Sandia National Laboratories

Bayesian Modeling of Self-Exciting Marked Point Processes with Missing Histories

PRESENTED BY

John Lewis

Department of Statistical Sciences Sandia National Laboratories

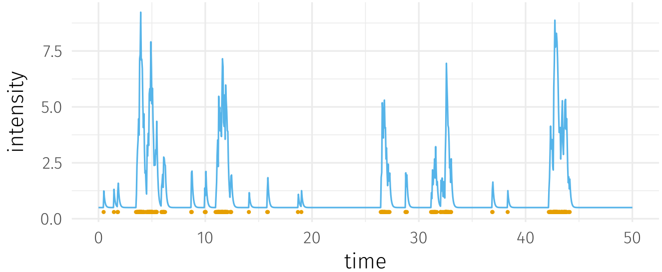
Contributors: J. Derek Tucker (PI), Lyndsay Shand, Stephen Rowe, Jonathan Lane, Kathy Simonson, John Rowe

UNM Statistics Colloquium - March 6, 2018

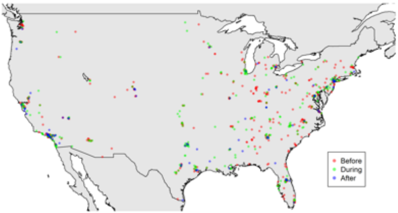
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. SAND NO. SAND2018-2214 PE

Introduction

- Many intelligence and surveillance applications require high degree of human interaction to interpret events
- Remains Difficult
  - increasing volume of data
  - points in time where data is missing
- Can we model a self-exciting process where the history has missing data?
- We develop missing data estimation for the Hawkes process using Bayesian methods



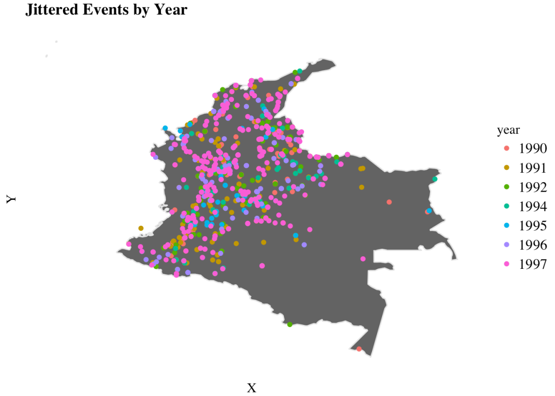
Twitter Examples



Left: Tweets during 2014 World Cup. Right: Tweets before, during, after Paris attacks on 11/13/2015.

Global Terrorism Database Example

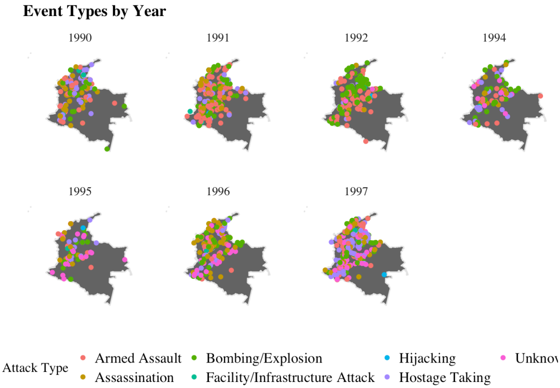
- Global Terrorism Database (2017) Open-source database including information on terrorism events around the world from 1970-2015



Terrorism Events in Colombia 1990-1997. **Notice anything strange about this data?**

Global Terrorism Database Example

- Including Marks - Auxiliary information about an event



Types of Terrorism Events in Colombia 1990-1997.

Self-Exciting Marked Point (Hawkes) Process

- A temporal point process $N(t)$ is characterized by its conditional intensity $\lambda(t)$ (Daley and Vere-Jones 2003),

$$\lambda^*(t) = \lim_{\Delta t \downarrow 0} \frac{E[N(t, t+\Delta t) | \mathcal{H}_t]}{\Delta t}$$

- $\mathcal{H}_t = \{t_i, \kappa_i\}_{t_i < t}$
- κ_i is the mark associated with t_i
- Often takes the form

$$\lambda^*(t) = \mu(t) + \sum_{t_k < t} \alpha(\kappa_k) g(t - t_k, \kappa_k)$$

- $\mu(t)$ is the immigrant intensity
- $\alpha(\kappa_k)$ is the total offspring intensity
- $g(t, \kappa)$ is the normalized offspring intensity

Likelihood

- $x = (t_1, \ldots, t_n)$ observed on $[0, T)$, ϕ - parameters
- $\gamma^*(\kappa|t) = \gamma(\kappa|t, \mathcal{H}_t)$ - marked distribution conditioned on past and current time.

$$p(x|\phi) = \left(\prod_{i=1}^{n} \lambda^*(t_i)\gamma^*(\kappa_i|t_i)\right) \exp(-\Lambda^*(T)),$$

$\Lambda^*(t) = \int_0^t \lambda^*(s)\, ds = M(t) + \sum_{t_k<t} \alpha(\kappa_i) G(t - t_k, \kappa_k),\ M(t) = \int_0^t \mu(s)\, ds$

- Bayesian inference: specify priors (proper), standard Gibbs sampling

Branching Structure

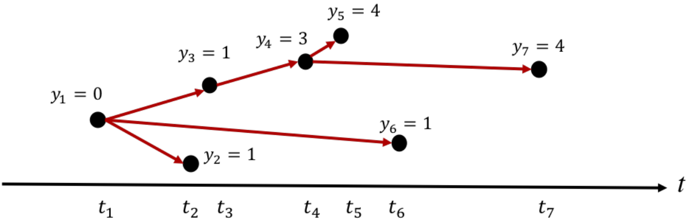
- Rasmussen (2013), Ross (2016)

Branching Structure

1. The parents I follow a Poisson process with intensity $\mu(t)$
2. Each parent $t_i \in I$ generates an independent cluster C_i
3. C_i consists of points of generation $n = 0, 1, \ldots$:
    - Generation 0 is just $\{t_i, \kappa_i\}$
    - Each point in generation n generates points in generation $n + 1$
    - $\{t_j, \kappa_j\}$ in generation n generates a marked Poisson process O_j with intensity $\alpha(\kappa_j)g(t - t_j, \kappa_j)$
4. The process is the union of all the clusters.

Branching Structure

- Let $Y = \{y_i\}$ be the branching structure
- $y_i = j$ if t_i is a child of t_j and $y_i = 0$ if t_i is an immigrant
- $S_j = \{t_i; y_i = j\}$ - set of children of t_j



Depiction of the branching structure.

Branching Structure Likelihood

- Speeds up likelihood computation
- Separates dataset into independent Poisson processes
  - Marked Immigrants with intensity $\mu(t)$
  - Marked Children of t_j with intensities $\alpha(\kappa_j)g(t - t_j, \kappa_j)$

$$p(x|\phi, Y) = p(I|\phi, Y) \prod_{j=1}^{n} p(O_j|\phi, Y)$$

$$p(I|\phi, Y) = \exp(-M(T)) \prod_{t_i \in I} \mu(t_i)\gamma_I^*(\kappa_i|t_i)$$

$$p(O_j|\phi, Y) = \exp(-\alpha(\kappa_j)G(T - t_j, \kappa_j)) \prod_{t_i \in O_j} \alpha(\kappa_j)g(t_i - t_j, \kappa_j)\gamma_O^*(\kappa_j|t_j)$$

Gibbs Step for Y

- Uniform prior on Y
- Make use of stochastic declustering (Zhuang, Ogata, and Vere-Jones 2002, Ross (2016))

$$p(Y_i = j | x, \phi) = \begin{cases} \frac{\mu}{\lambda(t_i)} & \text{if } j = 0 \\ \frac{\alpha g(t_j - t_i)}{\lambda(t_i)} & \text{if } j \in 1, 2, \ldots, i-1. \end{cases}$$

Bayesian Missing Data Augmentation

- Observe points t_{obs} on $[0, T_1] \cup [T_2, T)$
- Want $\pi(\phi | t_{obs})$
- Augment t_{obs} with missing data t_{miss}
- Gibbs Sampling
  1. $p(\phi | t_{obs}, t_{miss})$ - the full posterior
  2. $p(t_{miss} | \phi, t_{obs})$ - use MH
- Tanner and Wong (1987)

MH Step: $p(t_{miss}|\phi, t_{obs})$

- x_T - data up to time T
- Proposal: $p(t_{miss}|\phi, x_{T_1}) \propto p(x_{T_2}|\phi)$
- Sample t_{miss} using thinning method developed by (Ogata 1981)
- Target: $p(t_{miss}|\phi, t_{obs}) \propto p(t_{miss}, t_{obs}|\phi)$
- MH ratio can be computed.

Simulation

- Temporal Model with independent categorical marks.

$$\lambda(t) = \mu + \alpha \sum_{k:t_i<t} g(t - t_i)\gamma^*(\kappa_i|t_i) \quad (1)$$

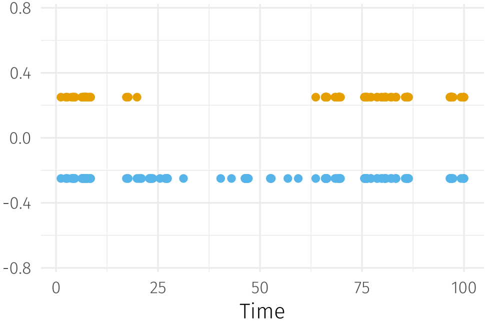
- $\kappa_i \in \{1, 2, \ldots, K\}$ is the categorical mark for event i.
- $m_k = \sum I(\kappa_i = k)$

$$m_1, \ldots, m_K \sim \text{mulitnomial}(n, p_1, \ldots, p_K)$$

- $p_1, \ldots, p_K \sim \text{Dirichlet}(\eta_1, \ldots, \eta_K)$
- $g(t) = \beta \exp(-\beta t)$

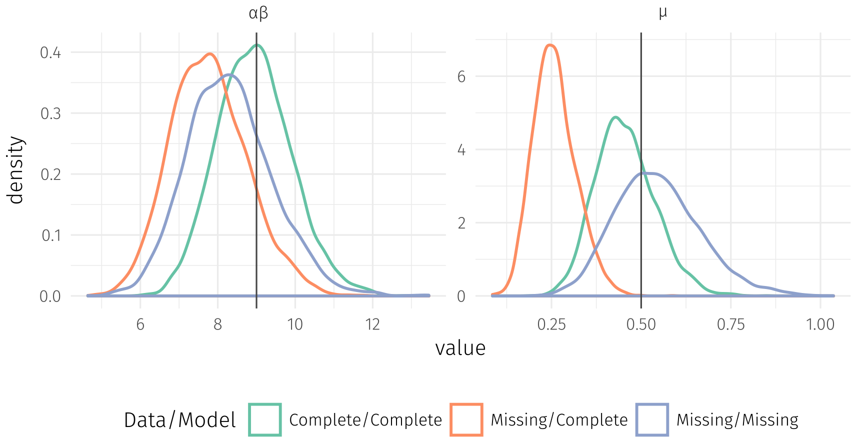
Simulated Data

- $\mu = .5, \alpha = 0.9$, and $\beta = 10$, time interval $(0, 100)$



- Arrival times for full data (blue) and arrival times with missing data (orange)

Parameter Estimates



- Missing data model: Covers true parameters, wider posteriors.

Stochastic Code Verification

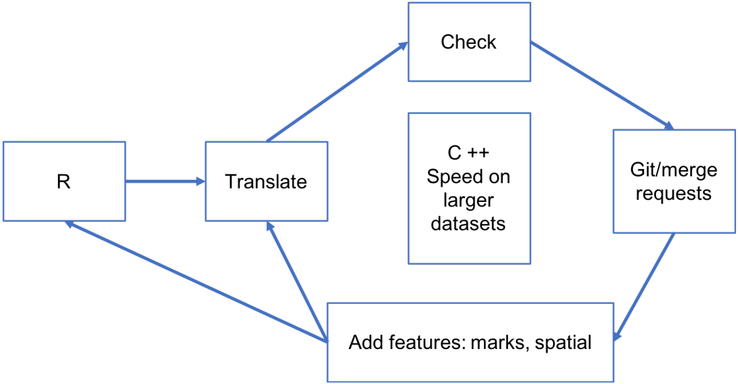


Figure: MCMC Code Development Process.

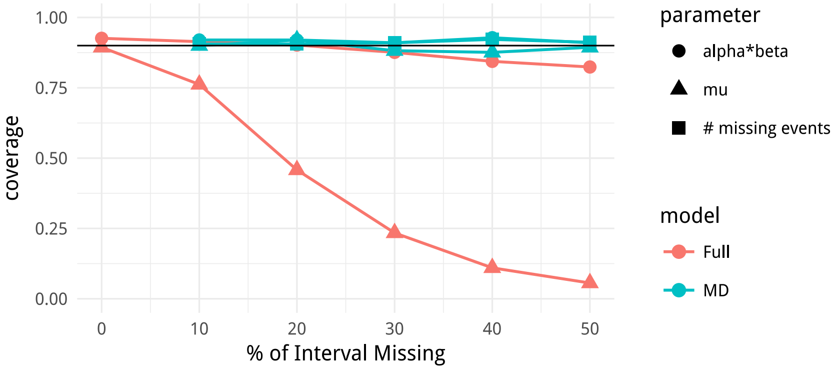
-How do we ensure we are fitting the correct model?

Stochastic Code Verification

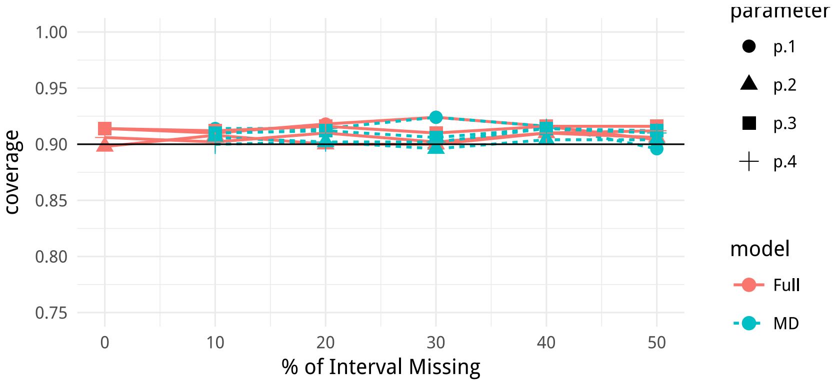
- How do we ensure we are fitting the correct model?
  - Single simulated dataset with fixed parameters - do you recover the parameters?
  - Multiple simulated datasets - frequentist coverage rates?
- Using theory: Cook, Gelman, and Rubin (2006)
  - $\phi_{true} \sim \pi(\phi)$
  - $x \sim p(x|\phi)$
  - If the code is correct, then the posterior quantiles of ϕ_{true} are uniform.

Frequentist Coverage

- $\mu \sim \Gamma(5, 10), \alpha \sim \Gamma(4.8, 12), \beta \sim \Gamma(20, 10), \eta_k = 1, K = 4$
- Time interval: (0,100), 500 simulated datasets.



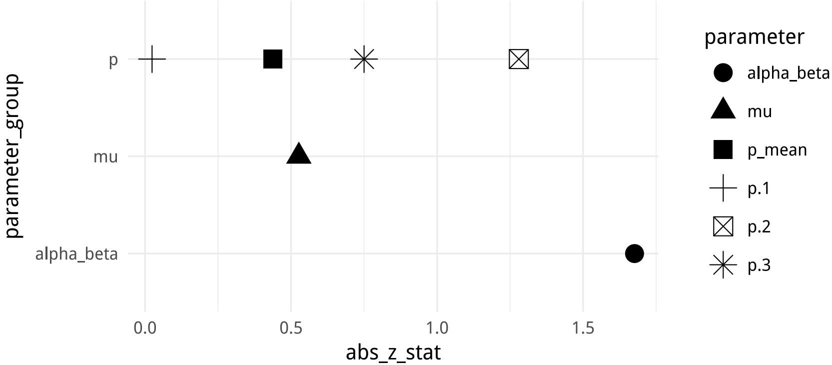
Frequentist Coverage



Distribution of Quantiles of True Parameters

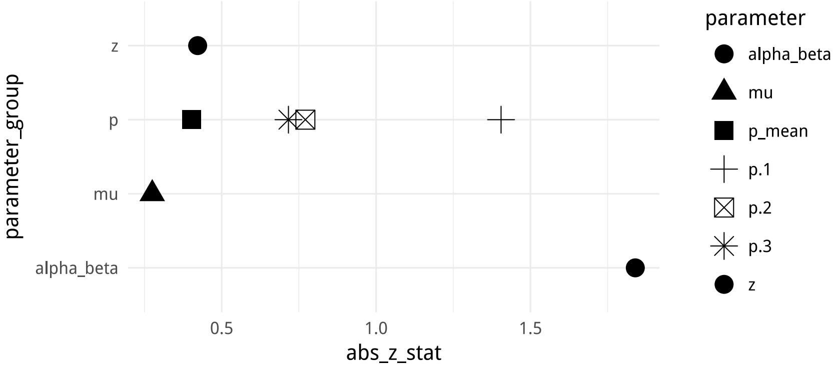
- $q_{true} = \int_{-\infty}^{\phi_{true}} \pi(\phi|x) d\phi$
- Estimated with MCMC samples using empirical quantiles
- Many parameters - test each q_{true}?
- Recommendation
  1. Group similar parameters together
  2. For each group - form a new parameter equal to the mean of the group
  3. For each group compute p-value p of chi-square test: $X^2 = \sum_i^N \Phi^{-1}(q_{true}^i) \sim \chi_N^2$
  4. Perform Bonferroni correction on p-values - multiply by number of groups.
  5. Plot z-transformed p-values $\left|\Phi^{-1}(p)\right|$ to identify extremes.

No Missing Data



50% of Interval Missing

- z - number of missing observations



Spatial-Temporal Model

- $\kappa = (\kappa_x, \kappa_y)$ - spatial locations viewed as marks
- Similar to ETAS model of Ogata and Zhuang (2006)

$$\lambda(t) = \mu + \sum_{k:t_k<t} \alpha g(t - t_k)\gamma(\kappa_k | t_k, \{t_{pa}, \kappa_{pa}\}) \tag{2}$$

$$\gamma(\kappa | t_k, \{t_{pa}, \kappa_{pa}\}) = \frac{1}{2\pi\sigma^2} \exp\left\{\frac{-||\kappa - \kappa_{pa}||^2}{2\sigma^2}\right\},$$

- Spatial location of offspring centered at parent location with Gaussian decay.
- Parents - homogeneous Poisson process.
- $\sigma^2 \sim IG(a, b)$

Simulated Data

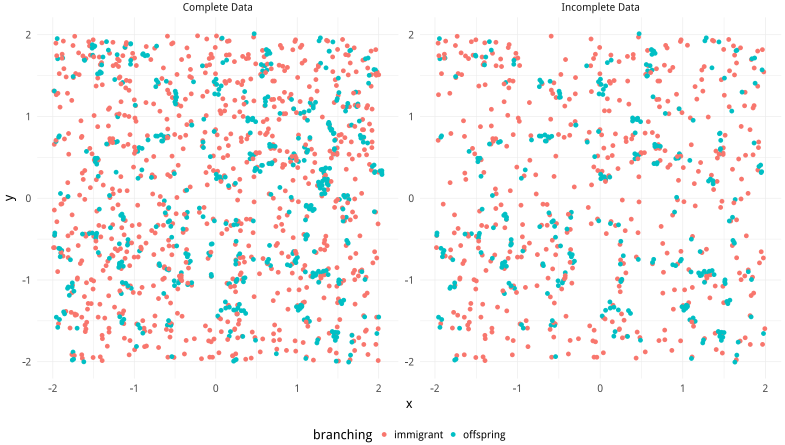
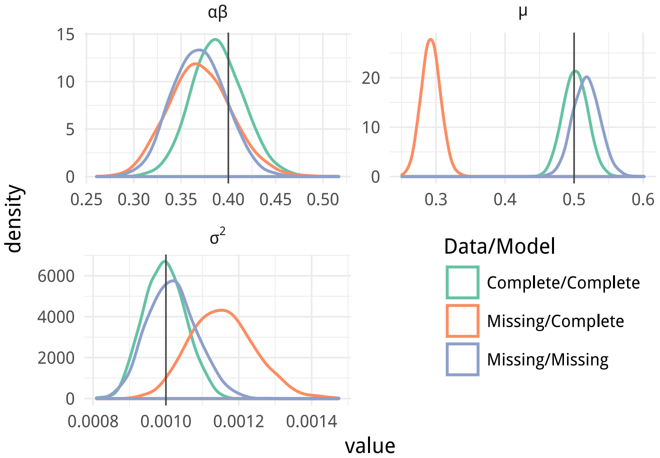


Figure: complete data, missing data on (0,20).

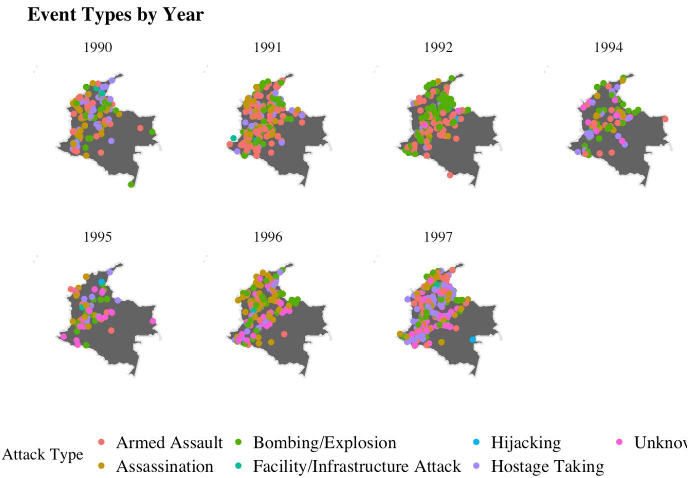
Parameter Estimates



Global Terrorism Database

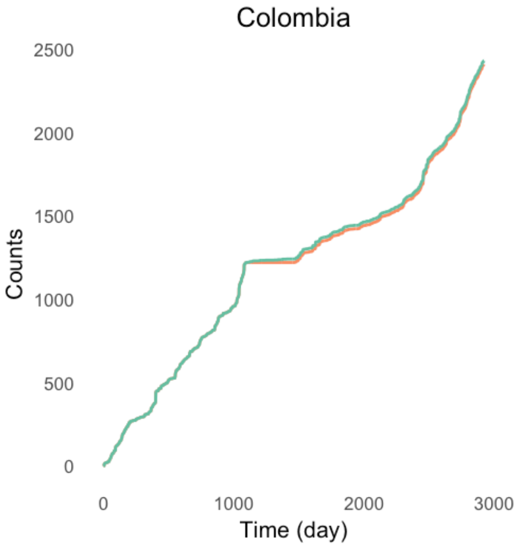
- The Global Terrorism Database is an open-source database including information on terrorism events around the world from 1970-2015
- We will look at the year span between 1990-1997 in Colombia which had multiple problems with guerrillas, paramilitaries, and narcotics
- The database is missing records for the entire year of 1993
- Records were 'partially' recovered (21 events)

Terrorism Events in Colombia

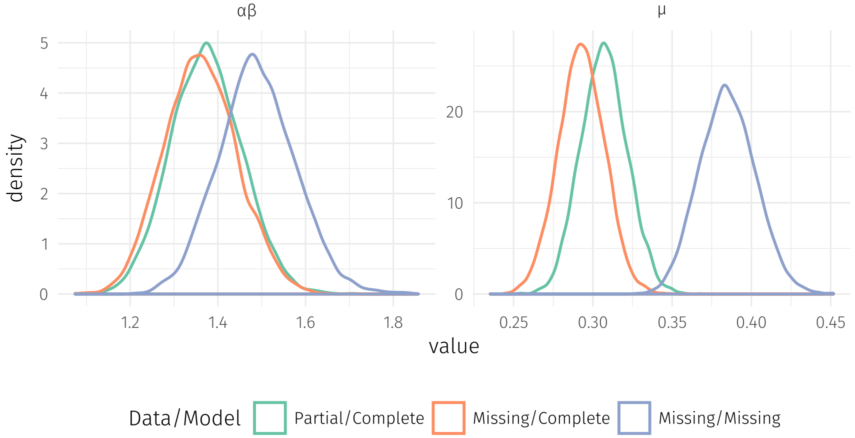


- Seven attack types: Armed Assault, Assassination, Bombing/Explosion, etc.

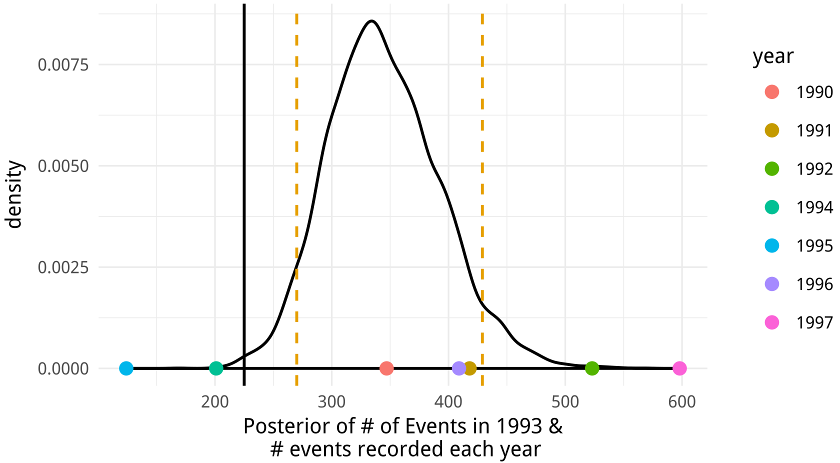
Cumulative Number of Events



Parameter Estimates



Inference on Missing Events in 1993



- GTD's estimate of the number of missing events is 225.
- 90% credible bound is (270, 429).
- Average number of events is 374

Inference on Event Probabilities

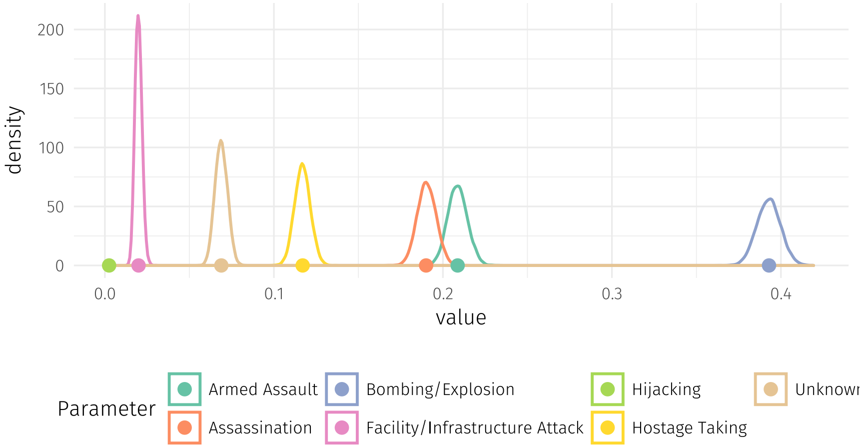


Figure: Posterior of probabilities of each event type. Points are raw frequencies from the data. Density for Hijacking events left out.

Spatial Model Results

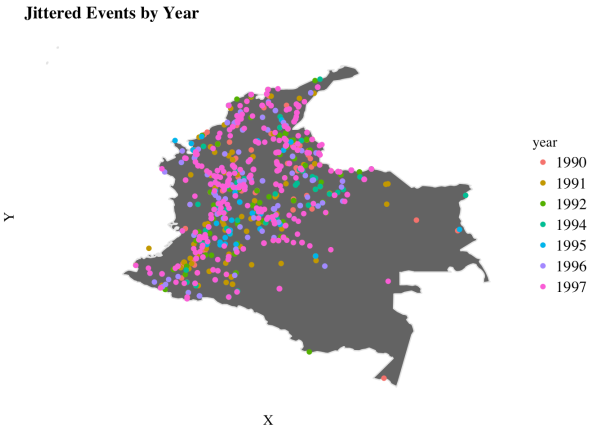


Figure: Terrorism Events in Colombia 1990-1997.

Locations of Missing Data?

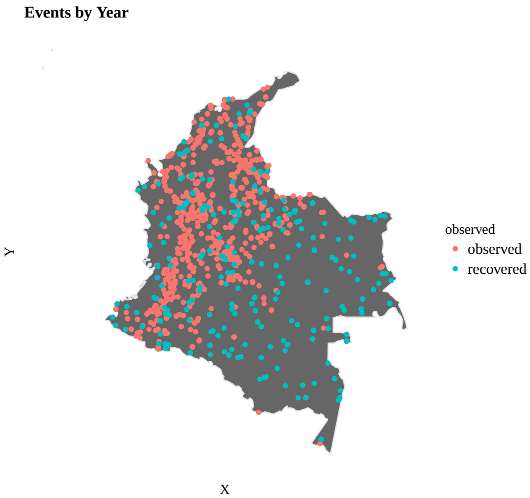


Figure: A sample of missing events.

Need Inhomogeneous Mean

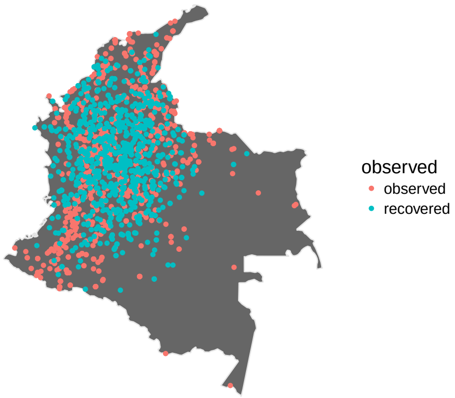


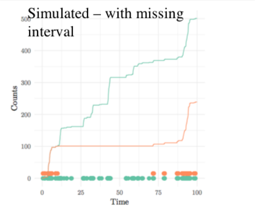
Figure: A sample of missing events.

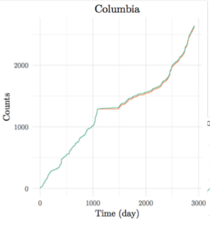
Summary

- Developed Bayesian Model for Hawkes processes with missing time histories
  - Temporal model with categorical marks
  - Treated Spatial locations as a mark
- Implemented formal stochastic code verification strategies
- Demonstrations on GTD:
  - Estimate of the number of missing events larger than GTD
  - GTD presumably used auxiliary information that we did not
- Future Work:
  - Continued development of inhomogeneous mean models

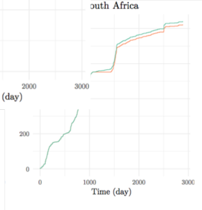
Questions?

Simulated – with missing interval



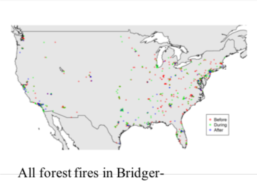


Terrorism events from GTD

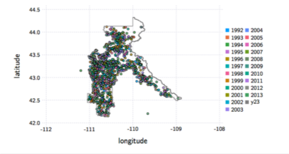


Tweets during the 2014 World Cup

Tweets by location shown for the four hours before, the four hours during, and the four hours after the Paris attacks on 11/13/2015.



All forest fires in Bridger-Teton National Forest from 1992-2013



References I

Cook, Samantha R, Andrew Gelman, and Donald B Rubin. 2006. “Validation of Software for Bayesian Models Using Posterior Quantiles.” *Journal of Computational and Graphical Statistics* 15 (3). Taylor & Francis: 675–92.

Daley, D., and D. Vere-Jones. 2003. *An Introduction to the Theory of Point Processes*. 2nd ed. New York:Springer.

Global Terrorism Database. 2017. “National Consortium for the Study of Terrorism and Responses to Terrorism (START).”

References II

https://www.start.umd.edu/gtd.

Ogata, Y. 1981. "On Lewis' simulation method for point processes." *IEEE Transactions on Information Theory* 27 (1): 23–31.

Ogata, Y., and J. Zhuang. 2006. "Space-time ETAS models and an improved extension." *Tectonophysics* 413 (1-2): 13–23.

Rasmussen, J. G. 2013. "Bayesian Inference for Hawkes Processes." *Methodology and Computing in Applied Probability* 15 (3): 623–42.

Ross, G. J. 2016. "Bayesian Estimation of the Etas Model for Earthquake Occurrences." *Http://Www.gordonjross.co.uk/Bayesianetas.pdf*.

Tanner, Martin A, and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398). Taylor & Francis Group: 528–40.

Zhuang, J., Y. Ogata, and D. Vere-Jones. 2002. "Stochastic Declustering of

References III

Space-Time Earthquake Occurrences." *Journal of the American Statistical Association* 97 (458): 369–80.