Measurement Models for Event Data

Will Lowe[*]
MZES, University of Mannheim

Event data, the sequences of dated categorised events between identifiable actors typically harvested from newswire, are a vital source of information about interacting political actors (Quarterly, 1983; Schrodt and Gerner, 2001; Schrodt, 2012). They offer a finer time scale and greater actor and spatial specificity than perhaps any other systematic data source. From a statistical perspective they are also undeniably awkward (see Schrodt, 1994, for useful discussion).

As data, events are inherently nominal – an event of some type occurs or it does not. When extracted from newswire, events are typically observed on at most a daily basis. In dyadic analyses, with which this paper is mostly concerned, event data therefore constitute *multivariate time series count data*. This kind of data is notoriously hard to model directly.

When the variety of possible events is relatively small then some existing count data time series models can be applied directly or with mild extensions (e.g. Brandt and Sandler, 2012). However, for most researchers counts of events are seldom interesting in themselves, so some form of scaling and aggregation is applied to create interpretable quantities of interest, e.g. the level of cooperation in a dyad (Azar, 1980; Goldstein, 1992; Shellman, 2004a) or the conflict carrying capacity of a state (Jenkins and Bond, 2001). But the scaling process as it is currently performed, and the awkward nature of the data lead to immediately to questions: how should events be aggregated? should they be numerically scaled? and if so how? And there are well-known paradoxes: e.g. how can any number of pessimistic comments be equivalent to a bombing according to any event scale? What *is* the relationship between numerical conflict scores and the event data to which it is applied? Can the scaling schemes currently in use be derived or validated from event data itself?

This paper attempts to shed some light on these paradoxes and provide principled answers to the questions by sketching a statistical framework for thinking about event data based on measurement modelling. In the first section of the paper review existing event aggregation problems and paradoxes. The next section shows in two steps how measurement considerations might defuse them and offers state space time series models as a practical realisation of the approach. In the final section of the paper I use measurement models originally designed for legislative text analysis to infer a conflict scale from mildly aggregated event categories and show that it gives nearly identical results to the standard Goldstein (1992) scale that was derived from expert judgements.

Treating event data analysis as a measurement problem is not novel – indeed the approach is familiar from other parts of the political science landscape (Clinton et al., 2004; Pickup, 2009), but previous event research has worked with very specific models (Schrodt, 2006; Bond et al., 2004; Schrodt, 2007, 2011, e.g.), when it has not simply tried to force event data analysis into a regression based time series framework. Rather than present one more model this paper is an attempt to map a wider range of measurement possibilities and highlight the consequences of a measurement perspective for event data analysis in two areas: event aggregation and inductive conflict scale construction.

[*] Will Lowe is Senior Researcher at The Mannheim Centre for European Social Research (MZES). He can be reached at will.lowe@uni-mannheim.de This paper has benefited from comments at the Turkish Event Data Workshop in December 2011.

1 Puzzles about event data

When multiple actors interact intermittently in multiple places in multiple ways some form of aggregation is always necessary. Yonamine (2011) is a comprehensive review of existing actor, action, and temporal aggregation possibilities. Spatial (dis)aggregation issues are well discussed in Shellman (2004b). Here I concentrate on time and action aggregation issues.

1.1 The necessity and hazards of aggregation

To perform a time series analysis it is necessary to choose a suitable time unit over which to model actors' interactions, e.g. using a vector autoregression (VAR, Lütkepohl, 1990). Unfortunately, it is well-known that VAR models give different substantive results when temporal aggregation is altered, e.g. from week to month to quarter (Shellman, 2004b, for examples and discussion). This should not be surprising; lower frequency observations typically mask actor dynamics at higher frequencies. Indeed it is quite rare for time series structure to be maintained at multiple aggregation levels. Consequently, methodological advice is to disaggregate as far as possible so as not to lose this information (e.g. Freeman, 1989).

However, VAR approaches, in common with all methods that define relationships in terms of events and lags, also seem to *require* some aggregation in order to avoid missing data problems. Models for higher frequency observations will require more lags than those formulated at lower frequencies. This can be problematic for event data because it is very likely that there are no observed events in some set of previous time periods. In a VAR framework this means that *all* observations in that time step are dropped from the analysis. Decreasing temporal aggregation therefore leads to increasing missing data problems. What then to do about missing data?

Current event data practise is often to 'live with' the possibility of missing or underestimating the strength of relationships (and perhaps to gesture hopefully at contemporaneous correlations, Goldstein and Pevehouse 1997, e.g.). Alternatively, assert that the variable of interest is *net* conflict or cooperation, replace time periods of no observations with a zero and fit a VAR model regardless. Or both. Replacing empty periods with a zero in non-count data is effectively imputing missing data with a constant so it is hard to see how it can avoid serious bias (Honaker and King, 2010).

1.2 Criticism

Most scaled event data is summed or averaged within a time period before analysis, so we treat criticisms of these two practices in turn. The practice of aggregating scaled event data and then running VAR-type models on it has been criticised on methodological grounds almost since its inception. An early example is the Folk Criticism, named because it is a correct observation but no one can quite remember where they saw it first.

Folk Criticism: the 'sum problem'

The 'Folk Criticism' (FC) notes that in summed scaled event data, multiple less conflictual events may be numerically identical to or greater than a single highly conflictual event. This is a problem when the numbers generated by the aggregated scaling process are intended to represent some continuous measure of an actor relationship at a time. The FC is also an example of what Yonamine (2011) describes as 'the sum problem'

> Consider two months, one with little dialogue but actual violence, and one with no violence but considerable negative dialogue. By summing the Goldstein values, the

> latter month could appear more conflictual than the first, even though it experience no actual conflict.

Or as Schrodt (2007, and elsewhere) more succinctly puts it in a discussion of COPDAB scaling: “three riots equals one thermonuclear war”.

In general, the FC simply points out the inconsistency between the substantive interpretations of numerical conflict-cooperation measures at different levels of analysis and it will apply to *any* procedure that assigns numbers to individual event codes and also sums them.

The mean problem

A natural alternative to summing is to take an average of the event scores. This leads to a related problem, labelled by Yonamine as the ‘mean problem’:

> Since it is obvious that a month with 90 “-10 events” is more conflictual than a month with only 30 “-10” events, taking the mean score can lack external validity.

Taking an average does maintain the event-level substantive interpretation at all levels of analysis, but at the cost of throwing away information. The nature of the information thrown away is considered in more detail below.

It is interesting to note that the ‘mean problem’ occurs even in the *construction* of scores for event types. In Goldstein (1992) experts disagreed about whether surrendering and yielding position (WEIS event codes 011 and 012 respectively) were strongly cooperative or strongly conflictual. This yielded an expert average score that was mildly positive and expert judgement standard deviations that were around six times larger than any other event code. Subsequent researchers have been sanguine about this, despite the relatively high frequency of surrendering and retreating occurring in conflict dyads.

Another aspect of basically the same problem is Yonamine’s ‘single scale problem’:

> “non-injury destructive action” (a -8.3 on the Goldstein scale) and a “extend military assistance” (a +8.3 on the Goldstein scale) occur between the same actors. The sum and the mean of these two events equals 0, which is the same score that a day with no events receive.”

There are actually *two* problems here: first, the scaled events cancel out under summing or averaging. Moreover, this must be a problem for any final value that is itself interpretable on the event scale, not just zero. In the second, there is no way to distinguish between the absence of an event coded as zero, a single actual event of neutral valence, and an arbitrary set of score-balanced events.

Yonamine’s solution – splitting into two scales, one for conflictual events and one for cooperative events – apparently solves this problem and has been argued for by Pevehouse (2004) on theoretical grounds. However, exactly the same issues arise with separately scaled quantities. An absence of cooperative events may still be coded as zero on the new cooperation scale but zero is also interpretable as the *minimum possible* level of cooperation. This cannot be a solution; now the constant that is imputed is complete lack of cooperation, rather than a neutral action on the original single scale.

To review, the Folk Criticism notes that there is no stable interpretation of summed scaled event data that also holds for individual events. The mean problem indicates that averaging instead loses information. Temporal aggregation hides actor dynamics, but disaggregation leads to missing data, which is imputed with zeros. How then should we treat scaled event data to avoid these issues?

2 A measurement approach to scaled event data problems

The first step to solving these problems is to identify what scaled event data is data *about*. Consider a single dyadic time series. In applying, e.g. the Goldstein scale, we are implicitly using event data to learn about the underlying level of conflict and/or cooperation in a dyad. This is never observed directly, not only because we rely on a news agency and then an information extraction system to recover the events, but also because events themselves are not conflict level, they are only its indicators. This is why we wanted a scale.

The Goldstein scale itself is a simple but incomplete measurement model; it takes an event as input and generates a decontextualised conflict score valid for that event type. As a stipulative measurement model the Goldstein scale is non-generative, has no parameters, and was 'fitted' directly to a small number of expert judgements rather than a large amount of conflict data, but it is a measurement model nevertheless. It is incomplete because event data arrive in time units that often contain multiple events. A factor analysis model applied to subject' survey responses will generate a subject score from any number of completed items, but we have to decide for ourselves how to combine multiple events and promptly get into the summing and averaging troubles described above.

Turning to the time periods within which events are observed, when we choose the length of a period we are, again implicitly, trying to identify a sampling period within which the underlying conflict level is stable. Making this decision allows us to treat multiple events in a period as repeated measures of a fixed underlying quantity.

Slightly more formally, assume that there exists an unobserved time series of true conflict levels x_t for $1 \leq t \leq T$ observed at suitable time intervals. We observe a sequence of sets of Goldstein scored events $\{y_1, \ldots, y_{N_t}\}_t$ where N_t is the number of events observed at t, possibly equal to zero when nothing is observed.

It is plausible to assume that x_t causes the events that are reported at t. In the best case, events and therefore their scores are conditionally independent given x. This is a standard measurement assumption sometimes described as 'local independence'. Local independence implies that any number of events can appear and larger values of N_t offer more precision in about x. At the other extreme, if no events occur the measurement framework leaves no doubt that there *is* some level of conflict; we simply do not at that moment get to observe it. These observations immediately provide a way to diagnose some event scaling problems.

2.1 Summed scaled scores

Summing scores are simply inappropriate: x_t has, by definition, the same interpretation and range as a single Goldstein score. But no sum involving multiple scaled events will be interpretable this way because its possible range is between N_t times the maximum or minimum of the original scale[1].

Why then did summed scores seem like a good idea? Presumably for their similarity with the raw form of the data as a multivariate count data. Indeed if we want to model count data, then exactly opposite conclusions apply: We *should* sum occurrences because they are naturally counts (although we should not scale them) and imputing zero when no events of a particular type are observed is in this case potentially correct.[2] It *is* possible to infer scalar x from a vector of events counts, and I do so in the second part of the paper, but simple summing is not the right way to do

[1] Actually we *can* work with a sum of scaled events at t, provided we take into account the fact that (again under independence) this quantity will have time-dependent observation measurement variance $v_t N_t$.

[2] A physics analogy may be helpful: counts are *extensive* quantities so they can be added but conflict scores are, if the argument above is correct, *intensive* so they should be averaged. There is also the 'stock' versus 'flow' terminology from systems engineering if you prefer an econometric version of a similar distinction (Harvey, 1991).

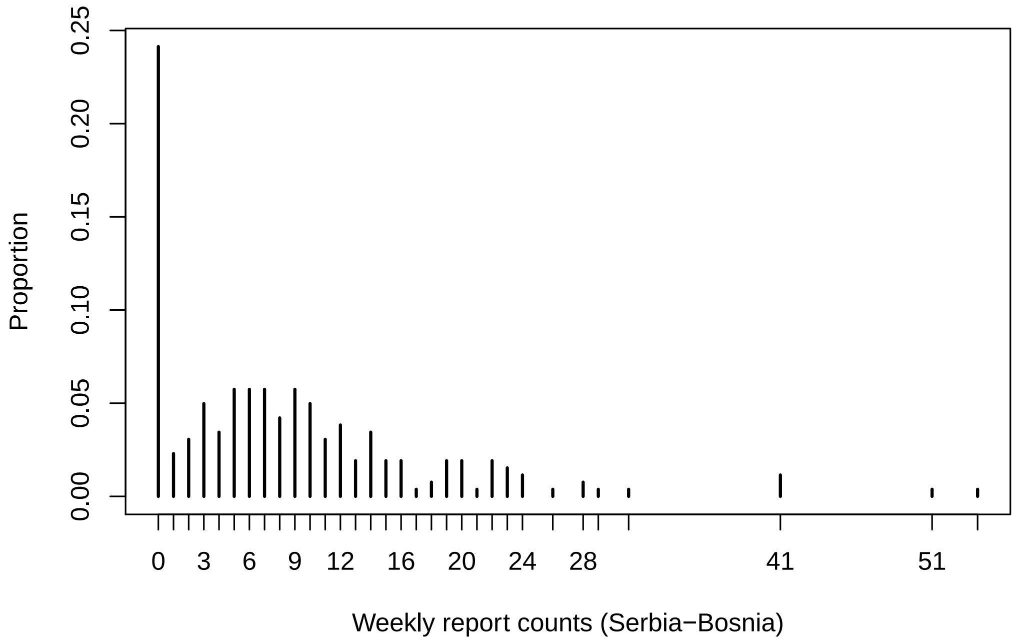


Figure 1: Weekly report counts (N_t) for Serbian actions to Bosnia, January 1991 to December 1995.

it. Mechanically it even involves a weighted average of event scale scores. Just not this one.

2.2 Averaged scaled scores

Consider instead averaging scaled event scores. Regardless of any model of x itself, as N_t increases we have more information to estimate it. In particular, if we summarise the observations at $\{y_1, \ldots, y_{N_t}\}_t$ with a mean $\bar{y}_t$ and denote the variance of each event measurement is v_t, then a reasonable level of uncertainty (assuming independence) for $\bar{y}_t$ is v_t/N_t.

By working instead with $\bar{y}_t$ we seem to fall directly into the mean problem, but in fact provide an analysis of it. In Yonamine's example a month with $N_t = 3$ events scored -10 is still estimated to be on average just as conflictual as a month with $N_t = 30$ identically scored events. What differs between these two cases is that we should be ten times more certain that x_t is close to -10 in the second than in the first case. The mean problem is therefore only a problem if the time-dependent reduction of observation measurement uncertainty is not represented in subsequent analysis.

2.3 Consequences

Will this matter in real data? Using data on Serbia's interactions with Bosnia fromGoldstein and Pevehouse (1997), Figure 1 shows that in the Serbia-Bosnia dyad at weekly aggregation there is considerable variation in N_t across the conflict period. While most of the weeks with no events are at the beginning of the period the remaining variation is spread across the subsequent conflict. Consequently the conflict level is measured *much* more precisely in some weeks than others.

A simple way to determine the consequences of the lack of variable N_t is to reanalyse some

Table 1: Bosnian conflict ‘threats’ phase (2/1994–12/1994) under different measurement assumptions. The dependent variable dyad is marked above each column of coefficients. Dyads are labelled S: Serbia, I: International, B: Bosnia. Bold face dyads are predicted by dyad scores in the previous week. The second and third columns are regression coefficients and T-test p-values for a VAR(1) model.

$\bar{y}_t \times N_t$			N_t			$\bar{y}_t$			N_t-weighted $\bar{y}_t$		
IS			**IS**			**IS**			**IS**		
IS	-0.19	0.39	IS	0.32	0.08	IS	0.07	0.67	IS	-0.02	0.91
SI	-0.36	0.45	SI	-0.39	0.23	SI	-0.17	0.36	SI	-0.26	0.16
SB	0.93	0.00	SB	0.82	0.01	SB	0.11	0.36	SB	0.39	0.00
SI			**SI**			**SI**			**SI**		
IS	0.03	0.78	IS	0.15	0.20	IS	0.05	0.79	IS	0.14	0.34
SI	-0.24	0.29	SI	0.18	0.37	SI	-0.39	0.04	SI	-0.39	0.01
SB	0.22	0.07	SB	0.15	0.43	SB	0.01	0.92	SB	0.13	0.22
SB			**SB**			**SB**			**SB**		
IS	-0.24	0.09	IS	-0.08	0.45	IS	-0.36	0.11	IS	-0.25	0.25
SI	0.46	0.14	SI	0.17	0.39	SI	0.10	0.66	SI	0.05	0.82
SB	0.56	0.00	SB	0.44	0.02	SB	0.46	0.00	SB	0.56	0.00

existing data. Here we replicate[3] some of Goldstein and Pevehouse’s Table 3. This is a dyadic analysis of the interaction between actors in the Bosnia conflict. Tables 1 and 2 show four analyses of the periods 2/1994-12/1994 and 12/1994-7/1995 respectively. The paper’s original analysis is in the first column, treating weekly summed Goldstein event scores (by definition $\bar{y}_t N_t$) as the dependent variable of a VAR(1) model. The second analysis tests a claim by Schrodt that “the frequency of coded events alone is the primary factor that differentiates the major political features of the data ”(Schrodt, 1994, p.17). Here the dependent variable is simply N_t. The next column treats $\bar{y}_t$ as the dependent variable without taking into account the extra precision implied by N_t, and the final column is a weighted least squares analysis that takes into account that the dependent variable $\bar{y}_t$ is an average of N_t terms.

Comparing the results in Table 1, the N_t-model does show predictable similarities to the original analysis because N_t is a large part of the variation in the observations. However it misses significant negative autocorrelation in Serbia’s treatment of international actors dyad that all other models pick up, presumably because it cannot distinguish positive and negative, only fewer and more events. When events are of mixed type results predictably diverge. Turning to Table 2 both models that use N_t find positive reciprocity between SI and IS but neither average based model does. Conversely both average based models find a coordination between international actors’ and Bosnian treatment of Serbia that no N_t based model does. While these differences are not huge, they matter substantively. Predictably these analyses also suggest that the $\bar{y}_t$ models that do not take into account variable event coverage are less efficient.

What these tables hide are model diagnostics. Goldstein and Pevehouse focus on serial autocorrelation and lag specification tests. However, ordinary diagnostics e.g. prediction vs residuals, QQ-plot, leverage and influence statistics all show that N_t-models are badly specified. This is not surprising – constructing an conditionally Normal data series out of N_t, which are counts and

[3] The replication data does not replicate the exact numbers in the paper, probably due to a discrepancy in the actor coding. The replication materials have more actors and events than mentioned in the paper.

Table 2: Bosnian conflict ‘promises’ (12/1994–7/1995) phase under different measurement assumptions: The dependent variable dyad is marked above each column of coefficients. Dyads are labelled S: Serbia, I: International, B: Bosnia. Bold face dyads are predicted by dyad scores in the previous week. The second and third columns are regression coefficients and T-test p-values for a VAR(1) model.

$\bar{y}_t \times N_t$			N_t			$\bar{y}_t$			N_t-weighted $\bar{y}_t$		
IS			**IS**			**IS**			**IS**		
IS	0.21	0.45	IS	0.54	0.08	IS	0.31	0.10	IS	0.13	0.35
SI	0.01	0.97	SI	-0.07	0.80	SI	0.14	0.48	SI	0.18	0.31
SB	0.09	0.72	SB	0.09	0.77	SB	0.16	0.24	SB	0.12	0.33
BS	0.26	0.67	BS	-0.75	0.22	BS	0.20	0.12	BS	0.32	0.00
SI			**SI**			**SI**			**SI**		
IS	0.54	0.00	IS	0.72	0.00	IS	0.20	0.40	IS	0.32	0.15
SI	0.05	0.80	SI	0.13	0.53	SI	-0.03	0.92	SI	-0.15	0.61
SB	-0.15	0.37	SB	-0.50	0.04	SB	0.13	0.47	SB	0.01	0.93
BS	0.12	0.74	BS	-0.07	0.87	BS	0.10	0.54	BS	0.07	0.59
SB			**SB**			**SB**			**SB**		
IS	0.15	0.43	IS	0.18	0.37	IS	0.41	0.11	IS	0.34	0.18
SI	0.19	0.44	SI	-0.02	0.90	SI	-0.11	0.70	SI	-0.08	0.78
SB	0.27	0.16	SB	0.42	0.05	SB	-0.04	0.83	SB	-0.01	0.95
BS	0.57	0.20	BS	0.01	0.97	BS	0.08	0.66	BS	0.15	0.31
BS			**BS**			**BS**			**BS**		
IS	-0.05	0.56	IS	-0.15	0.08	IS	0.68	0.03	IS	0.51	0.04
SI	-0.01	0.90	SI	0.17	0.04	SI	-0.55	0.11	SI	-0.30	0.29
SB	0.16	0.05	SB	0.22	0.02	SB	0.23	0.31	SB	0.30	0.09
BS	-0.26	0.15	BS	0.12	0.52	BS	-0.19	0.38	BS	-0.22	0.21

therefore have variance increasing with mean, creates strongly heteroskedasticity and clear outliers when important heavily reported events occur. In this data outlying and high leverage weeks are connected to surges of violence in Srebrenica, Sarajevo, Gorajde, and to NATO use of force. In contrast, the $\bar{y}_t$-based measures pass all the regression diagnostics.

A weighted average analysis only solves half the problem though. What to do about periods when no events occur? In the models above these observations are simply dropped. A better strategy is to model both the underlying level and the measurement process together.

2.4 Models for intermittently observed scaled event data

The measurement considerations above suggest a simple class of models for conflict levels derived from the state space time series framework (Durbin and Koopman, 2001; Harvey, 1991). In the simplest such model x is given Markov dynamics to define a dynamic measurement model that takes into account aggregation. The stream of events for a single dyad, assuming a fixed but possibly unknown variance v for scaled observations is then

$$\begin{aligned} x_{t+1} &= x_t T + \eta_{t+1} & \qquad \eta &\sim \text{Normal}(0, Q) \\ \bar{y}_t &= F x_t + \epsilon_t & \qquad \epsilon &\sim \text{Normal}(0, v/N_t) \end{aligned}$$

where in this simple situation $F = 1$. The 'local level' version of this model also sets $T = 1$ which giving x random walk dynamics. For known Q and v the 'state' x is estimated by Kalman filtering to get $x_{1\ldots t} \mid \bar{y}_1 \ldots \bar{y}_t$ and smoothing to get $x_{1\ldots T} \mid \bar{y}_1 \ldots \bar{y}_T$, which is the best estimate of the path of x. Kalman filtering provides an iterative way to compute the likelihood of the data under the model, so it can also be used to estimate the free parameters in the model T, Q, and v using Newton or Expectation Maximisation methods (Shumway and Stoffer, 2000).

This model takes into account that fact that x_t has a value, regardless of whether there are observations at t by separating out a dynamic model for conflict in the first line from a measurement model for observations in the second line. This setup is designed for a predetermined scale like Goldstein, so the mapping (defined with F) is trivially linear and the weekly reports affect measurement uncertainty by a factor of $1/N_t$.

Figure 2 shows the result of applying this model to estimate conflict levels in the Serbia-Bosnia dyad. Pointwise uncertainty is marked in grey. Average conflict scores are shown as grey points. The very large uncertainty at the beginning of the series indicates a diffuse prior on conflict level before the series begins which is not much constrained by events until about half way through 1992. Nevertheless the figure shows fairly clearly the course of Serbian action to Bosnia, from a period of persistently hostile interactions from mid 1992 starting with a first dip at the beginning of the siege of Sarajevo through mid 1993 then moving to an alternating pattern of cooperation, renewed hostility and broken peace plans and ceasefires, to attacks on Gorazde in March/April 1994. The final dip in 1995 is the fall of Zepa, Tusla and the the capture of and massacre at Srebrenica, followed shortly by NATO air strikes which apparently lead to decreasing hostility until the peace accords at the end of the year.

Unsurprisingly the standard deviation of posterior is tightly connected to the density of coverage (r=-0.89, p<.001) with more certainty expressed about peaks of cooperation and conflict since these are more densely reported. Increasing coverage also forces the estimated conflict level to follow the observations more closely and smooth over time periods less, as can be seen by comparing mid-1992 to mid-1993 with the much more variable remainder of the series as Western powers become more involved, more force is applied, and reporting is densest.

It is useful to review how missing data problems are dealt with in this formulation. Because x varies regardless of whether it is observed, if there are no observations at, say k, then the estimate of x_k is not a posterior combination of the prior prediction $\hat{x}_{k|k-1}$ from the previous week

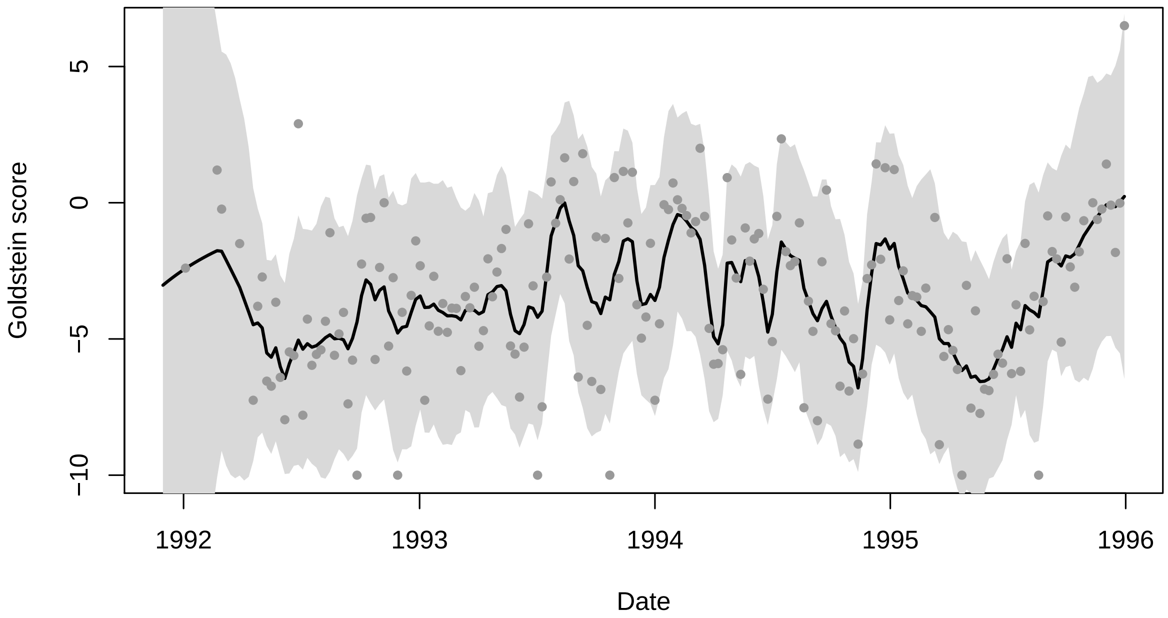


Figure 2: Conflict cooperation levels for the Serbia-Bosnia dyad. Bold black line is the estimated mean conflict level and grey regions are 95% posterior intervals.

and data $\bar{y}$ with variance ν/N_k from the current week but simply the prediction from last week. If time periods pass without data $\hat{x}$ follows the (here trivial) dynamics specified by T but with uncertainty increasing with every observationless period. The consequence is that missing observations are assigned a distribution. This can be used for any purpose, and the missing observations are effectively integrated out of inferences about other parameters. State distributions can also be used for multiple imputation (Tusell, 2012).

There are in fact four weeks in this data where no events occur, but these are sufficiently few that the expansion of the posterior is not easily visible in graph.

2.5 Model assumptions

The purpose of this paper is not to construct and fit a complete multivariate formulation of the simple dyadic model above, although that is future work, but rather to make clear the general consequences of a measurement approach to event data. The model above provides enough structure to show the solutions to the basic conceptual problems and paradoxes of aggregation described earlier, so I turn now the assumptions that are required.

Markov dynamics Scalar T implies particularly if it is set to 1. To relax this assumption T can be given more structure using a matrix with suitably transformed F. Variations on this theme to add extra series, seasonal terms, etc. constitute most of the work in state space time series analysis and we do not pursue them further here.

A natural extension would be to model the vector of mean conflict scores and estimate a full rank T. Off diagonal terms estimate cross dyad interactions and diagonal terms autoregressive tendencies. This would be the equivalent of the VAR(1) model only taking into account variable measurement error. This is future work.

Scale observation variance is constant The model assumes that v is fixed but unknown. In this formulation separate event scale estimates cannot be estimated because the Goldstein (1992) score simply stipulates that certain events get certain scores. It is, however, possible to use the expert judgement variances extracted from Goldstein's Table 1 instead, although with only eight experts they are unlikely to be precise, and are in any case subject to the caveat regarding event codes 011 and 012 noted earlier.

A more promising alternative is to build a less trivial measurement process that connects a vector of event counts observed weekly, to x. This is what the second half of the paper shows how to do outside the time series model context.

Report density is unrelated to conflict level The model also assumes that the sequence of event counts $N_1 \ldots N_T$ is exogenous. Prime facie selection bias considerations suggest this will not hold (see e.g. Reeves et al., 2006; Woolley, 2006, for reviews). However, the Bosnia data the hostility of events, as measured by $\bar{y}_t$, and intensity of coverage as measured by N_t are *not* significantly correlated[4] (r=-0.11, p=0.16).

Substantively, this corresponds to a scenario where the news agency in question (here Reuters) maintains a more or less continuous presence in the former Yugoslavia and sends regular reports. Of course, the exogeneity claim is, like all such claims, very hard to test directly.

Conflict level is scalar We also assume that x is continuous, rather than, e.g. a discrete unobserved state. If x were a nominal variable with K possible values, the model above would be renamed a Hidden Markov model (Zucchini and MacDonald, 2009), but its fundamental measurement features would be unchanged. Indeed all the measurement problems described above would have the same solutions: the $\bar{y}$ and N sequences would still be the fundamental forms of data, conditionally normal given x as above, missing data would be dealt with the same way by providing a distribution over possible values of x_k each element of which would become closer to $1/K$ as observation free time periods passed.

3 Measurement models for constructing event scales

The previous sections treated Goldstein's scale as given and considered how best to deal with it in a time series context to avoid aggregation problems. I now turn to the question of how to determine a scale from event data itself without polling experts. The raw data is now not expert-scaled events but the vector of event counts aggregated in each time period and the task is to learn a conflict scale.

Some previous work has taken a measurement modelling approach. In particular Schrodt (2007) used Rasch and Item Response Theory (IRT, Baker and Kim, 2004) models to determine a scale inductively from event category counts. Because these model work with 0/1 data this required transforming the stream of event category counts into zeros and ones. This was done by thresholding the raw counts based on whether they exceeded their monthly means, applying an IRT model and treating the latent variable as the induced scale values for each observation.

[4] in levels or logs of N_t

While Schrodt reported ‘mixed results’, and correlations with existing scales were weak this is nevertheless a very promising direction. I show below that with a minor change of measurement model class, event counts can be used directly to recover the Goldstein scale.

3.1 Measurement models for inducing a conflict scale

The assumption of all models is, as discussed above, that the elements of the vector of observations for each time period are conditionally independent given the unobserved x. Unlike the simple observation model above where the expected value of y_t is simply x_t with variance proportional to N_t, IRT models assume individual logistic regressions of observations on x_t. Unit changes in x then lead to slope parameter-sized changes in the log odds of seeing a 1 rathe than a 0. In Schrodt’s work that is, of a particular event category occurring more often than its monthly average.

There are two broad problems with this approach. First, events are multivariate count data, so reducing them to 0/1 loses information. In particular, it looses the information that increased reporting density offers. Second IRT models assume a Guttman-type structure in the relationship between event classes and x, that is, that there is an ordering of event categories such that the odds of seeing each event increases with x. As Schrodt observes, this may not be reasonable for event codes: conflict level and event occurrences are probably not structured in the same way as ability and question difficulty: easier questions are always answered by more able students but more cooperative actions are not guaranteed between high conflict actors.

A more promising measurement assumption from unfolding or ideal point models that assume that both x and the event category scores exist on the same scale and that as x moves closer to an event category score, that event appears more often.

Fortunately, unfolding-style models for count data exist and are applied widely for the task of extracting policy positions from text. The two problems are very similar. In event data scaling we are interested in extracting a scalar conflict measure from a set of event counts and in text scaling we are interested in extracting a measure of left-right position from a set of word counts. Suitable models in either domain explain both what the mapping is from x to observed counts and provide values for the unobserved x, ideally with some measure of uncertainty. Two such models are Wordfish (Slapin and Proksch, 2008; Monroe and Maeda, 2004), previously studied by (Goodman, 1979, 1985) as the ‘RC Model’ and Wordscores (Laver et al., 2003) which is a constrained version of correspondence analysis (Lowe, 2008).

3.2 Two scaling models for events

The connection between the multinomial unfolding model and the Wordfish text scaling model is not well-known in the literature, so I review it briefly here. Wordfish models a contingency table C of documents and words. When scaling events these are time periods and event classes respectively, so they are indexed t and j. Each row sums to N_t. Wordfish assumes that

$$\log\ \mathrm{E}[C_{tj}] = \alpha_t + \psi_j + x_t\beta_j$$

The term α_t ensures that N_t is reflected in the model’s fitted values. Consequently, since N_t is known (and treated as exogenous, as discussed above) then we can condition on it directly. The resulting form is multinomial with new parameters linearly connected to the ones above and arranged into logits of j versus j' (see Lowe and Benoit, 2011, for an complete development).

Estimates of x are not affected by working with the model as a multinomial logistic regression with ‘independent’ variable x or in the form above with no row sum constraint but nuisance parameter α_t. (Indeed when x *is* observed, the formulation above is a computationally easier way to

Table 3: Unit normalised event category parameters (β) from Wordfish and CA models. Categories are constructed by aggregating the WEIS cue categories in the second column.

Category	WEIS cue categories	Beta (Wordfish)	Beta (CA)
material cooperation	01, 03, 06, 07	0.80	0.99
verbal cooperation	02, 05, 07–10	0.59	0.55
verbal conflict	11–17	0.23	-0.28
material conflict	18–22	-1.42	-1.27

fit a multinomial logistic regression, sometimes known as the 'Poisson trick'.) Unsurprisingly, larger values of N_t lead to more precise estimates of x_t.

Also not well known is that correspondence analysis (CA, Greenacre, 1993) is a least-squares version of the same model. Correspondence analysis assumes (Greenacre, 1993, appendix A) that

$$C_{tj}/N = r_t c_j (1 - x_t \beta_j)$$

where $N = \sum_t N_t$. (To see that this is an approximation to the Wordfish model, consider taking the log of both sides.) It has the advantage of being easier to estimate and also allows us to investigate possible multidimensionality in x.

To fit these models to the Serbia-Bosnia dyad event counts, again in weekly aggregation, I first aggregate events into the event categories Schrodt recommends working with: material cooperation, verbal cooperation, verbal conflict, and material conflict. Running Wordfish and CA models on the weekly event counts gives a model with normalised β parameters that are interpretable as event category scores as shown in Table 3.

The parameters order the higher level categories as we would expect if a conflict cooperation scale had been induced. Further evidence that the model is capturing a conflict cooperation scale is that the both models' weekly estimates (x parameters) are highly correlated with weekly average Goldstein scores (r=0.85, p<.001 for Wordfish and r=0.9, p<0.001 for correspondence analysis). Figure 3 shows Wordfish and CA estimates against the average weekly Goldstein scores[5]

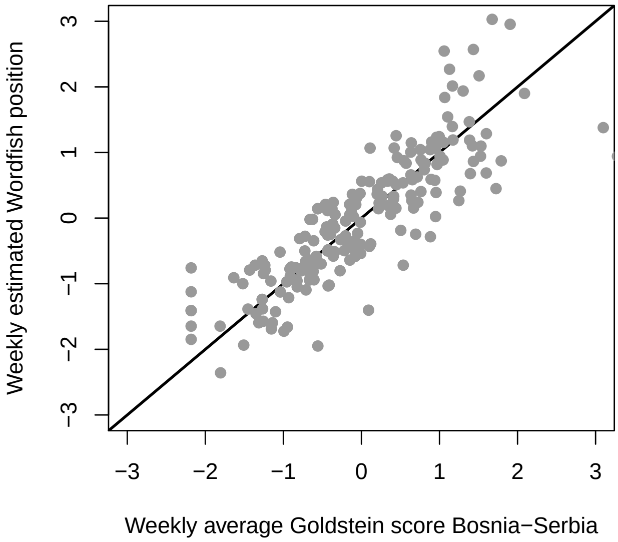
3.3 Scale dimensionality

One of the advantages of taking an inductive approach to measurement is the possibility of addressing questions of scale dimensionality. Since it seems clear that the main dimension when scaling weekly event counts is something very close to the Goldstein conflict-cooperation scale, we add another dimension to see what other structure might be lurking. Table 4 suggests that the second dimensional distinguishes verbal from material actions.

Finally, Figure 4 shows the category scores in both dimensions and the weeks numbers plotted between them. Since all weeks are composed of some proportion of each of the event categories they find positions on the simplex, distances within which are represented approximately in the two dimensional space defined by the correspondence analysis. The bulk of weeks move up and down the northwest to south east event axis with occasional forays toward material cooperation towards the end of the sequence.

Alternative analyses using all the event categories gives a less clear picture due to lack of data. High levels of aggregation are necessary to get enough covariance information to order event categories, but this leads to fewer observations. The several hundred weeks in the data set is apparently

[5] Wordfish estimation required a small constant to be added to each count to generate stable scores. However, correspondence analysis (using the `corresp` function from the R package MASS) worked on this data without any adjustment.



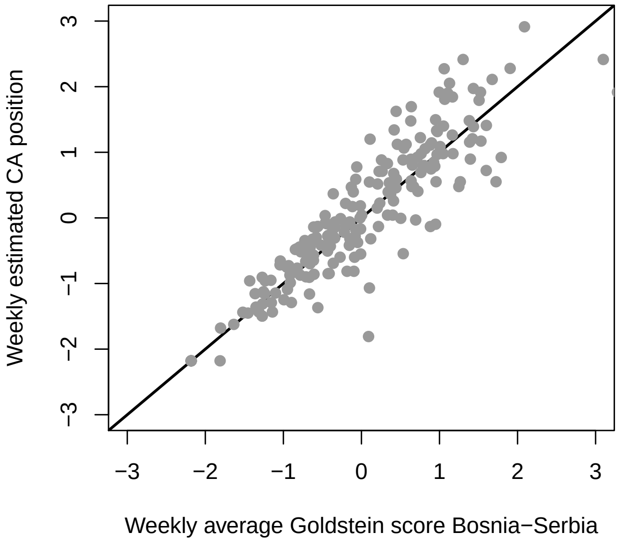


Figure 3: Weekly averaged Goldstein scores versus Wordfish (top) and correspondence analysis (bottom) estimates.

Table 4: Unit normalised event category parameters (β) for two dimensional correspondence analysis (CA) model.

Category	CA Dim. 1	CA Dim. 2
material cooperation	0.99	1.29
verbal cooperation	0.55	-1.01
verbal conflict	-0.28	-0.5
material conflict	-1.27	0.23

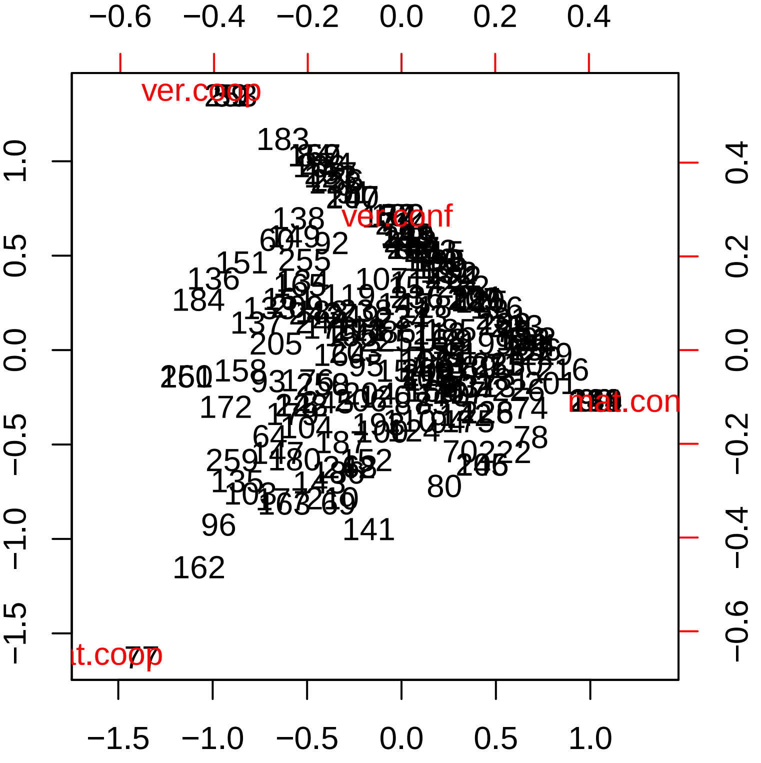


Figure 4: Week numbers and event categories from a two-dimensional correspondence analysis. From the top clockwise, the categories in red are verbal cooperation (‘ver.coop’), verbal conflict (‘ver.conf’), material conflict (‘mat.conf’) and material cooperation (‘mat.coop’).

not enough to reliably order 110 WEIS categories, and provides a rather messy analysis of the 22 cue categories. However, the success of the scaling analyses seems to confirm that an unfolding rather than a Guttman scale structure fits event data.

This model can, of course, only be preliminary, because it does not take into account the time series structure of x. Each week is treated as an independent containing N_t event counts.

4 Conclusion

As above, only shorter.

References

Azar, E. E. (1980). The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24(1):143–152.

Baker, F. and Kim, S. H. (2004). *Item Response Theory*. Wiley, New York NY, 2nd edition.

Bond, J., Petroff, V., O'Brien, S., and Bond, D. (2004). Forecasting turmoil in indonesia: An application of hidden markov models. Paper presented at ISA 2004.

Brandt, P. T. and Sandler, T. (2012). A Bayesian Poisson vector autoregression model. *Political Analysis*, 20(3):292–315.

Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call voting: A unified approach. *American Journal of Political Science*, 98(2):355–370.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Freeman, J. R. (1989). Systematic sampling, temporal aggregation, and the study of political relationships. *Political Analysis*, 1(1):61–98.

Goldstein, J. S. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36(2):369–385.

Goldstein, J. S. and Pevehouse, J. C. (1997). Reciprocity, bullying and international cooperation: Time-series analysis of the Bosnia conflict. *American Political Science Review*, 91(3):515–529.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, 74(367):537–552.

Goodman, L. A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, 13(1):10–69.

Greenacre, M. J. (1993). *Correspondence Analysis in Practice*. Academic Press, London UK.

Harvey, A. C. (1991). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Honaker, J. and King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(3):561–581.

Jenkins, J. C. and Bond, D. (2001). Conflict-carrying capacity, political crisis, and reconstruction. *Journal of Conflict Resolution*, 45(1):3–31.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Lowe, W. (2008). Understanding Wordscores. *Political Analysis*, 16(4).

Lowe, W. and Benoit, K. R. (2011). Estimating uncertainty in quantitative text analysis. Paper presented at MPSA 2011.

Lütkepohl, H. (1990). *Introduction to Multiple Time Series Analysis*. Springer Verlag, Berlin.

Monroe, B. and Maeda, K. (2004). Talk's cheap: Text-based estimation of rhetorical ideal-points. POLMETH Working Paper.

Pevehouse, J. C. (2004). Interdependence theory and the measurement of international conflict. *Journal of Politics*, 66(1):247–266.

Pickup, M. A. (2009). Measure twice, model once: Measurement methods for better longitudinal modelling. *Electoral Studies*, 28(3):349–353.

Quarterly, I. S. (1983). Symposium: Events data collections. *International Studies Quarterly*, 27.

Reeves, A. M., Shellman, S. M., and Stuart, B. M. (2006). Fair and balanced or fit to print? Occasional paper, University of Georgia, School of Public and International Affairs.

Schrodt, P. (1994). Statistical characteristics of event data. *International Interactions*, 20(1):35–53.

Schrodt, P. A. (2006). Forecasting conflict in the Balkans using hidden markov models. In Trappl, R., editor, *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*, pages 161–184. Kluwer, Dordrecht, Netherlands.

Schrodt, P. A. (2007). Inductive event data scaling using Item Response Theory. Paper presented at Polmeth 2007.

Schrodt, P. A. (2011). Forecasting political conflict in Asia using Latent Dirichlet Allocation models. Paper presented at European Political Science Association, 2011.

Schrodt, P. A. (2012). Precendents, progress and prospects in political event data. *International Interactions*, 38(4):546–569.

Schrodt, P. A. and Gerner, D. J. (2001). Analyzing international events. Available from http://eventdata.psu.edu/papers.dir/automated.html.

Shellman, S. M. (2004a). Measuring the intensity of intranational political events data: Two interval like scales. *International Interactions*, 30(109–141).

Shellman, S. M. (2004b). Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, 12:97–104.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and its Applications*. Springer Verlag, New York.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Tusell, F. (2012). Multiple imputation of time series with an application to the construction of historical price indices. MS.

Woolley, J. T. (2006). Using media-based data in the study of politics. *American Journal of Political Science*, 44(1):156–173.

Yonamine, J. (2011). Working with event data: A guide to aggregation choices. MS.

Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Chapman and Hall/CRC.