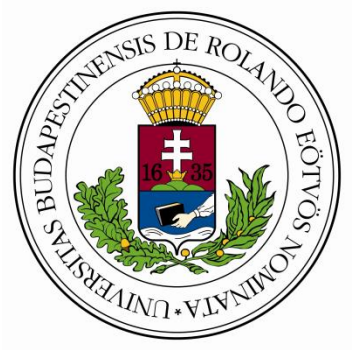
Investigation of Seriation on Chemical Data

Thesis work
Chemistry BSc

SASAN AMARIAMIR

Supervisor: Gergely Tóth Assoc. Professor
ELTE, Laboratory of Chemical Informatics



Institute of Chemistry, Faculty of Science,
ELTE Eötvös Loránd University, Budapest, Hungary

Place of defence: Department of Physical Chemistry

2017

Contents

Introduction

Seriation is an investigative combinatorial data analysis technique to order objects and/or variables into a sequence along one, two or more dimensions to help reveal regularity and patterning among the entire series. Seriation can be imagined where data mining, information visualization, and network science meet. Sometimes called “sequencing”, it is the attempt to arrange all objects of a set in a linear order given available data and some loss or merit function in order to find a suitable linear order for the set of objects to reveal structural information. Seriation can also be thought as a series of links between objects, which is a non-redundant and optimal presentation of a possibly very long list of linking rules; it introduces structural context to those relationships (Liiv, 2010).

Together with cluster analysis and variable selection, seriation is an important problem in the field of combinatorial data analysis (Arabie, et al., 1996). Data scientist Shneiderman has written ‘most books on data mining have only a brief discussion of information visualization and vice versa’ and that ‘the process of combining statistical methods with visualization tools will take some time because of the conflicting philosophies of the promoters’ (Shneiderman, 2002). Seriation is not easy to solve for all but very small sets due to the problem's combinatorial nature. Nevertheless, both exact solution methods and heuristics are available (Hahsler, et al., 2008). Using seriation and matrix reordering can lead to finding patterns at local fragments level of relationships, pairs of organized local fragments of relationships, and a global structural pattern at the same time. A seriation and matrix visualization results in clustering of data with further information about how one cluster is related to another, what the linking objects or variables are and in what direction the objects inside a cluster change.

After searching in chemistry papers, it seems like chemists have not heard of this idea as there are barely any chemistry papers found on this topic. The only article name that I found which mentioned seriation in chemical papers was published in 1990 (Bartel, 1990) and yet I could not even find the body of the article. There is another paper in chemometrics published in this university by my advisor on this very subject without using the name seriation which has been discussed in detail and used further in this work (Tóth, et al., 2009). For the purpose of this investigation the old method was extended and picked up new features including sorting data either diagonally or anti-diagonally, clustering methods based on the object distances, covariance matrix of variables and permuting row and

columns either dependently or independently. Also a pre-processing of data was taken up in most cases as it proved very useful, if not necessary, to modify the data by scaling or processing the rank matrix instead of the raw data.

This all seems like a waste of a potentially very useful tool for analysis; the opportunities are virtually limitless: whenever a chemist is facing a sizable matrix full of data with not very clear correlation links, seriation can come in handy (at least begin to) save the day and suggest which variables or objects should be investigated together more closely. In this work for example we try to find a hidden pattern in ancient Hungarian coins based on the metal content, order samples of wine based on their ion content and have guesses about which ions would affect the aroma of wine heavily and in the end try to get close to the correct sequence of the elemental reactions of combustion mechanism where the ordered reaction-component matrix might hint to a hidden aspect on this type of large chemical data sets of combustion chemistry.

The main goal of this work is to talk briefly about the history of seriation by going through the literature and looking for the use of the method in a wide range of disciplines from archaeology to biology, suggest available tools for implementing it and finally try some of them for seriating a few sets of chemical data and present the result graphically. By doing so we should present enough information to let the reader decide for themselves if seriation is a tool they would like to keep handy in the future as they deal with data sets.

The calculations in this work were done my supervisor and myself. Most of the work and decisions were managed in joint sessions yet I focused more on R-package seriation while my supervisor contributed more in his method (TS method) of calculation.

Literature Review

The opportunity to carry out seriation is mostly in to object-to-object and object-to- variable data tables, or by Tucker's terminology (Tucker, 1964) we are concentrating on two-way one-mode *(*$N \times N$; square tables, where rows and columns refer to the same objects) and two-mode ($N \times M$; rectangular tables, where rows and columns refer to two different sets of entities, as in objects and variables) data tables.

From these matrices, tables or plots, it is still not the most straightforward task to figure out the underlying relationships in the data, look for the patterns and guess the overall trend. Objects in such an adjacency matrix normally are ordered subjectively, possibly in the order of data acquisition/generation, which might indeed prove useful at times, or just sorted alphabetically by labels or names. Therefore, permuting the order of rows and columns, does not change the meaning of the data: there are $N!$ permutations of the same symmetrical matrix (or $N!*M!$ in case of a $N \times M$ matrix) will explicitly return the identical meaning of the system under observation. This can also be thought of in the perspective of a single element (cell), the position of which can be changed with the constraint that it must always be moved together with the whole row or column—making it somewhat similar to the classical game of Rubik's cube. (Liiv, 2010) To seriate is to find such a permutation, to reorder the objects from the same mode in a sequence so that it best shows regularity and patterning among the entire series.

Solving problems in combinatorial data analysis entails the solution of discrete optimization problems which generally involves evaluating all plausible solutions. Since the number of possible solutions grows with problem size (as in $N!$ or $N!*M!$) due to the combinatorial nature, a brute-force enumerative approach becomes implausible for all but very small problems. To solve larger problems (currently with up to 40 objects), partial enumeration methods can be used (Hahsler, et al., 2008). For example Arabie (Arabie, et al., 1996) propose dynamic programming and Brusco and Stahl (Brusco, et al., 2005) use a branch-and-bound strategy. For even larger problems only heuristics can be used. Also there are global optimization methods including some biologically inspired ones e.g. neural networks.

A search for implementation of seriation or similar practices of reordering tables to discover hidden patterns in the literature reveals that this idea has been taken advantage of historically in several disciplines such as: archaeology and anthropology; cartography,

graphics, and information visualization; sociology and sociometry; psychology and psychometry; ecology (where the idea is known as Ordination); biology and bioinformatics; cellular manufacturing; and operations research. (Liiv, 2010) Techniques related to these applications are still popular. For instance, on the R programing platform, several R-packages related to these fields already exist, e.g., ade4 (Chessel, Dufour, and Dray 2007; Dray and Dufour 2007) and vegan (Oksanen, Kindt, Legendre, and O'Hara 2007) cited by (Hahsler, et al., 2008) plus the R-package simply called "Seriation" which was used later in this work.

It seems that the first person who used Seriation as a formal method was an archaeologist named Petrie (Petrie, 1899). He used it to work out a chronological order for graves found near the Nile based on the objects found there. He used a cross-tabulation of grave sites and objects. He then went forward to reorder the table using row and column permutations so that the larger values were as close to the main diagonal as possible. In this reordered table, graves with objects in common were grouped with each other. Presuming that different objects regularly come into and go out of use, one can conclude the order of graves in the reordered table estimated a chronological order. Back then, the reordering of rows and columns of the table was done manually, the reliability was estimated subjectively by the researcher using visual perception. Actually, this was how it was done all the way up to the 1960s and 1970s by a research group directed by a French cartographer Jacques Bertin (Bertin, 1981) (p. 47), who stated that, with assistants and mechanical devices, 'it only takes three days to construct a matrix and three weeks to process and interpret it more deeply'. This has now become easier thanks to computers. Robinson (Robinson, 1951), Kendall (Kendall, et al., 1971) and others proposed measures of agreement between rows to quantify optimality of the resulting table. A comprehensive description of the development of seriation in archaeology is presented by Ihm (Ihm, 2005).

Bertin was working on data visualization for decades. In 1967 in a paper called Semiology of Graphics (Bertin, 2011) he presents illustrative examples on seriation alongside the main arguments. His idea was to propose a readable matrix - matrice ordonnabl)- as a handy tool for analysing systems of data. He attempted reorganizing rows and columns in what we now call the two mode manner and presented the results visually. He was aware of the significance of the readability of the three information levels in visual displays: 1- the

details of data tabulated in rows and columns, 2- the local patterns found in the data, 3- global patterns found in the data.

In Sociology, L. Moreno wrote a paper "Who Shall Survive?" (Moreno, 1934) in 1934 which possibly founded the branch of Sociometry by emphasizing on utilization of quantitative and mathematical methods for understanding social phenomena. His work inspired Forsyth and Katz who were the first the make use of reordering rows and columns of a "sociomatrix" to better present their data (Forsyth, et al., 1946) They recognized the new form of data representation superior to the verbal descriptions and relationships listings, even though confusing to the reader when the number of subjects is large. Also in psychometry, which might be sociometry's younger sibling, adaption of seriation methods has been known to happen. Hubert (Hubert, 1974) was one of first who implemented seriation in psychology who contemplated a subjects by item response matrix. He used binary matrix data –1 for existence of a relation and 0 for no relation- in both one mode and two mode seriation using the algorithms which already existed for archaeological seriations.

In the field of biology, it probably were Sokal and Sneath (Sokal, et al., 1963) who sparked the idea of reordering rows and columns in their paper Numerical Taxonomy. This would go on to suggest changes to the traditional way of creating taxonomies in biology. This paper talked about "differential shading of the similarity matrix" and called the result by the appropriate name of "a cleaned up diagram". They followed the method developed by Robinson (Robinson, 1951) in search for the "optimum structure in the system". Closer to our time in 2005 an ecology paper was published by our university researchers about rearrangement of ecological data matrices using Monte Carlo simulations which referencing seriation methods in ecology (Miklós, et al., 2005).

In the field of Chemometrics the notion of permuting rows and columns to visualize new hidden patterns in the data does not seem to have been explored by many. I stumbled upon a thesis work in our university which uses seriation to investigate sequence alignment of proteins (Szabó Attila, 2010). The thesis uses a method introduced in a paper published a year before by Tóth and Szepesváry in 2009 (Tóth, et al., 2009). In this chemometric paper the idea of using a diagonal measure and a local distance matrix to order the matrix data has been explored. „Proper permutation of data matrix rows and columns may result in plots showing striking information on the objects and variables under investigation" claim

the authors. They argue that since visual aid is very influential in human perception, sorting the data rows and columns arbitrarily is not the most pragmatic way of presenting the data. A known examples in chemistry could be “the correlation matrix of the objects in the vector space spanned by the variables can be used effectively to show similarities of chemical systems measured by chromatography” ((Gyseghem, et al., 2006), cited by (Tóth, et al., 2009)). The diagonal measure *D* used in this paper is a scalar that represents the distribution of large and small elements of the matrix relative to the main diagonal. The local distance matrix on the other hand, explores the relation between the rows and columns of a data matrix. By optimizing this matrix one can group (cluster) similar elements, and their respective rows and columns together. By minimization of D the local distance matrix by row and column changes of the original data matrix, the similar objects are clustered close together and also the variables leading to this similarity are grouped together near the diagonal.

Throughout the literature, the most common ways to present the data are using a matrix, a double-entry table with labels, and a colour-coded graphical (or shaded) graphical plot. Seriation is clearly correlated to clustering, although there does not exist an agreement across the disciplines about defining their distinction (Liiv, 2010). To link the two methods one could think of clustering with optimal leaf ordering, which is the procedure ‘to order the clusters at each level so that the objects on the edge of each cluster are adjacent to the object outside the cluster to which it is nearest’ (Gunnar Gruvaeus, 1972) cited by (Liiv, 2010).

Methodology

How to Seriate?

To seriate a set of *N* objects, one mode, we normally start with a Dissimilarity matrix D.

$$D = (d_{ij}),\ i \text{ is in the range } [1, N] \quad (1)$$

in data analysis "dissimilarity" and "distance" point to the same concept. They are related to "similarity" with the equation (2) (Bajusz D., 2015):

$$similarity = \frac{1}{1 + distance} \quad (2)$$

D represents the dissimilarity between the *i*-th and *j*-th object. The elements on the diagonal are equal to zero (an object is not dissimilar to itself). A permutation matrix/function P is defined. P rearranges the objects in D by permuting rows and columns. Different "loss" or "merit" functions have been defined to judge the usefulness of these permutations in visualizing the clusters and their relations. The loss functions indicate the better sets of permutations by generating smaller values while merit functions do this by getting larger values. The object of seriation is to find the particular permutation which minimizes a loss or maximises a merit function.

This dissimilarity matrix of course represents only one set of objects (one mode) even though they are tabulated in two dimensions (two mode). Seriation for two mode data is also possible if the matrix resembles a dissimilarity matrix (has no negative entity). The optimization of the merit of loss functions in the two mode data depends on both the columns and rows. These can be dependent on one another in some functions or be independent in others, which allows snapping the problem into two: optimising the order of rows and the order of columns. In the latter scenario one can calculate the dissimilarity matrix for each mode and solve them as one modes. Further still, seriation can be generalized to *k*-way *k*-mode data in the form of a *k*-dimensional array by breaking the problem down into several lower dimensional independent problems (Hahsler, et al., 2008). The advantage of breaking the problem into smaller chunks is to significantly lowering the number of possible enumerations, ergo simplifying the combinatorial problem. In a *k*-mode problem with *N* number of objects where each mode can be calculated separately, the total

number of possible permutations is $\sum_1^k N!$, while if in the same problem the merit/loss functions depend on all the *k* modes, the total number of possible permutations will be $(\sum_1^k N)!$ which is incomprehensively bigger for any k and *N* bigger than 4 or 5.

Common Merit and Loss Functions

A. Column/Row Gradient Measures: An perfect anti-Robinsonian matrix is a one mode dissimilarity matrix where the values *only* increase as they move away from the main diagonal (Robinson, 1951). In other words:

Within rows $$d_{ik} \leq d_{ij} \quad for \quad 1 \leq i < k < j \leq n \tag{3}$$

Within columns $$d_{kj} \leq d_{ij} \quad for \quad 1 \leq i < k < j \leq n \tag{4}$$

In an anti-Robinsonian matrix the smallest dissimilarities are close to the diagonal. This provides a path to seriation. An appropriate merit function which indicates 'divergence' from the anti-Robinsonian was worked out by Hubert (Hubert, et al., 2001):

$$M(D) = \sum_{i<k<j} f(d_{ik}, d_{ij}) + \sum_{i<k<j} f(d_{kj}, d_{ij}) \tag{5}$$

Where f is a function defining the violation or satisfaction of a gradient condition. *f* can be defined as f(*z*,*y*) = sign (*y*-*z*) which will result in the raw number of triples satisfying the gradient constraints (Hahsler, et al., 2008). It also can be defined while considering the weight of each element in the form *f*(*z*,*y*) = | *y*-*z* | sign (*y*-*z*) = *y*-*z*

B. Anti-Robinsonian Events: This loss function is built in the same way as the gradient measures but has fewer outputs ergo is simpler.

$$f(z\,, y) = I\,(z, y) = \begin{cases} 1 & \text{if } z<y \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

By counting only the violations (Chen, 2002). To account for the weights one can use $f(z\,, y) = |y - z| I\,(z, y)$ instead.

C. Hamiltonian Path Length: This function operates on the basis of presenting the dissimilarity matrix as a finite weighted graph (Hahsler, et al., 2008) G = (O,E) where O are the vertices (objects) and $e_{ij} \in E$ represents the edges between the *i*-th and *j*-th object with the weight w_{ij} which represents d_{ij} in the dissimilarity. This graph can be used for seriation (Hubert, 1974). Minimizing the Hamiltonian path (a path through which each node is visited only once) results in a seriation loss function through considering the dissimilarity between neighbouring objects:

$$L(D) = \sum_{i=1}^{n-1} d_{i,i+1} \tag{7}$$

D. Measure of Effectiveness (ME): This merit function was first defined in the paper "Problem Decomposition and Data Reorganization by a Clustering Technique" (Mccormick, et al., 1972)

$$M(X) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}[x_{i,j+1} + x_{i,j-1} + x_{i+1,j} + x_{i-1,j}] \tag{8}$$

With the convention X is a $N \times M$ data matrix and $x_{0,j} = x_{n+1,j} = x_{i,0} = x_{i,m+1} = 0$. ME is maxed if elements are closely related to their four neighbours.

It should be noted that this equation can be separated into two parts for row and columns. This means that the merit functions for rows and columns are independent.

E. Stress: Compares the values in a matrix with their neighbours in a two mode matrix. Niermann (Niermann, 2005) defined two types of neighbourhoods for two mode matrices:

The Moore neighbourhood:

$$\sigma_{ij} = \sum_{k=\max(1,i-1)}^{\min(n,i+1)} \sum_{l=\max(1,j-1)}^{\min(m,j+1)} (x_{ij} - x_{kl})^2 \tag{9}$$

And the Neumann neighbourhood:

$$\sigma_{ij} = \sum_{k=\max(1,i-1)}^{\min(n,i+1)} (x_{ij} - x_{kj})^2 + \sum_{l=\max(1,j-1)}^{\min(m,j+1)} (x_{ij} - x_{il})^2 \quad (10)$$

It must be clear that in the Neumann neighbourhood, unlike Moore's, the influence of rows and columns are independent.

In both of these cases a global stress measure can be built up into a loss function by summing up all the elements:

$$L(X) = \sum_{i=1}^{n}\sum_{j=1}^{m} \sigma_{ij} \quad (11)$$

F. Tóth-Szepesváry Diagonal and Local Distance Matrices: In this a diagonal measure and a local distance matrix are defined to be optimize. Optimizing these simultaneously puts the similarities close to the main diagonal as well as seriating the data (Tóth, et al., 2009).

The diagonal measure is a scalar:

$$D = \sum_{i}^{N}\sum_{j}^{M} t_{ij}|x_{ij}| \quad (12)$$

Where t_{ij} is a measure of the graphical distance of the x_{ij} element from the main diagonal. The unit of distance is related to the numbering of rows and columns.

$$t_{ij} = 1 + \frac{\left|\frac{j-0.5}{M} - \frac{i-0.5}{N}\right|}{\sqrt{\frac{1}{N^2} + \frac{1}{M^2}}} \quad (13)$$

The local distance matrix L works on grouping the similar rows and columns together:

$$L_{ij} = \frac{\sum_{k=i-I}^{i+I} \sqrt{\sum_{I=j-J}^{j+J} (x_{iI} - x_{kI})^2}}{2I} \quad (14)$$

usually I = J = 1.

This local distance matrix (LDM) is different from Stress in the sense that instead of working with stress on all directions, the short (length of 2) horizontal vectors above, below and the same row of an element are calculated and then averaged. Also unlike Stress, the LDM does is not a second power parameter.

Computation

Seriation Tools

R-Package Seriation

In this package of the R language, data structures and some algorithms for seriation are provided. A big part of the seriations carried out in this work have been done by taking advantage of this package. This package provides the means for two-way one-mode (class **dist**), two-way two-mode (class matrix) and k-way k-mode (class array) forms of seriation. However, we primarily used the two-way two-mode functions of this package as they proved the most pragmatic in practice.

A basic two-mode seriation goes this way in this package: A sequence of data in matrix form is read into R. This can be the raw data or scaled/ranked matrix (see data processing). If there is any nominal column/rows in the data matrix they need to be cut off prior to the calculation for the algorithm will not work on any but numeric data. If the nominal data is presented in numeric forms, e.g. number of streets, this too need to be separated from the data as the nominal numbers do not present the mathematical meaning that the algorithm would blindly assume. These nominal columns/rows can later on be reattached after the seriation has been carried out.

The **seriate()** function is then applied to the said matrix; a seriation method needs to be fed into the **seriate()** function if any method but the default one is desired. The result which is the seriated sequence is saved into a variable like **o.**

To get the heat-map of the reordered matrix, the command **pimage()** is used. The arguments to be fed to the function are the original matrix, the permutation vector **o** and the title of the figure, respectively; if the heat-map of the original data is desired, the seriated sequence is omitted.

To get the permuted matrix –the actual result of the seriation- one needs to use the **permute()** function. The function takes the original matrix and the new sequence, operates the permutation on the data and saves the result in a new variable which can be written as an output to an excel table.

A one-mode example from the paper ‘Getting Things in Order’ (Hahsler, et al., 2008) follows:

```
R> library("seriation")
```

```
R> data("iris")
R> x <- as.matrix(iris[-5])
R> x <- x[sample(seq_len(nrow(x))), ]
R> d <- dist(x)
R> o <- seriate(d)
R> pimage(d, main = "Random")
R> pimage(d, o, main = "Reordered")
```

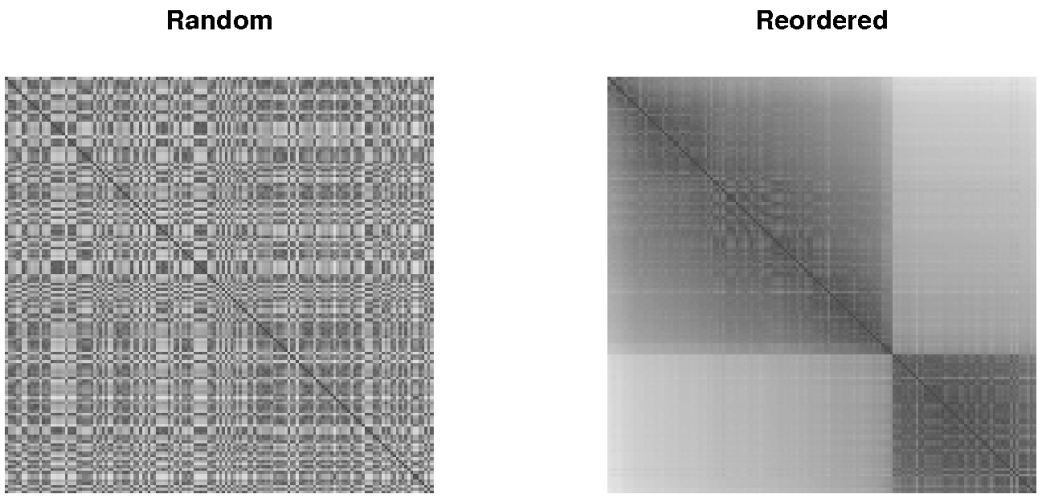


Fig 1-R-seriation (Hahsler, et al., 2008)

LDM_Seriation Code

A new version of the code in the Tóth-Szepesváry paper (Tóth, et al., 2009) was used to carry out some of the calculations. This new version included extensions which allowed for a number of new options such as data pre-processing (see data processing), choice between diagonalization and anti-diagonalization by minimizing and maximizing the scalar *D*, one-mode seriation based on the distance matrix for objects (rows) as option 1, one-mode seriation based on the covariance matrix for variables (columns) as option 2, seriation based on matrix data for rows and columns separately as option 3, and seriation based on the matrix data with dependant row and column permutations as option 4. Also as a new feature we could fix a number of rows or columns before the calculation so they would not move

in the permutations, giving us a better visual of the transition from/to states of interest. This new version of the code can be found here (Tóth, et al., 2017).

Seriation Methods

It should be noted that this work is not an algorithmic study on different seriation methods; we do not intend to investigate the details of each of the used methods but rather are going to apply them to different chemical data and observe the results.

Below the methods used in this investigation are briefly described.

Tóth-Szepesváry Method

This method was first published in 2009 (Tóth, et al., 2009). In that time the authors were not aware of previous investigation if the problem in the literature under the name Seriation; they had identified the problem in their own academic field and tried to come up with a solution. This ironically might have helped them from a certain "tunnel vision" as their method does present a certain advantage over the more familiar seriation methods. As mentioned above in the literature review and merit/loss functions, Tóth-Szepesváry (TS) method, takes advantage of two separate measures that guide the permutations toward a more seriated representation of the data. That is the diagonal measure D expresses the size relation of the matrix elements, the elements are weighted by their distance from the main diagonal. The local distance matrix focuses on the neighbours of each element and their similarity; permutations leading to higher overall similarities will then cluster the similar rows and columns together., It is important to note that the "optimizations" of these two happen in parallel to each other–in oppose to happening in consequently- Ergo, not only one could cluster the objects, which is not new in seriation, but they also would: a) cluster the variables responsible for the similarity between objects as the large objects are placed on the diagonal or b) cluster the objects of importance (and their variables) on the opposite corners of the matrix in the anti-diagonal case; this comes in handy when there are many zeros (or similar minimums) in data points. Monte Carlo technique with importance sampling is used to achieve the optimization goals of the diagonal measure. The procedure starts with calculating the matrix to be optimized, B_{old}, which can be the original matrix, a distance matrix or a covariance matrix from the raw data. Then the diagonal measure of this matrix is calculated called D_{old}. Two rows or columns of the original data are exchanged. The new diagonal measure is then calculated D_{new}. The change is accepted if D_{new} is varied in the desired direction; $D_{new}>D_{old}$ is desired for putting the larger values on

top-right and bottom-left corners of the matrix (ant-diagonal optimization) while $D_{new}<D_{old}$ is required for placing the large values close to the main diagonal (diagonal optimization). If the change is not desired, the result is accepted only with the probability of exp ($|D_{old}-D_{new}|/F$) (F is temperature factor in importance sampling).

R-Package Method i) Bond Energy Algorithm

The bond energy algorithm BEA (Mccormick, et al., 1972) is a method concerning with reordering rows and columns of matrices in a manner to increase the similarity between each element and its four closer neighbours (for the element of the *i*th row and *j*th column, they would be the elements x_{i-1j}, x_{i+1j}, x_{ij-1}, x_{ij+1}). This method works on the basis of maximizing the ME (Measure of Effectiveness) merit function (Hahsler, et al., 2008). Rows and columns can be worked out independently in this method. Three steps are carried out in this procedure:

Firstly, a column is place in a position randomly.

Secondly, each remaining column is placed randomly on different positions around the first column and the change in ME function is calculated. After trying all permutations are calculated, the one which gives the highest increase in ME is kept.

Thirdly, the same notion is applied on the rows.

R-Package Method ii) Travelling Sales Person to Optimize ME

As mentioned before, solving the Travelling Sales Person (TSP) problem is carried out by minimizing the length of the Hamiltonian path through a graph. In the seriation case with $n+1$ rows/columns (cities), $n!$ round trips need to be checked. Yet, even with this large searching space, small cases can be solved adequately using dynamic programming and larger cases of several hundred objects can be solved using branch-and-cut algorithms (Hahsler, et al., 2008).

R-Package Method iii) First Principle Component Analysis, First Two Principal Components (Angle)

These two methods take advantage of the Principle Component Analysis (PCA) statistical procedure. They investigate the linear correlation of each variable (columns) together in order to group them in bigger factors or cluster the objects with higher correlation with each other. The seriation package basically operates as a visual/permuting output of the principle components.

Data Pre-Processing

Seriation can be carried out on raw data. Nevertheless, we had good reason to pre-process the data in most cases before the seriation. Firstly, as the seriation algorithms for distance matrices work on the basis of comparison of the size of the elements, a sort of scaling was called upon to be able to get meaningful comparisons between variables. Otherwise, one sort of data e.g. altitude in meters in mountains will always trump the data of temperature in degrees Celsius for the same object one being in the scale of thousands and the other in the scale of (negative) tens. To this end "internal standardization" might be the first tool that comes to mind which is subtracting by the mean and dividing by the estimated standard deviation. But it turns out most seriation algorithms will not work with negative numbers. So, in order to scale the data in a reasonable positive range another scaling method was implemented on each column: the minimum of entity of each column was subtracted from every elements of the column and the result was divided by the difference of the minimum and the maximum of the column. This successfully scaled our data in the range of [0,1] allowing the seriation procedure to run correctly.

Secondly, on the few instance that raw data preferred to be used in the seriation process, or those where the nominal data columns were in numeric form, this scaling had to performed after the seriation and before plotting the heat-maps. Otherwise, the off range columns would shift the shading end of the graph to such an extreme that the other columns were barely noticeable.

There were a few cases that it seemed practical to convert the data into a binary input: where the investigation of objects processing a shared variable was of interest rather than the intensity of the variable, the data was converted into a matrix of 0s and 1s; zero meaning the object does not have the variable of that column and 1 being affirmative on possessing the variable.

In one or two cases, we experimented of converting the data matrix into the rank matrix before performing the seriation to see whether the clustering of similar objects would improve.

[1] It should be noted that most of the data in this work has been scaled before computation. In rare cases, like the Iris data set, where scaling was not necessary to seriation, it still had

to be carried out before plotting the heat-maps. So the heat-maps only serve as visual tools and the entities are unit-less.

Data Sets

IRIS

The well-known Iris benchmark data base (Anderson, 1935)is commonly used for introduction of the basic concepts of data analysis. Data scales (nominal) must be accounted for because before starting the seriation operations as mentioned before. Numerical data was represented in matrix form of this data.

This data frame consists of five columns and 150 rows. Each row represents a flower under study. The columns represent the number of the species, petal length, petal width, sepal length and the sepal width respectively. The first 5 rows of the data set are:

species petal-l petal-w sepal-l sepal-w

1 1 5.1 3.5 1.4 0.2

2 1 4.9 3.0 1.4 0.2

3 1 4.7 3.2 1.3 0.2

4 1 4.6 3.1 1.5 0.2

5 1 5.0 3.6 1.4 0.2

Wine Chemical Components

This data base is from a study which has applied a multivariate regression method to the chemical measurements of Pinto Noir wine samples (Frank, et al., 1984).

This data frame is a 38*19 matrix. The first row is nominal data, numbering of the wine sample being processed. The last column is “Aroma”, given a subjective scalar for how strong the aroma of each wine sample is. The columns in the middle present the amount of different elements in each wine sample; the elements being Cadmium, Molybdenum, Manganese, Nickle, Copper, Aluminium, Barium, Chromium, Strontium, Lead, Boron, Magnesium, Silicon, Sodium, Calcium, Phosphorus and Potassium.

Coins Composition Data

This data set describes the concentration of different components (metals) in a series of silver coins belonging to the Hungarian Árpád Dynasty, AD 997–1301 (Christie, et al., 2014). The data was originally published in 2013 (Rácz, et al., 2013). In the 2014 paper the

authors attempt to reclassify the coins based on their chemical data and statistical methods; not unlike what we hope to achieve here.

The concentration of different metals in the mix can reveal interesting glimpses to the history of the coins. The first correlation that comes to mind might be that one can classify the coins with similar compositions to estimate which ones were from the same origin e.g. silver mine. However, taking into account that each new king would regularly order to melt the old coins and press new ones, one could potentially follow the trail of impurities to say something about the original era when the coins were made. Or there could be an estimation about the economical situation of a certain area/era if the silver content of coins begins to drop.

The data frame comes in a 258*13 matrix form with the last three columns being the nominal data. The rest are the concentrations of the metals Titanium, Iron, Nickel, Copper, Zinc, Silver, Tin, Antimony, Lead and Bismuth.

Tamil Nadu Natural Radioactivity Data

This data set was collected from a study on determining the natural radioactivity and its hazards in sands in Tamil Nadu, India (Hariprasath, et al., 2016). In that region, construction projects are using the sand from the river beds as raw material. The radiological impact of these sands were measured by gamma-ray spectroscopy. The results are presented in a data frame which we attempt to seriate here.

The data frame comes in 30 rows and 14 columns apart from the nominal data. The rows are named after the places were the radiation measurements were conducted. The columns cover a wide range of radiological measures and activities. They are specific activities (sa) of ^{238}U, ^{232}Th and ^{40}K and their absorbed dose rates (adr) and the total radiation dose as well as other radiological parameters relating to radiation hazard including radium equivalent activity (Reaeq/ $Bq.Kg^{-1}$), annual effective dose equivalent (AEDE/ $\mu Sv.y^{-1}$) for outdoors and indoors, external radiation hazard (Hex), internal radiation hazard (Hin), gamma activity concentration index (I) and finally Annual gonadal dose equivalent (AGDE/ $\mu Sv.y^{-1}$).

Toxin Data Set

This data set comes from a study on structure-activity and toxicity of certain compounds (Arthur, et al., 2017). The research uses 112 anticancer compounds (rows) and develops

QSAR and QSTR models to predict the activity and toxicity of newly designed compounds. The data includes both 4 pairs of columns dedicated to different factors of toxicity; one of which being the predicted value and the other the experimental measure. Should the prediction been valid, one would expect a high correlation between the two members of each pair, ergo similar seriation clustering.

Combustion Reactions

This set of data comes from the BSc Thesis of Gergely Juhász (Juhász, 2015). He studied the application of biodiesel fuels and the mechanism if their combustion reaction in an attempt to reduce their numbers by selecting out "redundancies". The discovered reaction mechanisms of this sort are too complicate for pragmatic purposes such as calculation of designing the reactor. So it is worthwhile to try and reduce the number of reactions in a way that the simulated results of the reduced and original form are almost identical. In his thesis work Juhász took a data base of reactions consisting of 61 species and 172 reactions (Wesbrook, et al., 2011), applied the SEM-CM mechanism reduction method to them and achieved a similar system with 33 species and 95 reactions (Juhász, 2015). It is the reduced mechanism which we took on for seriation.

This data first needed to be recorded in the matrix form. This was set up in the following configuration: the reactions were placed as row names and the compounds as column headers. Taking into account that these are so called "elemental reactions" and almost all the reaction coefficients are one, the binary form of the data was used: each specie present in a particular would be assigned the value 1, otherwise it would be assigned 0. Furthermore, one-mode matrices from this data was also produced. For the rows (reactions) they were assigned 1 when the was a component common between them, and for the columns (components) they were assigned 1 if there was a reaction they both took part in; otherwise the cell would be assigned the number 0.

Results and Discussion

In the section below our results are presented in the form of heat-maps. As a reminder to help the reader from confusion, here is a list of abbreviations written below the figures:

- sc_ data has been scaled between 0 and 1
- rank the rank matrix has been fed to the seriation process instead of the

data matrix.

If not specified, raw data matrix has been used.

- TS Tóth-Szepesváry method
  - 1 option 1, object-object distance matrix has been used
  - 2 option 2, correlation matrix for variable has been used
  - 3 option 3, L calculated on data matrix was anti-diogonalized
  - 3min option 3min, L calculated on data matrix was diogonalized
  - 4 option 4, like option 3 but rows and columns permute simultaneously
- R R-package
  - PCA First Principal Component method
  - BEA Bond Energy Algorithm method to maximize ME
  - BEAT Traveling Sales Person to maximize ME

IRIS

We start with the Iris data set. This data set was actually already clustered into three different species, maybe not the most beneficial for the seriation method. Yet it seemed like a very good candidate for the first demonstration of our work. Figure 2 shows the data in its original order before seriation. As mentioned before, the data was not scaled before the seriation as all the columns were in close range; however, scaling was needed to be carried out before plotting here and after as the column numbers (nominal data in general) is not in the vicinity of our data points.

Figure 3 demonstrates our first attempt at seriation. The Bond Energy Algorithm method was implemented on the two-mode data. It is clear that column permutation is not very useful in this data set due to the very low number of columns.

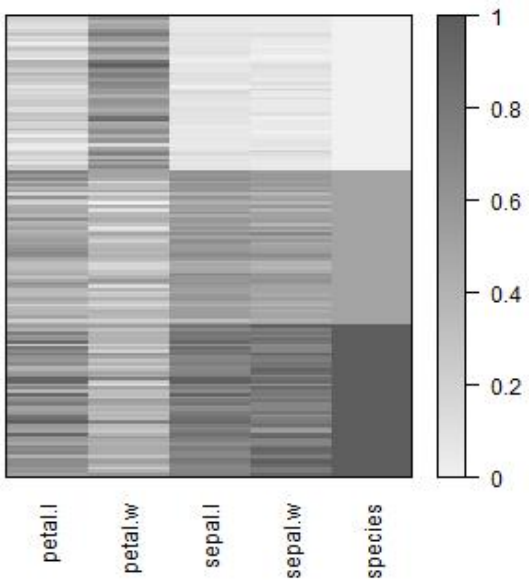


Fig 2-original_Iris

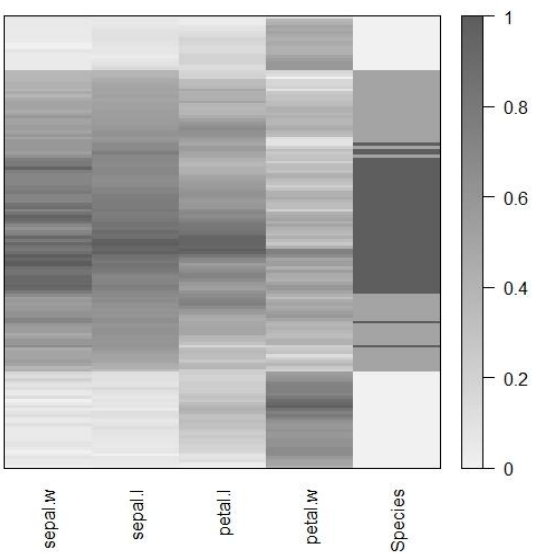


Fig 3- R-BEA Iris

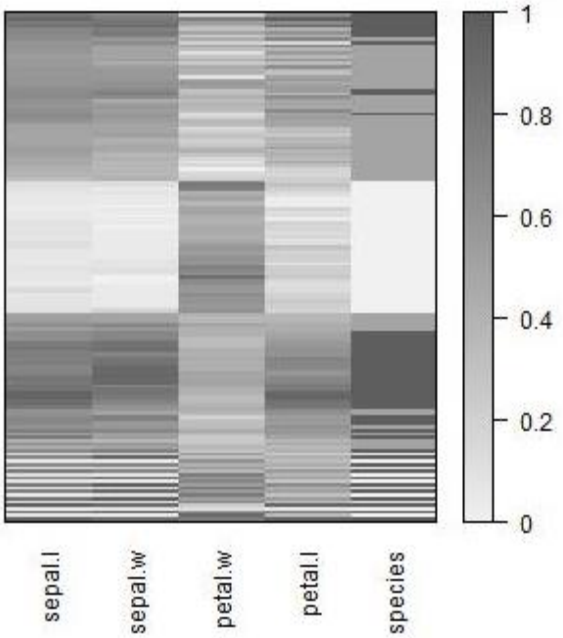


Fig 4-TS3 Iris

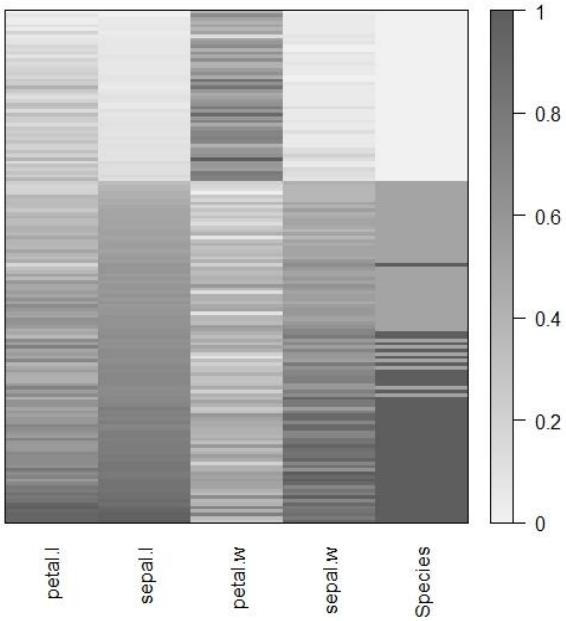


Fig 5-R-PCA Iris

In figure 4 the TS method with the third option and anti-diagonalization has been used. The middle of the graph seems very well ordered. Figure 5 is the graph related to the Principle Component Analysis method for the clustering. The first (lightest) specie has been fully

separated here. It is possible that if the petal area instead of the petal length was considered we would get better clustering.

Seriation does not necessarily need to happen between different clusters. As explained in the beginning of this work, seriation also attempts to order the objects inside the clusters in the manner which would demonstrate the pattern of change of variables from one local extreme to another. To show this, figure 5 has the Iris flowers seriated *within* the species by the TS method.

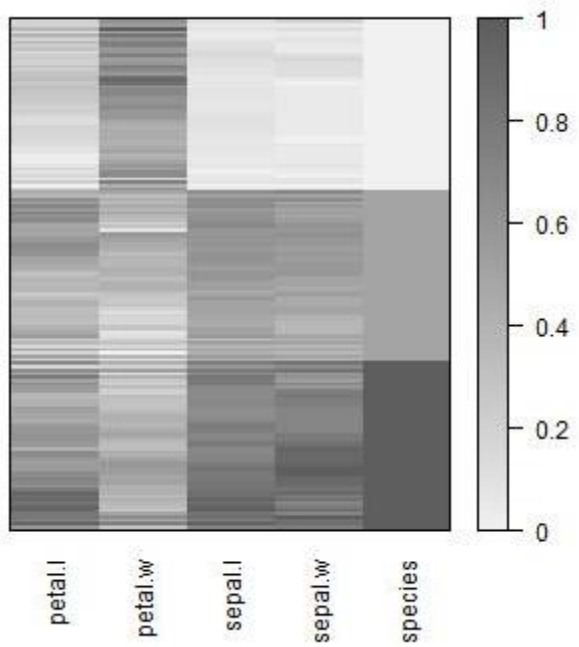


Fig 6-TS-Seriation within Iris Species

This "partial seriation" with keeping the species clusters intact has resulted in more homogeneous data fields and less intense stripes with clear trends compare to the original data.

Wine

In the wine data set we start with looking at the unseriated data; in the raw and [0,1] scaled form:

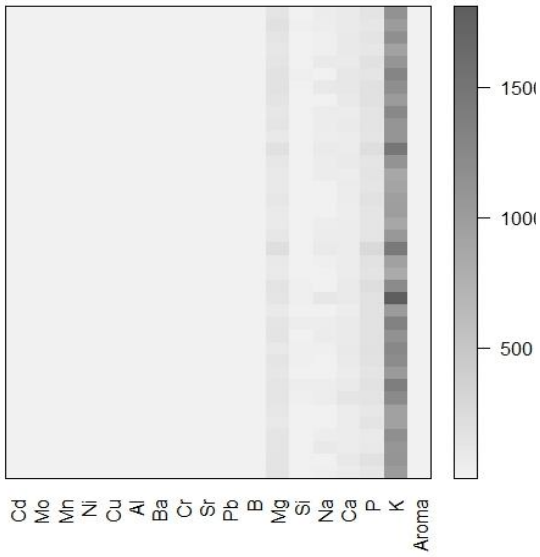


Fig 7-R-original wine

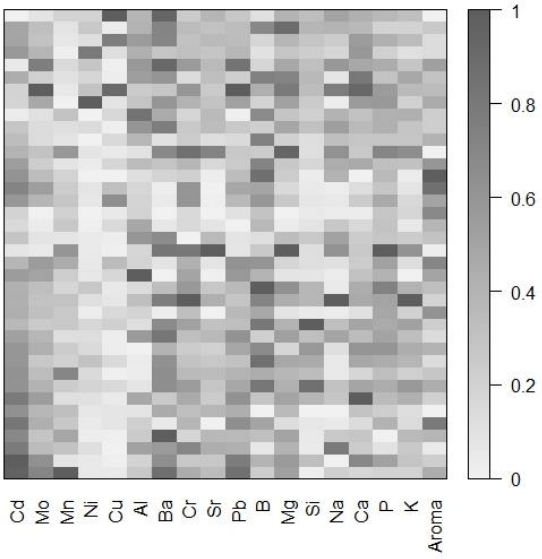


Fig 8-R-sc_original wine

One rapidly could notice that in the case of the raw data, the amount of Potassium is quite high in compare to the other elements; this would heavily weigh on the seriation process or at least, it dominates the visual field in the heat-map. In the case of the scaled data on the other hand, Potassium is as important as any other column. So using scaled graphs is favourable not to seriate almost solely the Potassium column.

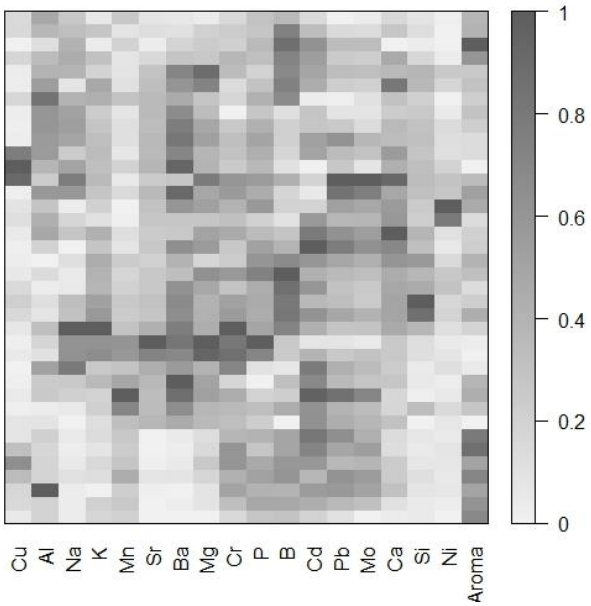


Fig 9-R-sc_BEA wine

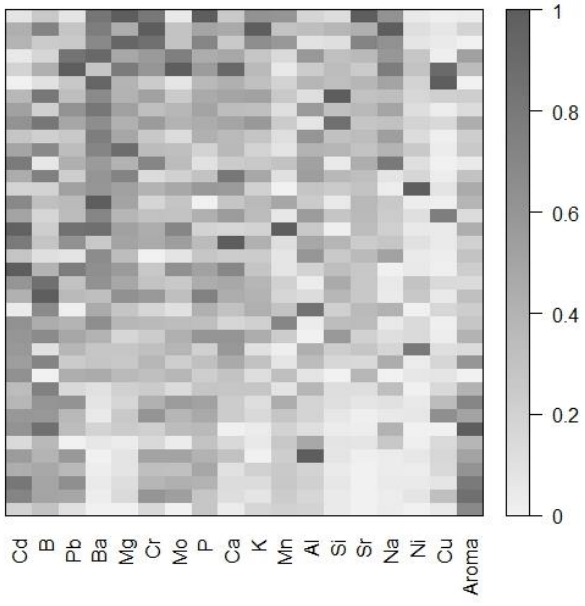


Fig 10-R-sc_PCA wine

Figure 9 demonstrates the h-map for the scaled data seriated by the BEA method and PCA method's result is shown in Figure 10. In both cases one could see that where high aroma numbers are clustered together (the subjective aroma column was not part of the seriation process!) some metals e.g. Sr, Ba, Ni, Si and maybe Ca have very low contents, suggesting they might affect the aroma of the wine negatively.

In the following figures 11 and 12, the scaled and unscaled matrixes have been seriated using the third option in the TS method; third option means that seriation of the rows and columns were independent and data was anti-diogonalized. This is an idea similar to the Moore's function except there is not (anti)-diagonalization in the Moore's method. It is worth mentioning the broad diagonal band in the case of scaled data in figure 11. TS method has seriated this matrix in the way so that similar entities would gather around the diagonal. It is again possible to draw connections between the wine aroma and certain ion contents.

Running the rows and columns seriation separately will usually have a satisfying result on the one-mode data but not so impressive when implemented on the two-mode. As an example, the working graphs of for the scaled rows of the wine data before and after seriation is included in figures 13 and 14. The graphs demonstrate the distance matrix of the wine samples where the seriated version is severely more homogeneous. The back substitution of this data on the two-mode object-variable matrix, however, does not such clear heat-maps.

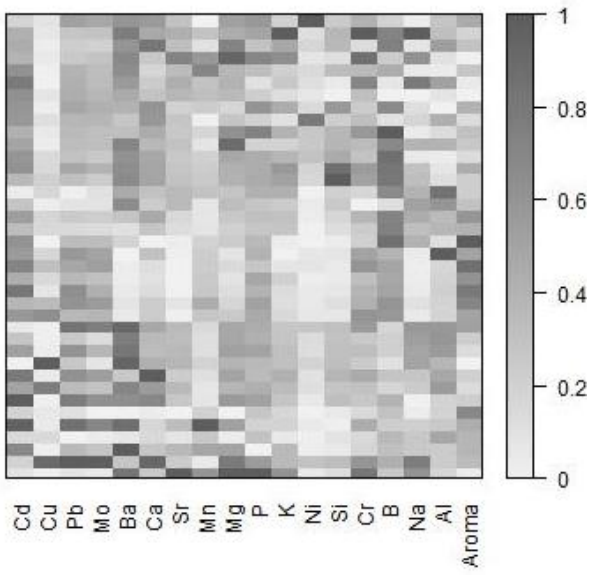


Fig 11-sc_TS3-antidiagonal wine

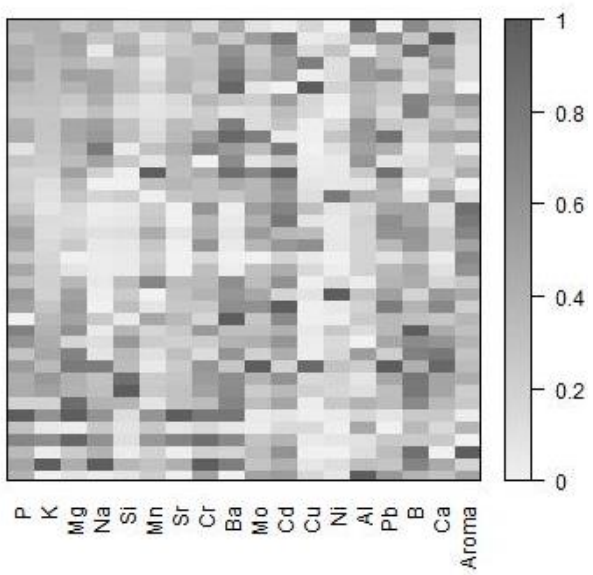


Fig 12-TS3-antidiagonolized wine

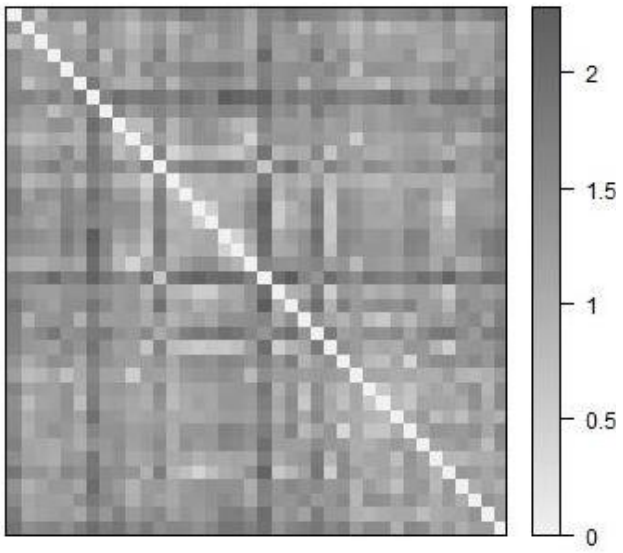


Fig 13-working matrix-TS1 original wine

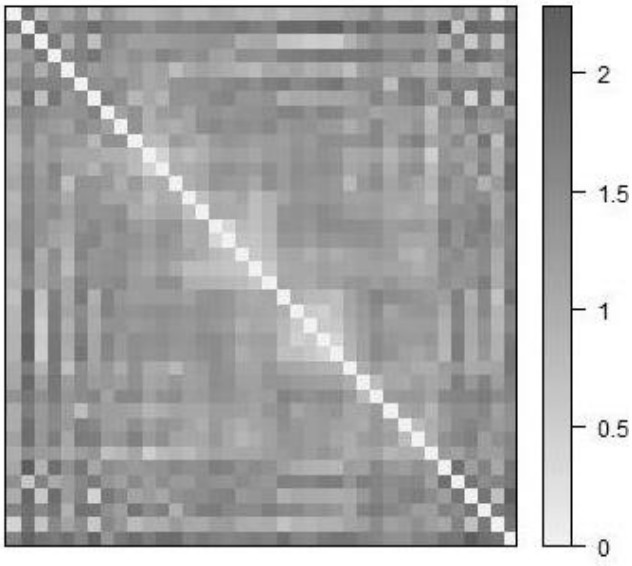


Fig 14-working matrix-TS1 wine

Coins

Scaling was quite important for this data set since we are dealing with metal contents of silver coins; the high percentage of the silver metal will paralyse the seriation if not scaled before computation. The original order of the scaled data is shown in figure 15.

The three nominal columns on the right were not involved in the seriation process but only in generating the heat-maps to help understand the effect of seriation.

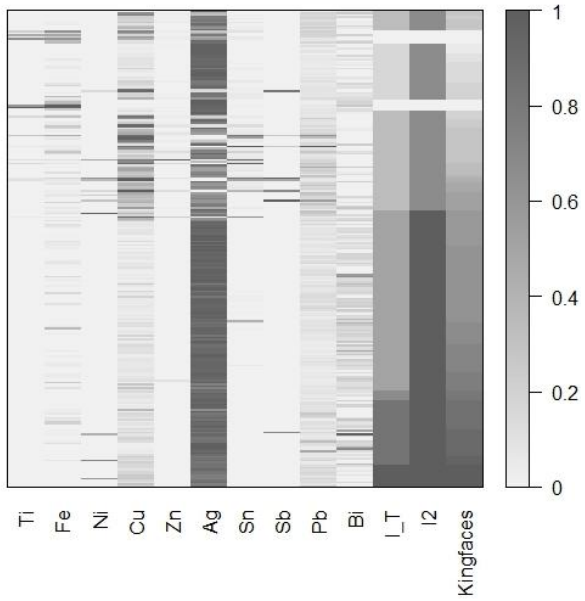


Fig 15-R-sc_original coins

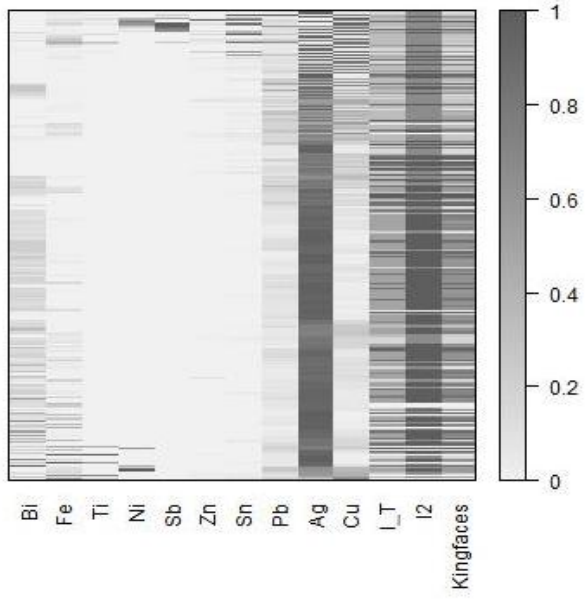
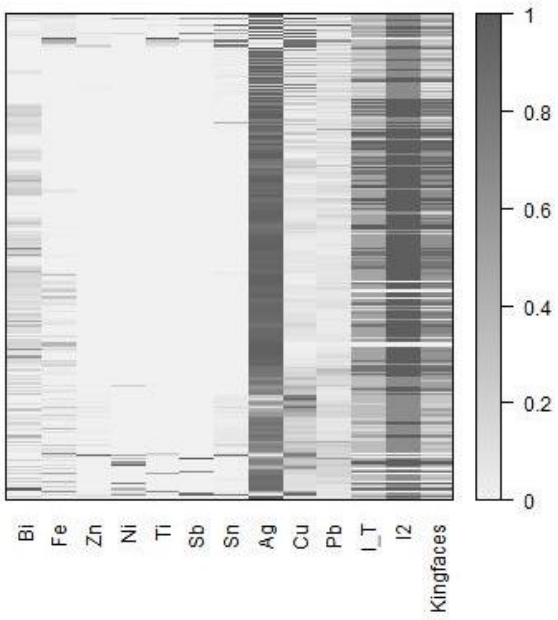


Fig 16-sc_TS3-antidiagonolized coins

In figure 16 the data (after being scaled) has been seriated by the anti-diagonalization two-mode (third) option of the TS method. It could be said that the I2 column which is a nominal data related to the era of different kings has been ordered to a noticeable degree.



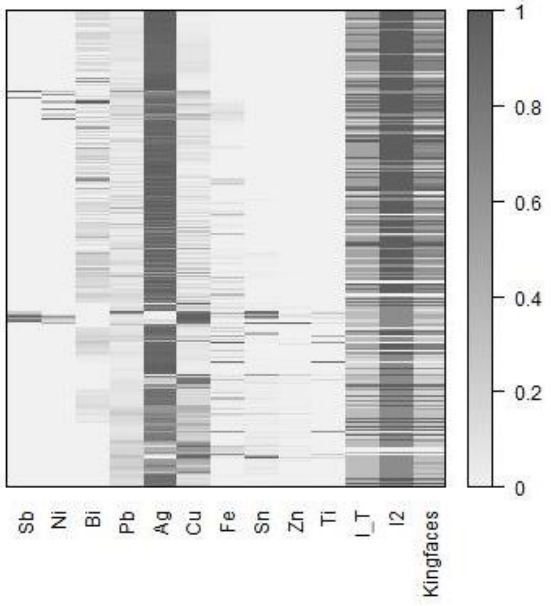


Fig 17-rank_TS3-antidiagonolized coins Fig 18-rank_TS3-diagonolized coins

In the above figures the data matrix has first been converted into the rank matrix, then gone into seriation. In Figure 17 the matrix is anti-diogonalized. Ranking as a robust method sounds like a good choice for such a scattered data set. In the first glance seriation might not look very successful yet the h-map seems to suggest an interesting notion: some of the metals in this study e.g. Pb and Cu might not be decisive in this ordering as they do not seem to correlate with the rest of the data. In Figure 18 the matrix is diogonalized as an example for the benefit of this approach when there are many zeros in the data: the less important entities are pushed to the corners of the matrix.

The two figures above are seriated by the R-package and are included for comparison. Figure 19 is seriated using PCA and figure 20 using BEA. The data in both cases were scaled before seriation.

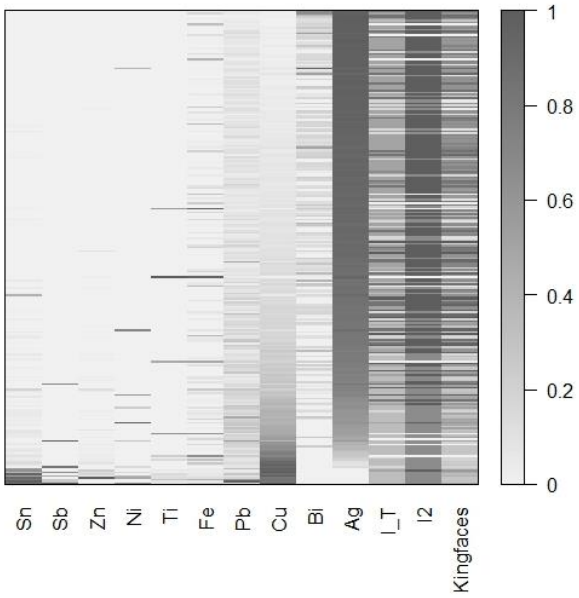


Fig 19-R-sc_PCA coins

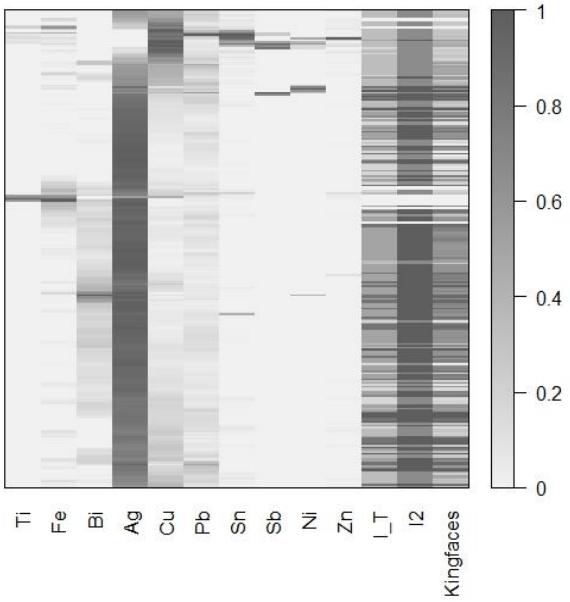


Fig 20-R-sc_BEA coins

It should be noted that our seriation procedure was unable to find hidden information between historical eras and King's faces on the coins, while such information was dug out and discussed in the paper where the original data came from (Christie, et al., 2014) using advanced chemometric methods. This shows that the uncertain and almost random metal content (contaminations) do not correlate simply with eras and historical acts like remelting and pressing the coins. Even the basic classifications based on not so reliable history are not heavily supported by the chemical data in the first look. That said, there eras which were seriated rather neatly.

Tamil Nadu Natural Radioactivity Data

The first characteristic that is seen in this data set is that some of the variables seem to change together which suggests redundancy. That is not very surprising as the radiological parameters are different methods to measure the same (or at least very similar) effects. Below are charted the scaled original data heat map and the scaled seriated data by the R-package with BEA Travelling Sales Person (BEAT). One could see that the seriated columns to the right look as if they were in a one-mode seriation since the data in these columns are heavily correlated.

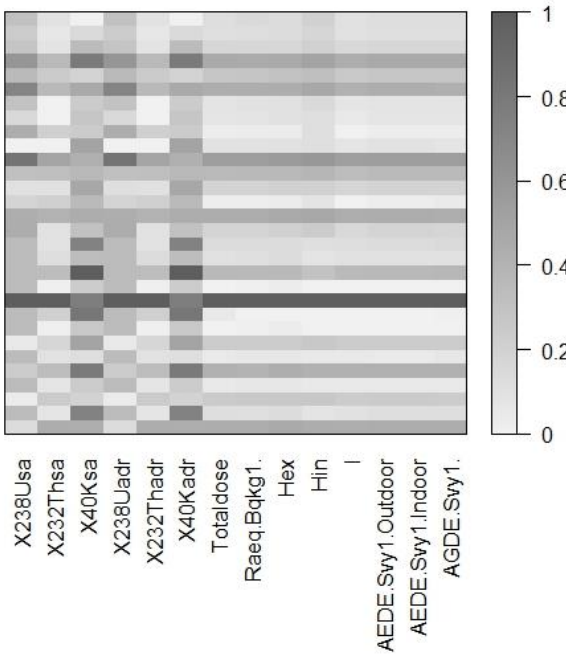


Fig 21-sc_original Tamil

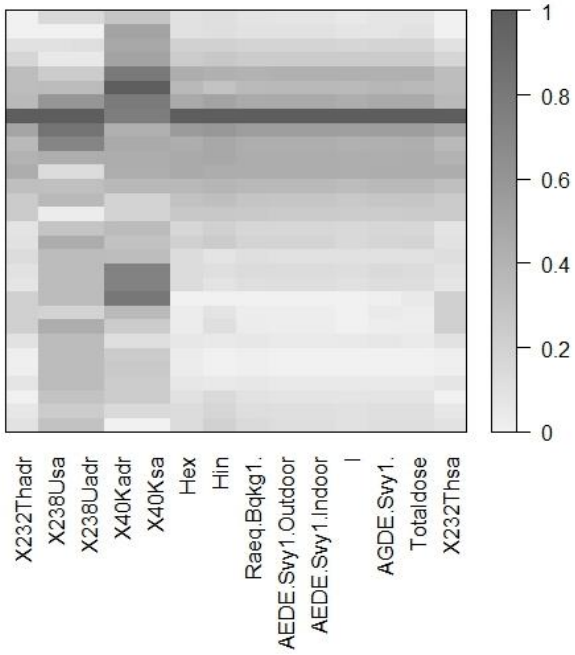


Fig 22-R-sc_BEAT Tamil

It is worth noting that scaling was very handy in this data set as the entries were not all in the same range. Below two figures are plotted both seriated by the TS method option 3. But Figure 23 was scaled before seriation and figure 24 was scaled after the seriation (before the plotting). The data on figure 23 was the best ordered in this data set, using the TS3 method on scaled data. We could create a negative uniform gradient for almost 10 of the 14 data columns.

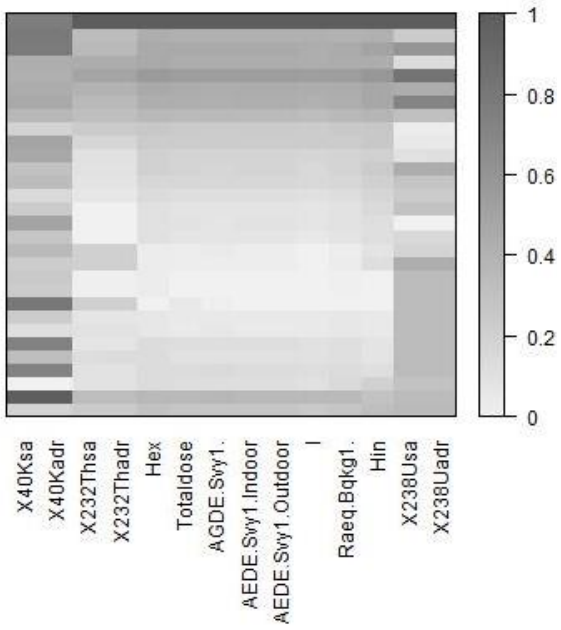


Fig 23-sc_TS3 Tamil

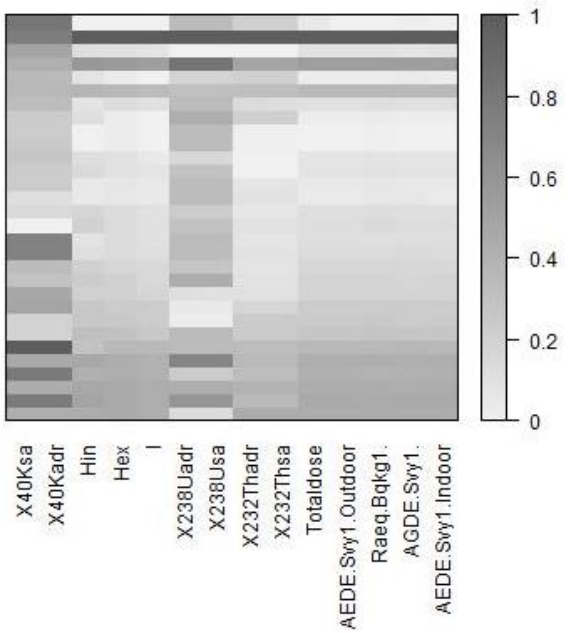


Fig 24-TS3 Tamil

The covariance matrix was not very useful in this seriation due to both the low number of columns and the high correlation between those few. Below the one-mode working matrix for the objects before and after seriation by the TS method can be found. The object distance matrix in figure 26 was anti-diogonalized which is drastically smoother than figure 25.

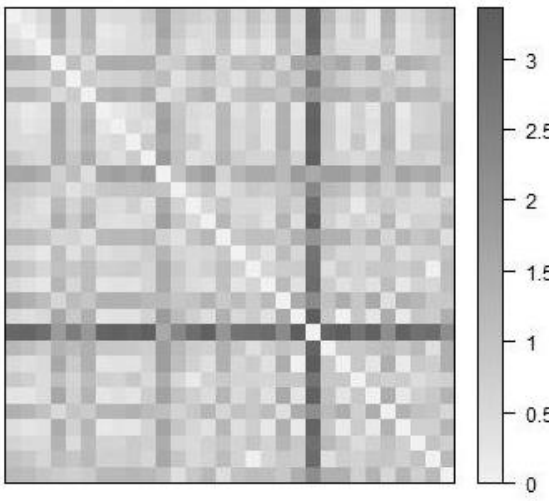


Fig 25- working object matrix TS1 original Tamil

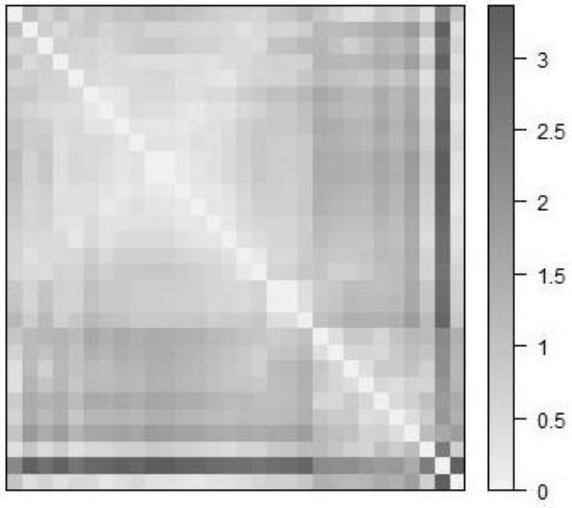


Fig 26-Working object matrix TS1 Tamil

Toxin Data Set

In this data set there are 4 pairs of columns as for each variable there is a predicted and an experimental data point. The seriation methods, however, do not seem to always class these pairs together. On the next page are plotted heat-maps of the scaled original matrix in figure 27 and the R-package PCA method which was one of the best results on this data set in figure 28.

Scaling before seriation in this data set does not make any tangible difference since the elements of the matrix are all in close range. For the sake of comparison two seriated heat-maps by the TS1 method are plotted below. Figure 29 was scaled before the seriation was carried out and figure 30 afterwards.

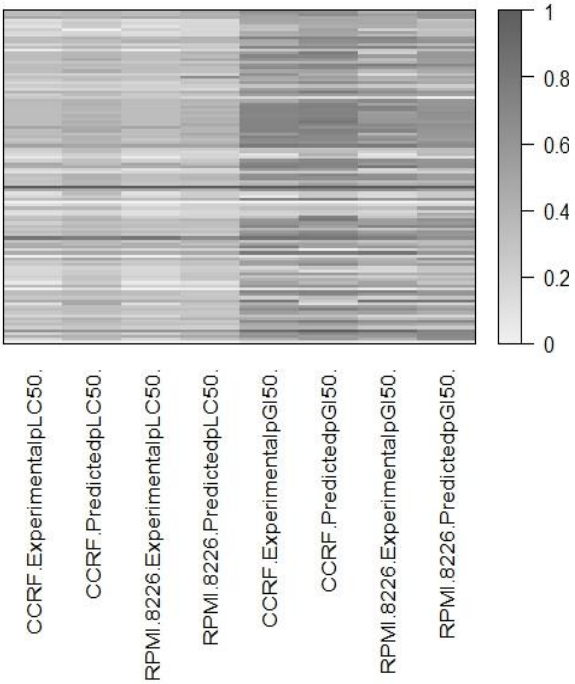


Fig 27-R-original Toxin

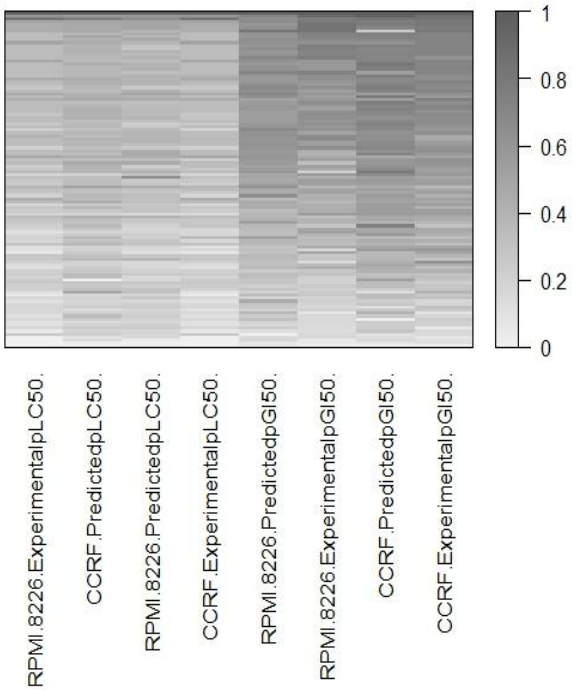


Fig 28-R-PCA Toxin

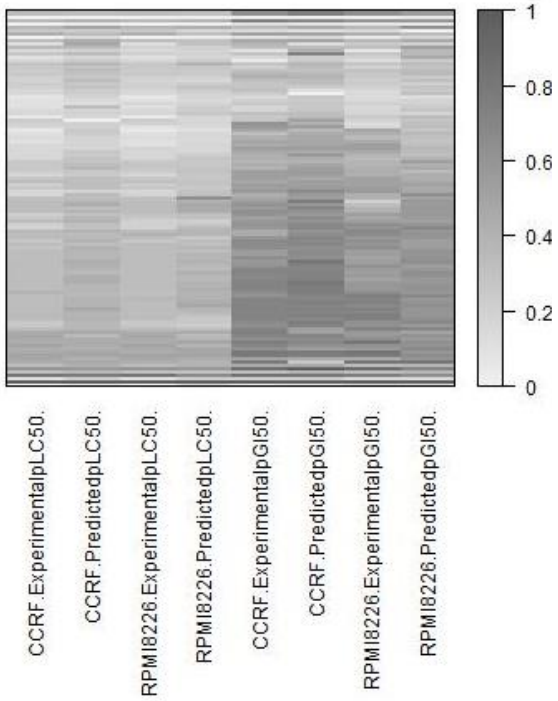


Fig 29-sc_TS1-Toxin

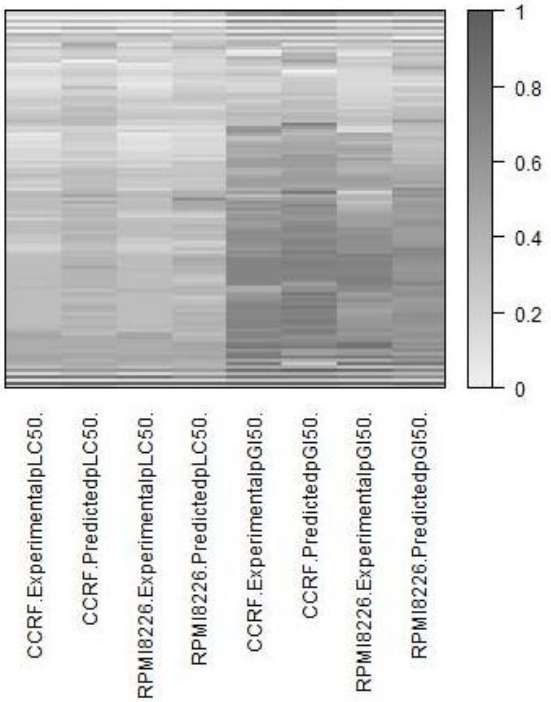


Fig 30-TS1-Toxin

The curious reader might be interested in the working graph of one of these similar seriations; so the unordered and ordered working graphs –object distance matrices- for objects of the raw data are plotted in figures 31 and 32 respectively.

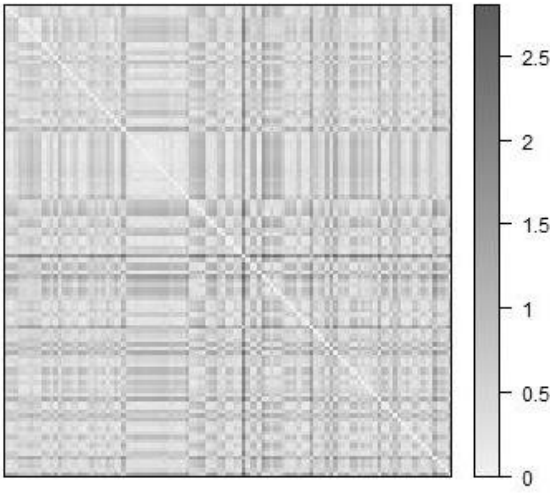


Fig 31-working object matrix-TS1 original Toxin

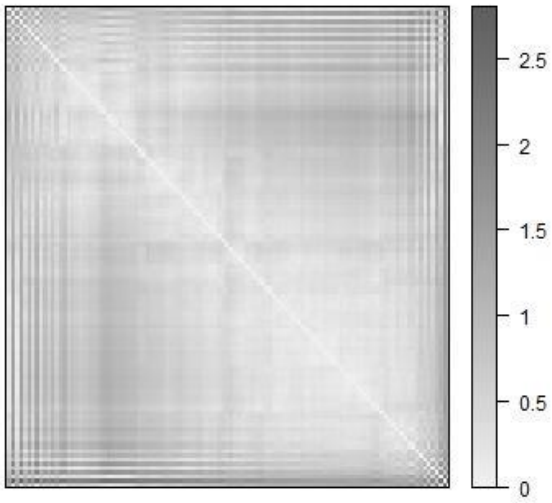


Fig 32- working object matrix-TS1 Toxin

Combustion Reactions

This data set was our attempt to work on a larger data set than before and try seriation on one-mode and two-mode data of this size. R-package proved quite problematic and not so effective with this data, there seemed to be some problems with one-mode analysis; so different options with TS method was explored.

First two kinds of two-way one-mode matrix was produced: the component matrix and the reactions matrix. The component matrix had elements with value 1 where the components crossing were both present in a reaction. Below the original and the one-mode seriated component one-mode matrices are plotted.

The reactions matrix had elements with value 1 where the crossing reactions had at least one component in common. Below are the original and seriated reaction plots.

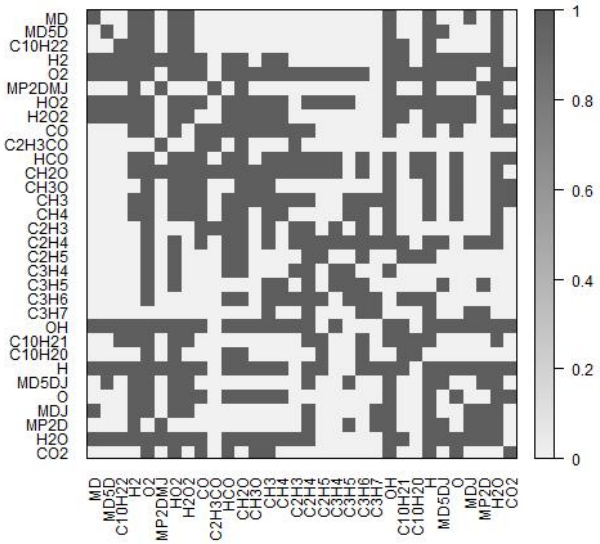


Fig 33-TS-original components matrix-combustion reactions

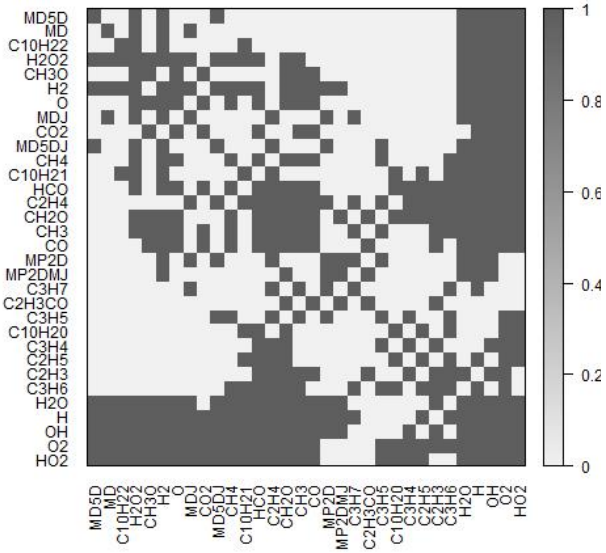


Fig 34- TS-seriated components matrix-combustion reactions

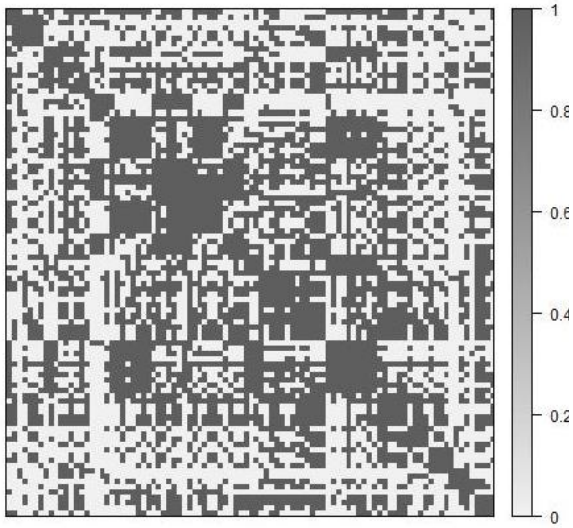


Fig 35- TS-original reaction matrix-combustion reactions

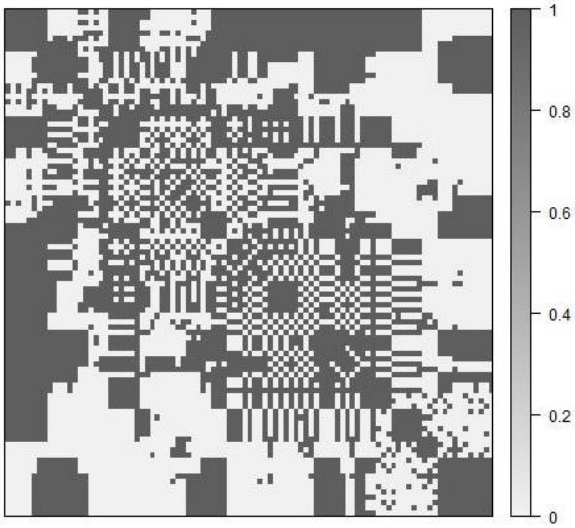


Fig 36- TS-seriated reaction matrix-combustion reactions

Now the results of these one-mode seriations projected back to the original reaction-component two-mode binary data matrix are shown down below. In Figure 37 the component seriation is at play and in figure 38 the reactions seriation.

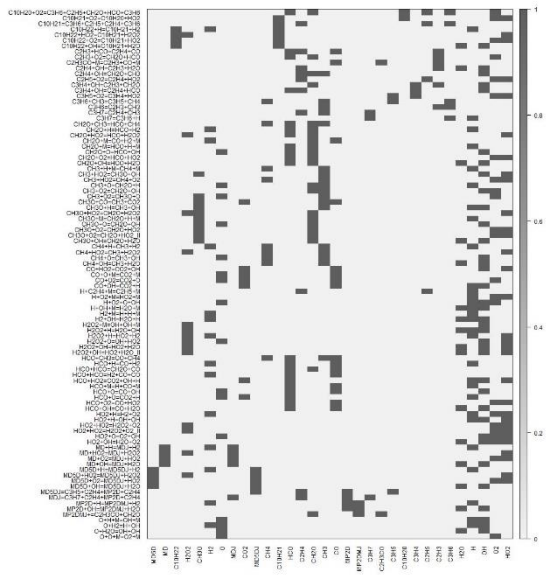


Fig 37-TS-components seriation combustion reactions

Fig 38-TS-reactions seriation combustion reactions

There are some clustering patterns visible yet it is not easy to pick up any global trend with one look.

In the hopes of getting better results, all the options 1, 2, 3 and 4 of the code with both diagonalization and anti-diagonalization was tried out on this data. The third option (local distance matrix on the reaction-component binary data) with diagonalization clearly resulted in the best visually clear trend. The original data and the result of this seriation are printed in the following two pages for comparison:

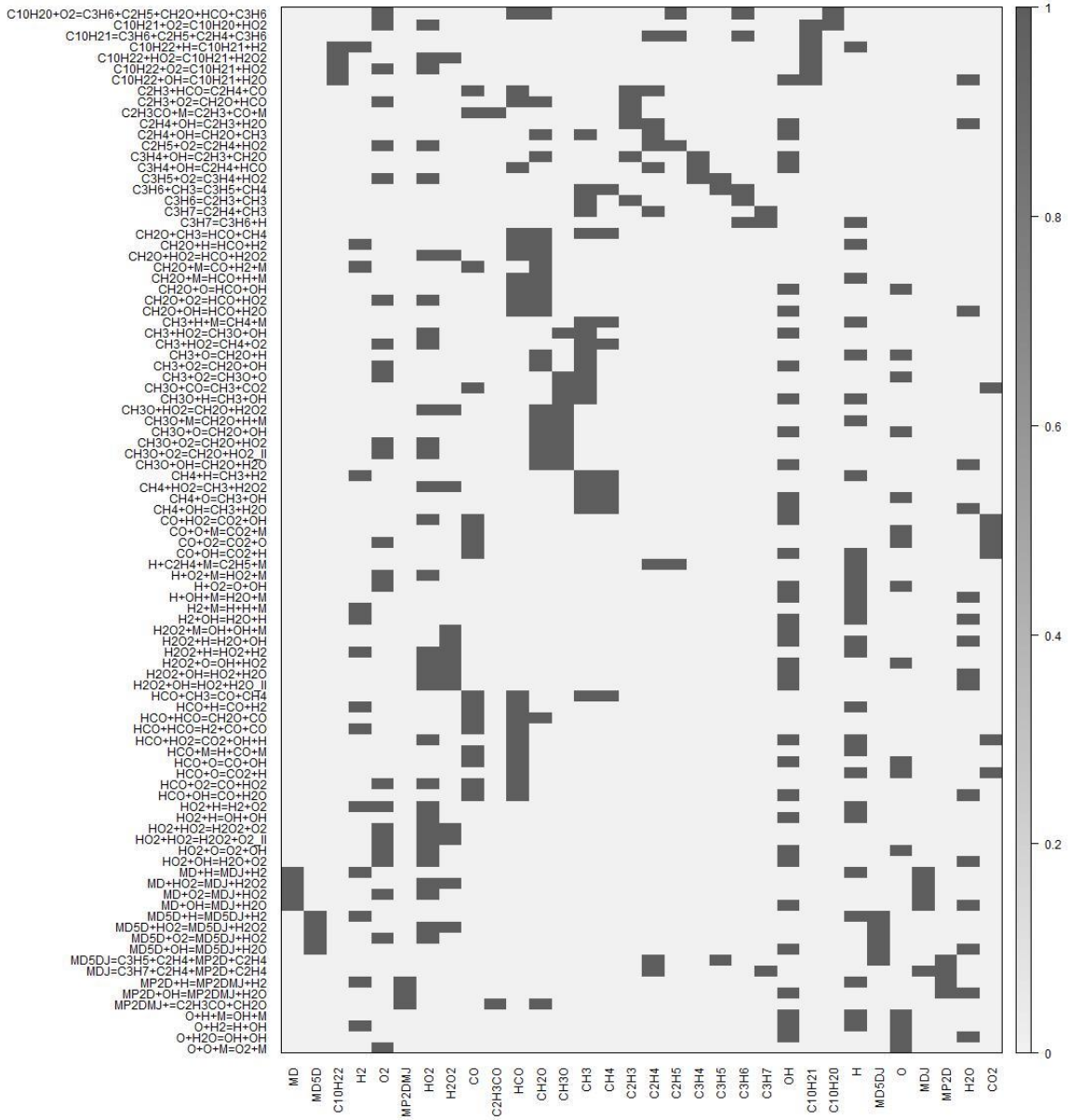


Fig 39-TS-original data combustion reactions

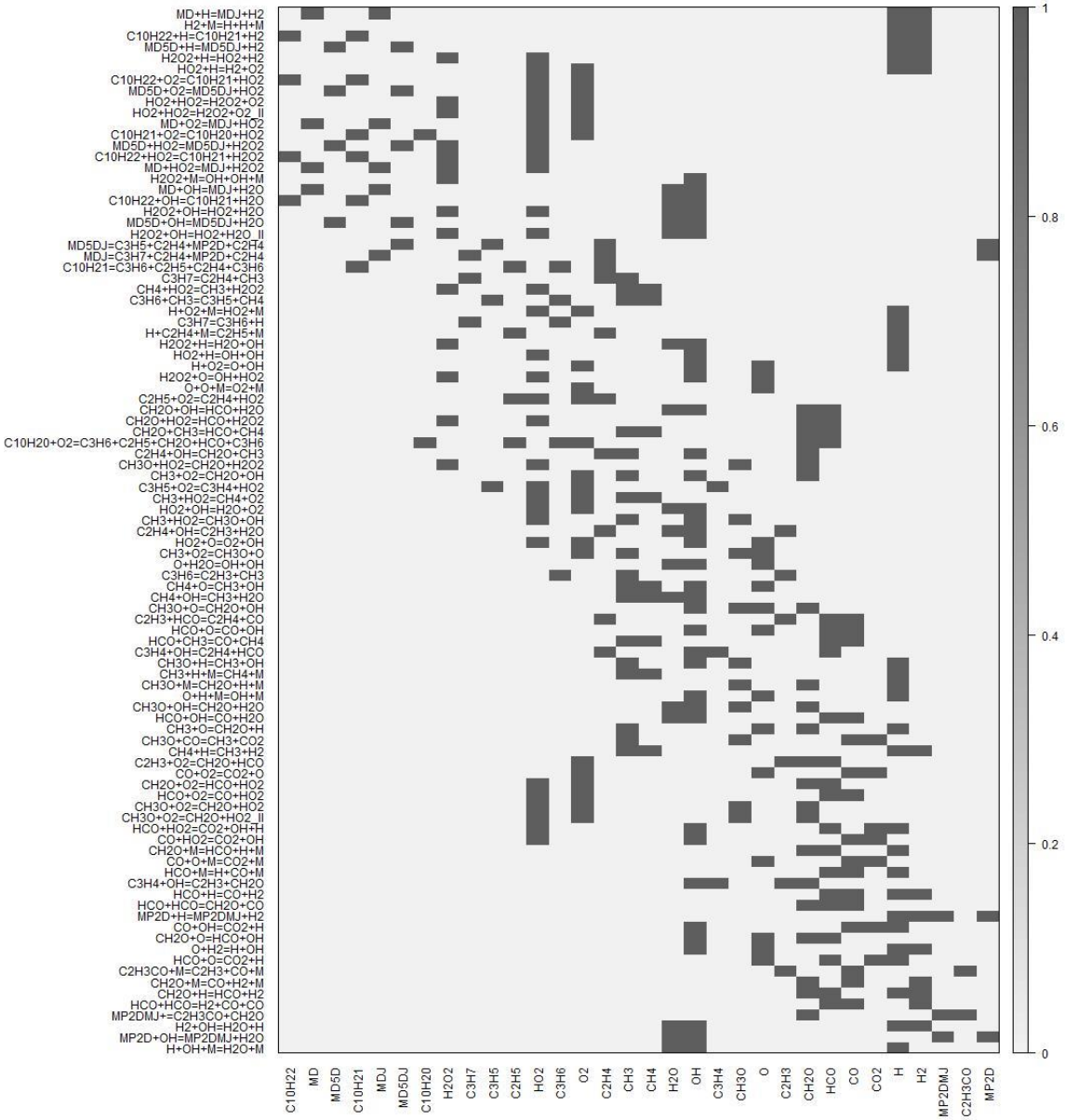


Fig 40-TS3-diagonalization- combustion reactions

Here you can see that the data was diogonalized. As mentioned earlier, diagonalization comes in handy when there are many zeros in our input matrix, for they are similar and will be put aside to the corners.

This graph looks well worth effort of computing to the author. One can trace the way the components are taking part in the reaction visually with the least difficulty possible. If I had to look for specific data points about certain compounds or reactions this would help me read it much faster rather than an arbitrarily ordered matrix. Beside the visual advantage, it is possible that a mathematic numeric code may run more efficiently if applied onto the seriated data rather than an unordered one.

There was also one more result that looks interesting to the author:

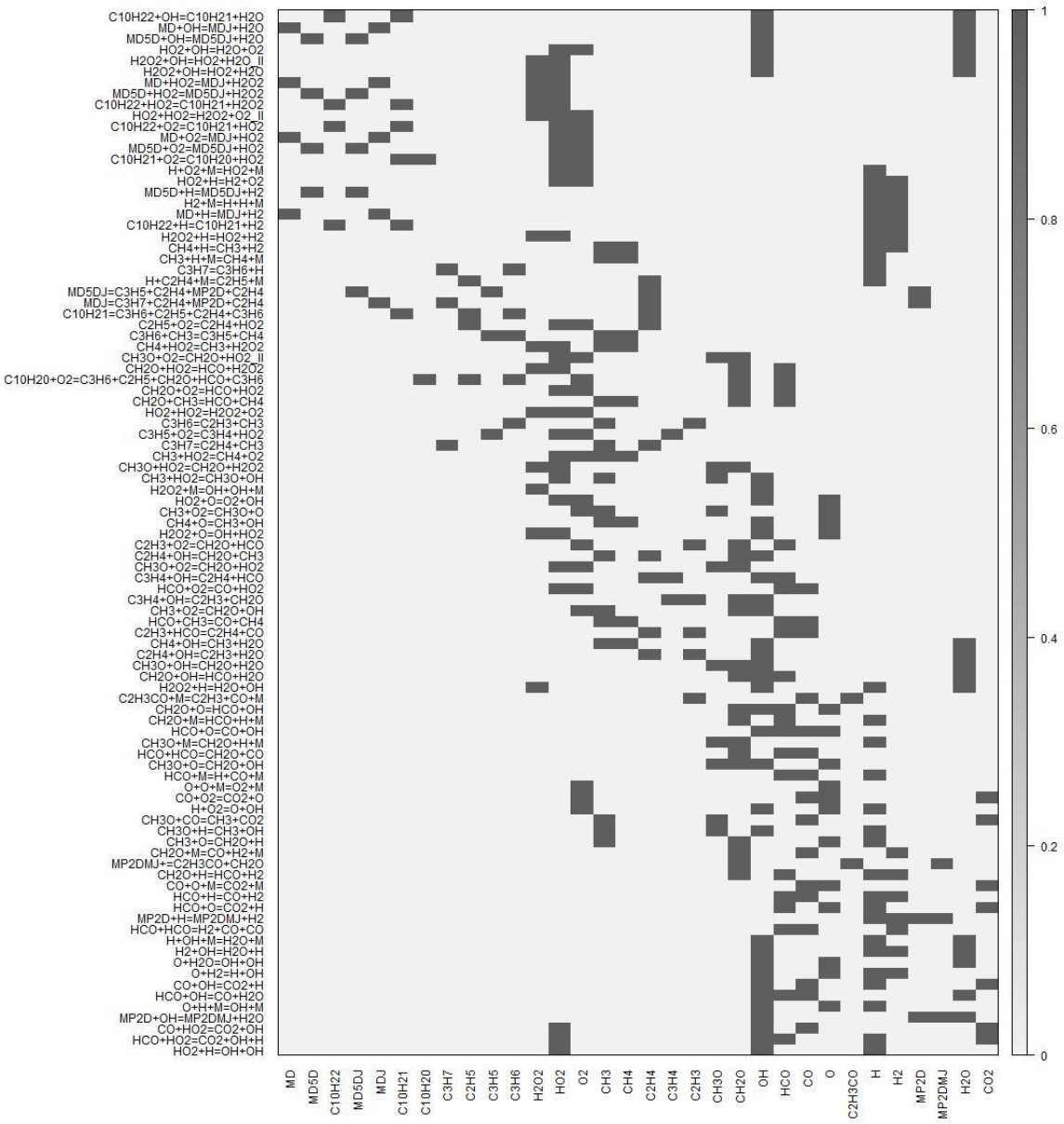


Fig 41-TS3-diagonalization-fix ends-combustion reactions

In this version the theoretically know beginning and ending components have been fixed at their position before the seriation and the permutation process has been applied only for the data in between. In the first look it might not look as well patterned as the graph in figure 40 yet it might help better imagine the real combustion reaction system and guess the direction at which the reaction moves.

Conclusion

Seriation is the act of permuting rows and columns of a data matrix in a meaningful way so that similar entities cluster up in the visual field. This non-redundant method while preserving all the data can hint in hidden local or global patterns without having to simplify the findings. There is no uniform method or priority of methods when it comes to seriation probably due to two different reasons. Firstly, different sorts and sizes are data respond better to different seriation methods and functions. Secondly, seriation historically suffers from being on the inspection of different fields, namely information visualization and data mining; the experts in these two disciplines tend not to care much for the other due to different perspectives. Yet, to develop better seriation methods and algorithms great care toward both sorting the data and presenting it is needed.

While being possible for *k*-way *k*-mode data arrays, in reality seriation is mostly used in ordering two-way one-mode and two-way two-mode data matrices. Data is usually presented in an arbitrary order in these matrices which can trick the eye and hide patterns and clusters within. Seriation attempts to liberate this information.

There have been many different methods designed for seriation of data each with their own loss or merit functions to be optimized. As one begins to try them out it quickly becomes clear that there cannot be a single method suitable for all data, as there are different types of patterns and correlations to be revealed in each matrix. A sufficient seriation package/software can be imagined which includes the option for using the most useful methods for different data sets, letting the user choose what they deem best for their need.

In this work we experimented with the R-package "Seriation" and the extended version of TS method designed for this investigation. The following might be the most remarkable notions we have extracted from this practice:

- When there is high correlation between the variables, using the Principle Component Analysis (PCA) is worth a try. See figure 5 and figure 28.
- Even in the case of pre-clustered matrices based on nominal data –as it were for the species in Iris data set– it is still not without benefit to try seriation within each cluster to reveal internal patterns. See figure 6.
- Pre-processing the data can be very useful in seriation. In fact, scaling the data proved vital in many cases: Firstly, quite a number of loss/merit functions' domain

does not include negative values, ergo scaling to the above zero range is necessary to begin with. Secondly, it happens more often than not the range of the data entries vary a lot from one another e.g. the altitude of a spot in mountains and its temperature. For the seriation function and also the heat-map function to relate these variables meaningfully, data needs to have been scaled.0; compare figures 7 and 8. In this work we scaled our data between 0 and 1. Thirdly, even in cases where pre-scaling was not necessary, it could be the case that the nominal data in numeric form exits in data; if one is interested in observing the nominal (numeric form) pattern in the heat-map, they need to be in the same scale as the rest of the data. Beside scaling, using the rank matrix instead of the raw data can be a robust pre-processing trick in seriation. See figures 17 and 18.

- Introducing diagonalization and anti-diagonalization, as was possible in the TS method, proved quite a useful feature which seems absent from the previous literature. Normally, one would like to focus the similar objects on the diagonal while clustering them. The beauty of this method is that by clustering similar objects and similar variables (based on local distance matrix) at the same time we can see our clustered object and the reason for their similarity both on the diagonal. There is however an exception when the reverse approach proves fruitful: in case of sparse matrices (where most of the elements are zeros) the many zeros provide artificial zero local distances. This way, zeros –which are similar- are pushed back to the corners and the interesting part of the data (non-zero elements) is again on the diagonal. A good example on it can be seen in Figures 40-41.
- In larger matrices where the data is presenting a stepwise phenomenon, as it was the case with our combustion reaction data, it could be a good idea to fix the beginning and final steps of the data on the opposite ends of the matrix (if they are identified) and then carry out the seriation process. This can lead to an insight on the direction of the progress in each step (Figure 41.).

We hope that the reader have found this work informative or at the very least has been convinced that seriation is a technique worth being available to many more users than it is now if we are to make the best of our data presentation.

Summary

Investigation of Seriation on Chemical Data.

SASAN AMARIAMIR, BSc student in Chemistry

Laboratory of Chemical Informatics, Institute of Chemistry, Eötvös University, Budapest

Place of defence: Physical Chemistry Department

Supervisor: Gergely Tóth, Assoc. Prof

Seriation is the practice of performing row and column permutations on data matrices in order to reveal clusters and hidden patterns within and between them. Seriation is a data mining problem with combinatorial nature aiming at enhancing visual clarity.

Seriation has been a known problem to tackle in the literature for over a century. Many different disciplines have used the method in order to find hidden patterns and clusters in the data such as archaeology, biology, ecology, sociometry etc (Liiv, 2010).

In chemistry, unlike those disciplines, there have been less than a handful of papers who have worked on this technique and its potential for chemical data. There is one paper written by my supervisor (Tóth, et al., 2009) which attempts to study the same problem of using matrix permutation to enhance data presentation without mentioning the name 'seriation' as the authors were not aware of its existence and literature background.

In this work we set out to find available seriation tools and methods, implement them on chemical data and report the result. With the help of my supervisor the old C code of their paper in 2009 was extended to include new functions for this investigation. I also picked up working with the R language and explored one of its available packages called "Seriation" (Hahsler, et al., 2008). This package offered different methods of seriation at one place.

We used a few different data bases to check the versatility of our methods. The data sets were Iris, ion contents in samples of wine, compositions of old Hungarian coins, natural radiation data of sand in Tamil Nadu, structure and toxicity relationship of molecules and elemental reactions mechanism for a biofuel combustion reaction.

In case of both of R-package and TS tools our investigation suggests that scaling the data was needed for meaningful seriation and plotting (we scaled between 0 and 1), using the rank matrix can be useful at times, PCA method is a good choice for data with high correlation, usually diagonalization of data helps perception except for the sparse matrixes where the opposite is true. We think that it is worthwhile for anyone working matrices to at least have a basic knowledge in seriation.

Summary References.

Hahsler, M., Hornik, K., & Buchta, C. (2008). Getting Things in Order: An Introduction to the R Package seriation. Journal of Statistical Software, 25(3).

Liiv, I. (2010). Seriation and Matrix Reordering Methods: An Historical Overview. Wiley Periodicals, 3(2). doi:10.1002/sam.10071

Tóth, G., & Szepesváry, P. (2009). A diagonal measure and a local distance matrix to display relations between objects and variables. Journal of Chemometrics. doi:10.1002/cem.1267

STATEMENT

Name: SASAN AMARIAMIR

ELTE Faculty of Science, Field of studies: Chemistry

Title of thesis: Investigation of Seriation on Chemical Data

I hereby declare, as the author of this thesis, that it is a product of my own and that it contains my own ideas. I used the standard rules for references and quotations consistently. I never used other people's ideas without proper reference.

2017.05.25, Budapest

_____

signature

References

Anderson E. The irises of the Gaspe Peninsula [Journal] // Bulletin of the American Iris Society. - 1935. - Vol. 59. - pp. 2-5.

Arabie P. and Hubert L. J. Clustering and Classification [Book]. - [s.l.] : World Scientific, 1996. - pp. 5-63.

Arthur Ebuka David [et al.] Quantitative structure-activity and toxicity relationship study of CCRF-CEM and RPMI 8402 cell line apoptosis with some anticancer compounds [Journal] // Chemical Data Collections. - Zaria : [s.n.], 2017. - Vols. 7-8. - pp. 8-50.

Bajusz D. Rácz A., Héberger Károly Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? [Journal] // Journal of Chemoinformatics. - 2015. - pp. 7-20.

Bartel H.G. seriation to describe some aspects of generalized evolution and its application in chemical informatics [journal] // systems analysis modelling simulation. [Journal] // SYSTEMS ANALYSIS MODELLING SIMULATION. - 1990. - 7 : Vol. 7. - pp. 557-565.

Bertin Jacques Graphics and graphic information-processing [Book] / trans. Scott W. J. Berg and P.. - Berlin : De Gruyter, 1981.

Bertin Jacques Semiology of graphics: diagrams, networks, maps [Book]. - Redlands, CA : ESRI Press, 2011.

Brusco Michael and Stahl Stephanie Branch-and-Bound Applications in Combinatorial Data Analysis [Book]. - [s.l.] : Springer Science+Business Media, Inc., 2005.

Chen C. H. Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices [Journal] // Statistica Sinica. - 2002. - Vol. 12. - pp. 7-29.

Christie OLAV H. J. [et al.] Classification and unscrambling a class-inside-class situation by object target rotation: Hungarian silver coins of the Árpád Dynasty,ad997-1301 [Journal] // Journal of Chemometrics. - 2014. - 4 : Vol. 28. - pp. 287-292.

Forsyth Elaine and Katz Leo A Matrix Approach to the Analysis of Sociometric Data: Preliminary Report [Journal] // Sociometry. - 1946. - 4 : Vol. 9. - pp. 340-.

Frank I. E. and Kowalski Bruce R. prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling [Journal] // Analytica Chimica Acta. - 1984. - Vol. 162. - pp. 241-251.

Gunnar Gruvaeus Howard Wainer Two Additions To Hierarchical Cluster Analysis [Journal] // British Journal of Mathematical and Statistical Psychology. - 1972. - 2 : Vol. 25. - pp. 200-206.

Gyseghem E. Van [et al.] Evaluation of chemometric techniques to select orthogonal chromatographic systems [Journal] // Journal of Pharmaceutical and Biomedical Analysis. - 2006. - 1 : Vol. 41. - pp. 141-151.

Hahsler Michael, Hornik Kurt and Buchta Christian Getting Things in Order: An Introduction to the R Package seriation [Journal] // Journal of Statistical Software. - 2008. - 3 : Vol. 25.

Hariprasath R. [et al.] Determination of natural radioactivity and radiological hazards of sediment sands in Tiruchirappalli district, Tamil Nadu, India [Journal] // Chemical Data Collections. - 2016. - Vol. 2. - pp. 1-9.

Hubert Lawrence Problems of seriation using a subject by item response matrix. [Journal] // Psychological Bulletin. - 1974. - 12 : Vol. 81. - pp. 976-983.

Hubert Lawrence Some Applications Of Graph Theory And Related Non-Metric Techniques To Problems Of Approximate Seriation: The Case Of Symmetric Proximity Measures [Journal] // British Journal of Mathematical and Statistical Psychology. - 1974. - 2 : Vol. 27. - pp. 133-153.

Hubert Lawrence, Arabie Phipps and Meulman Jacqueline Combinatorial data analysis: optimization by dynamic programming [Book]. - Philadelphia, PA : Society for Industrial and Applied Mathematics, 2001.

Ihm Peter A Contribution to the History of Seriation in Archaeology [Journal] // Studies in Classification, Data Analysis, and Knowledge Organization Classification — the Ubiquitous Challenge. - 2005. - pp. 307-316.

Juhász Gergely Reduction of a biodiesel combustion reaction mechanism BSc thesis work [Book]. - Budapest : Eötvös Loránd University, Institute of Chemistry, Department of Physical Chemistry, 2015.

Kendall D. G., Hodson F. R. and Tautu Petre Mathematics in the archaeological and historical sciences [Book]. - Edinburgh : Edinburgh University Press, 1971.

Liiv Innar Seriation and Matrix Reordering Methods: An Historical Overview [Article] // Wiley Periodicals. - Tallinn : [s.n.], 2010. - 2 : Vol. 3.

Mccormick William T., Schwitzer Paul J. and White W. Thomas Problem Decomposition and Data Reorganization by a Clustering Technique [Journal] // Operations Research. - 1972. - 5 : Vol. 20. - pp. 993-1009.

Miklós István, Somodi Imelda and Podani János rearrangement of ecological data matrices via markov chain monte carlo simulation [Journal] // Ecological Society of America. - 2005. - 12 : Vol. 86. - pp. 3398-3410.

Moreno J. L. Who shall survive? A new approach to the problem of human interrelations [Book]. - [s.l.] : Nervous and mental disease Publishing Co., 1934.

Niermann Stefan Optimizing the Ordering of Tables With Evolutionary Computation [Journal] // The American Statistician. - 2005. - 1 : Vol. 59. - pp. 41-46.

Petrie W. M. Flinders Sequences in Prehistoric Remains [Journal] // The Journal of the Anthropological Institute of Great Britain and Ireland. - 1899. - 3/4 : Vol. 29. - p. 295.

Rácz Anita [et al.] Classification of Hungarian medieval silver coins using x-ray fluorescent spectroscopy and multivariate data analysis [Journal] // Heritage Science. - 2013. - 1 : Vol. 1.

Robinson W. S. A Method for Chronologically Ordering Archaeological Deposits [Book]. - [s.l.] : American Antiquity, 1951. - Vol. 16 : pp. 293-301.

Shneiderman Ben Inventing Discovery Tools: Combining Information Visualization with Data Mining [Journal] // Information Visualization. - 2002. - 1 : Vol. 1. - pp. 5-12.

Sokal Robert R. and Sneath Peter H.A. Principles of numerical taxonomy [Book]. - San Francisco : Freeman, 1963.

Szabó Attila Ordering of Matrices Containing Sequences by Chemometric Methods, Master thesis work. [Book]. - Budapest : Eötvös Loránd University, Institute of Chemistry, Department of Physical Chemistry, 2010.

Tóth Gergely and Kakatics Máté Homepage of Gergely Tóth [Online]. - 2017. - http://www.chem.elte.hu/departments/elmkem/toth/.

Tóth Gergely and Szepesváry Pál A diagonal measure and a local distance matrix to display relations between objects and variables [Journal] // Journal of Chemometrics. - 2009.

Tucker L. R. The extension of factor analysis to three-dimensional matrices [Journal] // Contributions to mathematical psychology. - New York : Holt, Rinehart and Winston, 1964. - pp. 110-127.

Wesbrook C. K. [et al.] Detailed chemical kinetic reaction mechanisms for soy and rapeseed biodiesel fuels [Journal] // Combustion and Flame. - 2011. - 4 : Vol. 158. - pp. 742-755.