Classifying Samples According to Location using HCA and PCA

By Yusri Yusup

Table of Contents

1. Introduction

The identification of the source of a sample can be useful for the investigator when studying the scene of the event of interest. Collecting samples at different locations and chemically analyzing them could provide enough information on the distinct characteristic of each location through pattern recognition techniques such as hierarchical cluster analysis (HCA) or principal component analysis (PCA). Three locations were sampled, and chemical analysis was conducted to quantify the concentration of 12 elements. Using this dataset, patterns were discerned using HCA and PCA.

2. Methods

2.1 Data collection and treatment

The data, in the csv format, was downloaded from the IEG301/3 Environmental Forensics course elearn@USM portal on the 25th of March 2019. The data consists of elemental concentration observations at different locations. Three locations were available: car, plant, and scene. The samples were presumably analysed for twelve elemental concentrations: Mg, P, K, Ca, Fe, Cu, Zn, Rb, Sr, Mo, Cs, and Ba. The units for the concentration are unknown. There were nine observations for car, five for plant, and six for scene, giving a

total of 20 observations. The data was tidy, and thus did not need further data treatment for analysis.

2.2 Exploratory data analysis

The data was imported and analysed in R and RStudio (R Core Team, 2015). The data was explored by checking the summary of its statistics and some plots were drawn. This is done to check for missing data, number of observations available for each location, and determine the variables of the elemental concentrations that may be needed in further analysis.

A cursory look at the data showed that there were no missing data in the dataset. Scatter plots and histograms were drawn for the three locations to ascertain any distinct characteristics at the different locations. Analysis of variance or ANOVA was then used to test the significance of the difference between three randomly chosen elemental concentrations. This is followed by the hierarchical cluster analysis and principal component analysis on the concentrations to discriminate samples between the car, plant, and scene locations.

3. Results

Boxplots and scatterplots show that the samples collected from plants have higher concentrations than car and scene for some elements (e.g., **Figs. 1-2** for Ca and Zn) but the opposite trend for other elements (e.g., **Fig. 3** for Cs). These plots were produced from randomly choosing three elements to check the distribution of each element for similarity.

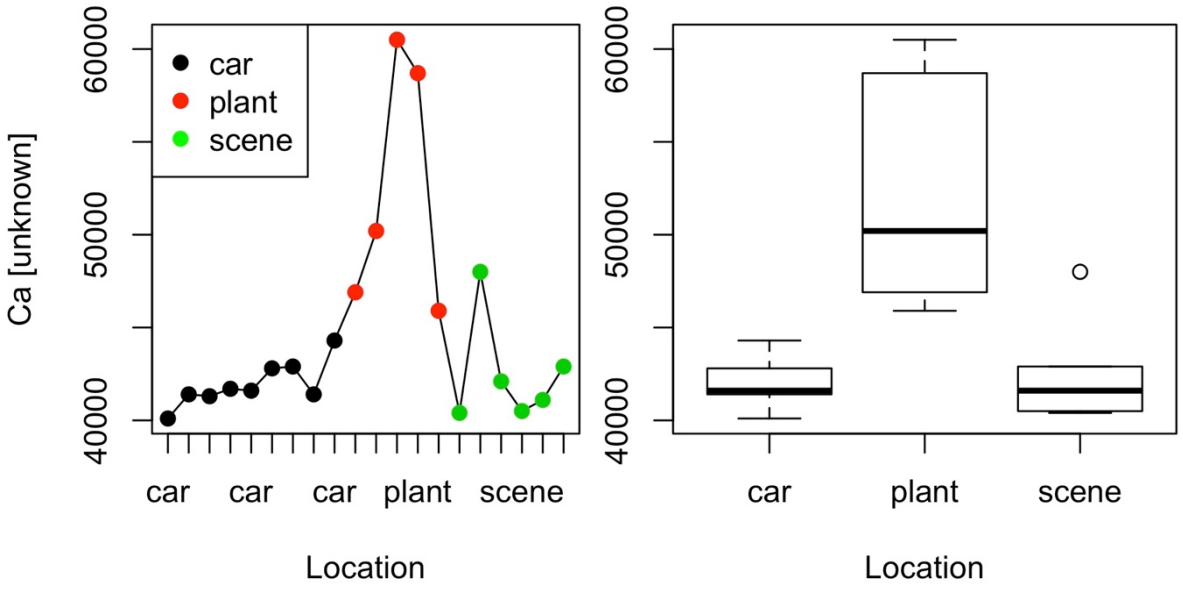

Fig. 1 Left panel Scatterplot and **right panel** boxplot for Ca with unknown units at car, plant, and scene locations.

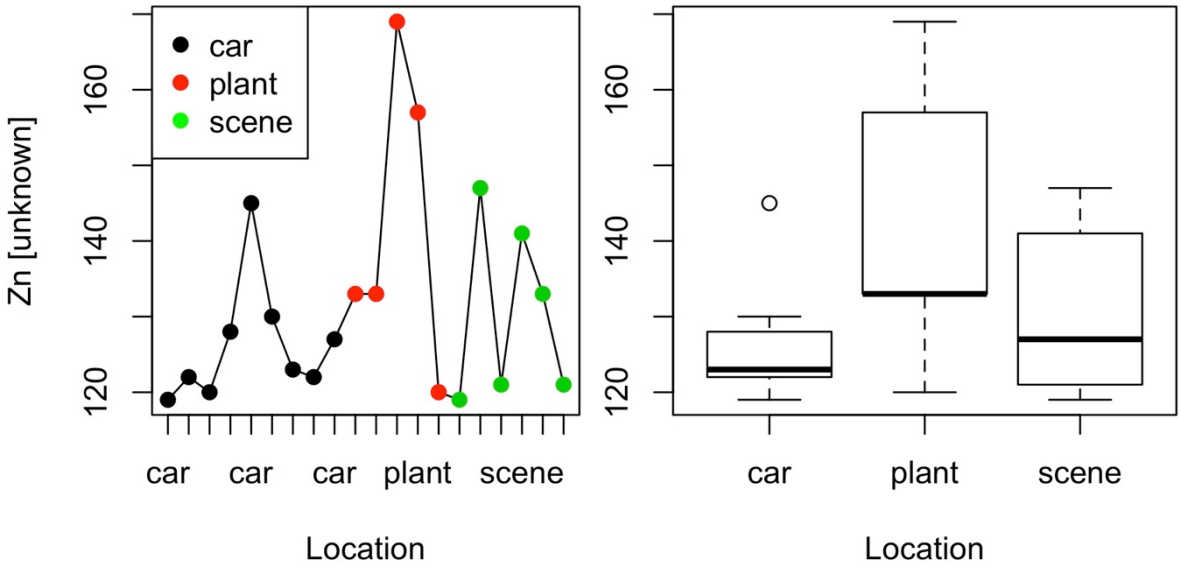

Fig. 2 Left panel Scatterplot and **right panel** boxplot for Zn with unknown units at car, plant, and scene locations.

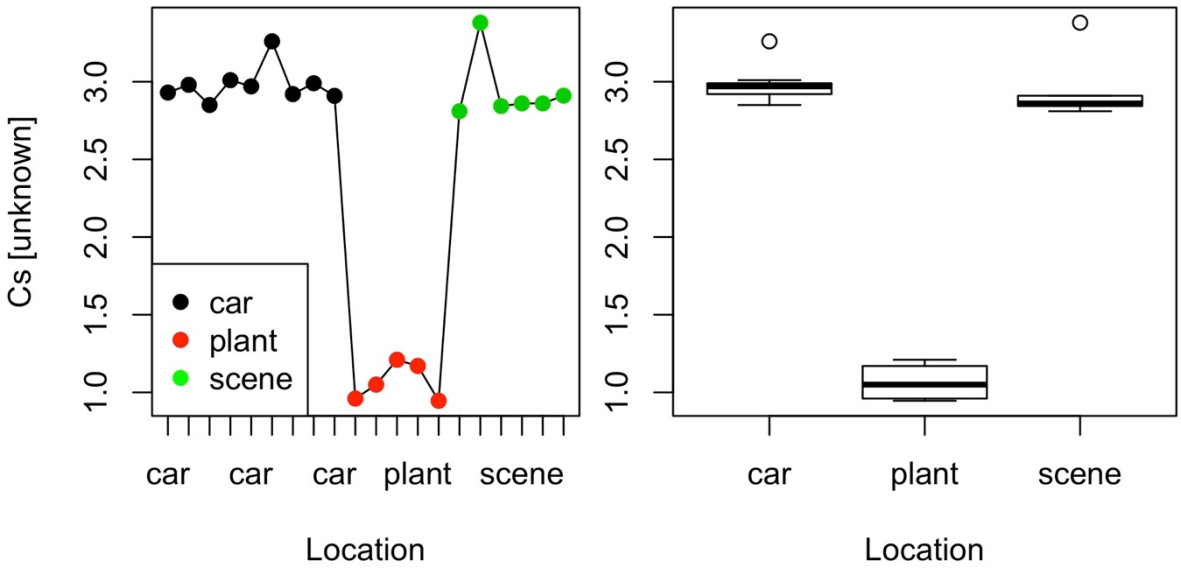

Fig. 3 Left panel Scatterplot and **right panel** boxplot for Cs with unknown units at car, plant, and scene locations.

ANOVA for Ca (p-value < 0.001) and Cs (p-value < 0.001) confirmed that the mean for plants and are significantly different than either car or scene but not significantly different for Zn. However, the boxplots show that the concentrations are not normally distributed and the number of observations for each location is not more than the arbitrary statistic consensus of 30. Hence, more data is needed to confirm the significance of the difference. However, the

results suggest distinct variations of elemental concentration among the locations, which could be distinguished by hierarchical cluster or principal component analyses.

Cluster analysis with heatmap (**Fig. 4 top panel**) shows that the locations could not be separated into the three main clusters. There are overlaps between car and scene while plant is distributed into two different clusters. Different methods, “complete” and “average” yield the same result (**Fig 4 bottom left panel** and **bottom right panel**). However, the heatmap does reveal that the elemental concentrations of Ca, P, and K are much higher than the other elements, thus the two top main clusters, which could add bias in the clustering algorithm due to the large distance between some clusters. This bias could be reduced by scaling the concentrations.

The elemental concentrations were scaled and centred before performing the PCA. The analysis reveals two distinct clusters: car-scene and plant (**Fig. 5**). The car and scene clusters overlap which suggest that the two locations have the same elemental characteristics, and thus could originate from the same source. This also explains the recurring overlap between car and scene for the cluster analysis.

Cluster analysis was not able to distinguish between the locations due to the large differences in elemental concentrations because it relies on distance to group the data. PCA was able to circumvent this limitation by scaling the variables before the analysis and that it is based on the best linear fit among the variables. However, cluster analysis could perform better if the data was scaled before the analysis is performed.

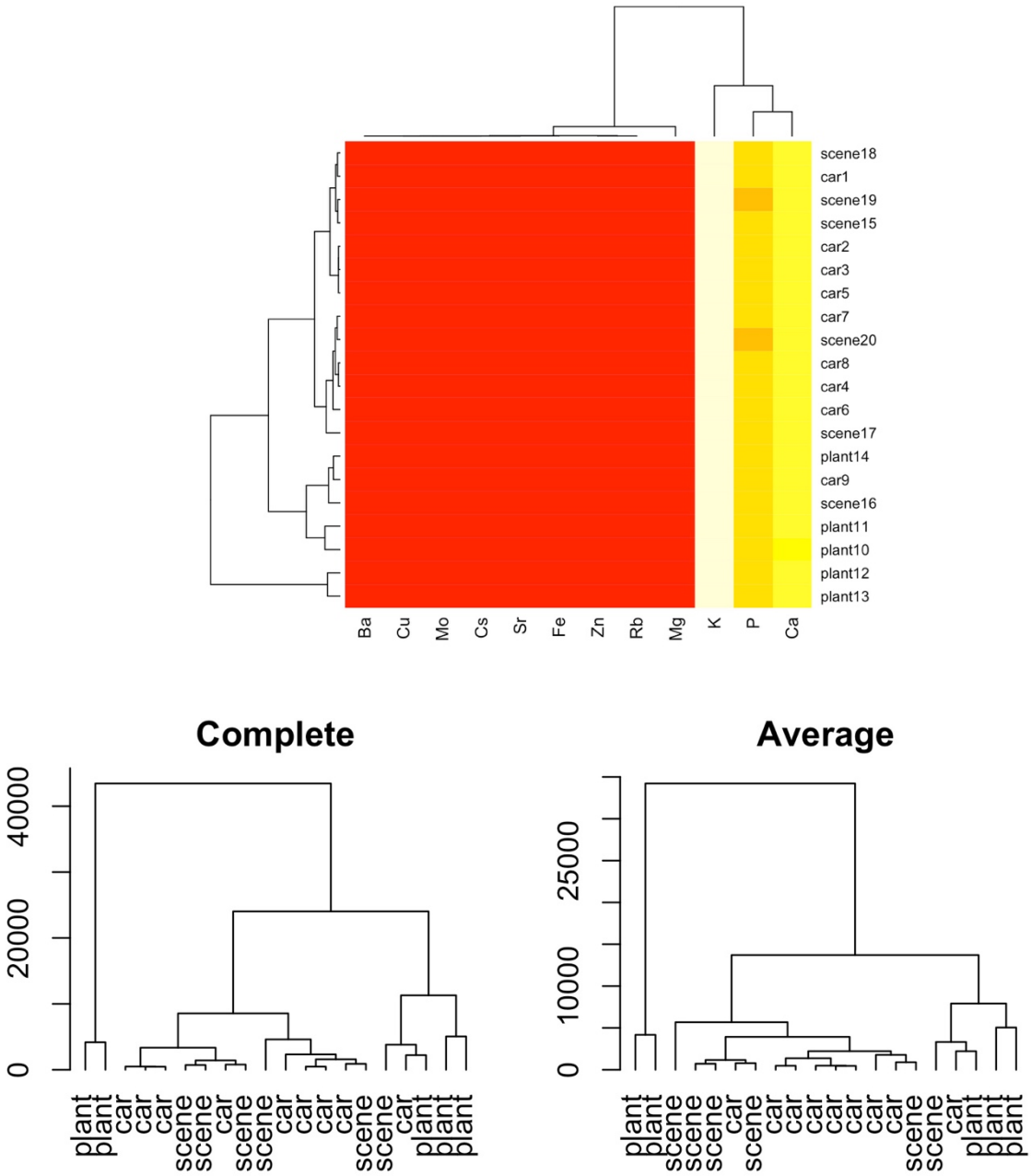

Fig. 4 Top panel Heatmap and dendrogram **bottom left panel** "complete" and **bottom right panel** "average" methods for the elemental concentrations at car, plant, and scene locations for the different observations.

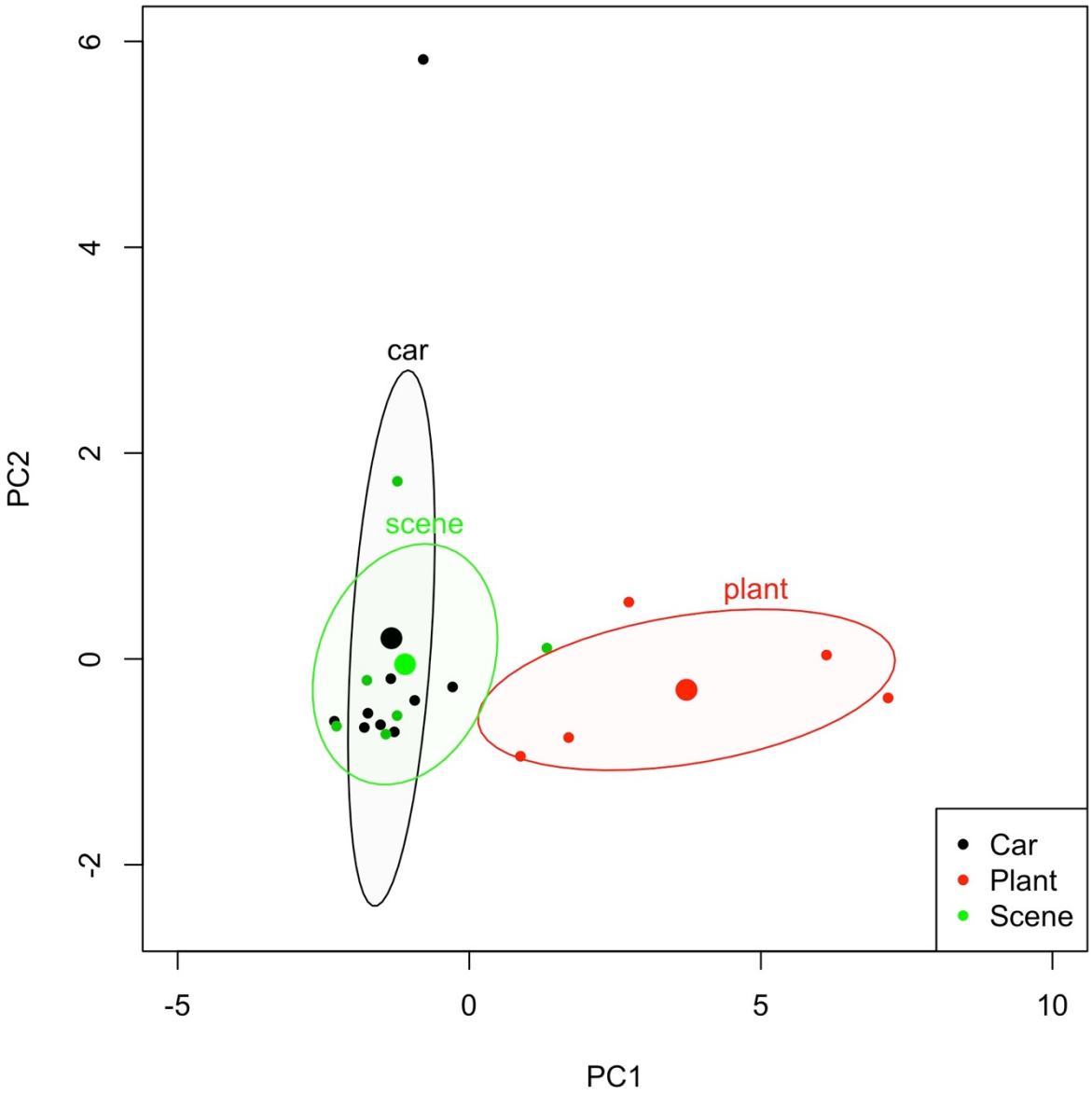

Fig. 5 Scatterplot between principal component 1 (PC1) against principal component 2 (PC2). Ellipses show the different location groups for car, plant, and scene.

4. Conclusions

Scatterplots and boxplots revealed the unique distribution of concentration among the variables that hinted the possibility of discriminating the samples based on the location of the sample. Cluster analysis was not able to distinguish between the locations with many overlaps between the location groups. Principal component analysis was able to separate the location groups into two main groups: car-scene and plant. This suggests that the car and scene samples share similar characteristics and source.

5. References

R Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing , Vienna, Austria. URL http://www.R-project.org/.