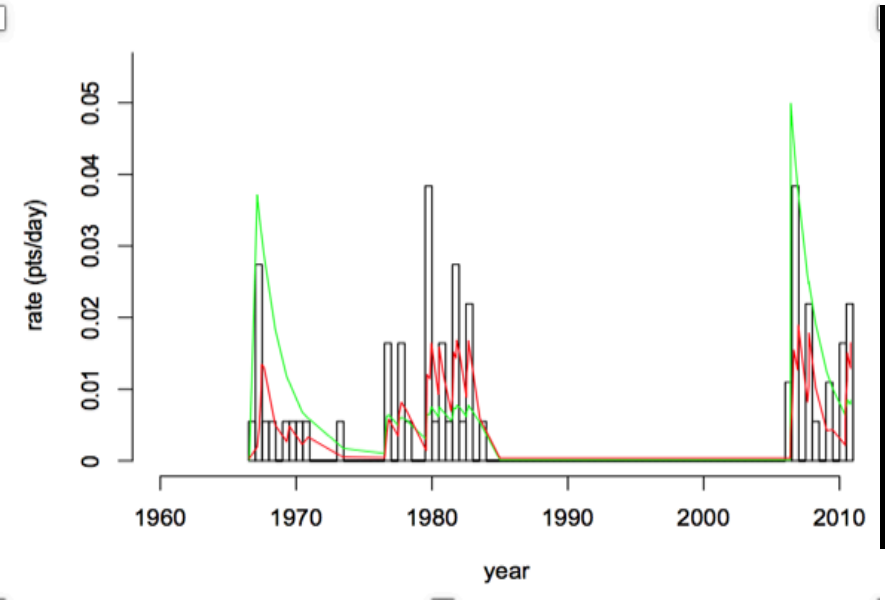
A recursive point process model for infectious diseases.



Frederic Paik Schoenberg, UCLA Statistics

Collaborators: Marc Hoffmann, Ryan Harrigan.

Also thanks to: Project Tycho, CDC, and WHO datasets.

1. Background and motivation.
2. Hawkes models and ETAS.
3. Proposed recursive model.
4. Properties of the recursive model.
5. Application to Rocky Mountain Spotted Fever data.
6. Conclusions.

1. Background and motivation.

* Epidemics have traditionally been modeled by grid based compartmental models like SIR and SEIR models (e.g. Althaus 2014).

* More recently, point process models have been used, especially Hawkes models, offering greater forecast precision (Law et al. 2009, Chaffee 2017).

* With Hawkes processes, the productivity or expected number of disease transmissions triggered directly by a given infected persion, is static. In the case of Hawkes models applied to earthquakes (e.g. Ogata 1988, Ogata 1998), the basic Hawkes model was extended to allow the productivity of an earthquake to depend on its magnitude, but still not to depend on the time or location of the event, nor on the number of previously occurring events.

* With diseases, one may want the productivity to vary.

2. Hawkes models and ETAS.

Probabilistic models for point processes typically involve modeling the conditional rate

$\lambda(t,x,y,m)$ = expected rate of accumulation of points at time t, location (x,y), and magnitude m, given the history of all previous events.

Hawkes (1971) modeled $\lambda(t) = \mu + K\sum_{i:ti<t} g(t-t_i)$.

$\mu \geq 0$ is the background rate, K is the productivity, $0 \leq K \leq 1$, and g is the triggering density satisfying $\int_0^\infty g(u)du \doteq 1$

Ogata (1988) proposed the Epidemic-Type Aftershock Sequence (ETAS) model, which is like a Hawkes model but where the productivity can depend on magnitude.

$\lambda(t) = \mu + K\sum_{i:ti<t} g(t-t_i; m_i)$, with $g(u; m_i) = (u+c)^{-p} \exp\{a(m_i-M_0)\}$.

2. Hawkes models and ETAS.

Hawkes and ETAS models have been extended to space-time, and also have been extended in some other ways.

For example, the HIST-ETAS model (Ogata et al. 2003, Ogata 2004) assumes the parameters in the ETAS model are locally constant within small spatial-temporal cells.

Harte (2014) allows the ETAS model's productivity parameter to vary smoothly in space and time.

Here we extend the model in a different way, allowing the productivity to vary as a function of λ.

3. Proposed recursive model.

When the prevalence of a disease is low in a region, as is the case when the epidemic has never struck before or has not struck in considerable time, then the conditional intensity λ is small and one would expect the rate of transmission for each infected person to be quite high. A carrier of the disease may be expected to infect many others.

When the epidemic is at its peak and many subjects have contracted the disease, on the other hand, λ is large and one might expect the rate of transmission to be lower due to human efforts at containment and intervention of the disease, and because many subjects may have already been exposed and thus might be recovered and immune to further infection, or deceased.

This suggests a model where the productivity for a subject infected at time t is inversely related to the conditional intensity at time t. Since the conditional intensity in turn depends critically on this productivity, we call the model *recursive*.

3. Proposed recursive model.

We may write this model $\lambda(t) = \mu + \int_0^t H(\lambda_{t'})\, g(t - t')\, dN(t')$,

where $\mu>0$, $g>0$ is a density function, and $\lambda_{t'}$ means $\lambda(t')$.

We focus in particular in what follows on the case where

$H(x) = k\, x^{-\alpha}$, so that $\lambda(t) = \mu + \kappa \int_0^t \lambda_{t'}^{-\alpha}\, g(t - t')\, dN(t')$.

We will refer to the special case where $\alpha = 1$, i.e. where

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'}\, dN(t')$, as the *standard recursive model*.

The triggering density g may be given e.g. by an exponential density, $g(u) = \beta \exp(-\beta u)$.

3. Basic properties of the recursion model.

$\lambda(t) = \mu + \kappa \int_0^t \lambda_{t'}^{-\alpha} g(t - t')\, dN(t')$ for the *recursive* model.

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'}\, dN(t')$ defines the *standard recursive model*.

(i) Existence. For any $\alpha > 0$, we can construct a simple point process with the intensity above.

(ii) Mean and variance.

Using the martingale formula, for a recursive process *N* with triggering density *g*,

$$1/T\, EN(0,T) \to \mu + \kappa /T\, E \int_0^T \lambda_{t'}^{1-\alpha}\, dt \text{ as } T \to \infty.$$

For the *standard recursive model*, $\alpha = 1$, so this reduces simply to $\mu + \kappa$.

3. Basic properties of the recursion model.

$\lambda(t) = \mu + \kappa \int_0^t \lambda_{t'}^{-\alpha} g(t - t')\, dN(t')$ for the *recursive* model.

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'}\, dN(t')$ defines the *standard recursive model*.

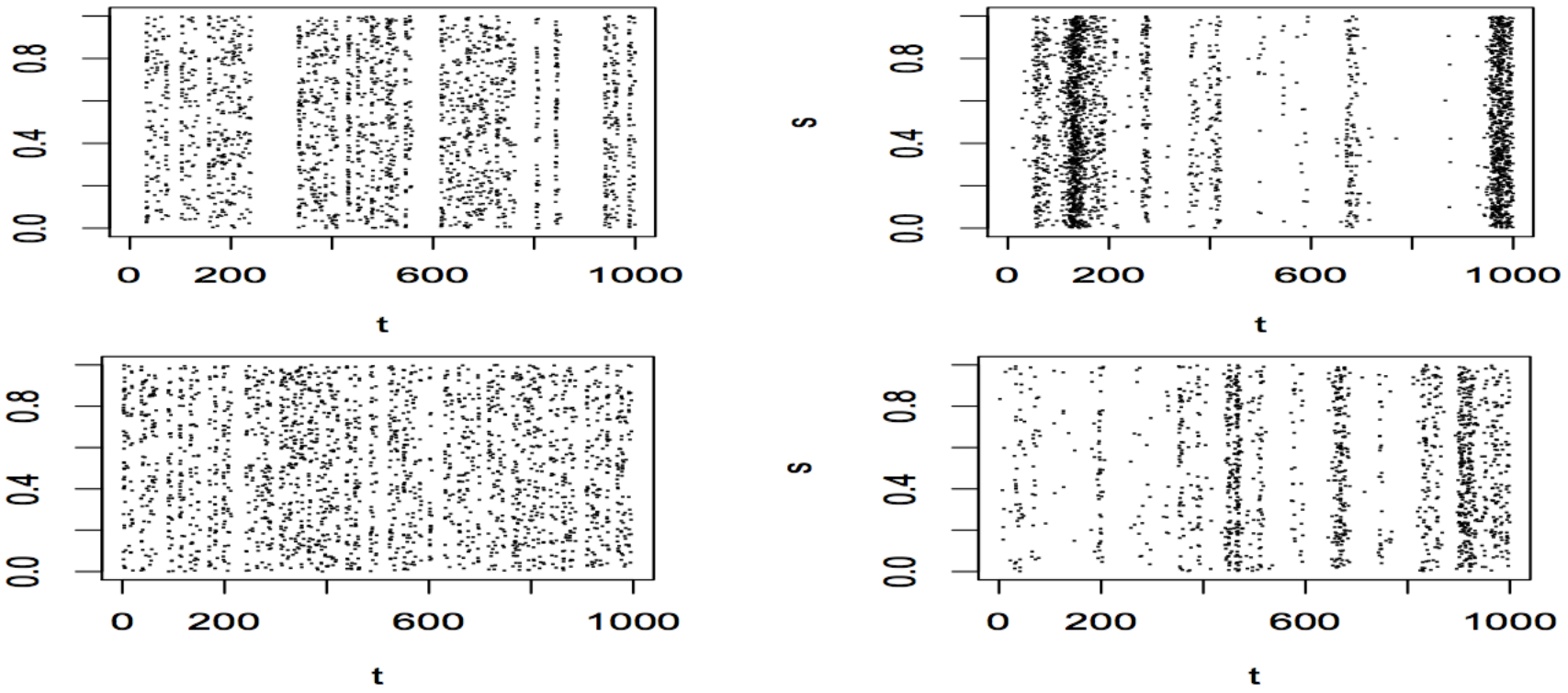
$1/T\, EN(0,T) \rightarrow \mu + \kappa$.

This highlights a major difference between Hawkes models and recursive models. For a Hawkes process, doubling the background rate amounts to doubling the total expected number of points, but this is far from true for the recursive process.

If $\mu = 0.1$ and $\kappa = 2$, doubling μ would only increase the total expected number of points by less than 5%.

If $\mu = 0.01$ and $\kappa = 2$, doubling μ would increase the total expected number of points by less than 0.5%.

3. Basic properties of the recursion model.



(a) Simulation of a std. recursive model with $\mu = 0.05$, $\kappa = 2$, and g exponential with $\beta = 0.8$.
(b) Simulation of a Hawkes model with the same g and μ as in (a), and with $K = \mu/(\mu+\kappa)$ so that the processes in (a) and (b) have the same expected number of points.
(c) Simulation of a standard recursive model with $\mu=0.1$, $\kappa = 2$, and g exponential with rate 1.
(d) Simulation of a Hawkes model with the same g and μ as in (c), and with $K = \mu/(\mu+\kappa)$ so the processes in (c) and (d) have the same expected number of points.
All 4 simulations are over the same temporal domain [0, 1000]. The points are spread uniformly over the y axis for ease of visualization.

3. Basic properties of the recursion model.

$\lambda(t) = \mu + \kappa \int_0^t \lambda_{t'}^{-\alpha} g(t - t')\, dN(t')$ for the *recursive* model.

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'}\, dN(t')$ defines the *standard recursive model*.

(iii) Law of Large Numbers.

Assuming $\limsup_{T \to \infty} \sqrt{T} \int_T^\infty g(t)dt < \infty$,

$N(0,T) / T$ converges to $\mu + \kappa$ as $T \to \infty$ with convergence rate $\sqrt{T}$ in L^2.

3. Basic properties of the recursion model.

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'} \, dN(t')$ defines the *standard recursive model*.

(iv) Productivity.

The productivity of a point τ_i is defined as the expected number of first generation offspring of the point τ_i. For a Hawkes process, the productivity of each point is simply K.

For the recursive model, the productivity of any point τ_i is $H\{\lambda(\tau_i)\}$.

For the standard recursive process, the expected value of the total productivity over all points is

$$\kappa E \int_0^T \frac{1}{\lambda_t} \, dN_t = \kappa E \int_0^T \frac{1}{\lambda_t} \lambda_t \, dt = \kappa T.$$

The avg. productivity of a point converges almost surely to $\kappa / (\mu + \kappa)$.

3. Basic properties of the recursion model.

$\lambda(t) = \mu + \kappa \int_0^t g(t - t') / \lambda_{t'} \, dN(t')$ defines the *standard recursive model*.

(iv) Productivity.

The avg. productivity of a point converges almost surely to $\kappa / (\mu + \kappa)$.

This highlights another difference between the recursive and Hawkes models.

For a Hawkes process, the points all have constant productivity, K.

For a standard recursive process, the productivity for the first point is κ/μ, which is larger than the productivity for any subsequent point.

The productivity of the points ultimately averages $\kappa/(\mu + \kappa)$.

3. Basic properties of the recursion model.

(v) Declustering.

Zhuang et al. (2002) proposed stochastic declustering for ETAS processes where for each pair of points, one computes the probability that earthquake j was triggered by earthquake i, and the probability that each earthquake is a mainshock, according to the fitted model.

With epidemics we are interested in the probability that person i may have infected person j.

For any points $\tau_i < \tau_j$, the probability that subject j was infected by subject i is given by $H(\lambda_{\tau_i})g(\tau_j - \tau_i) / \lambda(\tau_j) =$

$H(\lambda_{\tau_i})g(\tau_j - \tau_i) / [\mu + \int_0^{\tau_j} H(\lambda_{t'})g(\tau_j - t')dN_{t'}]$ which is easy to compute.

4. Estimation of the recursive model.

One can estimate the parameters in the model by MLE, where the integral term $\int\lambda(t)dt$ may readily be approximated in the standard way (see e.g. Schoenberg 2013). Assuming g(t) has negligible mass for $t > T - \tau_i$, one may use the approximation

$\int\lambda(t)dt \sim \mu T + \sum_i H(\lambda(\tau_i))$, which is trivial to compute.

The parameter vector θ maximizing the loglikelihood can then be estimated by standard Newton-Raphson optimization routines.

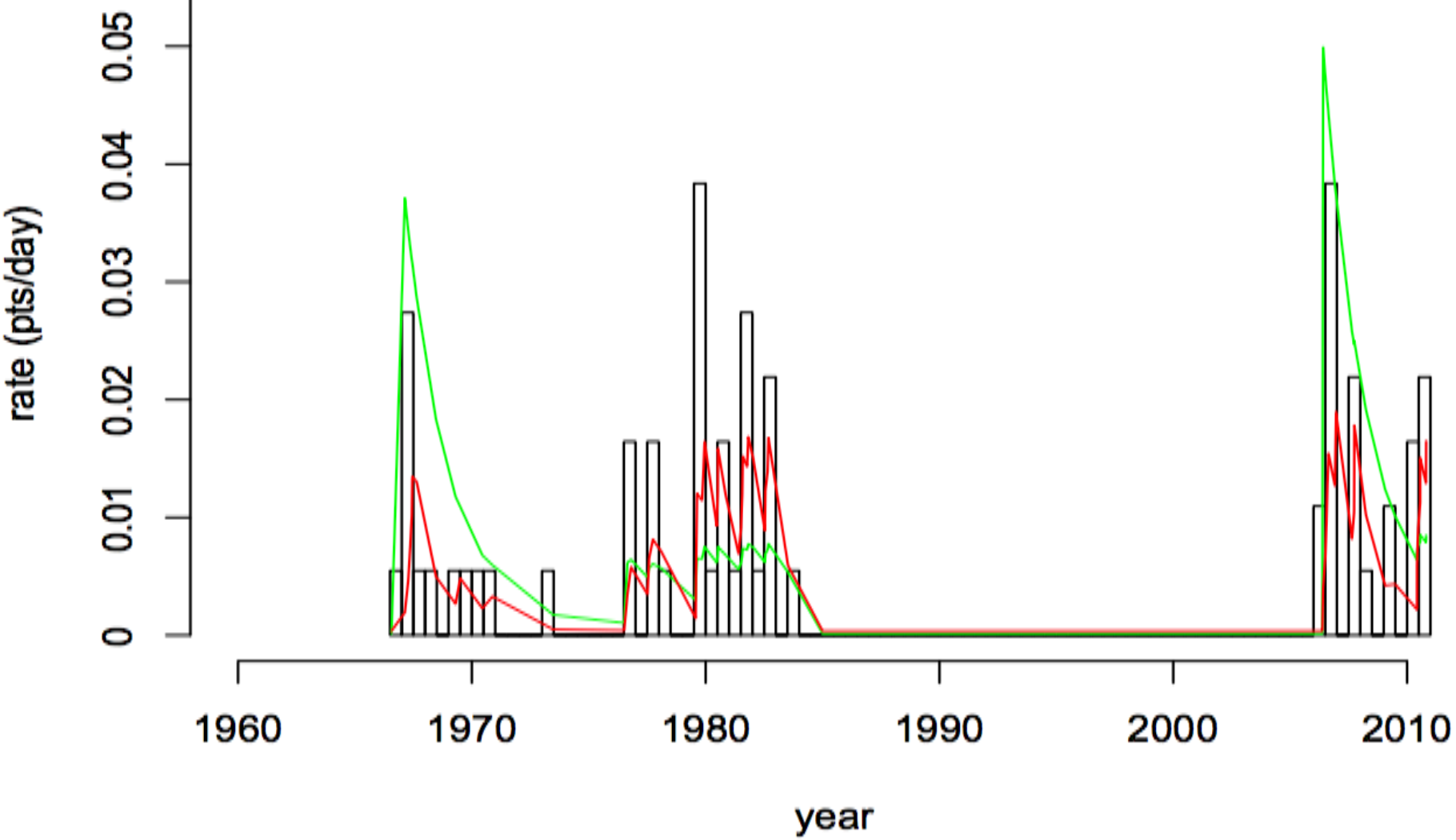
Approximate standard errors can be derived via the diagonal elements of the inverse of the Hessian of the log-likelihood (Ogata 1978), or by repeatedly simulating and re-estimating by MLE as suggested by Harte (2010).

5. Application to CA Rocky Mountain Spotted Fever cases.

Recorded cases of Rocky Mountain Spotted Fever in California from Jan 1, 1960 to Dec 31, 2011 were collected by the CDC and catalogued by Project Tycho, www.tycho.pitt.edu.

Weeks with no data over this period were treated as having zero confirmed cases. When the temporal resolution of the data is by week, onset times were drawn uniformly within each week, as e.g. in Althaus (2014) and Chaffee (2017).

Estimates of (μ, κ, β, α) are (0.000139 pts/day, 0.00205 triggered pts/observed pt, 0.00151 pts/day, 1.09), with corresponding standard error estimates (0.000144, 0.0403, 0.00630, 0.0731).



Histogram of confirmed Rocky Mountain Spotted Fever cases in California from 1/1/1960 to 12/31/2011, along with the estimated rate of the recursive model (green) and Hawkes model (red), each with exponential triggering function and fit by maximum likelihood.

5. Application to CA Rocky Mountain Spotted Fever cases.

The log-likelihood for the fitted Hawkes model is -385.1, or 19.9 less than the log-likelihood of the recursive model. As these are nested models, the difference in log-likelihoods is approximately $\chi 2$-distributed, and based on the Akaike Information Criterion (Akaike 1974), the improvement in fit using the recursive model is statistically significant.

5. Application to CA Rocky Mountain Spotted Fever cases.

In order further to assess the fit of the model, we used super-thinned residuals (Clements et al. 2013). In super-thinning, one selects a constant b, thins the observations by keeping each observed point τ_i independently with probability $b/\lambda(\tau_i)$ if $\lambda(\tau_i) > b$, and superposes points from a Poisson process with rate $(b-\lambda)1_{\lambda \leq b}$, where 1 denotes the indicator function. A default choice for b is the mean of $\hat{\lambda}$ at the observed points, as suggested in Gordon et al. (2015). The resulting super-thinned residuals form a homogeneous Poisson process with rate b iff. $\hat{\lambda}$ is the true conditional rate of the observed point process (Clements et al. 2013).

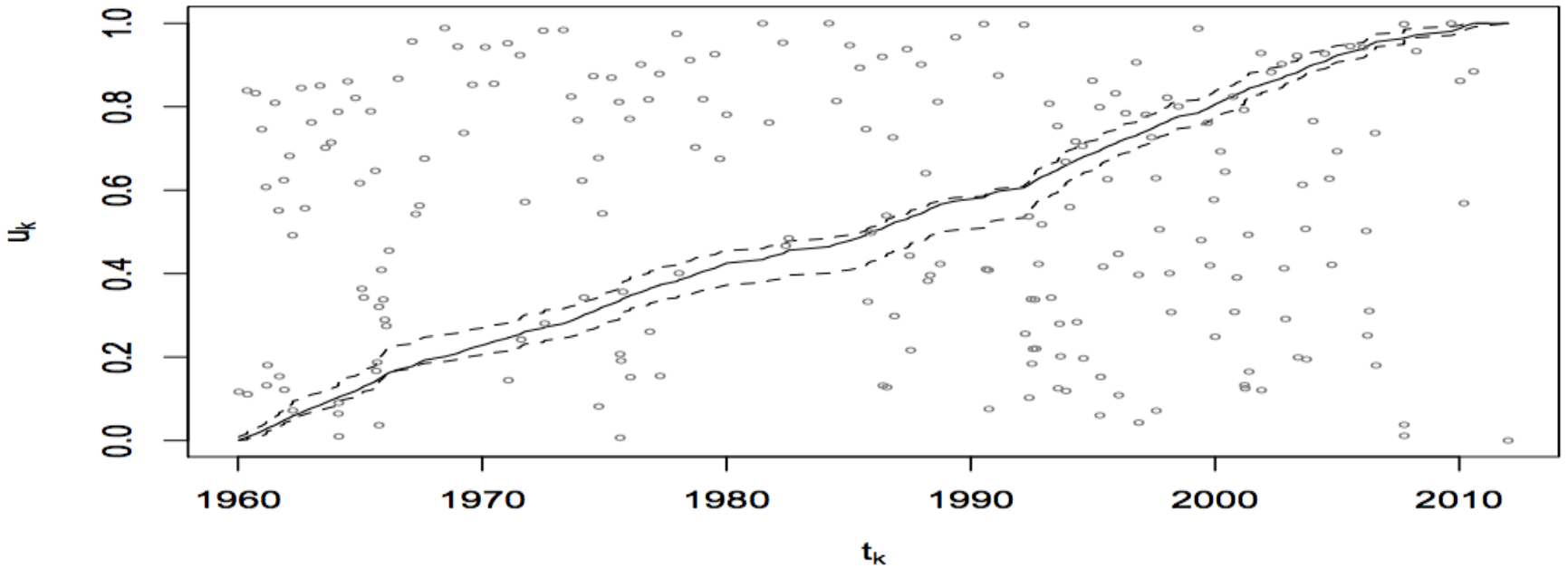
5. Application to CA Rocky Mountain Spotted Fever cases.

If t_i are the times of the super-thinned points, one may consider the interevent times, $r_i = t_i - t_{i-1}$ (with the convention $t_0 = 0$), which should be exponential with mean 1/b if the fitted model for λ is correct.

One can inspect the uniformity of the standardized interevent times $u_i = F^{-1}(r_i)$, where F is the cumulative distribution function of the exponential with mean 1/b.

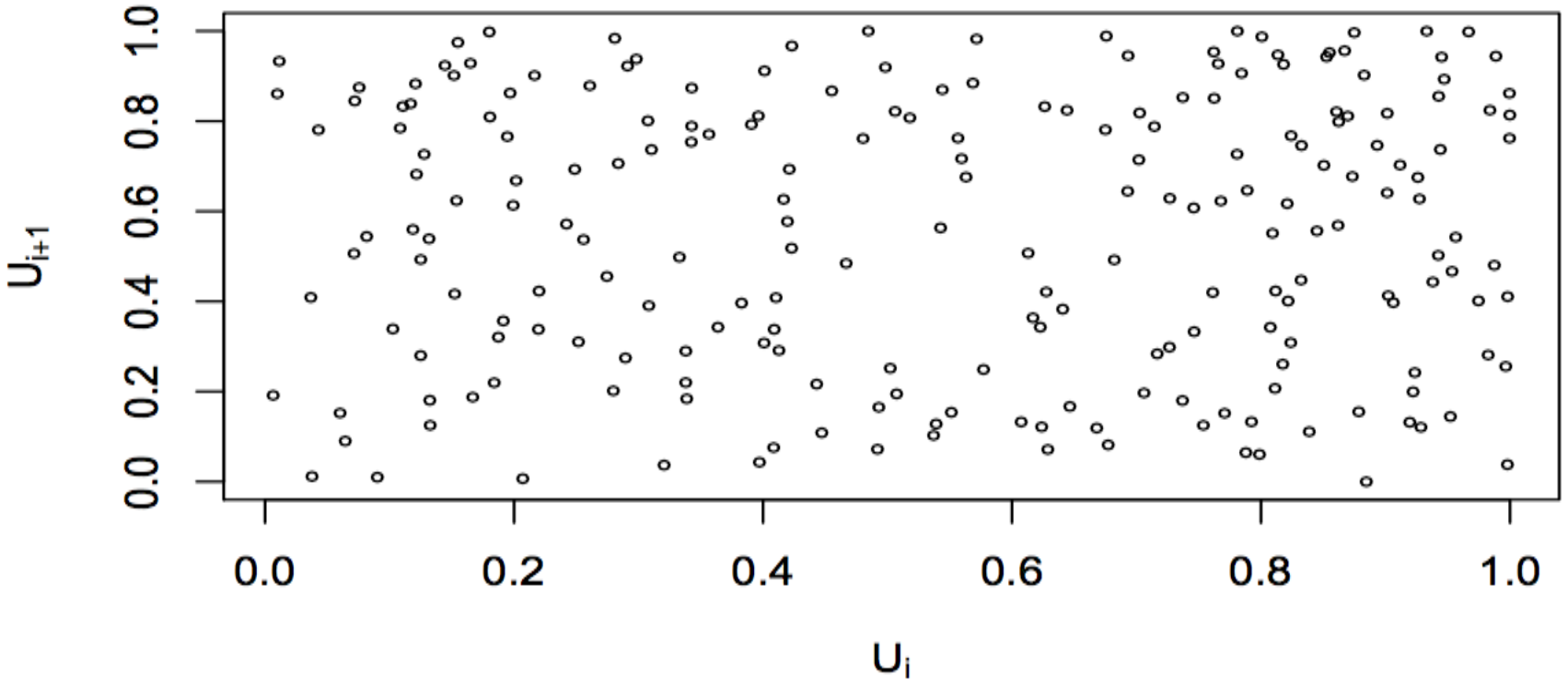
The Figure on the next slide shows the super-thinned residuals t_i along with their corresponding standardized interevent times u_i, as well as the cumulative sum of the standardized intervent times, and the individual 95% confidence bounds based on 1000 simulations of an equivalent number of uniform random variables.

5. Application to CA Rocky Mountain Spotted Fever cases.



Super-thinned residuals t_k and their corresponding standardized interevent times u_k. The solid line shows, for each value of t_k, the normalized cumulative sum of u_k. There are fewer small interevent times than expected, especially between 1979 and 1985. Otherwise the interevent times appear to be well scattered.

5. Application to CA Rocky Mountain Spotted Fever cases.

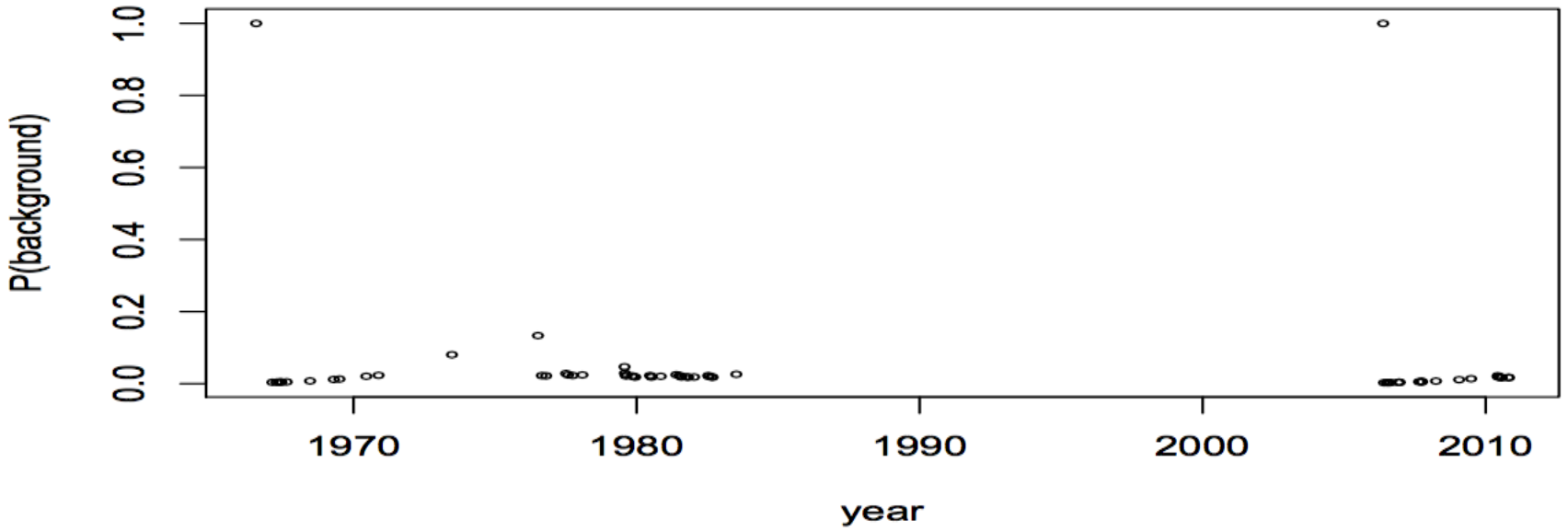


Lag plot of the standardized interevent times u_i of the super-thinned residuals using b = 100 points/year.

5. Application to CA Rocky Mountain Spotted Fever cases.

The proposed recursive model can also be compared with the Hawkes model in terms of predictive efficacy. We separated the California Rocky Mountain Spotted Fever data into a training set (1/1/1960-12/31/2006) on which the models would be fit by MLE, and saved the last 5 years of this dataset for testing. For the fitted recursive model, the loglikelihood over the test data was 65.8, and the loglikelihood for the fitted Hawkes model was -128.8, for a difference of 194.6. Imagine using as a threshold the 95th percentile of $\hat{\lambda}(t_i)$, evaluated at the observed points t_i in the training set.

For the recursive model, 13 out of 23 observed points (**56.5%**) in the test set occurred when the estimated value of $\hat{\lambda}$ exceeded the threshold, and only 9.6% of days in the test set were false alarms, i.e. days when the threshold was exceeded yet no points occurred. With Hawkes, the corresponding 95th percentile threshold was lower. 26.2% of days in the test set would have been false alarms, and only 7 out of 23 observed points (**30.4%**) in the test set occurred when $\hat{\lambda}$ exceeded the threshold.



Stochastic declustering of the Rocky Mountain Spotted Fever data in CA using the fitted recursive model. Most pts are attributed to contagion rather than new outbreaks. Two particular points in 1966 and 2006 are assigned near certainty of being attributed to new outbreaks, and two points in 1973 and 1976 are assigned higher probability of being attributable to new outbreaks rather than contagion from one of the other points in the dataset, according to the fitted model.

8. Concluding remarks.

The improvement in fit from the recursive model relative to the Hawkes model is significant and cannot be explained as overfitting, as even when the models were fit using a training dataset (1/1/1960 to 12/31/2006) and then assessed on a separate testing time period (1/1/2007-12/31/2011), the recursive model significantly outperformed the Hawkes model using this data on Rocky Mountain Spotted Fever in California.

The model seems particularly useful for epidemics where gaps are followed by big clusters of cases.

The dataset has potential problems, however, esp. misdiagnoses and missing data, and the model should be fit to other epidemic data in the future.