# Event Data in Space: Statistical Models of Machine Coded Vs. Human Coded Data *

Logan Stundal,† Benjamin E. Bagozzi,‡

John R. Freeman,§ and Jennifer S. Holmes¶

January 28, 2020

**Abstract**

Political event data are widely used in studies of inter- and intra-state political processes such as social protests, militarized interstate disputes, and civil wars. Recent years have seen notable advances in the automated coding of political event data from international news sources. Yet, the validity of machine coded event data remains disputed, especially in the context of event geolocation. Taking a cue from recent model-based event data validation studies (that do not consider geolocation), we estimate a series of discrete and continuous spatial error models using both machine and human coded event data. We then evaluate whether these separately estimated human- and machine-coded event data models produce comparable inferences, and similar estimates of spatial random effects, under varying conditions of spatial aggregation and remoteness. In doing so, we leverage data on subnational human rights violations in Colombia, as derived from the (Human Coded) Uppsala Conflict Data Program's Georeferenced Event Dataset (GED) and the (Machine Coded) Integrated Crisis Early Warning System (ICEWS). The spatial predictions derived from these data are validated against a collection of independently collected, gold standard records of human rights violations from Colombia. Our validation findings demonstrate the accuracy of

*This research is supported by NSF Grant Number SBE-SMA-1539302.We thank Clayton Webb and Phil Schrodt for comments.

†Department of Political Science, University of Minnesota. Email: stund005@umn.edu.

‡Department of Political Science & International Relations, University of Delaware. Email: bagozzib@udel.edu.

§Department of Political Science, University of Minnesota. Email: freeman@umn.edu .

¶School of Economic, Political and Policy Sciences, University of Texas, Dallas. Email: jholmes@utdallas.edu.

machine coded event data and their geolocations, while also providing scholars with guidance about how best to spatially aggregate and model subnational political event.

Scholars agree that text is a valuable source of political data (Grimmer and Stewart 2013, Wilkerson and Casas 2017). They also agree that machines potentially are better able to extract information about the location and timing of political events than humans, especially in the context of large scale event data collection efforts (King and Lowe 2004). This is predominantly because machines can filter large amounts of text more quickly and consistently than humans (Beiler et al. 2016, Schrodt and van Brackle 2013). Using machines for event data coding, we therefore should be able to make significant progress in understanding, for example, the diffusion and dynamics of civil war (Brandt et al. 2008) and narco terrorism (Osorio 2015) as well as in forecasting international conflict (Brandt et al. 2014).

What is at issue is whether machine coding is as valid as human coding for the measurement of political events. Measurement validity can be defined as the degree to which one's codings meaningfully reflect a corresponding concept (Adcock and Collier 2001). With regards to event data derived from text, there are two kinds of validity, internal and external. The former evaluates whether machine and human coders extract the same information from the same text (Grimmer and Stewart 2013: 279). The latter assesses whether the information extracted from the text by the machine and human coders corresponds to ground truth or to what actually happened at some location at a particular time. Internal validation alone is a pyrrhic victory if the text on which it is based is itself inaccurate and(or) incomplete.[1]

The results of validation efforts are mixed. Some studies find that machines and humans do an equally good job of coding events (Schrodt and Gerner, 1994; King and Lowe 2004). Others argue that machines do an especially poor job of coding events, particularly the location of events. For example, in introducing the Uppsala Conflict Data Program's (UCDP's) Georeferenced Event Database [GED] Sundberg and Melander (2013: fn. 4) argue that machine geocoding is not fruitful because machines cannot distinguish locations of events when there are multiple cities across the world with the same names (see also Hammond and Weidman 2014 and Althaus et al. 2018). As regards external validity, several researchers found

[1]See the appendix for schematics of how Schrodt and Gerner (1994) and King and Lowe (2004) each attempt to achieve one or both kinds of validity.

a *remoteness problem* in human and/or machine coded event datasets. Due to journalistic practice and other factors, human and machine coded event data have been shown to be less accurate the more remote an event is from major urban centers (Davenport and Ball 2002, Hammond and Weidman 2014, Weidman 2016). Researchers also discovered, not *aggregation problems*. Codings are more valid the higher the level of unit aggregation. That is, both human and machine coding appear more valid the higher the administrative unit and more temporally aggregated one's data. However, at the same time, report aggregation, both in creating outcome (dependent) and predictor (independent) variables, results in information loss and papers over certain kinds of measurement error (Cook and Weidman 2019).

Two research designs are used in extant evaluations. In one, researchers use experts to establish what is assumed to be valid codings. These experts then train a group of individuals to code text. The individuals' error rate is gauged in a pilot study. Then a corpus of text is assembled and the coding by the trainees is compared to the coding by a particular piece of software (Althaus et al. 2018, Raytheon BBN Technologies 2015). In some cases, the design includes assessments of the accuracy of the locations of events reported in the corpus by humans and software. The problem with this design is that the original expert coding is treated as the "gold standard." The accuracy of the expert codings is not questioned. And in the case of location coding it is assumed that the trainees have extensive geographical knowledge. Yet even the creators of GED admit human coders often lack this knowledge (Croicu and Sundberg 2015: 14). External validity often is not assessed. Neither the human (trainees) nor the machine coding is compared to an independent source. And, if such comparisons are made, the effects of remoteness and aggregation are not assessed.[2]

[2]Althaus et al. compare, for the same corpus of reports about Boko Haram in Nigeria, human coding from the Cline Center's Social, Political and Economic Event Database (SPEED, Nardulli, Althaus, and Hayes 2015) to the Python Engine for Text Resolution and Related Coding Hierarchy [PETRARCH-2]. Raytheon BBN Technologies (2015) is an assessment of the human vs. World Wide Integrated Crisis Early Warning System [ICEWS] machine coding of a comparable corpus. The former study includes some assessment of geo-coding. The latter does not.
The idea of using independently collected data to gauge the external validity of event data sets is becoming more common. See, for instance, Zammit-Mangion et al. (2012), Weidman (2016) and von Borzyskowski and Wahman (forthcoming)

The other design asks if the inferences based on the estimates from a single statistical model of human and machine coded data are the same *and* if the two models are equally good at predicting an independently collected set of events. An early example of this approach is Schrodt and Gerner's (1994) comparison of cross correlations and periodograms for human and machine coded data.[3] More recently Bagozzi et al. (2019) used Cook et al.'s (2013) binary misclassification model to gauge underreporting bias in the human coded GED, the human coded Social Conflict in Analysis Data [SCAD; Salehyan et al. 2012], and the machine coded World Integrated Crisis Early Warning System data set [ICEWS; Boschee et al. 2016]. They found remarkable similarities in the patterns and statistical significance of the coefficients in binary models of machine and human coded data, as well as in the coefficients in the auxiliary equations that explain the tendency of the codings to underreport events. In addition, Bagozzi et al.'s statistical model employing machine coded event data performed as good or better than their models employing human coded event data in predicting independently collected data on human rights violations. Neither of these studies explicitly addressed the accuracy of locational coding or the aggregation issue, however. The one study which did evaluate geo-location by human and machines by means of spatial statistical analysis, as we explain below, is of limited usefulness.

In light of these shortcomings, this paper poses the following research questions. First, do spatial models estimated from machine and human coded data produce the same inferences and estimates (maps) of random effects? Second, do fitted spatial models based on the two kinds of data tell the same stories about the impact of remoteness and aggregation? And, finally, are fitted spatial models based on human and machine coded data equally successful in predicting independently collected event data, data that are closer to ground truth?

Our test bed is a collection of data on rebel-perpetrated violence against civilians in Colombia during the period 2002-2009, data coded by humans in the GED and by machines

[3]To be more specific, Schrodt and Gerner (1994) examined the statistical properties of time series based on machine coded event data—specifically the Kansas Event Data System's coding of the World Event/Interaction Survey Data (WEIS)—in comparison to the (human coded) Conflict and Peace Data Bank (COPDAB) dataset's coding of the same WEIS corpus.

in the ICEWS data sets. Using a sequential cross-sectional design, we compare the estimates from both neighborhood (discrete) and geostatistical (continuous) spatial models of the two kinds of data.[4] After comparing the inferences about the impacts of theoretically important covariates and the impact of aggregation and remoteness of these inferences, we then compare estimates for the models of the independently collected data of the Centro de Investigación y Educación Popular (CINEP).

Briefly, we find...

The ensuing discussion is divided in to three parts. In the first section below we describe human and machine geo-coding for political event data and briefly discuss two approaches to spatial statistical modeling. The estimation and comparison of the Colombian spatial models is presented in part two. Part three outlines directions for future research including the need to evaluate additional human and machine coded data sets, and how to incorporate the temporal dimension into our research design.

# 1 The Production and Evaluation of Geolocated Political Event Data

## 1.1 Spatial coding of events

Both human and machine coded event datasets primarily code events (in terms of who did what to whom, and where/when) from international newswire reports. Based upon the location mentions within these newswire reports, human coders use a wide range of supplemental sources when assigning locations to events. GED coders employ the database of the National Geospatial Intelligence Agency (NGA), for example. They supplement this source with Google Earth, maps produced by aid agencies and field atlases. They employ the

[4]Sequential cross- sectional designs are often used to evaluate spatial interdependence. Examples include Baller et al. (2001) and Cho (2003). Single cross sections sometimes are employed in the study of spatial patterns of local violence as well, e.g., DeJuan (2013).

Global ISO 31662 standard for assigning administrative divisions. GED also reports a seven point precision scale for its geo-locations (Sundberg and Melander 2014: 526; UCDP Codebook pps. 14, 22-24). Another example is the Political Instability Task Force's Worldwide Atrocities Dataset (PITF). The PITF primarily relies on human coder lookups of identified place names in news articles via the GeoNames geographical database, alongside additional resources when necessary such as Google Earch (PITF 2009). Location is assigned with village- or city-level precision unless otherwise noted. A third example is the Cline Center's Social Political Economic Event Dataset (SPEED). SPEED similarly relies on the GeoNames database for geo-location (Nardulli et al. 2019). It proceeds first in an automated fashion by identifying place names in relevant articles with natural language processing and then passing these place names to GeoNames to obtain confidence scores associated with potential locations for each event. Human coders are presented with the latter information via drop down menus for actual event geo-location.

Machines fully automate the geo-location process described for SPEED above. Lee, Liu, and Ward (2018) characterize this automation as following three general steps. First, named entity recognition (NER) is used to identify the words in a given news article that correspond to location names. Second, each location name that is identified within a given news article is disambiguated to establish that location name's most likely true geographic location. GeoNames or similar geographic databases can be employed during this second step–such as in the case of the Mordecai geo-parsing software (Halterman 2017)–alongside additional contextual information from the original news text. Third, the (potentially multiple) disambiguated location name(s) recovered in step two are then evaluated to identify the proper geographic location for a given news article's (separately macine coded) political event.

Several automated routines have been proposed to implement the three steps described above. One common approach employs the CLIFF/CLAVIN geo-location software (D'Ignazio et al. 2014), oftentimes within additional automated event extraction software such as PETRARCH-2 (Norris, Schrodt, and Beiler, 2012). CLIFF/CLAVIN is currently imple-

mented in both the historical and real-time Phoenix event data projects (OEDA 2016, Althaus et al. 2017). Lee, Liu, and Ward (2018: 5) characterize the CLIFF/CLAVIN approach as one that leverages the frequency of location mentions to associate the proper disambiguated geolocation with a particular event. Halterman (2018: 3) critiques this CLIFF/CLAVIN-based Phoenix geo-location process[5] because it identifies a single "top" geo-location for an event in a manner that does not leverage any additional available information pertaining to the event itself. Lee, Liu, and Ward (2018: 5) point out that ICEWS' internal geo-location software—which is similar to that of CLIFF/CLAVIN—likewise only selects a single most likely geo-location for each event based upon a statistical ranking of all NER-identified locations in a news article. Given a number of potential inaccuracies in this single-shot geo-location assignment approach, recent research has proposed improved machine-based geolocation steps that leverage additional natural language processing (NLP) and machine learning tools to better disambiguate and assign geo-locations to events (Halterman 2017, 2018; Lee, Liu, and Ward 2018). However, these innovations have not yet been widely integrated into existing machine coded event data projects.

While these machine-based geo-location approaches identify a single latitude-longitude 2-tuple for relevant events, these coordinates do not always correspond to the city or village level. They can instead correspond to a municipality, department, or country centroid, depending on the location identified by NER and the event dataset being considered. As such, several machine coded event datasets report the name of the most spatially accurate geographic unit associated with each event, separately from the latitude-longitude 2-tuple recorded. For example, ICEWS includes latitude-longitude 2-tuples for all geo-located events, but additionally reports separate variables for "city," "district," "province," and "country." For these latter four variables, location names only appear for the relevant levels of geographic accuracy for each event, and the location names are missing otherwise. This allows one to back out a level of geo-coding precision for each event to produce the kind of precision

[5]As well as the geolocation steps implemented within the Global Database of Events, Language and Tone (GDELT) dataset (Leetaru and Schrodt 2013).

estimate that is provided by GED, albeit with less granularity.

Past evaluations of the internal validity of geo-located event data find that machine coded data are less accurate than human coded data. For example, Althaus et al. (2018:20ff) report that, in comparison to their SPEED (human) coders, PETRARCH-2's CLIFF/CLAVIN machine coder missed country level geo-location information in roughly 30%-70% of event reports pertaining to recent protests, curfew impositions, and suicide bombings for Nigeria. They found that CLIFF/CLAVIN's performance was even poorer at the state and province levels in this context. In a parallel study, Lee, Liu, and Ward (2018), compared the performance of their new two stage supervised machine learning algorithm for geolocation with the performances of the the machine geo-location coders used by the creators of the Phoenix and ICEWS data sets relative to human-coded geo-locations for the same collections of news stories regarding China, Syria, the Democratic Republic of the Congo and Colombia. Lee et al. show that their new geolocation algorithm is more accurate than the internal geo-location routines currently used by ICEWS and Phoenix, though none of these machine coders is able to accurately classify their human-labeled geo-locations at levels of accuracy greater than 90% (2018, Table 3).[6] Hammond and Weidman (2014) likewise compared the geo-location of violent events in twenty five African countries between 1997 and 2008 as reported in two human coded databases, ACLED[7] and GED and the machine coded Global Database of Events, Language and Tone (GDELT) dataset (Leetaru and Schrodt 2013). The GDELT data and the GDELT geolocations were not produced by CLIFF/CLAVIN. Rather, GDELT used a simple, automated look up of place names for each event (Leetaru and Schrodt 2013). Hammond and Weidman's results were similar to those of Althaus et al. in that GDELT's geo-tagged data compared much more favorably to human coded geo-tagged data at the national level than at the subnational level of aggregation. We discuss Hammond and Weidman's work in more detail below.[8]

[6] Pro and con on Lee et al.?

[7] The Armed Conflict Location and Event Dataset (Raleigh et al. 2010).

[8] Halterman 2019 as another study which reached positive results re potential of machine geo-coding? Weave into paragraph somehow?

These studies are of tremendous importance in gauging the internal validity of machine coded data and in advancing automation software. However, as we argued in the Introduction, all of them treat the expert (training set) coding as ground truth. Uncertainty in expert coding in not incorporated into the comparisons of human and machine coding in a systematic way. The estimation uncertainty from classifiers is ignored. And variance in performance in cross validation–for selected classifiers and across classifiers–also is not reported. Spatial statistical analysis of the two kinds of data therefore potentially is of much use in gauging the validity of machine geo-location.[9]

## 1.2 Spatial data analysis as validation

The familiar "neighborhood approach" (Anselin 1996: 113) analyzes spatial dependence in a variable of interest between *discrete* units like census tracts, municipalities, electoral districts, provinces, and countries. It therefore analyzes areal or lattice data.[10] The neighborhood approach usually employs "diffuse" pretests such as Getis and Ord statistics and(or) Moran's I. On the basis of the results, a connectivity matrix is prespecified; this matrix implies a cutoff or step function for where the relationship between units ends. A particular spatial model then is posited such as a spatial distributed lag (SDL) or spatial error model (SEM). Either type of model may contain both spatial and aspatial covariates (Cho 2003). Sometimes"focused" Lagrangian multiplier tests are used in specifying a model. In this case, the residuals from a simple linear model with covariates are analyzed and inferences are made

[9]Lee et al. (2018) appear to have hand-coded only those stories that ICEWS/Phoenix had already coded so they are to some degree ensuring that the most ambiguous stories (among the relevant news stories) are excluded. Moreoever, by only coding stories from the CAMEO categories "fight" and "protest" (*Ibid.* 2-3) they are ensuring that the action types of their handing coded stories are fairly clear cut and the respective locations are unambiguous. As regards estimation uncertainty Lee et al.'s classifiers yield proabilistic outputs (*Ibid.* p. 12). Their evaluations (Table 3, Fiure 3) apparently are based on predictions of location words with the highest predicted probabilities for the events in each story. But these classifiers are subject to estimation uncertainty. This uncertainty is increases as the number of parameters, trees, etc. grows. How this uncertainty is incorporated in the evaluation is never explained (if it is incorporated). Finally, the sample size for the evaluation is relatively small. Lee et al. appear not to have enough data to perform their final cross validations much less a separate set of human related data for searching for the best tuning parameters.

[10]These are data random aggregate values over an areal or lattice with well defined boundaries, a countable collection of given spatial units (Blangiardo and Cameletti 2015: 173.)

about the underlying functional form of the discrete spatial model. Post-tests on the residuals from the chosen models are used to determine model adequacy (Baller et al. 2001, Cho 2003, Barros et al. 2004). Out-of sample forecasting also is used (Ward and Gleditsch 2002).

SDLs and SEMs differ in how they conceive and analye spatial dependence. Of two, conceptually, the spatial error model is best suited to answer our questions about the validity of human and machine coded events data. Whereas SDLs represent the "mechanisms" that connect some units' behaviors such as interacting agents and social interaction (Anselin 2006: Section 26.2.1), SEMs capture data problems like the fact that model errors for neighboring units cluster together–"smaller(larger) errors for observation i ... go together with smaller [larger] errors for [neighbor] j" (Ward and Gleditsch 2019: 76)–and errors are correlated because of the mismatch between the spatial scale of a process and the discrete spatial units of observations (Anselin 2006: 907). These patterns of errors correspond respectively to what researchers call remoteness and aggregation problems in events data analysis. For example, remoteness means that a model's underestimates of violence in a unit distant from a capital city correlate with underestimates of violence in a neighboring unit which is also distant from the same city.

In its simplest form, the SEM model can be written

$$y_i = \beta x_i + \epsilon_i + \lambda w_i \xi_i \tag{1}$$

where i is the unit of observation; y and x are vectors of the variable of interest and the covariates, respectively; $\beta$ is a vector of coefficients on the covariates; $\epsilon$ is a spatially uncorrelated error term which may or may not be heteroscedastic, $w_i$ is the pre-specified, spatial weight row vector for unit i; $\lambda$ is a parameter that represents the degree of spatial correlation of the errors; and $\xi$ is the spatial component of the error term in unit i (Ward and Gleditsch 2019: Chapter 5). As regards estimation, the SEM is a special case of a model with a nonspherical error term; as such, conditional on the type of error process that exists,

a feasible generalized least squares estimator is employed. In contrast to the method used in time series analysis, however, a numerical method must be used to first to estimate $\lambda$ (Anselin, 2006: section 26.5.2.2).[11] Our questions are whether the estimates for $\lambda$ and the patterns of residuals from SEM models for human and machine coded data are similar.

To our knowledge, only one study has employed a discrete spatial model to evaluate human and machine geo-tagged events: Hammond and Weidman (2014). Hammond and Weidman use a SDL model with a temporal lag to explain a binary dependent variable indicating the occurrence of violence in African countries. The occurrence of violence was recorded by humans in ACLED and GED and by machine in GDELT. Briefly, they find that only the ACLED and GED models indicate that there is spatial dependence in patterns of violence. Moreover, the fitted SDL models for ACLED and GED confirm the conventional wisdom that violence is more likely in remote parts of countries. GDELT suggests the opposite. Hence they conclude that "there is clear evidence for a capital-centric geo-coding pattern [bias] in GDELT" (2014: 5).[12]

Hammond and Weidman make an important contribution. The problem with their piece is that it evaluates a single machine coded database (GDELT) whose transparency and false positive rates have been widely criticized (Ward et al. 2013; Wang et al. 2015). The machine coded ICEWS database is much more widely accepted and used.[13] As regards the specifics of their design, Hammond and Weidman apparently performed no pretests on their data. They also did not report the results of any specification tests for their model. They posited

[11]For a derivation of form of the variance-covariance matrix for three generic kinds of spatial errors–spatially autoregressive, conditionally autoregressive and spatial moving average–see Anselin 2006: Sections 26.3.1-3). The same source shows how these variance-covariance matrices are used in the maximum likelihood estimation of the SEM model (*Ibid.* Section 26.5.2.2). Attention must also be paid to the possibility of nonnormality and heteroscedasticity. See, for example, Cho 2003: fn. 13, p. 375. More complex methods for fitting autologistic spatial models are reviewed in Ward and Gleditsch (2002).

[12]The covariates in the Hammond and Weidman's model are: Conflict lagged one month, a Spatial Lag lagged one month, distance to the capital, population, and per cent mountainous; they include a constant and country fixed effects as well. Their test bed is 25 African countries and 136 country years between 1997 and 2008. They aggregate the ACLED, GED and GDELT data at the level of cell-months using uniform grid cells of about 55 km on each side of the PRIO-GRID data set. They therefore have 5897 unique cells and a total of 547,812 monthly observations (2014: 3).

[13]To this end, we note that ICEWS is considered to be one of the most accurate event datasets currently available (D'Orazio et al. 2011: 4).

a SDL model with temporal lag without describing (justifying) the connectivity matrix on which the spatial lag is based. The usefulness of the more conceptually appropriate SEM and what it tells us about the spatial error correlation patterns in the human and machine coded data apparently were not explored by Hammond and Weidman. Hammond and Weidman give no details about how their model was estimated, let alone any details about the fit of their model. Finally, they preform no external validity tests of their fitted models on independently coded event data.

The larger problem with the neighborhood approach to spatial modeling is the danger of "inappropriate discretization" (Lindgren and Rue 2015: 3). Applications like Hammond and Weidman's employ a prespecified connectivity matrix that treats the influence of spatial dependence as a step function–uniformily the same for some group of units and nonexistant for another group. In reality, dependence between units' behaviors is likely to vary *continuously* in space. The chosen connectivity matrix may misrepresent this process; the same might be true of robust checks with alternative connectivity matrices. Modeling this continuous process is a second approach to spatial analysis, an approach frequently employed in geo-statistics (Anselin 1996: 114). It *estimates* the distance at which units' behaviors are related. In addition, the second approach mitigates the aggregation problem (Cho and Gimpel 2007: 258) and it provides additional insights into the spatial correlation between model errors.

Geostatistical models analyze point referenced data. These models are based on the idea of a continuous spatial domain. For example, even though terrorist events are observed at specific locations and therefore "inherently discrete" these events can be interpreted as realizations of a continuously indexed space time process (Python et al. 2016, 2017,2018).[14] A variety of tools are available to test for spatial dependence in such data. One which has been applied in political science is the semivariogram (Cho and Gimpel 2007).

[14]The data, say y(s), is a random outcome at a specific location and the spatial index, s, can vary continuously in a fixed domain; s is two dimensional vector with latitudes and longitudes (three dimensional if altitudes are considered)

One geostatistical approach to analyzing these kind of data is Continuous Domain Bayesian Modeling with Integrated, Nested Laplacian Approximation (INLA).[15] Briefly, this approach "does not build models solely for *discretely observed data* but for approximations of *entire processes* defined over continuous domains" (Lindgren and Rue 2015:3, emphasis in the original). It assumes that the data generating process is a Gaussian field, $\xi(s)$, where s denotes a finite set of locations, $(s_1 \dots s_m)$. As such it suffers from a "big n problem;" analyzing the Gaussian field is costly computationally (Lindgren et al. 2011) Therefore, a particular linear, stochastic partial differential equation is assumed to apply to the Gaussian field:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau\xi(s)) = W(s), \qquad s \in D \tag{2}$$

where $\Delta$ is a Laplacian, $\alpha$ is a smoothness parameter such that $\alpha = \lambda + 1$ (for two dimensional processes), $\kappa > 0$ is a scale parameter, $\tau$ is a precision parameter, the domain is denoted by D, and W(s) is Gaussian spatial white noise. The solution of this equation is a stationary Gaussian field with the Matérn covariance function:

$$Cov(\xi(s_i), \xi(s_j)) = \sigma^2_{\xi_i} \frac{1}{\Gamma(\lambda)2^{\lambda-1}} (\kappa \mid\mid s_i - s_j \mid\mid)^{\lambda} K_{\lambda}(\kappa \mid\mid s_i - s_j \mid\mid) \tag{3}$$

where $\mid\mid s_i - s_j \mid\mid$ denotes the Euclidean distance between locations $s_i$ and $s_j$, $\sigma^2_{\xi_i}$ is the marginal variance, $\Gamma(\lambda) = \lambda!$, $K_{\lambda}$ is the modified Bessel function of the second kind and order $\lambda > 0$. The distance at which the spatial correlation becomes negligible (for $\lambda > .05$) is the range, r. The range is determined *empirically* from the scale parameter as $r = \frac{\sqrt{8\lambda}}{\kappa}$. In other words, in contrast to the neighborhood approach to spatial analysis, the range is *estimated* not pre-specified in a connectivity matrix or by a distance cutoff parameter.[16] This Gaussian field can be represented (approximated) by a Gaussian Markov Random Field. A finite element method using basis functions defined on a Constrained Refined Delaunay

[15]The following description draws from Blangiardo and Cameletti 2015 Chapter 6 and especially the passage on pps. 234-5 of Python et al. 2017

[16]See, for instance, Anselin 2006: Section 26.3.3 on nonparametric distance models.

Triangularization (mesh) over the corresponding shape file is used for this purpose.[17]

A hierachical Bayesian framework is used to model the data. For dichotomous data like the observation of a human rights violation in some time period, three equations are employed:

$$y_i \mid \eta_i, \theta \sim Bernoulli(\pi_i), \qquad i = 1 \ldots m \tag{4}$$

$$\eta_i \mid \theta = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{k,i} + \xi_i, \qquad i = 1 \ldots m \tag{5}$$

$$\theta = p(\theta) \tag{6}$$

where $y_i$ is the observation at point $i$, $m$ is the number of vertices in the Delaunay Triangularization, the second equation is the linear predictor, $\eta_i = logit(\pi)$, with spatially explicit covariates $z_{k,i}$, $\xi_i$ the Gaussian field as defined by equations (2) and (3) and approximated by the GMRF at point $i$, and equation (5) assigns the hyperparameters $\theta = (\kappa, \sigma_\xi^2)$. The same set-up can be used for count data using a Poisson link function (Python et al. 2016, 2018).

The choice of the priors is based, in part, on understandings of how the media covers events.[18] Because it is more efficient than Monte Carlo Markov Chain methods, INLA is used to estimate the model. Besides estimates of the effects of spatially explicit covariates on the probability of events, this geostatistical approach produces estimates of the parameters in the GMRF, in particular, the mean and standard deviation of the latent field at each point in the data set. In the simple static model above, these estimates tell us about the impact of uncertainty associated with both the scarcity (absence) of data and measurement

[17]This geostatistical approach is closely related to kriging, a method that has been applied in political science by Cho and Gimpel 2007 and more recently by Monogan and Gill (2015). For a discussion of how the present approach is related to kriging and several other geostatistical methods see Lindgren and Rue 2015: 1-2; see also Blangiardo and Cameletti 2015. Weidman and Ward (2010: 885ff) use GMRFs for lattice data in their application of the autologistic model. Chyzh and Kaiser (2019) use the GMRF concept in their graph theoretic analysis.

[18]In setting penalized complexity priors for the parameters of the Matérn covariance function Python et al. (2018): 8-9 based their choices on the assumption that "terrorist events should not substantially influence each other beyond relatively large distances (approximately 1911km), e.g., through demonstration and imitation processes promoted by the media."

error. Thus they will tell us how human and machine coded data compare in terms of the aggregation and remoteness problems without relying, as in the neighborhood approach, on an inappropriate discretization.[19]

## 1.3 External Validity and Design Overview

Table 1 summarizes our research design. The first row of Table 1 illustrates our *internal validation* assessment. For this assessment, we estimate separately a SEM (neighborhood model) for human-coded (GED) and machine coded (ICEWS) event data. We then repeat this GED vs. ICEWS comparison for the Continuous Domain Bayesian (geostatistical) model described above. Our assessment of internal validity evaluates how closely one's spatial inferences align when modeling human coded and machine coded data in these two ways. However, this analysis does not tell us the degree to which our spatial inferences, and any differences therein, are at all valid in relation to an *external* record of actual events. The second row in Table 1 describes our *external validation* assessment. Here we again separately compare inferences based on SEM and Continuous Domain Bayesian models of our human coded (GED) and machine coded (ICEWS) data. These comparisons are based on spatial models of a gold standard event database, namely, a collection of events collected independently by Colombia's Centro de Investigatción y Educación Popular (CINEP 2008). We describe the CINEP data below.[20]

[19]A more complex model assumes space-time separability. It also decomposes the stochastic part of the model into a GMRF and Gaussian white noise both of which are time dependent. The Gaussian white noise component then is interpreted as measurement error. See Python et al. 2016: 7, 2018: 7-8. We return to the possibility of using these more complex geostatistical models to compare machine and human geo coded data in the Conclusion.

[20]Two examples are Weidman (2016), von Borzyskowski and Wahman (2018). The former compares human coded events in an early version of GED to events reported in the United States Army database for Afghanistan (SIGACTS); the latter compares (human coded) SCAD- and ACLED-derived reports of electoral violence in Malawai and Zambia to reports from independent electoral monitors in these two countries, specifically, the Malawi Election Monitor Survey (MEMS) and the Zambia Election Monitor Survey (ZEMS).See also Zammit-Mangion et al. (2012) who compare results for data from the Afghan war diary to results for data collected by an independent NGO.

| | Neighborhood Model | Neighborhood Model | Geostatistical Model | Geostatistical Model |
|---|---|---|---|---|
| Internal Validity Assessment | Human Coded Data [GED] | Machine Coded Data [ICEWS] | Human Coded Data [GED] | Machine Coded Data [ICEWS] |
| External Validity Assessment | Human Coded Data [GED,CINEP] | Machine Coded Data [ICEWS,CINEP] | Human Coded Data [GED,CINEP] | Machine Coded Data [ICEWS,CINEP] |

Table 1: Research Design

# 2 Colombia as a Testbed for Evaluating Human and Machine Coded Text

With a domestic insurgency that has now spanned over five decades, Colombia has been the scene of an egregious number of human rights violations. These violations have been charted at the subnational level by numerous researchers and nongovernmental organizations (e.g., Holmes et al. 2007, Guberek et al. 2010, Lum et al. 2010, Bagozzi et al. 2019). In keeping with much of this past research, our human rights violation validation efforts focus on rebel perpetrated instances of violence against civilians. Separate human and machine coded databases contain comparable geo-tagged records of such violations. In particular, both GED and ICEWS code rebel perpetrated human rights violations against civilian targets using similar ontologies, and they use many of the same news source(s) to code events.[21]

There also exists several independent organizations in Colombia who monitor human rights violations. Through the data archived by one of these organizations, CINEP, we have created a spatially aggregated database of rebel human rights violations for validation purposes (CINEP 2008). CINEP is unlikely to exhibit many of the measurement problems that are common in (human- and machine-coded) event data sets. It has been documenting the conflict in Colombia for over 40 years; it has created a curated archive with an extensive collection of (Spanish language) national and regional Colombian newspapers and associated

[21] We provide additional details on these ontologies in relation to human rights violations, and news sources, further below and in the Appendix.

reports. These sources—which additionally include victim testimony, non-governmental organization reports, and government sources—are far more exhaustive in their coverage of potential Colombian human rights violations than international newswire reports. For these reasons, CINEP's records of rebel perpetrated human rights violations in Colombia are likely to be substantially more accurate than those of either ICEWS or GED, thus ensuring its role as an ideal external validation source in these regards (Bagozzi et al. 2019).

For our validation assessments, we aggregate the GED and ICEWS data on rebel perpetrated violence toward civilians for Colombian municipalities during the years 20XX-20XX. This time window is motivated by (1) past analyses of political violence in Colombia,[22] (2) data availability on our human, machine, and validation event data,[23] and (3) relevant political and social processes occurring within Colombia itself.[24] The specifics on the source actor categories, event types, and geo-location precision that we used to aggregate our GED and ICEWS data are explained in the appendix. We aggregate these Colombian GED and ICEWS events in two separate, parallel manners. First, we retain and aggregate all relevant instances of Colombian rebel violence against civilians for any and all primary sources recorded in GED and ICEWS. Note in this regard that GED and ICEWS overlap in the primary (e.g., newswire) sources that they record, but they are not identical in this regard. Second, we then separately retain and aggregate our rebel perpetrated violence events for these same two event datasets, but only for events identified according to the following four newswire sources–Associated Press, Agence France-Presse, Reuters, and Agencia EFE. These four newswire sources each have substantial coverage in ICEWS and GED for our period of

[22]For instance, Bagozzi et al. (2019) consider instances of rebel and paramilitary violence against civilians in Colombia during the years 2000–2009 whereas Lum et al. (2010) consider lethal violence in Colombia's Casanare department for the 1998–2007.

[23]The GED database reportedly underwent significant improvements about this time (Croicu and Sundberg, 2015: 14), and focusing on the post-2000 period allows us to avoid potential instability in the number of underlying sources coded within the ICEWS data.

[24]Colombia broke off three years of peace talks with the FARC in early 2002, leading to a sustained multi-year increase in violence, including a May 2002 rebel-perpetrated massacre of approximately 119 civilians in Bojayá. The year 2002 also marks the start of Álvaro Uribe's two four year terms as Colombia's President, and of the associated hard-line stance that the Colombian government took towards the FARC and ELN prior to the initiation of FARC peace talks by Juan Manuel Santos Calderón, Colombia's subsequent President.

interest thereby facilitating more controlled comparisons.

After combining the GED and ICEWS events described above into a pair of municipality-year datasets,[25] we next combined these data with our municipality-year CINEP (validation) data on rebel violence against civilians for use in external validation. We then merge a relevant set of of control variables to the combined municipality-year Colombia event data described above. These control variables are drawn from the same ICEWS and GED sources mentioned above, as well as from PRIO-GRID (Tollefsen, Strand, and Buhaug 2012). They are described briefly here, and are fully defined in the appendix....

Visual presentations of GED, ICEWS, and CINEP data via GEODA (as in Bagozzi et al. APSA ms. 2016). Discuss patterns in figures and visual evidence of remoteness problem. Could also present department level data and discuss the two types of aggregation problems here. Potentially introduce any standardization approach applied to the data here.

The SEMs are based on a model of rebel violence perpetrated towards Colombian civilians. There is no well accepted model of this kind. Alternative specifications–discuss.

### 2.1 Discrete SEMS for GED and ICEWS data

### 2.2 Continuous SEMS for GED and ICEWS data

## 3 Discussion

## 4 References

Adcock, R. and D. Collier. 2001. "Measurement validity: a shared standard for qualitative and quantitative research" *American Political Science Review* 95(3): 529-546.

Althaus, S., B. Peyton, and D. Shalmon. 2018. "Spatial and temporal dynamics of Boko Haram activity across six event data generation pipelines: testing a new approach to event data validation." Paper presented a the annual meeting of the American Political Science Association, Boston, MA.

Althaus, S., J. Bajjalieh, J. F. Carter, B. Peyton, and D. A. Shalmon. 2017. "Cline Center

[25] Municipalities are Colombia's second administrative unit, with 1,102 municipalities in total.

Historical Phoenix Event Data. v.1.0.0." Distributed by the Cline Center for Advanced Social Research. http://www.clinecenter.illinois.edu/data/event/phoenix/.

Anselin, L. 2006. "Spatial Econometrics." Chapter 26 in *Palgrave Handbook of Econometrics Volume 1 Econometric Theory* Basingstone: Palgrave MacMillan, pps. 901-969

Anselin, L. 1996. "The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association." Chapter 8 in *Spatial Analytical Perspectives on GIS.* M. Fisher, H. J Scholten, and D. Unwin eds. London, England: Taylor and Francis.

Ball, P. T. Guberek, D. Guzmánn, A. Hoover, and M. Lynch. 2008. "Asessing Claims of Declining Lethal Violence in Colombia." Working paper of the Human Rights Program of the Benetech Initiative.

Bagozzi, B., P. T. Brandt, J. R. Freeman, J. S. Holmes, A. Kim, A. Palao and C Potz-Nielsen. 2019 "The Prevalence and Severity of Underreporting Bias in Machine and Human Coded Data." *Political Science Research and Methods* 7(3): 641-649.

Bagozzi, B. , P. T. Brandt, J.R. Freeman, J.S. Holmes, A. Kim, and A. Palao. 2016. "External validation of event data." Paper presented at the annual meeting of the American Political Science Association, Philadelphia, PA.

Ball, P., J.Asher, D. Sulmont and D. Manrique. 2003. "How many Peruvians have died? An estimate of the total number of victims killed or disappeared in the armed internal conflict between 1980 and 2000." Washington, D.C.: American Association for the Advancement of Science.

Baller, Robert D. et al. 2001. "Structural Covariates of U.S. County Homocide Rates: Incorporating Spatial Effects." *Criminology* 39(3): 561-590.

Barros, Carlos Pestana, João Ricardo Faria, and Ari Francisco de Araujo, Jr. (2014) "Brazilian Land Tenure Conflicts: A Spatial Analysis" *Journal of International Development* 26: 409-421.

Beiler, John, Patrick T. Brandt, Andrew Halterman, Philip A. Schrodt, and Erin M. Simpson. 2016. "Generating Political Event Data in Near Real Time: Opportunities and Challenges." In *Computational Social Science: Discovery and Prediction* R. Michael Alvarez Editor. New York: Cambridge University Press.

Benoit, K.R.,D. Conway, B.E. Lauderdale, M. Laver, and S. Mikhaylov. 2016. "Crowd sourced text analysis: reproducible and agile production of political Data." *American Political Science Review.*

Berman, E., J.N. Shapiro, and J. H. Felter. 2011. "Can hearts and minds be bought? The economics of counterinsurgency in Iraq" *Journal of Political Economy* 119(4): 766-819.

Best, R. H., C.Carpino, and M.J.C.Crescenzi. 2013. "An analysis of the TABARI coding system." *Conflict Management and Peace Science* 39(4): 335-348.

Boshee, E., J. Lautenschalger, S. O'Brien, S. Shellman, J.Starz, and M. Ward. 2016. "ICEWS Coded Event Data." `http://dx.doi.org/10.7910/DVN/28075.HarvardDataverse.`

Brandt, P. T., J. R. Freeman, and P. A. Schrodt 2014 "Evaluating forecasts of political dynamics." *International Journal of Forecasting* 30(4): 944-962

Brandt, P. T., M. Colaresi, and J. R. Freeman 2008 "The dynamics of reciprocity, accountability, and credibility." *Journal of Conflict Resolution* 52(3): 343-374

Braybeck , B., W.D. Berry, and D. A. Siegel, 2011. "Strategic Theory of Policy Diffusion Via Intergovernmental Competition." *Journal of Politics* 73(1): 232-247

Brysk, A. 1994. "The politics of measurement: the contested count of disappeared in Argentina." *Human Rights Quarterly* 16: 676-692.

Bueno de Mesquita, E., C.C.Fair, J. Jordan, R.B. Rais, and J.N. Shapiro. 2015. "Measuring political violence in Pakistan: Insights from the BFRS dataset." *Conflict Management and Peace Science* 32(5): 536-558.

CINEP. 2008. "Marco Conceptual: Banco de Datos de Derechos Humanos y Violencia Política." Centro de Investigación y Educaión Popular.

Cigranelli, D. L. and T. E. Pasquarello. 1985. "Human rights practices and the distribution of U.S. foreign aid to Latin American Countries." *American Journal of Political Science* 29(3): 539-563.

Cho, Wendy K. Tam. 2003. "Contagion Effects and Ethnic Contribution Networks" *American Journal of Political Science* 47(2): 368-387.

Cho, Wendy K. Tam and James G. Gimpel. 2007. "Prospecting for (Campaign) Gold" *American Journal of Political Science* 51(2): 255-268.

Chyzh, O.V. and M.S. Kaiser 2019. "A Local Structure Graph Model: Modeling Formation of Network Edges as a Function of other Edges." *Political Analysis*.

Cook, Scott J. and Nils B. Weidman 2019. "Lost in Aggregation Improving Event Analysis With Report Level Data." *American Journal of Political Science* 63(1): 250-264.

Cook, S.J., B. Blas, R.J. Carroll, and S. Sinha. 2017. "Two wrongs dont make a right: addressing underreporting in binary data from multiple sources. *Political Analysis* 25(2): 223-240

Croicu, M. and R. Sundberg. 2015. UCDP Georeferenced Event Dataset Codebook. Version 18.1 (June 15)

Davenport, C. and P. Ball. 2002. "Views to a kill: Exploring the implications of source selection in the case of Guatemalan state terror, 1977-1995." *Journal of Conflict Resolution* 46(3): 427-450.

DeJuan, Alexander. 2013. "Long term Ecological Change and Geographical Patterns of Violence in Darfur 2033-2005." 2013. Paper presented at the Annual Meeting of the American Political Science Association, Chicago, August 29-September 1.

D'Orazio, V., J. E. Yonamine, and P. A. Schrodt. (2011). "Predicting intra-state conflict onset: an event data approach using euclidean and levenshtein distance measures." Paper Presented at the 69th Annual Midwest Political Science Association Meeting, Chicago, IL.

D'Ignazio, D. R. Bhargava, E. Zuckerman, and L. Beck (2014). "Cliff-clavin: Determining geographic focus for news." *NewsKDD: Data Science for News Publishing at KDD*.

Fariss, C.J. 2014. "Respect for human rights has improved over time: modeling the changing standard of accountability." *American Political Science Review* 108(2): 297-316.

Franzese, R.J., J.C.Hays, and S.J. Cook. 2016. "Spatial-and Spatialtemporal-Autoregressive Probit Models of Interdependent Binary Outcomes." *Political Science Research and Methods* 4(1): 151-173.

Grimmer, J. and B. M. Stewart. "Text as data: the promise and pitfalls of automated content analysis methods for political texts." *Political Analysis* 21(3): 267-297.

Guberek, T., D. Guzmán, M.Price, K. Lum, and P. Ball. 2010. "To count the uncounted: an estimation of lethal violence in Casanare." Benetech. Manuscript.

Guzmán, D. T. Guberek, A. Hoover, and P.Ball. 2007. "Missing people in Casanare." Unpublished manuscript.

Halterman, A. 2017. "Mordecai: full text geoparsing and event geocoding." *Journal of Open Source Software* 2(9), 91.

Halterman, A. 2019. "Geolocating political events in text." recarXiv: 1905.12713v1 [cs.CL] 29 May 2019.

Hammond, J. and N. B. Weidman. 2014. "Using machine coded event data for the micro level study of political violence." *Research and Politics* July-September: 1-8.

Hellmeier, S., N.B. Weidmann, and E.G. Rod (2018) "In the Spotlight: Analyzing Sequential Attention Effects in Protest Reporting." *Political Communication* 35: 587-611.

Holmes, J. S. and S. Amin Gutiérrez de Piñeres. 2014. "Violence and the state: Lessons from Colombia." *Small Wars and Insurgencies* 25(2): 372-403.

Holmes, J. S., S. Amin Gutiérrez de Piñeres, and K. M. Curtin. 2007. "A subnational study of insurgency: FARC violence in the 1990s." *Studies in Conflict & Terrorism* 30(3): 249-265.

King, G. 1989 "Variance specification in event count models: from restrictive assumptions to a generalized estimator." *American Journal of Political Science* 33(3): 762-784

King, G. 1988. "Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential poisson regression model" *American Journal of Political Science* 32: 838-863

King, G. and W. Lowe. 2004. "An automated information extraction tool for international conflict data with performance as good as human coders." *International Organization* 57(3): 617-642.

Krainski, Elias T., Finn Lindgren, Daniel Simpson and Håvard Rue. 2017. *The R-INLA tutorial on SPDE Models*

Leetaru, K. and Philip Schrodt. 2013. "GDELT: Global data on events, language and tone, 1979-2002." Presented at the Annual Meeting of the International Studies Association, San Francisco, California.

Lee, S. J., H. Liu and M. D. Ward. 2018. "Lost in space: geolocation in event data." *Political Science Research and Methods* 1-18.

Lindgren, Finn and Håvard Rue. 2015. "Bayesian spatial and spatio-temporal modeling with R-INLA." *Journal of Statistical Software* 63(19).

Lindgren, Finn, Håvard Rue, and J. Lindström. 2011. "An Explicit Link Between Gaussian Fields and Gausian Markov Random Fields: The Stochastic Partial Differential Equation Approach (with discussion)." *Journal of the Royal Statistical Society* Series B 73(4): 423-498.

Lowe, W. and K.R. Benoit. 2013. "Validating estimates of latent traits from textual data using human judgment as a benchmark." *Political Analysis* 21(): 298-313.

Lum, K., M. Price, T. Guberek, and P. Ball. 2010. "Measuring Elusive Populations with Bayesian Model Averaging for Multiple Systems Estimation: A Case Study of Lethal Violations in Casanare, 1998-2007." *Statistics, Politics, and Policy* 1(1).

Martinez, L. R. 2017. "Transnational insurgents: evidence from Colombia's FARC at the border with Chaávez's Venezuela." *Journal of Development Economics* 126: 138-153

Mikhaylov, S., M. Laver, and K.R. Benoit. 2012. "Coder reliability and misclassification in human coding of party manifestos" *Political Analysis* 20(): 78-91.

Monogan III, J.E. and J. Gill (2015) "Measuring State and District Level Ideology with Spatial Realignment" *Political Science Research and Methods* 4(1): 97-121.

Nardulli, P.F., A. Singh, M. Martin, D. Shalmon, B. Peyton, and S. Althaus (2019) "Social Political Event Data Set (SPEED): Liberia, Philippines, Sierra Leone (1979-2008) Codebook. v1.0.0. Champaign, Illinois. The Cline Center.

Nardulli, P. F., S. L. Althaus, and M. Hayes. 2013. "A progressive supervised learning approach to generating rich civil strife data." *Sociological Methodology* 45(1): 148-183.

Norris, C., P. A. Schrodt, and J. Beieler. 2012. "PETRARCH-2: another event coding program." *The Journal of Open Source Software*, 9(2).

OEDA. 2016. "Real Time Event Data/Phoenix." Available at http://eventdata.utdallas.edu/.

Osorio, J. 2015. "The contagion of drug violence: Spatiotemporal dynamics of the Mexican war on drugs." *Journal of Conflict Resolution* 59(8): 1403-1432.

PITF. 2009. "Political instability task force worldwide atrocities event data collection codebook version 1.0B2." Version pitf.world.19950101-20130930.csv.

Python, André, Janine B. Illian, Charlotte M. Jones-Todd, and Marta Blangiardo. 2018. "A Bayesian approach to modeling subnational spatial dynamics of worldwide non-state terrorism, 2010-2016" *Journal of the Royal Statistical Society Series* A. 1-22.

Python, André, Janine Illian, Charlotte Jones-Todd, and Marta Blangiardo. 2017. "Explaining the lethality of Boko Haram's terrorist attacks in Nigeria 2009-2014: A hierarchical Bayesian approach." In *Bayesian Statistics in Action* R. Argiento et al. eds.Springer.

Raleigh, C., A. Linke, H. Hegre, and J. Karlsen. 2010. "Introducing ACLED: An Armed Conflict Location and Event Dataset." *Journal of Peace Research* 47(5): 651-660.

Raytheon BBN Technologies. 2015. "BBN Accent Event Coding Evaluation."

Restrepo, J.A., M. Sagat, and Juan F. Vargas. 2006. "The severity of the Colombian conflict: cross-country data sets versus new micro-data" *Journal of Peace Research* 43(1): 99-115.

Richani, N. 2013. *Systems of violence: The political economy of war and peace in Colombia* Second Edition. SUNY Press.

Rosen, Jonathan D. 2014. *The Losing War: Plan Colombia and Beyond* NY. SUNY Press.

Salehyan, I., C.S. Hendrix, J.Hamner, C. Case, C.Linebarger, Emily Stull and J.Williams. 2012. "Social Conflict in Africa: A New Database" *International Interactions* 38: 503-511.

Schrodt, P.A. 2015. "Comparison metrics for large scale political event data sets." Paper presented at the Text as Data Meeting. NY. New York University, October 16-17.

Schrodt, P.A. and D. Van Brackle. 2013. "Automated coding of political event data." In *Handbook of Computational Approaches to Counterterroism*, V.S.Subrahmanian editor.

NY: Springer Press.

Schrodt, P. and D. Gerner. 1994. "Validity assessment of machine-coded event data set for the middle east, 1982-1992." *American Journal of Political Science* 38: 825-854.

Schutte, S. and Karsten Donnay 2014. "Matched wake analysis: finding causal relationships in spatio-temporal event data." *Political Geography* 41: 1-10

Shirk, D. and J. Wallman. 2015. "Understanding Mexico's drug violence." *Journal of Conflict Resolution* 59(8): 134-1376.

Sundberg, R. and E. Melander. 2013. "Introducing the UCDP georeferenced event dataset." *Journal of Peace Research* 50(4): 523-532.

Tollefsen, A. F., H. Strand, and H. Buhaug. 2012. "PRIO-GRID: A unified spatial data structure.." *Journal of Peace Research* 49(2): 363-374.

von Borzyskowski, I. and M. Wahman. forthcoming. "Systematic measurement Error in election violence data: causes and consequences" *British Journal of Political Science*

Wang, W., R. Kennedy, D. Lazar, and N. Ramakrishnan. 2016. "Growing Pains for Global Monitoring of Societal Events" *Science* 353(6307): 1502-1503.

Ward, M. D., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford. 2013. "Comparing GDELT and ICEWS event data," Working Paper. Available at: `https://benradford.github.io/images/publications/GDELTICEWS.pdf`.

Ward, M.D. and K. S. Gleditsch 2002. "Location, location, location: a MCMC approach to modeling the spatial context of war and peace" *Political Analysis* 10(3): 244-260

Ward, M.D. and K. S. Gleditsch. 2019. *Spatial Regression Models* Second edition. Sage.

Weidman, N. B. and M. Ward 2010. "Predicting Conflict in Space and Time" *Journal of Conflict Resolution* 54(6): 883-901.

Weidman, N.B. 2016. "A closer look at reporting bias in conflict event data." *American Journal of Political Science* 60(1): 206-218.

Weidman, N. B. 2015. "On the accuracy of media based conflict event data." *Journal of Conflict Resolution* 59(6): 1129-1149.

Wilkerson, J. and A. Casas. 2017. "Large-scale computerized text analysis in political science: opportunities and challenges" *Annual Review of Political Science* 20: 529-544.

Woolley, J.T. 2000. "Using media-based data in studies of politics." *American Journal of Political Science* 156-173.

Zammit-Mangion, A., M. Dewar, V. Kadiramanathan, and G. Sanguinetti. 2012. "Point process modelling of the Afghan war diary." *PNAS* 109(31): 12414-1224

# 5 Appendix

## 5.1 Past Event Data Validation Designs

Seminal works on validation of event data either ignore external validation or they do not offer a methodological framework to perform external validation. King and Lowe (2003: 619, 624) stress the importance of "independent evaluation" of event data which, in their case, is defined as validation by humans who did not develop the machine coding software.

They have the VRA Reader code 45,000 articles. King and Lowe assign the events in those articles to 157 bins, bins corresponding to the IDEA ontology. The bins then are divided into three groups: (i) bins containing at least five events, (ii) bins containing one to four events, and (iii) bins which are empty. To construct their internal validation test bed, King and Lowe randomly choose five events from each bin in the first group and all of the events in the bins in the second group. They also include in their test bed, twenty five randomly chosen events from among those for which the VRA Reader assigned a source and/or target but could not assign an IDEA category (King and Lowe 2003: 626-627). King and Lowe compare the codings by a handful of experts with those produced by one software routine and three undergraduates.[26] In their comparisons, King and Lowe ignore source-target information in the leads, focusing instead on the assignment of events or leads to the final event-categories.

Schrodt and Gerner (1994) evaluate the validity of machine and human coded events data. Their analysis of internal validity is based on the correlations between their machine coded events and those produced by human coders independently at the U.S. Naval Academy.[27] As for external validity, Schrodt and Gerner (1994) compare their machine coded data to death counts tabulated by the Palestinian Human Rights Information Center and to "critical shifts" in historical narratives of Israeli-Lebanese and Israeli-Syrian relations. King and Lowe (2003) and Schrodt and Gerner (1994) both recognize that aggregation (temporal, event class, etc.) affects validation. King and Lowe (2003) point out that higher aggregations across event class and time leads to better validation results. They internally validate with the IDEA and WEIS event code levels aggregated to the more general cue categories; they do this before examining how these comparisons vary across the conflict-cooperation dimension. Both King and Lowe's and Schrodt and Gerner's studies are of limited dimensionality. They compare human coding of a single body of text with the coding produced by a single piece of software.

The validation exercises in King and Lowe (2003) and in Schrodt and Gerner (1994) are illustrated in Figures 1 and 2. As these schematics show, internal and external validation are distinct concepts. And the texts that are used to validate event event data can be several levels removed from ground truth.[28]

## 5.2 Formatting Decisions for Colombia Data

We aggregate and combine GED, ICEWS, and CINEP (validation) data on Colombian human rights violations at the municipality-year level for the years 20xx-20xx. Due to distinct processes of geolocation and event coding, CINEP, ICEWS, and GED each exhibit different levels of spatio-temporal precision, have unique definitions of what ultimately comprises a human right violation event, and contain varying levels of specificity regarding the identities

[26]Based upon this approach, King and Lowe have 12 bins with no machine coded events/leads, and thus cannot provide any leads of these event types to their coders, although they do include a sample of leads that were not classified by the machine.

[27]Schrodt reports in personal communication that he does not know the design that was used to produce the human coded events for this comparison.

[28]Recent important articles on the production of political text by news sources include Cook and Weidman (2019) and Hellmeier et al. (2018). For a still more complex schematic of the sources behind newspaper reports of human rights violations see Davenport and Ball (2002).

Figure 1: **Validation in King and Lowe (2003)**

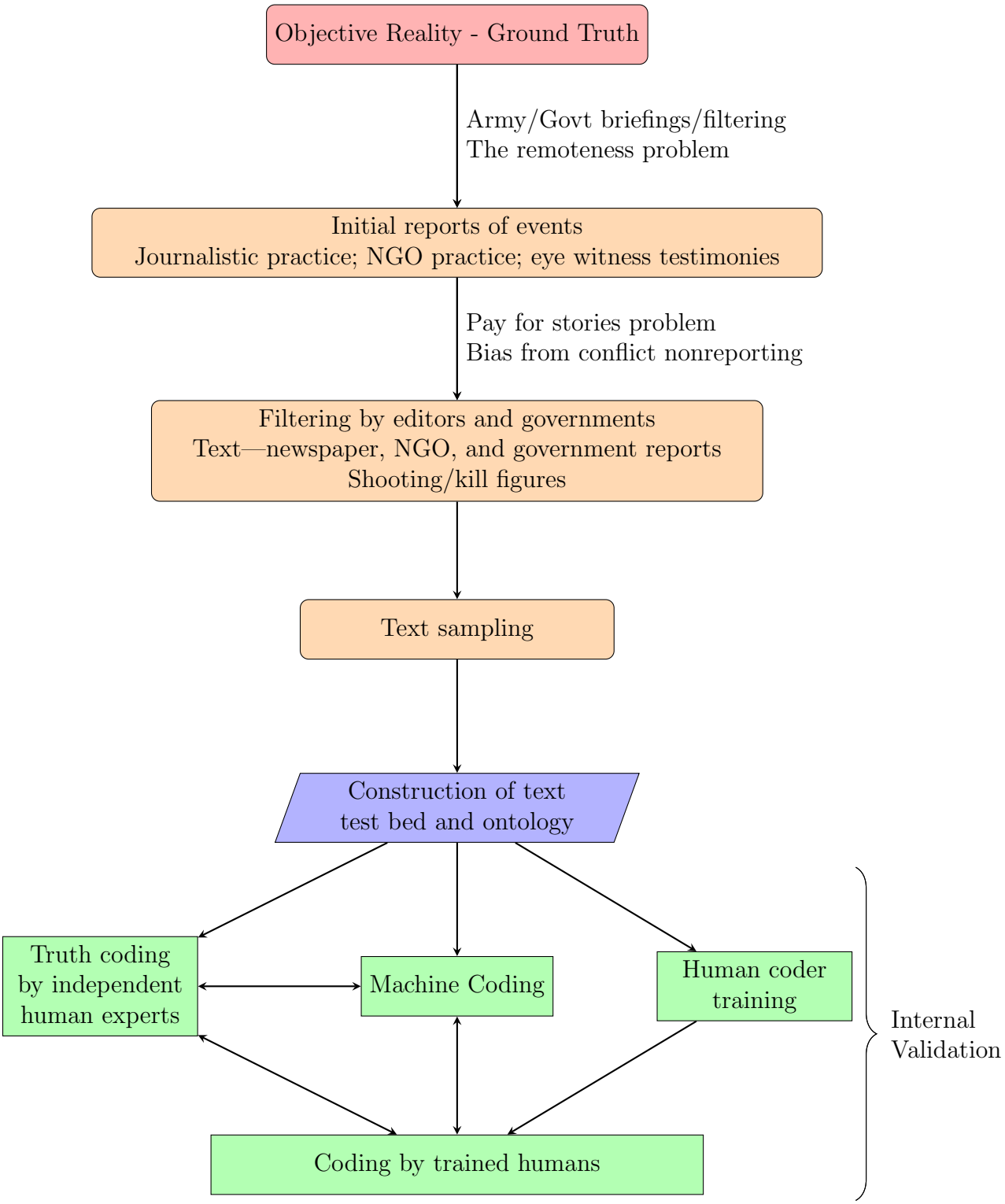
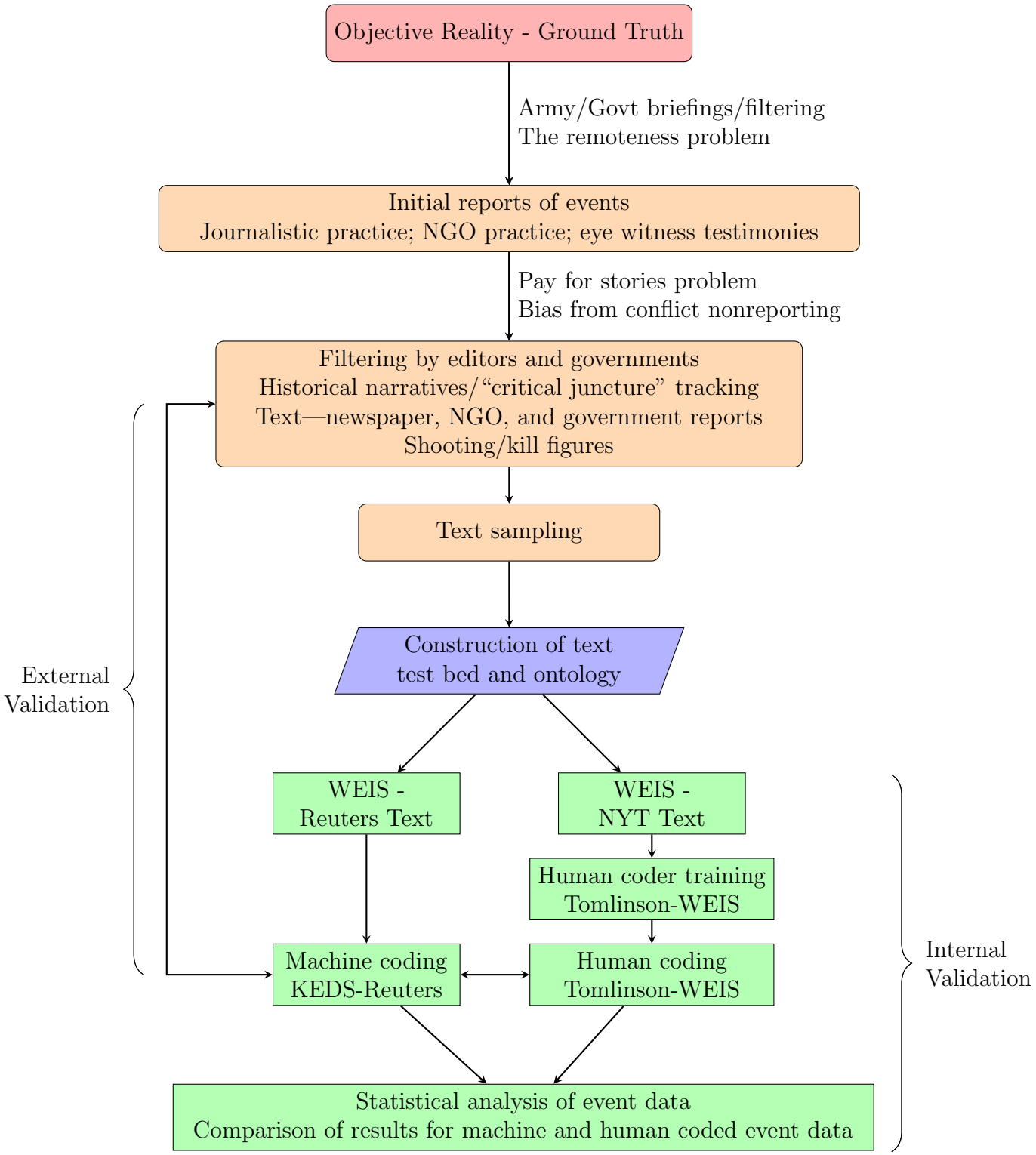


Figure 2: **Validation in Schrodt and Gerner (1994)**



of violence perpetrators and victims. These differences necessitate several important assumptions and decisions when spatially aggregating and combining these datasets for comparison.

We first formatted our (machine coded) ICEWS data (Boshee et al. 2016) to correspond as closely as possible to rebel perpetrated instances of material violence against civilians. As an initial step, this required that we identify and retain only those ICEWS events that were perpetrated by rebel actors in Colombia against civilian targets. To identify such events within the ICEWS data, we used the source and target actor designations contained within the ICEWS data to identify relevant actors. Herein, we treated actors designated as "rebel," "separatist," "insurgent," and "unidentified sources" as source actors, and those designated as "general population," "civilian," and "social" as our target actors. We next subset our identified rebel→civilian events to only include those events occurring in Colombia during the years 20XX-20XX that were geo-located to the city/town *or* municipality level(s) of geographic precision. After identifying events occurring in Colombian municipalities between these source and target actors based upon the above criteria, we retained only CAMEO category 18 (ASSAULT) and CAMEO category 20 (USE UNCONVENTIONAL MASS VIOLENCE) events with the following three or four digit CAMEO codes:

180: Use unconventional violence, not specified below
181: Abduct, hijack, or take hostage
182: Physically assault, not specified below
1821: Sexually assault
1822: Torture
1823: Kill by physical assault
183: Conduct suicide, car, or other non-military bombing, not specified below
1831: Carry out suicide bombing
1832: Carry out car bombing
1833: Carry out roadside bombing
184: Use as human shield
185: Attempt to assassinate
186: Assassinate
200: Use unconventional mass violence, not specified below
201: Engage in mass explusion
202: Engage in mass killings
203: Engage in ethnic cleansing

The above steps produced a set of all Colombian human rights violation events involving rebel actors and civilian targets that were coded to a municipality or city level of geographic precision for years 20xx-20xx. With these individual events in hand, we next turned to aggregating all remaining events. To do so, we first created two separate parallel versions of all remaining ICEWS rebel→civilian material violence events. The first version of these retained events included all such ICEWS events for Colombia, no matter the underlying news or newswire source used by ICEWS in the coding of each event. In the second version, we retained *only* those events that were coded by ICEWS from the following four newswire sources: Associated Press, Agence France-Presse, Reuters, and Agencia EFE. As noted in the main paper, the latter sample allows for more controlled comparisons to the GED data

that we aggregate below. Before aggregating each set of ICEWS events to the municipality-year level, we next applied a de-duplication criterion to ensure that only one event(-type) was recorded per day, source, and latitude-longitude coordinate. We implemented this step because ICEWS only does very mild de-duplication at the coding stage—effectively eliminating duplicate stories bearing the same publisher, headline, and date—but it still allows for some duplicate stories given, for example, variation in headlines, which can lead to over-reporting of many events (Schrodt, 2015: 14). We then aggregated that our deduplicated ICEWS rebel violence events to municipality-year counts by matching these events to maps of Colombia's municipalities via latitude longitude coordinates.

We formatted the GED (Sundberg and Melander 2013) in a comparable manner to the ICEWS data. The GED is a (near-global) human-coded event dataset that draws on both news(wire) sources and non-governmental organization reports for its coding of individual events. We started by subsetting the GED to encompass only Colombia-based events for the years 20XX-20XX. For these Colombian events, we next retained all nonstate perpetrated cases of violence against civilians (i.e., "one-sided violence") within the GED data, while taking care to exclude any instances of violence against civilians that were perpetrated explicitly by Colombian drug cartels, the Colombian military or police, and government-affiliated militia groups. We then split these events into two parallel versions of the GED that either recorded (i) all events based upon any source coded for Colombia by the GED and (ii) the GED instances of rebel violence against civilians that were coded from stories appearing in the same four newswire sources mentioned for ICEWS above (Associated Press, Agence France-Presse, Reuters, and Agencia EFE). The latter version of our subsetted GED—as was the case for our ICEWS data—only retains a small subset of the total instances of rebel-perpetrated violence against civilians included in the GED. With this caveat in mind, our two GED subsets were then aggregated and merged to Colombian municipality-year templates for our period of interest, while taking care to omit any GED events whose levels of geocoding accuracy were determined to be too ambiguous to fit within the municipality administrative level.[29] After these formatting and aggregation tasks were complete, we combined all the GED measures with the ICEWS measures described above.

Finally, we aggregated and merged our CINEP validation data (CINEP 2008) to the aforementioned municipality-year template. The CINEP data are originally stored at the event level, with information attached to each event for that particular event's perpetrator (source actor), year, and municipality—among other variables. The recorded events include only directed rebel (source) to citizen (target) violence events. Directed dyad interactions of this sort (1) facilitate the comparison of the events with the directed dyadic event information contained in ICEWS and GED, and (2) ensure that our analysis closely parallels the most common approach to event data coding and analysis within the discipline (i.e., dyadic relational interactions). Within CINEP's data, *source* actors are designated by the specific rebel group perpetrating a given human rights violation and the *target* of each event is inferred to be a civilian or group of civilians. To combine these data with our formatted ICEWS and GED events, we first collapse CINEP's recorded rebel-perpetrated HRV events to the unique event-ID level. We then subset CINEP's events to only include actual in-

[29]Specifically, we only retained events that GED indicated were either (i) geolocated within 25km of a known location or (ii) whose exact location was recorded with latitude and longitude coordinates.

stances of "material" human rights violations, rather than both material and verbal human rights violations.[30] After preprocessing our CINEP data in these manners, we aggregated all remaining events to the municipality-year level for only our years of interest, and merged these cases to our final GED and ICEWS data.

## 5.3 Extended Discussion of Predictors

As discussed in the main paper, we include a wide array of theoretically motivated predictor variables within our SEMs and SPDE models. This subsection provides an expanded discussion of the sources and operationalizations of each of these control variables.

A first set of predictor variables were coded from the same ICEWS and GED data described above. Here we created separate measures of government→civilian material violence and government→rebel material violence as such violence may affect rebels decisions to initiate violence against civilians. These government→civilian and government→rebel predictors are aggregated in the exact same manners as our primary rebel→civilian outcome variables. Specifically we, construct separate versions of these measures for (i) all relevant events coded from any source used in GED or ICEWS and (ii) only those relevant (ICEWS or GED) events that were coded from Associated Press, Agence France-Presse, Reuters, and Agencia EFE. These events are then aggregated to the municipality-year levels for the years noted above.

The subsetting of government→civilian material violence and government→rebel material violence in GED is straightforward. For GED we used the same rebel and civilian actor criteria that was described for GED above for these two actors. For government-initiated material violence events, we subset the GED data to include only violence events with government sources based upon GED's "Government" source actor designation. For ICEWS, rebel and civilian actors were again defined as above. Government source actors were defined as any actor whose source actor designation was recorded as "military," "government," "police," or the Colombian nation-state itself within the original ICEWS data.

[30] That is, we remove all non-material violence events (e.g., threats), including categories such as 'Threatens', 'Recruitment', and 'Collective Threats,' which altogether constituted 75% of all rebel-perpetrated violence events in CINEP for our years of analysis.