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:
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
Historical strategy is to fit parametric distributions and compare modern and archeological sands based on parameters.

- 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

- ORON SAYS CC DATA
  - RED
  - BLUE
    - MAGENTA
    - CYAN
    - YELLOW
  - GREEN
    - GREEN
    - GREEN
    - GREEN
Final Clustering
Scatterplots & Projection

| class     | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|-----------|------|------|-------|-------|-------|-----|-----|-----|-----|-------
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0
|          | 0.0  | 0.09 | 0.063 | 0.063 | 0.125 | 0.25 | 0.5 | 0.75 | 1.0 | 1.4+2.0

Scatterplots and Projection

- Scatterplots: Used to visualize the relationship between variables.
- Projection: A technique to reduce the dimensionality of the data.
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.
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:
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 - Horizontal

Hand-x

Hand-y

Hand-z

Forearm-x

Forearm-y

Forearm-z

Bicepex-x

Bicepex-y

Bicepex-z
Density Plots (ExplorN)
Scatterplots and Rotation (XGobi)
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:
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

Simplified Model of Pyramidal Brain Cells

Recording of Current (Soma)
Injection of Current (Soma)
Recording of Current (Apical Dendrite)

Basal Dendrite
Soma (Cell Body)
Apical Dendrite

Current Measurements
Simulated Physiological Response under 0.7 nA

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)
Smaller cells tend to be more excitable and have higher firing rates.
Visual Data Mining Using XGobi

- current
  - 2.0: ++++++++...+
  - 1.5: ++++++++...
  - 1.0: ++++++++...
  - 0.5: ++++++++...
  - 0.0: ++++++++...
- current
  - 2.0: ++++++++...+
  - 1.5: ++++++++...
  - 1.0: ++++++++...
  - 0.5: ++++++++...
  - 0.0: ++++++++...

- burst_rate
  - 6.0: X
  - 4.0: X
  - 2.0: X
  - 0.0: X

- plateau_rate
  - 3.0: 
  - 2.0: 
  - 1.0: 
  - 0.0: 

- spike_rate
  - 400: +
  - 200: +
  - 0.0: +

- Tools | Display | Info
- File | View: XYPlot | Tools | Display
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:
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: 
  \[ NDVI = \frac{Band2 - Band1}{Band2 + Band1} \]

- \( NDVI \approx 0.8 \) for Highly Vegetated Surfaces
- \( NDVI \approx 0.1 \) for Bare Soil
- Surface Radiant Temperature \( T_0 \): Band 4
- Surface Moisture Availability \( M_0 \)
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), $\hat{o} = 0.2$, $\Box = 0.4$, $\Delta = 0.6$, $\nabla = 0.8$, and $\times = 1.0$ (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

August

December
December

The City

August
Clouds and Forest in August

August
Reclassifying Clouds

August

Linked

December

August

December
Final Classification

December

December

Linked

August
Correlation of Temperatures
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 ???