Data Analysis

More Than Two Variables: Graphical Multivariate Analysis

Prof. Dr. Jose Fernando Rodrigues Junior

ICMC-USP

What is it about?

- ➔ More than two variables determine a tough analytical problem
- ➔ In particular, graphical methods quickly become impractical
- ➔ Although there are graphical techniques to display multivariate data, they can not deal with too many variables (typically, less than 15 – 25)

What is it about?

- Three-variables is a borderline case… there are several alternatives that work pretty well
  - False-color plots
- For a number of variables not much greater than three one may rely on multiple bivariate plots
  - Scatter plot matrices and co-plots
- For more variables
  - multidimensional visualization techniques
  - interaction

Three variables

➔For example, consider the data defined by **function**:

$$y = f(x, a) = \frac{x^4}{2} + ax^2 - \frac{x}{2} + \frac{a}{4}$$

that corresponds to **the three-variable setting y, x and a**

y the dependent variable, x and a the independent ones

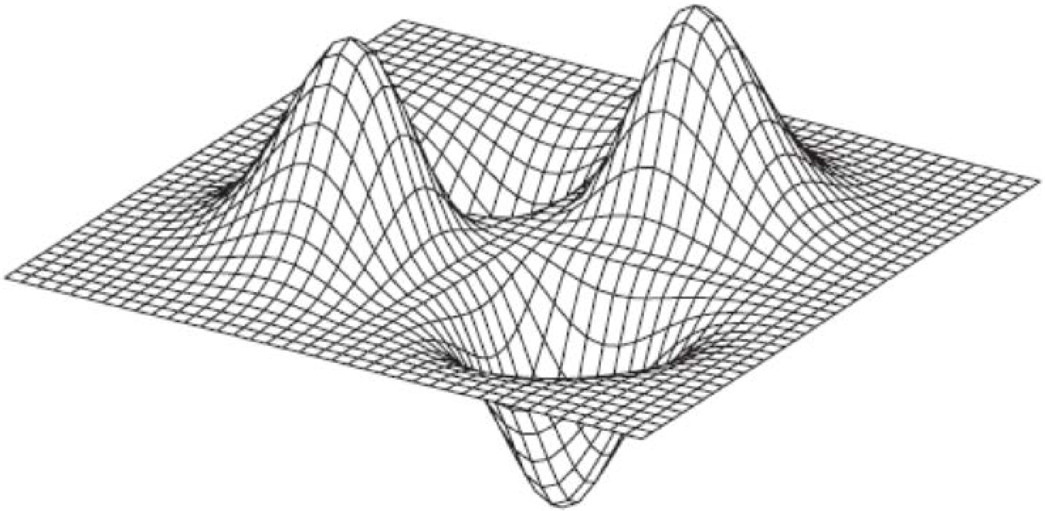
➔One way to analyze this is by means of a **surface plot**

Three variables

➔ For example, consider the data defined by **function**:

$$y = f(x,a) = \frac{x^4}{2} + ax^2 - \frac{x}{2} + \frac{a}{4}$$

➔ One way to analyze this is by means of a **surface plot**:



http://publicationslist.org/junio

Three variables

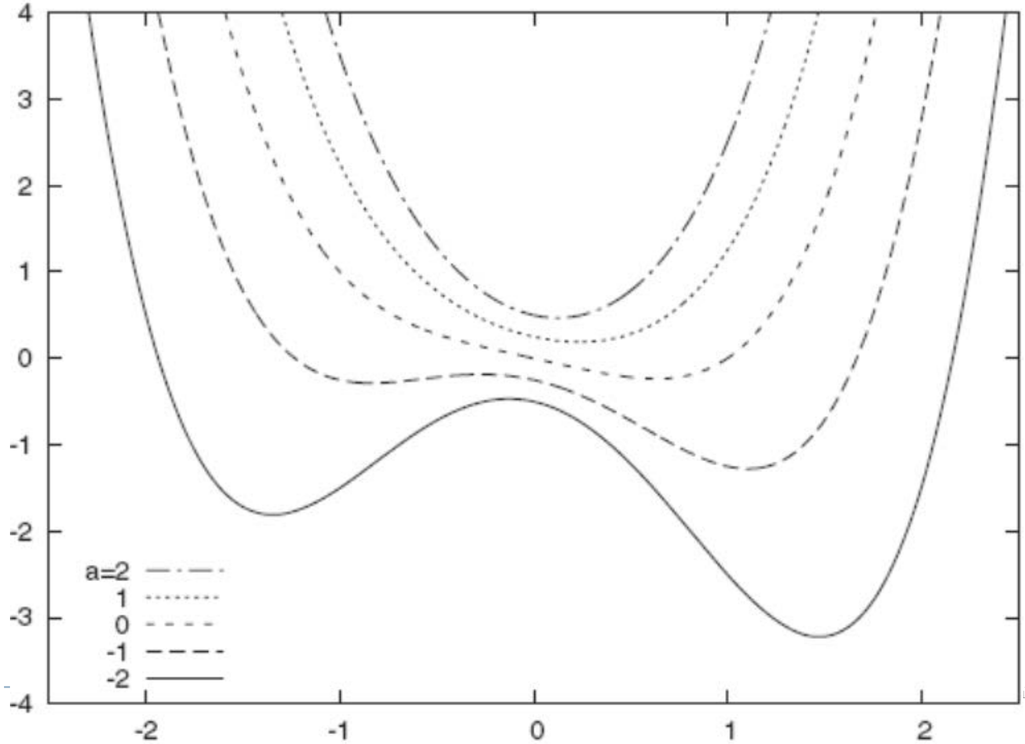
- Surface plots help build intuition for the overall structure of the data

- However, it is notoriously difficult to read off quantitative information from them, or develop a good sense for the behavior of the function

- Another way is to use a **two-dimensional xy plot** with **multiple curves**, one for each value of interest of one of the variables. This allows a more precise reading of quantitative information and a close inspection of the behavior of the function

Three variables

- Another way is to use a **two-dimensional xy plot** with **multiple curves**, one for each value of interest of one of the variables;
- In the previous example, variable *a* is considered for values 2, 1, 0, -1, -2



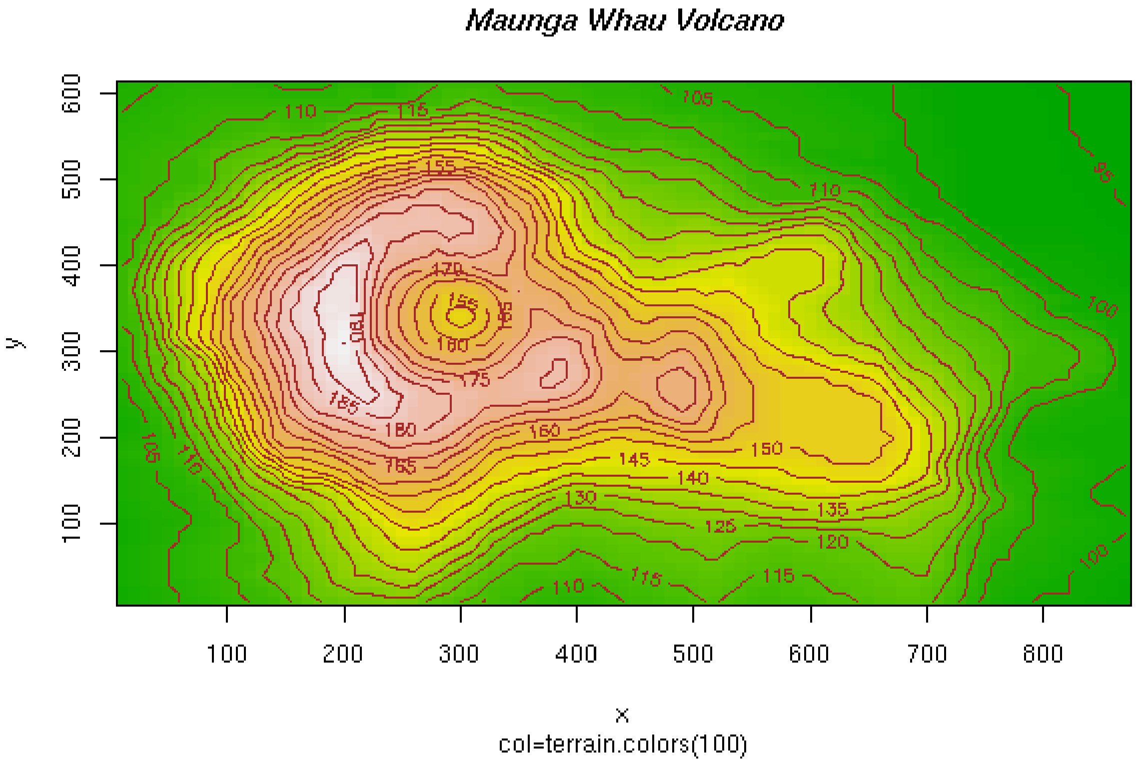
blicationslist.org/junio

Three variables

- Surface plots and multiple-curve xy plots can be used in **combination**, one providing an aesthetically appealing **overview**, the other providing **fine detail** for values of interest
- It is interesting to note that **surface plots go against the commonsense that 3D plots should be more informative than 2D plots**
- Yet another possibility is to **project the function into the base plane bellow the surface**, using either:
  - contour plots
  - false-color plots

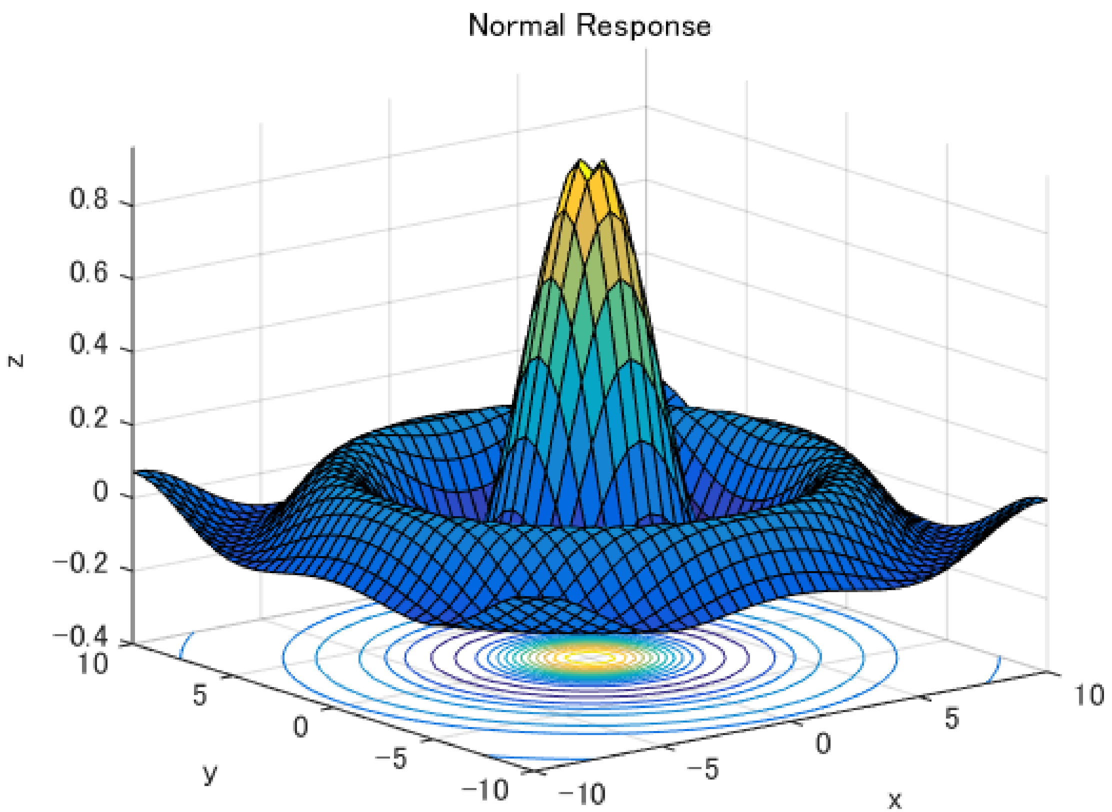
Three variables

- Contour plots: familiar from topographic maps
- Good to convey local properties, effective if the data is relatively smooth



http://publicationslist.org/junio

Surface plot + contours



http://www.mathworks.com/matlabcentral/fileexchange/35311-matlab-plot-gallery-surface-contour-plot/content/html/Surface_Contour_Plot.html

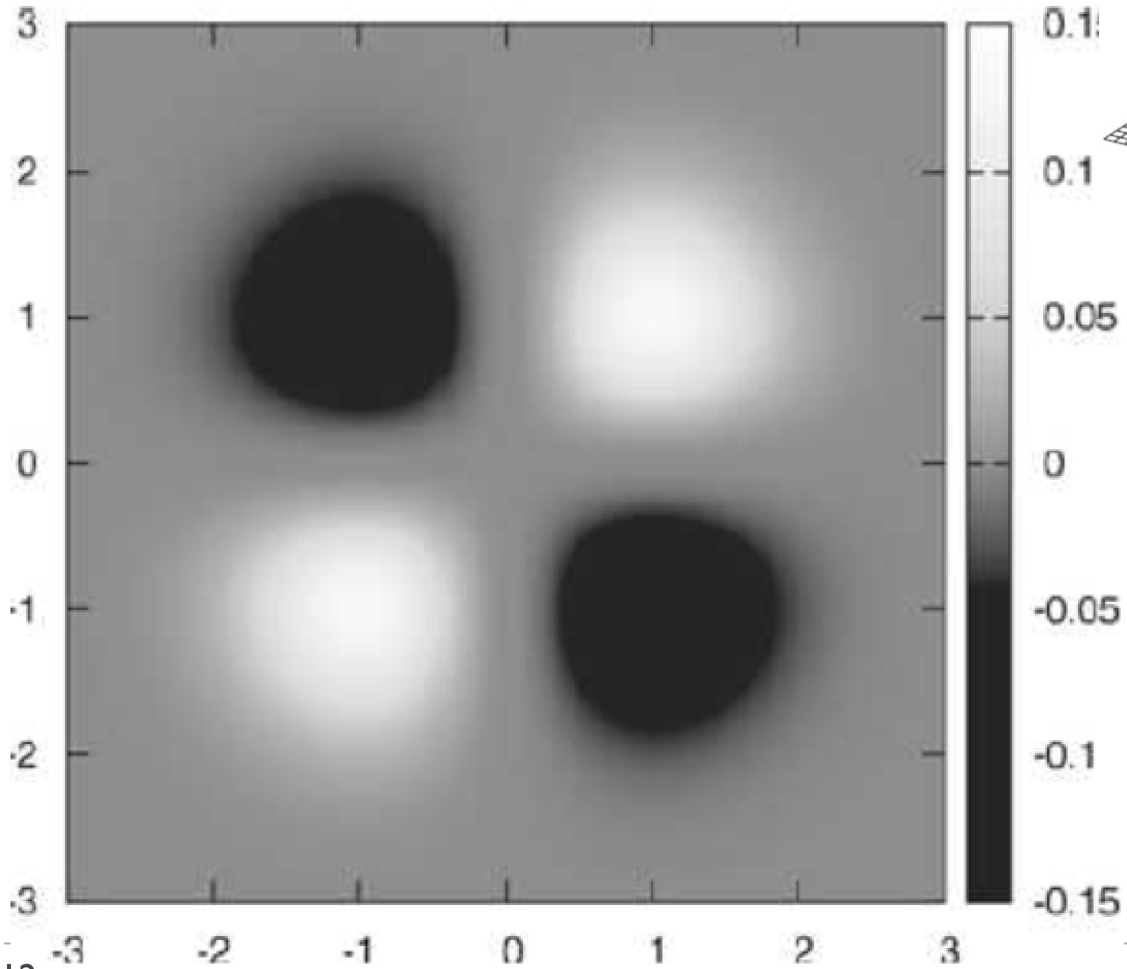
Three variables

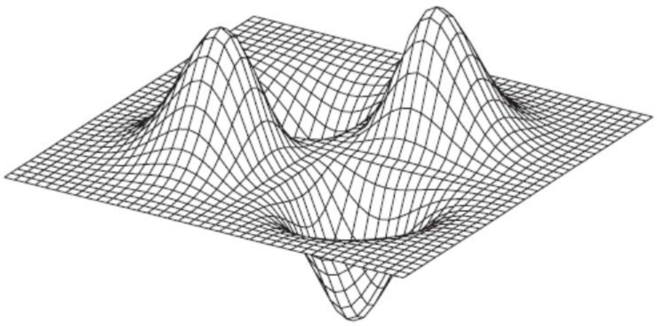
- The false-color plot is an alternative
  - Highly versatile: applicable in many different situations
  - Retains quantitative information

- Obtained by mapping all values of the dependent variable following a palette of colors

Three variables

- A **false-color** plot for function f(x, a)





http://publicationslist.org/junio

Three variables

- The fa… of the depen…
- The **fa**…

False-color plots are very effective for presenting quantitative information

It is important to note, however, that **its efficiency depends heavily on the color mapping**, which must be intuitive according to the task at hand

For an overview of color-mapping guidelines, see the textbook on page 104

http://publicationslist.org/junio

Parenthesis

- Actually, the choice of good color palettes when using color to convey information is a very relevant topic in data visualization

- Usually, novices in the field pay less attention to this topic than they should
  - Using whatever default is available in your system typically results in very bad results…

  - See http://colorbrewer2.org/
  - (a web tool for selecting colors for maps, not meant for general data analysis contexts, but still useful)

Parenthesis

- If color is used to map information, a color legend is obviously required!

- Color does not reproduce well across different media

More than three variables

- There are basically two ways to get more information on a plot
  - **Put similar graphs next to each other and vary the variables** in a systematic fashion from one subgraph to the next → multiplots
  - Make the graph elements themselves **richer with color, shape, and interaction**
- Multiplots
  - The most common forms of **multiplots** are the scatter-plot matrix, and the co-plot,

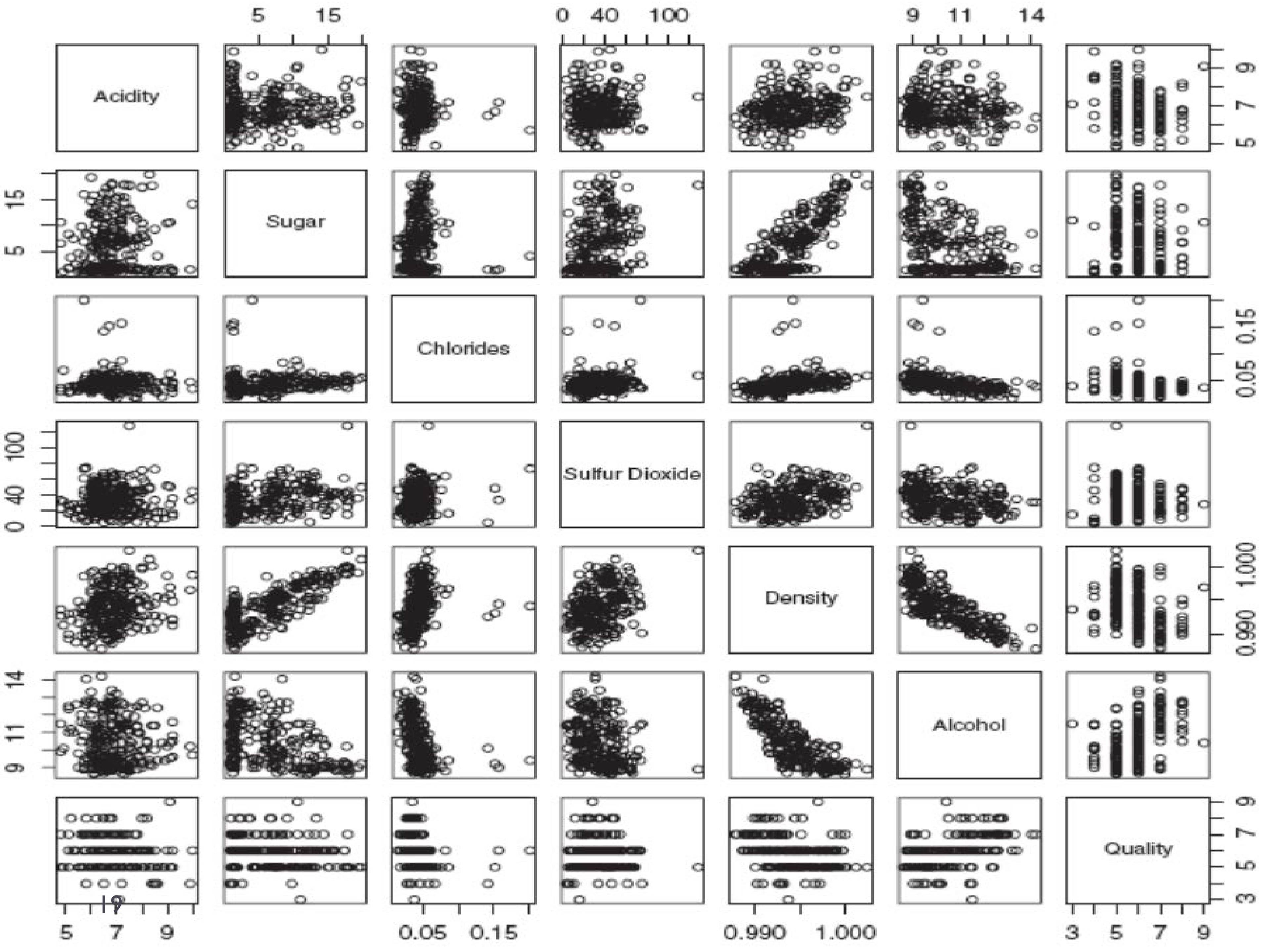
Scatter-plot matrix

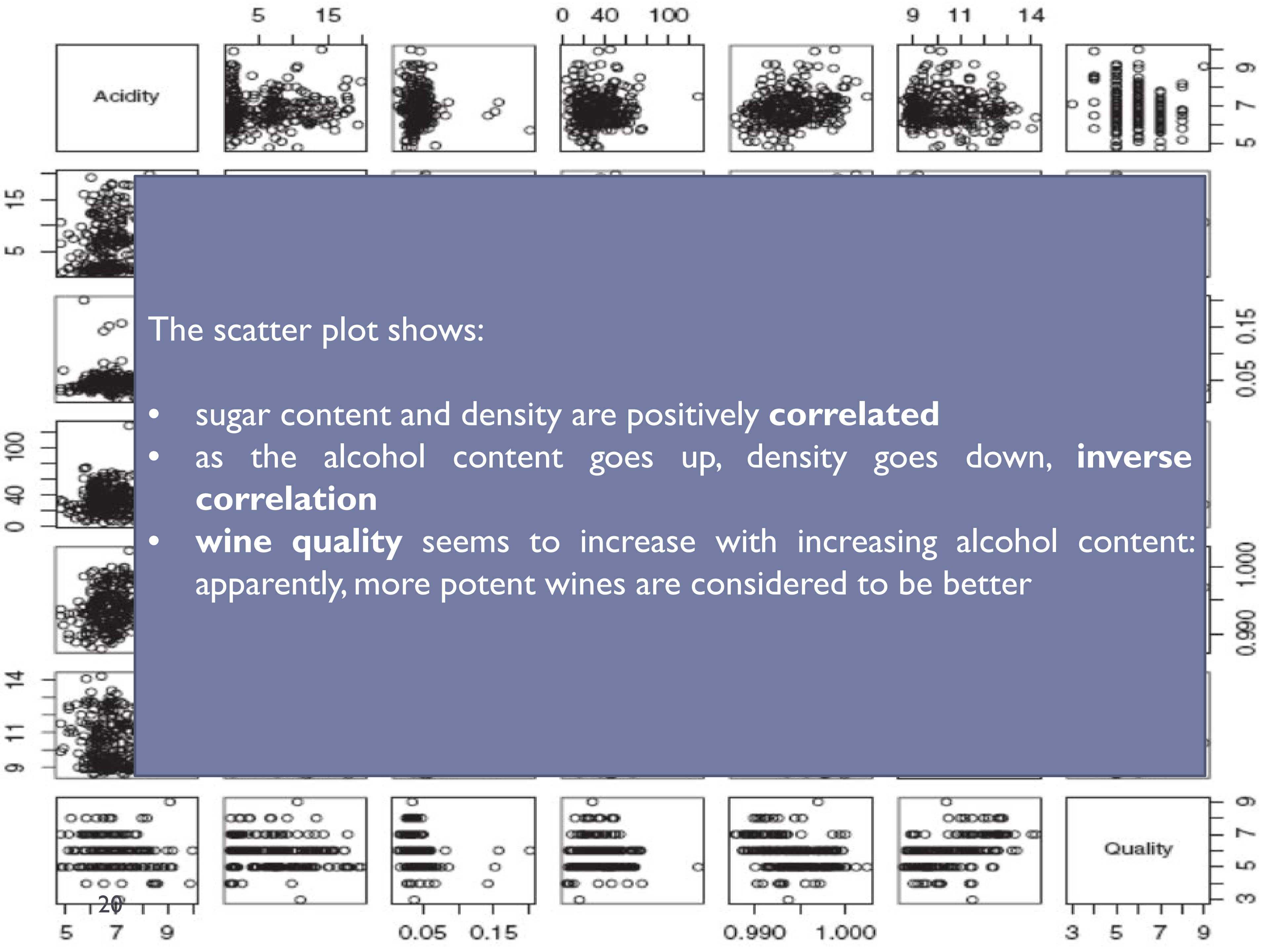
- The scatter-plot matrix is constructed considering **all the possible two-variable combinations** achieved from the set of variables
- **For each combination, a sub-region** of the space is reserved and all the combinations are put together according to a straight layout
- **The more variables, the bigger must be the screen**, limits start to manifest around 10 variables, the same for the number of data points, limited around 100

Scatter-plot matrix

➔ **For example**, consider a 250 wines data set consisting of seven different properties: acidity, sugar, chlorides, sulfur dioxide, density, alcohol, and quality

The data can be found in the “Wine Quality” data set, available at the UCI Machine Learning repository - http://archive.ics.uci.edu/ml/.

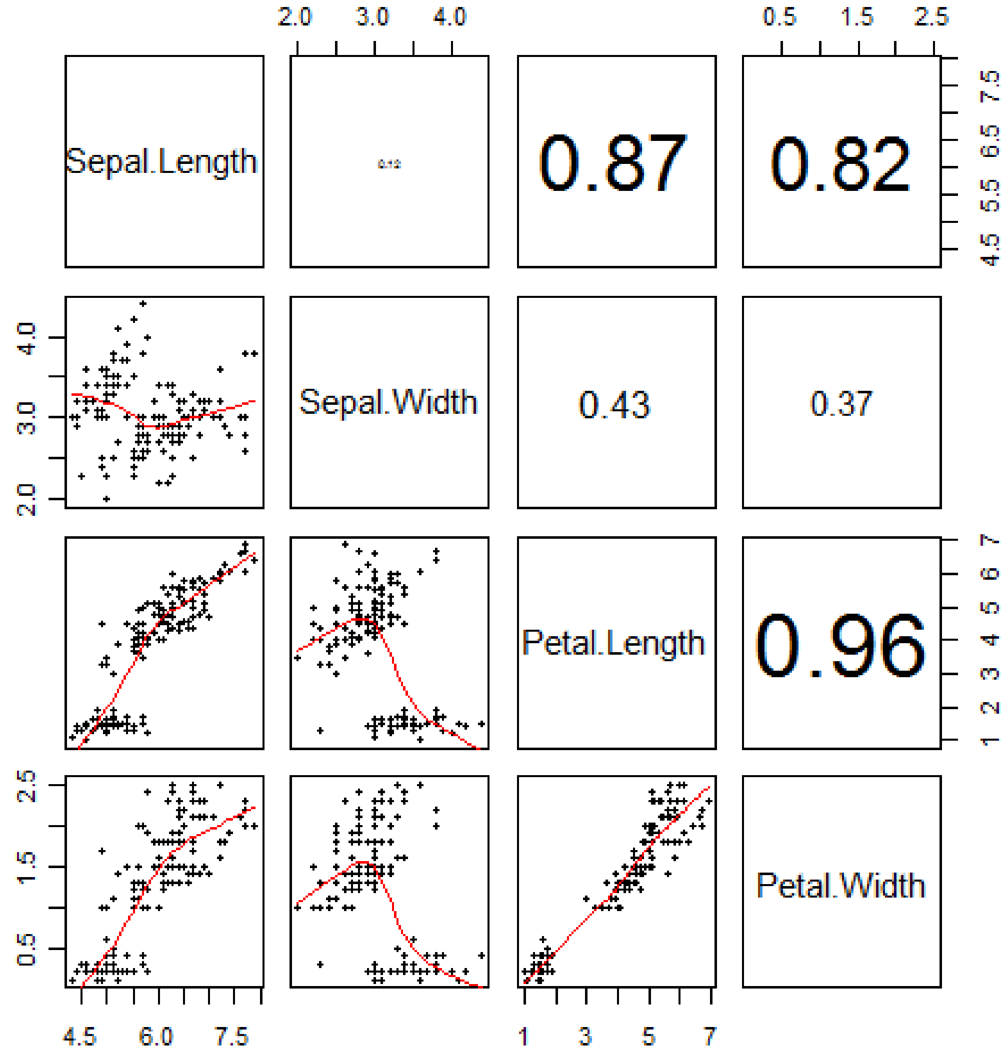

Acidity
Sugar
Chlorides
Sulfur Dioxide
Density
Alcohol
Quality



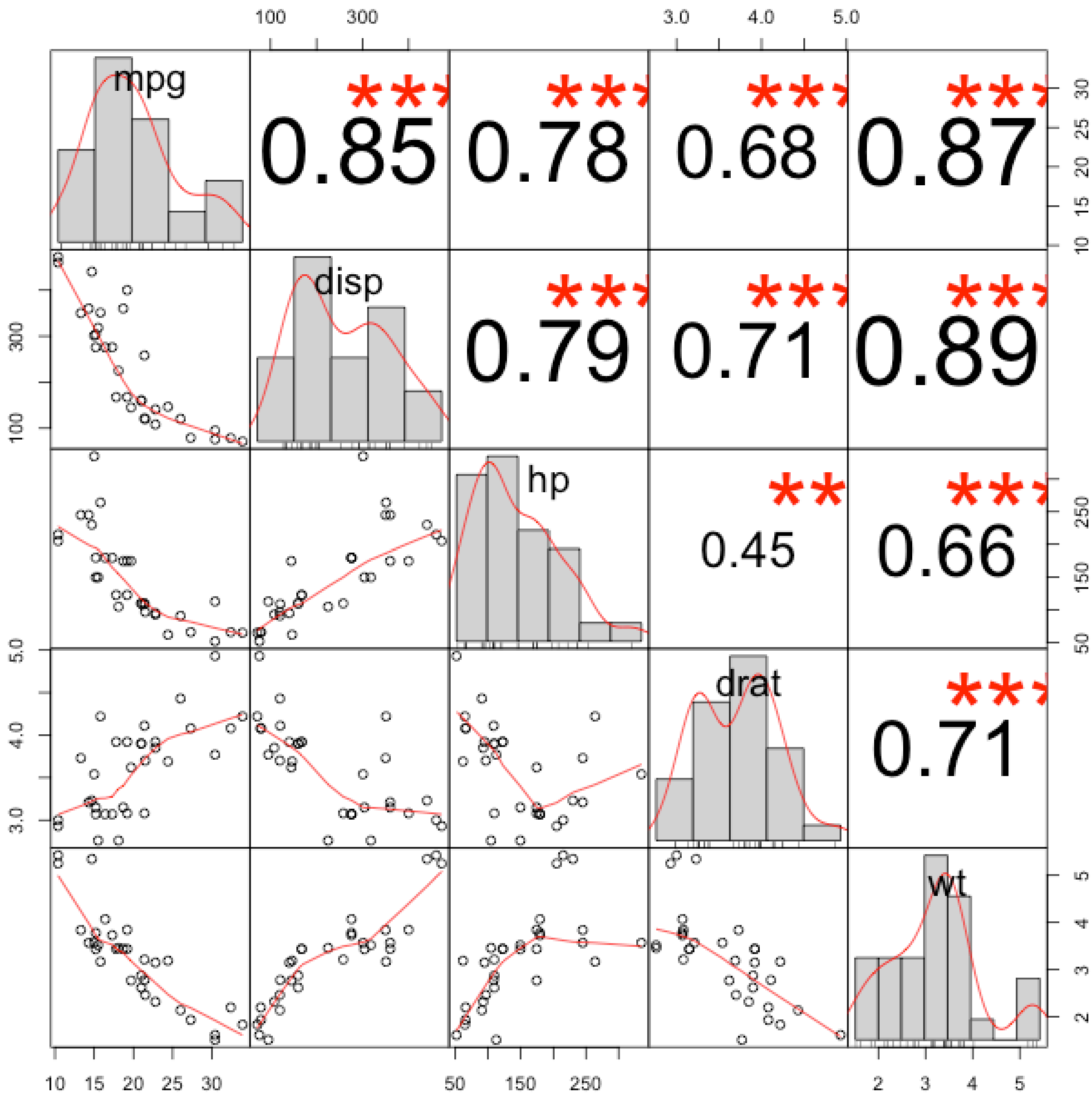
The scatter plot shows:

- sugar content and density are positively **correlated**
- as the alcohol content goes up, density goes down, **inverse correlation**
- **wine quality** seems to increase with increasing alcohol content: apparently, more potent wines are considered to be better

Iris Scatterplot Matrix



http://publicationslist.org/junio



http://publicationslist.org/junio

Co-plots

- Short for conditional plots or conditioning plots
  - A way of showing how a response (or ´control´ variable) depends on (two or more) other variables

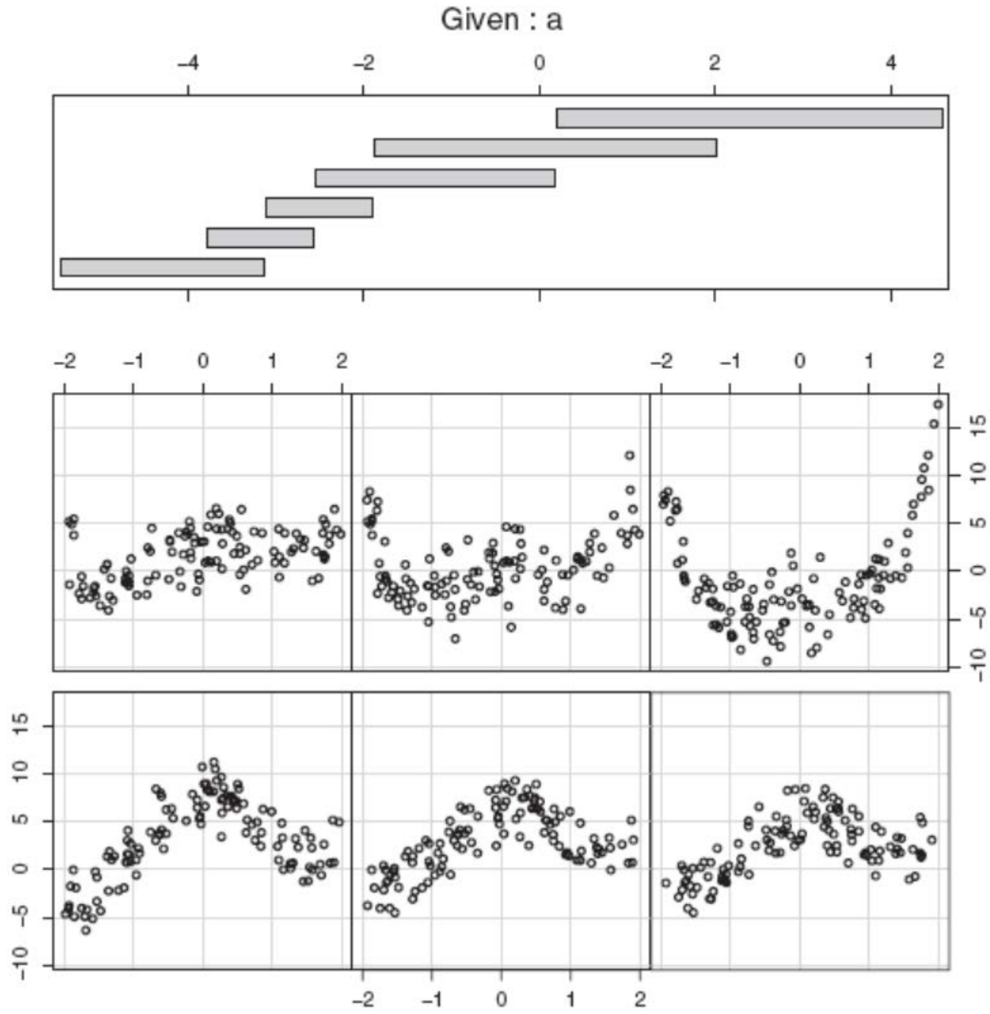
- Co-plots work by **partitioning** the data according to one of the variables (data slices) and plotting each partition in a **different plot**

Co-plots

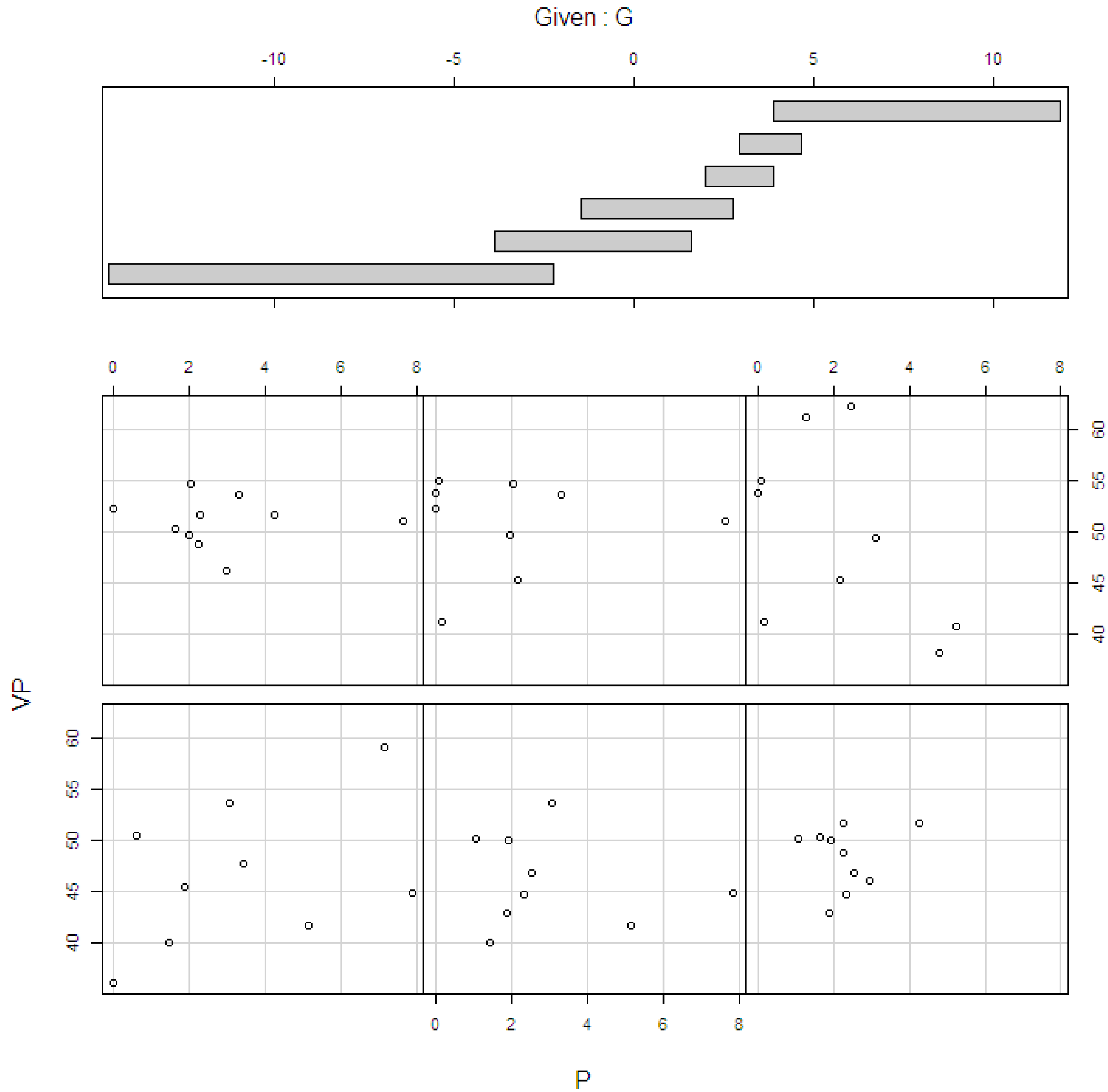
➔ In this example, consider the function y = f(x,a)

- ➔ the upper figure shows how one of the variables (a) was used to partition (slice) the data
- ➔ then x,y plots are shown for each interval

➔ Notice that the intervals **overlap** and have **different sizes** so that each plot has **the same number** of points



Co-plots



P = inflation rate,
VP = voting percentage
G = rate of growth

Coplot of an election data set. This is assessing the effect of P on VP conditional on varying values of G.

list.org/junio

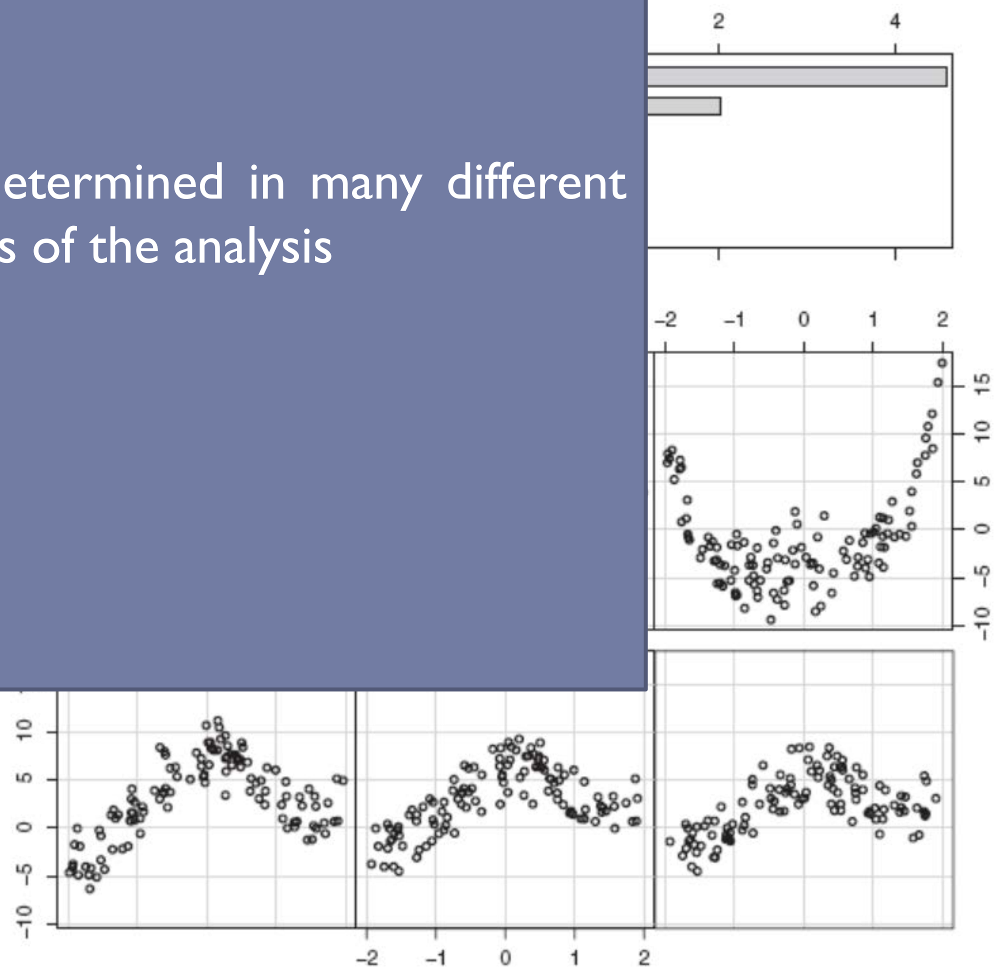
Co-plots

→ Co-plot[...] to one of the vari[...]ent plot

→ In this figure [...] the var[...] partitio[...]

The partitioning can be determined in many different ways according to the goals of the analysis

→ It is po[...] the interva[...] **different** [...] plot has **the same number** of points



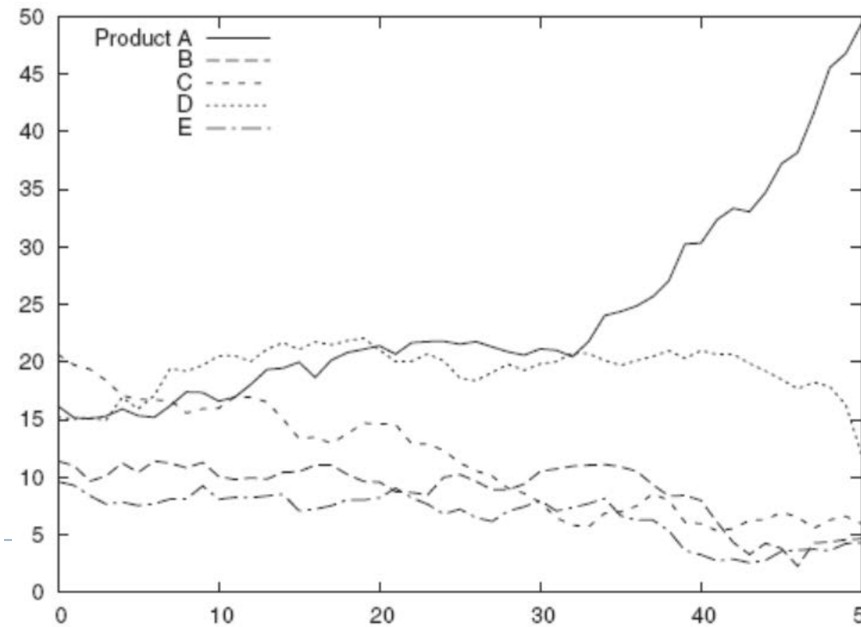
Composition

- Another way to visualize more than two variables is to **compose multiple plots** according to some of the variables
  - Suitable when the data describes how some overall quantity is composed out of parts

- **For example**: imagine a company that makes five products labeled A, B, C, D, and E, and two questions:
  - how many items of each kind are produced overall
  - how the item mix is changing over time

Composition

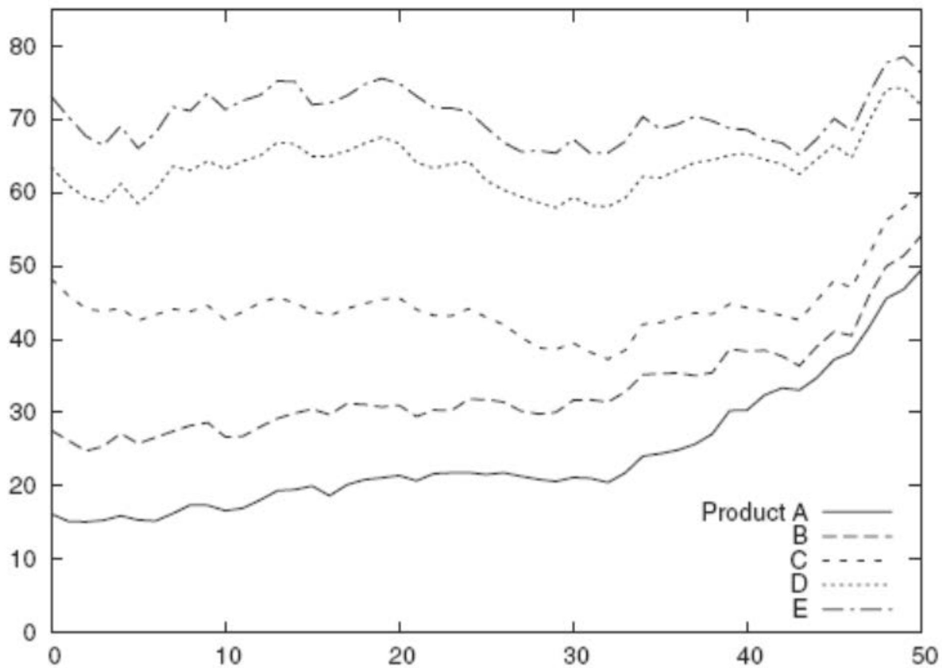
- **For example**: imagine a company that makes five products labeled A, B, C, D, and E, and two questions:
  - how many items of each kind are produced overall
  - how the item mix is changing over time
- A **simple** solution, but **not quite effective** is to plot (days x quantity) the different curves all together

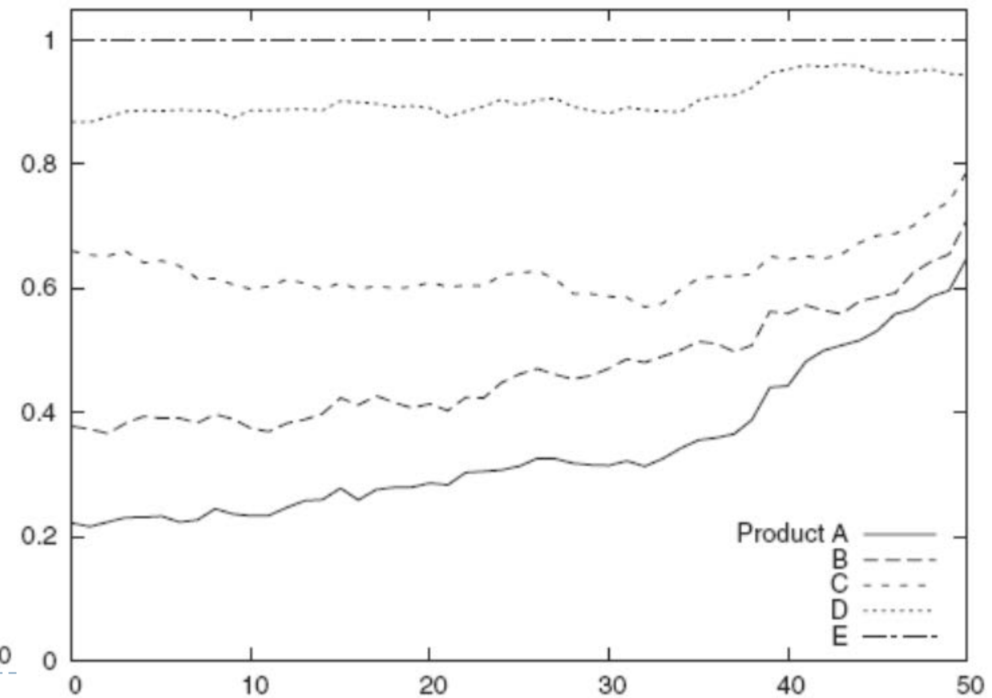


g/junio

Composition

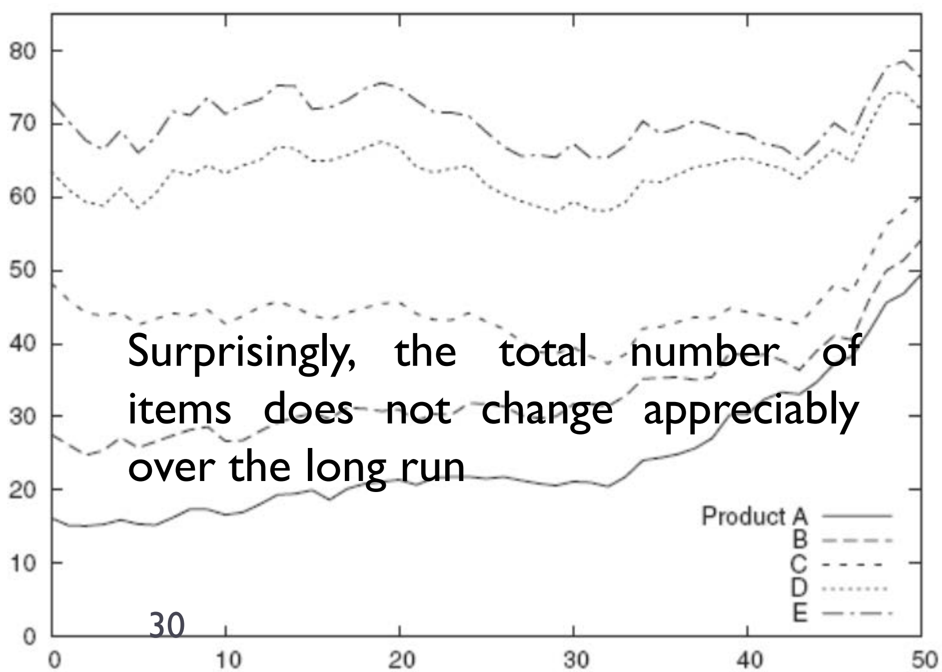
- **Another solution is to use the same plot but stacking the information, so as to have a notion of total**
  - In absolute numbers (left), or
  - In relative contributions (percentage)



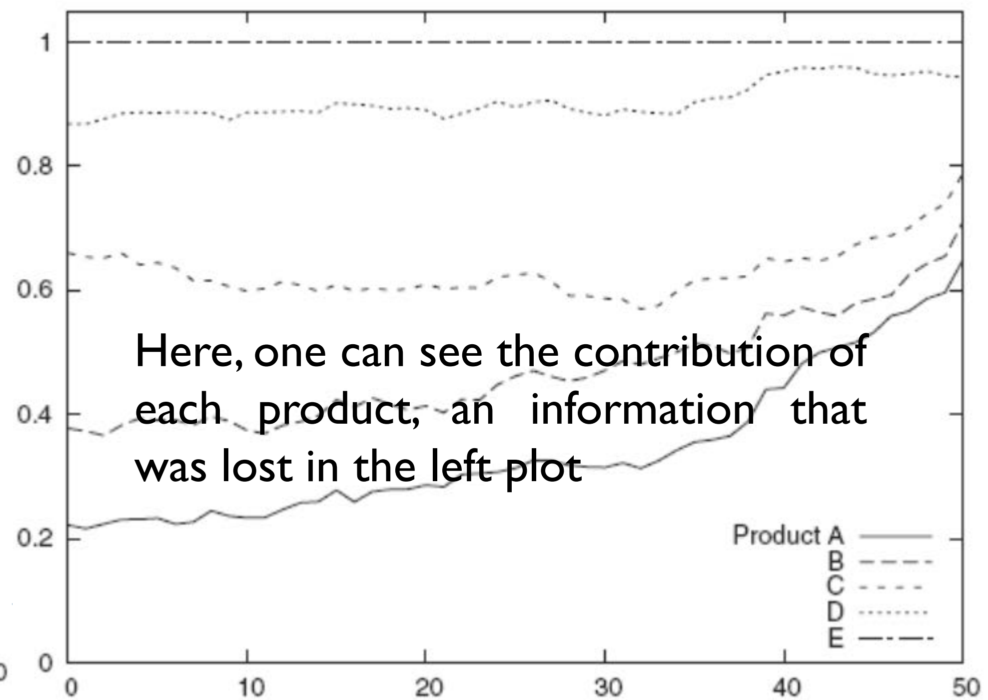


Composition

- Another solution is to use the same plot but **stacking** the information, so as to have a notion of total
  - In absolute numbers (left), or
  - In relative contributions (percentage)



Surprisingly, the total number of items does not change appreciably over the long run

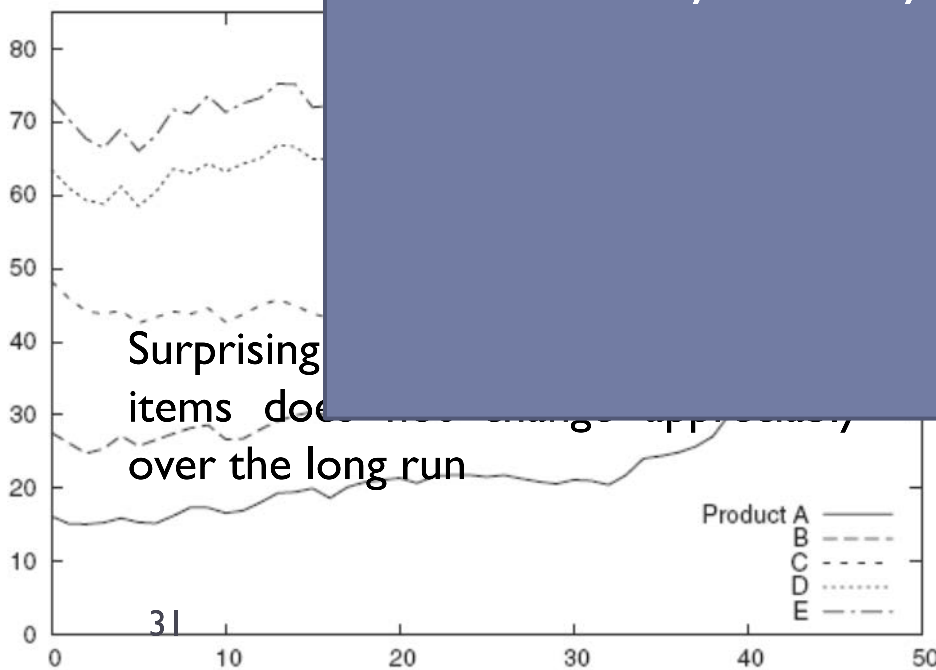


Here, one can see the contribution of each product, an information that was lost in the left plot

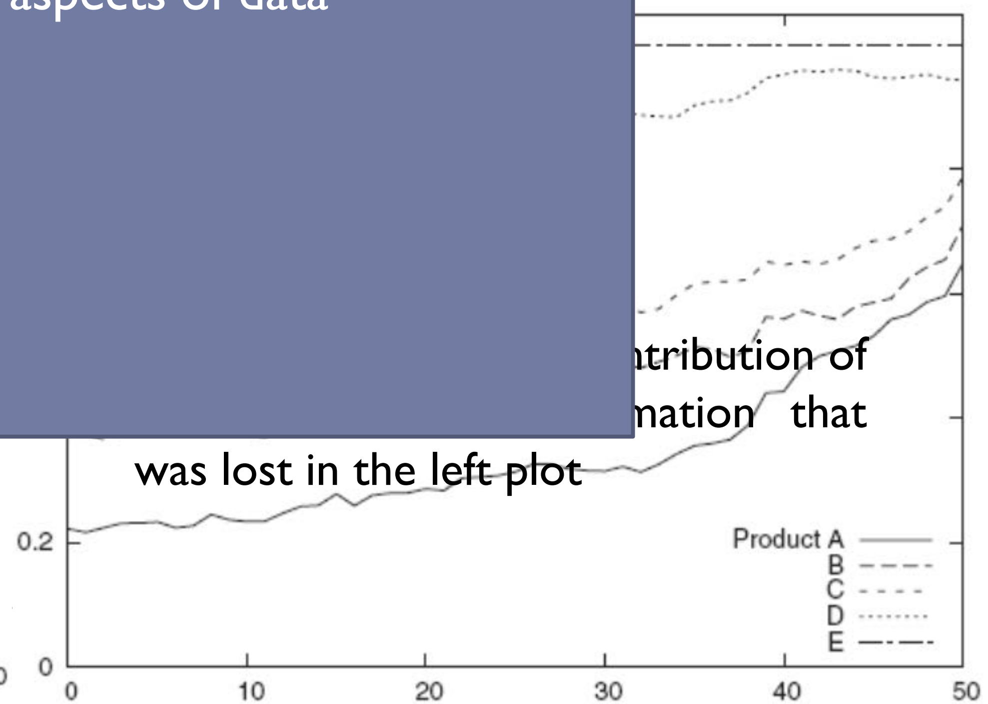
Composition

- Anothe… …cking the informa…
  - In abs…
  - In rela…

It is desirable to use multiple kinds of plots combined in order to verify the many aspects of data



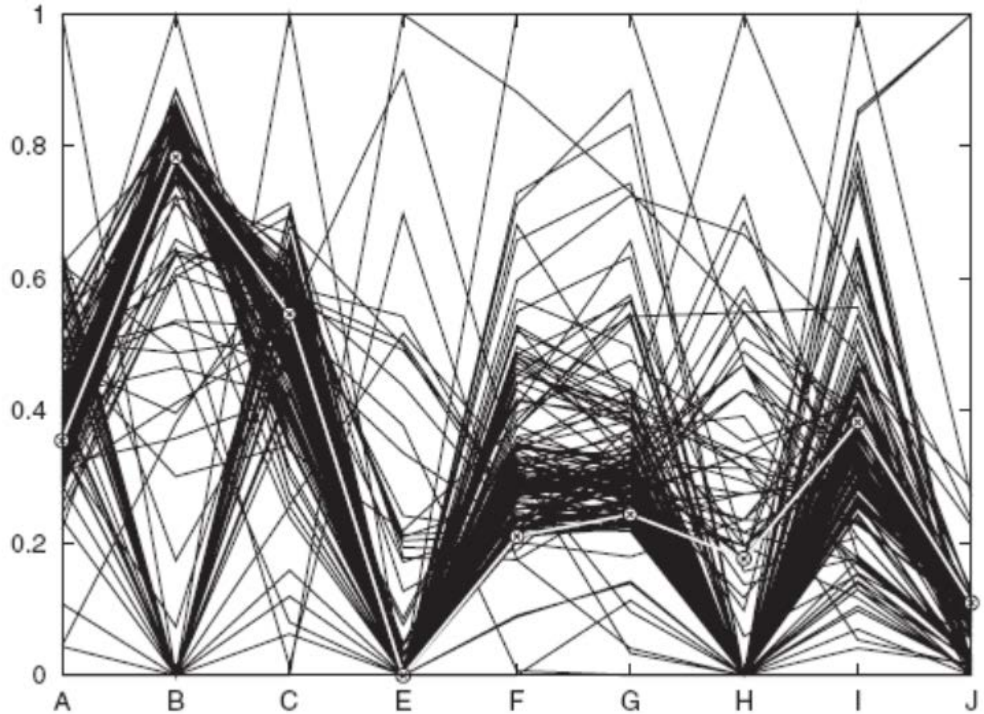
Surprising…
items doe…
over the long run



…tribution of
…mation that
was lost in the left plot

Parallel Coordinates

- In a parallel coordinate plot, the coordinate axes are **parallel** to each other
- For every data point, its value for each of the variables is **marked** on the corresponding axis, and then all these points are **connected with lines**



http://publicationslist.org/junio

Parallel Coordinates

- See https://syntagmatic.github.io/parallel-coordinates/

- Brushing
- Linking and brushing

Information Visualization

- Many other techniques are presented according the findings of the field known as **Information Visualization**:
  - Glyphs
  - Chernoff Faces
  - Tree-maps
  - Star coordinates
  - Table Lens
  - Multidimensional Projection
- And **many others**, all improved by means of **interaction techniques**:
  - Querying and zooming
  - Linking and Brushing
  - Combined projections, and so forth

References

- Philipp K. Janert, *Data Analysis with Open Source Tools*, O'Reilly, 2010.
- Wikipedia, http://en.wikipedia.org
- Wolfram MathWorld, http://mathworld.wolfram.com/