Survey Paper on Data Visualization on Multiple Dimensions

Jose Jimenez (6140508)
KFSCIS
Florida International University
Miami, Florida
jjime197@fiu.edu

***Abstract*—Critical data analysis consists of thorough analysis, and data in massive quantities. Visualizing data is a powerful asset when it comes to data analysis. Visualizing data on more than three dimensions using Cartesian Coordinates is not possible, especially when data is categorical, continuous, or mixed. Multiple methods are used in visualizing data analysis, such as Mosaic Plots, Parallel Coordinate Plots, and Trellis Displays.**

***Index Terms*—Cartesian Coordinates, Data Analysis, Artificial Intelligence, Machine Learning, Cartesian Planes, Mosaic Plots, Parallel Coordinate Plots, Trellis Displays**

I. Introduction

Data analysis has been a useful tool for analyze the unpredictable universe. It is a valuable asset in Artificial Intelligence (AI) and Machine Learning (ML). As a matter of fact, AI and ML use this as a method to predict and interpret incoming data. Besides algorithms analyzing data, people analyze data. For someone to interpret the data, one must visualize it using plots and diagrams. Beforehand the data is prepossessed into an interpretable (data) set.

Most common method of data visualization is plotting the data as individual points on a Cartesian plane. Cartesian planes, visually, have limits. For one to interpret and visualize the data, one must graph at least one dimensional data, and up to three dimensional data. At least one dimension must be an input dimension, so that the others may be used to give a result based on the input. Therefore, it leads to the data analyst or algorithm, up to two output data dimensions, or two input data dimensions.

Data that is given in the real world is more than likely to have more than two dimensions. Classical methods alone will not make the data easily interpreted. Method(s) must be created to solve the issue of lack of data entries. When data analyst need more data, they resort to a dataset, but with more dimensions. Data entries are also complex. The cause of a data entry being a certain value is factored by many variables (dimensions). To understand the vast complexity of data, multiple dimensions must be accounted for.

[1]In cancer research, data mining data with multiple dimensions are likely to be interpreted by a researcher. For the study to work, the researchers must acknowledge that cancer is a complex sickness that has many input values, that defines cancer as cancer. Although there are multiple data entries, the amount alone does not suffice for a conclusion.

II. Survey

In this paper, it is implemented the following multi-dimensional visualization techniques to a data set in R. The techniques are Mosaic Plots, Parallel Coordinate Plots, and Trellis Displays. These plots are different approaches to visualizing data that consist of multiple attributes. The most appropriate plot for a dataset will vary on the data.

A. Mosaic Plots

[2] After a painting style, Mosaic plots are plots that consist of multiple bar charts that vary in sizes, such that the size and shape of the bar chart consist of a dimension value itself. More specifically, the height of every bar in the chart is designed to be equal to 100%. Vertically, the bar contains at least two data categories, that each represent a portion of the 100%. Every portion will have assigned its own color. Every bar is a part of a certain attribute. Every attribute, share most if not all categories at different levels. Horizontally, the bar charts represent a single dimension or attribute of our data. Colors can be used to represent a separate dimension. It is useful to use Mosaic plots to display non-obvious patterns in one's data. The plot is designed to be pleasing to the reader(s).

B. Parallel Coordinate Plots

[3] Parallel Coordinate Plots help plot every attribute of a data point on a Cartesian plane in parallel with each other. It is used mostly to analyze data mining patterns, and identify and understand the data based on what the data has. If there is too much plotting, the data plot can be confusing due to the massive amounts of plotting lines stacked on each other; however, if most frequent areas can be highlighted - that is the most common areas that plots (lines) are presented - the plot can be useful to show the distribution between attributes, in other words, creating more visuals, in this case a heat map, will deliver to the reader more information in a compressed manner, especially for this type of plot. Since there will likely be plots stacked on each other, heat maps are a useful addition to the plot. There is virtually no limit to the amount of dimensions using this method, because every input variable (dimension) is plotted (usually) on the x-axis. This

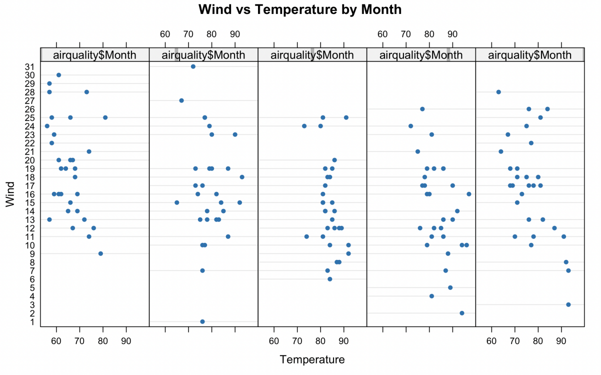

Fig. 1. Trellis Plots separating the months as every separable attribute that contain two dimensions

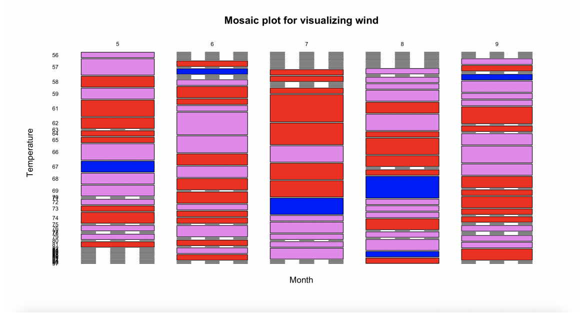

Fig. 2. Mosaic Plots used to represent the wind. Every box represents a data point. Colors are the dimensions of wind speed. Every column represent a month. Y-Axis represents the Temperature

plot is designed for readers to understand the data as a whole. Using this ideology, this plot can identify certain patterns, thus creating useful rules for our dataset.

C. Trellis Displays

[4] Similar to the Parallel Coordinate Plots, the data is plotted using lines. It is designed to plot every attribute on a separate plane called panels. Panels are not stacked on other panels, instead they are displayed in parallel with each other. It is designed for the reader to analyze every attribute carefully and separately. When readers struggle on interpreting stacked plots, such as Parallel Coordinate Plots, this plot would make it easier for the reader to understand the data attribute by attribute. It is well known plot to display outliers or miscalculated data. This plot is designed for readers to understand their data, attribute by attribute.

III. Analysis on a Dataset

In this section, an example dataset will be used to demonstrate the plots explained. The dataset "airquality," which represents New York's Air Quality, will be shown using the different methods. It is important to note that the data is large in scale. If we want these methods to work, we must have data on a large scale in order to see certain patterns. As shown below, the results have been analyzed by the author.

TABLE I
New York's Air Quality

	Ozone	Solar.R	Wind	Temp	Month	Day
Min	1.00	7.0	1.7	56.0	5.0	1.0
1st Qu.	18.00	115.8	7.4	72.00	6.0	8.0
Median	31.50	205.0	9.7	79.0	7.0	16.0
Mean	42.13	185.9	9.958	77.88	6.993	15.8
3rd Qu.	63.25	258.8	11.5	85.0	8.0	23.0
Max.	168.00	334.0	20.7	97.00	9.00	31.0

A. Process

In the creation of these data plots, multiple libraries were used. The programming language used to create the plot for this analysis is R. During development, many dependencies of libraries needed to be installed in this case. For the sake of this example, wind speed will be the asset we want to predict. First, we have to assume a model for this analysis to work. In this example, we assume that temperature, month, and ozone will affect the wind speed. On every plot, we must specify which axis belongs to the output variable - wind speed - and all of the other input variables - temperature, month, and ozone. In the mosaic plot the y-axis will be our output variable, while the first x-axis, the one that separates the bar charts, will represent every month in the data; therefore, every bar pair in the plot represents a month. The second axis, the one that separates the bars inside a single month attribute, will represent. Now that all of the data is visualized, we must use common data mining algorithms to analyze and interpret the data. In our "Trellis Plot," we see that in every attribute the points are clustered in on one area. Based on the author's experience in data mining and his visual interpretation, the best algorithm in this case, is the K-Nearest Neighbor Algorithm (KNN).

B. K-Nearest Neighbor Algorithm (KNN)

The way that KNN works, is that it will classify every data point on multi-dimesional plot. Once that is completed, the data point that is needed for prediction, will be based on it's neighbors. The closer the neighbor, the more likely the data point. "K" is a amount of selected neighbors used to identify the result for our prediction. The algorithm will choose "K" neighbors, and the selected neighbors will vote on what the prediction actually is. The vote will be based on the selected neighbor's value.

$$i = (d_1, d_2, ..., d_n) \tag{1}$$

$$KNN = \arg\max_i(euclidean(i)) \tag{2}$$

$$delta_w = (d_{1w} - \sum_{i=2}^{n}(d_{iw})) \tag{3}$$

$$euclidean(i) = \sqrt{\sum_{w=1}^{n}(delta_w)^2} \tag{4}$$

Equation (1) represents a coordinate on a multidimensional plot. Equation (2) is the indicator for KNN. The highest

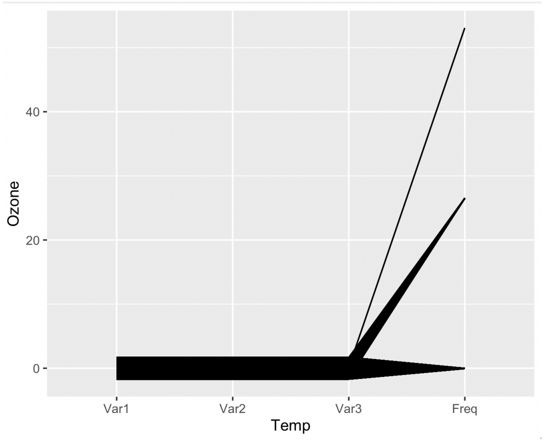

Fig. 3. Parallel Lines can be used to predict wind speed. Var1 represents the temperature, Var2 represents the Ozone levels, and Var3 represents wind. Frequency variable shows how often the three variables appear.

vote will be the prediction. Equation (3) is used to find the difference between all points on a dimension. ”w” is a variable - a placeholder - for dimension. Equation (4) is a vital function for KNN. It is used to score every attempt in KNN. This algorithm chooses a series of ”k” points, and uses that to vote what the prediction will be. What we have now is our training data for future incoming datasets.

C. Results

The results are hard to interpret if there is too much clustered data. In this case, there is data in large quantities. So, data is color coated to represent a separate dimensions. Based on the results of the example, we have clustered data to show wind speed. Every plot will have its own interpretation for getting the result. Usually, there is a better plot than others.

D. Forecasts and Predictions

Plotting all of the graphs together we see that Mosaic plot has shown where the wind has been blowing its hardest and its softest. Every color represents the wind speed, or the dimensional variable that we care about in this example. The color blue represents wind speeds less than 7.4 mph, slow wind speeds; the color violet represents wind speeds greater than 7.4, but less than 11.5 mph, medium wind speeds; and red represent high wind speeds, wind speeds greater than 11.5 mph. Using this as a heat map, we see that the wind is likely to be medium in the month of June (6) when temperatures are in between 60 and 68 degrees. This is also the case when it is July (7) and temperatures go above 75 degrees. On the lattice plot, we have a clustering of data points. Wind speed typically stand in one area. On the parallel lines plot, attributes stand in one area as well.

IV. Conclusion

In this paper, we have noted and analyze how to interpret multiple datasets on a Cartesian plane. Each of these plots have a different approach to plotting on multiple dimensions. The best method will vary, due to regression, clustering, and uncertainty. Also, the method on interpretation plays a role on the result. Interpretation can skew the data to make it seem far from what it actually is.

The example was a great example in showcasing the different ways that data is presented thus bringing in different interpretations. Although the interpretation are different, the results remain similar and useful for predicting wind speed.

All of these methods of plotting has helped science advance in ways it could not before. Visualizing multiple dimensions is realistic when it comes to data analysis. Cartesian plotting method alone could not do past three dimensions, unless if some modifications are made. Most data rules cannot be made on just three, two, or one dimensions, unless it is data that is considered causal.

Acknowledgment

The author would like to thank Prof. Dr. Antonio Hernandez for his support, knowledge, and guidance for helping me deliver this survey.

References

[1] P. A.-C. P. E. A.-R. C. O.-M. E. J. C.-O. A. P.-M. M. E. M.-P. F. Cesar Morales-Ortega Roberto, German Lozano-Bernal and M. Roca-Vides, “Method based on data mining techniques for breast cancer recurrence analysis,” in *Behavior research methods, instruments, computers: a journal of the Psychonomic Society, Inc*, 2002.

[2] C.-h. C.-A. U. Martin Theus, Wolfgang Karl Karl Härdle, “High dimensional data visualization,” in *Handbook of Data Visualization*, 2008, pp. 151–178.

[3] L. B. F. Matthew J Pastizzo, Robert F. Erbacher, “Multidimensional data visualization,” in *Behavior research methods, instruments, computers: a journal of the Psychonomic Society, Inc*, 2002.

[4] A. M. Namrata Jaiswal, Vishan Kumar Gupta, “Survey paper on various techniques of recognition and tracking,” in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015.