# Multivariate Data Visualization

Zdenko Sonicki, *Andrija Stampar School of Public Health*

**Abstract**

Medical images, maps, or signals are often converted into numerical data to be analyzed. On the other hand measured numerical data are analyzed with more or less common methods.

Is it possible to visualize measured numerical data, and convert it to signals or images? There are various examples of numerical data visualization with intention to avoid variable interaction loss within the system. With scatterplots two or three variable interactions could be plotted. The series of scatterplots representing one system are analyzed by AI procedures in order to select useful plots for further analyses from uninformative scatterplots. Unfortunately, despite very sophisticated process, every time a scatterplot is generated, two or three variables are selected from the system, and information about all interactions within the system is lost. Step further in preserving system interactions information is RadViz visualization, where variables are presented evenly on a unit circle. Visualization changes when order of neighboring variables is changed. Also, AI methods could be applied to select more informative visualizations for further analyses from uninformative plots. Despite improvement in preserving information of variable interactions within the system, some information is still lost.

Promising approach where all variable interactions within the system could be visualized, called Visual Co-Plot, applies superimposing of two plots in sequence. Approach is based on method of Multidimensional Scaling (MDS). First plot uses MDS to present distances between observations, while second plot, conditioned on a first one, displays vectors of relationships among variables. Those vectors describe correlations among the variables. Distances between each of the observations are calculated with the distance metrics called "city-block distance". Instead of distance matrix, CoPlot uses non-metric MDS based on Smallest Space Analysis (1, 2), to produce two-dimensional plot of n observations. A second plot, superimposed on first, consisting of vectors for each variable, is calculated by least-squares regression, so that the correlation of the values of variable and projections of each observation is maximized. The length of each vector is proportional to the correlation between the original data of that variable and the projections of the observations onto the vector. In order to describe how good the plot represents the observations and how good another plot represents the variables goodness of fit diagnostics are applied. Relative loss of information that appears when the multidimensional data are transformed into two dimensions is measured by "coefficient of alienation" (2). Also, a magnitude of maximal correlations could be calculated, and upon that, variables that have low magnitudes of correlations could be eliminated from the analysis.

This approach will improve description of the analyzed system in order to select useful predictor variable set for predictive modeling in medicine.