APPLICATIONS OF MULTIDIMENSIONAL SCALING WITH CO-PLOT ANALYSIS

Priyanka Poosapati

A Capstone Project (or Thesis) Submitted to the
University of North Carolina Wilmington in Partial Fulfillment
of the Requirements for the Degree of
Master of Science

Department of Computer Science
&
Congdon School of Supply Chain, Business Analytics, Information Systems

University of North Carolina Wilmington

2023

Approved by

Advisory Committee

Dr. Judith Gebauer	Dr. Yang Song

Dr. Yao Shi, Chair

Accepted By

Dean, Graduate School

Table of Contents

ABSTRACT

Exploratory graphs are a crucial element of statistical analysis and models that help us identify and interpret data patterns. As an example of exploratory graphs, multidimensional scaling (MDS) allows the visualization of large and complex datasets in a reduced dimensional space, while preserving pairwise differences and similarities between data points. Co-plot analysis extends the capabilities of traditional MDS and enables a multifaceted view of relationships between variables from different datasets. To develop a co-plot graph, the analyst starts out by standardizing the dataset and by generating a median absolute deviation correlation coefficient (MADCC) that subjects the dataset to MDS-embedding and helps to determine how the dataset fits the surface. This study demonstrates the use of non-metric multidimensional scaling (NMDS) and robust multidimensional scaling (RMDS) techniques to visualize dissimilarities and distances of the observations in a reduced dimensional space. The study also explains the use of shepard graphs as a second technique to visualize categorical variables in a reduced-dimensional space, with points closer together indicating higher similarity. Shepard graphs can assess the goodness of fit, which is indicated by the strength of a positive linear relationship. To demonstrate the application of co-plot analysis, two scenarios are used in the study. Scenario one graphically displays and compares MBA Programs in the U.S. Scenario two presents how co-plot analysis can be applied to describe and compare the patterns of ecological communities and species across multiple sites and environmental contexts. This study highlights the benefits of co-plot graphs in general as well as three different techniques in particular: Shepard analysis, NMDS, and RMDS techniques.

Keywords: Multidimensional scaling (MDS), Co-plot, Data visualization

CHAPTER 1: INTRODUCTION

Exploratory graphs are widely used in data analysis to suggest relevant statistical analyses and models to diagnose aspects of the data. In most situations, exploratory graphs do not call for presumptions about how the data will behave or how the process will work. Data analysis offers a wide range of graphical techniques for the treatment of large data sets [11].

Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) are the two widely applied techniques for dimensionality reduction and data visualization [6,7]. PCA aims to transform a high-dimensional dataset into a lower-dimensional space while retaining as much variance as possible. In contrast to PCA, MDS focuses on preserving pairwise distances and relationships between variables. Therefore, PCA is mainly used in data reduction, feature extraction, and visualization while MDS is typically used for interaction analysis. Although the above two techniques are widely used, Raveh [11] introduced Co-plot analysis as a complementary data visualization technique to address the limitations of the two methods. Co-plot analysis extends the capabilities of MDS by presenting additional scatter plots or line plots alongside the traditional MDS plot, effectively providing a multidimensional view of the relationships between variables. It is used for investigating connections between two or more sets of variables at once. Co-plot Analysis can explore complex relationships and dependencies between variables from different datasets simultaneously [9]. It provides a visual representation of the data structure and aids in identifying similarities, differences, and potential interactions between variables. This paper uses the graphical display technique called robust Co-plot [11] which includes both Non-Dimensional Scaling and robust Multidimensional Scaling for comparison and identifying the best solution [2,6]. The objective holds on how co-plotting Multidimensional Scaling may represent critical data points or outliers that can be crucial in understanding the underlying structure of the data. The focus is the efficiency and advantages of using Co-plot analysis as a method of multidimensional data visualization in MDS [2,6]. To

demonstrate the features of Co-plot analysis, this study applied the method to a sample dataset of ecological community species. A matrix of scatterplots that shows the correlations between variables is what is intended by Co-plot analysis [8,9,15]. It can investigate connections between two or more sets of variables at once. To view and examine large data structures, it is frequently used in conjunction with MDS. The method yields three outcomes: (1) similarity between observations based on the sum of all relevant variables; (2) the pattern of inter-variable correlations; and (3) the reciprocal links between the variables and the observations.

This study is structured as follows: we start with methodology in section 2 which briefly introduces the steps of Co-plot analysis, then present the application of Co-plot analysis in MBA Programs in the U.S and ecological community species analysis in section 3. We present conclusions and limitations in section 4.

CHAPTER 2: METHODOLOGY

Co-plot serves as a valuable tool for exploring relationships between multiple sets of variables. The method has been applied in many areas, for instance psychology, marketing, geography. In psychology, Co-plot analysis illustrated similarities or dissimilarities among the subject's perception of various stimuli [10]. In marketing, it helped to understand customer preferences based on multiple product attributes [11]. Co-plot helps geographers to examine relationships between various geographic factors such as temperature, rainfall, elevation, land use, economic indicators, etc. [8]. In business, co-plotting refers to cooperative strategic planning across various teams, departments, or stakeholders. It includes working together to plan, create, and carry out company strategies [17]. Diverse viewpoints and proficiencies combine to provide a thorough and logical strategic plan whereas Co-plotting, in IS/IT, describes teamwork in the planning, creation, and use of technological solutions or information systems. To guarantee that the system satisfies a variety of user needs and technological requirements, cross-functional teams work together [17]. This study attempts to apply Co-plot analysis in ecology to reveal the relationships between species based on various environmental factors. The Robust Coplot method generally consists of four major steps to obtain the robust graphs.

Step 1: The initial step is to obtain standardized data for the Co-plot analysis as variables measured at different scales do not contribute equally to the analysis. To ensure variables measured at various scales contribute equally to the analysis, the data needs to be standardized. Nevertheless, conventional standardization techniques that rely on sample mean and standard deviation are susceptible to outliers since even a single strong outlier may lead to distorted results [3]. The use of robust estimators, specifically the median and median absolute deviation (MAD), helps mitigate the influence of outliers on the standardization process. In Co-plot, the p dimensional n point data matrix X_{nxp} is transformed into the standardized matrix Z_{ixj} in a robust way as follows:

$$Z_{ij} = \frac{x_{ij} - MED(x_j)}{MAD(x_j)} \quad (1)$$

Where Z_{ij} is the i-th row and j-th column element of the standardized matrix Z_{nxp}, x_j is the j -th column of data matrix X_{nxp}, *MED* (.) is the median function, and $MAD(x_j) = 1.4286\, med(|x_j - med(x_j)|)$ stands for the median absolute deviation. MAD of any vector x of observations is always a median (abs(x - median(x))) multiplied by the default constant 1.4826 (scale factor for MAD for non-normal distribution), which is used to put MAD on the same scale as the data which assumes normally distributed data[1]

Step 2: In the second step of the Co-plot method [4,8], the p-dimensional dataset is projected onto a two-dimensional space, considering the dissimilarity metric obtained from the standardized data matrix. The goal is to find an appropriate embedding that represents the relationships within the data. Both metric (classic) and non-metric (ordinary) MDS techniques are commonly used in literature. However, non-metric MDS (NMDS) has shown that it can be negatively affected by outliers. To address this issue, the Co-plot method utilizes a robust MDS (RMDS) approach. RMDS introduces an outlier-aware cost function which offers several advantages [1,2,5].

Step 3: This step involves drawing p vectors on the graph that was created in the previous stage. A vector obtained from the center of gravity of the observations is used to represent each variable. The median absolute deviation correlation coefficient (MADCC) determines the direction and size of the vector. Each vector's direction is selected so that there is a maximum correlation between the related variable's initial values and its projections on the selected vector. MADCC for each variable is calculated, involving the robust principal variables and their projections. Correlation value measures how well vectors fit to a surface. Additionally, observations with high value in the vector are in the graph where the vector points to. MADCC given by [13] is defined as follows:

$$\rho_{j,MADCC} = \frac{MAD^2(u_j) - MAD^2(k_j)}{MAD^2(u_j) + MAD^2(k_j)} \quad (2)$$

where, u_j and k_j are the principal variables.

Step 4: Observations are colored in the Co-plot graph according to the chosen categorical variable. Two vectors near to each other and pointing in the same direction reflect two strongly correlated variables. If the correlation between the variables is negative, the corresponding vectors will point in the opposite direction. Two vectors that are perpendicular to one another represent two uncorrelated variables. Observations that are strongly influenced by a particular variable are clustered together. This mass is oriented in the same direction as the vector of the variable. One or more potential outlier observations are buried far from most observations. Outliers can cause distortion in the calculated distances between points. In MDS, the goal is to represent the dissimilarities or distances between data points accurately. Outliers, by their very nature of being extreme values, can disproportionately influence the computation of distances, leading to a skewed representation of relationships between points. In coplots generated from MDS analysis, outliers might appear as data points located far away from the main cluster or pattern of points. Their disproportionate influence on distance calculations might cause them to be visually highlighted, potentially misleading the interpretation of the overall relationships among many of the data points. Outliers, if not handled appropriately, might force the lower-dimensional representation to adjust significantly to accommodate their extreme values, affecting the overall structure of the visualization. Thus, we should always avoid outliers as much as possible.

CHAPTER 3: APPLICATIONS AND DISCUSSIONS

3.1 Scenario 1: MBA Programs in the U.S.

3.1.1 Description of the Dataset

We are mainly focused on the use of Co-plot analysis in Multidimensional Scaling and this method is demonstrated using graphical representation of a multidimensional dataset of an MBA Program which will include concentrations/specializations, months of education, location of the program and the universities which are offered in. We are looking to help students through this research to understand all the MBA Programs in a single graph so that they can choose which program they are interested in to decide. Easier for people to understand. All the programs are together, which can help to develop IT concentrations. Also, In the case of the MBA dataset, where multiple factors might affect the relationship between variables, coplots can reveal more nuanced patterns that traditional methods might overlook. Our motivation is to analyze relationships while controlling the influence of confounding variables, and the clusters of observations which are highly characterized by a particular variable is mapped together and located in the same direction as that of the variable's vector and main advantage of this representation is to allow the simultaneous consideration of both variables and observations. We use this method to identify the characteristics of IS/IT offerings in today's MBA programs and the curriculum patterns in current MBA dataset. By applying the method, we seek to obtain two major findings: (1) similarity among IT concentrations in MBA programs based on the composite of all criteria (concentration of the courses, education months of the program, location of the program) involved. (2) correlations among different universities of different states based on Education months and Concentration.

In the dataset, there are six variables for 55 universities which include: university Name (UNV), id of the University (ID), rank of the University (RNK), state in which the university locates (STATE), months of the education required in that program (EDUCATION MONTH), the amount

of concentrations in the MBA program (CONCENTRATION) and region category the university falls into (REGION), Regions are represented with values 1-5. Universities in northeast are coded by 1, southeast by 2, Midwest by 3, southwest by 4 and west by 5. Color code is used for region which is quite helpful in identifying clusters of observations and possible outliers. The graphs are plotted based on university names which could be identified based on ID. This data (shown in Table 1) is stored in csv file and processed by MATLAB (MathWorks Inc, 2016).

Table 1. Sample Dataset of MBA Programs in the United States

UNV (1)	ID (2)	RNK (3)	STATE (4)	EUCATION MONTH (5)	CONCENTRATION (6)	REGION (7)
Stanford University	1	1	1	20	9	5
University of Pennsylvania (Wharton)	2	1	5	20	18	5
Northwestern University (Kellogg)	3	3	6	22	8	2
University of Chicago (Booth)	4	3	6	21	13	2
Massachusetts Institute of Technology (Sloan)	5	5	16	20	3	1
Harvard University	6	6	16	21	10	3
University of California--Berkeley (Haas)	7	7	1	24	0	3
Columbia University	8	8	4	21	0	2
Yale University	9	9	29	24	0	2
New York University (Stern)	10	10	4	21	27	3
University of Virginia (Darden)	11	11	12	21	11	1
Dartmouth College (Tuck)	12	12	41	12	0	2
Duke University (Fuqua)	13	12	9	22	13	2
University of Michigan--Ann Arbor (Ross)	14	12	10	20	9	2
Cornell University (Johnson)	15	15	4	21	12	2

University of California--Los Angeles (Anderson)	16	16	1	22	15	1
University of Southern California (Marshall)	17	17	1	12	7	3
University of Texas--Austin (McCombs)	18	18	2	21	20	3
Carnegie Mellon University (Tepper)	19	19	4	21	15	1

3.1.2 Interpretation of Graphs

NMDS and RMDS Analysis

Co-plot Analysis is basically used to explore and visualize relationships between variables from the datasets. The package used here supports non-metric MDS analysis [1]. Non-metric Multi-dimensional Scaling (NMDS) is a way to condense information from multidimensional data (multiple variables/species), into a 2D representation or ordination. The interpretation of an NMDS graph involves examining the proximity of points. Objects or samples that are close together on the graph are more like each other, while those that are far apart are more dissimilar. For Example, as demonstrated in Figure 1, points 13 and 15 have similarities since they share the same region while 1 and 25 have differences since they are in different regions. 30 and 32 share the same region since they have the same blue color, they have many similarities in characteristics though they are far away from each other. It can be inferred that all the similar colors have similarities but if they are far from each other, they may not have many similar characteristics.

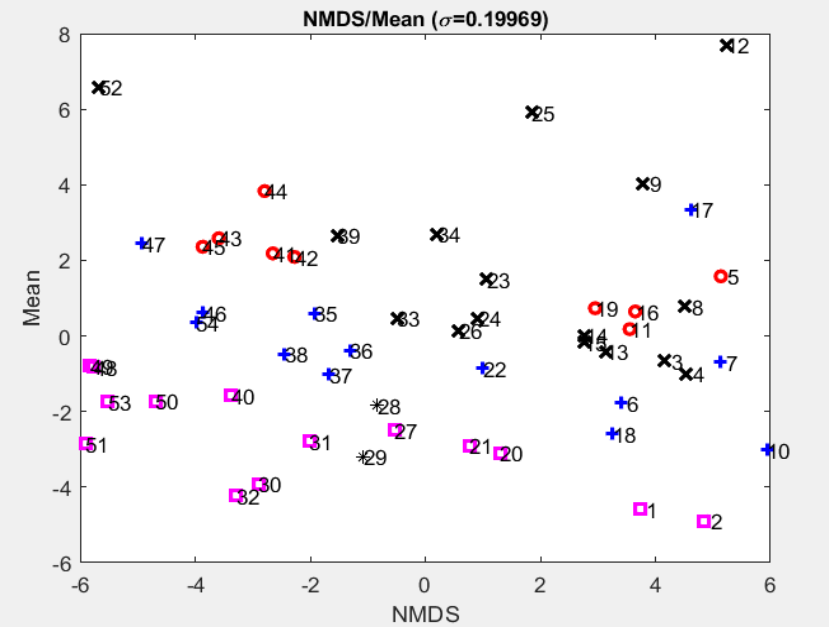

Figure 1. Non-metric MDS Analysis of MBA Dataset

Robust Multidimensional Scaling (RMDS), on the other hand, assumes that the dissimilarity matrix can be transformed into a meaningful distance matrix [2]. The dissimilarities are converted into distances while ensuring metric properties (symmetry, non-negativity, and the triangle inequality). RMDS assumes a linear relationship between dissimilarities and distances. We can use coplots to enhance the analysis of MDS results. Coplots can provide additional insights by allowing you to examine the relationships between variables while considering the effects of other variables. The combination of MDS plots and coplots provides a richer context for interpreting your data. For example, in Figure 2, point 14 and 15 have similarities since they share the same region and are nearer to each other, so they share the same characteristics as well. While point 2 and 10 have differences since they share a different region though they are nearer to each other.

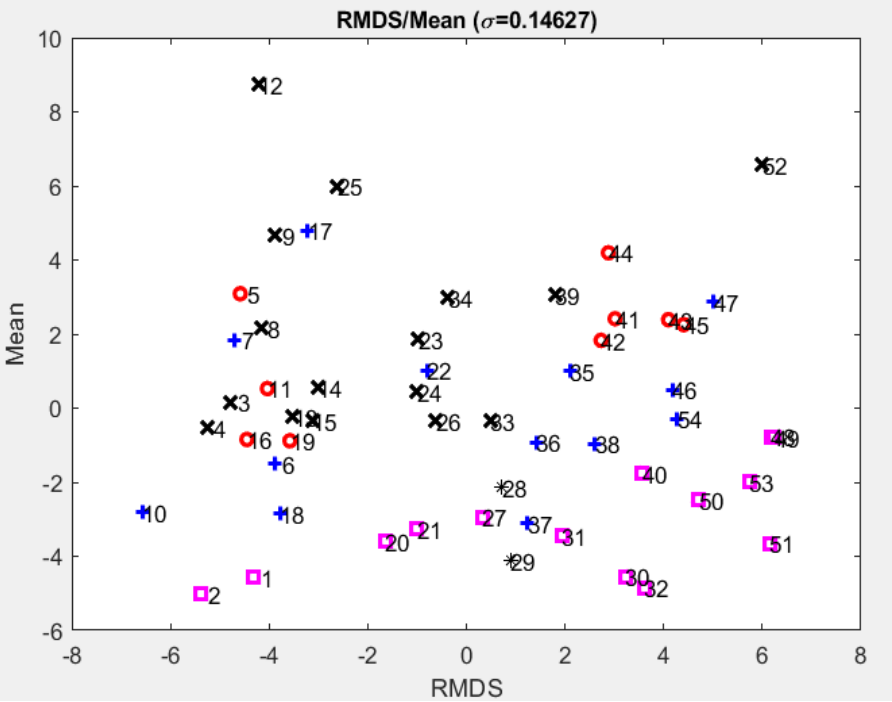

Figure 2. Robust MDS Analysis of MBA Dataset

We used Step 1 and Step2 to standardize the dataset and to change the scale and reduce the impact of the outliers.

Differences Between NMDS and RMDS

NMDS (Non-Metric Multidimensional Scaling) and RMDS (Robust Multidimensional Scaling) are both techniques used for dimensionality reduction and visualization of data in lower-dimensional spaces. NMDS focuses on preserving dissimilarity relationships between data points in the lower-dimensional space. It's used to represent the relative dissimilarity or similarity between data points while maintaining their rank order. RMDS, on the other hand, aims to perform MDS while accounting for the presence of outliers. It incorporates robust statistics to mitigate the impact of outliers on the resulting configuration.

Distances vs. Ranks. NMDS considers ranks rather than exact distances. It converts dissimilarities to ranks and then estimates a configuration based on these ranks. This transformation helps reduce the influence of outliers. RMDS uses the original dissimilarity or

distance matrix to construct the configuration. It aims to maintain the relative distances between data points as closely as possible.

Shepard Diagram Analysis

For the shepard diagram, a straight line should ideally be drawn between the observed proximities and the anticipated proximities in the shepard diagram, which is a scatter plot of the distances between points in the MDS plot against the observed proximities. A degenerate solution may be found if the shepard diagram resembles a stepwise or stair-case function. It also shows the relationship between the dissimilarity matrix and the distances in the reduced-dimensional space. The dissimilarities are plotted on the x-axis, while the corresponding distances in the reduced-dimensional space are plotted on the y-axis. The shepard graph is primarily useful for evaluating the best fit and estimating the dissimilarities of any solution. It checks the goodness fit, if it is a Linear diagonal graph, it is perfect fit. It shows us that points which align closer to the diagonal line indicate that they are a good fit while deviations suggest expansion or compression [12].

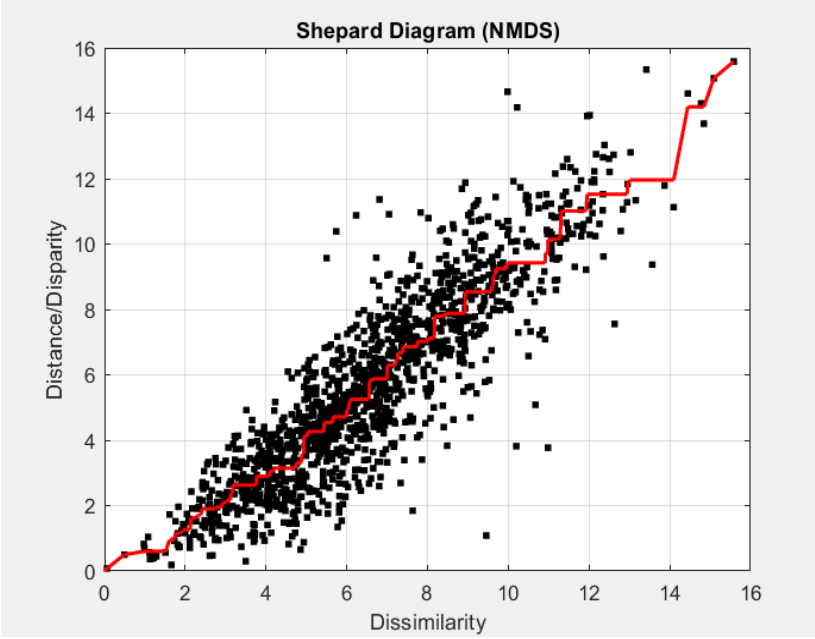

Figure 3. Shepard Diagram for NMDS Analysis of MBA Dataset

The Shepard graph is rarely used in non-metric MDS. Instead of explicitly retaining distances, NMDS uses rank or ordinal data resulting from differences. As a result, it is uncommon to use or regard the Shepard graph as a main tool for evaluating the goodness-of-fit. A linear Shepard graph is a positive outcome in MDS/NMDS analysis, indicating that the reduced-dimensional representation aligns well with the original dissimilarities. The inference in Figure 3 is that the distance and dissimilarity is not as directly applicable as the normal graph, which is generally used to evaluate clusters, patterns, and relationships among things. Deviations from the linear diagonal line indicate how well or poorly the dissimilarities are preserved. Points below the diagonal line suggest compression, where the reduced distances are smaller than the dissimilarities, indicating objects appear more similar in the reduced space. Points above the diagonal line indicate expansion, where the reduced distances are larger than the dissimilarities, suggesting objects appear more dissimilar in the reduced space [2].

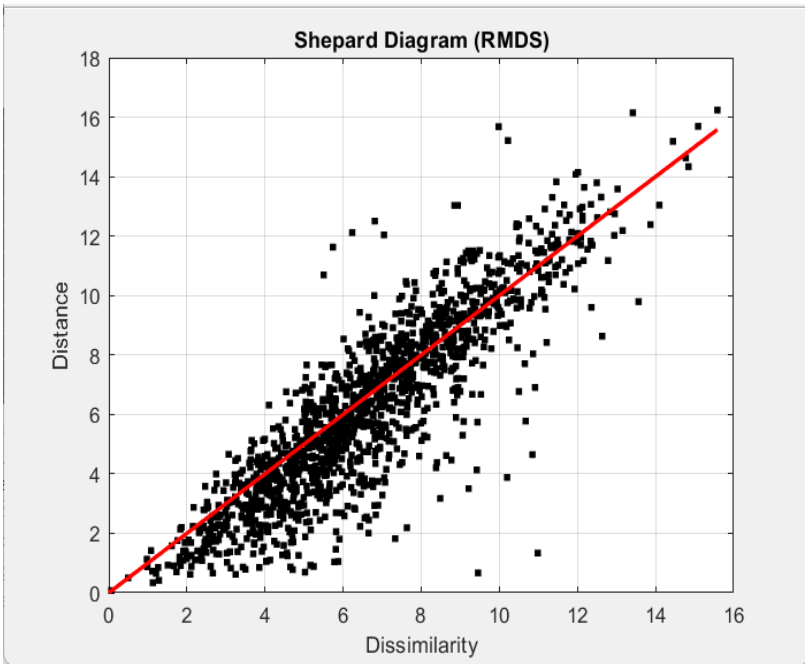

Figure 4. Shepard Diagram for RMDS Analysis of MBA Dataset

RMDS solution in the shown Figure 4 is a perfect fit if the graph's points ought to show a strong positive linear relationship. This implies that the mapped distance, which reflects the relative distances between objects, should grow as the observed dissimilarity increases. Here in

this case, the dissimilarity data may contain outliers or errors that cause certain points to diverge from the general linear trend. The link between observed differences and mapped distances is depicted visually in the Shepard graph in RMDS [2]. Thus, RMDS Shepard diagram is a good fit for MBA Dataset.

Co-Plot Analysis

The Coplot approach is used to reveal the relationships between a group of variables to get a better understanding of the factors that make up the MBA dataset. Coplot graph is used to decide which potential variables should be eliminated before beginning the traditional variable selection techniques [14]. Co-plot NMDS provides a detailed understanding of the connections between the Universities and the concentrations.

Co-Plot NMDS analysis in Figure 5 will provide a more thorough investigation and understanding of the connections between the Concentration, Education Month, University and Rank of the Universities. It explains that the Rank and Education Months of the Universities are not much co-related to each other. Similarly, the Universities and the Rank of the Universities are closely related to each other. It indicates that both of them are inversely related and analysis shares less angles with each other who shares similarities. The Rank and the Education Months share angles 180 degrees which means they have high level of dissimilarities. It can demonstrate graphically the similarities or differences between samples in a small-scale environment.

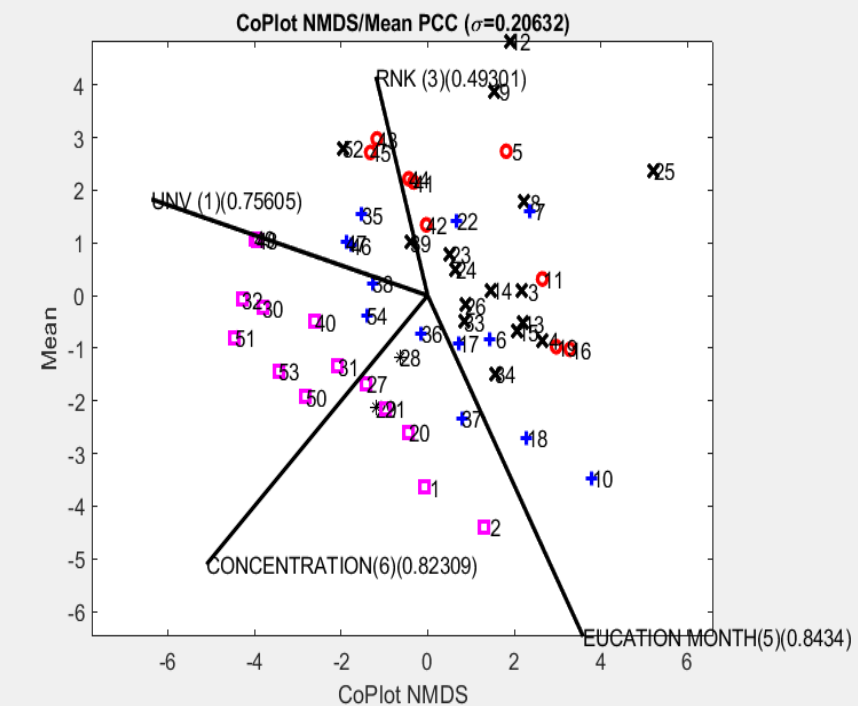

Figure 5. Classical Coplot Analysis of MBA Dataset

Co-Plot RMDS [2,4] statistics are suited to provide robustness against outliers. The RMDS analysis in Co-plot is like the NMDS where the variables are nearby [1]. RMDS analysis as shown in Figure 6 for Coplot is like NMDS as the interpretation is the same. The variables which are nearby and have less angles are mostly like each other and those which have opposite angles have dissimilarities and seem to be unrelated to each other. The visualization accurately reflects the relationships between the variables. It shows the connections between the Concentration, Education Month, University and Rank of the Universities. The Rank and the education month here share angles 180 degrees which means they have high level of dissimilarities. It can demonstrate graphically the similarities or differences between samples in a small-scale environment. The Rank and Education months are inversely related to each other.

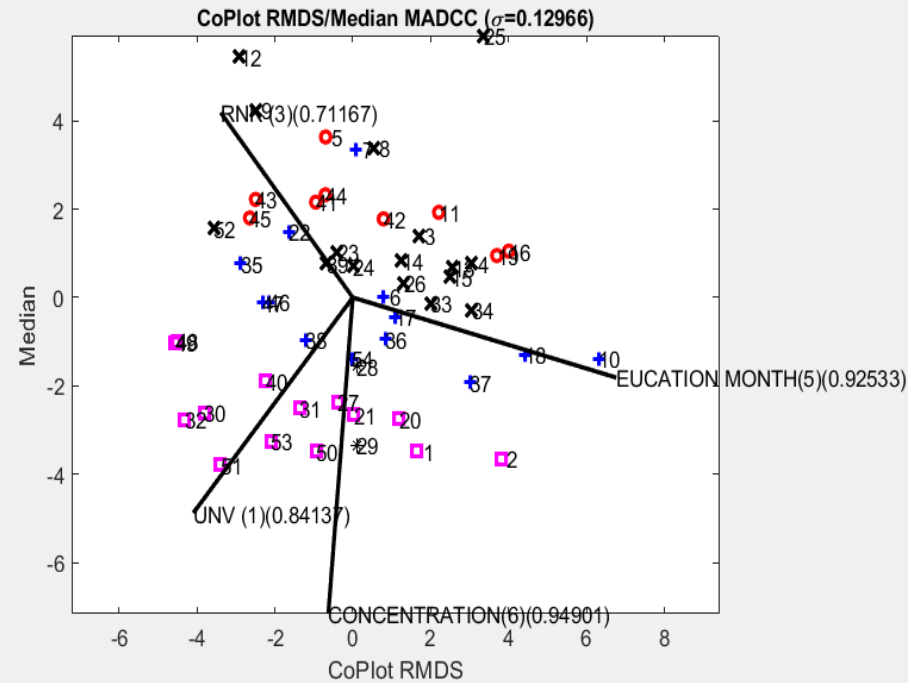

Figure 6. Robust Coplot Analysis of MBA Dataset

3.2 Scenario 2: Ecological Community Analysis

3.2.1 Description of the Dataset

Ecological communities are characterized by intricate interactions between species and their environment. Ecological community analysis often involves comparing different communities and their responses to environmental factors such as temperatures, moisture, etc. Ecological community analysis can help people gain the knowledge in species adaptations and sensitivities and address the questions like “how species relationships change across different conditions or environmental gradients?”, and “how different species respond to environmental gradients or disturbance regimes?” To assess the health and resilience of ecosystems and identify the patterns in species community across the U.S., we focus on addressing the following questions: (1) The similarity and dissimilarities of species across various communities. (2) The co-relation between the species of different communities. (3) The relationship of the environmental factors to the species. This sample dataset of the ecological species includes 14 observations (i.e., species) and 9 variables. The variables are Name of Species (SPNM), Number of Species (SPNR),

Species Richness (SR) identifying the abundance of species which is unique count of the species in a region, PH values (PH) measuring the soil's acidity or alkalinity whose values range from 3-13 where 3 is more acidic and 13 is more alkaline, Moisture content (MOISTURE) indicating the percentage of moisture in the soil, Elevation (ELEV) showing the elevation above sea level of each site in meters, Precipitation (PREC) indicating the amount of precipitation at each site in millimeters, Temperature (TMP) of the region where the species is located, and Region (REGION) representing the area of the species being located. Regions located in northeast are coded by 1, southeast by 2, Midwest by 3, southwest by 4, and west by 5. Regions are used as color code in identifying clusters of observations and possible outliers. This data is stored in csv file in (Table 2) and processed by MATLAB (MathWorks Inc, 2016).

Table 2: Sample Dataset of Ecological Species across Regions in the United States

SPNM (1)	SPNR (2)	SR (3)	PH (4)	MOISTURE (5)	ELEV (6)	PREC (7)	TMP (8)	REGION (9)
White-tail	10	15	3	20	500	800	25	1
Bison	8	12	5	40	200	600	28	2
American	12	20	7	82	300	1200	22	1
Gila Mo	7	10	9	10	800	100	30	3
Red Sq	9	14	5	30	600	700	26	2
Prairie Do	11	18	7	60	150	550	23	1
Great Blu	6	9	3	90	400	900	29	3
Gray Jay	10	16	5	25	700	1000	27	2
Muskrat	15	21	6	89	350	50	24	1
Bobolink	5	8	8	50	180	700	31	4
Desert Ter	8	12	9	95	950	50	39	5

American	14	19	10	87	350	650	32	4
Sidewinde	6	9	12	12	1000	800	40	5
Black-tail	11	17	13	77	180	1500	30	3

3.2.2 Interpretation of Graphs

NMDS and RMDS Analysis

Non-Metric Multidimensional Scaling (NMDS) analysis explains the relationships and patterns within high-dimensional data by projecting it into a lower-dimensional space [1]. It reduces the dimensionality of the data while retaining the most important information about the relationships between data points. The areas that share the same color and shape are in the same area. For example, as demonstrated in Figure 7 point 6 and 18 have similarities since they share the same region while 33 and 24 have differences since they are in different regions. 29 and 34 share the same region since they have the same blue color, they have many similarities in characteristics though they are far away from each other. It can be inferred that all the similar colors have similarities but if they are far from each other, they may not have many similar characteristics.

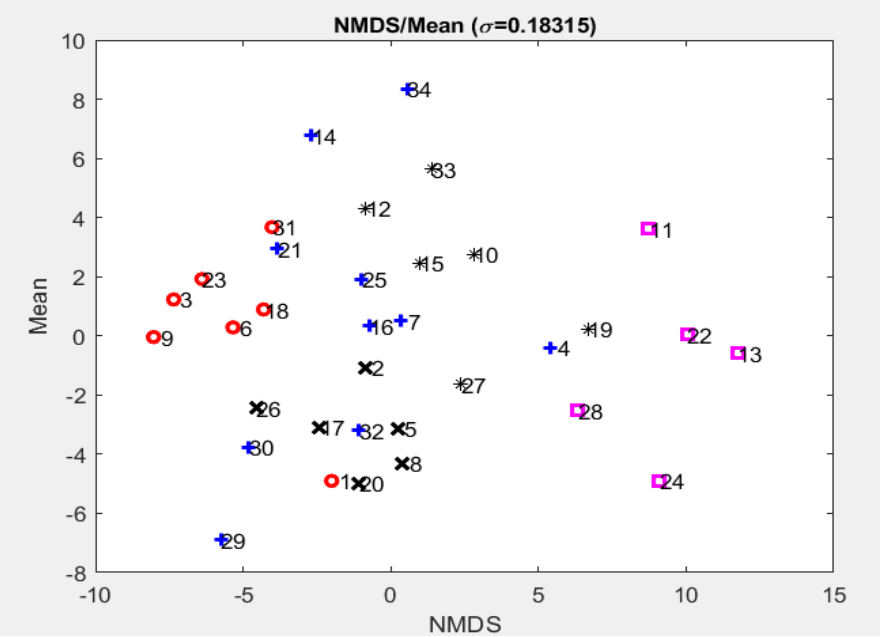

Figure 7: Non-metric MDS Analysis of Species Dataset

Robust Multidimensional Scaling (RMDS) explains the relationship between dissimilarities and distances. While the dissimilarities are converted into meaningful distances [2]. Points that are closer together in the MDS plot are more like each other in terms of the original dissimilarity matrix, while points that are farther apart are more dissimilar. In RMDS, the pairwise distances between the points in the lower-dimensional space approximate the dissimilarities in the original distance matrix as closely as possible. Despite RMDS graph and NMDS graph look similar, points in Figure 8 are plotted based on ranks whereas points in Figure 7 are plotted based on actual distances. We used Step 1 and Step 2 to standardize the dataset and to change the scale and reduce the impact of the outliers for both Non-metric MDS and Robust MDS Analysis. For example, in Figure 2, point 22 and 13 have similarities since they share the same region and are nearer to each other, so they share the same characteristics as well. While point 4 and 19 have differences since they share a different region though they are nearer to each other.

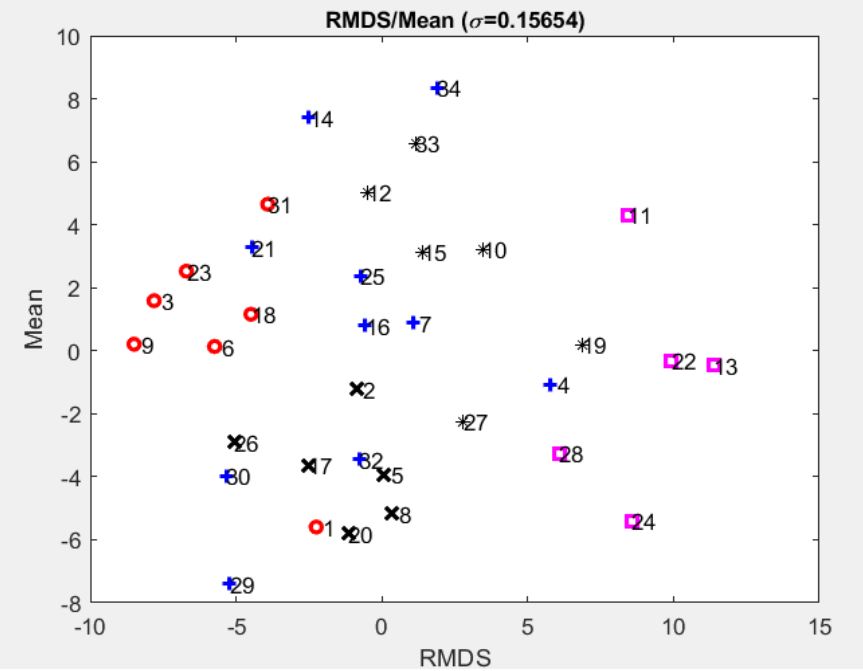

Figure 8:Robust MDS Analysis of Species Dataset

Shepard Diagram Analysis

Shepard diagram is a scatterplot where the x-axis represents the original dissimilarities, and the y-axis represents the corresponding residuals. Each point in the Shepard diagram corresponds to a pair of data points. The closer a point is to the diagonal line, the better the fit between the original dissimilarities and the distances in the MDS/NMDS plot. If the graph is a linear diagonal graph, the quality of fit is checked, and the fit is flawless. The deviations in the graph suggest expansion or compression. The points matching more closely to the diagonal line indicate that they are a suitable fit [12]. Expanding points are those where the reduced distances are greater than the differences, indicating that objects appear more dissimilar in the condensed space. In Figure 9 , there is a deviation from the diagonal line indicating a few numbers of points are dissimilar to each other. Deviations from the diagonal line indicate how well or poorly the dissimilarities are preserved. Figure 9 indicates that points below the diagonal line suggest compression, where the reduced distances are smaller than the dissimilarities, indicating objects appear more similar in the reduced space. Points above the diagonal line indicate expansion, where the reduced distances are larger than the dissimilarities, suggesting objects appear more

dissimilar in the reduced space. The Shepard diagram generally consists of Observed Dissimilarities which are based on measurements or observations made in the original dataset, the Shepard diagram often displays the differences or separations between objects or things while Fitted Dissimilarities which are the distances determined in the reduced dimensional space using the MDS technique. Thus, the closer the dots are to the diagonal line, the better the fit between the observed and fitted dissimilarities.

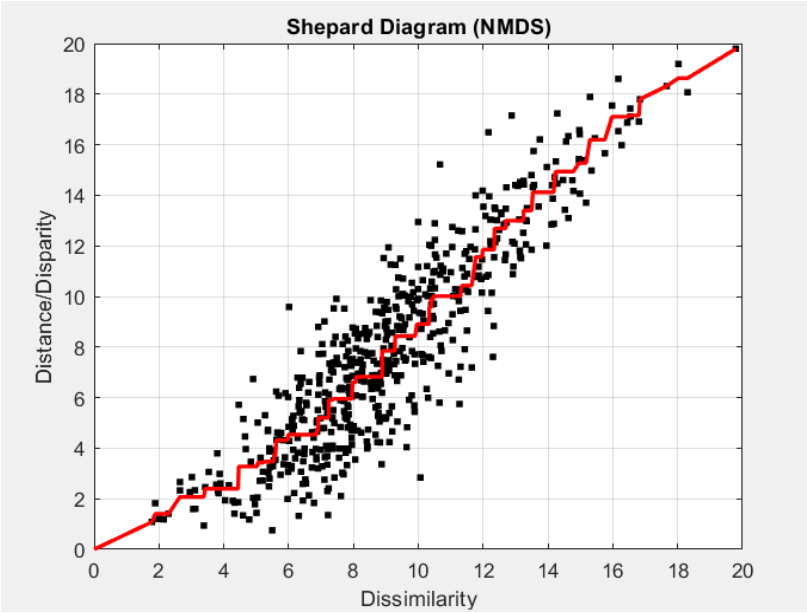

Figure 9: Shepard Diagram for NMDS Analysis of Species Dataset

If the shepard graph is linear, it indicates that the data may be represented in the specified dimensionality (2D or 3D) for the MDS analysis. The reduced-dimensional space maintains the fundamental structure of the data, according to the linear fit. A linear Shepard graph is a positive outcome in MDS/NMDS analysis, indicating that the reduced-dimensional representation aligns well with the original dissimilarities. It accurately represents the relationships and similarities among data points. The inference in the Figure 10 is that the distance and dissimilarity are closely related to each other as a linear graph [2]. Shepard graph may deviate from a perfect linear connection when there is noise or outliers in the data. It seems to be a good fit if the graph points to show a strong positive linear relationship. The mapped distance which reflects the relative

distances between objects, should grow as the observed dissimilarity increases. The link between observed differences and mapped distances is depicted visually in the shepard graph in RMDS as shown in Figure 10. Thus, the RMDS shepard diagram is a good fit for the species dataset.

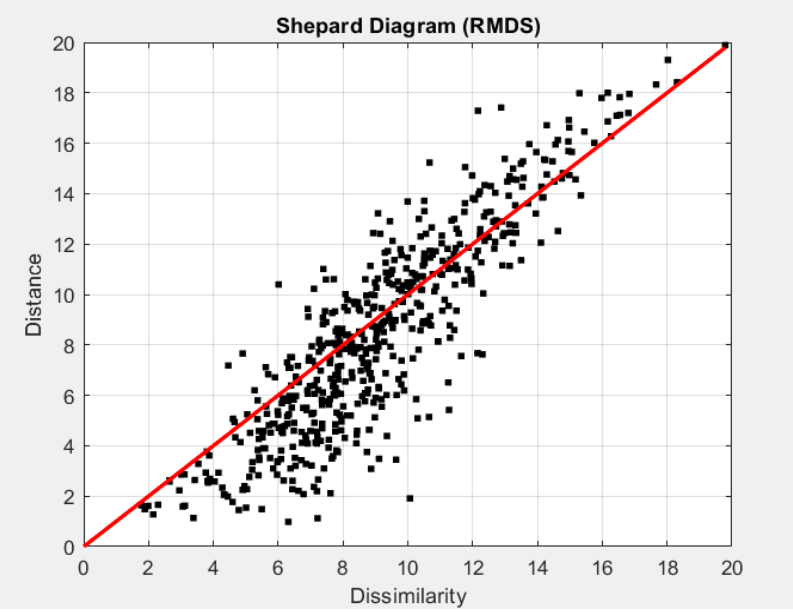

Figure 10: Shepard Diagram for RMDS Analysis of Species Dataset

Co-Plot Analysis

Co-plots explore the relationships between two variables and controlling for the influence of one or more additional variables. Co-plot NMDS provides a detailed understanding of the connections between the samples and the environmental variables [2]. Co-plot analysis NMDS in Figure 11 shows the connections between the Name of Species, Number of Species, Temperature, PH values of a soil, Moisture content in the soil, Precipitation, Elevation, Species Richness. The graph explains how Elevation and Species richness are like each other and how temperature and precipitation are related to each other. It indicates that with less temperature there is more precipitation and vice versa. and Analysis shares less angles with each other who shares similarities. The PH and the Moisture content in the soil share angles 180 degrees which means they have high level of dissimilarities. It can demonstrate graphically the similarities or differences between samples in a small-scale environment. The Co-plot offers a precise

illustration of the underlying relationships and patterns found in the collection. It is easier to spot outliers or unusual samples in the dataset. Samples known as outliers drastically depart from the general graph tends.

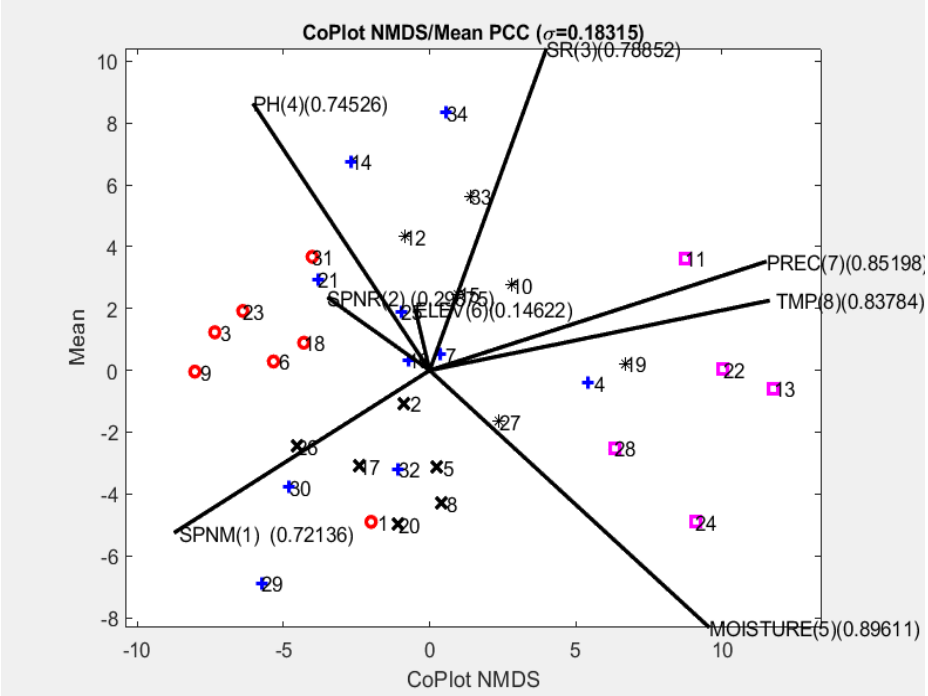

Figure 11: Classical Coplot Analysis of Species Dataset

Co-plot RMDS [2,4] are suited to provide robustness against outliers. The RMDS analysis in Co-plot is like the NMDS where the variables are nearby [1]. The visualization accurately reflects the relationships between the variables. The RMDS analysis has less angles which each other shares similarities as shown in Figure 12 The temperature and the species number are inversely related to each other. The Co-plot presents the connections between the Number of the species, PH value of the soil, Moisture Content in the soil, elevation of the region, temperature, and precipitation. The graph explains that the PH values of the soil and Number of the species are co-related to each other. Similarly, the PH values and the Moisture content in the soil are not so closely related to each other. It also infers that with the increase in Moisture content of the soil the PH values decreases and vice versa. The PH and the Moisture content in the soil share angles 180 degrees which means they have high level of dissimilarities. It can demonstrate graphically

the similarities or differences between samples in a small-scale environment. It can be observed that Species Richness is directly related to the Precipitation. The more amount of precipitation will increase the Richness of the Species.

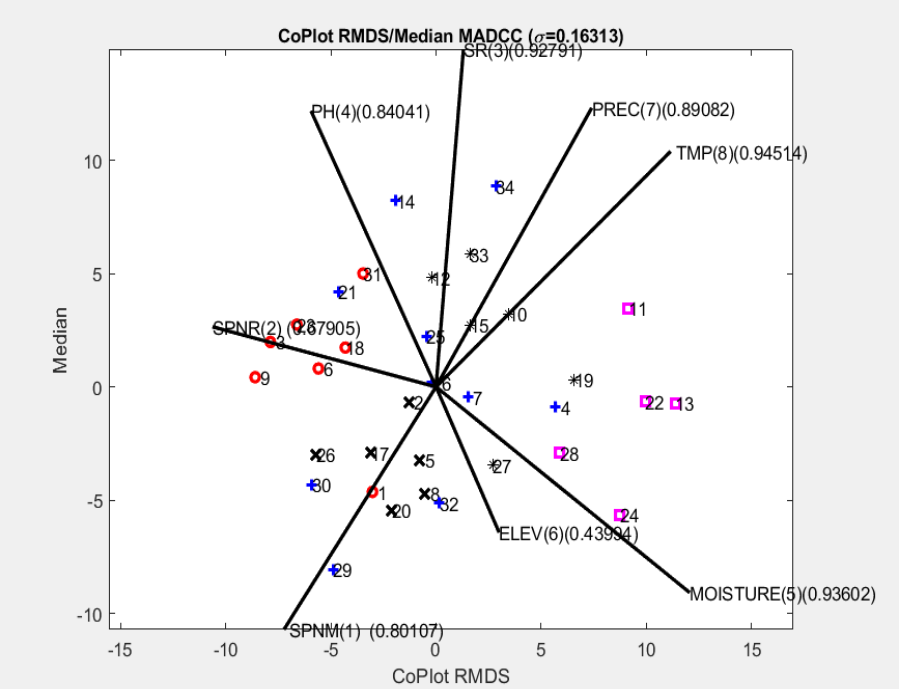

Figure 12: Robust Coplot Analysis of Species Dataset

CHAPTER 4: CONCLUSIONS AND LIMITATIONS

This paper suggests a graphical display method of Multidimensional Scaling. The main goal was to study how the Co-plot analysis works and how this analysis provides a visually informative way to explore the relationship between variables and observations.

In the Co-plot application in MBA dataset, we found out how relationships between key variables and how they change depending on various factors. In the ecological community data, we found how relationships between key variables change depending on demographic or other factors. This understanding can guide more effective decision-making. In many other domains, like agriculture and forestry, Co-plot analysis can be applied in resources management by considering how different factors (e.g., weather, soil type) affect crop yield or tree growth.

Co-plot analysis is often more effective with larger datasets and becomes less robust in small datasets. Co-plot may become complex, especially when variable has many categories or when the relationships between variables are intricate. Interpretation may require advanced statistical and domain knowledge. There are some limitations to using coplots in MDS: (1) Complexity of Visualization: Coplots can become complex and difficult to interpret, especially when dealing with high-dimensional data. MDS often reduces dimensions for visualization purposes, but coplots might not effectively capture all the nuances and relationships within the reduced dimensions; (2) Difficulty in Representing Higher Dimensions: MDS may reduce dimensions to two or three for visualization, but coplots might struggle to represent higher dimensions effectively. As a result, important information might be lost or not adequately portrayed in the visualization; (3) Overplotting: When dealing with many points or variables, coplots might suffer from overplotting, where data points overlap and make it challenging to discern individual relationships or patterns.

REFERENCES

1. Atilgan, Y. and Atilgan, E. (2017) RobCoP: A Matlab Package for Robust CoPlot Analysis. Open Journal of Statistics, 7, 23-35. doi: 10.4236/ojs.2017.71003.
2. Borg, I., Groenen, P. J., & Mair, P. (2018). Applied multidimensional scaling: A comparison of approaches and algorithms. Springer Science & Business Media.
3. Borg, I. and Groenen, P.J. (2005). Modern Multidimensional Scaling: Theory and Applications. Springer Science & Business Media, Berlin.
4. Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531-554.
5. Forero, P. and Giannakis, G. (2012) Sparsity-Exploiting Robust Multidimensional Scaling. IEEE Transactions on Signal Processing, 60, 4118-4134. https://doi.org/10.1109/TSP.2012.2197617
6. Giladi, R., Spector, Y., & Raveh, A. (1996). Multidimensional scaling: An analysis of 1980-1990 computers. European Journal of Operational Research, 95(2), 439-450. https://doi.org/10.1016/0377-2217(95)00296-0
7. Giladi, Ran & Spector, Yishay & Raveh, Adi, 1996. "Multidimensional scaling: An analysis of 1980-1990 computers," European Journal of Operational Research, Elsevier, vol. 95(2), pages 439-450, December.
8. Goldreich, Y. and A. Raveh, "Coplot Display Technique as an Aid to Climate Classification," Geographical Analysis, 25 (1993), 337-353
9. Lipshitz, G. and A. Raveh, "Application of the Co-plot Method in the Study of Socioeconomic Differences Among Cities: A Basis for a Differential Development Policy," Urban Studies, 31 (1994), 123- 135

10. Lipshitz, G. and Raveh, A. (1998) Socio-Economic Differences among Localities: A New Method of Multivariate Analysis. Regional Studies, 32, 747-757. https://doi.org/10.1080/00343409850119436
11. Raveh, A. (2000) Co-Plot: A Graphic Display Method for Geometrical Representations of {MCDM}. European Journal of Operational Research, 125, 670-678. https://doi.org/10.1016/S0377-2217(99)00276-3
12. Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. In Proceedings of the 1968 23rd ACM national conference (pp. 517-524). ACM.
13. Shevlyakov, G. and Smirnov, P. (2011) Robust Estimation of the Correlation Coefficient: An Attempt of Survey. Austrian Journal of Statistics, 40, 147-156.
14. Shoval, N. and Raveh, A. (2004) Categorization of Tourist Attractions and the Modeling of Tourist Cities: Based on the Co-Plot Method of Multivariate Analysis. Tourism Management, 25, 741-750. https://doi.org/10.1016/j.tourman.2003.09.005
15. Talby, D. (2015) The Visual Co-Plot, Version 5.5.
16. The MathWorks Inc. (2016) MATLAB—The Language of Technical Computing, Version 2016a. The MathWorks Inc., Natick.
17. CoPlot: A tool for visualizing multivariate data in medicine – https://doi.org/10.1002/sim.3078