Modeling Mortality in the Afghanistan War Logs: Combining topic-models and negative binomial recursive partitioning

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

Published: 01/01/2011

Document Version
Unknown

Modeling Mortality in the Afghanistan War Logs: Combining topic-models and negative binomial recursive partitioning

Paul Hofmarcher*,[a], Thomas Rusch [a], Kurt Hornik [a], and Reinhold Hatzinger[a]

[a]*Institute for Statistics and Mathematics, Department of Finance, Accounting and Statistics, Wirtschaftsuniversität Wien, Augasse 2–6, A-1090 Vienna, Austria*

[*]*Contact: E-mail: paul.hofmarcher@wu.ac.at*

GFKL 2011, Frankfurt Aug. 31 – Sep. 2, 2011

Content

Combine statistical methods from text mining and recursive partitioning for modeling fatality rates within this war.

- Literature on modeling mortality.
- Wikileaks War diaries allow to have a ground level look into the Afghanistan Invasion.
- Use topic-models, Latent Dirichlet Allocations (LDA) to extract themes of the written information.
- The single topics are used as additional covariates.
- Perform recursive partitioning to get a pattern of mortality.
- Results and Interpretation.

Literature I

- Marshall (1838) presenting a "Statistical Report on the Sickness, Mortality, & Invaliding among the troops in the West Indies".
- 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 in 1898.
- Seet (2000) look at fatality trends in UN peacekeeping missions from 1948 to 1998.

Literature II

- Haushofer et al. (2010) use vector-autoregressive OLS models to model the temporal dynamic of the Israeli–Palestinian conflict.
- OLoughlin et al. (2010), present an analysis of the spatial dynamics of the conflict in Afghanistan as portrayed in the Wikileaks data.
- Political science blogger Drew Conway (see `http://www.drewconway.com/zia/?p=2278`) provides an analysis of the reports filled over time and a spatial and temporal analysis of deaths.

The WikiLeaks War Diary I

The WikiLeaks Afghanistan war logs contains 76911 ground level reports about fatalities and the surrounding situations in the US led Afghanistan war covering the period from January 2004 to December 2009

- War logs contain thousands of mosaic stones describing events in Afghanistan from the perspective of the US forces.
- War logs themselves provide neither a coherent pattern of the war, nor do they contain any high level view like strategic decisions or a general picture.
- War diary was marked as the 21^{st} century equivalent of the Pentagon Papers of the 1970s released by Daniel Ellsberg.

The WikiLeaks War Diary II

- Each single report contains 32 columns with numerical and factorial variables such as id-numbers, reporting units, date of the mission, geographic location.
- Four columns represent the number of fatalities (*Host, Enemy, Friend, Civilian*).
- Each report contains a *report summary*, a short verbal description of what happened during incident.
- The report summaries tell us the how and why of the mission.

Latent Dirichlet Allocations

We are interested in extracting explanatory information from the reports.

- Assuming that the similarity between reports is reflected in the words contained in the summaries, we can use Latent Dirichlet Allocations (see Blei et al. (2003)).
- LDA is a powerful document generative hierarchical model for clustering words into topics and documents into mixtures of topics.
- Report summaries are preprocessed, i.e., stop words removed, stemming...

The Document Generative LDA Model I

LDA is a hierarchical model, in which each document is modeled as a mixture of a set of topics and each word in the document is chosen from the selected topic specific word distribution.

- Formally words w out of a vocabulary $W = \{w_1, \ldots, w_N\}$,are the basic unit.
- Denoting the latent topics with z_j $j = 1, \ldots, K$, and let K the predefined number of topics.

$$P(w_i) = \sum_{j=1}^{K} P(w_i|z_i = j)P(z_i = j), \tag{1}$$

- $P(w|z)$ is represented via K multinomial distributions ϕ over the words W.

The Document Generative LDA Model II

- $P(z)$ is represented by a set of D multinomial distributions θ over the K latent topics.
- LDA assigns a prior on θ.

$$P(W|\phi,\kappa) = \int P(W|\phi,\theta)P(\theta|\kappa)d\theta, \tag{2}$$

- topic proportions θ are drawn from a K dimensional Dirichlet distribution with parameter κ, i.e. $P(\theta) \sim Dir(\kappa)$.
- Setting κ to small values, e.g., 0.1 LDA allows to assign each single report a unique latent topic.
- Assigned topics are used as additional covariates in the recursive partition framework.

Recursive Partitioning I

- The observed fatalities are denoted by y_i, $i = 1, \ldots, n$, its associated random variable Y_i.
- Set of possible explanatory variables $\mathbf{x_i} = (x_{i1}, \ldots, x_{iP})^T$
- For recursive partitioning we assume the existence of a segmented model $\mathcal{M}_{\mathcal{R}}(Y, \boldsymbol{\vartheta})$ consisting of r segments $R_k, k = 1, \ldots, r$.
- The segments R_k arise from differences due to input variables $x_1, \ldots, x_p$.
- The 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 the whole segmented model over all segments.

Recursive Partitioning II – Regression Tree

- We model the conditional distribution $D(Y|.)$ with a tree like partition function f, i.e., $D(Y|\mathbf{x}) = D(Y|f(x_1, \ldots, x_m))$.
- 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 is the collection (or mixture) of all segment-specific models $\mathcal{M}_{\mathcal{R}}(Y, \boldsymbol{\vartheta})$.
- We assume the conditional distribution $D(Y|\mathbf{x})$ within each segment $R_k, k = 1, \ldots, r$ to be a negative binomial distribution with mean μ_k and dispersion parameter θ_k, i.e

$$P(Y = y; \mu_k, \theta_k, k) = \frac{\Gamma(y + \theta_k)}{\Gamma(\theta_k) y!} \left(\frac{\mu_k}{\mu_k + \theta_k}\right)^y \left(\frac{\theta_k}{\mu_k + \theta_k}\right)^{\theta_k} \quad (3)$$

Recursive Partitioning III – Estimation

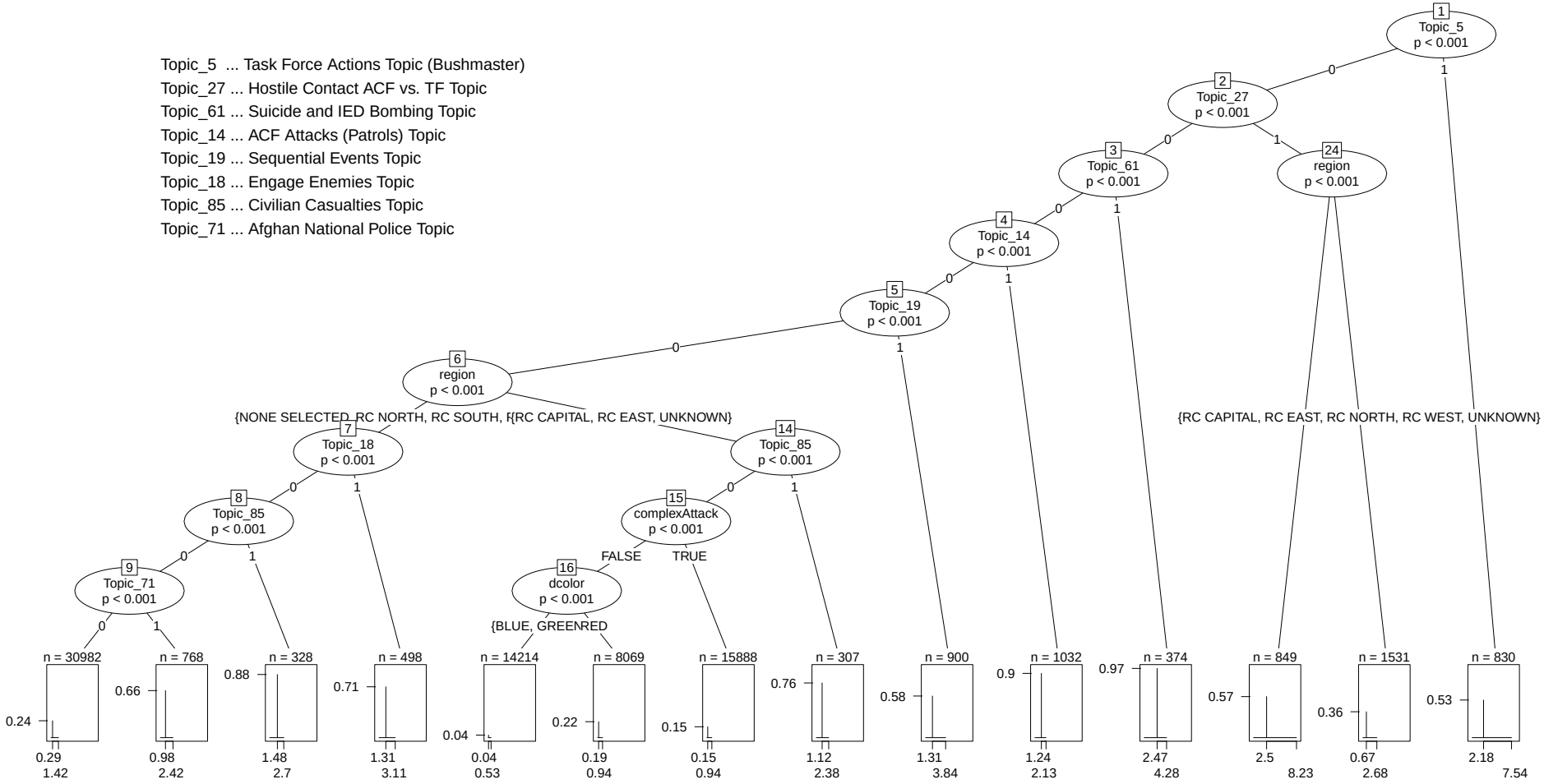
- Fit a negative binomial intercept-only model to all observations in the current node
- Assess instability of the mean parameter estimate $\widehat{\mu}_k$ with respect to the partitioning variables $x_1, \ldots, x_p$
- Choose the covariate associated with the highest instability for splitting
- Compute the binary split that, for all rival partitions, locally optimizes the sum of the partition specific negative log-likelihood functions
- Repeat recursively until no split variables are found or any other stopping criterion is fulfilled

Results

	Topic 5	Topic 18	Topic 61	Topic 85
numberDOC	830	508	378	638
CIV			x	x
ACF	x	x		
FRIEND			x	
HOST			x	
	tf	engag	suicid	ln
	bushmast	bda	bomber	wound
	fire	ground	deton	local
	forc	damag	vest	civilian
	isaf	mm	attack	hospit
	close	fire	nds	kill
	track	ah	khowst	injur
	friend	compound	explos	lns
	insurg	kill	svbi	child
	event	pid	kill	nation

Partition Tree

Topic_5 ... Task Force Actions Topic (Bushmaster)
Topic_27 ... Hostile Contact ACF vs. TF Topic
Topic_61 ... Suicide and IED Bombing Topic
Topic_14 ... ACF Attacks (Patrols) Topic
Topic_19 ... Sequential Events Topic
Topic_18 ... Engage Enemies Topic
Topic_85 ... Civilian Casualties Topic
Topic_71 ... Afghan National Police Topic



Segment TF Bushmaster

- First segment is governed by topic 5 from the LDA approach– *Task Force Bushmaster segment*.
  - Average number of fatalities is $\widehat{\mu} = 2.18$ per report.
  - Maximum number of deaths is 101.
  - In total 1808 fatalities observed within this segment, 1712 were ACF.
  - Clear segment of ACF fatalities, with third highest death rate μ

Segment Suicide Attacks

- Topic 61 (Suicide Attacks) governs segment 4.
  - Serves as splitting variable for Civilian, Allied Soldiers and Host.
  - It has the second highest mean death rate $\mu = 2.47$ and a median death number greater than 0.
  - In this segment we observe 924 deaths whereby 420 are civilian, 246 afghan soldiers and 233 ACF.
  - Next to topic 61, topic 14 "ACF attacks and subsequent fights" is inevitable for civilians.

Segments Civilian Fatalities

- Topic 85 and "region" govern two segments having a clear context to civilian fatalities. Segment 7 and 12.
  - It only serves for civilians as split variable. 81.2% of fatalities within this topic are civilians.
  - The mean deat rate μ differes whether event took place in region East & Capital or not. For East we get 1.12 and 1.31 otherwise.

Interpretation & Discussion

- We see a clear pattern between the fatality rates and the actions performed by allied soldiers respectively ACF.
  - For allied soldiers actions we observe rarely civilian fatalities.
  - ACF actions (suicide bombing) mainly results in civilian fatalities.
- We think that this approach works for e.g., data journalism, where the data consist of both statistical variables and written text.

References I

- Blei, D. M., Jordan, M. I., and Ng, A. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning*, 3:993-1022.
- Bortkiewicz, L. (1898). Das Gesetz der kleinen Zahlen [The law of small numbers].
- Haushofer, J., Biletzki, A., and Kanwisher, N. (2010). Both sides retaliate in the Israeli-Palestinian conflict. *PNAS*, 107(42):17927-17932.
- 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.

References II

- Marshall, H. and Balfour, T. G. (1838). Statistical report on the sickness, mortality, & invaliding among the troops in the West Indies.
- 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.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492-514.

Contact: *paul.hofmarcher@wu.ac.at*