Changepoints and Associated Climate Controversies

Robert Lund
Clemson Math Sciences
Lund@Clemson.edu

Coconspirators:
Alexander Aue, Colin Gallagher, Jaechoul Lee, Thomas Lee, Shanghong Li, Yingbo Li, QiQi Lu, Jaxk Reeves, Michael Robbins

May 12, 2015

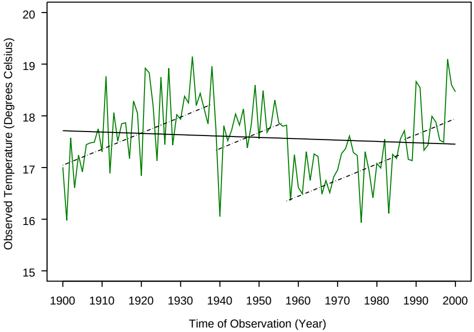
What is a Changepoint?

A changepoint is a discontinuity in the marginal distributions of a time-ordered sequence of data $X_1, X_2, \ldots$.

H_0 : The entire data sequence $\{X_i\}_{i=1}^{n}$ behaves via one model.

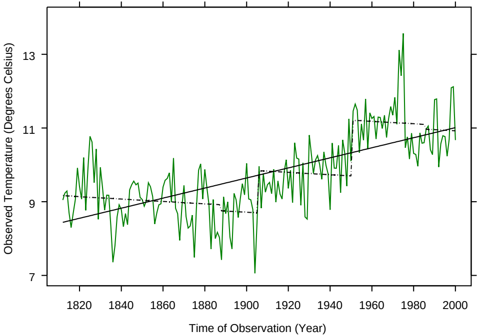
H_A : There is an unknown time c such that $\{X_i\}_{i=1}^{c}$ behaves via one model and $\{X_i\}_{i=c+1}^{n}$ behaves via a different model.

Changepoints Wreak Havoc on Inferences



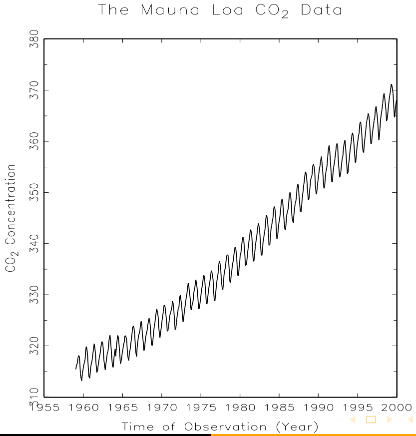
Yearly Temperatures at Tuscaloosa AL With Least Squares Trends

More Changepoint Shenanigans

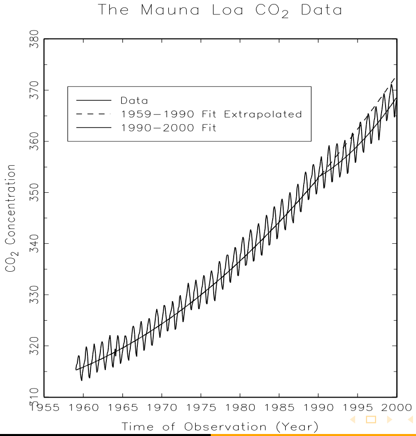


Yearly Temperatures at New Bedford MA With Least Squares Trends

Changepoints Arise in Unexpected Ways



See the Changepoint Now?



An AMOC Analyses

We present an asymptotic analysis for changes in the mean of a series for the at most one changepoint (AMOC) case.

While AMOC settings are seldom reality, climatologists believe that they can subsegment series down to those with either zero or one changepoint(s).

If the time of the change is known a priori, then changepoint tests are relatively simple, comprised largely of the usual t and z tests learned in a first statistics class.

Classical Background

The simplest regression setup is arguably

$$X_t = \mu + \Delta 1_{[t>c]} + \epsilon_t, \qquad 1 \leq t \leq n.$$

Take $\{\epsilon_t\}$ IID$(0, \sigma^2)$ for the moment; c is unknown.

The goal is to assess whether or not $\Delta = 0$, or whether $c = n$.

A CUSUM-based Changepoint Test

If the changepoint time c were known (documented), a simple test would compare differences in the means

$$\bar{X}_{1:c} = \frac{1}{c}\sum_{t=1}^{c} X_t \quad \text{and} \quad \bar{X}_{c+1:n} = \frac{1}{n-c}\sum_{t=c+1}^{n} X_t.$$

Weighting for the different segment sizes leads to consideration of

$$Z(c) = \frac{\bar{X}_{1:c} - \bar{X}_{c+1:n}}{\text{Var}(\bar{X}_{1:c} - \bar{X}_{c+1:n})^{1/2}}.$$

CUSUM Statistics

This entails examining the statistic

$$Z(c) = \frac{\text{CUSUM}(c)}{\sigma\sqrt{\frac{c}{n}\left(1-\frac{c}{n}\right)}},$$

where

$$\text{CUSUM}(c) = \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{c} X_t - \frac{c}{n}\sum_{t=1}^{n} X_t\right).$$

When c is unknown, we examine

$$Z_{\max}^2 = \max_{1 \le c \le n} Z^2(c).$$

CUSUM Asymptotics

Under the null hypothesis of no changepoints, as $n \to \infty$,

$$Z_{\max}^2 \to \infty.$$

Q: How does one make sense of this?

A: Truncate Boundaries!

CUSUM Asymptotics

Truncate the set of admissible changepoint times to all c such that $c/n \in [\ell, h] \subset (0,1)$.

Under a null hypothesis of no changepoints, as $n \to \infty$, MacNeill (1974, *Annals of Statistics*) shows that

$$\max_{\ell \leq \frac{c}{n} \leq h} Z^2(c) \xrightarrow{\mathcal{D}} \sup_{\ell \leq t \leq h} \frac{B^2(t)}{t(1-t)}.$$

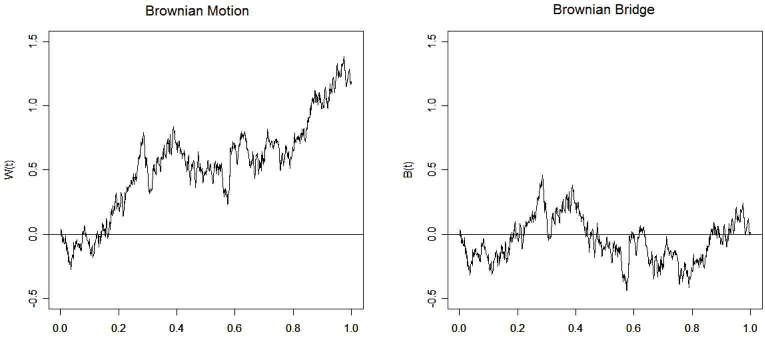
Here, $\{B(t)\}_{t=0}^{t=1}$ is a standard Brownian bridge process.

A Reminder on Brownian Bridges

If $\{W(t)\}_{t=0}^{t=1}$ is a Brownian motion, then

$$B(t) = W(t) - tW(1)$$

is a Brownian bridge over $t \in [0, 1]$.



Atlantic Basin Tropical Cyclone Counts and Strengths

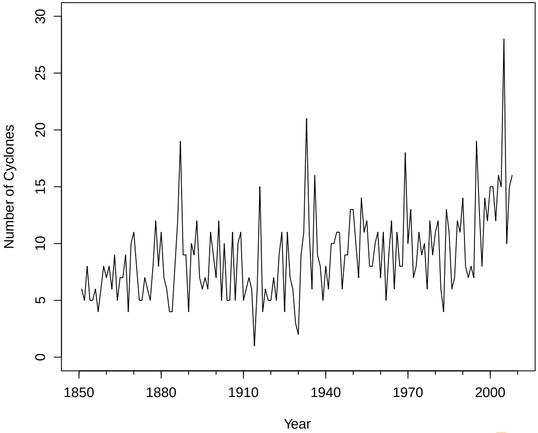
There is considerable controversy over whether or not tropical cyclone counts are increasing and/or if the strengths of the individual storms are increasing.

July 28, 2009 Senate testimonial from Dr. Kelvin Droegemeier (a climatologist from University of Oklahoma): North Atlantic tropical cyclone counts are not increasing but the individual strengths of the storms are.

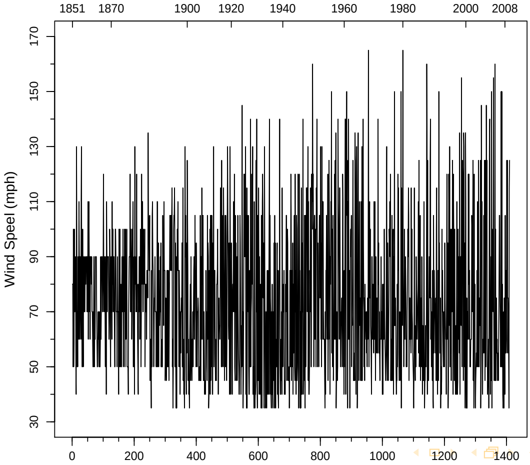
Hurricane data: HURDAT on NOAA's website. This has information on about 1500 tropical cyclones occurring from 1851-current.

The data, especially in regard to windspeed of the storms, can be unreliable.

Annual Number of Observed Cyclones



Peak Storm Windspeeds



Changepoint Tests for a Distributional Change

We now study AMOC techniques for marginal distributions.

The null hypothesis is that $X_1, \ldots, X_n$ (the storm windspeeds) are IID.

Our alternative hypothesis is that there is an unknown changepoint time c at which time the CDF of the X_is shift in an unknown way.

We devise a non-parametric test via χ^2 statistics.

Partition the X_t into the cells $\mathcal{I}_1, \ldots, \mathcal{I}_m$.

The $\chi^2_{\max}$ Statistic

If a changepoint occurred at time c, then

$$\chi^2(c) = \sum_{i=1}^{m} \frac{\left(O_{i,c} - \widehat{E[O_{i,c}]}\right)^2}{\widehat{E[O_{i,c}]}} + \sum_{i=1}^{m} \frac{\left(O^*_{i,c} - \widehat{E[O^*_{i,c}]}\right)^2}{\widehat{E[O^*_{i,c}]}}$$

should be statistically large — the two-sample χ^2 statistic.

Here, $O_{i,c}$ and $O^*_{i,c}$ are the observed category i counts before and after the changepoint time c.

We hence examine

$$\chi^2_{\max} = \max_{\ell \leq \frac{c}{n} \leq h} \chi^2(c).$$

A Categorical Asymptotic Theorem

(Robbins, Lund, Gallagher, Lu 2011, *Journal of the American Statistical Association*) Under the null hypothesis of no changepoints,

$$\max_{\ell \leq \frac{c}{n} \leq h} \chi^2(c) \xrightarrow{\mathcal{D}} \sup_{\ell \leq t \leq h} \frac{B_1^2(t) + \cdots + B_{m-1}^2(t)}{t(1-t)}.$$

Here, $B_1^2(t) + \cdots + B_m^2(t)$ is the sum of m independent squared Brownian bridges.

Boundaries must again be cropped.

Our “Saffir-Simpson” Partition of Windspeeds

Category 1: 40-73 mph (tropical cyclones)

Category 2: 74-95 mph hurricanes

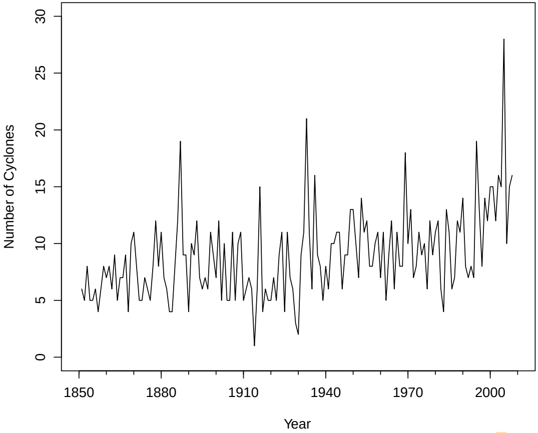
Category 3: 96-110 mph hurricanes

Category 4: 111-130 mph hurricanes

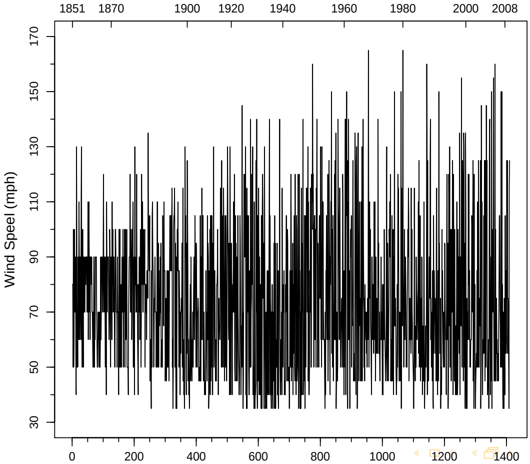
Category 5: 131+ mph hurricanes

We added some bells and whistles to the categorical changepoint test to jointly find changepoints in the above categories and, simultaneously, annual counts.

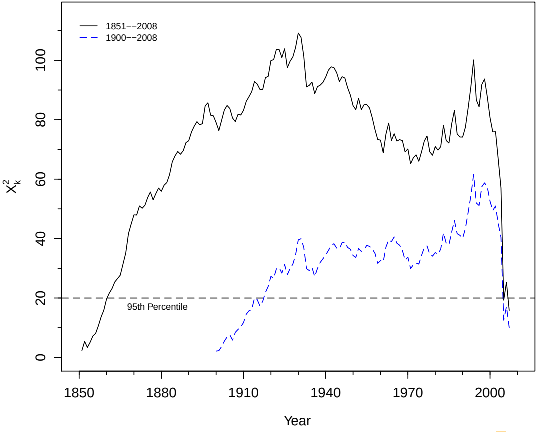
Annual Number of Observed Cyclones



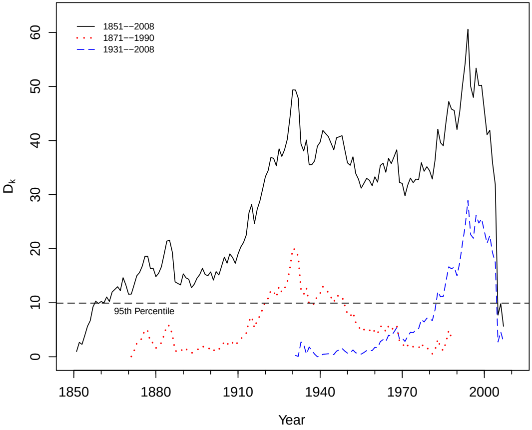
Peak Storm Windspeeds



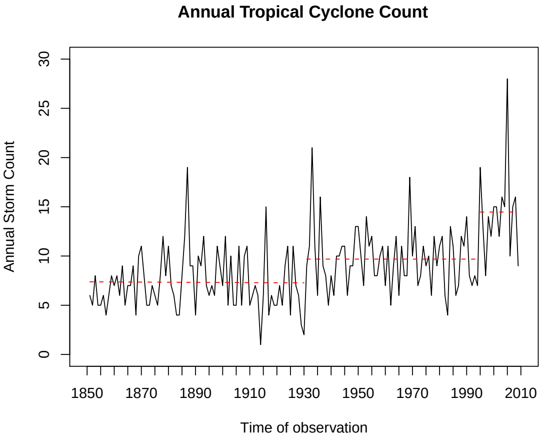
Count and Windspeed Changepoint Test (Joint)



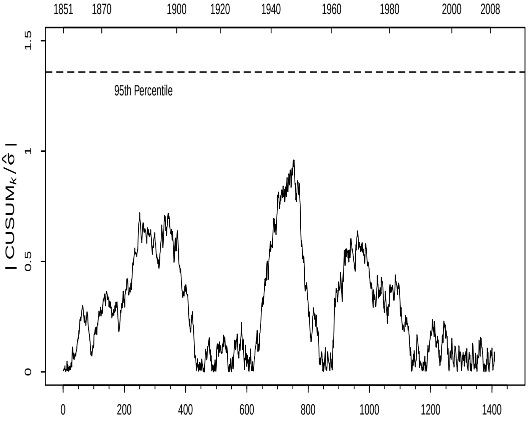
Counts Only Changepoint Tests



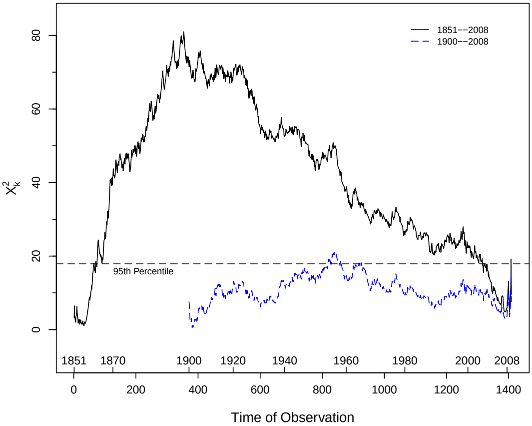
Optimal Count Segmentation



CUSUM Windspeed Changepoint Test



$\chi^2_{\max}$ Test for Windspeed Changepoints



What About Recent Counts?

Hurricane Conclusions and Comments

Hurricane counts have shown recent increases (circa 1995).

Windspeeds of the storms have not recently increased.

Not shown here: Almost every variable in this data set (longitude, latitude, duration,....) shows a changepoint around 1960.

The circa 1995 changepoint is hotly debated and is the subject of the 2006 popular book “Storm World” by Chris Mooney.

Key Questions

How many changepoints (call it m) are there?

Where are the m changepoints located — call them $\tau_1 < \tau_2 < \cdots < \tau_m$?

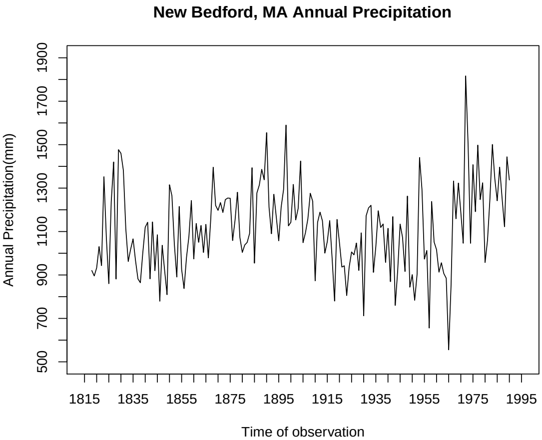
Three recent non-Bayesian references:

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).

New Bedford, MA Annual Precipitations



Minimum Description Length (MDL) Methods

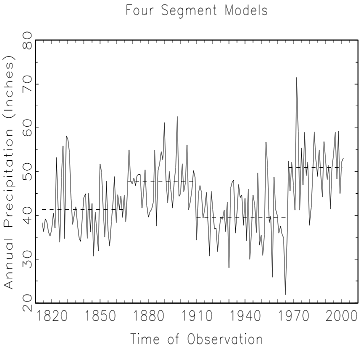
The MDL Objective Function:

$$\text{MDL}(m, \tau_1, \ldots, \tau_m) = -\log_2(L^*) + P(m, \tau_1, \ldots, \tau_m).$$

$L^* = L^*(m, \tau_1, \ldots, \tau_m)$ is an optimized model likelihood given the changepoint numbers m and location times $\tau_1 < \cdots < \tau_m$.

$P(m, \tau_1, \ldots, \tau_m)$ is a penalty for the number(s) and type(s) of model parameters and the changepoint configuration.

A Teaser for the Fit



The MDL Criterion

MDL methods penalize integer-valued parameters more heavily than real-valued parameters. MDL methods fundamentally differ from AIC and BIC methods, which penalize total parameter counts only.

- The penalty for a real-valued parameter estimated from k data points is $\log_2(k)/2$.
- The penalty for an unbounded integer I is $\log_2(I)$.
- The penalty for an integer parameter I that is known to be bounded by an integer B is $\log_2(B)$.
- The total penalty P is the sum of penalties for each parameter.

The Lognormal Likelihood

Lognormal distributions often describe annual precipitations well. Annual precipitation series often display correlation.

A reasonable model for annual precipitation series $\{X_t\}$ has

Lognormal marginal distributions.

A location parameter μ that shifts at each of the m changepoint times $\tau_1 < \cdots < \tau_m$.

A scale parameter σ that is constant over all regimes.

Lognormal Likelihood

The marginal density of X_t is

$$f(x) = \frac{\exp\left\{-\frac{(\ln(x)-\mu_{r(t)})^2}{2\sigma^2}\right\}}{x\sigma\sqrt{2\pi}}, \qquad x > 0.$$

Here, $r(t)$ denotes the time t regime number.

If X_t is independent in time t, the likelihood L of all N observations is

$$L = \prod_{t=1}^{N} f(X_t) = \frac{\exp\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(\ln(X_t) - \mu_{r(t)})^2\}}{(\sigma\sqrt{2\pi})^N \left(\prod_{t=1}^{N} X_t\right)},$$

Lognormal Parameter Estimators

For known changepoint numbers m and times at $\tau_1 < \cdots < \tau_m$, likelihood parameter estimators are

$$\hat{\mu}_\ell = \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{t \in R_\ell} \ln(X_t),$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} (\ln(X_t) - \hat{\mu}_{r(t)})^2.$$

Plugging in these optimizers gives $L^*(m, \tau_1, \ldots, \tau_m)$.

The Lognormal Likelihood

Because precipitation data is correlated, we modify the above to allow for AR(1) autocorrelation with lag-one autocorrelation parameter ϕ in (-1,1):

$$\ln(X_t) = \phi \ln(X_{t-1}) + Z_t.$$

Here, $\{Z_t\} \sim \text{WN}(0, \sigma^2)$.

The parameter estimators are more involved, but similar.

Again, we get $L^*(m, \tau_1, \ldots, \tau_m)$.

The MDL Penalty

An MDL penalty is obtained by adding penalties for each model parameter.

Real-valued parameters: the penalty for μ_ℓ is $\log_2(\tau_\ell - \tau_{\ell-1})/2$; the penalty for ϕ and σ^2 is $\log_2(N)/2$

Integer-valued parameters: the penalty for the number of segments is $\log_2(m+1)$; the penalty for τ_i is $\log_2(\tau_{i+1})$ since $\tau_i < \tau_{i+1}$.

The MDL Penalties

The MDL penalty is hence

$$P = 2\log_2(N) + \sum_{i=1}^{m+1} \frac{\log_2(\tau_i - \tau_{i-1})}{2} + \log_2(m+1) + \sum_{i=2}^{m} \log_2(\tau_i).$$

and the objective function is

$$\mathrm{MDL} = \frac{N}{2}\ln(\hat{\sigma}^2) + \sum_{i=1}^{m+1} \frac{\ln(\tau_i - \tau_{i-1})}{2} + \ln(m+1) + \sum_{i=2}^{m} \ln(\tau_i).$$

Simplifications: (1) all base two logarithms were changed to natural logarithms; (2) constant quantities are ignored (e.g., N and $X_1, \ldots, X_N$).

The Combinatorial Wall

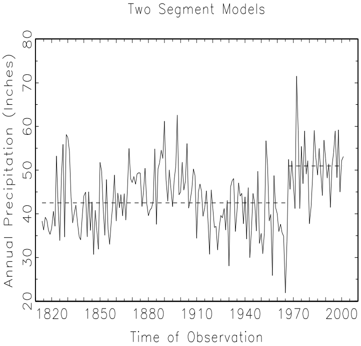
An exhaustive search over all models with m changepoints requires evaluation of $\binom{N}{m}$ MDL scores.

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

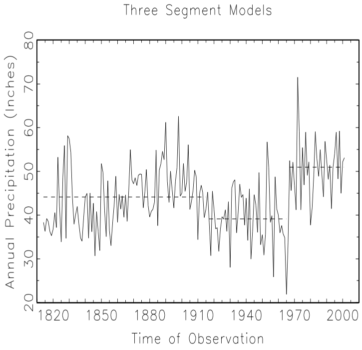
We devised a genetic algorithm for this task. A genetic algorithm is an intelligent random walk search.

No details here, but it works pretty well.

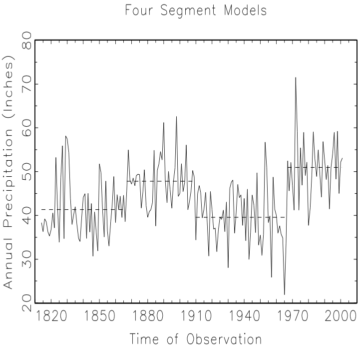
Two Segment Models



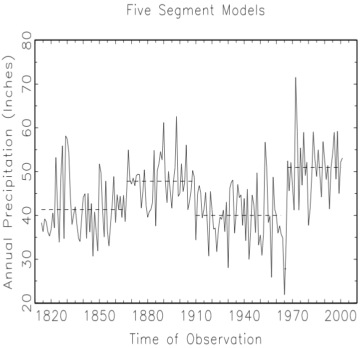
Three Segment Models



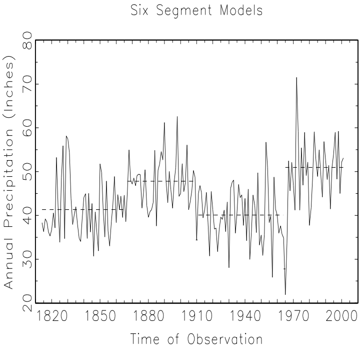
Four Segment Models



Five Segment Models



Six Segment Models



Simulations — Set I

Mimics the New Bedford Data with lognormal distributions:

1000 series of length $N = 200$ with no changepoints.
$\mu = 6.8$, $\phi = 0.2$, $\sigma^2 = 0.025$.

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

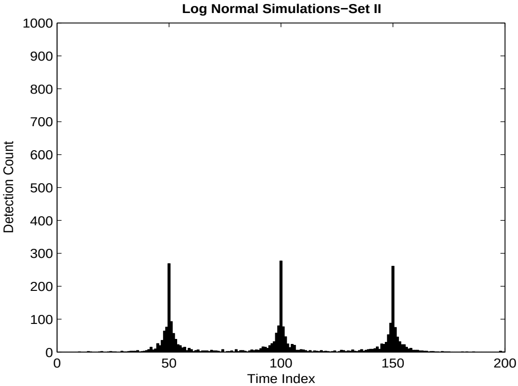


Figure: The detected changepoint times cluster around their true values of 50, 100, and 150.

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

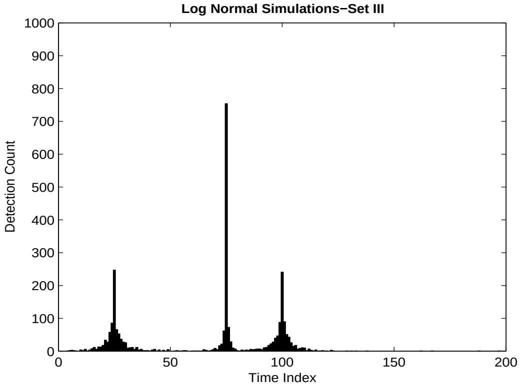
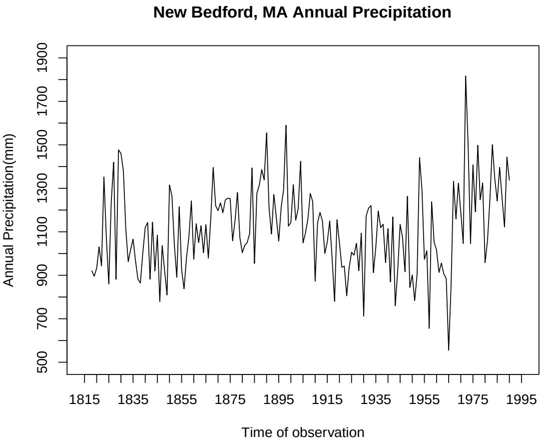


Figure: The detected changepoint times cluster around their true values of 25, 75, and 100.

New Bedford Precipitation Data



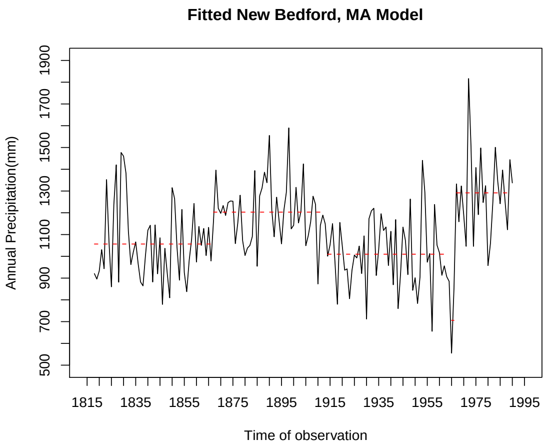
Optimal Model

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



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

USA Temperature Trends

Our last subsection studies US temperature trends in monthly high and low temperatures that takes into account changepoint features. The study is now out in Lee, Li, and Lund (2014) *Journal of Climate*.

Changes in average US temperatures are an essentially settled matter.

There are about 1000 data stations.

A monthly high temperature is the highest high temperature over all days during the month.

Station Locations

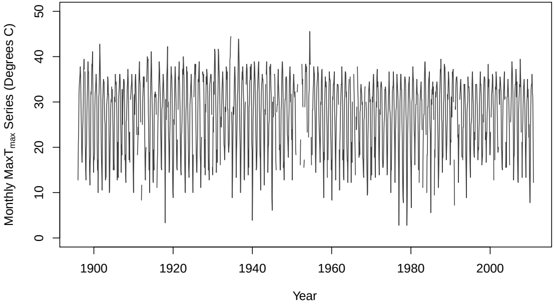


Figure: The Jacksonville, Illinois Maximum Record

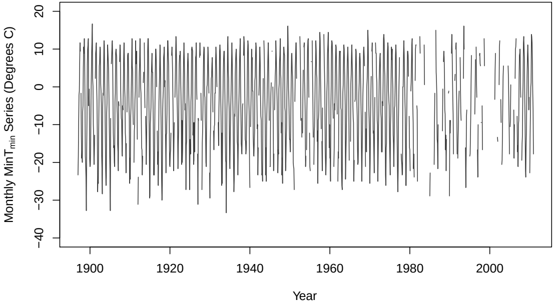


Figure: The Jacksonville, Illinois Minimum Record

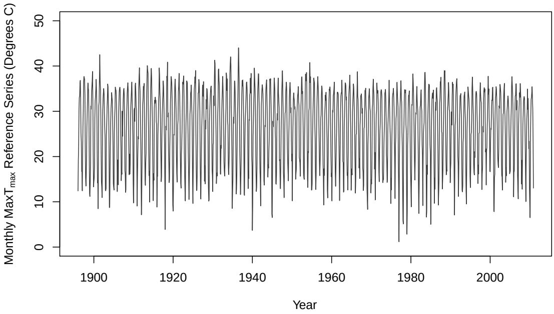


Figure: The Jacksonville, Illinois Maximum Reference Series

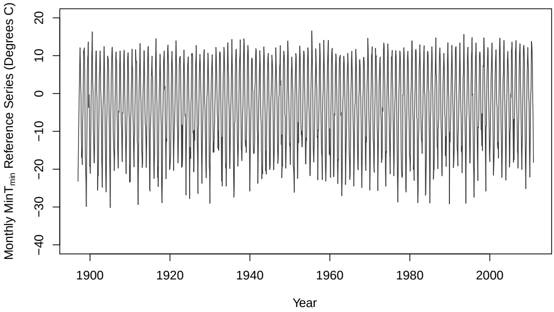


Figure: The Jacksonville, Illinois Minimum Reference Series

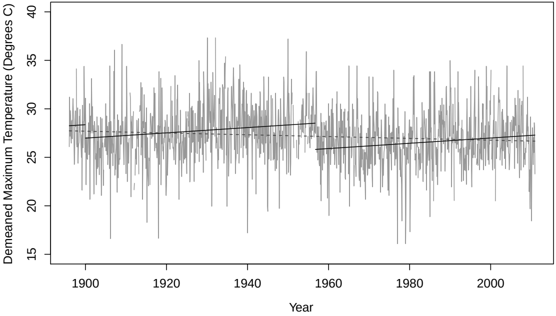


Figure: The Jacksonville, Illinois Maximum Changepoint Structure

GEV Extreme Distributions

For a fixed station, X_t, the month t extreme, is assumed to follow the generalized extreme-value distribution with CDF

$$F_t(x) = P[X_t \le x] = \exp\left[-\left\{ 1 + \xi \left(\frac{x - \mu_t}{\sigma_t} \right) \right\}_{+}^{-1/\xi} \right].$$

ξ is the all important shape parameter.

σ_t varies periodically with period 12 months.

μ_t is a time varying location parameter.

More on the Model

$$\mu_t = m_t + \alpha t + \delta_t.$$

Here, m_t is a monthly location parameter with period 12 months, α is a linear trend parameter, and δ_t is a regime mean allowing for m changepoints at times $\tau_1 < \cdots < \tau_m$.

$$\delta_t = \begin{cases} \Delta_1, & \text{if } t = 1, \ldots, \tau_1 - 1; \\ \Delta_2, & \text{if } t = \tau_1, \ldots, \tau_2 - 1; \\ \vdots & \quad\quad \vdots \\ \Delta_{k+1}, & \text{if } t = \tau_k, \ldots, N. \end{cases}$$

Details

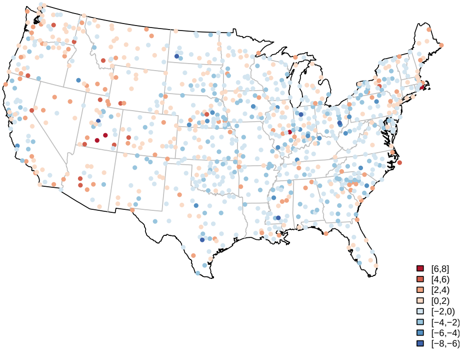
X_t and X_s are assumed independent when $t \neq s$:

$$L = \prod_{t=1}^{N} F_t'(X_t).$$

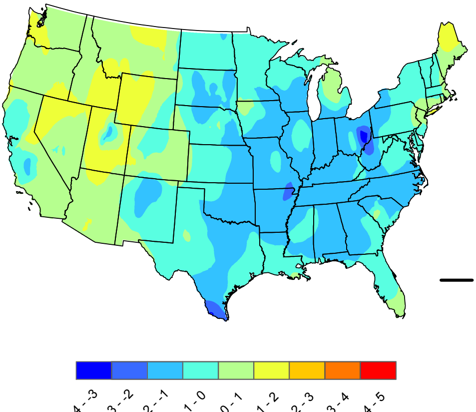
Temporal independence is probably not realistic.

The likelihood is optimized numerically with an MDL penalty. A genetic algorithm is used to optimize the penalized likelihood.

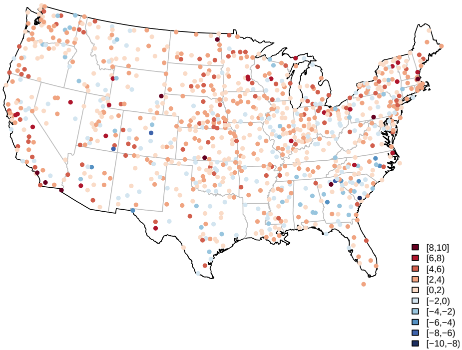
Raw USA Trends in Monthly Maximums



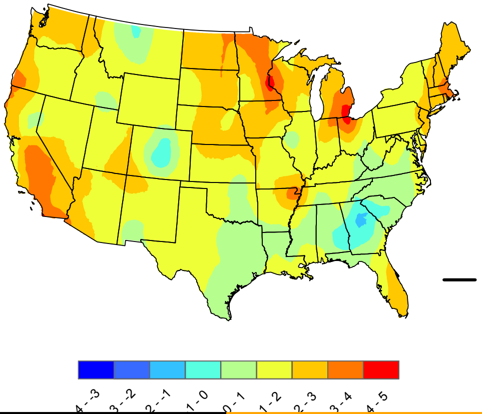
Smoothed USA Trends in Monthly Maximums



Raw USA Trends in Monthly Minimums



Smoothed USA Trends in Monthly Minimums



Conclusions

Low temperatures are warming more than high temperatures.

The Western US is warming more than the Eastern US.

Vancouver is warming!