* in these time series, i.e., that the autocorrelation function decays like a power law, $\rho(s) \sim 1/s^\alpha$, $0 < \alpha \leq 1$, as $s \to \infty$, rather than decaying faster, e.g., exponentially, $\rho(s) \sim e^{-s/\sigma}$, $\sigma > 0$, as $s \to \infty$. This would explain both the third and fourth observations we made at the end of the previous section, namely that the variance scales super-linearly with aggregation scale, and that the data are over-dispersed.

Figure 3 shows the results of least squares fits of the exponential function, $\rho(s) = Ae^{-s/\sigma}$, in blue, and the power law function, $\rho(s) = A/s^{-\alpha}$, in green, to

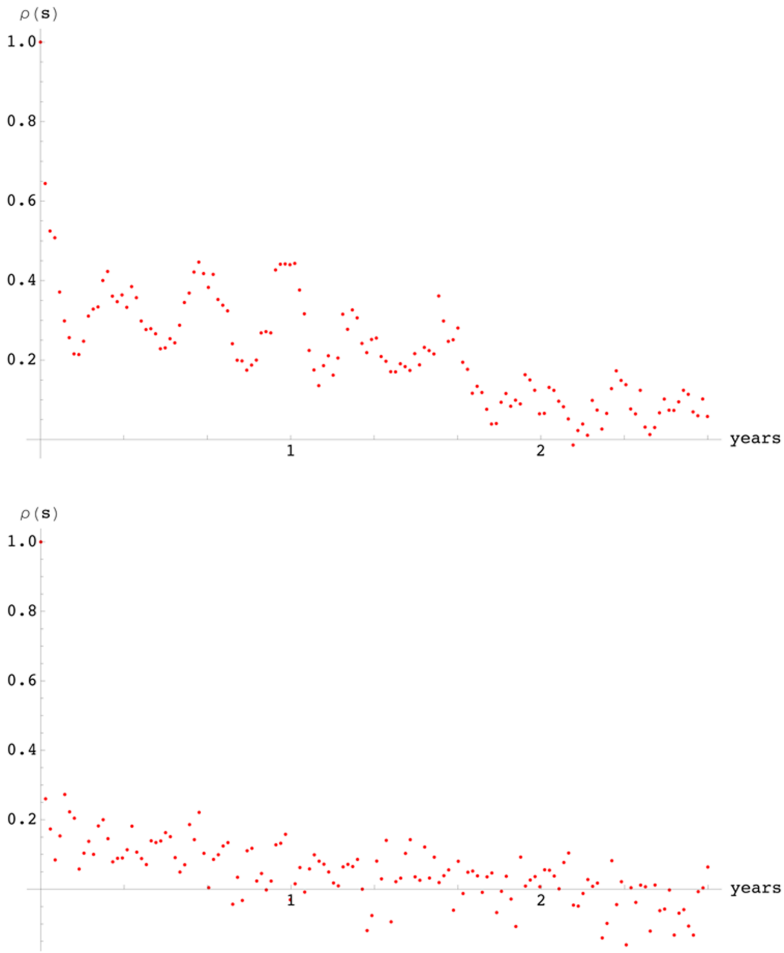

FIGURE 2 *The sample autocorrelation functions computed separately for the New Order (on the top) and post-New Order (below) sub-time series*

the empirical autocorrelation data for the New Order time series.[5] The decaying power law function fits the data much better: not only does it not rise above most of the data for lags s less than a year, as the exponential curve does, but it also goes to 0 slowly enough for large s that the relatively large values of $\rho(s)$ there are not impossibly unlikely. The power law exponent $\alpha \approx 0.35$ for this

5 These are computed for $\log \rho$, so for the latter we omit the data point at $s = 0$, and for both we fit only the data points out to the value of s after which the first negative correlation $\hat{\rho}(s)$ occurs, since the logarithms of non-positive values are not real numbers.

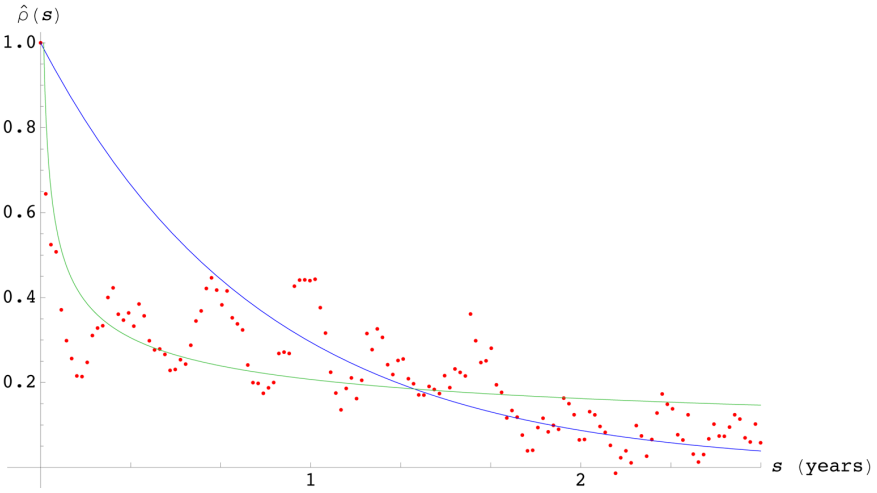

FIGURE 3 *Decaying exponential (blue) and power law (green) fits to the sample autocorrelation function for the New Order time series*

fit, supporting the existence of long-range dependencies in the time series of violent events and our observation that these events tend to be clustered in time, i.e., that the time series is "bursty".

That the numbers of violent events on different days are not independent is consistent with the qualitative observation that violent political events recur. Sometimes such violence is a singular explosion, an event that occurs once, but in most cases, whether they are rebellions, insurgencies, ethnic conflicts, civil wars, or international wars, violent conflicts are sustained and recur.[6] Even when there seems to be resolution, there are often renewed outbreaks. A substantial portion of countries experience repeated civil wars, for example; so much so, that one article's title asks the question, "Does conflict beget conflict?" (Walter, 2004). War, too, evinces a pattern of recurrence, in general (Moyal, 1949; Denton and Phllips, 1968), and in the relations of particular states that experience "enduring rivalries" (Diehl, 1998).

Our simplest understanding should also make us sensitive to the possibility of such dependencies in political phenomena. Studies of political violence, for example, talk of cycles of violence and retribution.[7] A common pattern is for

6 Personal violence, such as that between clans, is also often sustained (Diamond, 2008).

7 Revenge is listed as more important than material reasons for primitive war (Gat, 2006: 33). A more modern example is provided by Kaplan (2005: *xlvi*), who relates the story of the funeral service in 1940 for the reburial of the remains of 14 men brutally killed by King Carol's police in

a country to experience periods of extended political stability "separated by recurrent waves of internal war" (Turchin, 2012: 2).[8] Ethnic conflicts are long-standing, reflecting memories of even the distant past; violent events are timed to important anniversaries (Cairns and Roe, 2002).

Recurrences over such long times are plausibly explained by the fact that people are encoded with long-range memories. They are socialized so as to internalize the centrality of events they did not directly experience. Distant events are commemorated in public rituals. Scholars speak of "dedicated memory forms, which are purposefully created (and destroyed) by social groups in order to socially influence the process of collective remembering and forgetting", and include books, holidays, monuments, songs, poetry, artifacts, museums, symbols, and art (Devine-Wright, 2002: 12). Our empirical observation of long-range correlations, as shown in Figures 2 and 3, suggests that these qualitative observations of human behavior should be included into any dynamical model for violent events.

Seasonal Oscillations of Violence in Indonesia

Now let us address another prominent feature of Figure 2, namely the apparent 1/3 of a year periodicity in the data. This is unmistakable in the sample autocorrelation function of the New Order time series, and there seem to be hints of it even in the sample autocorrelation function of the post-New Order time series.

Periodic components in time series can be identified using the Fourier transform (Appendix A.3). A periodicity in the sample autocorrelation function $\hat{\rho}$ corresponds to a periodicity in the time series itself. Figure 4 shows the *periodogram* (Schuster, 1898), i.e., a plot of power as a function of frequency for the New Order sample autocorrelation function in Figure 2, and also the (smoothed[9]) log periodogram. The peak at $\nu_k \approx 0.058$ ($k \approx 25.5$) corresponds to an approximately 120 day period, or about 1/3 of a year.

As a first step towards identifying the cause of this periodicity, we can partition the events by category. Violent events classified as ethno-communal (which includes religious) or economic constitute a relatively small fraction,

1938, at which worshipers heard a voice recording of the dead Romanian Legionnaire leader, Codreanu, telling them, "You must await the day to avenge our martyrs".

8 Turchin argues that "episodes of internal warfare often develop in ways similar to epidemics or forest fires" (see also Turchin, 2006b: Ch. 9).

9 Using a modified Daniell (1946) filter of radius 2.

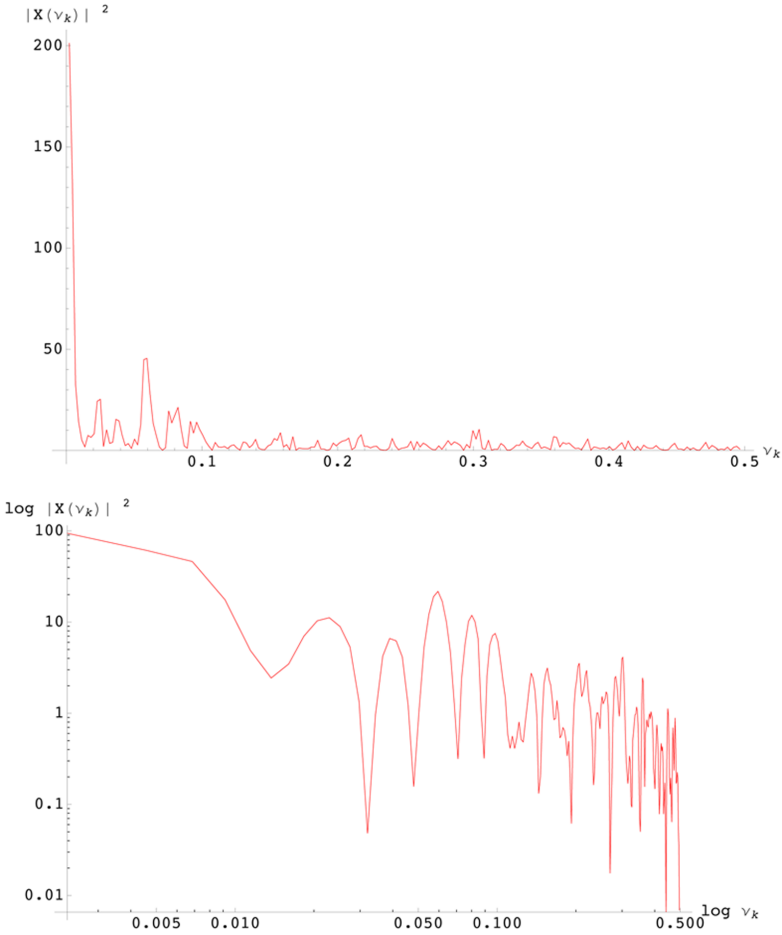

FIGURE 4 *The power spectrum for the New Order time series, log-log plotted in the lower graph*

17.9%, of the events on Java, and they *do not display this* 1/3 *year periodicity*. Events in the other two categories, state versus community and other—which includes "*dukun santet* (killings of persons who allegedly practice *santet*/black magic)", intergroup/intervillage brawls and "popular justice" (Varshney et al., 2008)—constitute the majority of the time series, and they do display the 1/3 year periodicity. In fact, recomputing the period using only these data gives approximately 121 days, slightly closer to a third of a year.

Computing the correlation function between shifted cosine functions with this period and the New Order time series, we find that the maxima occur at approximately 23 April, 22 August and 22 December (with minima equally spaced between these dates, of course). Although the first of these is close to

the dates of Easter, while the last is close to Christmas, as noted in the previous paragraph, events classified as arising from religious conflicts on Java *did not* show the 1/3 of year periodicity we would like to explain.

A more plausible explanation is seasonal: 22 December occurs near the peak of the rainy season and 22 August is in the middle of the dry season (Yasunari, 1981). More directly to the point, there are three growing seasons per year for irrigated *sawah* (rice paddy)[10] on Java (Heytens, 1991), and these three dates lie approximately at the end of the harvest of one season and the beginning of the planting for the next (Falcon et al., 2004).

Furthermore, "the limited job opportunity between planting and harvesting rice" (Hugo, 1982) leads to substantial circular migration in Indonesia. Migrant agricultural labour, either to or from the fields, creates possibilities for violence: Arson in fields in East Java was attributed to Madurese migrant workers' conflicts with owners and foremen as early as 1891 (Hussun, 1997). And even harvesting opportunities have become increasingly limited due to agricultural practices associated with the Green Revolution, i.e., the replacement of the *ani-ani* (a small knife traditionally used to harvest rice stalk by stalk) and the *bawon* system in which anyone can join a harvest and receive a share of the proceeds, with the *arit* (sickle) and the *tebasan* system in which farmers sell their rice before harvesting to *penebas* (traders) who supervise the harvest, *restricting the number of labourers*, or even by an *upah* (wage) system (Kanazawa, 1993). These changes have, on occasion, led to violent conflict; Collier et al. comment in a footnote that "a *penebas* was severely beaten by women harvesters because they could not join his harvest" (1973).

Thus we suspect that there is a substantial seasonal component to "routine forms of group violence" (Varshney et al., 2008), very possibly involving migrant agricultural workers.

Discussion

The statistical analyses above required a long dataset: 14 years of daily numbers of violent events in Indonesia. Without such long data, we would have been restricted in the scales over which we could aggregate to notice the over-

10 With decreasing control over the irrigation, one or two of the growing seasons are devoted to *palawija* (other crops, typically soybeans, corn, or vegetables), rather than to rice. With sufficiently poor irrigation, or none, there are only two growing seasons, but this is estimated to describe only about 20 % of the cultivated land on Java (Heytens, 1991).

dispersion of the data, and in the length of the lags over which we could compute the sample autocorrelation function, from which we (crudely) estimated the parameter α of the long-range dependence.

Since the time series does have such long-range dependence, the observed over-dispersion follows from the fact that the variance of the mean value $X^{(n)} = (X_1 + \cdots + X_n)/n$ satisfies $\lim_{n\to\infty} n^{\alpha}\mathsf{Var}[\bar{X}^{(n)}] = \sigma^2 c_\rho$, for some $c_\rho > 0$ (see, e.g., Beran, 1994). This explains the growth of the variances in Table 1 as $n^{2-\alpha}$, rather than linearly with the aggregation scale n, as they would if events were independent.

As we noted above, the empirical fact of long-range dependence in political violence time series is not unexpected, although rarely, if ever, stated explicitly or demonstrated quantitatively. Our observation of long-range dependencies in these Indonesian violence data, therefore, supports and extends qualitative analyses of such phenomena. Quantifying long-range dependence with estimates of the parameters α and c_ρ is necessary in order correctly to assess changes in the mean value of a time series which, as we described at the beginning, was one motivation for Varshney et al.'s (2008) collection of these data. The distribution of the normalized mean value $X^{(n)}$ of a time series with long-range dependence,

$$Z = \frac{\bar{X}^{(n)} - \mu}{\sigma_\mu} n^{\alpha/2},$$

where

$$\sigma_\mu^2 = \frac{2\sigma^2 c_\rho}{(2-\alpha)(1-\alpha)},$$

converges to the standard normal distribution as $n \to \infty$ (Beran, 1991). If we let $\bar{X}^{(-n)} = (X_{-(n-1)} + \cdots + X_{-1} + X_0)/n$, then $\mathsf{E}[\bar{X}^{(m)}|\bar{X}^{(-n)}] = \bar{X}^{(-n)}$. And then similarly, the distribution of

$$Z_{n,m} = \frac{\bar{X}^{(m)} - \bar{X}^{(-n)}}{\sigma_{\mu_1,\mu_2}(q)} n^{\alpha/2},$$

where $q = m/n$ and

$$\sigma_{\mu_1,\mu_2}(q)^2 = \sigma_\mu^2\left(1 + q^{-\alpha} - \frac{1}{q}\left((q+1)^{2-\alpha} - (q^{2-\alpha}+1)\right)\right),$$

converges to the standard normal distribution as $n \to \infty$, so we can use it to assess the significance of a deviation of $\bar{x}^{(m)}$ from $\bar{x}^{(-n)}$, *which will be smaller for smaller values of* α (Beran, 1989).

Applying this last formula requires good estimates for α and c_ρ. While the least squares fits illustrated in Figure 3, and the empirical correlation function itself, clearly indicate that the autocorrelation has a long tail, such least squares fits to apparently power law data suffer from well-known problems (Clauset et al., 2009), and furthermore cannot be expected to give good parameter estimates because the values being fit in this case are not independent. A variety of estimation methods have been devised, several of which are asymptotically (i.e., as the length of the time series becomes infinite) unbiased and efficient (Beran, 1994). For finite, even long, data these estimates tend to be biased, so our results here call for good estimation methods for realistic data. Derivation of such methods is a topic of current research (Malyarenko et al., 2015). Even without precise estimates for α and c_ρ for the UNSFIR Java time series, the empirical fact of long-range dependence adds a new reason (to the several adduced by Varshney et al.) to discount "the pessimism about the future of the country under a democratic dispensation" of some analysts, due to "the violence of post-Suharto years" (2008: 362), since it reduces the significance of the observed difference in mean levels of violence in the data UNSFIR collected.

A long time series was also required in order for the Fourier transform to reveal the periodic component we found and analysed here. This observation illustrates the power of analysing big data, even by relatively simple methods. To our knowledge, no-one has observed or proposed the existence of such a seasonal oscillation in group violence on Java previously. We provided a plausible explanation for it in terms of violence associated with migrant agricultural labour, but this calls out for further analysis, combining additional quantitative analysis of the UNSFIR data (e.g., including the spatial features of the data) with new qualitative investigation of this possibility.

Our analysis of "long data" on collective violence in Indonesia has not only revealed novel aspects of, and raised novel questions about, this specific milieu, but it also illustrates how big data can be used to enrich more traditional kinds of information studied in the social sciences. While mere correlation, or even successful prediction, without understanding is not science, we believe that analysis of big data, *combined with qualitative and quantitative modeling*, will increasingly provide insight into the workings of the social world.

References

Anderson, Chris (2008) "The end of theory: The data deluge makes the scientific method obsolete". *Wired Magazine* 16(7), 23 June.

Beran, Jan (1989) "A test of location for data with slowly decaying serial correlations". *Biometrika* 76(2):261–269.

——— (1991) "*M*-estimators of location for Gaussian and related processes with slowly decaying serial correlations". *Journal of the American Statistical Association* 86:704–708.

——— (1994) *Statistics for Long-Memory Processes*. Boca Raton, FL: Chapman & Hall/CRC.

Bertrand, Jacques (2004) *Nationalism and Ethnic Conflict in Indonesia*. Cambridge: Cambridge University Press.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte and Etienne Lefebvre (2008) "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment* (October):P10008/1–12.

Bucicovchi, Orest, Rex Douglass, David A. Meyer, Megha Ram, David Rideout and Dongjin Song (2013) "Analyzing social divisions using cell phone data". *NetMob 2013*, MIT, Cambridge, MA, 1–3 May 2013, UCSD preprint.

Cairns, Ed and Micheál D. Roe (eds.) (2002) *The Role of Memory in Ethnic Conflict*. New York: Palgrave Macmillan.

Clauset, Aaron, Cosma Rohilla Shalizi and M.E.J. Newman (2009) "Power-law distributions in empirical data". *SIAM Review* 51(4):661–703.

Collier, William L., Gunawan Wiradi and Soentoro (1973) "Recent changes in rice harvesting methods. Some serious social implications". *Bulletin of Indonesian Economic Studies* 9(2):36–45.

Cukier, Kenneth Neil and Viktor Mayer-Schoenberger (2013) "The rise of Big Data: How it's changing the way we think about the world". *Foreign Affairs*, May/June.

Daniell, Percy John (1946) Discussion of "On the theoretical specification and sampling properties of autocorrelated time-series". *Supplement to the Journal of the Royal Statistical Society* 8:88–90.

Denton, Frank H., and Warren Phillips (1968) "Some patterns in the history of violence". *Journal of Conflict Resolution* 12(2): 182–195.

Devine-Wright, Patrick (2002) "A theoretical overview of memory and conflict", in Ed Cairns and Micheál D. Roe (eds.) *The Role of Memory in Ethnic Conflict*, New York: Palgrave Macmillan, pp. 9–33.

Diamond, Jared (2008) "Annals of anthropology: Vengeance is ours". *The New Yorker*, 21 April.

Diehl, Paul F. (ed.) (1998) *The Dynamics of Enduring Rivalries*. Urbana: University of Illinois Press.

Djiwandono, J. Soedradjad (2005) *Bank Indonesia and the Crisis: An Insider's View*. Singapore: Institute of Southeast Asian Studies.

Editing Committee of China's Military History (1985) *Zhongguo Lidai Zhanzheng Nianbiao* [*The Chronology of Warfare in Dynastic China*]. Beijing: People's Liberation Army Press.

Einav, Liran and Jonathan Levin (2014) "The data revolution and economic analysis", in Josh Lerner and Scott Stern (eds.) *Innovation Policy and the Economy*, Volume 14. Chicago: University of Chicago Press, pp. 1–24.

Falcon, Walter P., Rosamond L. Naylor, Whitney L. Smith, Marshall B. Burke and Ellen B. McCullough (2004) "Using climate models to improve Indonesian food security". *Bulletin of Indonesian Economic Studies* 40(3):355–377.

Gat, Azar (2008) *War in Human Civilization*. Oxford: Oxford University Press.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant (2008) "Detecting influenza epidemics using search engine query data". *Nature* 457(7232):1012–1014.

Heytens, Paul (1991) "Rice production systems", in Scott R. Pearson, Walter Falcon, Paul Heytens, Eric A. Monke and Rosamond Naylor (eds.) *Rice Policy in Indonesia*. Ithaca, NY: Cornell University Press, pp. 38–57.

Hugo, Graeme J. (1982) "Circular migration in Indonesia". *Population and Development Review* 8:59–83.

Husson, Laurence (1997) "Eight centuries of Madurese migration to East Java". *Asian and Pacific Migration Journal* 6(1):77–102.

Jia, Ruixue (2014) "Weather shocks, sweet potatoes and peasant revolts in historical China". *The Economic Journal* 124(575):92–118.

Kanazawa, Natsuki (1993) *Southeast Asian Rice Farming and Farmers in Transition*. Regional Research Institute of Agriculture in the Pacific Rim, College of Agriculture and Veterinary Medicine. Tokyo: Nihon University.

Kaplan, Robert D. (2005) *Balkan Ghosts: A Journey Through History*. New York: Picador.

King, Gary, Jennifer Pan and Margaret E. Roberts (2013) "How censorship in China allows government criticism but silences collective expression". *American Political Science Review* 107(2):326–343.

van Klinken, Gerry (2007) *Communal Violence and Democratization in Indonesia: Small Town Wars*. New York: Routledge.

Lambiotte, Renaud, Vincent D. Blondel, Cristobald de Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda and Paul Van Dooren (2008) "Geographical dispersal of mobile communication networks". *Physica A: Statistical Mechanics and its Applications* 387(21):5317–5325.

Lazer, David, Ryan Kennedy, Gary King and Alessandro Vespignani (2014) "The parable of Google Flu: Traps in big data analysis". *Science* 343:1203–1205.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy and Marshall Van Alstyne (2009) "Computational social science". *Science* 323(5915):721–723.

Malyarenko, Anna, David A. Meyer and Nan Zou (2015) "Semiparametric estimators for finite length long-range dependent data". In preparation.

Moyal, J.E. (1949) "The distribution of wars in time". *Journal of the Royal Statistical Society. Series A* (*General*) 112(4):446–449.

Pearson, Scott R., Walter Falcon, Paul Heytens, Eric A. Monke and Rosamond Naylor (eds.) (1991) *Rice Policy in Indonesia*. Ithaca, NY: Cornell University Press.

Schuster, Arthur (1898) "On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena". *Terrestrial Magnetism* 3(1):13–41.

Tadjoeddin, Mohammad Zulfan (2002) "Anatomy of social violence in the context of transition: The case of Indonesia 1990–2001". *Politics, Administration, and Change* 38:1–35.

——— (2013) "Educated but poor: Explaining localized ethnic violence during Indonesia's democratic transition". *International Area Studies Review* 16(1):24–49.

Tadjoeddin, Mohammad Zulfan, Anis Chowdhury and Syed Mansoob Murshed (2012) "Routine violence in Java, Indonesia: Neo-Malthusian and social justice perspectives", in Jürgen Scheffran, Michael Brzoska, Hans Günter Brauch, Peter Michael Link and Janpeter Schilling (eds.) *Climate Change, Human Security and Violent Conflict: Challenges for Societal Stability*. Hexagon Series on Human and Environmental Security and Peace, Volume 8. New York: Springer Verlag, pp. 633–650.

Tadjoeddin, Mohammad Zulfan and Syed Mansoob Murshed (2007) "Socio-economic determinants of everyday violence in Indonesia: An empirical investigation of Javanese districts, 1994–2003". *Journal of Peace Research* 44(6):689–709.

Tufte, Edward R. (2006) *The Cognitive Style of Power Point: Pitching Out Corrupts Within*. 2nd ed. Cheshire, Connecticut: Graphics Press.

Turchin, Peter (2006) *War and Peace and War: The Life Cycles of Imperial Nations*. New York: Pi Press.

——— (2012) "Dynamics of political instability in the United States, 1780–2010". *Journal of Peace Research* 49(4):577–591.

Varshney, Ashutosh, Mohammad Zulfan Tadjoeddin and Rizal Panggabean (2008) "Creating datasets in information-poor environments: Patterns of collective violence in Indonesia, 1990–2003". *Journal of East Asian Studies* 8(3):361–394.

Walter, Barbara F. (2004) "Does conflict beget conflict? Explaining recurring civil war". *Journal of Peace Research* 41(3):371–388.

Welch, Bernard Lewis (1947) "The generalization of 'Student's' problem when several different population variances are involved". *Biometrika* 34(1/2):28–35.

Williamson, John (2004) "The years of emerging market crises: A review of Feldstein". *Journal of Economic Literature* 42(3):822–837.

Yasunari, Tetsuzo (1981) "Temporal and spatial variations of monthly rainfall in Java, Indonesia". *Southeast Asian Studies* 19(2):170–186.

Appendices

A.1. *Welch's t-Statistic*

Let $N_1 = 3063$ and $N_2 = 2049$ denote the number of days in the time series during and after the New Order, respectively, $\hat{\mu}_i$ the corresponding average number of violent events per day, and $\hat{\sigma}_i^2$ the corresponding variances. Then Welch's t-statistic for assessing the significance of the difference $\hat{\mu}_2 - \hat{\mu}_1$ in mean values is:

$$t = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2}}.$$

Under certain assumptions, this is approximately distributed according to a t-distribution with

$$\nu = \frac{(\hat{\sigma}_1^2/N_1 + \hat{\sigma}_2^2/N_2)^2}{\hat{\sigma}_1^4/N_1^2(N_1 - 1) + \hat{\sigma}_2^4/N_2^2(N_2 - 1)}$$

degrees of freedom (Welch, 1947); thus the probability p of the value of t being at least as large as observed, *if there were no difference in the true level of violence*, can be computed; these numbers are shown in Table 1.

A.2. *The Autocorrelation Function*

Assuming a time series $\{X_t \mid t \in \mathbb{Z}\}$ is *second-order stationary*, so that $\mathsf{E}[X_t] = \mathsf{E}[X_{t+s}]$ and $\mathsf{Cov}[X_t, X_{t'}] = \mathsf{Cov}[X_{t+s}, X_{t'+s}]$, for all $t, t', s \in \mathbb{Z}$, then the *autocorrelation function*,

$$\rho(s) = \mathsf{Corr}[X_t, X_{t+s}] = \frac{\mathsf{Cov}[X_t, X_{t+s}]}{\mathsf{Var}[X_t]},$$

is well-defined, with $\rho(-s) = \rho(s)$; $\rho(s)$ is the correlation between the numbers of events s time units apart. The *sample autocorrelation function* for a second-order stationary empirical time series $\{x_t \mid 1 \le t \le N\}$ is:

$$\hat{\rho}(s) = \frac{\sum_{t=1}^{N-s}(x_t - \bar{x}_s)(x_{t+s} - \bar{x}_{-s})}{\sqrt{\sum_{t=1}^{N-s}(x_t - \bar{x}_s)^2 \sum_{t=1}^{N-s}(x_{t+s} - \bar{x}_{-s})^2}},$$

where $\bar{x}_s = \sum_{t=1}^{N-s} x_t/(N-s)$ and $\bar{x}_{-s} = \sum_{t=1}^{N-s} x_{t+s}/(N-s)$, for $s \ge 0$.

A.3. *The Fourier Transform*

Periodic components in time series can be identified using the Fourier transform. Let the estimated covariance function (the renormalized numerator in the definition of $\hat{\rho}$) be

$$\hat{\gamma}(s) = \frac{1}{N}\sum_{t=1}^{N-s}(x_t - \bar{x}_s)(x_{t+s} - \bar{x}_{-s}).$$

Then the Fourier transform of $\hat{\gamma}$ is the (*power*) *spectral density*

$$\hat{f}(\nu) = \sum_{s=-N+1}^{N-1}\hat{\gamma}(s)e^{-2\pi i\nu s},$$

for $|\nu| \leq 1/2$. This name derives from the fact that at $\nu_k = k/N, \hat{f}(\nu_k)$ is equal to the *power* at frequency ν_k in the discrete Fourier transform of the time series, i.e., to

$$I(\nu_k) = |X(\nu_k)|^2 = \frac{1}{N}\left|\sum_{t=1}^{N} x_t e^{-2\pi i\nu_k t}\right|^2.$$

Thus a periodicity in the sample autocorrelation function $\hat{\rho}$ corresponds to a periodicity in the time series itself.