11-2018

Is there space for violence?: A data-driven approach to the exploration of spatial-temporal dimensions of conflict

Tin Seong KAM
Singapore Management University, tskam@smu.edu.sg

Vincent ZHI
Singapore Management University, vincent.zhi.2016@mitb.smu.edu.sg

Is There Space for Violence?

A Data-driven Approach to the Exploration of Spatial-Temporal Dimensions of Conflict

Vincent Z.W. Mack
S. Rajaratnam School of International Studies
Nanyang Technological University
Singapore
isvzwmack@ntu.edu.sg

Tin Seong Kam
School of Information Systems
Singapore Management University
Singapore
tskam@smu.edu.sg

ABSTRACT

With recent increases in incidences of political violence globally, the world has now become more uncertain and less predictable. Of particular concern is the case of violence against civilians, who are often caught in the crossfire between armed state or non-state actors. Classical methods of studying political violence and international relations need to be updated. Adopting the use of data analytic tools and techniques of studying big data would enable academics and policy makers to make sense of a rapidly changing world.

KEYWORDS

geospatial autocorrelation, Africa, political violence, hotspot detection, knowledge discovery

ACM Reference format:

Vincent Z.W. Mack, Tin Seong Kam. 2018. Is There Space for Violence: A Data-driven Approach to the Exploration of Spatial-Temporal Dimensions of Conflict. In *Proceedings of 2nd ACM SIGSPATIAL Workshop on Geospatial Humanities (ACM SIGSPATIAL'18). ACM, Seattle, Washington, USA, 10 pages.* https://doi.org/10.1145/1234567890

1 INTRODUCTION

Recent studies have shown the rise in political violence throughout the world (RiskMap 2018; Raleigh & Moody 2017). According to a report by Control Risks, there has been an increase of political violence and crime incidents globally in 2017 (by 17% compared to 2016). Violent incidents have increased by 63% in Europe, 51% in the Asia Pacific region, 39% in Africa, and 26% in the Americas. The exception is the Middle East and North Africa, where though conflict incidents have decreased by 20%, the region remains fraught with violence. The increase threat of political violence also provokes the armed responses of states, as evident in the domination of counterterrorism rhetoric in foreign policy, creating a "permissive environment" towards the use of state-sanctioned violence for political aims (Malley 2018). Coupled with that is the rise in waves of protest worldwide, and the frequency of the mobilizations of large-scale protests suggests a fundamental shift has occurred in the way global civil society is operating (Youngs 2017).

Of particular concern is the occurrence of violence against civilians. In wars or civil conflicts, civilians' non-combatants have often been targeted as they are perceived as a proxy for the opposing faction. In this research study, we ask the following questions: Where does violence against civilians commonly occur? Are we able to detect where the hotspots of violence occur? Or are they a randomly distributed phenomenon? Do they tend to cluster together in specific locations or along border regions where conflict is rife? Also, how do they diffuse in a geographical location over time?

The answers to these questions have much application to the role of UN peacekeeping enforcement operations, where one of the top priorities is to minimize violence against civilians and causalities of war. Thus, understanding where violence against civilians is situated and how it might diffuse through geography would be crucial for the UN peacekeepers to develop and operationalize intervention strategies.

1.1 Conceptualizing the Study of Conflict Spaces

The main aim of this study is to deepen and update our understanding on the dynamics of political violence and its diffusion over spatial boundaries. While much of the literature on armed conflict explain violence as a result of factors exogenous to the process such as lack of democracy, poor government institutions, poverty, and ethnic divisions, few have attempted to understand how the spatial-temporal dimension of violence may in fact lead to more violence (Schutte & Weidmann, 2011).

The field of conflict suffers from state-centrism (Chojnacki et. al., 2009) and a simplistic understanding of geography. Most studies of war in international relations, qualitative or otherwise adopt the assumption that states are the primary geopolitical actors in conflicts, with the country-dyad-year as the main unit of analysis, and war is often interpreted as the aggregation of violence or conflict events occurring between political entities. Explanations

ACM SIGSPATIAL'18, November, 2018, Seattle, Washington, USA

https://doi.org/10.1145/1234567890

of the causes of war, conflict, and violence rely on the attributes and bilateral interstate relations of states as main independent variables with the goal of "uncovering general patterns of behavior that are treated as universal tendencies" (Flint et al., 2009). For e.g. quantitative studies of conflict rely systematic statistical analyses of large datasets that span long time periods as well as geographical contexts, studying the interactions between the two countries in a particular year, focusing for e.g. on trade, conflict, or alliance relations. In many of these studies however, the spatial dimension is often ignored as "noise".

An update to this understanding of the role of space in conflict is needed, which will not only result in a more sophisticated appreciation of the effect of geography on conflict, but also alter the assumptions of how power is conceptualized and operationalized through spatial dimensions. The international relations study of conflict could benefit much from the work of political geographers, who have been applying quantitative methods to the study of armed conflicts for some time.

1.2 Outline of Paper

The rest of the paper is divided into 4 main sections. First, we review the available methods of discovering hotspots or clusters of violence and conflict from geospatial data. Second, we discussed data methodology and preparation, including the data description. The results and analysis section follows, where we first conducted the Initial Exploratory Data Analysis (EDA) to discover insights and generate hypothesis about the distribution of violence against civilians from the data through John Tukey's EDA approach. These insights and hypotheses are explored using the Exploratory Spatial Data Analysis (ESDA) method, and the results of the analysis is interpreted and discussed. Finally, the findings are summarized in the conclusion section.

2 RESEARCH METHODOLOGY

2.1 Current Methods

The relationship between objects on a map can be encapsulated by Tobler's (1970) First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things."

In recent times, some effort has been made to incorporate geospatial data into the analysis of conflict. However, attempts are fledgling at best. Scholars that have attempted to conceptualize location and space in their analysis do so in either relative or absolute terms; i.e., space is explored on the level of contiguity between state actors, where the status of whether a state in the dyad unit are neighbors are used as part of the explanation of war or conflict between them (O'Loughlin and Anselin 1992; Buhaug and Gleditsch 2008). Such an understanding is crude at best, as physical contiguity between states does not address the complexity in the process of how states go to war, especially since more neighboring states are not in conflict with each other.

Also, this approach does not provide much information regarding the local dynamics of conflict, nor is it able to explain when violent conflict occurs between non-state actors in areas where state authority is weak or non-existent. It is also unable to account for cases when these actors carry out acts of violence against civilians, which often occurs in environments where the state has collapsed and does not own the monopoly to violence, with multiple non-state actors such as rebel groups, warlords, and local and ethnic militias dominating the landscape.

More sophisticated attempts to operationalize the geospatial study of conflict have focused on the utilization of event-level data and geographic information systems (GIS) to aid the identification and mapping of patterns of violence. These methods sought to more realistically represent the area of conflict, identifying hotspots of conflict activity located within the boundaries of states. One such approach is the Standard Ellipse method, which is employed by studies that attempt to visualizes the distribution of point data by drawing an ellipse over the approximate area of conflict, centering upon the average centre of locations where violence events have occurred with the size of the ellipse determined by either the covering point data location where the conflicts occurred (Benson, 2018) or by the densities of conflict, measured by the frequencies at which the conflict occurred in that location (Chojnacki et. al., 2009). While these approaches are a step up, they still remain unsatisfactory, as the ellipse method is too coarse to gain any insight, especially if we are interested in how diffusion of violence happens.

2.2 Exploratory Spatial Data Analysis (ESDA)

Exploratory Spatial Data Analysis (ESDA) provides a solution to this challenge. According to Anselin (1998), ESDA is a suite of descriptive techniques used in the identification of spatial outliers and discovery of clustering patterns and hotspots through visualization of the spatial distribution of data. Central to ESDA is the notion of spatial autocorrelation, which is defined as the "correlation among values of a single variable strictly attributable to their relatively close locational positions on a two-dimensional (2-D) surface" (Griffith 2009). As a subset of Exploratory Data Analysis (EDA), ESDA is used in knowledge discovery and pattern detection in spatial datasets, leading to the formulation and testing of hypotheses based on the geography of the data.

2.2.1 Measuring Spatial Autocorrelation. There are two measures of spatial autocorrelation that is used for this project: 1) Moran's I statistic (Moran 1950) - one of the oldest indicators of spatial autocorrelation; and 2) Geary's C Ratio (Geary, 1954) - a statistic similar to Moran's I.

2.2.1.1 Moran's I Statistic. Generally, Moran's I is calculated as follows:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i-\mu)(x_j-\mu)}{\sum_i (x_i-\mu)^2} \quad (1)$$

where w_{ij} is a spatial weight matrix which compares the closeness between location i and location j, x_i is the frequency of incidents where violence occurred in area i, x_j is the frequency of incidents where violence occurred in area j, μ is the average frequency of

violence, and *N* is the total number of areas. The *spdep* package provides the *moran.test()* function to calculate the Moran's I scores.

As a measure of spatial autocorrelation, Moran's I statistic ranges from -1 to 1, where a value of -1 indicates perfect negative autocorrelation and +1 indicates perfect positive autocorrelation. A value of 0 indicates that there is no autocorrelation.

The presence of a positive spatial autocorrelation score implies that the occurrence of violence tends to cluster together, whereas the presence of a negative spatial autocorrelation score implies that the occurrence of violence against civilians in Africa are more dissimilar compared to its nearby neighbors.

2.2.1.2 Geary's C Ratio. Another indicator of spatial autocorrelation is the Geary's contiguity ratio (or Geary's C). The Geary's C ratio is based upon a paired comparison of juxtaposed map values and is calculated as follows:

$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2(\sum_i \sum_j w_{ij}) \sum_i (x_i - \mu)^2} \quad (2)$$

Geary's C is inversely related to Moran's I and all the terms calculating C are the same as defined for the Moran's I. Unlike Moran's I which lies between 1 and -1, the value of Geary's C lies between 0 and some unspecified value greater than 1. Values below 1 indicate positive spatial autocorrelation and values above 1 indicate negative spatial autocorrelation. Also, Geary's C is more sensitive than Moran's I to the absolute differences between neighboring values due to the squared term in the numerator. If there is no spatial autocorrelation, the value of C would be equals to 1. Likewise, the Geary's C score can be calculated with the *spdep* package using its *geary.test()* function.

2.2.2 Studying Spatial Autocorrelation at the Local Level. While the global Moran's I score and the Geary's C ratio can tell us whether violence against civilians tends to cluster or not on the map, it does not provide any information on the distribution of spatial dependence of violence against civilians and is unable to identify the location of hotspots and clusters. For that, we require the use of more localized methods - Anselin's Moran Scatterplot and the Local Indicator of Spatial Autocorrelation (or LISA) method.

2.2.2.1 Moran Scatterplots. The Moran Scatterplot allows us to study the local spatial instability of the distribution of violence against civilians in Africa. The *spdep* package provides the *moran.plot()* function to help us to plot the Moran Scatterplot. The various regions are distributed across the scatterplot, with spatially lagged values of violence in these regions plotted on the y-axis against the original values of violence on the x-axis. The Moran scatterplot is divided into four areas, with each quadrant corresponding with one of four categories: (1) High-High (HH) in the top-right quadrant; (2) High-Low (HL) in the bottom right quadrant; (3) Low-High (LH) in the top-left quadrant; (4) Low-Low (LL) in the bottom left quadrant. In the context of this capstone project, the meaning of the categories are as follows: 1) High-High (HH): indicates high spatial correlation where incidents of violence against civilians are clustered closely together. 2) High-Low (HL): where areas of high frequency of incidents where violence against civilians occurred are located next to areas where there is low frequency of incidents where violence against civilians occurred. 3) Low-High (LH): these are areas of low frequency of incidents where violence against civilians occurred that are located next to areas where high frequency of violence against civilians occurred. 4) Low-Low (LL): these are clusters of low frequency of incidents where violence against civilians occurred. However, the Moran Scatterplot has one drawback - it does not indicate whether these regions are significant or not.

2.2.2.2 Local Indicators of Spatial Association (LISA). We address this limitation of the Moran Scatterplot in our analysis by using the LISA method. LISA will not only allow us to identify the hotspot locations, but also the statistical significance of the hot spots in the dataset. The following equation describes calculates the local Moran's I scores for each region used in the LISA method:

$$I_i = \frac{x_i - \mu}{\sum_i (x_i - \mu)^2 / N} \sum_j w_{ij} (x_j - \mu) \quad (3)$$

The *spdep* package provides the function *localmoran()* to assist us in calculating the local Moran's I scores.

According to Anselin (1995), LISA can be used to locate "hot spots" or local spatial clusters where the occurrence of violence against civilians is statistically significant. Thus, in addition to the four categories described in the Moran Scatterplot, the LISA analysis includes an additional category: (5) Insignificant: where there are no spatial autocorrelation or clusters where violence against civilians occurred.

Unlike the previous calculations, computing the LISA scores and visualizing them for output is not readily provided by the *spdep* package. In order to facilitate computation in the RMarkdown, we wrote a function to compute the LISA scores and generate the requisite maps. All we need to do to calculate the LISA is to input the data frame, the row-standardized weight matrix, and the significance or alpha level we want to test it at, as can be observed in the following function:

Function 1: Calculating LISA scores

Input: A data frame *df* of frequency of violence events organized by type and year, with the respective country/administrative region names and geocoded location (latitude/longitude) data; a row-standardized weight matrix of neighbors *rswm* for aforementioned countries; the significance or alpha level α for testing.

Output: The coefficient matrix in table form of the local Moran's I scores; the choropleth visualization of the local Moran's I scores; the choropleth visualization of the p-values; the choropleth visualization of the LISA

classifications of the country/administrative regions; the working data frame from which these visualizations are derived from.

1 For each country/administrative region, local Moran's I scores are calculated with the *localmoran()* function from the *spdep* package, using the frequency of the event and the row-standardized matrix as the inputs. The output data frames include the coefficient matrix, the local Moran's I scores and the p-values of each score.

2 Choropleth of local Moran's I from **2** is plotted with the *tmap* package.

3 Choropleth of p-values of local Moran's I from **2** is plotted with the *tmap* package.

4 For each country/administrative region, the factor *DV* is calculated by deducting the mean of all event frequencies from individual event frequency scores.

5 For each country/administrative region, the factor *C_mI* is calculated by deducting the mean of all local Moran's I scores from individual local Moran's I scores.

6 For each country/administrative region, LISA classification of the scores are assigned as follows:

- If $DV > 0$ & $C\_mI > 0$; "High-High" is assigned
- If $DV < 0$ & $C\_mI < 0$; "Low-Low" is assigned
- If $DV < 0$ & $C\_mI > 0$; "Low-High" is assigned
- If $DV > 0$ & $C\_mI < 0$; "High-Low" is assigned

7 For each country/administrative region, "non-significance" label is assigned if p-value > α.

8 Choropleth of p-values of LISA classifications from **6** and **7** is plotted with the *tmap* package.

3 DATA METHODS AND TOOLS

3.1 Data Sources

The conduct of the exploratory spatial data analysis (ESDA) is centered around the ACLED (Armed Conflict Location and Event Data Project) dataset downloaded from https://www.acleddata.com/. Although ACLED provides political violence event data for Africa, the Middle East, and South and Southeast Asia, for the purposes of this capstone project we will be limiting the analysis to the Africa dataset as it has the longest time range (1997-2018). In this data set, a total of 49 countries are covered. There are 161,939 events observed in this data set ranging from the years 1997 to 2018; and event date data is available down to the actual day of occurrence. 9 different event types are coded in the dataset: 1) Violence against civilians; 2) Battle-No change of territory; 3) Remote violence; 4) Riots/Protests; 5) Strategic Development; 6) Battle-Government regains territory; 7) Non-violent transfer of territory; 8) Headquarters or base established; and 9) Battle-Non-state actor overtakes territory.

We plotted the conflict event data onto the African map through the use of shapefiles – a geospatial vector data format for geographic information systems (GIS), software. Developed by the Environment Systems Research Institute (ESRI), shapefiles store points, lines and polygons that are linked to data attributes such as countries, provinces, districts etc. which allow GIS software to render into graphical maps. Shapefiles of Africa were downloaded from the database of Global Administrative Areas (GADM) at https://gadm.org/data.html - a high resolution spatial database containing country regions and their respective administrative areas.

3.2 R Packages Used

Due to the dynamic nature of the data and the analysis process, we had decided to work in RMarkdown in order to document precisely and accurately the steps taken with the goal of creating reproducible data science. For general data wrangling and visualization purposes, we used the *tidyverse* suite of packages, which amongst others includes *ggplot2*, *tibble* data frames, and *dplyr*. This was supplemented with *readxl* and *lubridate* for reading excel files and performing date field manipulation, and a suite of other packages based on Hadley's Grammar of Graphics such as *ggmap*, *ggplus*, *ggforce*, *ggraph*, *ggally* etc. For mapping purposes, we used the *sf* package to translate the shapefile into the common architecture of Simple Feature Access in the International organization for Standardization (ISO) standard ISO 19125. Other packages used for geospatial analytics and map visualization include *tmap, rgdal, spatstat, geofacet, rgeos* and *lwgeom*. The data application programming interface (API) packages *wbstats* and *GADMTools* were used to download World Bank data and GADM shapefiles respectively.

3.3 Data Preparation and Cleaning

Several data quality challenges were encountered in the data preparation process. Due to Africa's colonial history, certain names of geographical regions and locations were spelt differently, incorrectly, or in a different language/alphabet. Location data between ACLED and GADM were not always analogous: a large number of conflict events were found to be missing or unassigned, or assigned to the wrong countries. Some of the locations of the data points even fell outside the boundary of the country/administrative region polygons. Due to the nature of political conflict, country/administrative boundaries and borders can sometimes be fluid, and names of countries and administrative areas were found to have changed; either disaggregated into new countries/administrative areas or previously active but now defunct administrative areas were agglomerated and upgraded into higher tier administrative areas. Given that the *GADMTools* package uses "iso3c" country codes to download country shapefiles, and these "iso3c" codes are used by World Bank's World Development Indicators, we had to join the

ACLED dataset up with the World Bank data to find out what the respective country codes are.

4 RESULTS AND ANALYSIS

4.1 Initial Exploratory Data Analysis

Based on 1997-2017 data from the ACLED dataset, the number of incidences of violence against civilians is perhaps the most numerous (See Figure 1). Violence against civilians in Africa have also been on the rise in previous years, rising sharply from 2009 onwards (See Figure 2). So as to explore any patterns of spatial autocorrelation, we mapped the distribution of violence/conflict events in Africa and perform an initial exploratory data analysis (ESDA) on Violence against civilians.

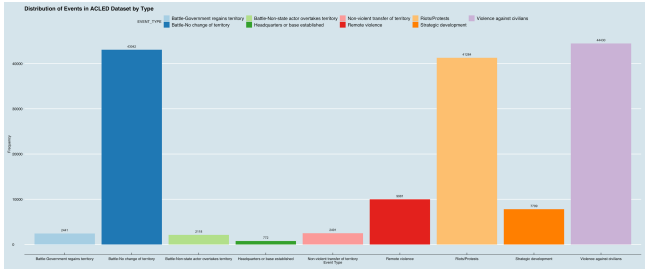

Figure 1: Bar plot Distribution of ACLED event types.

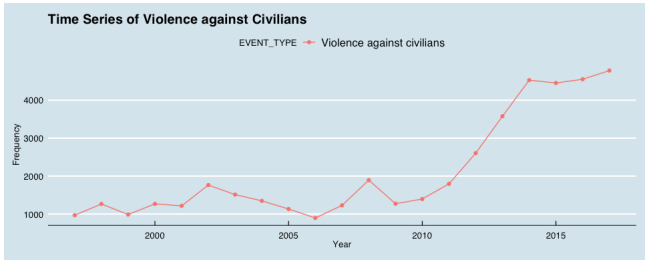

Figure 2: Violence against civilians rising over time in Africa

Figure 3 shows a choropleth plot of the incidences of violence against civilians at the country level, partitioned at equal intervals of 1000. The distribution of violence against civilians is not even, with certain countries like Sudan, Nigeria, Democratic Republic of Congo, Somalia and Zimbabwe having a higher concentration of violence events than the others. Nevertheless, there seem to be areas where they are clustered - i.e. around the Central and Eastern region of Africa.

However, not all countries had the same distribution of conflict incidents. Most countries had few incidences where violence against civilians occurred throughout the years and the distribution of violence over the time series is quite varied (see Figure 4). For countries like Algeria, Angola, Guinea, Liberia, Namibia and Sierra Leone, violence against civilians occurred more frequently in the earlier years of the dataset but dropped in the years thereafter. Others such as Burundi, Central African Republic, Democratic Republic of Congo, Egypt, Ethiopia, Kenya, Libya, Nigeria, Somalia, South Africa, Sudan and South Sudan, were observed to have more incidences of violence against civilians in the later years.

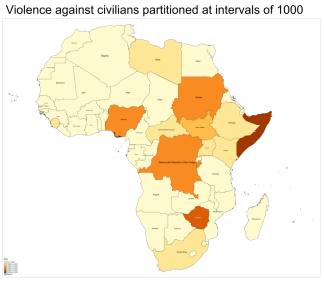

Figure 3: Distribution of violence against civilians amongst countries in Africa

Figure 4: Trellis plot of violence against civilians over time

The visualizations of the choropleth at the level of each individual year from provide a more locational context to the time series plots above. From Figure 5, it is easier to observe which countries have the most frequency of violence for each year. We note that the top areas with highest frequencies of violence are not constant; they change from year to year, though certain countries dominate for certain periods.

From 1997 to 2000, violence against civilians in Sierra Leone and Angola seems to be an all-time high, before lessening in 2001. There also appears to be some indication of clustering, where

certain groupings of neighboring countries exhibit similarly high levels of violence against civilians in certain years. For e.g., from 1999, Angola and Democratic Republic of Congo exhibited higher than average levels of violence, which appeared to have spread to the cluster of countries consisting of South Sudan, Sudan, Ethiopia, Kenya and Somalia in 2000. The levels of violence fluctuate a little from 2001 onwards and seemingly spread to neighboring countries such as Central African Republic, Libya and Egypt. However, this visualization method suffers from a few weaknesses. First, while neighboring regions show similar levels of violence over the same time periods is indicative that violence might have diffused from one country to another, we cannot tell conclusively if diffusion of violence has taken place.

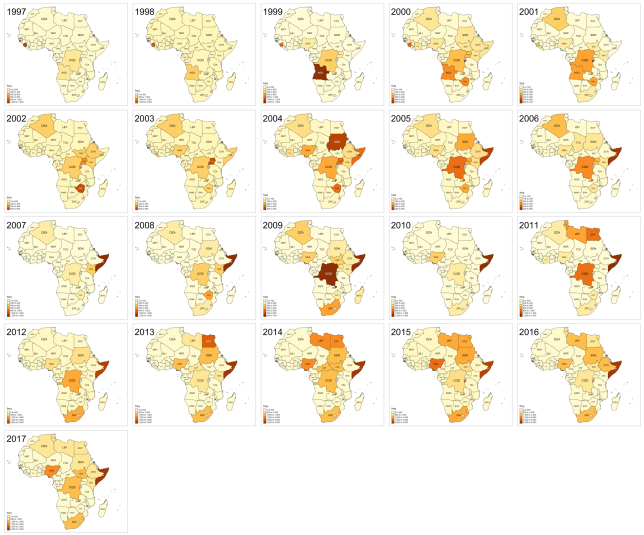

Figure 5: Choropleth of violence against civilians in Africa from 1997 to 2017: partitioned by "pretty" intervals

This is because despite the high levels of violence against civilians is observed in certain countries that share boundaries, it is unclear if the violence that occur is located along the boundaries between countries. Africa is a large place; while country sizes within Africa varies, the average area for each country is quite high, at 54,343,623 km^2. In order to get a better sense of the diffusion effect, we will need to study the distribution of violence at the administrative level 1 (see Figure 6).

Second, given that the algorithm that the *tmap* packages uses to decide which intervals to be plotted with which hue operates at a relative level. As can be seen in Figure 6 below, plotting the choropleth of Admin 1 regions using *tmap's* "pretty" style of creating regular intervals rounded off at the hundreds yields us less information than if we were to plot it using the quantile style of creating intervals, where given that there are more regions, we are able to get a better sense of the relative levels of violence via a finer distinction level as opposed to the "pretty" interval scale.

4.2 Exploratory Spatial Data Analysis

The exploration using the choropleths above show that regular EDA methods are limited in their ability to represent geospatial data in a useful and methodologically rigors way. As such, we will proceed to using ESDA methods where we attempt to establish whether the data is in fact spatially autocorrelated.

4.2.1 Measuring Global Spatial Autocorrelation. In this section, we test the hypothesis H_1: that the distribution of violence against civilians in Africa from 1997-2017 is spatially dependent through calculating the Moran's I statistic and Geary's C ratio for the data set, as opposed to the null hypothesis H_0 which states that the distribution of violence against civilians is a random phenomenon. To establish the p-values of these statistics, we ran the Monte Carlo simulations for Moran's I and Geary's C for each weight matrix for each year over the period of 1997 to 2017, with the inference based on Anselin's (1995) permutation approach with 1000 permutations.

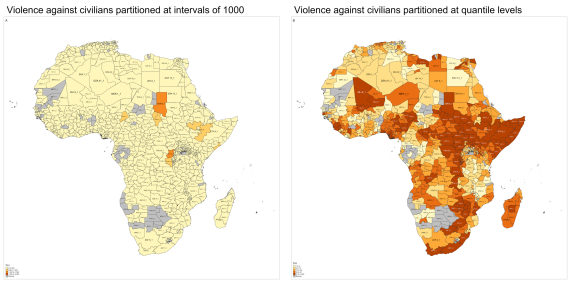

Figure 6: Distribution of violence against civilians amongst countries in Africa at the Admin 1 level: Comparing "pretty" partitions and quantile partitions

4.2.1.1 Contiguity Weight Matrices. Originally, w_ij as formulated by Moran (1950) is a contiguity or adjacency matrix where $w_{ij} = 1$ if area i and area j are adjacent. If they are not adjacent, $w_{ij} = 0$. Cliff and Ord (1973) have noted that the weight matrix w_{ij} could be generalized to fit any kind of weight criteria.

There are various methods of determining adjacency. The Rook method (named for the chess piece) considers areas adjacent if they are directly located horizontally or vertically on a 2-dimensional plane. The Queen method considers diagonally adjacent locations in addition to the Rook. The spatial weight matrices derived from these methods are W_r and W_q respectively.

Table 1: Rook vs Queen Contiguity Matrices

	W_r	W_q
No. of regions without links	48	27
Average number of links per region	3.4704	4.1395

Unfortunately, neither matrix yields a good set of neighbors. The Rook contiguity map fails in cases where the adjacent polygon is not clearly located north, south, east or west from the source polygon, or that there is more than one contender for one of the 4 polygons. Thus, certain administrative regions are left "neighbor-less". While W_q has better performance than W_r – with lesser number of regions without neighbors (see Table 1) – the same basic set of problems remain: there are still regions with on

neighbors and the number of neighbors across the dataset is not consistent. Also, for both matrices, Madagascar is separated from the mainland as are other islands such as São Tomé and Príncipe off the West Coast of Africa.

4.2.1.2 Distance-based Matrices. One method to overcome the limitations of the Rook and Queen contiguity matrix is to use a distance-based algorithm to determine adjacency. This involves calculating area adjacency through setting a distance parameter, where as long as the distance between the centroids of areas x_i and x_j fall within the specified distance, $w_{ij} = 1$, else $w_{ij} = 0$.

Using this algorithm, neighbors falling within the preset distance are marked as neighbors on a spatial weight matrix. The function we used to determine the neighbors is *dnearneigh()* from the *spdep* package. In order to run the *dnearneigh()* function, we will need to specify the lower and upper distance bounds. Using the value of 0.8 for the lower bound (derived using the *min()* and *nbdists()* functions), and the following values for the upper bound: *max_dist = {1,3,5,6,7}*.

Table 2: Distance-based Matrices

	$W_{d=1}$	$W_{d=3}$	$W_{d=5}$	$W_{d=6}$	$W_{d=7}$
No. of regions without links	494	54	27	3	0
Average number of links per region	0.9314	14.4941	31.9267	41.8345	52.2411

From Table 2, with minimum distance = 0.8 and maximum distance = 1, there are 494 regions with no neighbors. When the maximum distance is increased to 5, its performance seems to be similar to that of W_q with the same 27 regions with no neighbours. When we increased the maximum distance to 6, the number of regions with no neighbors dropped to 3. Finally, the number of regions without neighbors disappeared when we increased the maximum distance to 7 as in $W_{d=7}$. However, the average number of links per region went up to around 52, which may skew the analysis.

4.2.1.2 K-Nearest Neighbors Matrices. One solution to resolve these problems presented by the contiguity and the fixed-distance matrices is through the k nearest neighbours algorithm. This is an adaptive distance method where we first preset a fixed number k of neighbours desired per area, and the algorithm selects the k number of "adjacent" neighbours based on a proximity measure of the calculation of distance between each region. Compared to the previous contiguity matrices (i.e. W_r and W_q) all the regions are connected in the proximity matrices and there are no 0-neighbour regions. However, a high k value has the potential issue where second, third or even fourth order and higher neighbors become incorporated into the calculation of the spatial autocorrelation score, which will have the effect of smoothening out the values.

While this may not be too big of an issue for smaller, urbanized administrative regions as the effect of violence in a second order neighbor may diffuse into it, it presents a methodological problem for the larger, more rural areas, especially those located near the Sahara Desert region in the Northwestern region of Africa. Given their comparatively larger sizes, it may be unrealistic to assume that violence events occurring in second or third order neighbors may cause affect it. Therefore, we err on the side of caution and use $W_{k=8}$ - the smallest spatial weight matrix - as the main weight matrix for the analysis. Nevertheless, we ran the rest of the weight matrices through the Moran's I and Geary's C calculation for robustness testing and calibration purposes.

4.2.1.2 Calculating Global Spatial Autocorrelation. We found that the Moran's I and Geary's C statistics are significant at 0.1, suggesting that the distribution of violence against civilians over all of Africa is naturally clustered each year over the entire period from 1997 to 2017. Thus, we can reject the null hypothesis that the distribution of incidences of violence against civilian is random. This observation is echoed in the line plot of the Moran's I and Geary's C scores across the years (see Figure 7 and 8).

The increase in the number of neighbors have the effect of smoothening out the Moran's I scores. Moran's I scores are highest when $k = 8$ and lowest at $k = 28$. All the scores indicated positive spatial autocorrelation.

While the Global Moran's I scores appear to vary consistently and somewhat evenly across the different weight matrices, the same cannot be said for the Geary's C scores. Based on the time series visualization for the Geary's C scores, the weight matrices can roughly be grouped into three categories.

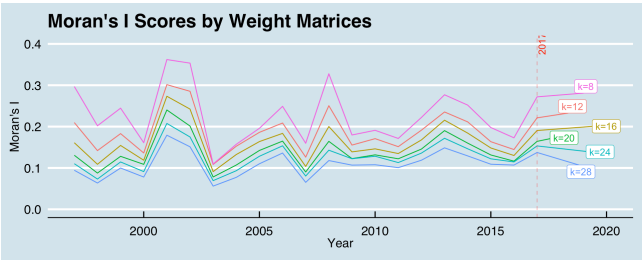

Figure 7: Calibration Plot: Moran's I scores across the years

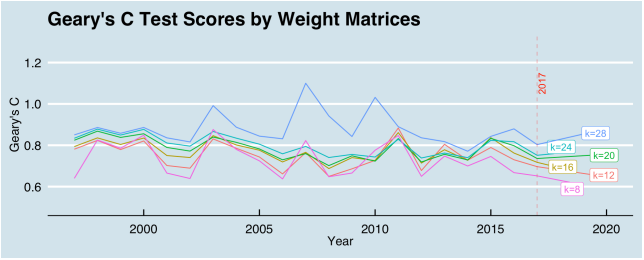

Figure 8: Calibration Plot: Geary's C scores across the years

The first group contains $k = 8$ and $k = 12$, the second consists of $k = 15, k = 20$ and $k = 24$ which has smaller variations compared to the other two groups. The final group contains one weight matrix $k = 28$, which is also the only matrix with a Geary's C score that was close to 1 in 2003 and 2010, and was more than 1 in 2007 i.e. signs of negative spatial autocorrelation.

4.2.2 Measuring Local Spatial Autocorrelation. Given that we have established that spatial clustering of violence against civilians does occur in Africa from 1997 to 2017, we proceed to discover if any hot spots of violence can be observed using Anselin's Moran Scatterplot and the Local Indicators of Spatial Autocorrelation (LISA) method.

4.2.2.1 Identifying Violence Hotspots with Moran Scatterplots. As mentioned earlier, the Moran scatterplot is divided into four quadrants, each corresponding with one of four categories. The direction and magnitude of global autocorrelation can be observed in a Moran Scatterplot, as the slope of the linear regression of the lagged values of violence vs the original frequencies of violence is equivalent to the Moran's I score.

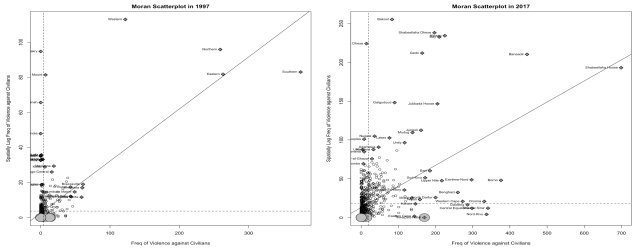

Figure 9: Moran Scatterplot for t = {1997, 2017}

The output of the sample of Moran Scatterplots agrees with the Moran's I scores in the Tables 3 and 4. However, we are not able to see which of these areas are significant. For that, we turn to the LISA analysis.

4.2.2.3 Mapping the Local Moran's I. To get a more intuitive sense of the local Moran's I scores, we also plotted out the values on a choropleth (see Figure 10). From these plots, we observe where the areas with positive and negative spatial autocorrelation (i.e. $I_i > 0$ and $I_i < 0$) are distributed respectively. This could be very informative when studied in conjunction with the Moran Scatterplots.

Table 3: Coefficients of Local Moran's I for 1997 for the first 10 admin 1 regions

Region Code	Local Moran's I (Ii)	Expected Ii	Variance of Ii	z-score	p-value
AGO.1_1	0.0312	-0.0012	0.0928	0.1063	0.4577
AGO.10_1	0.0772	-0.0012	0.0928	0.2571	0.3985
AGO.11_1	-0.0248	-0.0012	0.0928	-0.0776	0.5309
AGO.12_1	1.0896	-0.0012	0.0928	3.5780	0.0002
AGO.13_1	0.0312	-0.0012	0.0928	0.1063	0.4577
AGO.14_1	0.5655	-0.0012	0.0928	1.8598	0.0315
AGO.15_1	0.1813	-0.0012	0.0928	0.5990	0.2746
AGO.16_1	-0.0284	-0.0012	0.0928	-0.0894	0.5356
AGO.17_1	1.6796	-0.0012	0.0928	5.5167	0.0000
AGO.18_1	0.2124	-0.0012	0.0928	0.7009	0.2417

Table 4: Coefficients of Local Moran's I for 2017 for the first 10 admin 1 regions

Region Code	Local Moran's I (Ii)	Expected Ii	Variance of Ii	z-score	p-value
AGO.1_1	0.0941	-0.0012	0.1161	0.2796	0.3900
AGO.10_1	0.1175	-0.0012	0.1161	0.3483	0.3638
AGO.11_1	-0.0349	-0.0012	0.1161	-0.0988	0.5394
AGO.12_1	-0.0004	-0.0012	0.1161	0.0023	0.4991
AGO.13_1	0.0081	-0.0012	0.1161	0.0273	0.4891
AGO.14_1	0.0977	-0.0012	0.1161	0.2903	0.3858
AGO.15_1	0.1067	-0.0012	0.1161	0.3167	0.3757
AGO.16_1	0.1135	-0.0012	0.1161	0.3365	0.3682
AGO.17_1	0.0715	-0.0012	0.1161	0.2132	0.4156
AGO.18_1	0.0959	-0.0012	0.1161	0.2849	0.3879

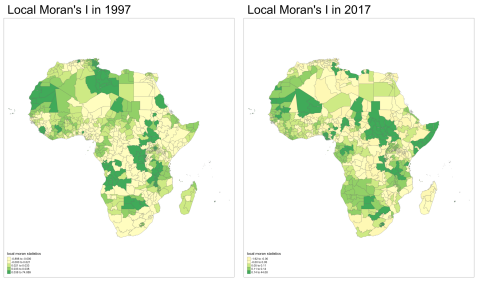

Figure 10: Local Moran's I plot for 1997 and 2017

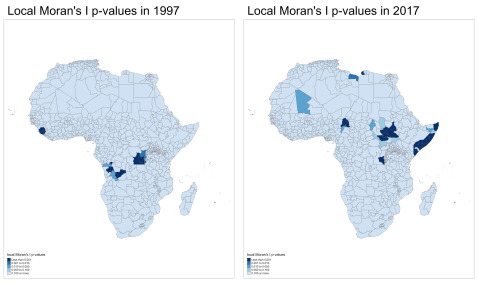

Figure 11: Plots of p-values for 1997 and 2017

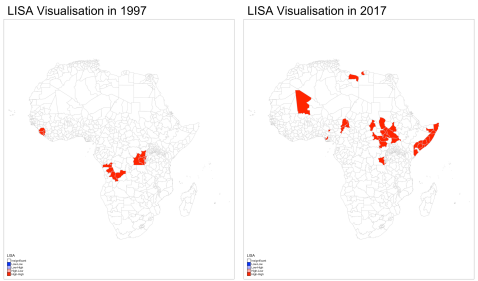

Figure 12: LISA cluster maps for 1997 and 2017

4.2.2.4 Mapping the p-values of the Local Moran's I. Next, we plot the p-values for these areas on a choropleth so that we can observe which areas are significant (see Figure 11).

4.2.2.5 Creating LISA Cluster Maps. While Figures 10 and 11 are informative, the strength of the LISA analysis can truly be observed when the Local Moran's I values are overlaid with the significance levels in the p-level map above. As mentioned earlier, this LISA cluster map also recodes each region into one of five categories: (1) High-High (HH); (2) High-Low (HL); (3) Low-High (LH); (4) Low-Low (LL); and (5) Insignificant. (See the Figure 12). Based on the initial analysis of the first and last year in the dataset, we see that the incidences of violence against civilians in Africa does indeed have a clustering pattern. This suggests that these areas featured above are potential hotspots of violence.

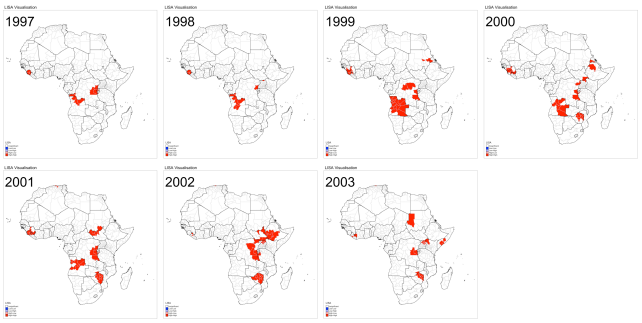

Figure 13: LISA Cluster Map for Phase I – from 1997 to 2003

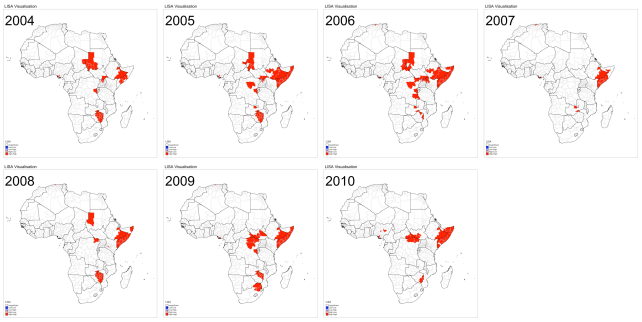

Figure 14: LISA Cluster Map Phase II – from 2004 to 2010

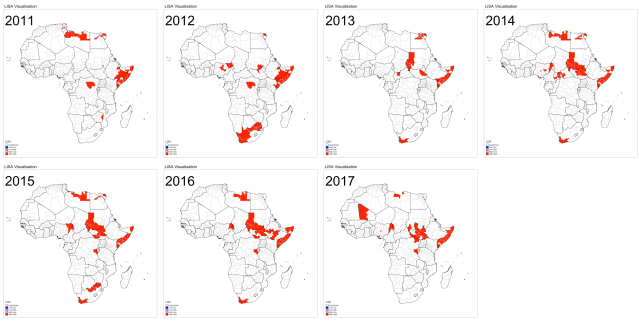

Figure 15: LISA Cluster Map Phase III – from 2011 to 2017

4.2.2.6 Interpretation of Results. From the LISA cluster plots, the only significant classification that is observed in the whole dataset is the High-High clusters. Nevertheless, this method is useful in highlighting violence "hot-spots" and show how they develop over time. Also, violence against civilians does appear to concentrated and recurrent in a few countries. In particular, a lot of the violence hotspots tend to occur near the boundaries i.e. borders between countries. In the first phase, violence is observed to congregate in clusters that gradually spread out to neighboring administrative regions over time. Key clusters to note are: (1) Sierra Leone cluster in the western part of Africa; (2) the conflict bordering Angola and Southwest Democratic Republic of Congo; and (3) the pockets of violence in the region consisting of Sudan, South Sudan, Ethiopia, Kenya, Uganda and the northeastern border of Democratic Republic of Congo.

In Phase I, Sierra Leone was identified as the centre of a hotspot of violence against civilians from 1997- 2001. From 1997 to 1998, violence was concentrated within the borders of Sierra Leone. It appeared to have spread over to Guinea in 1999 in small pockets of violence along the Sierra Leone-Guinea border and split over to Liberia in 2000. The violence appeared to surge in 2001 before completely disappearing from Sierra Leone in 2002, with some remnants lingering in Liberia in 2002 and 2003.

Interestingly the hotspot of violence in Sierra Leone highlighted by the LISA method overlapped with the Sierra Leone Civil War, which began on 23 March 1991 and ended in 18 January 2002. This civil war was notorious for the flagrant human rights abuses that had occurred, with the mass killings of civilians, war rape and sexual slavery (Human Rights Watch, 1999; Martin 2003).

We also observe that the violence clusters in (2) first started in the region near the northern border Angola and southwest of the Democratic Republic of Congo in 1997 and gradually spread over the years southwards to eventually engulf the whole of Angola in 1999, before subsiding over the next two years and eventually disappearing by 2002. This period coincided with the Angolan civil war which began in 1975 and concluded in 2002. Other notable events that happened in this period is the First Congo War (from 1996 to 1997) which resulted in the formation of the Democratic Republic of Congo as a sovereign nation of which Angola played a role (Turner 2002).

The cluster of hotspots identified by the LISA plot in the countries in (3) mostly took place in small pockets between borders, which grew into larger clusters, culminating in a somewhat contiguous corridor of violence as observed in 2002. This period of violence coincided with the Second Congo War, a.k.a. the Great African War, which began in the Democratic Republic of Congo in 1998 and ended in 2002, involving along the way the massacre of civilians including the horrific systematic extermination of the Bambuti pygmy people in the Democratic Republic of Congo by rebel forces (Penketh 2004).

This crescent shaped corridor is of interest as it appears to be a regular theatre of conflict, and significant pockets of violence against civilians tend to recur sporadically over time, especially in the Figure 14 of Phase II of the LISA plots above from 2004 to 2010. Most notably, it had also appeared again in 2006, with the significant pockets of violence occupying more or less the key regions as seen in 2002.

Interestingly, this crescent shape corresponds with the the general shape of the main rivers in Africa (see Figure 16). This observation resonates with the distinct pocket of violence in Phase II occurring in the Ethiopia-Somalia border, which actually seemed to have begun in the last year of Phase I in 2003 and slowly grew to a large cluster of violence that sustained throughout Phase II from 2004 to 2012 and even continued to the start of Phase II till 2012.

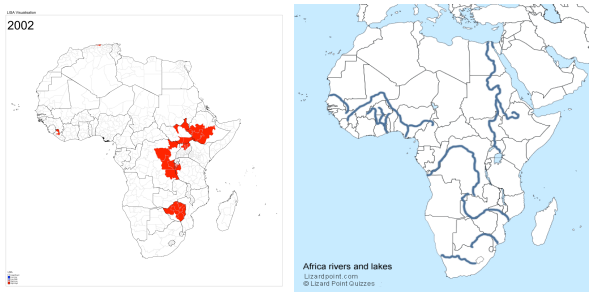

Figure 16: Corridor of violence vs Africa's main rivers? (Copyright Lyndsey McCollam/Lizard Point Quizzes)

On 2013 onwards, the violence against civilians was localized in Somalia, which history of conflict with Ethiopia spans decades. This is most likely caused by the Ethiopian-Somalian "border dispute" that has been going on for around half a century (Kendie 2007), where the main drivers of this conflict are economic interests and insecurity, and the struggle for resources, including contestation of grazing lands and water wells for the nomadic tribes which traverse the Oromia-Somali borders (Solomon 2017), as well as control over the Juba and Wabi Shebeli rivers.

In Phase III, there is a notable pocket of violence that occurs in the Sudan-South Sudan border region from 2014 to 2017. Interestingly, violence here hardly ever occurred in the northern part of Sudan but mostly along its south and south western border. Like Ethiopia and Somalia, a history of conflict between Sudan and South Sudan exists, especially since the latter's independence on 9 July 2011. However, the border between these two nations have never been formally demarcated. As the region has the most fertile land of both countries coupled with the rich oil resources present, the land has been rife with conflict (Craze 2013).

5 CONCLUSION

The study of the spatial distribution of violence against civilians in Africa from 1997 to 2017 using the approach of Exploratory Spatial Data Analysis (ESDA) had generated much insights from a large body of data. The LISA analysis in particular can be used to locate statistically significant hotspots of violence against civilians in a dataset like the ACLED, more rigorously than using the traditional choropleths. We also managed to validate our findings with actual historical cases. This demonstrates that the ESDA and LISA method has tremendous potential as a knowledge discovery tool and methodology. Nevertheless, while ESDA relies on formal statistical models and mathematics, much of it remains art and cannot be applied without deliberation. Geospatial analytics methods such as ESDA are an important supplement to traditional analysis of political violence as they provide the locational context for the data on violence.

REFERENCES

[1] A. Cliff, and J. Ord (1973). Spatial Autocorrelation. Pion: London.
[2] A. Penketh (2004). "Extermination of the pygmies". The Independent. Retrieved 2 August 2017.
[3] C. Flint , P. Diehl , J. Scheffran , J. Vasquez and S.h. Chi (2009) Conceptualizing ConflictSpace: Toward a Geography of Relational Power and Embeddedness in the Analysis of Interstate Conflict, Annals of the Association of American Geographers, 99:5, 827-835, DOI: 10.1080/00045600903253312 https://doi.org/10.1080/00045600903253312
[4] C. Raleigh and J. Moody (2017). REAL-TIME ANALYSIS OF AFRICAN POLITICAL VIOLENCE(Publication No. 55). Retrieved https://reliefweb.int/sites/reliefweb.int/files/resources/ACLED_Conflict-Trends-Report-No.55-February-2017-pdf.pdf
[5] D. A. Griffith, (2009). Spatial autocorrelation. International encyclopedia of human geography, 2009, 308-316.
[6] D. Kendie (2007) "Towards Resolving the Ethiopia-Somalia Disputes" International Conference on African Development Archives. Paper 104. http://scholarworks.wmich.edu/africancenter_icad_archive/104
[7] H. Buhaug, and K. S. Gleditsch. 2008. Contagion or confusion? Why conflicts cluster in space. International Studies Quarterly 52:215–33.
[8] Human Rights Watch. July 1999 "Getting Away with Murder, Mutilation, Rape: New Testimony from Sierra Leone". 1999. https://www.hrw.org/legacy/reports/1999/sierra/SIERLE99-03.htm Retrieved 17 August 2018.
[9] J. Craze (2013). Dividing lines, grazing and conflict along the Sudan-South Sudan border.
[10] J. Benson (2018). The Geography of Violence against Civilians: Implications for Peace Enforcement. An OEF Research Discussion Paper. OEF Research http://dx.doi.org/10.18289/OEF.2018.024
[11] J. O'Loughlin and L. Anselin. (1992). Geography of interna- tional conflict and cooperation: Theory and methods. In The new geopolitics, ed. M. D. Ward, 11–38. Philadelphia: Gordon and Breach.
[12] K. Martin (June 2003), Disarmament and Demobilisation in Sierra Leone, Humanitarian Exchange, Humanitarian Practice Network https://odihpn.org/magazine/disarmament-and-demobilisation-in-sierra-leone/
[13] L Anselin (1995) Local indicators of spatial assoc iation-LISA . *Geographical Analysis* 27:93-115
[14] L. Anselin (1998). Exploratory spatial data analysis in a geocomputational environment. Pp. 77–94 in Geocomputation: A Primer, edited by P.A. Longley, S.M. Brooks, R. McDonnell, and W. Macmillian. New York: Wiley and Sons.
[15] P. A. P. Moran (1950). Notes on continuous stochastic phenomena. Biometrika, 37, 17-23.
[16] R. Geary (1954). The contiguity ratio and statistical mapping. The Incorporated Statistician, 5, 115-145.
[17] RiskMap 2018: Political violence and crime rise across the world in 2017(Rep.). (2018, January 30). Retrieved https://www.controlrisks.com/riskmap-2018/articles/political-violence-and-crime
[18] R. Malley (2018). 10 Conflicts to Watch in 2018. Retrieved from http://foreignpolicy.com/2018/01/02/10-conflicts-to-watch-in-2018/
[19] R. Youngs (2017, October 2). What Are the Meanings Behind the Worldwide Rise in Protest? Retrieved from http://carnegieeurope.eu/2017/10/02/what-are-meanings-behind-worldwide-rise-in-protest-pub-73276
[20] S. Giest, (2017) Policy Sci 50: 367. https://doi.org/10.1007/s11077-017-9293-1
[21] S. Chojnacki, M. Grömping and M. Spies (2009) Armed Conflict Beyond the State: Spatial and Temporal Patterns of Non-State Violence in Somalia, 1990-2007. Paper for the Joint CSCW WG3/GROW-Net workshop: Environmental Conflicts and Conflict Environments, Oslo, 11–12 June 2009
[22] S. Schutte and N. B. Weidmann, (2011). Diffusion patterns of violence in civil wars. Political Geography, 30(3), 143-152.
[23] S. Solomon (2017) What's Driving Clashes Between Ethiopia's Somali, Oromia Regions? Voice of America's Africa Division. https://www.voanews.com/a/whats-driving-clashes-ethiopias-somali-oromia-regions/4050017.html
[24] T. Turner (2002) Angola's Role in the Congo War. In: Clark J.F. (eds) The African Stakes of the Congo War. Palgrave Macmillan, New York
[25] W. R. Tobler 1970. A computer movie simulating ur- ban growth in the Detroit region. Economic Geography 46:234–40.