Analysis of Astrophysics and Particle Physics Data using Optimal Segmentation

Jeffrey.D.Scargle@nasa.gov

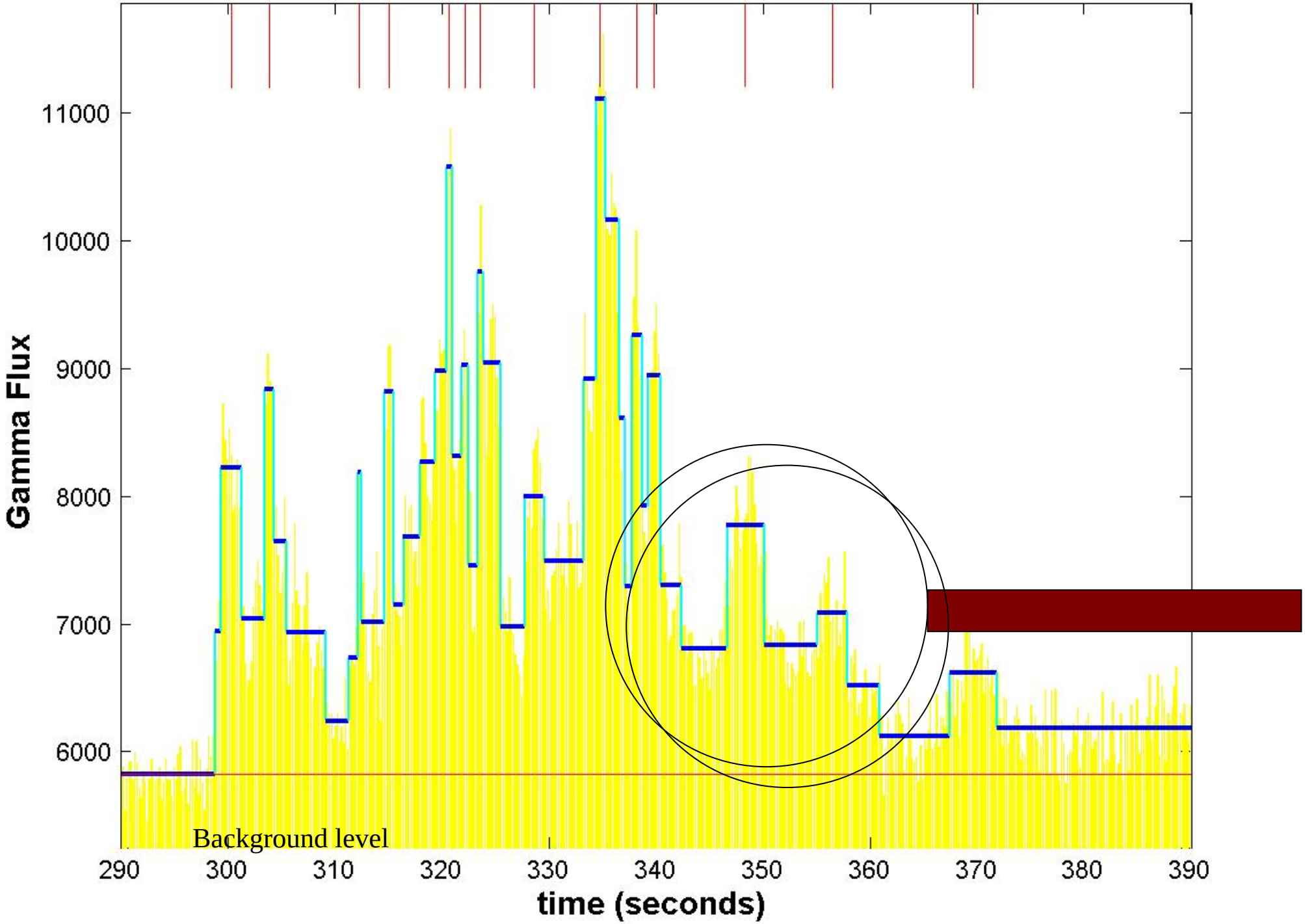
Space Science Division
NASA Ames Research Center

Santa Cruz Institute for Particle Physics
May 17, 2005

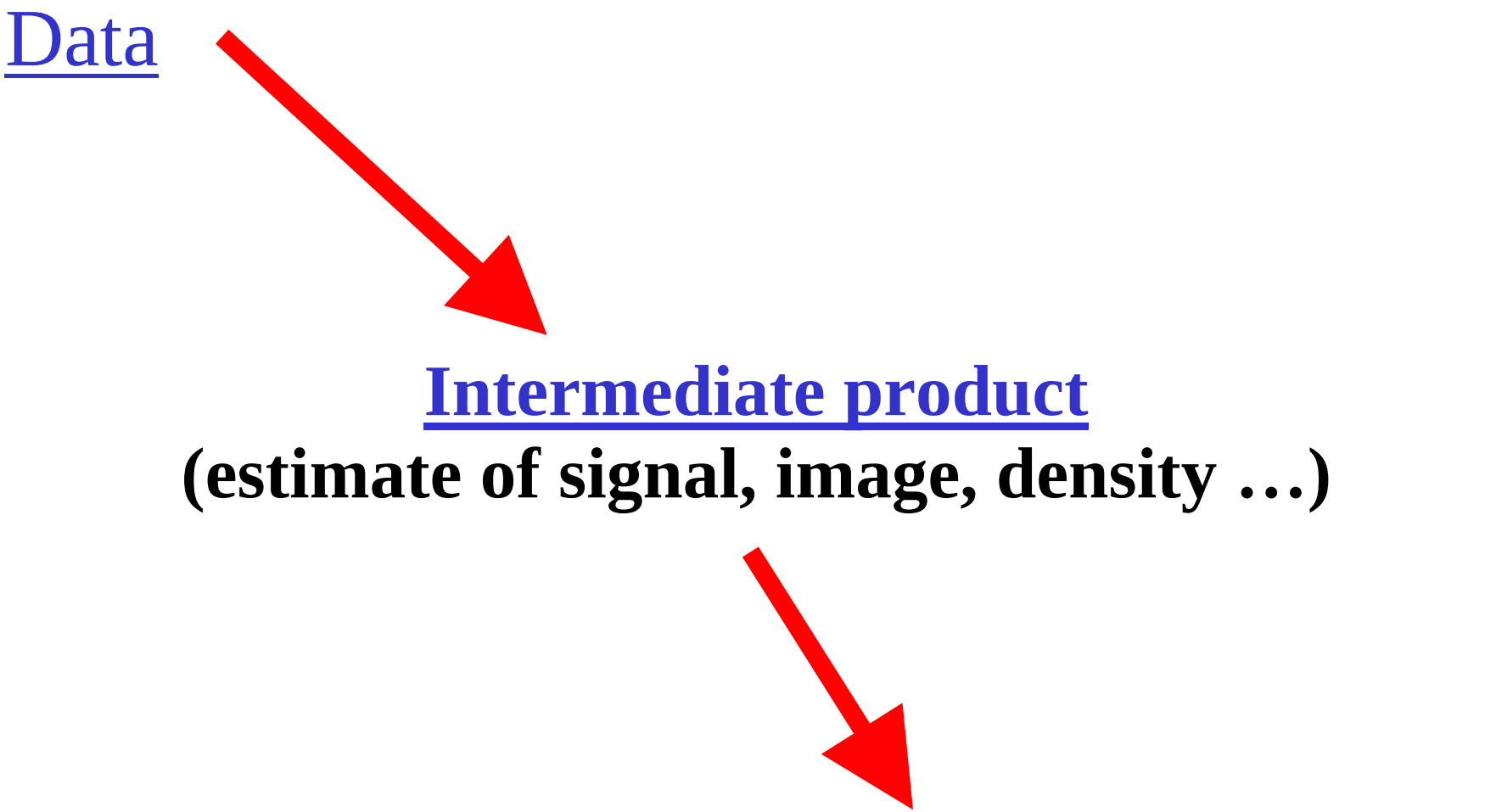
Outline

- Goal: Detect/Characterize Local Structures
- Data Cells
- Piecewise Constant Models
- Fitness Functions
- Optimization
- Error analysis
- Interpretation
- Extension to Higher Dimensions

The Main Goal is to Detect and Characterize Local Structures



From Data to Astronomical Goals



Smoothing and Binning

Old views: the best (only) way to reduce noise is to smooth the data
the best (only) way to deal with point data is to use bins

New philosophy: smoothing and binning should be avoided because they ...

- discard information
- degrade resolution
- introduce dependence on parameters:
  - degree of smoothing
  - bin size and location

Wavelet Denoising (Donoho, Johnstone) multiscale; no explicit smoothing
Adaptive Kernel Smoothing

Optimal Segmentation (*e.g.* Bayesian Blocks) Omni-scale -- uses neither explicit smoothing nor pre-defined binning

Data: Measurements Distributed in a *Data Space*

Independent variable (data space)
e.g. time, position, wavelength, …

Dependent variable
e.g. event locations, counts-in-bins, measurements, …

Examples: time series, spectra
images, photon maps
redshift surveys
higher dimensional data

DATA CELLS: Definition

***data space*:** set of all allowed values of the independent variable

data cell: a data structure representing an individual measurement

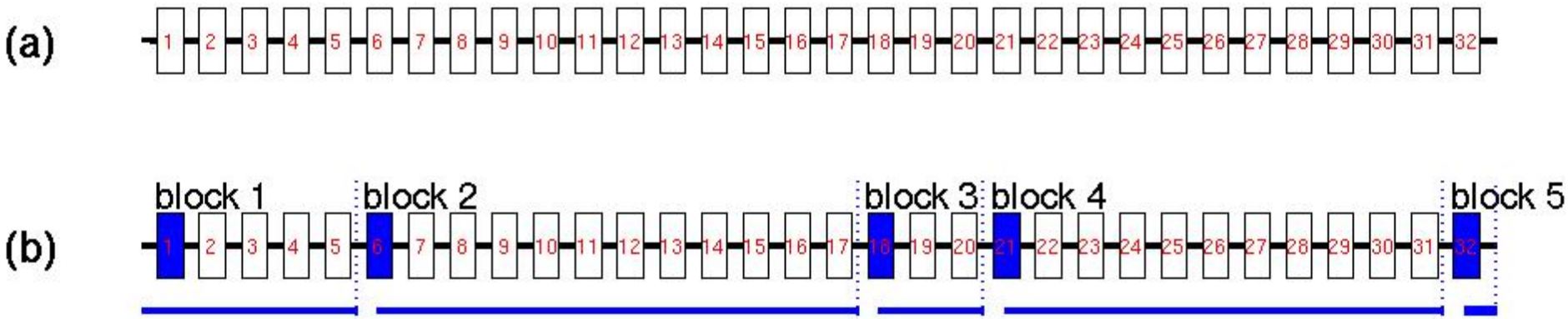
For a segmented model, the cells must contain all information needed to compute the model *cost function*.

The data cells typically:

- are in one-to-one correspondence to the measurements
- partition the entire data space (no gaps or overlap)
- contain information on adjacency to other cells

… but any of these conditions may be violated.

Simple Example of 1D Data Cells and Blocks



Fitness Functions

- Block likelihood = product of likelihoods of its cells

- Block Likelihood depends on
  - N = The Number of Events in the Block
  - M = The Size of the Block

- Model likelihood = product of likelihoods of its blocks

- Remove the dependence on the block event rates:
  - Marginalize, or
  - Maximize the Likelihood

- Adopt prior distribution for N_b, the number of blocks.
  (Parameter of this distribution acts like a smoothing parameter.)

- Take log to yield an additive fitness function.

The Optimiser

```
best = []; last = [];
for R = 1:num_cells
   [ best(R), last(R) ] = max( [0 best] + fitness( cumsum( data_cells(1:R, :) ) );

     if first > 0 & last(R) > first    % Option: trigger on first significant block
          changepoints = last(R); return
     end

end

% Now locate all the changepoints
index = last( num_cells );
changepoints = [];

while index > 1
   changepoints = [ index changepoints ];
    index = last( index - 1 );
end
```

Do not use at home: a few details omitted!

Bootstrap Method:
Time Series of N Discrete Events

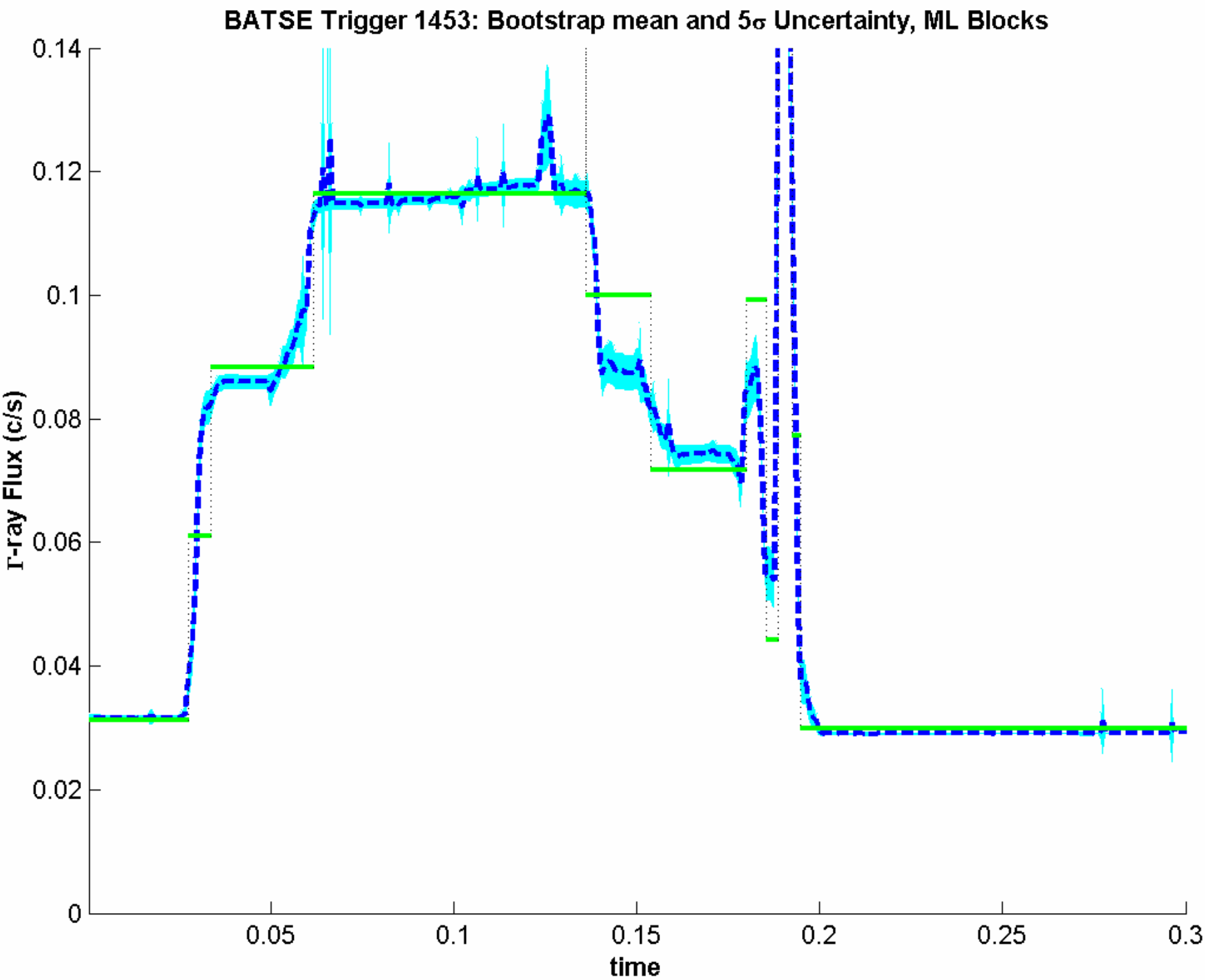
For many iterations:

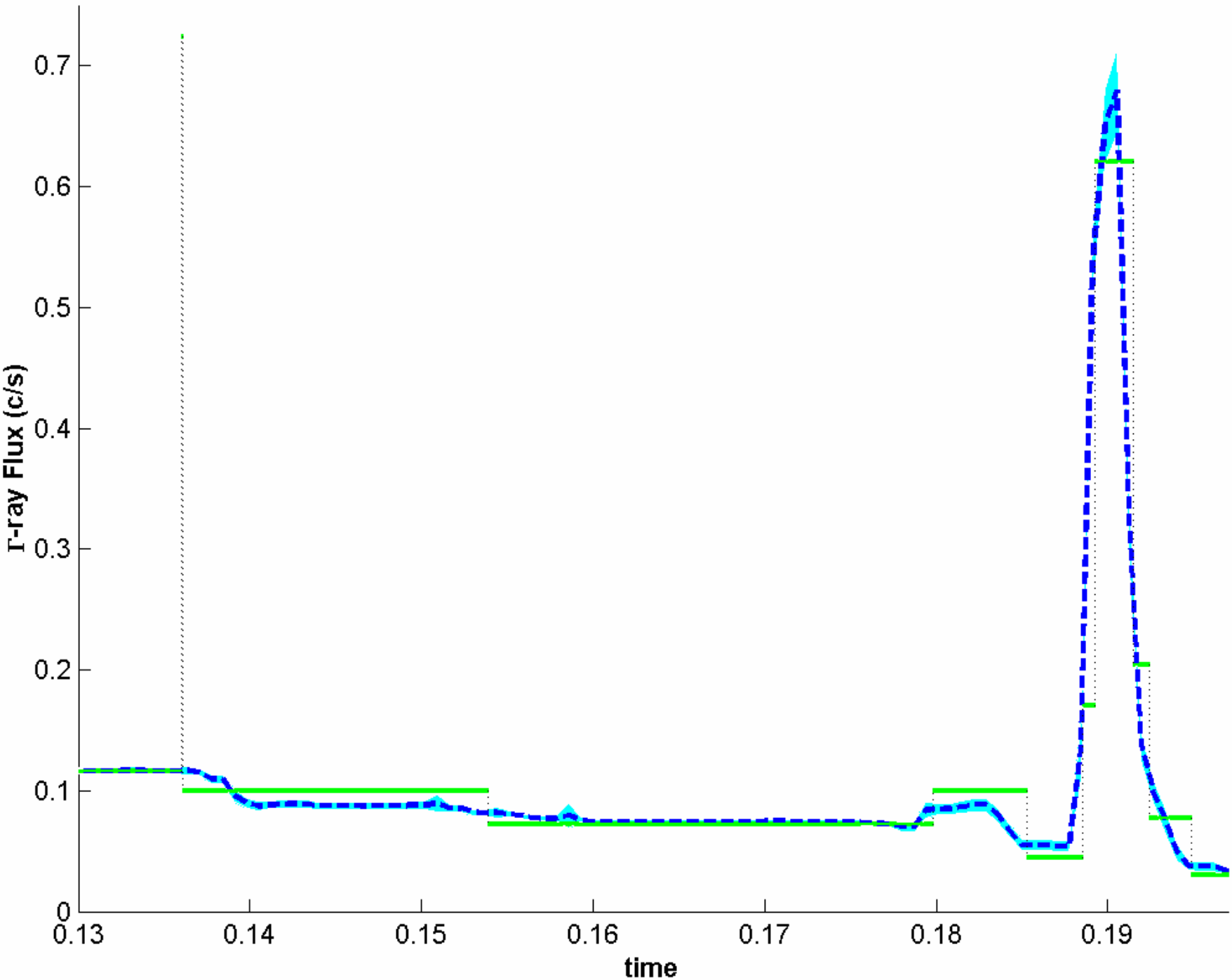
- Randomly select N of the observed events *with replacement*
- Analyze this sample just as if it were real data

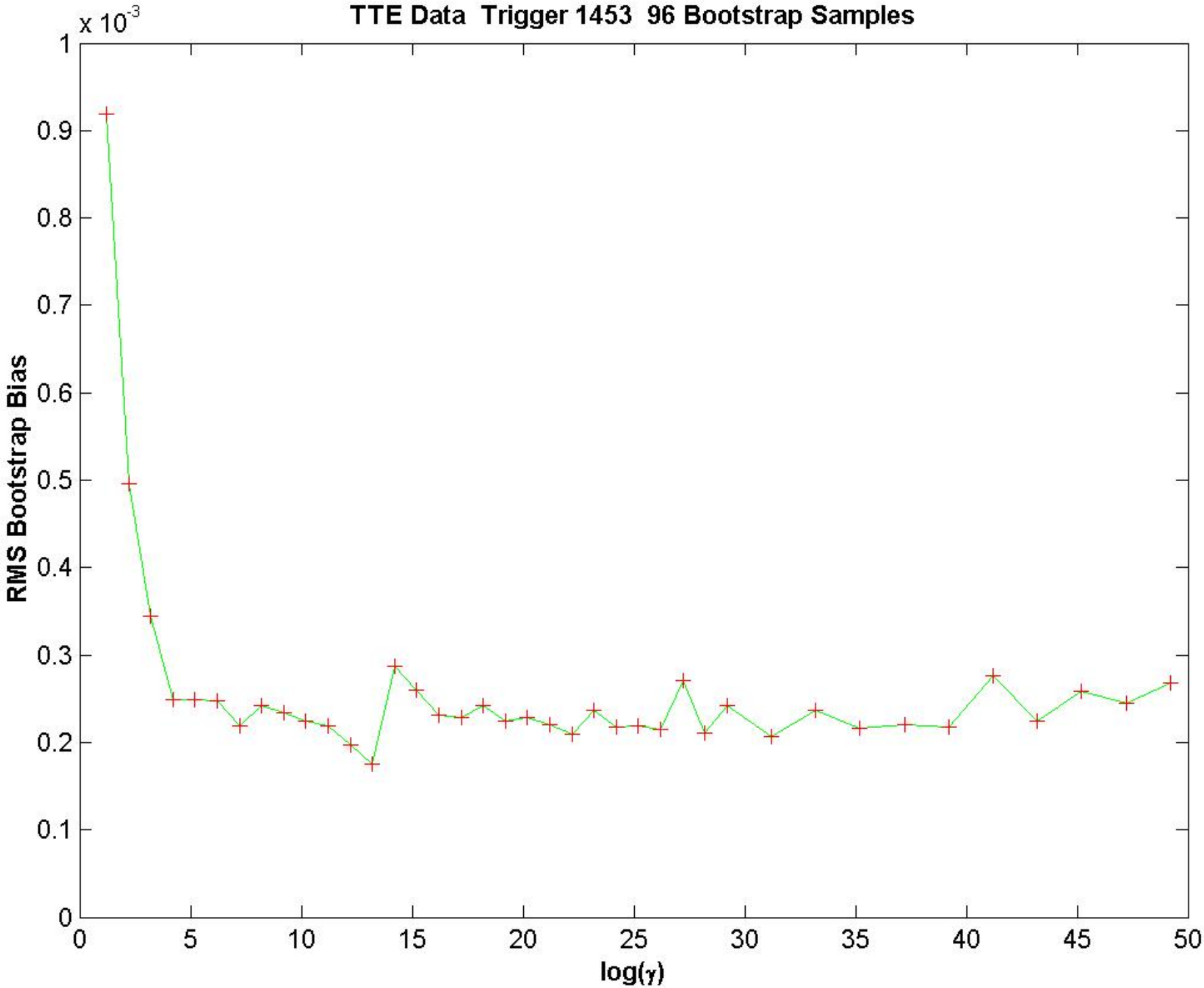
Compute mean and variance of the bootstrap samples

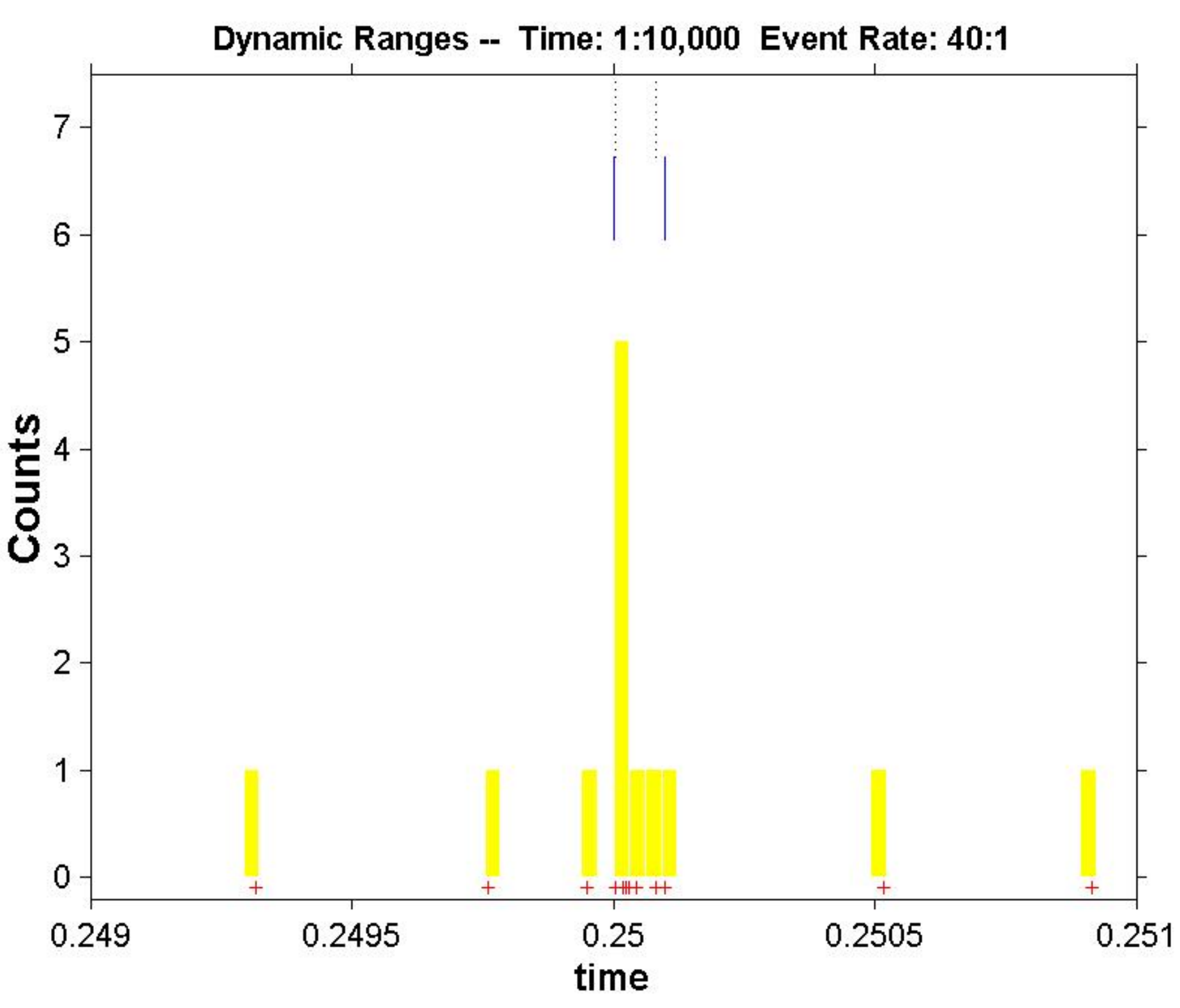
Bias = result for real data – bootstrap mean
RMS error derived from bootstrap variance

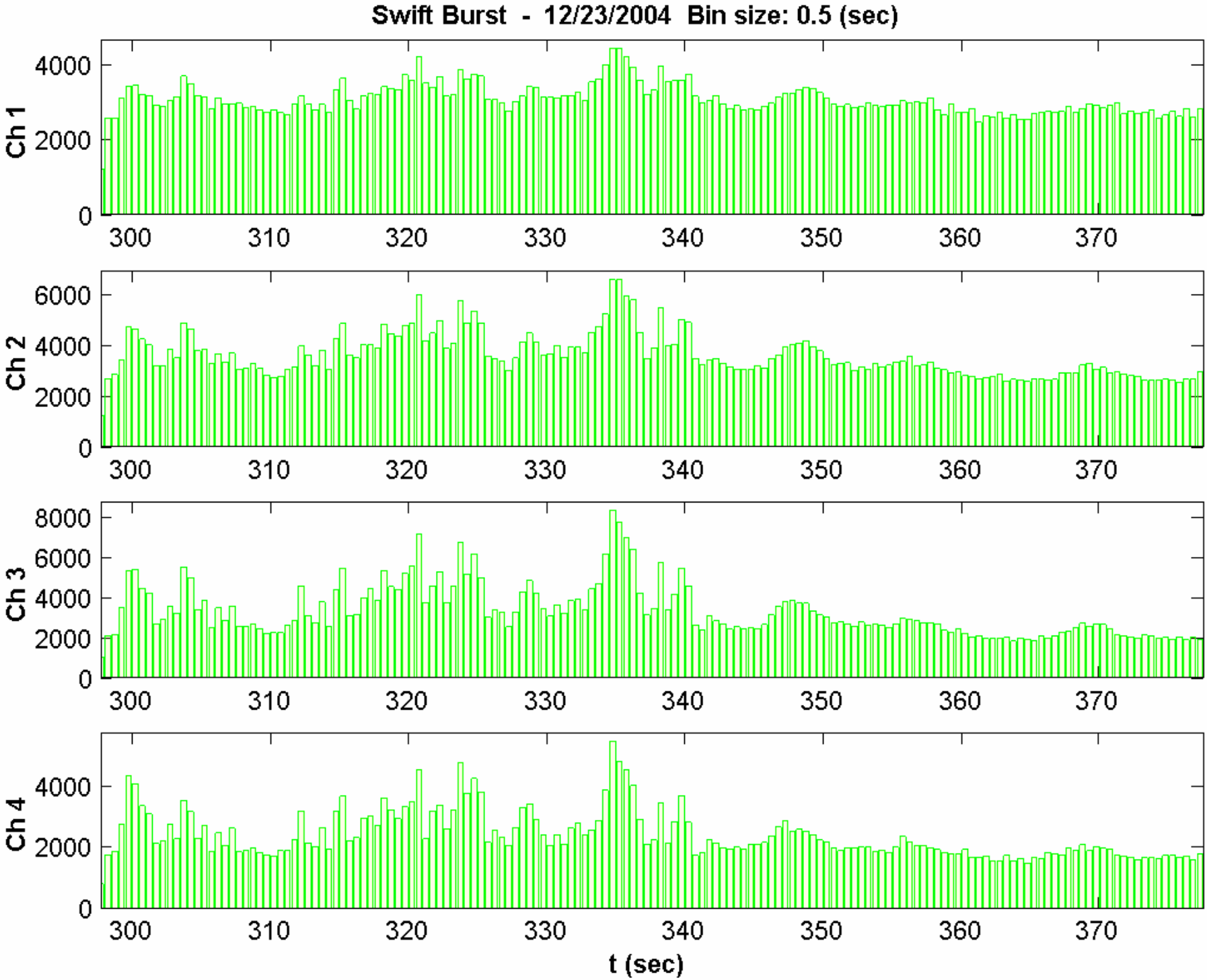
Caveat: The real data does not have the repeated events in bootstrap samples. I am not sure what effect this has.

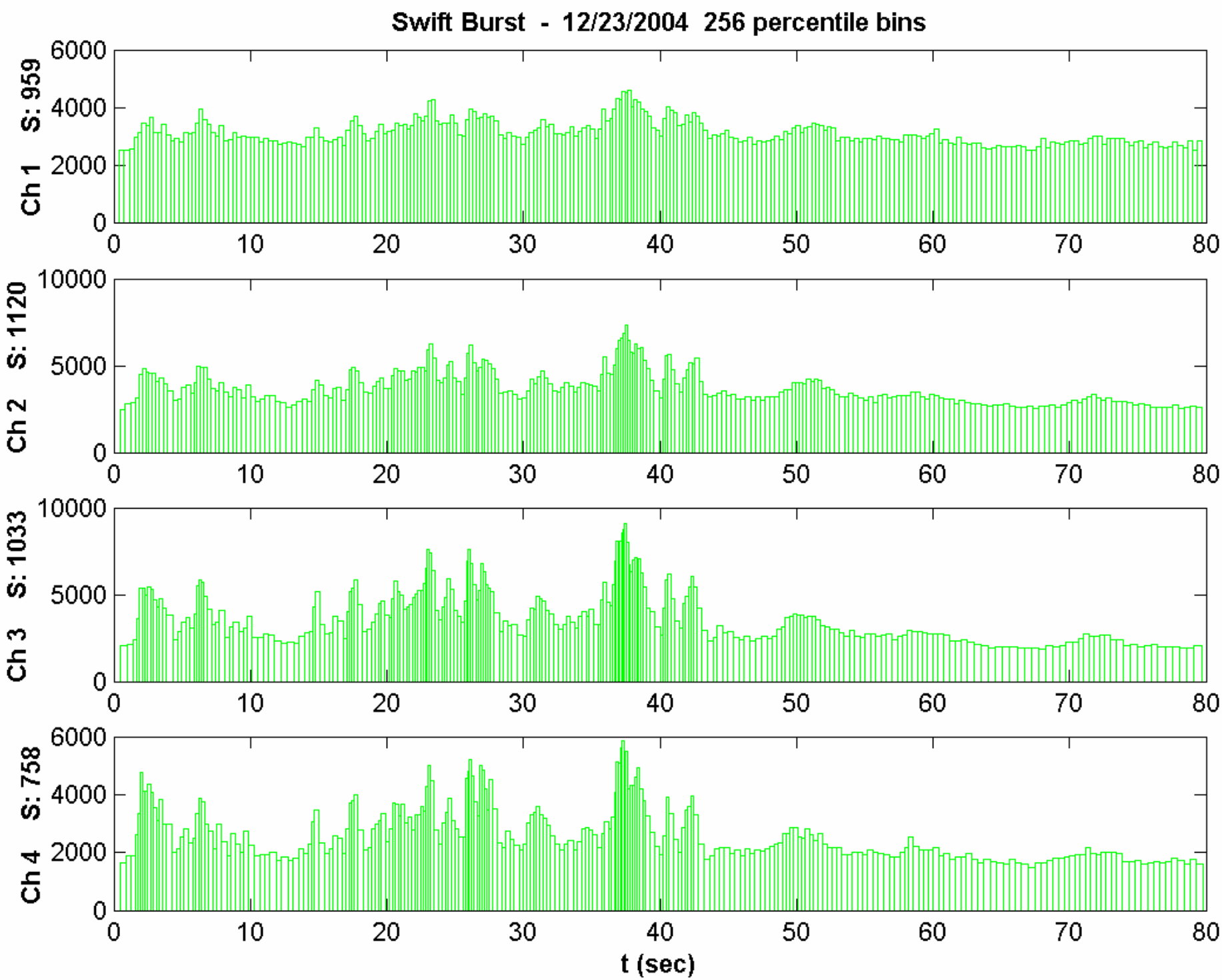

BATSE Trigger 1453: Bootstrap mean and 5σ Uncertainty, ML Blocks
Γ-ray Flux (c/s)
time
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.05
0.1
0.15
0.2
0.25
0.3

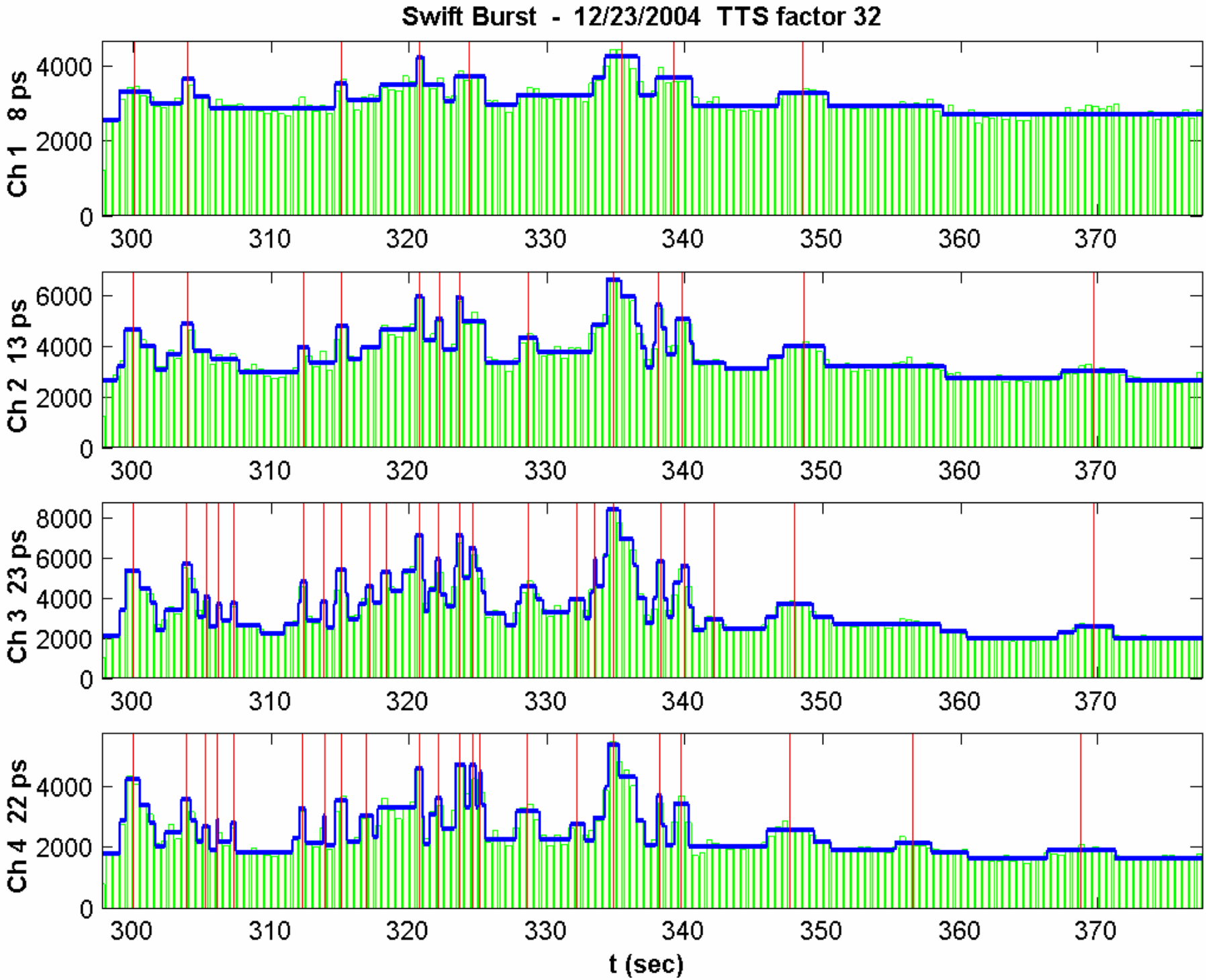

BATSE Trigger 1453: Bootstrap mean and 5σ Uncertainty, ML Blocks
Γ-ray Flux (c/s)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.13
0.14
0.15
0.16
0.17
0.18
0.19
time

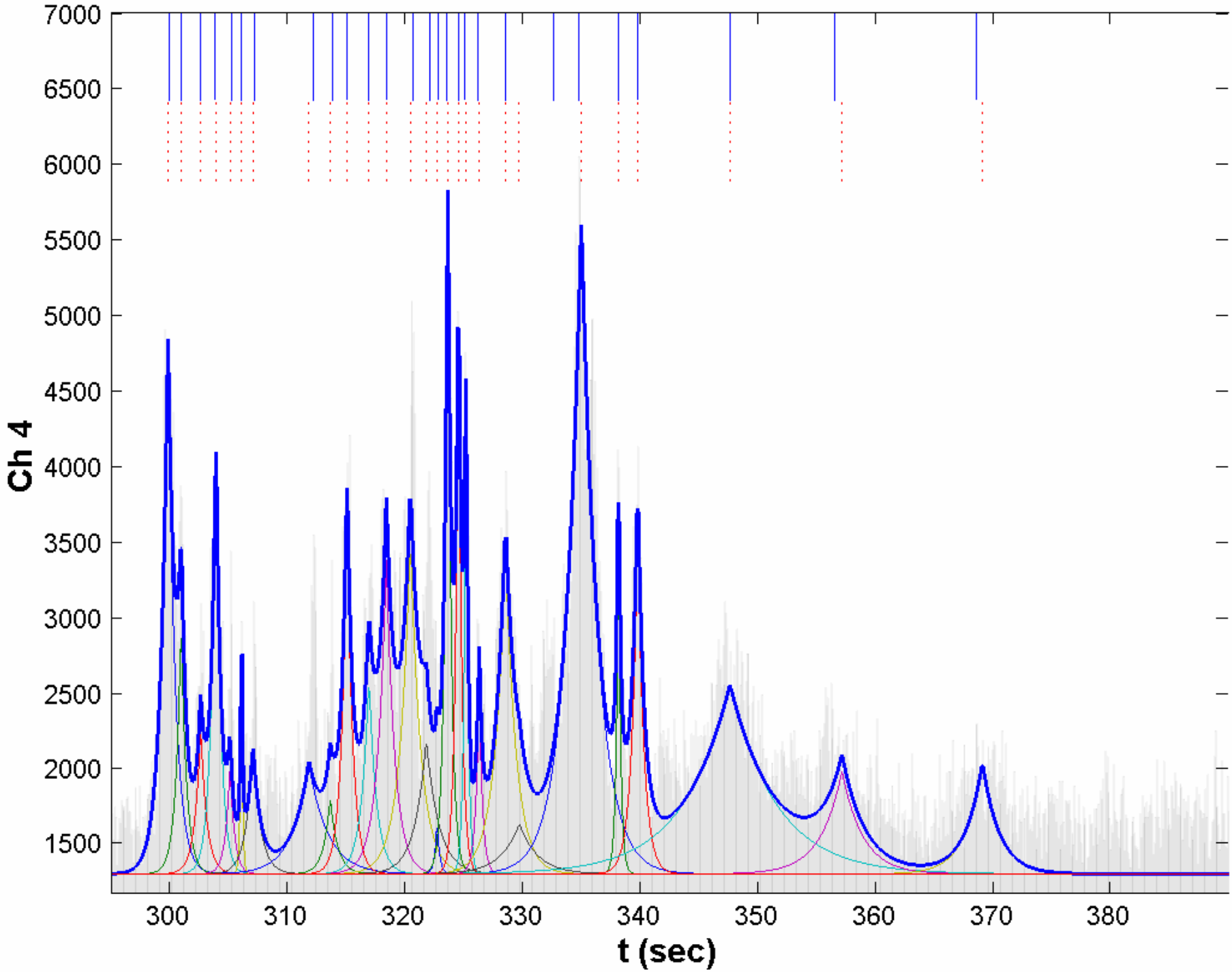

TTE Data Trigger 1453 96 Bootstrap Samples
x 10^-3
RMS Bootstrap Bias
log(γ)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
5
10
15
20
25
30
35
40
45
50

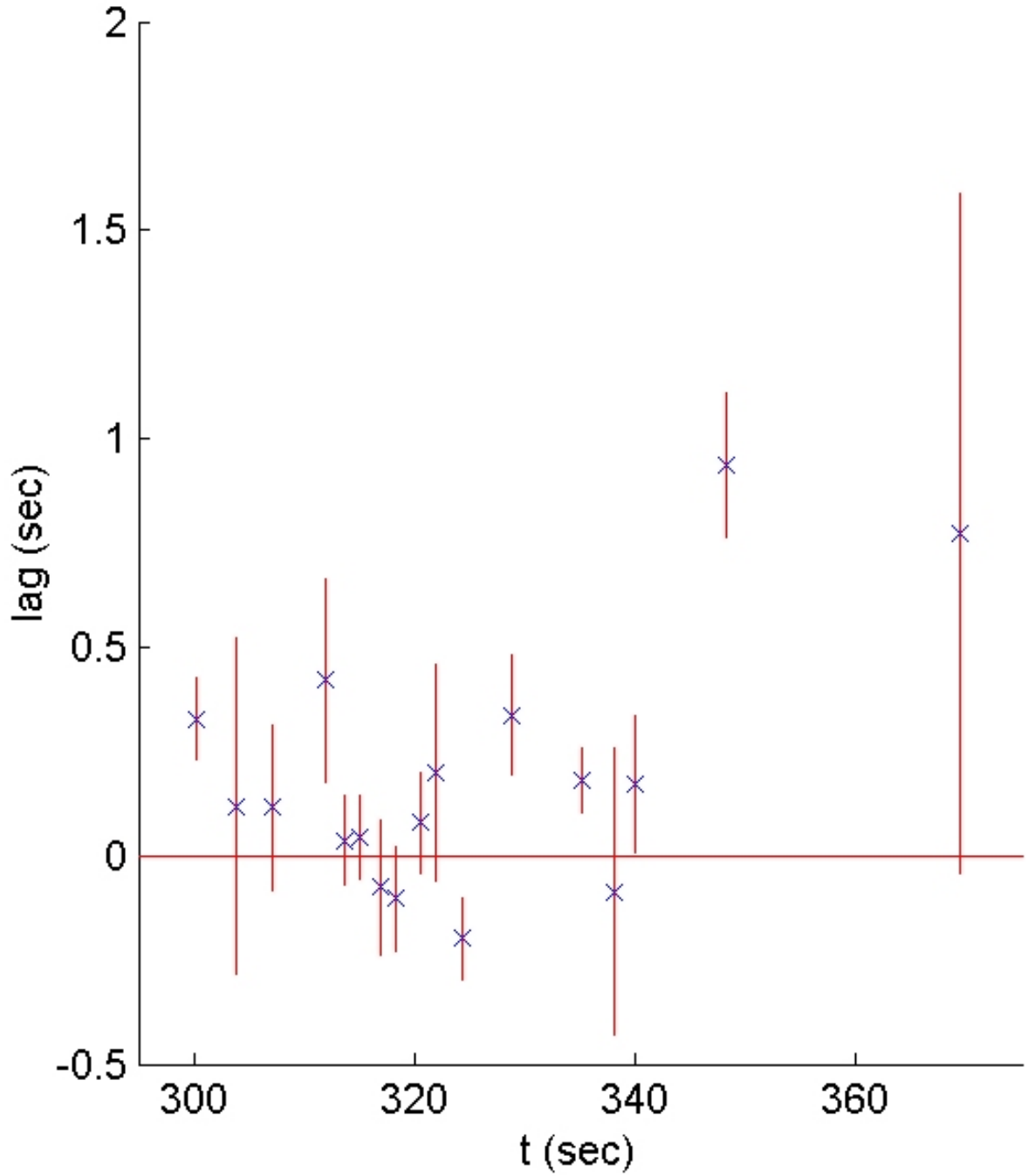

Dynamic Ranges -- Time: 1:10,000 Event Rate: 40:1
Counts
0
1
2
3
4
5
6
7
0.249
0.2495
0.25
0.2505
0.251
time


Swift Burst - 12/23/2004 Bin size: 0.5 (sec)
Ch 1
Ch 2
Ch 3
Ch 4
0
2000
4000
6000
8000
300
310
320
330
340
350
360
370
t (sec)

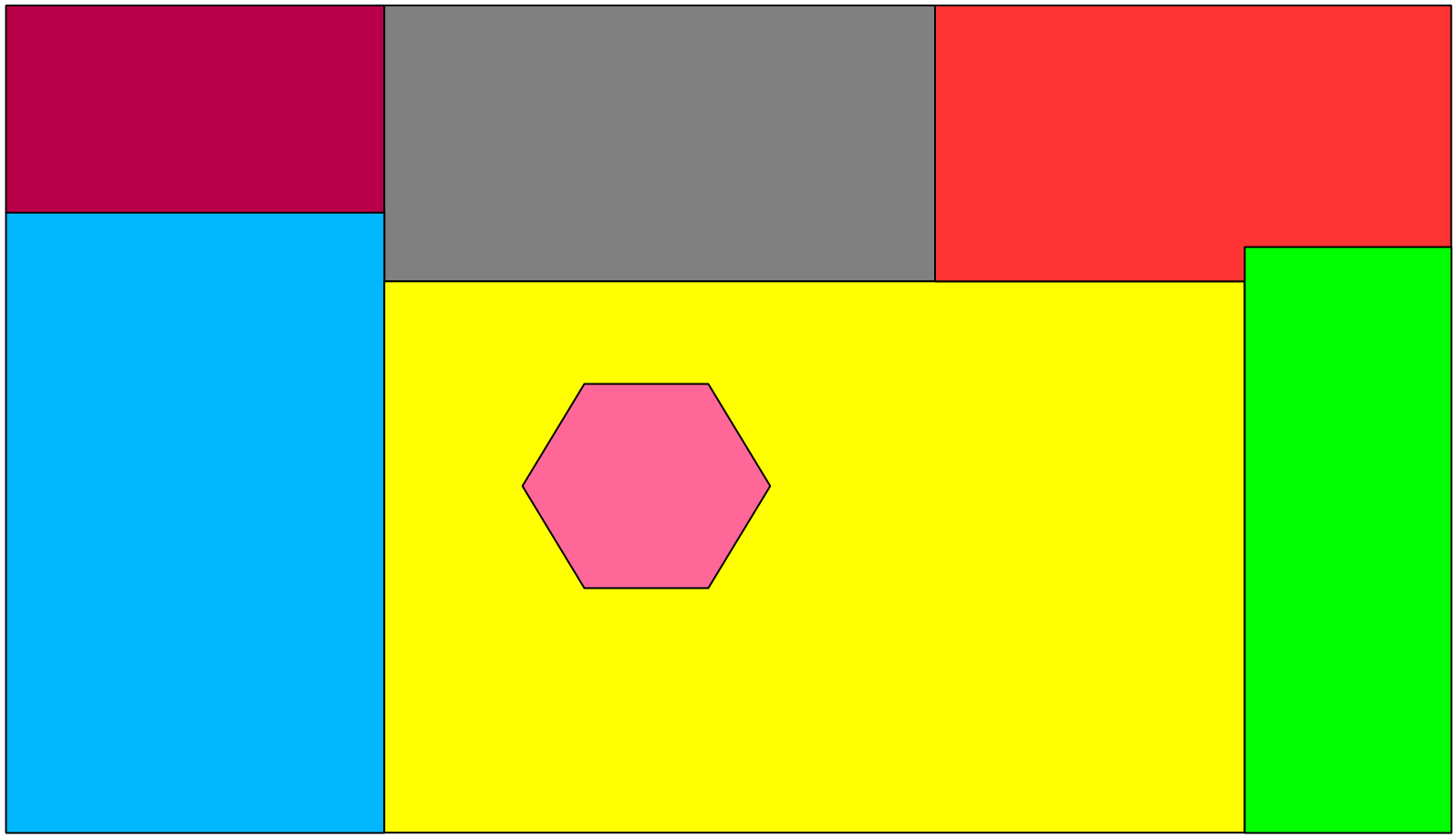

Swift Burst - 12/23/2004 256 percentile bins
Ch 1 S: 959
Ch 2 S: 1120
Ch 3 S: 1033
Ch 4 S: 758
0
2000
4000
6000
5000
10000
10
20
30
40
50
60
70
80
t (sec)


Swift Burst - 12/23/2004 TTS factor 32
Ch 1 8 ps
Ch 2 13 ps
Ch 3 23 ps
Ch 4 22 ps
0
2000
4000
6000
8000
300
310
320
330
340
350
360
370
t (sec)


7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
Ch 4
300
310
320
330
340
350
360
370
380
t (sec)


2
1.5
1
0.5
0
-0.5
lag (sec)
300
320
340
360
t (sec)

Piecewise Constant Model
(partitions the data space)



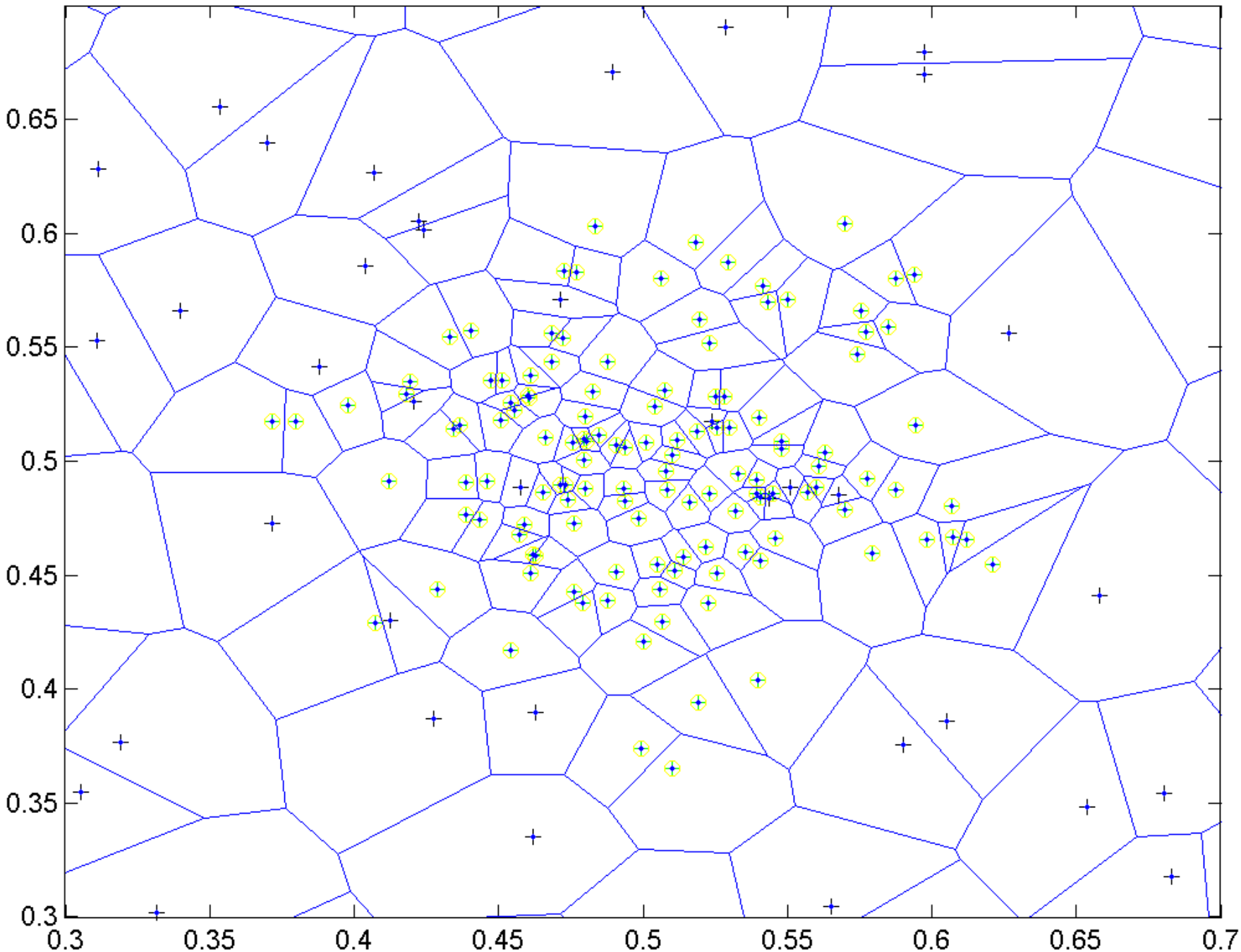
Signal modeled as constant over each partition element (block).

Optimum Partitions in Higher Dimensions

- Blocks are collections of Voronoi cells (1D,2D,...)
- Relax condition that blocks be connected
- Cell location now irrelevant
- Order cells by volume

Theorem: Optimum partition consists of blocks that are connected in this ordering

- Now can use the 1D algorithm, $O(N^2)$
- Postprocessing: identify connected block fragments


Data: Voronoi Tessellation
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

Blocks

Block: a set of data cells

Two cases:

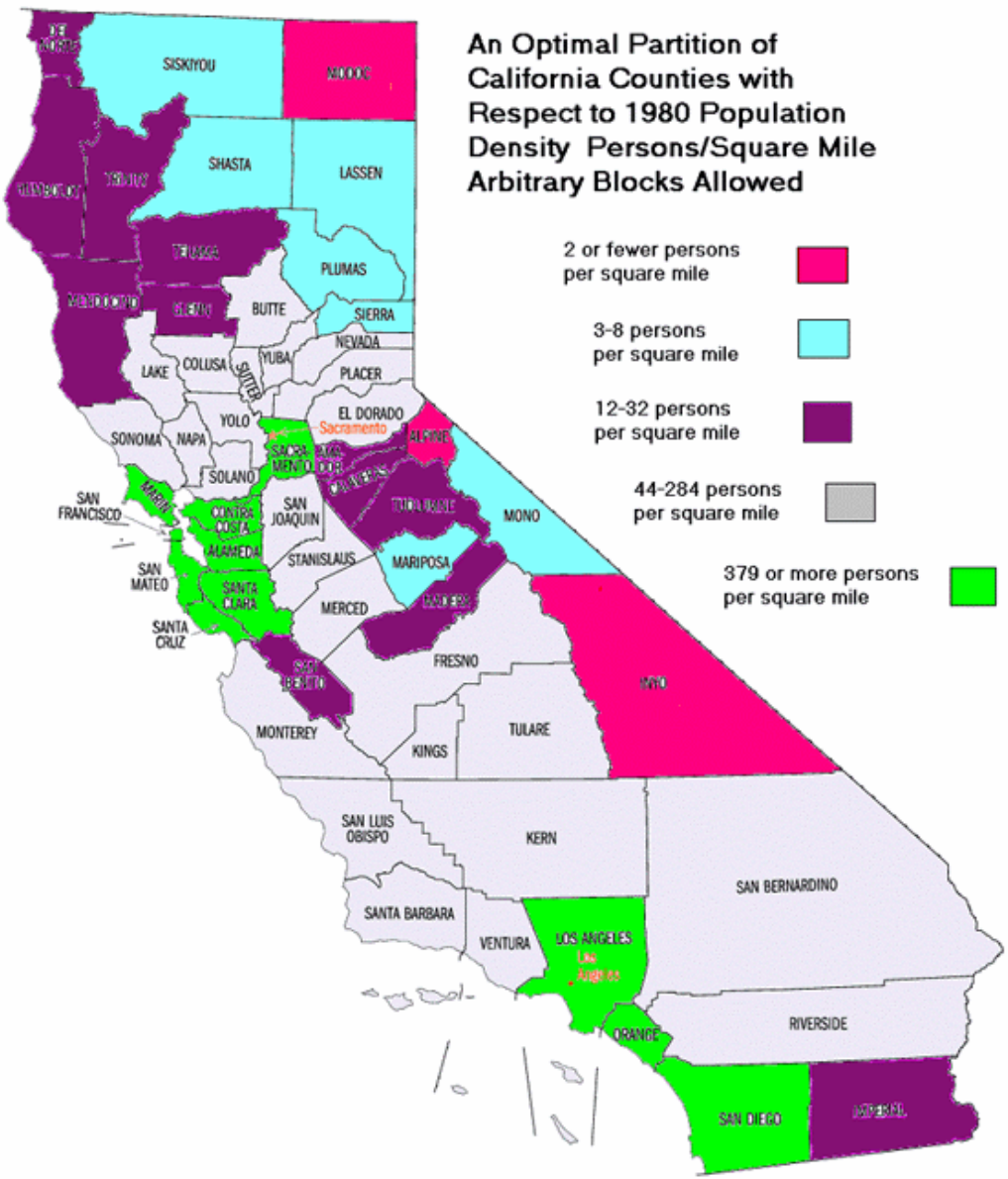
- Connected (can't break into distinct parts)
- Not constrained to be connected

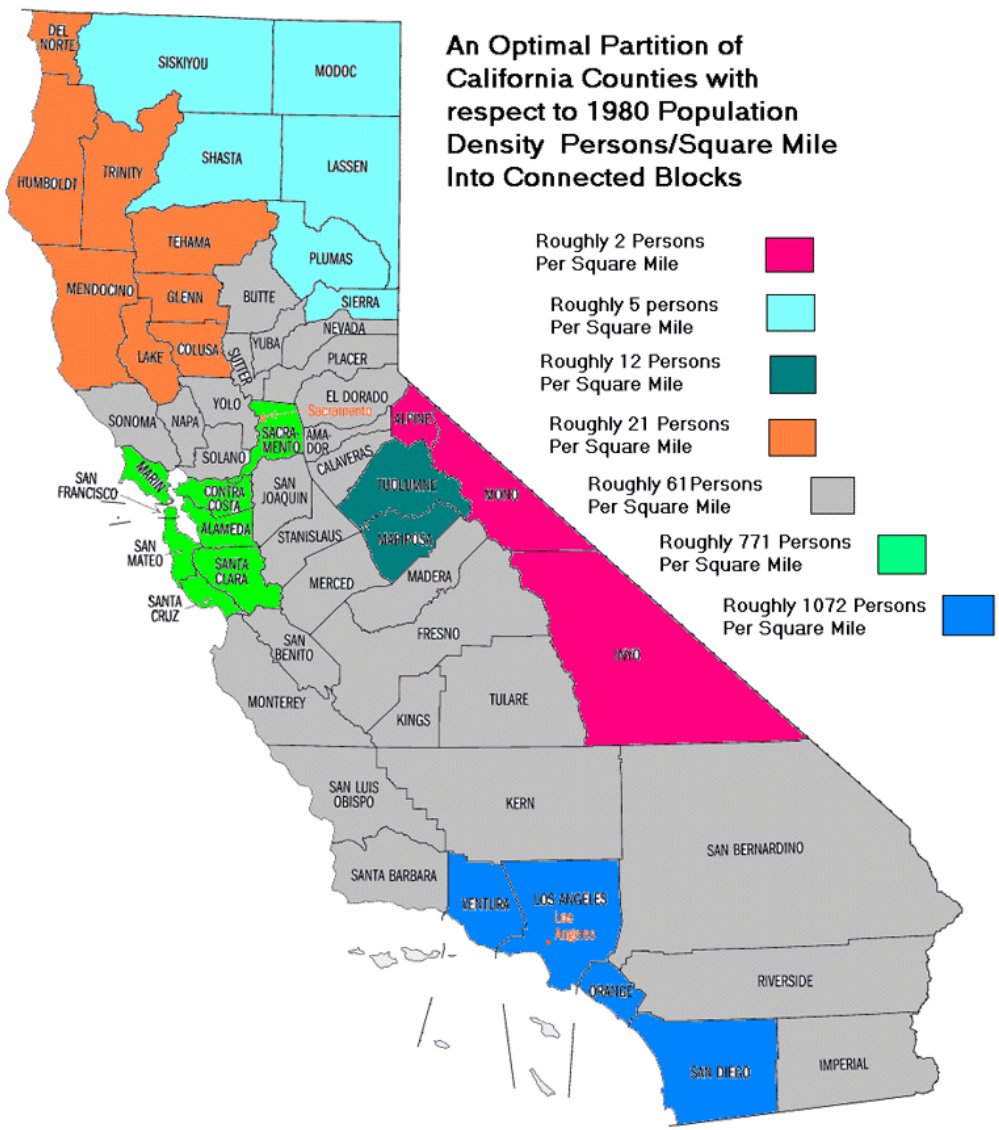
Model = set of blocks

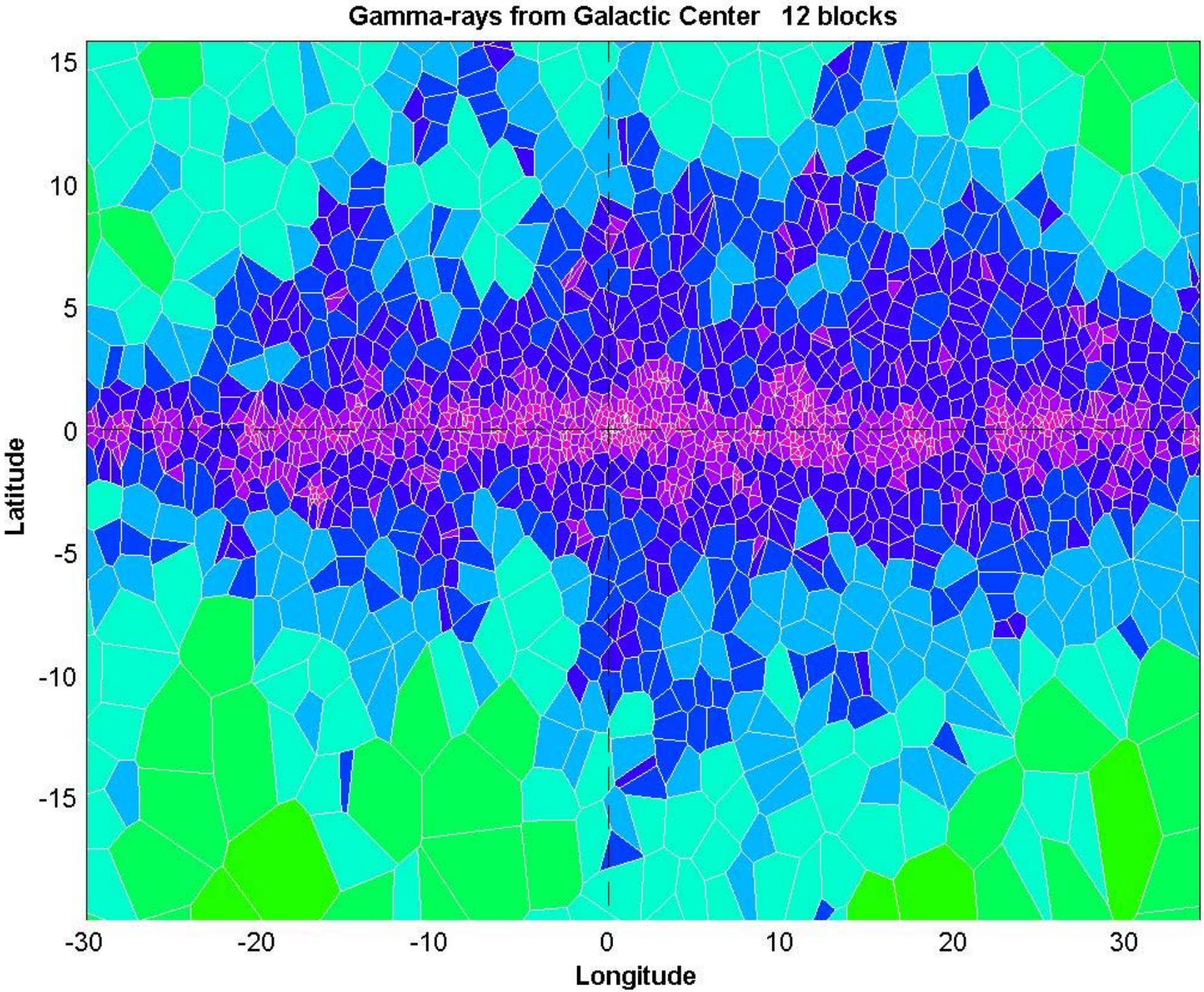
Fitness function:

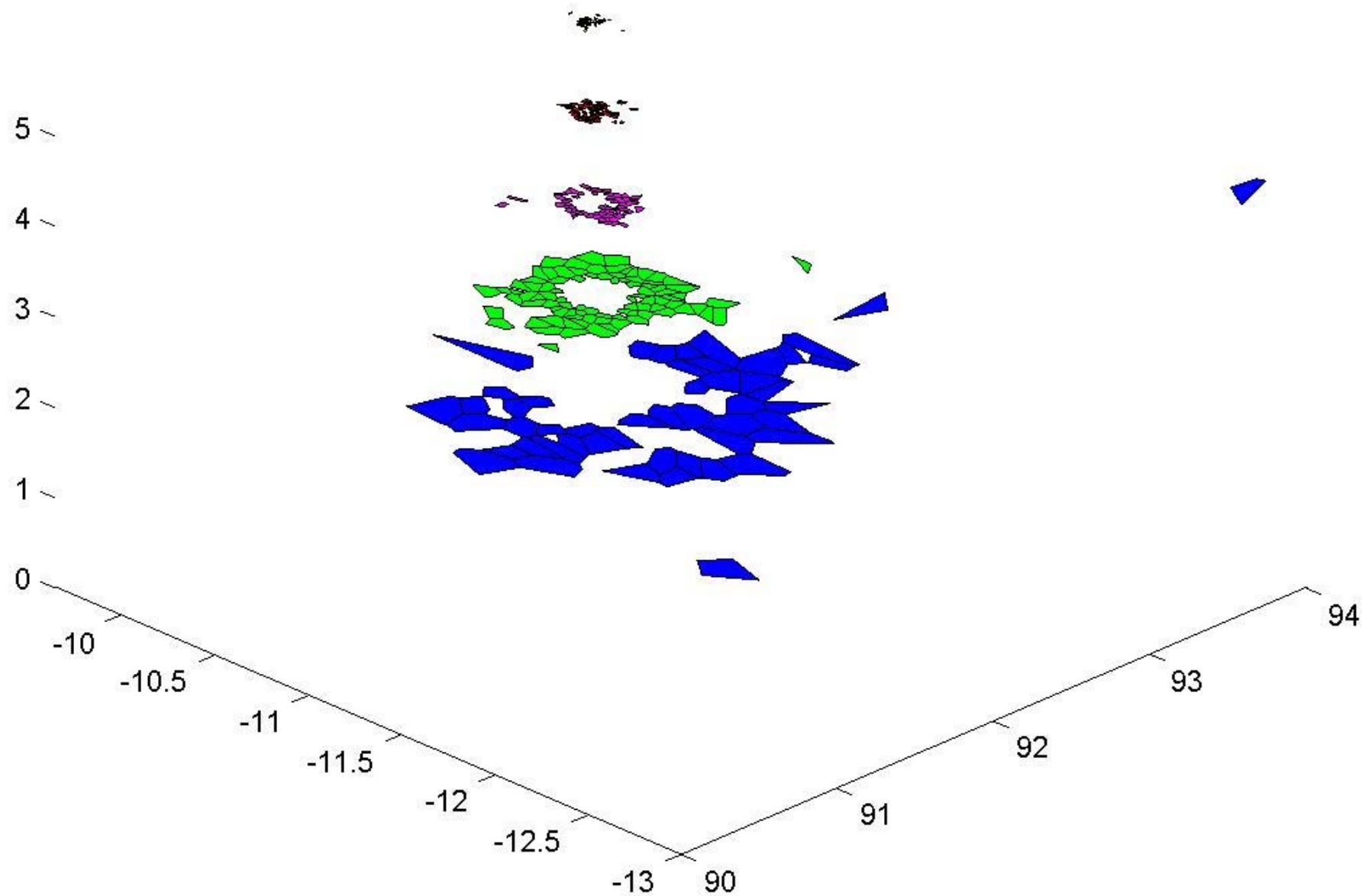
F(Model) = sum over blocks F(Block)

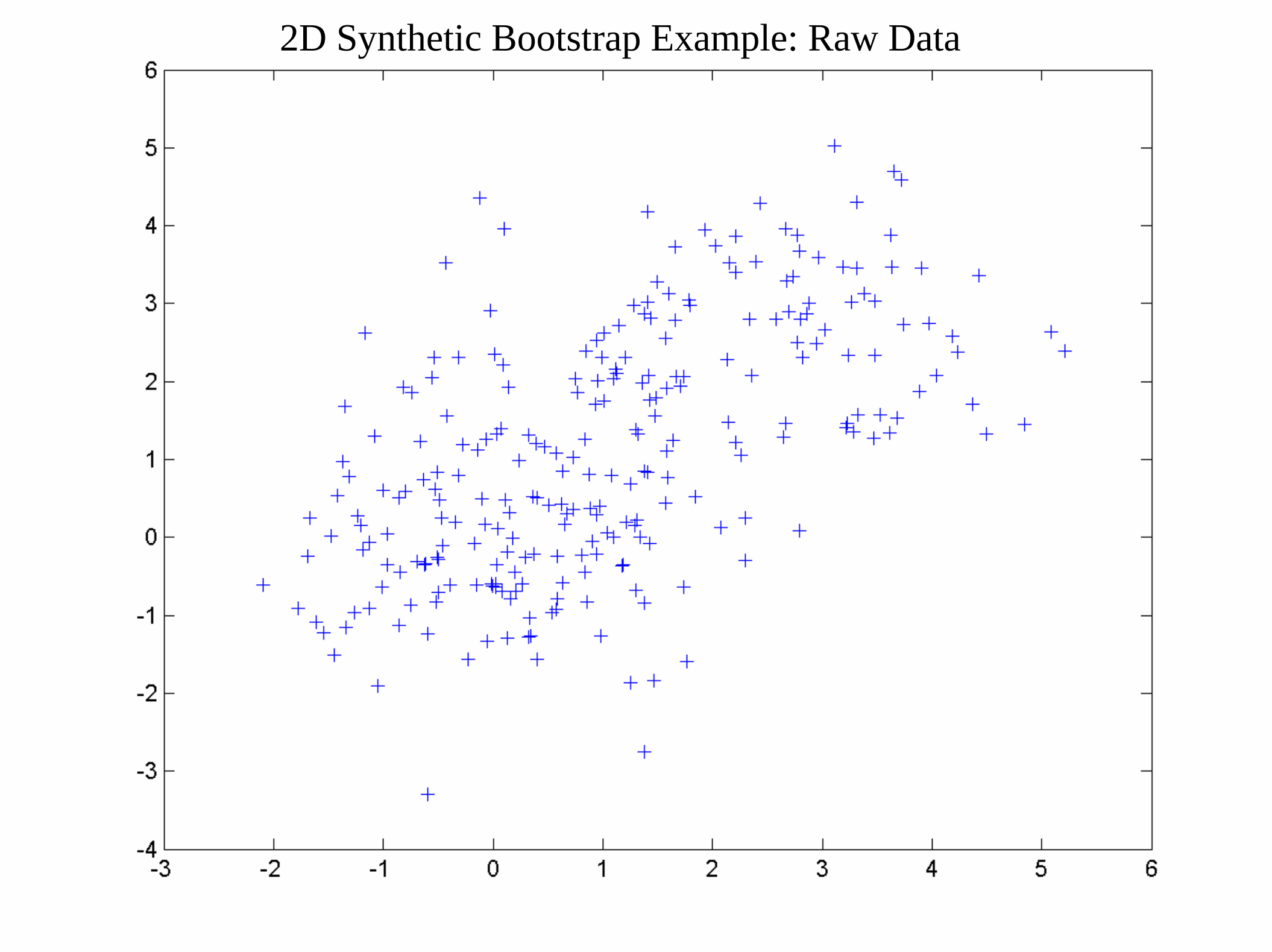
Connected vs. Arbitrary Blocks

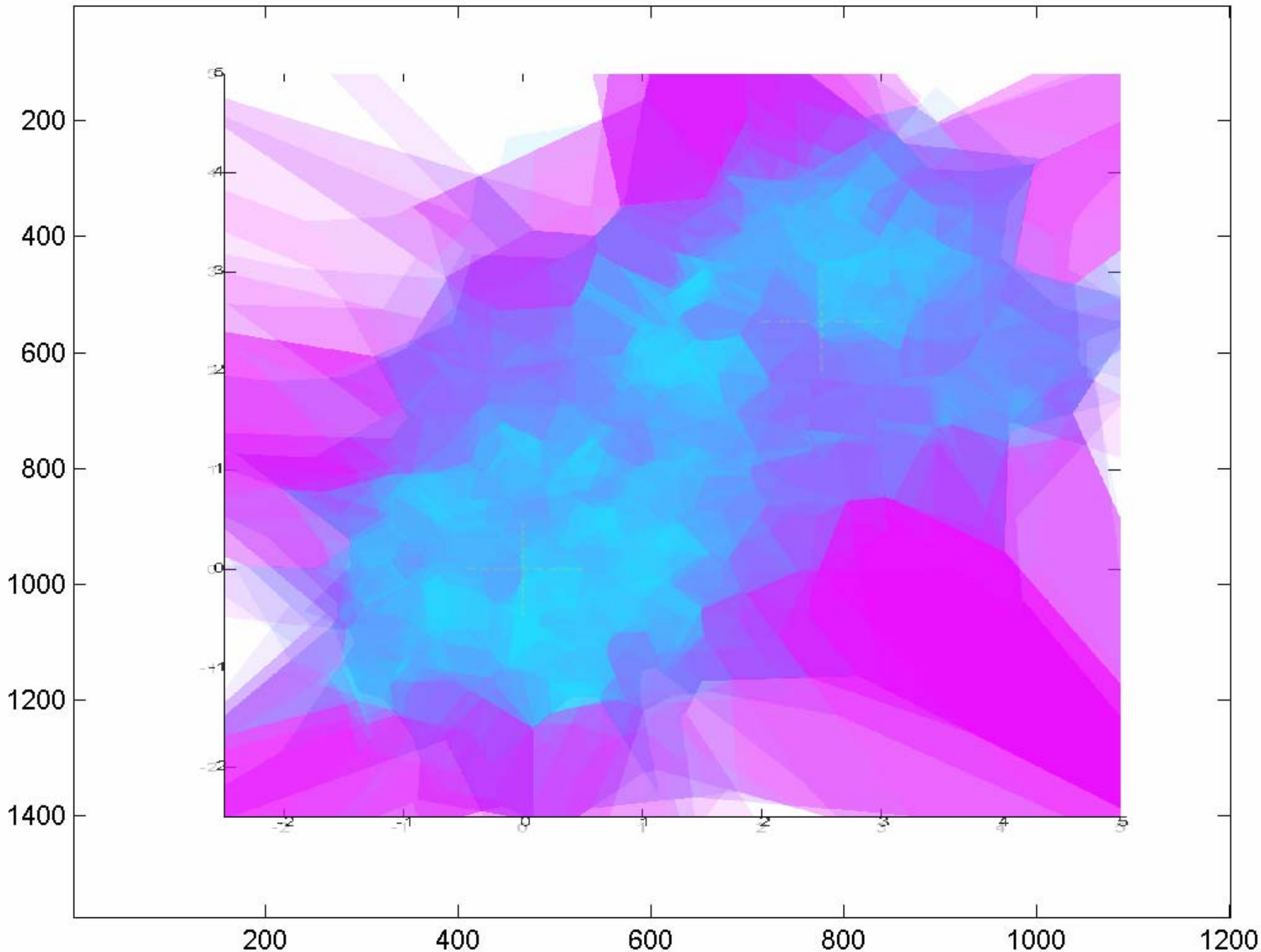


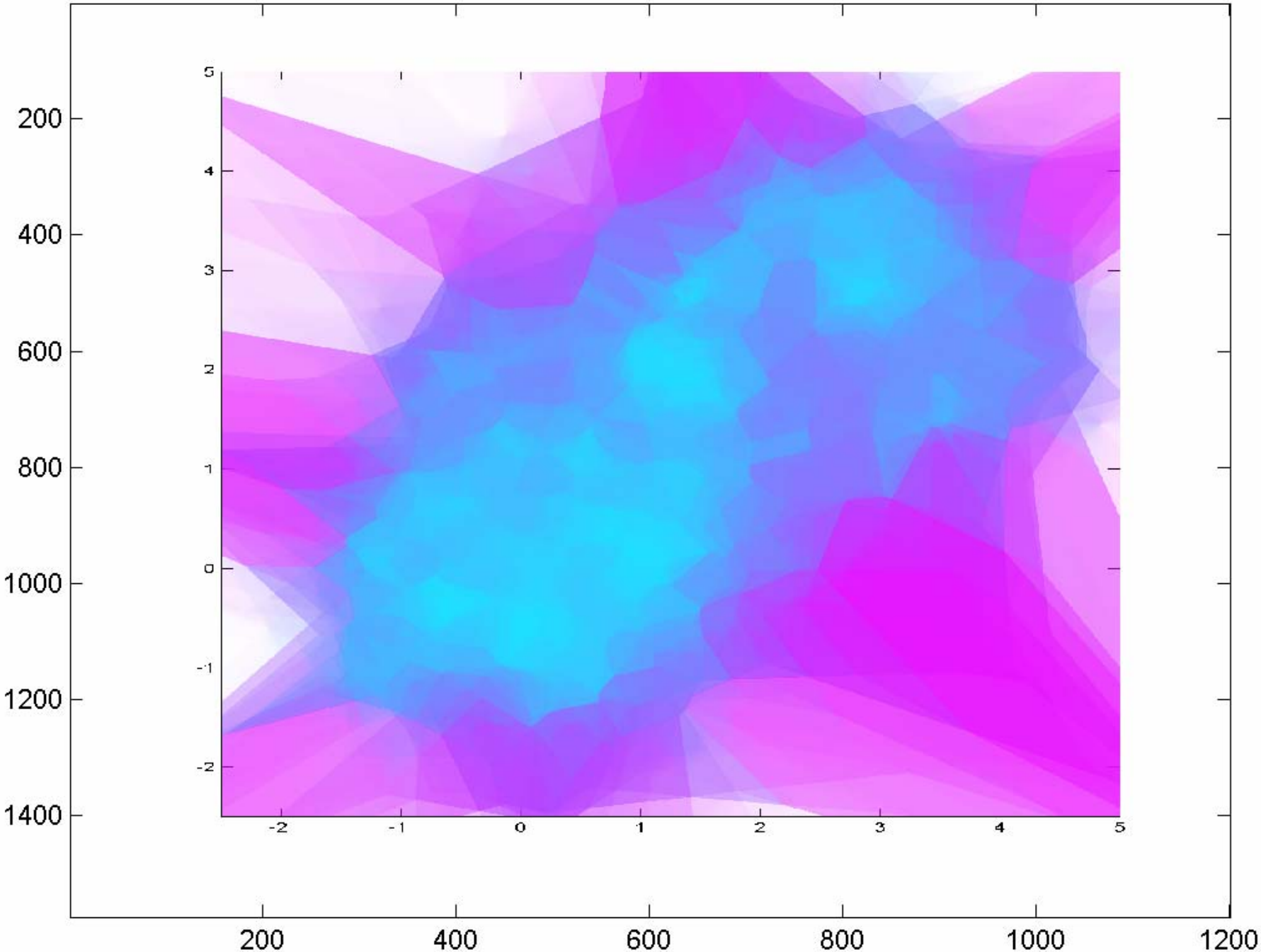


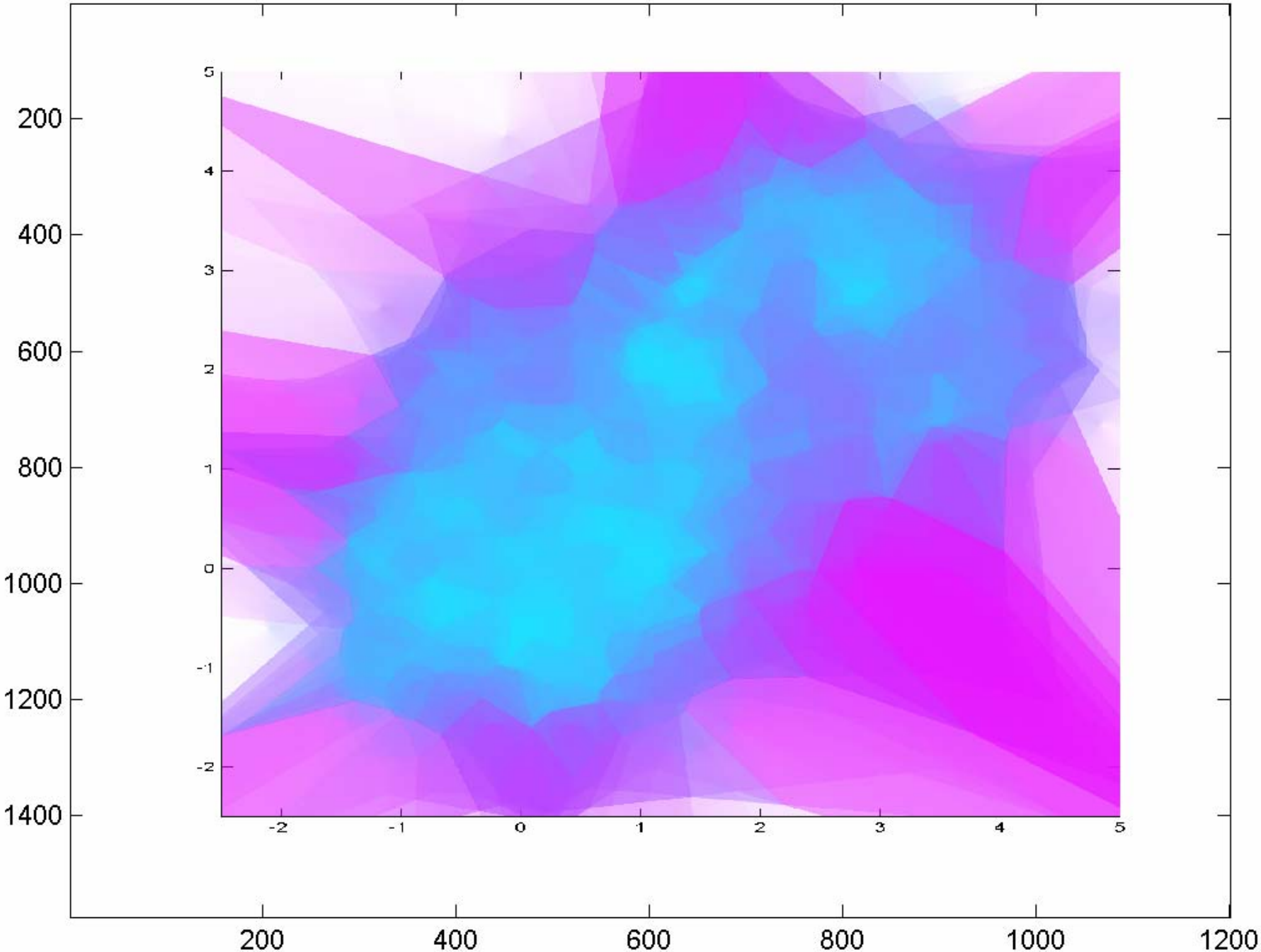

Gamma-rays from Galactic Center 12 blocks
15
10
5
0
-5
-10
-15
Latitude
-30
-20
-10
0
10
20
30
Longitude

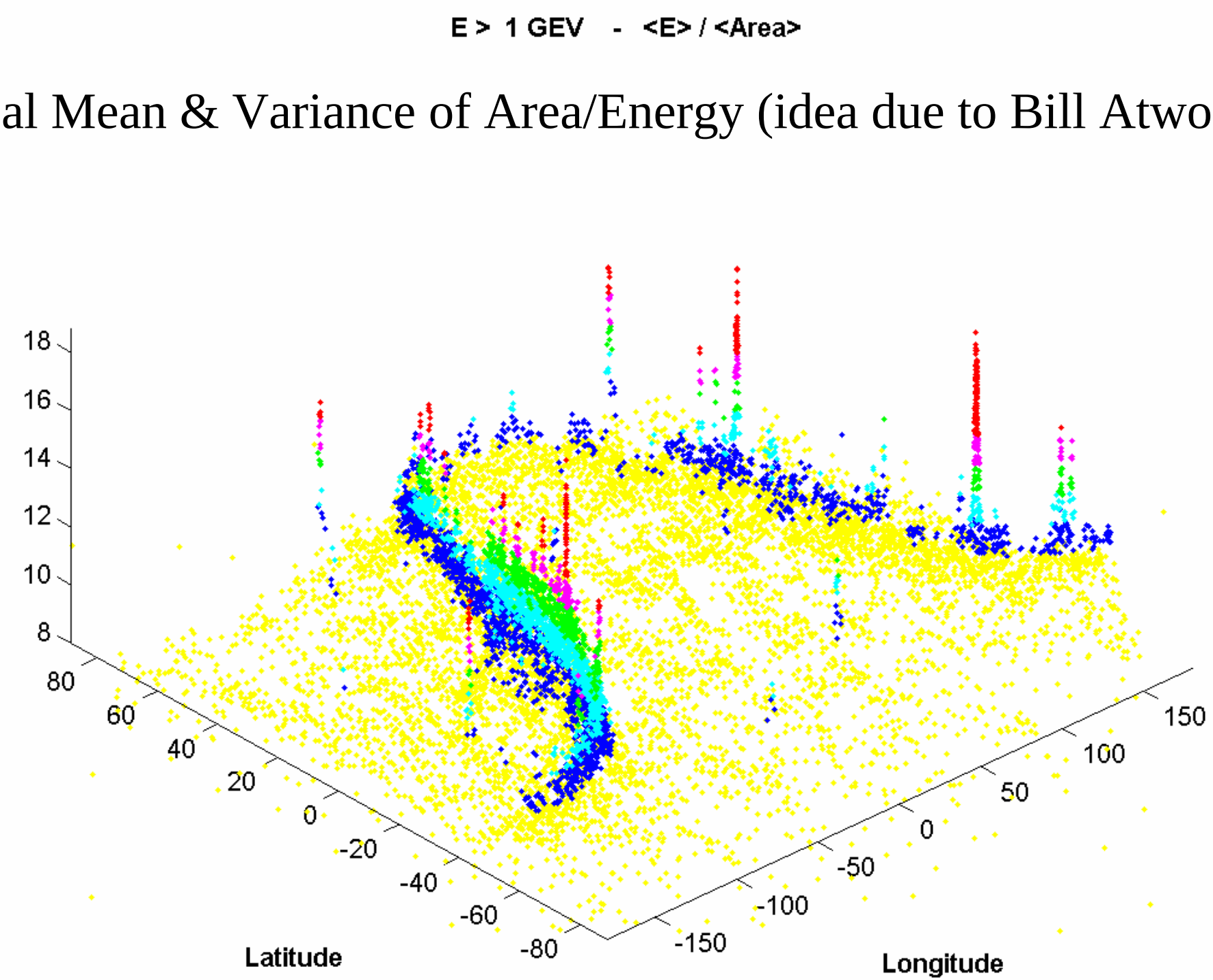

5
4
3
2
1
0
-10
-10.5
-11
-11.5
-12
-12.5
-13
90
91
92
93
94


2D Synthetic Bootstrap Example: Raw Data
6
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3
4
5
6

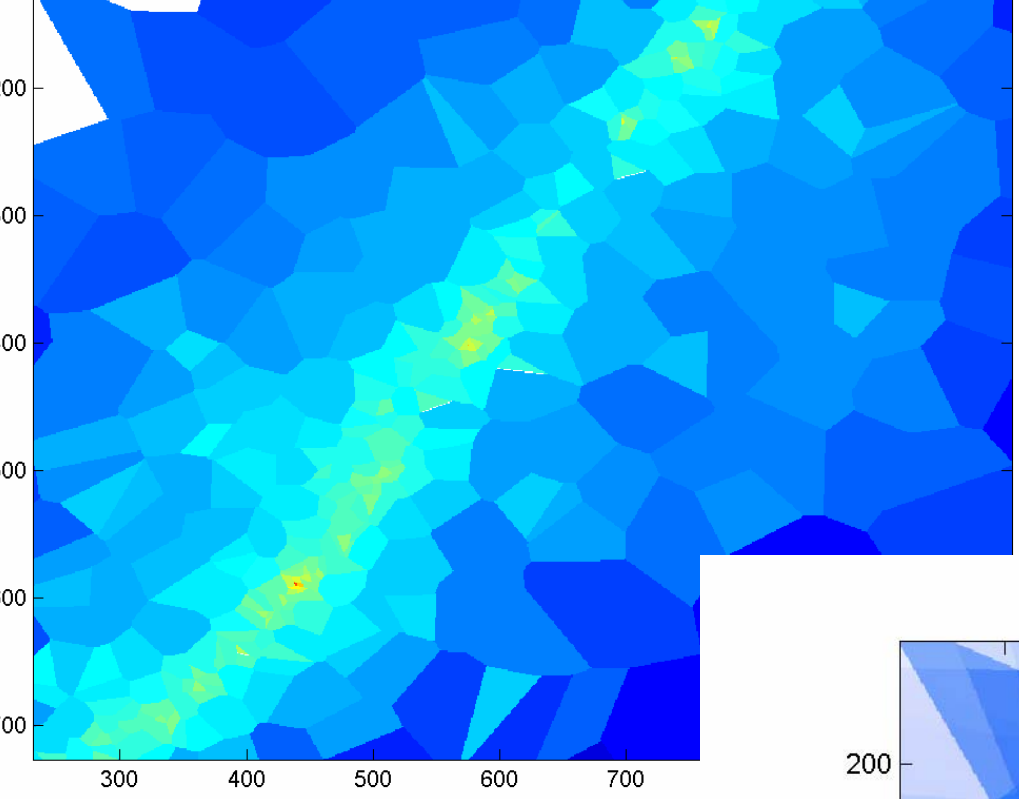

200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200

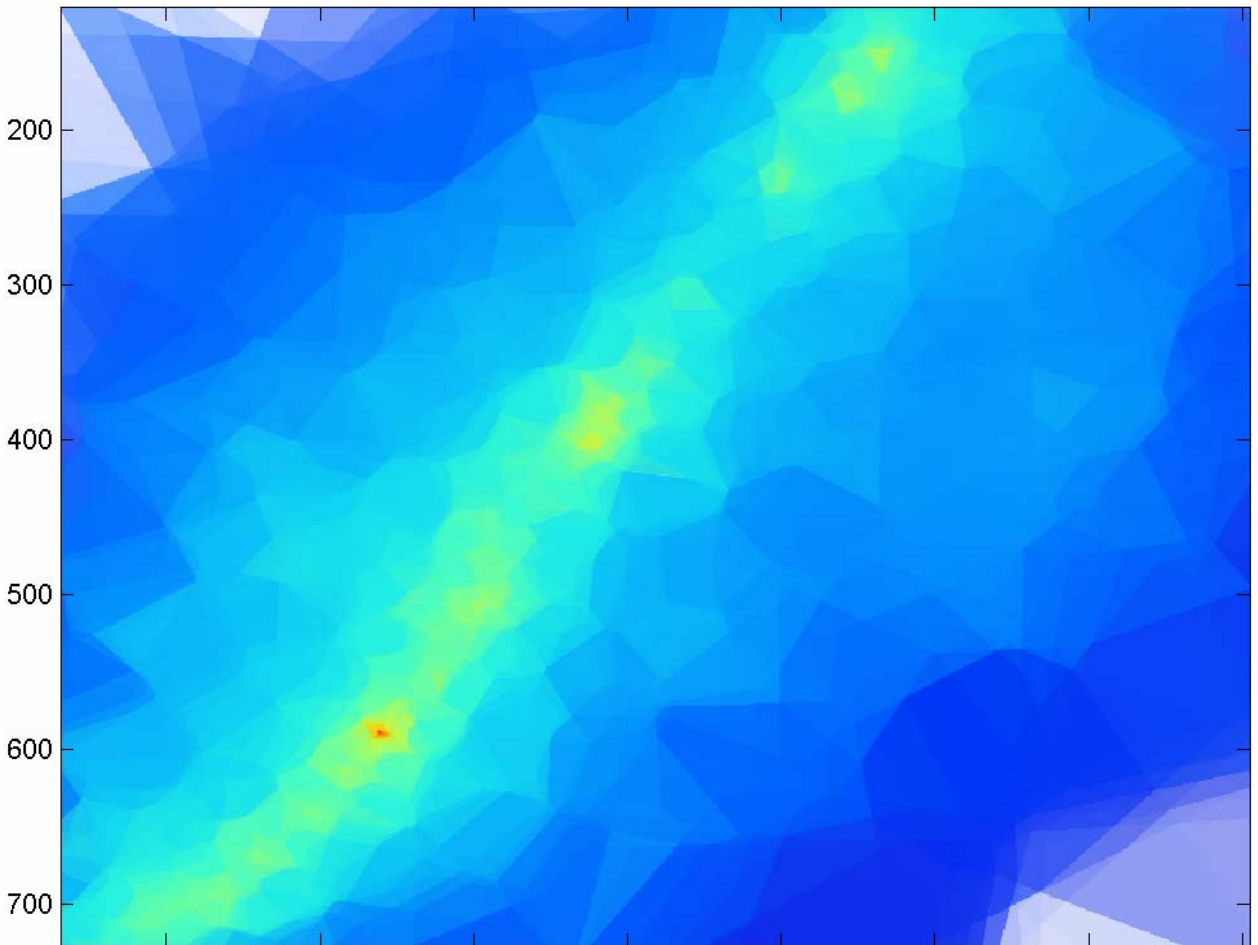

200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5


200
400
600
800
1000
1200
1400
200
400
600
800
1000
1200
5
4
3
2
1
0
-1
-2
-2
-1
0
1
2
3
4
5



Local Mean & Variance of Area/Energy (idea due to Bill Atwood)


300
400
500
600
700


200
300
400
500
600
700