An Investigation of How Conflicts End in Africa

Yan-Yu Chen* Ju-Sheng Hong†

Abstract

Developing effective strategies for conflict resolution and prevention requires a comprehensive understanding of how conflicts come to an end. This report presents a novel approach to clustering the conflicts that occurred in Somalia between January 1st, 2019, and February 10th, 2023 from Armed Conflict Location & Event Data (ACLED). Our method clusters conflicts by utilizing `Word2Vec` and `t-SNE` to adapt the sparse interaction in `actor`. By doing so, the categorical variable `actor` is transformed into a continuous variable and embedded into a low-dimensional space while preserving the interaction relationship. We also incorporate domain knowledge by weighting the location information twice to the others. The goodness-of-fit of our proposed clustering is evaluated by a F-statistics, which quantifies the importance of variables. Our implementation shows that the proposed method outperforms three baseline models, demonstrating its ability in clustering conflicts. Overall, our approach offers a valuable tool for gaining insight into conflict dynamics and informing conflict resolution and prevention strategies.

1 Introduction

Conflicts are a means of rationally resolving disputes over resource allocation (Fearon, 1995). Theoretical economic models provide some explanation of the relationship among the causes,

*This author is in charge of EDA and spectral clustering for the final model.

†This author is in charge of the `word2vec` model, `t-SNE` analysis, and comparison of different models.

behavior, and outcomes of violent conflict. One popular model suggests that actors use violence to achieve better outcomes than peaceful compromise, while an alternative model suggests that actors engage in violence to attribute it to randomness beyond their control. What's more interesting is that a new hypothesis suggests that violence can also be used to de-escalate situations and maintain stability (Abreu et al., 1990). To verify these models and hypotheses, multiple organizations collect datasets about armed conflict events worldwide (Raleigh and Dowd, 2015). Our initial exploratory work attempts to make the research questions concrete and quantitative and develop an informative investigation to assess those models and hypotheses. Understanding how conflicts ultimately end up helps develop effective strategies for conflict resolution and prevention.

One of our main contributions is to define a meaningful corpus of actors. During data preprocessing, we use `Word2Vec` and t-distributed stochastic neighbor embedding (`t-SNE`) to transform `actor` variable. Introduced by Mikolov et al. (2013), `Word2Vec` has since become one of the most popular neural network-based approaches for generating word embeddings in natural language processing and machine learning. The model works by training a neural network on a large corpus of text data, such as a collection of news articles or books, and using the network to predict the likelihood of a given word occurring in the context of other words in the same corpus and dense vector representations of words that capture their semantic and syntactic meaning. The other model, `t-SNE`, is a popular dimension reduction technique that has been widely used in machine learning and data visualization since its first appearance in Van der Maaten and Hinton (2008). By modeling the pairwise similarities between data points in the high-dimensional space and then mapping these similarities to the lower-dimensional space, `t-SNE` is particularly effective for visualizing high-dimensional data in a lower-dimensional space, such as the 2D or 3D space. In this report, `Word2Vec` is used to capture the meaning of `actor` in a 20-dimensional space, as `Word2Vec` works better when the embedding dimension is high (Mikolov et al., 2013). The issue of the curse of dimensionality, however, might appear when directly dealing with the 20-dimensional representation. We use `t-SNE` to further reduce the dimension to 3.

The rest of this report is organized as follows. We first review the previous works from the client in Section 2, and then define the research problem based on the client's domain interest in Section 3. The attribute of Armed Conflict Location & Event Data (ACLED, Raleigh and Dowd (2015)) is described in Section 4. In Section 5, we embed `actor`, establish a customized clustering that takes the importance of variables into account, and evaluate the clustering. Lastly, we demonstrate the advantage of our clustering to the three baseline models in Section 6.

2 Background

The previous works on ACLED utilized variables such as approximate longitude and latitude, and occurrence time of events. At the beginning of prepossessing, an equidistant projection is applied to minimize distortion to distances. The clustering algorithm, ordering points to identify the clustering structure (OPTICS, Ankerst et al. (1999)), with only temporal and spatial information does not guarantee a desirable result. Previous literature indicates that extra work in exploratory data analysis to incorporate `actor` variable is in need. However, extracting information from the category variable, `actor`, is challenging. The enemies network showed that several events have only a single actor. Meanwhile, some events have multiple actors. Additionally, the information from the actors may overlap with information from the temporal and spatial features. A proper clustering algorithm shall be chosen for this specific task. The preliminary investigation suggests that we find (1) another clustering algorithm that is more suitable for ACLED; (2) a way to incorporate the information from the actor into event time.

3 Research Problem

In this study, we focus on analyzing ACLED to gain insights into conflict resolution. Our objective is to propose an approach for identifying clusters of violent incidents within the dataset. The client has indicated that a desirable cluster would group events occurring in similar locations, involving comparable actors, and lasting for a similar period of time. In particular, we have three specific goals: (1) establish clear criteria for an ideal cluster, grounded in domain-specific knowledge; (2) emphasize the location information when doing clustering; (3) evaluate the importance of variables of clustering.

Our contribution is to achieve these three goals. An ideal cluster should exploit all the information including the approximate longitude and latitude, occurrence time of events, and the actor's interaction, to help understand how conflicts come to an end. We create the neighborhood dataset from the temporal and spatial information for estimating duration and embedding `actor`. A row estimate of the duration of conflict events is proposed based on the grouped events in the neighborhood dataset. We embed `actor` into a low-dimensional space readily to do clustering while preserving the sparse interaction relationship by utilizing `Word2Vec` and `t-SNE`. As client's request, we also emphasize the location information by weighting it twice to the others during the clustering process. Last, the goodness-of-fit of our proposed clustering is evaluated by a F-statistics, which quantifies the importance of variables. Our proposed method is shown to be superior in clustering conflicts and beneficial to gain insight into conflict dynamics.

4 Data Description

Containing around 300,000 samples with 30 attributes, ACLED itself is a complex, observational dataset with features across temporal and spatial. Previous literature revealed the importance of preprocessing the actor variable and studying their inner structure. With a preliminary study revealing 7,246 distinct actors in the dataset, a scatter plot matrix (Figure 1) displays the interaction between actors, with the size of the dots representing the number of fights, the areas enclosed by red and blue lines denoting conflicts involving ”Military” and ”Civilian” actors respectively, and a vertical strip represents attackers and a horizontal strip represents defenders. We observe that (1) the ”Civilian” is attacked by almost all the other actors and they barely start any conflicts and (2) the ”Military” start conflicts most of the time but they also get attacked. Moreover, the network of actors is very sparse, out of around 45,000,000 possible pairs of actors, only 0.038% of them were involved in the same conflicts.

5 Methodology

We utilize the spectral clustering algorithm to identify desirable conflict clusters from various variables, including space, time, and actors from ACLED. Spectral clustering has emerged as a popular technique for clustering data points from non-convex and high-dimensional clusters (Donath and Hoffman, 1973; Ng et al., 2001). It leverages the spectral properties of a similarity matrix computed from the data points to group them into distinct clusters. We incorporate the time and actor variables to the event duration and actor embedding vectors from both dyads, which are not directly available in ACLED. For those additional variables, the neighborhood dataset is introduced in Section 5.1. We then create event duration and embed actors based on that in Section 5.2

5.1 Create Neighborhood Dataset

Let conflict A be an event in ACLED. We define the event B to be in the neighborhood of event A if two conditions are satisfied: (1) event B occurs after event A and (2) event B takes place in a location and time that are close to those of event A. We calculate an additional variable, called `duration` of conflict A, by finding the median interval, measured in days, between conflict A and the events in its vicinity. To illustrate this concept, we use a subset of ACLED data shown in Table 1, and group conflicts that satisfy (1) and (2) with the occurrence within 30 kilometers and 30 days to create the neighborhood dataset presented in Table 2. Each row in Table 2 represents a pair of events, involving four actors in total. For instance, the duration of conflict 1 is 13.5, which is the median of 4 and 23, and is stored along with the embedding vectors of `actor`, as shown in Table 4. The neighborhood dataset provides us with not only the `duration` variable but also the corpus text data of actors.

We introduce a specific function called `CreateNeighborhood(inputPath, outputPath = NULL, distance = 30, timeSpan = 30)` in `R`, which generates the neighborhood dataset. This function takes the path of the original dataset downloaded from www.acleddata.com (`inputPath`) as an input and produces the neighborhood dataset based on the criteria specified by `distance` (in kilometers, default value is 30) and `timeSpan` (in days, default value is 30). Furthermore, if provided, the function saves the resulting dataset in `outputPath`.

5.2 Embed `actor` Variable

The previous network analysis depicts the interaction between actors; yet, as we have mentioned, the sparsity of the enemy network makes network analysis difficult to use. One-hot encoding is inappropriate as well due to the lack of correlation between the encoding vectors. For instance, Table 1 contains 5 distinct actors. The one-hot encodings of those actors are

$$\begin{aligned}\text{Military} &= (1,0,0,0,0);\\ \text{SSC} &= (0,1,0,0,0);\\ \text{Solo} &= (0,0,1,0,0);\\ \text{Al-Shabaab} &= (0,0,0,1,0);\\ \text{Civilians} &= (0,0,0,0,1),\end{aligned}$$

and the correlation of any pair of the encoding vectors is 0, which is unreasonable since actors interact. To overcome this issue, we utilize `Word2Vec` to embed actors into Euclidean space. While `Word2Vec` typically takes sentence inputs, generating sentences of two words for each event does not yield satisfactory results due to the short sentence lengths. Instead, we define a sentence by including all actors involved in a particular event and its surrounding area in the neighborhood dataset, as outlined in Table 2. This approach generates the sentences listed in Table 3, which are used to train `Word2Vec`. In addition to the actor embedding, considering the duration of a conflict is crucial for clustering analysis, we utilize the neighborhood dataset once again. The processed data are categorized into three groups based on the variables: `location`, `actor`, and `time`, displayed in Table 4.

In the attached file, we define `formSentenceFromFile(inputPath = NULL, outputPath, data = NULL)` in `R` to create sentences from the neighborhood dataset. To use this function, either `inputPath` or `data` must be specified. If `inputPath` is used, it should contain the path of the neighborhood dataset, while if `data` is used, it should contain the output from `CreatNeighborhood()` function. The outputs generated by `formSentenceFromFile()` are stored in `outputPath`, if specified. Furthermore, we include a class called `Word2VecTrainer` in the file `trainer.py`. This class takes `inputPath` as the path of the sentence file output from `formSentenceFromFile()` function and saves the resulting embeddings in `outputPath`.

To train a `Word2VecTrainer` object, we can simply call the method `.train()`.

5.3 Importance of Variables

As domain knowledge suggests that some variables may be more important than others to suit the downstream application, we adjust the relative importance of `location`, `time`, and `actor` in our analysis by manipulating the standard deviation of each category. The standard deviation is an important ingredient because it significantly affects the results of spectral clustering, as the method relies on computing the similarity matrix from pairwise distances between data points. Since the Euclidean distance metric is sensitive to the scale of the data, variables with larger variances tend to have a greater impact on the calculation of pairwise distances. According to the domain knowledge from the client, we believe that `location` should have a greater impact than `time` and `actor`. Therefore, we set the standard deviation of `location` to 2, while keeping the standard deviations of `time` and `actor` at 1.

Scaling the variables yields clusters that align more closely with existing literature, and we can, furthermore, quantify the influence of each variable in a posterior analysis. For this pursuit, we use the F-statistics and define the contribution of each variable as

$$\exp\left[-\frac{\text{within group SSE}/(n-k)}{\text{between group SSE}/(k-1)}\right],$$

where n is the total number of conflicts, and k is the number of clusters. We calculate the importance of each variable as the percentage of the corresponding contribution. The results of our clustering analysis indicate that `location`, `time`, and `actor` account for 38.4%, 21.4%, and 40.2% of the clustering, respectively.

6 Result

In this section, we demonstrate the effectiveness of the proposed clustering and compare our method with three baseline models. Only models including `location` are considered in the comparison since `location` is relatively more important than the others. Table 5 summarizes all the models, which all use spectral clustering with 50 clusters but analyze the different subsets of variables. The upper plots in Figures 2–5 show the clustering results for conflicts that occurred in Somalia between January 1st, 2019 and February 10th, 2023. Each conflict is represented by a dot, with the color indicating its assigned group. Note that the color similarity does not provide meaningful information.

Although `Baseline1` initially appears to perform well as there is no overlap between clusters, further examination reveals that `Baseline1` fails to generate meaningful clusters regarding the `actor` and `time` variables (Table 6). Similar issues are observed with models `Baseline2` and `Baseline3`. To further investigate the information loss from ignoring the other variables, we focus on conflicts that occurred near Mogadishu. The lower plots in Figures 2–5 show that the clustering results with three groups selected for each model. Tables 6–9 provide examples of conflicts in each group under different models. In the lower plot of Figure 2 and Table 6, we observe that `Baseline1` results in the three groups being geographically separated but fails to capture the similarity of conflicts involving similar actors and duration. The overlapping duration of conflicts within each group undesirably prevents the downstream application of understanding how conflicts come to an end. Analogously, `Baseline2` in Table 7 and `Baseline3` in Table 8 result in failing to differentiate conflicts involving similar actors and involving similar duration, respectively. The information loss by ignoring the other variables undermines the possible interpretation of meaningful clusters of conflicts. In contrast, our proposed method in Figure 5 and Table 9, which considers all variables, is able to bring events that occurred in similar locations, involved similar actors, and lasted for a similar period of time together. The comparison among the clustering results showcases the superiority of our proposed method and guarantees that incorporating `actor` using our method in clustering conflicts helps gain a deeper insight into conflict dynamics.

References

Dilip Abreu, David Pearce, and Ennio Stacchetti. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica: Journal of the Econometric Society*, pages 1041–1063, 1990.

Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

William E Donath and Alan J Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5):420–425, 1973.

James D Fearon. Rationalist explanations for war. *International organization*, 49(3):379–414, 1995.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Andrew Ng, Michael Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

Clionadh Raleigh and Caitriona Dowd. Armed conflict location and event data project (acled) codebook. *Find this resource*, 2015.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Appendix

The code used is attached in a separated `.zip` file.

Figures & Tables

	location	date	actor 1	actor 2
event 1	0	01/01/2022	Military	SSC
event 2	26	01/05/2022	Military	Solo
event 3	20	01/24/2022	Al-Shabaab	Civilians
event 4	39	02/04/2022	Military	SSC
event 5	63	02/10/2022	Al-Shabaab	Military

Table 1: Subset of ACLED

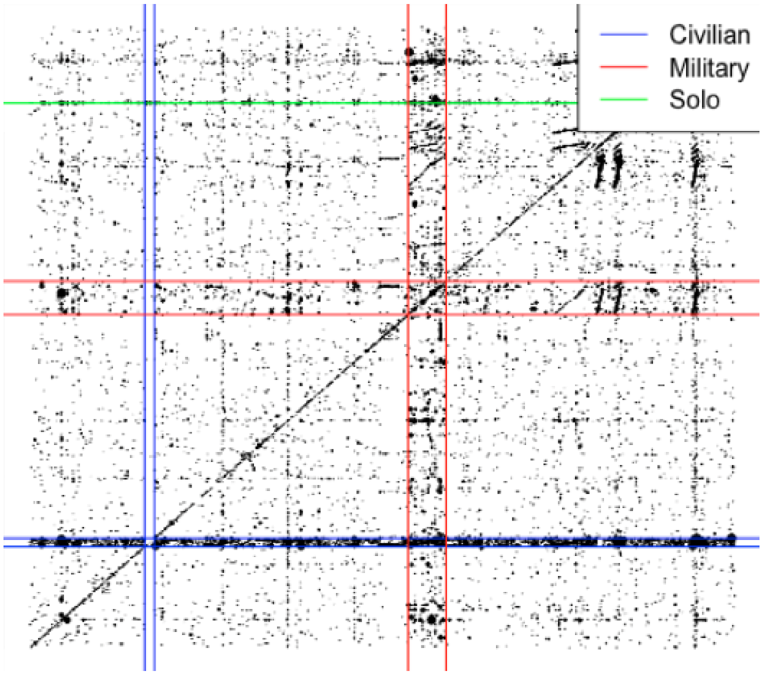

Figure 1: Scatter plot of actor1 versus actor2. A dot is marked if actor1 and actor2 fight with each other. The dot size reflects the number of conflicts between certain actors.

	the former event		the latter event		
	actor 1	actor 2	actor 1	actor 2	interval
event 1-2	Military	SSC	Military	Solo	4
event 1-3	Military	SSC	Al-Shabaab	Civilians	23
event 2-3	Military	Solo	Al-Shabaab	Civilians	19
event 2-4	Military	Solo	Military	SSC	30
event 3-4	Al-Shabaab	Civilians	Military	SSC	10
event 4-5	Military	SSC	Al-Shabaab	Military	6

Table 2: Neighborhood data based on Table 1

sentence 1	Military SSC Military Solo Military SSC Al-Shabaab Civilians
sentence 2	Military Solo Al-Shabaab Civilians Military Solo Military SSC
sentence 3	Al-Shabaab Civilians Military SSC
sentence 4	Military SSC Al-Shabaab Military

Table 3: Generated corpus text data of actor based on Table 2

	location		actor 1			actor 2			time
	latitude	longitude	E1	E2	E3	E1	E2	E3	duration
event 1	1.7	44.3	10.5	1.0	7.2	5.8	3.3	2.1	13.5
event 2	0.5	42.3	10.5	1.0	7.2	3.2	5.1	2.9	24.5
event 3	-1.2	45.2	7.2	5.3	6.1	1.3	1.2	6.3	10
event 4	0.8	46.0	10.5	1.0	7.2	5.8	3.3	2.1	6

Table 4: Example of the processed data with 3 embedding dimensions for each actor. E stands for embedding.

	location	actor	time
Baseline 1	✓	×	×
Baseline 2	✓	×	✓
Baseline 3	✓	✓	×
Proposed	✓	✓	✓

Table 5: Baseline models and proposed model.

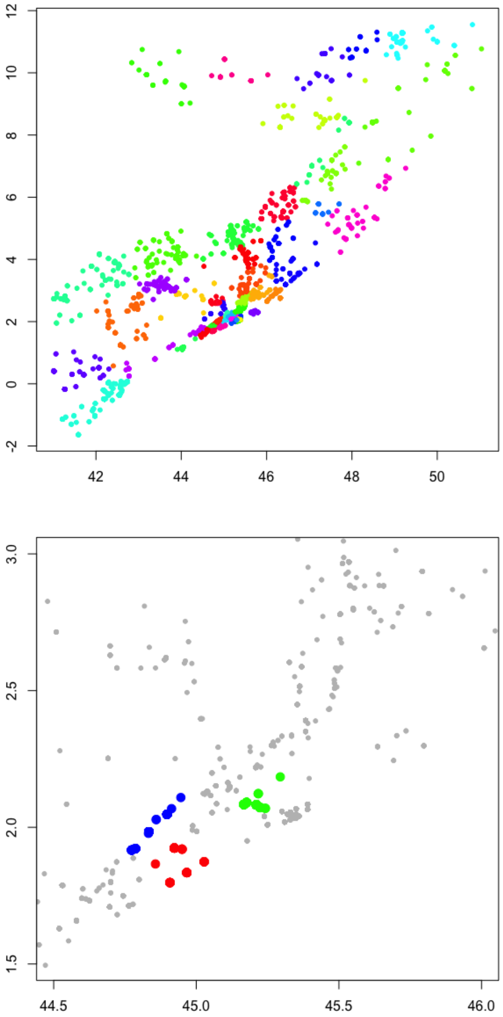

Figure 2: Clustering from `Baseline1`.

Blue	actor 1	actor 2	duration
#1	Al Shabaab	Military (17 –)	4
#2	Al Shabaab	Military (17 –)	10.5
#3	Al Shabaab	Military (17 –)	15
#4	Al Shabaab	Military (17 –)	11
#5	Al Shabaab	Military (17 –)	1
Green	actor 1	actor 2	duration
#1	Al Shabaab	Military (17 –)	29
#2	Al Shabaab	Military (17 –)	15.5
#3	Al Shabaab	Military (17 –)	1
#4	Al Shabaab	Military (22 –)	2
#5	Al Shabaab	Military (22 –)	17.5
Red	actor 1	actor 2	duration
#1	Al Shabaab	AMISOM	10
#2	Al Shabaab	AMISOM	28
#3	Al Shabaab	AMISOM	26
#4	Al Shabaab	Military (17 –)	21
#5	Al Shabaab	Military (17 –)	22

Table 6: Examples of conflicts from blue, green, and red groups near Mogadishu.

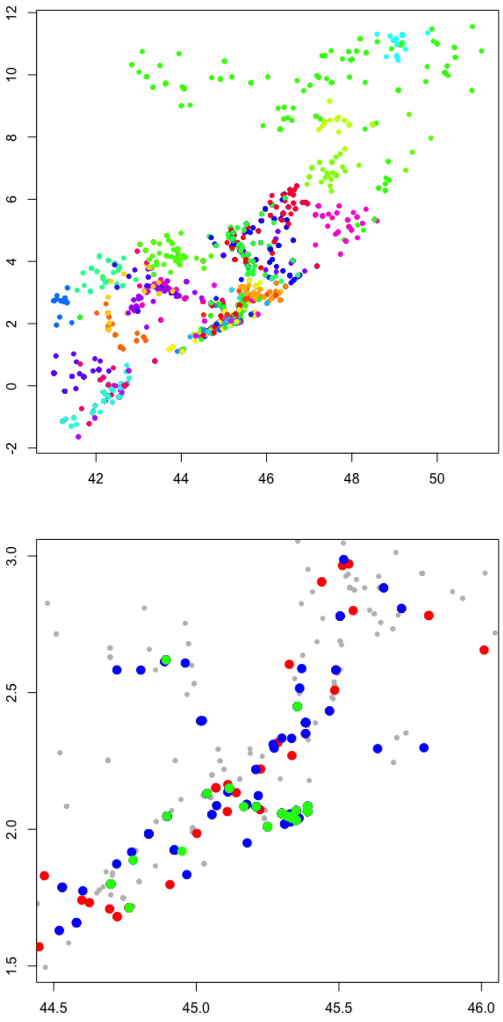

Figure 3: Clustering from `Baseline2`.

Blue	actor 1	actor 2	duration
#1	Unidentified	Civilians	9
#2	Al Shabaab	Military (17 –)	9
#3	Al Shabaab	Military (17 –)	9.5
#4	Al Shabaab	Civilians	10
#5	Al Shabaab	Military (22 –)	10
Green	actor 1	actor 2	duration
#1	Al Shabaab	Military (22 –)	29
#2	Al Shabaab	Police (17 –)	30
#3	Al Shabaab	AMISOM	30
#4	Al Shabaab	Military (22 –)	30
#5	Al Shabaab	Military (22 –)	29
Red	actor 1	actor 2	duration
#1	Al Shabaab	Police (17 –)	14.5
#2	Al Shabaab	AMISOM	14.5
#3	Al Shabaab	Police (17 –)	14.5
#4	Al Shabaab	Military (22 –)	14.5
#5	Al Shabaab	Military (17 –)	14.5

Table 7: Examples of conflicts from blue, green, and red groups near Mogadishu.

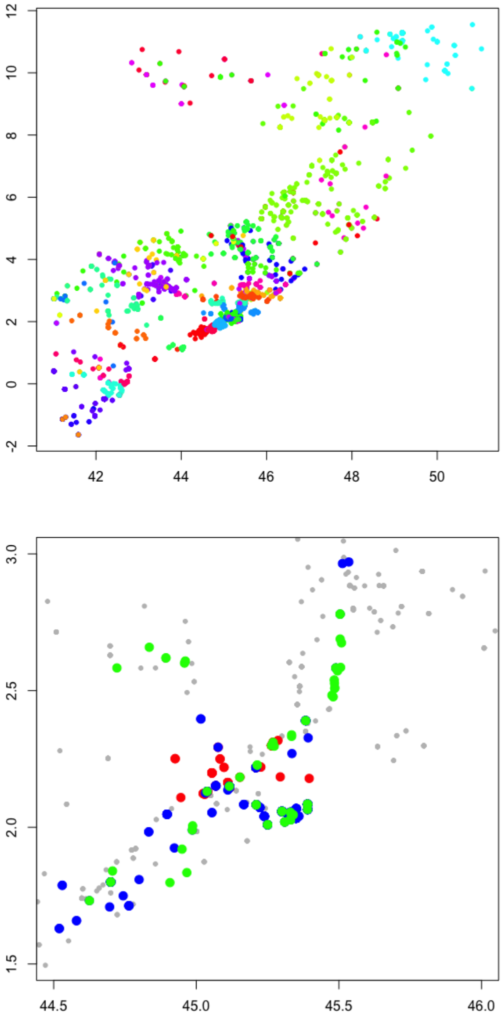

Figure 4: Clustering from `Baseline3`.

Blue	actor 1	actor 2	duration
#1	Military (17 –)	Civilians	7
#2	Al Shabaab	Civilians	12.5
#3	Military (17 –)	Military (17 –)	12
#4	Al Shabaab	Military (17 –)	13.5
#5	Al Shabaab	Military (17 –)	18
Green	actor 1	actor 2	duration
#1	Military (22 –)	Civilians	12
#2	Military (22 –)	Civilians	14
#3	ATMIS	Al Shabaab	15
#4	Military (22 –)	Civilians	7.5
#5	Military (22 –)	Al Shabaab	1
Red	actor 1	actor 2	duration
#1	Al Shabaab	ATMIS	30
#2	Al Shabaab	ATMIS	16.5
#3	Al Shabaab	ATMIS	14
#4	Al Shabaab	Military (22 –)	12
#5	Al Shabaab	Military (22 –)	20

Table 8: Examples of conflicts from blue, green, and red groups near Mogadishu.

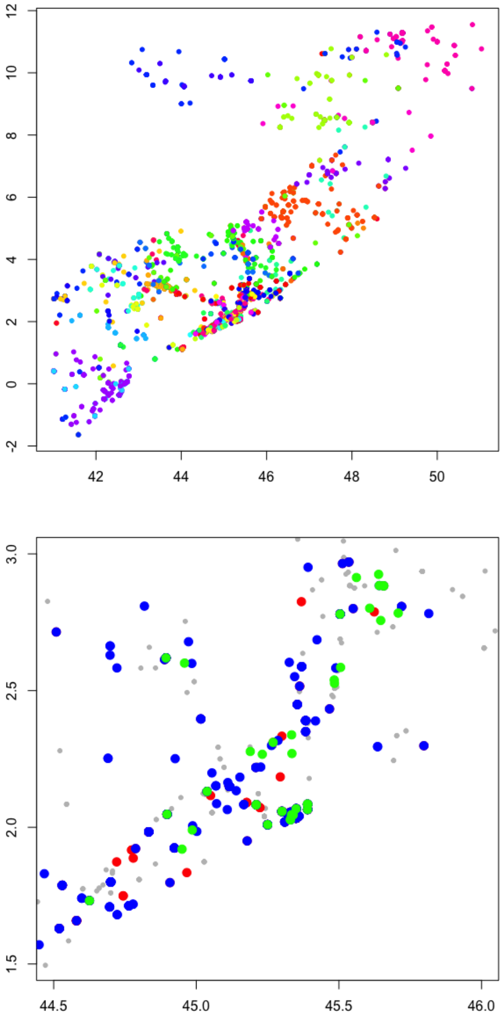

Figure 5: Clustering from `Proposed method`.

Blue	actor 1	actor 2	duration
#1	Al Shabaab	Military (17 –)	12.5
#2	Police (17 –)	Al Shabaab	18
#3	Military (17 –)	Al Shabaab	10.5
#4	Al Shabaab	Military (17 –)	12
#5	Al Shabaab	Military (17 –)	11.5
Green	actor 1	actor 2	duration
#1	Military (22 –)	Al Shabaab	15
#2	Military (22 –)	Al Shabaab	14
#3	Military (22 –)	Al Shabaab	16.5
#4	Military (22 –)	Al Shabaab	16
#5	ATMIS	Al Shabaab	16.5
Red	actor 1	actor 2	duration
#1	Al Shabaab	Military (17 –)	9
#2	Unidentified	Military (17 –)	9
#3	Military (17 –)	Civilians	8
#4	Unidentified	Police (17 –)	9
#5	Al Shabaab	Police (17 –)	9

Table 9: Examples of conflicts from blue, green, and red groups near Mogadishu.