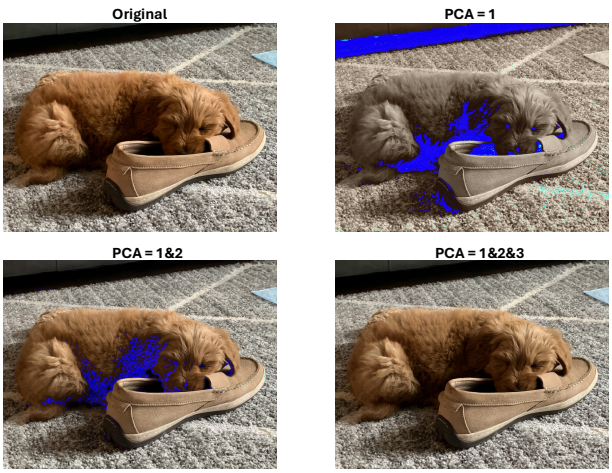
Graphical Data Analysis 02

EN5423 | Spring 2024

w03_graphical_02.pdf
(Week 3)

Contents

(Tomy Tjandra)



1 Scatterplots and Enhancements

The two-dimensional scatterplot is a well-known tool for analyzing the relationship between two variables. It typically raises three key questions:

• The nature of the relationship: Is it linear, curved, or piecewise linear?

• Consistency across groups: For data from different groups, defined perhaps by location or time, does the relationship hold constant or does it vary?

• Variability: Does the spread in the relationship remain steady across the data range?

Enhancements like smoothing curves can help clarify these aspects, providing a clearer understanding than scatterplots alone. These enhancements and their applications to scatterplots are further explored in subsequent sections.

1.1 Evaluating Linearity (*Exercise 1*)

• Scatterplots reveal a strong but complex relationship between the natural log of nitrate concentration and discharge in the Iowa River, suggesting non-linearity (**Figure 1**).

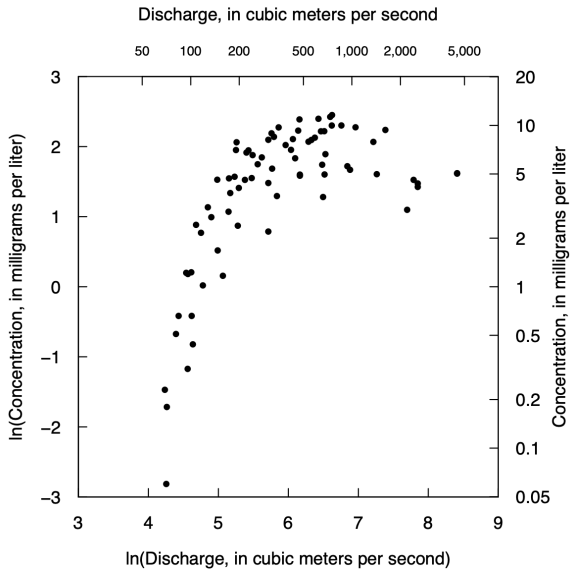


Figure 1. Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, for the months of June, July, August, and September of 1990–2008.

• Smooth curves enhance scatterplots by providing a locally adjusted trend line, highlighting the central trend without undue influence from outliers. Smoothing emphasizes data near each point rather than distant values, avoiding misinterpretation common in regression models (**Figure 2**).

• Preferred smoothing method: Local Polynomial Regression Fitting ("loess"; Locally Estimated Scatterplot Smoothing), effective for large datasets and uncovering hidden patterns (Figure 2).

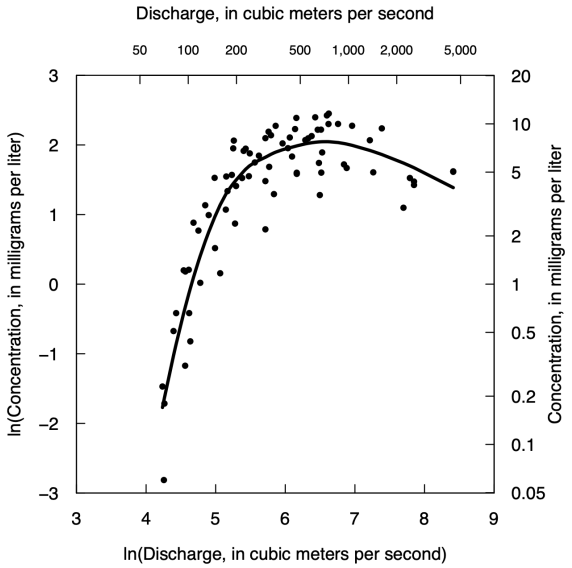


Figure 2. Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September. The curve represents a loess smooth of these data.

• Loess smoothing on the Iowa River data shows a linear relationship up to about 200 m^3/s discharge, then becomes less steep, indicating a change in the rate of increase in nitrate concentration with discharge.

• Beyond about 1,000 m^3/s, nitrate concentrations appear to decrease with increasing discharge, a pattern more easily observed with the smooth curve.

• Loess smoothing provides an exploratory tool that requires no predetermined model, allowing data to dictate the relationship's shape.

• Smoothed scatterplots are valuable for both analyzing and presenting data, especially when comparing multiple groups or large datasets, as shown in studies of arsenic concentration and pH in groundwater (**Figure 3**).

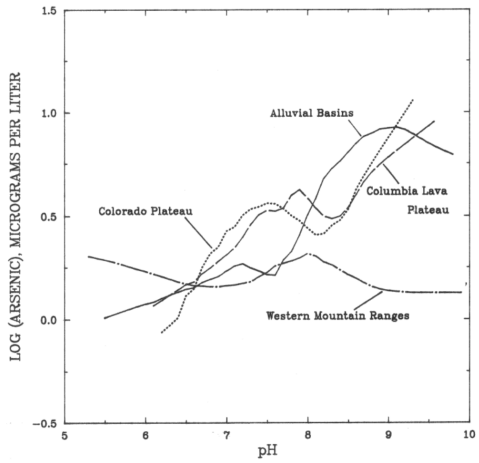


Figure 3. Loess smooths representing dependence of log(As) on pH for four areas in the western United States (from Welch and others, 1988).

1.2 Evaluating Differences in Central Tendency on a Scatterplot (*Exercise 2*)

• A scatterplot of conductance versus pH from low-flow samples in Ohio's coal mining region shows variations across streams draining unmined land, reclaimed land, and abandoned mined land, as depicted in **Figure 4** (Helsel, 1983).

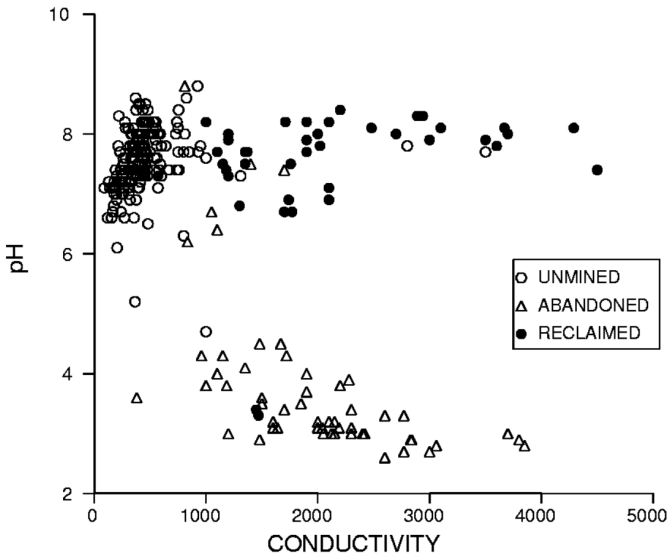


Figure 4. Scatterplot of water-quality measures draining three types of upstream land use (from Helsel, 1983).

• Polar smooths, which encapsulate 50% or 75% of data for each land-use type, are used for clarity, revealing distinct patterns for each group without assuming a specific shape, as shown in **Figure 5**. This method involves transforming data into polar coordinates, applying a loess smooth, and then converting back to original units.

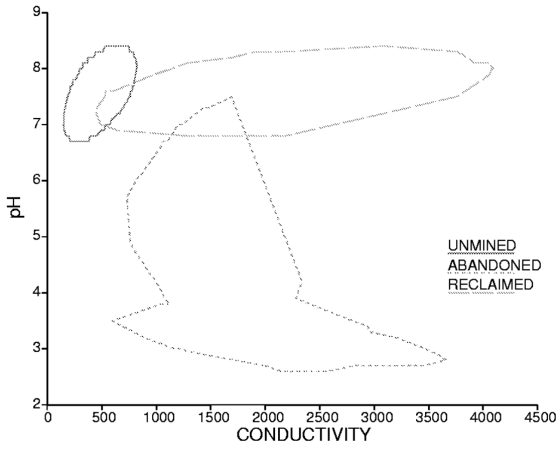


Figure 5. Polar smooths with 75 percent coverage for the three groups of data seen in figure 2.24, from Helsel (1983).

• The polar smooth of abandoned lands indicates possible subgroups, with differences in pH linked to underlying geologic units, suggesting the importance of including geologic type in chemical behavior analyses.

• Polar smooths prove beneficial in exploratory data analysis, especially with large datasets where traditional scatterplots become too cluttered, as exemplified in **Figure 5**, providing clearer differentiation between groups.

• Applying this technique to NO23 concentration data for the Illinois River across different seasons reveals distinct concentration patterns for warm versus cold months at various discharge levels, with loess smooths highlighting differences and similarities across seasons in **Figures 1**, **2**, and **6**.

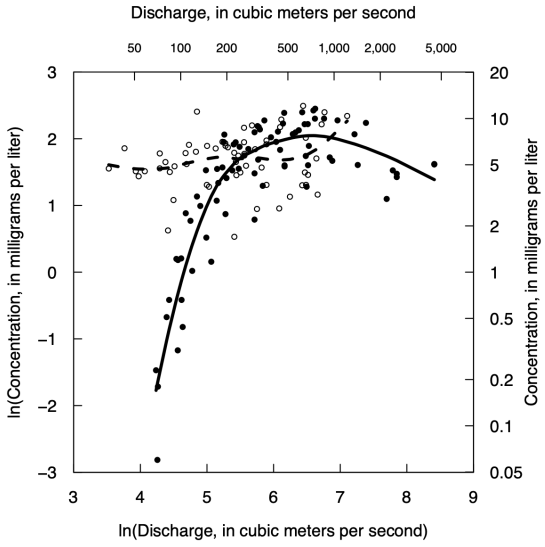


Figure 6. Dissolved nitrate plus nitrite concentration as a function of discharge, Iowa River, at Wapello, Iowa, water years 1990–2008 for the months of June, July, August, and September (filled circles) or the months of January, February, March, and April (open circles). The solid curve is a loess smooth of the warm season data; the dashed curve is a loess smooth of the cold season data.

• Using smooths to separate datasets on the same plot clarifies patterns that might be obscured in crowded scatterplots, aiding in the selection of flexible analysis methods that account for temporal and seasonal variations, thereby underscoring the need for adaptable approaches like Weighted Regressions on Time, Discharge, and Season (WRTDS).

1.3 Evaluating Differences in Spread (*Exercise 3*)

• Understanding the spread of data on scatterplots is crucial, as constant variance (*homoscedasticity*) is a key assumption in regression analysis and many parametric hypothesis tests. Variability changes can highlight differences in data repeatability more than method bias.

• Judging changes in spread is challenging due to visual perceptions; the presence of outliers and variations in data density across the plot can falsely suggest variability changes. The eye tends to misjudge the correct vertical distance between data points and the central trend line.

• Chambers et al. (1983) suggest a graphical method to evaluate spread changes by first applying a loess smooth to the data to establish a central trend, then calculating the absolute differences between each data point and this smooth. This approach quantifies variability and helps visually identify changes in spread across the scatterplot.

• In **Figure 2**, a smooth is computed using loess or some other smoothing method. For our purposes here we will call this the middle smooth. The absolute values of differences di between each data point and the smooth at its value of $\boldsymbol{x}$ is a measure of spread:

$$d_i = |y_i - l_i| \tag{1}$$

Where, l_i is the value of the loess smooth at x_i, and y_i is the true value at x_i.

• For the Iowa River NO23 data, plotting these absolute differences against discharge revealed that variability decreases with increasing discharge up to about 200 m^3/s, beyond which it remains constant. This pattern of *heteroscedasticity* indicates that models assuming constant error variance may not be suitable for this dataset, pointing towards more flexible statistical approaches like WRTDS model (Weighted Regressions on Time, Discharge, and Season [WRTDS] introduced in later chapter).

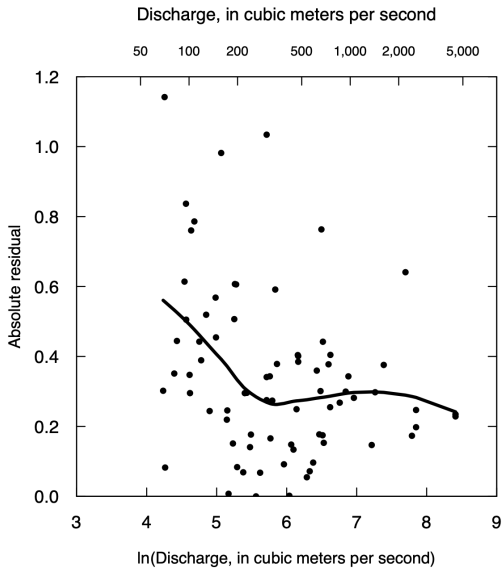


Figure 7. Absolute residuals from the loess smooth of ln(NO23) concentrations versus ln(discharge), Iowa River at Wapello, Iowa, for the warm season (June, July, August, and September) 1990–2008.

2 Graphs for Multivariate Data

• Boxplots are great for showcasing data characteristics of a single variable and identifying outliers. Scatterplots excel in displaying relationships between two variables, highlighting unusual x-y relations.

• Analyzing relationships involving *more than two variables* is often necessary, especially for comparing *groups* or *identifying outliers* within multivariate data.

• Graphical methods provide valuable insights into multivariate relationships, complementing formal hypothesis testing.

• (Advanced) In water-quality studies, Stiff and Piper diagrams are commonly used multivariate graphical methods. For a deeper exploration of multivariate graphical techniques, refer to works by Chambers et al. (1983) and Everitt (2007).

2.1 Parallel Plots

• Parallel plots (or profile plots) use separate, parallel axes for each variable, connecting each observation's variables with straight lines to form a profile, allowing for easy comparison of observations.

• Each characteristic on a parallel plot is scaled from its minimum to maximum value across the dataset, facilitating direct comparison between sites, as shown in Figure 2.28 for two contrasting sites.

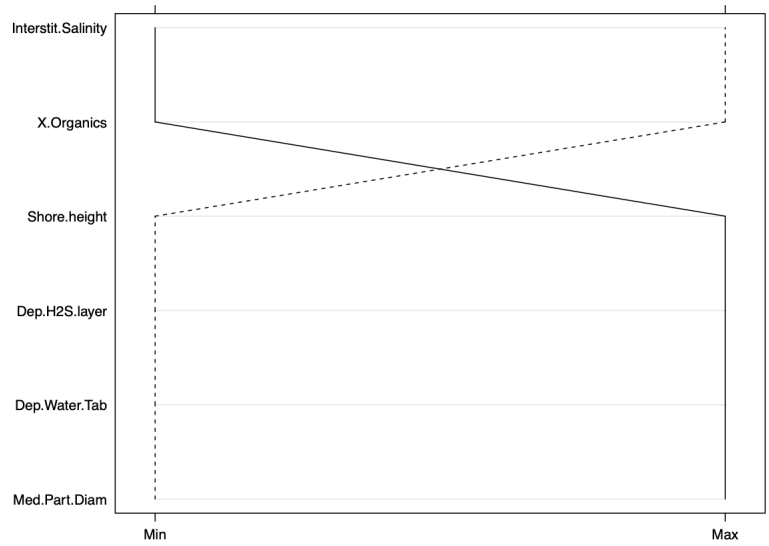


Figure 8. Parallel plot of six basin characteristics at a low salinity site (solid line) and a high salinity site (dashed line) (from Warwick, 1971).

• When plotting all sites, as seen in **Figure 9**, parallel plots can resemble "spaghetti plots" due to overlapping lines, highlighting groups with distinct profiles and outliers but complicating comparisons.

• The arrangement of characteristics on the axes significantly affects the plot's clarity, with adjacent characteristics being easier to compare than those placed further apart.

• Stiff diagrams, a variant of parallel plots used in water quality studies, plot milliequivalents of major water constituents on either side of a centerline, allowing for shape-based comparison of water samples, as demonstrated with groundwater samples from Wyoming in **Figure 10**.

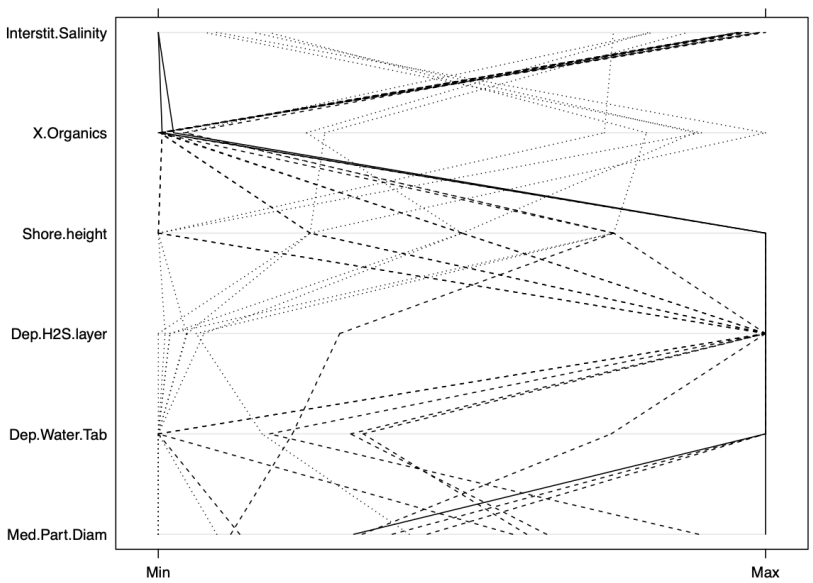


Figure 9. Parallel plot of six basin characteristics at the 19 sites of Warwick (1971).

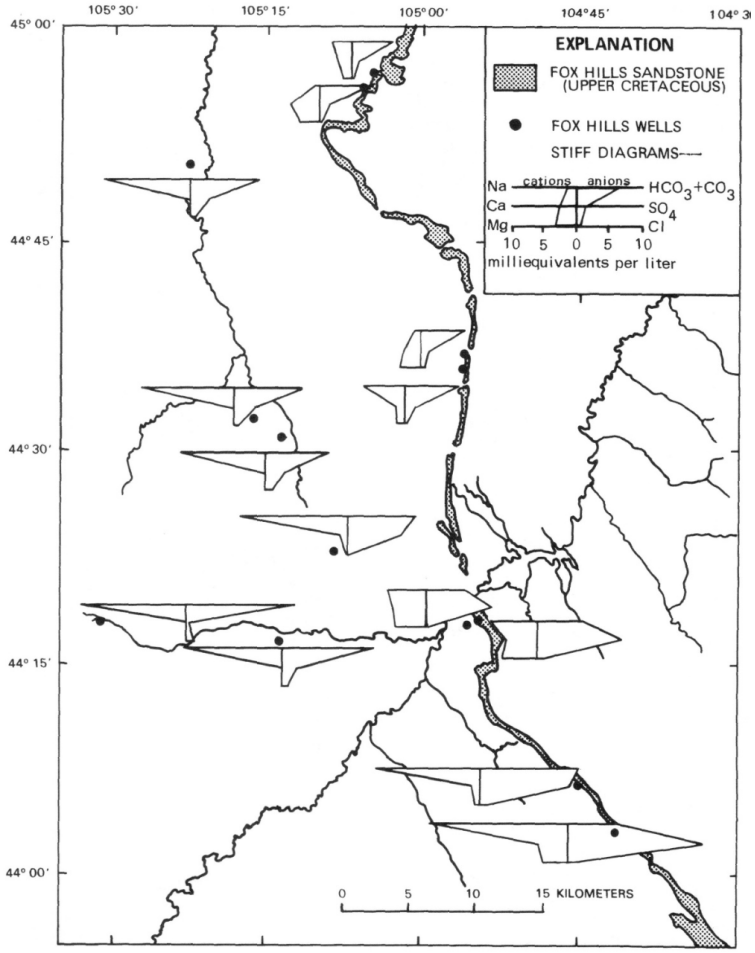


Figure 9. Stiff diagrams used to display differences in water quality in the Fox Hills Sandstone, Wyoming (from Henderson, 1985).

2.2 Star Plots

• Star plots represent another method for visualizing multivariate data, where axes radiate from a central point, connecting each variable's value for an observation to form a star-like pattern.

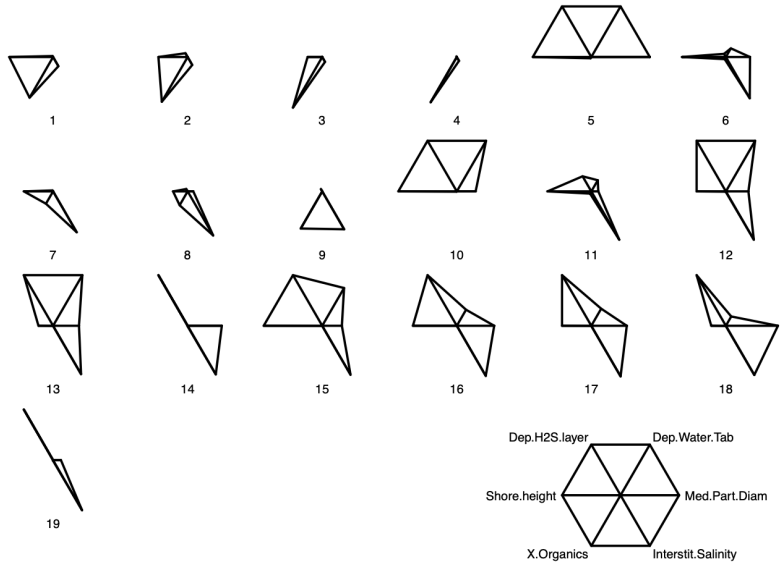


Figure 10. Star plots of site characteristics for 19 locations along the Exe estuary (from Warwick, 1971). Outliers such as sites 5 and 10 are seen to differ from the remaining sites owing to their low values for both interstitial salinity (Interstit.Salinty) and percent organics (X.Organics) composition.

• The angles between the plot's rays are determined by dividing 360° by the number of axes (variables), ensuring each variable is equally spaced around the circle.

• For clearer visual differentiation, related characteristics should be plotted on adjacent rays, making it easier to spot patterns or unusual observations within the dataset.

• Observations with distinct characteristics will appear as stars with notably different shapes, highlighted by variations in ray lengths.

2.3 Trilinear Diagrams (*Class Excersize and Quiz*)

• Trilinear diagrams represent a method used in geosciences since the early 1900s to display data where three variables sum to 100 percent, shown as a point on a triangular diagram.

• An example from the USGS's Groundwater Ambient Monitoring and Assessment Program illustrates cation compositions in California's Sierra Nevada study unit on a trilinear diagram (**Figure 11**), calculating percentage compositions for each cation based on milliequivalents.

2.4 Scatterplot Matrix (*Exercise 4*)

• For multiple variables (p), visualizing pairwise relationships involves creating scatterplots for each of the $\frac{p(p-1)}{2}$ variable pairs.

• These pairwise scatterplots are collectively displayed in a matrix format, allowing for an overview of all potential relationships.

• Although individual plots in the matrix provide limited detail, this approach enables identification of related variables, distinguishing between linear and nonlinear relationships.

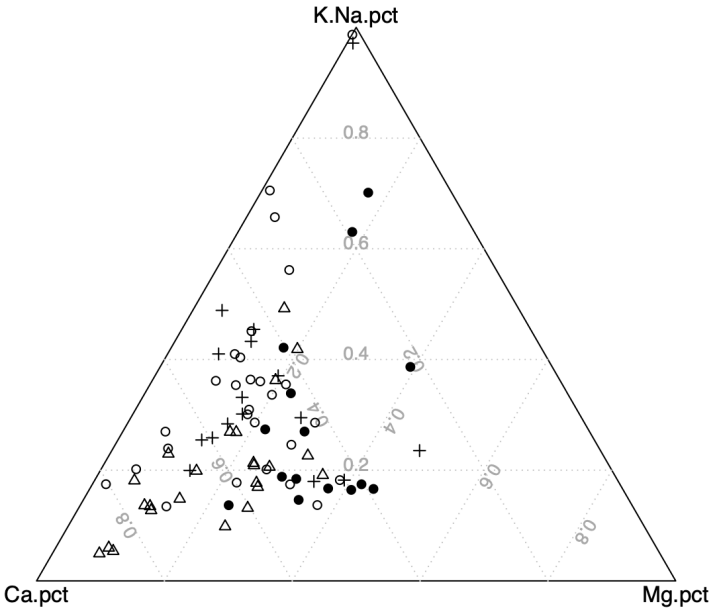


Figure 11. Trilinear diagram for groundwater cation composition in four geologic zones (each shown with a different symbol) of the Groundwater Ambient and Monitoring Assessment (GAMA) Program Sierra Nevada study unit (from Shelton and others, 2010). Units are percent milliequivalents (pct).

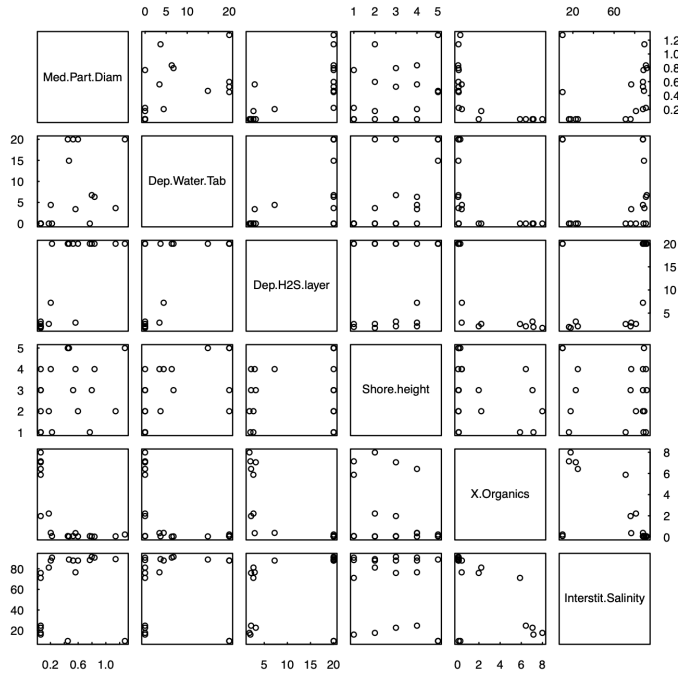


Figure 12. Scatterplot matrix showing the relations between six site characteristics from Warwick (1971).

2.5 Biplots of Principal Components (*Exercise 4*)

• Principal components analysis (PCA) reduces multiple variables to two main axes for a simplified scatterplot, identifying the main patterns in multivariate data without losing crucial information.

• PCA transforms original variables into a new set of uncorrelated axes (principal components) that sequentially capture the most variance, starting with the axis that explains the most variance and proceeding in order of diminishing variance explained.

• Each observation can be located on the new set of principal component (*pc*) axes. For example, suppose principal components were computed for four original variables, the cations Ca, Mg, Na, and K. The new axes would be linear combinations of these variables, such as:

$$pc1 = 0.75\text{Ca} + 0.80\text{Mg} + 0.10\text{Na} + 0.06\text{K} \text{ a calcareous axis?}$$
$$pc2 = 0.17\text{Ca} + 0.06\text{Mg} + 0.60\text{Na} + 0.80\text{K} \text{ a Na + K axis?}$$
$$pc3 = 0.40\text{Ca} - 0.25\text{Mg} - 0.10\text{Na} + 0.10\text{K} \text{ a Ca versus Mg axis?}$$
$$pc4 = 0.05\text{Ca} - 0.10\text{Mg} + 0.10\text{Na} + 0.20\text{K} \text{ residual noise}$$

• Each observation in the dataset can be plotted based on its values on the principal component axes, revealing patterns, similarities, or outliers among observations. Principal components are calculated as linear combinations of the original variables, enabling interpretation of the new axes in terms of the original variables' contributions.

• A biplot combines two types of information: the locations of observations (or scores) on the first two principal component axes and vectors indicating the directions and magnitudes of the original variables, aiding in understanding the relationship between variables and observations.

• Observations plotted near each other in a PCA biplot share similar characteristics, while the direction and length of vectors for original variables show how these variables contribute to the principal components and their inter-correlations.

• PCA biplots are effective for visualizing complex multivariate relationships in a two-dimensional space, facilitating easier interpretation of how observations and variables relate within the dataset.

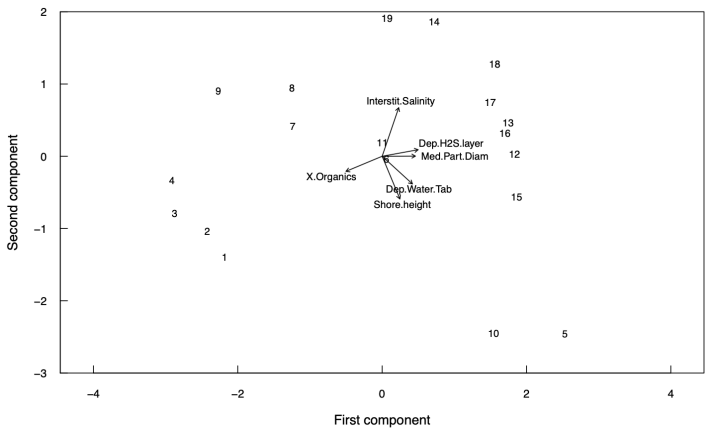


Figure 13. Principal component analysis (PCA) biplot of site characteristics along the Exe estuary (Warwick, 1971).

Note

When to Use PCA for Fitting Data:

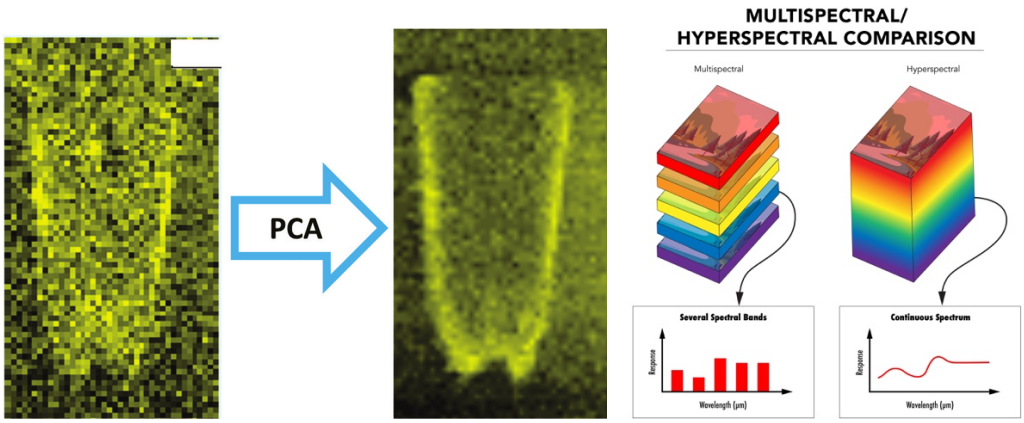
• **Noise Reduction**: PCA can be used to remove noise from data by discarding components that capture the noise rather than the signal.

• **Data Compression**: PCA reduces the dimensionality of data, which can be seen as a form of lossy data compression, making data storage and processing more efficient.

• **Visualization**: Lower-dimensional representations (e.g., using the first two or three principal components) allow for visual exploration of data.

• **Preprocessing for Other Analyses**: Reduced-dimension data can be used as input for other machine learning algorithms, improving performance by eliminating redundant features and focusing on the most informative aspects of the data.

It is important to remember that while PCA can simplify and denoise data, the fit to the original data using only a subset of components will not capture 100% of the variance unless all components are used. The choice of how many components to keep should balance the need for simplicity and efficiency against the desire to retain as much information as possible.



(Edmond optics)

2.6 Nonmetric Multidimensional Scaling

• Nonmetric Multidimensional Scaling (NMDS) was initially developed for psychology but has become widely used in ecology, offering an alternative visualization similar to PCA biplots.

• Unlike PCA biplots that use the ***two principal components*** with the most variance, NMDS utilizes ***all variables' information***, presenting a comprehensive view of data relations.

• A key difference from PCA biplots is that NMDS measures ***distances in ranks*** rather than original scales, resulting in arbitrary x and y axes without defined scales, making NMDS more of a data relations sketch.

• The NMDS plot for Warwick (1971) site characteristic data illustrates outliers and clusters effectively, indicating sites with similar characteristics and distinguishing them from others based on variables like depth to the water table and percentage organics. NMDS and PCA biplots complement each other in visualizing multivariate data, offering insightful first looks at complex datasets.

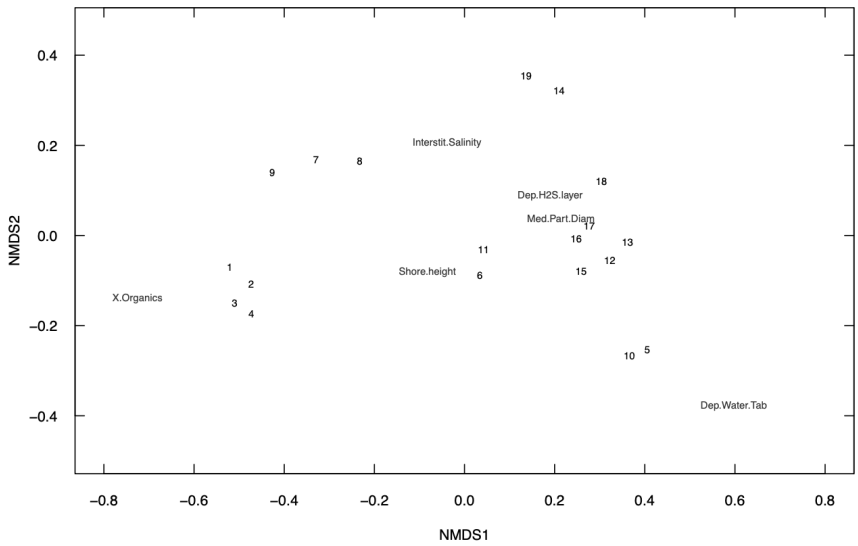


Figure 13. Nonmetric multidimensional scaling showing the relations among sites, and between sites and variables, using the six site characteristics of Warwick (1971).

Note
• **Use PCA** when dealing with quantitative data with linear relationships, aiming for dimensionality reduction while retaining the ability to interpret components in terms of original variables. • **Use NMDS** when exploring data with non-linear relationships, especially for ecological or psychological datasets, or when dealing with dissimilarity matrices, aiming for a more flexible visualization of data structure without the linear constraints of PCA. • Ultimately, the choice between PCA and NMDS will depend on the specific characteristics of your dataset and what you aim to achieve with your analysis. It's also not uncommon to try both methods to see which provides more insight into your particular dataset.

2.7 Three-dimensional Rotation Plots

• **Figure 14** showcases two orientations from the Exe site characteristic data, illustrating how different perspectives can reveal unique patterns and relationships among the variables.

• Rotating the data along all three dimensions provides insights that might not be apparent in static two-dimensional representations, enhancing understanding of the data's structure and dynamics.

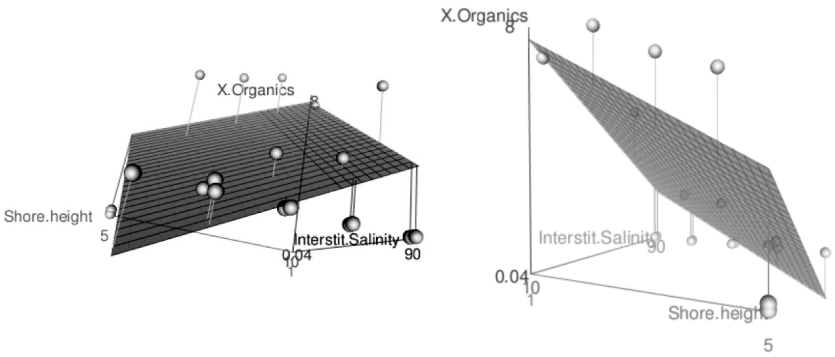


Figure 14 Two three-dimensional plots of the site characteristics data of Warwick (1971).

2.8 Methods to Avoid

• Stacked bar charts and multiple pie charts are generally not recommended for comparing groups of data due to their limited ability to distinguish differences between segments effectively.

• These methods allow only for coarse discrimination between segments, making it challenging to discern only large differences within a bar or pie chart.

• The difficulty in visually comparing magnitude differences, such as the percent K+Na between lithologic zones in a stacked bar chart, highlights the method's limitations.

• While stacked bar charts and pie charts can aggregate data across many sites or categories, they offer less visual distinction and clarity compared to other analytical methods discussed in the chapter.

• lternative visualization techniques are recommended to enhance data analysis, offering improved insight and usefulness.

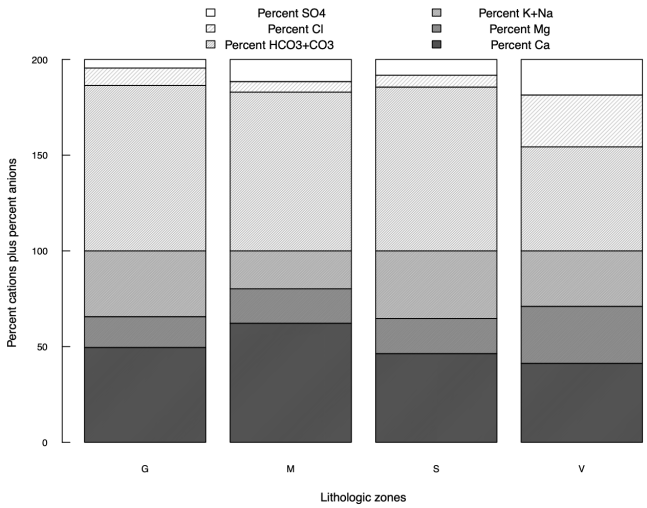


Figure 15. Stacked bar charts of mean percent milliequivalents of anion and cations within the four Groundwater Ambient Monitoring and Assessment (GAMA) Program lithologic zones of Shelton and others (2010).