Modeling Mortality Rates In The WikiLeaks Afghanistan War Logs

Rusch, Thomas; Hofmarcher, Paul; Hatzinger, Reinhold; Hornik, Kurt

Published: 01/09/2011

Document Version
Publisher's PDF, also known as Version of record

Modeling Mortality Rates In The WikiLeaks Afghanistan War Logs

Thomas Rusch, Paul Hofmarcher, Reinhold Hatzinger, Kurt Hornik

Research Report Series
Report 112, September 2011

Institute for Statistics and Mathematics
http://statmath.wu.ac.at/

Modeling Mortality Rates In The WikiLeaks Afghanistan War Logs

Thomas Rusch*, Paul Hofmarcher, Reinhold Hatzinger, Kurt Hornik

Institute for Statistics and Mathematics
WU Vienna University of Economics and Business
Vienna, Austria

September 19, 2011

Abstract

The WikiLeaks Afghanistan war logs contain more than 76000 reports about events and resulting fatalities in the US led Afghanistan war, covering the period from January 2004 to December 2009. In this paper we use those reports to build statistical models to help us understand the mortality rates associated with specific circumstances. We choose an approach that combines Latent Dirichlet Allocation (LDA) with negative binomial based recursive partitioning. LDA is used to process the natural language information contained in each report summary. We estimate latent topics and assign each report to one of them. These topics, in addition to other variables in the data set, subsequently serve as explanatory variables for modeling the number of fatalities of the civilian population, ISAF Forces, Anti-Coalition Forces and the Afghan National Police or military as well as the combined number of fatalities. Modeling is carried out with manifest mixtures of negative binomial distributions estimated with model-based recursive partitioning. For each group of fatalities, we identify segments with different mortality rates that correspond to a small number of topics and other explanatory variables as well as their interactions. Furthermore, we carve out the similarities between segments and connect them to stories that have been covered in the media. This provides an unprecedented description of the war in Afghanistan covered by the war logs. Additionally, our approach can serve as an example as to how modern statistical methods may lead to extra insight if applied to problems of data journalism.

Keywords: WikiLeaks; Afghanistan; topic models; model-based recursive partitioning; mixture models; negative binomial; fatalities; data journalism; count data

1 Introduction

The analysis of fatalities in wars and armed conflicts is an important subject of scientific investigation. Many of those have been conducted, mostly in a historical context, often retrospectively estimating the number of and circumstances under which fatalities of war occurred. To name a few, Gooch (2010) investigated fatality numbers in The White War, Cirillo (2008) looked at fatalities from disease and combat in America's principal wars from 1775 onwards and Seet and Bunham (2000) studied fatality trends in UN peacekeeping missions from 1948 to 1998. Lerner (2000) discussed the connection between psychiatry and casualties of war in Germany in WWI

*corresponding author: thomas.rusch@wu.ac.at

and Hirschman et al. (1995) as well as Barnett et al. (1992) estimated the number of Vietnamese and US American casualties in the Vietnam war. There are literally hundreds of historical investigations into numerous wars, see e.g. Garfield and Neugut (1991) for a review of the last 200 years. Retrospective investigations even went back into the time before civilized societies emerged (Keeley, 1996).

Notwithstanding such efforts, contemporary systematic scientific investigation into the number of fatalities in wars are much rarer and more closely tied to the emergence of statistics and epidemiology as disciplines rather than to the discipline of history. As one of the first examples we could find, in 1838, Marshall and Balfour presented a "Statistical Report on the Sickness, Mortality, & Invaliding among the troops in the West Indies". Just as noteworthy were the investigations into fatalities in the Crimean war and their causes a couple of years later by Nightingale (1863). While these investigations were still firmly rooted in descriptive statistics, statistical modeling of the number of fatalities was about to become imperative as Bortkiewicz (1898) published his seminal work on the use of the Poisson distribution for rare events which he motivated by the analysis of horse-kick deaths of Prussian soldiers. To our knowledge this was the first instance of a parametric approach to analyze war fatalities. Contemporary investigations into the number and circumstances of casualties of war that made use of statistical modeling next to descriptive approaches increased much since then, for example Spiegel and Salama (2001), Thomas et al. (2001), Lakstein and Blumenfeld (2005) or Holcomb et al. (2007).

In the modern age their number seems to peak[1] arguably because data on war fatalities are much easier to come by. Recent work in this field includes the paper by Haushofer et al. (2010) who used vector-autoregressive OLS models to model the temporal dynamic of the Israeli–Palestinian conflict or Degomme and Guha-Sapir (2010) who investigated patterns of mortality rates in the Darfur conflict with Quasi-Poisson models. Buzzell and Preston (2007) discussed the mortality rates of American troops in the Iraq war between 2003 and 2006 and Burnham et al. (2006) performed a cross-sectional cluster sample survey on the mortality in Iraq after 2003. For the war in Afghanistan, there are the studies on child casualties by Bhutta (2002) and on military fatalities by Bird and Fairweather (2007). Wars with US involvement are particularly well investigated, see e.g. Leland and Oboroceanu (2010) for a comprehensive list of US war fatalities from the American Independence wars to "Operation Enduring Freedom" in Iraq.

In July 2010 the availability of data on a specific war became unprecedented, as whistleblower website WikiLeaks released a massive amount of military classified war logs from the Afghanistan war into the public. These documents constitute a "war diary" of the US led military operation in Afghanistan, containing a detailed description of what happened in each event for which a report was filed, including counts of killed and wounded people, local and administrative information, temporal and spatial information and a short written description of each particular incident. The reports themselves stem from a database of the US army and along the lines of WikiLeaks, they do not generally cover any top-secret operations or European or other ISAF operations. In total, the war logs consist of 76911 reports and cover the time period between January 2004 and December 2009. They provide an unprecedented view of the war in Afghanistan with an information abundance that has previously been unknown and has only been topped by the release of the Iraq war logs some months later.

The disclosure of these documents has started a debate about the legitimacy of publishing such data.[2] The German news magazine *Der Spiegel* wrote that the editors-in-chief of *Der Spiegel*, The New York Times and The Guardian were "unanimous in their belief that there is a justified public interest in the material" (Gebauer, 2010) and the war diary was marked as the 21st century equivalent of the Pentagon Papers from the 1970s. However, while the Pentagon Papers have

[1] According to a quick survey in the ISI Web of Knowledge citation database, searching for "war casualties" found 1476 records, 840 of which were published after 2000. 580 of those were published no later than 2005.

[2] A Congressional Law Service expertise rendered usage of the published data lawful.

provided an aggregated view on the war in Vietnam, the WikiLeaks war diary is an account of the events in Afghanistan, containing thousands of mosaic tiles describing incidents from the perspective of the US forces, day in, day out, written by (thousands) of soldiers, sometimes accurate and often possibly subjective. The war logs themselves neither contain information on strategic decisions nor do they provide a coherent, general picture of the war. Hence, each media outlet had to write its own stories based on the material (see O'Loughlin et al., 2010). This has been praised as data-driven journalism in action (see Rogers, 2010), a type of journalism which allows stories to enfold from data.

The scientific community also approached the data. For example, O'Loughlin et al. (2010) presented an analysis of the spatial dynamics of the conflict in Afghanistan as portrayed in the WikiLeaks data. Political science blogger Drew Conway provided an analysis of the reports filed over time (Conway, 2010b) and a spatial and temporal analysis of deaths much like O'Loughlin et al. (Conway, 2010d). He also engaged in modeling strategies by investigating if Benford's law may be underlying the reported data (Conway, 2010a) and presented a visualisation of word stems for a subset of the report summaries of the war logs over time based on Latent Dirichlet Allocation (Conway, 2010c). While providing many interesting insights, all those analyses - journalistic and scientific alike - have remained mostly on a descriptive level. This may be due to the sheer bulk of the data.

One of the peculiarities of the war log and its main challenge is that the data at hand stem from a database and that the information is captured in both numeric variables as well as written text. To neglect the written text in a statistical evaluation of such data sets would often come along with discarding important if not crucial information. Especially in the WikiLeaks data nearly all detailed information about the events is stored as written text. Thus it is essential for statistical evaluation to incorporate that information.

Modern statistical and data mining procedures provide tools to handle and analyze such data sets appropriately and to allow a deeper investigation. In this paper we will make exemplary use of such statistical learning approaches to analyze the number of fatalities in the war logs in a deeper way and to build statistical models. By combining two promising new ideas, topic models and model-based recursive partitioning, our analysis allows to get a bigger picture of the war given the thousands of mosaic tiles.

The idea of our approach is as follows: Each single entry in the WikiLeaks war logs contains several variables but also a written report summary containing a short description of what happened in this particular incident. We are interested in extracting explanatory information from the reports, some type of meta information that aggregates reports with similar content. Assuming that this similarity is reflected in the words contained in the summaries, we make use of Latent Dirichlet Allocations (LDA; Blei et al., 2003) to cluster written report summaries together into latent topics. In a second step, we then use the generated topic assignments as further explanatory variables in modeling the number of fatalities in this data set. Since there is a high degree of overdispersion present, we chose to model the number of fatalities with the negative binomial distribution. To allow for a flexible, non-linear functional relationship between explanatory variables and the fatality numbers, which also focuses on interactions, we chose a recursive partitioning approach (Zeileis et al., 2008). Since the model in each segment is a distribution, we call this a manifest (i.e. based on explanatory variables) mixture model.

The remainder of this paper is organized as follows: Section 2 contains a description of the WikiLeaks war logs. The methodological Section 3 presents the methods used in the present effort. The results for all groups of considered fatalities are described in Section 4, while Section 5 provides an overarching discussion of the obtained results. We finish with conclusions in Section 6.

	Allied	Host	Civilian	ACF	Total
killed	1146	3796	3994	15219	24155
wounded	7296	8503	9044	1824	26667

Table 1: The number of casualties by group.

2 The WikiLeaks Afghanistan War Logs

The release of 76911 individual war logs by `WikiLeaks.org` represents a milestone in the possibility to take a look at an ongoing war. The war logs cover the period from January 2004 to December 2009 and each event for which a report has been filed corresponds to a single document. Figure 1 displays the number of filed reports per month. While for the first years of the military operation we can find only a few hundreds of reports per month, this number increases up to more than 3500 in mid 2009.

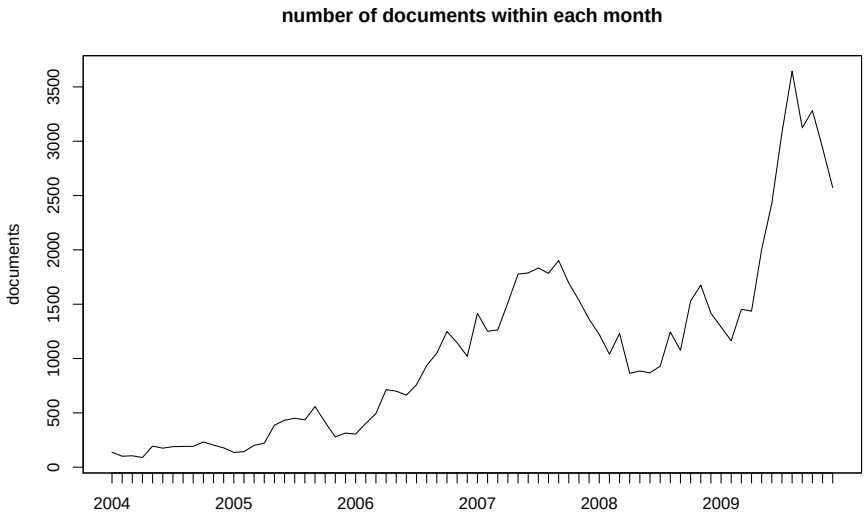

Figure 1: Monthly quantity of filed reports.

The report documents contain 32 columns with numerical and factor variables such as ID number, reporting unit, date or geographic location in latitude/longitude, number of fatalities for different groups and so on. As our dependent variables we use the four columns listing the number of *Civilian*, *Enemy*, *Friend* and *Host* fatalities within each report, as well as the sum of all fatalities. Troops fighting against coalition troops are referred to as "Enemies". We adopt the term "Anti-Coalition Fighters" (ACF) to describe this variable. The "Friends" column refers to ISAF forces including the NATO countries and the US military, while "Host" stands for local (Afghan) military and police. We subsume the former under "coalition troops" or "allied forces" and the latter under "Afghan or host forces".

Table 1 provides summary statistics for the casualties and Figure 2 displays a plot of the number of fatalities over time for each group during the observation period. In total we find 24155 fatalities in the war logs. 63% of the fatalities have been labeled as ACF. The second highest fatality number (16.54%) has been observed for civilians, closely followed by 15.72% Afghan soldiers and policemen and 1146 or 4.74% killed allied soldiers. Palpably are the two peaks for killed insurgents in late summer 2006 and 2007 in Figure 2. They account for 943 killed ACF

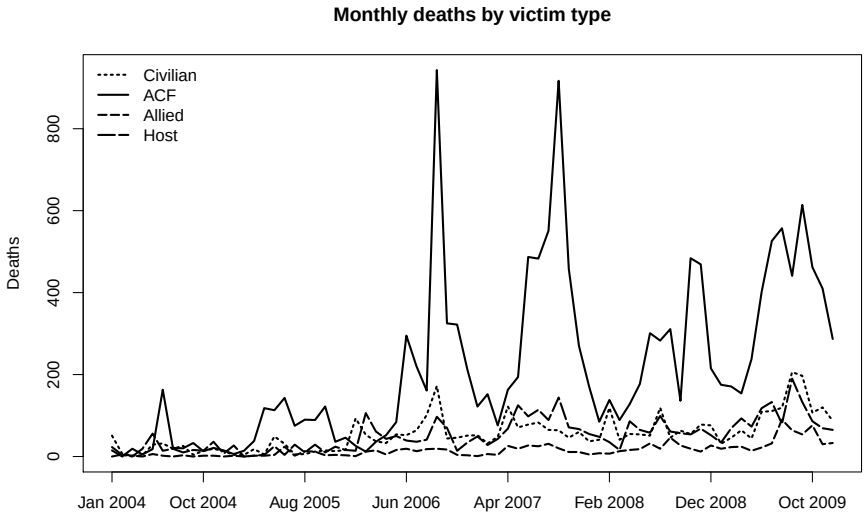

Figure 2: Frequency of fatalities by group per month

.

fighters during September 2006 and for 917 in September 2007. The former peak corresponds to "Operation Medusa", an operation that had the aim to establish government control over areas of Kandahar province. The latter marks operations near Kandahar in an effort to remove insurgents who have returned to this area. Mid to late 2009 is the bloodiest period for civilians, coalition soldiers and ACF. Between May 2009 and December 2009 we observe 1056 (26.4%) out of 3994 (see Table 1) civilian fatalities. In August 2009, during the period of the presidential elections (August 20) we observe 206 civilian victims and 190 killed ACF[3]. For both groups, this has been the highest death toll within one month. Roughly the same pattern can be observed for allied soldiers. Here the monthly maximum 90 has happened in July 2009 and from May 2009 to December 2009 the data account for 346 (30.2%) killed allied soldiers[4].

Additionally, the report documents contain 28 columns with numerical and factor variables that serve as possible explanatory variables. We restrict ourselves to describe only those explanatory variables that were of special relevance for our analysis.

The factor `attackOn`, with its levels `FRIEND, NEUTRAL, ENEMY, UNKNOWN` encodes the US militaries point of view on whom an "attack" (action) has been directed during the incident. O'Loughlin et al. (2010, p. 474 ff) state that this variable seems to have been mislabeled and should have been named "attackBy". However, after inspection of the war logs we believe that `attackOn` does not contain information about who carried out a certain action but rather contains information about on whom the action described in the report has been directed. For instance, leaflets of Anti-Coalition Forces (ACF) calling for attacks against the US forces have been categorized as `attackOn=NEUTRAL`, fire fights between ACF and allied soldiers as `attackOn=ENEMY` and friendly fire has been labeled as `attackOn=FRIEND`.

The categorical variable `Dcolor` controls the display color of the message in the messaging system and map views. Messages relating to enemy activity have the color `red`, those relating to

[3]The UN Assistance Mission in Afghanistan (UNAMA) and Afghanistan Independent Human Rights Commission (AIHRC) stated that on election day Afghanistan had suffered the highest number of attacks and intimidation the country had seen in some 15 years (see Wikipedia, 2011a).

[4]For wounded people the pattern differs. Here we observe the lowest fraction for ACF with 6.84% and the highest for civilians (33.91%). Hence every third wounded person within the war logs has been a civilian.

friendly activity have been colored `blue` and `green` stands for neutral. This variable can be seen as the one coding by whom an action has been carried out ("attackBy").

Another important variable for our analysis is `region`, roughly describing where an event took place. It has levels `RC NORTH, RC EAST, RC WEST, RC SOUTH, RC CAPITOL, UNKNOWN` and `NONE SELECTED` (RC stands for "Regional Code"). It is not clear what the difference between the levels `UNKNOWN` and `NONE SELECTED` is, we have therefore treated them as qualitatively different.

Next there is `complexAttack`, a binary variable that encodes the complexity of an attack. The US military states an attack as complex if it has been well organized and executed, if soldiers have made use of heavy artillery and the troops have been able to withdraw from the battlefield in an organized fashion (see Roggio, 2009).

2.1 The Report Summaries

The variables described above that serve as explanatory variables for modeling the number of fatalities, only allow for a rather limited view into the events of each report and therefore the circumstances under which fatalities have happened. We can however find additional information about the context of the various incidents in the provided report summaries. These summaries contain a short verbal description of what has happened during the incident. Often, these summaries are full of military acronyms and hard to understand for readers not familiar with this jargon. A self compiled list of meanings of the acronyms can be found in Appendix B[5]. To give an example for such a report, on 11-Feb-2004 we can find a report describing an ambush on a convoy that did not result in fatalities:

> At 110740ZFEB04 CJSOTF-A reported an oda convoy was ambushed ivo geresk; The ambush was initiated by an ied which detonated behind the first vehicle in the convoy (convoy consisted of three vehicles), the convoy was hit simultaneously with small arms fire. The convoy returned fire and moved out of the kill zone. CJSOTF-A reported there were no fatalities and no damage to equipment.

The report summaries tell us the how and why of the mission in a very detailed way, something the provided situational variables cannot. Thus the report summaries and their content are at the core of evaluating the ongoings of this war as portyed in the war logs as well as gaining insight into mortality in different situations. Disregarding these summaries in evaluating the war logs would be equivalent to discarding the most important information.

However, making use of this information is challenging. First as we mentioned, the summaries are plain natural language text filled with military acronyms which we need to process. Second, the sheer bulk of reports makes processing of the summaries by humans (who are most apt to process natural language) rather difficult. A person would have to read or process more than 79600 texts. If each summary takes a minute to read and file or process in any way, it would amount to approx. 1282 hours of work (or 160 work days if a work day consists of 8 hours).

There are three possible strategies to deal with that: Either the reports are processed by crowdsourcing them to a high number of people. Or, if there is an *a priori* defined category system, one may classify the reports into these categories with a supervised approach. But neither did we have such a category system nor did we want to crowdsource it. We needed an approach to get some kind of meta information that aggregates reports with similar content and at the same time generates the category system. The resulting meta information could then be used as explanatory variables. This led us to Latent Dirichlet Allocation (LDA; Blei et al., 2003)) or "topic models".

[5] see `http://www.armysignalocs.com/docs/War\%20Log\%20Glossary.pdf`

3 Method

3.1 Using Topic Models To Build Explanatory Variables From Report Summaries

Latent Dirichlet Allocation (LDA; Blei et al., 2003) is a powerful document generative probabilistic model for clustering words into topics and documents into mixtures of topics. A detailed description of LDA can be found in Blei et al. (2003) or Blei and Lafferty (2009). Assuming that the similarity between reports is reflected in the words contained in the summaries, we can use LDA to assign reports based on their summaries to a number of topics. This allocation of each report to (one or more) latent topics can be seen as a task of complexity reduction or as a preprocessing step.

According to Blei and Lafferty (2009), topics are automatically discovered from the original texts and we do not require any *a priori* information about the existence of a certain theme. We just need to fix the number of topics within the whole set of documents (corpus). The resulting topics are shared across the whole set of documents. For example, for the WikiLeaks report summaries LDA might find a topic that can be called "medical" and one that can be called "military convoys"[6]. The "military convoy" topic will have a higher probability for words like vehicle or highway. The "medical" topic likewise will have high probabilities for words like e.g. patient or wounded. Both topics might have a fairly similar probability to contain the word "crash". Vice versa, a report summary often containing the word "vehicle" would then have a higher probability to belong to topic "military convoy", whereas a summary listing "hospital", "patient" and "operation" will have a high posterior probability to belong to the "medical" topic. Please note that in general the topic distribution of each report does only include non-zero probabilities.

Boyd-Graber et al. presented results on measuring the interpretability of a topic model compared to human classification. They concluded that "humans are able to appreciate the semantic coherence of topics and can associate the same documents with a topic that topic model does" (Boyd-Graber et al., 2009, p. 8). Griffiths and Steyvers (2004, p. 5228) note that "the extracted topics capture meaningful structure in the data, consistent with the class designations provided by the authors". This makes LDA well suited for our purpose.

3.1.1 The Document Generative LDA Model

Following Blei and Lafferty (2009) and Blei (2011), LDA specifies the data-generating process as a probabilistic model, in which each document is a mixture of a set of topics and each word in the document is chosen from the selected topic specific word distribution.

More formally, let q denote the size of a vocabulary (unique words within the considered corpus of documents) and let s be the number of topics $\boldsymbol{\beta}_t, t = 1, \ldots, s$. Each topic $\boldsymbol{\beta}_t$ is a q-dimensional symmetric Dirichlet distribution over the vocabulary with scalar parameter η. The only observed variables are words $\boldsymbol{w}_{1:h}$, where h denotes the number of documents and $w_{d,m} \in \{1, \ldots, q\}$ denotes the $m-th$ word of document d. The documents $d,\ d = 1, \ldots, h$ are sequences of those words of varying lengths q_d. Each document d is assigned to a topic with the assignment being denoted by z_d and and the topic assignment of each of its words $w_{d,m}$ is denoted by $z_{d,m}$. Each document is seen as a mixture of topics and hence each document has a vector of topic proportion denoted by $\boldsymbol{\pi}_d$ with $\pi_{d,t}$ denoting the proportion of topic t in document d. The distribution of $\boldsymbol{\pi}_d$ is a s-dimensional symmetric Dirichlet distribution with scalar parameter κ. Hence the generative model for LDA is

$$P(\boldsymbol{W}_{1:h}, \boldsymbol{\beta}_{1:s}, \boldsymbol{\pi}_{1:h}, \boldsymbol{Z}_{1:h}|\eta, \kappa) = \prod_{t=1}^{s} P(\boldsymbol{\beta}_i|\eta) \prod_{d=1}^{h} P(\boldsymbol{\pi}_d|\kappa) \left(\prod_{m=1}^{q_d} P(z_{d,m}|\boldsymbol{\pi}_d) P(W_{d,m}|\boldsymbol{\beta}_{1:s}, z_{d,m}) \right),$$

[6] Note that naming is somewhat arbitrary because it can be difficult to assign an exclusive name to a topic.

where the conditional distributions of the topic assignments and the words are assumed to be multinomial, i.e. $P(Z_{d,m}|\boldsymbol{\pi}_d) \sim \text{Multinomial}(\boldsymbol{\pi}_d)$ and $P(W_{d,m}|\boldsymbol{\beta}_{1:s}, z_{d,m}) \sim \text{Multinomial}(\boldsymbol{\beta}_{z_{d,m}})$. Inference for LDA can be done by the variational EM-Algorithm (see e.g. Grün and Hornik, 2011) or the model can be estimated using a Bayesian approach Blei et al. (2003).

Since we need LDA to generate topics and assign each document to one of them, we are interested in the posterior distribution of the latent topics, the topic assignment and the topic proportions given the documents,

$$P(\boldsymbol{\beta}_{1:s}, \boldsymbol{\pi}_{1:h}, \boldsymbol{Z}_{1:h}|\boldsymbol{w}_{1:h}, \eta, \boldsymbol{\kappa}) = \frac{P(\boldsymbol{W}_{1:h}, \boldsymbol{\beta}_{1:s}, \boldsymbol{\pi}_{1:h}, \boldsymbol{Z}_{1:h})}{P(\boldsymbol{W}_{1:h})},$$

and the conditional expectations $\hat{\boldsymbol{\beta}}_{t,u} = E(\boldsymbol{\beta}_{t,u}|\boldsymbol{w}_{1:h})$, $\hat{\boldsymbol{\pi}}_{d,t} = E(\boldsymbol{\pi}_{d,t}|\boldsymbol{w}_{1:h})$ as well as $\hat{z}_{d,t} = E(Z_d = t|\boldsymbol{w}_{1:h})$ with $u = 1, \ldots, q$.

For our analysis, we *a-priori* specified $s = 100$ latent topics. In addition we set κ to very small values, e.g. 0.001, in order to ensure that the estimated topic distribution for each document will assign a probability of nearly one to a single topic and very small probabilities to all other topics. Such topic distributions allow to classify the documents into topics without loss of information by switching from soft to hard assignments. The constraint enables that the topic of each document is uniquely determined. The resulting dummy variables that encode if a document belongs to a topic or not served as possible explanatory variables for subsequent modeling of the fatality numbers.

3.1.2 Preprocessing Report Summaries

In order to make the report summaries accessible for text mining techniques and Latent Dirichlet Allocation, we used stemming functions to reduce derived words to their stem. For instance, we reduced the words “friendly”, “friend” or “friends” to their stem “friend”. Additionally, we eliminated stop words (terms that are extremely common and are not relevant for content of the sentence, e.g. a, an, and, or, because).

After stemming and stop word removal, we built a Document-Term Matrix (DTM) from the report summaries, which served as input for estimating topics of the documents. Here each row of the DTM stands for a single report summary and the columns contain the terms within the corpus of all report summaries. Each entry in this matrix represents the frequency of a specific term in a specific document.

3.2 Manifest Negative Binomial Mixtures

To model the observations $Y_i, (i = 1, \ldots, n)$, with realisations y_i we have chosen a flexible and non-linear approach that allows us to incorporate information of p observed explanatory variables $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^T$. To achieve this we look for a segmented model $\mathcal{M}_{\mathcal{R}}(Y, \boldsymbol{\vartheta})$ consisting of r segments $R_k, k = 1, \ldots, r$. Here, Y stands for the random variable giving rise to the observed values and $\mathcal{R}$ denotes the set $\{R_k\}_{k=1,\ldots,r}$. The segments R_k arise from differences due to input variables $x_1, \ldots, x_p$. The (local) model in each segment R_k, $\mathcal{M}_k(Y, \boldsymbol{\vartheta}_k)$, has its specific parameter vector $\boldsymbol{\vartheta}_k$. The vector of all segment-specific vectors $\boldsymbol{\vartheta} = (\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_r)^T$ therefore denotes the combined parameter vector of global model over all segments.

Since the distribution of Y will be modeled solely by a negative binomial distribution in each segment, we call this model class *manifest mixture models* - as opposed to latent mixture models - since our approach identifies clusters based on information from explanatory variables that form the segments and need not be specified *a priori*. Latent and manifest mixture models for overdispersed count data have been used before in various contexts. For example, Deb and Trivedi (1997) used finite latent mixture negative binomial models to model demand for medical care by the elderly or Ramaswamy et al. (1994) proposed latent class negative binomial regressions for purchase behavior. Both approaches did not include explanatory information for building the classes. Covariate driven

Quasi-Poisson tree approaches for modeling overdispersed count data have been proposed by Choi et al. (2005). Their approach is similar to ours, but we have used negative binomial distributions to account for overdispersion and a different tree algorithm. Using trees to estimate such mixture models has the additional benefit of inherent variable selection.

3.2.1 Model

Let the observed values be denoted by $y_1, \ldots, y_n$ where each is a realisation of the random variable Y. The conditional distribution of Y, $D(Y|\cdot)$, is modeled with a tree-like partition function f depending on the state of p input vectors (explanatory variables), $\mathbf{x} = (x_1, \ldots, x_p)$ stemming from the sample space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$. This means we have a tree model $\mathcal{M}_{\mathcal{R}}(Y, \boldsymbol{\vartheta})$ of the form,

$$D(Y|\mathbf{x}) = D(Y|f(x_1, \ldots, x_p))$$

where the function f partitions the overall covariate space $\mathcal{X}$ into a set of r disjoint segments $R_1, \ldots, R_r$ such that $\mathcal{X} = \bigcup_{k=1}^{r} R_k$. In each segment R_k, a model for the conditional distribution, denoted by $\mathcal{M}_k(Y, \boldsymbol{\vartheta}_k)$ is assumed to hold. The overall model, $\mathcal{M}_{\mathcal{R}}(Y, \boldsymbol{\vartheta})$, is the collection (or mixture) of all segment-specific models.

Our model for the conditional distribution $D(Y|\mathbf{x})$ within each segment $R_k, k = 1, \ldots, r$, $\mathcal{M}_k(Y, \boldsymbol{\vartheta}_k)$, is a negative binomial distribution with mean μ_k and dispersion parameter θ_k, i.e having the probability mass function

$$P(Y = y; \mu_k, \theta_k, k) = \frac{\Gamma(y + \theta_k)}{\Gamma(\theta_k) y!} \left(\frac{\mu_k}{\mu_k + \theta_k}\right)^y \left(\frac{\theta_k}{\mu_k + \theta_k}\right)^{\theta_k}$$

with $y \in \{1, 2, \ldots\}$, and $\Gamma(\cdot)$ denoting the gamma function. Mean and variance of Y for each segment R_k are

$$\mathrm{E}(Y) = \mu_k \qquad \mathrm{Var}(Y) = \mu_k + \mu_k^2 \theta_k^{-1} \tag{1}$$

Please note that the above formulation pays dues to interpreting the negative binomial as a gamma mixture of Poisson distributions (Aitkin et al., 2009) and thus essentially being a Poisson model that can account for extra variation, which is also reflected in the mean-varaince identities. It can be seen as a two-stage model for the discrete response Y in each segment R_k (cf. Venables and Ripley, 2002),

$$Y|V \sim \mathrm{Poisson}(\mu_k V), \qquad \theta_k V \sim \mathrm{gamma}(\theta_k). \tag{2}$$

Here V is an unobserved random variable having a gamma distribution with mean 1 and variance $1/\theta_k$. However, the marginal mean-variance identities for Y in (1) hold whenever V is a positive-valued random variable with mean 1 and variance θ_k^{-1} and V needs not necessarily be gamma-distributed (Lawless, 1987). Our approach integrates conceptually well with other approaches of modeling fatalities that use Poisson or Quasi-Poisson models and in principle these models might also be used for $\mathcal{M}_k(Y, \boldsymbol{\vartheta}_k)$. Using the negative binomial has the advantage over a Poisson model to account for extra variation and over Quasi-Poisson to integrate nicely into a maximum likelihood framework (see Venables and Ripley, 2002).

3.2.2 Estimation

To estimate the manifest mixture model, we employ the model-based recursive partitioning framework of Zeileis et al. (2008). We consider an intercept-only model estimated from a negative binomial likelihood which is then recursively partitioned based on the state of the partitioning covariates. In our case, the algorithm of Zeileis et al. (2008) becomes as follows (cf. Rusch and Zeileis, 2011):

1. Fit a negative binomial intercept-only model to all observations in the current node

2. Assess instability of the mean parameter estimate in the curent node, $\hat{\mu}_k$, with respect to permutation or ordering of each partitioning variable $x_1, \ldots, x_p$

3. Choose the covariate associated with the highest instability for splitting

4. Compute the binary split that, for all rival partitions, locally optimizes the sum of the partition specific negative log-likelihood functions

5. Repeat recursively until no split variables are found or any other stopping criterion is fulfilled

This algorithm ensures that we gets unbiased splits (Hothorn et al., 2006; Kim and Loh, 2001). Stability of the parameter estimates in Step 2 is assessed by means of generalized M-fluctuation tests (Zeileis and Hornik, 2007). Their behavior can be controlled by the global significance level α and this can be regarded as pre-pruning to avoid overfit. Additionally, using pre-pruned trees has the advantage of inherent variable selection. The depth of the tree can be further controlled by specifying the minimum number of observations a terminal node should contain. Please note that splitting is carried out with an inferential procedure based on the cumulative empirical process of the score function deviations for the mean parameter only, because we are solely interested in $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_r)$, whereas the dispersion parameters $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_r)$ are regarded as nuisance parameters. They are, however, estimated for each segment via maximum likelihood. Because of treating them as a nuisance they influence the splitting process indirectly via the likelihood. It is therefore possible that splitting occurs even if the means are practically the same due to a difference in θ[7].

Eventually we get a classification of all observations into a set of partitions $\mathcal{R} = \{R_1, \ldots, R_r\}$. The negative binomial distributions in these partitions are characterized by the parameter estimates $\hat{\mu}_k$ and $\hat{\theta}_k, k =, 1 \ldots, r$ and the estimated overall mixture model, $\mathcal{M}_{\mathcal{R}}(Y, \hat{\boldsymbol{\vartheta}})$, by $\hat{\boldsymbol{\vartheta}} = ((\hat{\mu}_1, \hat{\theta}_1)^T, \ldots, (\hat{\mu}_r, \hat{\theta}_r)^T)$.

3.2.3 Pre-pruning The Trees

To find sensible values for the significance level of the parameter stability test as well as for the minimal number of observations per node, we fitted different models using a grid of the two algorithm metaparameters. Specifically, we used global significance levels α of 1×10^{-7}, 5×10^{-7}, 1×10^{-6}, 5×10^{-6}, 1×10^{-5}, 5×10^{-5}, 1×10^{-4}, 5×10^{-4}, 1×10^{-3}, 5×10^{-3}, 1×10^{-2} and 5×10^{-2}. Very low values for α were chosen because of the size of the data set (using significance levels of around 0.01 might lead to spurious significances due to sample size). For the minimum number of observations per node we used values of 52, 100, 200, 300, 400, 500, 600 and 700. We then fitted a manifest mixture model for all $12 \times 7 = 84$ combinations of metaparameters and chose the tree that enabled the best explanation.

4 Results

In our analysis the modeled responses were the number of fatalities of the ACF, of Coalition troops, of troops of the host nation (Afghan police and soldiers), of civilians and all fatalities combined for every incident. In the following sections we discuss each of these cases separately and use tabular and graphical representations of the manifest mixture model. Without loss of generality, we label the segments $k = 1, \ldots, r$ in an increasing from right to left as they are displayed in the plots. This is of course arbitrary and should not imply a natural ordering of the k segments (terminal

[7] Although we treat θ as a nuisance parameter, strictly speaking it is not. In a negative binomial model where both the dispersion and mean parameter are estimated, they are not orthogonal. As it is included in the likelihood and therefore in the score function, it can influence the splitting process indirectly even though we do not explicitly look for instability in estimates of θ.

nodes). Each terminal node (leaf) k is associated with a negative binomial distribution with parameter estimates $\hat{\mu}_k$ and $\hat{\theta}_k$ and the vector of all parameters in the terminal nodes combined is the parameter vector of the final model $\mathcal{M}_{\mathcal{R}}(Y, (\hat{\mu}_k, \hat{\theta}_k)^T),\ k = 1, \ldots, r$.

We visualize the negative binomial distribution in each terminal node with a parsimonious plot of the magnitudes of the mean and the standard deviation. The vertical line in each panel marks the location of the mean, the horizontal line shows the distance between zero and one standard deviation (cf. Friendly, 2001). The height of the vertical line is the deviance divided by the degrees of freedom and indicates goodness of fit of the intercept-only model in the node. A smaller height means better fit.

A presentation of the selected estimated latent topics, the most frequent keywords, how many reports were assigned to them and for which fatality group they served as a splitting variable can be found in Tables 7 and 8. For instance, the report summary from Section 2.1 belongs to Topic 16, "Convoy Attacks (Kandahar)". In Table 7 the ten most frequent words of this (and all other topics) are listed. One can see that this topic describes events related to ambushed convoys or vehicles. Additionally, we can see in the first row of Table 7 (`numberDOC`) that overall 533 incidents have been assigned to this topic and that this topic has been used as a splitting variable only when modeling civilian fatalities. In Section 4.2 we discuss this topic in greater detail.

First we start with an overall analysis of all fatalities combined. Later we look at the number of fatalities for specific groups, namely fatalities of the civilian population, fatalities of Anti-Coalition Forces, fatalities of US or allied forces as well as fatalities of police or military of the host nation.

4.1 All Fatalities Combined

For all fatalities combined, we find $r = 14$ segments (with a global significance level for the fluctuation tests of $\alpha = 1 \times 10^{-4}$ and a minimum number of observations in each terminal node of 300). For each segment, Table 2 contains the segment number (Segment), parameter estimates in the transformed space ($\log(\hat{\mu}_k)$ and $\hat{\theta}_k$) and standard errors (se($\log(\hat{\mu}_k)$ and se($\hat{\theta}_k$)), degrees of freedom ($n_k - 1$, df), deviance (dev), the maximum number of fatalities (max) and the percentage of incidents with no fatalities (%zero).

Segment	$\log(\hat{\mu}_k)$	$\text{se}(\log(\hat{\mu}_k))$	$\hat{\theta}_k$	$\text{se}(\hat{\theta}_k)$	df	dev	max	% zero
R_1	0.779	.120	.089	.007	829	436.36	101	75.4
R_2	−0.399	.102	.069	.006	1530	554.37	68	84.8
R_3	0.917	.113	.096	.008	848	486.90	186	72.4
R_4	0.904	.090	.386	.038	373	361.19	36	42.8
R_5	0.215	.053	.468	.037	1031	926.77	31	53.8
R_6	0.269	.098	.128	.011	899	523.48	70	73.1
R_7	0.114	.121	.275	.039	306	234.08	43	63.2
R_8	−1.882	.049	.032	.002	15887	2418.40	25	94.6
R_9	−1.635	.054	.055	.003	8068	1801.90	28	92
R_{10}	−3.227	.113	.006	.001	14213	513.4	67	98.7
R_{11}	0.269	.106	.205	.022	497	353.50	56	66.3
R_{12}	0.389	.101	.373	.046	327	288.75	35	52.7
R_{13}	−0.016	.089	.199	.019	767	504.83	21	70.2
R_{14}	−1.238	.028	.048	.001	30981	7324.10	80	91

Table 2: Segmentwise statistics for all fatalities combined. The first column gives the segment. For each segment we listed the logarithm of the estimated mean $\log(\hat{\mu}_k)$, its standard error $\text{se}(\log(\hat{\mu}_k))$, the estimated dispersion parameter $\hat{\theta}_k$ and its standard error $\text{se}(\hat{\theta}_k)$ the degrees of freedom (df), the residual deviance (dev), the highest number of fatalities reported (max) and the percentage of reports with zero fatalities (% zero).

The first segment consists of $n_1 = 830$ incidents with an average number of fatalities of $\hat{\mu}_1 = 2.18$ per report. The maximum number of deaths is 101. 75.4% of the reports report no fatalities. This segment is characterized by reports that belong to Topic 5 "Taskforce Bushmaster". Table 7 and Table 8 respectively display the most frequent words in the summaries of this and subsequent topics. Inspection of summaries of reports assigned to this topic indicates that it refers to directed actions against Anti-Coalition forces primarily performed by Task Force (TF) unit "Bushmaster". For instance, on 28-Aug-2007, after attempting to ambush a TF Bushmaster convoy, 100 Taliban fighters were killed in a requested air support. According to the report summary there were no civilian fatalities but one killed Afghan soldier (see Tran, 2007). This is the highest number of observed killed Anti-Coalition fighters for reports belonging to segment R_1 and the third highest in the whole war diary. All in all 1808 deaths are reported for this segment, 1712 of those have been ACF.

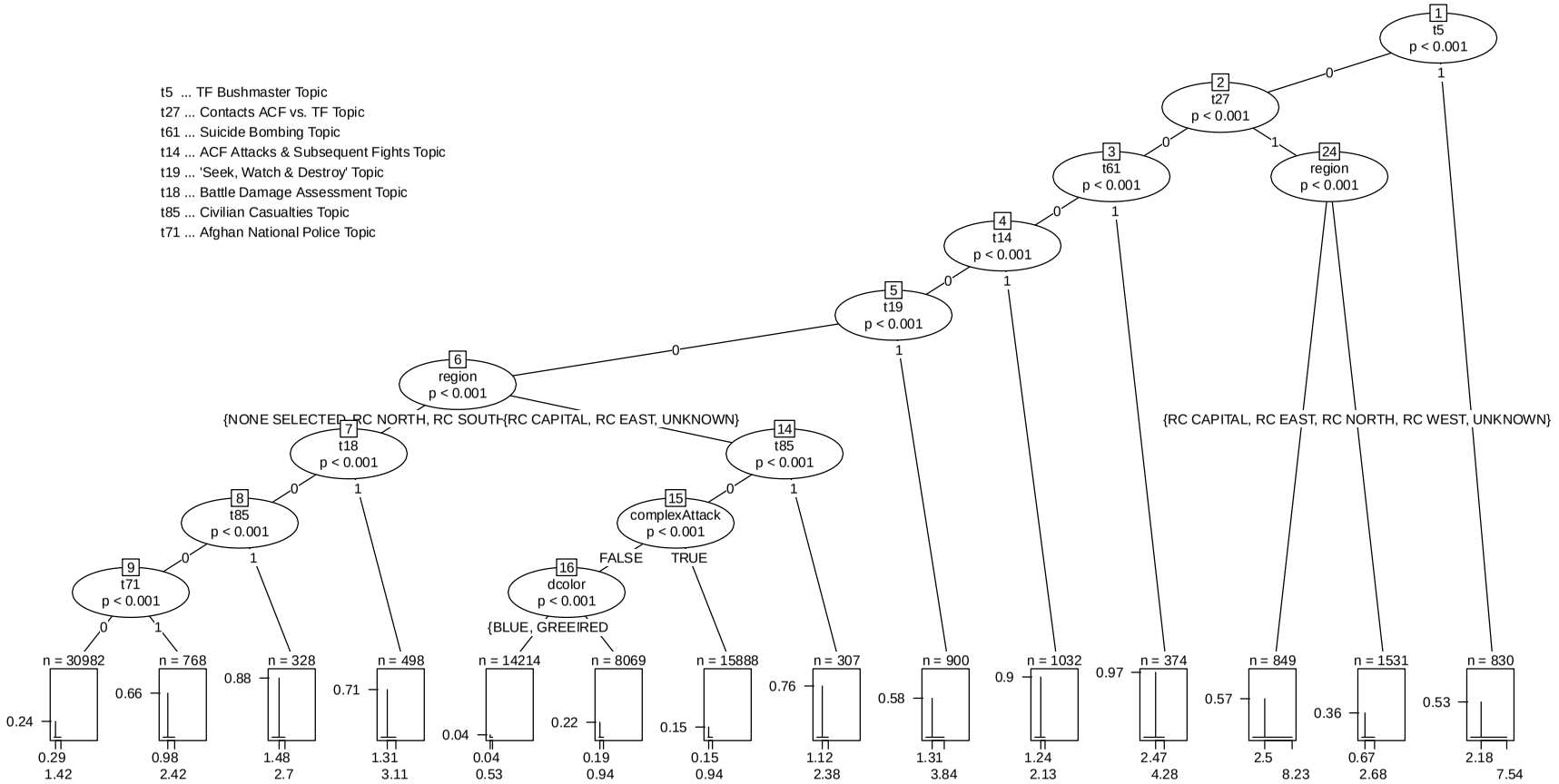

Figure 3: The manifest negative binomial mixture model tree for the combined fatalities. In the terminal nodes the vertical line marks the mean, the horizontal line the length between zero and one standard deviation and the height of the vertical line is deviance/df.

The next two segments are governed by Topic 27 "Contacts ACF vs. TFs" and differ in terms of the region they took place. The topic describes incidents where task forces or troops had enemy contact with fire fights taking place (individual combat with small arms, see Table 7). Excluded from this topic are operations performed by TF "Bushmaster" (Topic 5). Events assigned to this topic are further split according to the region where the events took place. The right branch in Figure 4.1 contains events around Kabul (`RC Capital`), `RC East`, `RC West`, `RC North` and unknown regions, as collected in segment R_2. These are associated with a death rate of $\hat{\mu}_2 = 0.671$ deaths per report. Of these 1531 incidents the maximum number of fatalities is 68 and 84.8% report no fatalities.

The 849 events belonging to Topic 27 "Contacts ACF vs. TFs" that happened in the south of Afghanistan (mainly province Kandahar, `RC South`) however show a much higher estimated fatality rate of $\hat{\mu}_3 = 2.501$. This is the highest estimated death rate of the whole analysis. Reports in this segment (R_3) have a maximum number of fatalities of 186 (the bloodiest incident in the whole war log) on 09-Sep-2006. This report (the incident being part of "Operation Medusa") notes 181 killed ACF fighters, one killed coalition force soldier and four killed Afghan soldiers 10 km southwest of Patrol Base Wilson, in Kandahar province's volatile Zhari district. This is the highest number of killed ACF fighters (or overall death) in the whole data within a single war log entry. Moreover, segment R_3 is the segment with the highest ACF fatalities (see Section 4.3). For 72.4% of events in this segment no fatalities are reported.

The next three segments consist of incidents exclusively attributable to a single topic each. First, there is Topic 61 "Suicide Bombing" with corresponding segment R_4. It describes incidents that are related to suicide bombing attacks (cf. Table 8). For example, one report assigned to Topic 61 and dated with 18-Feb-2008 reports 30 killed civilian due to a suicide bomb attack near Kandahar. The first lines of the associated report summary reads:

> TF Kandahar reported that a suicide vehicle born IED detonated at checkpoint 62D at 42R TV 551 276 in the Spin Buldak district, Kandahar province. A building was reported as on fire...

The segment's $n_4 = 374$ reports showed no fatalities in 42.8% of the cases, the only segment with a median death number higher than 0. The maximum number of killed people is 36. Accordingly, the estimated mean death rate for this segment is $\hat{\mu}_4 = 2.471$. It is the second highest overall death rate per incident, closely matching the results from R_3. However, in R_4 fatalities are mostly civilian or forces of the Afghan police forces, whereas deaths in R_3 are mostly ACF fighters. In R_4 we observe 924 deaths, 420 of those have been civilians, followed by 246 killed afghan soldiers and 233 killed ACF fighters.

Topic 14 "ACF Attacks & Subsequent Fights" gives rise to segment R_5 with an average number of deaths per incident of $\hat{\mu}_5 = 1.241$. In total, we observe 1287 deaths in the $n_5 = 1032$ reports (53.8% of whom had no deaths reported) in this segment. It is somewhat hard to identify the governing topic with an unique theme like "suicide attacks" for Topic 61, but inspection of a sample of report summaries indicates that this topic collects reports which describe smaller fights or incidents following attacks by the ACF mainly aimed at Afghan forces, resulting battle damage assessment (bda) and medical evacuation. Most victims of this segment have therefore been Afghan soldiers (529), but we also observe 326 killed civilians, 170 ACF and 262 killed allied soldiers. In contrast to Topic "TF Bushmaster" or "Contacts ACF vs. TFs" we do not find a report with an extremely high number of fatalities, the maximum number of reported fatalities being 31. On 10-Sep-2007 we can read in the associated report summary:

> TF HELMAND reported a large explosion on the road in between two markets at 41R PR 48565 20919, 4.8km northeast of FOB PRICE, NAHRIJ SARIAJ in HELMAND province. An unknown number of non combatants were KIA and WIA IVO GSK

attending a market....It appears to be a suicide attack intended to target a Police Chief Aram Attulah.

Segment R_6, is constructed of incidents that have been assigned to Topic 19 "Seek, Watch & Destroy". The most frequent terms are `update`, `att` (at this time), followed by `aaf` (Anti-Afghan Forces). Hence in this segment we find reports which describe a sequence of events (marked by `update` in the report summaries) within a mission, often monitoring instant happenings, primarily focused on Anti-Afghan forces (`aaf`). This segment has a mean death toll of $\hat{\mu}_6 = 1.309$ per report. Of the $n_6 = 900$ reports, 73.1% have been without fatalities. In total we observe 1061 killed ACF fighters within this segment. For civilian, Afghan soldiers and allied soldiers we find in total 117 fatalities in this segment. The highest death toll within this segment, has been observed on 26-Jul-2008, a report which accounts for 68 killed ACF fighters and 2 killed ANP

> ...no mercy reports enemy took cover in a qalaat at grid wb 48356 74598 no mercy was clear to engage w/ hellfire.at 2233 no mercy engaged qalaat approximately 3 to 4 pax came running out of the qalaat. at 2250z bearcat is breaking station to refuel and cm to rengage enemy. still have eyes on w/ sijan. update: at 2355z sijan maintain eyes on approximately 100x aaf headed south.

Of the remaining 71054 incidents not described so far, there is significant instability for the mean estimate based on the region they happened in. The first branch collects incidents in `RC Capital`, `RC East` and in `UNKNOWN` locations. Four segments result by further partitioning of these data. Segment R_7 are those $n_7 = 307$ incidents in the East, in the capital or unknown region associated with Topic 85 "Civilian Casualties". In Table 8 we see the clear context of civilian fatalities of this topic. Out of the ten most frequent terms of this topic, six are synonyms respectively acronyms of civilians. These are: `ln` (local national), local(s), civilian, `lns` (local nationals), child, nationals. The other four terms are clear synonyms of casualties, namely `wound`,`injur` (injury), `kill`, `hospit` (hospital). The mean number of deaths in this segment is $\hat{\mu}_7 = 1.12$ for $n_7 = 307$ reports. The maximum number of fatalities here is 43 and there are 63.2% of reports that reported no fatality at all.

Incidents in `RC Capital`, `RC East` and `UNKNOWN` locations not associated with Topic 85 "Civilian Casualties" can be distinguished by the complexity of the attack and by whom they have been carried out (`dcolor`, see Section 2). For complex attacks (segment R_8 with $n_8 = 15888$) the mean fatality number is $\hat{\mu}_8 = 0.152$. No fatalities are reported in 94.6% of the cases and the highest fatality number in this segment is 25. For incidents that are not classified as complex attacks, and are flagged as `red` (segment R_9) we estimate a mean of $\hat{\mu}_9 = 0.195$. Here, 92% of reports have recorded no deaths and the highest number of fatalities in a report has been 28.

Those not flagged as `red` (segment R_{10}) report a much lower average death toll of $\hat{\mu}_{10} = 0.040$. 98.7% of those incidents are not connected with someone's death. This segment has the third lowest fatality rate of all segments. However, the maximum number of fatalities for these reports is 67.

Those incidents not collected within segments R_1 to R_{10} share the regions they happened in: `RC North`, `RC South` and `RC West` (this also includes those incidents with an unspecified location). For those, three topics are used for further segmentation, Topics 18, 85 and 71. With Topic 18 "Battle Damage Assessment" one further split-topic in modeling all fatalities places particular emphasis on battle damage and battle damage assessment (see Table 7). Two out of the four most frequent terms are `bda` (battle damage assessment) and `damage`. Such battle damage assessment may come along with requested airstrikes, e.g. helicopter attacks (`ah`). The resulting segment, R_{11}, has a mean number of $\hat{\mu}_{11} = 1.309$ reported fatalities. 66.4% of those reports contain no death toll and of those who do, the maximum is 56 deaths on 03-Mar-2009:

> ...after ensuring that no civilians were in the vicinity, com prt kdz authorized an airstrike. at 2119z, an f-15 dropped 2x gbu 38 bombs. at 2158z, bda conducted

> by f-15/rover was that 56x ins kia (confirmed) and 14x ins fleeing in ne direction. the 2x fuel trucks were also destroyed...

All fatalities are stated to be ACF Fighters in the war log. In the media however, this event has been named Kunduz airstrike (RC North) and the killed people were civilians (see guardian.co.uk, 2010) who had been invited by the Taliban to take fuel from the trucks (see Amnesty International, 2009). The Taliban had stolen the two fuel trucks and the resulting airstrike against the fuel trucks had killed those 56 civilians.

Segment R_{12} (governed by events from Topic 85 "Civilian Casualties" happening in the South, North, West or in a non-specified regions) has an estimated mean of $\hat{\mu}_{12} = 1.476$. The percentage of reports without killings is 52.7% and the highest death toll is 35. The governing topic, Topic 85, has already appeared earlier as the governing topic of R_7. Therefore R_{12} and R_7 are corresponding topic-wise and differ in terms of their location. It is interesting to see that R_{12} has a moderately higher fatality number per incident, probably due to events in the south. Incidents in Kabul and the East (R_7) are associated with lower death numbers and a higher percentage of reports with zero deaths. However, the report with the highest fatality number for this topic is part of R_7, describing an attack on the Indian Embassy in Kabul. 42 civilians and one Taliban have been killed.

In the regions `RC North`, `RC South`, `RC West` or unspecified regions, Topic 71 "Afghan National Police" gives rise to segment R_{13} with nearly one death per incident on average ($\hat{\mu}_{13} = 0.984$). Of the $n_{13} = 768$ events 70.2% have not resulted in deaths. Topic 71 can be categorized as describing events with an involvement of the Afghan National Police (ANP). Often, these have been attacks on ANP checkpoints or police stations. The highest number of victims within segment R_{13} has been observed on 28-Aug-2006 in Helmand Province (South):

> AT 0847Z TFH received information from a local reporter that a SIED targeted a former police chief named KHANO, who was possibly KIA. The incident, still unconfirmed ATT,...

In total, this report lists 21 victims, all civilian. All 30982 reports not included so far constitute the last segment R_{14}. The mean number of fatalities here is $\hat{\mu}_{14} = 0.290$. The maximum number of reported fatalities in this segment is 80, while 91% of the reports do not list any deaths.

4.2 Civilian Fatalities

The manifest negative binomial mixture model of civilian fatalities is visualized in Figure 4.2. We used a significance level of $\alpha = 5 \times 10^{-6}$ and a minimum number of incidents of 300 per segment. The resulting model is a mixture of fourteen negative binomials distributions and again each terminal node (segment) R_k is associated with a negative binomial distribution with estimated parameters $\hat{\mu}_k$ and $\hat{\theta}_k$. Table 3 contains segmentwise indices, descriptive statistics, parameter estimates and goodness-of-fit statistics.

Segment	$\log(\hat{\mu})$	$\text{se}(\log(\hat{\mu}))$	$\hat{\theta}$	$\text{se}(\hat{\theta})$	df	dev	max	%zero
R_1	0.521	.099	.384	.046	325	295.76	42	50.3
R_2	−0.885	.152	.212	.046	309	164.79	14	79.4
R_3	0.116	.156	.116	.016	373	199.36	30	76.2
R_4	−1.152	.123	.074	.010	1031	313.90	19	88.5
R_5	−0.861	.177	.070	.012	529	171.25	50	87.2
R_6	−1.067	.222	.071	.016	342	105.86	19	88.1
R_7	−1.665	.145	.060	.010	1046	242.92	15	91.8
R_8	−1.491	.160	.169	.043	403	159.40	7	86.6
R_9	−2.134	.209	.021	.004	1275	139.77	21	96.1
R_{10}	−1.787	.193	.096	.025	441	123.00	8	90.7
R_{11}	−1.855	.263	.036	.010	491	81.41	20	94.1
R_{12}	−3.547	.426	.003	.001	1873	39.05	37	99.3
R_{13}	−1.116	.281	.034	.008	411	81.46	25	92.2
R_{14}	−3.845	.057	.006	.000	67759	1875.10	67	99.1

Table 3: Segmentwise statistics for civilian fatalities. The first column gives the segment. For each segment we listed the logarithm of the estimated mean, its standard error, the estimated dispersion parameter and its standard error, the degrees of freedom (df), the residual deviance (dev), the highest number of fatalities reported (max) and the percentage of reports with zero fatalities (% zero).

We can see that the first segment R_1 is governed by Topic 85 "Civilian Casualties" and `dcolor=red`. The most frequent terms (see Table 8) of Topic 85 suggest that it describes civilian fatalities and casualties that have no other context than being that: civilian fatalities (see Section 4.1 for a more thorough discussion of this topic). Together with the flag for "enemy action", it is clear that this segment stands for fatalities of the civilian population in actions of the ACF. With an estimated mean of $\hat{\mu}_1 = 1.684$ it has the highest average number of civilian fatalities per incident of all segments. The maximum number of fatalities reported for this segment is 42 and a mere 50.3% of reports list no civilian fatalities. This is by far the lowest percentage of reports without deaths and accounts for the high mortality.

Incidents in R_2 also belong to Topic 85 and therefore refer to civilian casualties but have been flagged as `blue` or `green` which refers to actions of allied forces or neutral ones. Here, the average number of fatalities drops to $\hat{\mu}_2 = 0.413$ which is the fourth highest overall rate. This segment's highest reported death toll is 14, with 79.4% reports listing no fatalities.

Besides Topic 85, 10 other topics are sequentially selected as splitting variables. The topics are 61, 14, 16, 57, 11, 86, 71, 79, 21 and 29. By looking at the most frequent terms for the topics, one can grasp their meaning quite clearly (see Table 8). Topic 61 for example (see also Section 4.1) describes suicide attacks. The associated segment, R_3, has an average number of civilian fatalities of $\hat{\mu}_3 = 1.123$ per incident. This is the second highest value for civilian fatalities. The highest death toll of this segment within a single report has been observed on 18-Feb-2008. 30 civilians and one suicide bomber have been killed in Kandahar province:

> ...The incident site was a busy market with an estimated 100-150 Local Nationals (LN) within a 50 m radius of the SUV on detonation. It was reported that LN casualties were; 30 killed and 37 injured. CF received a credible warning of a possible suicide attack within Spin Buldak from the Afghan Border Police (ABP) Commander...

76.2% of the reports in this segment listed no civilian fatalities.

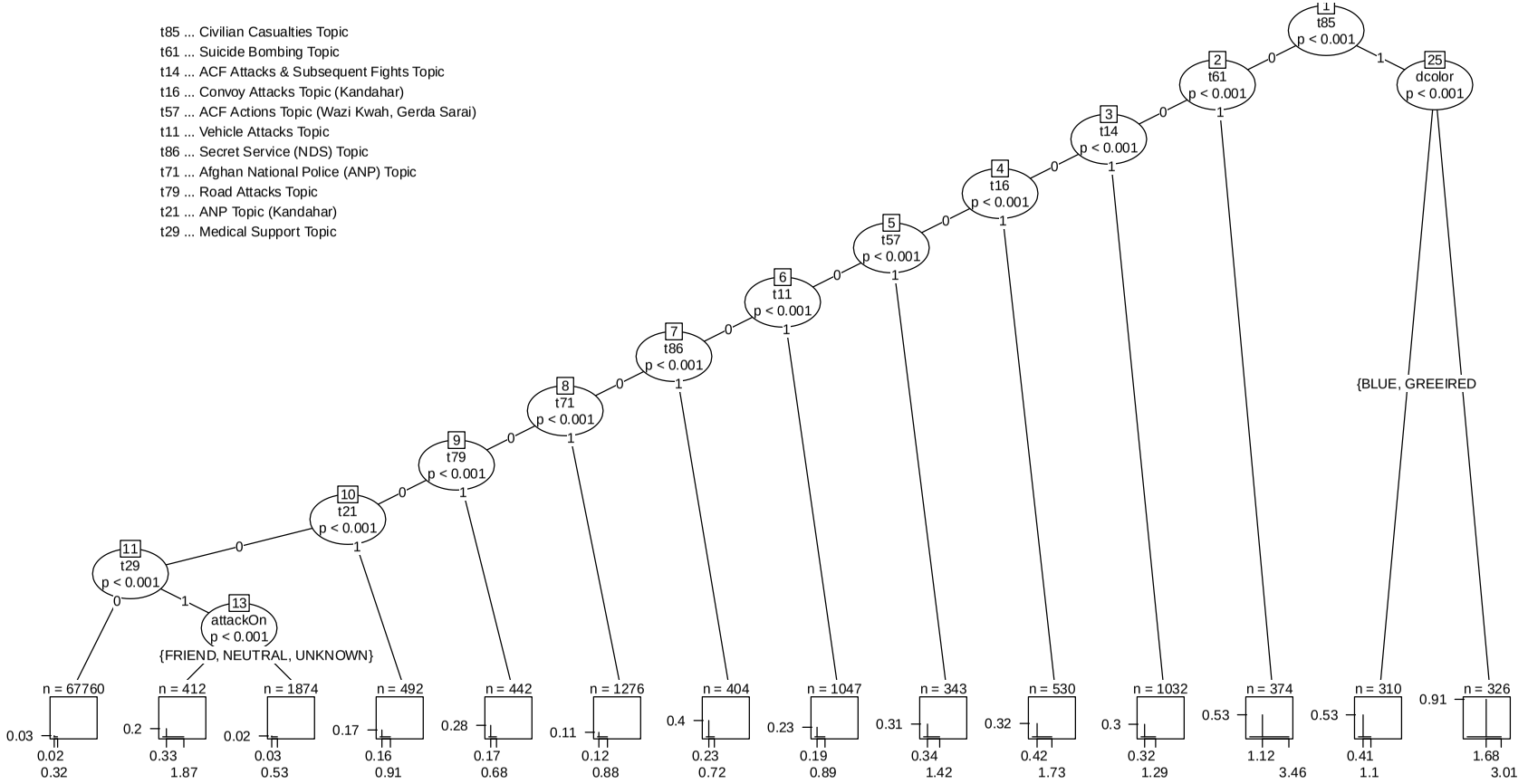

Figure 4: The manifest negative binomial mixture model tree for civilian fatalities. In the terminal nodes the vertical line marks the mean, the horizontal line the length between zero and one standard deviation and the height of the vertical line is deviance/df.

Topic 14 "ACF Attacks & Subsequent Fights" governs the next segment (see Section 4.1). The associated incidents in segment R_4 have $\hat{\mu}_4 = 0.316$ civilian deaths on average, with the highest reported number being 19. Overall, 88.5% reports do not mention civilian fatalities.

Next, there is segment R_5 which corresponds to Topic 16 "Convoy Attacks". It collects incidents associated with ambushes on vehicle convoys especially around Kandahar. The mean fatality rate for this segment R_5 is $\hat{\mu}_5 = 0.423$, the third highest average number of fatalities. In this segment the maximum number of civilian deaths reported is 50. The associated report summary notes:

> Explosion kills 50 near Qaba Mosque in Spin Boldak.

and there are 87.2% reports without killed civilians.

Segment R_6, induced by Topic 57 "ACF Actions (Wazi Kwah, Gerda Sarai)", refers once again to actions by Taliban fighters, primarily happening in the regions *Wazi Kwah* and *Gerda Serai* for which the mean number of civilian fatalities has been estimated as $\hat{\mu}_6 = 0.344$. 88.1% of the reports list no civilian fatalities. Of those who do, the most deadly incident has led to 19 killed civilians. The associated report summary, dated on 22-Sep-2006, describes an explosive device attack on a truck:

> RC(S) reported a truck carrying approx. 22x local workers doing work on border check points was hit by an RCIED. 19x LN were killed and 3x LN were wounded. Many of those that survived the IED blast were later killed by overwat.

Segment R_7 ($\hat{\mu}_7 = 0.189$) is governed by Topic 11 "Vehicle Attacks", describing (alleged) incidents with (improvised) explosives associated with vehicles and convoys sometimes attacked by US or allied troops; hence it is once more an ACF attack topic. Here, the highest number of fatalities reported is 15 described in a report which is dated on 03-Aug-2006:

> (DELAYED REPORT) TF Orion reports and IED STRIKE IVO Panjwayi. ANP reports 10-15 civilian KIA and 10-15 civilian WIA. A civilian car advanced rapidly to the Orion 79 convoy, who where inroute to support a TIC in Panjwayi. Troops stopped the car and after a few minutes the vehicle advanced again and detonated.

91.8% of reports assigned to this segment list no civilian death count.

The next topic that splits off a segment (R_8) is Topic 86 "Secret Service (NDS)". Incidents belonging to this topic are more or less associated with the Afghan intelligence agency (NDS) and seem to be partly concerned with specific operations against certain people. Inspection of the reports which account for civilian victims within this segment suggests that these fatalities are not connected to allied forces action. Additionally the reports are not necessarily connected to war-like situations either, but rather are information of events of interest to the US like assassination of local leaders. For instance, in a report summary assigned to Topic 86 we can find a description of the assassination of Mohammed Anwar, Chief Mullah of the pro government Mullah Council, by the Taliban fighters (07-Apr-2007). Detailed information about this attack has been provided by NDS. The $n_8 = 404$ incidents in this segment have on average $\hat{\mu}_8 = 0.225$ civilian fatalities. The maximum number of civilian deaths reported is 7. There are 86.6% reports not mentioning civilian fatalities.

The next segment is constituted of incidents related to the Afghan National Police and attacks or events against or by them, specifically happening at ANP checkpoints or with vehicles (Topic 71 "Afghan National Police", see Section 4.1). The average number of civilian deaths in this segment is $\hat{\mu}_9 = 0.118$. The highest loss of civilian life reported is 21 due to an incident at 28-Aug-2006:

> explosion in vicinity of (IVO) the bazaar ((Laskar Gah, Helmand province)...TFH received information from a local reporter that a SIED targeted a former police chief named KHANO, who was possibly KIA.

96.1% of these reports are not listing fatalities.

Topic 79 "Road Attacks" gives rise to another segment, R_{10}, with on average $\hat{\mu}_{10} = 0.167$ fatalities per incident. Fatalities within this topic are usually due to attacks on or with trucks that are used as mobile bombs, mostly aimed at convoys or happenings alongside roads. This topic is similar to other IED topics like 61 and 11. For this segment 90.7% of the $n_{10} = 442$ filed reports list no civilian deaths. Of those who do, the maximum number reported is 8 due to a detonated Toyota Corolla (13-Mar-2008):

> The Convoy consisted of one Ford F350 truck in the lead and one Land Cruiser following. The F350 noticed a slow moving Corolla in the inside northbound lane. As the Excursion moved to pass the Corolla on the right, the Corolla detonated. The F350 bore the brunt of the explosion and caught fire. The Land Cruiser was also damaged but drivable...

Another segment, R_{11}, consists of incidents connected to Topic 21 "ANP Kandahar" which is related to Kandahar and/or the Afghan National Police. The five most frequent terms within this topic are `anp` (Afghan national police), `polic`(e), `chief`, `district`, `aup` (Afghan uniformed police). Here, the average number of civilian fatalities per incident is $\hat{\mu}_{11} = 0.156$, with the maximum number of deaths in a report being 20 and 94.1% mentioning no fatalities. The report which accounts for those 20 killed civillians happened at 01-Jun-2005:

> TF Bayonet reported an explosion in Kandahar city (41R QR 56500 00700) at 0430Z. The explosion occurred at a funeral. 20 civilians were killed and 44 wounded (12X minor injuries and 32X hospitalized). Repeated offers of assistance were made to the Kandahar governors office...

For the last three segments, one more topic has been used for splitting. This topic, Topic 29 "Medical Support", describes primarily incidents with the need for medical support not only due to fights but also e.g. traffic accidents. The segments themselves differ in at whom the action has been aimed at. Segment R_{13} describes attacks on enemy targets ($\hat{\mu}_{13} = 0.328$). Here, 92.2% of reports list no deaths while the maximum number of killed civilians in this segment is 25. Its sister segment R_{12} (action towards friend, neutral or unknown targets) has a much lower average number of civilian fatalities of $\hat{\mu}_{12} = 0.029$ also because 99.3% of the reports mention no civilian fatalities. The highest death toll within this segment is accounted for by a demonstration in Jalalabad City at 11-May-2005 where 37 civilians have been killed:

> TF THUNDER reported that the size of the demonstration is 250x pax. The crowd is becoming unruly and is throwing rocks, burning tires and vandalizing buildings. Gunshots were also fired...LN was struck by vehicle...CJTF76 approves medevac mission 05-11A at 0745Z. Medevac is canceled due to PTS Status. PT is being casvac amp;apos; amp;apos;d to JBAD hospital. At 0810Z the UN JEMB security operations manager in Kabul requests the immediate evacuation of their JBAD provincial staff.

The segment with the lowest overall mean number of civilian fatalities, R_{14}, is the segment that includes all incidents not assigned to any of the topics mentioned so far and it is also be far the largest (with size $n_{14} = 67760$ or 88.1% of the war logs). The average number of civilian fatalities reported is $\hat{\mu}_{14} = 0.021$. 99.1% of these reports list no civilian deaths. However, this segment also sporadically contains reports with a high number of fatalities, such as 67 fatalities as a result of a natural disaster. On 01-Apr-2007 the war logs report about 67 civilians buried in a mud slide. This is the maximum civilian death toll in the war logs for a single incident.

4.3 Anti-Coalition Forces (ACF)

For fatalities of Anti-Coalition Forces (ACF), the resulting tree model can be found in Figure 4.3. We have used a configuration of $\alpha = 5 \times 10^{-7}$ and a minimum number of incidents per segment of 200. There are twelve segments resulting.

Interestingly, only five topics have been selected to build the manifest mixture model, less then for the other fatality groups. The most frequent terms of those five topics are displayed in Table 7. The first split topic, Topic 5 "TF Bushmaster", defines segment R_1 and describes incidents related to ambushes on or by Task Force "Bushmaster". This segment corresponds one-to-one to segment R_1 for all fatalities. See Section 4.1 for a more thorough discussion of this topic. On average, $\hat{\mu}_1 = 2.063$ ACF deaths occur for these $n_1 = 830$ incidents, the second highest number of all segments in the ACF tree. This value is roughly the same as the one observed for the corresponding segment R_1 for all fatalities (see Section 4.1). The highest number of deaths reported for this segment is 100. Of all the reports 78.4% list no fatalities of the ACF.

The highest ACF fatality rate per incident is $\hat{\mu}_3 = 2.379$ for segment R_3 which comprises of incidents that belong to Topic 27 "Contacts ACF vs. TF" and have happened in `RC South`. Again we refer to Section 4.1 for a more thorough discussion of the topic. Within this segment we find the highest death toll of ACF in a single incident with 181 deaths. 23.4% of the reports list ACF fatalities. In other parts of Afghanistan, incidents associated with Topic 27 are less bloody, on average $\hat{\mu}_2 = 0.572$ deaths per incident. Here, 88.4% of reports mention no ACF fatalities, while the highest reported number is 67.

Two other topics that are relevant for this model are Topics 19 "Seek, Watch & Destroy" and 18 "Battle Damage Assessment" and have already been discussed in Section 4.1. Their most frequent terms are again displayed in Table 7. They govern segments R_4 and R_5 respectively. The means of the negative binomial distributions in both segments are estimated as $\hat{\mu}_5 = 1.231$ and $\hat{\mu}_4 = 1.179$ respectively. R_4 contained a maximum number of 68 ACF fatalities in a single incident, with 78.8% of reports not listing any. For R_5 the respective numbers are 56% and 68.1%. Incidents taking place in `RC Capital`, `RC East` or `RC West` as well as in `UNKNOWN` regions and which are characterized as complex attacks directed at enemies or unknown targets constitute segment R_6. The average number of ACF fatalities for these $n_6 = 13914$ reports is $\hat{\mu}_6 = 0.127$. 95.9% of those list no ACF fatalities. The highest number of reported deaths for this segment is 30. Complex attacks in these regions directed at friendly or neutral targets on the other hand are collected in segment R_7. Its mean is estimated to be $\hat{\mu}_7 = 0.080$. Of all $n_7 = 4323$ reports, 97.7% list no fatalities of ACF. The maximum death toll mentioned is 18.

Non-complex attacks in `RC Capital`, `RC East` or `RC West` as well as in `UNKNOWN` regions constitute segments R_8 through R_{11}. The first of those, R_8, is associated with Topic 12 "Combat Outpost Attacks". For this topic we find the most frequent terms to be `fire`, `mm` (Military message) and `cop` (Combat Outpost). A strong involvement of TF "Eagle" is also suggested. Its estimated mean is $\hat{\mu}_8 = 0.509$. This is the highest mean number of ACF fatalities in non-complex attacks in these geographical regions. The highest death toll within this segment has been reported on 22-Jun-2007:

> C/1-503 reported seeing 20x PAX at WB 3835 0896 with the JLENS. Reports ACM setting up rockets. And using caves as stageing areas. At 1348z TF Eagle declared imminent threat. By 1410z, 2x A-10s (C/S) were on station. They dropped 2x MK82 airburst and 3x GBU12 on same group of pax at WB 3906 0908....

55 ACF (and 10 civilians) have been killed in this incident.

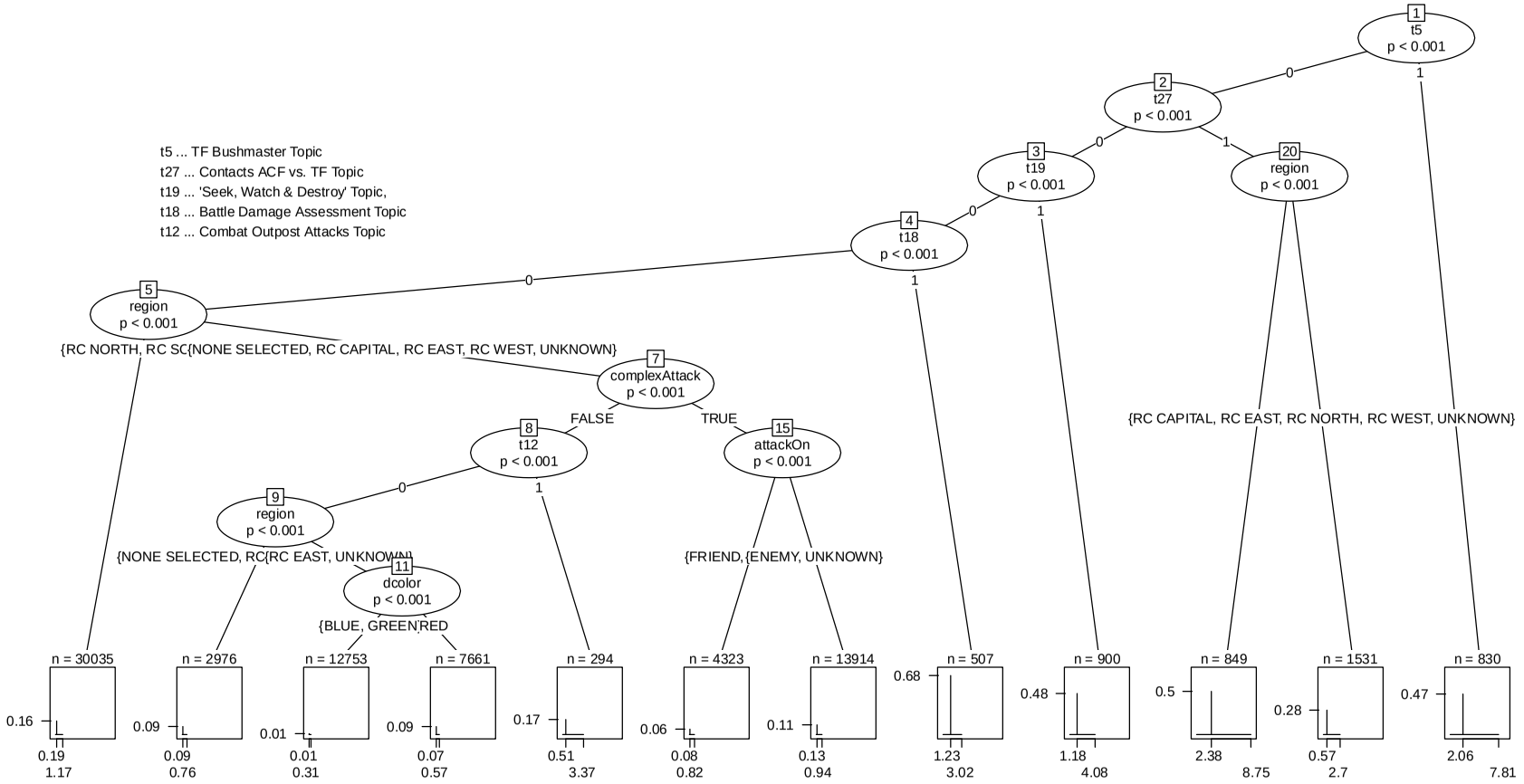

Figure 5: The manifest negative binomial mixture model tree for ACF fatalities. In the terminal nodes the vertical line marks the mean, the horizontal line the length between zero and one standard deviation and the height of the vertical line is deviance/df.

Non-complex attack incidents in `RC East` or in an `UNKNOWN` region that are not associated with any of the five topics described before constitute segments R_9 and R_{10}. Those two segments have the lowest average death rate for ACF fighters. The segment with the lowest death toll is segment R_{10} with an estimated mean of $\hat{\mu}_{10} = 0.012$ and $n_{10} = 12753$. Those reports are flagged as `blue` or `green` and therefore refer to actions of allied forces or the ISAF or are of neutral origin. The bloodiest incident in this segment has led to 13 ACF deaths and 99.6% of the reports reported none.

For those flagged as `red` (ACF action), collected in R_9, the mean number in this regions rises to $\hat{\mu}_9 = 0.072$ which can still be considered relatively low. Here we find the maximum number of killed ACF fighters to be 25. 96.9% of these reports list no ACF fatalities.

The segment of non-complex attacks that refers specifically to incidents in `RC Capital` and `RC West` is identified as R_{11} with an average of $\hat{\mu}_{11} = 0.093$. While being higher than in `RC East` and `UNKNOWN` regions it is still relatively low compared to other segments. We find 96.7 of reports to not list deaths of ACF troops. The highest reported death toll is 50 for this segment.

The last segment R_{12} collects all incidents not associated with any of the five topics and which are events that took place in the regions of `RC North` and `RC South`. In total we observe $n_{12} = 30035$ reports for this segment or roughly 39% of all reports. The average number of ACF fatalities for this segment is $\hat{\mu}_{12} = 0.189$. The maximum is reached by a report listing 80 ACF deaths. 94.2% of reports in this segment mention no ACF fatalities.

Segment	$\log(\hat{\mu})$	$\text{se}(\log(\hat{\mu}))$	$\hat{\theta}$	$se(\hat{\theta})$	df	dev	max	%zero
R_1	0.724	.132	.072	.007	829	389.23	100	78.4
R_2	−0.560	.121	.049	.005	1530	433.38	67	88.4
R_3	0.867	.126	.076	.007	848	424.49	181	76.6
R_4	0.165	.116	.090	.008	899	427.68	68	78.8
R_5	0.208	.109	.192	.021	506	346.10	56	68.1
R_6	−2.063	.063	.021	.001	13913	1565.70	30	95.9
R_7	−2.531	.156	.011	.001	4322	272.68	18	97.7
R_8	−0.680	.388	.024	.007	293	51.12	55	92.9
R_9	−2.621	.090	.021	.002	7660	708.45	25	96.9
R_{10}	−4.391	.225	.002	.000	12752	135.05	13	99.6
R_{11}	−2.367	.148	.018	.003	2975	278.21	50	96.7
R_{12}	−1.667	.036	.030	.001	30034	4733.50	80	94.2

Table 4: Segmentwise statistics for fatalities of Anti-Coalition Forces. The first column lists the segment and for each segment its node number in the tree, the logarithm of the estimated mean, its standard error, the estimated dispersion parameter and its standard error, the degrees of freedom (df), the residual deviance (dev), the highest number of fatalities reported (max) and the percentage of reports with zero fatalities (% zero).

4.4 Allied Forces

The analysis of fatalities of allied forces leads to the tree depicted in Figure 4.4. Here we have used a relatively high (compared to the available sample size) α of 0.01 and a minimum number of observation in each terminal node of 100. Lower significance values or higher number of observations would combine nodes 20, 22 and 23 together into one segment. The estimated parameter values, standard errors and goodness-of-fit as well as simple descriptive statistics are listed in Table 5.

Segment	$\log(\hat{\mu})$	$\text{se}(\log(\hat{\mu}))$	$\hat{\theta}$	$se(\hat{\theta})$	df	dev	max	%zero
R_1	−1.294	.125	.218	.046	524	244.41	5	83.8
R_2	−1.620	.192	.104	.027	398	122.74	10	89.5
R_3	−1.019	.280	.176	.069	107	51.02	6	82.4
R_4	−1.788	.426	.012	.004	471	39.05	16	96.8
R_5	−1.843	.133	.680	.354	441	217.94	4	86.7
R_6	−4.595	.078	.007	.001	39017	902.98	10	99.4
R_7	−3.110	.378	987	30227	156	43.538	1	95.5
R_8	−4.997	.302	145	3484	1626	109.85	1	99.3
R_9	−1.740	.169	.257	.093	335	137.95	4	87.5
R_{10}	−2.596	.524	.020	.011	227	20.887	4	96.9
R_{11}	−2.743	.346	.051	.027	294	41.305	4	95.9
R_{12}	−4.975	.091	.008	.001	33010	641.98	7	99.5

Table 5: Segmentwise statistics for fatalities of allied and ISAF forces. The first column lists the segment and for each segment its node number in the tree, the logarithm of the estimated mean, its standard error, the estimated dispersion parameter and its standard error, the degrees of freedom (df), the residual deviance (dev), the highest number of fatalities reported (max) and the percentage of reports with zero fatalities (% zero).

We can see that Topic 14 "ACF Attacks & Subsequent Fights" is associated with the first split. Again, its most frequent terms are found in Table 7. A more thorough discussion can be found in Section 4.1. Due to our specification of a high significance level (0.01) and a low number of minimum reports in each node, the reports belonging to this topic are divided into three segments based on the variables `complex attack` and `region`. For incidents that are not classified as a complex attack, segment R_3, we estimate the highest mean fatalities of allied soldiers to be $\hat{\mu}_3 = 0.361$. The highest number of deaths in this segment is reported to have been 6 allied soldiers. 82.4% of incidents report no fatalities.

The segments governed by Topic 14 "ACF sttacks and subsequent fights" which are characterized by complex attack incidents, R_2 and R_1, differ by the region they happened in. As before, `RC West` and `RC South` are contrasted with the other regions `RC East`, `RC Capital` and `RC North`. The former has an average fatality rate per incident of $\hat{\mu}_2 = 0.198$, the latter of $\hat{\mu}_1 = 0.274$. R_2's highest reported fatality number is 10 and of all incidents 89.5 report no deaths. For R_1 the numbers are 5 and 83.8%.

The next topics that lead to a segmentation are the Topics 34 "Aircraft (Bagram Airfield)" for R_4 and 94 "Medical Topic (Fights)" for R_5 with respective mean number of fatalities in the segments, $\hat{\mu}_4 = 0.167$ and $\hat{\mu}_5 = 0.158$. One can see that while there is a very similar mean rate in both segments, the two topics describe very different incident types. Topic 34 "Aircraft (Bagram Airfield)" can be summarized as describing events which refer to incidents around Bagram Airfield (`baf`) or aircraft incidents. For instance, on 28-Jun-2005, we can find a report summary within this segment describing the crash of a helicopter with 16 people on board (see CNN, 2005). This is the highest fatality number reported for R_4. In total we observe 79 coalition fatalities for this segment, while 96.8 reports mention no ISAF of similar fatalities. Most of the victims within this segment result from aircraft crashes. The reports with the seven highest death tolls account for 67 victims. All of the victims here have died in aircrafts accidents.

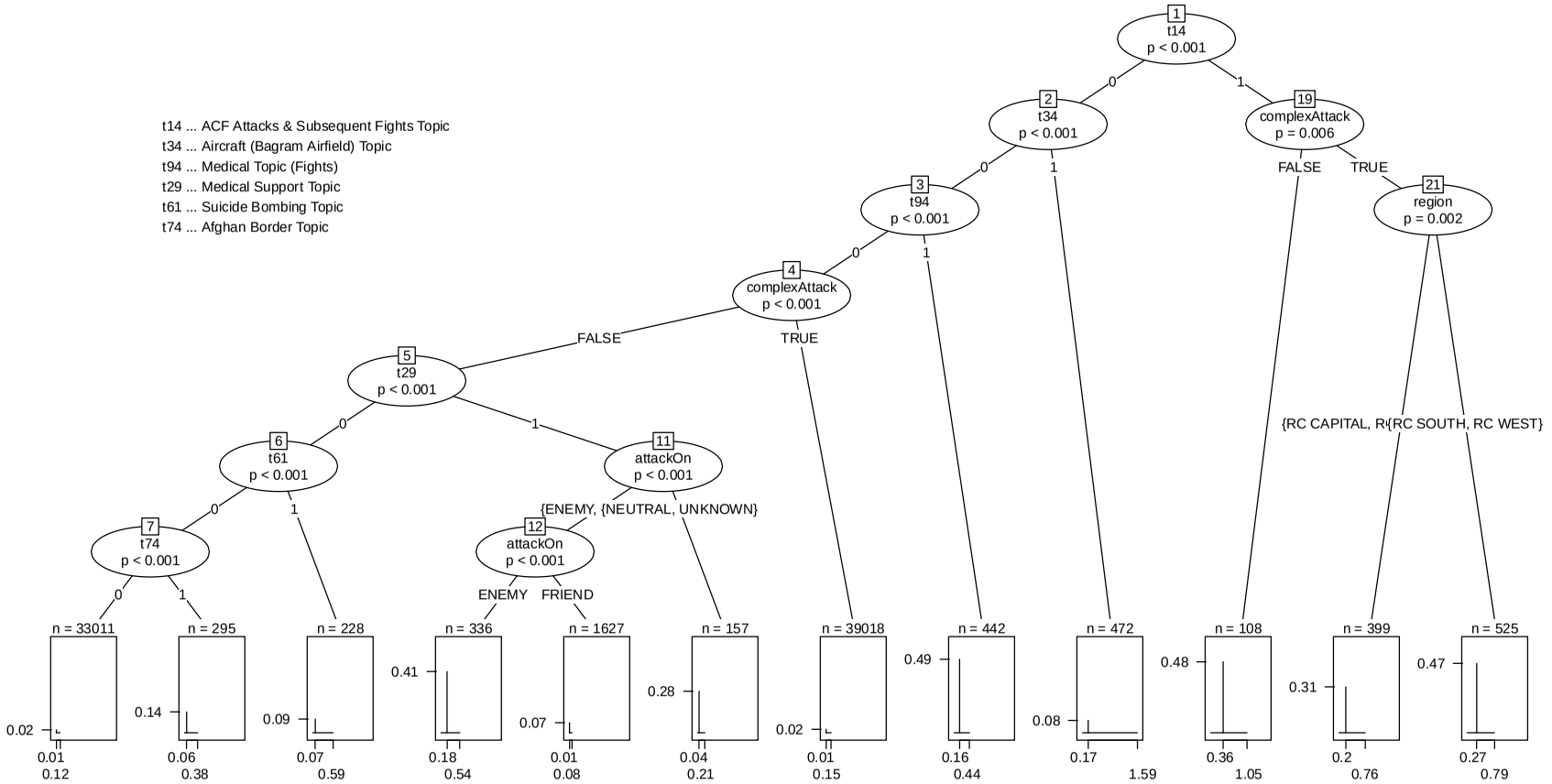

Figure 6: The manifest negative binomial mixture model tree for fatalities for allied and ISAF forces. In the terminal nodes the vertical line marks the mean, the horizontal line the length between zero and one standard deviation and the height of the vertical line is deviance/df.

Topic 94 "Medical Topic (Fights)", which governs segment R_5, may be categorized as a medical topic (see Table 8). In contrast to another medical topic, Topic 29 "Medical Support", it seems to be more focused on injuries caused by fights. Within the the most frequent terms of this topic we find `cat` (category patient, ranging from A to C), `wia` (wounded in action) and `action`. For instance, in a report which accounts for the maximum number of reported fatalities in this segment, 4 killed allied soldiers, we can read:

> ...urgent marine was transported to BSN role 3, currently in critical condition. the (1) priority marine is currently stable at DWYER STP and will later be transported to BSN...

This segment has a fatality rate of $\hat{\mu}_5 = 0.158$ and 86.7% of reports mention no fatalities.

Further segmentation depends strongly on whether the incident is classified as a complex attack or not. The 39018 incidents with complex attacks, that are not associated with the topics mentioned before, are classified together into segment R_6. This huge segment has a very low mean number of fatalities of $\hat{\mu}_6 = 0.01$. This is also reflected in the percentage of reports that reported no fatalities (99.4%). As for segments R_1, R_2 and R_3 complex attacks has a lower average death rate.

For the remaining incidents, certain topics become relevant again. First there is Topic 29 "Medical Support" (see Section 4.2). In combination with the information about towards whom the attack has been directed, it gives rise to two segments, R_8 and R_9. If the action has been directed at allied forces (including friendly fire incidents), the number of mean fatalities is very low, $\mu_8 = 0.007$. This finds its correspondence in the maximum number of reported deaths to be 1 and 99.3% of reports not mentioning any allied fatalities. On the other hand, for actions directed at enemy forces (R_9), it is considerably higher with a mean of $\mu_9 = 0.176$. Here the maximum number of fatalities for the allied forces is 4 and 87.5% of the reports list none.

Of all incidents not categorized in any of the above mentioned segments, two more segments are split off by topics: Topic 61 "Suicide Bombing" for R_{10} and Topic 74 "Afghan Border" for R_{11}. The means are $\mu_{10} = 0.064$ and $\mu_{11} = 0.075$ respectively. The topics most frequent terms can be found in Table 8. As already mentioned, Topic 61 refers to suicide attacks (see Section 4.1). The associated segment R_{10} has 4 allied deaths listed as its maximum and overall 96.9% of non-fatal incidents (for coalition troops).

Topic 74 "Afghan Border" contains reports in which the terms `afghan`, `border`, `force`, `coalition` and `afghanistan` appear most frequently. The corresponding segment R_{11} has an estimated mean death toll of $\hat{\mu}_{11} = 0.074$. The maximum number of killed allied soldiers within this segment is 4, and 95.9% of the reports do not report any fatalities. The report summary associated to the event which accounts for 4 killed allied soldiers is fairly uninformative. It only states that 4 coalition soldiers were killed in a search operation on 12-Feb-2004.

The remaining 33011 incidents constitute segment R_{12} which has a mean fatality rate of $\hat{\mu}_{12} = 0.007$. 99.5% of those reports have no ISAF or coalition fatalities listed. The maximum number of deaths is 7.

4.5 Afghan Troops

The last group of fatalities we investigate are those of forces of the host nation, such as police and Afghan military. Our recursively partitioned negative binomial mixture approach yields nine segments when using a minimum number of incidents in each node of 100 and a global significance level of the parameter instability tests of $\alpha = 0.005$. The resulting tree is visualized in Figure 4.5 and the according values can be found in Table 6.

Segment	$\log(\hat{\mu})$	$\text{se}(\log(\hat{\mu}))$	$\hat{\theta}$	$se(\hat{\theta})$	df	dev	max	%zero
R_1	−0.245	.118	.216	.030	427	274.09	19	72
R_2	−0.914	.114	.118	.015	847	350.18	20	84
R_3	−0.671	.080	.214	.023	1034	587.58	21	77.2
R_4	−0.419	.199	.076	.013	373	141.27	24	84.2
R_5	−0.449	.227	.178	.047	138	78.44	15	76.3
R_6	−2.738	.356	.165	.155	169	37.48	2	94.7
R_7	−1.456	.175	.126	.030	398	142.09	11	87.5
R_8	−3.390	.044	.012	.001	61164	3011.40	27	98.4
R_9	−4.876	.185	.004	.001	12062	165.95	13	99.6

Table 6: Segmentwise statistics for fatalities of Afghan police force and Afghan national troops. The first column lists the segment and for each segment its node number in the tree, the logarithm of the estimated mean, its standard error, the estimated dispersion parameter and its standard error, the degrees of freedom, the residual deviance, the highest number of fatalities reported and the percentage of reports with zero fatalities.

The first two segments that are split off contain incidents associated with Topic 71 "Afghan National Police" which is described in more detail in Section 4.1. It refers to incidents associated with the ANP. The two segments result according to the regions the incident has happened in, `RC SOUTH, UNKNOWN` and `RC CAPITAL, RC EAST, RC NORTH`. For segment R_1, containing those $n_1 = 428$ reports that have happened in `RC SOUTH` or an `UNKNOWN` region, we have an estimated mean value of the negative binomial of $\mu_1 = 0.783$. The highest reported number is 19 and 72% of incidents have no Afghan police or military fatalities listed. Segment R_2 contains those $n_2 = 848$ incidents happening in the other regions, namely `RC CAPITAL`, `RC EAST`, `RC NORTH` and `RC WEST`. This segment has a mean value that is roughly half of the former one, $\mu_2 = 0.401$. The maximum number of fatalities is 12 and in 84% of the cases incidents were void of Afghan forces fatalities.

The Topics 14 "ACF Attacks & Subsequent Fights" and 61 "Suicide Bombing" give rise to the next two segments. Segment R_3 arises from Topic 14 which is characterized by attacks by the ACF and fights that followed (see Section 4.1). It has an estimated fatality rate of $\mu_3 = 0.511$. 21 is the highest number of deaths in this segment and 77.2% of the reports mention no death toll.

Topic 61 "Suicide Bombing" governs segment R_4. It has already been discussed in detail in Section 4.1. This segment's mean is estimated to be $\mu_4 = 0.658$. In 84.2% of the cases no fatalities of Afghan forces are reported. Of incidents for which there were any, the maximum is 24.

Of the remaining incidents, those 12063 flagged as `green` (neutral) operations constitute segment R_9 with the average rate of $\mu_9 = 0.008$ deaths per incident. Here the maximum number of deaths reported is 13 and in 99.6% of the cases no deaths are listed.

For those flagged `red` (ACF action) or `blue` (ISAF action), four more segments emerged, depending on Topic 94 "Medical Topic (Fights)" and Topic 30 "Village Attacks". Those incidents related to neither topic, define the huge segment R_8 with $n_8 = 61165$ (79.5% of the war logs) and an average fatality rate of $\mu_8 = 0.034$ per incident. The maximum number of fatalities for host nation troops is 27 in this segment which also is the highest overall. Additionally, 98.4% of reports mention no fatalities.

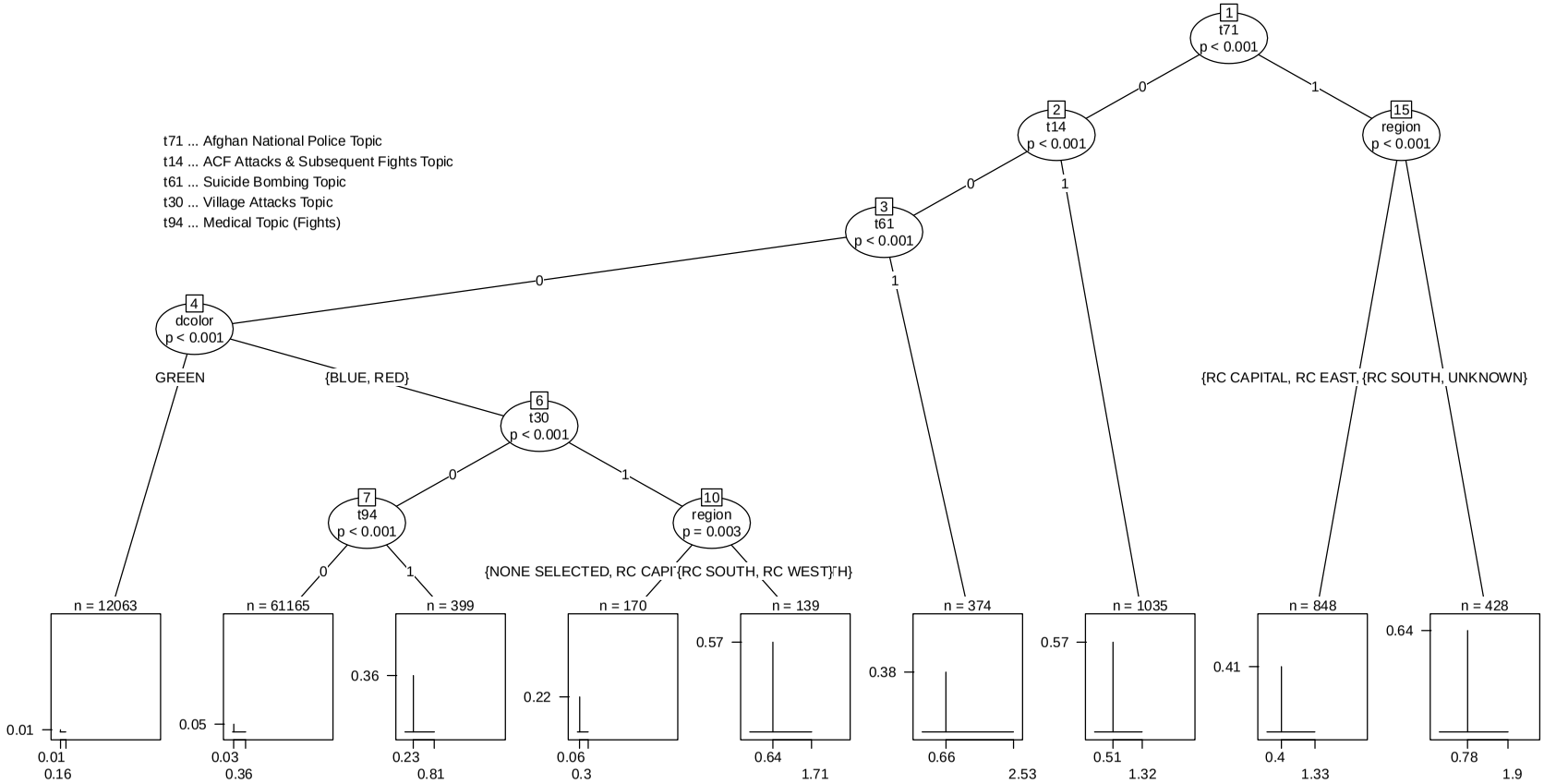

Figure 7: The manifest negative binomial mixture model tree for fatalities for Afghan police and military. In the terminal nodes the vertical line marks the mean, the horizontal line the length between zero and one standard deviation and the height of the vertical line is deviance/df.

The $n_7 = 399$ incidents that belong to Topic 94 "Medical Topic (fights)" constitute segment R_7 with $\mu_7 = 0.233$. This topic is categorized as medical help for fighting units (see 4.4). This segment's fatality maximum for the host nation forces is 11 and 87.5% of the reports mentioned no deaths.

Splitting according to Topic 30 "Village Attacks" leads to two more segments, once more differing by region, R_6 and R_5. The most frequent terms of topic 30 are `village`, `acm` (anti coalition militia), `attack`, `anp`, `taliban`. Segment R_5, with mean $\mu_5 = 0.64$, includes incidents from the south and west, whereas its sister segment R_6 with incidents taking place elsewhere has only about 1/10 of the fatalities per incident compared to R_5 ($\mu_6 = 0.065$). Additionally, the latter (R_6) has as a maximum death toll of only 2, while the former has 15 listed. Furthermore, R_6 consists to 94.7 of reports that mention no fatalities of the Afghan forces. Incidents in R_5 do not list fatalities in only 76.3% of the cases. The 15 victims in segment R_5 were bodyguards assigned to protect an Afghan province governor (03-May-2004):

> bodyguards assigned to protect Zabol province governor killed by Taliban militia during ambush: as many as 15 bodyguards assigned to protect Zabol Province governor Mohammad Hosayni Khial were killed on 3 May 2004 in an ambush in Shajoy (3231N 06725E). Governor Khial, who was traveling in a separate vehicle ahead of his bodyguards, was not involved in the attack.

5 Summary And Discussion

Here we summarize and highlight some of the results described in Section 4 and embed them into a broader context.

Fatalities Of Civilians: When looking at the results for civilian fatalities we can see a clear domination of topics as important explanatory variables. We can identify 11 types of situations for which different mortality rates were found. Those situations are often not surprising. For example suicide attacks figure prominently among them (Topic 61, segment R_3). This segment has the second highest mortality rate for civilians (1.1 per incident), only surpassed by mortality in actions of the ACF against civilians or where civilians were "collateral damage" (segment R_1, 1.7 deaths per incident). Both segments have in common that the attacks have been overwhelmingly carried out by the ACF and have been directed at places where there is a high number of the civilian population present, such as busses, bazars or markets. Segment R_1 has a mere 50.3% of reports that listed no civilian fatalities. This is by far the lowest percentage of reports without deaths and accounts for the high mortality. In contrast, for incidents in R_2 which contains actions of ISAF troops also belonging to Topic 85 "Civilian Casualties", we have about 25% of the former rate (0.41 deaths per incident in R_2, the fourth highest overall rate for civilians). Thus ACF action is associated with a fourfold increase in the average number of civilian fatalities for reports belonging to Topic 85 "Civilian Casualties". It is a clear pattern that actions of the ACF come along with a higher civilian death toll than actions of the allied forces.

The other segments corresponding to independent topics are quite clearly attributable to specific circumstances and we refer to Section 4.2. One thing that all of these segments (R_4 through R_8 and R_{10}) have in common is that they are again mostly connected to attacks by the ACF often with improvised explosive devices. The lowest mortality rate of 0.02 civilian deaths per incident (or 1 every 50 incidents) can be found in segment R_{14} which is the segment that includes all incidents that do not belong to any of the 11 independent topics. This segment contains about 88.1% of all logs with a reported number of civilian deaths. Given the circumstances, this can be seen as positive since by far most of the every day happenings in this war come along with a

low civilian death toll. Only in case of certain events this number increases and these events are mainly connected to the Taliban and other ACF groups who aim at or tolerate civilian casualties.

Fatalities Of Anti-Coalition Forces: Here topics still play an important explanatory role, but compared to fatalities of other groups they are not as prominent. There are other variables that are selected for the model, namely the region the incident happened in, if the report described a complex attack, at whom the attack was directed (`attackOn`) and whose action it was (`dcolor`). The topics that are relevant usually give rise to a single segment, only Topic 27 "Contacts ACF vs. TF" is further split based on the region the incidents happened in. The highest death toll of ACF is observed in segment R_3 with 2.4 fatalities per incident on average. This segment contains incidents that can be described as individual combat, fire fights with small arms by ground troops and they all happened in the south of Afghanistan (especially around Kandahar and in Helmand). This segment contains - among others - events from Canadian-led "Operation Medusa", which began on September 2, 2006 and lasted until September 17 (see Wikipedia, 2010). For incidents that are topic-wise equivalent but happened somewhere else, the mean death rate drops to 0.6 (R_2). This can be explained by the South - especially the province of Kandahar - being Taliban heartland and their stronghold. It is therefore heavily attacked by coalition troops (see O'Loughlin et al., 2010). This pattern of high death rates for incidents happening in the South is recurrent for all groups of fatalities. Overall incidents in this two segments are by far the most deadly for fighters of the ACF.

Another very prominent segment describes actions connected to task force "Bushmaster" (incidents collected in Topic 5). They have a comparably high average fatality number of 2.1 per incident. TF "Bushmaster" is a task force consisting of Afghans and American green beret soldiers, the latter being a synonym for the United States Army Special Forces. According to Wikipedia (2011b) they have "six primary missions: unconventional warfare, foreign internal defense, special reconnaissance, direct action, hostage rescue, and counter-terrorism. The first two emphasize language, cultural, and training skills in working with foreign troops. Other duties include combat search and rescue (CSAR), security assistance, peacekeeping, humanitarian assistance, humanitarian de-mining, counter-proliferation, psychological operations, manhunts, and counter-drug operations."

When looking at incidents that do not belong to the segments R_4, R_5 or the ones described, those happening in the North and South form a huge segment (30035 incidents) with on average 0.2 deaths of ACF fighters. Of those not happening in the South and North, the fact that the event has been planned to be a complex attack is important. If it has been one, attacks on allied or neutral targets have - on average - 62.9% of ACF fatalities compared to similar incidents directed at ACF or neutral targets. Thus, from the coalition forces point of view, defending (which often might come along with a withdrawal) is associated with less ACF fatalities than attacks on ACF. The lowest ACF death toll however is in segment R_{10} with around one death per 100 incidents. This segment are those 12753 incidents that happened in the East or in unknown regions, are not part of the other segments mentioned so far and are characterized to be non-complex attacks flagged as being actions of allied forces or the host nation (blue and green). If the flag labeled the event as being an ACF action (red), the death rate increased sevenfold (7 in 100).

Fatalities Of Allied Forces: Generally, the number of fatalities of coalition troops are the lowest of all fatality groups we looked at. Please note, however, that the war logs contain no ISAF or UN or top-secret operations. The highest average death rate is roughly 0.4 fatalities per incident in Segment R_3. This segment describes incidents that we have identified as ACF attacks with subsequent fights (Topic 14) that have not been complex attacks. This means that complex attacks are associated with lower fatality numbers for coalition troops (R_1 and R_2). This is interesting in comparison to ACF deaths, where non-complex attacks are usually associated

with lower fatalities. Hence complex attacks benefit coalition forces for this topic.

It is however not possible to generally call complex attacks safer for coalition troops, as segments R_6 through R_{12} show. R_6 ($n_6 = 39018$) contains all incidents that have not been complex attacks and in R_1 through R_5. On average, there is one fatality per 100 incidents. Segments R_{12} contains another vast majority of observations (about as many as R_6) and both have a similar mortality rate of 0.02 per incident. R_{12} however collects incidents that have not been complex attacks. The same holds for the weighted sum of fatality rates in R_7 through R_{11} which is actually very close to that too.

The lowest death toll is 0.007 per incident in segment R_8. Once again we can see that the South and West are associated with a higher fatality rate.

Fatalities Of Afghan Military And Police: For this group we see a repetition of some of the patterns already discussed. There is an especially high congruency with the patterns found for civilian fatalities. The bloodiest segment R_1 has on average 0.78 fatalities reported per accident. The topic of those incidents suggests involvement of the Afghan National Police (ANP) often describing attacks on ANP checkpoints or police stations. Not surprisingly R_1 is associated with `RC SOUTH`. We therefore can see once again that incidents in the South have on average a higher fatality rate than those in other parts, a pattern repeated when comparing segments R_5 (South and West) and R_6 (other regions). Two further topics stick out when looking at the results for police and military of Afghanistan: Topics 61 "Suicide Bombing" and Topic 14 "ACF Attacks & Subsequent Fights". These topics are also very important for civilian fatalities and fatalities of allied forces. More or less they refer to actions of the ACF either with suicide attacks (Topic 61) or other attacks on or at non-military targets such as markets, highways or busses (Topic 14). These attacks often come along with a high fatality rate of civilians and Afghan troops. We discussed these topics already when looking at civilian fatalities. Topic 30 "Village Attacks" is somewhat similar to these topics but only plays a role for fatalities of the Afghan troops. It leads to the mentioned two segments R_5 and R_6 with "Southern segment" R_5 having a mean of 0.65 deaths per incident. Once again, the majority of reports are clustered in segments that have a comparatively low fatality rate (R_8 and R_9) with $\mu_8 = 0.03$ and $\mu_9 = 0.01$.

All Fatalities: The analysis of all fatalities combined shows that the resulting tree is dominated by the fatalities of the ACF and those of the civilian population. This can be seen for example by the fact that the first three segments ("TF Bushmaster Topic" as well as "Contacts ACF vs TF Topic" in the South and elsewhere) in Figure 4.1 are the same as those in Figure 4.3. Segment R_5 for the ACF fatalities and segment R_6 for all fatalities are the same as well ("Seek, Watch & Destroy Topic"). Hence, those segments for the combined fatalities are dominated by ACF fatalities. Furthermore, the split according to region (inner node 6) only appears in the tree for ACF fatalities in the same fashion.

A similar picture can be found for segments R_4 and R_5 of all fatalities and their correspondence to segments R_3 and R_4 of civilian fatalities ("Suicide Bombing Topic" and "ACF Attacks & Subsequent Fights Topic"). Furthermore, the splits after the mentioned regional split due to ACF fatalities are then dominated by topics and segments appearing in the tree for civilian fatalities or ACF fatalities, namely segments governed by Topics 85 ("Civilian Casualties Topic") and 71 ("Afghan National Police Topic") and the ACF split topic 18 ("Battle Damage Assessment Topic"). Fatalities of allied forces and the troops of the host nation play a minor role for the overall number of deaths due to the comparatively small number of those fatalities (especially of allied forces) and the high congruency of civilian deaths and deaths of host nation troops. Because of this, the points made in the paragraphs on civilian and ACF fatalities are also mostly accurate for the overall number of fatalities in this war. It should also be noted (and that pattern is consistent throughout all the fatality groups) that those segments that contain by far the largest number of

reports have on average relatively low death rates per incident.

6 Conclusion And Outlook

Undoubtedly, innovations like the internet have changed the supply of potential data of interest for science as well as journalism and it is unavoidable to gather, manage and process this bulk of information. Central to this is "understanding" and interpreting written text documents with an automated procedure. The increase of available written information, e.g. in the world wide web, will increase the need for such methods. At least partly, this has nourished data journalism where the database becomes the center of journalistic work. The present effort tried to illustrate how modern statistical procedures can aid in extracting relevant information from bulk of written text documents or from a database and how they may help in processing and structuring this information to facilitate interpretation of the data.

In this paper, text mining and topic models (LDA) were used to analyze written text automatically, by assigning overarching themes to the single documents. This allowed to get a view on the data which is hard to obtain by manual processing and may find connections between documents which may not be at all obvious. The assignment of topics to the single documents further offered the opportunity to use those topics as explanatory variables in subsequent data analysis. One has to bear in mind, however, that the assignment of topics to documents is by far not absolute. At any rate, we saw that explanatory variables generated by LDA preprocessing proofed to be very important in subsequent modeling, whereas the variables that were already available played a minor role. Hence, discarding the information stored in the report summaries would have led to completely different models or interpretation.

Model based trees were then used to model the data flexibly and non-linearly as well as providing an intuitive interpretation. A representative model (here the negative binomial distribution) was formulated to answer a theory-driven research question. Instead of simply calculating the arithmetic mean of the dependent variable, the underlying model takes a whole likelihood for overdispersed count data, suitable for describing rare events, into account when estimating the mean fatality rates. Pre-pruning of the trees with an inferential splitting procedure helped to build a tree that does not overfit our data by becoming too branchy but retains useful explanatory power (judging by how clear topics could be named and how the segments could be connected to specific events reported in the media) as well as fit the data at hand very well. The model-based approach we chose offered additional insight as to how the mortality rate for specific incidents looked like, something that has not been done so far for this war. Generally, we think model based recursive partitioning offers a wide range for research in socio-economic sciences (see Zeileis et al., 2008; Kopf et al., 2010; Rusch and Zeileis, 2011).

We think that our approach works very well for but is not limited to data journalism, where the data consist of both statistical variables and written text which has to be analyzed. The recursive partitioning framework helps to find smaller groups of observations based on the information generated from the text to whom certain structural similiarities (such as rates, counts, frequencies) apply. This is equivalent to looking at a collection of local models in segments of the data that might provide better explanation and fit than an overarching global model would. Each segments can then be described by itself which may be very useful for journalistic work.

In the future we also want to analyze the Iraq war logs, possibly by adapting the approach presented here. The Iraq war log is even more challenging as there are over 300 000 records to be processed. We are confident though, that modern statistical techniques in modern statistical packages like R combined with cloud computing as well as parallel computing infrastructure will lead to interesting insights beyond anything that was possible before.

Computational Details: All calculations have been carried out with the statistical software R 2.12.0-2.13.1 (R Development Core Team, 2011) on `cluster@WU` (FIRM, 2011). Topic models were estimated with the extension package **topicmodels** 0.0-7 (Grün and Hornik, 2011). Further packages used were **slam** 0.1-18 and **tm** 0.5-4.1. Recursive partitioning infrastructure was provided by the function `mob()` (Zeileis et al., 2008) from the package **party** 0.9-99991. Further packages used were **strucchange** 1.4-3. The negative binomial family of models used for `mob()` was based on the implementation of `glm.nb()` in package `MASS` 7.3-7 (Venables and Ripley, 2002). The code can be obtained from the corresponding author until it is included in a version of **party**.

Acknowledgment: The authors want to thank Bettina Grün and Achim Zeileis for useful discussions and expert advice.

A Terms Of The Latent Topics

	Topic 5	Topic 11	Topic 12	Topic 14	Topic 16	Topic 18	Topic 19	Topic 21	Topic 27	Topic 29
numberDOC	830	1062	484	1035	533	508	900	498	2382	2287
ALL	x			x		x	x		x	
CIV		x		x	x			x		x
ACF	x		x			x	x		x	
FRIEND				x						x
HOST				x						
	tf	explos	fire	wia	convoy	engag	updat	anp	fire	medevac
	bushmast	ie	mm	ie	vehicl	bda	att	polic	enemi	request
	fire	vehicl	cop	kia	ambush	ground	aaf	chief	contact	pt
	forc	deton	attack	strike	damag	damag	pax	district	tf	wd
	isaf	damag	round	bda	hwi	mm	saf	aup	tic	wu
	close	tf	wb	cat	close	fire	event	tf	element	tf
	track	injuri	icom	medevac	event	ah	contact	ie	acm	patient
	friend	convoy	observ	struck	casualti	compound	station	wolfpack	arm	approv
	insurg	blast	eagl	vehicl	kandahar	kill	vc	offic	receiv	soldier
	event	strike	zerok	isaf	compass	pid	fire	kandahar	saf	mc

Table 7: The 10 most frequent terms of the estimated latent topics. To be continued in Table 8. The first five rows indicate whether the topics served as a split (`x`) or not.

	Topic 30	Topic 34	Topic 57	Topic 61	Topic 71	Topic 74	Topic 79	Topic 85	Topic 86	Topic 94
numberDOC	482	472	347	378	1288	320	442	638	407	443
ALL				x	x			x		
CIV			x	x	x		x	x	x	
ACF										
FRIEND		x		x		x				x
HOST	x			x	x					x
	villag	baf	district	suicid	anp	afghan	truck	ln	nds	cat
	acm	tf	attack	bomber	cp	forc	jingl	wound	jan	wound
	attack	land	polic	deton	attack	border	driver	local	arrest	action
	anp	site	wazi	vest	event	afghanistan	vehicl	civilian	abdul	wia
	local	aircraft	kwah	attack	close	nation	fuel	hospit	mohammad	medevac
	inform	ac	gerda	nds	qrf	coalit	burn	kill	mullah	bsn
	taliban	crash	road	khowst	isaf	releas	attack	injur	offic	result
	jm	ch	serai	explos	checkpoint	inform	secur	lns	khan	gbr
	individu	eagl	incid	svbi	ie	airfield	convoy	child	attack	mms
	activ	air	taliban	kill	wia	press	road	nation	name	ff

Table 8: The 10 most frequent terms of the estimated latent topics (continued).

B Acronyms

Acronym	Meaning
aaf	Anti Afghan Forces
ac	aircraft (a/c)
acm	Anti-Coalition Militia
ah	attack helicopter(ah-1w)
ak	assault rifle (ak-47)
anp	Afghan National Police
ansf	Afghan National Security Forces
att	At tis time
aup	Afghan uniform police
baf	Bagram Air Field
bda	battle damage assessment
bsn	(Camp) Bastion
cas	Close Airport Support
cat	Category (C) patient - priority
cf	Coalition Forces
ch	chief ? (CHOPS– chief of operation)
cop	COP Combat outpost, Chief of police (CoP)?
ff	Friendly Forces
fob	Forward Operating Base
gbu	Guided Bomb Unit
gerda	gerda serai
hvi	high value individual
hwy	highway
icom	radio
ie(d)	improvised explosive device
isaf	International Security Assistance Force
jingle (truck)	Brightly decorated trucks covered in bells common across central asia
jm	joint mission??
khowst	Khowst Province (City)
kia	killed in action
km(tc)	kabul military (training center)??
kwah	wazi kwah (province)

Table 9: Glossary of military terms. To be continued in Table 10.

Acronym	Meaning
ln	local national
lns	local nationals
medevac	Medical evacuation
mc	Medical ??
mm	Military Message
mms	Military Messages
nds	Afghan intelligence
pax	passenger, people
pid	positive i.d.
pkm	Russian-made machine gun
pt	Patient
qrf	Quick Response Force
rpg	Rocket propelled grenade
saf	small ams fire
serai	Gerda serai
sied	Suicide ied
svbi	suicide vehicle-borne IED
tf	Task Force
tic	Troops in Contact
vc	vehicle check
wazi	Wazi Kwah district
wb	???wheels broken
w/d	wheels down
wia	wounded in action
wu	wheels up
zerok	Location in Paktika state

Table 10: Glossary of military terms (continued).

References

Aitkin, M., Francis, B., Hinde, J., and Darnell, R. (2009). *Statistical Modelling in R*. Oxford University Press, New York.

Amnesty International (2009). Afghanistan: German Government must investigate deadly Kunduz Airstrikes. `http://www.amnesty.org/en/news-and-updates/news/afghanistan-german-government-must-investigate-deadly-kunduz-airstrikes-20091030`. [Online; accessed 07-March-2011].

Barnett, A., Stanley, T., and Shore, M. (1992). America Vietnam casualties - victims of a class war. *Operations Research*, 40:856–866.

Bhutta, Z. A. (2002). Children of war: the real casualties of the Afghan conflict. *British Medical Journal*, 324:349–352.

Bird, S. and Fairweather, C. B. (2007). Military fatality rates (by cause) in Afghanistan and Iraq: a measure of hostilities. *International Journal of Epidemiology*, 36:841–846.

Blei, D. (2011). Introduction to probabilistic topic models. *Communications of the ACM*, in press. Online available at `http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf`[accessed 07-Sep-2011].

Blei, D. M., Jordan, M. I., and Ng, A. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning*, 3:993–1022.

Blei, D. M. and Lafferty, J. D. (2009). Topic models. In Srivastava, A. and Sahami, M., editors, *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Press.

Bortkiewicz, L. (1898). *Das Gesetz der kleinen Zahlen [The law of small numbers]*. B.G. Teubner, Leipzig.

Boyd-Graber, J., Chang, J., Gerrish, S., Wang, C., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*.

Burnham, G., Lafta, R., Doocy, S., and Roberts, L. (2006). Mortality after the 2003 invasion of Iraq a cross-sectional cluster sample survey. *Lancet*, 368:1421–1428.

Buzzell, E. and Preston, S. H. (2007). Mortality of American troops in the Iraq war. *Population and Development Review*, 33(3):555–566.

Choi, Y., Anh, H., and Chen, J. (2005). Regression trees for analysis of count data with extra poisson variation. *Computational Statistics & Data Analysis*, 49:893–915.

Cirillo, V. (2008). Two faces of death - fatalities from disease and combat in America's principal wars, 1775 to present. *Perspectives in Biology and Medicine*, 51:121–133.

CNN (2005). Rescuers try to reach downed helicopter. `http://articles.cnn.com/2005-06-28/world/afghan.crash_1_military-helicopter-helicopter-crashes-medical-evacuation-mission?_s=PM:WORLD`. [Online; accessed 1-March-2011].

Conway, D. (2010a). Benford's law tests for Wikileaks data. `http://www.drewconway.com/zia/?p=2234`. [Online; accessed 07-March-2011].

Conway, D. (2010b). The evolution of report summaries in Wikileaks data over time. `http://www.drewconway.com/zia/?p=2278`. [Online; accessed 07-March-2011].

Conway, D. (2010c). Wikileaks Afghanistan data. `http://www.drewconway.com/zia/?p=2226`. [Online; accessed 07-March-2011].

Conway, D. (2010d). Wikileaks attack data by year and type projected on Afghanistan regional map. `http://www.drewconway.com/zia/?p=2268`. [Online; accessed 07-March-2011].

Deb, P. and Trivedi, P. K. (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12:313–336.

Degomme, O. and Guha-Sapir, D. (2010). Patterns of mortality rates in Darfur conflict. *Lancet*, 375(9711):294–300.

FIRM (2011). Cluster@wu. `http://www.wu.ac.at/firm/cluster_folder`. [Online; accessed 24-March-2011].

Friendly, M. (2001). *Visualizing categorical data*. SAS publishing, Cary, North Carolina.

Garfield, R. M. and Neugut, A. I. (1991). Epidemiologic analysis of warfare - a historical review. *Journal of the American Medical Association*, 266:688–692.

Gebauer, M. (2010). Explosive leaks provide image of war from those fighting it. `http://www.spiegel.de/international/world/0,1518,708314,00.html`. [Online; accessed 07-March-2011].

Gooch, J. (2010). The White War: Life and death on the Italian front 1915-1919. *Journal of Military History*, 74:267–268.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101:5228–5235.

Grün, B. and Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

guardian.co.uk (2010). Afghanistan war logs: 56 civilians killed in Nato bombing. `http://www.guardian.co.uk/world/afghanistan/warlogs/826B488C-EA6F-A132-511610DB68C2EDBD`. [Online; accessed 07-March-2011].

Haushofer, J., Biletzki, A., and Kanwisher, N. (2010). Both sides retaliate in the Israeli-Palestinian conflict. *PNAS*, 107(42):17927–17932.

Hirschman, C., S, P., and Loi, V. M. (1995). Vietnamese casualties during the American war: A new estimate. *Population and Development Review*, 21:783–812.

Holcomb, J. B., McMullin, N. R., Pearse, L., Caruso, J., Wade, C. E., Oetyen-Gerdes, L., Champion, H. R., Lawnick, M., Farr, W., Rodriguez, S., and Butler, F. K. (2007). Causes of death in US special operations forces in the global war on terrorism - 2001-2004. *Annals of Surgery*, 245:986–991.

Hothorn, T., Zeileis, A., and Hornik, K. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Keeley, L. H. (1996). *War Before Civilization: the Myth of the Peaceful Savage*. Oxford University Press, Oxford.

Kim, Y. and Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96:589–604.

Kopf, J., Augustin, T., and Strobl, C. (2010). The Potential of Model-Based Recursive Partitioning in the Social Sciences – Revisiting Ockham's Razor. Technical report, Ludwig-Maximilians University, Munich.

Lakstein, D. and Blumenfeld, A. (2005). Israeli army casualties in the second Palestinian uprising. *Military Medicine*, 170:427–430.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15:209–225.

Leland, A. and Oboroceanu, M.-J. (2010). *American War and Military operations casualties: Lists and Statistics*. Congressional Research Service, Washington.

Lerner, P. (2000). Psychiatry and casualties of war in germany, 1914-18. *Journal of Contemporary History*, 35:13–28.

Marshall, H. and Balfour, T. G. (1838). *Statistical Report on the Sickness, Mortality, & Invaliding among the troops in the West Indies*. W. Clowes and Sons, London.

Nightingale, F. (1863). *Notes on Hospitals*. Longman, Green, Longman, Roberts, and Green, London, 3rd edition.

O'Loughlin, J., Witmer, F. D. W., Linke, A. M., and Thorwardson, N. (2010). Peering into the fog of war: The geography of the Wikileaks Afghanistan war logs, 2004–2009. *Eurasian Geography and Economics*, pages 1–24.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramaswamy, V., Anderson, E. W., and DeSarbo, W. S. (1994). A disaggregate negative binomial regression procedure for count data analysis. *Management Science*, 40:405–417.

Rogers, S. (2010). Wikileaks' Afghanistan war logs: how our datajournalism operation worked. `http://www.guardian.co.uk/news/datablog/2010/jul/27/wikileaks-afghanistan-data-datajournalism`. [Online; accessed 07-March-2011].

Roggio, B. (2009). US, Afghan troops beat back bold enemy assault in eastern Afghanistan. `http://www.longwarjournal.org/archives/2009/10/us_afghan_troops_bea.php`.

Rusch, T. and Zeileis, A. (2011). Gaining insight with recursive partitioning of generalized linear models. Technical Report 109, Research Report Series, Institute for Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna.

Seet, B. and Bunham, G. M. (2000). Fatality trends in United Nations peacekeeping operations, 1948-1998. *Journal of the American Medical Association*, 284:598–603.

Spiegel, P. and Salama, P. (2001). War and mortality in Kosovo, 1998-99: an epidemiological testimony. *Lancet*, 355:2204–2209.

Thomas, T. L., Parker, A. L., Horn, W. G., Mole, D., Spiro, T. R., Hooper, T. I., and Garland, F. C. (2001). Accidents and injuries among US Navy crewmembers during extended submarine patrols, 1997 to 1999. *Military Medicine*, 166:534–540.

Tran, M. (2007). US kills 100 'insurgents' in Afghanistan battle. `http://www.guardian.co.uk/world/2007/aug/29/afghanistan.usa`. [Online; accessed 1-March-2011].

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Wikipedia (2010). Operation Medusa — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Operation_Medusa`. [Online; accessed 20-December-2010].

Wikipedia (2011a). Afghan presidential election, 2009— Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Afghan_presidential_election,_2009`. [Online; accessed 07-March-2011].

Wikipedia (2011b). Special forces (united states army) — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Special_Forces_(United_States_Army)`. [Online; accessed 05-June-2011].

Zeileis, A. and Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, 61:488–508.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492–514.