Statistical Models for Recurrent Events with Application to Earthquake Data Analysis

W. Zachary Horton   Athanasios Kottas

University of California, Santa Cruz

Recurrent Event Models

The goal when modeling recurrent events is to learn patterns about when the next event will occur. The amount of time between events is called the "inter-arrival time".

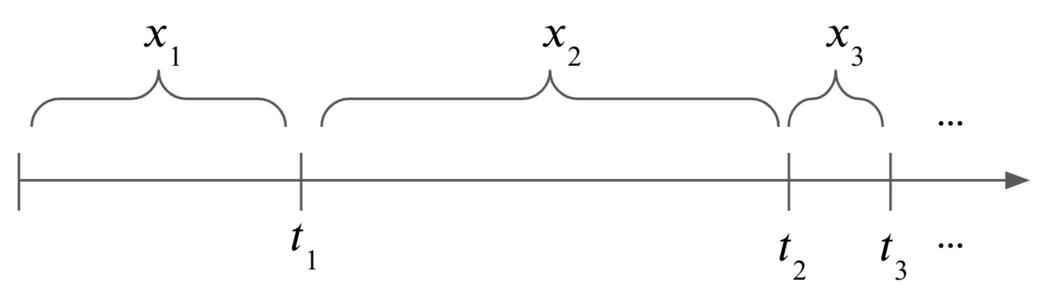

Figure 1. Inter-arrival times (x) shown with corresponding event times (t)

- **Simple Example:** Inter-arrival times of an office printer malfunctioning
- Times between failures are independent and identically distributed.
- In other words, $x_1, x_2, \ldots$ all come from the same distribution $f(\cdot)$

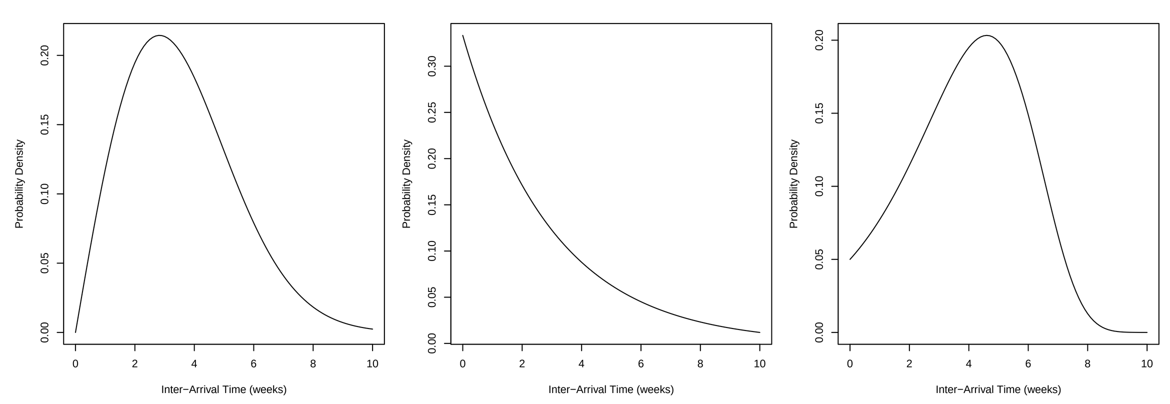

Figure 2. Some simple choices for inter-arrival distribution

Different types of data warrant different inter-arrival distribution shapes. Common choices include Weibull, Gamma, and Lomax distributions.

Mixture Models

In many cases, a simple distribution shape is insufficient. A *mixture model* combines several simple distributions $f_1(\cdot), \ldots, f_L(\cdot)$ into a more complex shape $f^*(\cdot)$ through:

$$f^*(x) = \sum_{\ell=1}^{L} \omega_\ell f_\ell(x)$$

where $\omega_1, \ldots, \omega_L$ are weights that sum to one. Examples:

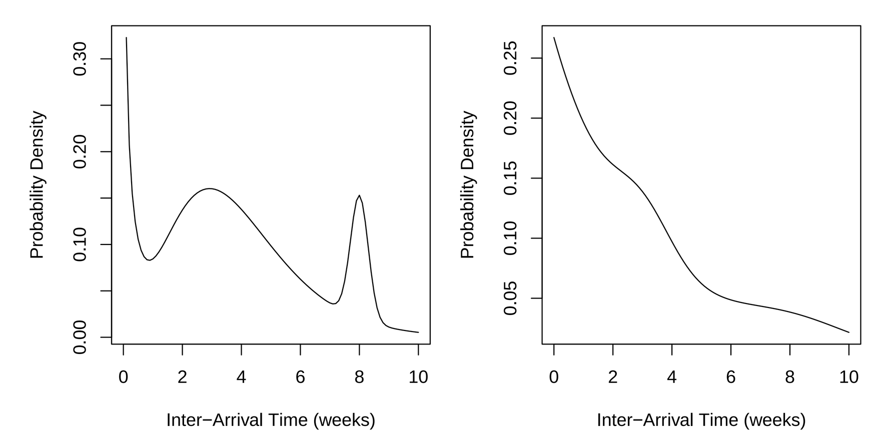

Figure 3. Some possible mixture distributions

Mixture models are able to support a wide variety of shapes. However, *overfit* can be a problem, especially with small amounts of data. Mixtures also exacerbate computational problems for recurrent event models.

Earthquake Data

Motivating our project is a dataset of earthquake occurrences from Jan. 1 1900 to Dec. 31 2018 in the western hemisphere, split into 4 distinct regions [3, 5].

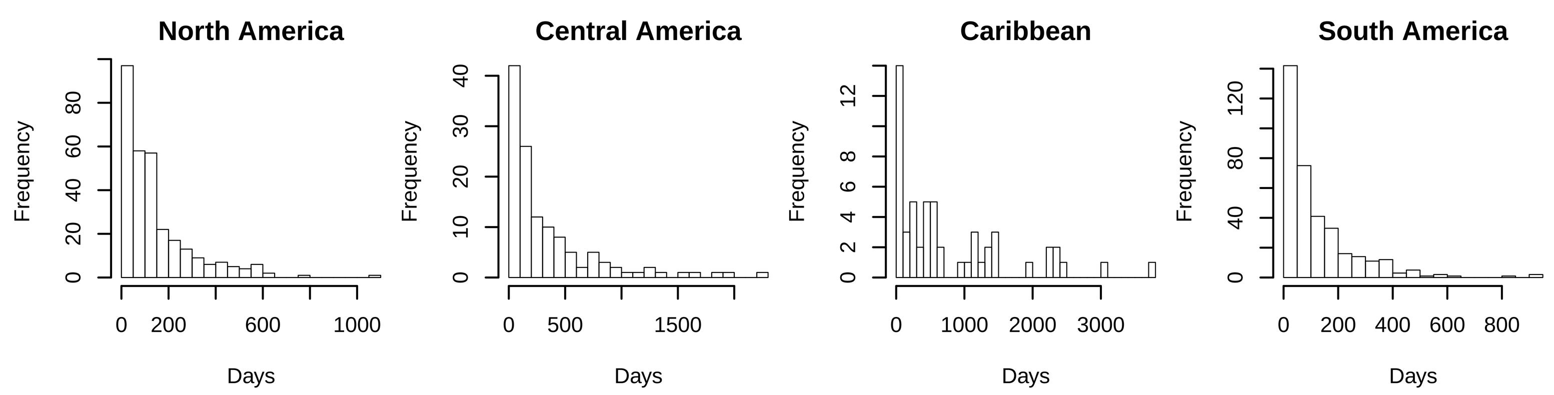

Figure 4. Inter-arrival times of earthquake dataset

We want to model these using *decreasing* inter-arrival distributions, but we also want the flexibility that mixture models provide.

Research Objectives

Build a recurrent event model which accomplishes the following:

1. Guaranteed decreasing distribution shape.
2. Shape flexibility within the decreasing constraint.
3. Overfit alleviation.
4. Easily handled computation.

Key Mathematical Result

To achieve full shape flexibility, we rely on the following important result. Any decreasing function $g(x), x \in \mathcal{R}^+$ can be represented by:

$$g(x) = \int_0^\infty U(x|0, \theta) g^*(\theta) d\theta$$

where $U(x|0, \theta)$ is the uniform or flat distribution on the range $(0, \theta)$ and $g^*(\theta)$ is some weighting function of θ [1]. Below is an example showing how a curve can be decomposed into an infinite mixture of flat distributions with a weighting curve.

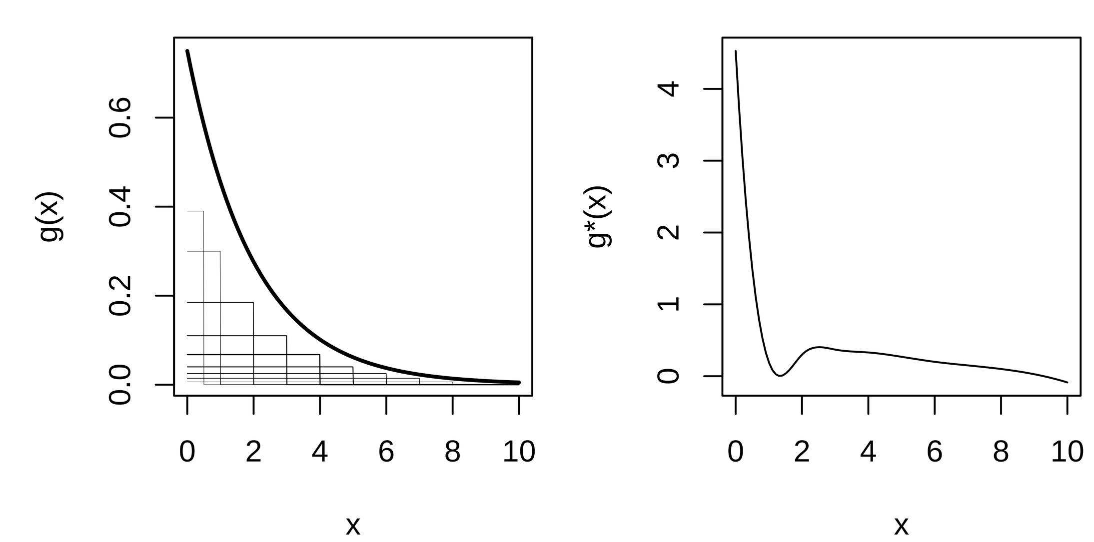

Figure 5. Decomposition into infinite mixture of flat distributions (left) with corresponding weighting function (right).

This result tells us that a big enough mixture of uniform blocks can represent any decreasing shape! This alone achieves the first two research objectives. Additionally, the uniform nature will partially fulfill the third objective.

Proposed Model

We propose the following Bayesian hierarchical model to achieve all our objectives:

$$\begin{aligned}
& x_i | z_i, \boldsymbol{\theta}, N(T) \stackrel{iid}{\sim} U(0, \theta_{z_i}) && i = (1, \ldots, n) \\
& P(N(T) = n | z_{n+1}, \boldsymbol{\theta}) = 1 - F_U(T - t_n | \theta_{z_{n+1}}) && \\
& P(z_i = \ell | \boldsymbol{\omega}) = \omega_\ell && i = (1, \ldots, n+1); \ell = (1, \ldots, L) \\
& \quad \boldsymbol{\omega} | \alpha \sim GDir_L((1, \ldots, 1), (\alpha, \ldots, \alpha)) && \\
& \quad \theta_\ell | b_\theta \stackrel{i.i.d.}{\sim} IG(a_\theta, b_\theta) && a_\theta > 2 \\
& \quad \alpha \sim Ga(a_\alpha, 1/b_\alpha) && \\
& \quad b_\theta \sim Ga(a_0, b_0) &&
\end{aligned}$$

Note that this structure supports mostly closed-form Gibbs sampling. **This model combines both the uniform result and a computational trick with the Dirichlet Process mixture framework [2, 4] to create a powerful class of decreasing-probability recurrent event models.**

Earthquake Data Analysis

Applying the model to the earthquake dataset yields these results. Note that along with distribution shape, we also include inference on the *K-function*[5], which measures the level of event clustering as time passes.

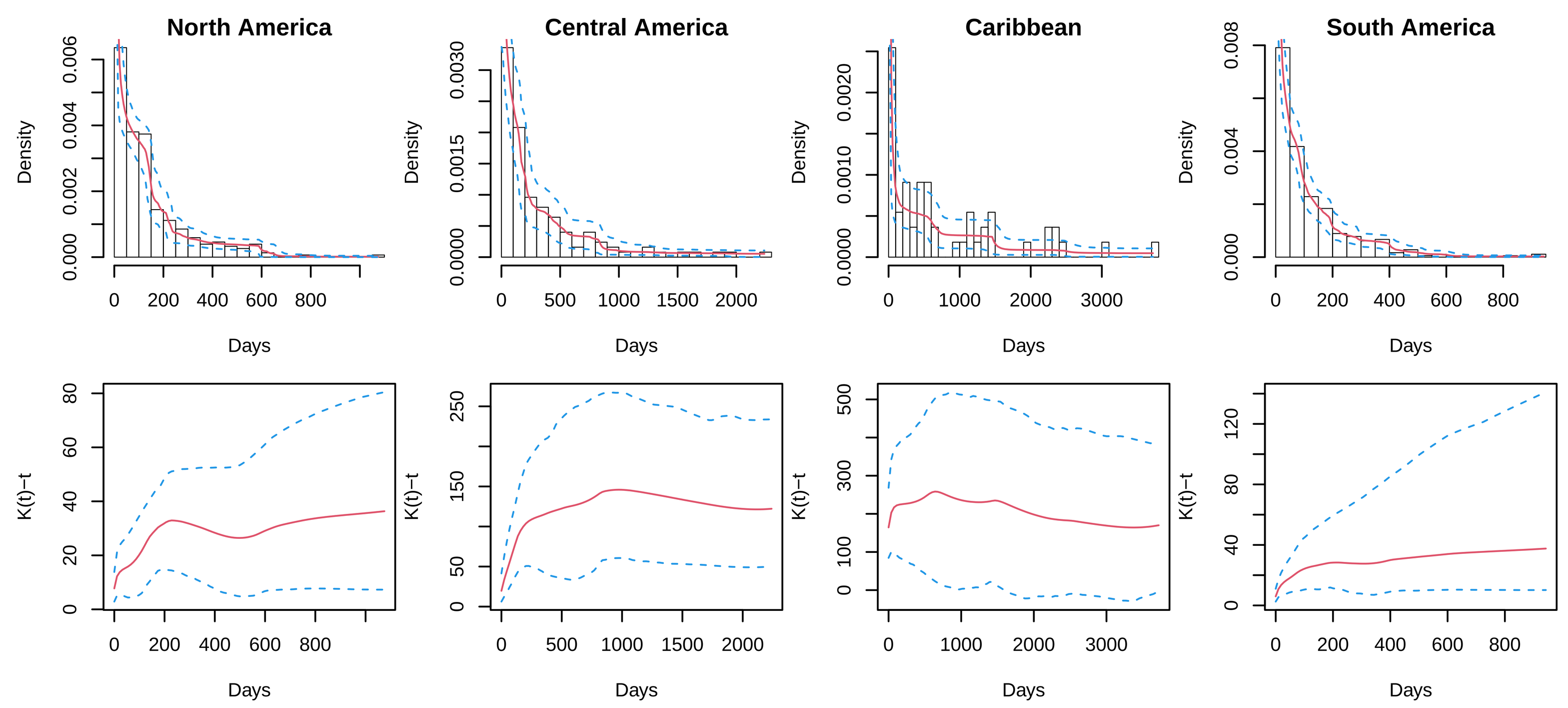

Figure 6. Earthquake data analyzed. Distribution shapes in top row, K-functions in bottom row. Red lines are estimates and dashed blue lines are corresponding 95% confidence intervals.

Clustering behavior appears to vary across region, with equatorial regions exhibiting stronger cluster behavior. This could help engineers appropriately design buildings in the different areas.

Model findings:

- No sign of overfit, even in low data regions.
- All shape estimates are decreasing, yet flexible.
- Computations took only a few minutes on a laptop.

References

[1] Lawrence J. Brunner and Albert Y. Lo. Bayes Methods for a Symmetric Unimodal Density and its Mode. *The Annals of Statistics*, 17(4):1550 – 1566, 1989.

[2] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[3] National Geophysical Data Center/World Data Service (NGDC/WDS). Ncei/wds global significant earthquake database, 2019.

[4] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[5] Sai Xiao, Athanasios Kottas, Bruno Sansó, and Hyotae Kim. Nonparametric bayesian modeling and estimation for renewal processes. *Technometrics*, 63(1):100–115, 2021.

Scan for more.

ARCS ADVANCING SCIENCE IN AMERICA®

ARCS FOUNDATION INC NORTHERN CALIFORNIA CHAPTER