Journal of Chromatography, 350 (1985) 151-168 Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROM. 18 102

QUALITATIVE ORGANIC ANALYSIS

I. IDENTIFICATION OF DRUGS BY PRINCIPAL COMPONENTS ANALY-SIS OF STANDARDIZED THIN-LAYER CHROMATOGRAPHIC DATA IN FOUR ELUENT SYSTEMS

GIUSEPPE MUSUMARRA*, GIUSEPPE SCARLATA and GIUSEPPE CIRMA

Dipartimento di Scienze Chimiche, Università di Catania, Viale Andrea Doria 6, 95125 Catania (Italy) GUIDO ROMAN0 and SILVANA PALAZZO

Istituto di Medicina Legale e delie Assicurazioni. Universitd di Catania, Via Biblioteca 4, 95124 Catania (Italy)

and

SERGIO CLEMENT1 and GIANFRANCO GIULIETTI

Dipartimento di Chimica, Università di Perugia, Via Elce di Sotto 10, 06100 Perugia (Italy) (Received August 6th, 1985)

SUMMARY

Principal component analysis of standardized R_F values in four eluent systems [ethyl acetate-methanol-30% ammonia (85:10:15), cyclohexane-toluene-diethylamine (65:25:10), ethyl acetate-chloroform $(50:50)$ and acetone, with the plate dipped in potassium hydroxide solution] provided a two-component model which accounts for 73% of the total variance. The "scores" plot allowed the restriction of the range of inquiry to a few candidates. This result is of great practical significance in analytical toxicology, especially when account is taken of the cost, the time, the analytical instrumentation and the simplicity of the calculations required by the method.

INTRODUCTION

The idea of using thin-layer chromatography (TLC) in qualitative organic analysis has long been pursued due to the simplicity, the low cost, the rapidity and the sensitivity of this analytical technique. Obviously a single retention factor, R_F , is not sufficient for the identification of any organic compound and it is evident that more measurements are needed. The R_F values in different eluent systems reported either in graphical representations such as the "chromatographic spectrum"' and the "chromatographic profile"², or in tables³ have been considered to be suitable for identification purposes. In this regard the choice of the minumum number of eluent systems containing different information is of crucial importance for the identification of unknowns and has been the topic of several statistical studies. The individual information provided by each eluent system and the correlation between such systems

have been investigated using the "discrimination power"³⁻⁶, while information theory⁷ and numerical taxonomy techniques^{8,9} have been used to evaluate the quality of TLC separations and for the selection of optimum sets of eluents.

In this context principal components analysis (PCA) has been proven to have great potential for the identification of basic drugs such as benzodiazepines, phenothiazines and opiates from their R_F values in eight eluents¹⁰ and for the evaluation and the selection of eluent systems in $TLC^{11,12}$. PCA has significant advantages over statistical methods based on the information provided by single systems due to the fact that it enables a direct measure of the properties of each system in combination with the others, and indicates both the minimum set of eluents needed and reliable statistical criteria for their selection. Following these criteria we have recently proposed a minimum set of four eluents containing virtually all the information obtainable from a larger set of 40 eluent mixtures¹². As a consequence of these results we here report the PCA of the R_F values of 362 drugs in the above set of four eluents with the purpose of achieving a drastic restriction in the range of inquiry and hopefully identification of unknown samples. The examined compounds (basic, neutral and a few acidic drugs), which include substances widely used in Italy for therapeutic purposes and well known drugs of abuse, are nitrogen bases which can be detected using the Dragendorff reagent and acidified iodoplatinate solution¹³.

EXPERIMENTAL

The drugs are named according to the Merck Index¹⁴; for substances not reported therein, either the nomenclature adopted by Clarke¹³ or that of Chemical Abstracts is used.

RF Measurements

The eluent compositions are reported in Table I, together with the $R_F \times 100$ values for four reference compounds in each system. These $R_F \times 100$ values were used to correct the experimentally determined $R_F \times 100$ values (see below). The corrected $R_F c \times 100$ values for compounds 1-362 in eluent mixtures I-IV are reported in Table II.

The drug (10 mg) was dissolved as the free base form or as the hydrochloride salt (different salts are explicitly stated in Table II) in methanol (5 ml), or extracted from an alkaline aqueous solution with ethyl acetate and prepared as a solution containing about 2 mg/ml of drug. No significant differences between the R_F of the free base and those of the salts (especially the hydrochlorides) were observed. All drug solutions were freshly made and aliquots, 2-3 μ l containing 4-6 μ g of drug (except where otherwise stated in Table II), were applied approximately 1 cm apart to 20 \times 10 cm silica gel 60 F₂₅₄ HPTLC plates (Merck). For eluent IV the plates were dipped in 0.1 M potassium hydroxide methanolic solution and dried before application of the drugs. The quantity of drug applied was strictly dependent on the sensitivity towards the detection reagents. Amounts of $4-6$ μ g were usually sufficient to obtain spots which were clearly visible and had the same intensities as those of the reference compounds: cases where higher quantities of drugs were needed are explicitly indicated by footnotes in Table II.

The standardization procedure suggested by Stead *et a1.3* for the correction of

TABLE I

STANDARDIZED TLC SYSTEMS

Silica gel 60 F_{254} HPTLC plates and saturated chambers were used throughout. Solutions of the four reference compounds were prepared to give a concentration of approximately 2 mg/ml of each compound.

* Plates were dipped in 0.1 M potassium hydroxide methanolic solution and dried.

 $R_F \times 100$ values was adopted throughout. A solution (2 μ) containing an appropriate mixture of reference compounds (see Table I) was applied at three separate positions along the baseline of each plate, together with the solution of the drugs. The solvents (100 ml) were placed into TLC tanks, which were sealed and allowed to equilibrate for at least 30 min before use. The systems were allowed to migrate 5 cm from the baseline. The use of shorter distances has been shown not to produce significant changes either in the corrected R_F values or in the reproducibility¹⁵. The solvent front was marked and the plates were air-dried. The drug detection was achieved by spraying first with 10% sulphuric acid, then with the Dragendorff spray reagent^{3,13} and finally with acidified iodoplatinate solution^{3,13}.

The R_F values were measured independently in two laboratories where the eluent mixtures were freshly prepared using commercial solvents often provided by different companies; values were corrected according to the standardization procedure of Stead *et al.*³. The experimentally determined $R_F \times 100$ values were converted into the corrected values $(R_Fc \times 100)$ by a graphical method, using a six-point correction graph including the $R_F \times 100$ values of the four reference compounds, together with the 0,0 and 100,100 points. The $R_{FC} \times 100$ data for compounds 1-362 in eluent mixtures I-IV reported in Table II are averages of four determinations (two in each laboratory).

Principal components analysis

The PCA using the soft independent modelling of class analogy (SIMCA) method¹⁶⁻¹⁹ and its applications for the identification of drugs by TLC in different

TABLE II

CORRECTED $R_r c \times 100$ VALUES IN ELUENTS I-IV, PRINCIPAL COMPONENTS SCORES, θ_1 , θ_2 , AND RESIDUAL STANDARD DEVIATIONS s_k AFTER 2 PC, FOR COMPOUNDS 1-362

TABLE II *(continued)*

(Continued on p. 156)

TABLE II (continued)

TABLE II *(continued)*

(Continued on p. 158)

TABLE II *(continued)*

TABLE II *(continued)*

(Continued on p. 160)

TABLE I1 *(continued)*

161

TABLE II *(continued)*

* Nomenclature according to the *Merck Index*¹⁴ except where otherwise stated. Quantity of drug in the range $4-6 \mu$ g except where otherwise stated.

** Nomenclature according to Chemical Abstracts.

*** Nomenclature according to Clarke¹³.

[§] Quantity of drug in the range 7–15 μ g.

^{§§} Quantity of drug in the range 16-30 μ g.

 $\frac{555}{100}$ Quantity of drug in the range 30–50 μ g.

 \dagger The chromatographic spot showed an elongated shape.

eluent systems¹⁰⁻¹² have been presented in detail. In the present instance, the matrix Y with the elements y_{ik} contains R_Fc values, where index *i* is used for the eluent mixtures (variables) and index *k* for the compounds (objects). From this data matrix, the number of significant product terms, A, and then the parameters α_i , β_{ia} and θ_{ak} in eqn. 1 are estimated by minimizing the sum of the cross validated squared residuals, ε_{ik} :

$$
y_{ik} = \alpha_i + \sum_{a=1}^{A} \beta_{ia} \theta_{ak} + \varepsilon_{ik} \tag{1}
$$

In this model, α_i and β_{ia} are constants which are only dependent on the eluent mixtures and θ_{ak} are the compound-dependent parameters. The deviations from the model are expressed by the residuals ε_{ik} , which include also the experimental errors in the determination of the R_Fc values.

Before the PCA computation, the eluent parameters were autoscaled (see, e.g., ref. 19), *i.e.,* the variables were given the same variance (unity). With this scaling, all variables were given the same initial importance in the PCA, so that the model chooses the relative importance of each eluent system when defining the components according to their information content. We are aware that other authors prefer to weight variables according to the measurement precision. However, we have given reasons for our choice: "the use of multivariate methods such as PCA, where all objects are to be described at the same time by all variables, renders less dramatic the problem of reproducibility, since the experimental error gets lost in the residuals together with the error due to approximation of the mathematical model"20.

RESULTS AND DISCUSSION

The use of HPTLC plates for the R_F determination has many advantages such as a higher reproducibility, a shorter analysis time $(5 \text{ min}$ as compared to $20-30 \text{ min}$) due to an acceptable separation over a developing distance of only 5 cm, which will result also in an improved sensitivity $(3-4 \mu g)$ as compared to $5-10 \mu g$ due to the reduced diffusion of the spot. The reproducibility using the corrected R_F values according to the standardization procedure of Stead et al..³ (cf., Experimental) is always $\leq 7\%$ for non-biological samples.

The examination of extracts from biological fluids and tissues or from postmortem samples in various stages of decomposition is complicated by the interference from the biological matrix which alters the values of the chromatographic data²¹. In this case we suggest a preliminary chromatographic purification (and separation) using eluent I.

Thus, after elution, it is possible to scratch the spots from the plate, to separate from the silica gel by extraction with methanol, to concentrate the solution and to perform the TLC analysis for each substance in all four eluent systems. In order to carry out the PCA, the R_F c values were arranged into a matrix (see Table II) with the compounds as "objects" and the eluent mixtures as "variables". Each of the 1448 elements of the matrix is indicated in eqn. 1 as v_{ik} .

The variables $(R_Fc$ values for each eluent mixture) were first autoscaled¹⁹. Each element was multiplied by the weighting typical of the eluent (the reciprocal of the variable standard deviation) in order to give unit variance to each eluent mixture. The weightings for the individual variables I-IV are recorded in Table III.

The PCA of the data matrix gave a model comprising two significant principal components. A third component, still significant according to the cross validation technique¹⁸, was not taken into account because of the small number of original variables. The first component explains 47% of the total variance and the second one a further 26%; the planar model thus accounts for 73% of the total variance. The values of α , β_1 and β_2 are recorded in Table III, while θ_1 and θ_2 values (the "scores" for compounds l-362) are listed in Table II, together with the residual standard deviations, s_k , after two principal components.

In this paper we do not use the refinement procedure based on reweighting

TABLE III

WEIGHTINGS, α , β_1 AND β_2 FOR VARIABLES (ELUENT MIXTURES) I-IV

Variable (eluent mixture)	Weighting	α	ρ,	β_2
	0.0448	2.843	0.576	-0.295
\mathbf{I}	0.0444	1.110	0.420	-0.662
ш	0.0617	0.561	0.434	0.597
IV	0.0359	1.478	0.551	0.344

Fig. 1. Plot of β_2 vs. β_1 for variables (eluents) I-IV; 0 indicates origin (0,0).

with modelling powers as adopted in our previous work. We are aware that any reweighting procedure is somewhat arbitrary, and in view of the slight improvement in the identification ability of the model we now suggest that the analysis be limited to simple $PCA²⁰$.

Fig. 1, a plot of β_2 vs. β_1 , shows that eluents I-IV, which lie along different directions with respect to the origin (O,O), have, in the present instance, different information contents, paralleling the trend already observed for 55 drugs¹².

Fig. 2, a plot of θ_2 vs. θ_1 for the 362 compounds examined, is the basis for the identification of unknowns (see below). However, a careful inspection of this figure provides also interesting insights into the "zones" where substances characterized either by analogous chemical structures or by similar pharmacological activities are grouped. Benzodiazepines are characterized by θ_1 values in the range -0.5 to 3.2 and θ_2 values in the range 0.7-2.1. Their location in the plot depends on the nature of the substituent, especially when attached at the l-position.

A peculiar behaviour is shown by flurazepam (143), which has a much lower θ_2 value, probably due to the presence of a terminal diethylamino group at nitrogen-1. Low θ_2 values are exhibited by a great number of substances (almost all antihistamines and phenothiazinic tranquillisers and many antidepressants) containing similar groups as "predominant" substituents. Phenothiazinic tranquillisers where

Fig. 2. Plot of θ_2 vs. θ_1 for compounds 1-362 (0) and of t_1 and t_2 for pseudo-unknowns X_1-X_{10} (\triangle). For X_5 and

QUALITATIVE ANALYSIS. I. 165

confidence rectangles" are also reported.

the terminal diethylamino group is replaced by an hydroxyl (69, 110, 142, 258 and 259) exhibit much lower θ_1 values, being shifted towards the zone typical of substances with hydroxyl groups.

Compounds having the same base skeleton as morphine (217, 156, 94, 74 and 127) and synthetic derivatives with similar structures such as 176 and 308 lie along the same line which represents the lower left limit of the populated zone of Fig. 2. The position along this line is determined by the presence of one or more hydroxyl groups, which causes a shift toward the left, *i.e.*, lower θ_1 values. The lack of an hydroxyl group or partial or total transformation into the corresponding acetyl derivatives (213, 90, 3) causes a shift towards higher θ_1 values.

The replacement of a nitrogen methyl group in compounds 217, 156 and 176 with an allyl characteristic of antagonists (in 223, 224 and 175 respectively) shifts the latter derivatives towards higher values of both θ_1 and θ_2 .

Identification of unknowns

The identification of unknowns, provided the unknown is one of the 362 compounds in the data set, can be attempted by measuring the corrected $R_F c$ values in the four eluents and fitting them with the PC model.

The t_1 and t_2 values for each unknown are given by eqns. 2 and 3 respectively

$$
t_1 = 0.576 (0.0448 \times 100 R_{FC_1} - 2.843) + 0.420 (0.0444 \times 100 R_{FC_{II}} - 1.110) + 0.434 (0.0617 \times 100 R_{FC_{III}} - 0.561) + (2) + 0.551 (0.0359 \times 100 R_{FC_{IV}} - 1.478)
$$

$$
t_2 = -0.295 (0.0448 \times 100 R_{FC_1} - 2.843) - 0.662 (0.0444 \times 100 R_{FC_{II}} - 1.110) + 0.597 (0.0617 \times 100 R_{FC_{III}} - 0.561) + (3) + 0.344 (0.0359 \times 100 R_{FC_{IV}} - 1.478)
$$

which can easily be simplified to:

$$
t_1 = 0.0258 (100 R_F c_1 - 63.48) + 0.0186 (100 R_F c_{II} - 25) ++ 0.0268 (100 R_F c_{III} - 9.09) + 0.0198 (100 R_F c_{IV} - 41.17) (4)
$$
t_2 = -0.0138 (100 R_F c_1 - 63.48) - 0.0294 (100 R_F c_{II} - 25) ++ 0.0368 (100 R_F c_{III} - 9.09) + 0.0123 (100 R_F c_{IV} - 41.17) (5)
$$
$$

The values for the unknown substance can be fitted into the "scores" plot (Fig. 2) to select the candidates for its identification.

The selection of candidates is done by defining a region of statistical relevance around the t values obtained for the unknown. For this purpose we measured the R_F values for 43 pseudo-unknowns representative of "good" (141) and "bad" (101) ones as well as of compounds which can hardly be distinguished by TLC (19 and 90). For each of the pseudo-unknowns we then determined the differences between their experimental t_1 and t_2 values and their "true" values reported in Table II. The averages

TABLE IV

 R_r x 100 VALUES IN ELUENTS I-IV, $t₁$ AND $t₂$ VALUES AND "CANDIDATES" FOR UNKNOWN SAMPLES $X_1 - X_{10}$

	Unknown $R_{\rm F}c \times 100$			t_{1}	t_{2}	"Candidates" at 99% confidence level	Compound	
		$\boldsymbol{\mathit{II}}$	Ш	IV				
X_1	54	21	0	28	-0.823	-0.254	19, 90, 339, 268, 185, 110	19
X_2	56	6	9	53	-0.314	0.799	256, 47, 51, 96, 20, 147, 109	47
X_3	85	51	5	62	1.342	-0.943	187, 73, 340, 122, 335, 138	73
X_4	75	54	2	20	0.228	-1.526	336, 82, 44, 101, 166, 98, 201, 114, 160, 331, 330, 78	82
X_5	59	20	0	16	-0.950	-0.438	8, 90, 19, 41, 247, 278, 154, 267	90
X_{ϵ}	82	53	$\bf{0}$	25	0.435	-1.601	244, 114, 98, 82, 241, 336, 201, 44, 55, 101, 160, 18	101
\mathbf{X}_2	83	19	45	79	2.104	1.705	141	141
$X_{\rm R}$	86	60	8	55	1.477	-1.197	122, 340, 287, 174, 187	174
$X_{\rm o}$	79	25	$\bf{0}$	44	0.213	-0.505	161, 254, 210, 248, 120	254
X_{10}	79	9	$\bf{0}$	44	-0.084	-0.034	180, 128, 290, 196	290

of these differences are 0.075 in both cases and their standard deviations are 0.050 and 0.058 for t_1 and t_2 respectively. Consequently, by means of appropriate student t values, we can conclude that there is a 95% probability of finding the "true" compound within ± 0.16 t₁ and ± 0.17 t₂ from the position of the unknown on the "scores" plot (Fig. 2) and that this probability is increased to 99% when the interval is ± 0.20 t₁ and ± 0.22 t₂.

A few illustrative examples of identification of unknowns (a complete list of t_1 and t_2 values for all 43 pseudo-unknowns is available on request from the authors) are reported in Table IV, which also lists the possible candidates included in the 99% "confidence rectangle" defined as before. The unknowns are reported as triangles in Fig. 2, which also depicts the 99% "confidence rectangles" for X_5 (90) and X_7 (141). The number of candidates obviously depends upon the number of compounds with similar TLC properties included in the set.

CONCLUSIONS

This work confirms the validity of PCA as a suitable statistical approach for the treatment of TLC data in different eluent systems aimed at the identification of drugs. The application of PCA, which reduces the number of variables, allows a graphical representation of all compounds in a two-dimensional space, *i.e.,* Fig. 2, and represents a great advantage over previous approaches using graphical representations^{1,2} or tables³.

Standardized R_F data in four eluent systems appropriately selected to extract the maximum information available from TLC data¹² are not sufficient to achieve unambiguous identification.

However, in the present instance, the reduction of the range of inquiry to a few candidates is, in the authors' opinion, satisfactory when account is taken of the cost, the time and the analytical instrumentation required by TLC measurements and of the simplicity of the calculations involved. This approach is of great practical importance when financial resources, time and sophisticated analytical instrumentation are not available. Measurements of a different nature, such as gas chromatographic (GC) data, are needed to attempt unambiguous identification of unknowns. Further work on the application of PCA to both TLC and GC data is in progress.

ACKNOWLEDGEMENT

We thank Ministero della Pubblica Istruzione for financial support of a research project on chemometrics.

REFERENCES

- 1 J. Franc and Z. Stransky, Collect. *Czech. Chem. Commun., 24 (1959) 3611.*
- *2* V. Milana, G. Roman0 and G. Scarlata, *Minerva* Medicoleg., 99 (1979) 15.
- 3 A. H. Stead, R. Gill, T. Wright, J. P. Gibbs and A. C. Moffat, Analyst *(London), 107 (1982)* 1106; and references cited therein.
- 4 A. C. Moffat and K. W. Smalldon, J. *Chromatogr., 90 (1974) 9.*
- 5 P. Owen, A. Pendlebury and A. C. Moffat, J. Chromatogr., 161 (1978) 187.
- 6 P. Owen, A. Pendlebury and A. C. Moffat, J. *Chromatogr., 161 (1978) 195.*
- *7* D. L. Massart, J. *Chromafogr., 79 (1973) 157.*
- *8* D. L. Massart and H. De Clercq, *Anal. Chem., 46 (1974) 1988.*
- *9* H. De Clercq and D. L. Massart, J. *Chromatogr.,* 115 (1975) 1.
- 10 G. Musumarra, G. Scarlata, G. Roman0 and S. Clementi, J. *Anal. Toxicol., 7 (1983) 286.*
- 11 G. Musumarra, G. Scarlata, G. Romano, S. Clementi and S. Weld, J. *Chromatogr. Sci., 22 (1984) 538.*
- *12 G.* Musumarra, G. Scarlata, G. Cirma, G. Romano, S. Palazzo, S. Clementi and G. Giulietti, J. *Chromatogr., 295 (1984) 31.*
- *13* E. G. C. Clarke, *Isolation and Identification of Drugs,* Vol. 1, The Pharmaceutical Press, London, 1978.
- 14 *The Merck Index,* Merck and Co. Inc., Rahway, NJ, 10th ed., 1983.
- 15 J. P. Franke, P. Schepers, J. Bosman and R. A. de Zeeuw, J. *Anal. Toxicol., 6 (1982) 131.*
- 16 S. Wold and M. Sjöström, *ACS. Symp. Ser.*, 52 (1977) 243.
- *17 C.* Albano, G. Blomquist, D. Coomans, W. J. Dunn III, U. Edlund, B. Eliasson, S. Hellberg, E. Johansson, B. Norden, M. Sjöström, B. Sördeström, H. Wold and S. Wold, in A. Hoskuldson, K. Esbensen *etal.* (Editors), *Proceedings Symposium i avent statistik, Kipenhanm, January 22nd, 1981,* NEUCC, RECKAV and RECKV, Copenhagen, 1981, p. 183.
- 18 S. Wold, *Technometrics, 20 (1978) 397.*
- *19 S.* Weld, in D. L. Massart, A. Dijkstra and L. Kaufman (Editors), *Evaluation and Optimization oj Laboratory Methods and Analytical Procedure,* Elsevier, Amsterdam, 1978, pp. 385, 423.
- 20 G. Musumarra, G. Scarlata, G. Roman0 and S. Clementi, *J. Anal. Toxicol., 9 (1985) 187.*
- *21* M. Bogusz, M. Klys, J. Wijsbeek, J. P. Franke and A. R. de Zeeuw, *J. Anal. Toxicol., 8 (1984) 149.*