Solvents Classification using a Multivariate Approach: Correlation and Principal Component Data Analysis.

1. INTRODUCTION

Already in 1985, in a paper published in Acta Chemica Scandinavica[1], Carlson and coworkers, applied Principal Component Analysis[2] to eighty-two (82) different solvents each characterized by eight common physicochemical descriptors (or *variables*).
Almost concurrently, Chastrette and coworkers published another paper [3] also dealing with solvents classification using multivariate analysis. Even in this case the Authors used eight descriptors, but of a different kind with respect to those chosen by Carlson. However, in both cases, the Authors' attention was focused on the practical application, to solvent selection, of the results obtained from data analysis. Unlike the previous papers, this post concerns the details of the data analysis process that is shown here using R/RStudio for both explorative multivariate data analysis and visualization.

2. EXPERIMENTAL SECTION

For the purpose of this study, it have been considered the solvents listed in Carlson's paper[1] (see Table 1), each characterized by following eight descriptors (or *active variables*): *melting point* (mp), *boiling point* (bp), *dielectric constant* (dc), *dipole moment* (dm), *refractive index* (ri), E_T (ET), *density* (d) and *log P* (logP).
The abbreviations in brackets will be used, from now on, to refer to descriptors in graphs, *etc*.

The dataset used is listed here below.

Table 1

solvent	mp	bp	dc	dm	ri	ET	d	logP
Water	0.0	100.0	78.39	6.07	1.333	63.1	0.9982	-1.38
Formamide	2.5	210.5	111.0	11.24	1.4475	56.6	1.1134	-1.51
1,2-Ethanediol	13	197.3	37.7	7.61	1.4318	56.3	NA	-1.93
Methanol	-97.7	64.7	32.2	5.67	1.3284	55.5	0.7914	-0.77
N-Methylformamide	-3.8	180.5	182.4	12.88	1.4319	54.1	1.01	NA
Diethylene glycol	-6.5	244.8	31.69	7.71	1.4475	53.8	1.109	NA
Triethylene glycol	-4.3	288.0	23.69	9.97	1.4561	53.5	NA	NA

Table 1 (*cont.*)

solvent	mp	bp	dc	dm	ri	ET	d	logP
2-Methoxyethanol	-85.1	124.6	16.93	6.81	1.4021	52.3	0.065	0
N-Methylacetamide	30.6	206.7	191.3	14.65	1.4286	52.0	0.957	-1.05
Ethanol	-114.1	78.3	24.55	5.77	1.3614	51.9	0.789	-0.31
2-Aminoethanol	10.5	171.0	37.72	7.57	1.4539	51.8	1.018	-1.31
Acetic acid	16.7	117.9	6.15	5.60	1.3719	51.2	1.0492	-0.17
Benzyl alcohol	-15.3	205.5	13.1	5.54	1.5404	50.8	1.042	1.10
1-Propanol	-126.2	97.2	20.33	5.54	1.3856	50.7	0.804	0.25
1-Butanol	-88.6	117.7	17.51	5.84	1.3993	50.2	0.8098	0.89
2-Methyl-1-propanol	-108	107.7	17.93	5.97	1.3959	49.0	0.794	0.83
2-Propanol	-88.0	82.3	19.92	5.54	1.3772	48.6	0.786	0.05
2-Butanol	-114.7	99.6	16.45	5.54	1.3972	47.1	0.8080	0.61
3-Methyl-1-butanol	-117.2	130.5	14.7	6.07	1.4071	47.0	0.8092	1.16
Cyclohexanol	25.2	161.1	15.0	6.20	1.4548	46.9	0.962	1.23
4-Methyl-1,3-dioxol-2-one	-48.8	241.7	65.1	16.7	1.4209	46.6	1.204	NA
2-Pentanol	NA	119.0	13.82	5.54	1.4064	46.5	0.810	NA
Nitromethane	-28.6	101.2	35.87	11.88	1.3812	46.3	1.137	-0.33
Acetonitrile	-43.8	81.6	37.5	11.48	1.3441	46.0	0.7857	-0.34
3-Pentanol	-75	115.3	13.02	5.47	1.4103	45.7	0.8201	1.21
Dimethylsulfoxide	18.5	189.0	46.68	13.0	1.4783	45.0	1.101	-1.35
Aniline	-5.98	184.4	6.89	5.04	1.4863	44.3	1.0217	0.90
Sulfolane	28.5	287.3	43.3	16.05	1.4920	44.0	1.262	NA
Acetic anhydride	-73.1	140.0	20.7	9.41	1.3904	43.9	1.0820	NA
2-Methyl-2-propanol	25.8	82.4	12.47	5.54	1.3877	43.9	0.789	0.37
N,N-Dimethylformamide	-61	152.3	37.0	12.88	1.4269	43.8	0.925	-1.01
N,N-Dimethylacetamide	-20	166.1	37.78	12.41	1.4384	43.7	0.937	-0.77
Propionitrile	-92.8	97.4	27.2	11.91	1.3658	43.7	0.782	0.16
1-Methyl-2-pyrrolidone	-24.4	204	32.0	13.64	1.4700	42.2	1.026	NA
Acetone	-94.7	56.3	20.70	9.54	1.3587	42.2	0.790	-0.24
Nitrobenzene	5.8	210.8	34.82	13.44	1.5500	42.0	1.204	1.85
Benzonitrile	-12.8	191.1	25.20	13.51	1.5282	42.0	1.010	1.56
1,1-Diaminoethane	11.3	117.3	12.9	6.34	1.4568	42.0	0.899	NA
1,2-Dichloroethane	-35.7	83.5	10.36	6.20	1.4448	41.9	1.235	1.48
2-Methyl-2-butanol	-8.8	102.0	5.82	5.7	1.4049	41.9	0.806	1.36
2-Butanone	-86.7	79.6	18.51	9.21	1.3788	41.3	0.835	0.29
Acetophenone	19.6	202.0	17.39	9.87	1.5342	41.3	1.0281	1.58
Dichloromethane	-95.1	39.8	8.93	5.17	1.4242	41.1	1.33	1.25
1,1,2,2-Tetramethyl-urea	-1.2	175.2	23.45	11.58	1.4493	41.0	0.969	NA

Table 1 (*cont.*)

solvent	mp	bp	dc	dm	ri	ET	d	logP
Hexamethylphosphoric triamide	7.2	235	29.6	18.48	1.4584	40.9	1.024	0.28
Cyclohexanone	-32.1	155.7	18.3	10.04	1.451	40.8	0.9478	0.81
Pyridine	-41.6	115.3	12.4	7.91	1.5102	40.2	0.982	0.65
Methyl acetate	-98.1	56.3	6.68	5.37	1.3614	40.0	0.933	0.18
4-Methyl-2-pentanone	-84.0	116.5	13.11	NA	1.3957	39.4	0.7978	NA
1,1-Dichloroethane	-97.0	57.3	10.0	6.61	1.4164	39.4	1.176	1.79
Quinoline	-14.9	237.1	9.00	7.27	1.6273	39.4	1.093	2.03
3-Pentanone	-38.9	102	17.00	9.41	1.3923	39.3	0.8138	1.91
Chloroform	-63.6	61.2	4.81	3.84	1.4429	39.1	1.48	1.92
Triethylene glycol dimethyl ether	NA	222	7.5	NA	1.4233	38.9	NA	NA
Diethylene glycol dimethyl ether	NA	159.8	NA	6.57	1.4097	38.6	NA	NA
Dimethoxyethane	-58	85	7.20	5.70	1.3796	38.2	0.8629	NA
1,2-Dichlorobenzene	-17.0	180.5	9.93	7.57	1.5515	38.1	1.305	3.38
Ethyl acetate	-84.0	77.1	6.02	6.27	1.3724	38.1	0.900	0.73
Fluorobenzene	-42.2	84.7	5.42	4.90	1.4684	38.1	1.023	2.27
Iodobenzene	-31.3	188.3	4.63	4.64	1.6200	37.9	1.831	3.25
Chlorobenzene	-45.6	131.7	5.62	5.15	1.5248	37.5	1.106	2.84
Bromobenzene	-30.8	155.9	5.40	5.17	1.5571	37.5	1.495	2.99
Tetrahydrofuran	-108.5	66	7.58	5.84	1.4072	37.4	0.889	0.46
Anisole	-37.5	153.8	4.33	4.17	1.5170	37.2	0.996	2.11
Ethyl-phenyl-ether	-29.5	170.0	4.22	4.54	1.5074	36.4	0.967	2.51
1,1,1-Trichloroethane	-30.4	74.0	7.53	5.24	1.4379	36.2	1.339	2.49
1,4-Dioxane	11.8	101.3	2.21	1.50	1.4224	36.0	1.034	-0.27
Trichloroethylene	-86.4	87.2	3.42	2.7	1.4746	35.9	1.464	2.29
Piperidine	-10.5	106.7	5.8	3.97	1.4525	35.5	0.861	0.85
Diphenyl ether	26.9	258.3	3.69	3.87	1.4763	35.3	1.075	4.21
Diethyl ether	-116.3	34.6	4.34	4.34	1.3524	34.6	0.714	0.77
Benzene	5.5	80.1	2.28	0.0	1.5011	34.5	0.8787	2.15
Diisopropyl ether	-85.5	68.3	3.88	4.20	1.3681	34.0	0.7251	2.03
Toluene	-95.0	110.6	2.38	1.43	1.4969	33.9	0.867	2.73
Di-n-butyl ether	-95.2	142.2	3.08	3.94	1.3992	33.4	0.7689	NA
Triethylamine	-114.7	89.5	2.42	2.90	1.4014	33.3	0.7275	1.44
1,3,5-Trimethylbenzene	-44.7	164.7	2.28	0.0	1.4994	33.1	0.865	3.42
Carbon disulfide	-111.6	46.2	2.64	0.0	1.628	32.6	1.263	NA
Carbon tetrachloride	-23.0	76.8	2.24	0.0	1.4574	32.5	1.59	2.83
Tetrachloroethylene	-22.4	121.2	2.3	0.0	1.5057	31.9	1.623	2.60
Cyclohexane	6.5	80.7	2.02	0.0	1.4262	31.2	0.778	3.44
n-Hexane	-95.3	67.8	1.88	0.0	1.3749	30.9	0.66	NA

As for some solvents (18) among the 82 listed in Table 1, one or more quality descriptors are missing, these solvents have not been considered and removed from the dataset using the function *na.omit ()* of R *stats* package. The excluded solvents were:

Excluded Solvents because of Missing Data				
1,2-Ethandiol	N-Methylformamide	Diethylene glycol	Triethylene glycol	4-Methyl-1,3-dioxol-2-one
2-Pentanol	Sulfolane	Acetic Anhydride	1-Methyl-2-pyrrolidone	1,1-Diaminoethane
1,1,2,2,-Tetramethyl-urea	4-Methyl-2-pentanone	Triethylene glycol dimethyl ether	Diethylene glycol dimethyl ether	Dimethoxyethane
Di-n-butyl ether	Carbon Disulfide	n-Hexane		

As each solvent in the diagrams later on displayed is identified using a number, Table 2 here below allows to quickly trace from that number to the corresponding chemical entity:

Table 2

Solvents actually considered					
No.	**solvent**	**No.**	**solvent**	**No.**	**solvent**
1	Water	23	N,N-Dimethylformamide	45	Iodobenzene
2	Formamide	24	N,N-Dimethylacetamide	46	Chlorobenzene
3	Methanol	25	Propionitrile	47	Bromobenzene
4	2-Methoxyethanol	26	Acetone	48	Tetrahydrofuran
5	N-Methylacetamide	27	Nitrobenzene	49	Anisole
6	Ethanol	28	Benzonitrile	50	Ethyl-phenyl-ether
7	2-Aminoethanol	29	1,2-Dichloroethane	51	1,1,1-Trichloroethane
8	Acetic acid	30	2-Methyl-2-butanol	52	1,4-Dioxane
9	Benzyl alcohol	31	2-Butanone	53	Trichloroethylene
10	1-Propanol	32	Acetophenone	54	Piperidine
11	1-Butanol	33	Dichloromethane	55	Diphenyl ether
12	2-Methyl-1-propanol	34	Hexamethylphosphoric triamide	56	Diethyl ether
13	2-Propanol	35	Cyclohexanone	57	Benzene
14	2-Butanol	36	Pyridine	58	Diisopropyl ether
15	3-Methyl-1-butanol	37	Methyl acetate	59	Toluene
16	Cyclohexanol	38	1,1-Dichloroethane	60	Triethylamine
17	Nitromethane	39	Quinoline	61	1,3,5-Trimethylbenzene
18	Acetonitrile	40	3-Pentanone	62	Carbon tetrachloride
19	3-Pentanol	41	Chloroform	63	Tetrachloroethylene
20	Dimethylsulfoxide	42	1,2-Dichlorobenzene	64	Cyclohexane
21	Aniline	43	Ethyl acetate		
22	2-Methyl-2-propanol	44	Fluorobenzene		

The decision of excluding the solvents incompletely described was determined by the difficulty of finding missing data from the same sources used by Carlson and by the fact that other methods (*e.g.*, replacing missing values with the averages of the corresponding values for similar substances, adoption of estimation algorithms, *etc.*) did not seem adequate.

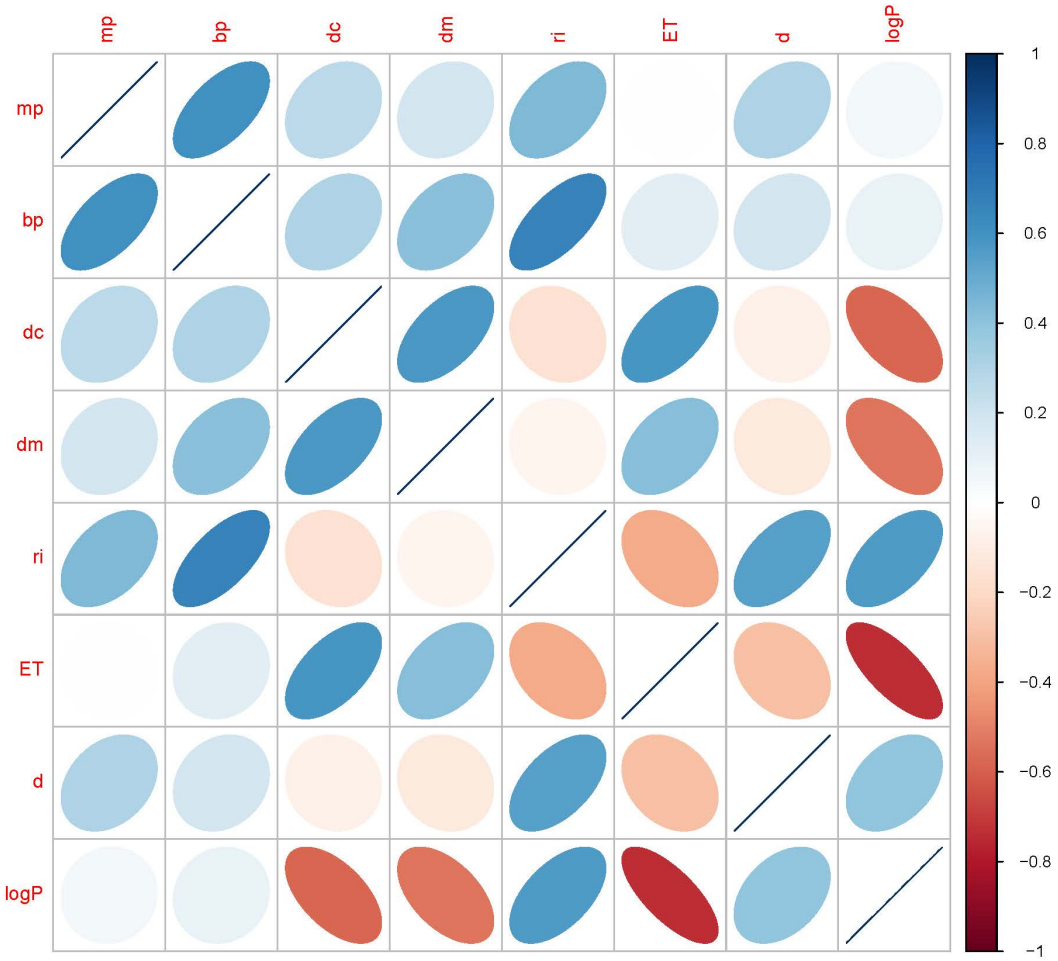
Both, data analysis and visualization, have been performed using RStudio version 1.1.414 and R version 3.5.0 (The R Foundation for Statistical Computing). In particular, the following specific R packages have been used:

- *tidyverse* (H. Wickham, RStudio Inc., Boston, USA)[4]
- *FactoMineR* (F. Husson, J. Josse, S. Le, J. Mazet, Agrocampus Ouest, Rennes University, France) [5, 6]
- *factoextra* (A. Kassambara, F. Mundt, HalioDx, Marseille, France and Pädagogische Hochschule, Karlsruhe, Germany) [7, 8, 9]
- *corrplot* (T. Wei, V. Simko, Fujian Agriculture and Forestry University (China) and FZI Forschungszentrum Informatik, Karlsruhe, Germany) [10, 11]
- *GGally* (B. Schloerke, J. Crowley, D. Cook, H. Hofmann, H. Wickham, Iowa State and Rice Universities, USA) [12]
- *PerformanceAnalytics* (B.G. Peterson, P. Carl, University of Washington, USA) [13]
- *scatterplot3D* (U. Ligges, TU Dortmund, Germany) [14]
- *cluster* (M. Mächler, ETH Zürich, Switzerland) [15]

3. RESULTS AND DISCUSSION

In Figure 1 is displayed the correlation plot obtained on the initial data autoscaled.

Figure 1



This diagram has been obtained using the function *corrplot()* of R *corrplot* package and, as already explained in the previous post, its elements are geometrical shapes that become more and more elliptical and intensely colored as the two initial variables gets strongly related each other. On the main diagonal, where the correlation is maximum (in fact the correlation of each element with itself is equal to one) the ellipses become a segment. Ellipses are right-oriented and blue colored if the two variables are positively correlated each other, while they are left oriented and red/brown colored if negatively correlated.

Figure 1 shows immediately a few strong correlations such as those between:

- boiling point and refractive index – positive
- E_T and log P - negative

and other, weaker, such as those between:

- boiling point and melting point– positive
- dielectric constant and dipole moment – positive
- dielectric constant and E_T – positive
- refractive index and density – positive
- refractive index and log P – positive
- dielectric constant and log P – negative
- dipole moment and log P - negative

This classification in “stronger” and “weaker” correlations between variable pairs is related to the correlation coefficient values in the *correlation matrix*, reported here below. The r_{ij} element of this matrix is the sample correlation coefficient (or Pearson correlation coefficient) between the *i*th and the *j*th variables and it is defined as:

$$r_{ij} = {}^{S_{ij}}\Big/_{S_i S_j}$$

Where S_{ij} is the sample covariance, standardized by the standard deviations S_i and S_j of the two variables.

The *correlation matrix* is symmetric because the correlation between S_i and S_j is the same as the correlation between S_j and S_i.

Correlation matrix

	mp	bp	dc	dm	ri	ET	d	logP
mp	1.00							
bp	0.58	1.00						
dc	0.27	0.30	1.00					
dm	0.18	0.37	0.57	1.00				
ri	0.45	0.66	-0.16	-0.06	1.00			
ET	0.00	0.13	0.59	0.43	-0.38	1.00		
d	0.31	0.20	-0.07	-0.12	0.55	-0.30	1.00	
logP	0.05	0.07	-0.58	-0.54	0.57	-0.73	0.39	1.00

As "strongly correlated" variables are generally considered those for which $r_{ij} \geq 0.7$ (absolute value) while as "moderately correlated variables" are those for which $0.3 \leq r_{ij} < 0.7$ (absolute value). As "weakly correlated" variables are generally considered those for which $r_{ij} < 0.3$ (absolute value).

It is interesting to observe that most of the correlations above highlighted belong to the fundamentals of Chemistry.

The correlation matrix has been calculated using the function *cor()* of the R *stats* package.
The above information can be visualized in a more easy and comprehensive manner using the scatterplot matrices shown in Figures 2 and 3 that provide pairwise comparison of multivariate data.
The *scatterplot matrix* provides a graphical display of the:

- anomalous values (*outliers*) both one-dimensional (a point far from the rest for that variable) and two-dimensional (a point far from the regression line that correlates two variables),
- relationship existing between each variables pair, showing if it looks linear or not,
- groups of individuals characterized by similar values of variables pairs and that look like neighboring points in the graphs.

Figure 2

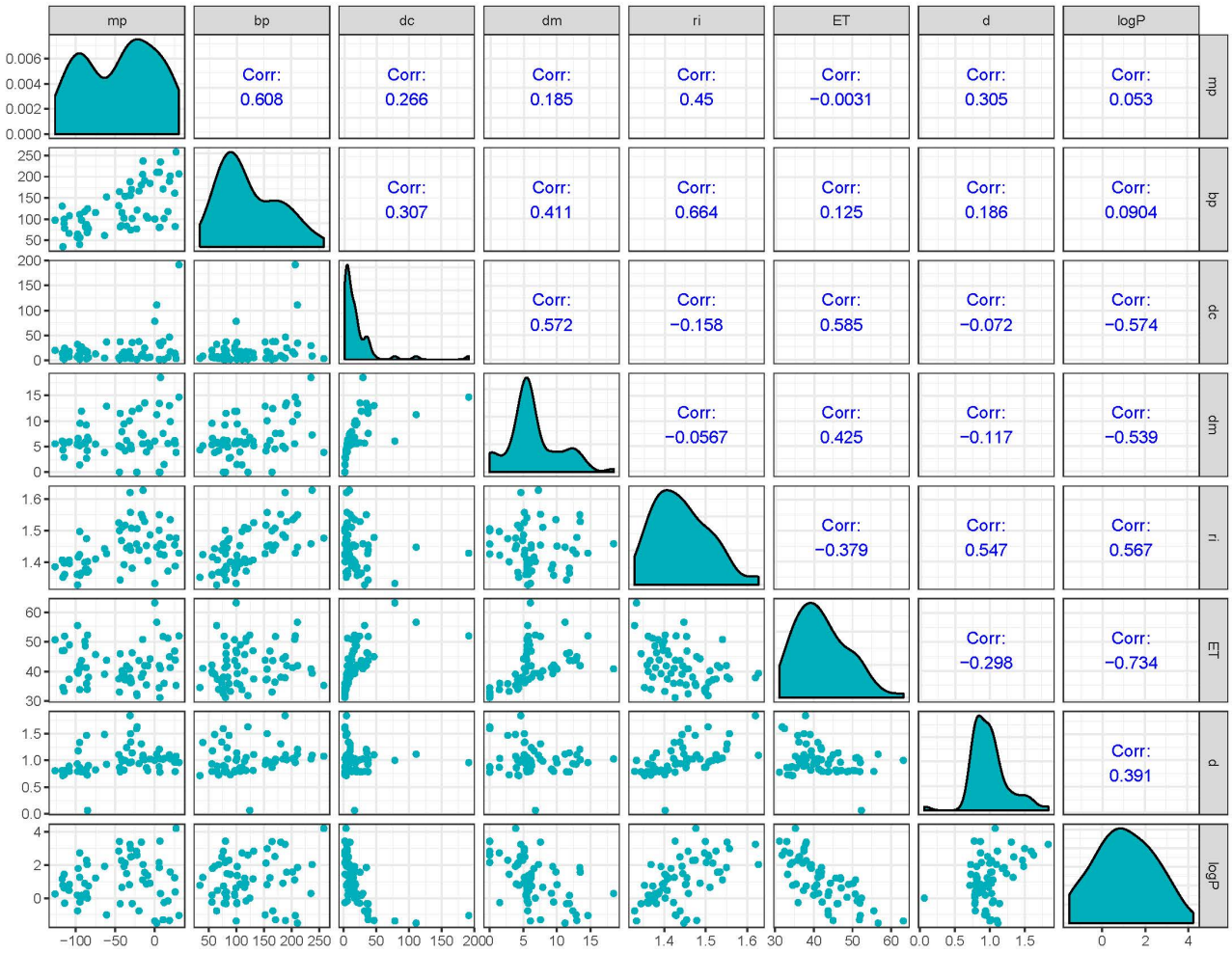
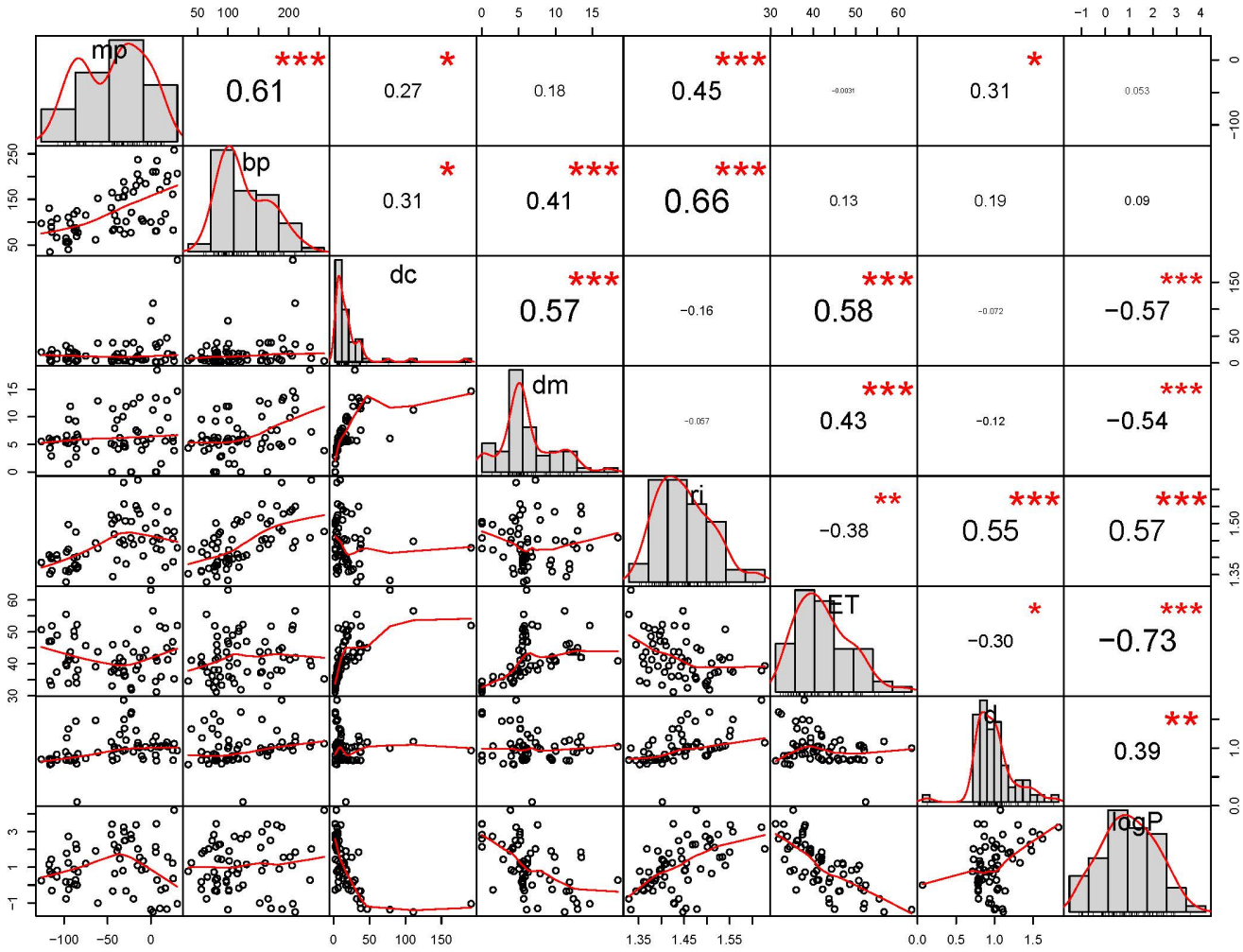

mp
bp
dc
dm
ri
ET
d
logP
Corr: 0.608
Corr: 0.266
Corr: 0.185
Corr: 0.45
Corr: −0.0031
Corr: 0.305
Corr: 0.053
Corr: 0.307
Corr: 0.411
Corr: 0.664
Corr: 0.125
Corr: 0.186
Corr: 0.0904
Corr: 0.572
Corr: −0.158
Corr: 0.585
Corr: −0.072
Corr: −0.574
Corr: −0.0567
Corr: 0.425
Corr: −0.117
Corr: −0.539
Corr: −0.379
Corr: 0.547
Corr: 0.567
Corr: −0.298
Corr: −0.734
Corr: 0.391

Figure 3



The diagrams shown in Figures 2 and 3 have been obtained, respectively, using the *ggpairs()* function of *GGally*[12] R package and the *chart.Correlation()* function of *PerformanceAnalytics*[13] R package.

Both these plots are not only graphically beautiful, but also full of information as they combine in one diagram the scatterplots of all variable pairs (lower triangle), the numerical values of the corresponding correlation coefficients (upper triangle), and the histograms/estimated density distributions of each variable (univariate) along the diagonal.

Professor Everitt and coworkers in their beautiful book on Cluster Analysis [16] state that:

… *scatterplots of each pair of variables can still be used as the basis of an initial examination of the data for informal evidence that the data have some cluster structure, particularly if the scatterplots are arranged as scatterplot matrix…*

… *it is generally argued that a unimodal distribution corresponds to a homogeneous, unclustered population and, in contrast, that the existence of several distinct modes indicates a heterogeneous, clustered population, with each mode corresponding to a cluster of observations. Although this is well known not to be universally true, the general thrust of the methods to be discussed in this section is that the presence of some degree of multimodality in the data is relatively strong evidence in favor of some type of cluster structure.*

… *the humble histogram is often a useful first step in the search for modes in data, particularly, of course, if the data are univariate.*

In light of what above, the histograms / estimated density distributions of each variable along the diagonals of Figures 2 and 3 suggest in some cases (*e.g.*, melting point, boiling point, dipole moment) the presence of two relatively distinct groups of observations in the data. A specific analysis will later verify the correctness of this statement.

In Figure 3, to gain more insight into the possible patterns of data, to each scatterplot is added a *lowess curve* (**LO**cally **WE**ighted **S**catter-plot **S**moother curve).

The *lowess smooth* is a local regression based on nearby points to *x* while the classical regression line is an overall linear fit. *Lowess curves* are typically used for:

- Fitting a line to a scatter plot or time plot where noisy data values, sparse data points or weak interrelationships interfere with your ability to see a line of best fit.
- Linear regression where least squares fitting does not create a line of good fit or is too labor-intensive to use.
- Data exploration and analysis.

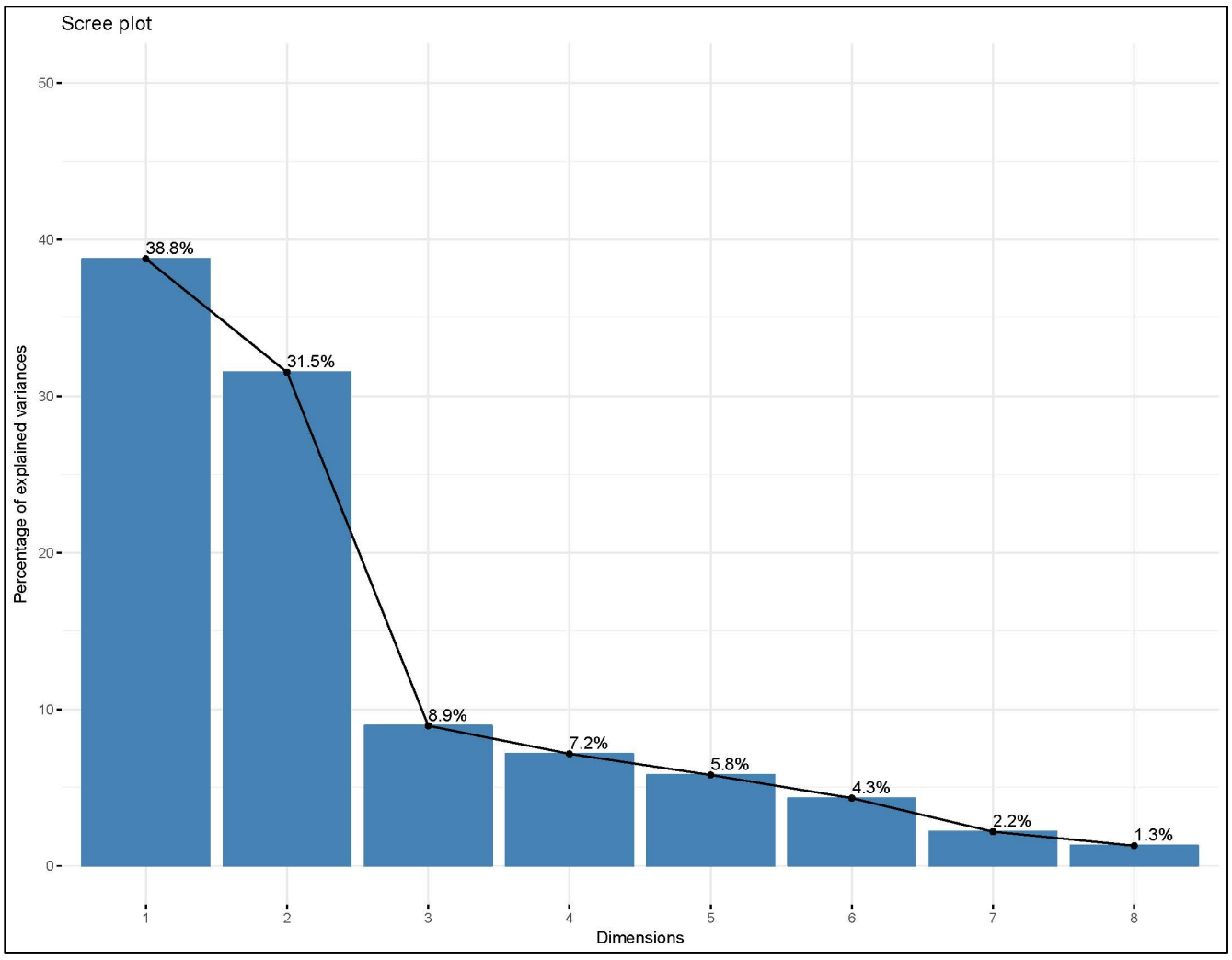
Lowess curves tend to linearity as data points tend to be arranged along a straight line.

Since previous analysis has shown that variables display correlation among them, a Principal Component analysis (PCA) [2] has been carried out to summarize and visualize the information contained in the dataset.

In Figure 4 is shown the scree plot, discussed and named by Cattell (1966) [17], which shows the fraction of total variance in the data as explained or represented by each principal component plotted in successive order from the largest to the smallest.

The scree plot has been obtained using the function *fviz_eig()* of R *factoextra* package.

Figure 4



Being variables well intercorrelated, as already observed examining Figure 1, the diagram in Figure 2 shoes a clear *elbow* just after the second component (*i.e.*, at about 70 % of explained variance). It is interesting to note that the first two components do not differ much from each other in regard to explained variance (~ 38.8% *vs.* 31.2%).
Table 2, here below, summarizes the main results for the first eight principal components, or dimensions.

Table 2

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	3.104	2.493	0.704	0.572	0.482	0.345	0.175	0.126
% of var.	38.794	31.157	8.800	7.148	6.029	4.307	2.193	1.571
Cumulative % of var.	38.794	69.952	78.752	85.900	91.929	96.236	98.429	100.000

In Table 2, the variances of the principal components are the eigenvalues of the correlation matrix while:

. % of var.: is the percentage of variance (or data variability) explained by each component. It provides a measure of the relative importance of each principal component.
. Cumulative % of var.: is the progressive addition, component by component, of the percentage of data variability. The progressive addition of individual contributions is made possible by the orthogonality of the axes (or components).

As after standardization, the original variables have variances of 1.0, the first principal component has a variance of 3.104 of original variables. The second principal component has a slightly smaller value, 2.493, while the other principal components account for far less variation. This confirms the importance of the first two principal components in comparison with the others.

To deepen the knowledge of principal components' structure, Table 3 summarizes the numerical compositions (eigenvectors) of the first five principal components that explain overall about 92% of the initial data variability.

Table 3

Variable	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
mp	0.41	**24.06**	0.03	46.99	19.55
bp	0.09	**31.27**	16.82	0.42	8.84
dc	**15.54**	9.28	7.91	0.34	0.30
dm	**12.34**	9.80	1.01	40.84	19.96
ri	**13.43**	17.46	2.42	3.79	7.18
ET	**22.01**	1.04	0.74	3.14	42.90
d	9.20	7.08	69.35	4.23	1.16
logP	**26.98**	0.01	1.73	0.26	0.11

To the first two principal components (*i.e.*, Dim. 1 and Dim. 2) contribute nearly all variables (seven out of eight) and each occurs with a relevant coefficient.

Simplifying, it can be stated that:

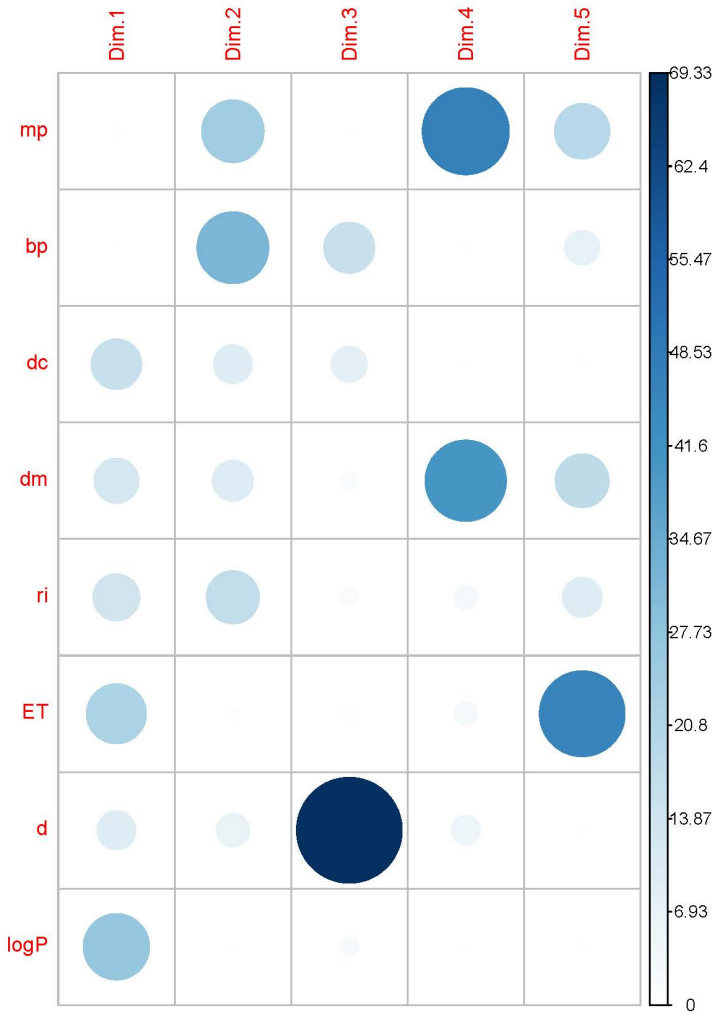
- in the first component are prevalent the contributions related to "polarity/polarizability" (~ 63% as sum of *dc*, *dm*, E_T and *ri*) and "lipophilicity" (~ 27% as per *log P*) of molecules while
- in the second component is prevalent the contribution related to "the strength of intermolecular forces" (~ 55% as sum of *mp*, *bp*).

This last one is obviously a very rough approximation as, for instance, boiling points are not just dependent from the functional groups present in the molecule, but they are also related to the number of carbon atoms and from the molecule branching.

The variable "density" that occurs, first, in the third component and dominates it (69.35%), is practically missing from the first two components.

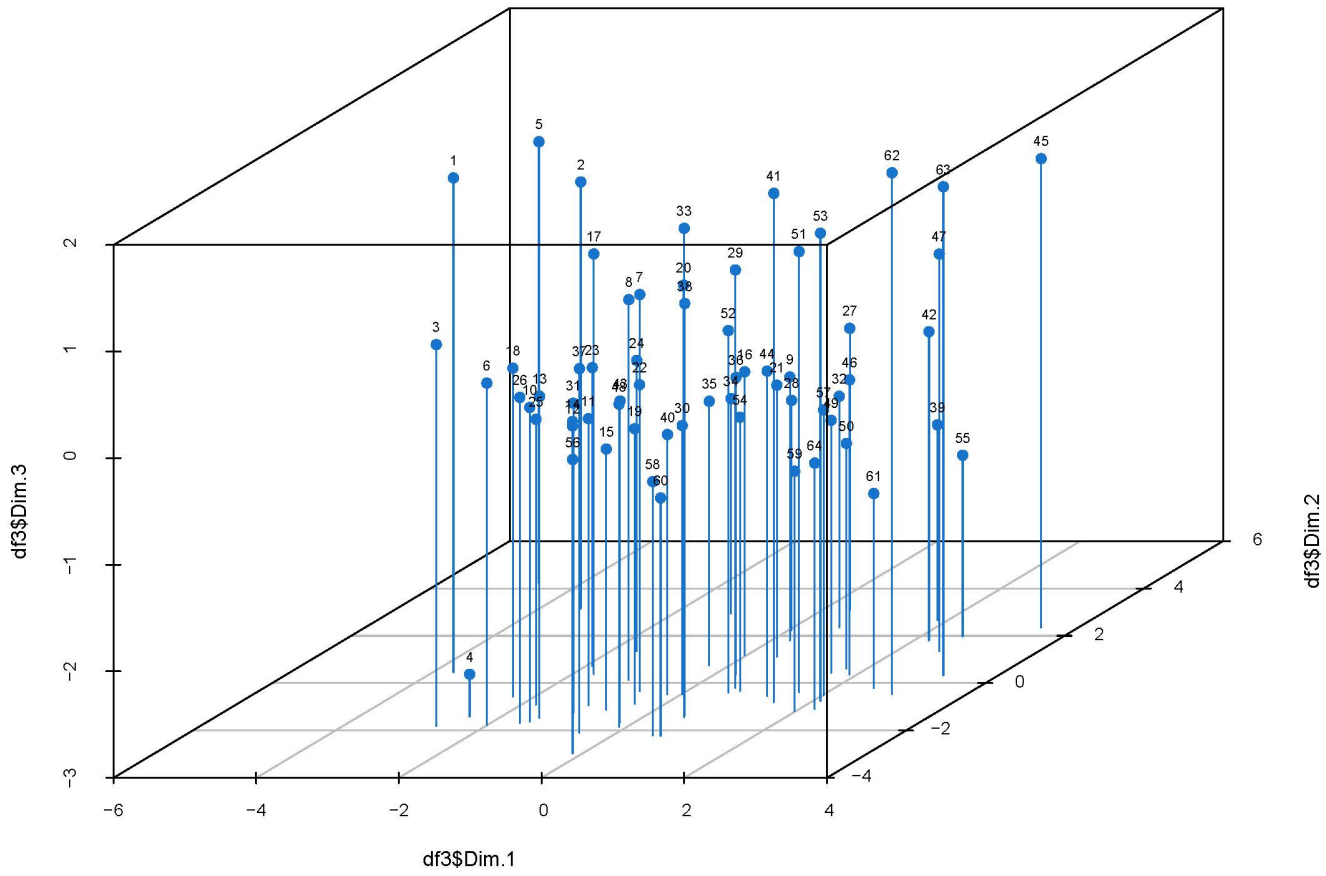
The remarks made analyzing the data listed in Table 3 are evident looking at the diagram shown in Figure 5 that displays, for the first five principal components, the contribution of variables to each principal component or dimension. The larger is the contribution, the darker blue and broader is the spot.

Figure 5



Scatterplot is the oldest and widely used static graphical technique to begin exploring data [18]. Considering the first three components (*i.e.*, about 79% of the total variation in the data), Figure 6 shows a 3D-scatterplot of the individuals obtained using the *scatterplot3d()* function of the R package *scatterplot3d* for visualizing Multivariate Data [14]. In this scatterplot, each point represents a single solvent.

Figure 6



In spite of the high percentage of total variation in the data considered, the diagram of Figure 6 does not visualize much about data distribution. For a more informative view it should be used a scatterplot matrix enhanced with contours of a 2d-density estimate such as those shown in Figures 7-9.

All these diagrams, obtained using in combination the *ggscatter ()* and the *geom_density2d ()* functions of the *ggplot2* R package [19], correspond to projections of the data points on two-dimensional sections of the 3D-scatterplot shown in Figure 6. Each section is cut along a plane defined by two axis each corresponding to a principal component, or dimension.

Figure 7

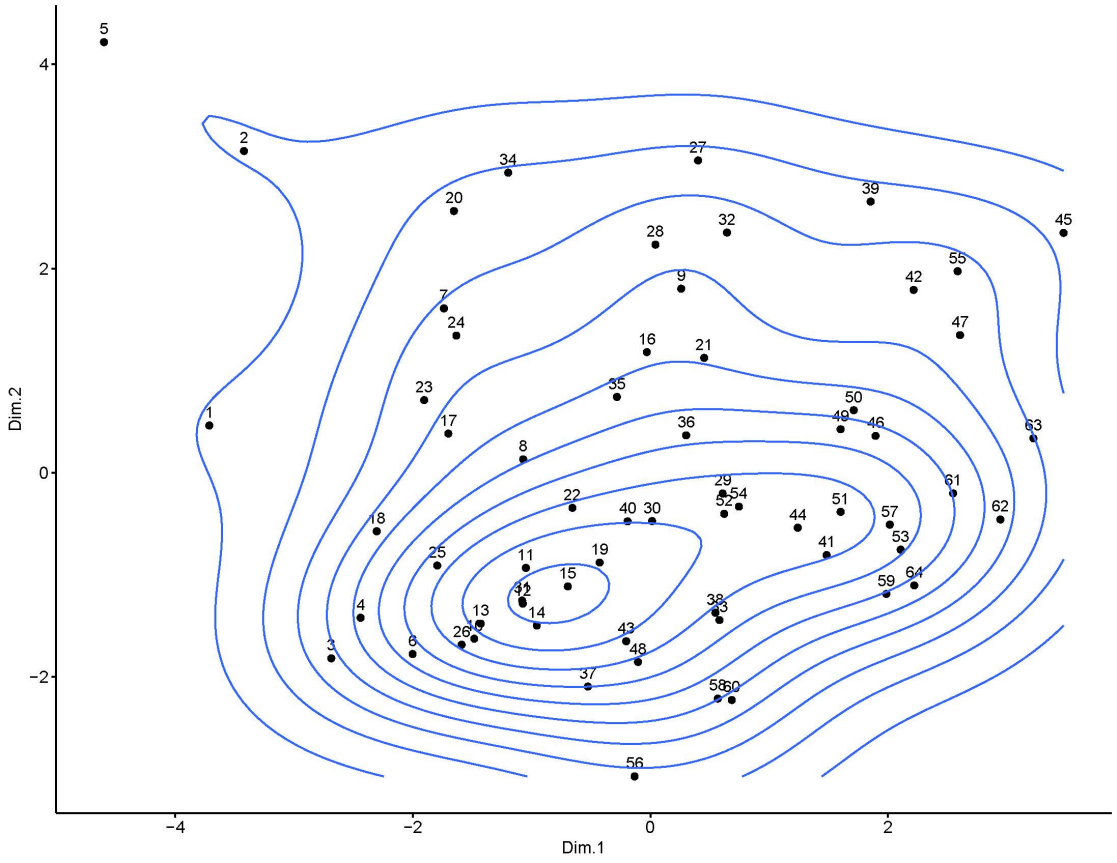


Figure 8

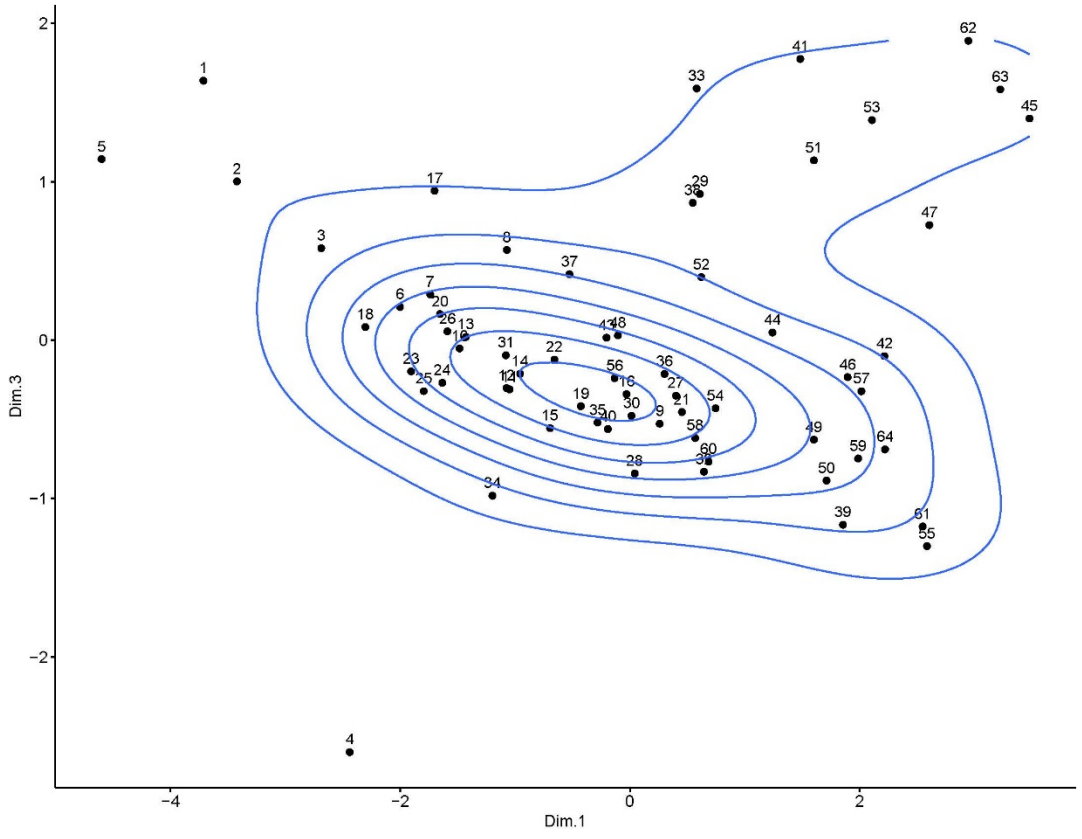
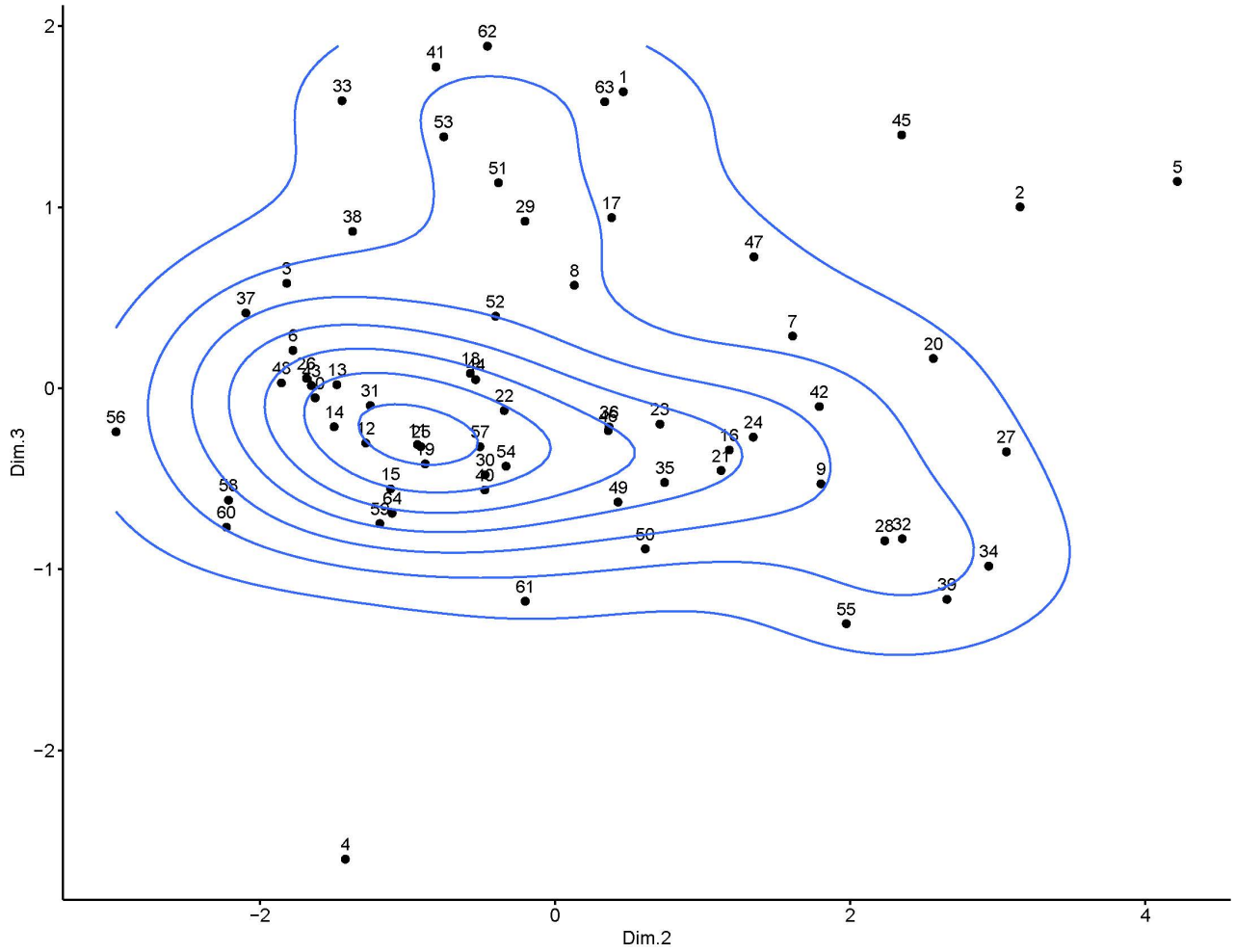


Figure 9



Among the three 2d-contour plots shown in Figures 7-9 that in Figure 7 is the one that better displays the cloud of data points. In the plane defined by the two first principal components, in fact, the cloud is projected in such a manner that the distortion of the swarm of points is minimized and, at the same time, it is captured the maximum variability.

The exam of Figure 7 shows:

- a well defined central kernel centered around data points 11,12,14,15, 31 and 19,
- a second and a third kernel, both just outlined, and centered, respectively, around data points 51 (second kernel) and 9, 28, 32 (third kernel) ,
- a few solvents unrelated with the remaining (1, 2, 5 and 45).

Figures 8 and 9 also capture the anomaly represented by these four last solvents and suggest a possible data point's disposition aggregated around three kernels.

In Table 4, here below, are summarized the data points of Figure 7 above mentioned with the corresponding solvent names and the contributions, to each individual, from the first five principal components (or dimensions). The data in Table 4 are a selection of those obtained using the function *res.pcaindcoord* of the R package *FactoMineR*.

Table 4

Kernel	Solvent No.	Solvent Name	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	11	1-Butanol	**-1.05**	**-0.93**	-0.31	0.04	1.03
	12	2-Methyl-1-propanol	**-1.07**	**-1.28**	-0.30	0.35	1.00
	14	2-Butanol	**-0.96**	**-1.50**	-0.21	0.43	0.88
	15	3-Methyl-1-butanol	**-0.70**	**-1.11**	-0.56	0.64	1.07
	31	2-Butanone	**-1.08**	**-1.25**	-0.10	0.71	-0.56
	19	3-Pentanol	**-0.43**	**-0.88**	-0.42	-0.05	0.56
2	51	1,1,1-Trichloroethane	**1.60**	**-0.38**	1.14	0.03	-0.67
3	9	Benzyl alcohol	**0.26**	**1.80**	-0.53	-0.42	1.59
	28	Benzonitrile	**0.04**	**2.23**	-0.84	1.02	-0.32
	32	Acetophenone	**0.64**	**2.35**	-0.83	0.01	-0.20
-	1	Water	**-3.71**	**0.46**	1.64	-1.81	0.88
-	2	Formamide	**-3.42**	**3.15**	1.00	-0.36	0.68
-	5	N-Methylacetamide	**-4.60**	**4.22**	1.14	-0.41	-0.68
-	45	Iodobenzene	**3.48**	**2.35**	1.40	0.97	1.14
-	56	Diethyl ether	**-0.13**	**-2.97**	-0.24	0.34	-0.74

All members of the first kernel are characterized by comparable values of the coefficients of both the 1st and the 2nd dimension. In the case of 3-Pentanol these values slightly deviate from those of the other members of the first kernel and in fact the corresponding data point (19) looks a bit shifted on the right in Figure 7.

To the first kernel practically belong just alcohols with the exception of 2-Butanone whose data point (31) practically overlaps to that of 2-Methyl-1-propanol (12). This is because of the great similarity existing between their coefficients for Dim. 1 and Dim. 2. Obviously, the remaining principal components, or dimensions, account for the differences existing between these two chemical entities. Table 5, an excerpt of Table 1, highlights the similarities existing between the physicochemical descriptors of these two chemical entities.

Table 5

solvent	mp	bp	dc	dm	ri	ET	d	logP
2-Methyl-1-propanol	-108	107.7	**17.93**	5.97	**1.3959**	**49.0**	**0.794**	0.83
2-Butanone	-86.7	79.6	**18.51**	9.21	**1.3788**	**41.3**	**0.835**	0.29

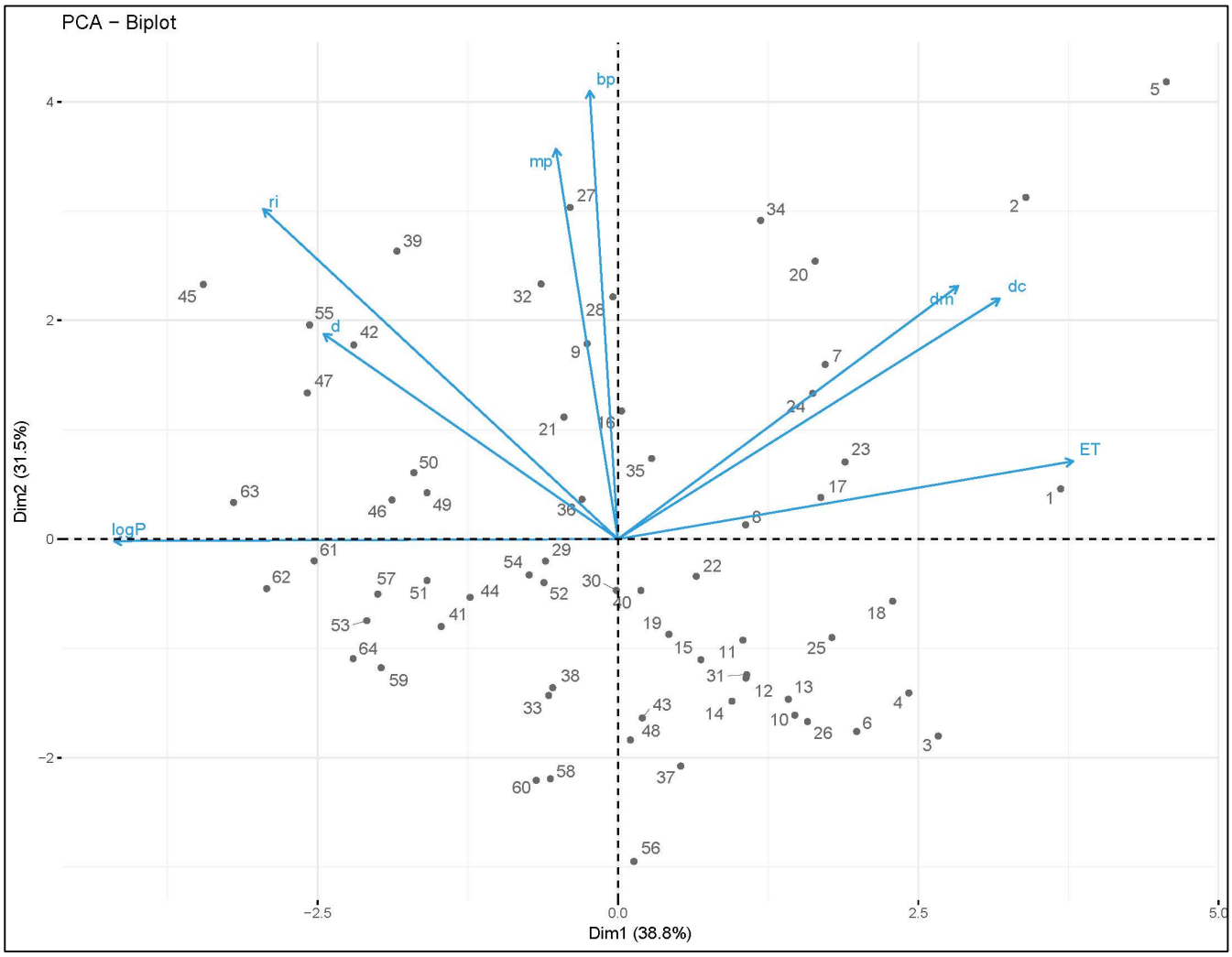
As above mentioned, the 2d-contours plot of Figure 7 suggests a second possible kernel centered around data point 51 corresponding to 1, 1, 1-Trichloroethane. The exam of Figures 8 and 9 substantiates this hypothesis. Figure 8 shows in fact the presence of data point 53, close to 51, while Figure 9 also shows data point 29. Data points 53 and 29 correspond to two chlorinated hydrocarbons: Trichloroethylene (53) and 1, 1-Dichloroethane (29). Figures 8 and 9, in fact, take into account the third principal component that is dominated by the descriptor "density" never considered until then. The third component accounts for an additional 9% about in the explained variability of initial data.

Concerning the third kernel suggested by the 2d-contours plot of Figure 7, Figure 9 strengthen this hypothesis and confirms data points 28 and 32 as its possible center. Data point 9, which looks close to 28 and 32 in Figure 7, is far from them in Figure 9. This finding is in line with the similarities existing between the contributions to the first five components for data points 28 and 32 that differ from those for data point 9 (see Table 4). In fact, data points 28 and 32 are described by practically identical coefficients for Dim. 2 and Dim. 3 and rather similar values for Dim. 5. This last principal component is heavily dominated by the descriptor E_T (~ 43% - see Table 3) and both, Benzonitrile and Acetophenone, have practically identical values for this physicochemical descriptor (42.0 *vs.* 41.3, see Table 1). Data point 9, apart from a kind of similarity in the coefficients for Dim. 2 and Dim. 3, strongly differs for the rest with data points 28 and 32.

Data points 1, 2, 5, 45, and 56 are placed at the borders of the field due to their chemical natures.

In Figure 10 is displayed a PCA-Biplot [20, 21] obtained using the function *fviz_pca_biplot ()* of R *factoextra* package

Figure 10



As already mentioned in the previous post, this type of graphs display simultaneously individuals (*i.e.*, solvents) and variables (*i.e.*, analytical quality descriptors). The Biplot in Figure 10 is drawn using first and second principal components and it can be interpreted as follows:

- positively correlated variables (*e.g.*, *dc* and *dm*, *mp* and *bp*, *d* and *ri*) are grouped together,
- variables negatively related are on opposite quadrants (*dc* and *log P*, E_T and *log P*),
- individuals with a similar profile are grouped together.

Data points in Figure 10 are displayed in a mirrored mode with respect to that shown in Figure 7, but, apart from that, there are no differences between the two arrangements. Even the few isolated points that occur at the borders of the quadrants of Figure 7 (*e.g.*, 1, 2, 5, 45 and 56) can be observed at the extremes of the field in Figure 10.

This difference in data point's arrangement between Figures 10 and 7 is just due to the R code used to draw the Biplot. In fact, what reported in Figure 10 has been obtained entering:

fviz_pca_biplot(prcomp(data, scale = TRUE), repel = TRUE, col.var ="#2E9FDF", col.ind = "#696969")

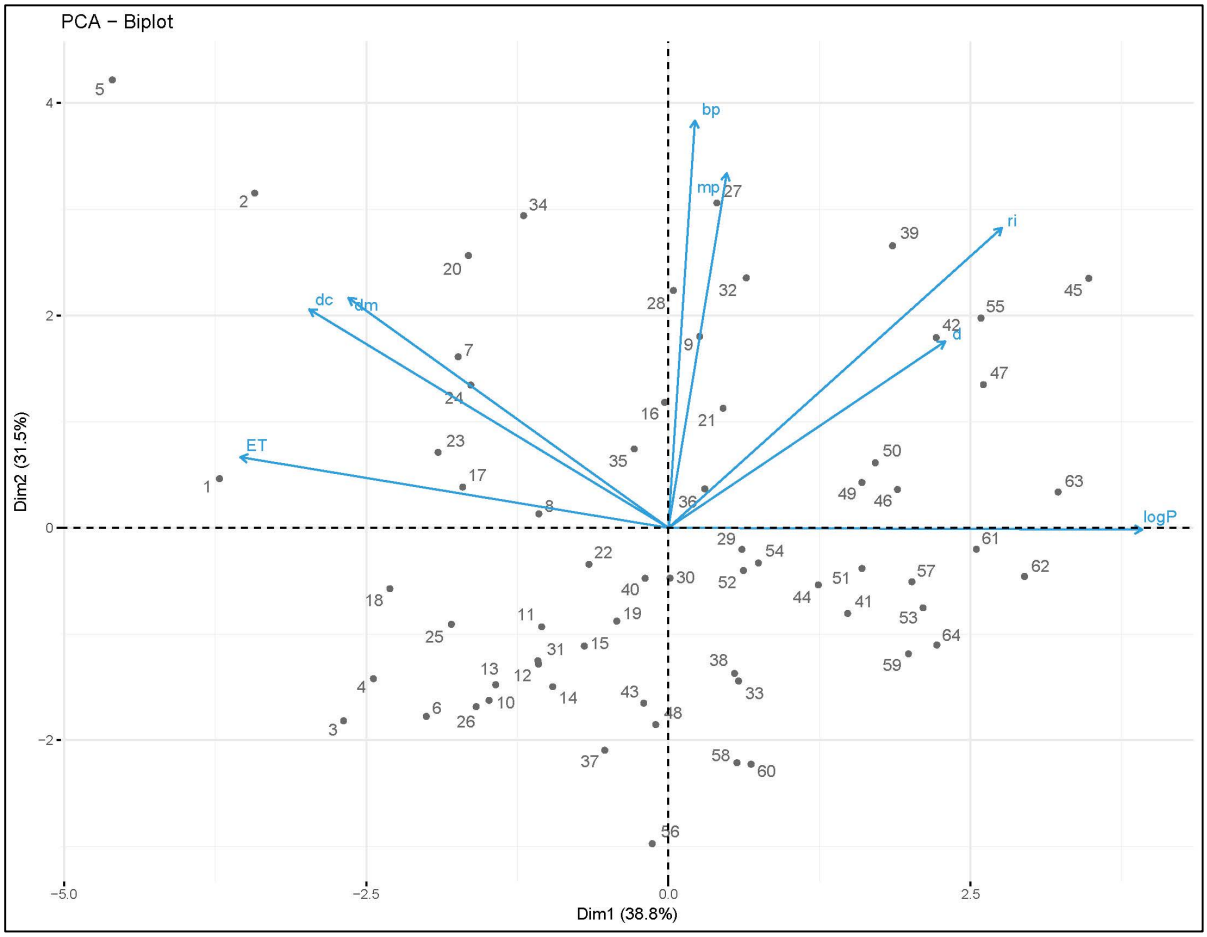
which includes the *prcomp()* function that calculates principal components starting from scaled initial data.

If, on the contrary, it is entered the R code:

fviz_pca_biplot(res.pca, repel = TRUE, col.var ="#2E9FDF", col.ind = "#696969")

that uses PCA results from FactoMineR, it is obtained the diagram shown in Figure 11 here below. The diagram of Figure 11 is the mirror image of that in Figure 10 and it displays the same data points arrangement of Figure 7.

Figure 11



Focusing, for the sake of simplicity, on this last diagram (Figure 11) and considering that:

- an individual that is on the same side of a given variable has a high value for this variable[8],
- any individual that is on the opposite side of a given variable has a low value for that variable[8],
- variables specifically related to "polarity" (*i.e.*, *dc*, *dm* and *ET*) are in the upper left quadrant while those more related to "polarizability" and "intermolecular forces" (*i.e.*, *ri*, *mp*, *bp*) are in the upper right quadrant,
- variable *log P*, that is related to "lipophilicity", is coincident with the positively oriented *x* axis and divides the upper right quadrant from the lower right one,

a few general remarks can be made and, in particular, that:

- solvents whose corresponding data points lay on the right side of the diagram are, in general, low-polar or non-polar while those laying on the left side are polar,

- in the lower right quadrant lay low polarity - apolar solvents (*e.g.*, 1, 2-Dichloroethane (29), Chloroform (41), Dichloromethane (33), Diisopropyl ether (58), Cyclohexane (64), Carbon tetrachloride (62), Benzene (57) *etc.*). All these solvents are characterized by low hydrophobicity or, alternatively, high lipophilicity. This characteristic was not unexpected due to the direction of the vector associated to variable *log P*,

- in the upper left quadrant, opposite to the previous one, lay polar-aprotic and dipolar aprotic solvents (*e.g.*, Water (1), Acetic acid (8), Dimethylsulfoxide (20), N, N-Dimethylformamide (23), *etc.*). Unlike the previous, these solvents are characterized by high hydrophobicity or, alternatively, low lipophilicity as logical in light of the direction of *log P* vector.

- in the upper right quadrant, close to the y axis, lay dipolar aprotic solvents (*e.g.*, Benzonitrile (28), Acetophenone (32), Benzyl alcohol (9), Aniline (21)).

- in the lower left quadrant also lay polar-aprotic and dipolar aprotic (*e.g.*, Acetone (26), Ethyl acetate (43), Acetonitrile (18), Methanol (3), Ethanol (6), 1-Propanol (10), 1-Butanol (11) *etc.*).

It is evident from these considerations than it cannot be identified quadrants that are specific, respectively, for polar-protic or dipolar-aprotic solvents. Nevertheless, the plot in Figure 11 reveals that chemical entities belonging to the same family tend to aggregate in specific areas of the diagram. For instance, several alcohols data points lay close each other in the lower left quadrant of the map shown in Figure 11 while the data points of several chlorinated hydrocarbons can be found in the lower right quadrant.

In light of this, it is reasonable considering *clustering* algorithms to investigate how groups of similar individuals tend to aggregate. However, due to the wide nature of this subject, it will be separately discussed in the next post.

4. CONCLUSIONS

The exploratory data analysis documented in this post, and carried out using the methods of Multivariate Statistical Analysis, has first shown intercorrelation among the physicochemical descriptors used to characterize the solvents under study. This allows to capture 70% of the initial data variability just using two principal components the first of which is related to "polarity/polarizability" and "lipophilicity" of molecules and the second to "strength of intermolecular forces".

The use of these two principal components suggests the possibility of grouping solvents into aggregates (or clusters) of similar individuals. This investigation will be covered in the next post and, due to the nature of the descriptors here used, it will probably lead to a solvent classification simpler than that in nine classes reported by Chastrette [2] which is however based on different types of descriptors.

5. ACKNOWLEDGMENTS

I wish to express my deepest gratitude and appreciation to the R Foundation, RStudio and to all Authors of the R packages that I have used in this post.

6. BIBLIOGRAPHY

1. R. Carlson, T. Lundstedt, C. Albano, *Screening of Suitable Solvents in Organic Synthesis. Strategies for Solvents Selection*, Acta Chemica Scandinavica B, 39 (1985) 79-91
2. I.T. Jolliffe, *Principal Component Analysis*, 2nd Edition, 2002, Springer
3. M. Chastrette, M. Rajzmann, M. Chanon, K.F. Purcell, *Approach to a General Classification of Solvents using a Multivariate Statistical Treatment of Quantitative Solvent Parameters*, JACS, 107 (1), 1985, 1-11
4. H. Wickham, G. Grolemund, *R for Data Science*, 2017, O'Reilly
5. F. Husson, S. Lê, J. Pagès, *Exploratory Multivariate Analysis by Example using R*, 2011, CRC Press.
6. F. Husson, J. Josse, J. Pagès, *Principal component methods – hierarchical clustering – partitional clustering: why would we need to choose for visualizing data?*, September 2010, Technical Report - Agrocampus.
7. A. Kassambara, *R Graphics Essentials for Great Data Visualization*, STHDA, 2017
8. A. Kassambara, *Practical Guide to Principal Component Methods in R*, STHDA, 2017
9. A. Kassambara, *Practical Guide to Cluster Analysis in R*, STHDA, 2017
10. M. Friendly, *Corrgrams: Exploratory displays for correlation matrices*, The American Statistician, 56 (2002) 316-324
11. D.J. Murdoch, E.D. Chow, *A graphical display of large correlation matrices*, The American Statistician, 50 (1996) 178-180
12. J.W. Emerson, W.A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, H. Wickham, *The generalized pairs plot*, Journal of Computational and Graphical Statistics, 22(1) 2012
13. https://cran.r-project.org/web/packages/PerformanceAnalytics/PerformanceAnalytics.pdf
14. U. Ligges, M. Mächler, *Scatterplot3d – an R Package for Visualizing Multivariate Data*, J. of Statistical Software, 8(11), 2003, 1-20
15. A. Struyf, M. Hubert, P.J. Rousseeuw, *Clustering in an Object –Oriented Environment*, J. Statistical Software, 1(4) 1997, 1-30

16. B.S. Everitt, S. Landau, M. Leese, D. Stahl, *Cluster Analysis*, 5th Edition, Wiley (2011) 16-24
17. R.B. Cattell, *The screen test for the number of factors*, Multivariate Behavioral Research, 1(1966), 140-161
18. M. Friendly, D. Denis, *The early origins and development of the Scatterplot*, J. History of Behavioral Sciences, 41 (2) 2005, 103-130
19. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*, Second Edition, Springer, 2016
20. K.R. Gabriel, *Biplot display of multivariate matrices for inspection of data and diagnosis*, in *Interpreting Multivariate Data*, Editor V. Barnett, Chichester Wiley, 1981, 147 – 173
21. J.C. Gower, D.J. Hand, *Biplots*, Chapman & Hall, London, 1996