Multiple Changepoint Detection in Climate Time Series

Robert Lund
Clemson Math Sciences
Lund@Clemson.edu

Joint work with Shanghong Li, Yingbo Li, and Hewa Priyadarshani

June 13, 2017

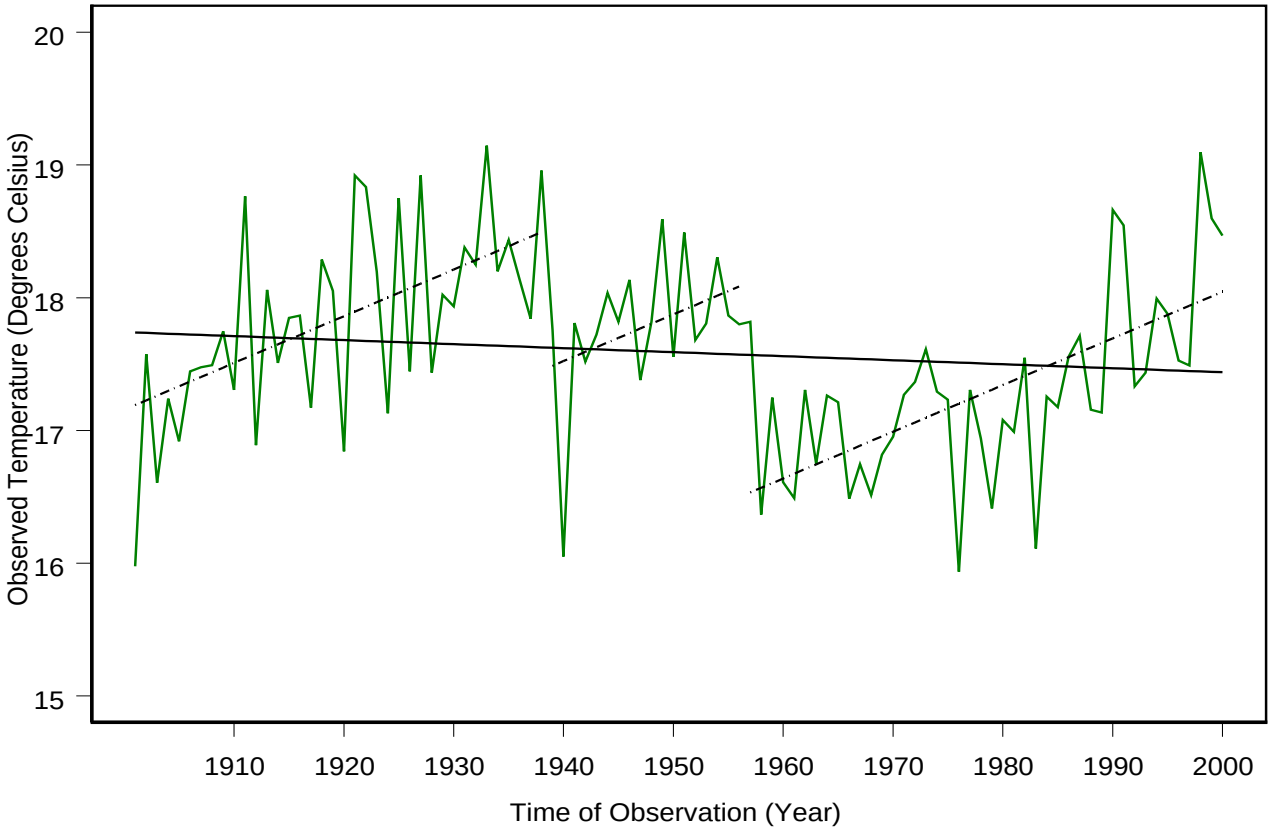
The Need to Detect Changepoints.

Changepoints are discontinuity times (inhomogeneities) in a time series. In climate settings, these can be induced from changes in observation locations, equipment, measurement techniques, environmental changes, etc.

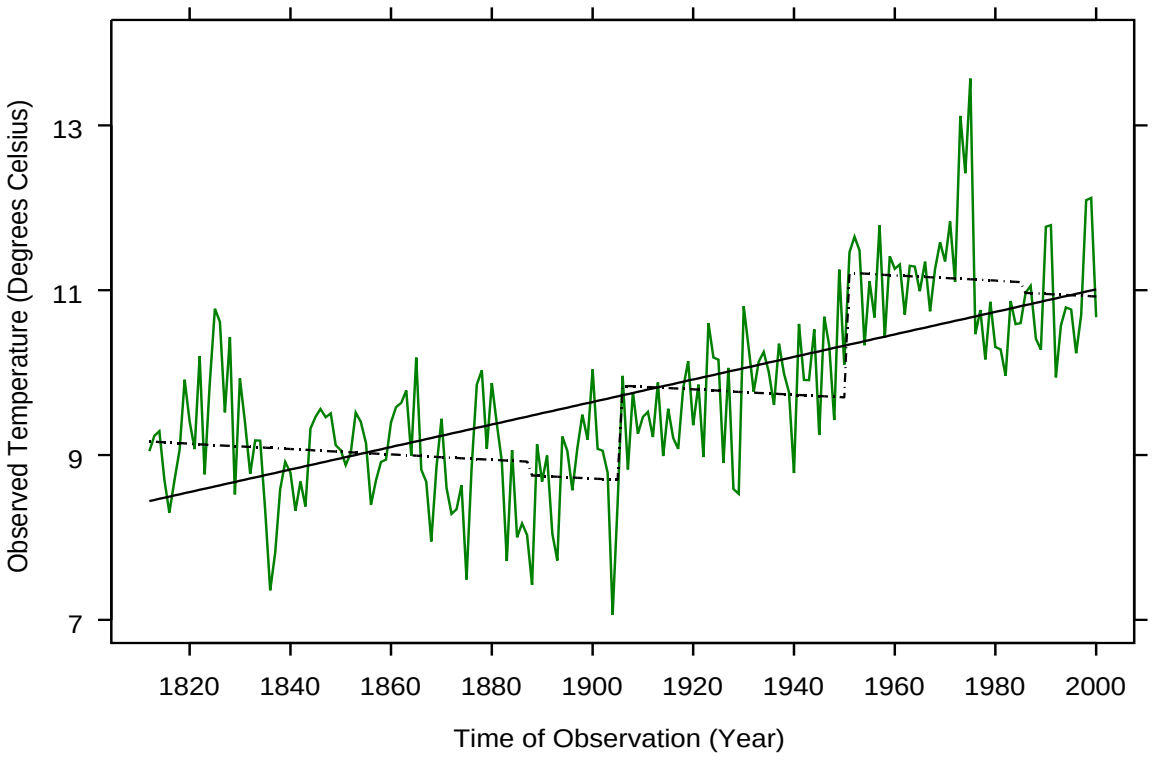
In this talk, a changepoint is a time where the mean of the series first undergoes a structural pattern change.

- Changepoint issues are critical when estimating trends.
- Many changepoints go undocumented.
- Changepoint techniques can help calibrate new gauges.

Tuscaloosa, AL Annual Temperatures



New Bedford, MA Annual Temperatures



Yearly Temperatures at New Bedford MA With Least Squares Trends

Key Questions

- How many changepoints are there?
- At what times do the changepoints occur?

Some recent penalized likelihood references:

1. Davis, Lee, and Rodriguez-Yam, *Journal of the American Statistical Association*, (2006).

2. Lu, Lund, and Lee, *Annals of Applied Statistics*, (2010).

3. Li and Lund, *Journal of Climate*, (2012).

Our Model

For annual ($T = 1$), monthly ($T = 12$), or daily ($T = 365$) data, our model for the data $\{X_t\}_{t=1}^{N}$ takes a time series regression:

$$X_{nT+\nu} = \mu_\nu + \alpha(nT + \nu) + \delta_{nT+\nu} + \epsilon_{nT+\nu}.$$

- The seasonal index $\nu \in \{1, \ldots, T\}$.
- μ_ν is the seasonal mean at season ν.
- α is a linear trend parameter, which may or may not be needed. Other trend functions are possible.

More on the Model

For annual ($T = 1$), monthly ($T = 12$) or daily ($T = 365$) data, our model for the data $\{X_t\}$ is a time series regression:

$$X_{nT+\nu} = \mu_\nu + \alpha(nT + \nu) + \delta_{nT+\nu} + \epsilon_{nT+\nu}.$$

The mean shifts are parametrized in $\{\delta_{nT+\nu}\}$:

$$\delta_t = \begin{cases} \Delta_1 = 0, & 1 \leq t < \tau_1, \\ \Delta_2, & \tau_1 \leq t < \tau_2, \\ \vdots & \\ \Delta_{m+1}, & \tau_m \leq t < \tau_{m+1}. \end{cases}$$

The errors $\{\epsilon_{nT+\nu}\}$ are a zero mean autoregressive process (this is periodic if $T > 1$).

The model for annual data

The model for annual data is

$$X_t = \mu + \alpha t + \delta_t + \epsilon_t.$$

- Location parameter: μ
- Linear trend: αt
- Piecewise constant mean shifts: δ_t
- Stationary but correlated errors $\{\epsilon_t\}$

Periodic Autoregressions

A zero-mean series $\{\epsilon_{nT+\nu}\}$ is called a periodic autoregression of order p (PAR(p)) if it satisfies the periodic linear difference equation

$$\epsilon_{nT+\nu} = \sum_{k=1}^{p} \phi_k(\nu)\epsilon_{nT+\nu-k} + Z_{nT+\nu}.$$

Here, $\{Z_{nT+\nu}\}$ is zero-mean periodic white noise with $\mathrm{Var}(Z_{nT+\nu}) = \sigma^2(\nu) > 0$ for all seasons ν.

$\phi_1(\nu), \ldots, \phi_p(\nu)$ are the PAR coefficients during season ν.

Such series are indeed "periodically stationary".

Penalized Likelihood Methods

A penalized likelihood for our model has form

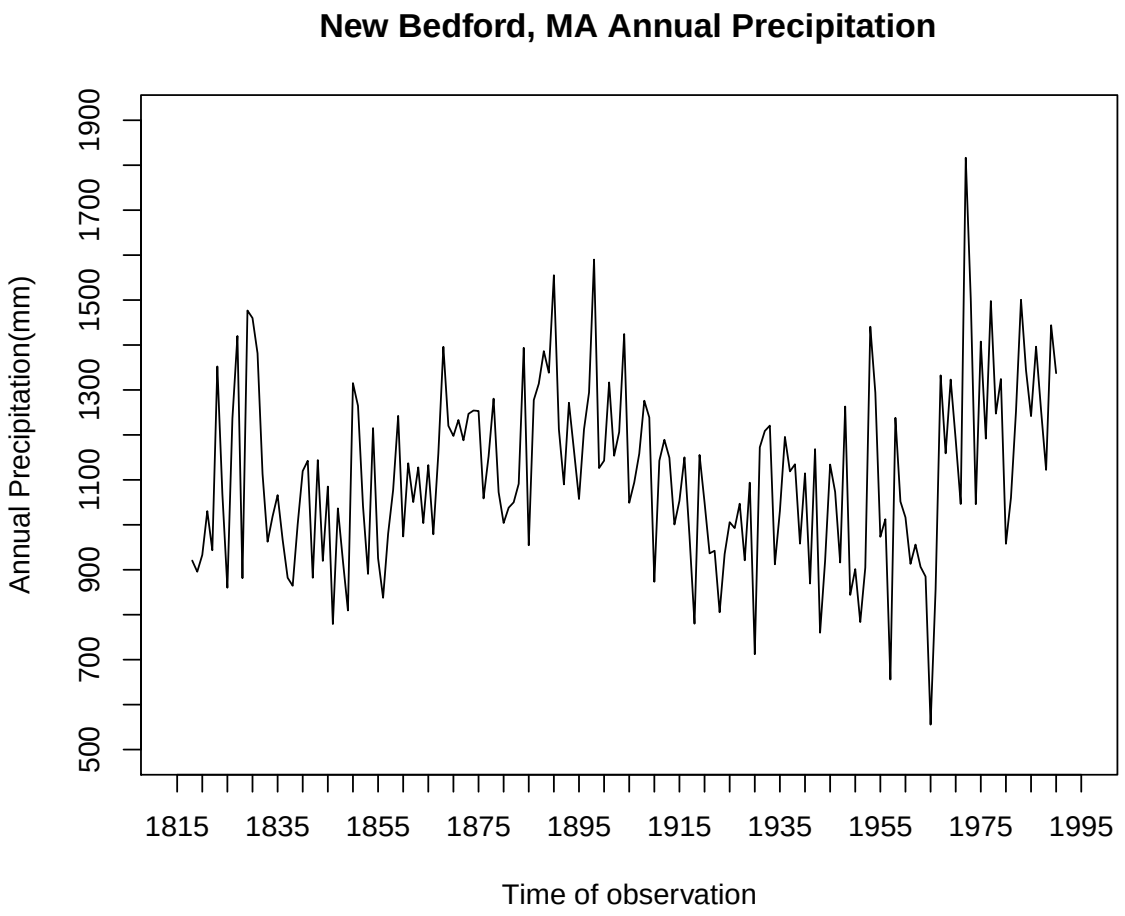
$$-\log(L^*(m, \tau_1, \ldots, \tau_m)) + \text{Penalty}(m, \tau_1, \ldots, \tau_m).$$

- $L^*(m, \tau_1, \ldots, \tau_m)$ is an optimized model likelihood given the changepoint count m and location times $\tau_1 < \cdots < \tau_m$.
- $\text{Penalty}(m, \tau_1, \ldots, \tau_m)$ is a penalty for the changepoint configuration.

Common $\text{Penalty}(m; \tau_1, \ldots, \tau_m)$ terms used:

- $\text{AIC} = 2m$.
- $\text{BIC} = m \ln(N)$.
- $\text{MDL} = \sum_{i=1}^{m+1} \ln(\tau_i - \tau_{i-1})/2 + \ln(m) + \sum_{i=2}^{m} \ln(\tau_i)$.

New Bedford, MA Annual Precipitations



Lognormal Annual Precipitation Setup

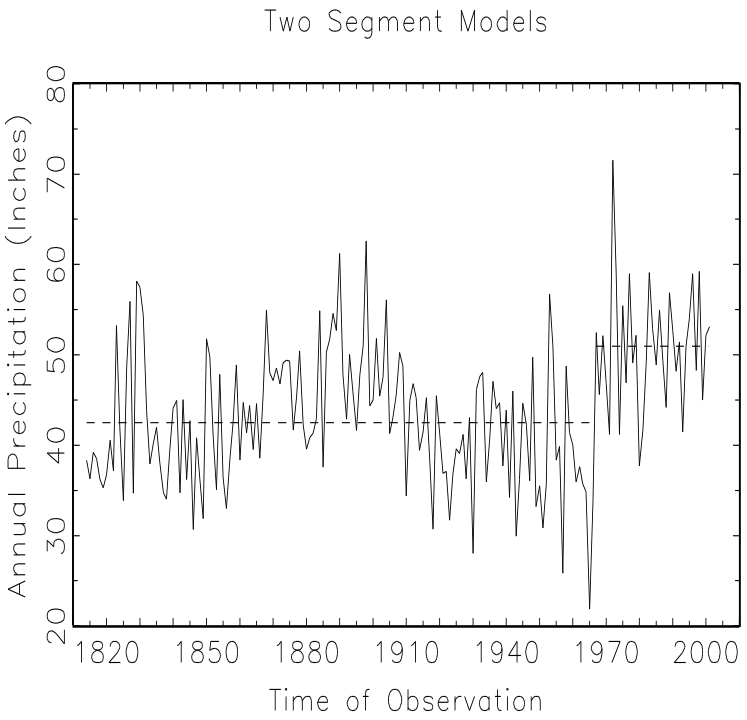
The logarithm of $\{X_t\}$ is modeled as a Gaussian time series with no trend, multiple mean shifts, and autoregressive errors (AR(p)). Here, $T = 1$: no periodicities.

For each changepoint configuration $(m; \tau_1, \ldots, \tau_m)$, we must

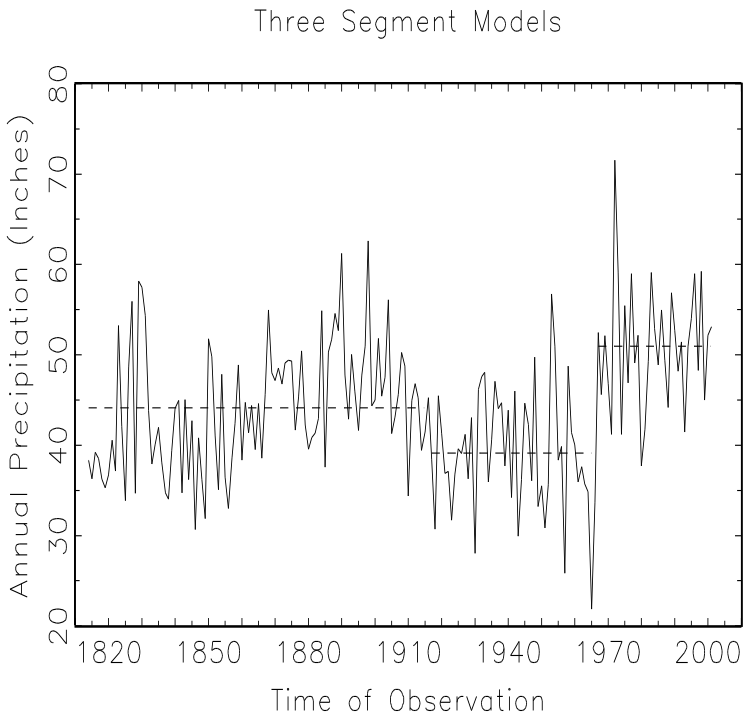
- Fit a time series model with optimal time series parameters and mean shift sizes.
- Compute the penalty

$$\mathrm{MDL}(m; \tau_1, \ldots, \tau_m) = \sum_{i=1}^{m+1} \ln(\tau_i - \tau_{i-1})/2 + \ln(m) + \sum_{i=2}^{m} \ln(\tau_i).$$

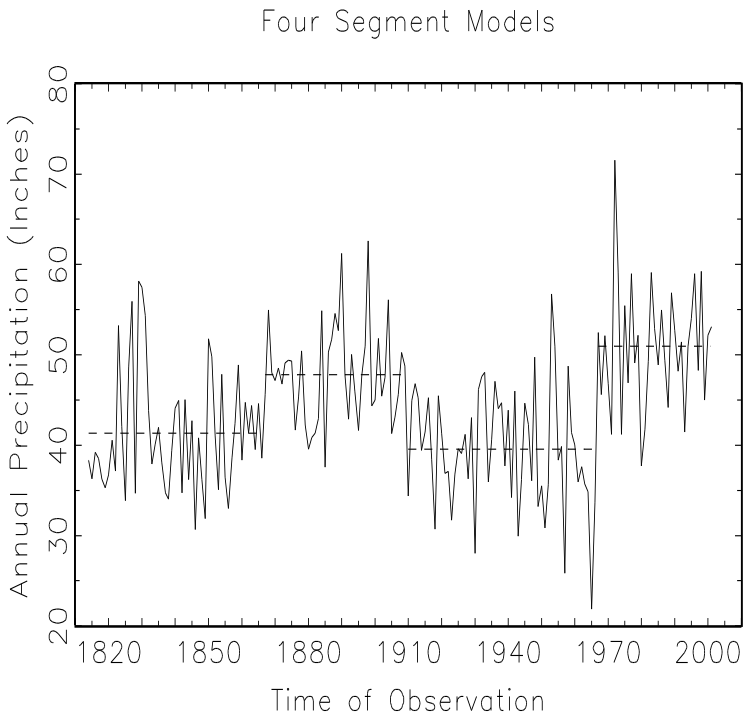
Two Segment Models



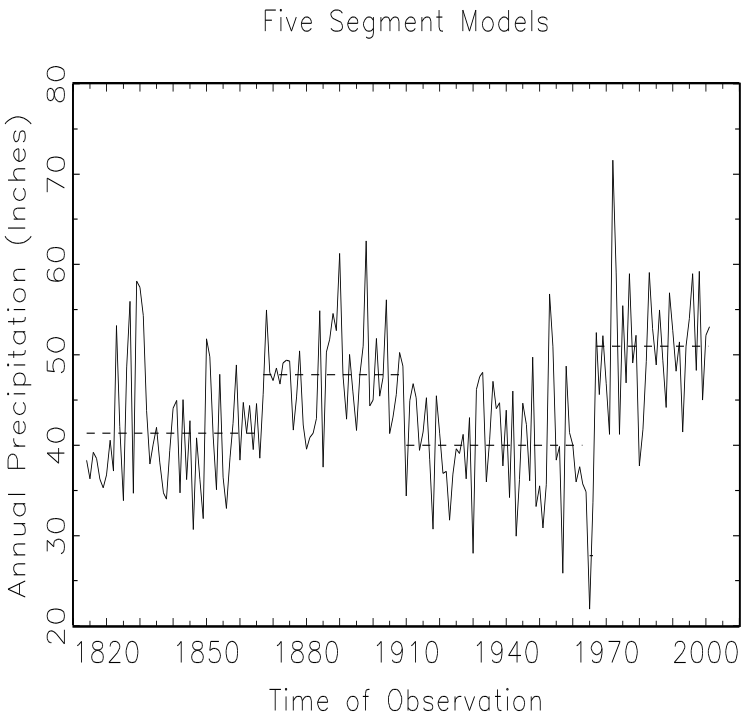
Three Segment Models



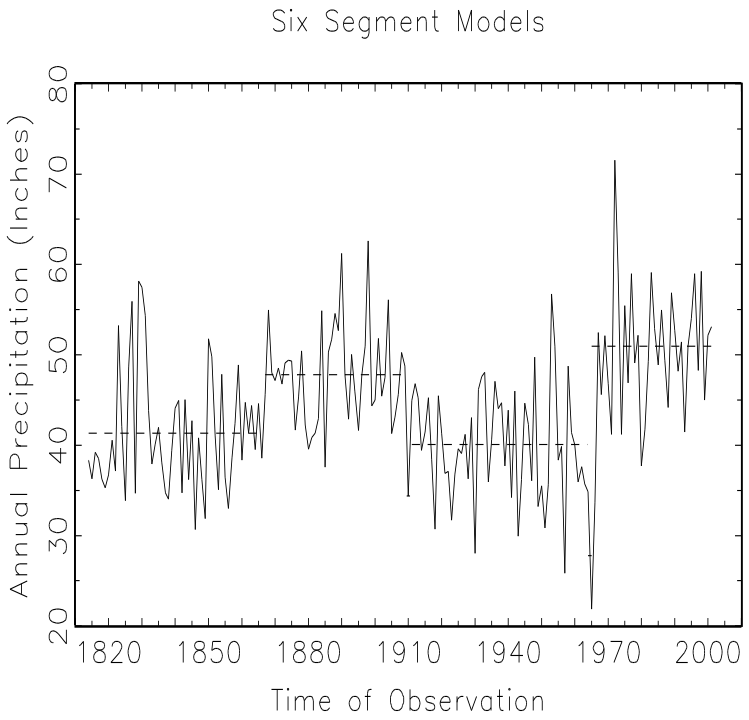
Four Segment Models



Five Segment Models



Six Segment Models



Summary

The table below shows optimum MDL scores for various numbers of model segments. These values were obtained by exhaustive search and are exact.

Table: Optimum MDL Scores

# Segments	Changepoint Times	MDL Score
1	—	-296.7328
2	1967	-303.8382
3	1917, 1967	-306.6359
4	1867, 1910, 1967	-309.2878
5	1867, 1910, 1965, 1967	-309.8570
6	1829, 1832, 1867, 1910, 1967	-308.2182

The Combinatorial Wall

We need to minimize

$$-\log(L^*(m, \tau_1, \ldots, \tau_m)) + \mathrm{MDL}(m, \tau_1, \ldots, \tau_m).$$

over all m and $\tau_1, \ldots, \tau_m$.

An exhaustive search over all models with m changepoints requires evaluation of $\binom{N}{m}$ MDL scores.

Summing this over $m = 0, 1, \ldots, N-1$ shows that an exhaustive optimization requires 2^{N-1} different MDL evaluations.

We now devise a genetic algorithm for this. A genetic algorithm is an intelligent random walk search.

Genetic Algorithms (GAs)

Chromosome Representation. Each changepoint configuration has the form $(m; \tau_1, \ldots, \tau_m)$.

Selection. Give mating preference to the fittest individuals, allowing them to pass their genes on to the next generation. Fitness is determined by the objective function

$$-\log(L^*(m; \tau_1, \ldots, \tau_m)) + \mathrm{MDL}(m; \tau_1, \ldots, \tau_m).$$

GA Details

Chromosome Crossover. Two individuals are chosen as parents from the current generation to breed — call these $(m; \tau_1, \ldots, \tau_m)$ and $(k; \eta_1, \ldots, \eta_k)$. A new chromosome $(\ell; \xi_1, \ldots, \xi_\ell)$, having traits of both parents, is created from these individuals to produce children for the next generation.

Mutation. Increases the diversity of the population, preventing premature convergence to poor solutions. Our mutation mechanism allows a small portion of generated children to have extra changepoints. After each child is formed from its parents, each and every non-changepoint time is independently allowed to become a changepoint time with probability p_m. Typically, p_m is small: $p_m = 0.003$.

GA Details

Algorithm Termination. Successive generations are simulated until a termination condition has been reached. Common terminating conditions are:

- A solution is found that satisfies a minimum criteria.
- A fixed number of generations is reached.
- The generation's fittest ranking member is peaking (successive iterations no longer produce better results).

The fittest member of the terminating generation is deemed as the solution.

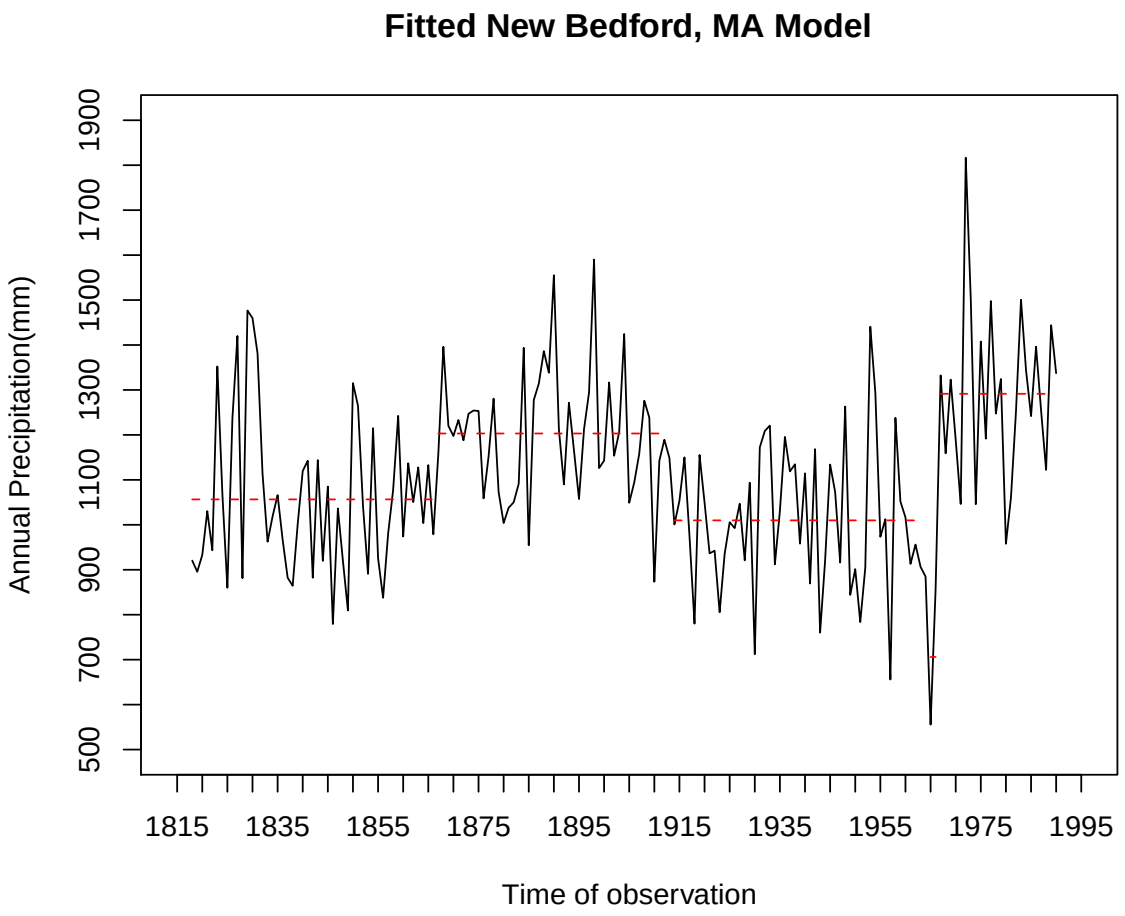
Optimal Model

The GA algorithm converged to a model with four changepoints at times $1867, 1910, 1965,$ and 1967.

The minimum MDL score achieved was -309.8570.

This segmentation is graphed against the data and appears visually reasonable.

Optimal Model Has Four Changepoints!



Simulations — Set I

Mimics the New Bedford Data with lognormal distributions:

1000 series of length $N = 200$ with no trend, seasonality, or changepoints: $\mu_t \equiv 6.8$. AR(1) errors $\{\epsilon_t\}$ with $\phi = 0.2$ and $\sigma^2 = 0.025$.

Table: Empirical proportions of estimated changepoint numbers. The correct value of m is zero.

m	Percent
0	99.0 %
1	0.4 %
2	0.5 %
3+	0.1 %

Simulations — Set II

$$\mu_t = \begin{cases} 6.8 & 1 \leq t \leq 49 \\ 7.0 & 50 \leq t \leq 99 \\ 7.2 & 100 \leq t \leq 149 \\ 7.4 & 150 \leq t \leq 200 \end{cases} .$$

Table: Empirical proportions of estimated changepoint numbers ($m = 3$)

m	Percent
0	0.0 %
1	3.6 %
2	28.8 %
3	63.1 %
4	4.3 %
5+	0.2 %

Count Detection Histogram

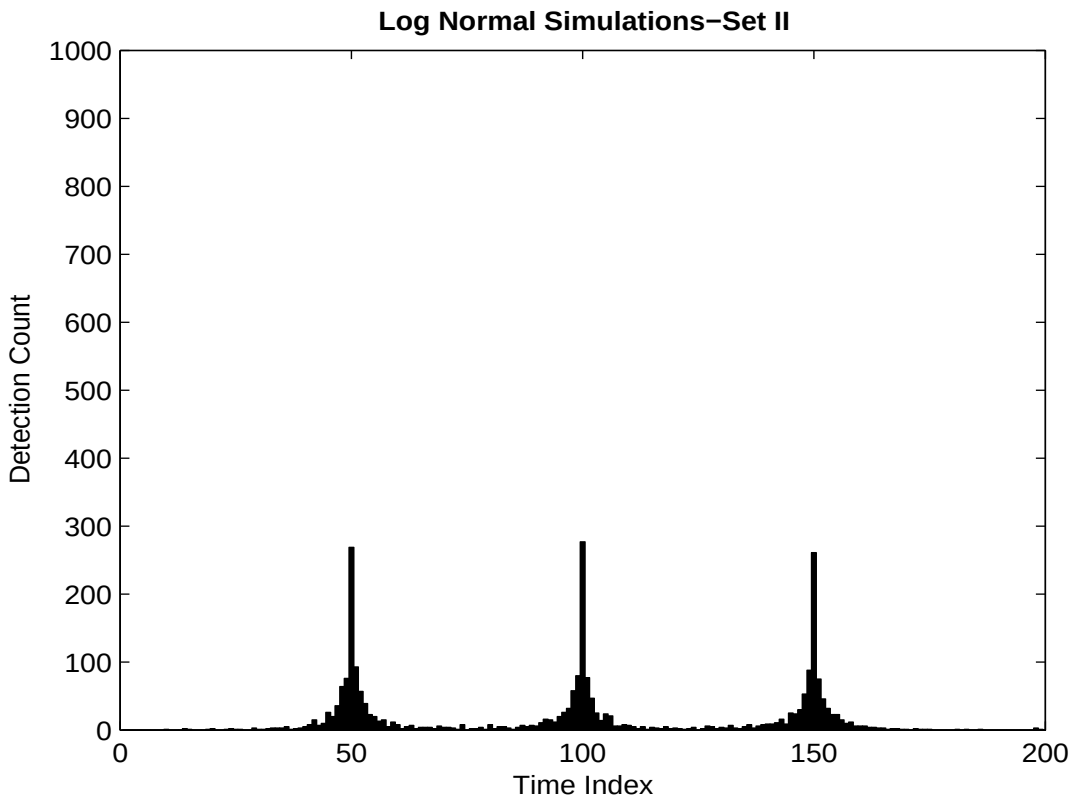


Figure: The detected changepoint times cluster around their true times of 50, 100, and 150.

Simulations — Set III

$$\mu_t = \begin{cases} 6.8 & 1 \leq t \leq 24 \\ 7.0 & 25 \leq t \leq 74 \\ 6.6 & 75 \leq t \leq 99 \\ 6.8 & 100 \leq t \leq 200 \end{cases} .$$

Table: Empirical proportions of estimated changepoints ($m = 3$)

m	Percent
0	0.0 %
1	6.0 %
2	19.5 %
3	69.2 %
4	5.1 %
5+	0.2 %

Count Detection Histogram

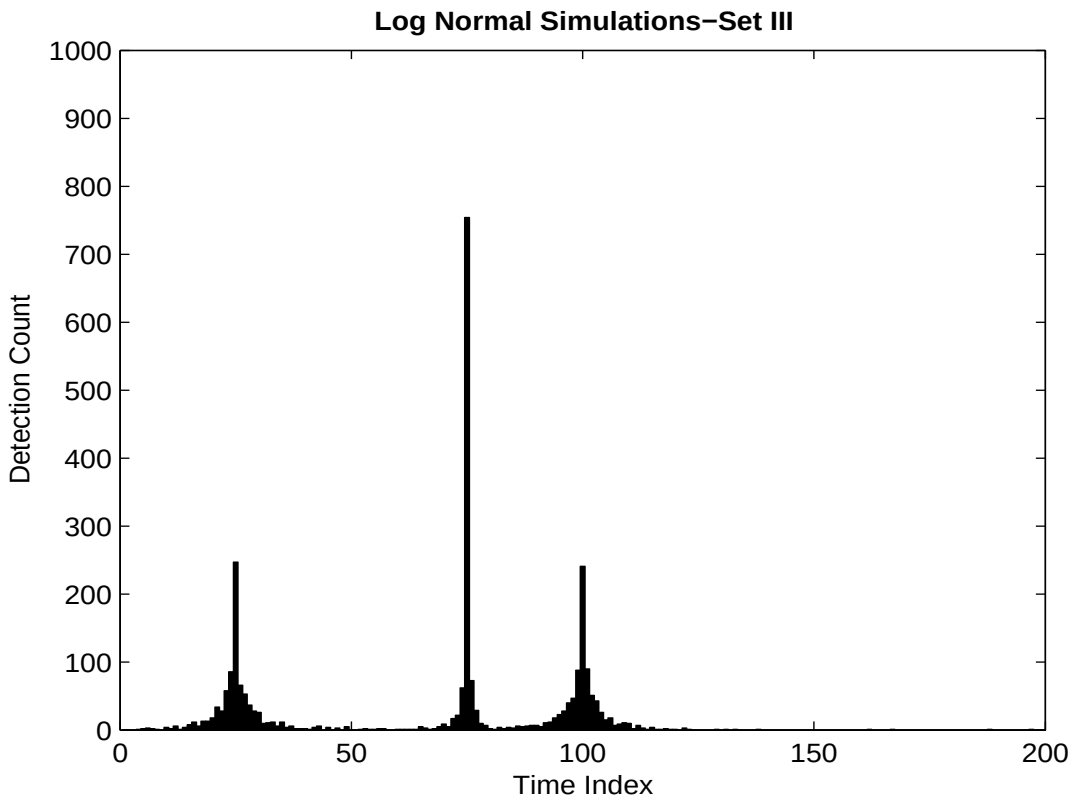


Figure: The detected changepoint times cluster around their true times of 25, 75, and 100.

Thank you! ☺