SFB 700
GOVERNANCE
IN AREAS OF LIMITED STATEHOOD

New Method, Different War?

Evaluating Supervised Machine Learning by Coding Armed Conflict

Christian Ickler/John Wiesel

SFB-Governance Working Paper Series • No. 39 • September 2012

DFG Sonderforschungsbereich 700 Governance in Räumen begrenzter Staatlichkeit - Neue Formen des Regierens?
DFG Collaborative Research Center (SFB) 700 Governance in Areas of Limited Statehood - New Modes of Governance?

SFB-Governance Working Paper Series

Edited by the Collaborative Research Center (SFB) 700 "Governance In Areas of Limited Statehood - New Modes of Governance?"

The SFB-Governance Working Paper Series serves to disseminate the research results of work in progress prior to publication to encourage the exchange of ideas and academic debate. Inclusion of a paper in the Working Paper Series should not limit publication in any other venue. Copyright remains with the authors.

Editorial assistance and production: Insa Eekhoff/Anna Jüschke/Sophie Perl

All SFB-Governance Working Papers can be downloaded free of charge from www.sfb-governance.de/en/publikationen or ordered in print via e-mail to sfb700@zedat.fu-berlin.de.

Ickler, Christian/Wiesel, John 2012: New Method, Different War? Evaluating Supervised Machine Learning by Coding Armed Conflict. SFB-Governance Working Paper Series, No. 39, Collaborative Research Center (SFB) 700, Berlin, September 2012.
ISSN 1864-1024 (Internet)
ISSN 1863-6896 (Print)

This publication has been funded by the German Research Foundation (DFG).

DFG Collaborative Research Center (SFB) 700
Freie Universität Berlin
Alfried-Krupp-Haus Berlin
Binger Straße 40
14197 Berlin
Germany
Phone: +49-30-838 58502
Fax: +49-30-838 58540
E-mail: sfb700@zedat.fu-berlin.de
Web: www.sfb-governance.de/en

Deutsche
Forschungsgemeinschaft
DFG

New Method, Different War? Evaluating Supervised Machine Learning by Coding Armed Conflict

Christian Ickler and John Wiesel

Abstract

The internet promises ad hoc availability of any kind of information. Conflict researchers seem to be bound only by the effort needed to find and extract the necessary information from international news sources. This begs the question of whether the sheer number of accessible news sources and the speed of the news cycle dictate an automated coding approach in order to keep up. Will the initial costs of implementing such a system outweigh the possible loss of information on violent conflict? We answer these questions in relation to the Event Data on Armed Conflict and Security project (EDACS) where we carry out both human and machine-assisted coding to generate spatiotemporal conflict event data. We use spatiotemporal comparability measures for quantitative and qualitative comparison of the two datasets. While the quality of human-coding exceeds a purely automated approach, a compromise between efficiency and quality results in a supervised, semi-automated machine learning approach. We conclude by critically reflecting on the possible discrepancies in the analysis of these resulting datasets.

Zusammenfassung

Das Internet verspricht ad hoc Verfügbarkeit jedweder Information. Konfliktforscher müssen daher dem Anschein nach nur noch die gewünschten Informationen finden und extrahieren. Dies wirft die Frage auf, ob die schiere Zahl verfügbarer Nachrichtenquellen und die Geschwindigkeit des Informationsflusses eine Maschinenkodierung zwingend notwendig machen? Und wiegen die initialen Kosten der Implementierung eines solchen Systems die Kosten des möglichen Informationsverlustes auf? Wir haben diese Fragen für das Event Data on Armed Conflict and Security Projekt (EDACS) beantwortet und im Zuge dessen, sowohl manuell als auch semiautomatisch, raumzeitlich desaggregierte Ereignisdaten eines bewaffneten Konflikts kodiert. In diesem Papier stellen wir beide Ansätze quantitativ und qualitativ mit Hilfe raumzeitlicher Vergleichsmaße einander gegenüber. Während die Qualität manuell kodierter Daten die maschinell erstellter Daten übertrifft, bietet die semi-automatische Variante einer überwachten Maschinenkodierung einen Kompromiss zwischen Effizienz und Qualität. Wir schließen mit einer kritischen Aufarbeitung möglicher Diskrepanzen in Analysen basierend auf den beiden Datensätzen.

Table of Content

1. Introduction

Spatially and temporally disaggregated event data has become the backbone of quantitative conflict science literature. A growing number of georeferenced datasets provides the necessary information on violent incidences in armed conflict (Chojnacki et al. 2012a; Dulic 2010; Melander/Sundberg 2011; Raleigh et al. 2010). These spatiotemporal disaggregated datasets enable researchers to analyze variations of violence in time and space (cf. Buhaug 2010; Raleigh et al. 2010; Weidmann et al. 2010).

The Event Data on Armed Conflict and Security project (EDACS) develops and maintains one of these datasets. EDACS focuses on violence in areas of limited or failed statehood. The dataset encompasses seven countries of Sub-Saharan Africa (Burundi, Democratic Republic of the Congo, Liberia, Republic of the Congo, Rwanda, Sierra Leone, and Somalia) between 1990 and 2009. While this effort has reached its final stages, it took years and several thousand working hours to complete the process. One of the most time-consuming and error-prone phases of data generation is the step of data transformation or "coding" (cf. Chojnacki et al. 2012a), which has been an almost entirely manual task. The rise in numbers and increased availability of news sources over the internet raises the question of whether the sheer number of accessible news sources and the speed of the news cycle dictate an automated coding approach in order to keep up. Will the possible gain of information on concomitants of violent events outweigh the initial cost of implementing such a system? Will such a system be able to achieve the necessary degree of data quality?

Based on these questions, we carry out an experiment using machine learning (ML) to generate spatially and temporally disaggregated event data of the armed conflict in Sierra Leone in the year 1999, and compare the resulting dataset with human-coded event data. In the first part of this paper, we will describe the Event Data on Armed Conflict and Security project (EDACS), referring to the relevance of computer-supported natural language processing to conflict event data projects in general and describing the implementation of our ML-based approach.

In the second part we will present the comparative experiment above, discussing the experimental design, the costs originating from the two different approaches, the level of data comparability, and the overall pros and cons of machine and human coding. In addition, we will briefly describe the methods to perform a step-by-step comparison of machine-learning and human-coded conflict event data according to their spatial and temporal attributes. We therefore first analyze the time series in both datasets, in search for similar trends. Second, we map and compare the spatial distribution of the two datasets. Third, we apply spatiotemporal K-functions and plot matching events defined by a narrow spatiotemporal threshold, which is based on an SQL query combined with the results of a spatiotemporal cluster analysis. In order to understand the task at hand, we begin by outlining the foundations of EDACS, along with its goal, scope, and core definitions.

2. Transformation of News Articles into Conflict Event Data

The transformation of natural language into structured event data requires a thorough conceptualization in order for the resulting database to achieve relevance. Any data project must also recognize the inherent challenges of the data generation process. We will begin by explaining the concepts behind EDACS and other data projects, then touch on the subject of data quality in the field of conflict research, and finally describe two crucial steps of the data generation process: the selection of source documents and subsequent extraction of events.

2.1 Conceptualization of Conflict Event Data

In EDACS, the basic unit of analysis is an event, defined as a violent incident with at least one fatality resulting from the direct use of armed force. Events are coded with their location (name of location, longitude and latitude coordinates) and timeframe (start and end date in the case of events lasting more than one day). Among others, the type of military action (fighting[1] or diverse forms of one-sided violence[2]) is coded in EDACS, as well as the violent or non-violent actors involved and any details on (civilian and military) fatalities. Events can be based on one or several news sources. For each event, the names and publication dates of all news articles used are indicated. Beyond that, EDACS coders mark an article as "biased" if its information originates from a source directly connected to a violent actor involved in the respective event.

EDACS is built on information retrieved from newspaper articles. These articles are gathered from the LexisNexis news portal. EDACS is based on a set of four predefined sources or media outlets that are used for all coded countries and years of observation. The archives of three international newspapers (Guardian, New York Times, and Washington Post) and the broad collection of translated local news reports by BBC Monitoring are searched by keywords through the news portal. In case of inconsistent information or missing data on one of EDACS's central variables (location and timeframe of event), the four mandatory sources are supplemented by other sources such as other news services (trust.org/alertnet, irinnews.org, crisisgroup.org, humansecuritygateway.com), and regional internet gateways (allafrica.com, africa-confidential.com, reliefweb.int).

Each news article found by the search engine is read by EDACS coders who extract the relevant information and enter it into a data entry form. In order to ensure inter-subjectivity and data consistency, a set of strict and conservative coding rules has been developed. Additionally, all data is coded and double-checked by two different coders and cross-checked by a supervising coder (cf. Figure 1 – Human Coding Procedure, p. 7).

1 Mutual violence is defined in EDACS as "armed interaction between two or more organized groups" (Chojnacki et al. 2012b: 4).

2 One-sided violence is defined in EDACS as "direct unilateral violence by organized groups aimed at civilian or military targets" (Chojnacki et al. 2012b: 4).

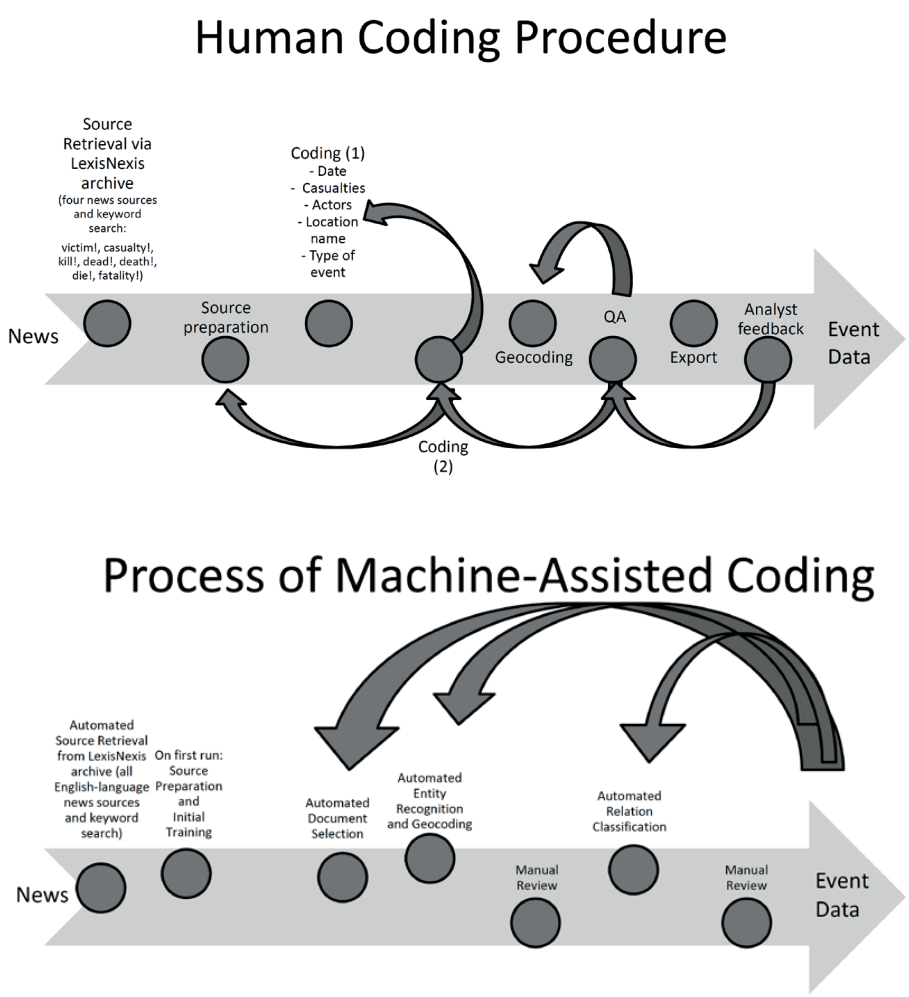


Figure 1: Human Coding Procedure and the Process of Machine-Assisted Coding

Events are localized with longitude and latitude coordinates (WGS 84) using the toponymic GEOnet Names Server (GNS) provided and maintained by the US National Geospatial-Intelligence Agency.[3] GNS data is an extensive settlement dataset that is easily accessible at no charge. In case of imprecise ("between town A and town B") or ambiguous (duplicate location names in the GNS-data) indications of event locations in a primary sources, EDACS coders consult additional map data such as GoogleEarth or Harvard's AfricaMap and/or apply a variety of standardized buffer rules.[4]

EDACS differs from other semi-automated georeferenced conflict event data projects, such as the Armed Conflict Location and Events Dataset (ACLED), in its stricter event definition. Among other variables, EDACS codes the number and type of fatalities, as well as the involved

3 The toponymic GEOnet Names Server (GNS) database is maintained by the US National Geospatial-Intelligence Agency (NGA) and provides location names and coordinates in the World Geodetic System 1984 (WGS 84) on a global level (http://earth-info.nga.mil/gns/html/, last accessed 24 September 2012). GNS provides an extensive settlement dataset that is easily accessible at no charge.

4 EDACS bias, buffering, and overall coding rules and procedures are described in the Codebook: www.conflict-data.org.

violent or non-violent actors.[5] Unlike the Uppsala Conflict Data Program Georeferenced Events Dataset (UCDP-GED), EDACS covers events with unknown actor participation and includes actors and dyads not surpassing the UCDP-GED-threshold of twenty-five conflict-related fatalities per year (Melander/Sundberg 2011). On the one hand, EDACS provides a more comprehensive view of patterns of violence in the observed countries by factoring in all actors, also those who remain unidentified; on the other hand, UCDP-GED data may be more reliable because it only considers events with clearly identifiable actors, surpassing the 25-fatality threshold, which might be more relevant to the armed conflict as such (Chojnacki et al. 2012a).

2.2 Challenges to Data Quality

Event data projects such as EDACS generally face four categories of challenges to data quality: errors and bias contained in (1) the source (news) or (2) the auxiliary data (maps, etc.), as well as faults in (3) the transformation processes from source data into event data (misinterpretation, oversights, etc.), or (4) the contextualization of the events using auxiliary data (event localization, actor identification, etc.) (Chojnacki et al. 2012a).

While machine coding techniques could, in theory, be used to address each of these challenges, we will specifically evaluate how automated source selection and data transformation impacts data quality. When measuring data quality, we will consider the following aspects: completeness, accuracy, consistency, and relevancy (Batini/Scannapieca 2006: 40; Thion-Goasdoué et al. 2007). For instance, machine coding can improve the speed of the coding process and thereby increase the completeness of the resulting data. But machine coding may be susceptible to errors contained in the auxiliary data and possible faults in the processes of spatiotemporal contextualization. Fuzzy specification of locations in the sources can deteriorate the performance of any geocoding approach (Pasley et al. 2007); ambiguous location names (ambiguous toponyms) can have the same effect (Clough 2005; Leidner 2007).

Furthermore, while machine coding aims to improve the accuracy and consistency of the transformation of raw text into structured data, results can be very misleading. As early as 2003, King and Lowe experimented with the machine coding of events, and event extraction using the proprietary software provided by Virtual Research Associates, Inc. called VRA-Reader (King/Lowe 2003). Although the authors reported "virtually identical" accuracy of machine coding to human coding, upon closer examination, the results were mixed: a high precision of 93% was accompanied by a false positive rate of 77%, meaning that 77% of sources were wrongly classified as containing an event (King/Lowe 2003: 632). Using an unfiltered corpus of documents, their approach would result in vast amounts of false data and would eventually render any automated coding result useless if applied to other event data projects. A further downside to using proprietary software for this purpose is that it may decrease the openness of the resulting database and deteriorate reliability, as argued by Kauffmann (2008: 108).

5 A more detailed description of the dataset can be found in the download section of our website: www.conflict-data.org.

Today, almost all data projects such as ACLED, Minorities at Risk, or UCDP either rely on manually coded data or use simple lists and lookup mechanisms for data generation (Nardulli et al.2011: 10). But adaptive ML techniques have been shown to outperform static natural language processing (NLP) approaches, in particular when facing "noisy" data such as news reports in the conflict domain (Carlson et al. 2009; Sarawagi 2007). Due to this finding and the progress in NLP, some event data projects such as the Integrated Crisis Early Warning System (ICEWS) (Schrodt 2011) and the Social Political and Economic Events Database project (SPEED) (Nardulli et al. 2011) are currently transitioning to adaptive NLP-based event extraction techniques. While ICEWS has not yet documented any results following their transition from the static, rule-based NLP software Textual Analysis By Augmented Replacement Instructions (TABARI) to a new system, SPEED has already implemented a system to identify possibly relevant articles and uses NLP to find proper nouns in text to help coders identify participating actors or location names. An adaptive, custom-trained, NLP addition to the system is under development. In this case study for EDACS, we have implemented and completed a similar system that takes this further step and makes use of artificial intelligence to learn how to identify and distinguish between actors, casualties, and locations, and directly annotate news articles. To prevent the system from generating too many false positives and to reduce the overall workload, it is necessary to preselect or filter the relevant from the irrelevant source documents.

2.3 Filtering the Filters

In the context of conflict research, it can be argued that achieving a certain level of representativeness is equivalent to a high degree of completeness or relevancy: "a sample of even 5% of [real] events would not be problematic if it were truly representative" (Earl et al.). Thus, to have a variety of databases on the same subject of investigation is a clear advantage, as it enables us to perform a comparative analysis across datasets (Chojnacki et al. 2012a). We will therefore compare two different sets of sources:

The first is retrieved by a simple keyword search of the LexisNexis archive and restricted to four media outlets (Guardian, New York Times, Washington Post, and the broad collection of translated local news reports by BBC Monitoring). The sources are then manually processed in their entirety and in chronological order.

For the second set, we repeat the retrieval but lift the restriction on the media outlets. From this expanded array of sources, roughly five times larger than in the first search, we use a machine-learning (ML) approach to select a sample of the sources for event extraction. We expect that both approaches' results will show some similarity if in both cases the generated database shows a certain degree of completeness and relevancy.

The second approach aims to classify the sources and choose documents that are relevant to the subject of interest. In contrast to projects like UCDP, which to a certain extent use the static approach of VRA (Gleditsch et al. 2002; Harbom/Wallensteen 2009), we employ an active, adaptive approach that learns over time to select the right documents for the user. We aim

to drastically reduce the number of documents that have to be reviewed manually. Adaptive document classification has become common and can be found in everyday email clients, but has only recently been applied in this field by Nardulli et al. (2011). Instead of finding events themselves, the aim is to select, or classify, the documents that are either unrelated or relevant to the conflict, to allow researchers to understand the dynamics of the armed conflict in Sierra Leone. A subset of these articles contains the conflict events. In order to perform this selection, we pair two different models, naive Bayes and boosted decision trees, to perform the candidate selection.[6] Classifiers based on ML algorithms need so-called "features" as input, normally a pre-selection of words, grammatical attributes, or similar characteristics. As feature selection in text classification tasks often delivers varying outcomes (Kim et al. 2006: 1460), we employ a straightforward bag-of-words approach, without filtering out common words (stop-word approach) or linguistic transformations such as stemming. Initially, since no trained model exists yet, random articles are selected from the documents. Beginning with the second coding session, models are trained from the documents that are discarded or used by the users at each new start. The next possible candidate for event extraction is then selected from these documents.

2.4 Event Extraction

We complement the automated document classification with a method for ML-based event extraction. In the section above we explained how, on a document level, classifiers are trained to select documents of interest from a large set of source documents, our corpus. In addition to this, we use ML on an event level, and use sequence labeling to identify phrases that signify an event according to the EDACS definition and the related entities such as actors, time, and place.

Requirement for an event in the definition of the EDACS project are casualties from a violent event. Inspired by Banko et al. (2008), we create a sequence tagger that performs as a casualty extractor to identify phrases in which casualties are mentioned. The underlying model is based on conditional random fields as that model has proven to be highly efficient, although the calculation needed is central processing unit (CPU) intensive.[7] After each coding session, a new, updated model is trained, incorporating the new training data into the model (see Figure 1 – Machine-Assisted Coding, p. 7).

We employ the same approach to the identification of phrases denoting actors, locations, and dates. While there may be proper nouns, there are also composite names such as "Liberian rebels" or more general locations such as "the border". The proper nouns may be found using existing, pre-trained taggers, but the common nouns are normally not found, not usually being part of the training data used to train the model. Furthermore, they are application domain specific. In a further step to support the annotation process, we combine both elements: a pre-

6 Both decision trees and the meta-classifier AdaBoost are part of Carnegie Mellon's MinorThird pakkage (Cohen 2004). The naive Bayes classifier is from LingPipe's natural language processing (NLP) libraries (Carpenter 2010): http://alias-i.com/lingpipe, last accessed 24 September 2012.

7 We use LingPipe's commercial implementation due to its high encapsulation and decent documentation, which is free for research use (Carpenter 2010).

trained tagger to identify actors and location names in text, and a custom tagger trained using the annotations provided by the pre-trained tagger combined with the manually revised annotations to recognize these specific forms of actor and location names. Inspired by Stanford's approach to sequence tagging for named entity recognition (Finkel et al. 2005) and the conclusions drawn for our casualty extractor, we use LingPipe's implementation for the custom tagger and pair it with Stanford's tagger, whose pre-compiled model achieved f-scores[8] of 86% on the 2003 corpus used at the Conference on Computational Natural Language Learning (Finkel 2007; Finkel et al. 2005). While sequence taggers can in principle be used to identify relationships occurring in sequences of words (Banko et al. 2008), related entities in this context do not necessarily appear in the same sequence, but may be sentences apart. When manually extracting events, relations become implicit when the user enters event by event. When annotating text, we have to explicitly define relations. We combine the sequence labeling approach outlined above with an idea by Ahn (2006) and use the "anchor" phrases provided by the casualty tagger for ML-supported relation extraction.

Drawing upon all of the entities and phrases identified above, the relation extractor classifies the relationships between these annotations. As per design, a casualty phrase forms the anchor of the event. This reduces the complexity of the task by evaluating all relationships between the entities in a text and a particular casualty phrase instead of considering all possible combinations. The relations are extracted as a binary classification task, using a maximum entropy classifier. An actor, a location, or a date is either related to the event in question, or not.

2.5 Process of Machine-Assisted Coding

First, the system iterates through all articles until one of the document classifiers identifies a candidate article. The program automatically executes the sequence taggers to identify possible casualty phrases in the text, as well as entities such as locations and dates. The example text in Figure 2 (p. 13) contains three incidents.

The first is an attack by rebels of the Revolutionary United Front (RUF). The casualties, our event anchor, are identified correctly. It therefore also meets the minimum criterion of one casualty and is a valid target for event extraction. Two other incidents are mentioned: the second is the beginning of an offensive by forces of the Economic Community of West-African States Monitoring Group (ECOMOG), while the third is the inauguration of a new Chief of National Security. Neither is related to the first event, and neither one is an event in itself according to the definition used here. Therefore, they should not be extracted.

Second, the user reviews the tags presented and corrects accordingly all annotations that relate to the event. This is a necessary step, as even well-trained annotators have been shown to in-

8 The f-score is most often defined as the harmonic mean of precision (defined as the ratio of relevant items retrieved and all retrieved items) and recall (defined as the ratio of relevant items retrieved and all relevant items) (Manning et al. 2008: 156).

troduce a "very high" number of spurious instances (Giuliano et al. 2007). This could adversely affect any further steps, in our case the following relationship classification. The text highlights are updated accordingly in a different highlight style.

Third, the user runs the relation classifier to determine which annotations are related. Lastly, the user reviews whether the relation classifier has performed its task correctly. In the given example, it performs without error. If necessary, the user adds or removes relations where appropriate using a context menu. After finishing an article, the program selects the next candidate article for annotation. All annotations are stored by a open-source software library, Apache's Unstructured Information Management Architecture (UIMA), which uses the open standard XML.[9]

After this data transformation of free text into semi-structured data, we use a simplified, automated version of our data contextualization procedure. We automatically geocode location names (toponyms) and set coordinates by performing a lookup in the GNS database (see Figure 1 – Machine-Assisted Coding, p. 7). The software also deduces dates from temporal descriptions and meta-information, such as the publication date, using a small set of rules. This final machine-based coding procedure is actually capable of producing an entire dataset in a matter of seconds. Alterations in the coding rules are achieved by simply changing the software code accordingly and running the software once, whereas changes in the coding rules in traditional dataset generation efforts – such as our own EDACS project – may make tedious manual recoding necessary. The results of this of procedure are then used as an ad-hoc set of data for comparison with our set of reference data, a subset of our manually generated and extensively reviewed EDACS database.

3. Evaluation

Event extraction based on ML has already proven to be a useful application of artificial intelligence. The central question is whether the effort needed to apply this technique to the domain of conflict research can be justified. We can answer this question in the context of the EDACS project through a simple experiment that enables us to compare the ML-generated events with the EDACS dataset, which has been extracted, geocoded, and proofed twice – all manually.

First, we evaluate the performance of our classifier-based approach to finding relevant documents by storing the numbers of correctly identified relevant and irrelevant articles per session. Next, we evaluate the performance of the overall system by recoding the time needed to extract the resulting set of events. Finally, we investigate the resulting data quality more closely by measuring the temporal and spatial similarity of the two datasets. But firstly, we will outline the approach of our experiment.

9 The Apache Software Foundation, Apache UIMA, http://www.apache.org/, last accessed 24 September 2012.

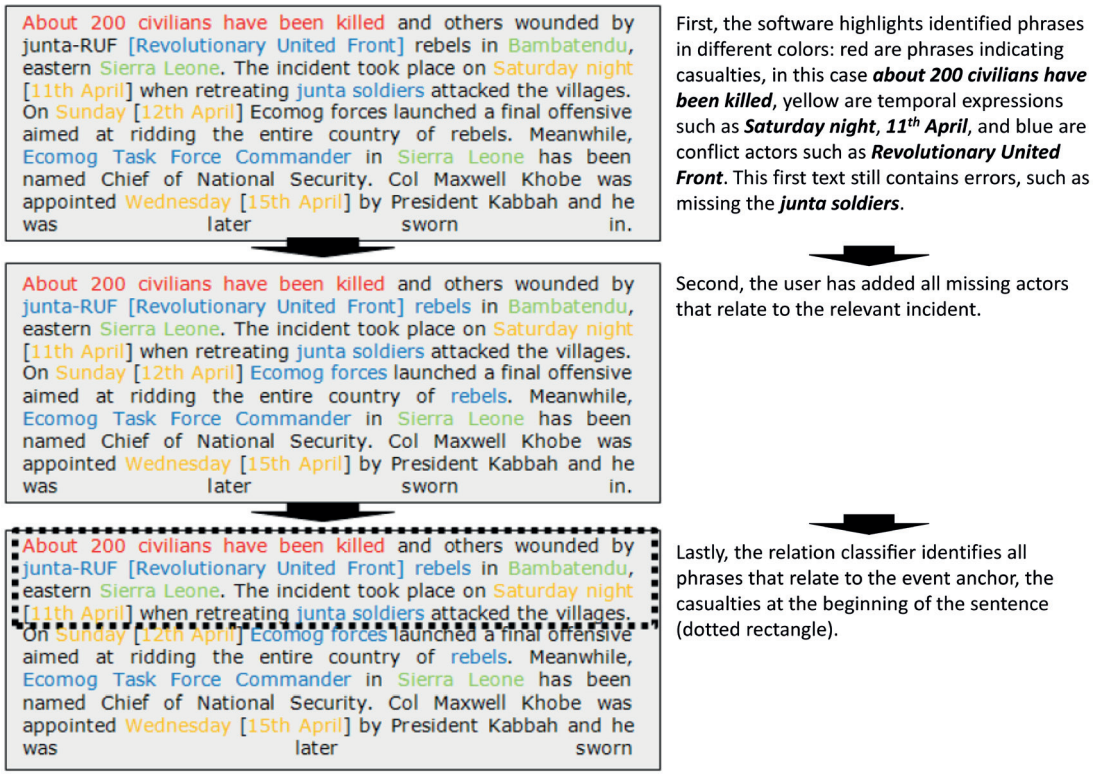


Figure 2: The machine-assisted coding consists of three steps. First, the software highlights identified phrases in different colors; second, the user adds all missing actors that relate to the relevant incident; and lastly, the relation classifier identifies all phrases that relate to the casualties at the beginning of the sentence (dotted rectangle).

3.1 Experiment Setup

For the purposes of this experiment, we restrict the scope of our sources to the year 1999 and extract only events for the case of Sierra Leone. For the machine-learning approach, we retrieve 7,000 articles from only English-language sources in the LexisNexis archive for Sierra Leone 1999, based on a simple keyword search. For the manual approach, EDACS originally retrieved 1,200 articles from the four media outlets. As there is no training data (i.e., extracted events) at the beginning, the software can only rely on pre-existing models that allow it to highlight actor names, locations, and dates. The software cannot yet identify common nouns as actor names, casualties, and relations, nor distinguish between relevant and irrelevant documents. Each session provides new training data, which the machine-learning algorithm uses to create a mathematical model; the software then employs this model to present the user with the next candidate document, highlighting the identified phrases of relevance.

The participating coders are experts in manual coding but completely unfamiliar with machine-assisted coding, and receive minimal training beforehand. In order to ensure that the generated dataset be spatially comparable to the human-coded EDACS data, we ask one of them to review the automatically assigned coordinates, as the geographic database used contains am-

biguous entries. We also restrict the comparison to events with exactly specified dates and settlements only.

3.2 Cost Evaluation

One crucial, limiting element for all research is the available budget. The cost generated within projects can truncate primal objectives strongly, and especially data projects depend on a sufficient budget – in particular during their setup phase. The budget limits the targeted coding project dimensions (number of coded cases and years, etc.) or potential retrospective refinement to the project design and coding rules.

We will outline the actual cost of coding for the case of Sierra Leone 1999, comparing approximate human coding costs with the costs of machine learning. We have chosen Sierra Leone 1999 because it represents an "average" swaying conflict year, with phases of escalation and de-escalation.

The costs deriving from the two different approaches can be divided into four major categories: facilities, development, coders, and output (respectively the number of events over time). The facilities and development include building occupancy expenses, office equipment, purchase of computers and software, and the provision of a database server (for remote coding and centralized access). The setup of a database, programming of a data entry form, etc., and development of the overall coding rules – things we subsume under "facilities and development" – are hard to quantify. The costs for facilities provided partly by the Collaborate Research Center 700 and the EDACS project amount to roughly 58,500 USD for the period of five years that the coding has been underway. The money and time spent on setting up the database and the first version of the data entry form only total to about 2,123 USD. Further development costs, including the salary of at least two research fellows (for five years) increase the expenses to about 320,579 USD.

In the case of machine learning, most of these costs also accumulate; the crucial difference is the specialized knowledge needed to implement such a system. There are no out-of-the-box solutions for ML-based event extraction, which makes custom development necessary. In this case, the actual software development of our prototype alone took up to half a year, while developing a full-fledged "classic" database and entry form software system took roughly 160 hours. Overhead such as planning, research, etc. is not included in these figures.

The costs of coding mostly depend on the research assistants' salary (in the case of EDACS, 14.39 USD/hour), which equals about 120 read news article pages per hour; for Sierra Leone 1999, consisting of 1,442 pages for the four sources used within EDACS, this adds up to 172.92 USD. The ML approach, in comparison, only requires about 66% of the working hours, costing roughly 115.28 USD. One has to bear in mind that every coded year has its own characteristics with regard to conflict dynamics and news coverage. Therefore, the number of events per source fluctuates immensely. Still, the experiment only spans about 0.8% of the entire source data

for the EDACS project. In a rough total, the ML approach could save more than 14,000 USD or 974 hours of manual work when applied to the entire project.

The generated output (the number of coded events per hour) is unequal: whereas the human coders needed twelve hours to complete the first round of coding, the prototype ML coding only required two-thirds of that time. But this prototypical comparison is only feasible because the datasets contain differing details. The human-coded dataset offers more information, but also requires a second round of coding. The lack of detail is further mitigated by the fact that the prototype ML approach does not identify duplicate events; every event is automatically annotated and then manually revised. The gross increase in events coded per hour by the ML approach compared to manual event coding was 156%.

3.3 Performance Evaluation

A prerequisite for efficient event extraction is the reduction of the source corpus to a manageable size. Restricting the source articles to four media outlets and texts containing certain keywords only barely accomplishes this task. Only about 3.2% of the articles retrieved from LexisNexis are actually used by EDACS for event extraction in the case of Sierra Leone. Over 95% – thousands of articles per country – have to be scanned and discarded manually. A fully automated event extraction approach used on an unfiltered corpus would lead to large amounts of false positives. By introducing a document classifier, we are able to significantly reduce the number of documents used for extraction. For the year 1999, a year with relatively intense fighting, about 90% of articles had to be discarded manually in the manual coding. When using the document classifier approach, on average, about 40% of articles selected by the classifier were deemed relevant by the human coder, and half of those contained an event. Although we used a simple and fast classification method, and applied it to a corpus seven times as large as during our manual efforts, the ratio of events per document doubled on average. This decreases the overall workload, since fewer documents have to be scanned manually to reach the same number of events. Figure 3 (p. 16) shows how the classifier improves over time beginning with session one, after a random subset of documents is manually processed and used as initial training data. Figure 3 also shows the ratio of documents deemed relevant to all documents presented to the coder, in comparison to the manual coding displayed here as an average. During the manual extraction of events, on average 89% of the documents are discarded (the lower, continuous line). Due to the low amount of training data, the algorithm achieved only 17% at first but improved significantly over time, averaging 43% overall.

Similar to the document classifier, event extraction began in session zero at minimal capacity. Random articles were reviewed and coded into events, if appropriate. On average, only three events were coded per hour, similar to the manual method. Already in the next session, where document classification preselected news articles and the taggers highlighted actors, locations, dates and casualties, the number increased to 8.4 events per hour. The coders averaged 9.4 events per hour. Accounting for duplicates, 5.2 unique events were coded per hour. This is an increase of 50% compared to the 3.67 unique events per hour that a trained and experienced

coder codes manually. Pictured below is the performance of the human-machine tandem, with the learning curve clearly visible.

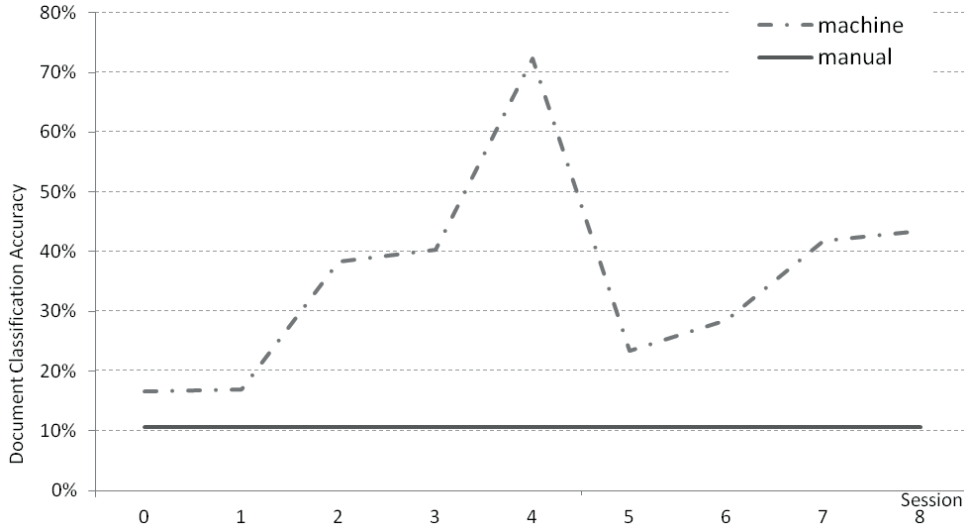


Figure 3: Document classification accuracy over time, beginning at session one. On average, 43% were deemed relevant by human coders in comparison to the approx. 10% baseline.

In summary, the overall throughput has greatly increased. The gross increase is 156%. However, this does not necessarily indicate that the system achieves similar quality to the manual approach. We will analyze and compare the data from a temporal, a spatial and a joint spatiotemporal perspective to determine whether there are similar trends and distributions.

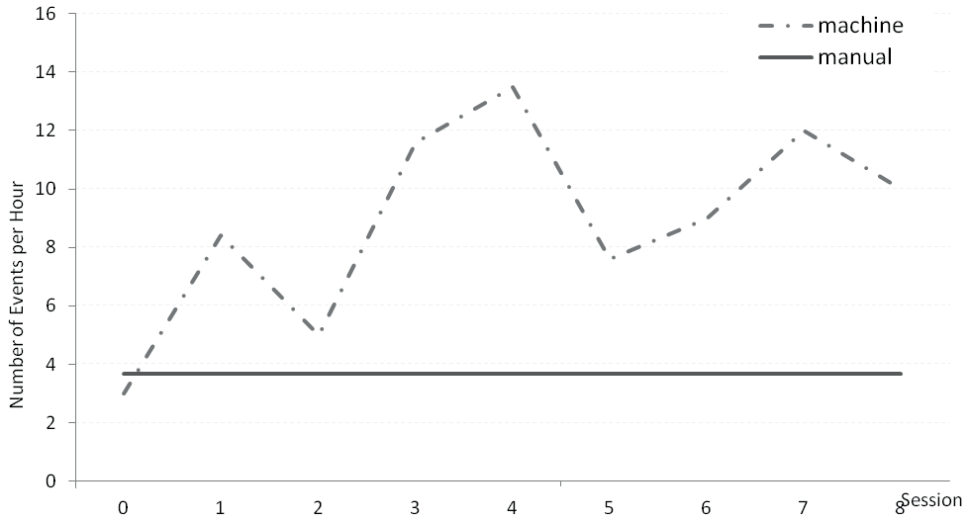


Figure 4: The number of machine-assisted extracted events per hour and session is shown in comparison the baseline, the average number of manually coded events per hour.

3.4 Data Quality Evaluation

Spatiotemporal precision is the key aspect for any quantitative conflict analysis based on georeferenced conflict event data. There is no gold standard of conflict event data to refer to, so we evaluate the spatiotemporal precision of the machine-learning dataset, and thereby its data quality, in relative terms by comparing it to the manually generated EDACS dataset. Below we measure the similarity of these two datasets temporally, spatially, and spatiotemporally.

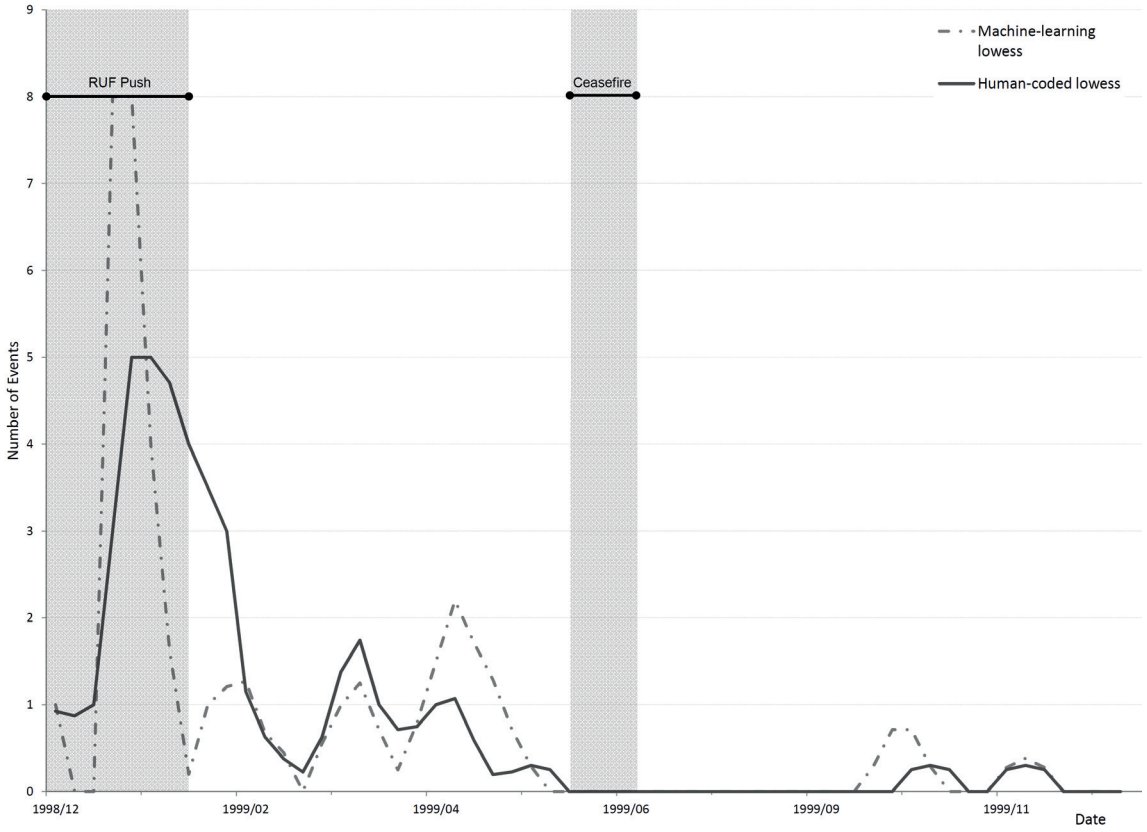


Figure 5[10]: Weekly Time Series of Events by Machine-Learning and Human-Coded EDACS Data for Sierra Leone 1999.

3.4.1 Temporal Comparison

The coding of the exact date of a violent event is a challenging task. In 53% of human-coded events, no exact date is provided by the source itself, and only approximate information is available (e.g., "two weeks ago," "over the last few days," etc.). In addition to imprecise temporal information regarding the circumstance of events, a substantial number of events are reported as aggregates (e.g., "over the course of the last two weeks"), and some sources give no temporal

10 The values are smoothed via a locally weighted polynomial regression (10% -window). The lowess function (Cleveland 1981) in the R-package {stats} has been applied for that.

information at all. For comparative reasons, we exclude the aggregated[11] events or events without any clear indicated date.

To measure temporal (dis-)similarity quantitatively, the violent events are aggregated into weekly sums and charted in a time-series graph for descriptive analysis. In a subsequent analytical step, we calculate the Granger causality and the cross-correlation between the two time series.

The overall frequency – the number of events detected per time window – seems mostly in unison, except for a substantive peak in January and a decrease in May 1999 (see Figure 5, p. 17). Both changes must be understood in the context of the historical events of the Sierra Leone Civil War. The escalation in violence at the turn of the year was rooted in a push by the rebels to retake Freetown, and in January 1999 they overran most of the city. The drop to zero events per week in May can be attributed to the ceasefire agreement between the forces of President Kabbah and the Revolutionary United Front that took effect on May 24 and finally led to the Lomé Peace Accord (United Nations 1999, 2000).

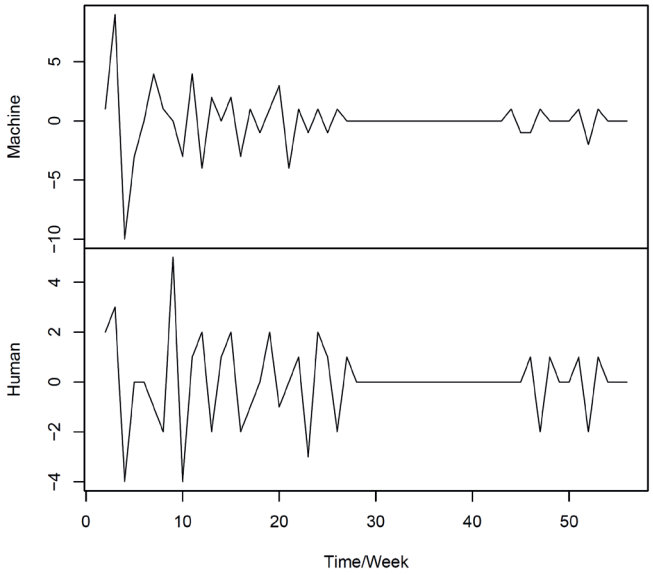


Figure 6: Detrended Time Series of Events by Machine-Learning and Human-Coded EDACS Data for Sierra Leone 1999.

In order to control for trending in the data, we detrend the data and perform seasonal adjustments. The detrended data, shown in Figure 6, points to the fact that both datasets capture the main conflict developments, but appear to have slightly varying characteristics. On the basis of the detrended data, we calculate the Cross-Correlation Function estimation (ccf12). The ccf time-series analysis shows positive, significant values. Especially the zero lag correlates signifi-

11 Events lasting more than thirty days are supposed to be aggregated, but this procedure does not guarantee eliminating a bias caused by event aggregates, so much as serve as an arbitrary threshold to minimize possible biases.

12 The cross-correlation function estimates the degree to which two univariate time series correlate. For calculations we use the ccf-function (Venables/Ripley 2002) in the R package {stats}.

cantly (see Figure 7), allowing us to draw the conclusion that there is a strong temporal resemblance between the two datasets.

The results of a Granger causality test of the two weekly aggregated, detrended, and seasonality adjusted datasets show a highly correlated reciprocal effect (human-coded > machine learning [0.0582*]; machine learning > human-coded [0.0101**]). The Granger causality suggests the finding that an increasing number of coded events at time t(-1), in both datasets, positively affect the number of events at time t(0) (Granger 1969). This also supports the view that both datasets comprise a similar temporal trend.

3.4.2 Spatial Distribution

The spatial distribution of machine-learning and human-coded data is utilized to evaluate their similarity purely within the spatial dimension. The overall spatial distribution (see Figure 8, p. 20) underlines the first impression gained by the temporal comparison and also seems to resemble the general course of conflict events outlined above. The events in both datasets are concentrated in the western part of Sierra Leone, but at least two distinct locations in each of the datasets deviate from this pattern. Near Magburka, human-coded events are present but machine-learning events are missing; in Kenema it is the other way around.

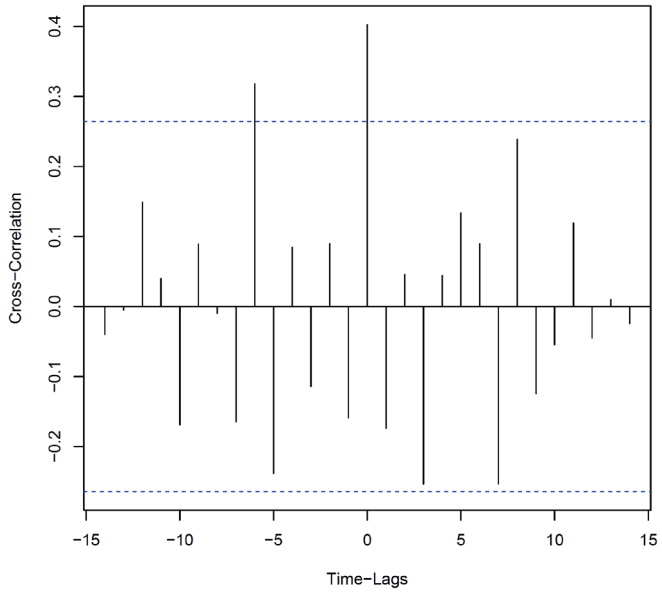


Figure 7: Cross-Correlation (CCF) of Machine-Learning and Human-Coded Event Data for Sierra Leone 1999.

In addition to the cartographic mapping, we also run Ripley's K clustering for spatial processes. The Ripley estimator summarizes spatial dependence (clustering or dispersion) over a range of distances, and displays changes in the spatial dependence with regard to neighborhood size. Therefore, the average number of neighboring events throughout the study area, evaluated with regard to their specific distance to one another, is compared to each event's neighborhood and

either considered clustered or dispersed (ESRI 2011). We simulate outer boundary values to correct for boundary effects, which can lead to an underestimation due to the number of neighbors for features near the edges of the study area of Sierra Leone (the simulated points are the duplicated points near the edges), and calculate ninety-nine permutations for the confidence envelopes (ESRI 2011).

Differences begin to emerge between the spatial clustering characteristics of machine-learning and human-coded data. The machine-learning data clusters gradually, decreasing with the increase in distance, whereas the human-coded event data clusters more locally, declines, and then finally levels after about 37 km distance (cf. Bivand/Gebhardt 2000; Ripley 1976; Rowlingson/Diggle 1993). This suggests that the machine-learning data is less clustered – especially on the local level – than the human-coded event data. Aside from a higher degree of spatial dispersion of the machine-learning events, the reason for this might be their smaller number.

Figure 8: Spatial Distribution of Machine-Learning and Human-Coded Event Data for Sierra Leone 1999.

3.4.3 Spatiotemporal Distance

Spatial and temporal analyses of the conflict event data gathered can only provide a partial picture of the actual data resemblance. Therefore, we also use three different spatiotemporal analysis approaches to evaluate the overlap between machine-learning and human-coded data: firstly, through a comparison of the spatiotemporal K-function; secondly by spatiotemporal permutation scan statistics; and thirdly via SQL-based spatiotemporal similarity matching queries.

3.4.4 Spatiotemporal K-function

We start with the space-time K-function (stK). The space-time K-function estimates the extent of space-time clustering as a function of spatial and temporal separation, based on second-order properties of a generally stationary, homogeneous, spatial-temporal Poisson point process. The space-time K-function is closely related to the Knox statistic and tests the null hypothesis of no spatial and temporal interaction. The basis for the test is the theoretical intensity of the expected number of events per spatial location and time unit, and the observed number of points within a space-time cylinder centered on the event (for further information, see Cressie/Wikle 2011: 210; Diggle et al. 1995: 125ff.; Gabriel/Diggle 2009: 45).

The space-time interaction of the two datasets shows a high degree of similarity (see Figure 9), whereas for the purely spatial cluster analysis (see above: 3.4.2 Spatial Distribution), the discrepancy between machine-learning and human-coded data is partially larger (Bailey/Gatrell 1995; Diggle et al. 1995; Rowlingson/Diggle 1993).

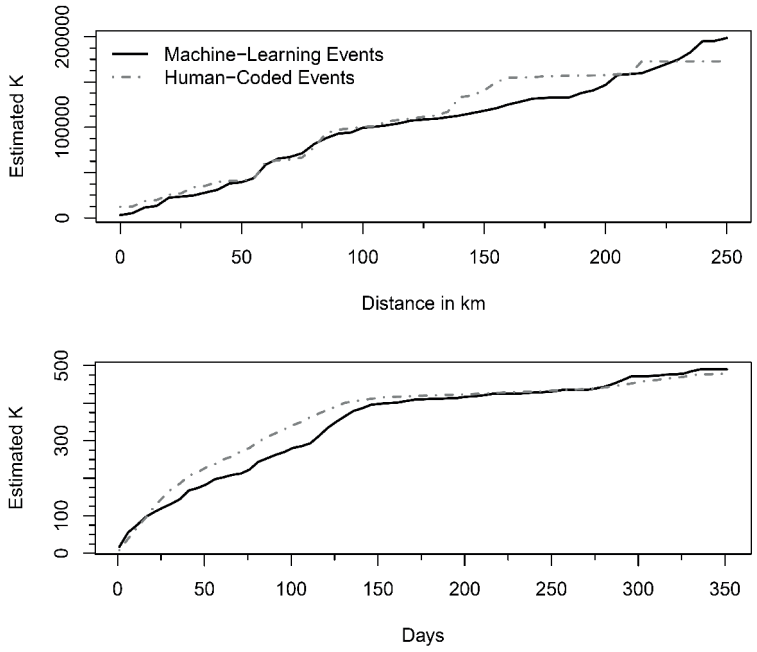


Figure 9: Spatial K-function (top) and Temporal K-function (bottom) Graph of Machine-Learning and Human-Coded EDACS Data for Sierra Leone 1999.

3.4.5 Spatiotemporal Permutation Scan Statistics

The K-function only provides a global measure of spatiotemporal similarity. This is why we run a further local spatiotemporal cluster analysis in order to identify similar clustering in the two datasets. These matching spatiotemporal clusters are statistically significant data-specific hotspots of violence, which again indicate similar trends – irrespective of the data source and data-gathering technique used. The spatiotemporal permutation scan statistics provides values of

local clustering[13] of violent events. The test statistic – and the determination of the cluster – is performed with the software SaTScan™. SaTScan creates a grid of centroids for the region and an infinite number of cylinders around each event location. The circular or ellipsoid radius of the cylinder reflects the portion of the events covered by the cluster; by default this does not exceed 50% of the total number of events. The height of the cylinder reflects time.

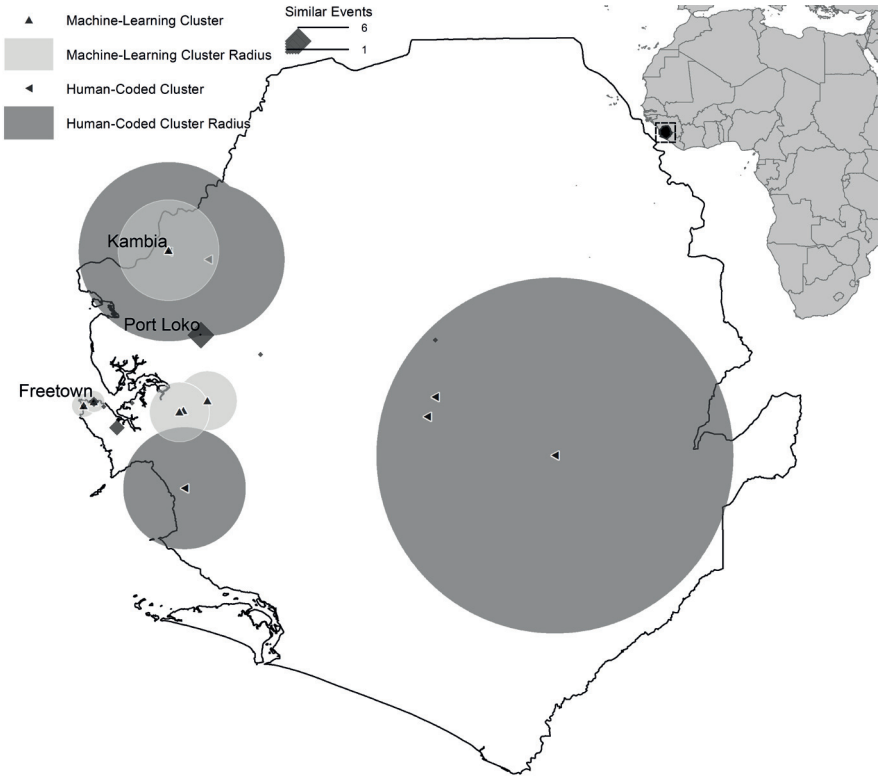


Figure 10: SQL-Query-Based Similarity Matching and Spatiotemporal Permutation SaTScan Statistics of Machine-Learning and Human-Coded Event Data for Sierra Leone 1999.

SaTScan calculates the likelihood function, obtaining the actual and expected number of events, considering all events within the cylinders, and testing for the Complete Spatiotemporal Randomness (CSTR).[14] During this process the "most likely clusters" are identified and 999 Monte Carlo simulations are run, ranking and comparing the most likely clusters with randomly generated data via a likelihood-ratio test. The Monte Carlo permutation procedure generates simulated datasets and envelopes of 95% confidence interval for assessing the significance of

13 A comparative discussion of different space-time clustering methods is provided by Norström et al. (2000). They make a case for Kulldorffs scan statistics because other test options, for instance the Knox test, needs space and time thresholds; furthermore, the Knox test and the Jacquez-k-nearest-neighbours-test assume that population size does not change over time, whereas Kulldorffs scan statistics can accommodate confounding covariates like population size (Norström et al. 2000).

14 Complete Spatiotemporal Randomness (CSTR) implies that there is no stochastic process present in space or time (see for example Cox/Isham 1980; Diggle 2003).

the spatiotemporal permutation statistics. In a final step, p-values for each cluster are calculated (Kulldorff et al. 2005; Kulldorff/Information Management Services Inc. 2009).[15]

The result of the spatiotemporal permutation SaTScan statistics is sobering. Only one significant cluster is detectable within both datasets (see Figure 10, p. 22). One reason for this result could be the small sample size. The available number of coded violent events is problematic with regard to the reliability of the presented test statistics. As a rule of thumb, at least thirty events should be included in the calculations (here: fifty-nine and forty-three events, respectively). This raises the question of the results' robustness, because the statistical results given here can only serve descriptive purposes and thereby only reveal approximate tendencies within the event data.

3.4.6 SQL-Based Spatiotemporal Similarity Matching

We conclude the spatiotemporal comparison by plotting matching results for illustrative reasons on the map of Sierra Leone. This complements the cluster analysis, which did not allow for robust statements on the actual local comparativeness of the datasets. We match the datasets with the help of an SQL query based on defined spatial and temporal thresholds, in order to cover not only statistically significant, but *all* similarities, independent of sample size. We then run the SQL query successively with an increment of one day and five kilometers, which leads to the optimal Euclidian distance threshold of twenty km and a time window of two days.

This narrow threshold gives us twenty-three events where both datasets agree, out of the fifty-nine manually coded and forty-three ML events. The results show similarities but also substantial differences between conflict event data produced by the two presented approaches. The location and extent of the computed spatiotemporal SaTScan clusters is also mapped in Figure 10 (p. 22). As we can see, there is little resemblance between the two datasets captured by the cluster analysis, although the total number of SQL-based matches gives a very different picture. Notice that in Figure 10 there are neither machine-learning nor human-coded space-time clusters around Port Loko, although the map shows a visual cluster derived from the SQL-query-based matching events. The SQL-based similarity matching shows that 53.49% of the ML-coded events can also be found in the human-coded set and 38.98% of all human-coded events are also in the ML data. This again emphasizes the result of almost all temporal and spatial evaluations measured above: that the conflict data events generated manually and via ML are very much alike.

First, the trends generated from the weekly aggregated, detrended, and seasonally adjusted datasets seem to be similar. Second, the results of a cross-correlation analysis and the computed Granger causality test are in line with this view. The purely spatial comparison suggests that

15 We search for high rates of event clustering and set the parameters for the cluster analysis to 35% of the population at risk, with a temporal window of 15% of the study period. We further use a circular spatial window shape, aggregate temporally by seven days, and set the temporal cluster size to one day. Finally, we run 999 Monte Carlo replications.

both datasets show related violent events, but with a few exceptions. Statistically valid answers to this guesswork delivered by global and local spatiotemporal cluster analyses corroborate this view. To meet robustness concerns due to the small sample size, the complimentary SQL-based spatial-temporal comparison broadens the basis of the assessment and leads to the final conclusion that machine learning and human coding produce – with restrictions – similar events.

4. Discussion and Conclusion

The creation of spatiotemporal disaggregated conflict event data opens up possibilities to unpack the black box of war, and to gain a more detailed view on conflict dynamics, actor constellations, and the processual nature of armed conflicts. This reflects the growing importance of event data analysis in peace and conflict research that relies on precise and reliable data. Likewise, the number of news sources and the speed of the information flow via modern ICT – even in remote areas and areas isolated by war – are rapidly increasing. We propose facing this challenge by implementing a semi-automated machine-learning event extraction approach.

The main goal for machine learning is to increase the throughput of event extraction. An important accompanying effect is a possible increase in openness and flexibility. We implemented an infrastructure that is based on open standards and stores all sources and their complete annotations which, copyright restrictions disregarded, allows for complete source transparency and makes ad hoc recoding possible at the same time. The entire method uses free, available libraries that enable an adaptive approach to information extraction that can also be applied to topics beyond conflict data generation.

During our evaluation, it turned out that the costs of implementation are marginal in comparison to the huge amount of time and money necessary to manually create a high-quality conflict database such as EDACS. The increase in flexibility and throughput clearly outweighs the costs of human coding. The ML-based approach increased the number of events generated per hour by 50% when accounting for duplicates. The gross increase was 156%. When extrapolated to the entire first coding rounds of EDACS, even a 50% increase would have saved about one thousand hours of manual coding. An enhanced ML approach that automatically detects duplicates would have saved more than two thousand hours, about one year's worth of manual work and the accompanying financial resources.

A necessary condition for the applicability of this method in the context of research is achieving a high degree of reliability. We evaluated this key element of data quality by performing in-depth robustness checks using manually generated data, and assessed the spatiotemporal comparability of the machine-learning dataset in contrast to the EDACS dataset. These evaluations showed a high degree of similarity between the two. The conducted temporal, spatial, and spatiotemporal comparisons indicated that the machine-learning dataset mirrors – to a large extent – the human-coded event data.

The discrepancies that became apparent in the analysis are marginal with respect to the spatiotemporal precision and revealed conflict trends. Still, the comparability of other variables is yet unknown. We artificially restricted the comparison to key variables; information such as the aggregation of events, "fuzzy" event location, or bias by the reporting agencies are all documented in the EDACS dataset but were not part of the evaluated ML data.

We are confident that creating a holistic ML approach for event extraction that incorporates these facts is a feasible and necessary step to achieve a higher degree of event data quality. But it would seem ill advised to neglect the human aspect of machine-assisted coding procedures. The minimal training our coders received was sufficient for the narrow scope of this experiment. To achieve even better data quality, good training is necessary. Ideally, well-trained coders should be teamed up with a well-adapted ML system that supports coders along each step with proposals for geocoding and ad hoc geo-information, context information such as conflict timelines, and real-time duplicate detection to generate the best conflict event data possible.

These advances in computer sciences enable us to push the envelope of what is possible. They can and should be improved to include more languages and be applied to other sources as well, as this is the only possibility to quantitatively prove the reliability of current conflict event data. Possible approaches include mining crowdsourced (e.g. crisismappers.net) or crowd-seeded data (e.g. Voix de Kivu).[16] Both approaches will have to answer the data quality challenges outlined in this paper: Crowdsourcing's participatory approach will have to tackle reliability issues, since oversight is not inherent to the approach and robustness of reports must be established. Crowd-seeding faces high initial costs, since intensive planning and distribution are necessary to ensure that a representative sample of sources has been selected.

Furthermore, storing semi-structured information in the form tags within the original articles, as proposed in this paper, would allow project researchers and – assuming copyright permission – data users and analysts to apply custom coding rules to the source data and generate custom datasets tailored to their research design. Using such an approach can make most simple changes to the coding rules instantaneous. To facilitate this, a common framework for coding, i.e., the quantification of semi-structured text-based data, would be a great step forward and would further the integration and comparison of data relevant to many different fields of social science. While the next generation of conflict event data projects should draw upon the lessons learned in this evaluation, current event data projects, too, should consider adopting state-of-the-art techniques as proposed in this paper to improve their existing quality.

16 Crowdsourcing and crowdseeding are both participative data gathering techniques. Examples of these are: crowdsourcing at the International Network of Crisis Mappers (crisismappers.net, last accessed 24 September 2012), crowdseeding at Voix des Kivus. The latter is conducted by staff of Columbia University in Eastern Congo: http://cu-csds.org/wp-content/uploads/2009/10/Voix-des-Kivus-Leaflet.pdf, last accessed 24 September 2012.

Literature

Ahn, David 2006: The stages of event extraction. Paper presented at the Proceedings of the Workshop on Annotating and Reasoning about Time and Events, Sydney.

Bailey, Trevor C./Gatrell, Anthony C. 1995: Interactive spatial data analysis (1. ed.), Harlow.

Banko, Michele/Etzioni, Oren/2008: The Tradeoffs Between Open and Traditional Relation Extraction. Paper presented at the Proceedings of ACL-08: HLT, Columbus, OH.

Batini, Carlo/Scannapieca, Monica 2006: Data quality: Concepts, methodologies and techniques (Data-Centric Systems and Applications), Secaucus, NJ.

Bivand, Roger/Gebhardt, Albrecht 2000: Implementing functions for spatial statistical analysis using the R language, in: Journal of Geographical Systems, 2: 3, 307-317.

Buhaug, Halvard 2010: Dude, Where's My Conflict? LSG, Relative Strength, and the Location of Civil War, in: Conflict Management and Peace Science, 27: 2, 107-128.

Carlson, Andrew/Gaffney, Scott/Vasile, Flavian 2009: Learning a Named Entity Tagger from Gazetteers with the Partial Perceptron. Paper presented at the AAAI Spring Symposium on Learning by Reading and Learning to Read, Stanford, CA.

Carpenter, Bob 2010: LingPipe 4.0.1, http://alias-i.com/lingpipe, last accessed 24 September 2012.

Chojnacki, Sven/Ickler, Christian/Spies, Michael/Wiesel, John 2012a: Event Data on Armed Conflict and Security: New Perspectives, Old Challenges, and Some Solutions. International Interactions, 38: 4, 382-401.

Chojnacki, Sven/Ickler, Christian/Schoenes, Katharina/Spies, Michael/Wildemann, Tim 2012b: EDACS Codebook Version 3.5, The Event Data on Armed Conflict and Security (EDACS) Project, Free University Berlin.

Cleveland, William S. 1981: LOWESS: A program for smoothing scatterplots by robust locally weighted regression, in: The American Statistician, 35: 1, 54.

Clough, Paul 2005: Extracting Metadata for Spatially-aware Information Retrieval on the Internet. Paper presented at the GIR, 05 Workshop on Geographic Information Retrieval, Bremen.

Cohen, William W. 2004: Minorthird: Methods for Identifying Names and Ontological Relations in Text Using Heuristics for Inducing Regularities from Data, http://minorthird.sourceforge.net, last accessed 24 September 2012.

Cox, David R./Isham, Valerie 1980: Point processes, London.

Cressie, Noel/Wikle, Christopher K. 2011: Statistics for Spatio-Temporal Data, Hoboken, NJ.

Diggle, Peter J. 2003: Statistical analysis of spatial points patterns (2 ed.), New York, NY.

Diggle, Peter J./Chetwynd, Amanda G./Häggkvist, Roland/Morris, Sarah E. 1995: Second-order analysis of space-time clustering, in: Statistical Methods in Medical Research, 4: 2, 124-136.

Dulic, Tomislav 2010: Geocoding Bosnian Violence: A note on methodological possibilities and constraints in the production and analysis of geocoded event data. Paper presented at the Annual meeting of the Theory vs. Policy? Connecting Scholars and Practitioners, New Orleans, LA.

Earl, Jennifer/Martin, Andrew/McCarthy, John D./Soule, Sarah A. 2004: The Use of Newspaper Data in the Study of Collective Action, in: Annual Review of Sociology, 30, 65-80.

ESRI 2011: ArcGIS Desktop: Release 10. Redlands, CA.

Finkel, Jenny R. 2007: Named Entity Recognition and the Stanford NER Software.

Finkel, Jenny R./Grenager, Trond/Manning, Christopher 2005: Incorporating Non-local Information Into Information Extraction Systems by Gibbs Sampling. Paper presented at the Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005).

Gabriel, Edith/Diggle, Peter J. 2009: Second-order analysis of inhomogeneous spatio-temporal point process data, in: Statistica Neerlandica, 63: 1, 43-51.

Giuliano, Claudio/Lavelli, Alberto/Romano, Lorenza 2007: Relation Extraction and the Influence of Automatic Named-entity Recognition, in: ACM Transactions on Speech and Language Processing, 5: 1, 1-26.

Gleditsch, Nils P./Wallensteen, Peter/Eriksson, Mikael/Sollenberg, Margareta/Strand, Havard 2002: Armed Conflict 1946-2001: A New Dataset, in: Journal of Peace Research, 39: 5, 615-637.

Granger, Clive W. J. 1969: Investigating Causal Relations by Econometric Models and Cross-spectral Methods, in: Econometrica, 37: 3, 424-438.

Harbom, Lotta/Wallensteen, Peter 2009: Armed Conflicts, 1946-2008, in: Journal of Peace Research, 46: 4, 577-587.

Kauffmann, Mayeul 2008: Enhancing Openness and Reliability in Conflict Dataset Creation, in: Kauffmann, Mayeul (ed.): Building and Using Datasets on Armed Conflicts. NATO Science for Peace and Security Series E: Human and Societal Dynamics, Vol. 36, Amsterdam.

Kim, Sang-Bum/Han, Kyoung-Soo/Rim, Hae-Chang/Myaeng, Sung Hyon 2006: Some Effective Techniques for Naive Bayes Text Classification, in: IEEE Transactions on Knowledge and Data Engineering, 18: 11, 1457-1466.

King, Gary/Lowe, Will 2003: An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design, in: International Organization, 57, 617-642.

Kulldorff, Martin/Heffernan, Richard/Hartman, Jessica/Assunção, Renato/Mostashari, Farzad 2005: A space-time permutation scan statistic for the early detection of disease outbreaks, in: Public Library of Science (PLoS) Medicine, 2: 3, 216-224.

Kulldorff, Martin/Information Management Services Inc. 2009: SaTScan™ v9.0: Software for the spatial and space-time scan statistics, http://www.satscan.org/, last accessed 24 September 2012.

Leidner, Jochen L. 2007: Toponym Resolution in Text Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Boca Raton, FL.

Manning, Christopher D./Raghavan, Prabhakar/Schütze, Hinrich 2008: Introduction to Information Retrieval. Cambridge, MA.

Melander, Erik/Sundberg, Ralph 2011: Climate Change, Environmental Stress, and Violent Conflict - Tests introducing the UCDP Georeferenced Event Dataset. Paper presented at the Annual meeting of the International Studies Association, Quebec.

Nardulli, Peter F./Leetaru, Kalev H./Hayes, Matthew J. 2011: Event Data, Civil Unrest and the Social, Political and Economic Event Database (SPEED) Project: Post World War II Trends in Political Protests and Violence. Paper presented at the Annual meeting of the International Studies Association, Quebec.

Norström, Madeleine/Pfeiffer, Dirk U./Jarp, Jorun 2000: A space-time cluster investigation of an outbreak of acute respiratory disease in Norwegian cattle herds, in: Preventive Veterinary Medicine, 47: 1-2, 107-119.

Pasley, Robert C./Clough, Paul D./Sanderson, Mark 2007: Geo-tagging for imprecise regions of different sizes. Paper presented at the Proceedings of the 4th ACM workshop on Geographical information retrieval, Lisbon.

Raleigh, Clionadh/Linke, Andrew/Hegre, Havard/Karlsen, Joakim 2010: Introducing ACLED: An Armed Conflict Location and Event Dataset, in: Journal of Peace Research, 47: 5, 651-660.

Ripley, Brian D. 1976: The second-order analysis of stationary point processes, in: Journal of Applied Probability, 13, 255-266.

Rowlingson, Barry S./Diggle, Peter J. 1993: SPLANCS: Spatial point pattern analysis code in S-plus, in: Computers in Geosciences, 19: 5, 627-655.

Sarawagi, Sunita 2007: Information Extraction, in: Foundations and Trends in Databases, 1: 3, 261-377.

Schrodt, Philip A. 2011: Automated Production of High-Volume, Near-Real-Time Political Event Data. Paper presented at the New Methodologies and Their Applications in Comparative Politics and International Relations, Princeton, NJ.

Thion-Goasdoué, Virginie/Nugier, Sylvaine/Duquennoy, Dominique/Laboisse, Brigitte 2007: An evaluation framework for data quality tools. Paper presented at the International Conference for Information Quality (ICIQ), Cambridge, MA.

United Nations 1999: Secretary-General Welcomes Ceasefire Agreement on Sierra Leone, http://www.un.org/sc/committees/1132/pdf/6998e.html, last accessed 24 September 2012.

United Nations 2000: Sierra Leone – UNOMSIL: Background, http://www.un.org/en/peacekeeping/missions/past/unomsil/UnomsilB.htm, last accessed 24 September 2012.

Venables, William N./Ripley, Brian D. 2002: Modern Applied Statistics with S, New York, NY.

Weidmann, Nils B./Rød, Jan K./Cederman, Lars-Erik 2010: Representing ethnic groups in space: A new dataset, in: Journal of Peace Research, 47: 4, 491-499.

Previously published Working Papers from the SFB-*Governance* Working Paper Series

Livingston, Steven/Walter-Drop, Gregor 2012: Information and Communication Technologies in Areas of Limited Statehood, SFB-Governance Working Paper Series, No. 38, Collaborative Research Center (SFB) 700, Berlin, September 2012.

Schüren, Verena 2012: Two TRIPs to Innovation. Pharmaceutical Innovation Systems in India and Brazil, SFB-Governance Working Paper Series, No. 37, Collaborative Research Center (SFB) 700, Berlin, June 2012.

Sonderforschungsbereich 700: Grundbegriffe der Governanceforschung, SFB-Governance Working Paper Series, No. 36, 2. revised edition, Collaborative Research Center (SFB) 700, Berlin, June 2012.

Eimer, Thomas R. 2012: When Modern Science Meets Traditional Knowledge: A Multi-Level Process of Adaption and Resistance, SFB-Governance Working Paper Series, No. 35, Collaborative Research Center (SFB) 700, Berlin, June 2012.

Kötter, Matthias 2012: Non-State Justice Institutions: A Matter of Fact and a Matter of Legislation, SFB-Governance Working Paper Series, No. 34, Collaborative Research Center (SFB) 700, Berlin, June 2012.

Koehler, Jan 2012: Social Order Within and Beyond the Shadows of Hierarchy. Governance-Patchworks in Afghanistan, SFB-Governance Working Paper Series, No. 33, Collaborative Research Center (SFB) 700, Berlin, June 2012.

Risse, Thomas 2012: Governance Configurations in Areas of Limited Statehood. Actors, Modes, Institutions, and Resources, SFB-Governance Working Paper Series, No. 32, Collaborative Research Center (SFB) 700, Berlin, March 2012.

Hönke, Jana, with Thomas, Esther 2012: Governance for Whom? – Capturing the Inclusiveness and Unintended Effects of Governance, SFB-Governance Working Paper Series, No. 31, Collaborative Research Center (SFB) 700, Berlin, April 2012.

Contreras Saíz, Mónika/Hölck, Lasse/Rinke, Stefan 2012*:* Appropriation and Resistance Mechanisms in (Post-) Colonial Constellations of Actors: The Latin American Frontiers in the 18th and 19th Century, SFB-Governance Working Paper Series, No. 30, Collaborative Research Center (SFB) 700, Berlin, April 2012.

Börzel, Tanja 2012*:* How Much Statehood Does it Take – and What For?, SFB-Governance Working Paper Series, No. 29, Collaborative Research Center (SFB) 700, Berlin, March 2012.

Prigge, Judit 2012: Friedenswächter. Institutionen der Streitbeilegung bei den Amhara in Äthiopien, SFB-Governance Working Paper Series, No. 28, Collaborative Research Center (SFB) 700, Berlin, January 2012.

Jacob, Daniel/Ladwig, Bernd/Oldenbourg, Andreas 2012: Human Rights Obligations of Non-State Actors in Areas of Limited Statehood, SFB-Governance Working Paper Series, No. 27, Collaborative Research Center (SFB) 700, Berlin, January 2012.

Schmelzle, Cord 2011: Evaluating Governance. Effectiveness and Legitimacy in Areas of Limited Statehood, SFB-Governance Working Paper Series, No. 26, Collaborative Research Center (SFB) 700, Berlin, November 2011.

Börzel, Tanja A./Hönke, Jana 2011: From Compliance to Practice. Mining Companies and the Voluntary Principles on Security and Human Rights in the Democratic Republic of Congo, SFB-Governance Working Paper Series, No. 25, Collaborative Research Center (SFB) 700, Berlin, October 2011.

Draude, Anke/Neuweiler, Sonja 2010: Governance in der postkolonialen Kritik. Die Herausforderung lokaler Vielfalt jenseits der westlichen Welt, SFB-Governance Working Paper Series, No. 24, Collaborative Research Center (SFB) 700, Berlin, May 2010.

Börzel, Tanja A. 2010: Governance with/out Government. False Promises or Flawed Premises?, SFB-Governance Working Paper Series, No. 23, Collaborative Research Center (SFB) 700, Berlin, March 2010.

Wilke, Boris 2009: Governance und Gewalt. Eine Untersuchung zur Krise des Regierens in Pakistan am Fall Belutschistan, SFB-Governance Working Paper Series, No. 22, Collaborative Research Center (SFB) 700, Berlin, November 2009.

Schneckener, Ulrich 2009: Spoilers or Governance Actors? Engaging Armed Non-State Groups in Areas of Limited Statehood, SFB-Governance Working Paper Series, No. 21, Collaborative Research Center (SFB) 700, Berlin, October 2009.

Mueller-Debus, Anna Kristin/Thauer, Christian R./Börzel, Tanja A. 2009: Governing HIV/AIDS in South Africa. The Role of Firms, SFB-Governance Working Paper Series, No. 20, Collaborative Research Center (SFB) 700, Berlin, June 2009.

Nagl, Dominik/Stange, Marion 2009: Staatlichkeit und Governance im Zeitalter der europäischen Expansion. Verwaltungsstrukturen und Herrschaftsinstitutionen in den britischen und französischen Kolonialimperien, SFB-Governance Working Paper Series, No. 19, Collaborative Research Center (SFB) 700, Berlin, February 2009.

Börzel, Tanja A./Pamuk, Yasemin/Stahn, Andreas 2008: The European Union and the Promotion of Good Governance in its Near Abroad. One Size Fits All?, SFB-Governance Working Paper Series, No. 18, Collaborative Research Center (SFB) 700, Berlin, December 2008.

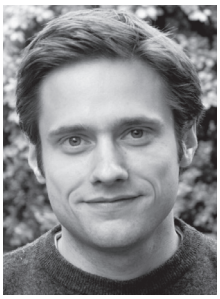
Koehler, Jan 2008: Auf der Suche nach Sicherheit. Die internationale Intervention in Nordost-Afghanistan, SFB-Governance Working Paper Series, No. 17, Collaborative Research Center (SFB) 700, Berlin, November 2008.

Beisheim, Marianne/Fuhr, Harald (ed.) 2008: Governance durch Interaktion nicht-staatlicher und staatlicher Akteure. Entstehungsbedingungen, Effektivität und Legitimität sowie Nachhaltigkeit, SFB-Governance Working Paper Series, No. 16, Collaborative Research Center (SFB) 700, Berlin, August 2008.

Buckley-Zistel, Susanne 2008: Transitional Justice als Weg zu Frieden und Sicherheit. Möglichkeiten und Grenzen, SFB-Governance Working Paper Series, No. 15, Collaborative Research Center (SFB) 700, Berlin, July 2008.

These publications can be downloaded from www.sfb-governance.de/publikationen or ordered in printed versions via e-mail to sfb700@zedat.fu-berlin.de.

The Authors



Christian Ickler is a research fellow at the Collaborative Research Center (SFB) 700 "Governance in Areas of Limited Statehood" in Berlin, Germany. He is a geographer with a special interest in geographical conflict science. His current focus is on spatiotemporal patterns of violence in civil war.

Contact: *c.ickler@fu-berlin.de*

John Wiesel is a research fellow at the Collaborative Research Center (SFB) 700 "Governance in Areas of Limited Statehood" in Berlin, Germany. He is a computer scientist with a special interest in databases, information systems, and machine learning. His research focuses on the combination of modern information technology and conflict research to create georeferenced micro-level conflict data.

Contact: *john.wiesel@fu-berlin.de*

Research Framework

Governance has become a central theme in social science research. The Collaborative Research Center (SFB) 700 *Governance in Areas of Limited Statehood* investigates governance in areas of limited statehood, i.e. developing countries, failing and failed states, as well as, in historical perspective, different types of colonies. How and under what conditions can governance deliver legitimate authority, security, and welfare, and what problems are likely to emerge? Operating since 2006 and financed by the German Research Foundation (DFG), the Research Center involves the Freie Universität Berlin, the University of Potsdam, the European University Institute, the Hertie School of Governance, the German Institute for International and Security Affairs (SWP), and the Social Science Research Center Berlin (WZB).

Partner Organizations

Host University:
Freie Universität Berlin

University of Potsdam

German Institute for International and Security Affairs (SWP)

Social Science Research Center Berlin (WZB)

Hertie School of Governance