Visualizing Distributions from Multi-Return Lidar Data to Understand Forest Structure

David L. Kao¹, Marc G. Kramer², Alison L. Love⁴, Jennifer L. Dungan² and Alex T. Pang³

¹ NASA Advanced Supercomputing Division, ² Earth Science Division, NASA Ames Research Center, ³ Computer Science Department, UCSC and ⁴ Adobe Systems, Inc.

Email: David.L.Kao@nasa.gov, kramerm@fsl.orst.edu, alove@adobe.com, Jennifer.L.Dungan@nasa.gov, pang@soe.UCSC.edu

Spatially distributed probability density functions (pdfs) are becoming more relevant to Earth scientists and ecologists because of stochastic models and new sensors that provide numerous realizations or data points per unit area. One source of these data is from multi-return airborne lidar, a type of laser that records multiple returns for each pulse of light sent towards the ground. Data from multi-return lidar is a vital tool in helping us understand the structure of forest canopies over large extents. This paper presents visualization tools to allow scientists to rapidly explore, interpret and discover characteristic distributions within the entire spatial field. The major contribution of this work is a paradigm shift which allows ecologists to think of and analyse their data in terms of full distributions, not just summary statistics. The tools allow scientists to depart from traditional parametric statistical analyses and to associate multimodal distribution characteristics to forest structures. Information on the modality and shape of distributions, previously ignored, can now be visualized as well. Examples are given using data from High Island, southeast Alaska.

Keywords: geovisualization, lidar, density estimation, mode finding, probability density functions, interactive visualization

1. INTRODUCTION

The development of remote sensing has issued in an era where large quantities of multidimensional and multivariate data are routinely analysed. The dimensions are spatial, spectral and sometimes temporal. Spectral data has also been considered multivariate, where each waveband represents a variable and coincident data on variables measured by some other means may also need to be concomitantly analyzed. In this paper we address a different aspect of some remote sensing data sets, their multi-valued nature. That is, we suggest that there are many situations where there exist multiple values of a single variable for each grid cell or spatial unit. Complete information on these multiple values is contained in the probability density function (pdf) of the variable at each location. These data sets are common products from remotely sensed image data and also from geophysical simulation.

Lidar (Light Detection And Ranging) is a remote sensing technology that yields multi-valued datasets. When lidar sensors are used to measure vegetated surfaces such as forests, they can yield a detailed understanding of the canopy structure across an entire study area rather than at a few select plots. Distribution data from raw multi-return lidar data from forests provides information on forest structure, tree size and density (Hyyppä et al., 2001; Yin and Chen, 1994). For example, forest plots recovering from natural disturbance tend to have unimodal distributions of stem sizes and canopy heights with low standard deviations, whereas older, less disturbed forest plots tend to have multimodal distributions (Kramer et al., 2001).

Previous work with lidar data has relied on statistical summaries that attempt to characterize the distribution with a small set of descriptors. The summaries reduce the dimensionality of the dataset and make visualization straightforward. This approach fails when the distributions are nonparametric, and especially when they are multimodal. In this paper, we propose an approach to visualizing lidar data that allows exploration, query and comparison of distributions. The approach should increase opportunities to query multi-valued data in new ways that better help us to understand the distributions of geophysical and ecological phenomena, both at single locations and across the spatial domain. Interpretation can be gained from field reconnaissance, expert knowledge or ancillary information.

DOI: 10.1179/000870405X57257
2. BACKGROUND

Historically, statistical descriptors such as the mean, median, standard deviation, skewness and kurtosis have been used to describe and compare the populations being sampled. While these descriptors work well for unimodal distributions, they often fail to capture the nature and dynamics of multimodal distributions. Yet multimodal distributions commonly occur in populations found in the natural environment. Multimodal distributions create a challenge for visualization because they are less amenable to statistical summaries and thereby render with a few variables or parameters. In a geospatial context, multimodal distributions provide an even greater problem. In this context, a distribution may exist at many locations in a spatial field, such as at every cell in a grid. If even some of the distributions are multimodal and cannot be summarized easily, the whole field cannot be rendered using conventional approaches.

The challenge to explicitly visualize spatial, multimodal distributions is due to the four dimensional nature of the problem. We need to consider the probability density functions (pdfs) over space — two dimensions are the orthogonal spatial dimensions, a third is the variable scale (in this case the height scale given by lidar) and the fourth is the frequency scale. Previously, we have reported on techniques for visualizing 4D spatial distribution data sets (Kao et al., 2001) using parametric statistics. That is, the pdf at every cell is characterized by a few statistical parameters such as mean, standard deviation, skewness, etc. and visualized. When some of the pdfs, particularly in mixed forest areas, have multimodal distributions, statistical summaries are not sufficient. To address this, we have also developed shape-based descriptors for distributions (Kao et al., 2002). The basic idea here is to describe the shape of a distribution using information about the number of modes, the location of the modes, the width and height of each mode, etc. These descriptive information are then mapped to visual parameters. We demonstrate how that approach can be brought to bear on the lidar data in Section 5.

More generally, we proposed an operator-based approach to visualizing spatial distribution data sets (Luo et al., 2003). The main idea here is to treat distributions as first-class objects with their own set of methods and operations. This allows one to compare, add, subtract, etc. two distributions. More complex operations can also be constructed from these simpler operations. The benefit of this approach is that standard visualization techniques such as contour lines, isosurfaces, streamlines, etc. can be extended to support the new data type.

Previous efforts to visualize lidar data (Eggleston et al., 2000; Watson et al., 2000) presented ways in which a user can navigate through forest lidar data sets within a virtual environment. This is essentially the creation of a digital elevation model of the canopy top. Unlike their approach, our approach looks at aggregated multiple lidar returns. Therefore the data at each cell location is actually a collection of height values. In this study, we visualized distributions from 0.1 hectare cells, the size of field plots for which forest stand measures exist. The techniques developed in (Kao et al., 2001; Kao et al., 2002; Luo et al., 2003) are brought to bear upon this problem.

3. DATA

Forest canopy height distribution data were collected using a multi-return lidar system. The system, the digital airborne topographic imaging system (DATIS-2; 3-Di Technology, MD, USA) is a small footprint lidar. The sensor is capable of retrieving multiple (up to five) returns of elevation and intensity for every shot as it passes through a forested canopy. Over wooded terrain, the first return measures forest canopy height, while the last return measures ground elevation. The laser fires at a rate exceeding 4000 pulses per second and scans across the aircraft flight path (see Figure 1). Since the speed of light is known, the reflection time of the laser light back to the aircraft is measured, allowing the distance to the terrain surface to be calculated. To locate the elevation points, the latitude and longitude of the aircraft are recorded with a high accuracy (< 1 cm) Global Positioning System (GPS). The accuracy of point locations is further increased by compensating for the aircraft’s attitude (pitch, roll and yaw), measured using an inertial measurement unit. DATIS-2 was flown in a Cessna 206 in May 2001. The data were initially collected at a density exceeding two shots per m². Raw data were processed into 81 measures of forest canopy height for each 0.1 hectare cell across the island, resulting in 1800 grid cells with a distribution for each cell. The data were collected above High Island (approximately 180 hectares), which is located in the middle of the Alexander Archipelago (Figure 2). The maximum elevation on the island is 150 m. The parent material is fractured basalt. Average annual precipitation is 1.9 m, with the wettest months during fall and winter. Extreme temperature fluctuations are infrequent due to the maritime influence. Cloud cover, precipitation, cool ambient air temperatures (4–10°C), and high relatively humidity (80%) are characteristic throughout the year. The island is dominated by productive western hemlock (Tsuga heterophylla (Raf.) Sarg.) with scattered Sitka spruce (Picea sitchensis (Bong.) Carr.).

4. ALGORITHMS

Prior to visualization, algorithms are applied to the raw data to estimate and characterize their distributions. In particular, density estimation is used to generate a probability density function from the 81 heights at each grid cell, a peak hunting algorithm is used to find all the modes in the pdfs and an operator is selected to allow distribution matching.

A. Density Estimation

For each grid cell in the field, there are multiple lidar returns, each with an associated height. These represent a sample of the full set of heights of all elements in the canopy. We use each sample to make an estimate of the “true” density, that is, the distribution of the full set of heights. One common density estimator, the histogram, does not produce a mathematically valid pdf and is very sensitive to the bin width used. There are many other estimators possible depending on the nature of the data (Silverman, 1986). In this application with lidar data, we selected a kernel estimator because it provides robust density estimation and is widely
Figure 1. Airborne lidar data acquisition

Figure 2. Vicinity map of High Island in Alaska where the lidar data were collected

Figure 2. Vicinity map of High Island in Alaska where the lidar data were collected
used. A kernel estimator is given by
\[
f(t) = \frac{1}{n} \sum_{i=0}^{n} \frac{1}{h} K \left( \frac{t-z_i}{h} \right)
\]  
(1)
where \( f(t) \) is the pdf estimated from the data values \( \{z_i, i=1,...,n\} \), \( b \) is a smoothing parameter and the kernel function \( K(t) \) satisfies the following property:
\[
\int_{-\infty}^{+\infty} K(t) dt = 1
\]  
(2)
If the kernel function \( K(t) \) is \( C^1 \) continuous, then kernel estimators are also \( C^1 \) functions; this is in contrast to histograms which are \( C^0 \) continuous. A function if \( C^k \) is for all \( r < k \), the \( ith \) derivative is defined. For this application, we used the Gaussian kernel function:
\[
K(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}
\]  
(3)
Kernel estimators are also influenced by \( b \) which controls the overall smoothness of the estimate. As \( b \) decreases, the kernel estimator becomes more sensitive to slight variations in the distribution; as \( b \) increases, contributions from more neighboring points are coalesced to form a smoother estimate. If \( b \) is not chosen appropriately, the shape of the estimate can vary significantly and even change the modality of the pdf (e.g. from unimodal to bimodal). So, rather than letting the user specify the value of \( b \), a data-dependent \( b \) can be derived (Silverman, 1986):
\[
b = 0.9 \times \text{Min} \left(\text{std.dev,interquartile range}/1.34\right) n^{2/5}
\]  
(4)
For the lidar data, we compute the kernel density estimate \( f_i(t) \) for each grid point \( (i, j) \) using the data-dependent smoothing factor given in Equation 4. Thus, the smoothing factor \( b \) varies across the field depending on the data values at each grid point.

B. Mode Finding

Given a distribution, there are a number of ways to characterize its modality, or how bumpy it is. For example, one may treat the distribution as a signal and apply Fourier analysis to extract the major frequency with the largest amplitude. However, one is still faced with the task of deciding what is a significant mode in frequency space. Alternatively, one may draw from data modeling and use Gaussian mixture models (Raja et al., 1998) to fit the data as a weighted combination of Gaussian distributions with different means and standard deviations. However, this approach requires significant a priori knowledge about the modality of the distribution in the first place. Instead, we use a descriptive approach that analyses the shape of the distribution. Yin and Chen (1994) suggested a way to identify modes in histograms of gray scale images. A hierarchical tree is associated with each histogram, which reveals the structure of the histogram. Peaks can be generated by traversing the tree with a selected threshold. Peaks with curve length higher than the threshold are considered significant. Their algorithm is efficient in terms of computational cost \( O(\log(n)) \), where \( n \) is the number of peaks. However, there can be improper peak concatenation in Ying and Chen’s algorithm.

Instead, we use the following mode finding algorithm that we proposed in Kao et al. (2002). Given a pdf obtained from a density estimator, our goal is to determine the number of peaks in the distribution and their respective positions. The distribution may be very bumpy; in other words it has many local maxima. First, we compute all the local maxima in the distribution. By our definition, a local maximum is an interval \([a,b]\) such that the density estimate is concave over \([a,b]\), but not over any larger interval (Silverman, 1986). We refer to the local maximum as a basic peak. The height of a peak is defined as the vertical distance between the local maximum and one of the two ends of the interval with higher density. A peak with a height higher than a user-defined threshold is considered significant. A concatenated peak, as its name implies, includes at least two basic peaks. We further classify both kinds of peaks into two types, type A and type B (See Figure 3). 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 (Figure 3, second case of type A). We use the following rules for concatenating the peaks.

These two types of basic peaks may be concatenated to form combined peaks according to the following rules:

\[
\text{concatenate}(A,A) = A
\]
\[
\text{concatenate}(A,B) = \begin{cases} B \text{ if the start of } A \text{ is } \leq \text{ the end of } B \\ A \text{ otherwise} \end{cases}
\]
\[
\text{concatenate}(B,B) = B
\]
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 and concatenation only takes place if at least one of the two adjacent peaks is not significant. Then we iteratively loop through all the basic peaks and determine if the adjacent peaks can be concatenated into a concatenated peak. The iterations stop when there is no further concatenation possible, that is, all pairs of adjacent peaks are considered as significant peaks. Finally, we count and record the locations and heights of these significant peaks.
C. Distribution Matching

There are a number of operators from different fields that can be applied as distance measures to compare distributions. The Euclidean distance is commonly used to compare distributions with spatial geometry properties. For example, if the collection of values is represented as a vector, then the Euclidean distance is an appropriate choice for comparing vectors. Absolute distance and correlation can also be used to compare vectors. Another distance measure is the Kullback-Leibler (KL) distance derived from information theory (Cover and Thomas, 1991; Kullback and Leibler, 1951) which represents the relative entropy between two probability distributions.

We used an operator that compares distributions in order to find ones that have shapes that match a distribution of interest. The operator returns a scalar value that indicates how similar the two distributions are. In this case study, the lidar data are univariate probability distributions. Although there is more than one operator that can be used for this purpose, we chose the KL operator (Cover and Thomas, 1991; Kullback and Leibler, 1951), because it specifically computes the distances between statistical distributions. Let $P$ and $Q$ be two distributions of variable $x$ to be compared. The Kullback-Leibler distance is defined as:

$$KL(P, Q) = \int_{-\infty}^{+\infty} P(x) \log \frac{P(x)}{Q(x)} \, dx$$

The greater the KL distance is, the less similar the two distributions are. We set a threshold to control how similar we want the search results to be to a distribution of interest or target distribution. All the distributions with a distance less than the threshold to the target distribution will be accepted as similar distributions to the target distribution. By relaxing the threshold, a larger number of similar distributions will be identified.

5. VISUALIZATION TOOLS

In the following subsections, we describe several visualization techniques designed to provide capabilities ranging from synoptic, general views of the full data set to more specific, localized and detailed query and display. The techniques allow extensions well beyond summary descriptors such as the quadratic mean, robust mean and chi-square or other ‘non-parametric’ summaries or clustering algorithms. Briefly, they are a map display of the number of modes for each distribution, an interactive data probe, mode exploration (finding locations that have distributions with modes specified by the user, finding distributions that have a specified number of modes and graphing distributions that have specific modes) and distribution exploration (matching complete distribution shapes). For each technique, we describe how it can assist the scientist in exploring distributions. We apply these tools to distributions of canopy height derived from the lidar data, focusing on three characteristic distributions that are of particular interest. Collectively, these tools can be used effectively to analyse distribution data. Figure 4 depicts relationships among the techniques, starting from the number of modes (that many different distributions may have in common), to querying a specific mode in a distribution (that a smaller subset will share), and finally to matching a distribution, that another, possibly even smaller subset of distributions will share. So, the queries work with increasing specificity.

It is expected that these visualization tools will provide scientists with new understanding of their data, previously not possible. The extraction features provided with many of these tools will allow creation of new data sets from specific distributions, or portions (e.g. modes) of distributions, that are of interest.

A. Displaying the Modality of the Distribution Data

The first look at the spatially explicit distribution data shows the number of modes for each cell, calculated using the algorithm described in Section 4B (Figure 5). This synoptic descriptor of all the data across the sample space provides the scientist with the first glimpse of new information related to the distribution. The modality of each cell gives some indication of patterns of multimodality. This is the first new information about the distribution that is not available through coarse statistical descriptors. This display helps answer questions such as what proportion of the data is unimodal or multimodal? Are the number of modes spatially clustered or concentrated in any one subregion of the field?

B. Interactive Data Probe

We have implemented an interactive data probe that allows the user to view the distribution of an individual cell at the current probe position set by the user. The interactive data probe is straightforward and useful for visualizing the pdf at any location in the field. It provides a per point basis query and shows the modality of the distribution. Only one density estimate is displayed at a given time. To begin gaining familiarity with the data the scientist can probe at different locations in order to have a good overall feel for the distributions in the study area. In addition, when viewing the distribution of a given cell, adjacent cells can be selected (through an up, down, left and right keyboard feature), thus allowing the scientist to visually traverse portions of the forest. This feature allows the scientist to view and relate distributions of particular forest regions of interest that he/she might already be familiar with through field reconnaissance or other ancillary data. Figure 6 shows the kernel density estimate and the histogram corresponding to the probe position shown on the right image of the figure. The image is colored by the mean canopy height.

C. Mode Exploration

The modes of a distribution can be explored in a variety of ways using mode exploration tools. For all of these tools, the mode is computed using the peak hunting algorithm as described in Section 4B. Meaningful mode exploration depends on using the proper density estimator, so that shapes in the distribution are real and not an artifact of low sampling density in the data. Conversely, it is important that all real information in the distribution be retained, so
the smoothing function must minimize the loss of real information contained in the distribution. Our mode exploration tools comprise the following processes: (1) mode query, (2) visualizing the distributions from the results of a mode query, and (3) visualizing the distributions and the spatial locations from the results of a mode query.

Figure 4. Diagram of the visual analysis of distribution data made possible by the techniques described in this paper

Figure 5. The spatial locations of unimodal and complex multimodal distributions in the High Island lidar data
query. Each of these processes is described in more detail in the following sections.

1) Mode Query

Mode query allows the scientist to specify queries that show: (1) the abundance, (2) the spatial location (possibly corresponding to ancillary information about that area or prior knowledge), (3) the spatial extent and (4) the spatial structure (e.g. random vs clustered) of distributions with a specific mode. This feature allows all the distributions that contain a particular mode of interest to be identified and displayed. For example, after using the interactive data probe and relating the distributions observed with field observations, the scientist may become interested in finding all unimodal distributions with a mode between, say, 117 and 194 feet. The mode query tool now allows for visual identification of all those distributions that match the query. Figure 7 shows the locations of the forest that have unimodal distributions with a canopy height mode between 117 to 194 feet and with the minimum density value of 0.05. These locations are grid cells marked by black outlines. It is not surprising that the mean field of the distribution data at these grid cells are relatively high as indicated (in the red and magenta colour range in Figure 7). For reference, the topography of the data are shown in Figure 8. Our tool allows the user to interactively change the range of values where the mode should lie within and the minimum density value, along with the number of modes for the entire distribution. When any one of these parameters changes, a new query is generated and the tool updates the screen by highlighting grid cells that satisfy the query.

2) Visualizing the Distributions from a Mode Query

From a mode query, there may be tens or hundreds of grid cells that match a specified mode criterion. The
visualization challenge here is how to display all of these density estimates effectively, so that the scientists can begin to explore the shapes and diversity of the distributions identified through the query. Furthermore, these similarities and differences need to be highlighted for analysis e.g. for those pdfs that are 'very' different, the information about which grid cells these pdfs represent should be shown/highlighted. Similarly, the grid cells of those conforming pdfs (pdfs that are similar) should be clustered or colored in the same group. If the pdfs are very different, then the scientist would like to know how they differ and where they differ.

The most common approach to view several pdfs is to simply plot them side by side for visual comparison. This can be done by plotting a set of pdfs, or as many pdfs that can possibly fit on the screen in multiple windows. If the query only found a few pdfs, then this method is ideal and effective for comparing these pdfs. However, if there were tens or hundreds of pdfs found by the query, the user would need to view so many graphs as to make this method impractical.

Another simple approach is to plot all of the pdfs in one single graph, giving the scientist a visual comparison of these pdfs. This method is only useful, however, if the scientist is interested in determining whether there are any pdfs that differ significantly from others. The scientist would be able to see the overall shapes of these pdfs using this approach. However, for more detailed comparisons of pdfs, this method would not be suitable since it is most likely that many pdfs would overlap in the graph making it difficult to distinguish the details. Figure 9 shows a graph of the pdfs that matched the query used in Figure 7. By displaying all of these pdfs in the same graph, the overall shape of these distributions can be seen to be very similar.

Visualizing the distributions identified through a mode query not only allows the scientist to inspect them, it also allows the scientist to look for the following features: (1) outliers (how are outliers shaped, how many modes do they have, etc.), (2) trends (how are most of the distributions shaped, how many modes do they have), (3) diversity (how different are they from one another), (4) homogeneity
Visualizing the Distributions and their Spatial Locations from a Mode Query

At this point, a scientist using the mode query tool has an idea of the locations and a graph of all the distributions that match the query. One source of information that is missing from the previous approach, shown in Figure 9, is that we do not know which grid cells the pdfs correspond to. In Figure 10, the pdfs from Figure 9 are plotted right above their grid cells. We found this technique to be effective also for revealing the pdfs found by the query. By plotting the pdfs right above the corresponding grid cells, we can easily see the spatial locations of the matching pdfs. Note that the pdfs are drawn such that the density estimates are plotted along the axis perpendicular to the image plane. We construct a pdf curve for each grid cell found by the mode query. A pdf curve is created by horizontally displacing points along a vertical line by the magnitude of the density estimate. The height of the pdf curve is determined by the number of evaluation points of the density estimate given in Equation 1. In our example, 150 evaluation points are used. The colour of each pdf curve presents the mean tree height of the distribution data at the corresponding grid cell. There are some parameters that the user can set for this approach. For example, the user can select another arbitrary direction to project the pdf curves and a different statistical descriptor to colour the pdf curves. Though at first glance, the ‘drop down’ pdfs shown in Figure 10 may appear to be somewhat cluttered as those shown in Figure 9. However, our tool allows the scientists to interactively rotate the graph shown in Figure 10 so that he/she may be able to view the ‘drop down’ pdfs from different angles and thus allow for a better comparison of these pdfs and their shapes.

Visualizing the distributions and the spatial locations identified through a mode query allows a scientist to inspect them and find new spatial/distribution trends or clusters of ‘like’ distributions. Conversely, the scientist can also find out where ‘different’ distributions are located and if there are any random versus clumped patterns. Lastly, it allows visual traversal of the distributions in a manageable way.

D. Distribution Exploration

Distribution exploration is performed by distribution matching and visualizing similar distribution shapes. This allows scientist to identify all distributions that are similar in their entirety rather than in just a mode as described in Section 5C. Our tool allows the user to be more restrictive or more relaxed in the specificity of finding ‘like’ distributions and allows all distributions to be ranked in terms of their similarity to the specified pattern. For example, matching could be restricted to certain data range, or only when the frequency is above a certain threshold. Likewise, matching could be relaxed by lowering an acceptance threshold or using more liberal similarity metrics. Since density estimates vary in their quality, the ability to relax or restrict the definitions of similarity with the query tool allows user flexibility in identifying a range of like distributions and their spatial locations.

Through using the interactive data probe, the scientist is able to ‘visit’ portions of the forest he was already familiar with (through field reconnaissance) and obtain three characteristic distributions of forest that are in various stages of recovery from major storm-driven disturbances. Using the distribution matching tool, all distributions that are similar in their entirety to those three distributions of interest were identified using contour lines as illustrated in Figures 11, 13 and 14. Identifying similar or matching distributions can be a powerful way to perform hypothesis testing.
Figure 11. Using the distribution matching tool, the scientist finds all distributions that are similar to the one shown on the right, which is characteristic of an area recovering from a recent disturbance event.

Figure 12. This figure shows a forest recovering from a catastrophic disturbance event which was found to have a unimodal distribution. Using the data probe tool, the scientist identified the characteristic distribution of this forest shown in Figure 11. Other forests recovering from more recent disturbance events were found to have this same characteristic distribution.

Figure 13. The distribution on the right represents tree heights of a forest clump recovering from a recent disturbance event with residual (surviving) trees. The image on the left shows contour lines of where such distributions can be found on the island. The image is color mapped according to the similarity metric used in matching the characteristic distribution on the right with the distribution at each pixel location.
testing, guide additional field work, and generate new data products of interest.

Once the contour lines are generated from the results of the distribution matching tool, an additional visualization tool provides yet another way of studying subtle differences in the matching distributions. This tool constructs color mapped characteristic distribution surfaces to depict the variations of the pdfs along the contour lines. For each grid cell along a contour line, a vertical line is plotted right above the corresponding grid cell. Then, a surface mesh is formed by connecting vertical lines from the adjacent points along the contour line. The surface mesh is colored by the density estimates. Since there are usually several contour lines, our tool would generate several disjoint characteristic distribution surfaces. As with the pdf curves shown in Figure 10, the height of the surfaces is determined by the number of evaluation points of the density estimate. Figure 15 shows the characteristic distribution surfaces of the matching distributions. As with the visualization tools provided in the modality exploration, this visualization tool

![Figure 14. Similar configuration as Figure 13, but using a characteristic distribution that implies no evidence of any recent disturbances](image1)

![Figure 15. This image shows the characteristic distribution surfaces for the contours lines shown in Figure 11. The surfaces are colored by the density estimates. The figure shows that the pdfs along the contours are mostly unimodal as indicated by the central magenta color band that runs across the middle section of the surfaces](image2)
provides even further refinement of relative homogeneity, heterogeneity and associated possible spatial patterning.

6. DISCUSSION

Several visualization techniques ranging from synoptic, general views of the full data set to more specific, localized and detailed query and display were described in this paper. The techniques allow extensions well beyond summary descriptors such as the quadratic mean, robust mean and chi-square or other non-parametric summaries or clustering algorithms. The utility of each technique fundamentally depends on the selection of an appropriate density estimator. The estimator determines how the data are smoothed and how modes are defined. Each estimator is different, and may be well-suited for one type of data but not another. The kernel estimator used in this application with lidar data, for example, may have smoothed possibly interesting features in the data. The appropriateness of a given estimator depends partly on the number of raw data values per grid cell. In general, the larger the number of raw values, the more robust a given estimator will be. The range of the data can also affect the size of the kernel used and the consequent smoothing of the distribution.

A key feature of these tools is their flexibility. Software that gives the scientist a choice of estimator and the ability to specify the parameters used in estimation will allow the accommodation of diverse data sets and exploratory data analysis. The kernel estimator we used in this study was selected because it is a robust, widely-used estimation technique but many other choices are possible.

Once an estimator is selected, the identification of modes is also not completely deterministic. Small bumps may be of little or no interest to the scientist, so what constitutes a mode in the display and query of modality can be user defined. Matching entire distributions is also a user-defined process. Success depends on increasing or decreasing the specificity of the distribution matching algorithm and having some meaningful criteria for doing so. Also, the distribution matching algorithm used is important. In this paper we used KL, but others are possible. Ultimately we envision a user-selection capability, so that various algorithms can be employed and their output assessed.

7. CONCLUDING REMARKS

Overall, our visualization tools provide new ways to query, visualize and compare distributions. The key contributions of this paper are as follows:

1. provide automated and interactive ways to analyse forest canopy distributions derived from lidar data,
2. make it easier for scientists to analyse distributions derived from lidar data, and
3. allow scientists to query distribution data for special features and then
   a) identify areas of the spatial field with similar distributions and
   b) discover potentially interesting distributions and their locations.

Although the application described in this paper deals specifically with multi-return lidar data, our tools can be easily used with distribution data sets from other applications. There are several open research problems in visualizing spatially varying distribution data sets, including the extension to distribution data that are sampled in a 3D domain and the extension to distribution data on more than one variable at a time.

8. ACKNOWLEDGEMENTS

This work is supported in part by the NASA Intelligent Systems Program Cooperative Agreement NCC2-1260 and NSF ACI-9908881. Additional support was provided to the second author through a National Research Council (NRC) postdoctoral research fellowship. We would like to thank Chris Hlavka and Michael Gerald-Yamazaki for their helpful comments and Anna Chen, Newton Der, Jose Renteria, Wei Shen, and Bing Zhang for help with programming and data preparation.

REFERENCES


