STATISTICAL STUDY OF A LARGE STRUCTURAL FILE BASED ON THE MENDELEEV TABLE

Michel PETITJEAN and Jacques Emile DUBOIS

ITODYS (Institut de Topologie et de Dynamique des Systèmes),
de l'Université Paris 7, associé au CNRS, 1 rue Guy de la Brosse, 75005 Paris, France

Received July 31, 1989
Accepted November 7, 1989

Dedicated to Professor Otto Exner on the occasion of his 65th birthday.

The atom or element content of a large structural file is considered through relations between the occurrences of the elements and their geometric distribution obtained by correspondence analysis over the Mendeleev periodic table, which is considered as a rectangular (7 × 32) contingency table. The potential of various geometric tools is explored with different CAS files.

The chemical knowledge inherent in large structural files is difficult to apprehend. The number of parameters increases drastically with precision. The simplest information contained in a file is the atom, element of a well defined classification. This highly generic parameter allows simple handling of large amounts of data. We explore a method which is an alternative and a complement to the univariate analysis consisting of the list of atoms and their occurrences, where the element positions in the Mendeleev table is omitted. Considering the type of input data, no new interpretation may be expected; the interest rather lies in pointing out classical features differently and in establishing a basis for further possible investigation.

Elemental composition statistics coming from the CAS file in 1967, 1974, 1979 and 1987 were published[1], giving the statistical weights of the elements. In this paper, a large CAS subfile available at the ITODYS and containing 3 424 428 compounds registered up to July 1978 is investigated (incompletely defined structures and coordination compounds were not taken into account in order to preserve homogeneity of handling and use). The distribution of each of the 103 elements has been considered.

The most interesting one is the carbon distribution (mean 16·985, standard deviation 9·561): see Table I. The two distributions defined with even and odd values are quasi-identical with a even/odd balance of about 53·0–47·0 (37 403 compounds without carbon are not included here).

The hydrogen distribution also offers a greater set of even values which may be explained by the abundance and the odd valency of hydrogen (according to graph

TABLE I

The carbon distribution in the 3 424 428 CAS compound file

Atom	Compounds	Atom	Compounds	Atom	Compounds	Atom	Compounds
1	7 866	43	4 204	85	143	127	25
2	15 103	44	5 562	86	183	128	29
3	20 660	45	3 690	87	135	129	36
4	37 579	46	4 112	88	217	130	36
5	47 194	47	2 495	89	127	131	37
6	84 992	48	4 069	90	199	132	51
7	93 749	49	1 973	91	125	133	34
8	129 229	50	2 756	92	121	134	50
9	146 038	51	1 787	93	118	135	30
10	187 807	52	2 244	94	106	136	59
11	175 058	53	1 372	95	121	137	40
12	202 276	54	2 145	96	176	138	31
13	181 448	55	1 344	97	107	139	38
14	197 990	56	1 664	98	144	140	36
15	184 784	57	1 208	99	121	141	29
16	187 214	58	1 285	100	122	142	50
17	157 197	59	783	101	89	143	31
18	159 743	60	1 407	102	98	144	47
19	135 080	61	727	103	68	145	31
20	140 901	62	1 002	104	83	146	29
21	120 291	63	752	105	63	147	29
22	111 367	64	911	106	73	148	36
23	85 749	65	604	107	56	149	23
24	82 790	66	793	108	111	150	45
25	59 191	67	395	109	58	151	40
26	57 846	68	682	110	72	152	23
27	47 725	69	440	111	60	153	28
28	45 445	70	567	112	51	154	36
29	32 856	71	315	113	49	155	21
30	36 803	72	704	114	67	156	18
31	22 545	73	300	115	46	157	15
32	24 253	74	330	116	47	158	20
33	16 332	75	289	117	55	159	23
34	18 537	76	436	118	46	160	14
35	11 373	77	234	119	36	161	9
36	14 993	78	370	120	68	162	11
37	8 459	79	208	121	41	163	10
38	9 759	80	344	122	45	164	12
39	6 548	81	244	123	39	165	6
40	8 958	82	267	124	43	166	4
41	5 260	83	140	125	41	167	3
42	7 534	84	280	126	37	168	11

TABLE I
(*Continued*)

Atom	Compounds	Atom	Compounds	Atom	Compounds	Atom	Compounds
169	8	176	2	183	3	196	2
170	8	177	6	185	1	198	1
171	3	178	3	187	1	199	1
172	2	179	1	188	1	200	1
173	7	180	6	189	3	206	1
174	1 .	181	3	190	3	208	1
175	4	182	1	191	1	220	1

theory, there is an even number of odd-connected nodes). The distribution of many data registered in the file has also been obtained: bonds, components, valencies, charges, isotopes, and other data depending or not on the recording mode of a compound.

When the total number of each element is obtained we suggest a statistical view of the Mendeleev table, giving more than the classical ordering of the 103 elements, from the most to the least abundant. Either alphanumeric or graphic presentations of the results are available with the technique described below.

METHODS AND RESULTS

Performing Correspondence Analysis

We show here the results of relations carried out on the most elementary level of chemical knowledge, that of the atom. The methodological tools issued from correspondence analysis are used in this paper for the first time in this field.

Correspondence analysis is a multivariate exploratory technique devoted to contingency table analysis (see refs[2,3] for theoretical aspects and mathematical results). This technique is similar to PCA (principal component analysis), but applied to categorical data; the contingency table is the Mendeleev table, described with two categorical variables (see Tables II and III): the period of the element (7 categories), and the chemical family (32 categories). This rectangular formatting of the periodic table is conventionally obtained by assigning a zero value to non-pertinent positions (this 2D-presentation shows the occurrences as a potential third dimension).

Compared to PCA, the contingency table may be considered as a set of 32 individuals described with 7 continuous variables, or a set of 7 individuals described with 32 continuous variables; both PCA (individuals are weighed, and variables are weighed: see refs[2,3]) would give the same 7 eigenvalues. Then eigenvalues and vectors are computed (Table IV). The highest eigenvalue is 1 for every contingency table, so there are only 7 – 1 = 6 factorial axes. Each of the 148 815 839 atoms of the file now has its 6-dimensional coordinates, but the coordinates of the atoms having the same atomic number are identical. There are then only 103 different points in the factorial space, each point being one of the 103 elements of the Mendeleev table.

TABLE II

Occurrence of the elements in the Mendeleev table

H	70 908 654	Kr	200	Lu	366
He	51	Rb	1 338	Hf	375
Li	7 668	Sr	1 138	Ta	628
Be	654	Y	626	W	1 746
B	66 864	Zr	1 219	Re	585
C	57 528 231	Nb	577	Os	333
N	5 820 786	Mo	2 295	Ir	265
O	10 568 323	Tc	234	Pt	542
F	767 626	Ru	435	Au	443
Ne	70	Rh	397	Hg	9 047
Na	48 281	Pd	557	Tl	1 796
Mg	4 522	Ag	2 441	Pb	3 477
Al	4 773	Cd	1 587	Bi	1 094
Si	124 453	In	625	Po	162
P	233 598	Sn	17 823	At	156
S	1 101 733	Sb	4 214	Rn	73
Cl	1 110 863	Te	3 050	Fr	79
Ar	86	I	94 361	Ra	108
K	14 964	Xe	317	Ac	84
Ca	4 649	Cs	1 711	Th	659
Sc	480	Ba	2 971	Pa	156
Ti	2 739	La	907	U	1 230
V	1 816	Ce	812	Np	302
Cr	2 625	Pr	669	Pu	321
Mn	1 649	Nd	776	Am	221
Fe	5 661	Pm	118	Cm	113
Co	2 681	Sm	751	Bk	90
Ni	2 444	Eu	596	Cf	98
Cu	3 934	Gd	584	Es	83
Zn	3 827	Tb	409	Fm	72
Ga	708	Dy	504	Md	56
Ge	8 895	Ho	397	No	61
As	12 560	Er	530	Lr	51
Se	20 508	Tm	340		
Br	257 583	Yb	499		

The projection of the points in the first factorial planes is given in Fig. 1; it is possible, just as for PCA, to interpret the factorial axes. The first axis shows an opposition between most and least abundant elements, and the second shows an opposition between low and high atomic numbers; the actinides group and the lanthanides group are far from other elements. The geometrical repartition of these 103 points is a picture of the statistical content of the Mendeleev table, suitable for comparison with other files or subfiles, and for following the chronological evolution of a file.

TABLE III

The contingency table: the 7 rows and 32 columns are exchanged for clarity

51	70	86	200	317	73	—
—	767 626	1 110 863	257 583	94 361	156	—
—	10 568 323	1 101 733	20 508	3 050	162	—
—	5 820 786	233 598	12 560	4 214	1 094	—
—	57 528 231	124 453	8 895	17 823	3 477	—
—	66 864	4 773	708	625	1 796	—
—	—	—	3 827	1 587	9 047	—
—	—	—	3 934	2 441	443	—
—	—	—	2 444	557	542	—
—	—	—	2 681	397	265	—
—	—	—	5 661	435	333	—
—	—	—	1 649	234	585	—
—	—	—	2 625	2 295	1 746	—
—	—	—	1 816	577	628	—
—	—	—	2 739	1 219	375	—
—	—	—	—	—	366	51
—	—	—	—	—	499	61
—	—	—	—	—	340	56
—	—	—	—	—	530	72
—	—	—	—	—	397	83
—	—	—	—	—	504	98
—	—	—	—	—	409	90
—	—	—	—	—	584	113
—	—	—	—	—	596	221
—	—	—	—	—	751	321
—	—	—	—	—	118	302
—	—	—	—	—	776	1 230
—	—	—	—	—	669	156
—	—	—	—	—	812	659
—	—	—	480	626	907	84
—	654	4 522	4 649	1 138	2 971	109
70 908 654	7 668	48 281	14 964	1 338	1 711	78

Convex Hulls and Peeling

It is difficult to provide a simple description of a 6-dimensional set of points without altering information. For a one-dimensional set, a possible description is the ordering of the points, pointing out the extremal values; for a multidimensional

TABLE IV
Eigenvalues and inertia percent

Eigenvalues (except trivial value 1)	Associated cumulated inertia percent
0·998074	45·054%
0·586131	71·512%
0·364415	87·962%
0·218656	97·833%
0·041272	99·696%
0·006742	100·000%

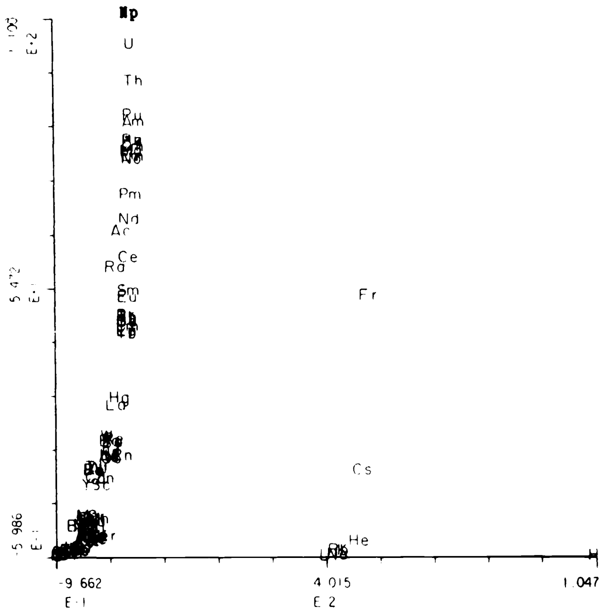

FIG. 1
The 103 atomic symbols in the first factorial plane (axis 1 vertical, axis 2 horizontal)

set, it is also possible to give extremal values and a partial ordering. Extremal values are, mathematically speaking, the extremal points of the convex hull of the set.

The convex hull of a set of points is the intersection of all the convex sets containing the points; it is also the smallest polyhedron containing the points. The vertices of this polyhedron are called the extremal points. This polyhedral hull offers a simple description of the shape of the set.

After the convex hull has been computed, the set of the internal points is considered. This new set also has a convex hull, enclosed in the first one; we then consider the new internal points, and so on, until there are no remaining points. This process, called peeling, has been used for multivariate data ordering[4-6], and is suitable for describing the wide set.

When the points are projected on a sub-space, it is known that the convex hull of the projections is also the projection of the convex hull. Thus every bidimensional convex hull computed in a factorial plane can provide a display of the projection of the multidimensional convex hull (however, the peeling of the bidimensional set does not give the projection of the peeling of the multidimensional set, because some extremal points may be on none of the factorial planes).

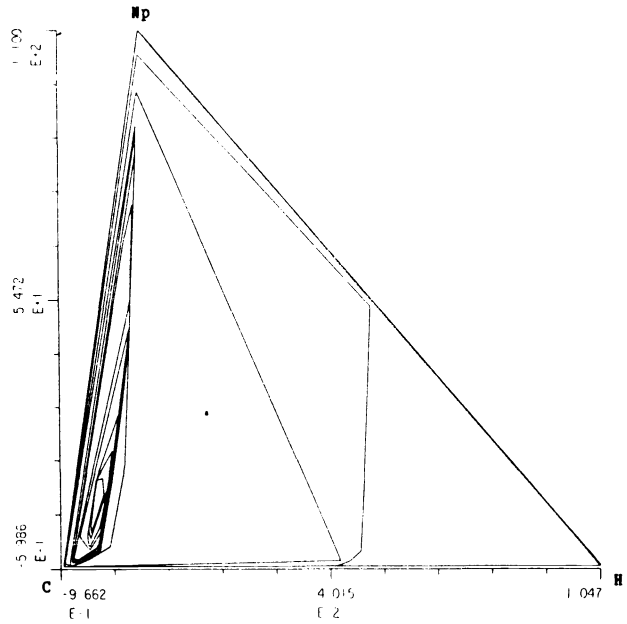

Fig. 2

Peeling of the 103 atomic symbols in the first factorial plane (axis 1 vertical, axis 2 horizontal)

The peeling in the first factorial plane is shown in Fig. 2; extremal points in this first factorial plane are ordered trigonometrically (Table V). Since all the 148 815 839 atoms of the file take only the 103 positions of the elements in the factorial space, the convex hull of the 103 elements is also the convex hull of the 148 815 839 atoms.

We point out that the peeling of the 148 815 839 atoms (and not of their 103 positions) provides a different set of successive convex hulls, together with an ordering of the 103 symbols. The least abundant on the external hull is the first symbol removed by the algorithm; then the new least abundant on the external hull is removed, and so on until no element remains. This procedure, which is the usual peeling procedure for data analysis, requires a special algorithm saving much computation time. The example below comes from the first factorial plane (see Table VI and Fig. 3).

A compound is a geometric mean of its atoms (a geometric mean is a convex linear combination), and every one of the 103 uniatomic compounds exists in the file. Thus, the external convex hull of the 103 elements is also the convex hull of the 3 424 428 compounds. The extremal compounds are those monoatomic compounds whose unique atom is extremal, such as C, H, or Np.

Comparison with Other CAS Files

The elemental composition statistics published[1] give reference data to be compared with the 1978 file. The 1974, 1979 and 1987 files, and the file defined by difference

TABLE V
Peeling in the first factorial plane

Number of symbols in the hull	List of symbols in each hull (from outermost to innermost hulls)
3	C H Np
8	N Li Na K He Cs Fr U
3	O Rb Th
11	B Si S Cl Br Kr Rn Yb No Fm Pu
11	F P Se Ar Xe Os Hg Er Lr Es Am
11	Sn Al Ge As Fe Ir Re Lu Md Cm Pa
9	Sb Te Ga Ne Co Pt Tm Cf Bk
13	In I Cu Ti Ni Mn Ta Gd Dy Ce Pm Ac Ra
8	Be Ag Ru V Hf Ho Tb Nd
11	Mg Zr Rh Ca Cr Au W Pr Eu Sm Pb
5	Sr Pd Zn La Bi
8	Mo Nb Tc Sc Ba Tl Po Y
2	Cd At

between 1987 and 1979 data, were treated with correspondence analysis (the 1967 file was not considered because many elements were missing: see ref.[1]). The coordinates of the symbols gave similar shapes for successive hulls, either with or without weighed symbols (see e.g. Fig. 4).

We show the influence of a perturbation starting from a probable printing error for the occurrence of W in the CAS 1974's file[1] (see Table VII). The value 134 149 is then improbable, and does not match with the value computed with the percentage: 23 768 W atoms. The cumulated sum of all the elements computed with the data also given in ref.[1] is then 118 283 553, which leads to 23 790 W atoms.

The data sets, differing only by the number of W atoms, are compared; the variations of the 6-dimensional coordinates of W are shown in Table VII. The relative variation on the last axis has the same magnitude as the relative variation of the number of W atoms, and the relative variation on the first axis has the same magnitude as the relative variation of the sum of all the 103 elements. Intermediate variations are observed on intermediate axes.

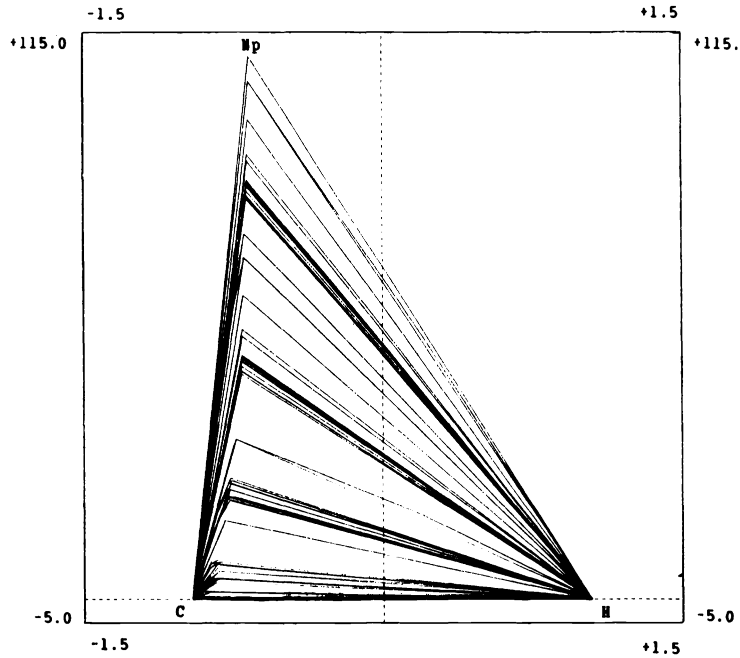

Fig. 3

Peeling of the 148 815 839 atoms in the first factorial plane (axis 1 vertical, axis 2 horizontal)

TABLE VI

Peeling of the weighed symbols in the first factorial plane

Number of symbols in the hull	List of symbols in each hull (from outermost to innermost hulls)	
3	C H Np	(Np to be removed)
4	C H Fr U	(Fr to be removed)
3	C H U	(U to be removed)
....	....	....
	(103 hulls were computed by the algorithm)	
....	....	....
3	C H Li	(Li to be removed)
2	C H	(C to be removed)
1	H	(H to be removed)

Ordering of the symbols, from the first to the last removed

Np, Fr, U, Th, Pu, Am, Pa, Bk, Es, Cf, Cm, Md, Lr, Fm, No, Ac, Pm, Ra, Nd, Ce, Sm, Eu, Pr, Tb, Ho, Dy, Gd, Tm, Lu, Er, Yb, La, Cs, Hg, W, Tl, Re, Pb, Bi, Ba, Po, Ta, Y, At, Pt, Rn, Hf, Ir, Au, Os, Cd, Be, Zn, Sc, Mo, Tc, Cr, Sr, Nb, Mn, Mg, Pd, Ca, V, He, Zr, Ni, Rh, Ag, Ti, Ru, Kr, Xe, In, Co, Cu, Ne, Fe, Ar, Ga, Rb, K, Sn, Sb, Te, B, I, Al, Ge, As, Br, Se, F, Cl, Na, P, S, Si, N, O, Li, C, H

TABLE VII

Influence of a perturbation on the coordinates

1974 file: 134 149 W atoms (0·020113% of the 118 173 172 atoms)
1979 file: 44 549 W atoms (0·022214% of the 200 537 175 atoms)
1987 file: 88 739 W atoms (0·022481% of the 394 730 177 atoms)

Axis	134 149 W atoms	23 790 W atoms	23 768 W atoms
1	−0·7714984 E + 00	−0·7732000 E + 00	−0·7732018 E + 00
2	0·1366503 E + 02	0·1117546 E + 02	0·1117443 E + 02
3	−0·3353852 E + 01	−0·4478106 E + 01	−0·4478546 E + 01
4	−0·7318581 E + 01	−0·6753468 E + 01	−0·6752183 E + 01
5	−0·1906342 E + 01	−0·3382746 E + 01	−0·3381835 E + 01
6	−0·3689612 E + 00	0·8848312 E + 00	0·8861454 E + 00

DISCUSSION AND CONCLUSION

This attempt to present a multivariate analysis of a large structural file shows how simple graphic displays may give characteristic pictures of the file intended for comparison with others. The information taken from the file was limited to atomic nature, giving a graphic representation of the Mendeleev table. A complete interpretation of the graphic Mendeleev table and its ordering of the symbols with peeling would require a 6-dimensional algorithm, numerically consolidated. Only 2-dimensional examples were presented, in order to have simple outputs and results. The technique can be easily extended to all information registered in the file (and not only to atomic nature), using multiple correspondence analysis.

No problem was encountered in handling large amounts of data. Every contingency table can be computed with an execution time proportional to the number of individuals, without using storage areas (except for the contingency table itself, which is small compared to the large number of individuals). Every computation needed for simple or multiple correspondence analysis can then be performed without rereading the file.

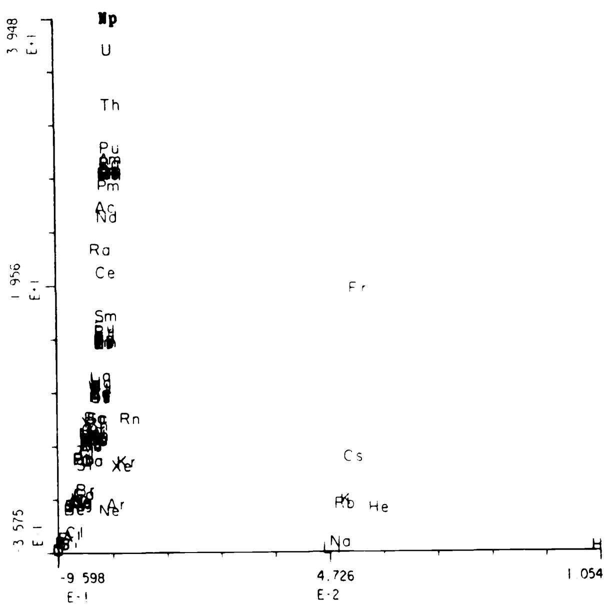

FIG. 4

The 103 atomic symbols in the first factorial plane (CAS 1987) (axis 1 vertical, axis 2 horizontal)

It is also possible to define the 6-dimensional coordinates of each of the 3 424 428 compounds. A compound is a group of atoms, each atom having one of the 103 6-dimensional coordinates. A correct representation of the compound will be the geometric mean of its atoms (this is a usual definition of groups in correspondence analysis); for example, every uniatomic compound will take the coordinates of its unique element. Moreover, there is a distance between every couple of compounds, so that a chemical synthesis can be represented by a positive valued graph. Now, unarbitrary numerical values are suitable for correlation attempts or classification algorithms.

The coordinates obtained here for compounds having the same elemental composition are identical, but a multiple correspondence analysis performed with variables describing expanded formulas will give separate points. This approach is possible each time a set of categorical variables is defined over a file, followed by multiple correspondence analysis. When structural descriptions of compounds are needed for QSAR or related correlation and classification problems, the problem is usually to convert these descriptions into continuous values, which are required for many analyses. This problem can be replaced by a new one: how to define a set of categorical variables to obtain a good representation of structural information. This new problem is easier to solve, because structural information has indeed a qualitative nature (e.g. fragments, chemical family, functional group), and not a numerical nature; multiple correspondence analysis can then be performed to give the expected continuous values.

REFERENCES

1. Stobaugh R. E.: J. Chem. Inf. Comput. Sci. *28*, 180 (1988).
2. Benzecri J. P.: *L'analyse des données*, Tome 2: ISBN 2-04-007225-X or 2-04-007335-3. Dunod, Paris 1973.
3. Greenacre M., Hastie T.: J. Am. Stat. Assoc. *82*, 437 (1987).
4. Barnett V.: J. R. Stat. Soc., A *139*, 318 (1976).
5. Barnett V.: *Interpreting Multivariate Data*, Chap. 1. Wiley, New York 1981.
6. Holmes-Junca S.: *Thesis*. Montpellier II University, Montpellier 1985.