Event Comparisons for Colombia Data to be Used in "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 12, 2020

[*] This research is supported by NSF Grant Number SBE-SMA-1539302.

[†] 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.

1 Colombia Data and Context

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.

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-perpetrated 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, and has an archive with an extensive collection of (Spanish language) national and regional Colombian newspapers and associated 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 are 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 2002-2009. This time window is motivated by (1) past analyses of political violence in Colombia,[1]

[1] For instance, Bagozzi et al. (2019) consider instances of rebel and paramilitary violence against civilians

(2) data availability on our human, machine, and validation event data,[2] and (3) relevant political and social processes occurring within Colombia itself.[3] Specifics on the actors, event types, and geo-precision used in aggregating our GED and ICEWS data appear below.

2 Formatting the Colombia Data

We aggregate our GED, ICEWS, and CINEP data on Colombian human rights violations at the municipality-year level for the years 2002-2009. Due to distinct processes of geolocation and event coding, these datasets 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 of violence perpetrators and victims. These differences necessitate several important decisions when spatially aggregating and combining these datasets for comparison. What follows is a detailed discussion of our efforts to format and combine each of these event datasets in a manner that ensures that our retained events are as comparable as possible across all three sources.

We first formatted our 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 first identify events committed against civilian targets. Here we subset ICEWS to include only those events designated as "general population," "civilian," and "social" as our target actors. Next, we sought to identify and retain only those ICEWS events that were perpetrated by rebel actors in Colombia. We evaluated three separate approaches:

in Colombia during the years 2000–2009, whereas Lum et al. (2010) consider lethal violence in Colombia's Casanare department for the 1998–2007.

[2] 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.

[3] 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.

1. We used the source actor designations contained within the ICEWS data to retain any events perpetrated by actors designated as: “rebel,” “separatist,” “insurgent,” and “unidentified sources.”

2. We used the source actor designations contained within the ICEWS data to retain any events perpetrated by actors designated as: “rebel,” “separatist,” and “insurgent.” (That is, we excluded “unidentified sources”).

3. We used the sourcename designations contained within the ICEWS data to retain any events that were perpetrated by the FARC. This accordingly excludes events that were perpetrated by Colombia’s main other rebel groups (ELN & EPL) as well as events in ICEWS that were only recorded as being perpetrated by “guerillas,” “communist guerillas,” “communist rebels,” “rebels,” and so on. The only way to properly identify which rebel-perpetrated events were attributed to the FARC, specifically, was to use ICEWS’ additional sourcename variable. This variable includes the entity name identified in an ICEWS-coded news article. These names are not standardized. Hence, for our relevant events, sourcename contains many variants of ’FARC’, alongside a number of individual rebel/paramilitary/cartel-member actor names. We manually identified the subset of the latter actors that were members of FARC via Google/Wikipedia, and then retained only those events with a FARC-associated sourcename.

For each rebel source-actor approach mentioned above, we subset our identified rebel → civilian events to only include those events occurring in Colombia during the years 2002-2009 that were geo-located to the city/town level(s) of geographic precision. This was achieved by dropping any events that contained no textual information within ICEWS’ “city” variable. Future iterations will consider also retaining cases that only contained textual information in ICEWS’ “district” variable. 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 generated the ICEWS-set of all Colombian human rights violation events involving rebel actors and civilian targets that were coded to a city level of geographic precision for years 2002-2009. With these individual events in hand, we next turned to aggregating all remaining events. Before aggregating each set of ICEWS events to the municipality-year level, we 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—while still allowin for some duplicate stories given (e.g.,) variation in headlines (Schrodt, 2015: 14). We then aggregated that our deduplicated ICEWS rebel violence events to municipality-year counts by matching these events to shapefiles of Colombia's municipalities via latitude longitude coordinates.

We next formatted the GED (Sundberg and Melander 2013) in a comparable manner to the ICEWS data described above. 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 2002-2009. 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) events from any retained rebel source actor (noting that GED, unlike ICEWS, does not exclude an unknown source actor designation) or (ii) the subset of those events that had GED's standardized "FARC" source actor designation, specifically. Our two GED datasets 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.[4] 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 field (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

[4] 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.

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 instances of "material" human rights violations, rather than both material and verbal human rights violations.[5] After preprocessing our CINEP data in these manners, we aggregated all remaining rebel-perpetrated CINEP events to the municipality-year level for only our years of interest, and merged these cases to our final GED and ICEWS data. We then repeated this process with the subset of our CINEP events that were designated as having a FARC source actor.

2.1 Comparisons

After combining our final GED, CINEP, and ICEWS events into a pair of municipality-year datasets, we seek to compare and validate our different event dataset measures of rebel violence against civilians in Colombia. We compare two different scalings of our final event datasets. First, we compare these event datasets' raw municipality[6] counts at the 2002-2009 period level. Second, we construct a logged, re-scaled version of each count. Specifically, our logged (ICEWS, GED, and CINEP) rebel-perpetrated violence event counts were re-scaled to the 0-1 range via min-max normalization:

$$\text{Scaled Event Count}_i = \frac{\ln(\text{Event Count}_i + 1) - \ln(\text{Event Count}_{min} + 1)}{\ln(\text{Event Count}_{max} + 1) - \ln(\text{Event Count}_{min} + 1)}$$

where i is a given municipality, ln is the natural logarithm, and min & max are the minimum and maximum municipality-level counts for a particular (ICEWS, GED, or CINEP) event dataset, across 2002-2009. This recaling is intended to address the potential that our three event datasets, on the whole, operate on different scales.

To re-cap: we are comparing three different event datasets at the municipality-level across the 2002-2009 period: ICEWS, GED, and CINEP. For each event dataset, we create

[5] 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.

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

separate versions that (i) include all rebel source actors recorded by that event dataset and (ii) only include FARC source actors. For item (i), we also evaluate two distinct ICEWS approaches: (a) an approach that treats ICEWS' "unidentified sources" as rebel perpetrators and (b) an approach that omits ICEWS' "unidentified sources" as rebel perpetrators. For each version of event data, we then consider raw counts and logged min-max normalized counts. We first consider our municipality maps (for the entire 2002-2009 period) when using all rebel-perpetrated (ICEWS, GED, and CINEP) events in Figures 1-2. The top rows of these Figures report raw counts, whereas the bottom rows of each Figure report our logged min-max normalized counts. The only difference between these Figure 1 and Figure 2 is that the ICEWS event counts in Figure 2 omit the aforementioned "unidentified sources" perpetrators, whereas Figure 1 includes relevant events perpetrated by "unidentified sources" source-actors as rebel perpetrators. For reference, also note that Figure 4 in the Appendix presents a higher resolution version of Figure 1 for easier visual interpretation.

Looking first across the raw counts (top rows) of Figures 1/4 and 2, we can note that our municipality-level rebel-perpetrated event counts are on a noticeably higher scale for ICEWS, in comparison to GED/CINEP. Most notably, these Figures indicate that ICEWS exhibits over 2,000 rebel-perpetrated events for Colombia's capital district (Bogotá) during the 2002-2009 period, whereas GED and CINEP each exhibit total event counts for Bogotá during this period that are closer to 20 and 30, respectively. As Figure 2 demonstrates, this difference in event count totals is not driven by the inclusion of "unidentified sources" source actors within the ICEWS aggregations, as omitting these "unknown actors" from our ICEWS counts has little effect on the patterns noted above. We *do* find (via the second rows of 1/4 and 2) that our scaling approach appears to address these scale-issues somewhat. However, the logged min-max normalized ICEWS and GED counts do appear to still exhibit a disproportionately higher event total in the capital district, relative to CINEP.

The remarkably higher (capital district) event counts for ICEWS seen in Figures 1/4 and 2 could be attributable to a number of factors. While Figure 2 omits "unidentified sources"-

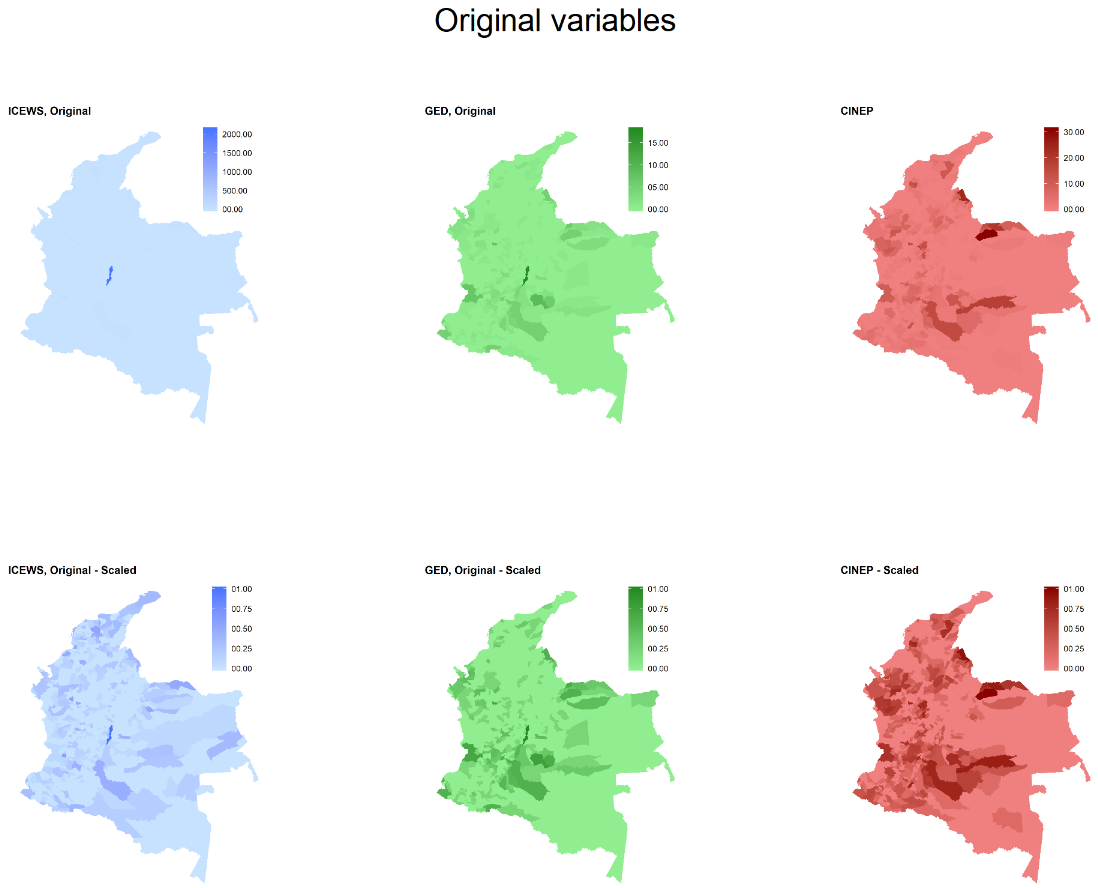


Figure 1: Comparisons of All Rebel-Perpetrated Events (Incl. ICEWS Unknown Actors)

perpetrated events from the ICEWS data (as GED and CINEP do by default), ICEWS also includes events that saw no casualties, whereas GED only records events yielding casualties. However, given that ICEWS was already subset to contain only events with CAMEO codes 18 and 20, this is unlikely to be the sole driver of the discrepancies identified in Figures 1/4 and 2. Another source of the discrepancy may be the underlying news sources that are coded across ICEWS and GED: ICEWS codes more underlying sources, and may end up capturing more events. This will be investigated at a later date. A final source of the discrepancy may be ICEWS' more extensive—but more ambiguous—rebel actor designations. ICEWS' more ambiguous actor designations may be (i) ensuring that ICEWS is capturing more true rebel-perpetrated events than GED but that (ii) ICEWS is also overreporting rebel-perpetrated

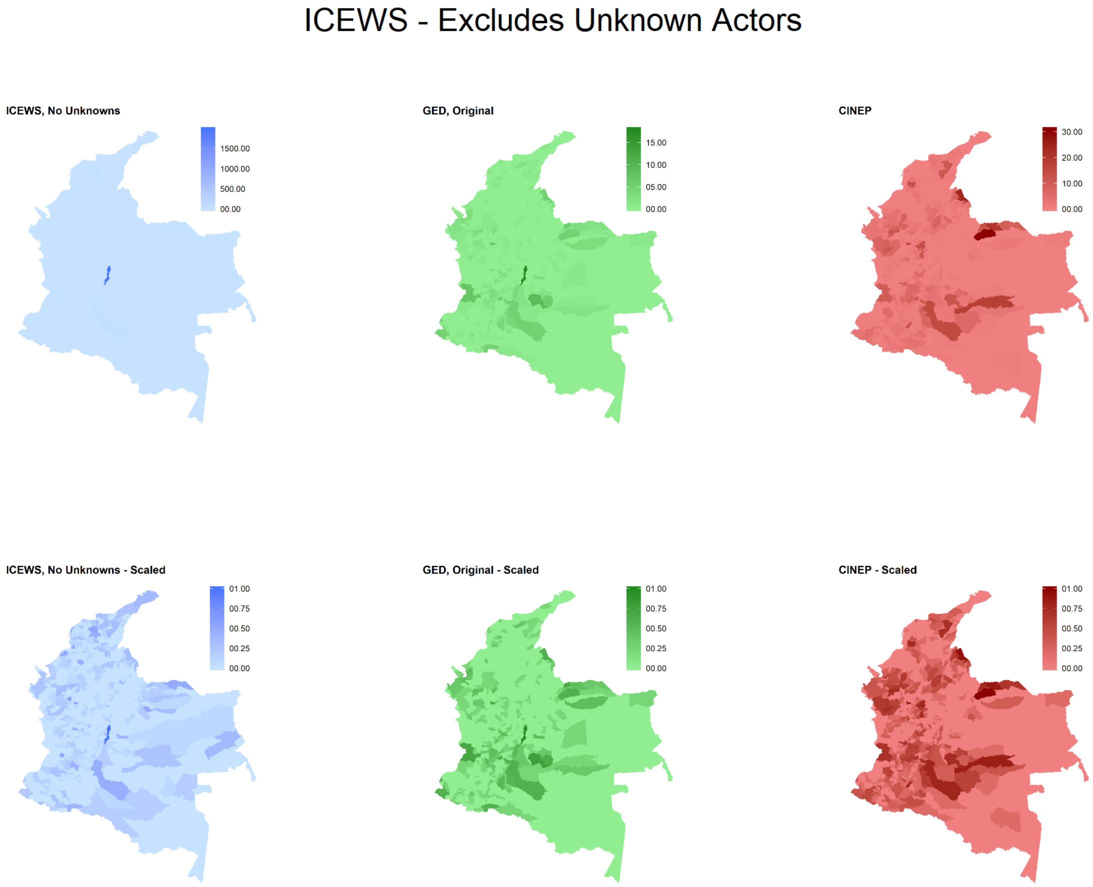


Figure 2: Comparisons of All Rebel-Perpetrated Events (No ICEWS Unknown Actors)

events (in manners that can't be addressed with one-a-day de-duplication) more so than GED. To examine this, we next turn to our final set of event data comparisons, which only focus on FARC-imitated (ICEWS, GED, and CINEP) instances of violence against civilians.

Turning to these comparisons in Figure 3, our ICEWS events now appear to be on a much more similar scale to our comparable GED and CINEP events. Herein, our GED and CINEP event totals remain relatively unchanged, with municipality-level maximums (for the capital district) of 15 and 30, respectively. However, our ICEWS event count maximum in Figure 3 is now reduced from the previous capital district-maximum (of over 2,000) to a new capital district-maximum of roughly 600 events. This suggests that a focus on FARC-perpetrated events ensures a clearer comparison across these three event data sources. That being said, for all of these comparisons, it does appear that a degree of rebel-perpetrated event overreporting within Bogotá—for GED and especially for ICEWS—is arising. Access to more information regarding ICEWS' geolocation process and geolocation assignments may help further filter out this overreporting in the ICEWS-context. Without such information, a focus on the FARC-perpetrated events, alongside the scaling approach presented here, is likely to best ensure that our ICEWS, GED, and CINEP data are comparable for our evaluations.

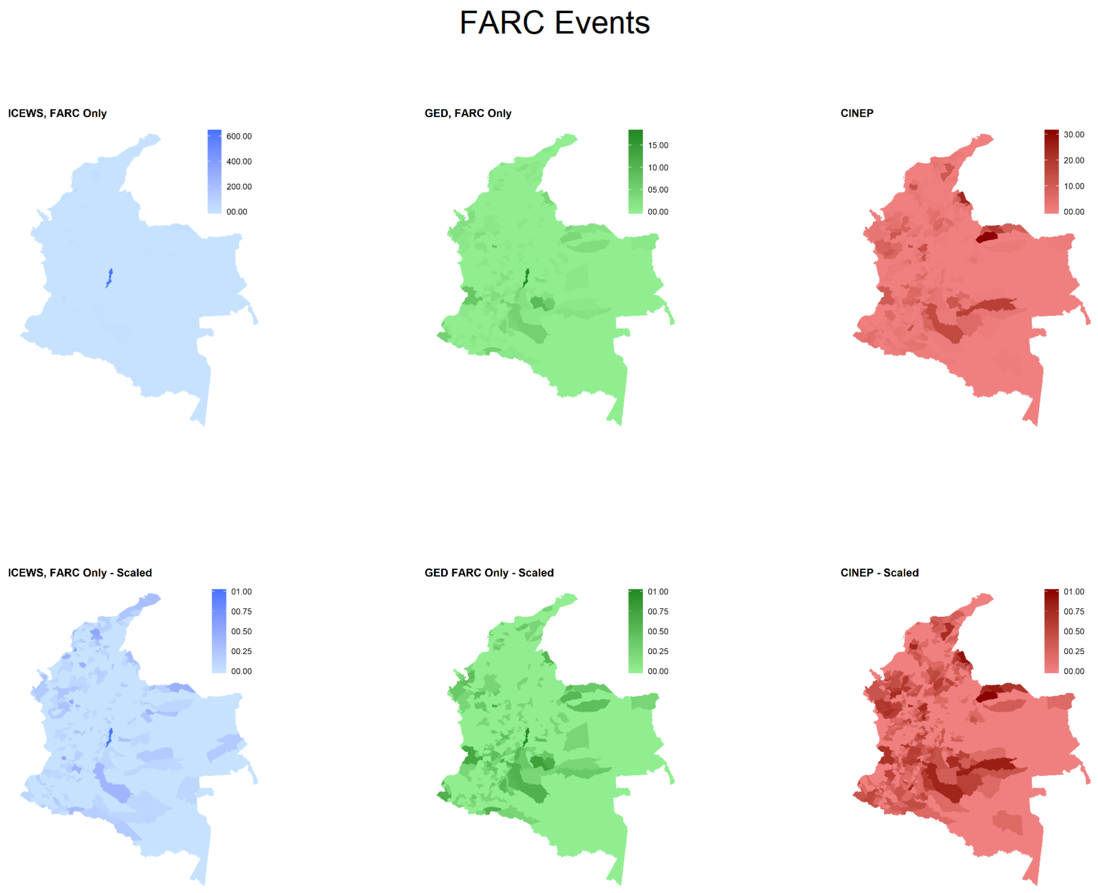
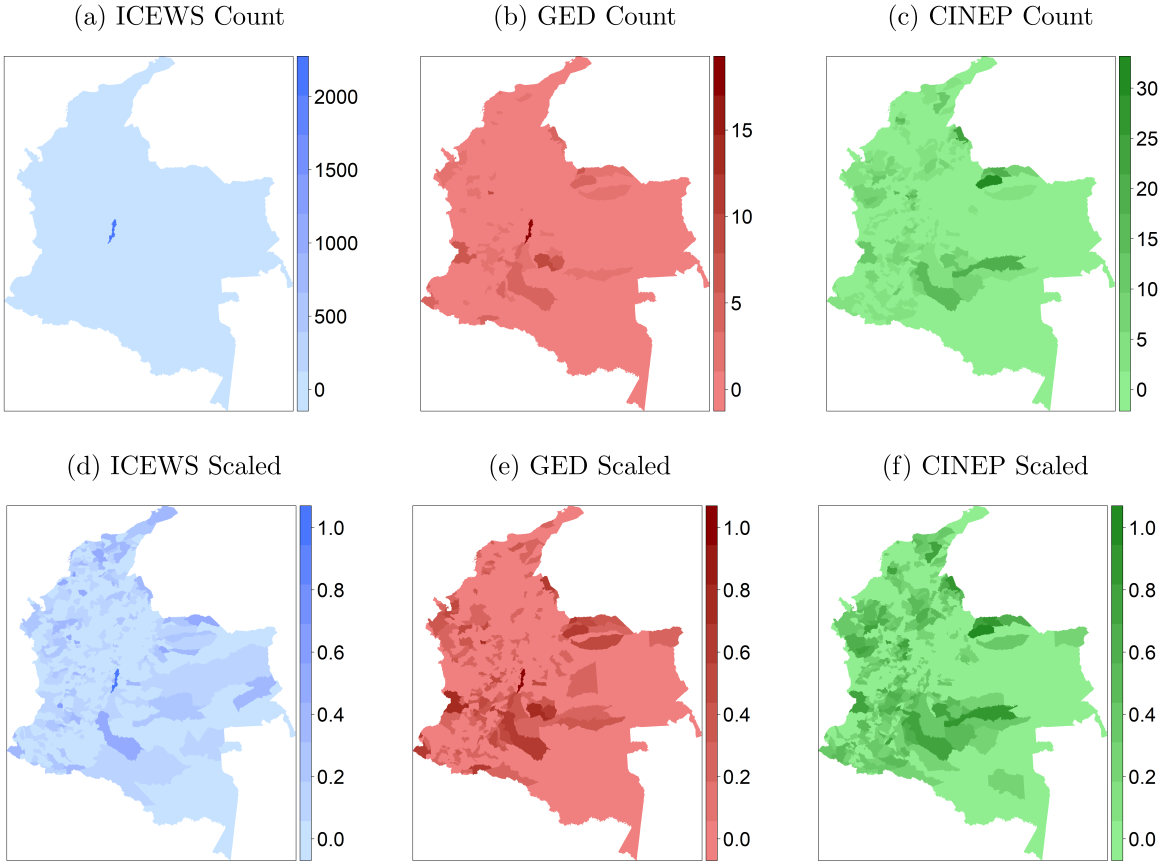


Figure 3: Comparisons of All Rebel-Perpetrated Events (Only FARC Actors)

3 Appendix

Figure 4: Comparisons of All Rebel-Perpetrated Events (Incl. ICEWS Unknown Actors)



4 References

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.

Boschee, E., J. Lautenschlager, S. O'Brien, S. Shellman, J. Starz, M. Ward, 2015. "ICEWS Coded Event Data." `https://doi.org/10.7910/DVN/28075` Harvard Dataverse, V25.

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

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

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.

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.

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

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.

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