Cluster Analysis and Data Mining

Task 1

In datafile (`Heavymetal.xls`) there are some measures of pollution with some metals (Pb and Hg) for a given industrialized country. Assume a complex survey design where strata are defined as 20 provinces. In each province we take 2 regions at random from all possible regions (*psu*) and then we take 2 cities (*ssu*) at random and 2 zones per *ssu*.

Consider an stratified two-stage cluster design, where the population number of *psu* are located in the variable `Npsu` and the population number of *ssu* are in the variable `Nssu`. Levels of Hg and Cu are measured in *nanograms* (`ng`) per Kg of soil.

Explain which are the main characteristics of this class of design and identify its components.

Estimate the total amount of both metals. Estimate confidence intervals (90%) for their mean values. Plot histograms and boxplots taking into account the sampling design.

Calculate the regression line between both metals taking into account the sampling design.

Task 2

Find out a set of association rules, using `arules` or using `Weka`, in any dataset located in the *UCI repository* of data:

`http://archive.ics.uci.edu/ml/`

Justify the obtained results and write what are the main **conclusions**.

Apply in your selected dataset tree-methods and k-means clustering procedures. You must consider one dependent variable in your data to run the tree-methods technique.