Applications of Visual Data Mining

Jürgen Symanzik

Utah State University, Logan, UT, USA

***e-mail: symanzik@math.usu.edu**

WWW: http://www.math.usu.edu/~symanzik

Examples

1) Archaeological Data
2) Human Motion Data
3) Neuroanatomical Data
4) Remote Sensing Data

Example 1: Archaeological Data

Published as:

Wilhelm, A. F. X., Wegman, E. J., Symanzik, J. (1999): Visual Clustering and Classification: The Oronsay Particle Size Data Set Revisited, Computational Statistics: Special Issue on Interactive Graphical Data Analysis, 14(1):109-146.

Oronsay Sand Particles

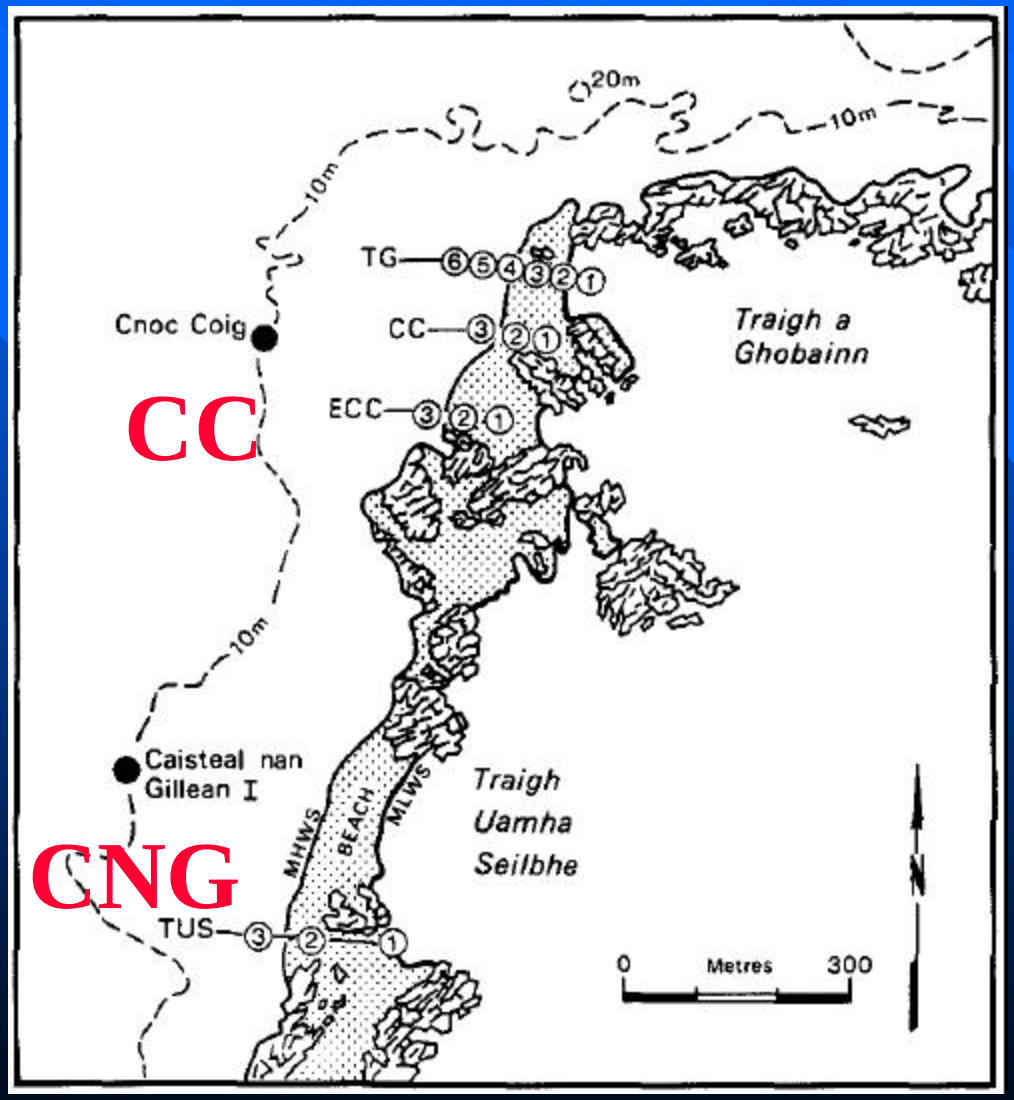
"The mesolithic shell middens on the island of Oronsay are one of the most important archeological sites in Britain. It is of considerable interest to determine their position with respect to the mesolithic coastline. If the sand below the midden were beach sand and the sand from the upper layers dune sand, this would indicate a seaward shift of the beach-dune interface."

Flenley and Olbricht, 1993

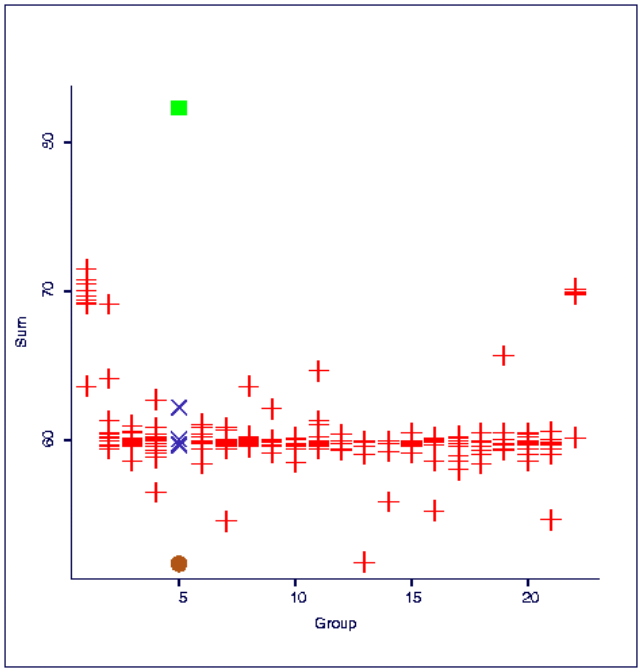
Objective of Study

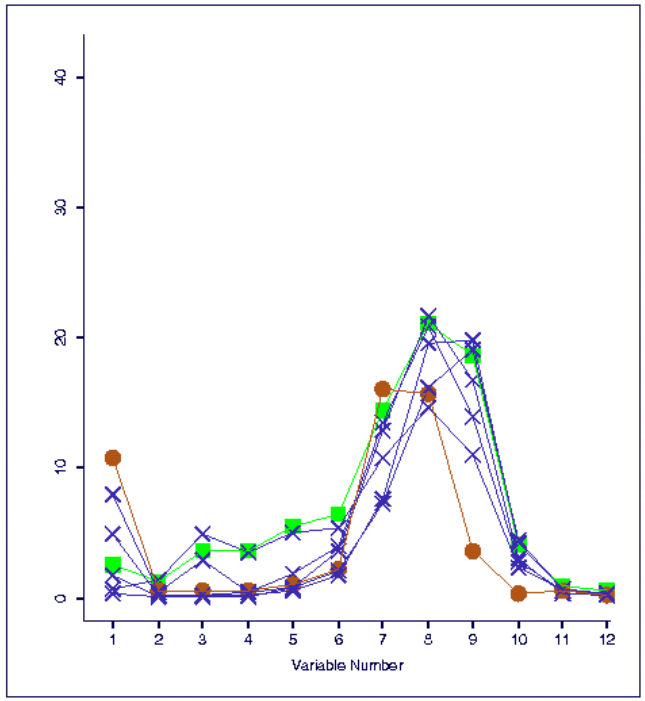
- Cluster samples of modern sand into “beach-like” or “dune-like” sand.
- Classify archaeological sand samples from middens as to whether they are beach sand or dune sand.

Oronsay - Geography



Oronsay - Data Problems





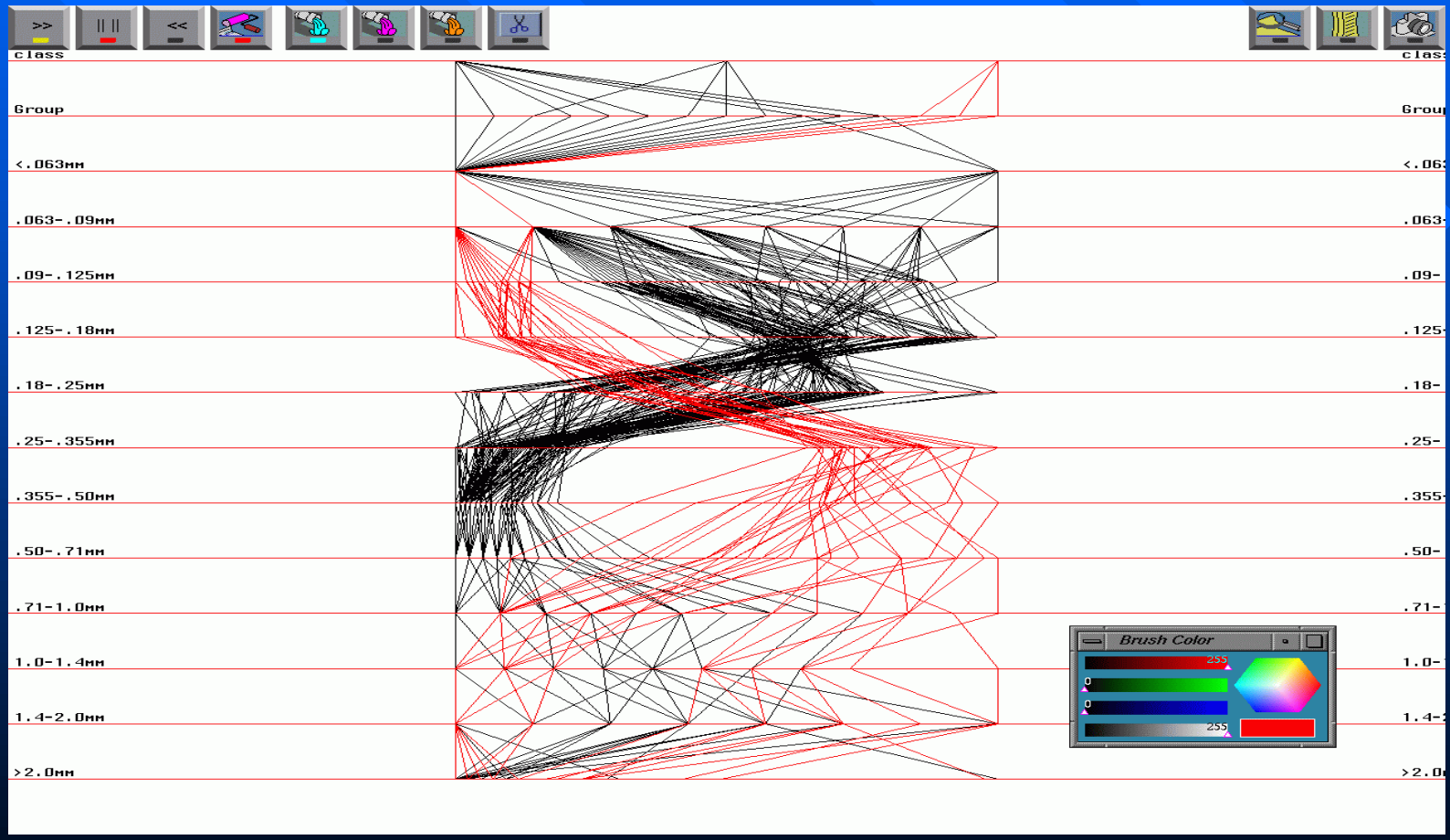
Oronsay - Parametric Analysis

- Historical strategy is to fit parametric distributions and compare modern and archeological sands based on parameters.
- Weibull, 1933; lognormal (breakage models), log-hyperbolic, log-skew-Laplace, 1937, Barndorff-Nielsen, 1977.
- Models 2 to 4 parameters, theory developed, practice problematic.

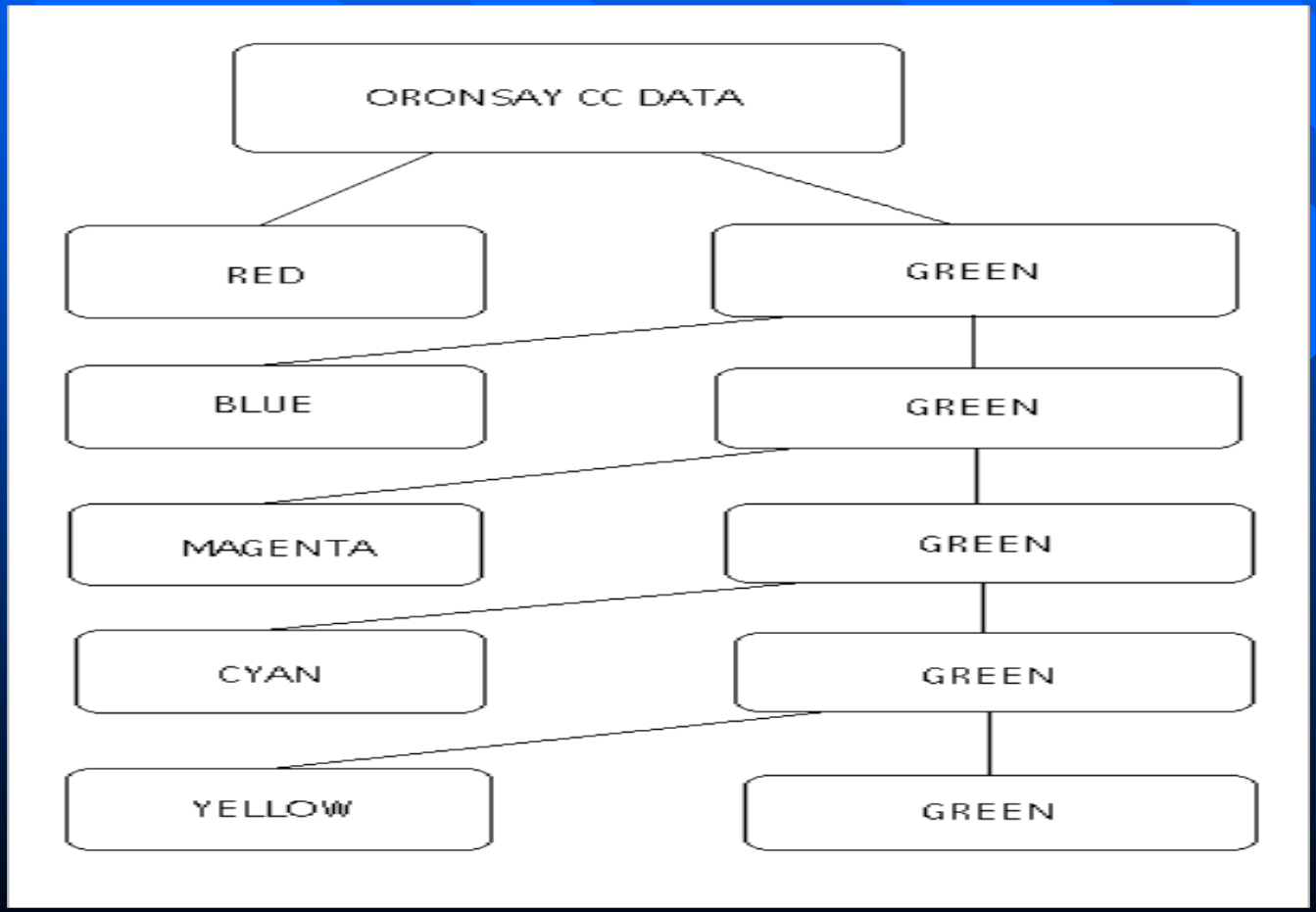
Oronsay - Visual Approach

- Multidimensional parallel coordinate display combined with grand tour.
- BRUSH-TOUR strategy
  - Clusters recognized by gaps in any horizontal axis.
  - Brush existing clusters with colors.
  - Execute grand tour until new clusters appear, brush again.
  - Continue until clusters are exhausted.

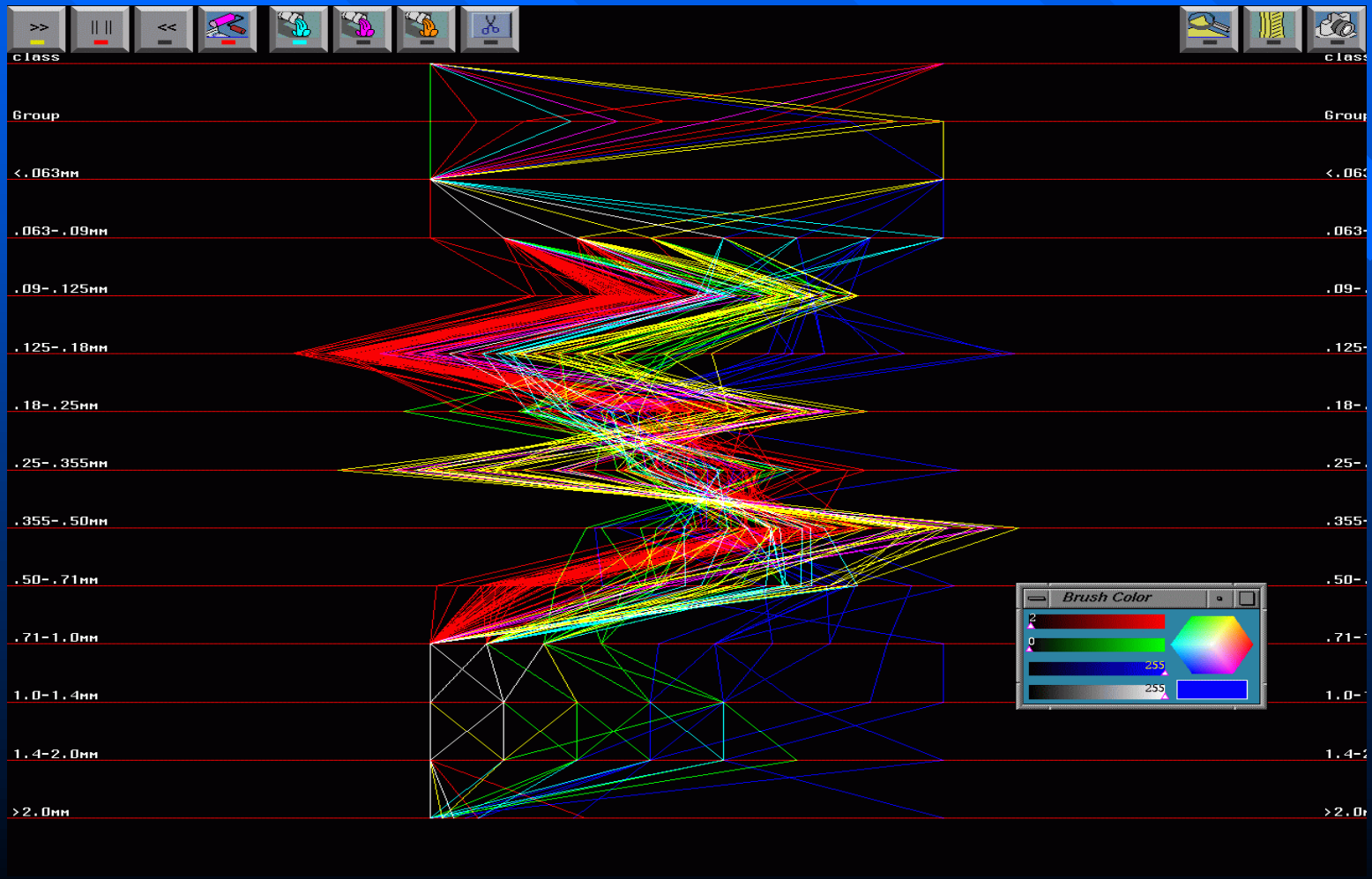
Beach & Dune Sand



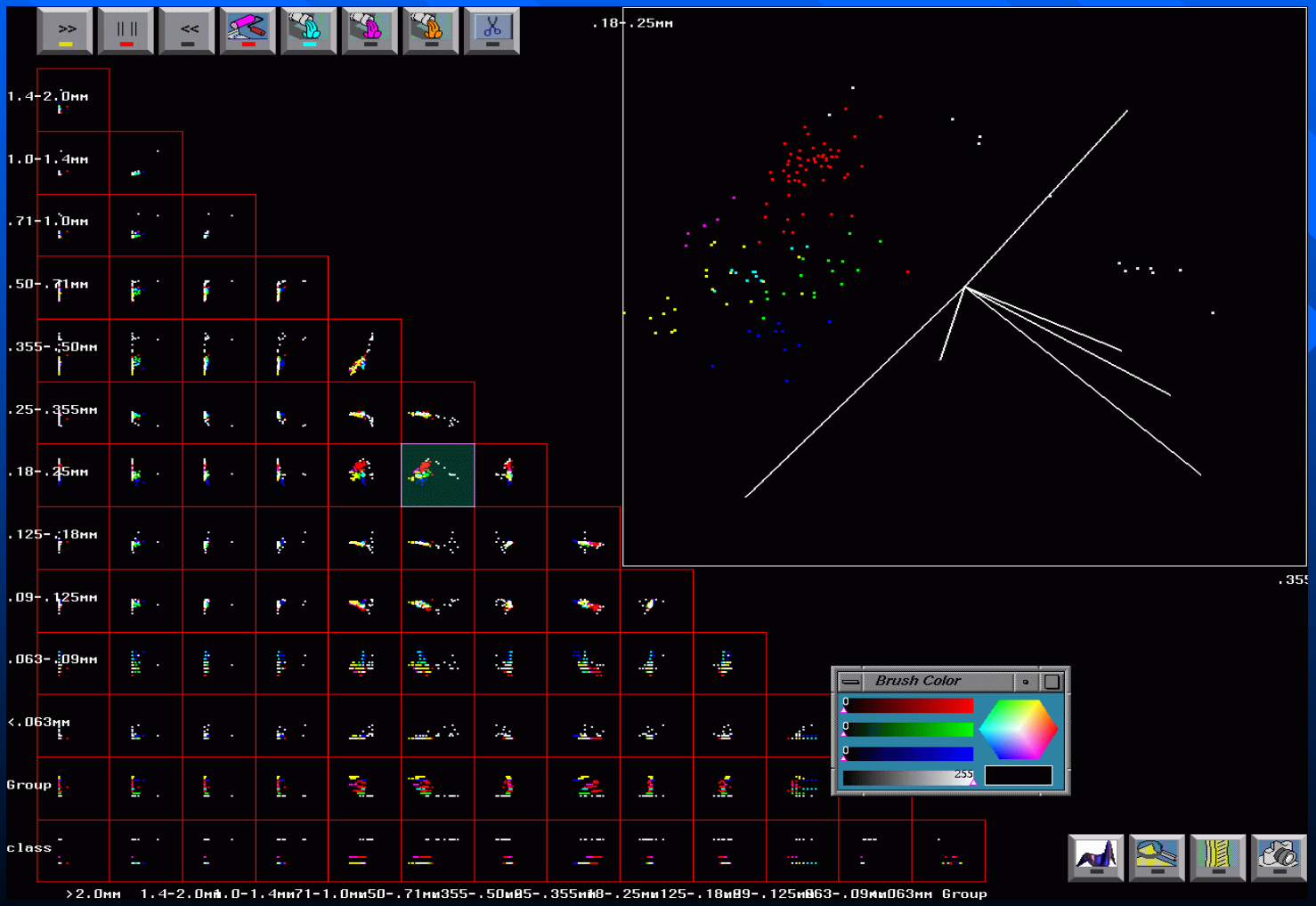
Separation of Clusters



Final Clustering



Scatterplots & Projection



Oronsay - Conclusions (1)

- Sands from the CC site and the CNG site have considerably different particle size distributions and cannot be effectively aggregated.
- Data at small and at large particle dimensions is too quantized to be used effectively.
- The visual based BRUSH-TOUR strategy is extremely effective at clustering.

Oronsay Conclusions (2)

- Midden sands are neither modern beach sands nor modern dune sands.
- Midden sands are more similar to modern dune sands.
- This result does not support the seaward-shift-of-the-beach-dune-interface hypothesis, but suggests the middens were always in the dunes.

Example 2: Human Motion Data

Published as:

Vandersluis, J. P., Cooke, J. D., Ascoli, G. A., Krichmar, J. L., Michaels, G. S., Montgomery, M., Symanzik, J., Vitucci, B. (1998): Exploratory Statistical Graphics for an Initial Motion Control Experiment, Computing Science and Statistics, 30:482-487.

Purpose of Experiments

- Rehabilitation of people after accidents
- Knowledge of adaptation of humans to perform mechanical tasks, e.g., arm movement
- Perfection of movements
  - Dancers
  - Ski jumpers
  - Piano players

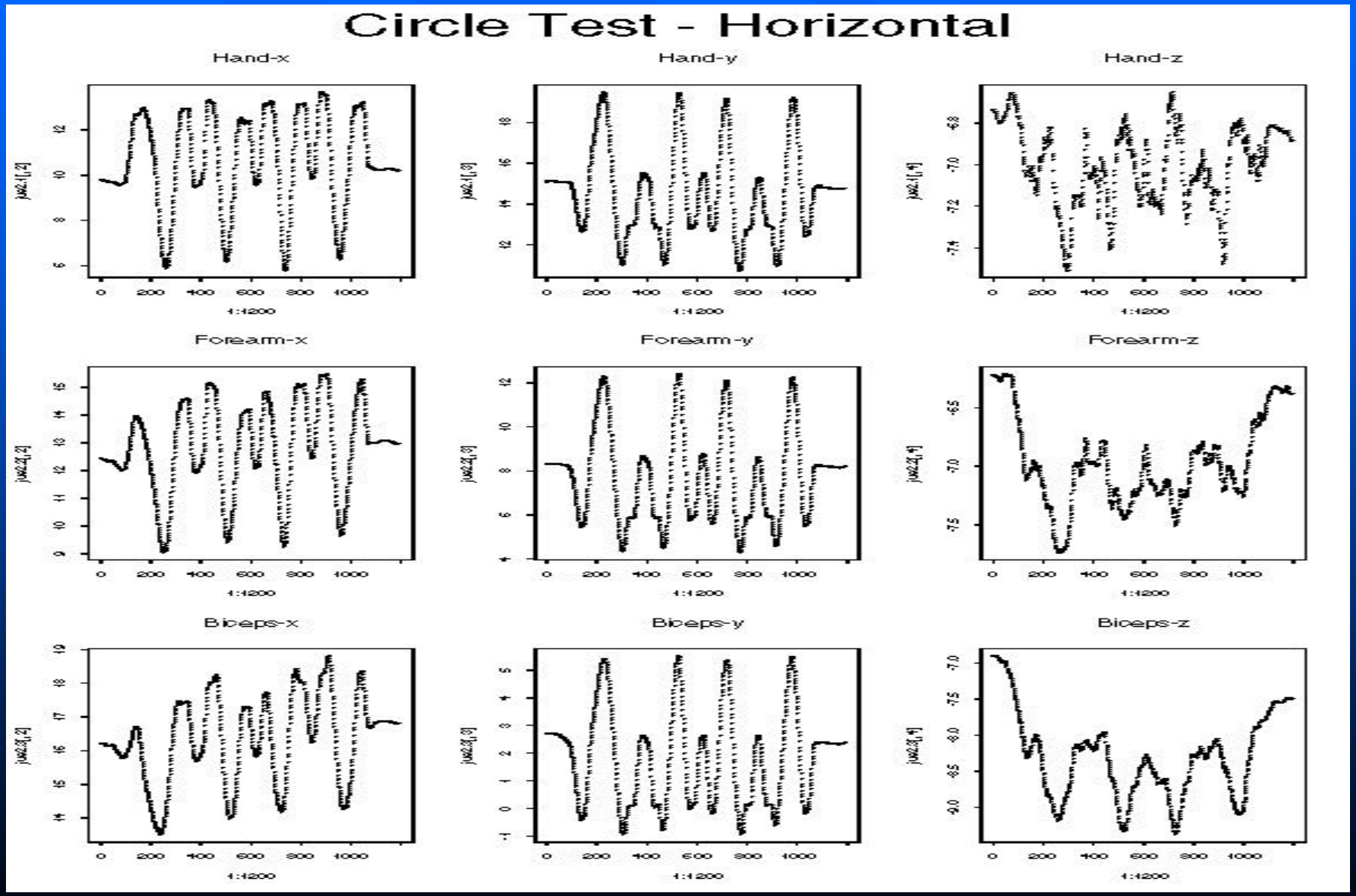
Aim of Preliminary Experiments

- Get used to sensors & other hardware.
- How does visualization help to understand the data?
- Need: Visualization during experiments
  - Complicated setup - impossible to redo once finished
  - Data plausible?
  - Data correctly recorded?

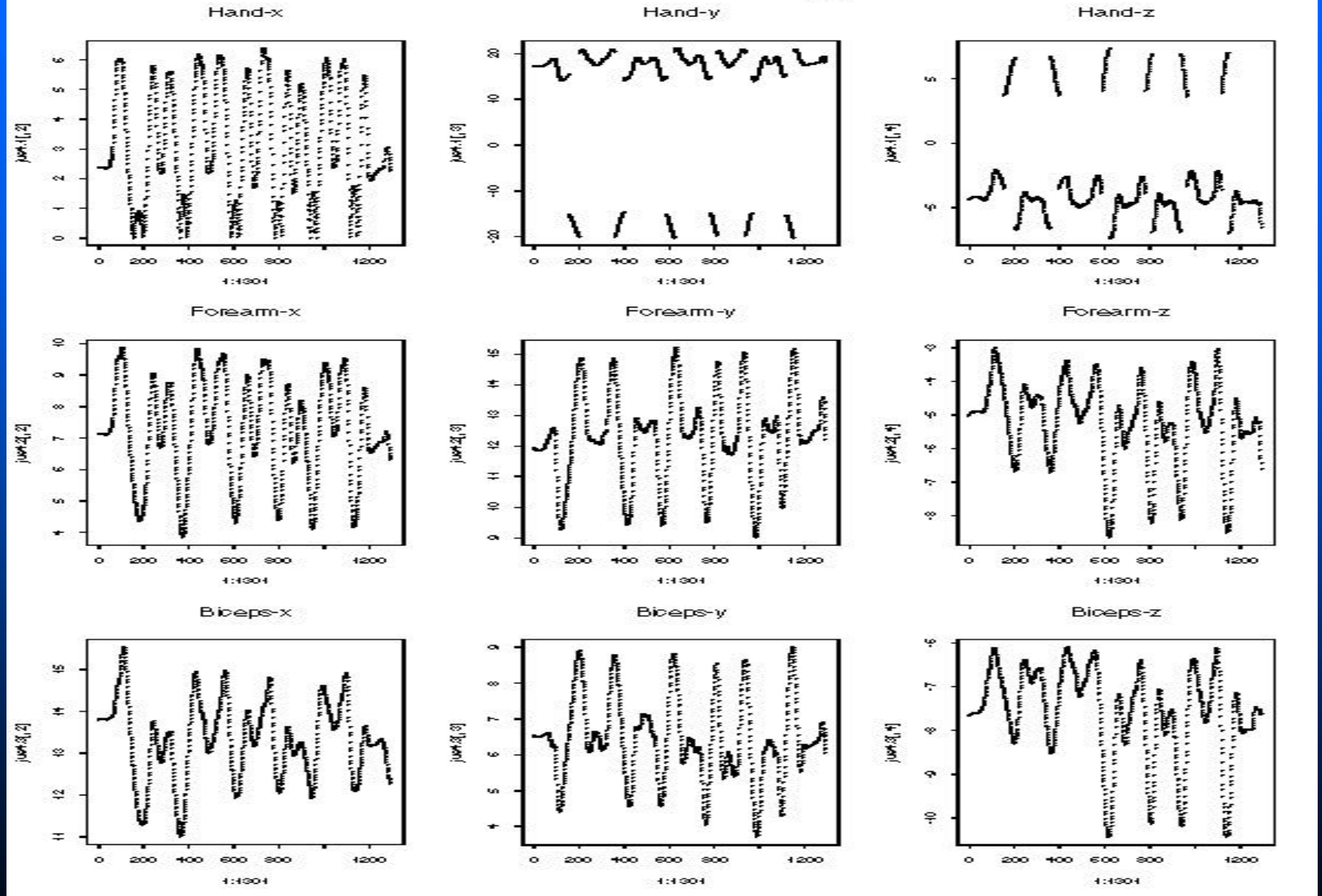
Data Collected

- Small to medium size data set:
  - 60 to 100 Hz
  - 30 to 120 sec
  - 6 x 3 Flock of Birds (FOB) sensors
  - Here: 25,000 to 40,000 Measurements

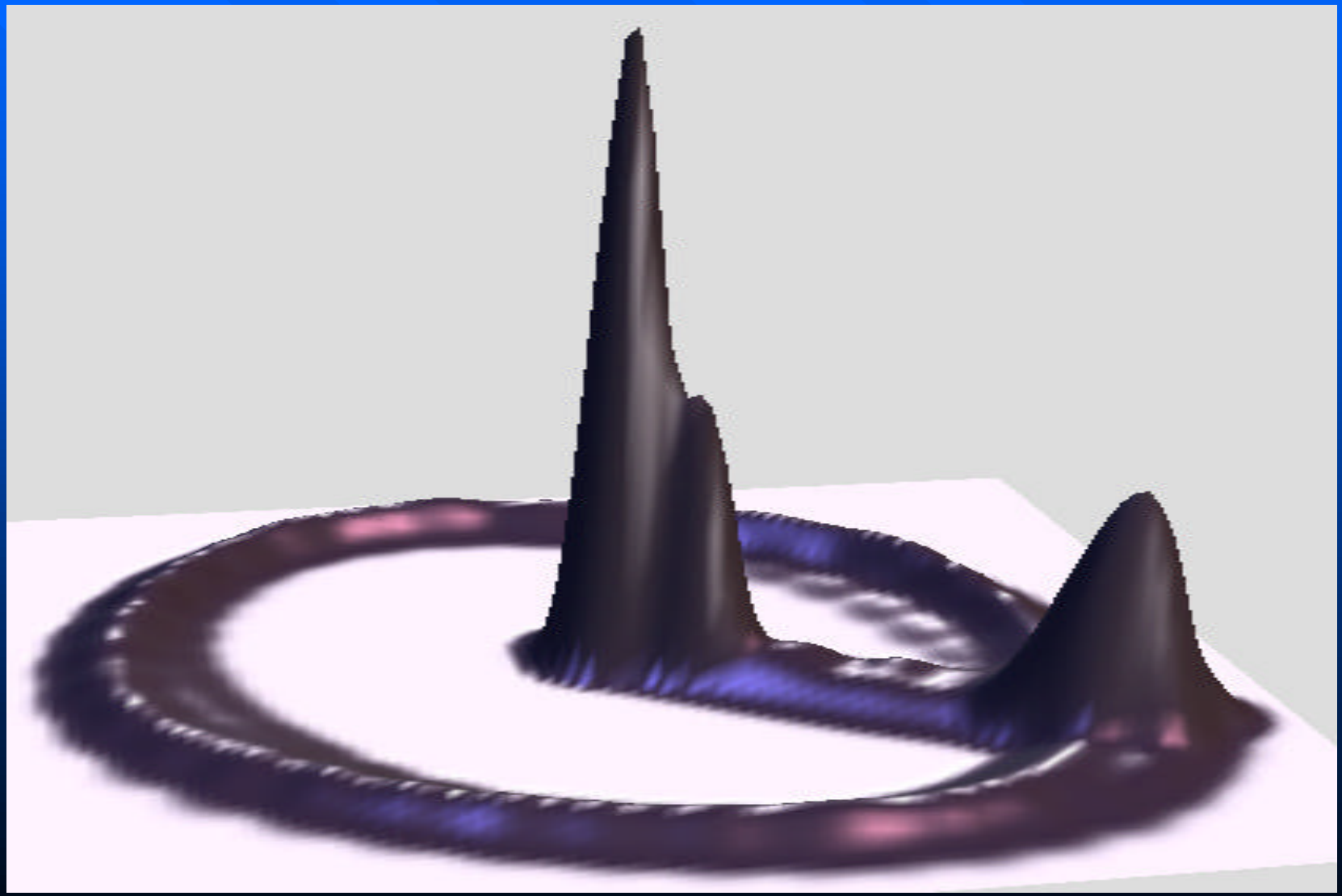
Timeseries Plots (S-Plus)



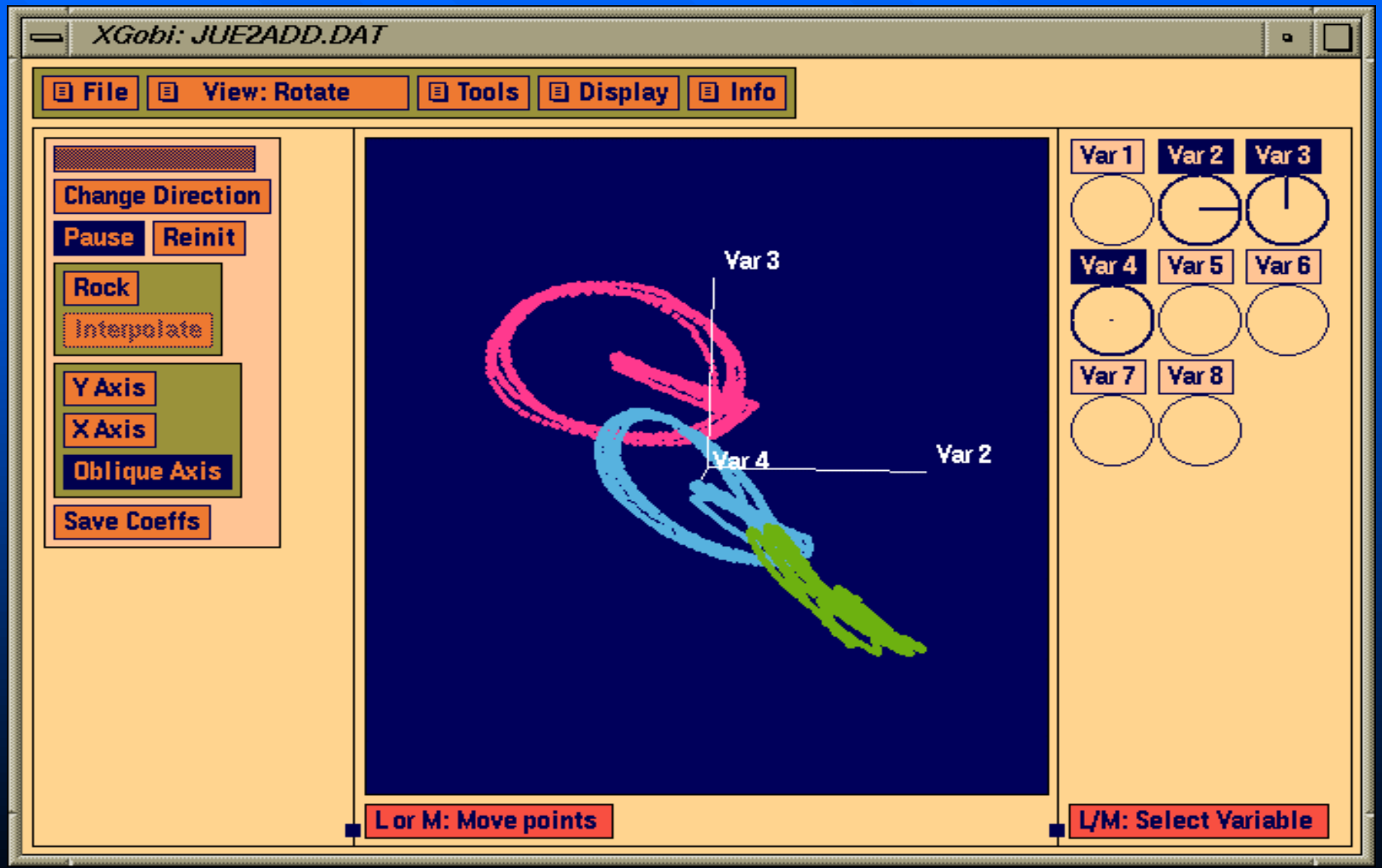
Circle Test - Angular

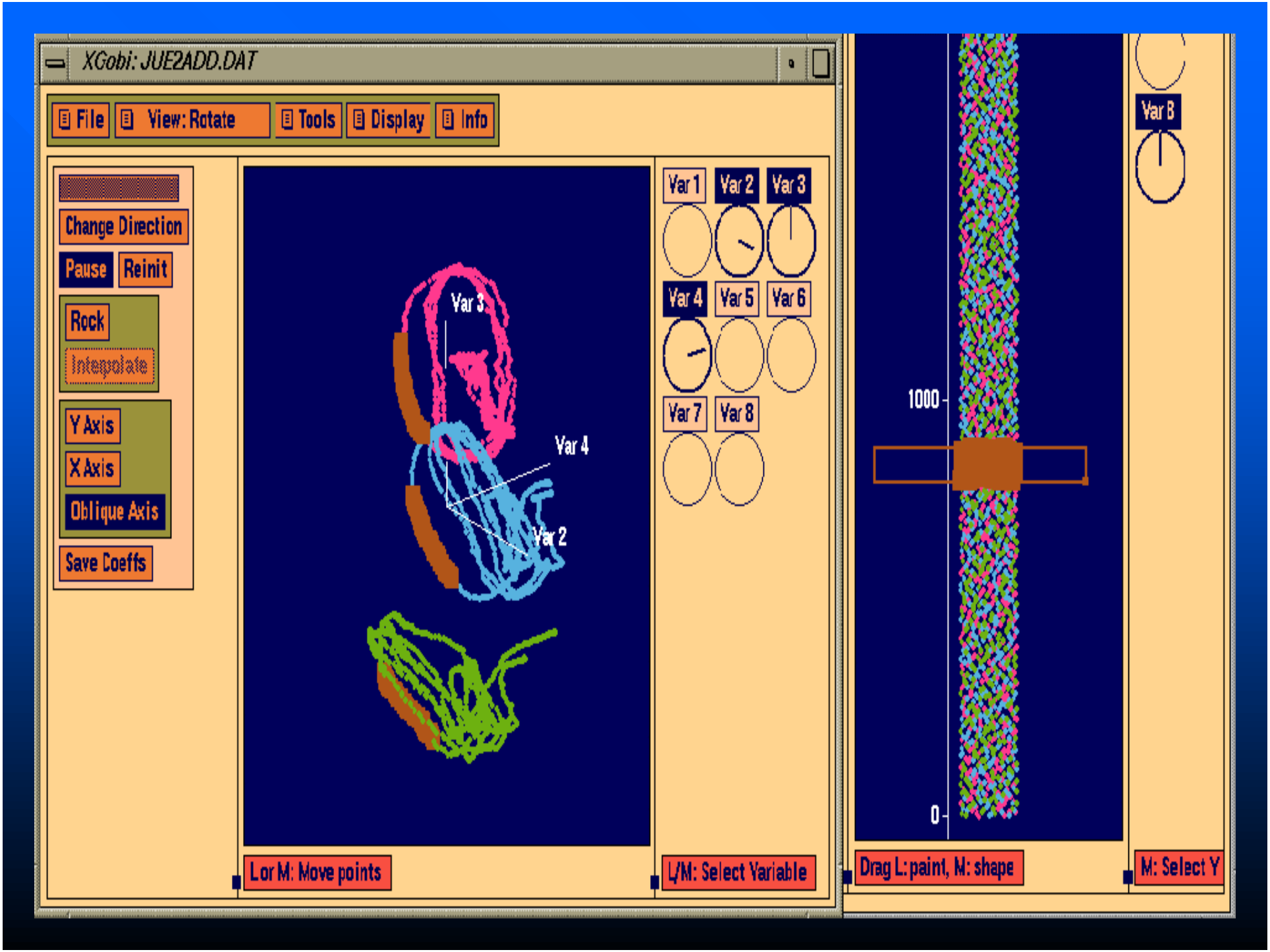


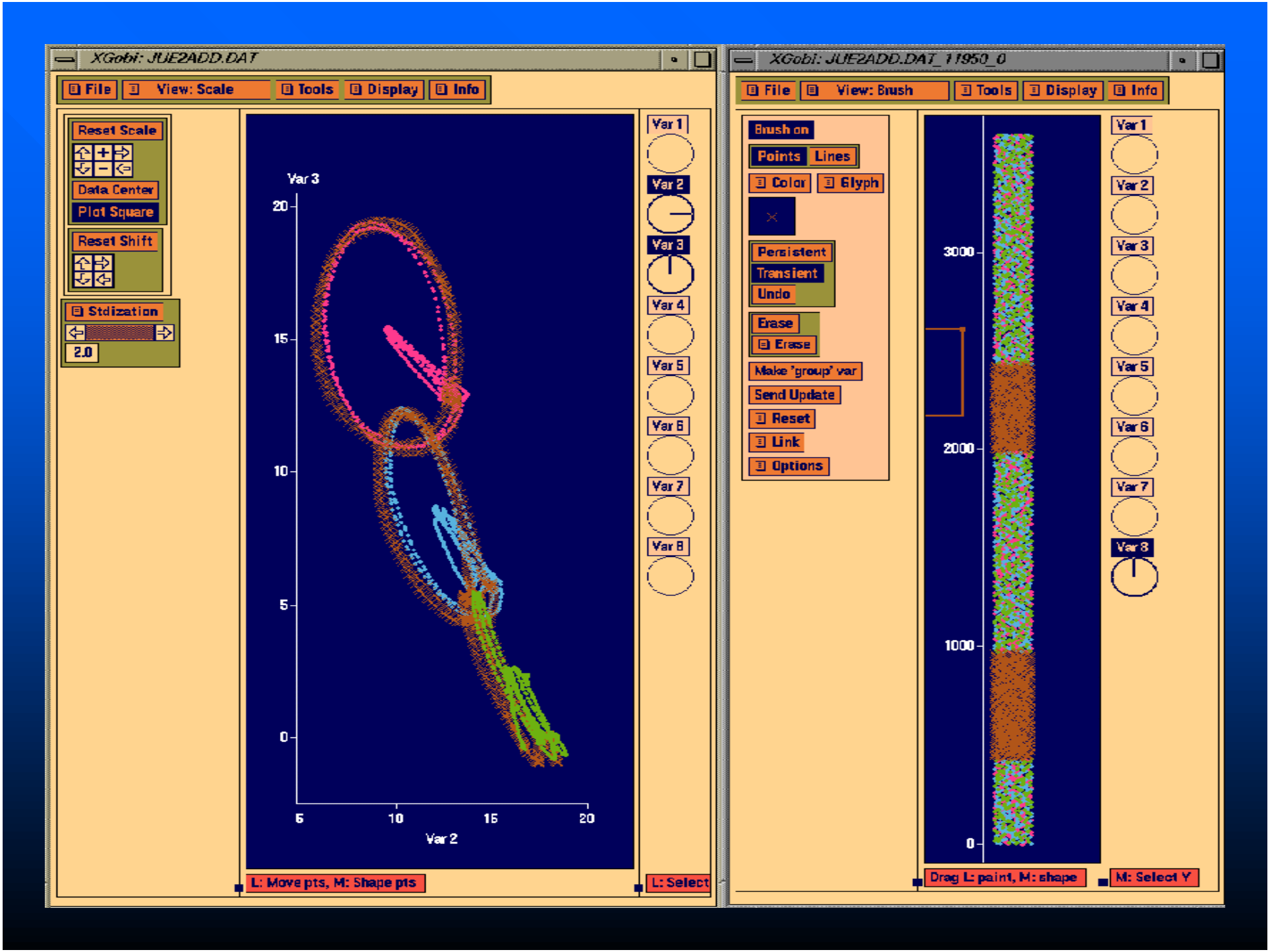
Density Plots (ExplorN)

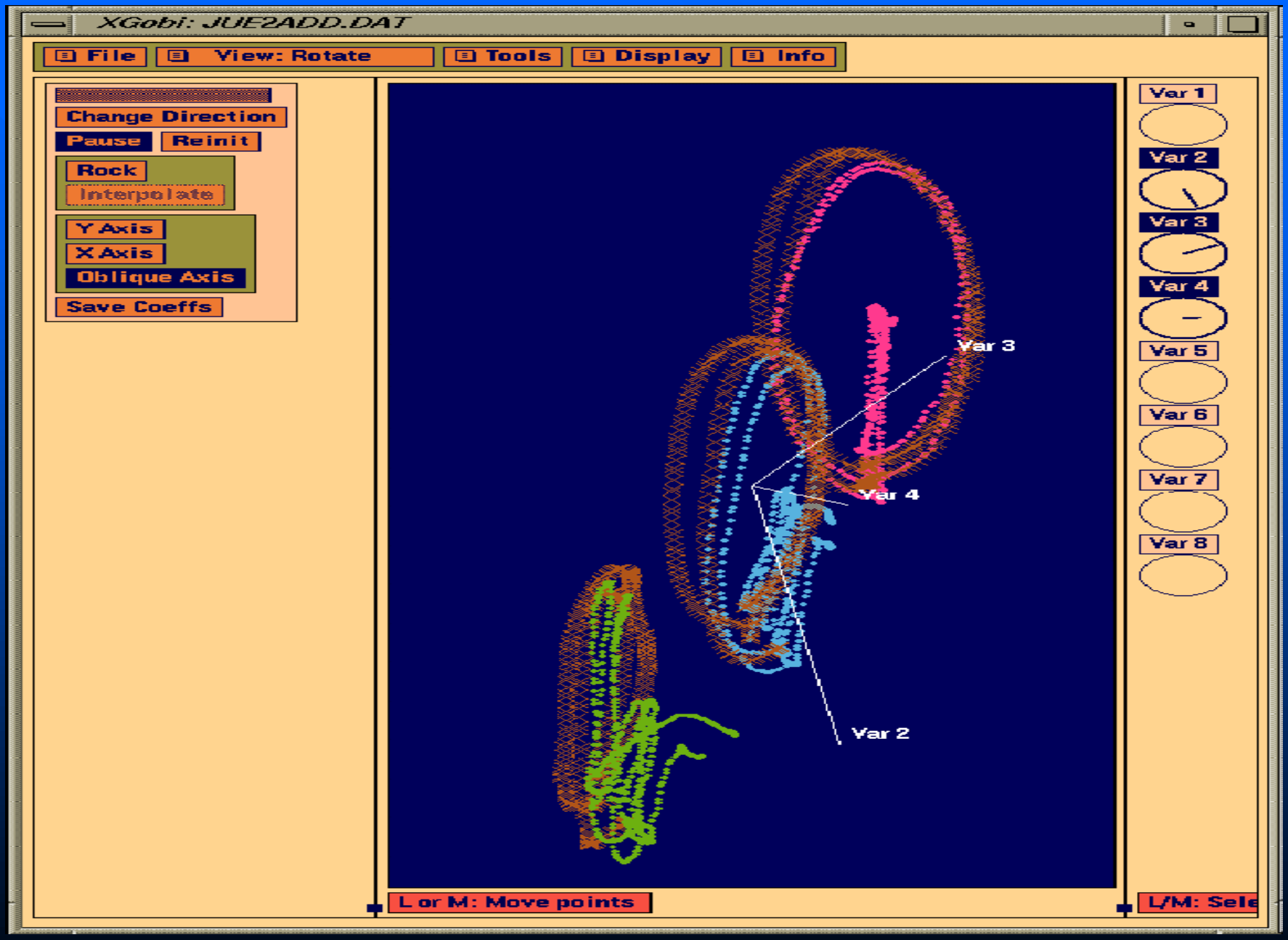


Scatterplots and Rotation (XGobi)




XGobi: JUE2ADD.DAT
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Var 3
Var 4
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
1000
0
Var 8
L or M: Move points
L/M: Select Variable
Drag L: paint, M: shape
M: Select Y


XGobi: JUE2ADD.DAT
File
View: Scale
Tools
Display
Info
Reset Scale
Data Center
Plot Square
Reset Shift
Stdization
2.0
Var 3
20
15
10
5
0
5
10
15
20
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
L: Move pts, M: Shape pts
L: Select
XGobi: JUE2ADD.DAT_11950_0
File
View: Brush
Tools
Display
Info
Brush on
Points
Lines
Color
Glyph
Persistent
Transient
Undo
Erase
Erase
Make 'group' var
Send Update
Reset
Link
Options
3000
2000
1000
0
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
Drag L: paint, M: shape
M: Select Y


XGobi: JUE2ADD.DAT
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Var 3
Var 4
Var 2
Var 1
Var 2
Var 3
Var 4
Var 5
Var 6
Var 7
Var 8
L or M: Move points
L/M: Sele

Motion - Conclusions

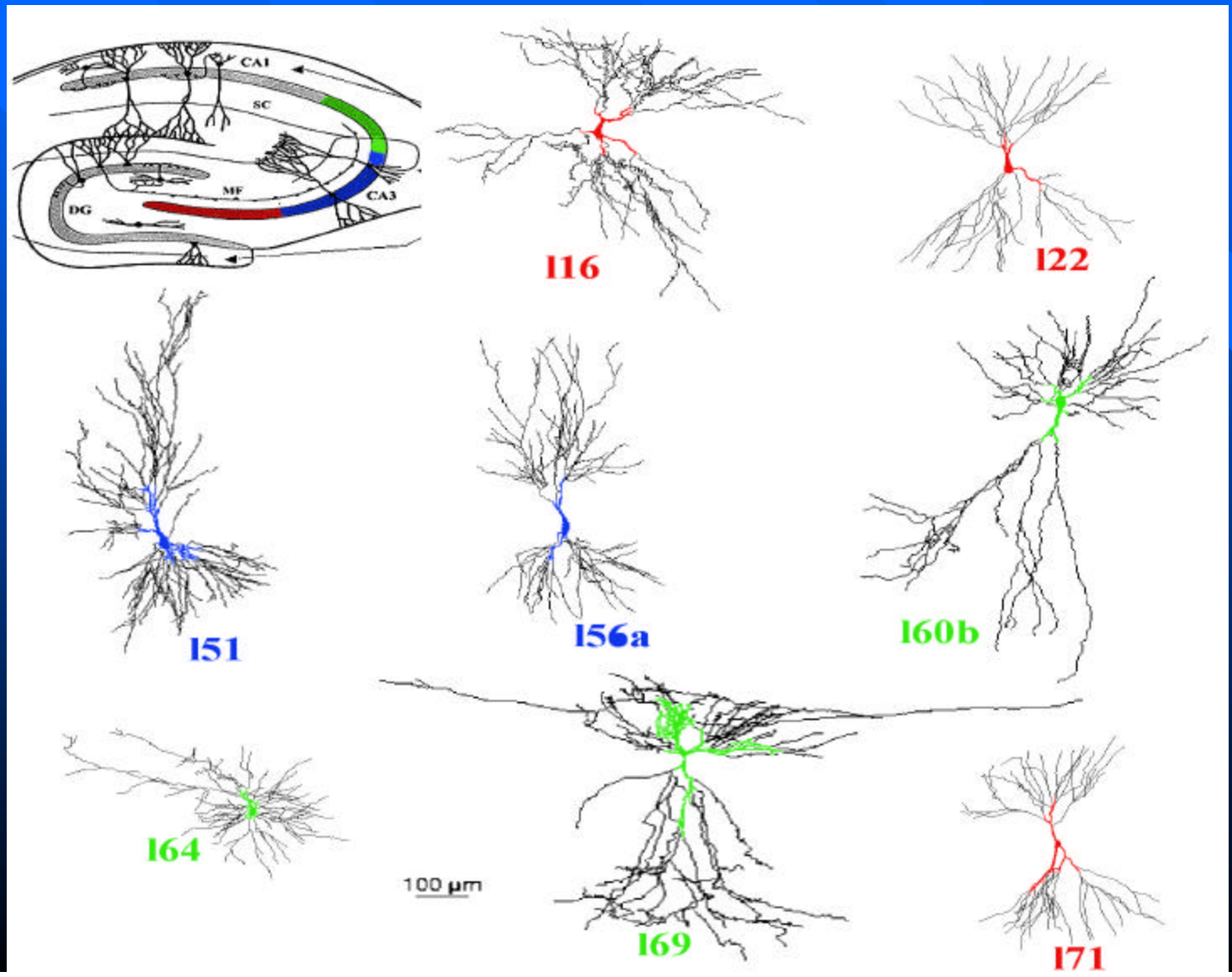
- Visualization helps to immediately check the correctness of the data.
- Realistic 3D Visualization helps to detect unexpected behavior.

Example 3: Neuroanatomical Data

Published as:

Symanzik, J., Ascoli, G. A., Washington, S. S., Krichmar, J. L. (1999): Visual Data Mining of Brain Cells, Computing Science and Statistics, 31:445-449.

Pyramidal Brain Cells



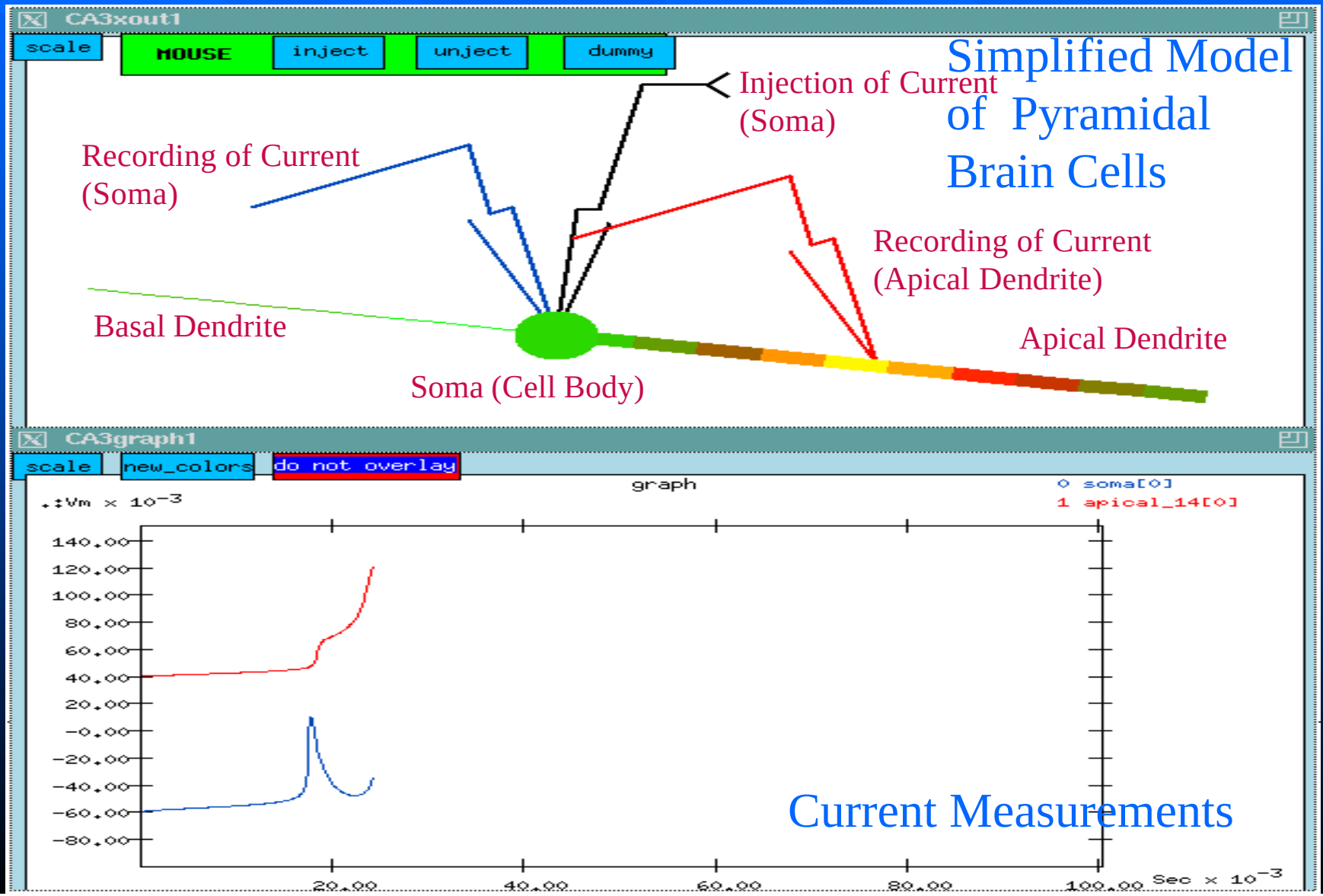
Morphological Parameters

- Apical Dendrite
- Basal Dendrite
- Distance from Soma
  - 50 um
  - 100 um
  - 150 um
  - 200 um
  - Entire Dendrite Tree
- Length
- Diameter
- Area
- Asymmetry
- Bifurcations
- Terminations

Aim of the Study

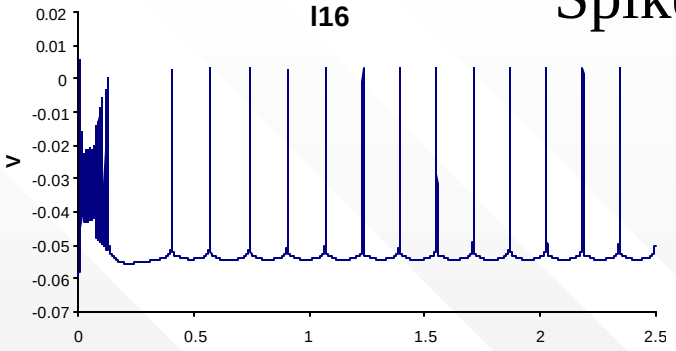
- Study the function of neurons by injecting current into a neuron and measure the neuron's response
- Here: Computational Simulator
- 16 sets of morphometric data used
- About 3 hours of computer time for 5 sec of neuron time on SGI Origin 200
- 10 injected currents per cell: 0.1 nA to 1.9 nA

Simulation

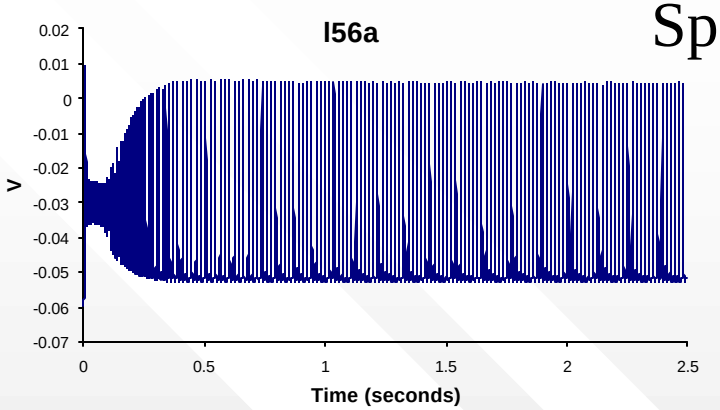

CA3xout1
scale
MOUSE
inject
unject
dummy
Simplified Model of Pyramidal Brain Cells
Injection of Current (Soma)
Recording of Current (Soma)
Recording of Current (Apical Dendrite)
Basal Dendrite
Apical Dendrite
Soma (Cell Body)
CA3graph1
scale
new_colors
do not overlay
graph
0 soma[0]
1 apical_14[0]
Vm x 10^-3
140.00
120.00
100.00
80.00
60.00
40.00
20.00
-0.00
-20.00
-40.00
-60.00
-80.00
20.00
40.00
60.00
80.00
100.00
Sec x 10^-3
Current Measurements

Simulated Physiological Response under 0.7 nA

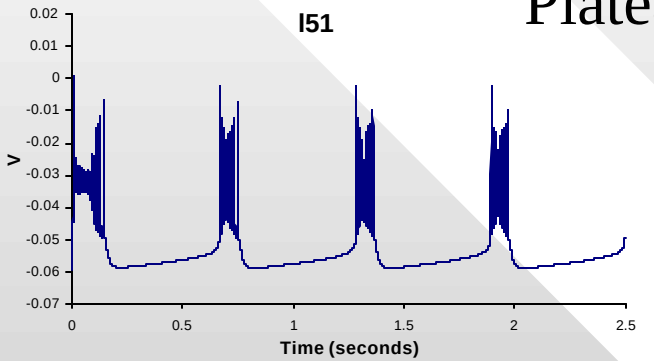
Spike



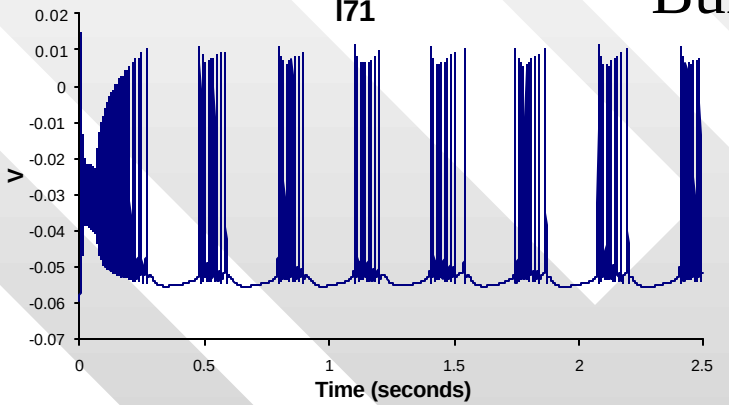
Spike



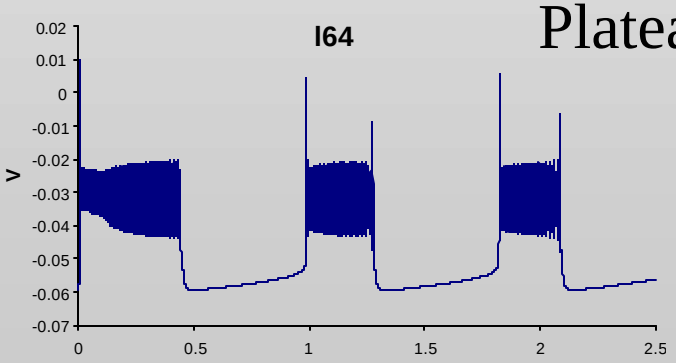
Plateau



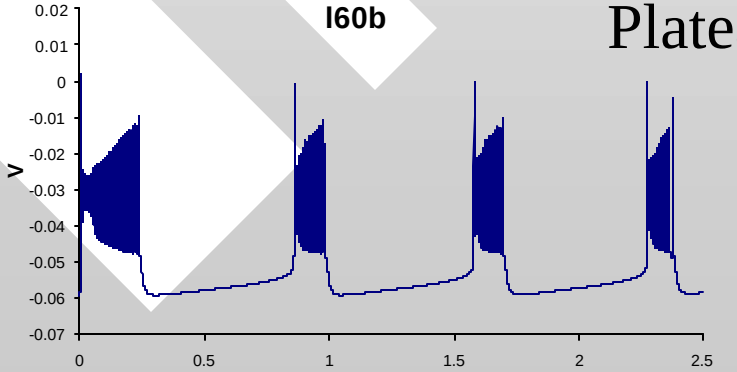
Burst



Plateau

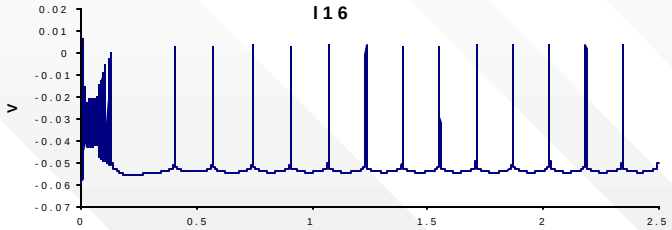


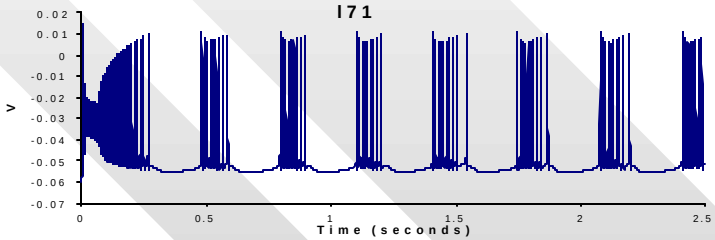
Plateau

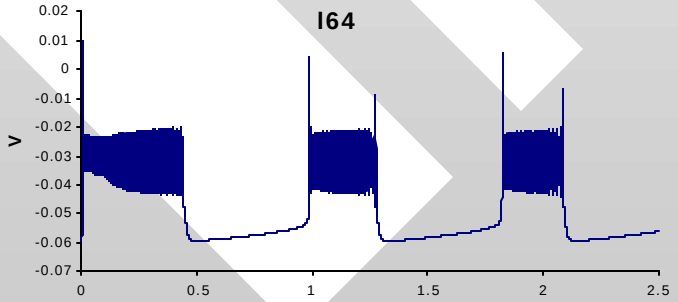


Response Parameters

- Spiking:
  - Spike Rate (Hz)
  - Spike Transition (nA)
- Bursting:
  - Burst Rate (Hz)
  - Interburst Interval (sec)
  - Spikes per Burst (Hz)
- Plateau:
  - Plateau Range (nA)
  - Plateau Rate (Hz)
  - Interplateau Interval (sec)
  - Spikes per Plateau (Hz)

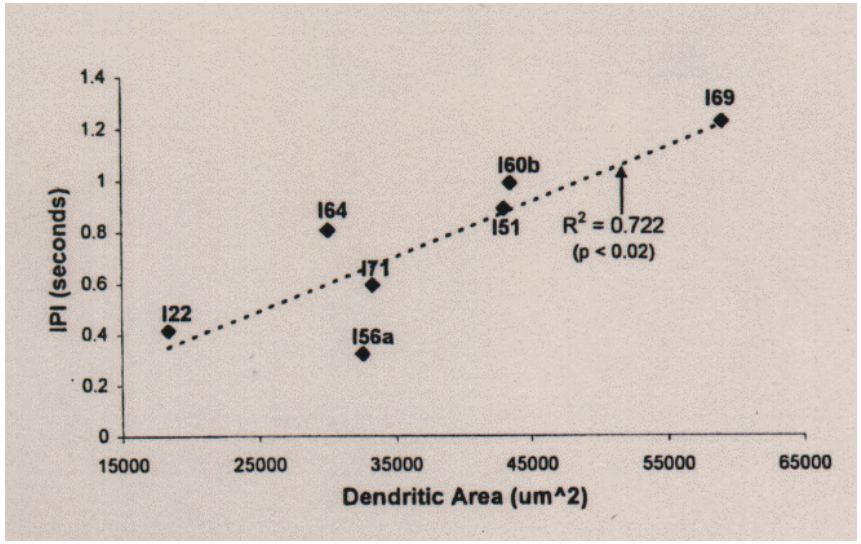






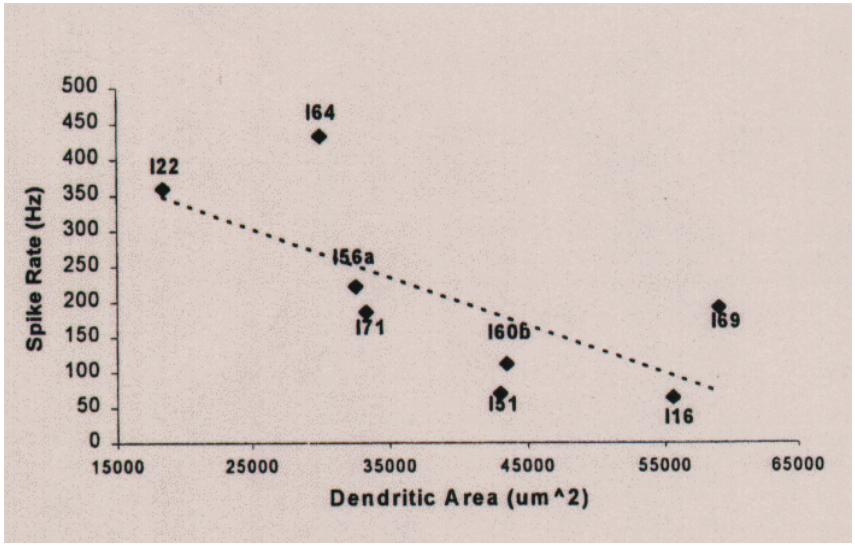
Influence of Dendritic Area on Firing Rate

Interplateau Interval vs Dendritic Area



Current: 0.5 nA

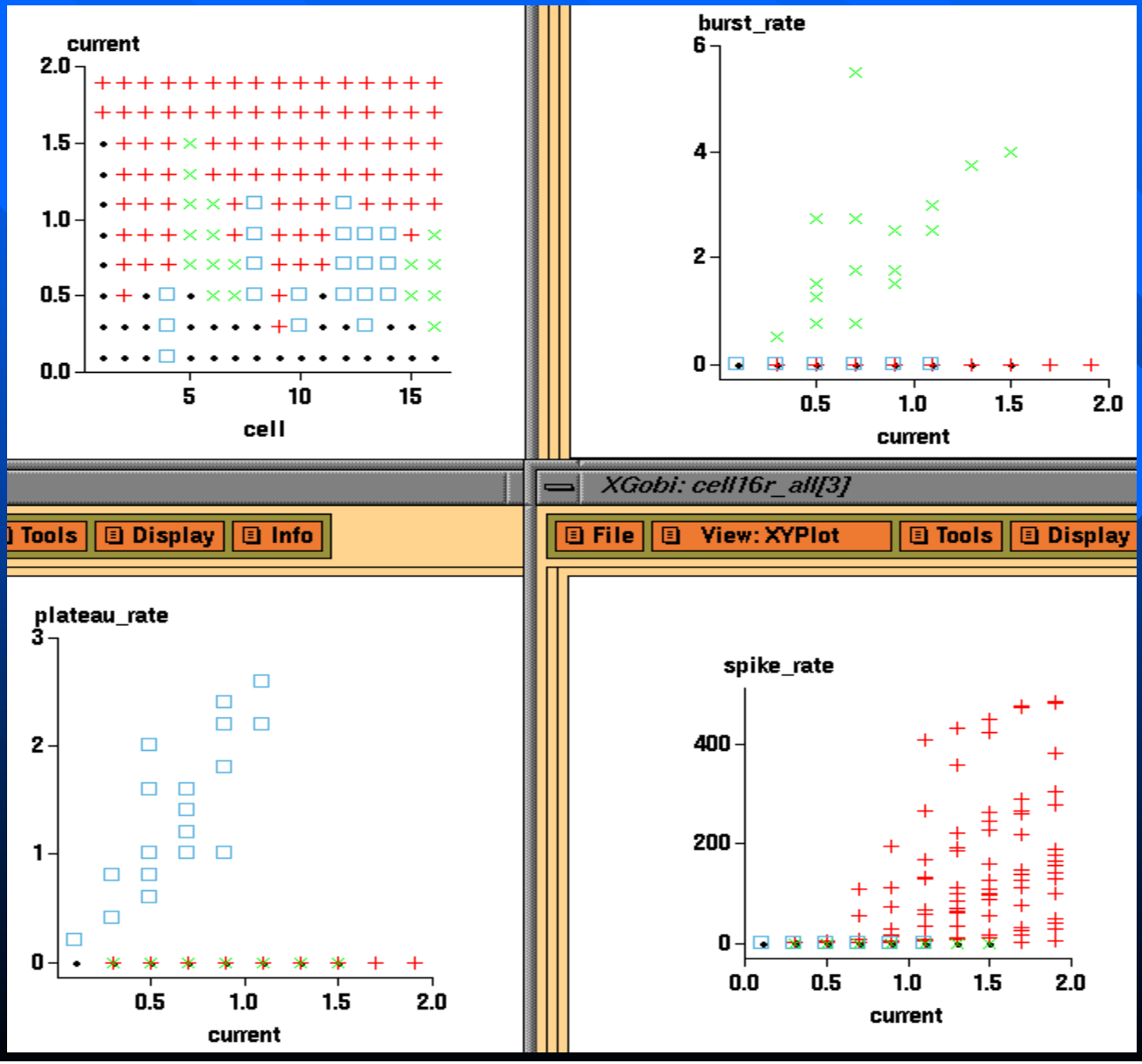
Spike Rate vs Dendritic Area



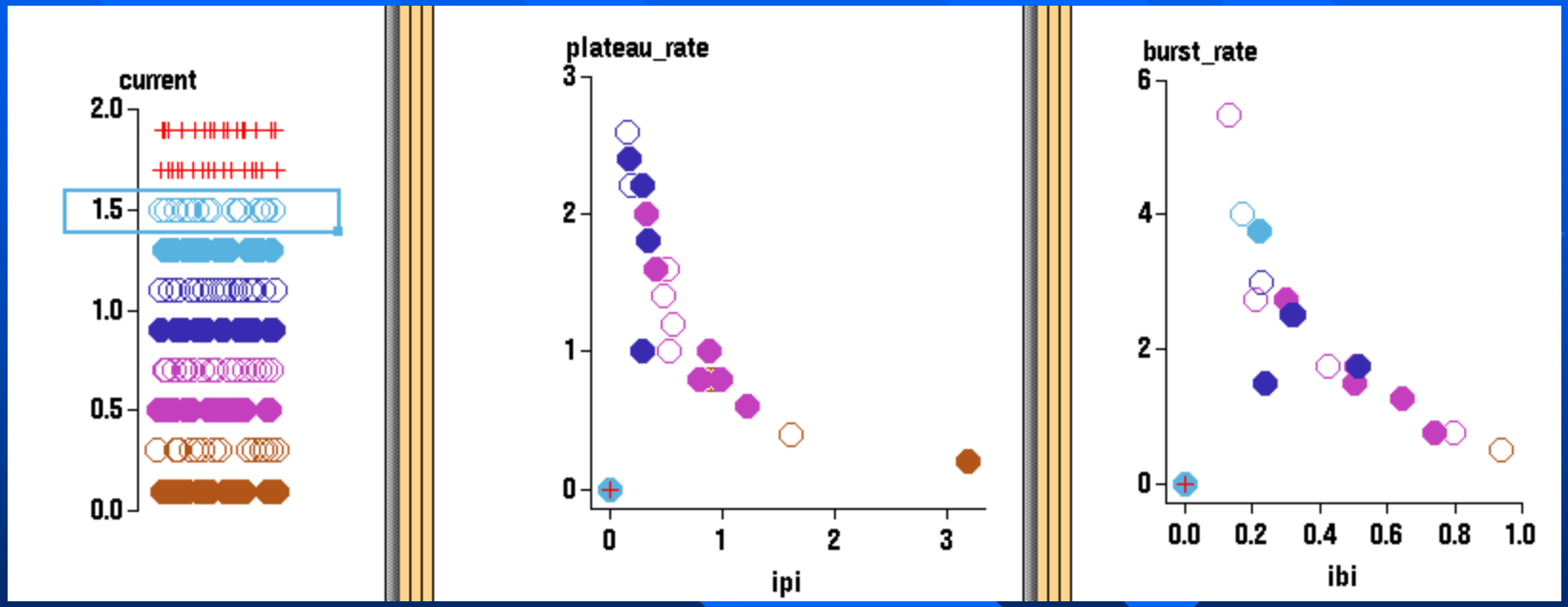
Current: 1.3 nA

- Smaller cells tend to be more excitable and have higher firing rates.

Visual Data Mining Using XGobi



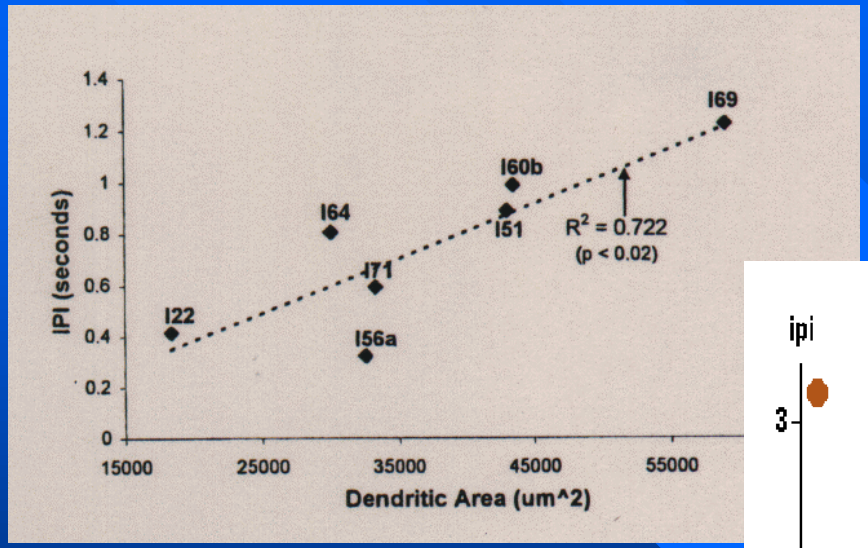
Visible Patterns

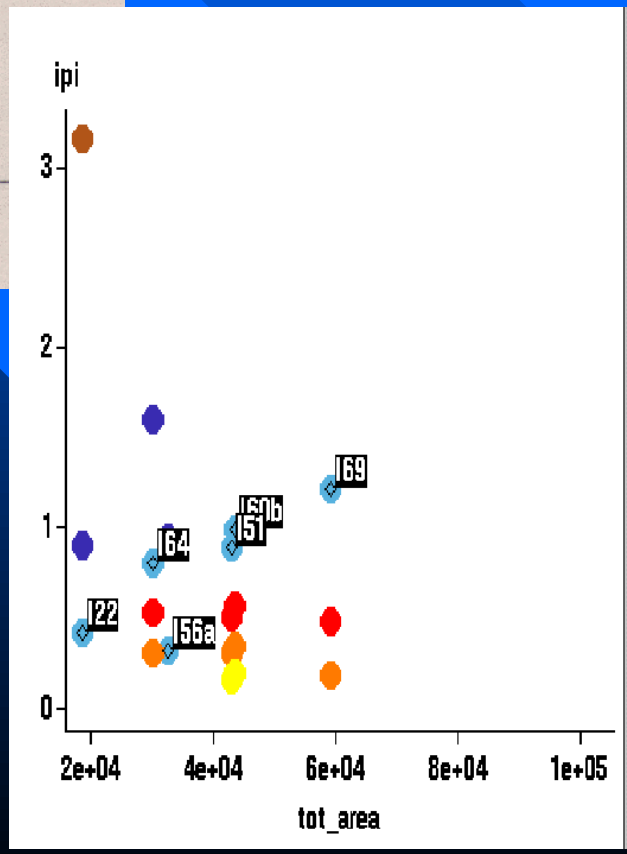


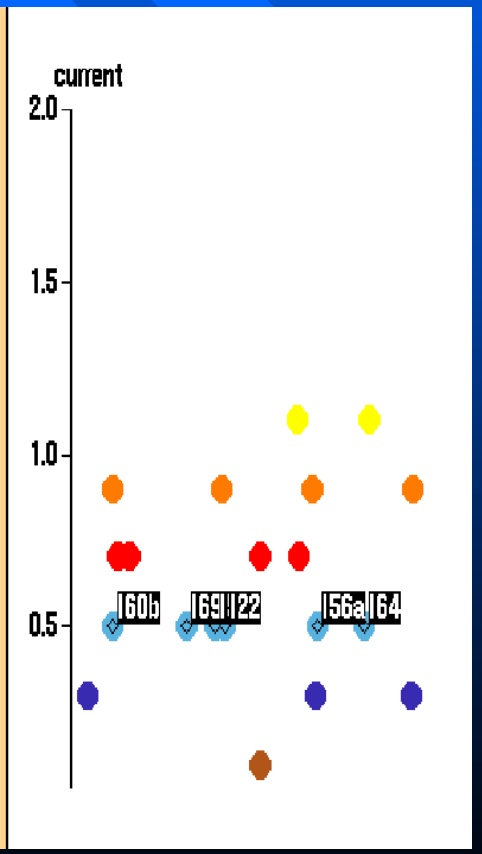
Interplateau Interval

Interburst Interval

Interplateau Interval vs Dendritic Area ???







Brain Cells - Conclusions

- Visualization suggests which cells to simulate/analyze next.
- Some prior assumptions may not hold or only hold under additional restrictions.

Example 4: Remote Sensing Data

Published as:

Symanzik, J., Griffiths, L., Gillies, R. (2000): Visual Exploration of Satellite Images, 2000 Proceedings of the Statistical Computing Section and Section on Statistical Graphics, American Statistical Association, Alexandria, Virginia, 10-19.

ArcView

ESRI™

- Desktop GIS with wide Range of Viewing and Data Manipulation Functions
  - Editing Features
  - Query Operations
  - Map Display
  - Interactive Interface
  - High Level Internal Scripting Language
  - ArcView has been linked to XGobi

The Data

- NOAA-14 Satellite (National Oceanic and Atmospheric Administration)
- AVHRR Sensor (Advanced Very High Resolution Radiometer):
  - Band 1: Red
  - Band 2: Near Infrared
  - Band 3: Mid Infrared
  - Band 4: Long Infrared
  - Band 5: (Very) Long Infrared
- Data from “NASA’s Project Atlanta”
- 18 Days from Jan 1997 to Dec 1997
- Resolution: 1 km x 1 km per Pixel
- Main Study Area: 70 km x 46 km

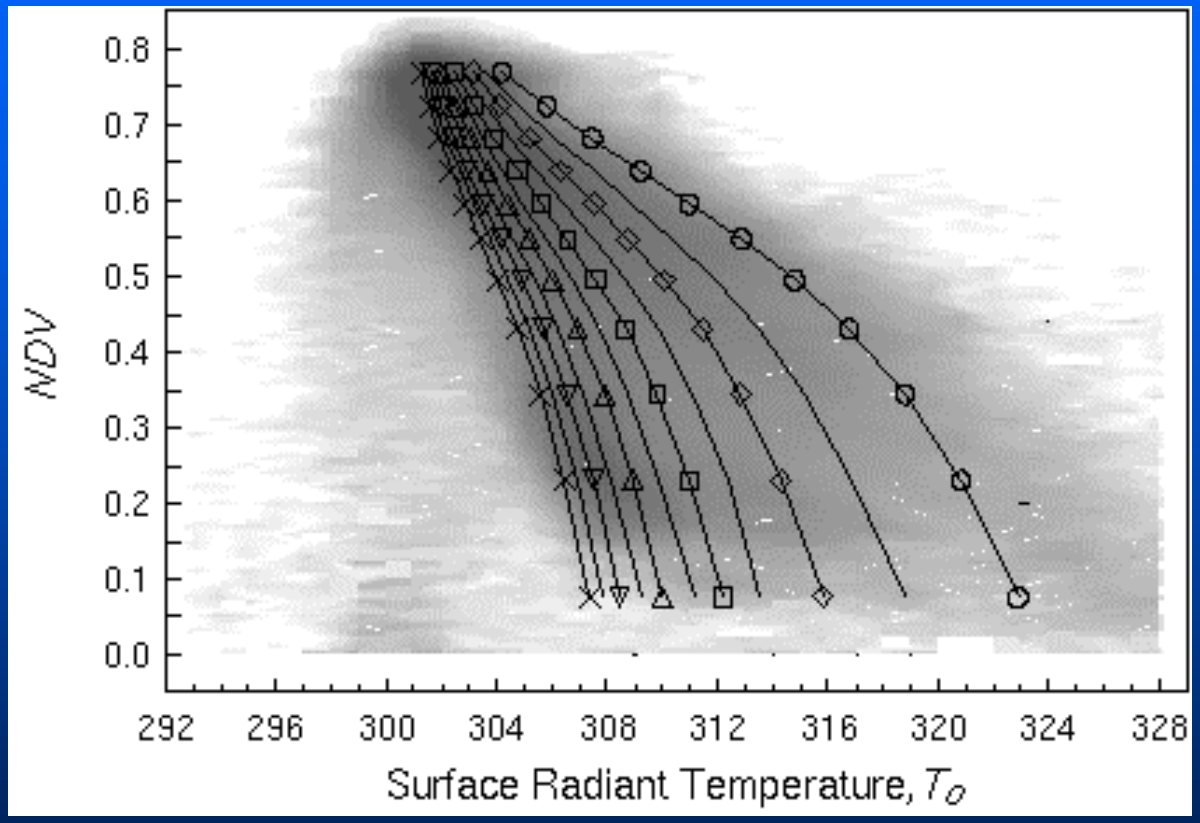
Some Definitions

- Normalized Difference Vegetation Index:

$$NDVI = \frac{Band2 - Band1}{Band2 + Band1}$$

- *NDVI* ~ 0.8 for Highly Vegetated Surfaces
- *NDVI* ~ 0.1 for Bare Soil
- Surface Radiant Temperature T_0: Band 4
- Surface Moisture Availability M_0

An Example

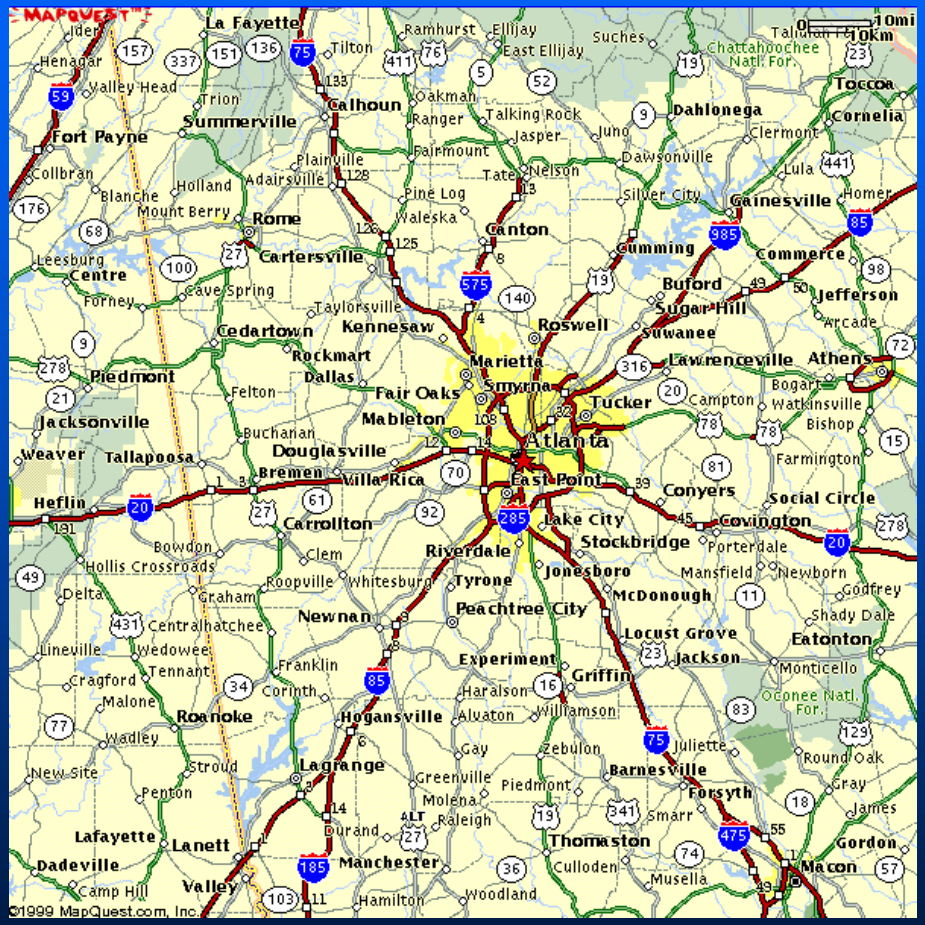


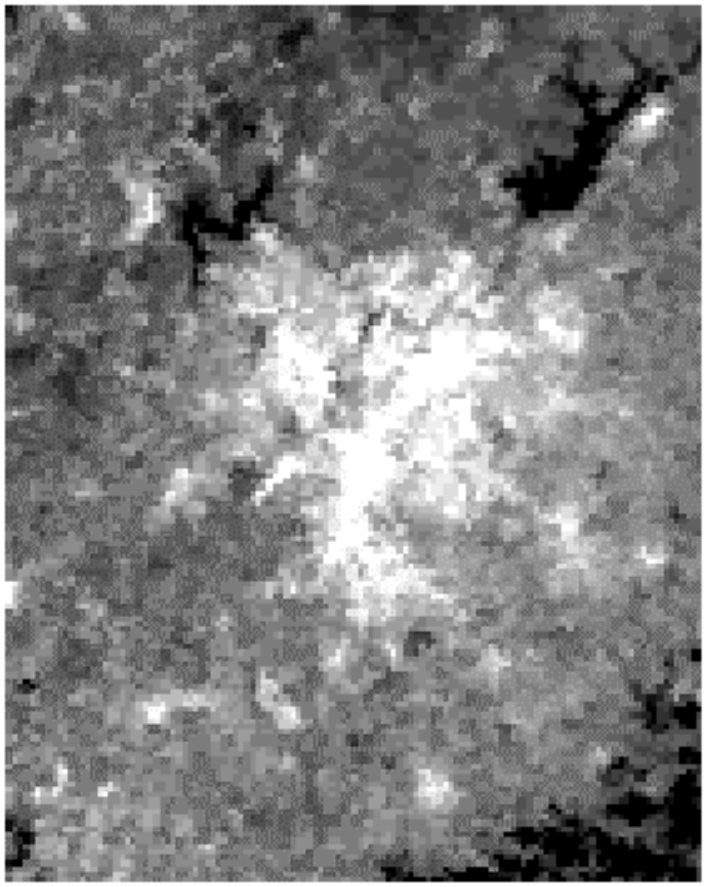
NS001-TMS derived T_o-*NDVI* scatterplot (gray spectral scaling) at a 5 meter spatial resolution for a 7 x 3 km area of the Mahantango Watershed, Pennsylvania. 18 July 1990, 1145 LST. Isopleths representing moisture availability index, M_o are overlaid with the legend, o = 0.0 ('warm' edge), ◊ = 0.2, □ = 0.4, Δ = 0.6, ∇ = 0.8, and × = 1.0 (cold edge).

Goal of the Study

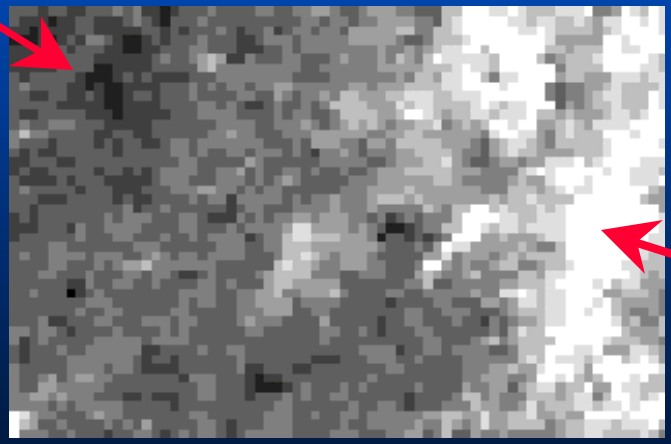
- Explore (and model) relationships between *NDVI*, T_0 and M_0 for different seasons
  - Specify wide-range behavior (e.g., for city, forest, water)
  - Find unusual places

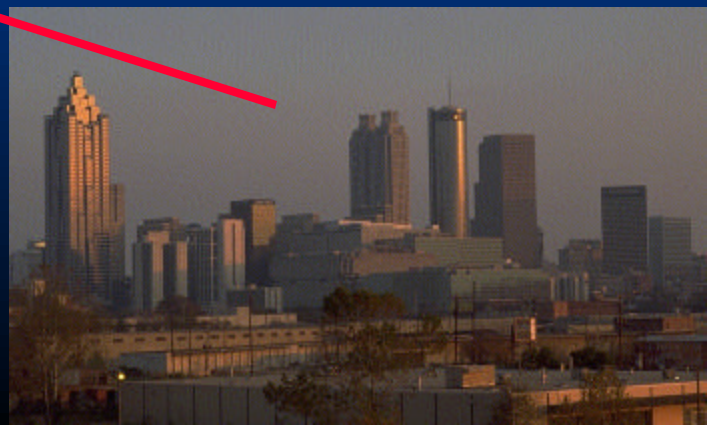
The Geographic Setting



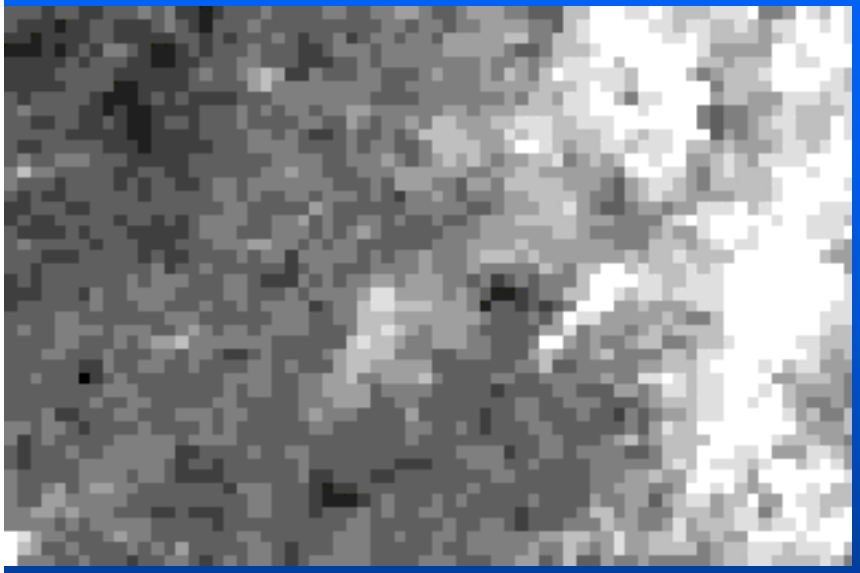


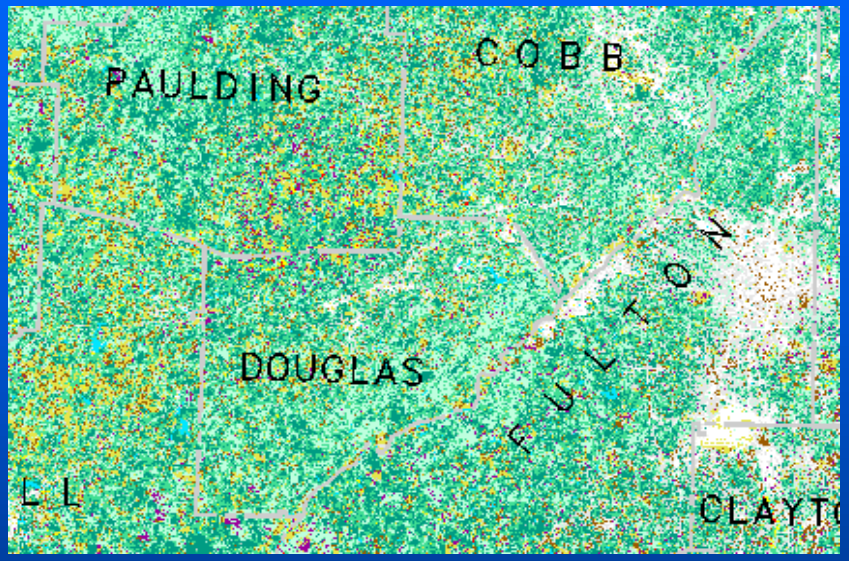
The Main Study Area





The Main Study Area - Landcover



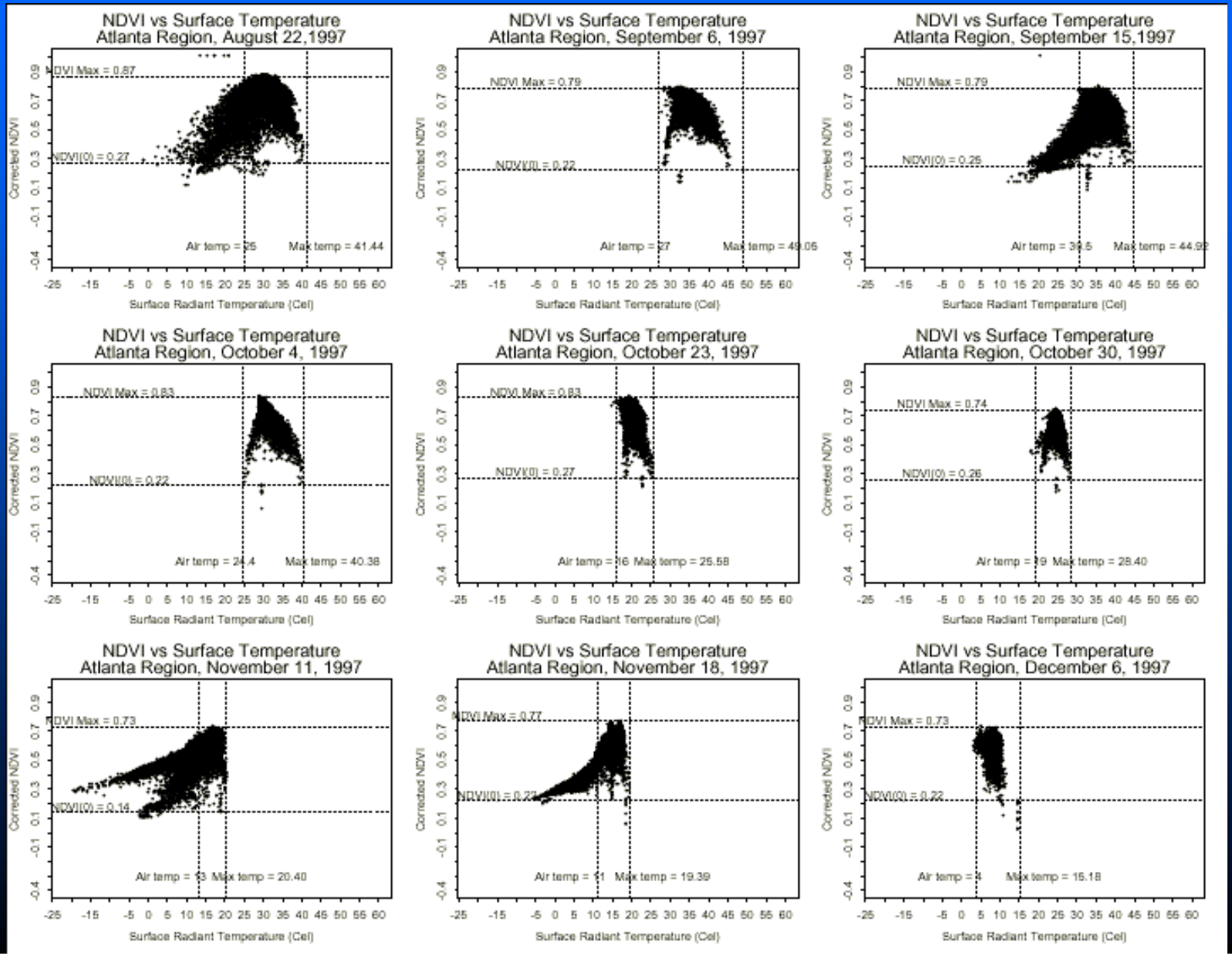


EXPLANATION

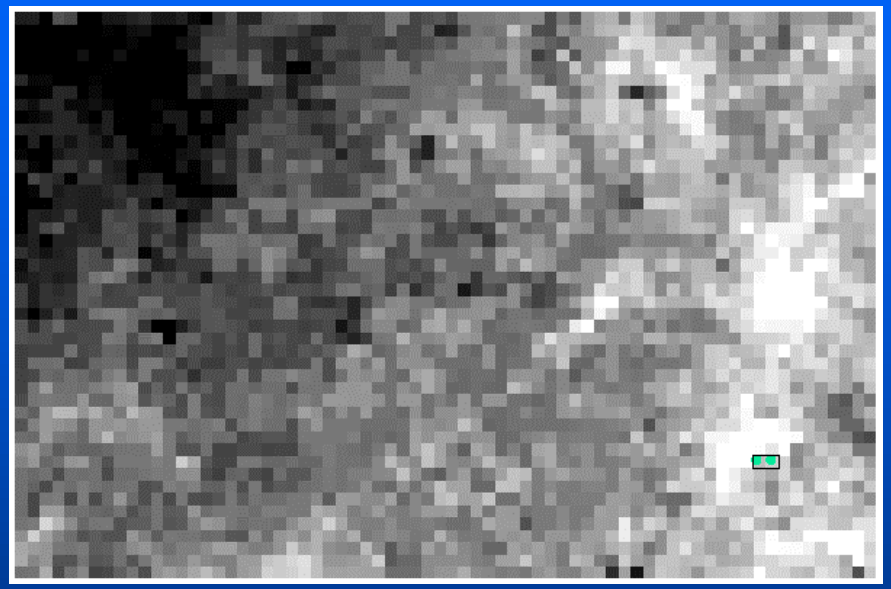
- Open Water
- Clear Cut / Young Pine
- Pasture
- Cultivated / Exposed Earth
- Low Density Urban
- High Density Urban
- Emergent Wetland
- Scrub / Shrub Wetland
- Forested Wetland
- Coniferous Forest
- Mixed Forest
- Hardwood Forest
- Salt Marsh
- Brackish Marsh
- Tidal Flats

Digital Landcover from Georgia Department of Natural Resources, Wildlife Resources Division, Natural Heritage Program, 200ft resolution.
Digital state and county data compiled from US Census TIGER/line files 1:100,000.
Digital shoreline data compiled from NOAA vector shoreline of the US 1:70,000.

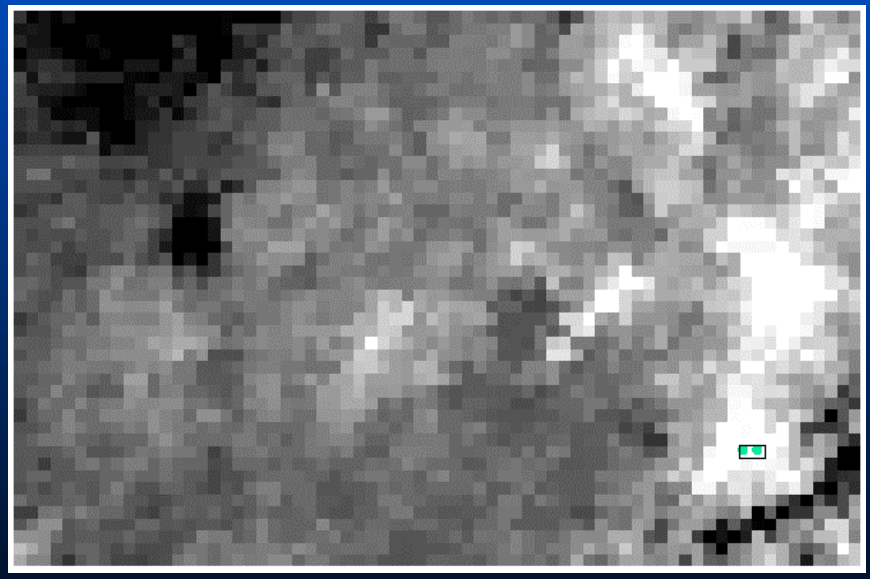
NDVI vs Surface Temperature



Two Months



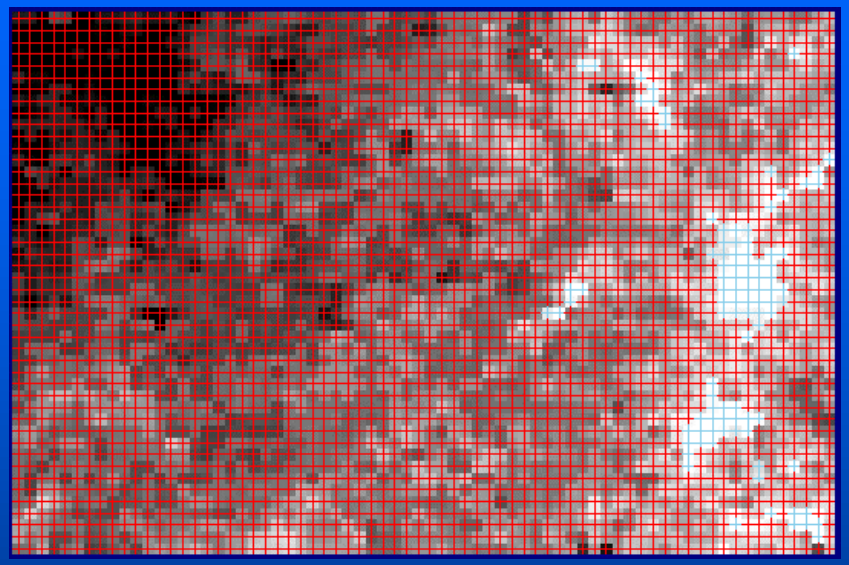
December

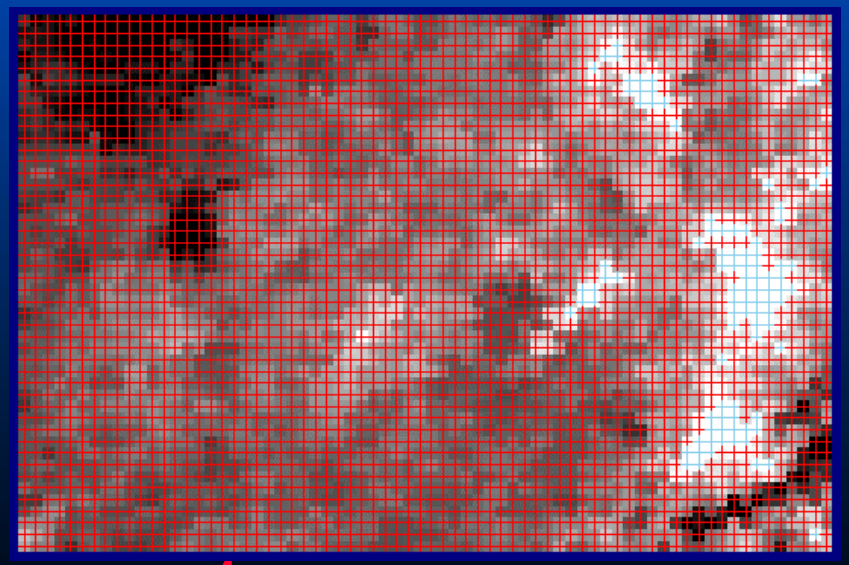


August

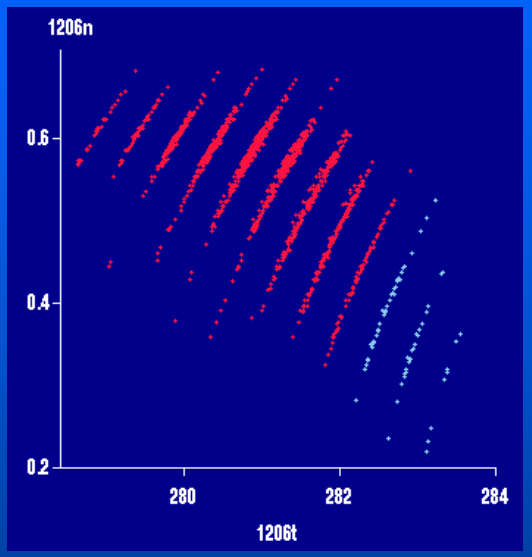
The City

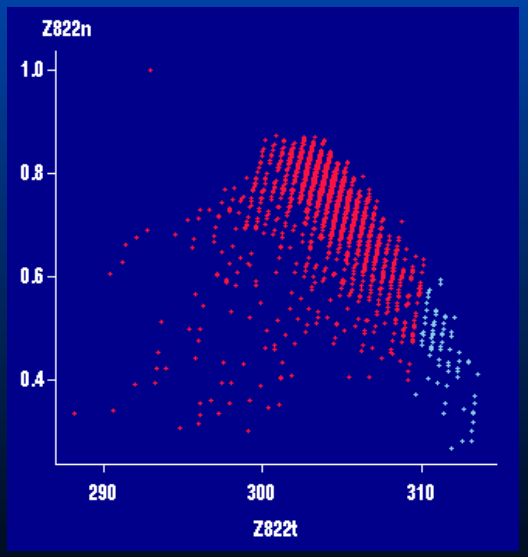
December



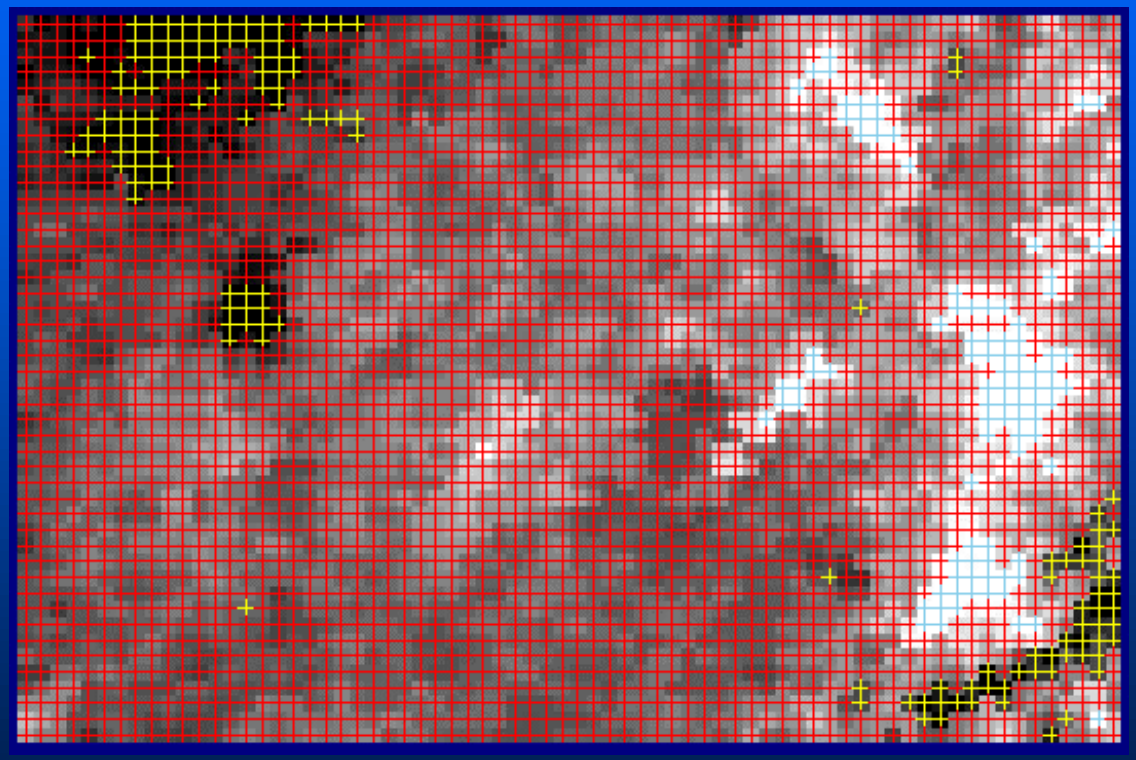


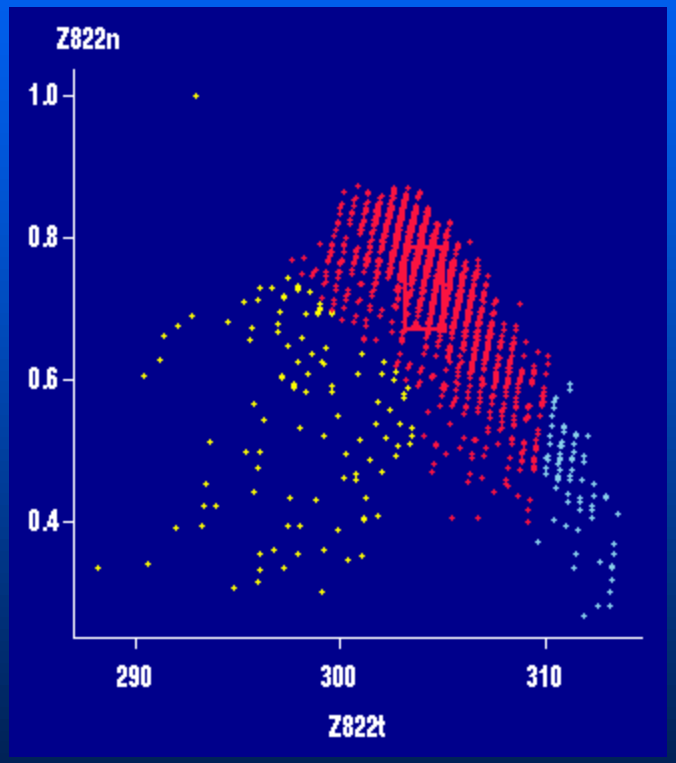
August





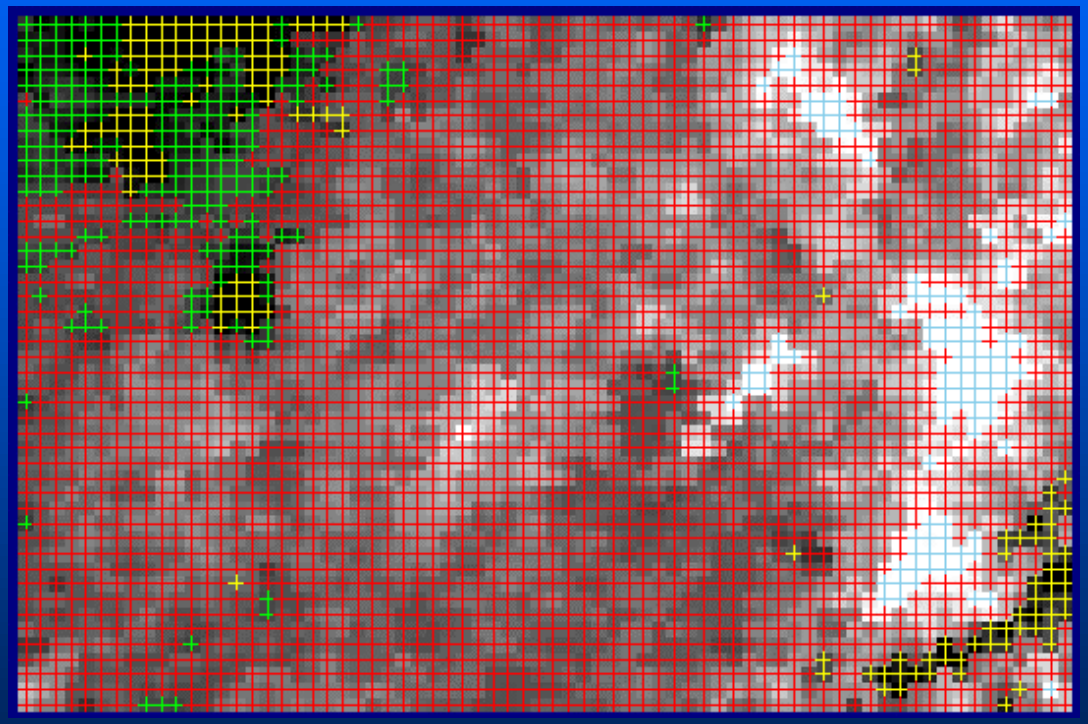
Clouds in August

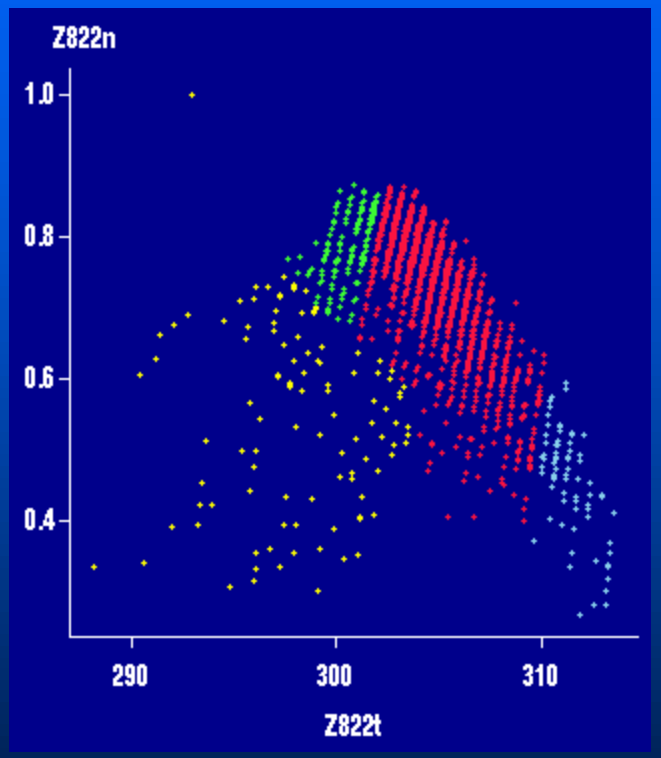




August

Clouds and Forest in August

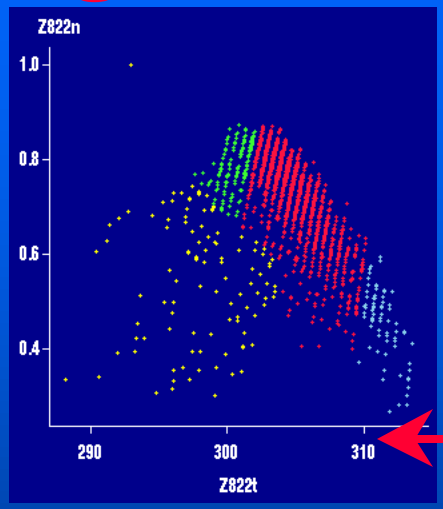




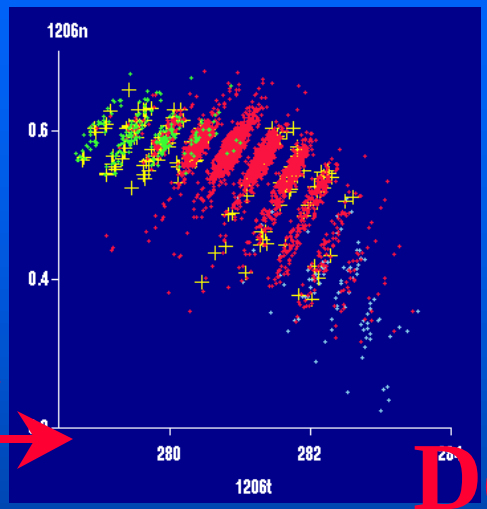
August

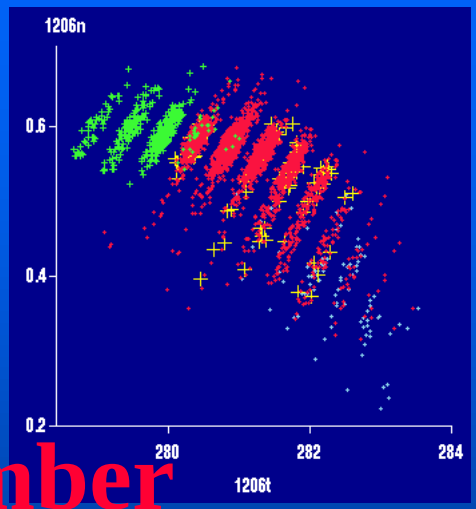
Reclassifying Clouds

August

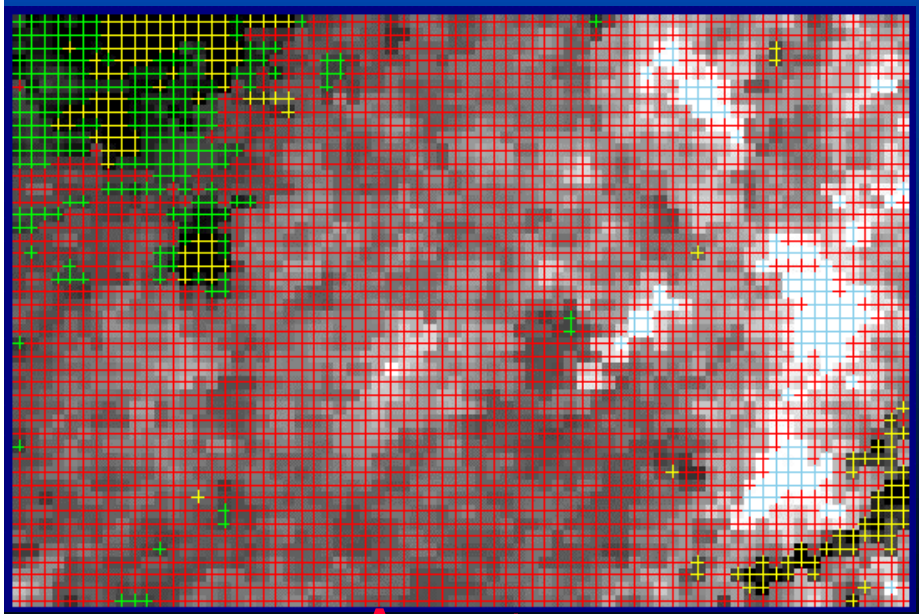


Linked

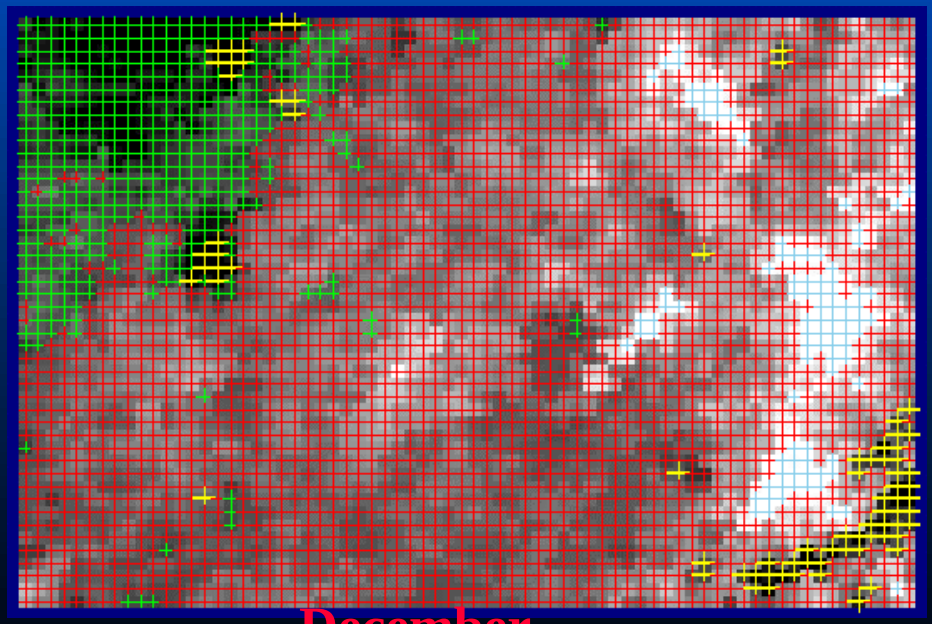




December

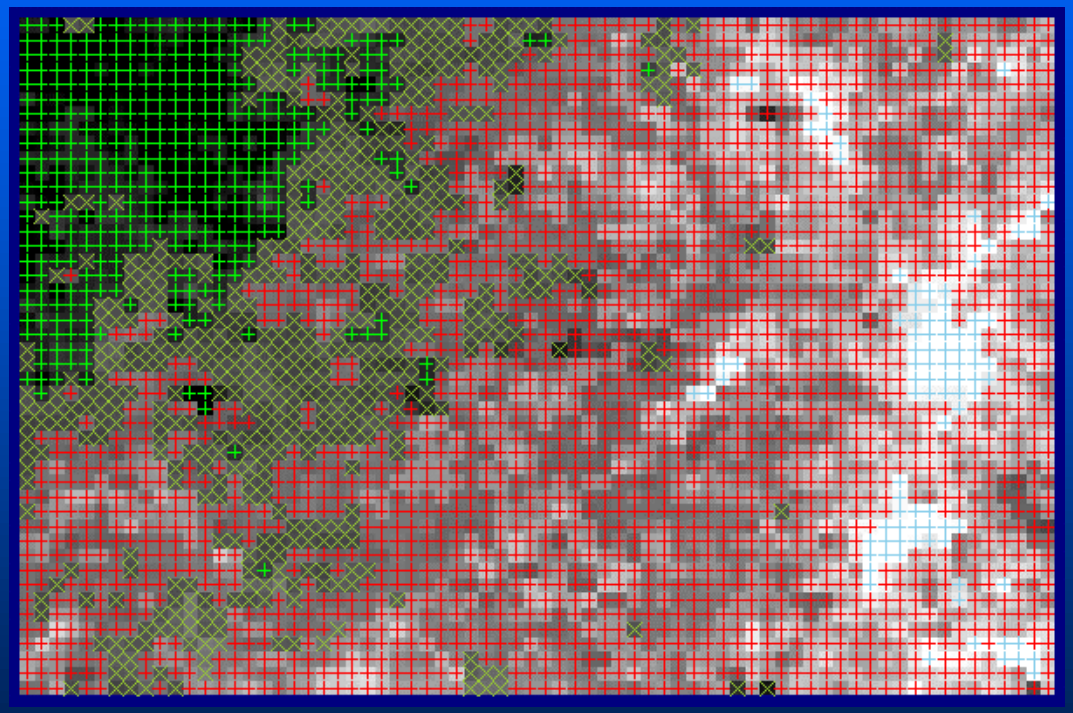


August

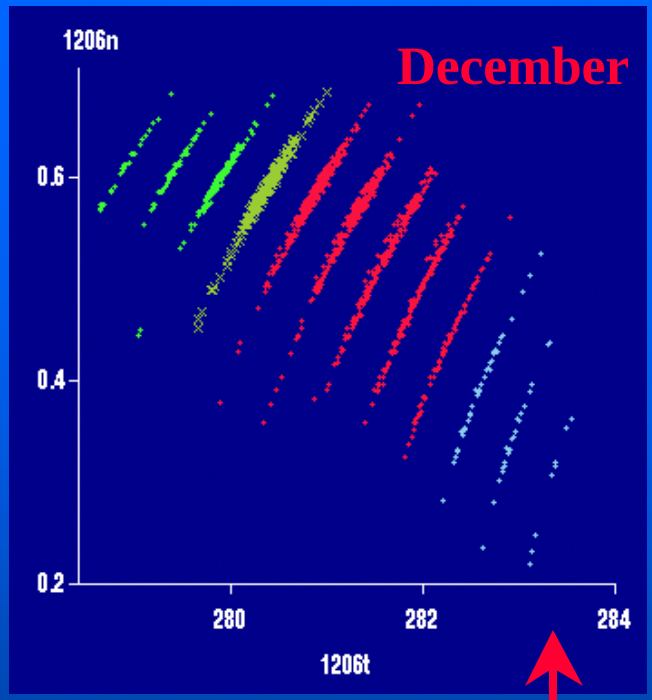


December

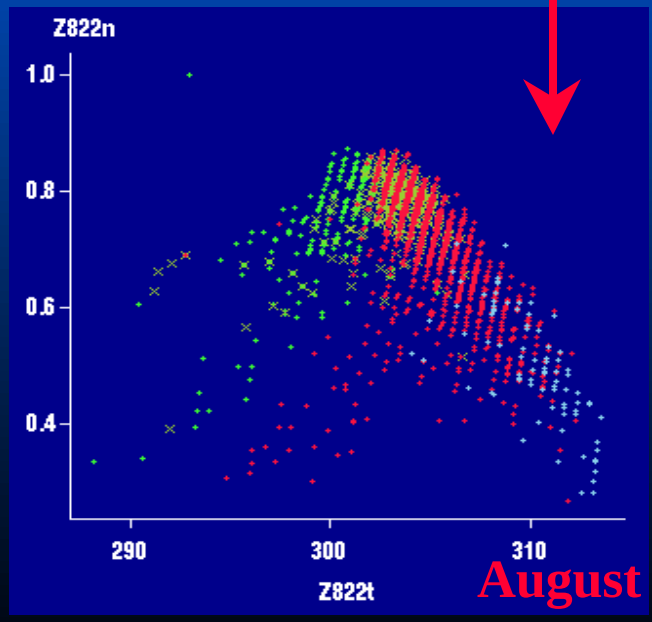
Final Classification



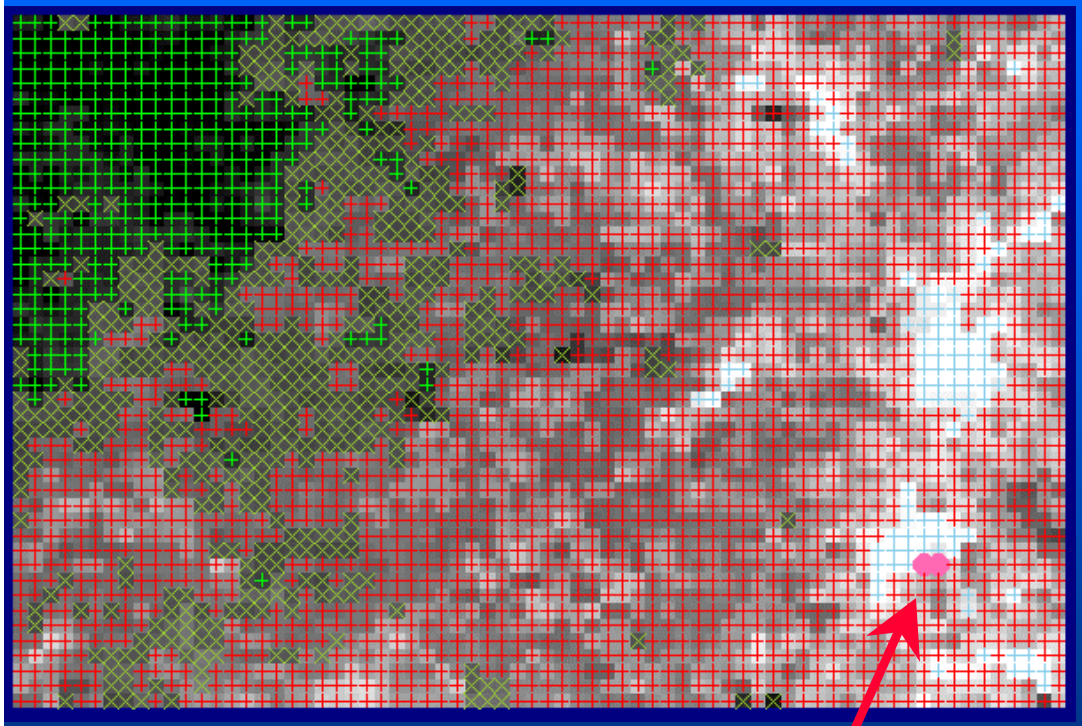
December



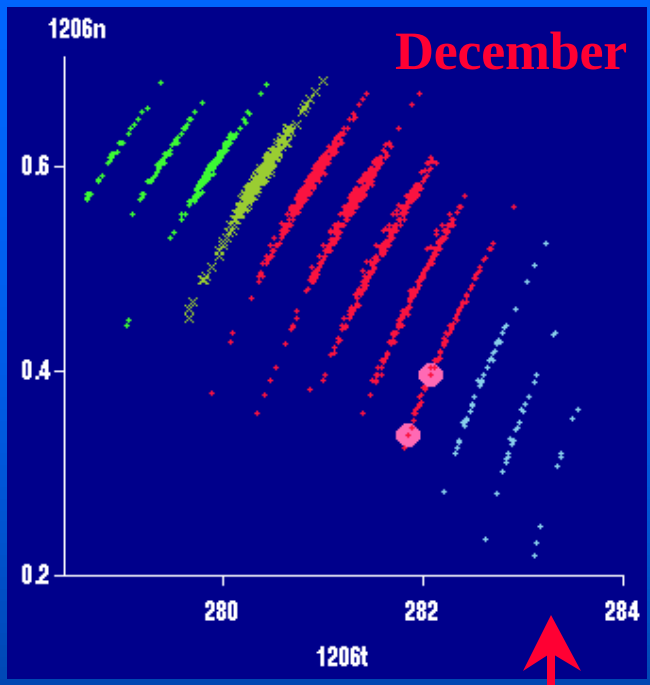
Linked



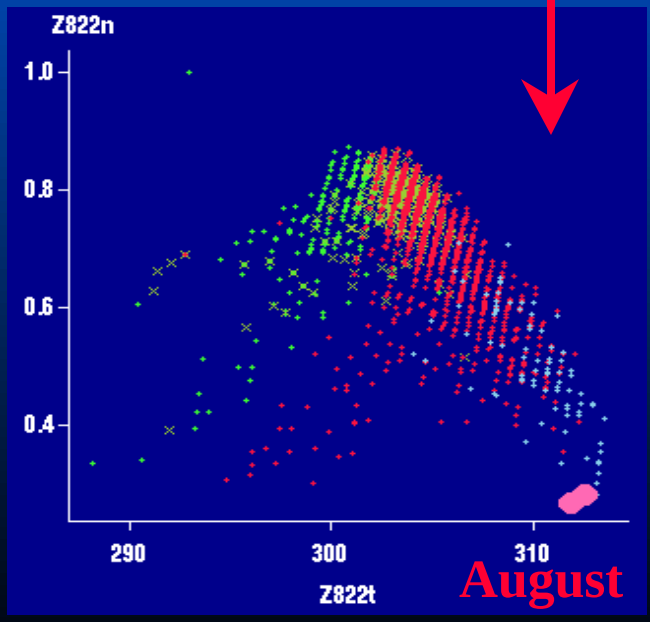
2 Pixels of Interest



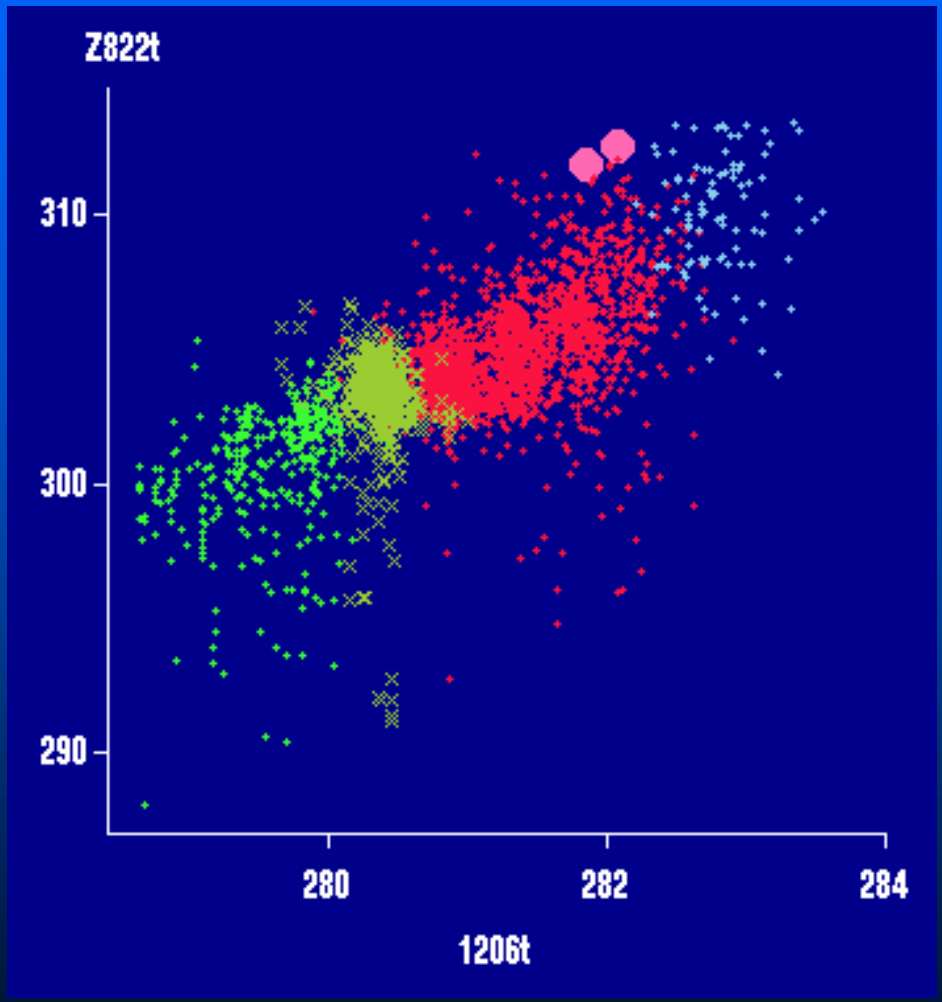
December



Linked



Correlation of Temperatures



Remote Sensing - Conclusions

- Visualization helps in classification of missing pixels.
- Visualization allows to detect unusual pixels.

Overall Conclusion

- Visual approach effective to see unexpected structure in data.
- Combination of different techniques most effective.
- Can be used for almost all types of data.

Questions ???