ELSEVIER

Thermochimica Acta 284 (1996) 103–108

thermochimica acta

Country of origin of peanuts: a comparison of statistical software for discriminant analysis of DSC results[1]

Susan M. Dyszel*

U.S. Customs Service, Washington, DC 20229, USA

Abstract

As personal computers (PC's) have become more powerful, sophisticated data analysis programs have evolved for this platform. SAS/STAT for Personal Computers, a derivative of the mainframe application, was leased by the Customs Research Laboratory to do canonical and discriminant analysis in November 1988. In January 1995, we acquired the Unistat 3.0 package. There are obvious differences in the ease of use of the two packages. This paper will discuss the analytical differences that were observed by processing a single data set in both packages.

Implications of using this type of data treatment on measurements from thermal analysis experiments will be discussed, along with the underlying analytical questions posed by the sample set. A basic question will be addressed: does the choice of the software affect the ultimate results of the analysis?

Keywords: Country of origin; DSC; Peanuts; Statistical analysis

1. Introduction

Discriminant and canonical analyses have been demonstrated to be very powerful for the classification and grouping of large data sets. Specifically, this laboratory has developed the use of discriminant and canonical analysis to determine the country of origin of natural products such as orange juice, macadamia nuts and peanuts by trace elements [1, 2], and tropical oils and coffee beans by DSC [3, 4].

With the evolution of powerful personal computers, data analysis software migrated from the large mainframe. Software was made "user-friendly" through a graphical user

* Correspondance address: US Customs Service, 610 S. Canal Street, Suite 850 Chicago, IL 60607, USA.

[1] Presented at the 24th North American Thermal Analysis Society Conference, San Francisco, CA, U.S.A., 10–13 September 1995

PII: S0040-6031(96)02861-4

interface, allowing for easier data movement from spreadsheet applications. "Sneaker-net" was replaced with a few clicks of a mouse. However, the improved ease of application does not remove the analyst's obligation to understand the data analysis system and its effect on data. Using software as a "black box" into which one puts data, turns the crank and gets results, may be fraught with serious pitfalls. In addition, it is possible that different conclusions may be drawn from the same data if different underlying equations in the statistical work-up are used. This paper will explore the use of two software packages in the analysis of DSC data for the determination of the country of origin of peanuts.

2. Experimental

Twenty-six peanut samples were obtained from the USDA in August 1994 and an additional five from U.S. Customs field locations in 1995. There were four countries of origin represented in the sample set: Argentina, China, USA and Mexico. With the exception of one of the Mexican samples obtained by the USDA, the origin of the Mexican peanut samples was only tentatively identified. The thirty-one samples were crushed by mortar and pestle before extraction with petroleum ether. The DSC runs used the oil that remained after the evaporation of the ether. Two of the USDA samples were extracted in duplicate, once when the USDA samples were received and once when the Customs samples were received. A minimum of three specimens from each peanut oil was run on the DSC. A single drop of the peanut oil was dispensed by glass pipette into a tared aluminum pan, sealed and weighed. Specimen weight ranged from 5 to 10 mg depending on how large a drop was dispensed.

The DSC run profile consisted of placing the specimen pan into the DSC sample chamber held at 233 K and waiting for the specimen to reach temperature equilibrium, approximately four minutes. The temperature was ramped from 233 to 325 K at 5 K min^{-1}, held for 5 min and cooled back to 233 K at 5 K min^{-1}. Curves for both heating and cooling were recorded and saved. The calibration of the DSC was verified using indium and methyl stearate.

The DSC runs were conducted on a PE DSC-2 (Perkin–Elmer, Norwalk, CT), with a PE TADS data station running standard DSC and Partial Areas software. The resulting data were tabulated in the Excel 5.0 spreadsheet (Microsoft, Redmond, WA). Either Unistat 3.0 for Windows (Megalon, Novato, CA) or SAS/STAT for Personal Computers (SAS Institute Inc., Cary, NC) was used for the discriminant and canonical analysis.

3. Discussion

To facilitate the following discussion, a common vocabulary describing the data set is necessary. A single specimen may have one or more DSC runs yielding data points of various types. These are measured experimental values such as temperature and heats of transition. Secondary or derived values, such as partial areas, can also be obtained.

Each of the experimental values belongs to a data variable. At the end of the experimental work, the values are tabulated in a spreadsheet matrix with rows of data, listed by specimen, arranged by the data variables in columns. The use of discriminant and canonical analysis enables the analyst to compare or consider more than just a few variables obtained from a single data run or combination of data runs. The question of which experimental data variables are significant may also be determined through these data manipulations.

The original data matrix has fifteen data variables: T_{max}, ΔHf_h, partial areas and area slice temperatures for the first three peaks in the heating profile, partial areas for heating curve peaks at temperatures of 250, 260, 270 and 280 K, and from the cooling

Table 1
Country of origin misclassifications

Specimen	from	to	SAS probability	from	to	Unistat probability
Argentina						
a 11	a	u	0.7994	a	u	0.7554
a 12	a	u	0.7653	a	u	0.3337
a 13	a	u	0.6369	a	u	0.2321
a 22	a	m	0.5295	a	m	0.3000
a 36	a	m	0.4810	a	m	0.1589
USA						
u 10	u	m	0.5306	u	m	0.6660
u 11	u	m	0.6076	u	m	0.6609
u 13	u	m	0.5670	u	m	0.6613
u 14	u	a	0.5546	u	a	0.5132
u 16	u	a	0.9941	u	a	0.0024
u 19	u	m	0.5199	u	m	0.3434
u 25	u	c	0.5404	u	c	0.0414
u 27	u	m	0.6154	u	m	0.3202
u 28	u	c	0.4546	u	c	0.8379
u 31	u	m	0.5573	u	m	0.7107
u 38	u	m	0.4714	u	m	0.7488
u 43	u	c	0.7059	u	c	0.3138
u 44	u	c	0.5438	u	c	0.0672
China						
c 2	c	u	0.6102	c	u	0.5944
c 3	c	u	0.6325	c	u	0.5529
c 4	c	u	0.6384	c	u	0.7151
c 8	c	u	0.8790	c	u	0.2654
c 17	c	u	0.5819	c	u	0.6208
Mexico						
m 7	m	c	0.8049	m	c	0.6819
m 8	m	c	0.9193	m	c	0.6390
m 10	m	a	0.7982	m	a	0.9016
m 11	m	a	0.9105	m	a	0.9941
m 12	m	a	0.7612	m	a	0.5420

curve, T_{min}, T_{onset} and ΔHf_c. Two means of selecting the data variables for the data analysis for the peanut samples were used: an intuitive approach based on the selection of data variables which favored the measured values, such as temperatures or heats of transition, and a mathematical approach that used a backwards elimination procedure with stringent criteria to determine the variables needed for the "best model."

The Excel spreadsheet contained the following column headings: code, id, CO, sample #, T_{max}, ΔHf_h, ta1, pa1, ta2, pa2, ta3, pa3, a250, a260, a270, a280, T_{min}, T_{onset}, and

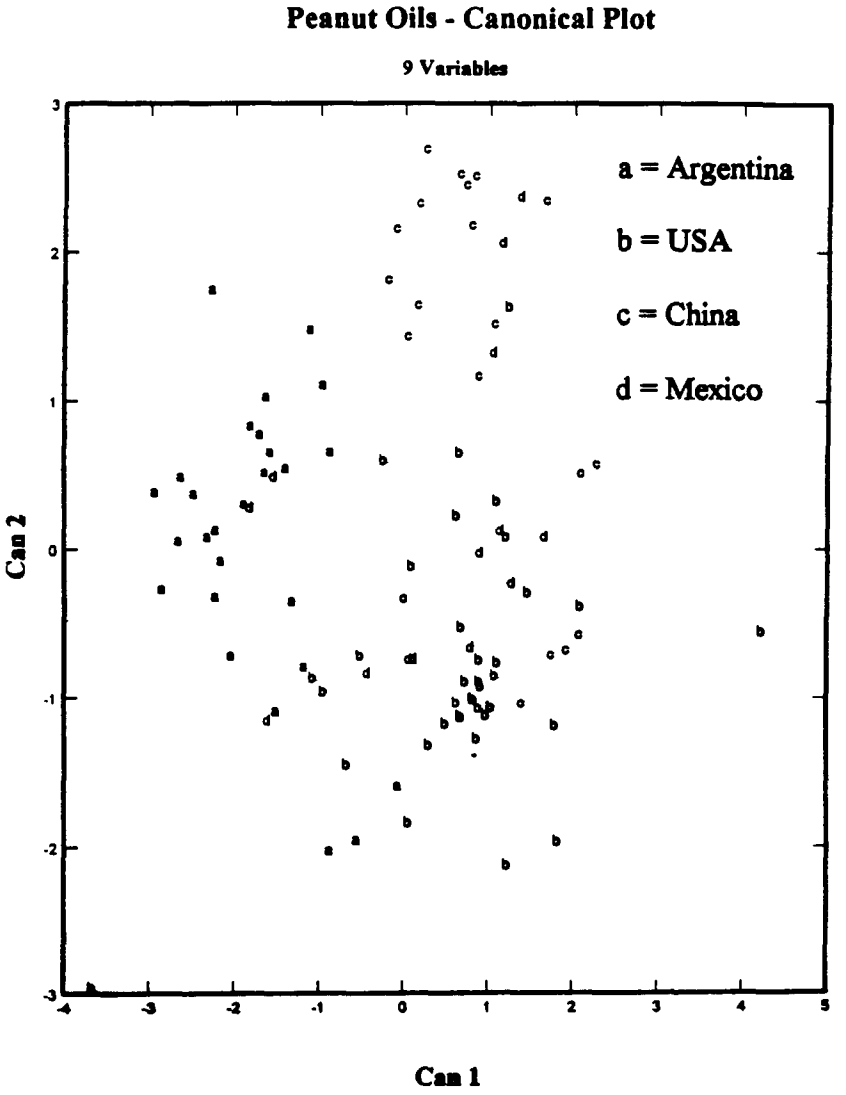

Fig. 1. Canonical plot for the nine-variable data sets from Unistat.

ΔHf_c. Unistat, using the default configuration, will only accept 100 rows and 18 columns in its data processor, so the data set was limited to 99 of the 128 specimens. Three specimens from each sample were included in the final data set. In addition, the sample # and third peak temperature and partial area (ta3, pa3) were deleted. The same 99-row data set was then used for the SAS analysis.

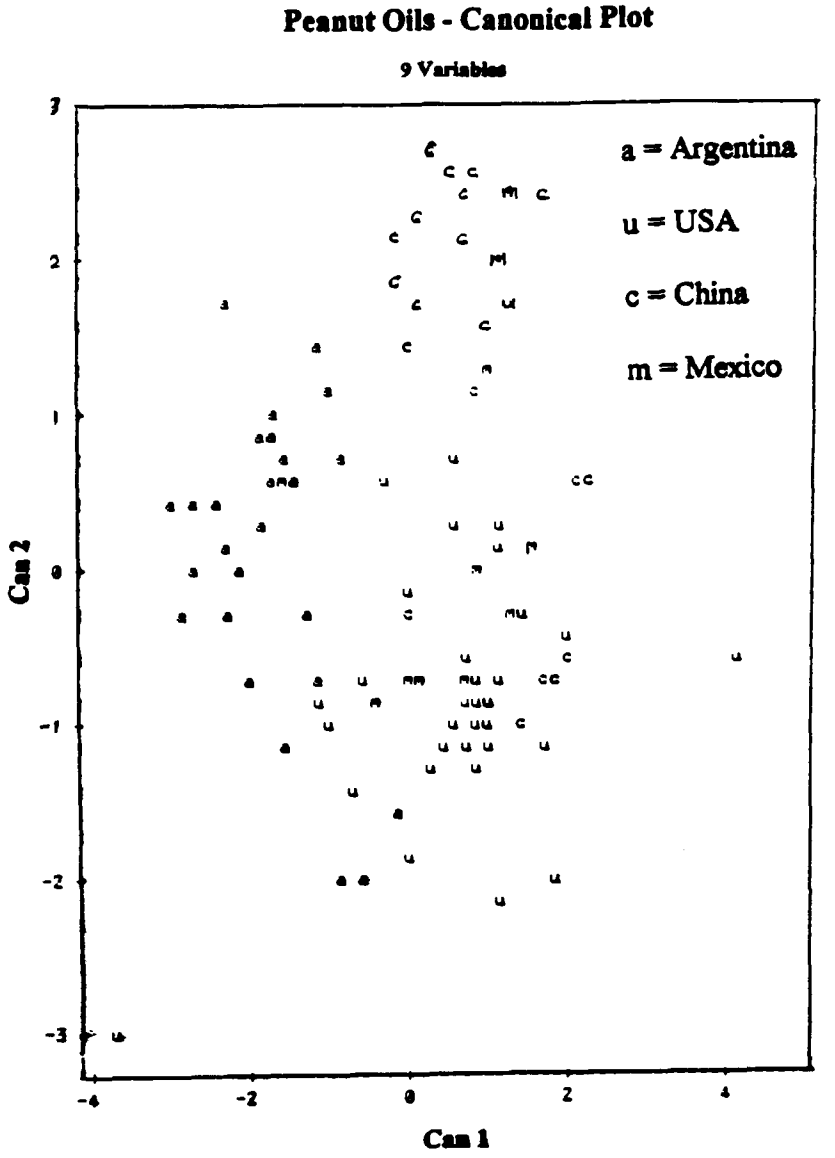

Fig. 2. Canonical plot for the nine-variable data sets from SAS.

Data analysis in SAS included the nine-variable list, the five-variable list, and the three-variable list. The three-variable list, was determined by the backward elimination procedure using a sensitivity level of 0.01 (tightly defined model). As above, the sample statistics, the probablity of membership and canonical plots were calculated and printed.

Data analysis by Unistat was conducted in four passes: first, using all thirteen variables; second, omitting pal, tal, pa2, ta2 (nine-variable list); third, using the five-variable list (a250, a260, a270, and a280 are deleted) and, fourth, using the three-variable list of only ΔHf_h, a280 and ΔHf_c. Sample statistics, membership classification and canonical plots were calculated and printed.

The resulting canonical plots and resubstitution classification tables were then ready for comparison. Table 1 compares the probability of classification by SAS and Unistat for those specimens which were misclassified. While the same specimens were misclassified, the probability of membership did not always agree, sometimes significantly. Fig. 1 and 2 are the canonical plots for the nine-variable data set from Unistat and SAS respectively. Allowing for slight differences in the size and relative scaling of the plots, the results are identical. In addition, a manual comparison of results indicated that, for the nine-variable set, the classification by origin was also the same.

4. Conclusion

We attempted to answer the question: Would two different software packages produce the same answers from a single data set? This experimental series showed that, given the same data set, two different statistical packages for discriminant analysis and canonicl analysis gave virtually identical results. This would indicate that this form of data analysis is dependent on the data set rather than on the mathematical computations performed. Thermal data performed in two different laboratories, analyzed by this type of data treatment, should be comparable to the limits set by the differences in the initial data collection.

Acknowledgements

I would like to thank Shantae Allen, an ACS Project Seed student during the summer of 1994, for her assistance on the DSC work for this project. Mention of specific products does not constitute endorsement by the U.S. Customs Service.

References

[1] R.S. Schwartz and L.T. Hecking, Customs Laboratory Bulletin, 1 (2) (1989) 3.
[2] R.S. Schwartz and L.T. Hecking, Customs Laboratory Bulletin, 2 (2) (1990) 51.
[3] S.M. Dyszel and S.K. Baish, Thermochim. Acta, 212 (1992) 39.
[4] S.M. Dyszel, in G.K.R. Williams (Ed.), Proceedings of the 22nd NATAS Conference, Denver, CO, 1993, p. 446.