588 126

Visualizing Spatially Varying Distribution Data

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

Abstract

Box plot is a compact representation that encodes the minimum, maximum, mean, median, and quartile information of a distribution. In practice, a single box plot is drawn for each variable of interest. With the advent of more accessible computing power, we are now facing the problem of visualizing data where there is a distribution at each 2D spatial location. Simply extending the box plot technique to distributions over 2D domain is not straightforward. One challenge is reducing the visual clutter if a box plot is drawn over each grid location in the 2D domain. This paper presents and discusses two general approaches, using parametric statistics and shape descriptors, to present 2D distribution data sets. Both approaches provide additional insights compared to the traditional box plot technique.

Key Words and Phrases: parametric statistics, shape description, uncertainty representation, probability density function.

1 INTRODUCTION

In the 1970's, John Tukey had a great influence on the visualization of data distributions by inventing the box plot [8]. The box plot was and remains an effective means to see how a set of data or a variable is distributed. Each box (also known as box and whisker) represents one distribution. It is a compact representation that encodes minimum, maximum, mean, median, and quartile information, summarizing what is essentially three dimensional information in two dimensions. From their origination in the statistical literature, box plots are now used in most scientific disciplines and are widely available in statistical software packages.

Geographical problems involve variables situated in space (2D). Modern mapping activities sometimes involve predicting the value of one or more variables for every spatial unit. For each spatial unit, a complete probabilistic statement about the variable at that location is desired. If there exists a probability distribution for each unit in the map, there are four dimensions of information. These are the spatial dimensions (x and y), the dimension of the variable being mapped and finally the probability dimension (the probability density is equal to or greater than 0). Can the insight used to come up with the box plot be extended to this geographical problem to visualize spatially varying probability distributions?

Prominent algorithms for generating probabilistic statements about geographical phenomena are geostatistical conditional simulations [4, 6] and Monte Carlo methods with physics-based models [3]. Probabilistic statements may also be formed from sensitivity analyses on different model-input parameters or other statistical methods that characterize uncertainty. Whatever the source, visualization of 2D distribution data sets is a new challenge. Simply extending the box plot technique to distributions over 2D domain is not straightforward. One challenge is reducing the visual clutter if a box plot is drawn over each grid location in the 2D domain.

The main visualization requirements that are usually desired from such spatial distribution data sets include: (a) a sense of what the distributions look like over the field and (b) any special features of the distribution data that may not be immediately obvious without some feature extraction step. This paper addresses the first requirement. More specifically, just as pseudo-coloring a scalar field gives an overall impression of a scalar field, we want to provide a similar overall impression of a 2D distribution data. This can also be seen as an extension of box plots over a 2D spatial domain.

Existing tools such as those from image processing and geographic information systems (GIS) packages typically do not support distribution data sets. For example, GIS packages deal with static 2D data primarily as layers that are displayed one at a time or "stacked" one on top of another. What is needed is the ability to process all the distributions as a single set. Furthermore, it is desirable to probe and query the set of distributions about the properties of features within a region.

This paper presents a number of methods and analyses to help visualize 2D distribution data sets. In the next section, we present some methods that assume the distributions can be described well by a few statistical parameters. We then describe some basic density estimators that can be used to construct distribution data sets from raw data. Subsequent sections explore more involved methods that describe the shape of the distribution at each point, where a few statistical parameters fail to do the task. We present these techniques using 2D distribution data sets from two application areas.

2 APPLICATION DATA

In this paper, we work with two distribution data sets from Earth science, a terrestrial and an oceanographic data set. The first data set is from a synthetic example constructed using a small region in the Netherlands imaged by the Landsat Thematic Mapper [2]. 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. A distribution data set was generated using this information: conditional co-simulation [1, page 124] using both ground measurements and the coincident satellite image. The data set consists of 101×101 pixels and 250 realizations. Values range from 0 to 255, rescaled from % cover.

Our second data set is from an ocean model covering the Middle Atlantic Bight shelfbreak which is about 100 km wide and extends from Cape Hatteras to Canada. Both measurement data and ocean dynamics are combined to produce a 4D field that contains a time evolution of a 3D volume such as temperature and salinity. To dynamically evolve the physical uncertainty, an Error Subspace Statistical Estimation (ESSE) scheme [7] is employed. This scheme is based on a reduction of the evolving error statistics to their dominant components or subspace. To account for nonlinearities, they are represented by an ensemble of Monte-Carlo forecasts. Hence, numerous 4D forecasts are generated and collected into a 5D field.

For this paper, we extract the top layer of the 3D ocean volume, and only look at the Monte-Carlo forecasts of this 2D slice for a given instant in time. This gives us the raw data for a 2D distribution data set. The field value is for sound speed and is derived from the other physical field values. The dimension of this data is set 65 × 72 pixels with 80 values at each point.

3 PARAMETRIC APPROACH

The problem can be stated as follows: given a 2D distribution data set $f(i, j, t)$, where $i = 1, \ldots, N$, $j = 1, \ldots, M$, and t is a real number, (a) analyze the probability density function at each pixel (i, j) and (b) give an overall impression of the entire $f(i, j, t)$.

A first step toward addressing this problem is to assume all the pdfs are parametric so that all the distributions can be summarized using a concise set of statistical parameters. For example, the normal or Gaussian distribution can be completely described by two parameters, its mean and standard deviation, and it has a a symmetric bell-shape with roughly 67%, 95%, 99% of the population within 1, 2, and 3 standard deviations. It is then relatively straightforward to visualize the summary statistics. Kao *et al.* [5] calculated first, second and third order statistics for each distribution and visualized them on different layers. 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 1 shows four statistics for the satellite-image derived distribution data set. 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 distribution data set, 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.

Summary statistics included mean, median, standard deviation, interquartile range, kurtosis and skewness. However, this approach is limited because distributions often deviate from parametric shapes. Summary statistics are still useful for describing some but not all aspects of nonparametric distributions. Cases where parametric summaries are less informative occur where distributions have more than one mode. For example, one can easily construct a bimodal distribution that has the same mean and standard deviation as the normal distribution.

Visualizing parametric statistics of 2D distribution data sets is relatively easy to implement, easy to understand since the concepts are well known and can be incorporated in most GIS packages. In this paper, we go beyond these parametric descriptions to seek robust technique that allow the visualization of the larger class of nonparametric distributions.

4 DENSITY ESTIMATION

Sometimes distribution data comes from sets of possible values that have been generated from a model or simulation. Raw data of this type is of the form $z(i, j, s)$, where i and j index the pixel and s is

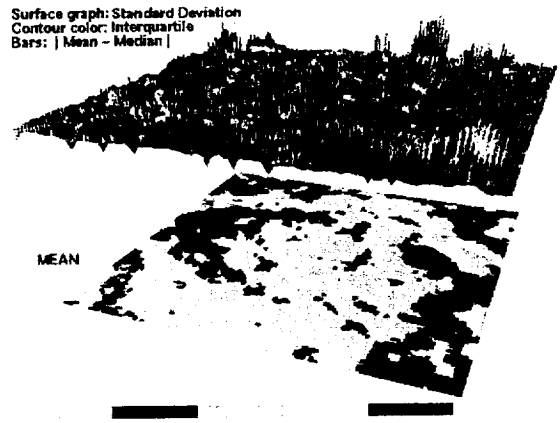


Figure 1: This figure illustrates how parametric statistics from the 2D distribution data set can be visualized in different layers. In the figure, four parametric statistics are visualized using the satellite-image derived distribution data set. The bottom plane is the mean field colored from non-forest (cyan) to dense 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.

a sample value (or realization) in the set of possible values. Various density estimators exist to create probability density functions (pdfs) from these sets of possible values. For example, the histogram is a very common density estimator, though it does not produce a mathematically valid pdf. More accurate methods include the naive estimator and the kernel estimator [9]. More formally, given a set of data values $\{z_i, i = 1, \ldots n\}$ density estimation is the construction of an estimate of the function $f(t)$ from $\{z_i\}$. Though the histogram is widely used, it is sometimes unsuitable for statistical analysis because of its C^0 continuous property and the fact that it is very sensitive to the bin width used. In this paper, we use a different class of estimators, the kernel estimators, which are C^1 continuous functions.

We first describe the naive estimator which resembles a histogram and provides a basis for understanding the kernel estimator. A naive estimator can be presented by

$$f(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w(\frac{t - z_i}{h}), \tag{1}$$

where $w(t) = 1/2$ if $|t| < 1$ otherwise, $w(t) = 0$. Like traditional histograms, naive estimators are C^0 continuous functions. They also depend on the choice of h, known as the smoothing parameter, just as histograms depend on the bin width.

Kernel estimators replace the function $w(t)$ in the naive estimator with a kernel function $K(t)$ that satisfies the following property:

$$\int_{-\infty}^{\infty} K(t)\, dt = 1 \tag{2}$$

If the kernel function $K(t)$ is C^1 continuous, then kernel estimators are also C^1 functions. Examples of kernel functions that satisfies the property above include:

Epanechnikov, for $|t| \leq \sqrt{5}$

$$K(t) = \frac{3}{4}(1 - \frac{1}{5}t^2)\sqrt{5} \quad \text{and} \tag{3}$$

Gaussian

$$K(t) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)t^2} \tag{4}$$

Kernel estimators are also influenced by the smoothing factor h. As the name implies, this parameter controls the overall smoothness of the estimator. As h decreases, the kernel estimator becomes more sensitive to slight variations in the distribution; as h increases, contributions from more neighboring points are coalesced to form a smoother kernel estimator. If h is not chosen appropriately, the shape of the estimator can vary significantly, and even change the modality of the pdf. e.g. from unimodal to bimodal

Rather than letting the user specify the value of h, a data dependent h can be derived [9]:

$$h = 0.9 \times min(std.\ deviation, interquartile\ range/1.34)\ n_s^{-\frac{1}{5}}. \tag{5}$$

For our given raw data $z(i, j, s)$ with n_s samples at each point, we calculate either a Gaussian or an Epanechnikov kernel estimate at each i, j location. We use Equation (5) to determine the smoothing factor h for each kernel. This approach of selecting the smoothing factor works well when we have a large number of kernel estimates to compute. These kernels can be evaluated at different values of t. If we evaluate it at k equally spaced t values, then $z(i, j, s)$ is transformed into an $N \times M \times k$ volume, V, where each voxel is a density estimate value. We refer to this 3D representation as the density estimate volume. Standard volume visualization techniques can then be applied to this density estimate volume.

5 VISUALIZING THE DENSITY ESTIMATE VOLUME

The 3D density estimate volume is a discrete sampling of the pdf at each pixel and will be the starting point for a number of visualization options presented here. All the figures in this section use the same data set as in Figure 1.

A simple and quick way to visualize the 2D distribution data is to allow the user to probe and interrogate the pdf at each point. Figure 2 shows three different representations of the data at a probe position identified by the cross-hair: (1) the raw data (the lower plot on the bottom), (2) the histogram of the pdf (the middle plot), and (3) the kernel density estimate of the pdf (the upper plot). These plots would change as the user moves the probe interactively. However, this approach can only depict a single pdf at a time and it may be difficult to determine how the distribution changes from neighboring pixels. This challenge can be tackled if we first treat the pdfs computed at all pixels as a volumetric data set and then use several existing volume visualization techniques on the density estimate volume.

5.1 Cutting Planes

The most straightforward method in volume visualization is to display cutting planes inside the density estimate volume V. Suppose $i = I$ for some constant I, then the cutting plane corresponding to $V(I, j, k)$ would show the pdf for pixels $(I, j = 1 \ldots M)$. Hence, from the cutting plane, we can easily see how the distribution changes along this row of pixels. Similarly, suppose $j = J$ for some constant J, then the cutting plane corresponding to $V(i, J, k)$

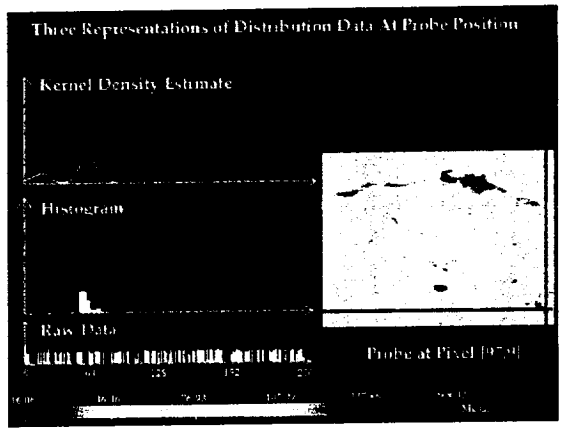


Figure 2: The image on the right shows the mean field of the 2D distribution data set. The three plots on the left show more information about the distribution at the cross-hair. The bottom plot is of the raw data values making up the distribution. Each line corresponds to one data point, and the line lengths are proportional to the data value. The middle plot is a histogram of those values. The top plot is a kernel estimate of those values.

would show the pdf for pixels $(i = 1 \ldots N, J)$. Hence, the cutting plane would show the variation of the distribution along this column of pixels. The cutting plane can be positioned anywhere in the volume. The most common choice is to position the plane right above the slice (cut). As the slice is swept across the field, the cutting plane would move together with the slice. Another possibility which we found useful for visualizing distribution data is to keep the cutting plane stationary on one of the faces of the density estimate volume. The natural choice is the farthest face from the current view that is parallel to the slice. This cutting plane appears as a wall on the volume where density estimates above the slice are displayed (see Figure 3). The height of the walls corresponds to the discretization granularity of the density estimators. For this data, there are 150 evaluations for each pdf $(k = 150)$.

Note that the axes of the density estimate volume are made up of two horizontal spatial axes and a third vertical sample value axis. Hence, one must exercise caution when interpreting cut planes of different orientations. With vertical cut planes such as those in Figure 3, the vertical axis corresponds to different sample values. With horizontal cut planes such as the one in Figure 5, we are at the density for a particular sample value over the entire 2D domain. Arbitrary cut planes are possible, but additional care must be used in their interpretation.

5.2 Local Surface Graphs

In studying 2D distribution data, one of the tasks is to visualize the modality of the distributions. Though cutting planes are fairly straightforward to understand, the modality of the distributions may not be depicted clearly. The presence of peaks in the distribution implies the presence of local maxima and minima. Using color mapped cutting planes, one would look for color streaks or narrow color bands to identify these peaks. Figure 3 shows a narrow color band that horizontally runs across the cutting planes and, at some locations, there is a shorter and fainter streak below the main color band. This indicates that the distribution is bimodal at these locations. Where there is only one color band, the distributions at those locations are unimodal. Though color mapped cutting planes may

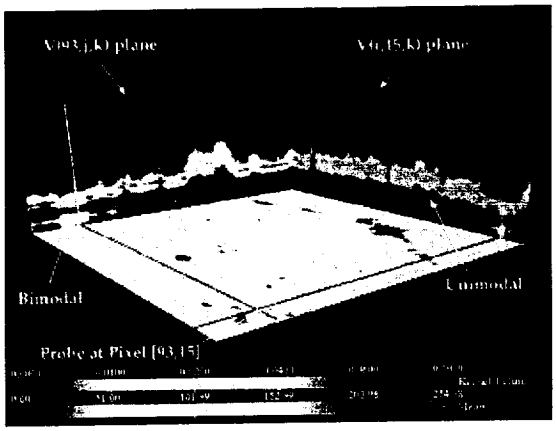


Figure 3: Cutting planes. The left wall shows the pdfs along the slice indicated by the blue line while the right wall shows the pdfs along the slice indicated by the red line. The bottom plane is the mean of the pdfs. The height of the walls is 150 which corresponds to the number of function evaluations for each pixel on the bottom plane. One can see that the distribution is mostly unimodal as indicated by the narrow color band that runs horizontally across both cutting planes. Also note that some of the pixels on the blue slice (left side of left wall) and also some pixels on the red slice (right side of right wall) appear to be bimodal. For example, the point under the cross-hair appears to be bimodal with a smaller peak near the bottom plane because the pdf for this point (along the white vertical line on the two walls) intersect two distinct non-blue regions.

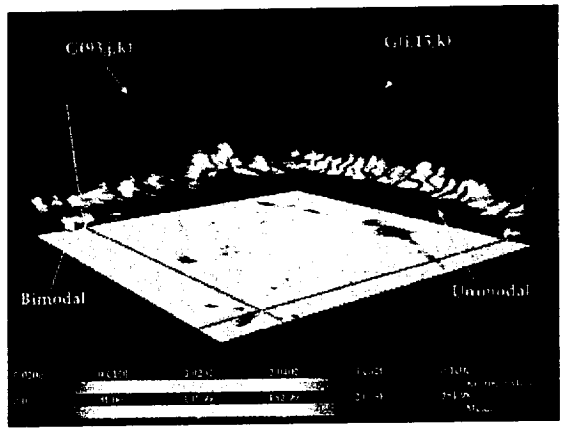


Figure 4: Local surface graphs. The surface graphs $G(93, j, k)$ and $G(i, 15, k)$ for the same cutting planes shown in Figure 3 are displayed. The surface graphs are colored using the density estimate. Note that the peaks (shown as ridges) are now much easier to see. The vertical curves drawn on top of the left and right surface graphs are the actual plots of the density estimate at the probe position (93, 15).

be useful to highlight the modality of the distribution, the technique is sensitive to the color map used. Depending on the distribution, some choices of color map may not be able to reveal the bimodal part of the distributions. An alternative approach is to construct a local surface graph so that the the modality of the distributions will be easier to see.

The surface graph is basically a displacement map. For a given cutting plane defined by $V(I, j, k)$, the surface graph $G(I, j, k)$ can be constructed by protruding points on the cutting plane by an amount proportional to the density estimate at that point. Let $G(I, j, k) = (scalefactor \times V(I, j, k), j, k)$, where $scalefactor$ is the height scale factor of the surface graph. This is illustrated in Figure 4. If we let $G(I, j, k) = (0, j, k)$, then the surface graph would be flat and it reverts back to the cutting plane defined by $V(I, j, k)$. Overall, the surface graph offers several advantages over the cutting plane technique: (1) it provides an accurate depiction of the roughness of the distribution, (2) it gives a 3D look and feel of the distribution, and (3) it allows the changes of the distribution across the row/column to be seen easily.

So far, the cutting planes have been lined up across a row or column of the 2D distribution. If one applies the cutting plane on the 3rd dimension of the density estimate volume, then we get a view of density estimate volume for the same evaluation point K for all pixels. Figure 5 shows a surface graph for the cutting plane $V(i, j, 40)$. The magnitude of the density estimates for all points at $k = 40$ are displayed as heights of the surface graph. Note that the density estimate at the cross-hair is bimodal as shown in a graph of the estimate shown in left upper plot.

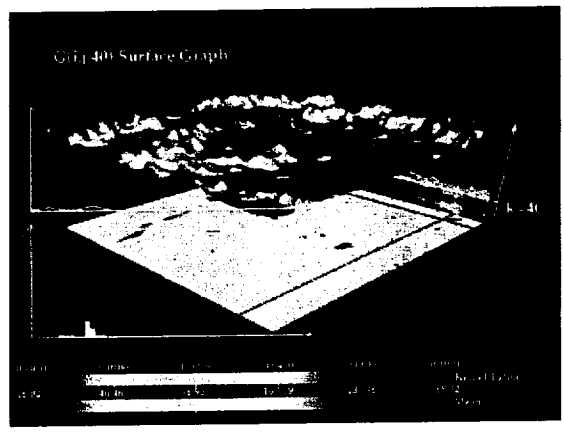


Figure 5: The cutting plane $V(i, j, 40)$ is depicted as a surface graph. The heights of the surface graph shows the relative magnitude of the density estimates for all the points at $k = 40$.

5.3 pdf Isosurfaces

Both the cutting plane and the local surface graph techniques allow the user to interrogate the density estimate volume interactively at some specified row, column, or slice. In addition to knowing how the distribution data changes along the selected row and column profiles, we are also interested in seeing the global attributes of the distribution data inside the estimate volume. These include the locations and the number of peaks. Suppose that the distribution data is unimodal and has roughly the same mean for all pixels in the data set. Suppose further that for some value C, $V(i,j,k_r) = C$ and $V(i,j,k_s) = C$ where $k_r < k_s$ are at the rising and descending part of the peak. Then, the isosurface defined by $V(i,j,k) = C$ consists of two surface layers. The lower layer corresponds to $V(i,j,k_r)$ (the rising part of the peak) and the upper layer corresponds to $V(i,j,k_s)$ (the descending part of the peak). The thickness of the isosurface (i.e. the distance between the two layers) corresponds to the width of the peak (see Figure 6).

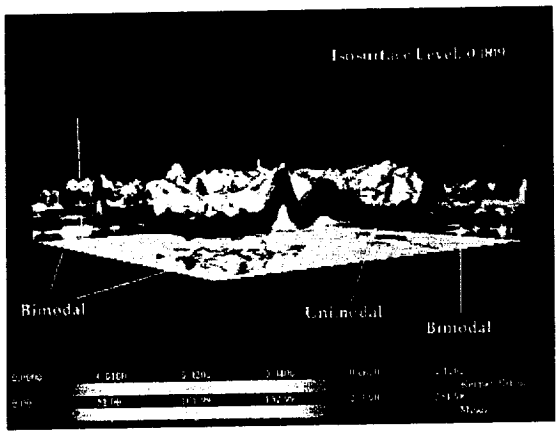
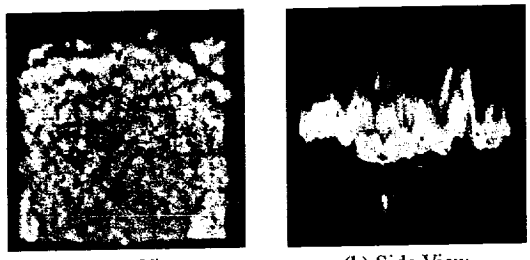


Figure 6: The isosurface defined by $V(i,j,k) = 0.009$ for the density estimate volume. Notice that there are smaller isosurfaces beneath the main isosurface. These locations show where the distributions are bimodal. The thickness, separation, and size of these isosurfaces also provide an overall impression of the density estimate volume.

5.4 Direct Volume Rendering

The density estimate volume is a good candidate for direct volume rendering. The data are scalar and usually on a regular grid. If the pdfs of neighboring points are quite similar to each other, then there is a good chance that direct volume rendering can identify such clustering (e.g. see Figure 7). Conversely, if the pdfs are spatially uncorrelated, then the direct volume rendering will not be able to find significant structures and therefore lead to such conclusions about the density estimate volume. Using the same data set as the other techniques presented in this section, we found that this technique allows us to see the spatial arrangement of the peaks. The ability to experiment with different color and opacity mappings is critical in extracting such features. An intuitive mapping for opacity is in direct proportion to the density estimate. However, a discontinuous mapping may also prove more useful in highlighting a range of values of the density estimate volume.

Although volume visualization techniques allow one to interrogate and analyze the modality of the density estimate volume, we are also interested in the shape description of the pdfs. In the next section, we propose to characterize a pdf by its roughness. We first describe the roughness parameters and introduce a peak hunting algorithm, then we propose two visualization methods for displaying the roughness parameters.



(a) Top View (b) Side View

Figure 7: Direct volume rendering of the density estimate volume. The transfer function attempts to map density estimates in a consistent fashion as other images in this section – blue for low values and magenta for high values. From the top view (a), we can see how the distribution changes within the image. The prominent arch-like feature located near the top image shows that the distributions are quite similar. The pdfs here tend to have a higher variance, longer tails, and hence lower density estimates. From the side view shown in (b), the bright blue spots underneath the main layer show locations where the distribution is bimodal.

6 ROUGHNESS DESCRIPTION

Given a pdf obtained from a density estimator, our goal is to characterize the pdf with a concise set of shape descriptors that can be mapped and presented visually. Towards this end, we define a roughness parameter for a pdf. The distribution may be very bumpy, in other words it has many local maxima. In this case we say it has a high roughness value. If the distribution has few local maxima, the roughness is low. We quantify roughness as the number of peaks in the distribution.

We define two kinds of peaks, basic peaks and concatenated peaks. A basic peak is an interval [a,b] such that density f is concave over [a,b], but not over any larger interval [9]. A concatenated peak includes at least two basic peaks. We further classify both kinds of peaks into two types, type A and type B (See Figure 8). Type A peaks have the minimum density at the start of the interval and type B peaks have the minimum density at the end of the interval. If the start and end of the interval have the same density, then the peaks are classified as type A. We propose the following procedure for finding significant peaks in a distribution: First, identify the basic peaks in the distribution. Next, we classify and combine these basic peaks into larger peaks. Finally, we count and record the locations and heights of these large peaks.

The concatenation rules between the two types of peaks proceed as follows,

$$concatenate(A, A) = A \tag{6}$$

$$concatenate(A, B) = \begin{cases} A & \text{if the start of A} \leq \text{the end of B} \\ B & \text{otherwise} \end{cases} \tag{7}$$

$$concatenate(B, B) = B \tag{8}$$

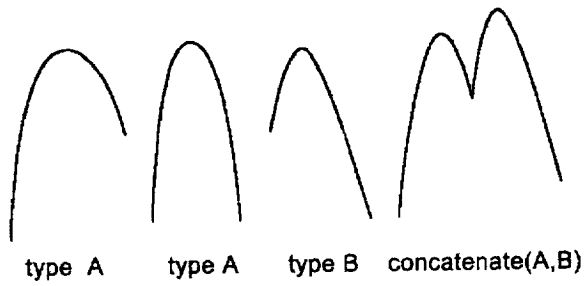


Figure 8: Basic and concatenated type A and type B peaks.

The concatenate(B, A) produces no new peaks. These operations apply to both basic and concatenated peaks. The peaks to be concatenated must be adjacent to each other.

Algorithm: *Peak Hunting Algorithm*

Find all the basic peaks $pk_1, pk_2, ..., pk_n$, n is the number of basic peaks
for All pairs of consecutive peaks (both basic and concatenated) **do**
  if they are both above the threshold **then**
    do not concatenate them
  else
    if they are both type A peaks **then**
      merge them into a concatenated peak of type A
    end if
    if the first one is type A and the second one is type B **then**
      merge them into a concatenated peak as defined in Equation 7
    end if
    if they are both type B peaks **then**
      merge them into a concatenated peak of type B
    end if
  end if
end for

The height of a peak is defined as the distance between the local maxima and the higher of the left and right minima. We look at the magnitude of the height to determine if a peak is significant or not. First, we determine the maximum height among all the basic peaks in the pdf. This is used as a point of reference for testing the significance of a peak. Basic peaks are concatenated until no more concatenation can take place. Then the heights of all the resulting peaks are compared to the reference height. If the height of a peak is more than a percentage of the reference height, the threshold, then it is a significant peak. The percentage is introduced as a user specified threshold and can range from 0 to 1. The number of significant peaks is then used as a measure for the roughness of the distribution at the pixel.

7 VISUALIZING ROUGHNESS

We use the data set from ocean modeling to illustrate how the roughness parameter of the distribution data can be visualized. First, we convert the ocean data set into a density estimate volume using an Epanechnikov kernel estimator with data dependent smoothing parameter h and evaluated at 300 points. The resulting density estimate volume is 65 × 72 × 300. The number, location, height and width of the peaks are obtained by applying the peak hunting algorithm to the density estimate volumes.

We propose two visualization methods for the roughness parameters. First, we quantify the roughness as simply the number of peaks $count_{i,j}$.

$$R(i,j) = count_{i,j} \tag{9}$$

where $R(i,j)$ is the roughness at pixel (i,j). Each pixel (i,j) is colored by $R(i,j)$. R_{max} and R_{min} are the maximum and minimum of the roughness among all pixels. The color assigned to pixel (i,j) is linearly interpolated based on the color map shown in (Figure 9).

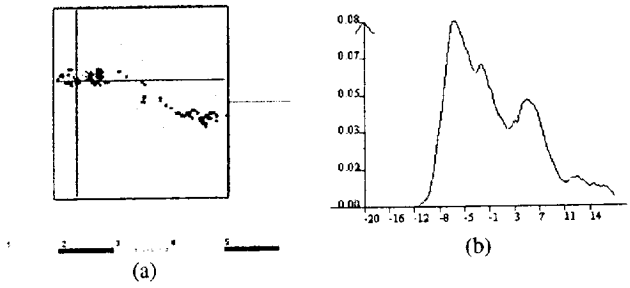


Figure 9: Roughness parameter using Epanechnikov kernel estimator with data-dependent smoothing factors. The number of peaks (in this case ranging from 1 to 5) are colored using distinct colors as indicated by the color map. We immediately note that most of the pdfs are unimodal except for those above the shelfbreak where there is more mixing and variability in the physical variables. The pdf for the pixel under the cross-hair is shown on the right. This pdf is counted as 2 peaks because the height of the small peak off the leftmost peak is below the threshold value of 36% of the maximum height among all the basic peaks.

The number of peaks alone is not a precise representation of the pdf's roughness. There are other factors that account for the shape description. For instance, given two pixels P and Q, both with one peak. The peak in the the first pdf may be narrow and tall while the peak in the second pdf may be wide and short. Both will however be assigned the same color using the visualization method described above. Therefore, we introduce a second visualization method which employs multiple parameters to show the roughness. Besides the number of peaks, we also show locations, heights and widths of the peaks.

The left and right ends of a peak give the location of the peak within a pdf. The height of a peak is the distance from that peak to the higher of the left and right ends of that peak. The interval between the two ends of a peak is the width of the peak. We combine this information using a line glyph representation to show the roughness of the density estimate volume. This is illustrated in Figure 10.

If the glyphs of all the pixels are shown together, it would be heavily clustered. Therefore, we separate the glyphs into different frames, each of which shows the line glyphs of pixels with the same number of peaks. In Figure 10 (a)-(e) we show the glyph visualization for pixels with one to five peaks respectively, and (f) is for all the pixels. From this figure, we can make the following observations:

1. The bluish line glyphs that are significantly longer e.g. in frames (a), (b) and (f), lie above pixels that are over the shelfbreak. The wider peaks (and multiple peaks) also implies higher standard deviation. This is consistent with the amount of variability expected in those regions.

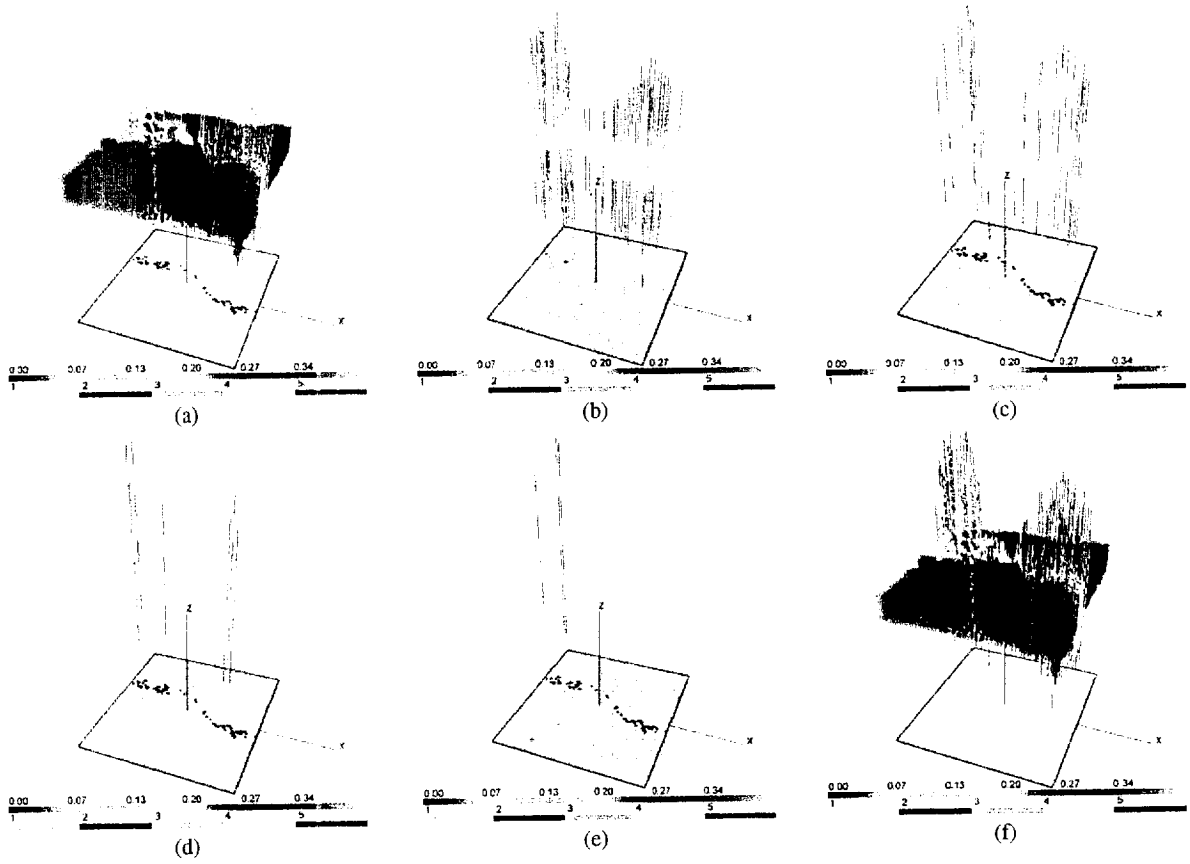


Figure 10: Glyph visualization of roughness . Each pixel has a one or more line glyphs showing the location and width of each peak. The lines are color coded by the height of the peak using the continuous color map below each frame. The bottom plane is similar to Figure 9 and is colored by the discrete color map on the bottom. The white areas on the bottom plane corresponds to pixels with line glyphs being drawn above it. Frames (a) to (e) show those pixels with 1 to 5 peaks in their pdf respectively. Frame (f) shows all the detected peaks in one image. We can make the following observations from these images. Even in the busy frame (a), we can see that majority of the pdfs are unimodal, and that most of them (the reddish ones) have peaks with similar heights and locations, and a smaller percentage (the greenish ones) have peaks which are flatter and wider. Frames (b) to (e) illustrates that fewer and fewer points have multiple peaks in their pdfs, and that those peaks are distributed over a wider set of sample values, and by necessity relatively flatter.

2. Excluding those pixels above the shelfbreak region, we can observed from frames (a) or (f) that majority of the peaks are of similar widths and across the same locations on the pdf.

3. In general, the lines of shorter length have redder color, and longer lines have bluer color. The integral of any pdf is always one. Therefore, among the pixels with the same number of peaks, the taller peaks are narrower. This can be seen in the far edge of frames (a) and (f).

8 CONCLUSION

We have presented a number of methods that give an overall impression of 2D distribution data sets. We go beyond histograms for constructing such data sets by employing a kernel density estimator with a data-dependent smoothing parameter and representing the resulting data as a volume. The representation has allowed a more complete description of pdfs on a grid using reliable visualization techniques such as iso-surface and volume rendering. In addition, we have implemented nonparametric summaries of the pdfs that will be helpful for multimodal data.

There are a number of improvements that we are working on. The peak hunting algorithm can still be improved. Towards this end, we have looked at identifying significant peaks in the frequency domain using Fourier transforms. However, finding the appropriate frequencies to band limit the signal automatically across different pdfs is not immediately obvious. A more fundamental challenge is to define an algebra or a set of operations on distributions that will provide us a more formal method for carrying out comparisons of distribution fields and feature extractions from distribution fields. Finally, we plan to extend our work to time varying distributions and to higher dimensionality.

9 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 David Draper for discussion on statistical methods, Pierre Lermusiaux for providing the ocean data, and HP Laboratories for making Ultravis publicly available. We would like to thank Newton Der, Wei Shen, and Bing Zhang for help with programming and data preparation.

References

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

[2] 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.

[3] S. W. Franks and K. J. Beven. Bayesian estimation of uncertainty in land surface-atmosphere flux predictions. *Journal of Geophysical Research*, 102:23991–23999, 1997.

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

[5] David Kao, Jennifer Dungan, and Alex Pang. Visualizing 2D probability distributions from EOS satellite image-derived data sets: A case study. In *Proceedings of Visualization 01*, pages 457–460, 2001.

[6] P.C. Kyriakidis. Geostatistical models of uncertainty for spatial data. In Hunsaker, Friedl, and Goodchild, editors, *Spatial uncertainty in Ecology: Implications for Remote Sensing and GIS Applications*, pages 175–213. Springer-Verlag, 2001.

[7] P.F.J. Lermusiaux. Data assimilation via error subspace statistical estimation, Part II: Middle Atlantic Bight shelfbreak front simulations and ESSE validation. *Monthly Weather Review*, 127(7):1408–1432, 1999.

[8] R. McGill, J. Q. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[9] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.