Are Many Sets of Eyes Better Than One?

Evaluating Multiple Databases of Armed Actors in Colombia

Javier Osorio*
University of Arizona

Abstract

In contrast to the pervasive scarcity of disaggregated data affecting sub-national conflict studies, Colombia concentrates a wealth of databases measuring armed actors. How comparable are these databases? What are the implications of their differences for statistical inference? This research compares seven prominent sub-national measures of armed actors in Colombia. Using the Jaccard Similarity Index, the analysis reveals low similarity between measures. At best, results show 28.7% similarity when considering aggregated actor types, but similarity drops to 14.4% when considering specific armed groups. These measures also yield diverging statistical results when used as dependent or independent variables. In addition to their conceptual and methodological differences, pervasive missing data seem to be driving estimate discrepancies. The nuances of these measurement sets make it difficult to categorically determine if this low similarity is an asset or a limitation for empirical research. Yet, the analysis provides clear prescriptions for researchers in data-abundant settings.

Keywords: Colombia; similarity; measurement; paramilitaries; insurgents

Author's ORCiD number: 0000-0002-4548-7794
Author's Twitter handle: `@javier_osorio`

*Javier Osorio, Assistant Professor, School of Government and Public Policy, University of Arizona. Email: `josorio1@airzona.edu`. Address: 125 Social Sciences Building, Tucson, AZ, 85721-0027.

The research reported here was funded in part by the Minerva Research Initiative (OUSD [RandE]) and the Army Research Office/Army Research Laboratory via grant W911-NF-17-1-0569 to George Mason University.

Introduction

The quantitative analysis of the micro-dynamics of conflict requires high-quality, disaggregated, and accurate data about the characteristics and behaviors or armed actors to enable precise description, valid inferences, and knowledge accumulation. To build confidence that the results are not artifacts driven by particular modeling strategies or data choices, researchers generally conduct robustness tests using different models and alternative measures (Neumayer and Plümper 2017; Weisber 2006). In contrast to the robustness tests firmly rooted in country-year conflict research (Sambanis 2004; Hegre and Sambanis 2006), sub-national analyses are often limited to a single database or variable, thus hindering the possibility of conducting robustness tests. Despite efforts to enrich sub-national conflict data repositories (Zhukov et al. 2019), data scarcity remains a considerable obstacle to study the micro-dynamics of violence (Eck 2012). Lack of sub-national data prevents knowledge accumulation based on robust findings, thus leaving researchers with the uneasiness of accepting results based on limited data or convincing robustness tests.

In contrast to the scarcity of sub-national data in many research environments, the wealth of conflict data on Colombia offers a unique opportunity to assess different sub-national databases. How comparable are these measures? What are the consequences of their differences for statistical analyses? This research evaluates the Colombian conflict data environment by comparing seven different databases and provides four main contributions. First, it evaluates substantive differences between measures and discusses their limitations. Second, it focuses on armed actor presence as a commensurable indicator to compare their similarity using the Jaccard Index. Third, to prevent missing data from distorting similarity scores, the study presents a dimensionality-reduction algorithm to accurately estimate Jaccard similarity. Finally, it evaluates the implications of using different databases for statistical inference.

The study compares measures of guerrilla and paramilitaries as two generic types of actors, and further disaggregates them into specific organizations: the Revolutionary Armed Forces of Colombia (*Fuerzas Armadas Revolucionarias de Colombia*, FARC), the National

Liberation Army (*Ejército de Liberación Nacional*, ELN), and the United Self-defense Forces of Colombia (*Autodefensas Unidas de Colombia*, AUC). The study compares the substantive and methodological characteristics of seven databases. The pairwise and aggregate comparisons indicate that the proliferation of databases is not a panacea as results reveal considerable discrepancies between databases. At best, the Jaccard Index reports 28.7% similarity for guerrillas and 27.5% similarity for paramilitaries. Agreement is even lower for specific groups such as ELN, with only 14.4% similarity, followed by FARC with 15.3%, and AUC with 23.9% similarity. Moreover, the statistical analysis reveals that these distinct databases yield varying statistical results.

A nuanced interpretation of the low similarity across databases makes it difficult to categorically determine if these differences are an asset that help analyzing different conflict dimensions or a shortcoming that prevents conducting meaningful robustness tests. In any case, the analysis shows the importance of researchers systematically assessing the similarity between measures to advance their research, replicate findings, or collect new data.

Measuring in the Midst of Conflict

Advancing quantitative research on the micro-dynamics of violence relies on the availability of high-quality data collection efforts at the sub-national or individual level. As researchers depart from country-year studies and analyze conflict in a disaggregated manner (Kalyvas et al. 2008), micro-level studies of conflict often face the problem of lack of available data (Eck 2012). With the gradual progression of the field, scholars have been producing temporally and geographically disaggregated databases to analyze the micro-dynamics of conflict.

Developing databases in the midst of conflict is not a trivial endeavor. Violence may distort, suppress, destroy, or limit access to sources, documents, or testimonies (Weidmann 2016; Bell-Martin and Marston 2019; Osorio 2013). Coding data may be subject to coder bias, or methodological and logistical challenges (Baumgartner et al. 1998), and investigators may suffer psychological and ethical burdens (Loyle and Simoni 2017). Research on data quality has identified problems of geo-location, lax actor attribution, over or under-reporting, coding ambiguity or rigidity, and challenges with multi-label categories (Donnay et al. 2019; Eck 2012; Ward et al. 2013; Douglass and Harkness 2018; Hammond and Weidmann 2014; Weidmann 2016). These problems are more acute when databases rely on a single information source (Davenport and Ball 2002; Earl et al. 2004).

To address some of these concerns, the standard recommendation is to use multiple information sources (Davenport and Ball 2002). Just as a database benefits from integrating several information streams, a field of study benefits from having multiple databases. In line with this idea, conflict scholars have produced subnational-level databases using dedicated coders (Salehyan et al. 2012; Raleigh et al. 2010; Hegre et al. 2018; Sundberg and Melander 2013) or computerized approaches (Schrodt 2006; Schrodt and Van Brackel 2013; Osorio et al. 2019). Some scholars favor one database over another (Eck 2012), while others develop large repositories of sub-national data for replication and robustness checks (Zhukov et al. 2019). Some prefer integrating data streams (Donnay et al. 2019) while others use multiple systems estimation (Lum et al. 2013). Unfortunately, different databases could yield mixed results. Although adversarial arguments are fundamental to knowledge accumulation (Kuhn 1962), the lack of robust results leaves open questions and unconfirmed theories.

Measurement Sets

Taking advantage of the wealth of data on the Colombian conflict, this study compares seven prestigious databases often used to analyze armed actor presence or violent behavior. These databases include: (i) paramilitary presence by Rutas del Conflicto (2019), RC; (ii) paramilitary and guerrilla presence by Claudia López (2010), CL; (iii) reports of narcoparamilitary presence by Indepaz (2019), IN; (iv) paramilitary and guerrilla attacks by the Centro de Estudios Sobre Desarrollo Económico (2019), CD; (v) paramilitary and guerrilla violent presence from Violent Presence of Armed Actors, ViPAA (Osorio et al. 2019), VI; (vi) paramilitary and guerrilla attacks by Restrepo et al. (2004), RE; and (vii) paramilitary and guerrilla violent events from UCDP (Sundberg and Melander 2013), UC. Some measurements are produced by local researchers (RC, CL, IN, CD) and others by international scholars (UC and VI). See Appendix A1 and A7 for details.

Figure 1 compares these data along seven categories including the type of variable (dichotomous or count), information gathered from news agencies, NGOs, government, and testimonies, as well as their internal quality controls, and replicability. The analysis reveals considerable differences (see Appendix A1). RC presents dichotomous data on paramilitary presence using local and national newspapers, NGOs, and victims. CL data comes from a prestigious conflict study (López 2010) coding paramilitary and guerrilla presence (dichotomously) using local and national newspapers, and interviews. IN provides dichotomous data on paramilitary presence using local and national newspapers, local NGOs, and government records. CD presents count data on paramilitary and guerrilla attacks exclusively relying on government statistics from the Police, Army, and the Colombian census authority. VI provides count data on paramilitary and guerrilla violent presence using computerized coding (Osorio et al. 2019) based on *Noche y Niebla*, a collection of political violence and human rights narratives gathered from local and national news, NGOs, and victims, produced by CINEP (2016). RE presents count data on paramilitary and guerrilla attacks manually coded from CINEP's *Noche y Niebla* using their own codebook. Finally, UC provides count data on organized violence events involving paramilitary and guerrilla based on international news and NGOs. The top row of Figure 1 presents the databases measuring armed actor presence and the bottom row are the databases measuring attacks.

[Figure 1 around here]

Although generally focused on the Colombian conflict, these databases have considerable conceptual and operationalization differences, and vary in the armed groups they include. There is only partial substantive overlap on their objects of study as some databases analyze different actors or behaviors than others. This could constitute an advantage as it may enable studying distinct conflict dimensions. However, partial overlap could also limit the feasibility of robustness tests. Despite their conceptual differences, databases seem to overlap on their information sources. With the exception of CD, all other databases use a combination of news and NGO reports; however, few rely on testimonies. Unfortunately, there are noticeable discrepancies in internal quality control and replicability. See Appendix A1.

These measures also have commensurability problems that hinder comparison. While IN, CL, and RC use dummy variables, VI, RE, UC, and CD use count data. To address this incommensurability, the similarity analysis in this study transforms count variables into dichotomous measures taking the value of 1 whenever count data reports non-zero values. Although this minimalist approach reduces the data variation, it enables the comparison across databases. The transformed dummy variables could be interpreted as indicators of armed actor's *violent presence*. However, since violence is a limited proxy of armed actor presence (Arjona 2011), particularly for groups holding monopolistic control, measures of violent presence of armed actors require careful interpretation.

Figure 2 reveals pervasive missing data affecting some measurement sets. Panel (a) shows that VI is the most complete database by covering the entire 1988-2017 period, followed closely by UC, with one year less. All other measures have truncated or sporadic coverage. Unfortunately, there is not a single year in which all databases overlap, which prevents comparing all metrics in a concurrent period. The exploration also reveals missing municipalities in some databases. Panel (b) shows the extent of overall missingness by type of actor per measurement set. VI is the least affected by missingness while RC's guerrilla suffers from the most missing data. Assessing the reasons for missing data is beyond the objectives of this study (Fariss 2014), but missingness posts considerable difficulties for data analysis.

[Figure 2 around here]

Similarity Assessment

To assess the consistency between measures, the study relies on the Jaccard Similarity Index (Jaccard 1901, 1912), a score often used in ecology to evaluate the similarity of measures tracking species in a study site (Ricotta and Pavoine 2015). Although Jaccard is gaining popularity as an evaluation metric (Niwattanakul et al. 2013; Fletcher and Isla 2018), its use in political science still is limited (e.g. Sanger and Warin 2019). The intuition behind Jaccard Similarity considers a measurement set, M, as a vector recording the presence of an entity (e) in location i at time t, such that $M = \{e_{it}, \ldots, e_{NT}\}$, where e takes the value of 1 if the entity is present, and 0 otherwise. If there is no measurement effort in a location-time, then the observation is recorded as missing, $e_{it} = \text{NA}$, since it is not possible to determine the entity's presence or absence. Now, consider a comparison set, C, containing multiple measurement sets, $C = \{M_1, M_2, \ldots, M_{\overline{N}}\}$.

The Jaccard Similarity Index (J) evaluates the similarity between two measurement sets, $M_1 = (e_{1it}, \ldots, e_{1NT})$ and $M_2 = (e_{2it}, \ldots, e_{2NT})$ by ranging from 0 to 1, where 0 indicates total dissimilarity and 1 perfect similarity. For a single time point t, the local Jaccard Similarity of two vectors is:

$$\mathrm{J}_t(M_1, M_2) = \frac{M_1 \cdot M_2}{M_1^2 + M_2^2 - (M_1 \cdot M_2)} = \frac{\sum_{i=1}^{n} e_{1i}e_{2i}}{\sum_{i=1}^{n} e_{1i}^2 + \sum_{i=1}^{n} e_{2i}^2 - \sum_{i=1}^{n} e_{1i}e_{2i}} \tag{1}$$

More intuitively, Jaccard Similarity is the ratio of the intersection between observations and their union:

$$\mathrm{J}_t(M_1, M_2) = \frac{|M_1 \cap M_2|}{|M_1 \cup M_2|} \tag{2}$$

Based on equation 2, it is possible to generate an Average Jaccard Index for two measurement sets across an entire time frame:

$$\overline{\mathrm{J}} = \frac{\sum_t^T \mathrm{J}_t(M_1, M_2)}{T} \tag{3}$$

The Jaccard Similarity Index can be extended to compare any number of measures in a given comparison set, C, which includes $C = \{\text{RC, IN, CL, CD, VI, RE, UC}\}$ in this study. However, as Figure 2 shows, missingness is a pervasive problem in some measures. Imputing missing data with $e_{it} = 0$ would erroneously assume that a measurement took place but the entity was not detected, thus distorting the similarity score. Zero-imputation could inflate similarity if other observations already have zeros or if the imputed zeros fill in multiple measurement sets. Zero-imputation could also reduce similarity if the zero-imputed values contradict observations marked with 1.

Rather than making about the missing data generation process, the study proposes a dimensionality-reduction algorithm to calculate Jaccard Similarity in the presence of missingness. The algorithm (detailed in Appendix A2) considers time (t), location (i), and measurement sets (M) as relevant dimensions to calculate the Jaccard similarity. The algorithm uses the available M and i in a given t to generate data subsets (s) and calculate their local Jaccard scores (J_{ts}). If missing data alters the subsets' dimensions in a given year, the algorithm calculates the weighted average local Jaccard using the proportion of observations in each subset ($w_s = n_s/N$), such that $J_{tw} = (\sum_s^S J_{ts}w_s)(1/S)$. If there are no subsets in a given year, the algorithm directly calculates the local Jaccard (J_t). After computing the local Jaccard scores for each t, the algorithm generates an aggregated Jaccard ($\overline{\mathrm{J}}$) averaging local similarity scores over time ($t = 1, ..., T$). In this way, the algorithm only uses data overlapping at common points in time without being affected by missing data.

Pairwise Similarity Assessment

The similarity assessment applies the Jaccard dimentionality-reduction algorithm on two levels. First, the type-level analysis focuses on *Paramilitaries* or *Guerrilla*, including all the groups classified in each database as paramilitaries or insurgents. The type-level comparison is largely forgiving as it only requires identifying the actor type without agreeing on the specific group. The second level analyzes three specific armed actors: the FARC and ELN guerrilla groups, and the AUC paramilitaries. The group-level assessment is more stringent as it requires detecting the same armed organization per municipality-year.

Figure 3 presents the Average Jaccard Similarity by actor type and specific groups for each measurement pair. The solid square groups databases measuring attacks, while the dashed square clusters databases measuring presence. Overall, the pairwise matrices reveal low similarity between measures. Across panels, similarity is considerably lower between data measuring presence (dashed clusters) than between data measuring violence (solid clusters). At the type-level, the highest similarity of paramilitary and insurgent groups is between VI and CD, each with 34% similarity (Panels a and b). At the actor-level, similarity is even lower. The highest FARC similarity is between CD and RE with only 26% (Panel c). Similarity of ELN reaches only 21% agreement between VI and CD (Panel d). Finally, VI and CD are the most similar measures of AUC with 34% similarity (Panel e).

[Figure 3 around here]

Databases CD and VI report the highest level of similarity in four out of five comparisons, yet they only agree in about three out of ten observations. The ontological and methodological differences between CD and VI could explain these differences (see Figure 1). On one hand, CD reports armed attacks by paramilitary or guerrilla groups based on government records. In contrast, VI reports violent presence of armed actors based on CINEP narratives of political violence and human rights violations coming from national and local news, and testimonies. Substantively, violent attacks measured in CD could be considered a subset within VI's broader conceptualization of violent presence that includes many other behaviors such as threats, kidnapping, rape, displacement, etc.

Similarity Across Measures

This section analyses Average Jaccard Similarity across measures rather than by pairs. The type-level assessment in Panel (a) of Figure 4 shows low similarity across databases with 28.7% for guerrillas and 27.5% for paramilitaries. The actor-level analysis in Panel (b) reveals even lower similarity across measures. The Average Jaccard Similarity for AUC is 23.9%, for ELN is 14.4%, and for FARC is 15.3%. Despite the decades-long centrality of these armed groups in the Colombian conflict, there is little agreement about these actors.

[Figure 4 around here]

Figure 5 reports Local Jaccard scores over time by actor type and group, with darker color indicating low similarity and abbreviations for the databases included in each year. Panel (a) shows variation in guerrilla similarity with a minimum of 11.4% in 1989 and a maximum of 44.3% in 2000. In contrast, the paramilitary similarity is generally lower but fluctuates more markedly with a minimum of 6.4% in 2006 and a 56.1% peak in 2013. Panel (b) in Figure 5 shows even lower similarity at the group level. AUC similarity shows broad variation, ranging from 0% similarity between 1988 and 1992, up to 52.9% in 2015. ELN similarity is the lowest, ranging from 0% in 1988 to 52.4% in 2012. Finally, FARC similarity is also lower than AUC but slightly better than ELN, ranging from 3.3% in 1989 and 33.9% in 2002. Overall, measures tracking the presence of Colombian armed groups have low levels of agreement. These discrepancies make it difficult to determine patterns of territorial presence or behavioral trends for different groups.

[Figure 5 around here]

Overall, the descriptive analysis reveals low similarity across measures. These discrepancies are not necessarily problematic nor surprising if we consider the relatively different objects of study, operationalization, and information sources used in these databases. This measurement diversity can help enrich our understanding of the Colombian conflict by analyzing distinct aspects of it. However, the lack of comparable data makes it difficult to implement robustness tests to corroborate trends across databases.

Empirical Implications

This section evaluates the statistical consequences of using different measures of armed actors in two ways. First, it replicates the Dube and Vargas (2013) causal inference model using each database as the dependent variable. The second assessment uses these measures as independent variables to explain homicides in a correlational model. In addition to the seven measurement sets discussed above, this section includes an integrative variable labeled *All*, taking the value of 1 when any other databases detect an armed actor (by type or group), and 0 otherwise. In contrast to more sophisticated data aggregation methods (Donnay et al. 2019; Lum et al. 2013), this integrative approach represents the simplest and most straight-forward amalgamation of individual measures.

Armed Actors as Dependent Variable

Dube and Vargas (2013) estimate the effects of oil and coffee price shocks on armed conflict using instrumental variables. Their paper analyzes paramilitary and guerrilla attacks as count data, but this study uses the dichotomous measures of actor presence mentioned above. Based on their study, oil shocks should increase armed actor presence by promoting predatory behavior. In contrast, coffee shocks should reduce armed actor presence through a labor substitution process diverting combatants away from fighting and into coffee production. This analysis replicates the Dube and Vargas (2013) model by substituting each measure of armed actor presence as the dependent variable using the following second-stage specification:

$$y_{irt} = \lambda(Oil_{rt} \times OP_t) + \rho(\widehat{Cof_{it} \times CP_t}) + \gamma Coca_{irt} + \phi X_{irt} + \alpha_i + \beta_t + \delta_{rt} + \epsilon_{irt} \quad (4)$$

where y indicates armed actor presence by type or group from each database in municipality i, region r, and year t; oil shocks are defined as oil production Oil_{rt} interacted with international oil prices OP_t; coffee shocks come from municipal coffee cultivation Cof_{it} interacted with domestic coffee prices CP_t as derived from the first stage; $Coca_{irt}$ indicates coca cultivation; X_{irt} represents controls; and the model includes several fixed effects. See Appendix A4.

The top row of Figure 6 presents the second-stage coefficients of oil and coffee shocks on armed actor types (Panels a and b) and specific organizations (Panels c-e), with the marker representing the sample size. Results indicate some variations with respect to oil shocks. Most paramilitary measures (Panel a) and all AUC indicators (Panel e) present the expected positive sign of oil shocks, but with different point estimates and only half of the models reaching significance. Contrary to the expectation, most guerrilla measures (Panel b) present negative effects of coffee shocks and half of them are statistically significant, while FARC and ELN (Panels c and d) show mixed coefficient signs with varying significance. Regarding the coffee substitution mechanism, coffee shocks have consistently negative effects on armed actor presence (Panels a-e), but coefficient magnitudes and their significance vary considerably.

[Figure 6 around here]

The pairwise comparison above indicates that CD and VI are the most similar databases. A closer inspection of these data sets in Figure 6 shows that the coefficients associated with CD and VI present some inconsistencies. The negative sing of the coffee shock coefficients in VI and CD remain largely consistent across Panels a-e. In contrast, the oil shock coefficient of VI is positive and statistically significant in most cases in Panels a-e; however, the CD oil coefficients flip sings across Panels a-e and only reach significance in two out of five models. These varying results indicate that event the most similar pair of databases could lead to statistically inconsistent results. See Appendix A4 for details.

Armed Actors as Independent Variable

This section evaluates different measures of armed actor presence as independent variables used to explain homicide rates. The model uses the following specification:

$$y_{it} = \alpha_i + \beta A_{it} + \delta X_{it} + \epsilon_{it} \tag{5}$$

where y_{it} is the homicide rate reported by the Colombian Police (Moreno 2014); A_{it} refers to the actor type or group by measurement; X_{it} are the controls described in Appendix A5; α_i are municipal fixed effects, and ϵ_{it} the disturbances.

The bottom of Figure 6 reports the effects of different measures of armed actor presence on homicide rates with the marker size representing the sample size. See Appendix A5 for details. In general, Panels f-j consistently show that armed actors' presence increases violence. However, coefficient magnitudes vary broadly. Panel (f) shows that the UC paramilitary measure is associated with the highest homicide rate of 52.1, while IN attributes paramilitaries a homicide rate of only 2.6. Similarly, Panel (g) shows broad discrepancies for guerrillas. Again, UC reports the strongest effect of guerrilla with a homicide rate of 25.1, while CL reports a homicide rate of 10.4. Coefficient discrepancies are even sharper for specific group estimates. Panel (h) shows that UC attributes AUC a coefficient of 52.1 homicides, while the variable All is associated with a 5.01 homicide rate. According to Panel (i), RC has the strongest effect of ELN with a homicide rate of 86.3, however it has the smallest sample size and widest confidence intervals. In contrast, All has the smallest effect of ELN on violence with 10.4. Finally, Panel (j) associates RC's measure of FARC with the strongest effect on homicides and All with the smallest effect, 45.2 and 14.7, respectively.

A closer look at CD and VI, the pair of databases with the highest similarity score, reveals that the sign and statistical significance of their coefficients is consistent across Panels f-j in Figure 6. However, the magnitude of the VI estimates is always larger than CD's coefficients. See Appendix A5 for details.

In general, different measures of armed actor presence yield distinct results. Using different metrics as the dependent variable to replicate the Dube and Vargas (2013) study fails to support the expected positive effect of oil shocks and offers limited confirmation of the negative effect of coffee shocks. Similarly, using different measures as independent variable provides limited agreement for the positive effect of armed actors on homicides due to wide-varying coefficients. The conceptualization and measurement differences of these databases could be driving result discrepancies. Moreover, given the pervasiveness of missing data in some measures, gaps in the data coverage are probably a key reason for the disparities in the statistical results.

The logic of territorial control outlined by Kalyvas (2006) could offer a potential substantive interpretation of the diverging statistical results between measures of presence and measures of violence used either as dependent or independent variables. According to this expectation, armed actors are likely to display distinct patterns of violence depending on the degree of control they hold on a territory. In this way, estimate discrepancies could be expected between measures of violence and those of armed presence. Unfortunately, the Colombian data on armed actor presence is not measured categorically to capture degrees of territorial control. Without the right kind of empirical support, the above-mentioned interpretation could be no different than mere speculation.

Conclusion

Contrary to the data scarcity pervasive in the micro-dynamics of conflict research, the long duration of the Colombian conflict enabled the production of multiple databases on non-state armed actors, thus providing a rich data environment to study conflict. This study compares seven different measurement sets of armed actors in Colombia at the municipality-year level between 1988 and 2017. The study provides five main lessons.

First, the descriptive analysis reveals some ontological and methodological differences across measures, as well as a pervasive problem of missing data. S Second, the study advances a novel algorithm for calculating Jaccard similarity that prevents distortions from missing data. This similarity assessment generally shows low agreement across databases. Third, the analysis compares the of similarity between each pair of databases.Pairwise comparisons indicate that VI and CD are the most similar databases; however, they only overlap in about 3 out of 10 cases. Fourth, moving from pairwise to aggregate comparisons across databases shows that, at best, there is 28.7% of similarity for measures of guerrilla presence and 27.5% for paramilitary groups. Similarity is even lower when considering specific armed groups: similarity across AUC measures is 23.9%, for the ELN insurgency is 14.4%, and for FARC is 15.3%. It is remarkable that despite the centrality of these three armed organizations in the Colombian conflict, different databases display such low levels of similarity. Finally, the analysis also shows that using distinct measures as dependent or independent variables is consequential for statistical inference. The regression results show limited statistical consistency across database with frequent instances of flipping coefficient signs, different estimate magnitudes, and varying confidence intervals when using the different measures of armed actors.

A careful and nuanced analysis of the data makes it difficult to make a definite interpretation of the similarity across databases as an asset or a shortcoming for quantitative analysis. If we consider that these databases measure inherently distinct phenomena with minor overlap, then the low similarity and statistical consistency should not be surprising nor concerning. The descriptive and inferential differences could be the results of distinct databases measuring different behaviors. In such case, the ontological and operationalization discrepancies of these databases could be an asset that helps researchers analyze different questions. In contrast, if we consider that these measures capture substantively interrelated phenomena, then the low data similarity and lack of robust statistical results would be problematic. In consequence, lessons derived from disparate data would hardly lead to meaningful robustness tests and knowledge accumulation.

This study shows that despite the wealth of data in some settings, the availability of sub-national databases is not a panacea. Researchers studying conflict in data-abundant settings would benefit from evaluating the similarity of the databases. This research advances methodological guidelines on how to assess data similarity in a systematic way. Researchers in low-similarity data environments should be particularly careful in assessing and using different measures to conduct robustness tests. Swapping one database for another one is not a sound decision without a proper assessment of their substantive, methodological, and coverage characteristics. Similarity assessments could also guide future data collection efforts while trying to maximize the overlap between existing databases and new ones. Given the pervasive problem of missingness, researchers could also focus on filling historical gaps of existing databases and updating truncated ones while keeping in mind substantive and methodological similarities.

References

Arjona, A. M. 2011. Presencia vs. violencia: problemas de medición de la presencia de actores armados en Colombia.

Baumgartner, F., B. Jones, and M. MacLeod. 1998. Lessons from the Trenches: Ensuring Quality, Reliability, and Usability in the Creation of a New Data Source. *The Political Methodologist 8*(2): 1–10.

Bell-Martin, R. and J. Marston. 2019. Confronting Selection Bias: The Normative and Empirical Risks of Data Collection in Violent Contexts. *Geopolitics*: 1–34.

Centro de Estudios Sobre Desarrollo Económico 2019. Catálogo de Datos.

Centro de Investigación y Educación Popular 2016. Banco de Datos.

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.

Donnay, K., E. Dunford, E. McGrath, D. Backer, and D. Cunningham. 2019. Integrating Conflict Event Data. *Journal of Conflict Resolution 63*(5): 1337–1364.

Douglass, R. and K. Harkness. 2018. Measuring the landscape of civil war: Evaluating geographic coding decisions with historic data from the Mau Mau rebellion. *Journal of Peace Research 55*(2): 190–205.

Dube, O. and J. Vargas. 2013. Commodity Prices Shocks and Civil Conflict: Evidence from Colombia. *Review of Economic Studies 80*(4): 1384–1421.

Earl, J., A. Martin, J. McCarthy, and S. a. Soule. 2004, aug. The Use of Newspaper Data in the Study of Collective Action. *Annual Review of Sociology 30*(1): 65–80.

Eck, K. 2012. In data we trust? A comparison of UCDP GED and ACLED conflict events datasets. *Cooperation and Conflict 47*(1): 124–141.

Fariss, C. 2014. Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability. *American Political Science Review 108*(2): 297–318.

Fletcher, S. and M. Z. Isla. 2018. Comparing sets of patterns with the Jaccard index. *Australasian Journal of Information Systems 22*: 1–17.

Hammond, J. and N. B. Weidmann. 2014, jul. Using machine-coded event data for the micro-level study of political violence. *Research & Politics 1*(2): 2053168014539924.

Hegre, H., M. Croicu, K. Eck, and S. Hogbladh. 2018. Introducing the UCDP-Candidate Events Dataset and the ViEWS Outcomes dataset.

Hegre, H. and N. Sambanis. 2006, aug. Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution 50*(4): 508–535.

Indepaz 2019. Informe sobre presencia de grupos Narcoparamilitares.

Jaccard, P. 1901. Distribution de la flore alpine dans le Bassin des Drouces et dans quelques regions voisines [Distribution of alpine flora in the Bassin des Drouces and in several neighboring regions]. *Bulletin de la Société Vaudoise des Sciences Naturelles 37*(140): 241–272.

Jaccard, P. 1912. The Distribution of the Flora in the Alpine Zone. *New Phytologist 11*(2): 37–50.

Kalyvas, S. N. 2006. *The Logic of Violence in Civil War*. New York: Cambridge University Press.

Kalyvas, S. N., I. Shapiro, and T. Masoud. 2008. *Order, Conflict and Violence*. Cambridge: Cambridge University Press.

Kuhn, T. 1962. *The Structure of Scientific Revolutions*. Chicago: Chicago University Press.

López, C. (Ed.) 2010. *Y Refundaron la Patria: De cómo Mafiosos y Políticos Reconfiguraron el Estado Colombiano*. Bogotá, Colombia: Corporación Nuevo Arcoíris.

Loyle, C. and A. Simoni. 2017. Researching Under Fire: Political Science and Researcher Trauma. *Political Science & Politics 50*(1): 141–145.

Lum, K., M. E. Price, and D. Banks. 2013. Applications of multiple systems estimation in human rights research. *American Statistician 67*(4): 191–200.

Moreno, J. 2014. Tasas de homicidios en Colombia.

Neumayer, E. and T. Plümper. 2017. *Robustness Tests for Quantitative Research*. New York: Cambridge University Press.

Niwattanakul, S., J. Singthongchai, E. Naenudorn, and S. Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *International Multi-Conference of Engineers and Computer Scientists*, Volume 1, pp. 380–384.

Osorio, J. 2013. Numbers Under Fire: The Challenges of Quantitative Data Gathering in Highly Violent Settings. *DSD Working Papers on Research Security*.

Osorio, J., M. Mohamed, V. Pavon, and S. Brewer-Osorio. 2019. Mapping Violent Presence of Armed Actors in Colombia. In *ICC 2019 Interntaional Cartographic Conference*, Tokyo, Japan, pp. 1–8. International Cartographic Association.

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.

Restrepo, J., M. Spagat, and J. Vargas. 2004. *The Dynamics of the Colombian Civil Conflict: A New Data Set*, Volume 21.

Ricotta, C. and S. Pavoine. 2015. A multiple-site dissimilarity measure for species presence/absence data and its relationship with nestedness and turnover. *Ecological Indicators 54*: 203–206.

Rutas del Conflicto 2019. Geografía del Paramilitarismo en Colombia.

Salehyan, I., C. S. Hendrix, J. Hamner, C. Case, C. Linebarger, E. Stull, and J. Williams. 2012. Social conflict in Africa: A new database. *International Interactions 38*(4): 503–511.

Sambanis, N. 2004. What Is Civil War?: Conceptual and Empirical Complexities of an Operational Definition. *Journal of Conflict Resolution 48*(6): 814–858.

Sanger, W. and T. Warin. 2019. Dataset of Jaccard similarity indices from 1,597 European political manifestos across 27 countries (1945-2017). *Data in Brief 24*(103907): 1–7.

Schrodt, P. and D. Van Brackel. 2013. Automated Coding of Political Event Data. In D. Subramanian (Ed.), *Handbook of Computational Approaches to Counterterrorism*, pp. 23–50. New York: Springer.

Schrodt, P. A. 2006. Twenty Years of the Kansas Event Data System Project. *The Political Methodologist 14*(1): 2–6.

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

Ward, M., A. Beger, J. Cutler, M. Dickenson, C. Dorff, and B. Radford. 2013. Comparing GDELT and ICEWS Event Data.

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

Weisber, M. 2006. Robustness Analysis. *Philosophy of Science 73*(5): 730–742.

Zhukov, Y. M., C. Davenport, and N. Kostyuk. 2019. Introducing xSub: A new portal for cross-national data on subnational violence. *Journal of Peace Research 56*(4): 604–614.

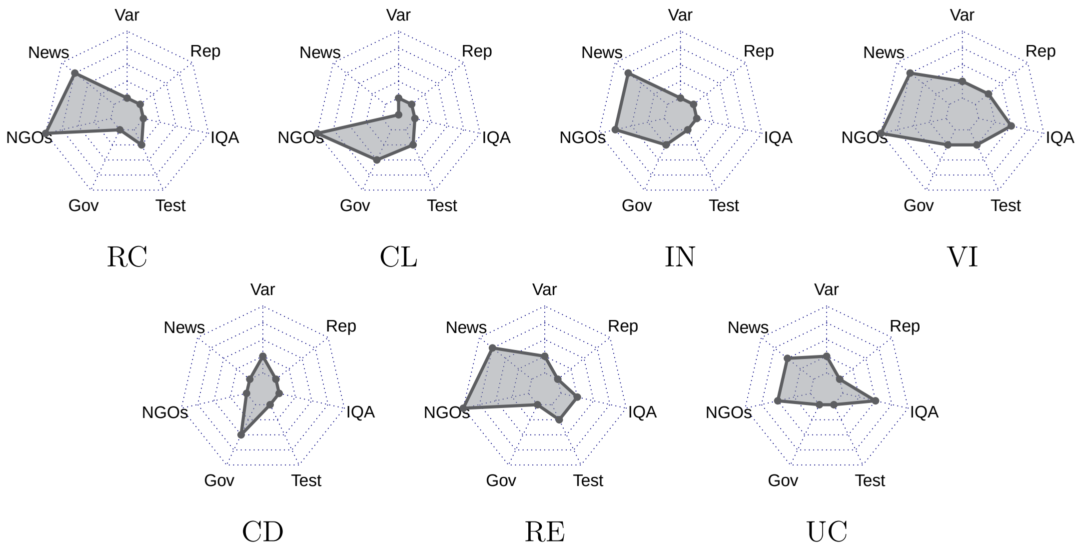
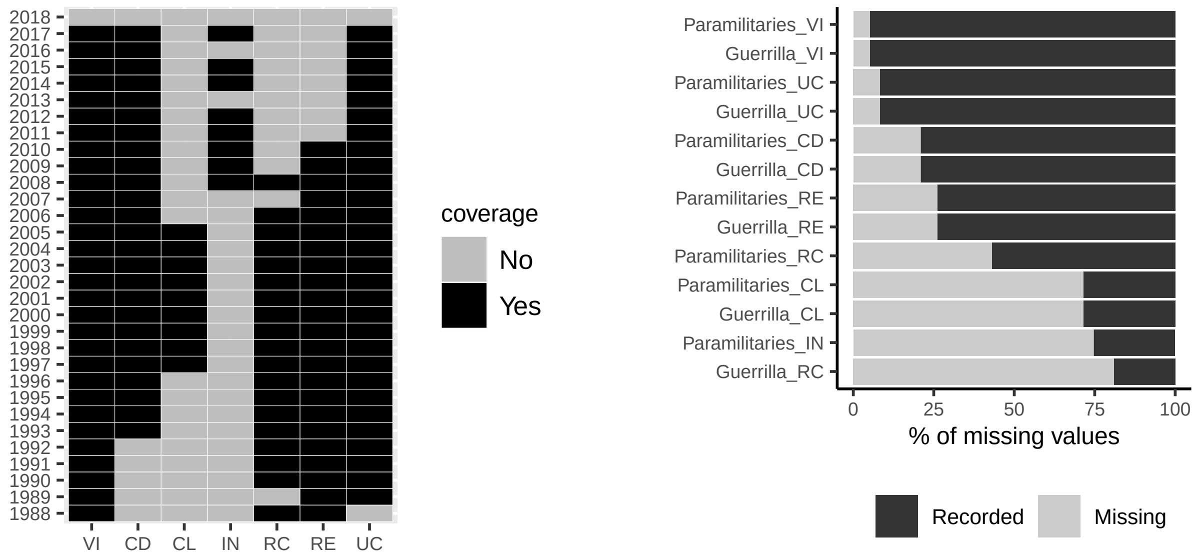


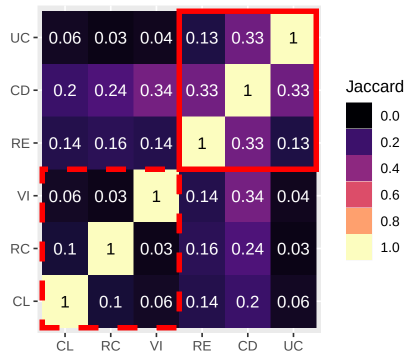
Figure 1: Measurement set characteristics



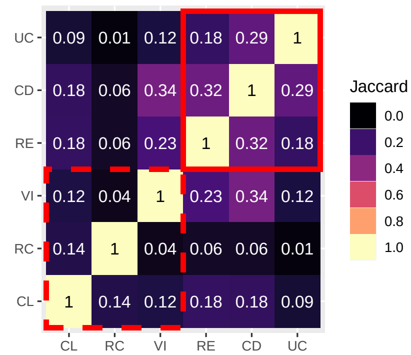
(a) Temporal Coverage

(b) Missing Data

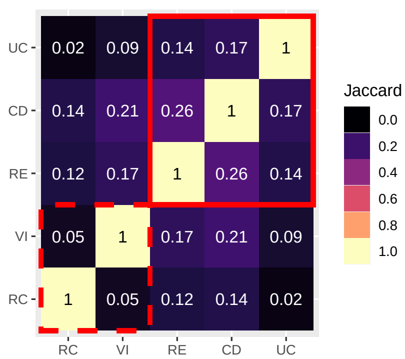
Figure 2: Coverage and Missingness



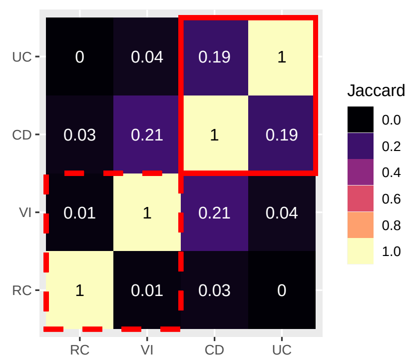
(a) Paramilitaries



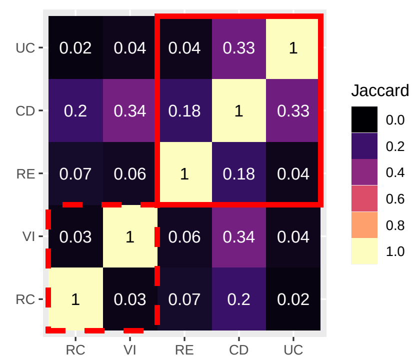
(b) Guerrilla



(c) FARC

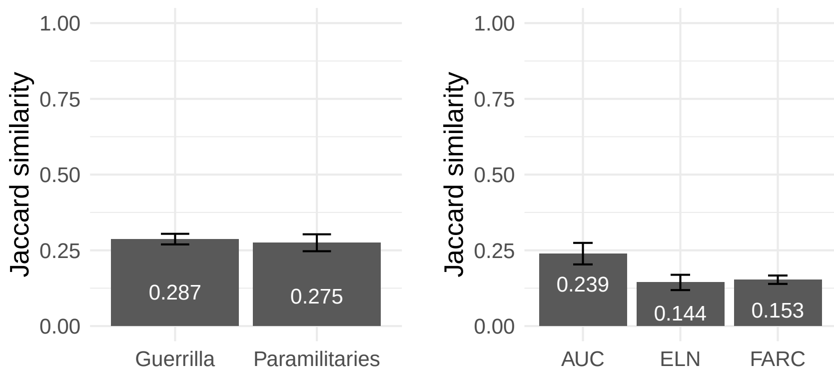


(d) ELN



(e) AUC

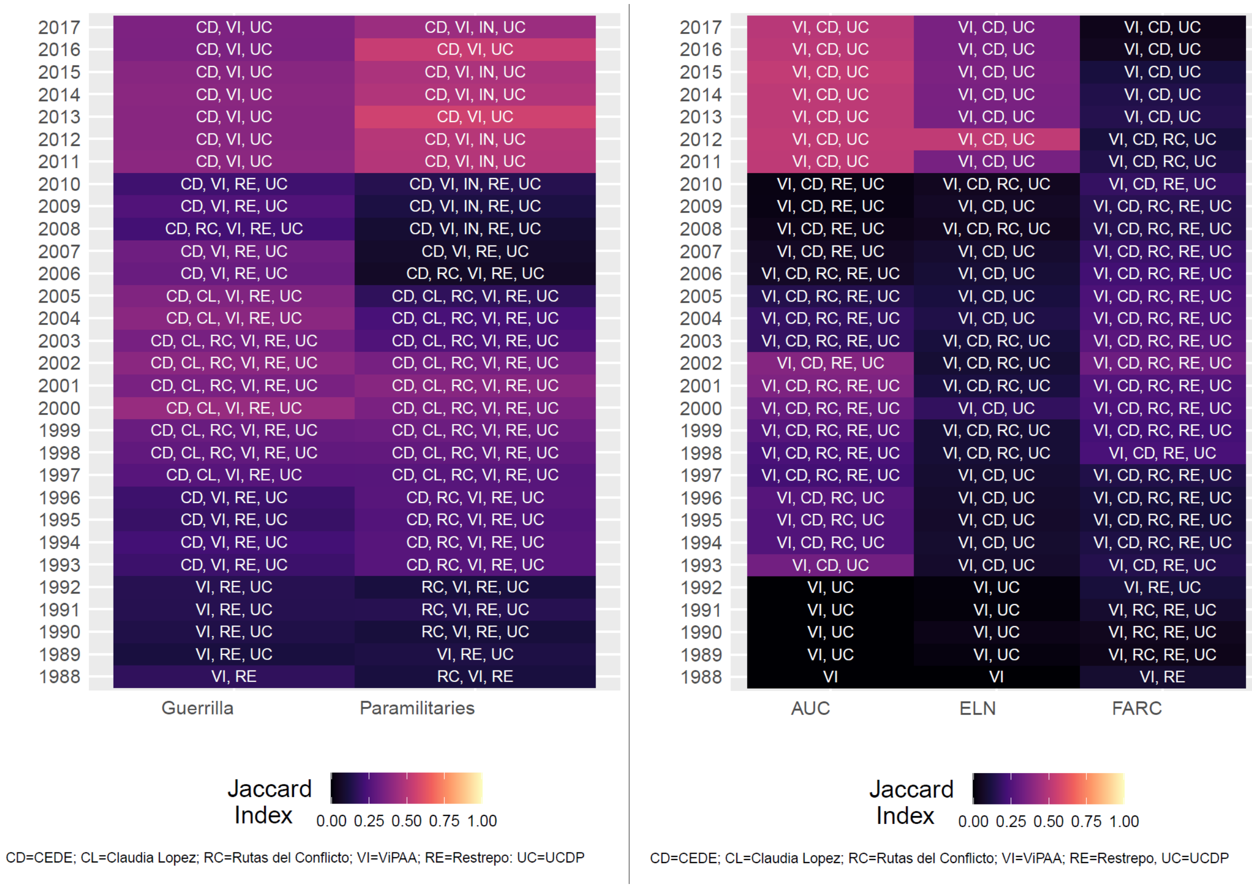
Figure 3: Pairwise Jaccard Similarity



(a) Actor Type

(b) Armed Group

Figure 4: Average Jaccard Similarity



(a) Actor type

(b) Armed groups

Figure 5: Local Jaccard Similarity

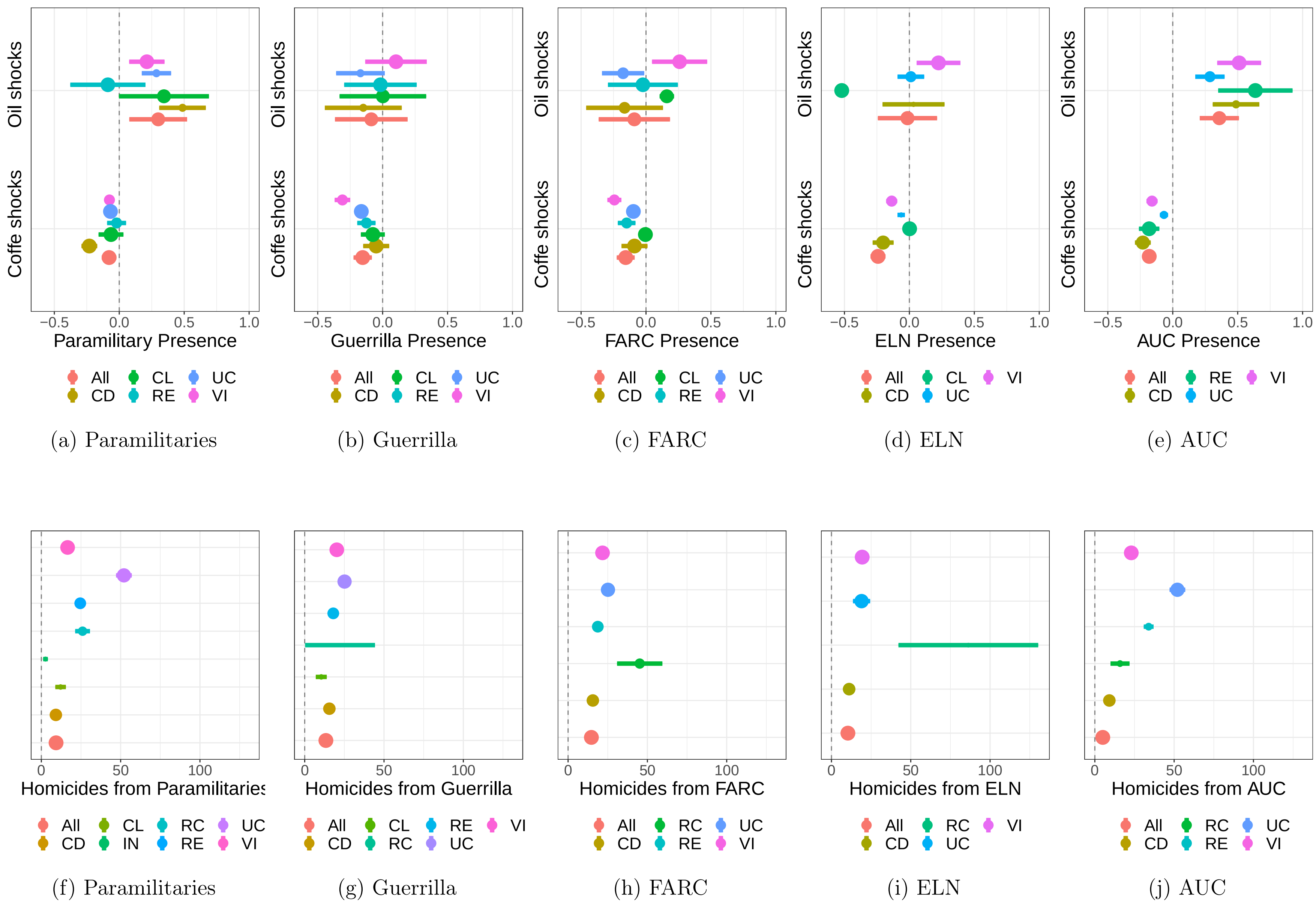


Figure 6: Predicted Effects Using Different Measures