Visual Analytics of High-dimensional Data Sets
A Hyperspectral Imagery Test Case

F. W. Gasdia and B. Pineyro
Department of Physical Sciences
Embry-Riddle Aeronautical University

H. J. Cho
Department of Integrated Environmental Science
Bethune-Cookman University

Abstract

Visualization and interpretation of big data poses new and unique challenges. As engineering students enter the work force, many will be tasked with analyzing increasingly large and complex data sets with which they have little experience. This paper presents simple heat map and multi-line plotting techniques used to select critical spectral attributes produced from data mining a hyperspectral satellite image for bathymetry mapping. Additionally, good graphic design practices regarding color choice and reducing visual distraction are suggested in order to more quickly and clearly communicate information to an audience. These techniques can be applied to all types of data visualization as an effective way of communicating data.

Introduction

Modern technology, such as social network sites or airborne and satellite remote sensing, produces a massive amount of data. And with 50 times more data expected to exist in 2020 compared to 2010, demands for individuals with big data analysis skills are growing rapidly (UMUC, 2015). Big data consists of not only social information gathered on users by companies like Facebook or Google, but also consists of sensor data used to monitor factors from environmental changes to stresses in airframes and the sequence of DNA. Although engineering students will frequently encounter two or three variable relationships in their classes, few are exposed to high dimensional and extremely large data. Unfortunately, the tools and techniques that work for visualizing a few variables does not carry over to more complex data sets.

In order to demonstrate the challenges and techniques of visualizing high dimensional data, data mining of a hyperspectral image is presented as a case study. The image data used in this project is from the Hyperspectral Imager for the Coastal Ocean (HICO) on the International Space Station and was obtained February 28, 2014 over the Indian River Lagoon (IRL) on the Atlantic coast of Florida. HICO provided 87 spectral bands in the visible through near infrared wavelengths. The image area was covered by about 34,000 target pixels each of which is 90m x 90m.

Methods

Spectral features which appeared to be strong predictors of water depth were determined by data mining the hyperspectral image and ground-truth sonar data of the same area. Rather than directly using the raw light intensity (pixel value) in every spectral band, several attributes were produced for each sample pixel combination: slope, average, ratio, log ratio, and log slope of every combination of the intensities over the spectra range. This results in a total of about 10,000 individual attributes. Dissecting the intensity curves along the spectra in this way facilitates the identification of physically meaningful features of the spectra.

By definition, big data cannot be analyzed using conventional data visualization and analysis techniques. There are simply too many data points or dimensions for standard scatter plots or bar graphs to effectively reveal patterns and relationships among variables. In order to understand the 10,000 dimensional data extracted from the HICO image, binned correlation matrix graphics and multi-line plots were implemented. We present these simple visualization methods alongside information visualization theory that dictates good practice for the design and presentation of information in graphics.

Information Visualization

Information Visualization Theory (InfoVis Theory) is the science of quantification, coding, and communication of information (Chen, 2010). InfoVis Theory includes objective measures, such as the proportion of ink used on a graphic for non-redundant data (Ink-Data Ratio), to evaluate how effectively data is presented based on cognitive psychology. The goal of InfoVis Theory is to make data interpretation easier and more efficient through minimization of redundant features and effective use of visual elements. Cleveland and McGill (1985) empirically verified a general hierarchical taxonomy of basic visual properties in human perception. Humans most accurately perceive the orientation and length of data visuals while least accurately perceiving the color, volume and density of data visuals. Although the brain has a stronger response to color and volume, these attributes can potentially add undesired dimensionality to a simple graphic.

To demonstrate the effects dimensionality has on graphics, two plots of the Indian River Lagoon bathymetric distributions were created for comparison. Each graph shows the distribution of sampled depths, which is useful in identifying sampling bias. If a depth is under represented in the training data, our model cannot accurately predict that depth from the hyperspectral image. The binned depths are shown in Figures 1 and 2.

Figure 1 uses color and volume as unnecessary features that distract the viewer from interpreting the data. These extra flashy features are known as “chartjunk”. The use of different colors is chartjunk since it is an unused dimension of the visual. Extra color adds ambiguity to the data which increases the chances of a subjective interpretation from the observer. For example, the

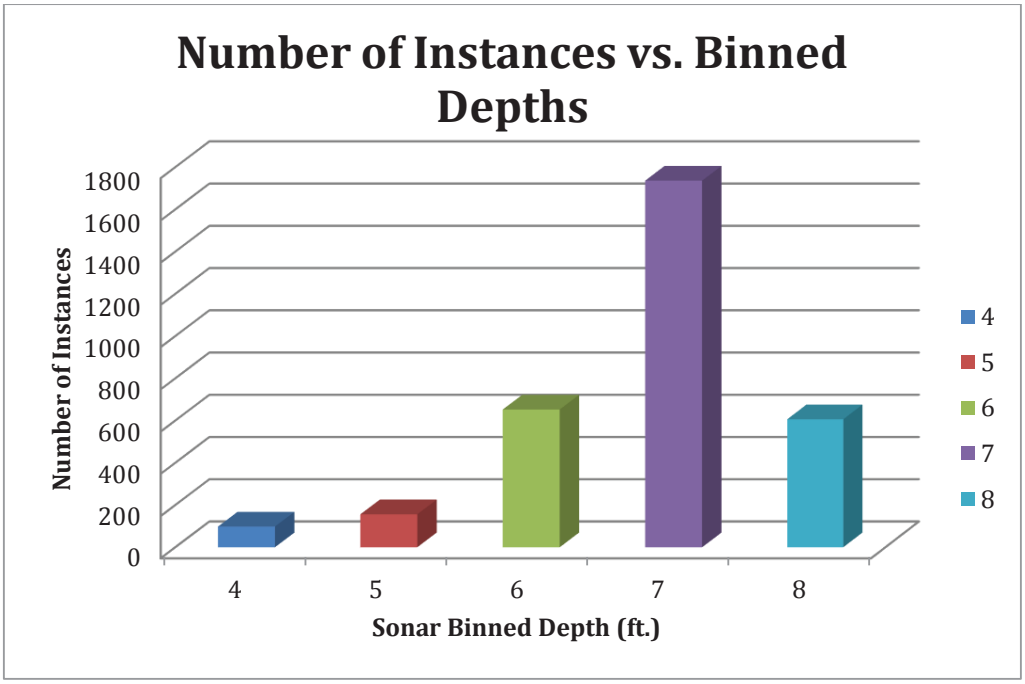

Figure 1. A visual with excessive dimensionality. The color and volume give no additional information about the bathymetric distribution which distracts the viewer from the objective of the graph.

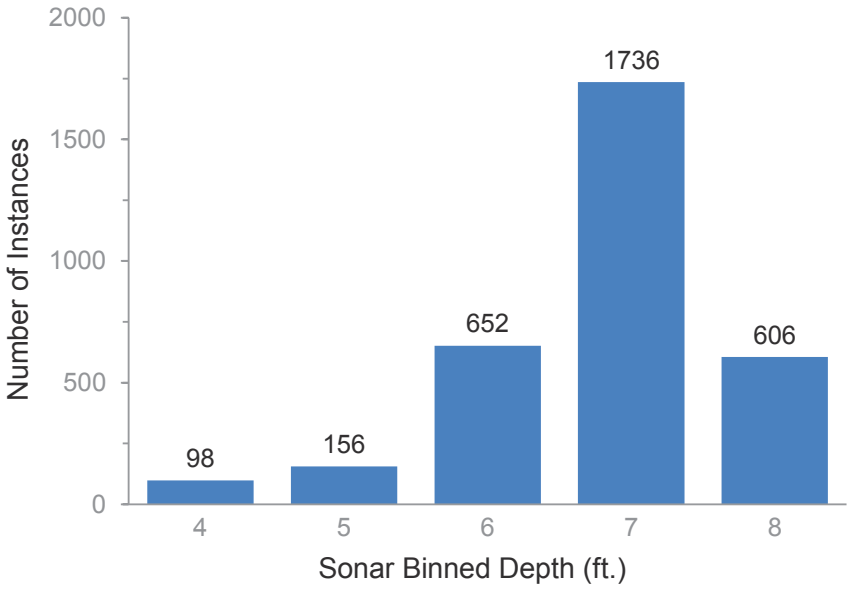

Figure 2. A visual that uses color and volume effectively. One color was used since the data was obtained from the same place. One color was used since the data was obtained from the same place. The numbers on top of the histograms remove ambiguity from the true value of the binned depths.

different colors could suggest the depths used in the analysis come from different bodies of water, when in fact all of the depths were collected in the same geographic extent. Volume of the bars is also chartjunk in Figure 1 since it does not help describe the representation of depths. Based on Figure 1, the number of instances for the depths cannot be clearly determined. The 8-ft sampled depth appears to have less than 600 instances due to the three-dimensional effects.

In order to create a better visual, the typeface, color, volume, and graph design were taken into consideration. Figure 2 shows a clearer representation of the binned depths. The typeface was chosen as a sans-serif because it is more legible and removes style which could be distracting. The volume and color over dimensionality from Figure 1 was removed in Figure 2. The number of instances is displayed over the bar graph which provides exact values without comprising the trend of the graph. These changes were made so that the viewer can quickly observe the trend of the data, gather as much information as quickly as possible with the least amount of confusion.

InfoVis theory differs with larger data sets. During the analysis of HICO hyperspectral profiles, it was known a priori that reflected light intensity decreases with depth due to increasing light attenuation through the water column. In order to test the proposed hypothesis 200 water depths between 4 and 8 feet were sampled from the thousands of measured spectral curves. Figures 3 and 4 provide another example of how data can become unclear with improper use of graphic elements and how InfoVis Theory can be used to form a clearer visual. The line color variation in Figure 3 represents the different depths but no pattern emerges at first glance and anomalous data is not easily detected. The lines in the graph are the default colors from the Matplotlib graphing software. Since most plotting programs have similar default color schemes, it is easily seen that the programs have not been designed for high-dimensionality data.

Ascending lightness values of a monotonic hue were used as a gradient which assigns the lighter colors to shallower depth and gradually darker colors to deeper depths. Pattern recognition is easier using Figure 4 compared to Figure 3 because of the simple presentation of the relationship between and depth. Each of the examples has drawbacks; and the key to InfoVis theory is to balance the trade-off of added dimensionality. Figure 4 shows the distribution of depths through the color ramp. Although it is easily seen how the depths vary, resolving each depth isn't as easily done. On some parts of the spectral profile, deeper depths have higher intensity values than the shallower depths. This can be observed in Figure 3 more easily than in Figure 4.

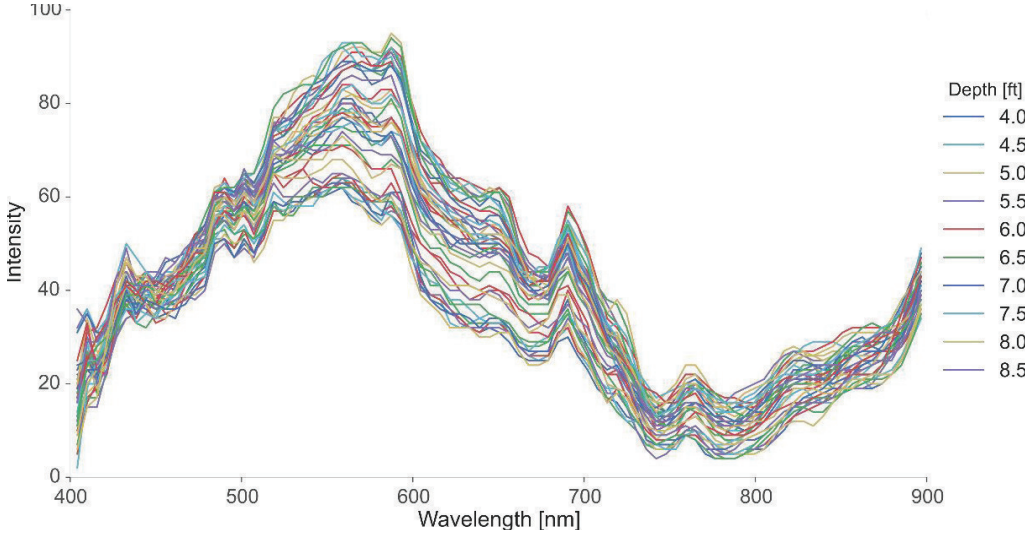

Figure 3. Sample depths from the Indian River Lagoon and their spectral profile.

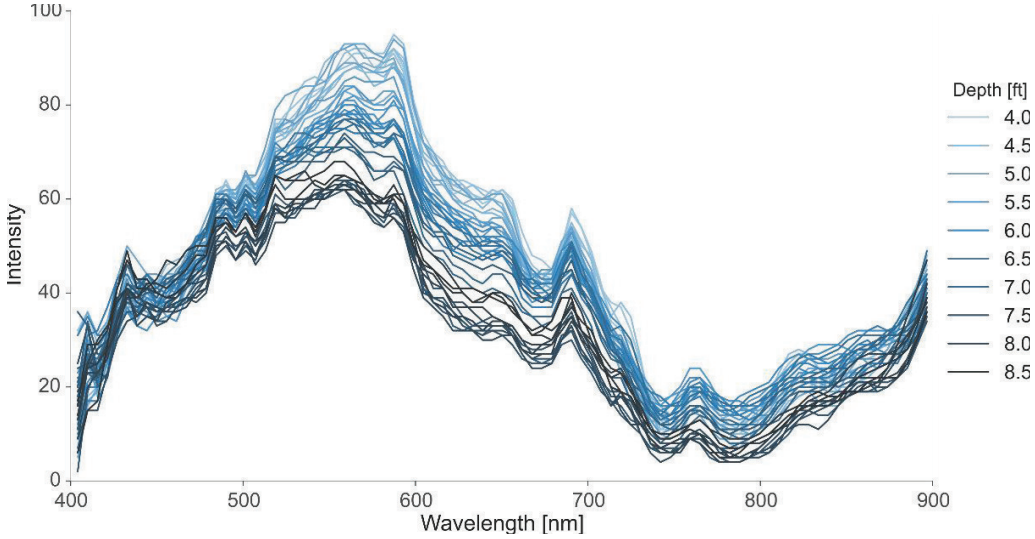

Figure 4. Sampled depths from the Indian River Lagoon and their spectral profiles. The color palette 'blues' makes use of color as a gradient which aids pattern recognition.

Heat Maps for Identifying Relationships in High Dimensional Data

The HICO instrument measured 87 unique spectral intensities at every pixel. In order to identify features of this spectral curve that are good predictors of water depth, 10,000 total combinations of average intensity, slope, ratio, log slope, and log ratio were calculated for every combination of the spectral bands. Although techniques such as Principal Component Analysis (PCA) identify unique features and help reduce the dimensionality of the problem, they rearrange the raw spectral information such that the information contained in the original intensities is lost in

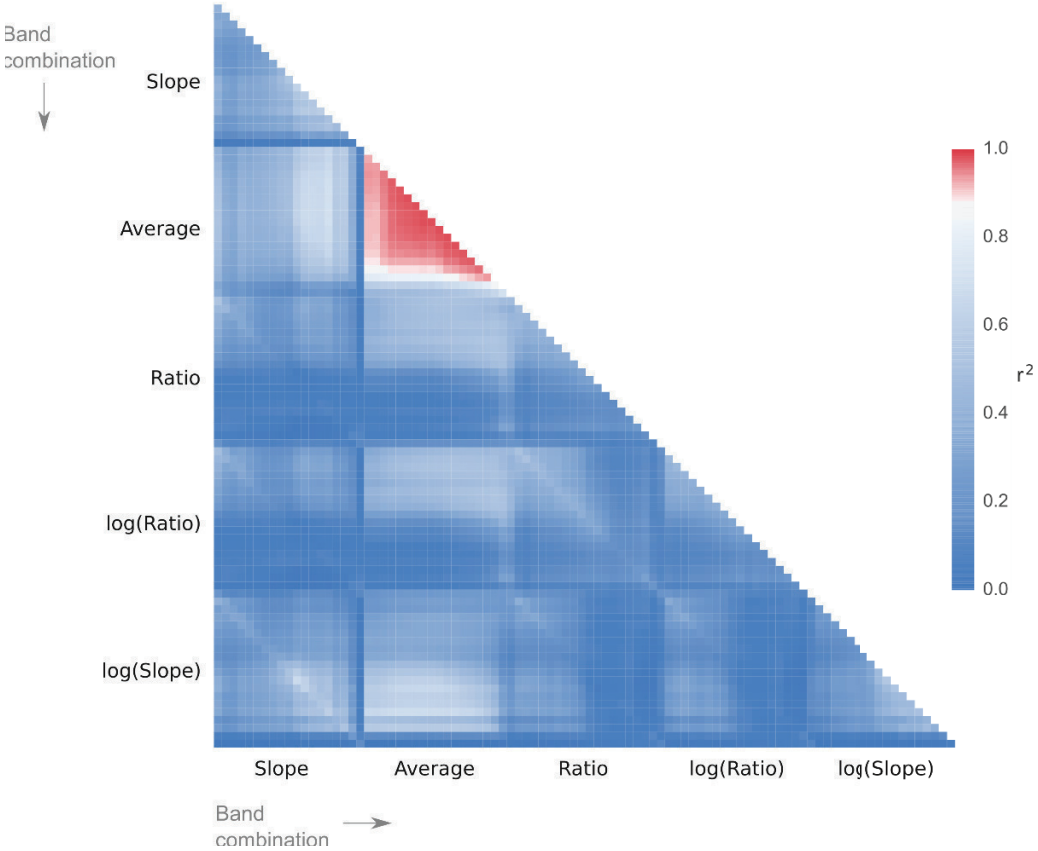

Figure 5. Binned correlation matrix for the 10,000 attributes which describe the spectral curves. The rejection threshold is at $r^2 = 0.9$. Each position in this matrix represents the average of 100 correlation coefficients between given attributes. As the viewer moves from left to right or up to down across the graphic, the spectral bands being compared are toward the redder wavelengths. This repeats for each of the 5 attribute types.

the PCA-derived attributes. If instead the curve shape is described by many single attributes, future bathymetry mapping efforts can use the strongest indicators directly and also give more insight into the physical processes of light energy propagation through a water column.

Once each of the 10,000 spectral curve features are calculated, it is necessary to remove redundant information. Because each attribute was calculated for every combination of spectral band (e.g. average intensity between band 4 and 7 and average intensity between band 4 and 8), many attributes are similar and can be removed without loss of critical information. In order to identify the most unique attributes, we calculate a correlation matrix and remove attributes for which their correlation with another attribute falls above some threshold. However, if each correlation coefficient were to be displayed, the user would need to look through 100,000,000 values. In order to quickly view patterns and identify outliers, correlation matrices are commonly displayed using heat maps: graphics in which each matrix position is colored with a color which represents the value

it holds. However, with a 10,000 square matrix, to view each point would require several computer monitors arranged in an array. Because of the slowly varying nature of our attribute values, we average every 100x100 matrix positions into a single bin. This greatly reduces the number of points plotted in the heat map while maintaining the overall attribute correlation trends.

Figure 5 demonstrates the binned heat map approach for analyzing the relationship between variables. To emphasize the uniqueness of each variable, a simple diverging color map was chosen and centered on the correlation threshold value for attribute rejection. By changing the correlation threshold, we can choose how many attributes are rejected before the next phase of our bathymetry mapping routine. Correlations that fall above the threshold appear as red while correlations below the threshold are blue. Additionally, stronger positive and negative correlations have more saturated colors. This color scheme enables the viewer to quickly distinguish between spectral attributes that are strong or weak identifiers of depth. For example, in Figure 5, average intensities contain little unique information and are clearly above the rejection criteria. Many patterns exist across the other attributes and are visible as white stripes or patches in the correlation matrix. Although these are not colored red, they are still just at the edge of rejection and the corresponding attributes are not very unique.

Conclusion

Data mining provides a unique way for many fields to monitor and analyze data. Given the nature of big data, it is necessary to apply special methods of visualization while avoiding the use of dimensionally ambiguous visual features. InfoVis Theory offers techniques to create effective visuals so a viewer can gather the maximum information in the least amount of time without compromising data or creating chartjunk. Visual analytics and InfoVis Theory can be used in school curricula and work force training to teach the sometimes non-intuitive skills necessary to create an effective data graphic.

References

Chen, M., Janicke, H. (2010). *An Information-theoretic Framework for Visualization.* IEEE Transactions on Visualization and Computer Graphics, Vol 16, No 6, pp 1206-1215.

Cleveland, W., McGill, R. (1985). *Graphical Perception and Graphical Methods for Analyzing Scientific Data.* Science, Vol 229, No 4716, pp 828-833.

University of Maryland University College (2015). *The Big Data Revolution is Here.* Adelphi, MD. Retrieved from: http://www.umuc.edu/analytics/about/big-data-job-growth-infographic.cfm