PA

# Human Rights Violations in Space: Assessing the External Validity of Machine-Geocoded versus Human-Geocoded Data

**Logan Stundal**[1], **Benjamin E. Bagozzi**[2], **John R. Freeman**[3] and **Jennifer S. Holmes**[4]

[1] *Department of Political Science, University of Minnesota, Minneapolis, MN, USA. E-mail: stund005@umn.edu*
[2] *Department of Political Science & IR, University of Delaware, Newark, DE, USA. E-mail: bagozzib@udel.edu*
[3] *Department of Political Science, University of Minnesota, Minneapolis, MN, USA. E-mail: freeman@umn.edu*
[4] *School of Economic, Political & Policy Sciences, UT-Dallas, Richardson, TX, USA. E-mail: jholmes@utdallas.edu*

## Abstract

Political event data are widely used in studies of political violence. 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. We analyze the frequencies of human- and machine-geocoded event data agreement in relation to an independent (ground truth) source. The events are human rights violations in Colombia. We perform our evaluation for a key, 8-year period of the Colombian conflict and in three 2-year subperiods as well as for a selected set of (non)journalistically remote municipalities. As a complement to this analysis, we estimate spatial probit models based on the three datasets. These models assume Gaussian Markov Random Field error processes; they are constructed using a stochastic partial differential equation and estimated with integrated nested Laplacian approximation. The estimated models tell us whether the three datasets produce comparable predictions, underreport events in relation to the same covariates, and have similar patterns of prediction error. Together the two analyses show that, for this subnational conflict, the machine- and human-geocoded datasets are comparable in terms of external validity but, according to the geostatistical models, produce prediction errors that differ in important respects.

*Keywords:* Event data, geocoding, spatial analysis, spatial regression, machine coding, external validity, human rights violations

## 1 Introduction

Scholars agree that text is a valuable source of political data (Grimmer and Stewart 2013; Wilkerson and Casas 2017). They also agree that machines are potentially 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). Using machines for event data coding, we therefore should be able to make significant progress in conducting subnational (micro level) analyses of terrorism and other forms of conflict.

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 regard 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.

***Political Analysis (2021)***

**DOI:** 10.1017/pan.2021.40

**Corresponding author**
Logan Stundal

**Edited by**
Xun Pang

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).[1] 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 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 Althaus, Peyton, and Shalmon 2021). As regards external validity, several researchers found a *remoteness problem* in human- and/or machine-coded event datasets. Due to journalistic practice and other factors, human—and especially 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 Weidmann 2014; Weidmann 2016).

Two research designs are used in extant evaluations. In one, researchers use experts to establish what is assumed to be valid codings. These experts—often project managers—then train a group of individuals to code text. Individuals' error rates are gauged in a pilot study. Then a text corpus is assembled and trainee coding is compared to the coding by a particular piece of software (Althaus, Peyton, and Shalmon 2021; 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. Confusion matrices and related tools are primarily used within such validation efforts.[2]

There are three problems with this design. First, 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 GED, for example, admits that human coders often lack this knowledge (Croicu and Sundberg 2015, 14). Moreover, in these cases, external validity often is not assessed. Neither the human (trainees) nor the machine coding is compared to an independent source.[3] Second, confusion matrices and tools used for the analysis of raw data do not tell us *where* coding errors are located, for instance, if they cluster far away from major cities (the remoteness problem). As such, these tools do not tell us *where* errors in the predictions made on the basis of each kind of data are most prevalent, and where the errors cluster spatially. A third problem is that such evaluations usually are based on a relatively small sample of events. The datasets we employ in this article allow for more than 8,000 comparisons of ground truth versus machine- and human-geocoded human (HRVs). Making this number of comparisons is not feasible by hand. And, again, by hand evaluations of a subset of complete datasets do not reveal statistically meaningful patterns synonymous with the remoteness problem and related sources of error.

The other validation design asks if the inferences based on the estimates from statistical models of human- and machine-coded data are the same *and* if the two event datasets 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. More recently, Bagozzi *et al.* (2019) used Cook *et al.*'s (2017) binary misclassification model to gauge underreporting bias in the human-coded GED and the widely used machine-coded Integrated Crisis Early Warning System dataset (ICEWS; Boschee *et al.* 2016). Bagozzi *et al.* (2019) found remarkable similarities in the patterns and statistical significance of the coefficients in models of machine- and human-coded event data as well as in the coefficients in the auxiliary

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

2 A more extensive description of this approach to establishing external validity can be found in the Supplementary Appendix.

3 Althaus *et al.* compare human- and machine-derived events coded for the same corpus of reports about Boko Haram in Nigeria. Raytheon BBN Technologies (2015) is an assessment of the human coding versus the Integrated Crisis Early Warning System (ICEWS) machine coding of a comparable corpus. The former study includes some assessment of geocoding. The latter does not. We also note that the idea of using independently collected data to gauge the external validity of event datasets is becoming more common. See, for instance, Zammit-Mangion *et al.* (2012); Weidmann (2016), and von Borzyskowski and Wahman (2021).

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 HRVs. Neither Schrodt and Gerner (1994) nor Bagozzi *et al.* (2019) explicitly modeled geolocation accuracy, however. Hence, they provide few insights into the usefulness of machine-coded data for investigating subnational patterns of conflict.

Both research designs are employed in this article. First, the frequency of agreement between reports of human right violations by an independent source and reports in the machine-geocoded and human-geocoded datasets are each compared for 8- and 2-year periods of conflict and for a small cluster of (non)remote units. Then, as an important complement to this evaluation, we fit geostatistical spatial error models (SEMs) of HRVs for each dataset's full set of events (over one thousand units over an eight year period and also for several subperiods). We employ SEMs because, conceptually, they are best suited to address the external validity issue (Anselin 1996, 907; Ward and Gleditsch 2019, 76). Geostatistical spatial error models are employed rather than neighborhood spatial error models because, as we explain below, the former are more informative than the latter. In particular, the most useful insights into the validity of the datasets are produced by spatial probit models for which the errors are assumed to be Gaussian Markov Random Fields (GMRFs). These models are estimated by means of integrated nested Laplacian approximation (INLA). They yield estimates of the *range* of spatial error dependence as well as *site specific* information about the mean and variance of the errors in machine-geocoded and human-geocoded events.[4]

Our test beds are cross-sections[5] of subnational data on FARC-perpetrated acts of violence against civilians in Colombia during the period 2002–2009, as coded by humans for GED and by machines via ICEWS. Both the GED and ICEWS data are widely used in the study of conflict.[6] In addition, ICEWS is considered one of the most accurate machine-coded datasets currently available (D'Orazio, Yonamine, and Schrodt 2011, 4). GED has likewise been shown to have superior geolocation accuracy compared to other prominent human-coded datasets (Eck 2012). The independently collected data from the Centro de Investigación y Educación Popular (CINEP) is used to assess the external validity of the GED and ICEWS data.

We find the external validity of our machine-geocoded (ICEWS) data compares favorably to that of our human-geocoded (GED) dataset. Predictions of FARC–HRV events based on these machine-geocoded and human-geocoded datasets depend on the same covariates—covariates that are, in some respects, indicative of the remoteness problem. These same covariates explain underreporting in the two datasets. Our geostatistical analysis based on these covariates reveals some differences in the spatial dependence of the respective model errors. For example, the ranges of spatial error dependence and marginal variance of these errors differ for models based on the ICEWS- and GED-based datasets. But in terms of predictive accuracy and other metrics, the two models are comparable. In delving deeper into these findings, some differences in the pattern of spatial error dependence is found for models of subperiods of the Colombia conflict. We attribute these differences to, among other things, the number of news sources on which ICEWS

4 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 Gill (2021). 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). Weidmann 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.

5 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, for example, DeJuan (2013).

6 A 2021 Google search for "Integrated Crisis Early Warning System" and "Georeferenced Event Database" produced approximately 4,100 and 5,200 results, respectively.

and GED are based. But overall then there is little evidence that, in our case, machine-geocoded data are less externally valid than human-geocoded data.

The ensuing discussion is divided in to three parts. In the next section, we discuss two approaches to external validity assessments, and especially the use of geostatistical models in external validity assessment. Part three explains the Colombian testbed, and presents our results. Part four outlines directions for future research including the need to evaluate additional human- and machine-coded datasets, and how to incorporate more complex kinds of spatial evaluation in our validity assessments.

## 2 External Validation of Event Data

Both human- and machine-coded event datasets primarily code events (in terms of who did what to whom, and where/when) from international news(wire) reports. The Supplementary Appendix reviews the current state-of-the-art for geolocation methods used within human- and machine-coded political event data. As noted above, the Supplementary Appendix also contains a description of how researchers employ experts to assess the validity of subsamples of these codings. Once more, spatial data analysis is a valuable complement to these latter assessments. Spatial data analysis enables researchers to evaluate the validity of complete sets of machine- and human-geocoded data, and, in the process, to illuminate statistically meaningful correlates of geocoding and, concomitantly, of underreporting.

There are two forms of spatial data analysis. The first is the familiar "neighborhood approach" (Anselin 1996). It analyzes spatial dependence in a variable between *discrete* units like census tracts, municipalities, electoral districts, provinces, and countries.[7] A connectivity matrix is *pre-specified* for neighborhood models indicating presumed relationships between the values of a variable in certain units or between the model errors of certain units. The models may contain spatial and(or) aspatial covariates (Cho 2003). Conceptually, the SEM is the best suited "neighborhood approach" for answering our questions about the validity of human- and machine-coded event data. SEMs capture model errors for neighboring units that cluster together—"smaller(larger) errors for observation $i$ ... go together with smaller [larger] errors for [neighbor] $j$" (Ward and Gleditsch 2019, 76). Errors also may be correlated because of the mismatch between the spatial scale of a process and the discrete spatial units of observations (Anselin 2006, 907). These error patterns correspond respectively to what researchers call the remoteness problem. 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. An example, is the spatial probit error model (SPEM). A technique based on the conditional log likelihood and variance–covariance matrix of the model can be used to estimate it (Martinelli and Geniaux 2017). The model provides estimates of a $\lambda$ parameter which, with a row-standardized connectivity matrix, indicates the average dependence in the errors of a prespecified set of neighbors on the estimation error in a unit of interest.

For several reasons neighborhood models like the SPEM are not well suited to externally validate our machine- and human-geocoded events. To begin, neighborhood models may suffer from "inappropriate discretization" (Lindgren and Rue 2015, 3). The prespecified connectivity matrices used in neighborhood models treats spatial dependence of errors as a step function—the same for some subset of units and nonexistant for another subset of units. In reality, the spatial dependence of errors at different sites may vary *continuously* in space. In addition, the $\lambda$ parameter produced by neighborhood error models is difficult to interpret; it does not tell us about the impacts of spatial error dependence at specific sites. And it is difficult to infer

7 That is, random aggregate values over areal units or a lattice with well defined boundaries; a countable collection of spatial units (Blangiardo and Cameletti 2015, 173); see also (Weidmann and Ward 2010, 884).

these impacts from the respective model estimates. The alternative—geostatistical spatial error models—produce *estimates* of both the range of spatial error dependence and of the site specific impact of unobserved factors including measurement errors on the unit of interest. For these reasons, we fit SPEM models for the CINEP, ICEWS, and GED datasets, but relegate the discussion of the respective results to the Supplementary Appendix.

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 are interpreted as realizations of a continuously indexed space-time process (Python *et al.* 2017, 2018).[8] One geostatistical approach to analyzing these kinds of data are Continuous Domain Bayesian Modeling with INLA.[9] 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, $(s_1, \ldots, s_m)$. As such it suffers from a "big n problem"; analyzing the Gaussian field is costly computationally (Lindgren, Rue, and Lindström 2011). Therefore, a particular linear, stochastic partial differential equation (SPDE) is assumed to apply to the Gaussian field:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}}(\tau\xi(s)) = W(s), \qquad s \in D, \tag{1}$$

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:

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

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 solution to the SPDE implies that the formula for the marginal variance is $\sigma^2 = \frac{\Gamma(\lambda)}{\Gamma(\alpha)(4\pi)^{\frac{d}{2}}\kappa^{2\lambda}\tau^2}$ where $d = 2(\alpha - \lambda)$. And the formula for the range is $r = \frac{\sqrt{8\lambda}}{\kappa}$. In this way, the Gaussian field can be represented (approximated) by a GMRF. A finite element method using basis functions defined on a Constrained Refined Delaunay Triangularization (mesh) over a corresponding shapefile of latitude–longitude event data is used for this purpose.

A hierarchical Bayesian framework can be used to model the data. For dichotomous data like the discrete observation of a human rights violation, three equations are employed:

$$y_i \mid \eta_i, \theta \sim Bernoulli(\pi_i), \qquad i = 1, \ldots, m, \tag{3}$$

$$\eta_i \mid \theta = \beta_0 + \sum_{k=1}^{n_\beta} \beta_k z_{k,i} + \xi_i, \qquad i = 1, \ldots, m, \tag{4}$$

$$\theta \sim p(\theta), \tag{5}$$

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

9 The following description draws from Blangiardo and Cameletti (2015, Chap. 6) and especially the passage on pp. 234–235 of Python *et al.* (2017).

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, here $\eta_i = probit(\pi)$, with spatially explicit covariates $z_{k,i}$, $\xi_i$ the Gaussian field as defined by equations (1) and (2) and approximated by the GMRF at point $i$, and equation (1) assigns the hyperparameters $\theta = (\kappa, \sigma_\xi^2)$.[10]

INLA is used to estimate the model. INLA performs numerical calculation of posterior densities and in this regard it is more efficient than Markov Chain Monte Carlo methods. Besides estimates of the effects of spatially explicit covariates on the probability of events and of the range of spatial error dependence for each dataset, this geostatistical approach produces useful estimates of the parameters in the GMRF—in particular, the mean and standard deviation of the latent field at each point in the dataset. These estimates tell us about the impact of uncertainty produced by both the scarcity (absence) of data and measurement error at each site. The estimates of the GMRF therefore tell us how human- and machine-coded data compare in terms of the remoteness problem.[11]

## 3 External Validation of Geolocated FARC HRVs in Colombia

### 3.1 The Testbed

Our external validation testbed corresponds to Colombia.[12] With a domestic insurgency that has now spanned over five decades, Colombia has been the scene of an egregious number of HRVs. These violations have been charted at the subnational level by numerous researchers and nongovernmental organizations (e.g., Holmes, de Piñeres, and Curtin 2007; Guberek *et al.* 2010; Lum *et al.* 2010; Bagozzi *et al.* 2019). In keeping with much of this past research, our HRV validation efforts focus on rebel—and specifically, FARC—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 FARC perpetrated HRVs against civilian targets using similar ontologies, and they use many of the same news source(s) to code events.[13]

There are also several independent organizations in Colombia who monitor HRVs. Through the data archived by one of these organizations, CINEP, we created a spatially aggregated database of FARC-directed HRVs for validation purposes (CINEP, 2008). CINEP is unlikely to exhibit many of the measurement problems that are common in (human- and machine-coded) event datasets. It has been documenting the FARC conflict in Colombia for over 40 years; it has created an extensive archive of (Spanish language) national and regional Colombian newspapers and associated reports. These sources—which additionally include victim testimony, nongovernmental organization reports, and government sources—are far more exhaustive in their coverage of potential Colombian HRVs than international newswire reports. For these reasons, CINEP's records of FARC perpetrated HRVs in Colombia are likely to be substantially more accurate than those of either ICEWS or GED, thus making it an ideal external validation source (Bagozzi *et al.* 2019).

For our validation assessments, we aggregate the CINEP, GED, and ICEWS data on FARC perpetrated violence toward civilians (hereafter, HRVs) for Colombian municipalities during the years 2002–2009. Details on the source/target actor categories, event types, and geolocation precision designations used in aggregating our GED, ICEWS, and CINEP data are explained in the Supplementary Appendix. Our choice of a 2002–2009 time window for analysis is motivated by three factors. First, past analyses of political violence in Colombia tend to focus on this period

10 For comparability to the SPEM, we employ the probit link function in our geostatistical evaluation of the ICEWS and GED data. In the Supplementary Appendix, we also report substantively comparable results for standard probit model with no spatial error component.

11 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.* (2018, 7–8). We return to these more complex geostatistical models in the Section 4.

12 The data used in this analysis are available on the Political Analysis Dataverse (Stundal 2021).

13 We provide additional details on these datasets' ontologies in relation to HRVs, past validation efforts, and news sources, further below and in the Supplementary Appendix.

of civil conflict in Colombia. 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 1998–2007. Second, data availability on our human, machine, and validation event data helped to inform our choice of the 2002–2009 time window. GED reportedly underwent significant improvements about this time (Croicu and Sundberg 2015, 14) whereas focusing on the post-2000 period allows us to similarly avoid potential instability in the number of underlying sources coded within the ICEWS data. Third, a wide range of relevant political and social developments occurred within Colombia itself making the 2002–2009 period an optimal window for analyzing FARC HRVs. Colombia broke off 3 years of peace talks with the FARC in early 2002, leading to a sustained multiyear 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 4-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, Colombia's subsequent President.

Following the work in political science (Cho 2003; Cho and Gimpel 2007), sociology (Tolnay, Deane, and Beck 1996) and criminology (Baller *et al.* 2001), we also conduct additional validity assessments for three subperiods: 2002–2004, 2005–2007, and 2008–2009. These three subperiods capture distinctly different conflict phases. During the years 2002–2004, the FARC made use of the government peace process and associated demilitarized regions established by the Andrés Arango administration to prepare for future insurgency. These early years were characterized by a noticeable uptick in violence and an overall rise in the number of clashes between FARC and government security forces. In contrast, during 2005–2007 subperiod the Uribe administration executed a counterinsurgency strategy which led to a sharp decline in overall FARC activity in comparison to the violence of first subperiod. The last subperiod, 2008–2009, was one of relative peace. The three subperiods thus allow us to evaluate ICEWS and GED over conflict phases of escalation, de-escalation, and relative peace.[14] Finally, we examine the accuracy of ICEWS and GED geocoding in terms of the ability to detect *at least one event per municipality* in the main cross-section and then in each subperiod. To save space, we focus on the 2002–2004 and 2005–2007 periods in the text. Our full analyses for the 2008–2009 subperiod is reported in the Supplementary Appendix.

## 3.2 Frequencies of Agreement in Geocodings of ICEWS and GED with CINEP

We begin with an assessment based on the full 2002–2009 datasets. Figure 1 displays the observed FARC HRVs obtained from the ICEWS, GED, and CINEP data, indicating which municipalities were reported to experience at least one reported FARC HRV between 2002 and 2009. There are several noticeable differences in the patterns in the maps. ICEWS records more events both in the north and south of Colombia than did GED. Looking at the map for CINEP, both ICEWS and GED appear to underreport FARC HRVs especially in the west and south of the country. Yet the confusion matrix, Table 1, indicates that there is roughly the same level agreement between ICEWS and CINEP (73.4%) as between GED and CINEP (78.5%). The Cohen's kappa scores for both datasets are in the "fair" range of 0.20–0.40. For all the municipalities in the entire study period then, the raw data do not show major differences in the geocoding of ICEWS and GED. Rather, they exhibit spatial similarities in coverage of FARC HRVs.

Because we are interested in the problem of underreporting as a function of remoteness, we next assess the accuracy of the datasets in a collection of journalistically remote and journalistically proximate locations. To be specific, we chose five journalistically remote and five journalistically proximate municipalities from among those which CINEP reported at least five

14 Once more, in the Section 4, we discuss an extension which allows for space time error analysis by municipality year.

**Table 1.** Municipality-event confusion matrix, 2002–2009 (confidence intervals in brackets).

| 2002–2009 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No event | Event | No event | Event |
| CINEP | No event | 736 | 92 | 763 | 65 |
| | Event | 205 | 83 | 175 | 113 |
| Precision accuracy (%) | | 73.4 [70.7, 76.0] | | 78.5 [76.0, 80.9] | |
| Cohen's Kappa | | 0.20 [0.13, 0.28] | | 0.36 [0.29, 0.43] | |

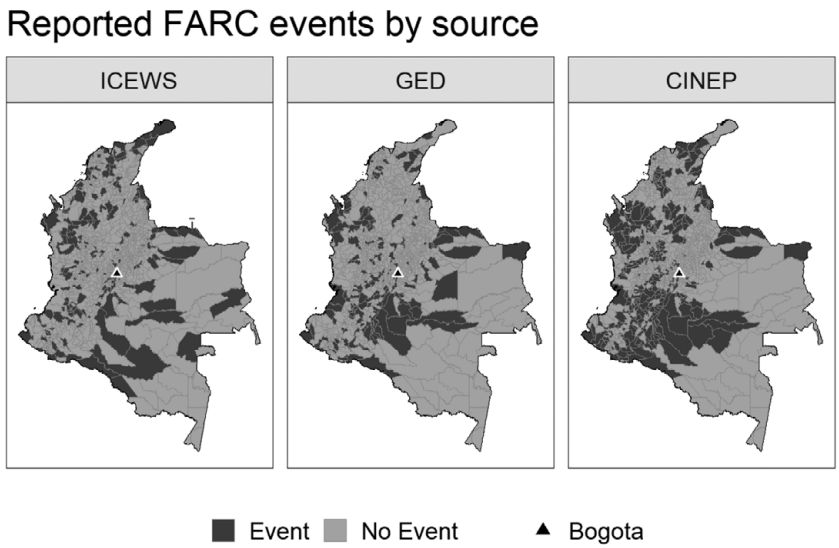


**Figure 1.** Observed FARC events.

FARC HRVs between 2002 and 2009. Five percent of all FARC HRVs reported by CINEP occurred in these ten municipalities in the study period. Journalistic remoteness and journalistic proximity (hereafter "remote" and "nonremote") were defined in terms of distance from the capitol and largest city in Colombia, Bogota. Bogota is the likely location of the journalists who supply the information to the newswires coded by ICEWS and GED. This is because in the study period Bogota was the safest location in Colombia and also the home of the country's international airport. Figure 2 depicts the selected municipalities.[15]

Table 2 contains these additional external validity assessments. The top panel reports the results for at least one reported FARC HRV in each of the eight years in each cluster of five municipalities. Both ICEWS and GED are slightly more accurate in reporting remote than nonremote FARC HRVs. But the confidence intervals for their accuracy statistics overlap. The same is true of the confidence intervals for their accuracy in reporting FARC HRVs in our selected nonremote municipalities. Only the confidence interval for the Cohen's kappa statistic for GED for the selected remote municipalities does not span zero. The middle and bottom panels repeat the assessment for the two subperiods, 2002–2004 and 2005–2007, respectively. Both ICEWS and GED are slightly more accurate in the remote than in the nonremote selected municipalities in the former subperiod. Once more the confidence intervals for their accuracy scores overlap; the confidence intervals for the Cohen's kappa statistics for the 2002–2004 subperiod are all wide and span zero. In contrast, ICEWS is a bit more accurate than GED in the remote selected municipalities in the 2005–2007 subperiod; the confidence interval for its Cohen's kappa statistic is wide but

15 The level of FARC HRVs and exact (centriod) distances of each of the ten selected municipalities from Bogota are reported in the Supplementary Appendix.

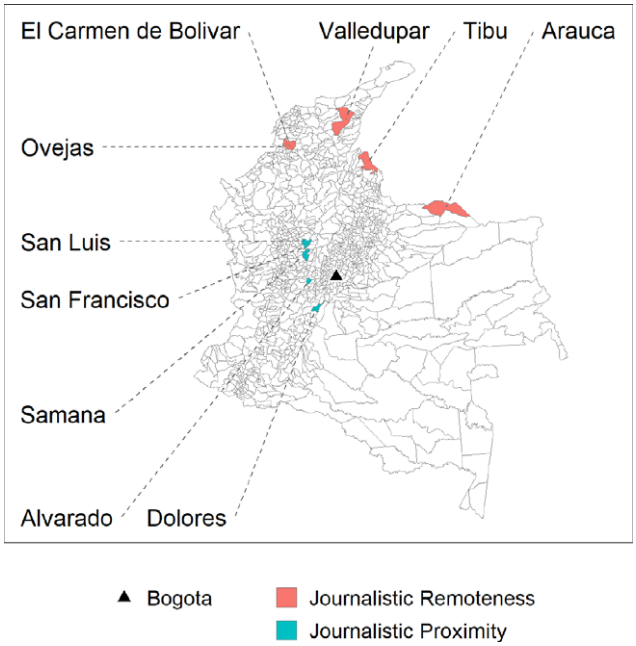


**Figure 2.** Selected journalistically remote and journalistically proximate municipalities.

in this case it does not span zero. GED hence is more prone to underreporting than ICEWS in the remote selected municipalities in 2005–2007. Otherwise the validity of the two datasets in 2005–2007 is indistinguishable. Overall this additional analysis of the raw data does not show a remoteness problem. It also does not show one of the datasets is more externally valid than the other.

While useful, this analysis of the raw data is limited to comparisons of geocodings for a small number (10) of selected (high FARC HRV) municipalities. We could examine individual misclassification errors at other specific sites (municipalities) for each dataset. But it would be difficult to sort out the effect of observed factors like each municipality's terrain and population from unobserved factors that might be responsible for geocoding errors. We also could examine the accuracy of ICEWS and GED over more subperiods of municipalities and years. If we examined ICEWS' and GED's accuracy in each municipality for the entire (pooled) 8-year time period we would have 1,116 comparisons with CINEP for each dataset. If we conducted the same municipality by municipality assessment for each year we would have $8 \times 1,116 = 8,928$ comparisons with CINEP for each dataset. These additional assessments would be not just time consuming but difficult to synthesize in a manner that offers insights beyond those provided in Tables 1 and 2. Our inferences about the impact of remoteness above are based on our twofold division of ten municipalities with high levels of FARV HRVs into journalistically remote and journalistically proximate categories. The average impact of distance on spatial error interdependence over all 1,116 municipalities, regardless of their level of FARC HRVs—what we called the "range"—is not revealed by these initial validity assessments. Spatial modeling gives us insights into the average impact of observed covariates and unobserved factors on geocoding accuracy. Spatial modeling also allows us to analyze the external validity of ICEWS and GED for the entire sample of municipalities in 2002–2009 and our subperiods including the determinants of underreporing (the remoteness problem). In these ways, spatial modeling is a valuable complement to the analysis of the raw data.

## 3.3 External Validity Assessment with the Geostatistical Spatial Error Model

Drawing upon earlier work on the remoteness problem in political violence event data measurement and analysis (Davenport and Ball 2002; Hammond and Weidmann 2014; Weidmann

**Table 2.** Municipality-event confusion matrices for selected municipalities, 2002–2009, 2002–2004 and 2005–2007 (confidence intervals in brackets).

**Selected Non-Remote Municipalities**

| 2002 - 2009 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 21 | 0 | 21 | 0 |
| | Event | 18 | 1 | 15 | 4 |
| Precision Accuracy (%) | | 55.0 [38.5, 70.7] | | 62.5 [45.8, 77.3] | |
| Cohen's Kappa | | 0.06 [-0.27, 0.38] | | 0.22 [-0.09, 0.53] | |
| Report at least one event over eight years. | | | | | |

| 2002 - 2004 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 3 | 0 | 3 | 0 |
| | Event | 11 | 1 | 10 | 2 |
| Precision Accuracy (%) | | 26.7 [7.8, 55.1] | | 33.3 [11.8, 61.6] | |
| Cohen's Kappa | | 0.04 [-0.26, 0.33] | | 0.07 [-0.26, 0.41] | |
| Report at least one event over three years. | | | | | |

| 2005 - 2007 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 10 | 0 | 10 | 0 |
| | Event | 5 | 0 | 4 | 1 |
| Precision Accuracy (%) | | 66.7 [38.4, 88.2] | | 73.3 [44.9, 92.2] | |
| Cohen's Kappa | | 0.00 [-0.72, 0.72] | | 0.25 [-0.38, 0.88] | |
| Report at least one event over three years. | | | | | |

**Selected Remote Municipalities**

| 2002 - 2009 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 18 | 3 | 21 | 0 |
| | Event | 11 | 8 | 12 | 7 |
| Precision Accuracy (%) | | 65.0 [48.3, 79.4] | | 70.0 [53.5, 83.4] | |
| Cohen's Kappa | | 0.28 [-0.02, 0.59] | | 0.38 [0.09, 0.67] | |
| Report at least one event over eight years. | | | | | |

| 2002 - 2004 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 1 | 1 | 2 | 0 |
| | Event | 8 | 5 | 7 | 6 |
| Precision Accuracy (%) | | 40.0 [16.3, 67.7] | | 53.3 [26.6, 78.7] | |
| Cohen's Kappa | | -0.05 [-0.48, 0.39] | | 0.19 [-0.25, 0.63] | |
| Report at least one event over three years. | | | | | |

| 2005 - 2007 | | ICEWS | | GED | |
|---|---|---|---|---|---|
| | | No Event | Event | No Event | Event |
| CINEP | No Event | 9 | 1 | 10 | 0 |
| | Event | 2 | 3 | 5 | 0 |
| Precision Accuracy (%) | | 80.0 [51.9, 95.7] | | 66.7 [38.4, 88.2] | |
| Cohen's Kappa | | 0.53 [0.05, 1.01] | | 0.00 [-0.72, 0.72] | |
| Report at least one event over three years. | | | | | |

2016), we collected data on both spatial and aspatial covariates. For the former, we used distance from the Colombian capitol of Bogota in logged kilometers.[16] For the latter, we used municipality population and a terrain ruggedness index (TRI).[17]

Centroids for our 1,116 municipalities served as the initial points for our Constrained Refined Delaunay Triangulation. We set the maximum edge length on the triangulation to 1.6 degrees which corresponds to about 180 km at the Equator; this produces a mesh of 196 vertices akin to the 141 vertices in the mesh employed by Python *et al.* (2017) in their study of terrorism in Nigeria.[18]

As in Python *et al.* (2017), the hyperparameters of our model essentially are the default values in R-INLA. We set $\lambda$ in equation (2) to 1, leading the smoothing parameter, $\alpha$, to equal to 2. R-INLA employs Neumann boundaries. At these boundaries, variances are inflated. To avoid this problem, a baseline range is calculated to ensure a range smaller than the size of the domain size (mesh). Lindgren and Rue (2015, 12) call this the "prior median for the spatial range." Using the solution to the SPDE, from this baseline range, baseline $\tau$ and baseline $\kappa$ parameters can be derived. The baseline $\tau$ and baseline $\kappa$ parameters are used in the parameter basis functions of the estimation. To the baseline values for $\tau$ and for $\kappa$ are added the $\theta$ 2-tuple in equation (5). Specifically, $\theta_1$ is added to $\log(\tau_0)$ and $\theta_2$ is added to $\log(\kappa_0)$. $\theta_1$ and $\theta_2$ are assumed to be jointly normally distributed. Our hyperparameter means are set to zero; and their precisions are set to 0.1 and 1, respectively. We assign Gaussian priors with mean 0 and precision 0.001 to the intercept and covariate coefficients.[19]

Figures 3 and 4 present the results of our geostatistical analysis with the three datasets for observed FARC HRVs and underreporting of these events respectively. The results for the entire period, 2002–2009, can be found in the first columns of these figures.[20] In terms of prediction of FARC-HRVs presented in Figure 3, the results for the ICEWS and GED datasets, like the model for the ground truth dataset, CINEP, do not depend on the distance from Bogota. Rather for all three datasets, predictions of FARC-HRVs depend positively on municipality population and TRI (although the 2.5% left credible interval for TRI for the ICEWS model is nearly zero). The determinants of underreporting in 2002–2009—the failure of ICEWS or GED to report a FARC-HRV when a FARC-HRV is reported by CINEP—are reported in the first column of Figure 4. The statistical results are qualitatively identical; they indicate that for both the machine- and human-geocoded datasets underreporting is related to population and TRI. So while there is some evidence that ICEWS' machine geocoding depends less on TRI than GED's human geocoding, the determinants of their predictions and underreporting are very similar.

There is evidence that the errors of the models based on the three datasets are spatially interdependent. For the period 2002–2009 of observed FARC–HRVs in Figure 3, none of the credible intervals for any of the GMRF parameters span zero. But the GMRF parameters tell somewhat

16 Distance was estimated using municipality centroids and the latitude and longitude for Bogota extracted from a shapefile projected with a South America Albers Equal Area Conic.

17 As taken from the WorldPop Global Population Data and the EarthEnv project (Amatulli *et al.* 2018), respectively. While others have favored using mountainous terrain (Hammond and Weidmann, 2017), we employ the more precise TRI, which better captures difficult terrain favored by rebel groups. Additionally, the small resolution of the data (< 1 km) allow us to meaningfully estimate average terrain ruggedness in small Colombian municipalities.

18 Python *et al.* (2017) is a good benchmark since (i) it studies a related phenomenon, terrorism, and (ii) Colombia is roughly the same size as Nigeria. Note that our results are not sensitive to the use of centroids. A random draw of locations within municipalities yielded the same result reported here.

19 Boundary effects inflate the variance by a factor of 2 along straight boundaries and by a factor of 4 near right angle corners (Lindgren and Rue 2015, 6). Recall that the SPDE solution included formulae for the marginal variance and range. Using these formulae, one can derive from the baseline range, the baseline $\log(\tau)$ and $\log(\kappa)$ values. In general, $\theta_1 = \log(\tau)$ and $\theta_2 = \log(\kappa)$. These hyperparameters are assumed to the jointly normally distributed. With the baseline range set at 1/5 of the mesh domain size, the precision for $\theta_2$ implies there is a 95% probability that the actual range is less than the domain size (Lindgren and Rue 2015, 6). We employ R-INLA version 21.2.23; it is available at https://www.r-inla.org/.

20 Tables containing the coefficient values and credible intervals behind Figures 3 and 4 are in the Supplementary Appendix. These figures present posterior median parameter estimates and 95% highest posterior densities. A parallel set of the results for the entire period analysis when using neighborhood-based SPEMs also can be found in the Supplementary Appendix.

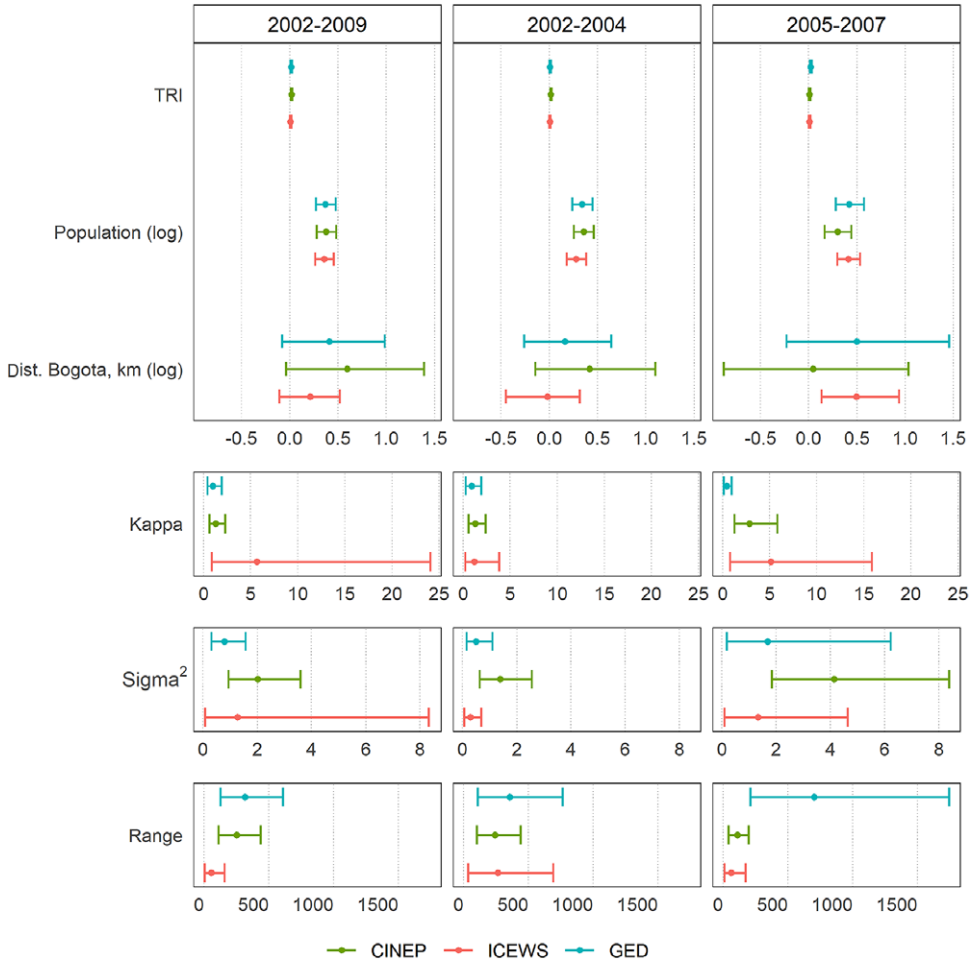


**Figure 3.** Coefficient and GMRF parameter values for geostatistical models based on ICEWS, GED, and CINEP datasets, 2002–2009, 2002–2004, and 2005–2007.

different stories about the machine versus human coding of FARC HRVs. The range parameter is indicative of the remoteness problem. And the median estimate of the range of spatial error dependence for the ICEWS dataset is closer to that of the errors of the CINEP model. The $\sigma_\xi^2$ parameter tells us about uncertainty in prediction due to data sparseness and measurement error. Again, for 2002–2009 the median estimate of this parameter for the ICEWS model is closer to the median estimate for CINEP model than that of the GED model.[21]

The left panel of Figure 5 depicts for the 2002–2009 period the receiver operating characteristic (ROC) curves for predictions of FARC–HRVs based on the ICEWS and GED datasets using the external CINEP dataset as ground truth. These curves and the associated area under the curve (AUC) statistics show that the performance of the models based on the machine- and human-coded datasets are roughly comparable. The AUC for the GED-based model is slightly higher than the AUC for the ICEWS model but their confidence intervals overlap. In sum, our geostatistical analysis for the Colombia case shows that there are spatially dependent errors in all three datasets. However, as in our analysis based on the selected (non)journalistically remote municipalities, there is no evidence that the machine-geocoded dataset is less externally valid than the human-coded dataset.

21 Error bands for these range estimates are reported in Figure A.9 of the Supplementary Appendix. Site specific estimates of the probability of FARC HRVs and of mean errors, and the marginal variances of these errors, are depicted in Supplementary Figure A.8.

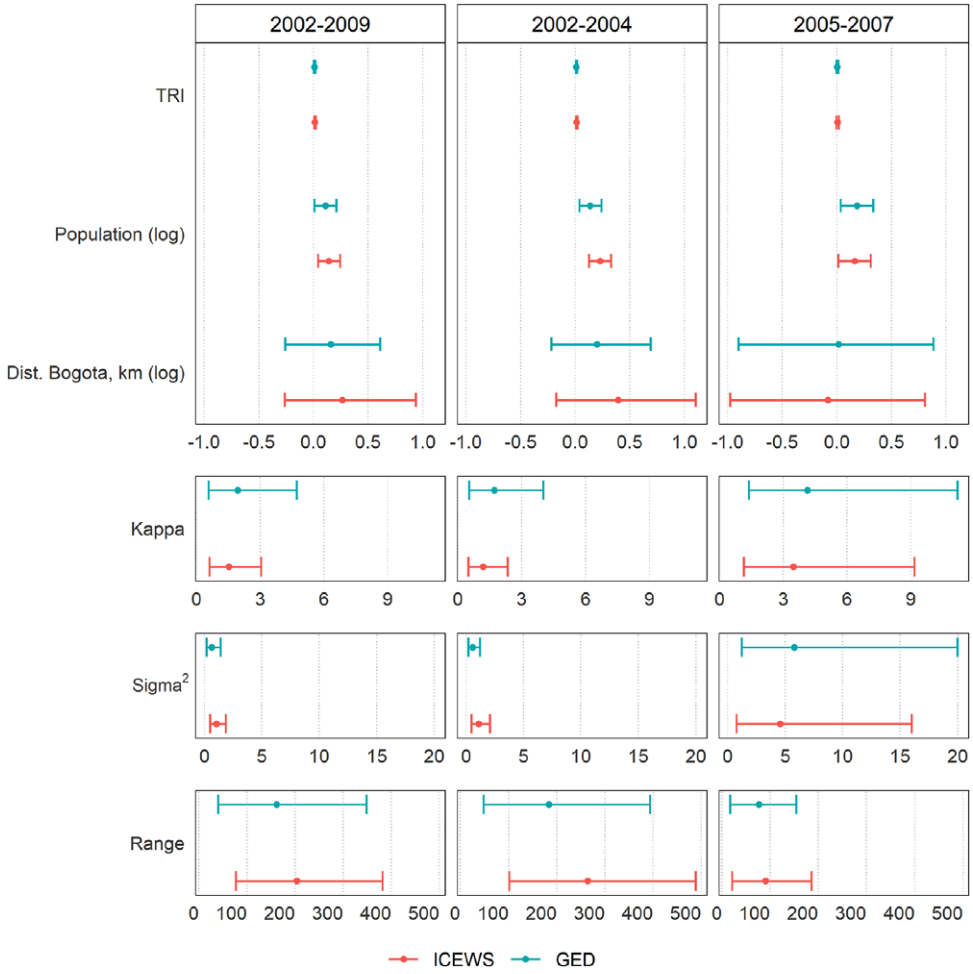


**Figure 4.** Coefficient and GMRF parameter values for geostatistical models of underreporting of ICEWS and GED compared to CINEP, 2002–2009, 2002–2004, and 2005–2007.

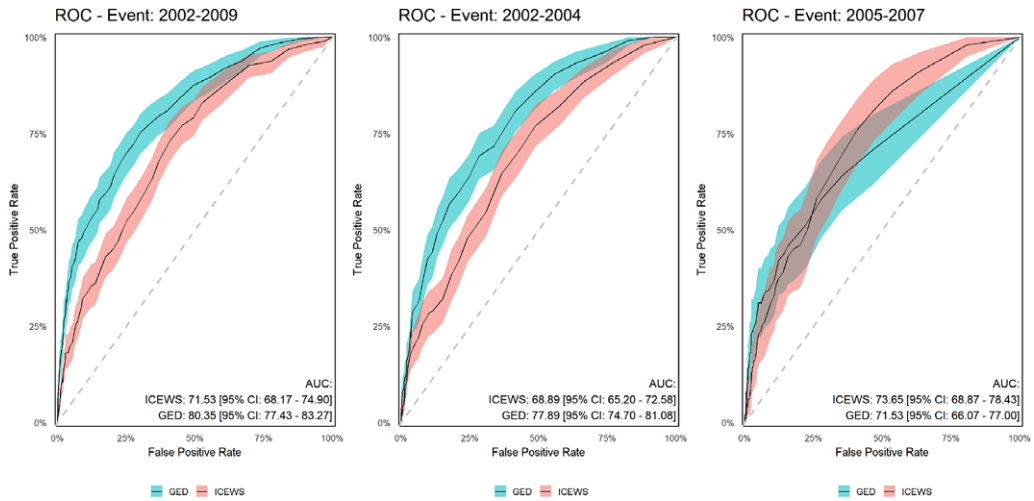


**Figure 5.** ROC curves and AUC statistics for ICEWS, GED, and CINEP models of FARC–HRVs.

Figures 3–5 also report the results for the geostatistical analysis of the datasets in the substantively important subperiods of the Colombian conflict, specifically, for 2002–2004 and 2005–2007.[22] Briefly, distance from Bogota again does not consistently predict FARC–HRVs in any of the

22 The results for the subperiod 2008–2009 are in the Supplementary Appendix.

three models. That being said, the estimate for this variable does not overlap zero in the 2005–2007 ICEWS model, possibly indicating period-specific factors during this more intense conflict period. The coefficients for population are consistent predictors for all three models across the subperiods with parmater values not overlapping with zero. The difference is that in the 2002–2004 subperiod the coefficient on TRI is not a reliable predictor for the ICEWS based model while it is for the GED and CINEP based models. In contrast, for the 2005–2007 subperiod, the coefficient on TRI is reliable for the GED model but not for the ICEWS and CINEP models. The determinants of underreporting by ICEWS and GED are qualitatively identical in both of these subperiods. As regards the GMRF parameters for the subperiod models, the median estimate of the range of spatial error dependence for ICEWS is closer to that of the model errors for CINEP in both subperiods; the median estimate for the spatial error dependence from the GED model in the 2005–2007 is much larger than the CINEP median range estimate. Finally, the ROC curves for the predictions based on the ICEWS and GED models in the two subperiods are very similar and the GED models' AUCs are only slightly better than those of the ICEWS models. Therefore, the subperiod analyses also do not show that the machine-geocoded data are less externally valid than the human-geocoded data.

Our results do suggest some subtle differences in the validity of these two datasets. GED appears to miss more HRVs than ICEWS because, as we explain in the Supplementary Appendix, ICEWS uses a larger collection of news sources and recovers more raw events. As a result, our model-based analysis of ICEWS-derived HRV events—and of the associated GMRF parameters discussed above—indicate that ICEWS has somewhat better event coverage, especially in some subperiods of the conflict. In our models' coefficient estimates and predictions, this superior coverage offsets ICEWS' potential shortcomings in machine-based geolocation accuracy. At the same time, our supplemental results—especially regarding the range estimates for the GMRF (Supplememtary Figure A.9)—imply that GED's sparser coverage still produces spatial model errors and spatial interdependence estimates that better match the spatial pattern of errors from models based on the ground truth dataset, CINEP. This is likely owing to GED's relatively more accurate human-based event geolocations (i.e., among the events that GED does capture), relative to ICEWS' machine-based geolocation routines. In this way, our results help researchers using both kinds of data to understand how remoteness affects the accuracy of their coding methods.

## 4 Discussion

This article develops and implements a framework for the subnational validation of machine-coded event data. Past research reviewed above and in the Supplementary Appendix suggests that the two event datasets evaluated here—ICEWS and GED—are currently considered among the most subnationally accurate, global machine- and human-coded event datasets available. As such our findings obtained from these two sources can be viewed as a "ceiling" for the spatial validity of currently available global machine- and human-coded datasets. Yet previous spatial validations of these two datasets have only been implemented at the premodeling stage (Eck 2012; Lautenschlager, Starz, and Warfield 2017). Like the confusion matrices discussed above, this does not allow one to assess their spatial validity in relation to, and conditional on, covariates such as remoteness. In implementing spatial validation at the modeling stage with the aid of ground truth data (CINEP) our article addresses this deficiency while providing a stringent test for evaluating subnational geolocation accuracy of machine- and human-coded event data.

Our findings also significantly improve upon previous external validation assessments of machine- and human-coded event data. Past spatial statistical comparisons of human- and machine-coded event data find that the machine-coded GDELT data are significantly less accurate than human-coded data due to remoteness problems, concluding with respect to GDELT that "[f]or geo-spatial analyses of violence, this may be reason to worry" (Hammond and Weidmann 2014, 5).

Yet these past analyses are limited to a very narrow—and arguably suboptimal—subset of spatial statistical models and machine-coded event datasets. In extending this spatial statistical analysis to a broader set of spatial statistical models—and the more accepted and widely used machine-coded event dataset (ICEWS)—we find that these earlier worries may no longer hold. Specifically, we demonstrate that with the most commonly used neighborhood (Supplementary Appendix) and geostatistical spatial models and for a relatively fine grained level of spatial (dis)aggregation, machine- and human-coded event data produce comparable inferences, often yield substantively indistinguishable predictions, and exhibit similar rates of (remoteness-induced) underreporting of FARC HRVs. Yet these models also offer unique insights into the different sources of spatial (in)accuracy in each dataset, where our findings suggest that ICEWS' noisier geolocation accuracy may be counterbalanced by its superior recall relative to GED when it comes to spatial model inference and prediction.

Additional validity assessments of human rights event data are clearly called for. Our binary, municipality-period event indicators collapse higher frequencies of HRVs in some instances. Such collapsing is commonplace in the subnational conflict literature and our subperiod findings suggest that this is not substantially altering our conclusions with respect to time. Nevertheless analyses of event frequencies or of more fine grained time periods is an important next step. Because our model-based validation framework is amendable to assessments of other event datasets and subnational contexts, comparable evaluations of other prominent machine- and human-coded event datasets should likewise be implemented to enhance the generalizability of our findings. Although not widely available, ground truth datasets like CINEP must be included in these assessments in order to gauge the external validity of the geocoded data. To this end, the SIGACTS dataset (e.g., Weidmann 2016) is an additional testbed for an extension of our current assessments. Event data's temporal dynamics also should be incorporated into future validity assessments. Taking a cue from our studies of subperiods in the Colombian case, this will be rigorously implemented for Iraq as a next step in our research agenda. In this extension, we will employ the full spatio-temporal version of our geostatistical models analyzing coding errors by both years and municipality (Python *et al.* 2018). Such an extension will produce a more complete understanding of the validity of machine- and human-coded event data.

## Acknowledgments

We thank Scott Cook, Jude Hays, Phil Schrodt, Clayton Webb, and Nils Weidmann for comments.

## Funding

This work was supported by the National Science Foundation [SBE-SMA-1539302].

## Data Availability Statement

The replication materials for this article are posted to the Political Analysis Dataverse (Stundal *et al.* 2021).

## Supplementary Material

For supplementary material accompanying this paper, please visit https://dx.doi.org/10.1017/pan.2021.40.

## Bibliography

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. 2021. "A Total Error Approach for Validating Event Data." *American Behavioral Scientist* , 1–22.

Amatulli, G., S. Domisch, M. Tuanmu, B. Parmentier, A. Ranipeta, J. Malczyk, and W. Jetz. 2018. "Data Descriptor: A Suite of Global, Cross-Scale, Topographic Variables for Environmental and Biodiversity Modeling." *Scientific Data* 5: 1–15.

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

Anselin, L. 2006. "Spatial Econometrics." In *Palgrave Handbook of Econometrics: Volume 1: Econometric Theory*, edited by T. C. Mills and K. Patterson, Chap. 26, pp. 901–969. Basingstone: Palgrave MacMillan.

Bagozzi, B. E., 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.

Baller, R. D., L. Anselin, S. F. Messner, G. Deane, and D. F. Hawkins. 2001. "Structural Covariates of U.S. County Homicide Rates: Incorporating Spatial Effects." *Criminology* 39 (3): 561–590.

Beiler, J., P. T. Brandt, A. Halterman, P. A. Schrodt, and E. M. Simpson. 2016. "Generating Political Event Data in Near Real Time: Opportunities and Challenges." In *Computational Social Science: Discovery and Prediction R*, edited by M. Alvarez. New York: Cambridge University Press.

Blangiardo, M. and M. Cameletti. 2015. *Spatial and Spatial Temporal Bayesian Models with R-INLA*. New York: Wiley.

Boschee, E., J. Lautenschalger, S. O'Brien, S. Shellman, J. Starz, and M. Ward. 2016. "ICEWS Coded Event Data." https://doi.org/10.7910/DVN/28075, Harvard Dataverse, V30, UNF:6:NOSHB7wyt0SQ8sMg7+w38w== [fileUNF].

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

Cho, W. K. T. and J. 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* 27 (4): 397–414.

CINEP. 2008. "Marco conceptual: banco de datos de derechos humanos y violencia política." Centro de Investigación y Educaión Popular.

Cook, S. J., B. Blas, R. J. Carroll, and S. Sinha. 2017. "Two Wrongs Don't 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).

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 MPSA Meeting, Chicago, IL.

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, A. 2013. "Long Term Ecological Change and Geographical Patterns of Violence in Darfur 2003–2005." Paper Presented at the Annual Meeting of the American Political Science Association, Chicago, August 29–September 1, 2013.

Eck, K. 2012. "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets." *Conflict and Cooperation* 47(1): 124–141.

Gill, J. 2021. "Measuring Constituency Ideology Using Bayesian Universal Kriging." *State Politics & Policy Quarterly* 21(1): 80–107.

Grimmer, J. and B. M. Stewart. 2013. "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." A Report by the Benetech Human Rights Program.

Hammond, J. and N. B. Weidmann. 2014. "Using Machine Coded Event Data for the Micro Level Study of Political Violence." *Research and Politics* 1(2): 1–8.

Holmes, J. S., S. A. G. 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. 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.

Lautenschlager, J., J. Starz, and I. Warfield. 2017. "A Statistical Approach to the Subnational Geolocation of Event Data." In *Advances in Cross-Cultural Decision Making*, edited by S. Schatz and M. Hoffman. Switzerland: Springer International Publishing.

Lindgren, F., and H. Rue. 2015. "Bayesian Spatial and Spatio-Temporal Modeling with R-INLA." *Journal of Statistical Software* 63 (19): 1–25.

Lindgren, F., H. 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.

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): 1–18.

Martinelli, D., and G. Geniaux. 2017. "Approximate Likelihood Estimation of Spatial Probit Models." *Regional Science and Urban Economics* 64: 30–45.

Python, A., J. Illian, C. Jones-Todd, and M. Blangiardo. 2017. "Explaining the Lethality of Boko Haram's Terrorist Attacks in Nigeria 2009–2014: A Hierarchical Bayesian Approach." In *Bayesian Statistics in Action*, edited by R. Argiento, E. Lanzarone, I. A. Villalobos, and A. Mattei. Cham: Springer.

Python, A., J. B. Illian, C. M. Jones-Todd, and M. 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* 182 (1): 1–22.

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

Schrodt, P., and D. Gerner. 1994. "Validity Assessment of Machine-Coded Event Dataset for the Middle East, 1982–1992." *American Journal of Political Science* 38 (3): 825–854.

Stundal, L., B. Bagozzi, J. Freeman, and J. Holmes. 2021. "Replication Data for: Human Rights Violations in Space: Assessing the External Validity of Machine Geo-coded Vs. Human Geo-coded Data." https://doi.org/10.7910/DVN/JYVGLU, Harvard Dataverse, V1.

Sundberg, R., and E. Melander. 2013. "Introducing the UCDP Georeferenced Event Dataset." *Journal of Peace Research* 50 (4): 523–532.

Tolnay, S. E., G. Deane, and E. M. Beck. 1996. "Vicarious Violence: Spatial Effects on Southern Lynchings, 1890–1919." *American Journal of Sociology* 102 (3): 788–815.

von Borzyskowski, I., and M. Wahman. 2021. "Systematic Measurement Error in Election Violence Data: Causes and Consequences." *British Journal of Political Science* 51 (1): 230–252.

Ward, M. D., and K. S. Gleditsch. 2019. *Spatial Regression Models*, 2nd edn. Thousand Oaks, CA: Sage.

Weidmann, N. B. 2016. "A Closer Look at Reporting Bias in Conflict Event Data." *American Journal of Political Science* 60 (1): 206–218.

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

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.

Zammit-Mangion, A., M. Dewar, V. Kadiramanathan, and G. Sanguinetti. 2012. "Point Process Modelling of the Afghan War Diary." *Proceedings of the National Academy of Sciences of the United States of America* 109 (31): 12414–1224.