Visualizing 2D Probability Distributions from EOS Satellite Image-Derived Data Sets: A Case Study

David Kao[1], Jennifer L. Dungan[1], and Alex Pang[2]
[1]NASA Ames Research Center
[2]Computer Science Department, UCSC

Abstract

Maps of biophysical and geophysical variables using Earth Observing System (EOS) satellite image data are an important component of Earth science. These maps have a single value derived at every grid cell and standard techniques are used to visualize them. Current tools fall short, however, when it is necessary to describe a *distribution* of values at each grid cell. Distributions may represent a frequency of occurrence over time, frequency of occurrence from multiple runs of an ensemble forecast or possible values from an uncertainty model. We identify these "distribution data sets" and present a case study to visualize such 2D distributions. Distribution data sets are different from multivariate data sets in the sense that the values are for a single variable instead of multiple variables. Data for this case study consists of multiple realizations of percent forest cover, generated using a geostatistical technique that combines ground measurements and satellite imagery to model uncertainty about forest cover. We present two general approaches for analyzing and visualizing such data sets. The first is a pixel-wise analysis of the probability density functions for the 2D image while the second is an analysis of features identified within the image. Such pixel-wise and feature-wise views will give Earth scientists a more complete understanding of distribution data sets. See www.cse.ucsc.edu/research/avis/nasa_js for additional information.

Key Words and Phrases: uncertainty, probability density function, geostatistics, conditional simulation, data assimilation.

1 INTRODUCTION

To advance scientific understanding of the entire Earth as a system, NASA's Earth Observing System (EOS) is providing regular, synoptic observations from low Earth orbit for a minimum of fifteen years. These observations are converted into biophysical and geophysical variables, for example surface reflectance, snow cover, temperature, and fraction of absorbed photosynthetically active radiation, using complex model-based algorithms. An important type of EOS product consists of maps of biophysical variables giving a "snapshot" of the state of a region at a given point in time. Increasingly, the need exists to comprehend a new dimension from these Earth science data and models. This dimension can be considered "probability space," connected with the frequency of occurrence of a biophysical value over time or a probability model, such as that representing uncertainty about the map, developed for a particular investigation. That is, for many investigations, there exists a probability density function (pdf) for every grid cell in the map representing a region or for every region over time (or both). We call these distribution data sets. In this paper, we examine 2D distribution data sets where we have a pdf at each pixel.

Distribution data sets are those that contain a set of values that can be represented as a distribution at each pixel. Data sets of this type arise in a number of Earth science applications and situations. Some examples include: distributions at each sample location in the 2D field based on a model with uncertainty components, distributions at each sample location representing the frequency of values occuring over a period of time and distributions of data from ensemble model runs and/or "fused" measurements from multiple satellite sensors [14]. Some EOS data products have formally defined uncertainty metrics [9, 10]. Uncertainty may be modeled as standard error, or a full pdf from a conditional simulation model [5, 6]. However, such uncertainty is rarely visualized in 2D.

Existing tools and packages that deal with image processing and geographic information systems (GIS) typically do not support distribution data sets. For example, GIS systems typically deal with static 2D data primarily as layers where users of these systems usually process one map at a time. What is needed is the ability to process all the distributions as a single set. Further, it is desirable to probe and query the set of distributions about the properties of features within the region, such as clumps with similar values of the biophysical variable. Distribution data sets are different from multivariate data sets in the sense that the values are for a single variable instead of multiple variables. This fact alone requires a re-examination of how popular multivariate analyses tools such as projection pursuit (PP) or principal component analyses (PCA) [11] can be utilized, if at all.

This case study focuses on how to visualize 2D distribution data sets generated through conditional simulation. The remainder of this paper includes a brief description of two distribution data sets followed by our progress to date in visualizing both pixel-wise and feature-wise summaries of these spatial distributions.

2 DATA SETS

A case study was made on two 2D distribution data sets created using conditional simulation. Conditional simulation, also called stochastic interpolation, is one way to model uncertainty about predicted values in a spatial field [12, 13]. It is a process by which spatially consistent Monte Carlo simulations are constructed given some data and assumptions [2]. Conditional simulation algorithms yield not one, but several maps, each of which is an equally likely outcome from the algorithm. Each equally likely map, called a *realization*, has certain properties. The properties are that

1. the values at sampled locations (locations where ground-truth exist) are identical to the sample values,
2. the frequency distribution of values from the map is the same as that deemed realistic from the sample data and
3. the spatial pattern of the values has the same spatial autocorrelation function as that deemed realistic from the sample data.

Taken jointly, these realizations describe the uncertainty space about the map. Any number of realizations may be generated by a particular conditional simulation algorithm; Chiles and Delfiner

[2] recommend at least 100. Conditional simulation was first applied in a remote sensing/Earth observation context by Dungan et al. [8]. To some extent, the application of this method has been hindered by limited and cumbersome visualization tools.

The case study data are from a synthetic example constructed using a small region in the Netherlands imaged by the Landsat Thematic Mapper [6]. Imagine that the biophysical variable to be mapped across this region is percent forest cover. Say there are ground-based measurements of forest cover from 150 well-distributed locations throughout this region as well as space-based measurements from Landsat of a spectral vegetation index. This spectral vegetation index is related to forest cover in a linear fashion but with significant unexplained variance. Further assume that the ground area represented by a field measurement is equal to the area represented by one pixel. Two distribution data sets were generated using this information: `Case1`, generated using a conditional simulation algorithm [4, page 170] taking into account ground measurements only; and `Case2` generated using conditional co-simulation [4, page 124] using both ground measurements and the coincident satellite image. Each data set consists of 101 $\times$ 101 pixels and 250 realizations. Values range from 0 to 255, rescaled from % cover.

The next section describes two approaches for visualizing `Case1` and `Case2`: on a pixel-by-pixel basis and on a per-feature basis.

3 PIXEL-WISE SUMMARIES

The problem here can be stated as follows: given a 2D distribution data set $F_r(i,j)$, where $i = 1,\ldots,N$, $j = 1,\ldots,M$, $r = 1,\ldots,R$, and R is the number of realizations of the $N \times M$ image, analyze the probability density function at each pixel. A simple approach would be to plot the histogram at each pixel. However, displaying a distribution at each pixel is not tenable when there are more than a very small number of image pixels. Or one could treat the data as a 3D volume $V(i,j,r) = F_r(i,j)$, where realization is treated as the third dimension. Then one can interrogate the volume data with existing visualization techniques such as cutting planes, isosurfaces, or even volume rendering. Isosurface representations (Figure 1) of `Case1` and `Case2` show dramatic structural differences between the two pairs of distribution data sets. These differences indicate that the addition of image data into the conditional simulation in `Case2` changed the values, though it is difficult to summarize exactly how. Also, the isosurfaces do not summarize uncertainty effectively. Straight application of isosurfaces as well as other techniques (e.g. volume rendering, non-orthogonal slices, and attempts at smoothing out the data) would somehow use the order of the realizations. But in the case of conditionally simulated realizations, order is not relevant, and smoothing is not desirable.

The central tendencies of the distributions from conditional simulation are often mapped to show what the most likely or common values are. The spread of the distributions are the most obvious way to summarize uncertainty. That is, the wider the distribution, the greater the range of possible values and therefore the less the certainty about that location. The best statistic to summarize a central tendency or spread depends on whether the distributions are parametric or non-parametric.

Parametric distributions can be fully described by one or more parameters of a known analytical function (such as Gaussian, Poisson, Beta, etc.). Mean and variance statistics are sufficient to fully describe Gaussian and Poisson distributions. Non-parametric distributions cannot be completely summarized by function parameters, but summary statistics still give useful information. Quantiles (such as the median) and quantile differences (such as the inter-quartile range) are good summaries of the central tendency and spread of non-parametric distributions while kurtosis describes the "flatness" of the distribution and skewness describes its asymmetry. Positive skewness signifies an asymmetric tail extending toward positive x. Since it may not be known in advance whether the distributions are best described as parametric or non-parametric, flexible tools are needed to summarize data sets.

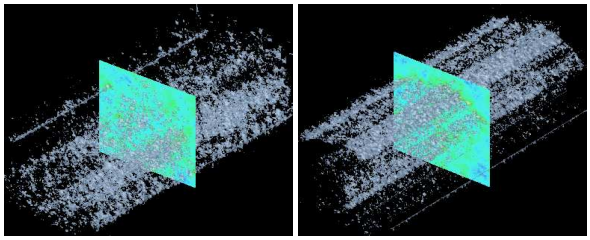
We calculated these standard statistical measures – mean, median, standard deviation, interquartile range, kurtosis and skewness – at each pixel across the realizations. In particular, standard deviation or interquartile range can be used as uncertainty metrics. The image plane can be colored according to any of these statistical measures or metrics and viewed separately. Alternatively, they can be simultaneously displayed in the same viewing space so that the scientist can study relations among the measures.

Figure 2(a) represents four statistics for `Case2`. The bottom image plane is colored based on the mean, the upper plane is deformed by the standard deviation and colored by the interquartile range, and the heights of the vertical bars represent the absolute value of the difference between mean and median values (only values above 3 are drawn). For reference, the vertical bars are also colored by the mean field shown in the image plane. Five color bands were used for the figure; cyan denotes low values of forest cover and red denotes high forest cover. The flexible selection of thresholds for the vertical bars allow the detection of extremes by different criteria, which would be application-specific. In this case, the regions with the lowest and highest values of forest cover also appear to be the most uncertain, judging from the "hills" in the deformed plane and the arched ridge that runs from left to right near the top of the image.

Figure 2(b), also using `Case2`, depicts the median as the colored image plane (using the same color map as shown in Figure 2(a)), the kurtosis field as the deformed upper plane which is colored according to skewness and the absolute value of the median-mean difference as vertical bars. The color of the vertical bar matches the color of the median field shown in the image plane. Three colors were used to depict skewness of the distribution at each pixel. The green regions of the surface graph represent negative skewness while the red regions represent positive skewness. The yellow regions represent no skewness, which only occur in a few areas. As in 2(a), the extreme values come from unusual distributions – those that are most skewed.

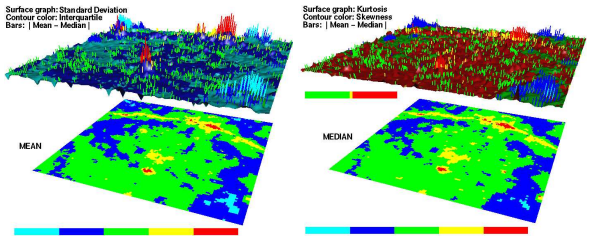
Another method that we have used to summarize the distribution data is to display a histogram cube. Conventionally, the scientist would plot the histogram at each pixel. In our approach, we generate a 3D histogram cube $H(i,j,b)$, where $b = 1,\ldots,B$, using the following steps. First, we specify B, the number of bins (i.e. the number of data value ranges) and then assign the data ranges of each bin based on the minimum and maximum values in the distribution. Next, we compute the histogram at each pixel and store the frequency count in each bin. Finally, we construct the cube by letting the first slice contain the frequency count of the first bin, and so on for the rest of the slices. Hence, the value at each pixel in a slice is the number of realizations with the same data range. We found that the histogram cube is very effective in visualizing the modality (i.e. the number of peaks) of the distributions. Figure 3 shows the horizontal cutting plane $H(i,50,b)$ and vertical cutting plane $H(60,j,b)$ of the histogram cubes for `Case1`, `Case2`, and a synthetic unimodal distribution data set. The latter is a data set generated by drawing randomly from a Gaussian distribution without any spatial autocorrelation. Both cutting planes are perpendicular to the realization plane $H(i,j,0)$ (the left facing plane) and show the frequency counts across all bins. In this figure, seven color bands were used and the colors vary from dark to bright in each band. Hence, variations of gray denote low frequency count while variations of red denote those with high frequency count. For this figure, 15 bins were used for the histogram (i.e. $B = 15$). `Case1` and `Case2` have narrower distributions and more spatially coherent patterns than does the random image. In addition, `Case1` has

a more spatially coherent pattern than does `Case2` because the use of satellite image data increased the spatial variation in the `Case2` simulation.



(a) (b)

Figure 1: Isosurface representations of two distribution data sets. (a) Isosurface at a value of 130 using one explanatory variable (`Case1`), and (b) isosurface at a value of 130 using two explanatory variables (`Case2`). The 250 realizations run along a diagonal axis from the lower left to the upper right corner. The pair of corresponding slices in the middle of (a) and (b) is one of the 250 realizations, added to help orient the data.



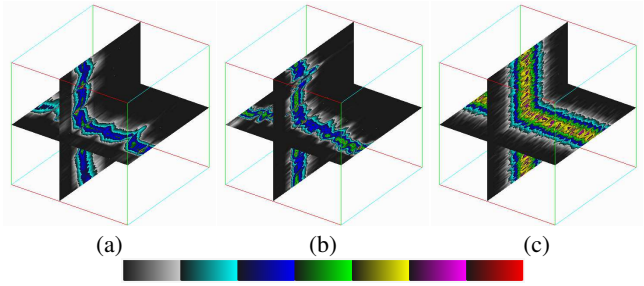
(a) (b)

Figure 2: Pixel-wise representations from the `Case2` data set. (a) The bottom plane is the mean field colored from non-forest (cyan) to closed forest (red). The upper plane is generated from three fields: the surface is deformed by the standard deviation field and colored by the interquartile range; and the heights of the vertical bars are from the absolute value of the difference between the mean and median fields colored according to the mean field on the lower plane. Only difference values exceeding 3 are displayed as bars to reduce clutter. (b) The bottom plane is the median field, while the upper plane is deformed by the kurtosis field, and colored by the skewness field (green denotes negative and red denotes positive). The deformed surface uses the colormap indicated in (b), while the vertical bars are colored by the median field below. The heights of the vertical bars are the absolute value of the difference between mean and median values as in (a). See color plates.

4 FEATURE-WISE SUMMARIES

Besides individual pixels, scientists are also interested in clumps of contiguous, similarly-valued pixels. These similar values may be recognized as a class, and spatial features representing this class are typically important. An example is percent forest cover identified by classification of spectral data in the visible and near infrared. Though percent cover is a continuous variable ranging from 0 to 100%, ranges of cover are often categorized into non-forest, open and closed forest classes. Scientists need to know the area covered by each class as well as the uncertainty about that area [7, 3]. The areas and shapes of individual forest patches or clumps are also of interest [1, 15].

To identify the class of interest, the scientist would first pick a pixel location p that has a value range within the class, i.e. 0-10% (for non-forest) or 80-100% (for closed forest). Then, the area of the clump surrounding pixel p, the contiguous region of pixels of the same class, can be determined by recursively searching and counting the neighboring pixels within the same value range as



(a) (b) (c)

Figure 3: Histogram cube of three distribution data sets: `Case1`, `Case2`, and a synthetic unimodal distribution. While the dark region in each color band may potentially be ambiguous, it is not the case based on the context of where they appear in the visualization. Experiments with other similar colormaps did not produce as much detail as this one. See color plates.

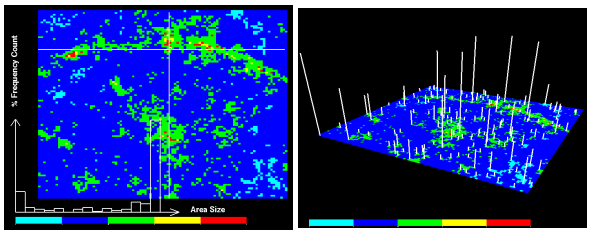
pixel p. The clump area calculation is first performed in the current realization. Next, because the scientist is interested in the variability of this clump, i.e. how much does the clump grow or shrink from one realization to another, the clump area around the chosen pixel is computed for the rest of the realizations, one at a time. It is possible that for some realizations, there is a different class at pixel p. In our approach we identify all the possible classes in the realizations and compute the clump areas of these classes for all realizations. Thus, for a given pixel location in any realization, we know the class that the pixel belongs to and its clump area. Next, we provide two interrogating tools for the scientists to analyze clump area statistics. We now discuss these tools and their usefulness.

We first provide a probe via a crosshair that allows scientists to interactively interrogate the clump area statistics. As the scientist moves the probe within the image of a single realization, our tool reports the current class and its clump area at the current probe position. In our implementation, we allow the scientist to move forward or backward to a desired realization and then move the probe within that realization. Our tool also reports the pixel-wise summary discussed in section 3. Thus, at the current probe location, the scientist can get the the mean, standard deviation, interquartile, skewness and kurtosis values. In addition, we compute the histogram of the clump area at the current probe position. The histogram of clump area at the current probe allows the scientist to get uncertainty metrics, such as standard deviation or interquartile range, of that feature. Figure 4(a) shows the probe in a location and the histogram of area surrounding that location. The x axis is the area size and the y axis represents the frequency as a percentage, i.e. the number of realizations found with the same class at the probe location divided by the total number of realizations (250). The histogram shows that the area of this clump, a clump within the "green" class here (scaled % cover values ranging from 101 to 152) is either quite small or, more likely, moderately large. Rather than naively choosing a mean area for this clump, the histogram indicates that the area is more likely to be accurate with a higher value (as indicated by the tallest rectangle towards the positive end of the x axis).

Our second tool highlights clump size and their locations. First, we determine the number of clumps in the realization. Then for each clump, we plot a vertical line whose length is proportional to the clump area. Thus, we would see many lines originating from the clumps. Using the lines to represent clump area gives a helpful perspective on the relative clump areas in the realization and one where the eye is less likely to be fooled by complex or circuitous shapes. Traditionally, the scientist would rely on color-encoded realizations only to visually assess or compare the clump areas for each class in the realization. With many clumps in each realization, this is often a tedious task. In realization 10, there are 513 clumps in total with

5 classes denoted by the color bar. Many (307) of these clumps consist of only one pixel (unit area). If these single pixel clumps were represented as lines, this would distract the viewer from more significant clumps. Therefore, we allow the scientist to "shave-off" these undesirable lines by not plotting clump areas that have unit area size. Figure 4(b) shows the clump area line bars for realization 10 of the `Case2` data set. In the figure, all lines with a clump area that is less than two pixels are not drawn. Somewhat surprisingly, the largest clumps (besides the "background" clump shown in blue) are in the semi-contiguous arch near the top of the image. Because these clumps are long and thin, it would be difficult without the lines to see that they have significant size. We also attempted to use rectangle blocks to represent clump area instead of lines (not shown here). However, because there are often many clumps in one realization, the rectangle blocks can easily obscure other rectangle blocks in the realization making it difficult to convey the clump area information.

Another interesting problem is where to plot the line within each clump area. A simple approach would be to pick any arbitrary pixel in the clump. We use a more meaningful approach that plots the line at the pixel with the lowest pixel-wise variance in the clump. The scientist can choose to use some other specified statistical values instead of variance. Furthermore, our tool can depict a symbol on top of the vertical line to quickly identify all clumps above or below a given value. Our tool can also loop through all the realizations and show how the clump area lines change across the realizations. We propose that the vertical lines help to clarify how clump areas change across the realizations. During animation, it can be difficult to judge the relative areas of clumps, but the vertical lines, as they lengthen and shorten, give a useful sense of areas both across the realizations and across the mapped region.



(a) (b)

Figure 4: Two representations of realization 10 of the `Case2` data set. (a) Interactive histogram plot (above the color bar) of clump area for the current probe (crosshair cursor) location in the realization. (b) Clump area line bars in the realization. A total of 166 line bars are drawn. See color plates.

5 CONCLUSION

Our case study shows the value of two ways of visualizing distribution data sets generated using conditional simulation. Such techniques also may be useful for portraying distributions from a long time period, such as values of a biophysical or geophysical variable on a given date at a given location over a period of several years. Multiple realizations may also come from alternative runs of a deterministic process model with varying parameters. The visualization tools presented in this paper are designed to help the scientist analyze distribution data from the Earth Observation System and EOS will provide directions for future visualization research. Our future work plans include: more extensive user studies following initial user feedback during the design and development phase of the project, extension to larger images to test the extent to which pixel-wise and feature-wise techniques scale (current implementation running on a high-end PC can support data up to 1000 x 1000 with 200 realizations), the development of novel means of visualizing all of $F_r(i, j)$ at once, study of time-varying distribution data sets and the comparison of distribution data sets.

6 ACKNOWLEDGEMENTS

This work is supported in part by the NASA Intelligent Systems Program, LLNL Agreement No. B347879 under DOE Contract No. W-7405-ENG-48, and NSF ACI-9908881. We would like to thank Alison Luo and Newton Der for improvements to the PDFVIZ program as well as members of the Advanced Visualization and Interactive Systems laboratory at Santa Cruz for some initial visualization of distribution data sets. Liane Guild and Steve Klooster at Ames provided helpful comments on a draft.

References

[1] T. R. Allen and S. J. Walsh. Spatial and compositional pattern of alpine treeline, Glacier National Park, Montana. *Photogrammetric Engineering and Remote Sensing*, 62:1261–1268, 1996.

[2] J. Chiles and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999.

[3] S. de Bruin. Predicting the areal extent of land-cover types using classified imagery and geostatistics. *Remote Sensing of Environment*, 74:387–396, 2000.

[4] C.V. Deutsch and A. G. Journel. *GSLIB: Geostatistical Software Library*. Oxford University Press, New York, 1998.

[5] J. L. Dungan. Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing*, 19:267–285, 1998.

[6] J. L. Dungan. Conditional simulation: An alternative to estimation for achieving mapping objectives. In F. van der Meer A. Stein and B. Gorte, editors, *Spatial Statistics for Remote Sensing*, pages 135–152. Kluwer, Dordrecht, 1999.

[7] J. R. Dymond. How accurately do image classifiers estimate area? *International Journal of Remote Sensing*, 13:1735–1742, 1992.

[8] J. L. Dungan *et al.* Alternative approaches for mapping vegetation quantities using ground and image data. In W. Michener, J. Brunt, and S. Stafford, editors, *Environmental Information Management and Analysis: Ecosystem to Global Scales*, pages 237–261. Taylor & Francis, London, 1994.

[9] Y. Knyazikhin *et al.* Estimation of vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from atmosphere-corrected MISR data. *Journal of Geophysical Research*, 103:32239–32256, 1998.

[10] Y. Knyazikhin *et al.* Synergistic algorithm for estimating vegetation canopy leaf area index and fraction of absorbed photosynthetically active radiation from MODIS and MISR data. *Journal of Geophysical Research*, 103:32257–32275, 1998.

[11] A. Ifarraguerri and Chein-I Chang. Unsupervised hyperspectral image analysis with projection pursuit. *IEEE Transactions on Geoscience and Remote Sensing*, 38:2529–2538, 2000.

[12] A.G. Journel. Geostatistics for conditional simulation of ore bodies. *Economic Geology*, 69:527–545, 1974.

[13] A.G. Journel. Modeling uncertainty and spatial dependence: Stochastic imaging. *International Journal of Geographical Information Systems*, 10:517–522, 1996.

[14] C. Pohl and J. L. Van Generen. Multisensor image fusion in remote sensing: Concepts, methods, and applications. *International Journal of Remote Sensing*, 19:823–854, 1998.

[15] R. A. Zampella and R. G. Lathrop. Landscape changes in Atlantic white cedar *(Chamaecyparis thyoides)* wetlands of the New Jersey pinelands. *Landscape Ecology*, 12:397–408, 1997.

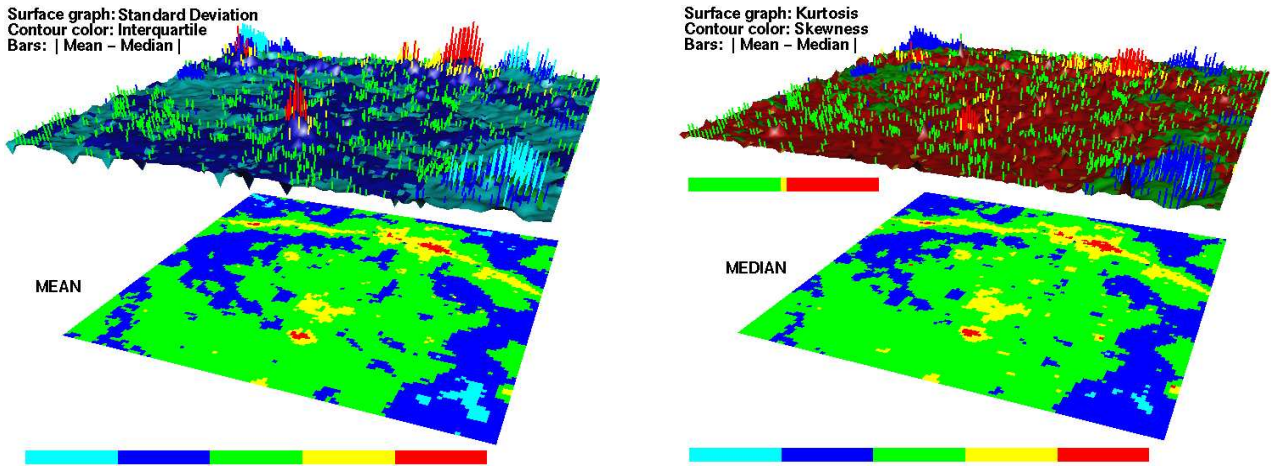


Figure 2:

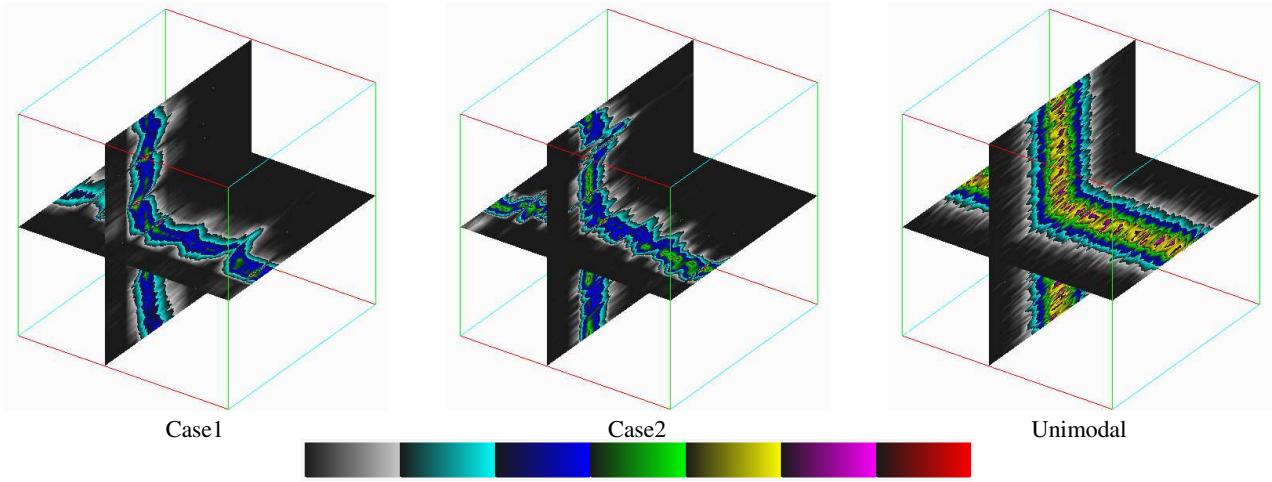


Figure 3:

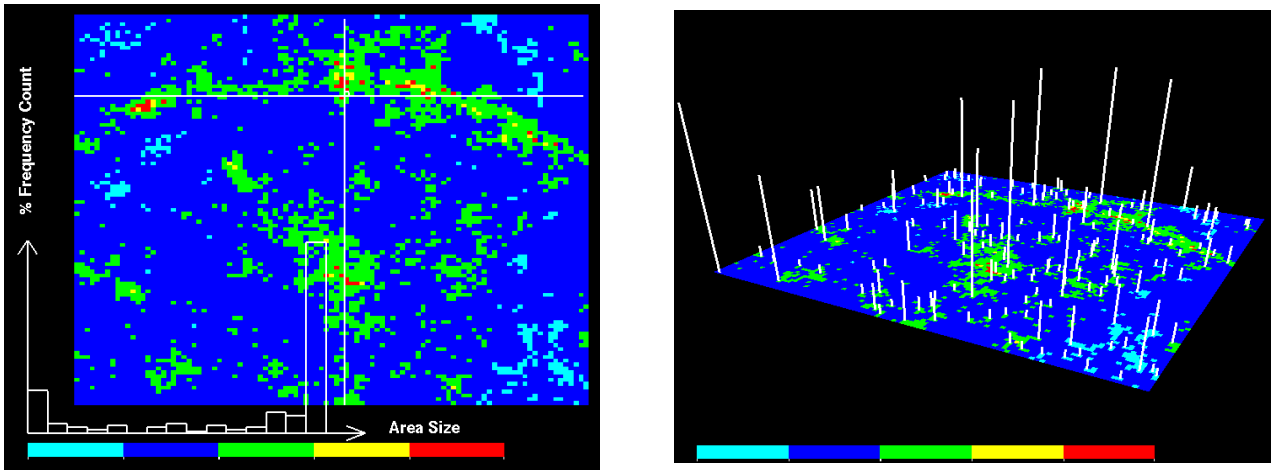


Figure 4: