Journal of Chromatography, 115 (1975) 293–298

CHROM. 8605

THE APPLICATION OF CHROMATOGRAPHIC DATA TO COMMODITY SOURCE IDENTIFICATION

A. M. C. DAVIES*

Research Department, Reckitt and Colman Food Division, Carrow, Norwich NR1 2DD (Great Britain)
(First received March 23rd, 1975; revised manuscript received July 7th, 1975)

SUMMARY

An empirical method of assigning samples to determined groups of any particular commodity, within the population of that commodity, is described.

A computer program is used to calculate and select ratios of variables to characterise groups from known members of the groups. A second program is used to test if an unknown sample is a member of a previously characterised group. The method is tested using published chromatographic data for spearmint oils and animal fats.

INTRODUCTION

The general problem of establishing the identity of a sample can be divided into three catagories: the unity, finite and infinite types. In the unity type it is assumed that all samples are different, an example being human fingerprints. In the finite type the sample is known to belong to one of a few defined groups, while in the infinite type the sample belongs to one of a large number of groups, some of which are defined and some are undefined. The finite type involves the allocation of the sample to one of the known groups while the infinite type involves testing whether the characteristics of the sample are consistent with any of the known groups.

In the unity type of problem any marked difference between the unknown sample and the reference sample is sufficient to disprove identity and the main problem is the classification of a large number of samples. An example of this type of problem is the identification of oil spills[1].

The method of discriminant analysis has been successfully applied to the finite type of problem. Powers and Keith[2] classified four lots of coffee, Hartmann and Hawkes[3] determined the geographical origin of peppermint oils from ten sources and Kawahara *et al.*[4] classified samples of asphalt and heavy residual oil. An example of the infinite type of problem is the detection of adulteration of Israel lemon oil studied by Lifshitz *et al.*[5]. This is considered to be an infinite problem because although

* Present address: Agricultural Research Council's Food Research Institute, Colney Lane, Norwich NR4 7UA, Great Britain.

there is only one defined group —genuine Israel lemon oil— there are an unlimited number of adulterated groups.

In order to proceed with a project to replace the classical use of pollen analysis for identifying the source of a honey it has been found necessary to devise a new method of classification. The method was designed to allow the application of amino acid analysis data to this infinite type of problem. Tests of the method using relatively simple gas–liquid chromatographic (GLC) data from spearmint oils and animal fats indicate that it could have application to both finite and infinite types of problems using data of greater complexity.

METHOD

The chromatographically determined variables were used to calculate ratios which were then used as the variables for the test. This approach was used because of the success of other workers using ratios of variables in preference to the actual variables[2–4]. Biggers *et al.*[6] found that the two types of coffee could be distinguished by the use of ratios of GLC peak areas but not by the intensity of single peaks or the combination of single peak areas. A further advantage of using ratios is that they are less affected by instrumental errors than peak areas. Powers and Keith[2] recommended that the use of reciprocal ratios in discriminant analysis should be avoided. The use of reciprocal ratios in the proposed test was avoided by numbering the input variables and only using those ratios in which the numerator variable number is less than the denominator variable number.

An identity group was characterised by taking values of peaks from several samples known to belong to the group. Any peak absent from a sample was assigned a value equal to the smallest detectable amount of that peak. All possible peak ratios, excluding reciprocals were calculated and then the average value, standard deviation and coefficient of variation for each ratio was determined. A specific number of ratios (<100) having the lowest coefficients of variation were selected as characterising the group. These calculations were repeated for each identity group under consideration so that each group was characterised by a list of defined ratios. A computer program (SOCIT) was written in Honeywell FORTRAN F to carry out the calculations and selections*.

To determine if an unknown sample belonged to any of the known groups, a test based on the Camp–Meidell inequality[7] was applied to the values of the same ratios. The purpose of this test was to try to disprove the hypothesis that the sample is a member of the group under consideration. Failure to reveal a difference which was significant at the 10–20% level of significance allowed the conclusion that the sample was consistent with the group. A significance level below 5% was regarded as proving the sample was not a member of the group. A second computer program (ITUSOC) was used to make the required calculations*.

Standardised ratio differences were obtained, for each of the selected ratios, by calculating the difference between the group mean for a ratio and the same ratio in the unknown sample and dividing by the standard deviation for that ratio. According to the Camp–Meidell inequality the probability of these values exceeding a value t

* Program listings are available from the author.

TABLE I

np VALUES FOR VARIOUS VALUES OF P AND C

Example of use of Table I (for $n = 60$). To obtain the value of t which 4 ratio differences (C) will exceed with a probability (P) not exceeding 1%. For $C = 4$ and $\%P = 1$, $np = 1.3$. Hence $t = \sqrt{(4/9) \cdot (60/1.3)} = 4.5$. Thus if 5 out of 60 standardised ratio differences were observed to exceed 4.5 then it could be stated that there is less than a 1% chance of the sample belonging to the group being tested.

%P	C				
	5	*4*	*3*	*2*	*1*
20	3.9	3.1	2.3	1.52	0.82
10	3.2	2.4	1.75	1.10	0.53
5	2.6	1.9	1.35	0.80	0.36
1	1.8	1.3	0.80	0.44	0.15
0.1	1.1	0.75	0.42	0.19	0.044

did not exceed $4/(9t^2)$ if the unknown sample was a member of that group. Thus, if a probability level percentage was fixed, values of t could be calculated which no more than 1, 2, 3, 4, 5 ratio differences should exceed. These values were calculated as follows: Let C = a specific number of differences exceeding t, n = number of ratios considered, p = proportion of differences exceeding t in universe to result in more than C exceeding t in sample of n with a probability not exceeding P%. Values of the product np were obtained for each value of C and probability P from a graph of the cumulative binomial distribution. Table I gives values of np for the levels of P and C which were used in the present work. t was calculated from the equation, $t = \sqrt{4/9 \cdot n/np}$.

The lowest value of P at which a C value was exceeded was taken as the level of significance of the difference between the test sample and the group. The test was repeated on the unknown sample for each group under consideration.

TABLE II

TESTS ON SPEARMINT VARIETIES; PROBABILITY LEVELS AT WHICH TEST WAS SIGNIFICANT

Samples tested, 11; correctly identified, 10; not identified, 1; incorrectly identified, 0.

Sample	*M. spicata*	*M. cardiaca*	*Identity*
SP16*	20	0.1	*M. spicata*
E0557	20	0.1	*M. spicata*
SP17	0.1	0.1	*M. spicata*
E0565	20	0.1	*M. spicata*
JCB	20	0.1	*M. spicata*
E0254*	0.1	20	*M. cardiaca*
SP13	0.1	20	*M. cardiaca*
E0739	0.1	20	*M. cardiaca*
E0237	0.1	20	*M. cardiaca*
E0567	0.1	20	*M. cardiaca*
E0570	0.1	20	*M. cardiaca*

* Sample used in characterising program.

APPLICATION OF THE METHOD

The method has been tested using relatively simple published data on spearmint oil and fatty acid analysis of animal fats. These data were chosen because the authors had shown that it could be used to distinguish between the different types of samples.

Spearmint oil

The data used are from a paper by Smith *et al.*[8]. Ten samples of *Mentha spicata* and thirteen samples of *Mentha cardiaca* were randomly selected for use in the SOCIT program. There were 16 variables giving 120 possible ratios. The program was set to select 60 ratios having the smallest coefficients of variation. The remaining four samples of *M. spicata* and five samples of *M. cardiaca* plus one sample of each variety used in the SOCIT program were tested against the data from the SOCIT program, using the ITUSOC program. The results are shown in Table II. This table shows that ten out of eleven samples were assigned the correct identity and one sample was unidentified.

Animal fatty acids

The data used are from a paper by Hubbard and Pocklington[9] which presented

TABLE III

TESTS ON ANIMAL FAT TYPES; PROBABILITY LEVELS AT WHICH TEST WAS SIGNIFICANT

Key to abbreviations for locations of cuts: LG, leg; R, rump; FL, foreloin; HL, hindloin; K, kidney; L, loin; C, cod; BK, brisket; F, flare; HD, head.

Sample	*Beef LG*	*Beef R*	*Pork FL*	*Pork HL*	*Lamb*
Beef LG ARG 2	20	1	0.1	0.1	0.1
Beef LG ARG 3	20	1	0.1	0.1	0.1
Beef R ENG 3	1	20	0.1	0.1	0.1
Beef R RHOD 2	0.1	0.1	0.1	0.1	0.1
Pork FL SWED 2	0.1	0.1	20	20	0.1
Pork HL ENG 3	0.1	0.1	20	20	0.1
Lamb LG ARG 2	0.1	0.1	0.1	0.1	20
Lamb K NZ 2	0.1	10	0.1	0.1	15
Beef K UkG 1	0.1	20	0.1	0.1	0.1
Beef L URG 1	1	20	0.1	0.1	0.1
Beef C EIRE	20	20	0.1	0.1	0.1
Beef BK SCOT	15	20	0.1	0.1	0.1
Pork F CAN 2	0.1	0.1	10	1	0.1
Pork F YUG 1	0.1	0.1	20	0.1	0.1
Pork F YUG 2	0.1	0.1	20	1	0.1
Pork F N. IR 1	0.1	0.1	20	1	0.1
Pork HD N. IR 1	0.1	0.1	20	20	0.1
Pork L N. IR 1	0.1	0.1	20	20	0.1
Chicken	0.1	0.1	0.1	0.1	0.1
Turkey	0.1	0.1	0.1	0.1	0.1
Duck	0.1	0.1	0.1	0.1	0.1
Rabbit	0.1	0.1	0.1	0.1	0.1

TABLE IV

SUMMARY OF TESTS ON ANIMAL FATS

Test	*Number of samples*	*Results of tests*			
		Identified		*Not identified*	
		Correctly	*Incorrectly*	*Correctly*	*Incorrectly*
Types	22	17	0	4	1
Cuts	16*	5	16	10	1

* Since each cut was tested twice there were 32 tests.

fatty acid analysis data for different cuts of beef, pork and lamb. Two cuts of beef and pork were chosen for testing and a mixed selection of lamb cuts were used. Not less than eight samples of each cut were used for characterising the cuts. There were sixteen fatty acids which appeared regularly giving 120 possible ratios. The SOCIT program was set to select 60 ratios. Tests were carried out using the selected ratios on further samples of the same cuts and also on different cuts of the same type of meat and on different types of meat. The results of these tests are shown in Tables III and IV.

DISCUSSION

Although the method has been shown to give good results with two different types of data and has been successfully applied to the problem of honey identification[10] it must be considered as an empirical method. The reason for this is that the method assumes that the ratios are independent variables. In a limited test of the data, the extent of correlation between the ratios was not large, but this might not be the case with other data. Another limitation of the method is that it assumes that the distributions of standardised ratios are symmetrical although the distributions of ratios are inherently skew. However, the distributions of standardised ratios in the groups so far encountered have been sufficiently symmetrical to have reasonable confidence in the validity of the Camp–Meidell inequality.

At the present time the number of ratios selected has been arbitrarily determined. When the number was increased from 60 to 100 no improvement of the animal cut results was achieved. It would be appropriate to determine the number of ratios by limiting the value of the coefficient of variation, which in these tests had a maximum value of 100%.

The spearmint oil data are examples of the finite type of problem because there are only two varieties of spearmint. Although it is likely that the problem could also be solved by discriminant analysis, the method used does not require the use of a large computer demanded by the Biomedical Computer Program[2,4]. The results of the test on animal fats types is considered satisfactory, only one sample (beef R Rhod 2), which should have been identified, failed to be identified. It seems reasonable to take the results of the test on Lamb K NZ 2 as indicating the sample was lamb and not beef rump. The four unidentifiable samples were convincingly distinguished from the groups under consideration. The results on cuts of beef and pork were unsatisfactory and the test could not be recommended for distinguishing between cuts. If the cut-type tests had been successful different selections of ratios would have been made for distinguishing between different animal types.

Lifshitz *et al.*[5] used a Chi-square test for detecting adulteration of Israeli lemon oil. The variables for the equation were chosen, from a number of variables, by consideration of the coefficients of variation and correlation coefficients of the variables. The application of this approach to infinite problems containing several known groups would be more difficult than the proposed method. The Chi-square test is also of limited value in considering large numbers of variables because the individual values of each variable are not taken into account.

Most infinite and finite types of identity problems contain several known groups. Chromatographic methods of analysis have made it possible to obtain readily data from many more variables than can be utilised by traditional classification methods. Despite the limitations imposed by the empirical nature of the proposed method it could be applicable to a number of this type of problem.

ACKNOWLEDGEMENTS

The author is grateful to Mr. A. D. Deans for statistical advice, Dr. D. G. Land for helpful discussion and the Directors of Reckitt and Colman for permission to publish this paper.

REFERENCES

1 J. S. Mattson, *Anal. Chem.*, 43 (1971) 1872.
2 J. J. Powers and E. S. Keith, *J. Food Sci.*, 33 (1968) 207.
3 N. Hartmann and S. J. Hawkes, *J. Chromatogr. Sci.*, 8 (1970) 610.
4 F. K. Kawahara, J. F. Santner and E. C. Julian, *Anal. Chem.*, 46 (1974) 266.
5 A. Lifshitz, Y. Stepak and H. B. Basker, *J. Food Sci.*, 34 (1969) 254.
6 R. E. Biggers, J. J. Hilton and M. A. Gianturco, *J. Chromatogr. Sci.*, 7 (1969) 453.
7 M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms*, Oliver and Boyd, Edinburgh, 1971.
8 D. M. Smith, W. Skakum and L. Levi, *J. Agr. Food Chem.*, 11 (1963) 268.
9 A. W. Hubbard and W. D. Pocklington, *J. Sci. Food Agric.*, 19 (1968) 571.
10 A. M. C. Davies, unpublished results.