

Is a sub 2 hour marathon in the near future? Modeling rare events in sports.

Rodney X. Sturdivant, Ph.D., Baylor University and Nick Clark, Ph.D.,
West Point



Outline

- ▶ Background
- ▶ Marathon Data
- ▶ Simple Model
- ▶ Self-Exciting Model
- ▶ Model Comparisons (preliminary results)
- ▶ Further Research
- ▶ SCORE

Background

The New York Times

Records Fell at the Track Worlds. A Trend? Not So Fast.

Some have referred to this era as a golden age of better and better times. But a deeper look at the data shows the simple shorthand conclusion is incomplete.

Figure 1: Golden Age?

- ▶ Are we living in a time of records?
 - ▶ Idea: seems like an increase in records falling, but is it just the nature of randomness?
- ▶ How can we address this question?
- ▶ What would randomness look like?



Figure 2: Rod Aloha 10K Run (San Diego, 2018), 2nd Age Group



Figure 3: Rod Last Marathon (LA, 2018), 1st, all city of Glendora CA runners

Marathon World Record Data

Men's Marathon world records since 1908

- ▶ 50 total
- ▶ First 5

Time	Name	Nationality	Date	Event/Place
2:55:18.4	Johnny Hayes	United States	July 24, 1908	London, United Kingdom
2:52:45.4	Robert Fowler	United States	January 1, 1909	Yonkers,[nb 5]United States
2:46:52.8	James Clark	United States	February 12, 1909	New York City, United States
2:46:04.6	Albert Raines	United States	May 8, 1909	New York City, United States
2:42:31.0	Henry Barrett	United Kingdom	May 8, 1909[nb 6]	Polytechnic Marathon, London, United Kingdom

- ▶ Last 5

Time	Name	Nationality	Date	Event/Place
2:03:59	Haile Gebrselassie	Ethiopia	September 28, 2008	Berlin Marathon
2:03:38	Patrick Makau	Kenya	September 25, 2011	Berlin Marathon
2:03:23	Wilson Kipsang	Kenya	September 29, 2013	Berlin Marathon
2:02:57	Dennis Kimetto	Kenya	September 28, 2014	Berlin Marathon
2:01:39	Eliud Kipchoge	Kenya	September 16, 2018	Berlin Marathon

Breaking News...

SPORTS

He was so fast, he had time to celebrate long before the second-place runner arrived

Updated September 25, 2022 · 2:36 PM ET ⓘ



Kenya's Eliud Kipchoge crosses the line to win the Berlin Marathon in Berlin on Sunday.
Christoph Soeder/AP

Figure 4: Eliud Kipchoge, Berlin, 2:01:09, September 25, 2022

Fun Facts

- ▶ Countries with most records
 - ▶ U.K. 12
 - ▶ U.S. 7
 - ▶ Kenya 6
 - ▶ Japan/Ethiopia 5
- ▶ Largest decrease:
 - ▶ James Clark, US, 1909
 - ▶ 2:46:52.8
 - ▶ Nearly 6 minutes faster than previous



Figure 5: Khalid Khannouchi, last US record holder (2:05:38, 2002)

Fun Facts

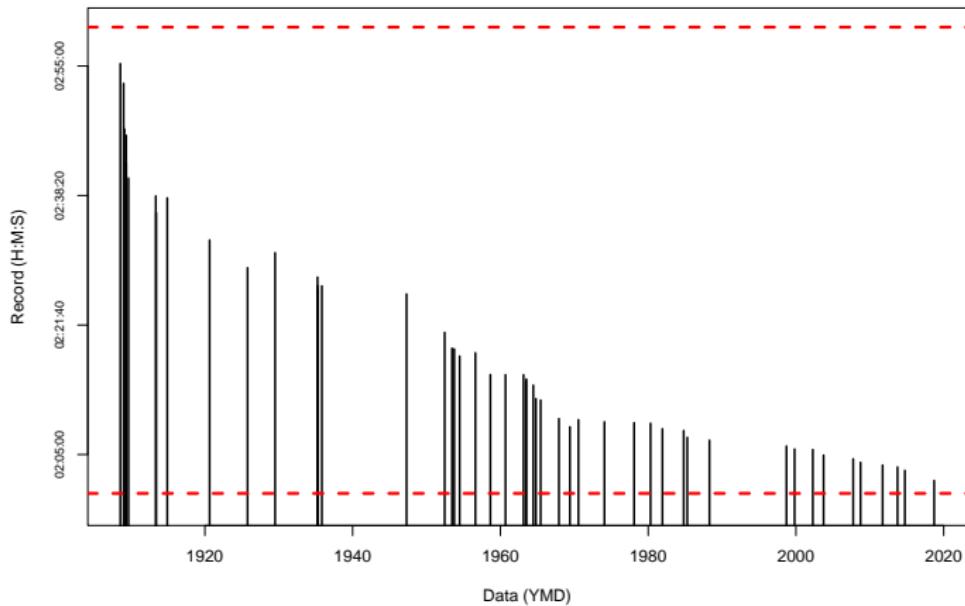


- ▶ Most records
 - ▶ Jim Peters, UK,
 - ▶ Set records 4 times from 1952-1954
- ▶ Locations with most records
 - ▶ Berlin Marathon - 9
 - ▶ Last 8 records
 - ▶ Polytechnic Marathon (London) - 8
 - ▶ Tokyo Marathon - 5

Figure 6: Polytechnic Marathon Program 1993 (Jim Peters)

Visualizing the data

- ▶ Two and three hour times shown as horizontal lines



- ▶ 2 hour marathon pace: 4:35 per mile
- ▶ 3 hour pace: 6:52 per mile

SIMPLE MODEL

POISSON PROCESS

A model for a series of discrete events where the average time between events is known, but the exact timing of events is “random” meeting the following criteria:

- ▶ Events are independent of each other. The occurrence of one event does not affect the probability another event will occur.
- ▶ The average rate (events per time period) is constant.
- ▶ Two events cannot occur at the same time.

Poisson Process Interarrival Times

The time between events (known as the interarrival times) follow an exponential distribution defined as:

$$P(T > t) = e^{-\lambda t}$$

- ▶ T is the random variable of the time until the next event
- ▶ t is a specific time for the next event
- ▶ λ is the rate: the average number of events per unit of time.

Note the possible values of T are greater than 0 (positive only).

Reasonableness of Exponential Interarrivals

The exponential distribution has certain attributes, for example:

$$E(T) = SD(T) = 1/\lambda$$

For the time between record data:

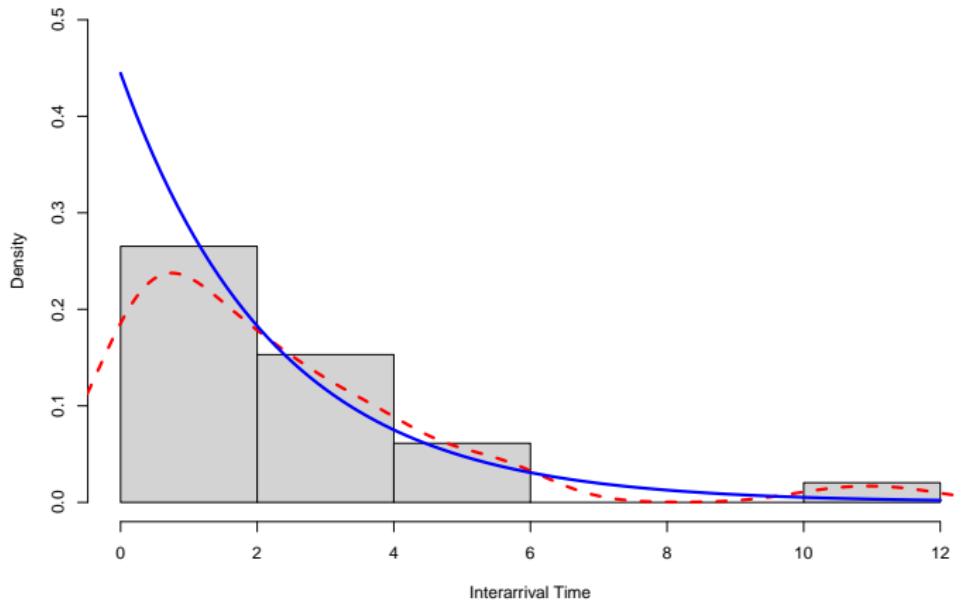
- ▶ Mean: 2.25 years
- ▶ SD: 2.43 years

Reasonable... slightly “overdispersed”

Estimating the Model

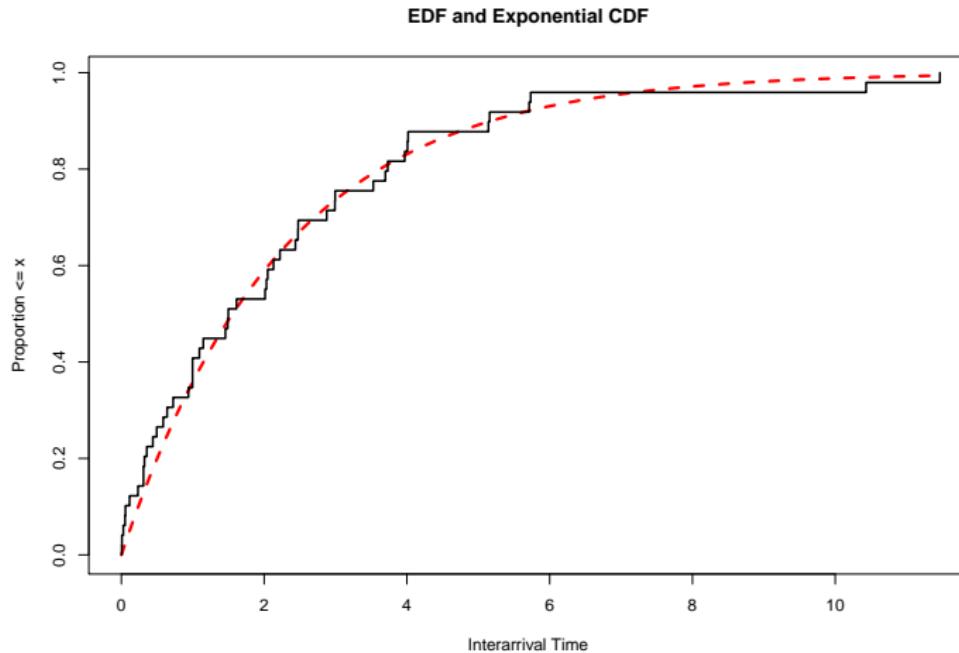
We estimate (MLE) $\lambda = 1/E(T) = 0.445$

Histogram, density curve and exponential model



Graphical Assessment of Fit

Empirical Distribution Function (EDF) and Cumulative Distribution Function (CDF)



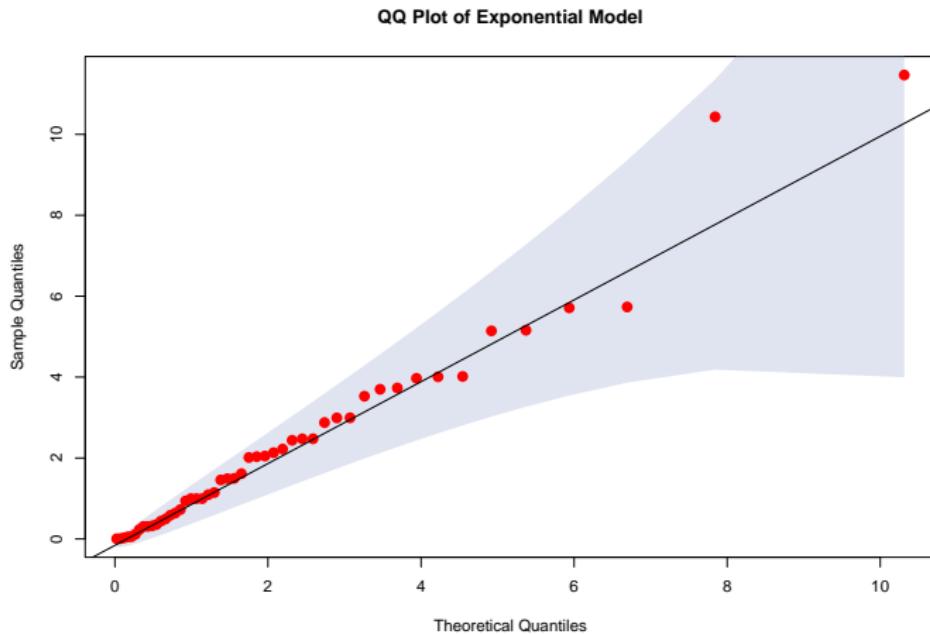
Testing Model Fit

Goodness of fit tests

- ▶ Kolmogorov-Smirnov: $p = 0.926$
- ▶ Cramer-Von Mises: $p = 0.855$
- ▶ Anderson-Darling: $p = 0.563$

All fail to reject the null hypothesis of model fit

Are records then random?



2022-09-30

What are the poorly fit points?

- ▶ 11.5 year gap

Time	Name	Nationality	Date	Event/Place
2:26:42	Sohn Kee-chung	Japanese Korea	November 3, 1935	Tokyo, Japan
2:25:39	Suh Yun-bok	Korea	April 19, 1947	Boston Marathon

- ▶ 10.4 year gap

Time	Name	Nationality	Date	Event/Place
2:06:50	Belayneh Dinsamo	Ethiopia	April 17, 1988	Rotterdam Marathon
2:06:05	Ronaldo da Costa	Brazil	September 20, 1998	Berlin Marathon

Have we answered the question?

- ▶ Model fit seems to suggest random events (no “golden age” of records)
- ▶ Assumptions such as independence very hard to assess
 - ▶ Could model miss periods of increased records and still look reasonable?
 - ▶ There may be factors that increase the probability of records...

A “Self-Exciting” Model

Self-Exciting Point Processes

- ▶ Events “trigger” more events
- ▶ Examples of use include earthquakes, crime waves

Hawkes Processes

- ▶ Let H_t be the history of events up to time t . The Hawkes (1971) model of the conditional intensity is:

$$\lambda(t|H_t) = \nu + \sum_{i:t_i < t} g(t - t_i)$$

where ν is the background rate of events and g is the “triggering function”.

Exponential Triggering Function

- ▶ The “triggering” function can be further decomposed:

$$g = \mu g^*$$

where g^* is a density function known as the “reproduction kernel” and μ is known as the “reproduction” mean.

- ▶ A common choice for the “reproduction kernel” is the exponential density given by:

$$g^*(t) = \beta e^{-\beta t}$$

Fitting the model

Parameter estimates for marathon data (exponential) Hawkes process, using MLE (Cheyson and Lang, 2020):

- ▶ baseline intensity 0.397
- ▶ reproduction mean 0.126
- ▶ exponential reproduction function rate 3.795

Note the baseline intensity is slightly lower than the constant model rate estimate of 0.445

The estimated reproduction function is then:

$$\begin{aligned}g(t) &= \mu g^*(t) = \mu \beta e^{-\beta t} \\&= 0.13 * 3.79 e^{-3.79t}\end{aligned}$$

Model implications

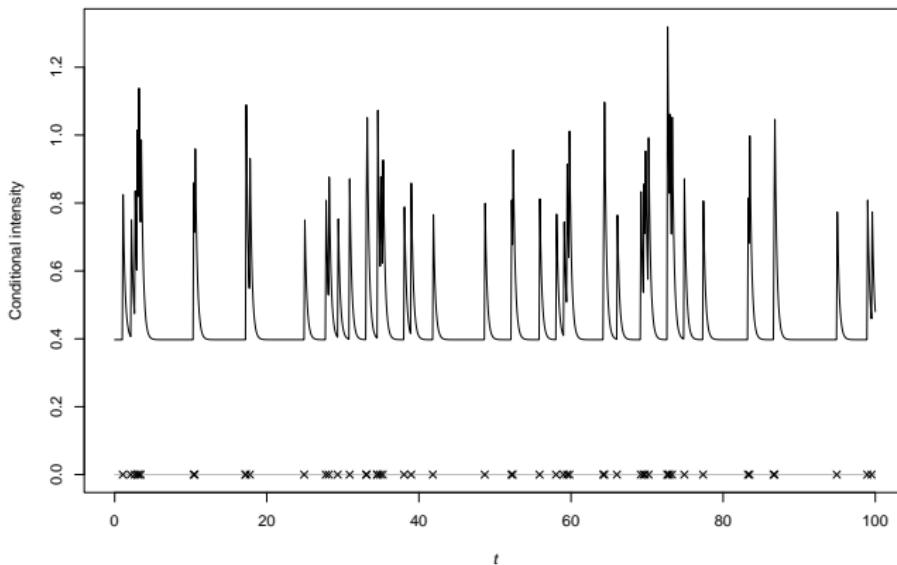
At the instant of the first event (world record), $t = t_1$ so $g(t - t_1 = 0)$ and the reproduction rate is:

$$g(0) = 0.13 * 3.79e^{-3.79*0} = 0.13 * 3.79 = 0.479$$

- ▶ The rate increases from the baseline rate of 0.397 by this amount at the moment of this occurrence
- ▶ The rate then decays back to baseline over time (unless a new event occurs)
- ▶ Each new event “excites” the rate to increase and then decay

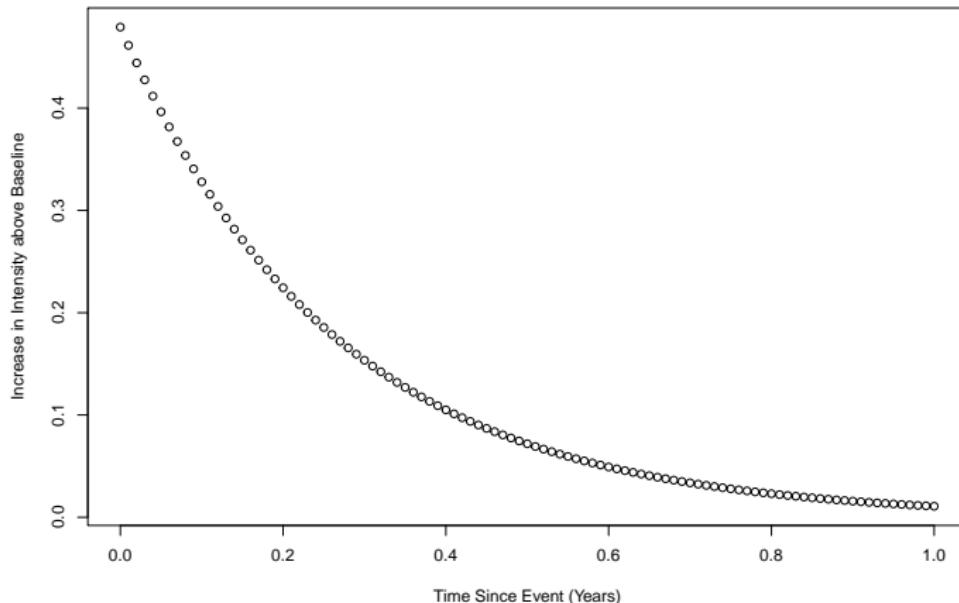
The Intensity Function over Time

- ▶ The intensity function gives the value of the rate at any time
- ▶ Here we simulate 100 years based on the fitted model.
 - ▶ The rate is at the baseline of 0.397 until a new event occurs
 - ▶ We see the jump in rate with each new event
 - ▶ The rate decays to baseline unless another event happens



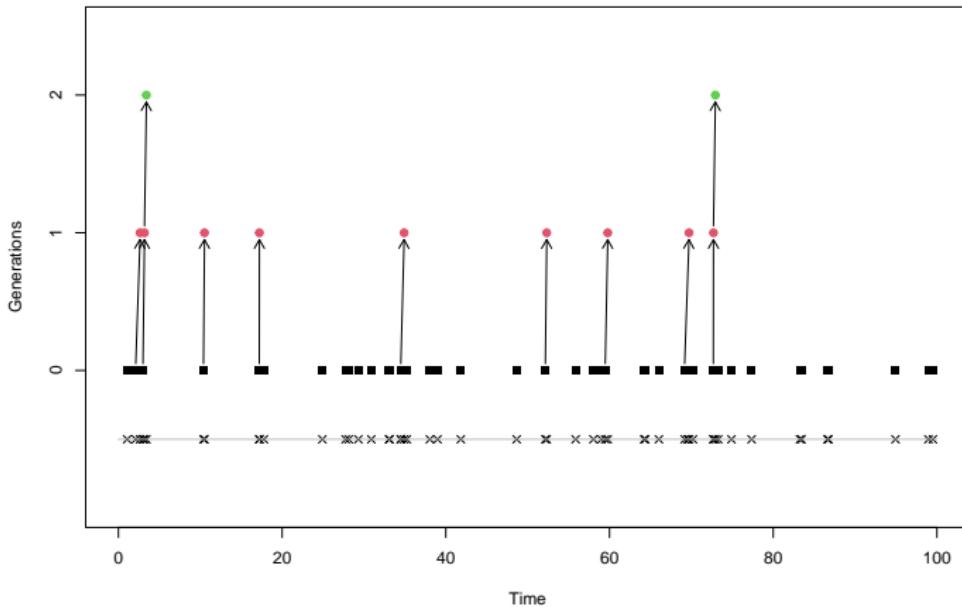
How Fast is the rate “decay”

- Within 1 year essentially return to baseline



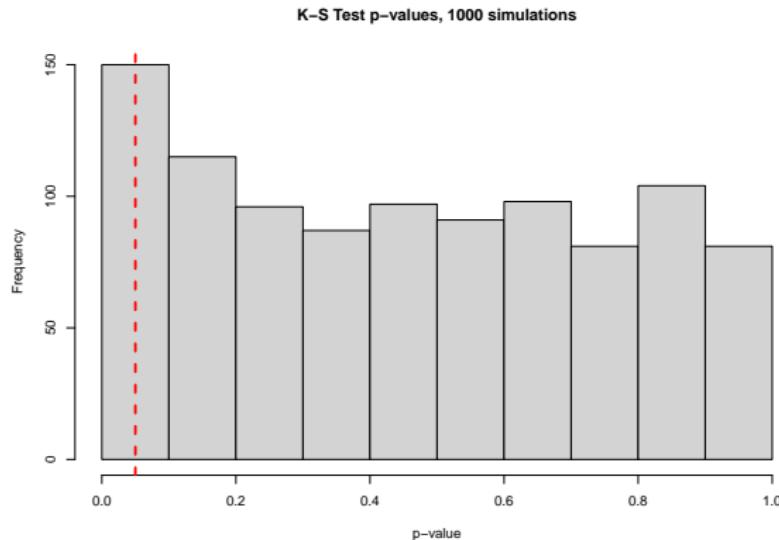
Process as “Generations”

- ▶ An event (1st generation) may spur immediate new event (2nd generation)... etc
- ▶ Cannot truly identify in our data; below is simulated



Would a Poisson Process model detect “excited” process?

- ▶ Simulated Hawkes process using estimates from our data
 - ▶ 1000 runs of 100 years each
- ▶ Fit exponential (constant) rate interarrival model
 - ▶ Tested fit using K-S Test



- ▶ Rejected fit in only 8.9 percent of simulations.

Inference for parameter estimates

- ▶ Ogata (1978) asymptotic result for the parameters (θ) :

$$M^{1/2}(\hat{\theta}_M - \theta_o) \xrightarrow{d} N(0, I(\theta_o)^{-1})$$

where M is the total time for the process and:

$$-M^{-1}H_M(\theta_o) \xrightarrow{p} I(\theta_o)$$

- ▶ We can estimate the Hessian, H_M , from ML estimation
 - ▶ Use the result to produce confidence intervals for the parameters.
 - ▶ Asymptotic result may not hold (Cavaliere et al, 2021)

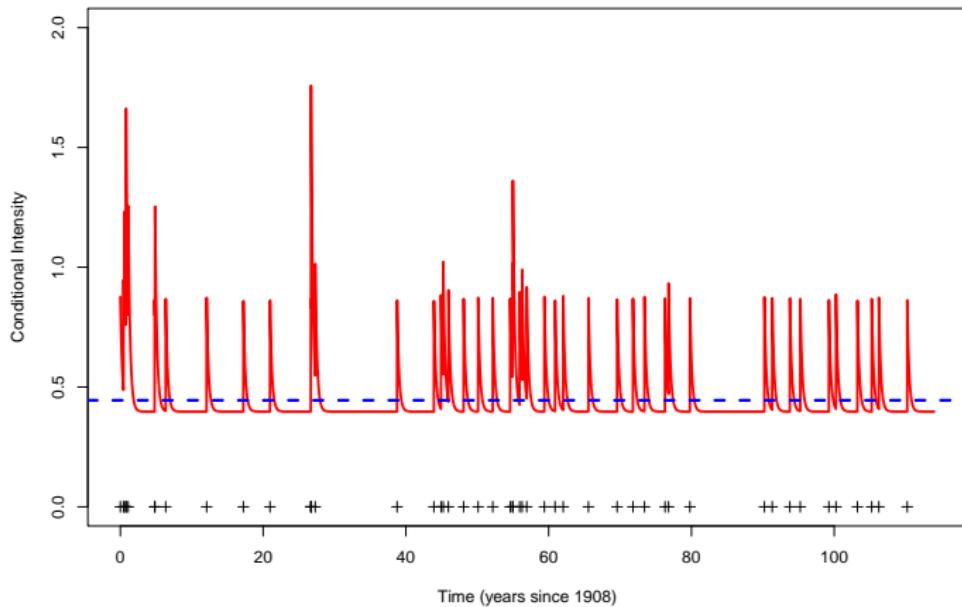
Estimated Ogata confidence intervals

95% confidence intervals based on Ogata for the three model parameters

- ▶ Baseline intensity estimate: 0.397
 - ▶ 95% CI: 0.253, 0.541
- ▶ Reproduction mean estimate: 0.126
 - ▶ 95% CI: -0.082, 0.334
- ▶ Exponential reproduction function rate estimate: 3.795
 - ▶ 95% CI: -7.193, 14.783

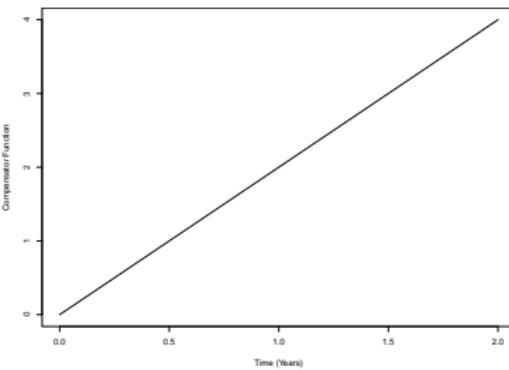
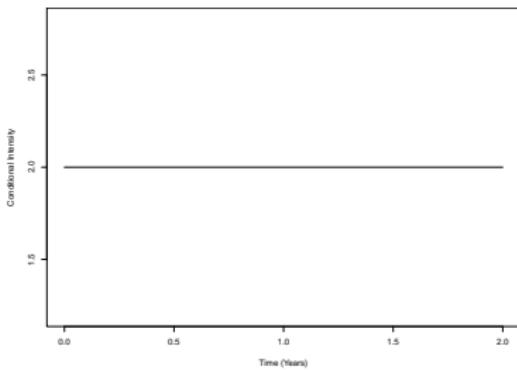
Intensity Functions for Both Models

- ▶ The Hawkes intensity function is the model applied to the actual data
- ▶ The function for the Poisson model is a constant rate of 0.445



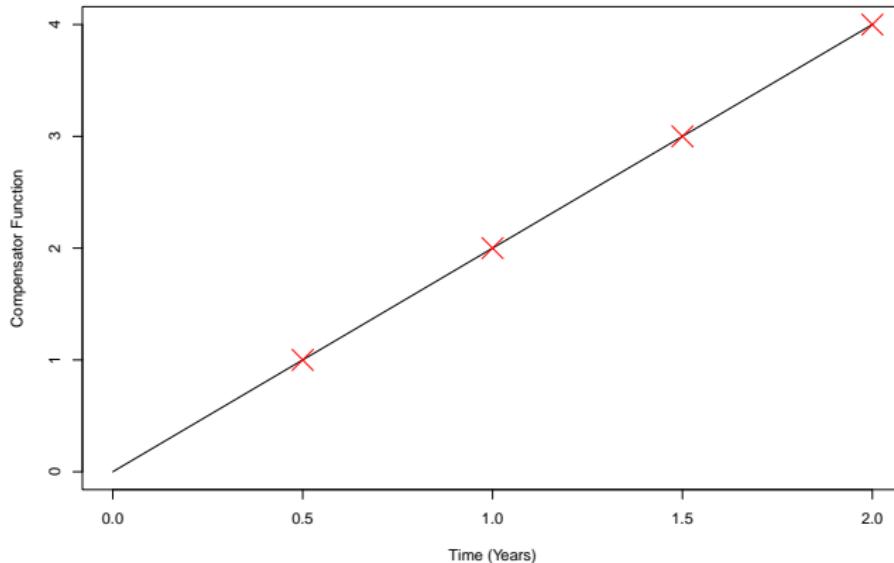
The Compensator Function

- ▶ Integrated intensity function... cumulative rate
- ▶ Simple example: constant intensity with rate 2 per year
- ▶ Compensator grows at constant rate (linear)



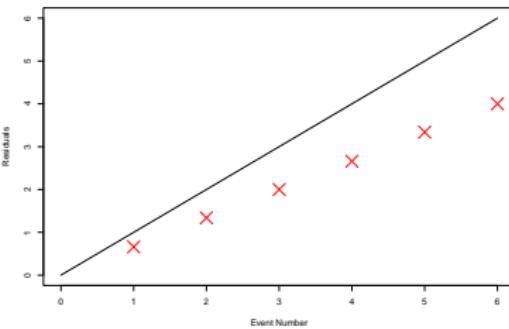
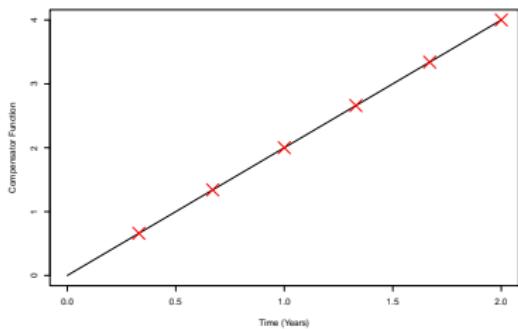
Arrivals Exactly on Schedule

- ▶ Two arrivals per year spaced perfectly at 6 months
- ▶ Values of compensator for points 1, 2, 3, 4 are 1, 2, 3, 4
- ▶ Compensator values at arrivals are “residuals” (Paparoditis, 2000)
 - ▶ if “on schedule” Poisson process with rate 1

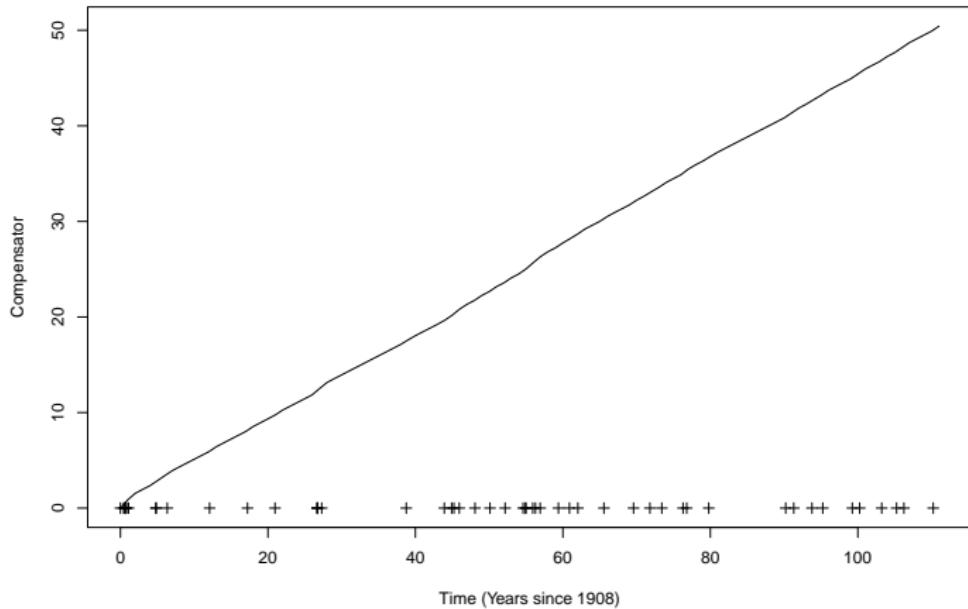


"Residuals" based on Compensator

- ▶ Suppose actual arrivals still constant but 3 per year (model is 2 per year)
- ▶ Residuals below the $y = x$ (Poisson rate 1) line

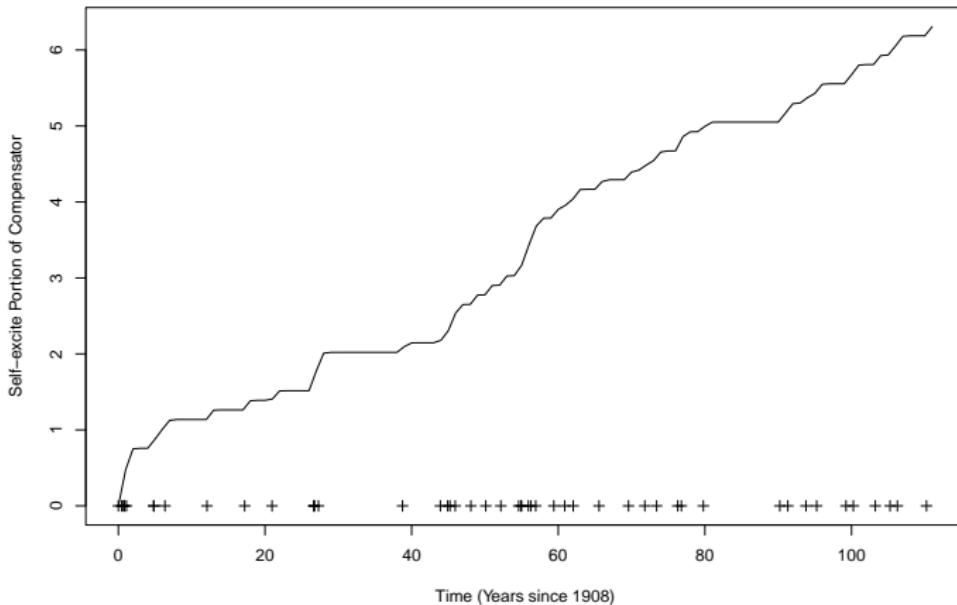


Compensator Function for Fitted Hawkes Model



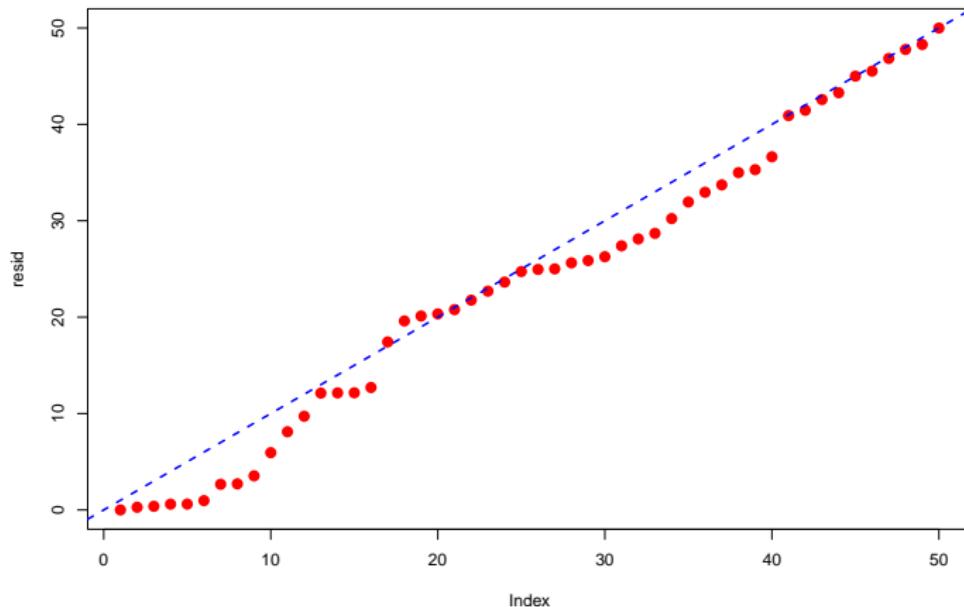
“Self Exciting” Portion of Compensator Function

- We remove the baseline (cumulative) rate to better see the Hawkes process

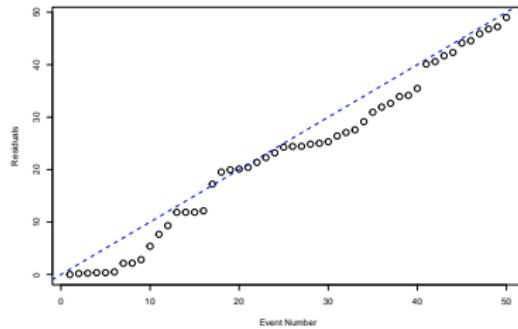
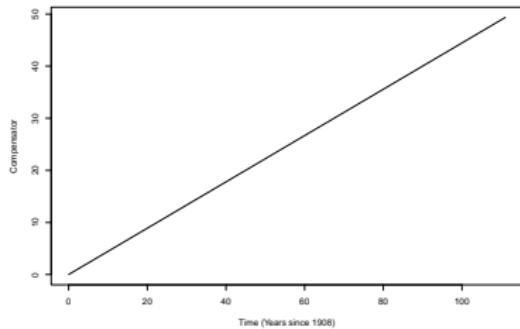


Residuals for Fitted Hawkes Model

- ▶ Arrivals at times “faster” than model
- ▶ Jump up event 17: 1947 record (after 12 year gap)
- ▶ Jump up event 41: 1998 record (after 10 year gap)

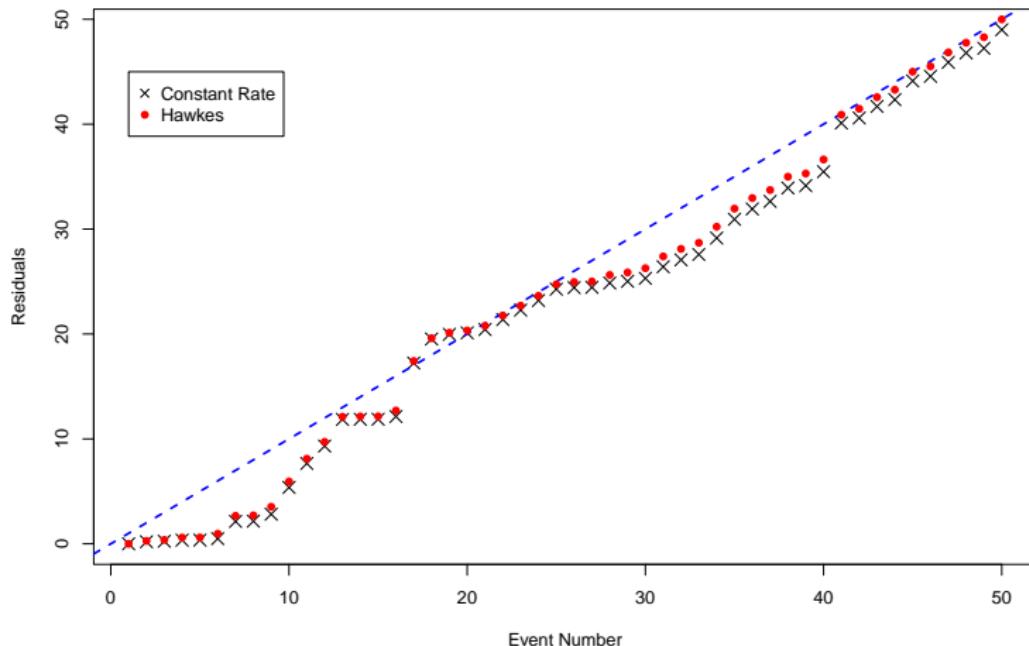


Constant Rate Model Compensator and Residuals



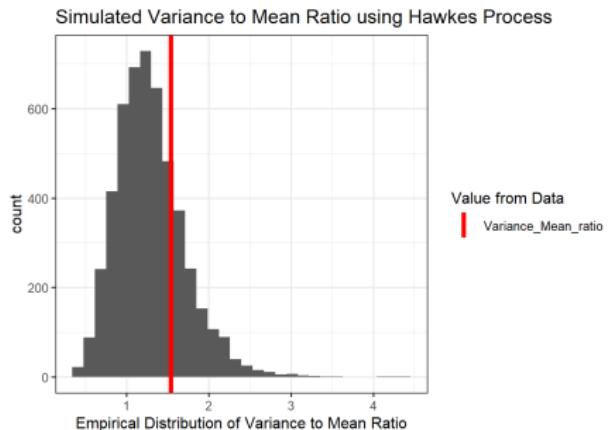
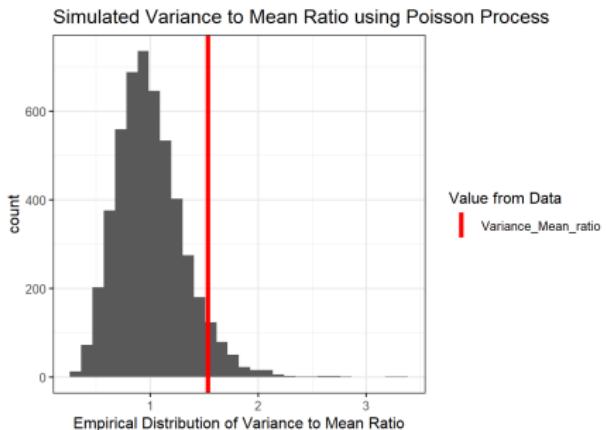
Comparing Model Residuals

- ▶ Hawkes generally “better”
- ▶ Two large gaps clear impact on both models
- ▶ Hard to tell degree of improvement



Comparing Models - Overdispersion

- ▶ “Posterior predictive P values” (Gelman et al, 1996)
- ▶ Simulate from based on the fitted model parameters
- ▶ Here we use the “overdispersion” (measured by variance/mean ratio)



Conclusions

- ▶ Men's marathon records reasonably modeled using Poisson process
- ▶ Some indication a “self-exciting” process could explain arrivals better
- ▶ Other factors not in model
 - ▶ As times get faster harder to break the record?
 - ▶ Two “outlier” periods impact the models
 - ▶ New technology (“super shoes”), training methods etc.
 - ▶ Magnitude of event: how much faster than previous (inverse?)

Further Work

- ▶ Explore models for other events: women's marathon, other distances, swimming
- ▶ Develop better metrics to compare and assess models

SCORE

- ▶ NSF Grant, 2022 - 2025
- ▶ SCORE with Data: Building a sustainable national network for developing and disseminating **S**ports **C**ontent for **O**utreach, **R**esearch, and **E**ducation in data science
- ▶ Carnegie Mellon, Baylor, West Point, University of Pittsburgh, St. Lawrence University, Yale
- ▶ Partners across sports industry
 - ▶ NBA, NFL, MLB, NHL teams
 - ▶ ESPN
 - ▶ Other sports related organizations (analytics, etc)

SCORE Goals

- ▶ Create a national network
 - ▶ academic, industry, media and government partners
 - ▶ elevating data science education
 - ▶ particularly in underrepresented populations and minorities
- ▶ Modules based on sports
 - ▶ Develop, implement, evaluate and disseminate an educational framework
 - ▶ Case-based Learning involving real-world problems and applications
- ▶ Educational research on data science education delivery modalities

SCORE Module

<https://isle.stat.cmu.edu/SCORE/marathons/>

The screenshot shows the "ATHLETE BIOS" section for Kate Sanborn. At the top, there's a logo for the "U.S. OLYMPIC TEAM TRIALS MARATHON FEBRUARY 29, 2020 ATLANTA, GA". Below it, a banner says "ATHLETE BIOS". A large red box highlights "KATE SANBORN". To the left, her bio includes: QT:2-HD4, Residence: Raleigh, NC, Homebase: Fayetteville, NC, College: United States Military Academy, North Carolina State University, Age on Race Day: 22, Affiliates: Raleigh Distance Project, Charlotte, Qualifying Race: Northern Marathon, Marathon: 2:35:18.4, Marathon Best: 2:34:40, Social Handles: Instagram/Twitter - @kate.sanb. Favorite inspirational quote: "For faith is a source of insurance." - Coach Mariano (her dad). On the right, there's a photo of Kate Sanborn in athletic gear standing in front of a red brick building with two towers.

Figure 7: Kate Sanborn, youngest qualifier US Olympic marathon trials, introduces module

The video player interface shows a woman speaking, identified as Clark Sturdivant. The video title is "Are We Living in an Age of Records?". The YouTube logo is visible at the bottom right of the video frame. The video has a play button icon in the center.

Are We Living in an Age of Records?

Sturdivant, Clark

Often it may seem that we see a lot of World Records happen all at the same time. But are we suffering from a [recency bias](#)? How would we even know?

DATA:

In this module we will explore the question, "Are Men's Marathon World Records Random Events?"

Below we can see the list of Men's Marathon World Records from Wikipedia.

A data visualization tool showing men's marathon world records. The interface includes a "Data" button and a "Rows: 50 (Total: 50)" indicator. The main area is a table with columns: ID, Time, Name, Nationality, Date, Event/Race, Source, Notes, and Time. Two rows of data are visible:

ID	Time	Name	Nationality	Date	Event/Race	Source	Notes	Time
1	2:35:18.4	Johnny Hayes	United States	July 24, 1988	London, United Kingdom	Wikipedia	Time was officially recor...	2:35:18.4
2	2:10:45.4	Robert Powell	United States	January 5, 1908	New York, United States	Wikipedia	None [2]	2:10:45.4

Figure 8: Students explore the data interactively in ISLE

Using the ISLE toolkit, explore the data.

Would it be appropriate to use a Normal Distribution to model this data?
Your answer:

Enter your answer here...

Show Solution

The Exponential Distribution

There are several steps one might use to determine if a model, such as exponential interarrivals, is appropriate. The exponential distribution has certain attributes, for example, if we say that $T \sim \text{Exp}(\lambda)$:

The expected value of T is:

And the Standard Deviation of T is:

SCORE Participation

- ▶ Participate as part of network
- ▶ Use and test modules
- ▶ Create new modules
 - ▶ Example modules soon available
 - ▶ “HandboK” with criteria in development
 - ▶ Peer-review similar to R package contribution

Contact us at:

Rod: Rodney_Sturdivant@baylor.edu

Nick: nicholas.clark@westpoint.edu

Notes

- ▶ Data sources:

Wikipedia scraped August 12, 2022

(https://en.wikipedia.org/wiki/Marathon_world_record_progression)

- ▶ Poisson process:

<https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459>

- ▶ Hawkesbow, R Package for Hawkes process:

Felix Cheysson (2021). hawkesbow: Estimation of Hawkes Processes from Binned Observations. R package version 1.0.2.

<https://CRAN.R-project.org/package=hawkesbow>

References

- Cavaliere, G., Lu, Y., Rahbek, A., and Østergaard, Jacob (2021). Bootstrap inference for Hawkes and general point processes. Discussion Paper Series, U. Copenhagen Economics. <http://dx.doi.org/10.2139/ssrn.3844552>
- Cheysson, F., and Lang, G. (2020). Strong mixing condition for Hawkes processes and application to Whittle estimation from count data. arXiv, March. <https://arxiv.org/abs/2003.04314>.
- Gelman, A., Meng, X.L. and Stern, H.S. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, 6, 733-807.
- Hawkes, Alan G. (1971). Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika* 58 (1): 83–90. <https://doi.org/10.2307/2334319>.
- Ogata, Y. (1978). The asymptotic behavior of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30(2), 243-261.
- Paparoditis, E. (2000). Spectral density based goodness-of-fit tests for time series models. *Scand. J. Stat.* 27 (1): 143–176. <https://doi.org/10.1111/1467-9469.00184>.