Multiscale Data Analysis through Optimal Segmentation

Jeffrey.D.Scargle@nasa.gov
Space Science Division
NASA Ames Research Center

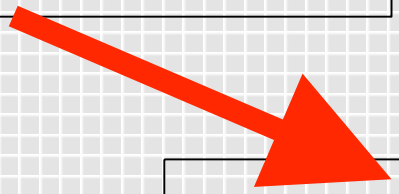
Brad Jackson, Mathematics Department, San Jose State University
Jay Norris, NASA Goddard Spaceflight Center
Michael Way, Paul Gazis, Chris Henze, Creon Levit, NASA-ARC
Adam Roy, San Francisco State University
Mahmoud Quweidar, University of Texas, Brownsville

Multiscale Geometric Methods in Astronomical Data Analysis
Institute for Pure and Applied Mathematics, UCLA
November 8-12, 2004

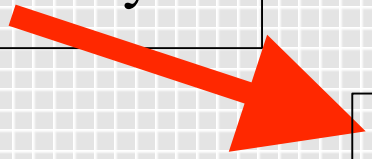
From Data to Astronomical Goals

Input data:
measurements
distributed in some space

sequential data (time series, spectra, ...)
images, photon maps
redshift surveys
higher dimensional data



Intermediate product:
estimate of
signal, image, density ...



End goal: estimates of values of scientifically relevant quantities

Astronomical Goals

Exploratory analysis: little prior information – few assumptions

Use simplest possible nonparametric model: *piecewise constant*
(allows exact calculation of marginalized likelihoods)

Take account of:

observational noise
exposure variations
arbitrary sampling, gaps
point spread functions (1D, 2D+ in progress)

} intrinsic

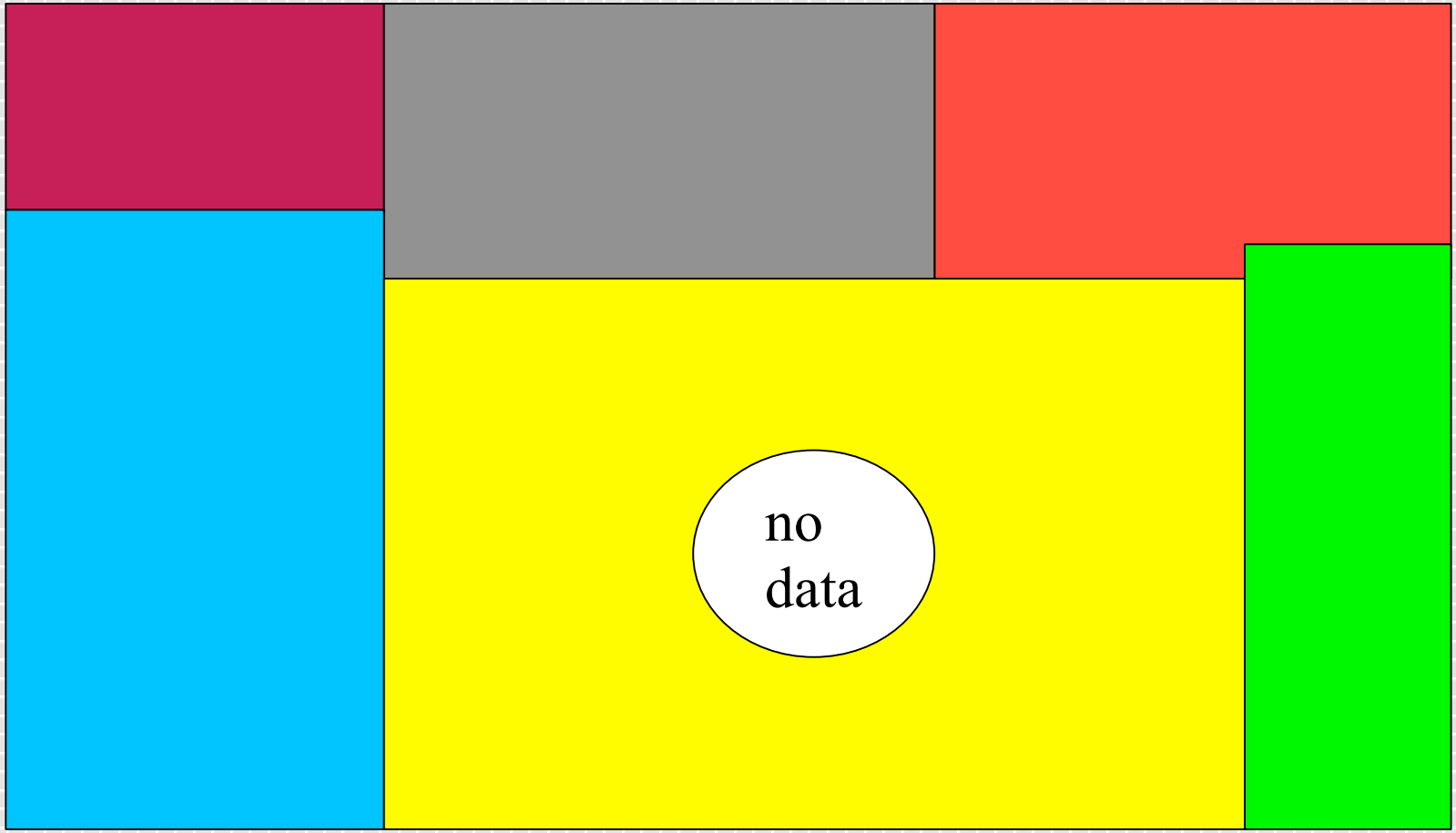
Just say no to:

pretty pictures; smoothing; continuous representations
bins or pixels (unless raw data in this form)
methods tuned for specific structures, *e.g.* beamlets
resampling
sensitivity to some global structures, *e.g.* periodic

} constructs

Piecewise Constant Model (Partition)

Can simply ignore gaps – model says nothing about signal there.



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; uses no explicit smoothing

Adaptive Kernel Smoothing

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

The Problem: Signal/Density Estimation

The data: Data cells: points, event counts, measurements, *etc.*
Noise: known distribution, not necessarily additive (*e.g.* Poisson)

Distributed over a data space:
- 1D: time series, sequential data, ...
- 2D: images, photon maps, star/galaxy catalogs, ...
- 3D: galaxy redshift surveys, energy-resolved photon maps
- xD: time & energy resolved photon maps; data mining

Goals: Nonparametric representation of the underlying signal or density

Easy analysis of signal structure (*e.g.* clusters in point density)

Suppression of noise (using prior knowledge of noise statistics)

Objective, automatic analysis of large data sets (data mining)

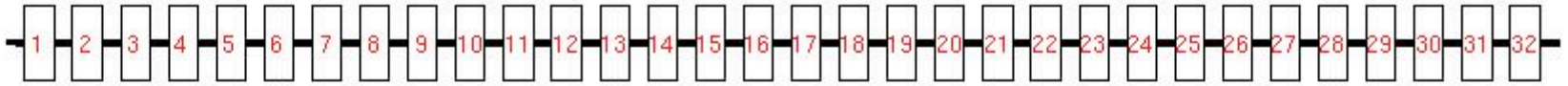
Signal Model: Piecewise Constant

Represent signal as *constant over elements of a partition of the data space*.

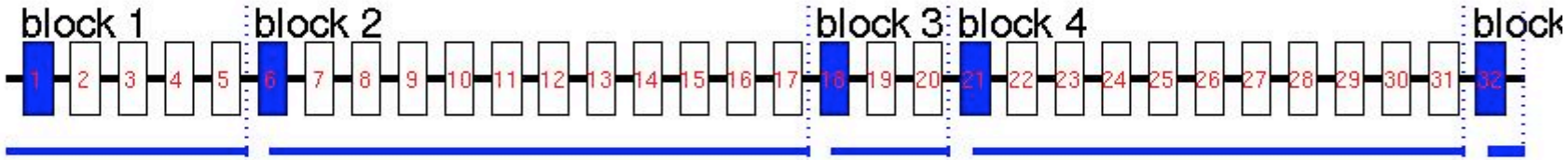
Optimize model by maximizing model fitness over all possible partitions.

- Nonparametric
- Few prior assumptions about the signal:
  - prior on signal amplitudes
  - prior on number of partition elements
- No limitation on the resolution in the independent variable.
- Representation, while discontinuous, is convenient for further analysis
- Local structure, not global

(a)


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

(b)


block 1
block 2
block 3
block 4
block
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

DATA CELLS: Definition

data space: the set of possible measurements in some experiment

d*ata 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*.

In a specific application, the data cells may or may not:

- be in one-to-one correspondence to the measurements
- partition the entire data space
- overlap each other
- leave gaps between cells
- contain information on adjacency to other cells

DATA CELLS: Event (Point) Data

Measurements: Point coordinates
Data Space: Space of any dimension
Signal: Point density (deterministic or probabilistic)
Data Cell: Voronoi cells for the data points

Suf. Statistics N = number of points in block
V = volume of block

Max Likelihood: $(N / V)^{N} e^{-N}$
Posterior: $N! (V - N)! / (V+1)!$

Example: any problem usually approached with histograms (1D)
positions of objects from a sky survey (2D)
positions of objects in a redshift survey (3D)

DATA CELLS: Serial Measurements

Measurements: Values and error distribution of dependent variable at given values of independent variable, e.g. $X(t) \sim N(x,\sigma)$

Data Space: Interval, area, volume, ...

Signal: Variation of dependent variable

Data Cell: Measurement point

Suf. Statistics $x_n = x_n(t_n)$, t_n, parameter(s) of error distribution: σ_n, ...

log posterior:

Example: time series, spectra, images, SOM output ...

DATA CELLS: Distributed Measurements

Measurement: Dependent variable averaged over a range of independent variable

Data Space: Space of any dimension

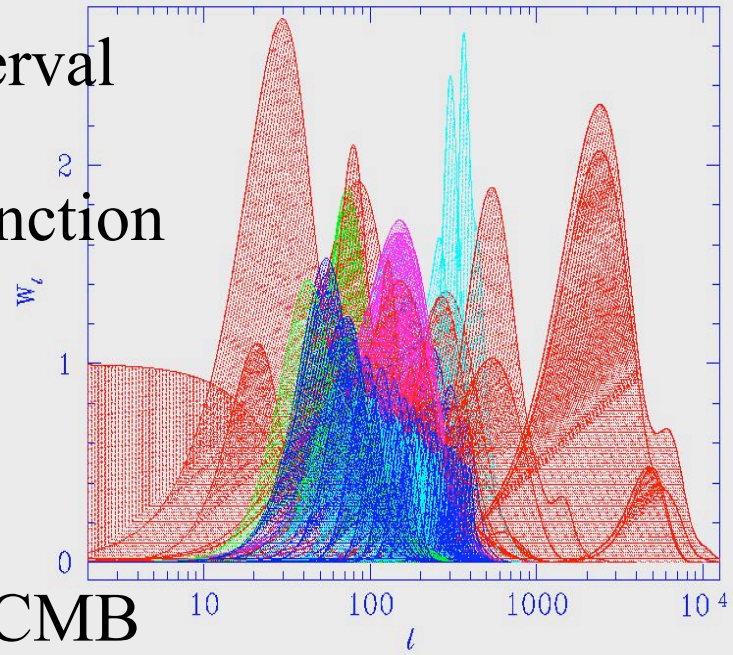
Signal: Physical variable

Data Cell: Measurement and its interval

Suf. Statistics: x, σ_x , W(t) = window function

Posterior: see Bretthorst, G-S orthogonalization

Example: spatial power spectra of CMB



Blocks

Block: a set of data cells

Two cases:

_ Connected (can't break into distinct parts)

_ Not constrained to be connected

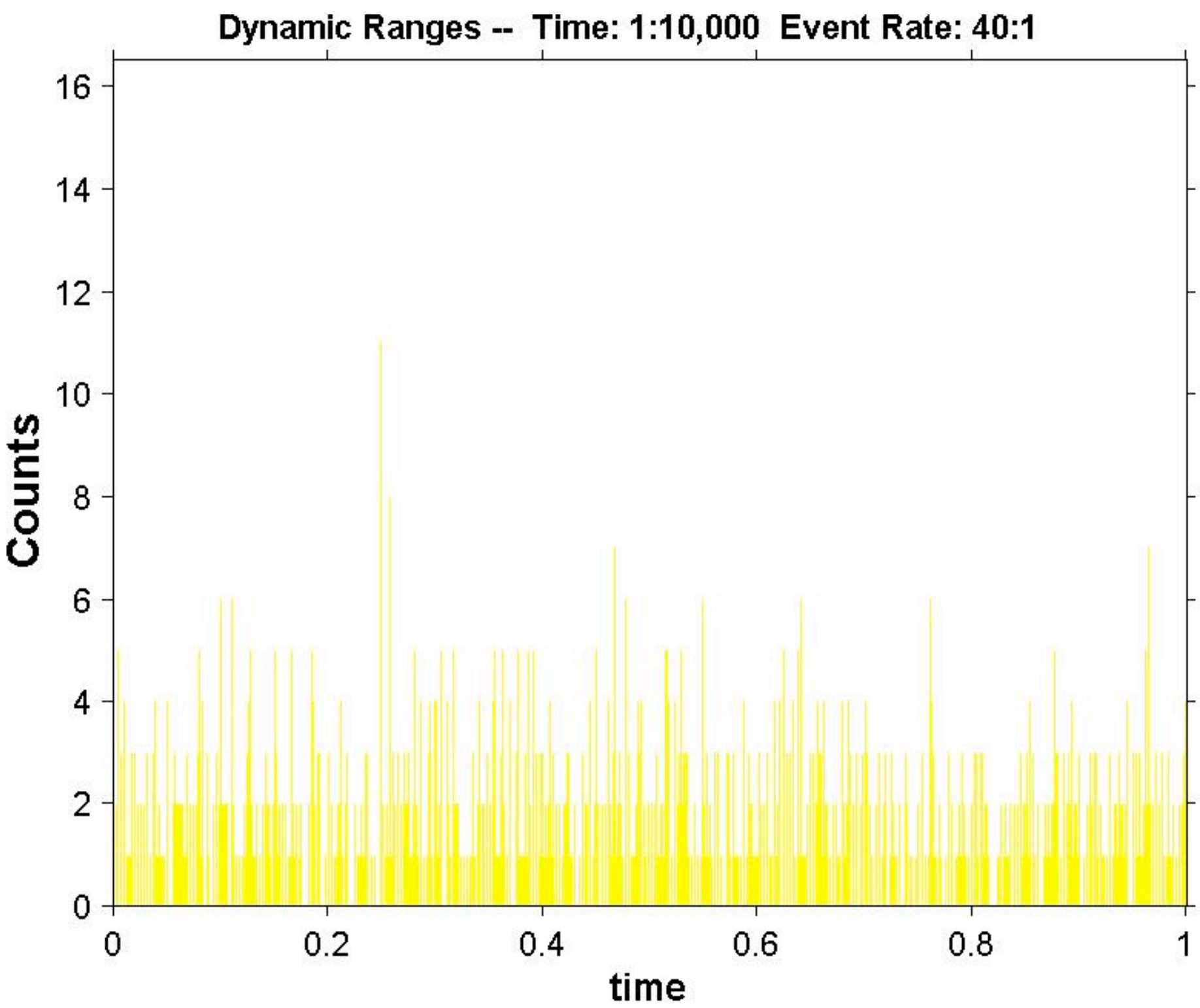
Model = set of blocks

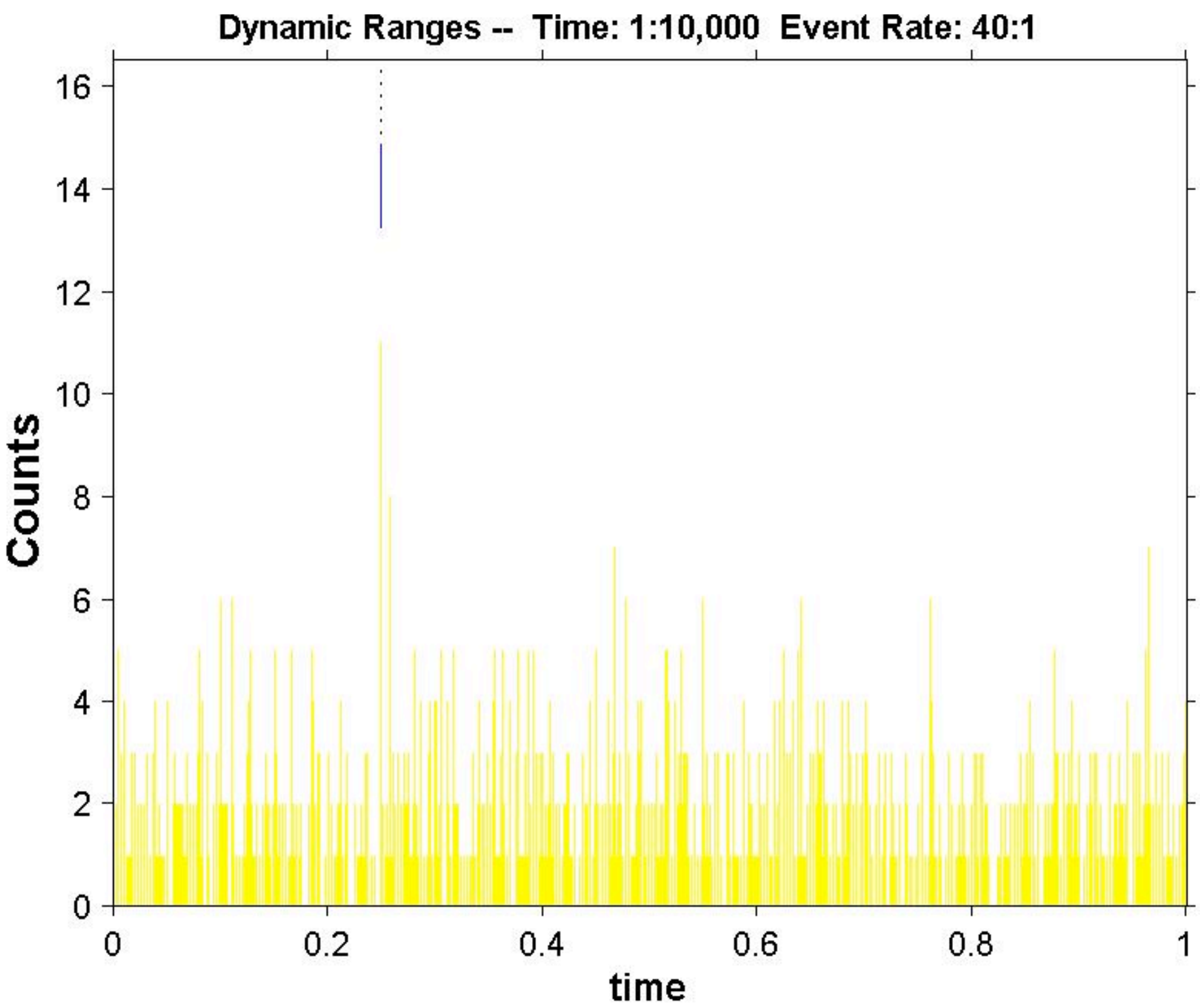
Fitness function:

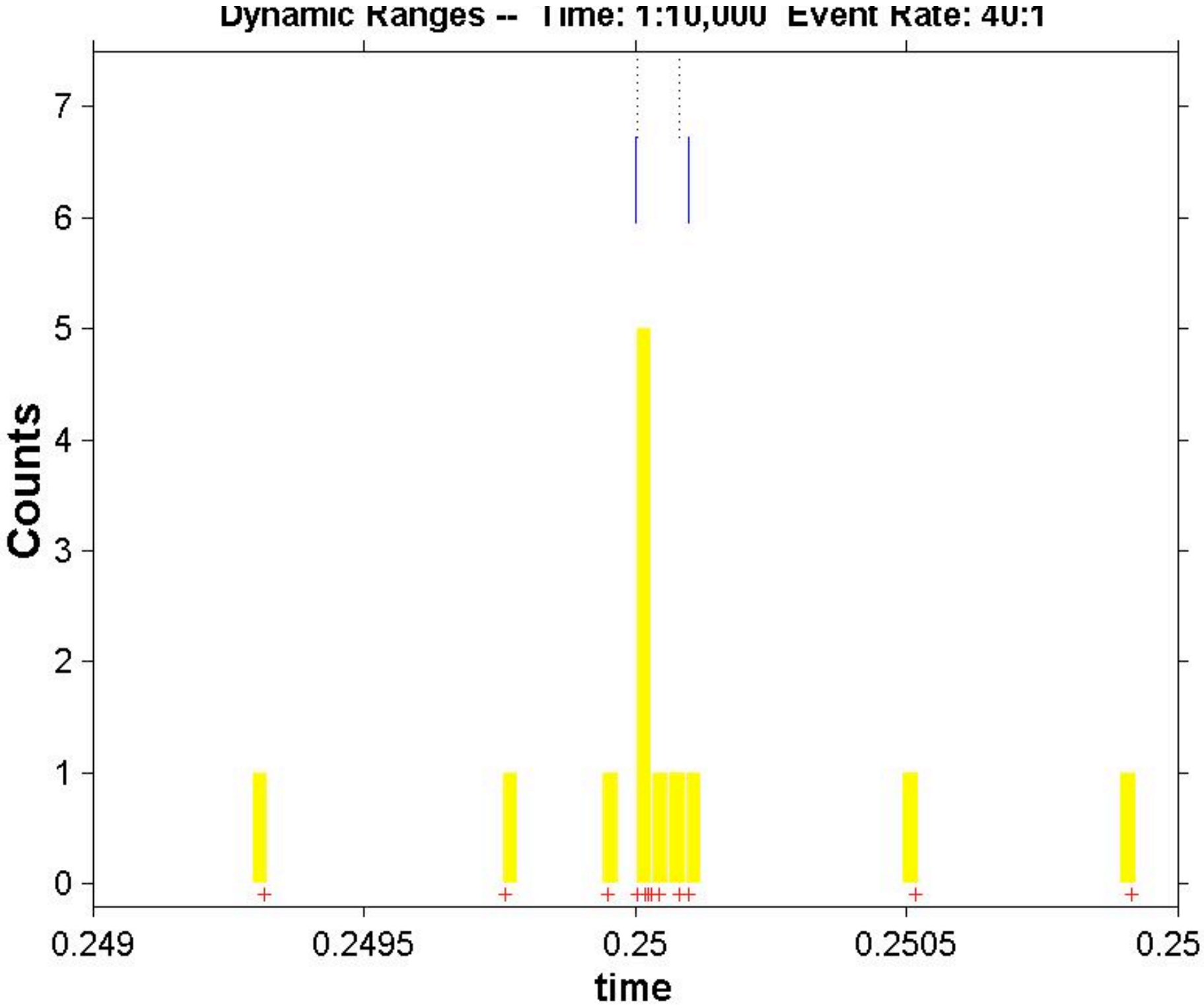
F(Model) = sum over blocks F(Block)

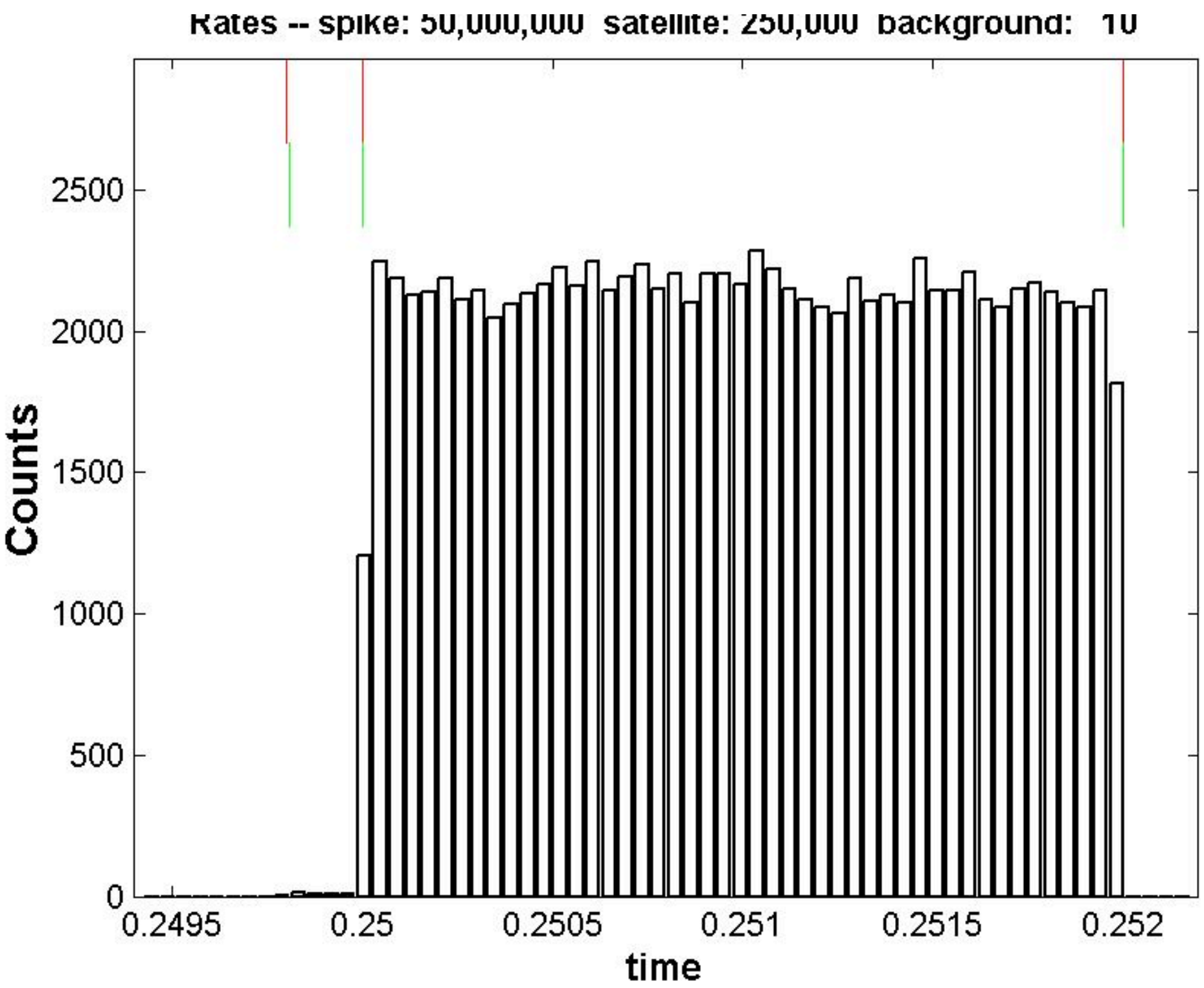
The Optimizer

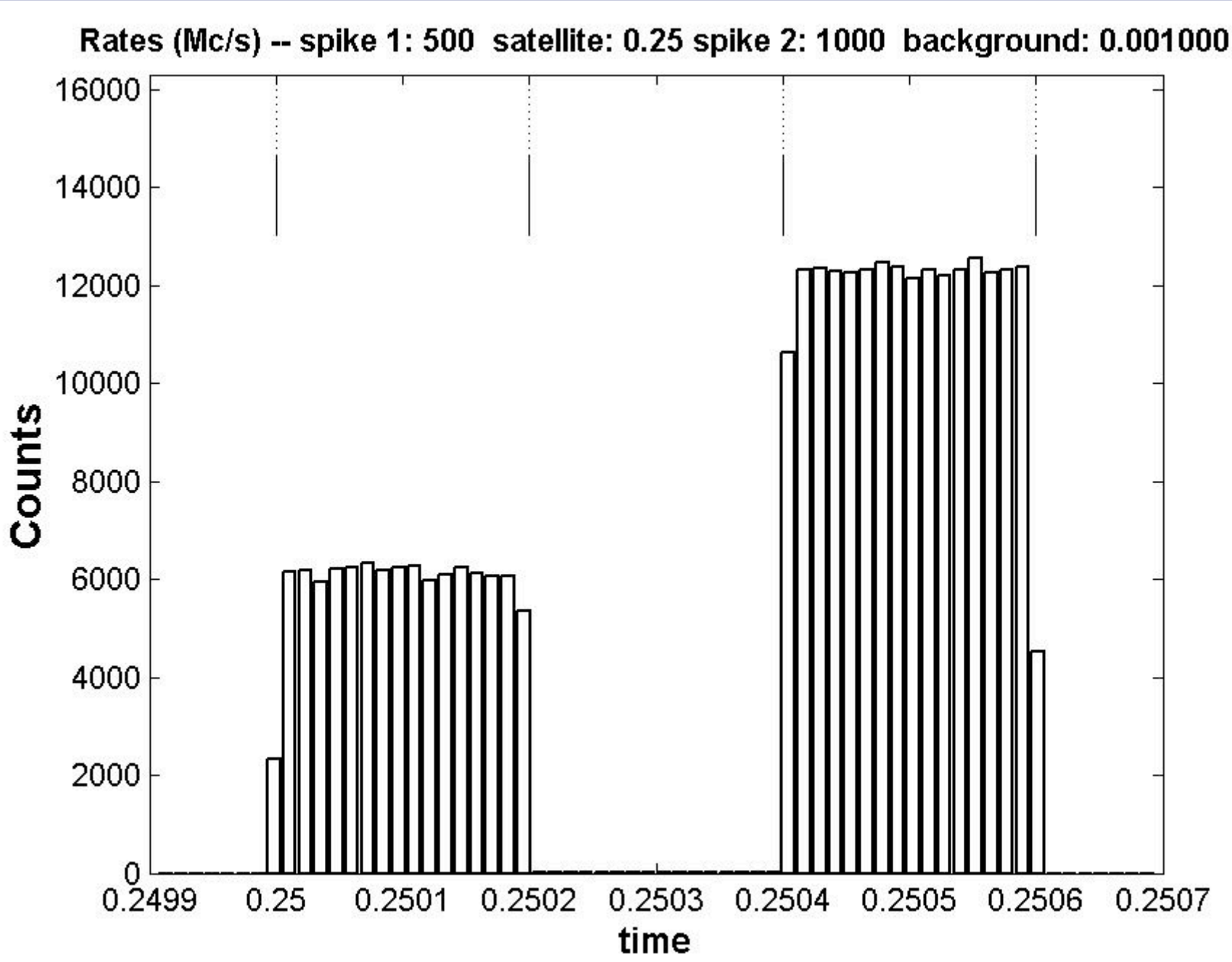
```
best = []; last = [];
for R = 1:num_cells
   [ best(R), last(R) ] = max( [0 best] + ...
         reverse( log_post( cumsum( data_cells(R:-1:1, :) ), prior,  type )  ) );

    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
```

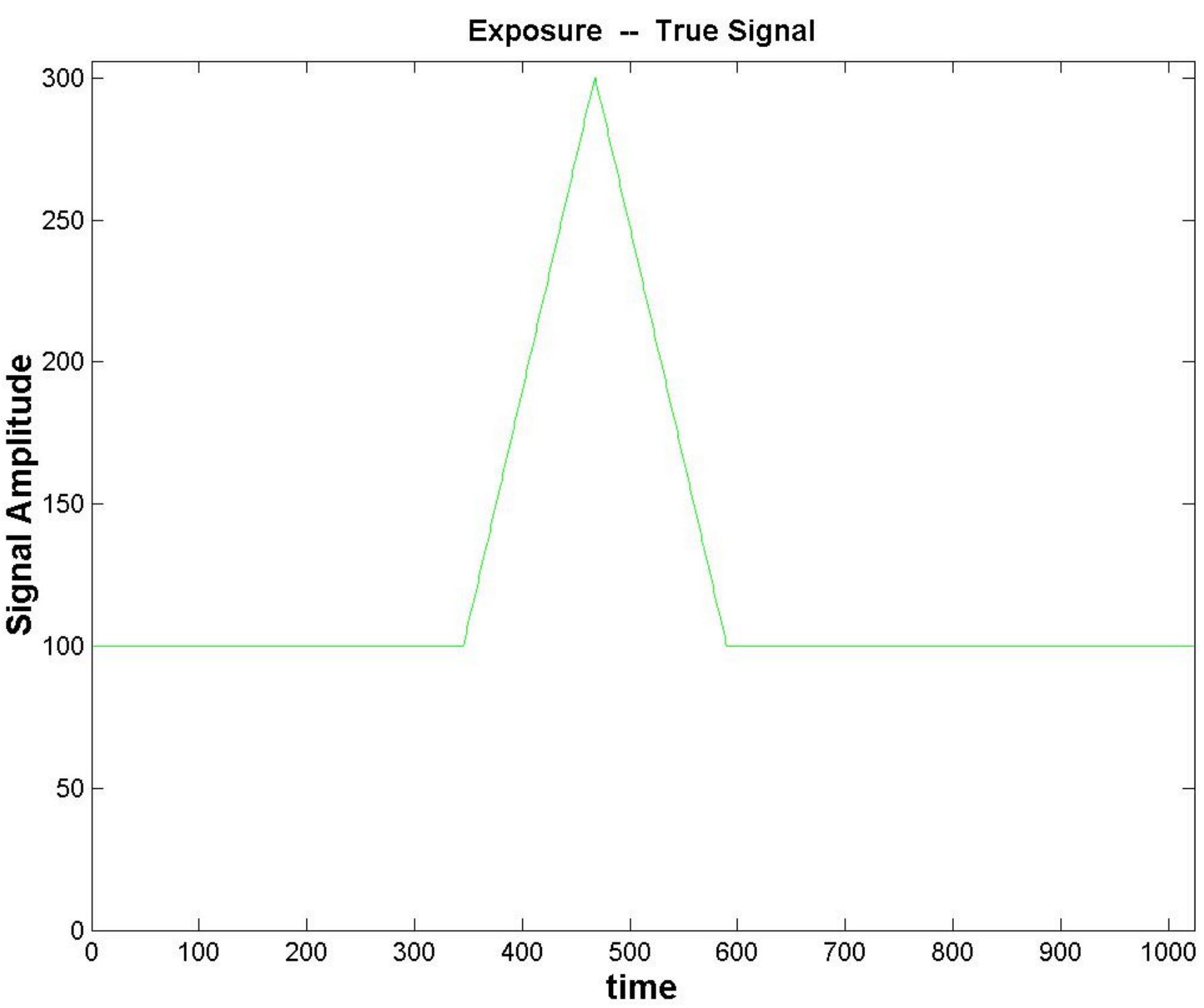

Dynamic Ranges -- Time: 1:10,000 Event Rate: 40:1
Counts
16
14
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
time

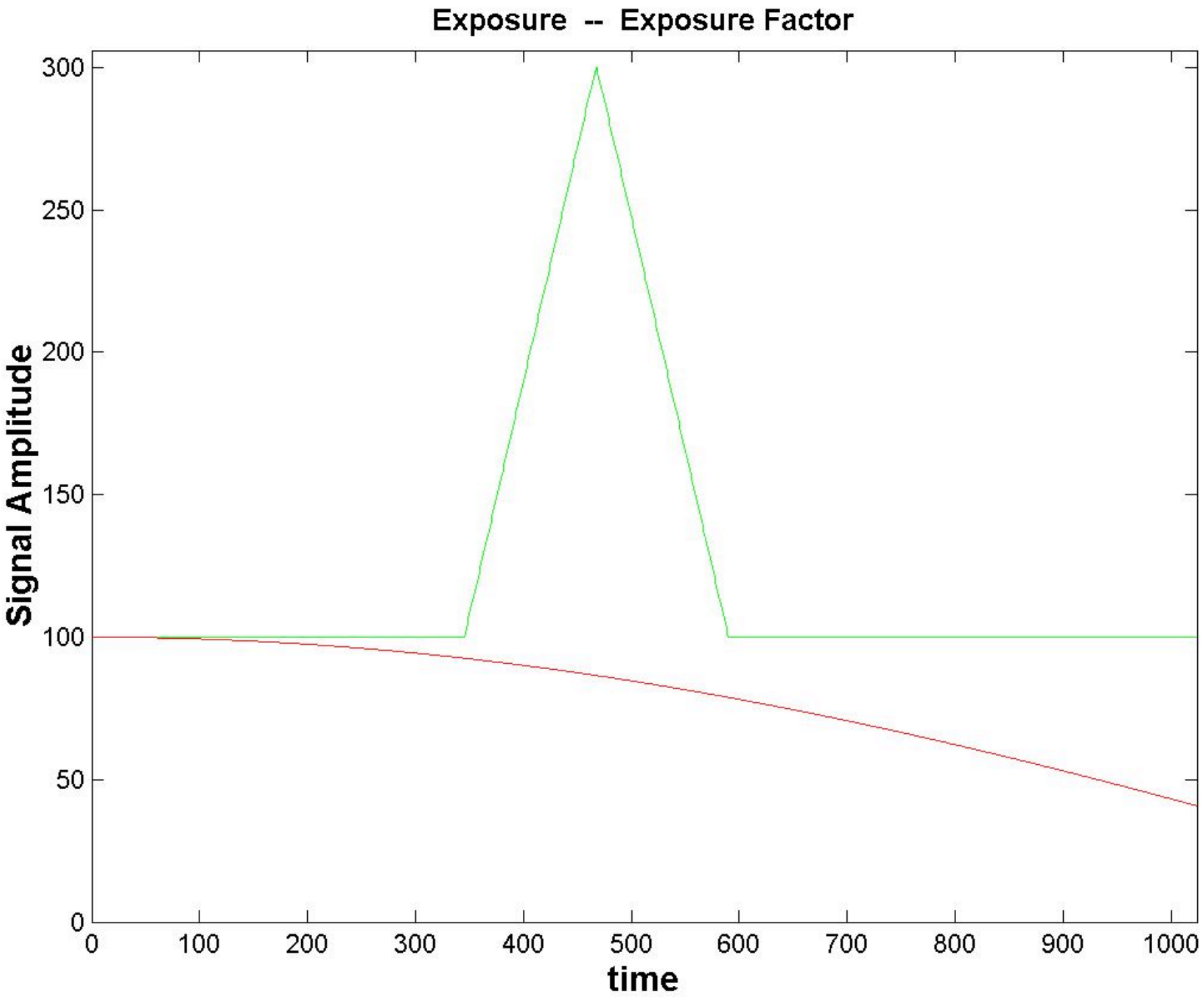

Dynamic Ranges -- Time: 1:10,000 Event Rate: 40:1
Counts
16
14
12
10
8
6
4
2
0
0
0.2
0.4
0.6
0.8
1
time

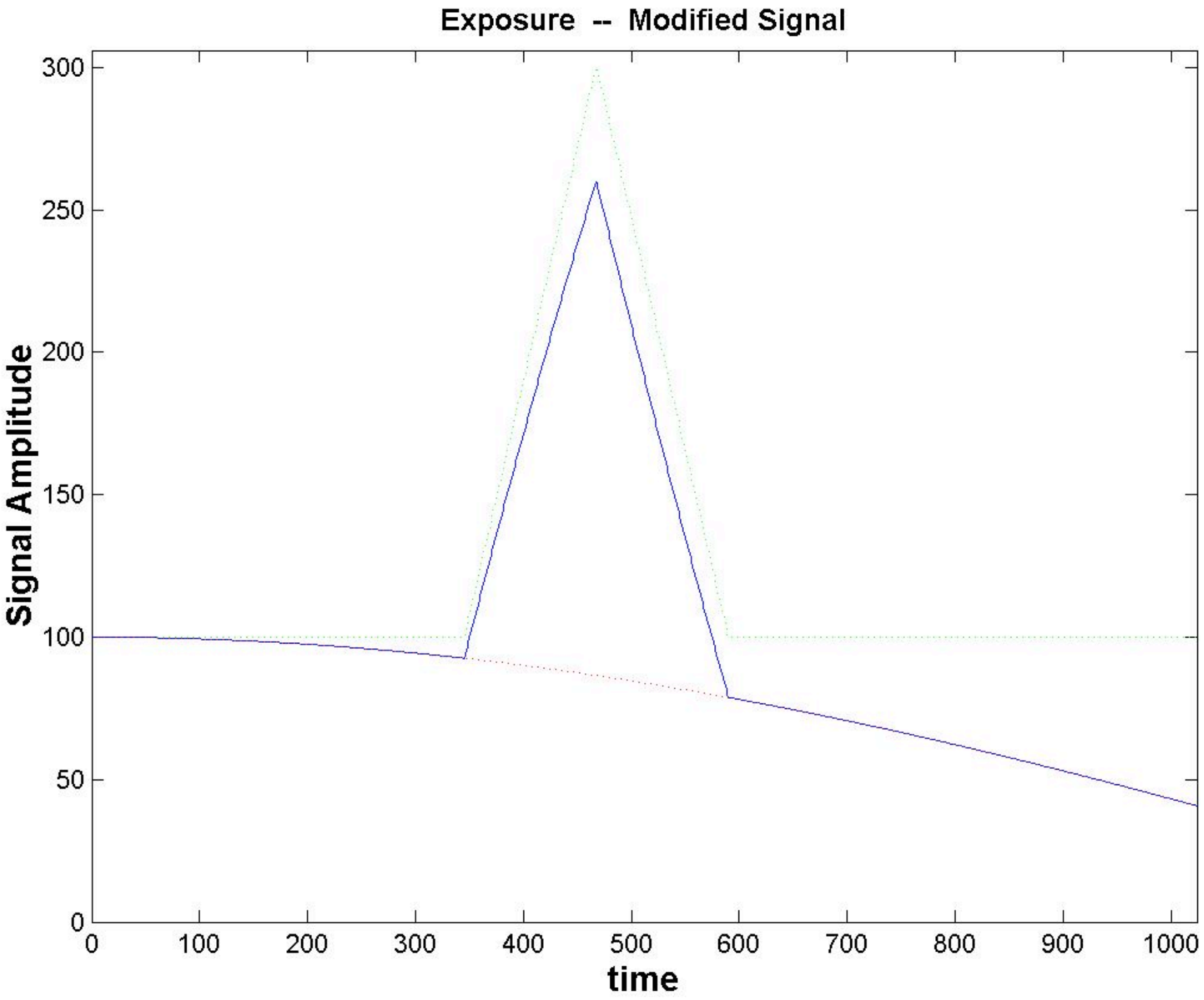

Dynamic Ranges -- Time: 1:10,000 Event Rate: 40:1
Counts
time
0
1
2
3
4
5
6
7
0.249
0.2495
0.25
0.2505
0.25

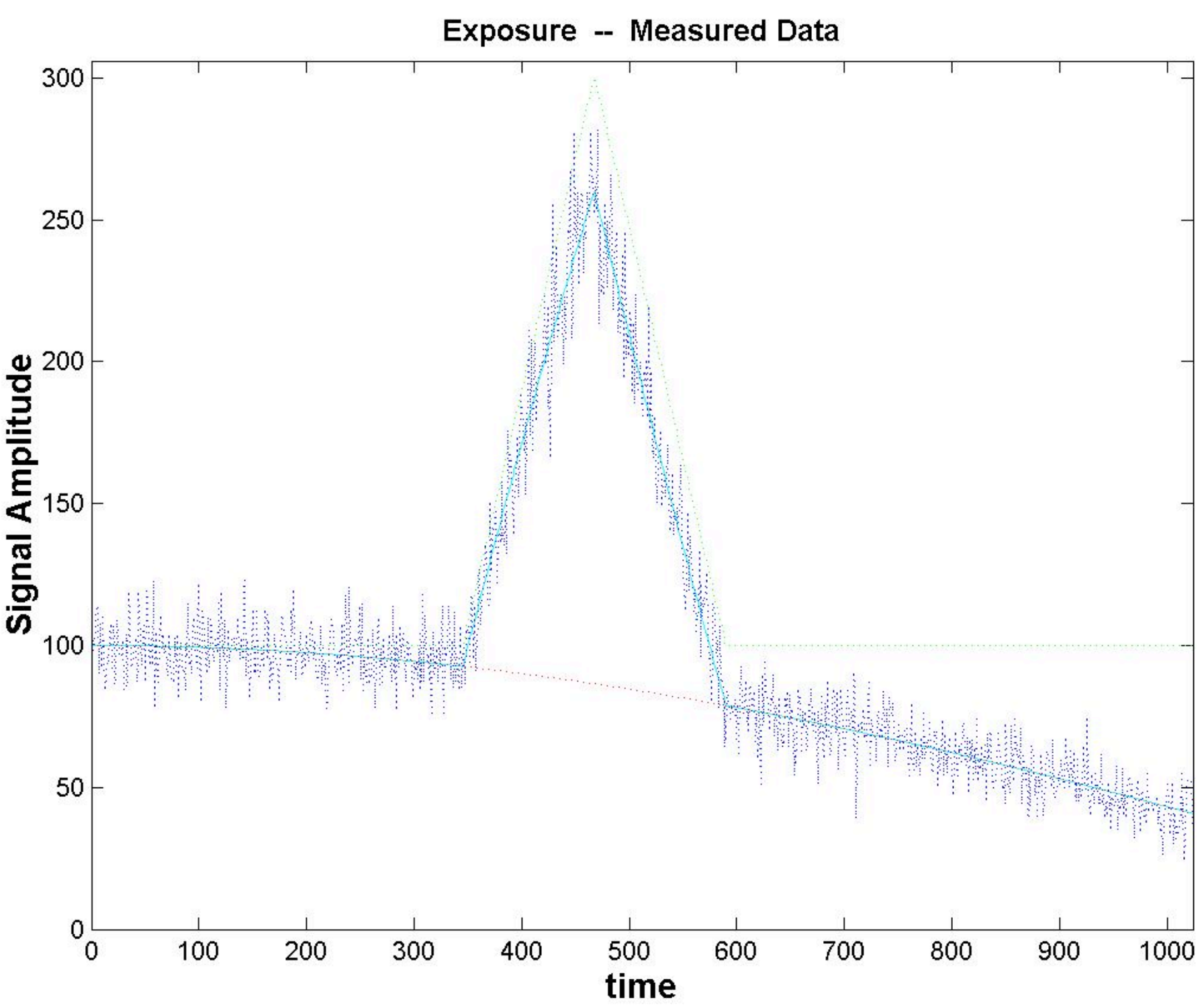

Rates -- spike: 50,000,000 satellite: 250,000 background: 10
2500
2000
1500
1000
500
0
Counts
0.2495
0.25
0.2505
0.251
0.2515
0.252
time

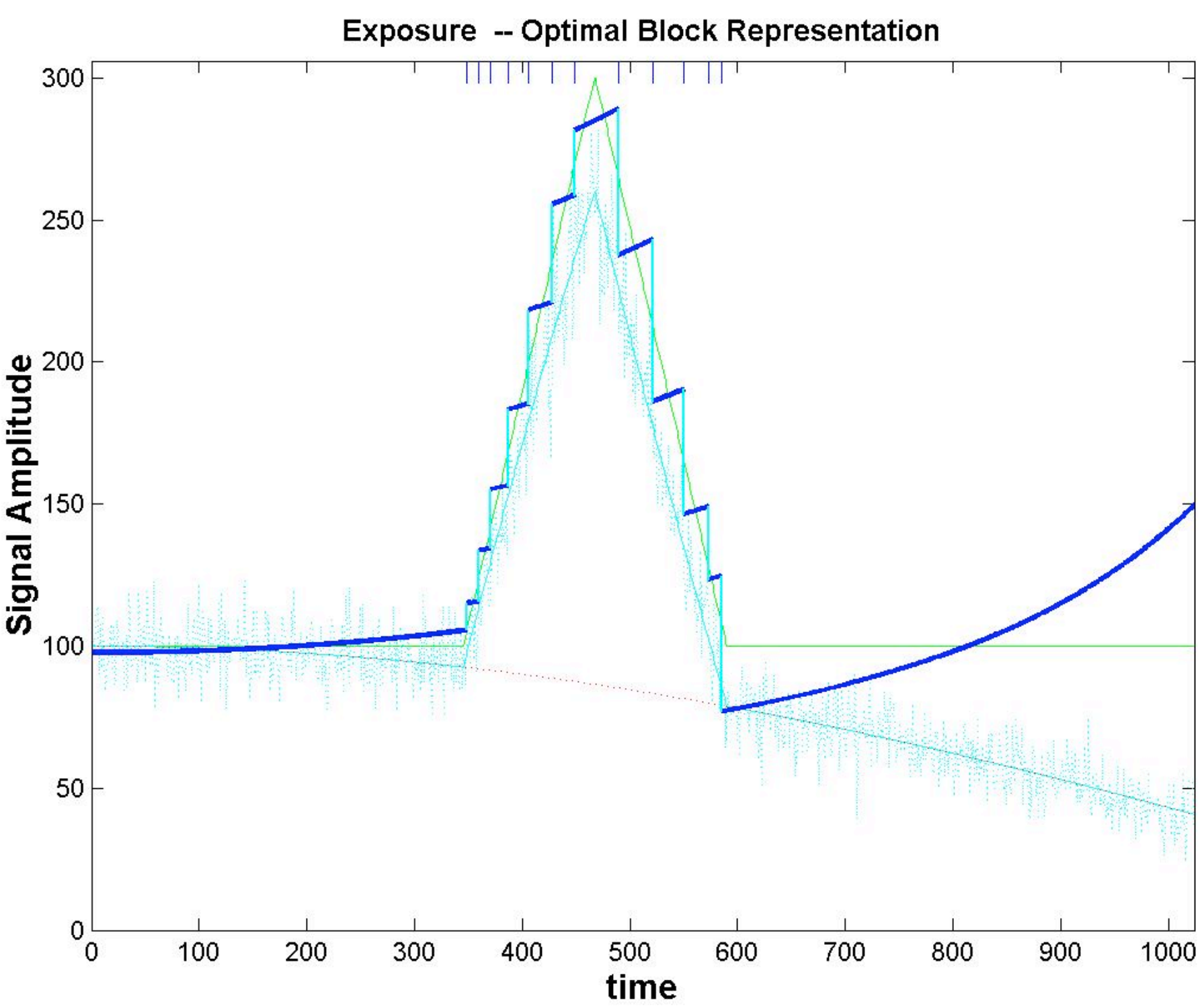

Rates (Mc/s) -- spike 1: 500 satellite: 0.25 spike 2: 1000 background: 0.001000
16000
14000
12000
10000
8000
6000
4000
2000
0
Counts
0.2499
0.25
0.2501
0.2502
0.2503
0.2504
0.2505
0.2506
0.2507
time


Exposure -- True Signal
Signal Amplitude
time
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800
900
1000


Exposure -- Exposure Factor
Signal Amplitude
time
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800
900
1000


Exposure -- Modified Signal
Signal Amplitude
time
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800
900
1000


Exposure -- Measured Data
Signal Amplitude
time
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800
900
1000


Exposure -- Optimal Block Representation
Signal Amplitude
time
0
50
100
150
200
250
300
0
100
200
300
400
500
600
700
800
900
1000

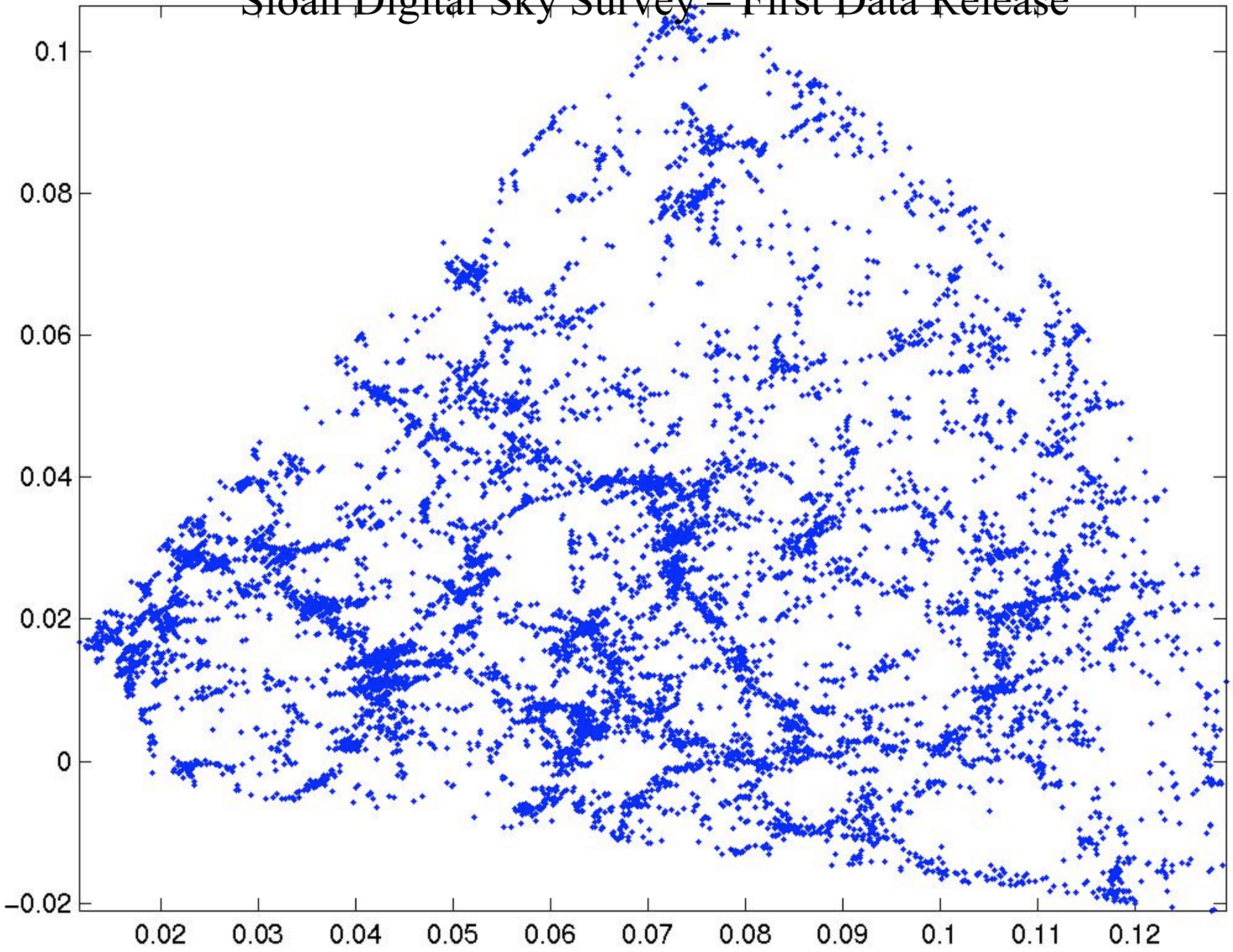
Distribution of Galaxies in 3-space (and time)

Nature of the astrophysical process:

- initial density fluctuations = (Gaussian?) random field
- fluctuations grow over time due to gravity
- end product = discrete galaxies

What is the best mathematical model of this process?

Sloan Digital Sky Survey – First Data Release



Distribution of Galaxies in 3-space (and time)

Structures observed in redshift surveys:

_ Voids (3D underdense cells) -- “voids”

_ Sheets (2D density excesses) -- “Zeldovich pancakes”

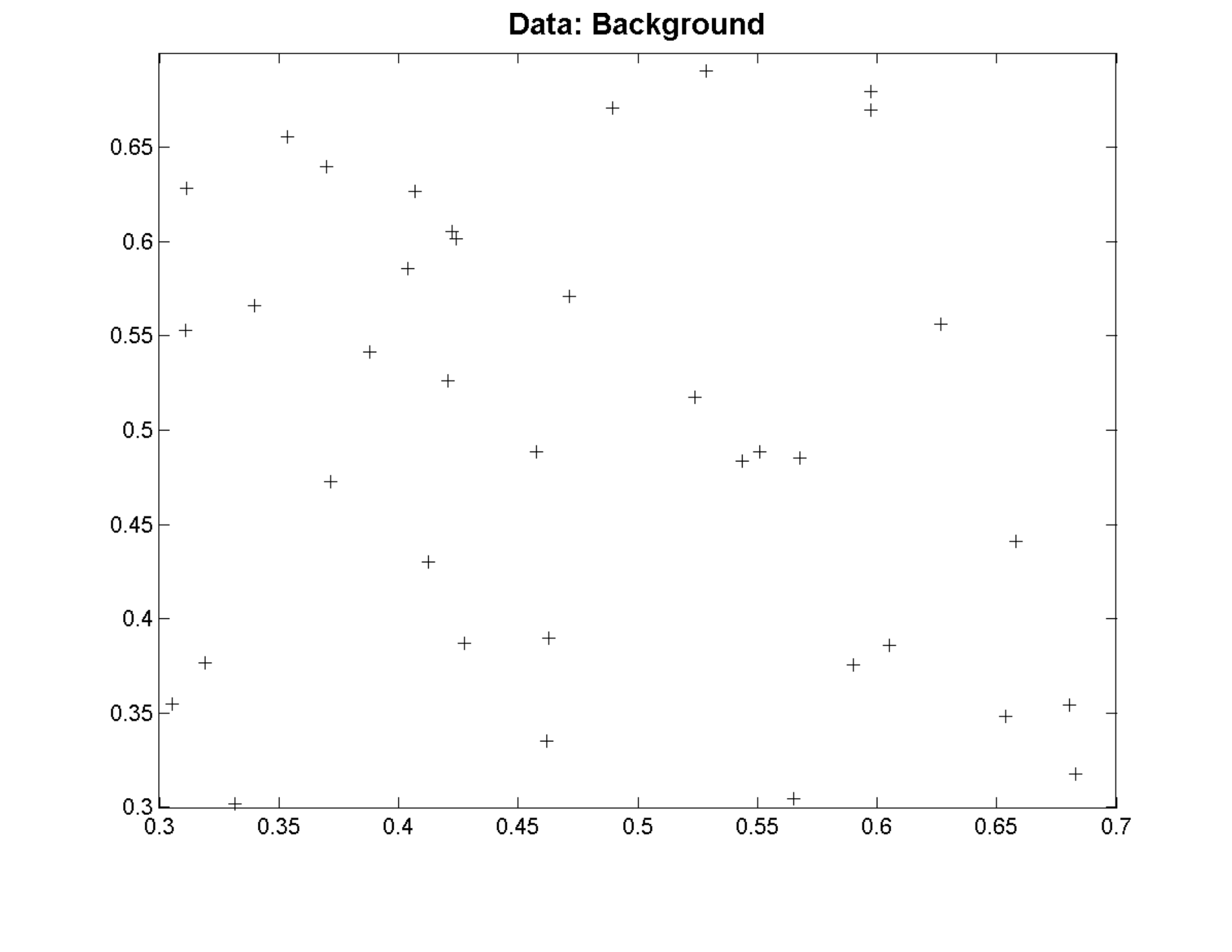
_ Nodes (1D density excesses) -- “classical clusters”

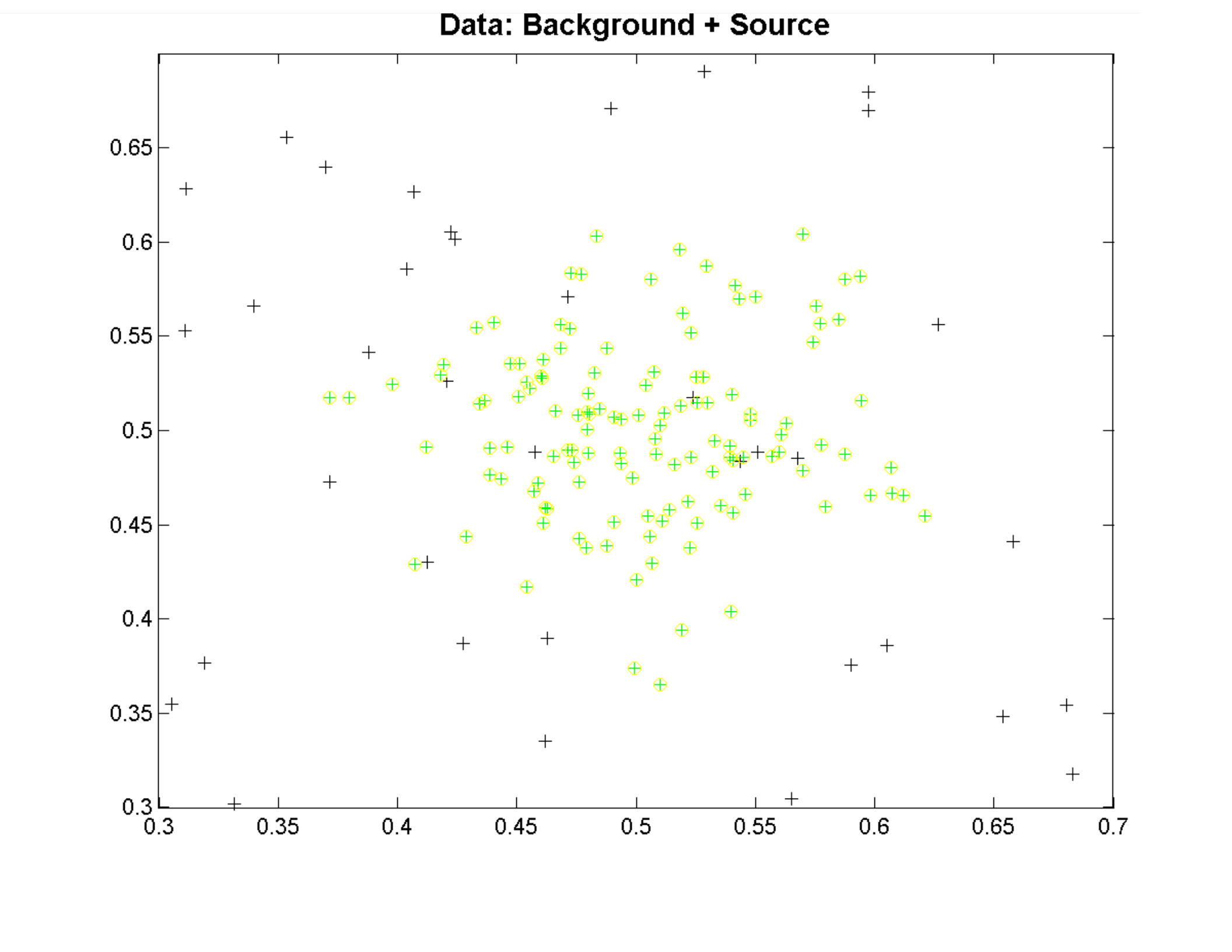
Optimum Partitions in Higher Dimensions

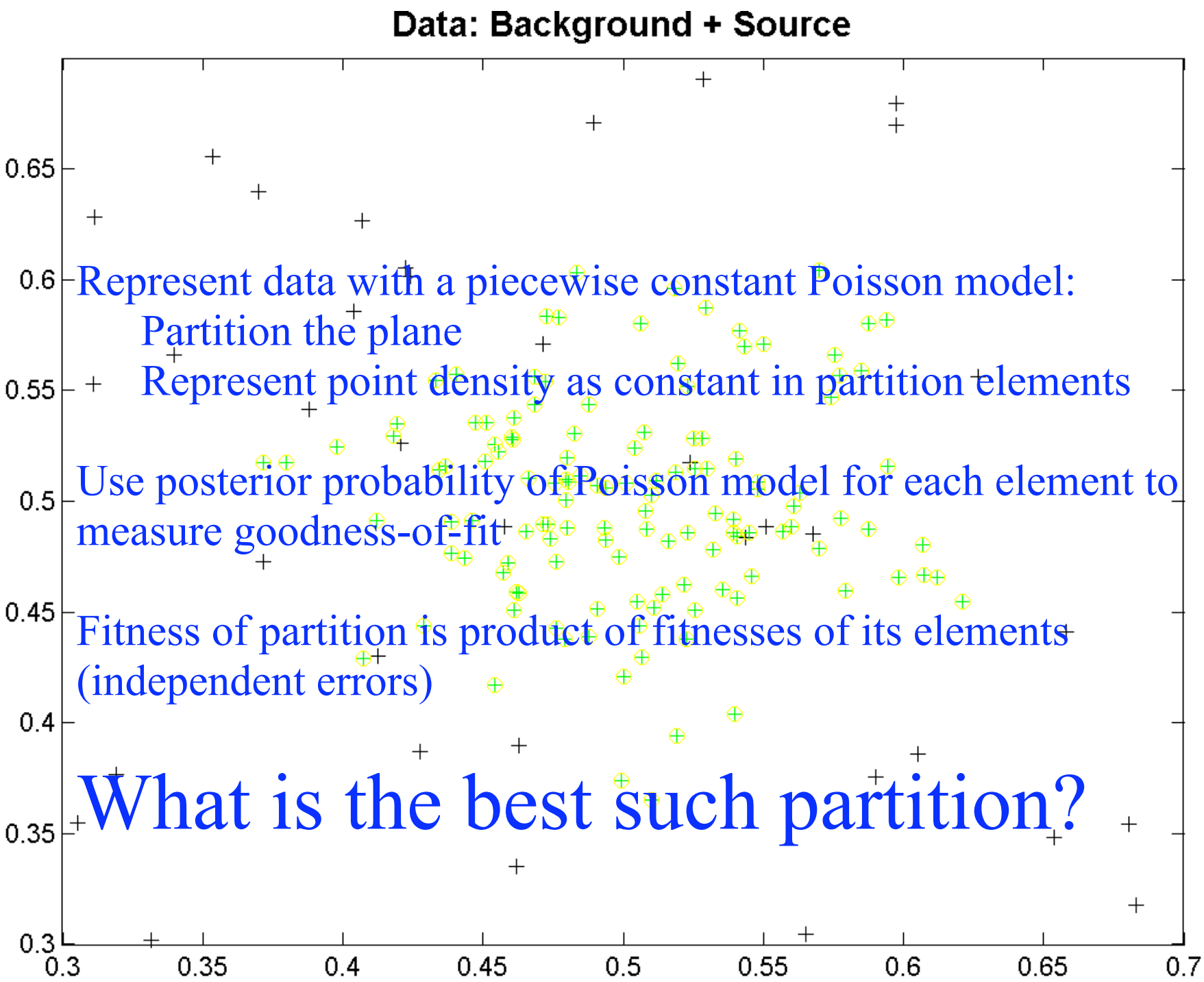
- Blocks are collections of Voronoi cells (1D,2D,...)
- Relax condition that blocks be connected
- Cell location now irrelevant
- Order cells by volume

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

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


Data: Background
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


Data: Background + Source
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



Represent data with a piecewise constant Poisson model:

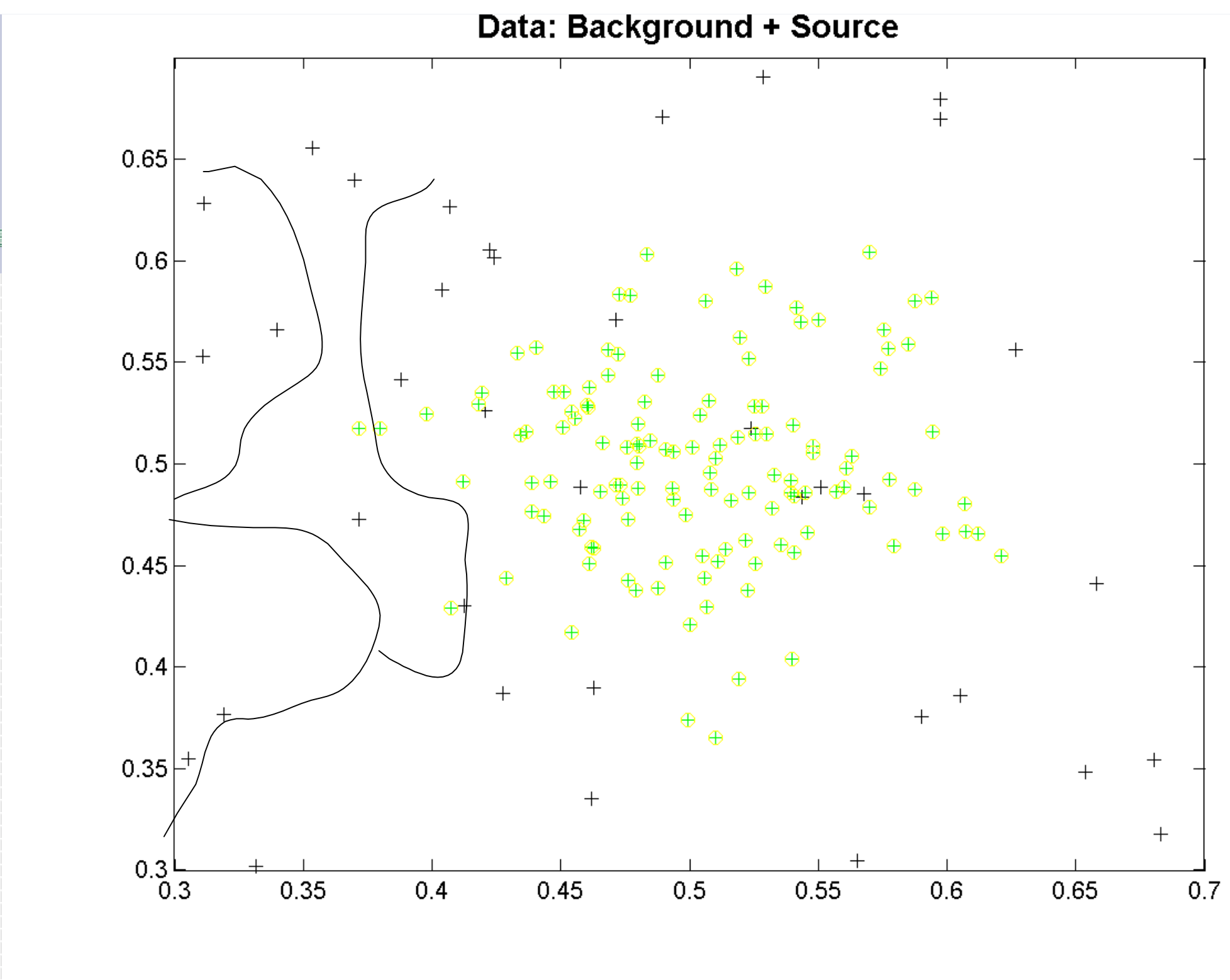
Partition the plane

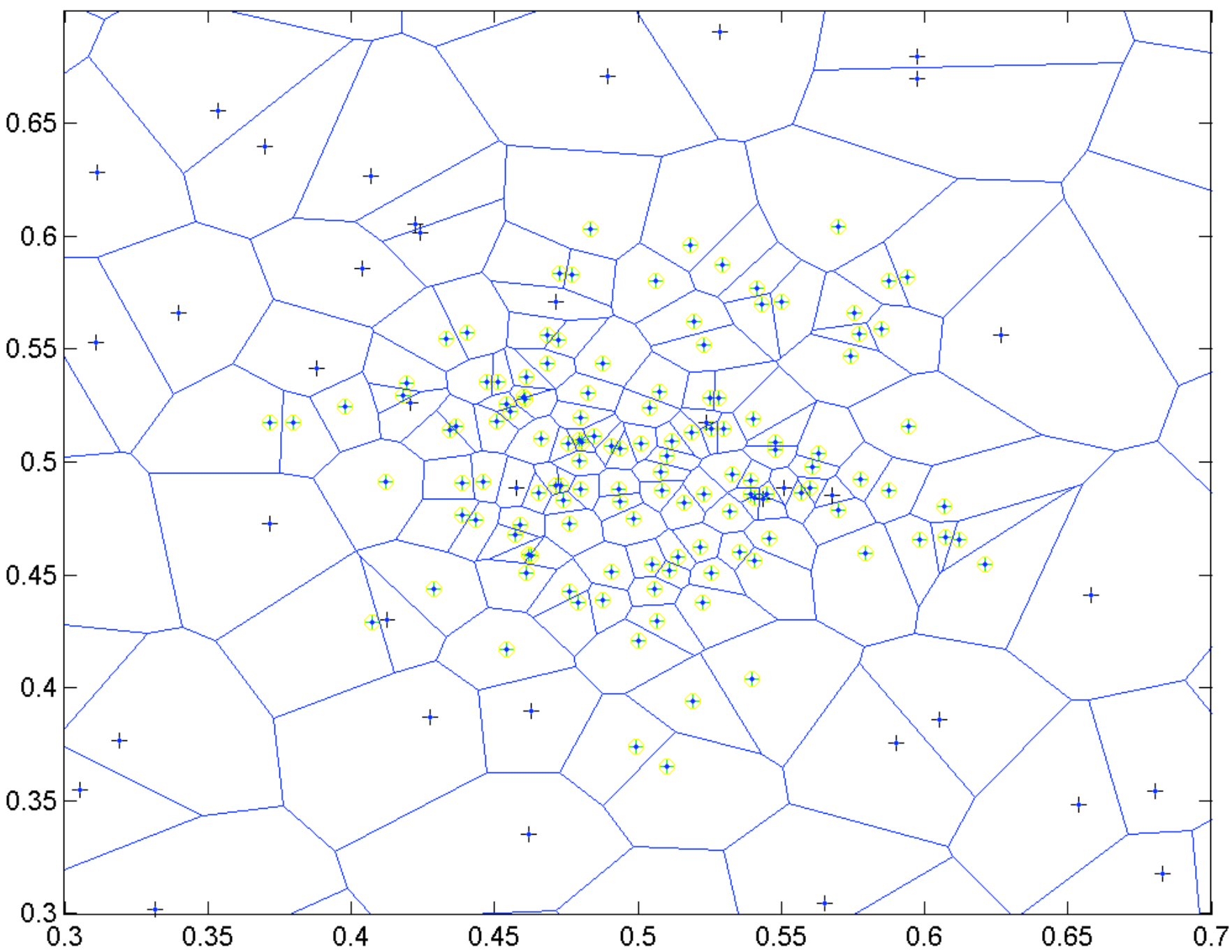
Represent point density as constant in partition elements

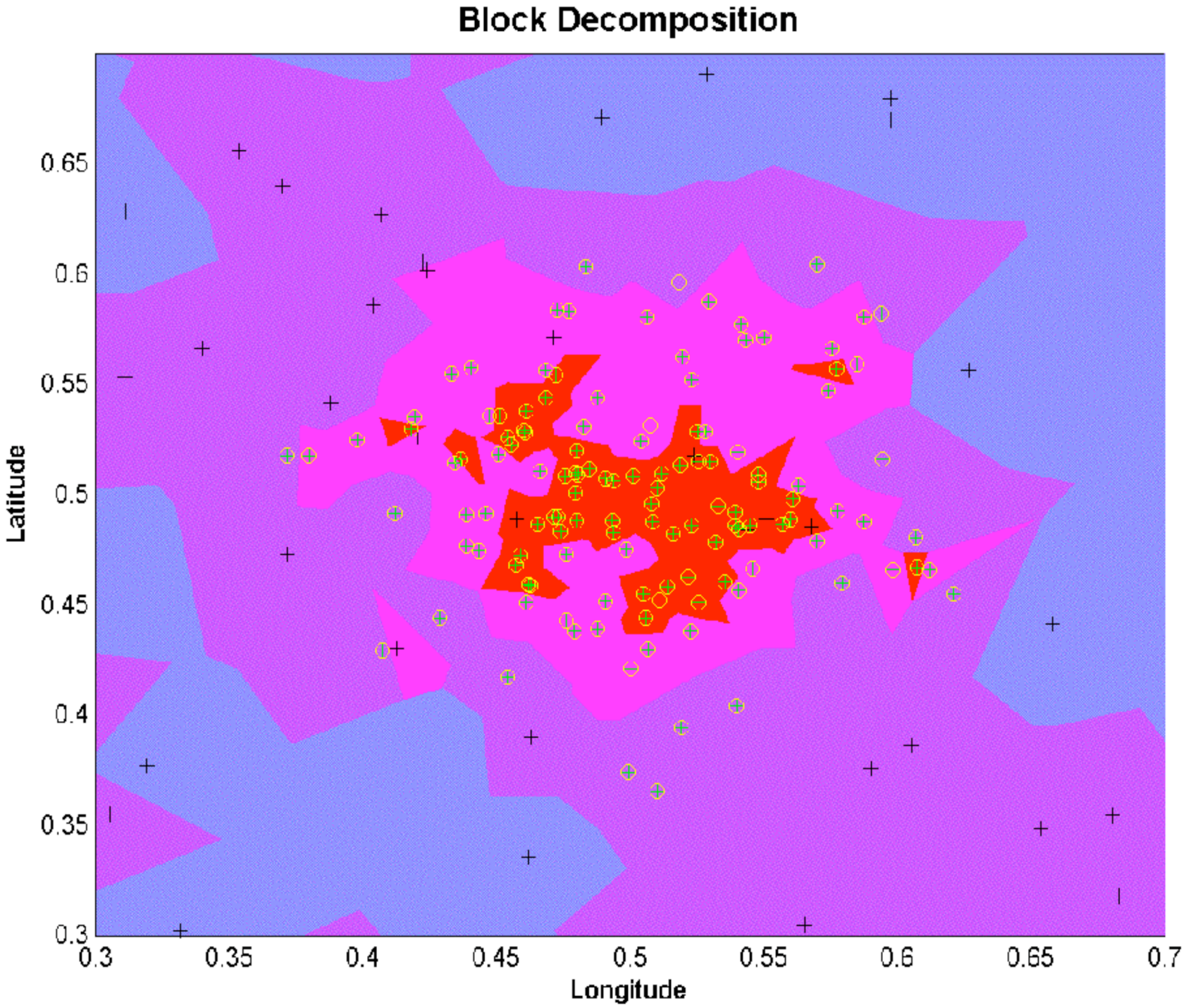
Use posterior probability of Poisson model for each element to measure goodness-of-fit

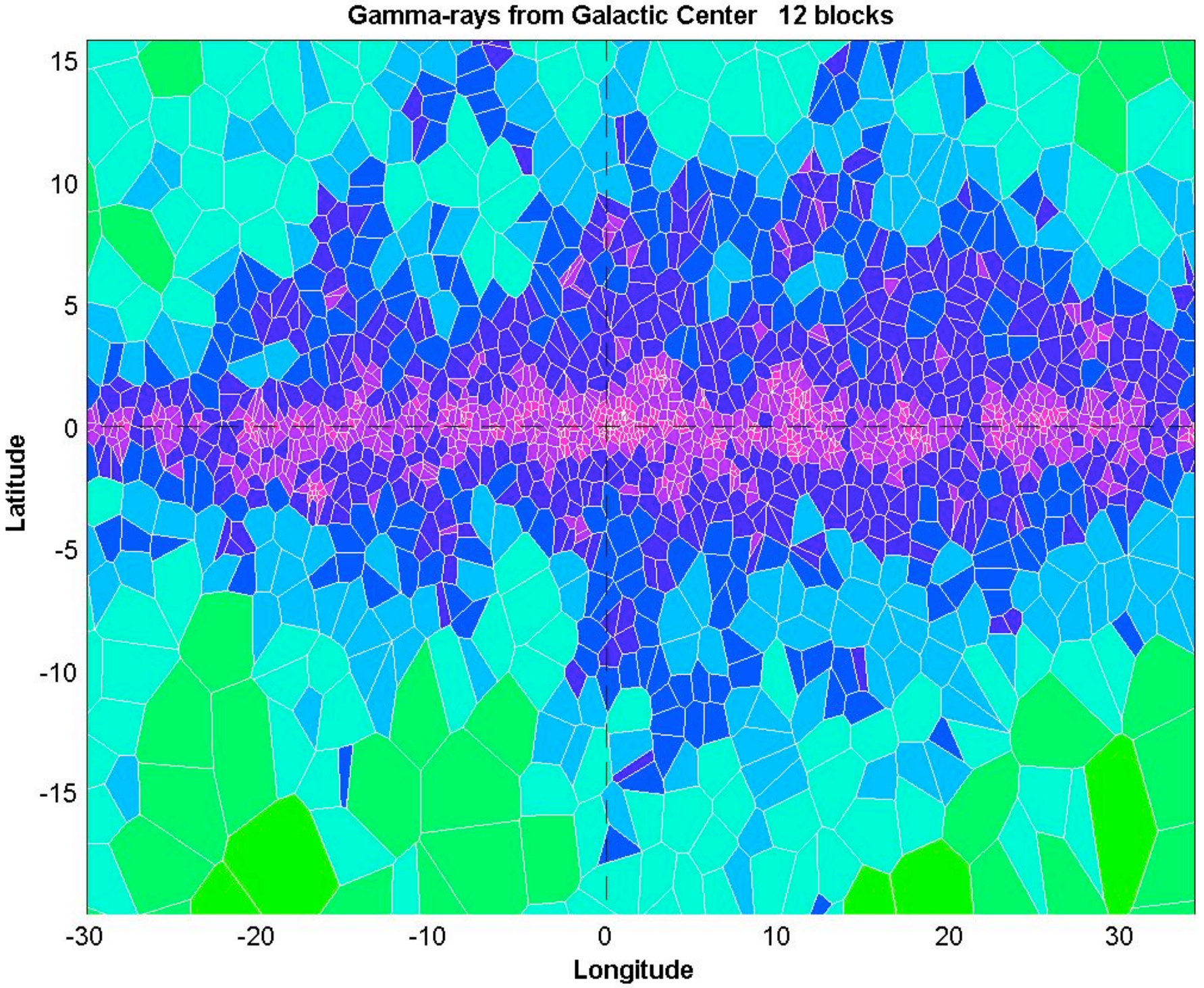
Fitness of partition is product of fitnesses of its elements (independent errors)

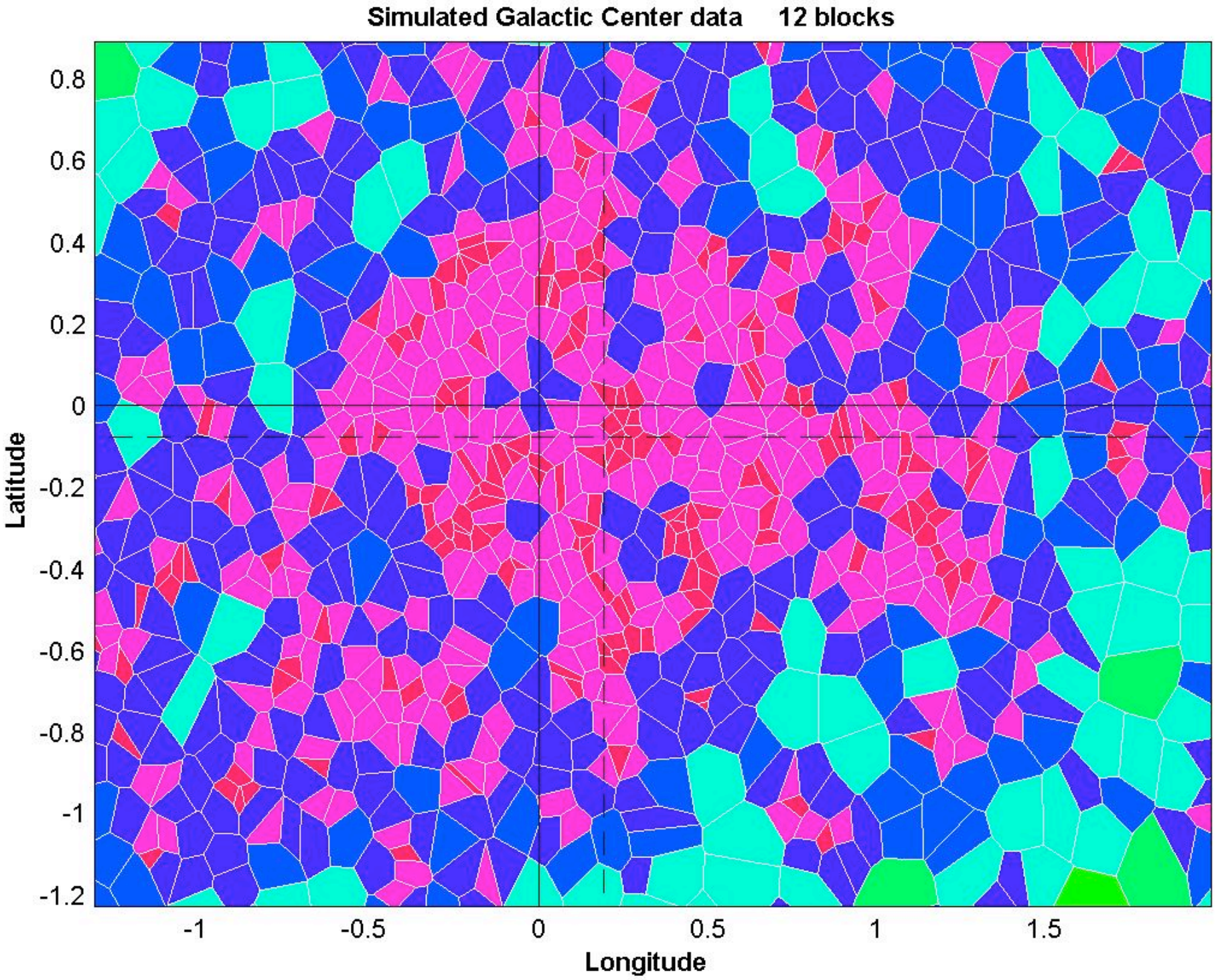
What is the best such partition?

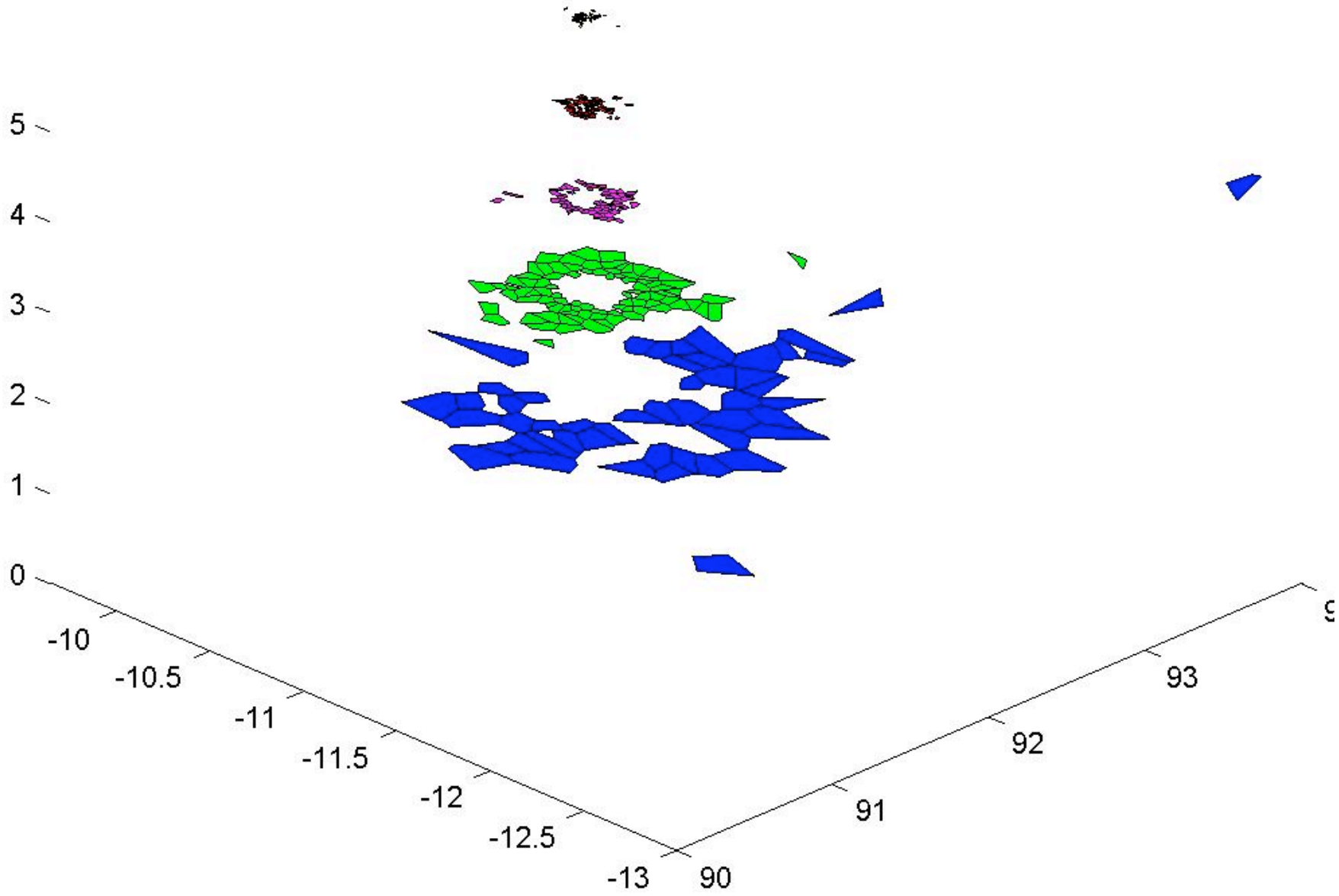

Data: Background + Source
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

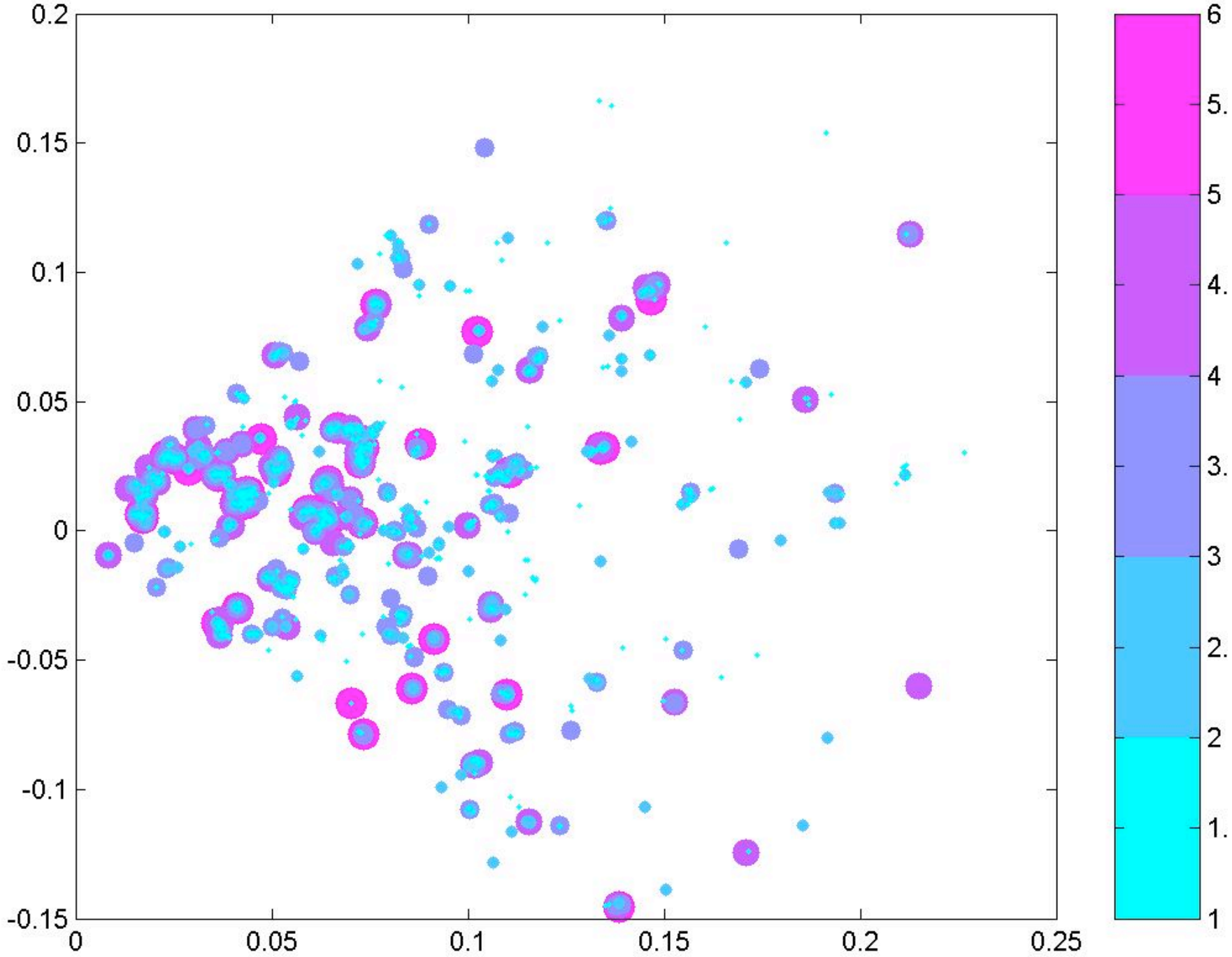

Data: Voronoi Tessellation
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

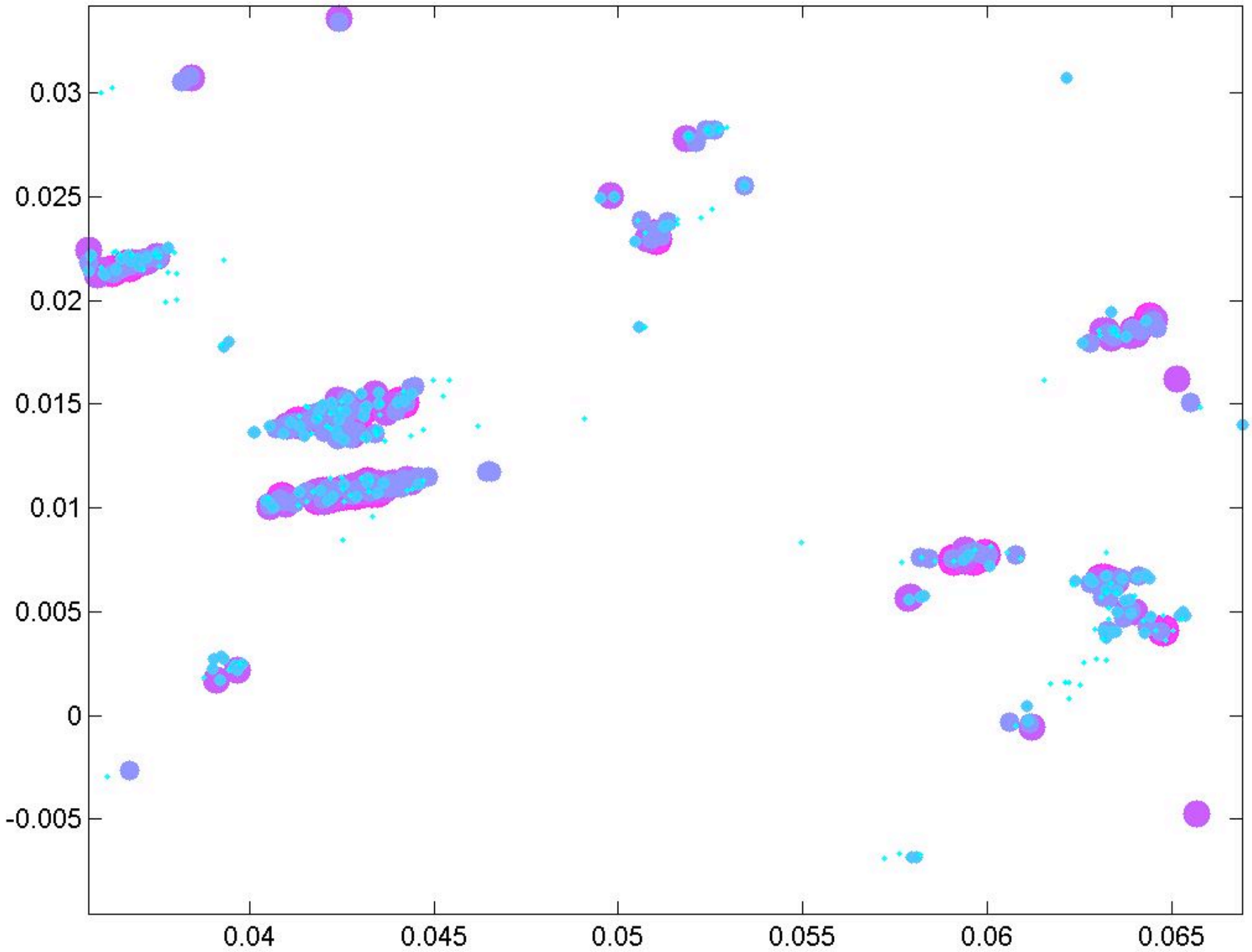

Block Decomposition
Latitude
Longitude
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7


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

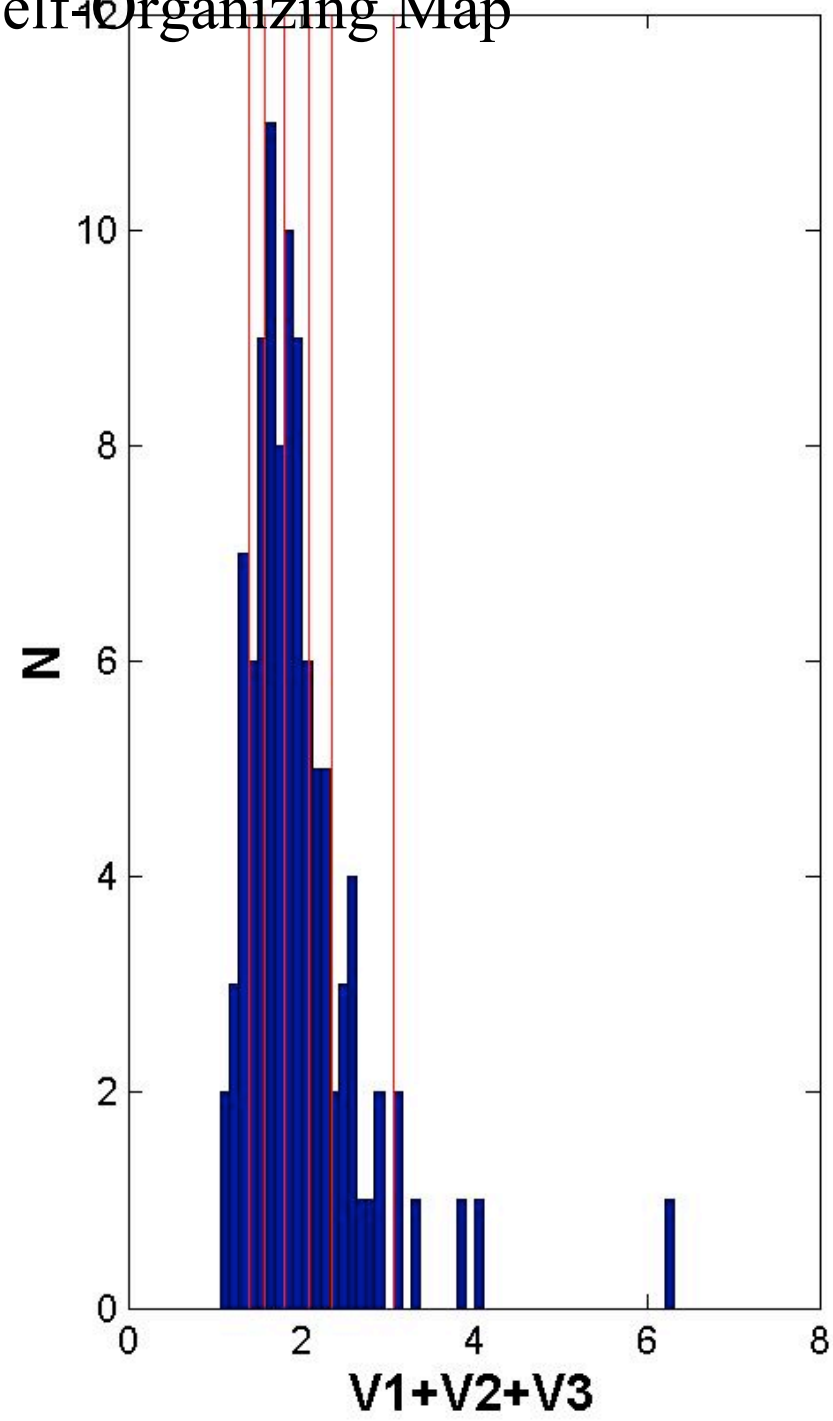

Simulated Galactic Center data 12 blocks
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
Latitude
-1
-0.5
0
0.5
1
1.5
Longitude

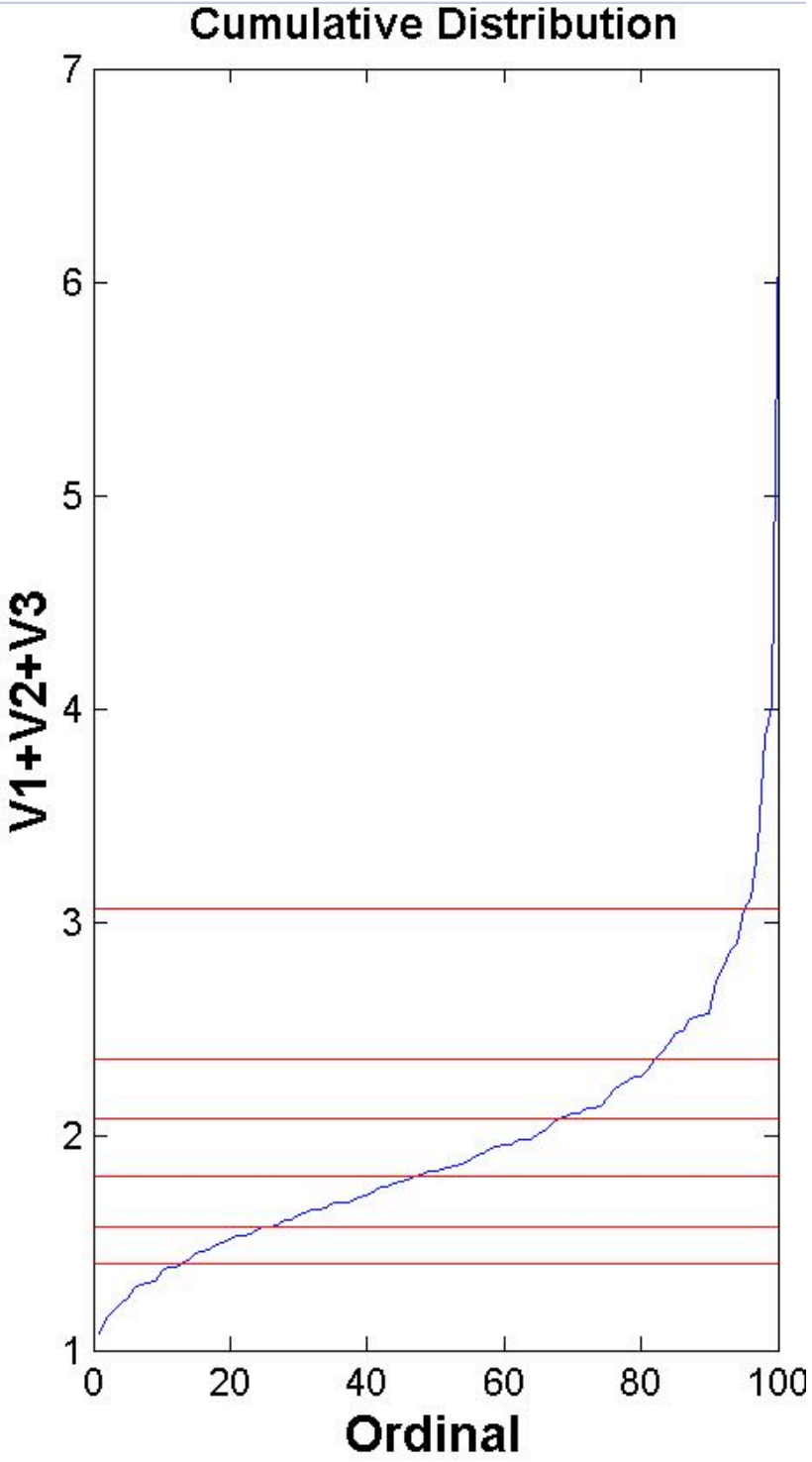

5
4
3
2
1
0
-10
-10.5
-11
-11.5
-12
-12.5
-13
90
91
92
93

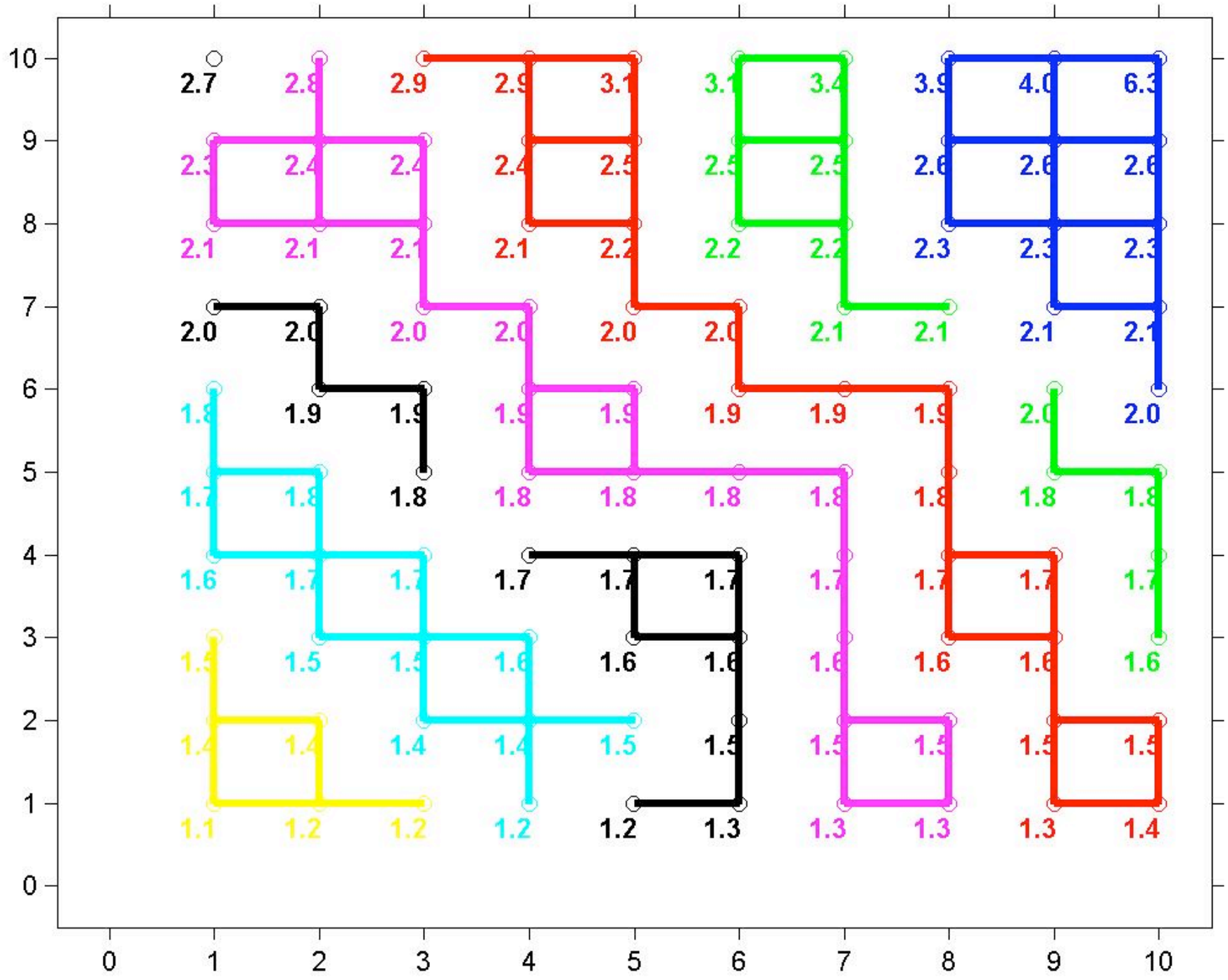

0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
0
0.05
0.1
0.15
0.2
0.25
6
5.
5
4.
4
3.
3
2.
2
1.
1

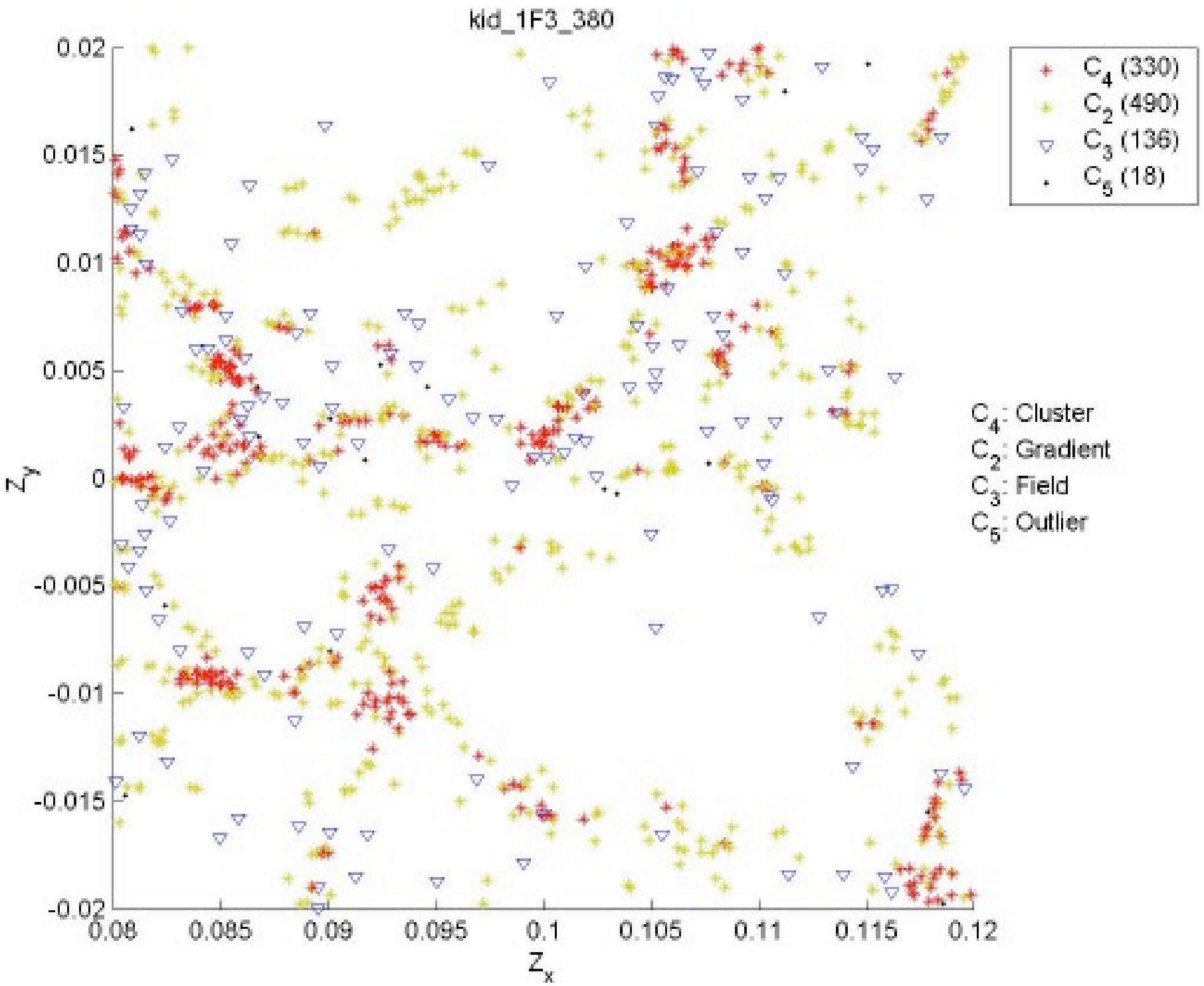

0.03
0.025
0.02
0.015
0.01
0.005
0
-0.005
0.04
0.045
0.05
0.055
0.06
0.065

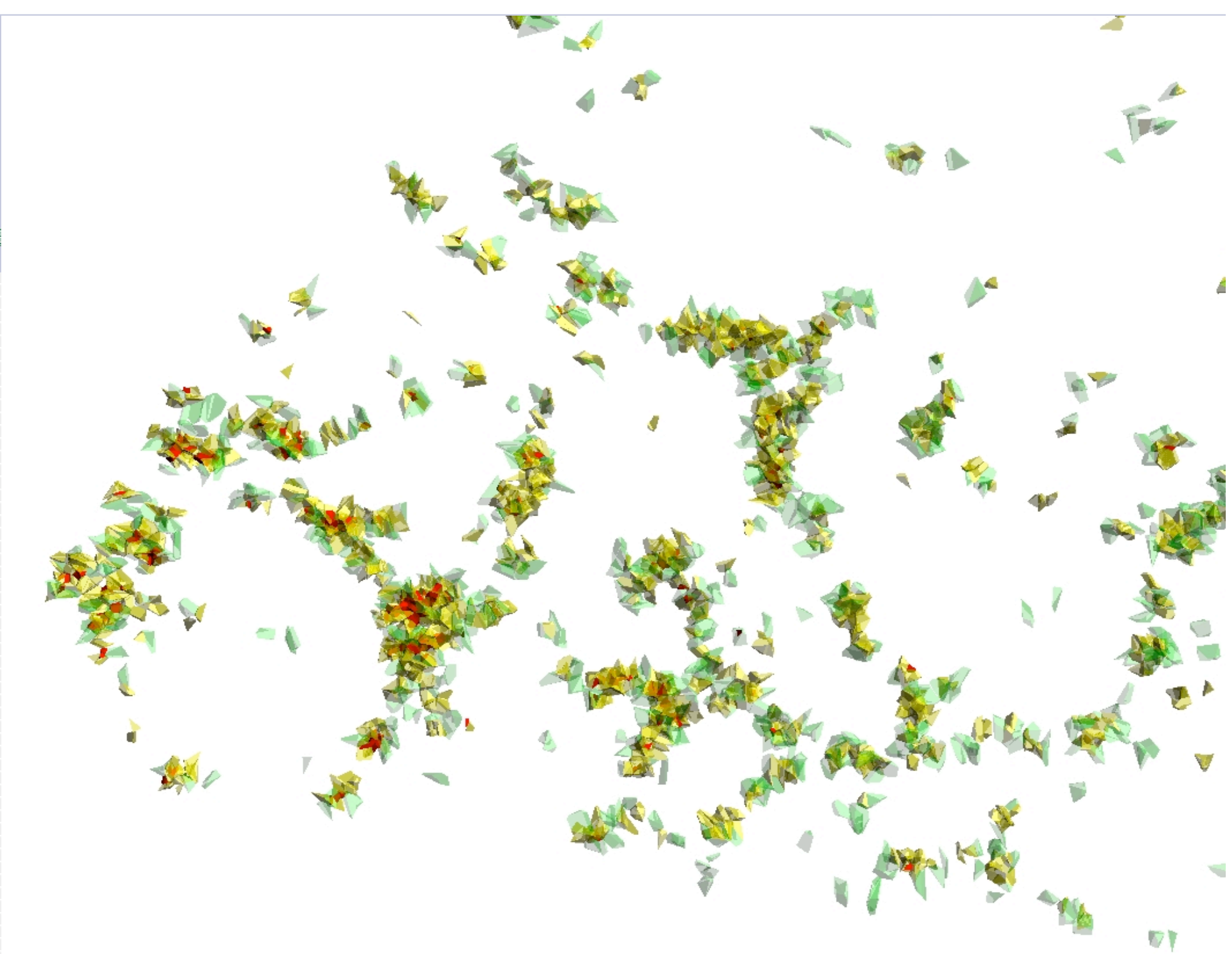
Self-Organizing Map

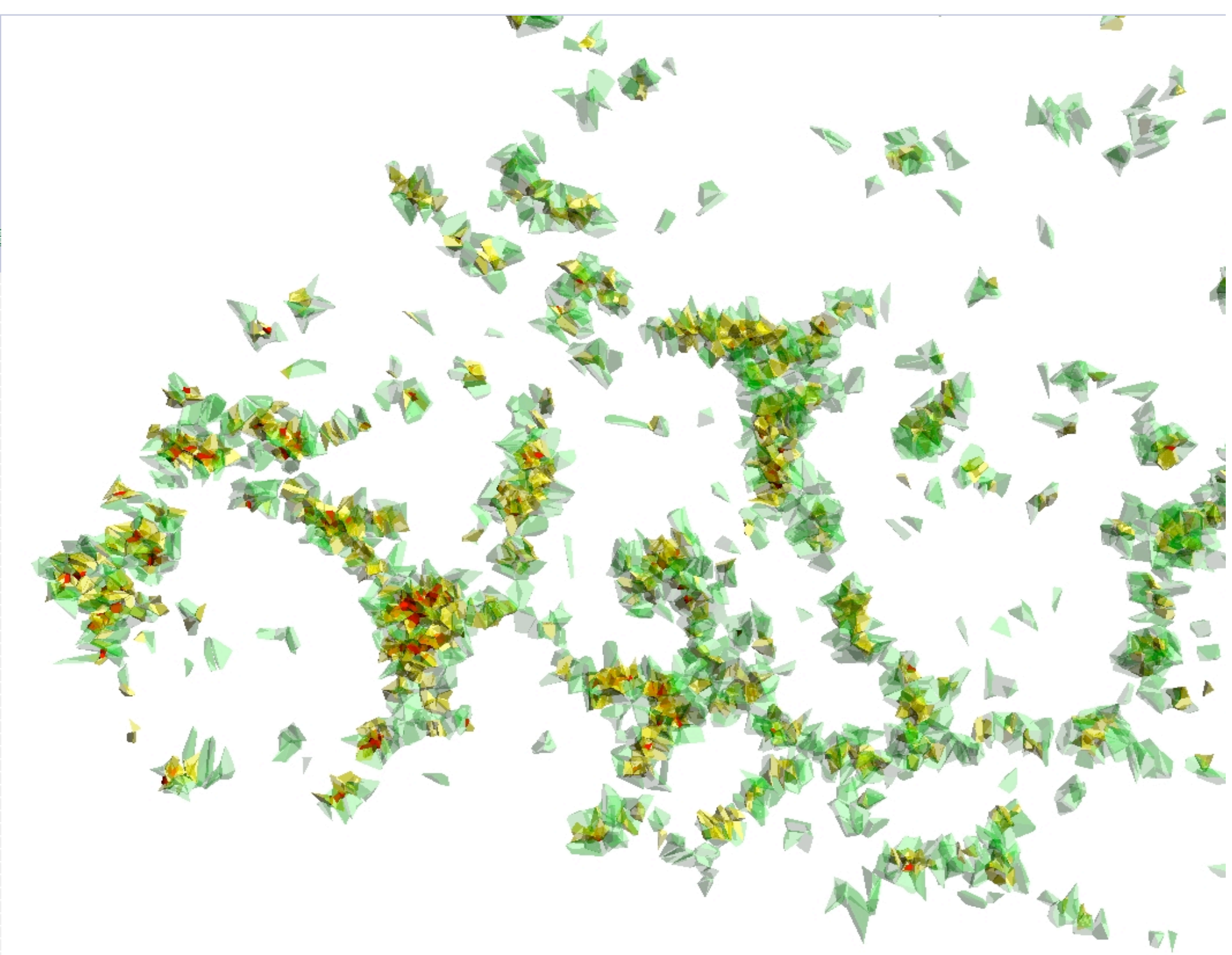


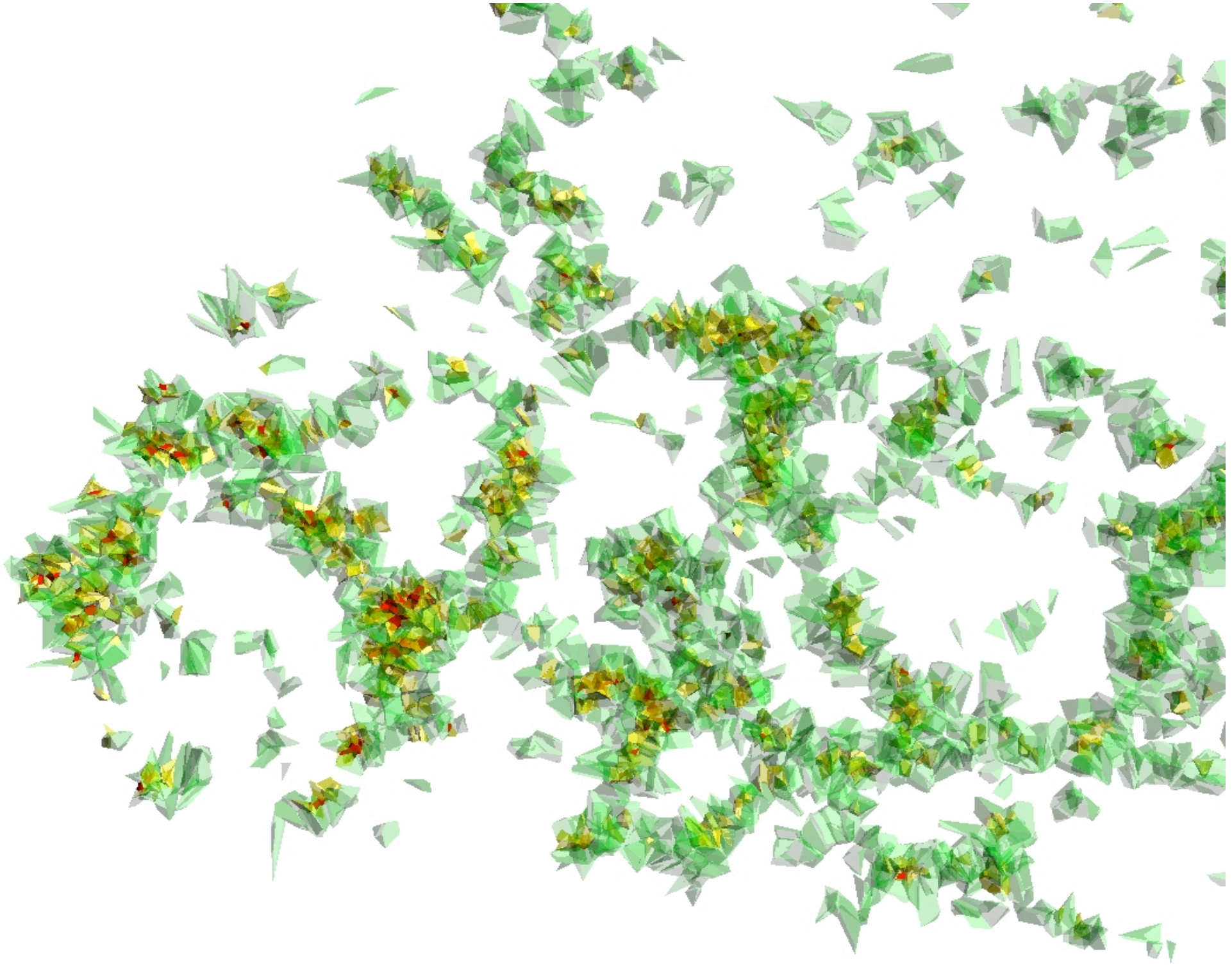


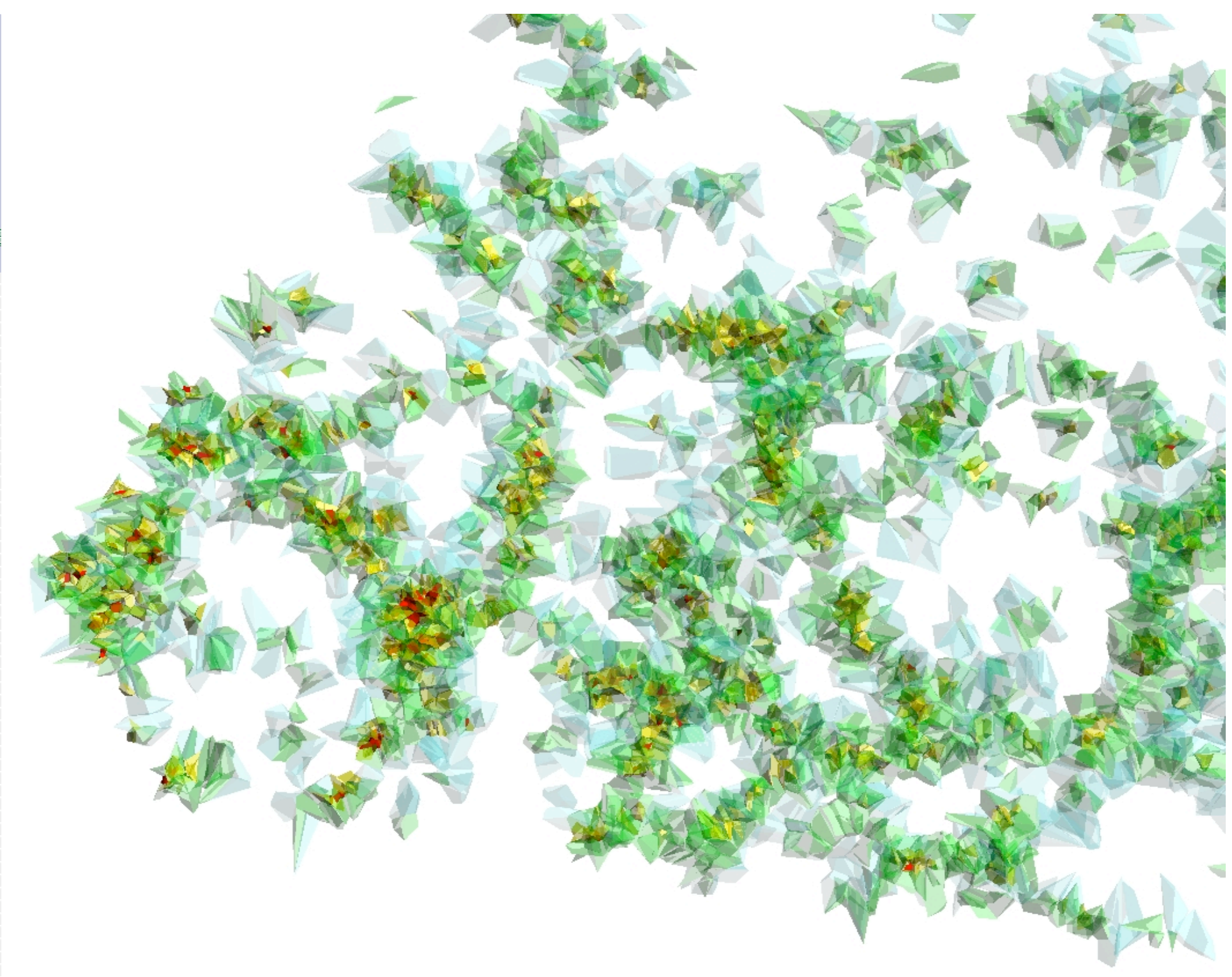

10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
10

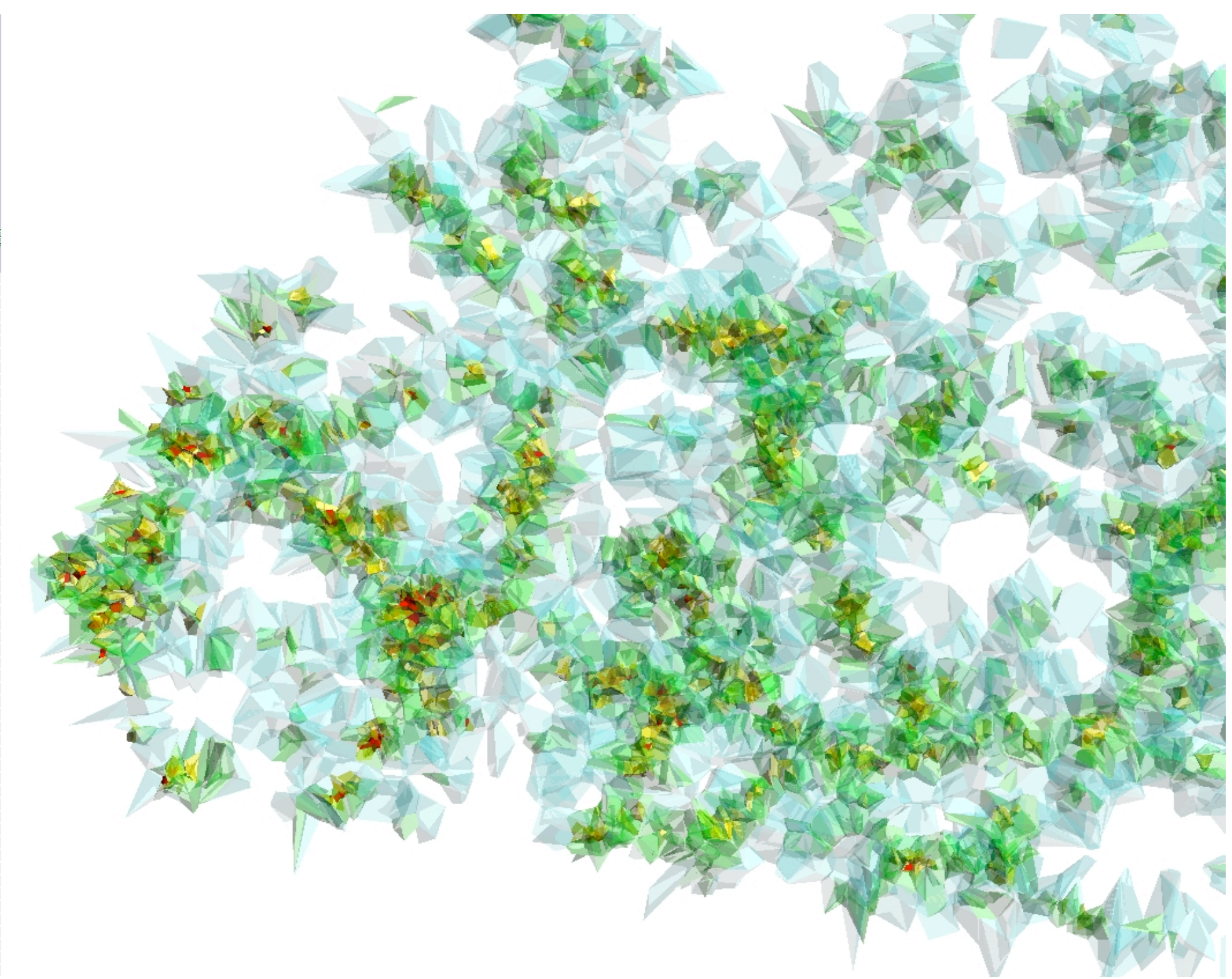

kid_1F3_380
0.02
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.02
Z_y
0.08
0.085
0.09
0.095
0.1
0.105
0.11
0.115
0.12
Z_x
C_4 (330)
C_2 (490)
C_3 (136)
C_5 (18)
C_4: Cluster
C_2: Gradient
C_3: Field
C_5: Outlier

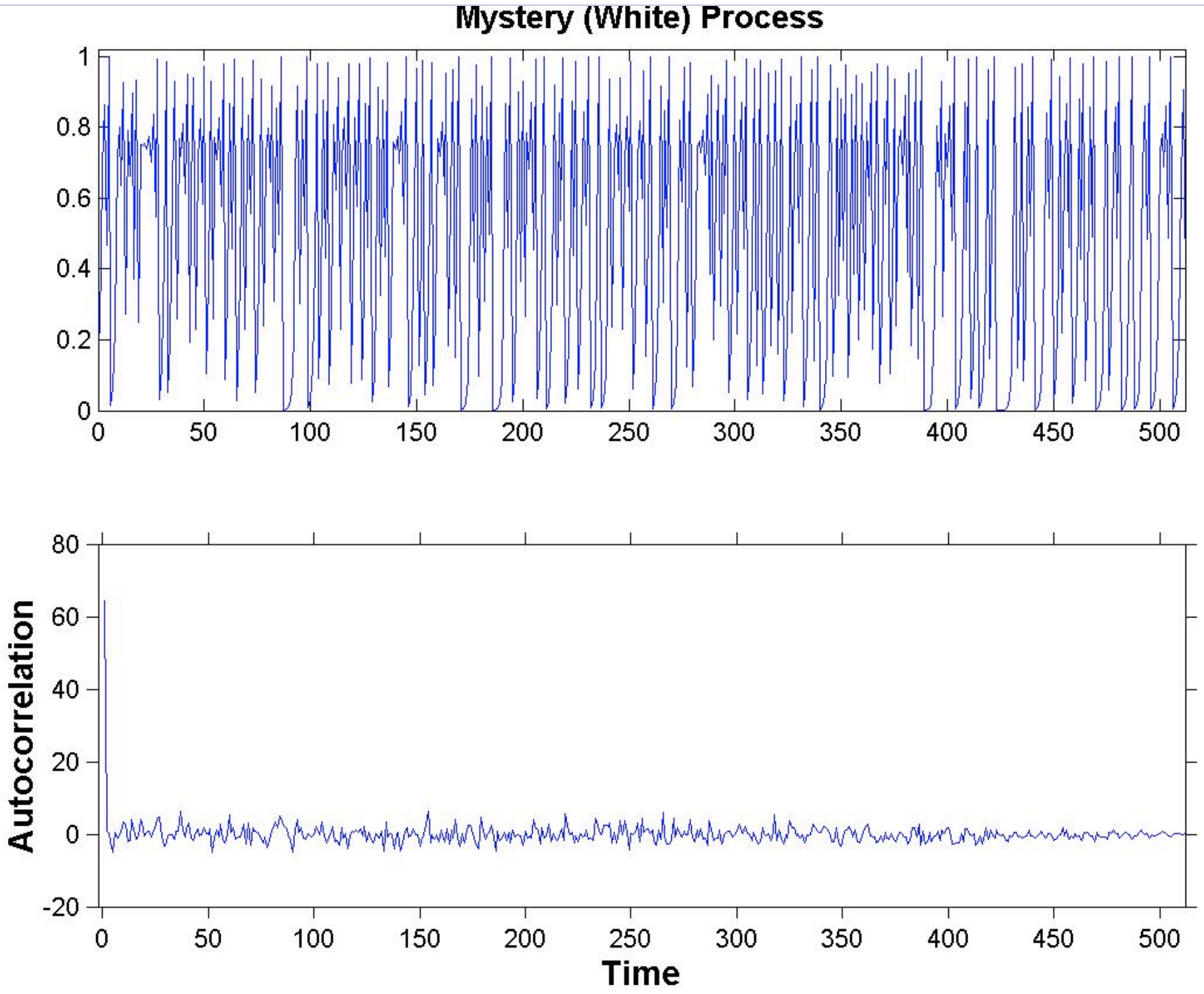


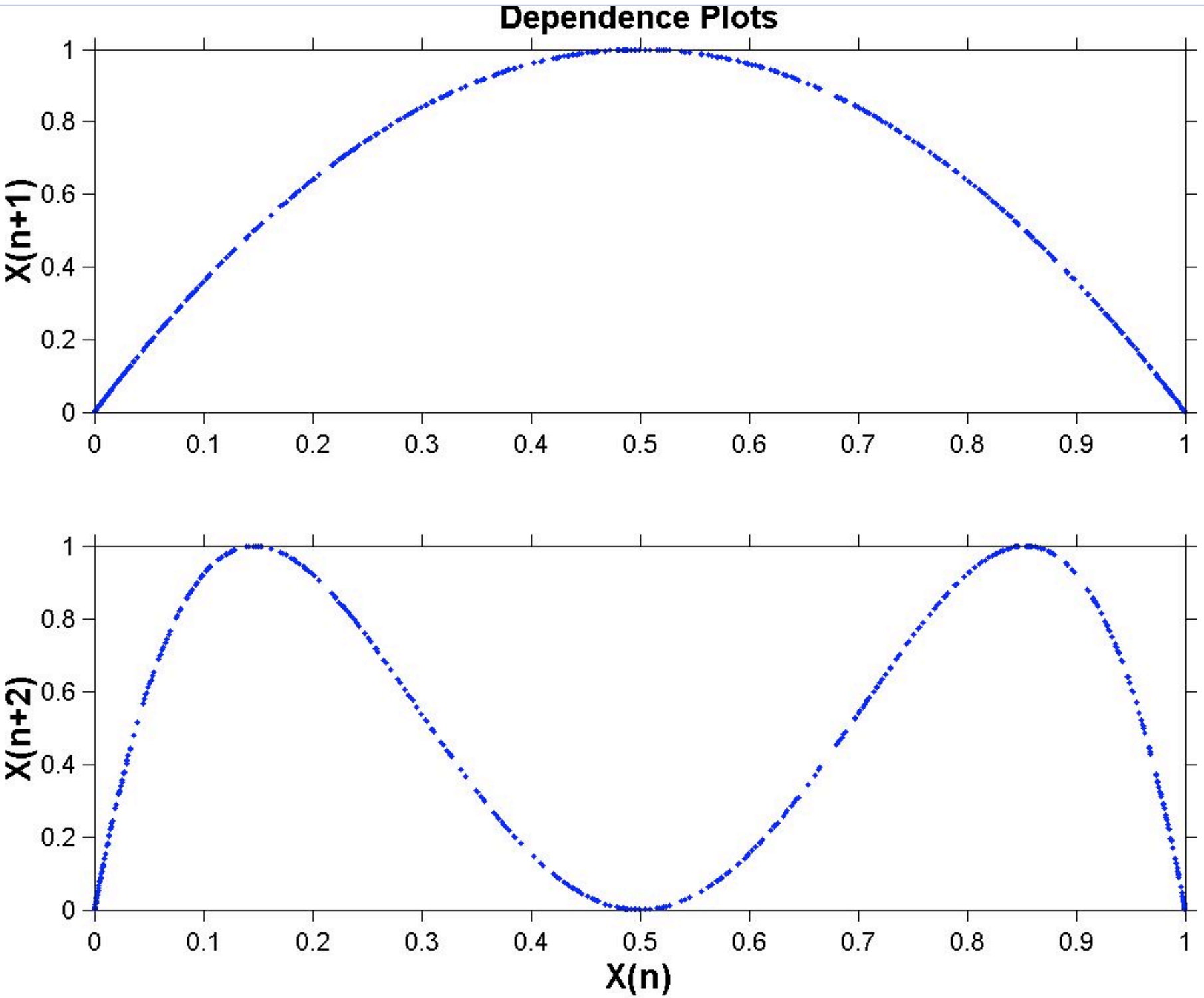


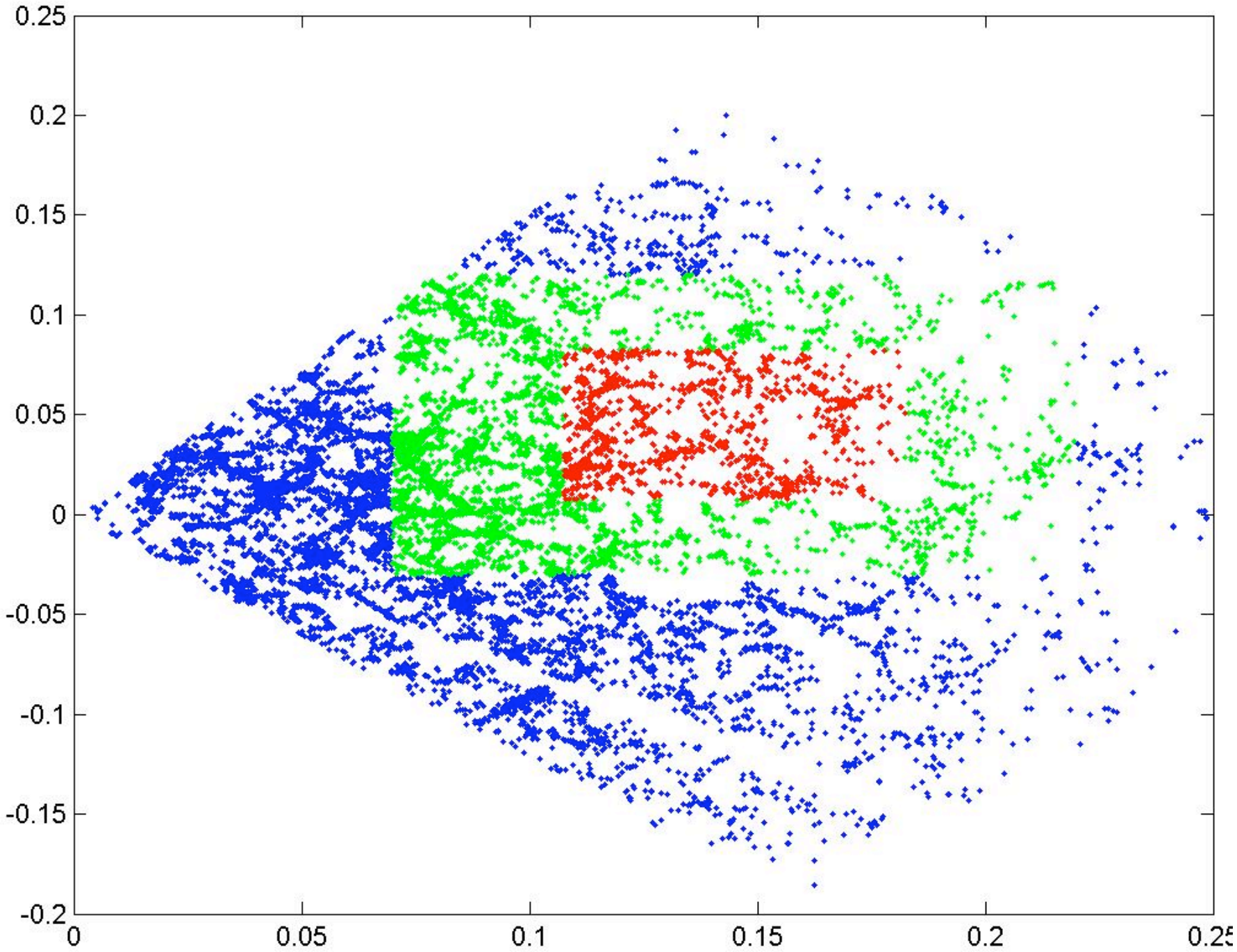


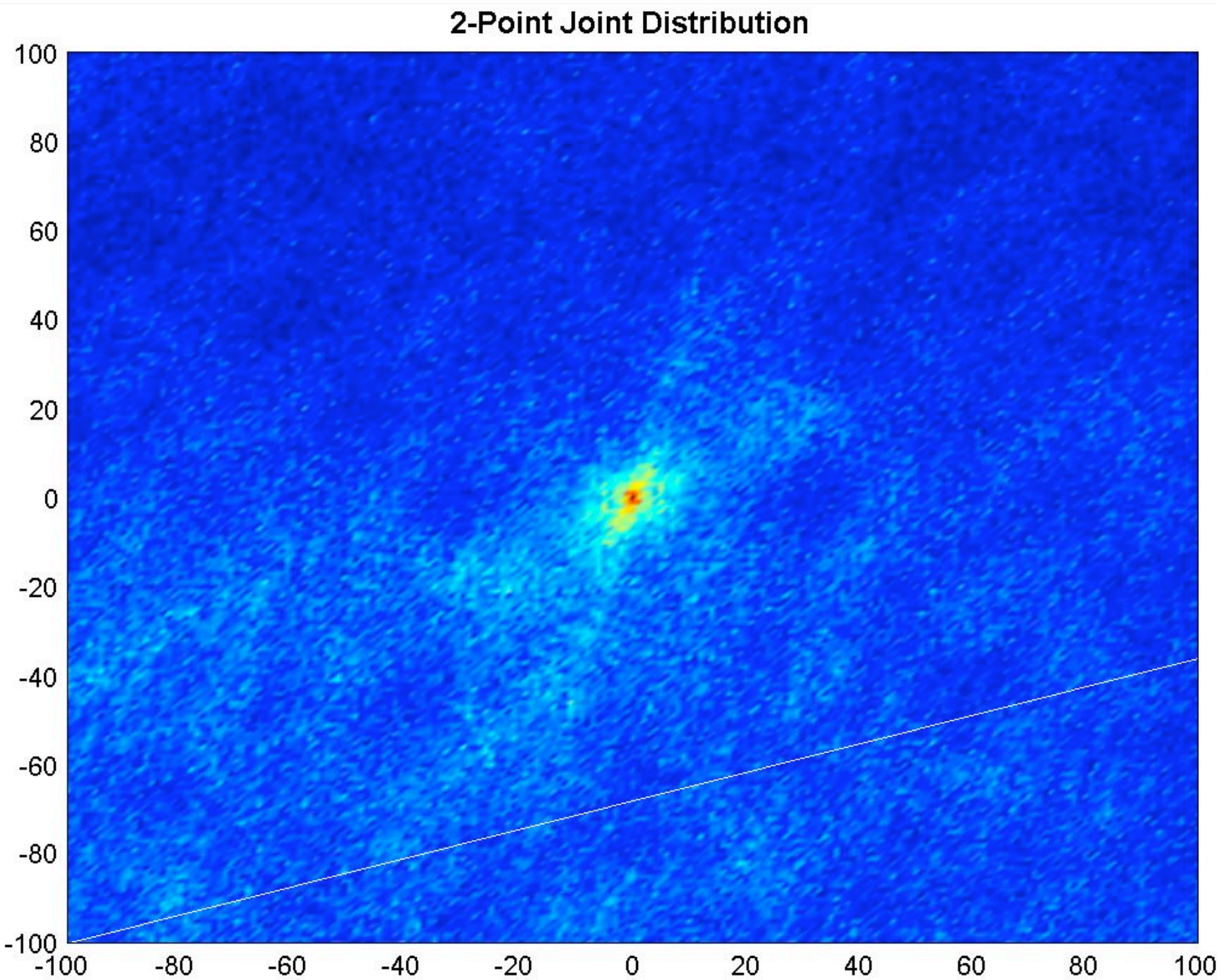


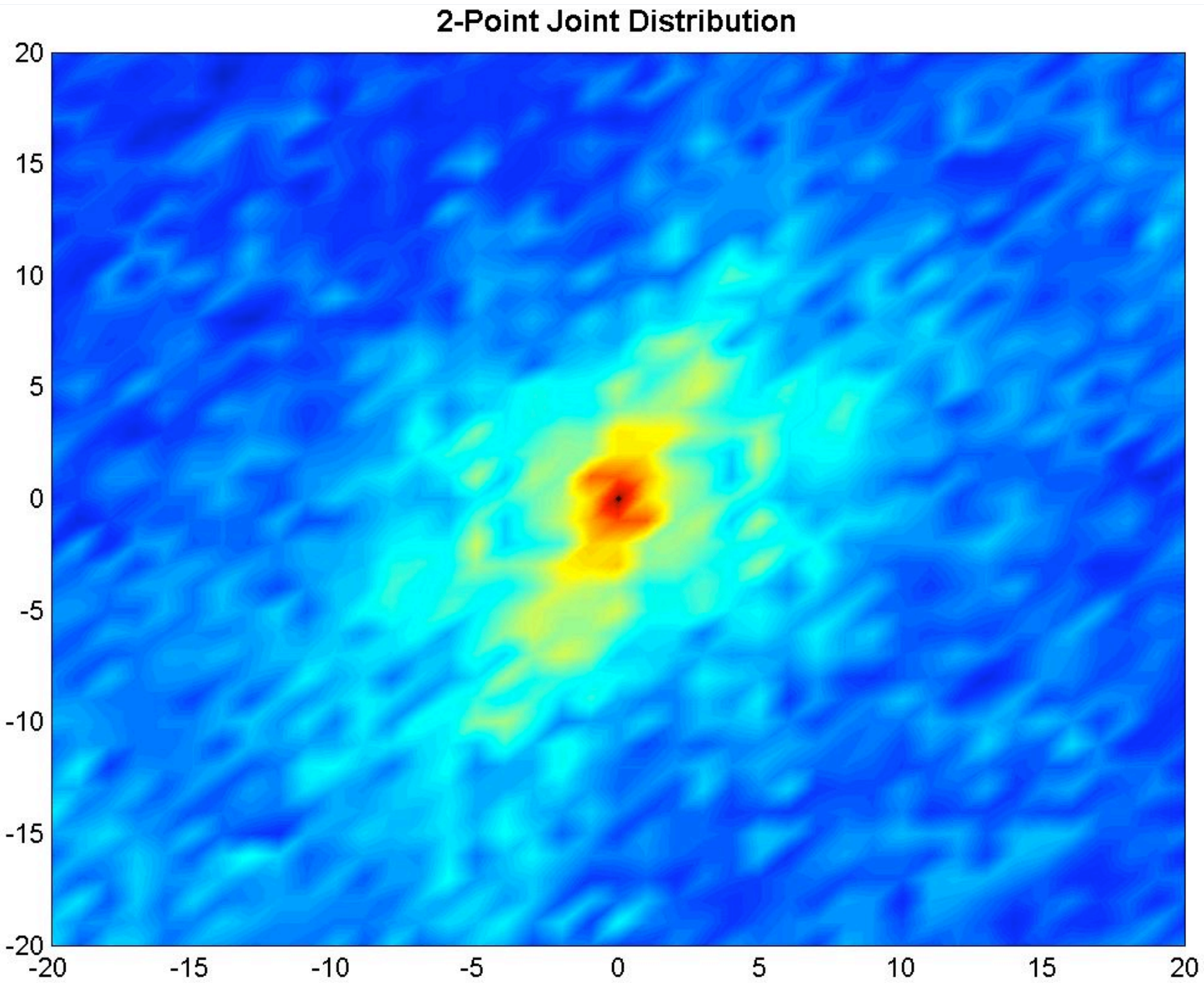



Mystery (White) Process
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350
400
450
500
Autocorrelation
80
60
40
20
0
-20
0
50
100
150
200
250
300
350
400
450
500
Time


Dependence Plots
X(n+1)
X(n+2)
X(n)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0
0.05
0.1
0.15
0.2
0.25


2-Point Joint Distribution
100
80
60
40
20
0
-20
-40
-60
-80
-100
-100
-80
-60
-40
-20
0
20
40
60
80
100


2-Point Joint Distribution
20
15
10
5
0
-5
-10
-15
-20
-20
-15
-10
-5
0
5
10
15
20