ORIGINAL PAPER

C. A. Andersson · L. Munck · R. Henrion · G. Henrion

Analysis of *N*-dimensional data arrays from fluorescence spectroscopy of an intermediary sugar product

Received: 13 January 1997 / Revised: 18 March 1997 / Accepted: 25 March 1997

Abstract Unwanted formation of colour takes place during the production of crystalline sugar. The degree of colouration depends partly on the necessary processing conditions, e.g. heating and pH, and partly on the initial composition and condition of the sugar beets used as raw material. Reducing sugars are formed during the process. These are reactive compounds forming a variety of coloured complexes and strong precursors to further formation of colour and many of these compounds contain fluorophores. In the present work it is discussed if spectrofluorometric screening of intermediary sugar products prior to the final heating stages combined with a multi-way chemometric approach can provide information that significantly reflects the condition of the process and the beets. The model used is the *N*-way PCA (Principal Component Analysis) which is an exploratory model, not necessitating explicit modelling of single parameters nor any assumptions towards parameter interaction. By use of a 4-way PCA of order (3,2,3,3) satisfactory classification of 47 thick juice samples belonging to 5 factories has been obtained from a spectrofluorometric screening method. Also, a temporal trend has been found to evolve during the time of production. The investigation substantiates the use of modern models from data analysis for extracting significant information from large and complex data sets.

Dedicated to Professor Dr. Gerhard Werner on the occasion of his 65th birthday

C. A. Andersson · L. Munck
Chemometrics Group, Food Technology,
Royal Veterinary and Agricultural University, Rolighedsvej 30,
DK-1958 Frederiksberg, Denmark

R. Henrion
Weierstrass-Institute of Applied Analysis and Stochastics,
Mohrenstrasse 39, D-10117 Berlin, Germany

G. Henrion
Institute of Chemistry, Humboldt University of Berlin,
Hessische Strasse 1–2, D-10115 Berlin, Germany

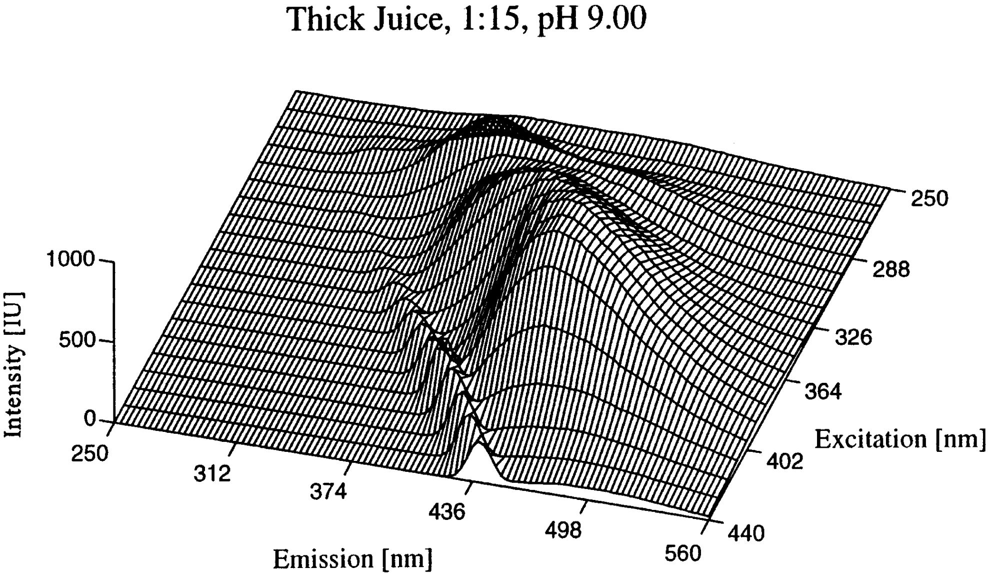
1 Sugar production

In northern Europe the most important source of sucrose for the production of crystalline sugar is the sugar beet, *Beta Vulgaris*. Harvesting of sugar beets and, immediately following, production of sugar is concentrated to a yearly period of approx. 4 months. This period is called the campaign and runs typically from October to January. During the campaign the factories continuously receive beets from many different beet farmers. Due to premises of growing, e.g. fall of rain, frost, soil characteristics, fertilizer type and harvesting machinery, there is a high variation between the truckloads delivered by the farmers. A consequence of this variation is that the parameters for the chemical unit processes are difficult to control with regard to securing a white and uniform final product (see [1] for an overview of the process). The quality class of the sugar is determined according to European standards in which colouration is a main parameter. The classification influences the price at which the product can be sold, hence there is a strong economical motivation for minimizing the formation of colour during the process. Chemometrics has successfully been applied to the prediction of selected quality parameters in sugar [2].

A spectrofluorometrically based screening method has been applied to samples taken weekly of a preliminary sugar product, *thick juice*. Data from this screening have been explored with multi-way, multivariate chemometric methods.

2 Experimental

Fluorescence intensity landscapes, or excitation-emission matrices, have been measured on 47 thick juice samples from the 1994 campaign. Five factories have contributed thick juice samples. Each sample has volumetrically been diluted 1:15 and 1:150 with NH_4Cl pH 9.00 buffer in doubly ion-exchanged and Si-free water. The buffer was made only once. Both of the dilutions were measured using 20 excitation wavelengths (250 nm–440 nm, 10 nm intervals) and 311 emission wavelengths (250 nm–560 nm, 1 nm intervals). Two typical landscapes for one sample are shown in Fig. 1.



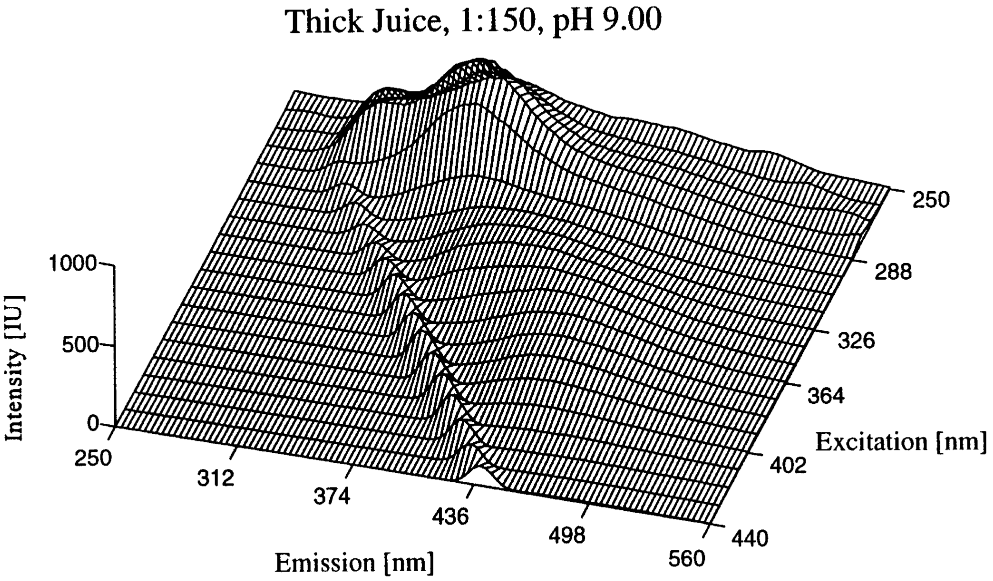


Fig. 1 Two fluorescence landscapes – one for each dilution – are measured per thick juice sample

Note that the peaks in the ultraviolet do not decrease with dilution, this is caused by concentration quenching, or inner-absorption effect, see [3]. At the excitation and emission sides 10 nm slits were used. The instrument was the Perkin Elmer LS50B spectrofluorometer. As indicated by Fig. 1, the combination of a narrow emission slit width and generally low turbidity allows neglecting the Rayleigh scattering. The 47 samples were measured in arbitrary order.

3 Analysis of *N*-way data arrays

Each intensity measurement in the collected data depends on four external parameters; the sample number (47 samples), the concentration (two levels of dilution), the detection wavelength (311 emission wavelengths) and the excitation wavelength (20 excitation wavelengths). Hence, the intensities measured constitute a 4-way data array of order (47,2,311,20).

Various models exist for analyzing three-way data sets, see [4]. In the present work we focus on the *N*-way principal component analysis (*N*-way PCA) which is a generalization of the 3-way Tucker3 [5] model to *N*-way data arrays. Taking a starting point in the 3-way case, Fig. 2 provides a basis for presenting the *N*-way PCA. The 3-way PCA model of a 3-way data array **X** of order r_1, r_2, r_3) is depicted in the figure. The array is decomposed into a significant systematic part and a non-significant residual depicted by **E**. The systematic part is described by orthogonal factors which are stored columnwise in matrices

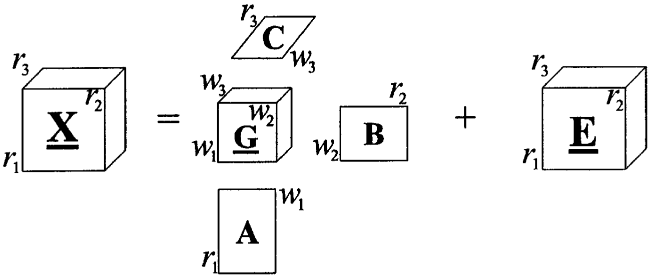


Fig. 2 The three-way principal component analysis (PCA) model

Table 1 Sum-of-squares explained by PCA models of different orders

Model order	Expl. SS [%]	Num. Par.
(1,1,1,1)	74.13	384
(2,1,2,2)	82.88	772
(2,2,2,2)	92.08	782
(3,2,3,3)	96.25	1201
(3,3,3,3)	96.24	1230
(4,2,4,4)	97.85	1656

A (r_1, w_1), **B** (r_2, w_2) and **C** (r_3, w_3). The number of factors in each of the three ways, i.e. w_1, w_2 and w_3, must be determined by the analyst from *a priori* knowledge about **X** or by evaluating models with different combinations of w_1, w_2 and w_3, choosing the order that gives the most accurate model of **X**. The array **G** of order (w_1, w_2, w_3), referred to as the core array, allows the factors to interact in the model of **X**. Interaction of factors is not encountered in conventional, i.e. bilinear, PCA but is only feasible for $N \geq 3$. After having estimated the orthogonal factors and the core array the squared entries in the core express how significant the factor combinations are for the model. The 4-way PCA can be conceived as an extension of the decomposition illustrated in Fig. 2 with a necessary introduction of an additional set of factors, **D**, and by extending **X** (r_1, r_2, r_3, r_4), **G** (w_1, w_2, w_3, w_4) and **E** (r_1, r_2, r_3, r_4) to be 4-way structures. The general *N*-way PCA may be formulated according to (1).

$$\mathrm{vec}\,\mathbf{X} \approx (\mathbf{A}_1 \otimes \cdots \otimes \mathbf{A}_N)\,\mathrm{vec}\,\mathbf{C} \qquad (1)$$

In (1) **X** represents the *N*-way data array of order (n_1, $\cdots$, n_N) and $\mathbf{A_i}$ (n_i, w_i) is the orthogonal component matrix belonging to the *i*th way. The array **C** of order (w_1, $\cdots$,w_N) designates the core array. ⊗ represents the Kronecker product. For details of the general *N*-way model the reader is referred to [6]. A tutorial on *N*-way PCA is given in [7]. A common algorithm calculating component matrices and core array from the data array in (1), is described in [8].

Factors from *N*-way PCA suffer from rotational ambiguity, i.e. the *N*-way PCA of **X** has an infinity of factors and cores, where one solution can be rotated into another giving the exact same fit to **X**. Returning to the exploratory power of the squared elements of the core, one can perform controlled transformations of a solution to give a core where only a few squared entries are significant, see [9]. Having only a limited number of significant core entries allows the analyst to focus on a few combinations of more significant and general factors. In contrast, having no significant combinations of factors, interpretation is rendered impossible due to the high number of non-significant factors that must be evaluated.

4 Principal component analysis of the 4-way data array

In order to find the optimal order of the 4-way PCA model, several combinations of different orders were investigated. Table 1 shows the relative increase in explained sum-of-squares (SS) as the orders of the models increase. The total number of parameters is shown in the rightmost column of Table 1. The findings from this table suggest that a model of order (3,2,3,3) should be chosen since 96.25% of SS explained seems appropriate in comparison with the models of higher orders. Also, the number of parameters should be kept as low as possible in accordance with the principle of parsimony. Parsimonious models involve as few parameters as possible, hence the risk for fitting non-systematic trends (noise) in **X** is minimized. Note, that the model does not improve in fit when using more than two factors in the second dimension, this is in concordance with the number of observations: One cannot derive three or more orthogonal solutions in a dimension that is only spanned by two variables.

In order to improve the interpretability of the (54 elements large) core array, the solution was transformed to yield maximum variance-of squares of the core as proposed in [9]. By transformation the variance-of-squares of the core array, which is an indicator of how few significant entries are present in the core, changed from 4.11×10^{20} to 5.46×10^{20}, i.e. an increase of 32%. The resulting profiles are plotted in Fig. 3A–D. Inspection of the variance-of-squares maximized core elements yields (with the involved factors of the four modes in parentheses) 2.36×10^{10} (1,1,1,1), 1.73×10^9 (1,1,2,2) 9.50×10^8 (1,2,1,3), 1.49×10^8 (1,2,2,3) and 1.03×10^8 (1,2,1,2). For convenience of the reader the squared elements of the core have been sorted, and the values of the 5 largest entries are depicted in Fig. 4. From this figure it is clear that the combination indicated by (1,1,1,1), being the first sample profile, first dilution factor and the first excitation and emission profiles, is most important in the model of **X**. Therefore we shall initially concentrate on explaining these factors since they are most general. In Fig. 3A, profile 1 shows that the main variation between samples is caused by two levels of the fluorescence intensities. In Fig. 3A the samples are arranged factory-wise in ascending week number such that samples 1–10 are from factory a, 11–18 from b, 19–28 from d, 29–36 from e and 37–47 from f. Hence, we may conclude that the samples from the last factory (number 37–47) generally have lower levels of intensity. Similarly, the major trend in the data is that the fluorescence intensities descend when the samples are diluted. This is deduced from Fig. 3B since the factor de-

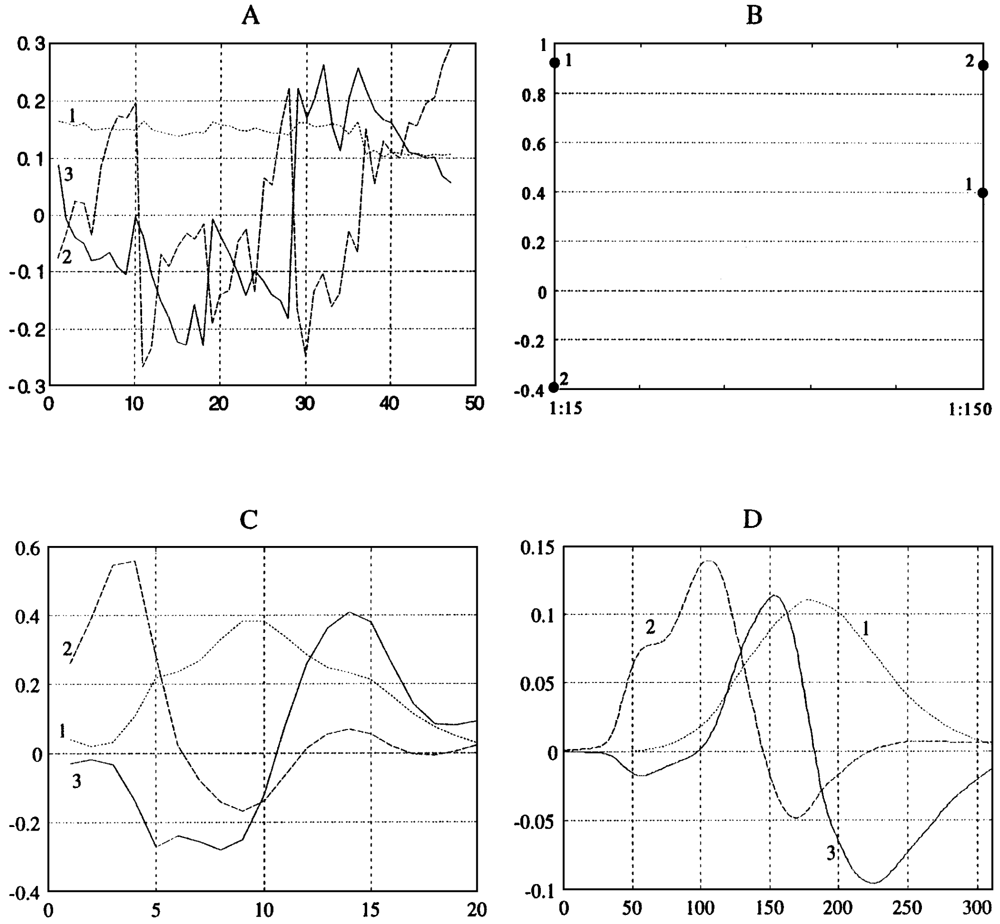


Fig. 3A–D The rotated factors from PCA on the 4-way data set. The sample profiles are shown in **A**. Emission and excitation profiles are shown in **C** and **D**, respectively. The factors explaining the variation caused by dilution are illustrated in **B**. *1–3* see text

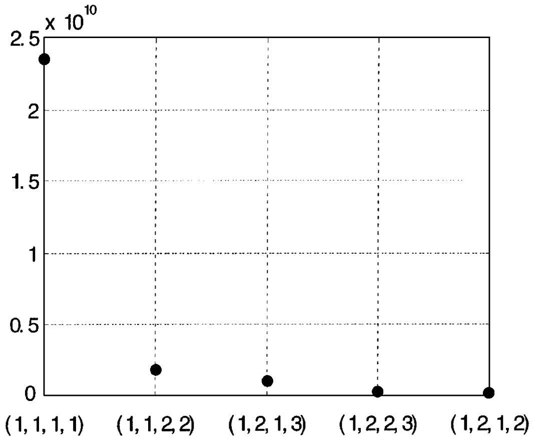


Fig. 4 The 5 largest squared elements of the core array. The remaining 49 elements are in the same range, or lower, than the lowest two elements shown here. Hence, the three most significant factor combinations are (1,1,1,1), (1,1,2,2) and (1,2,1,3)

creases from 0.93 to 0.40 upon dilution. The reason that this is not true for all samples, as indicated by factor two, may be due to concentration quenching, i.e. that the intensity does not decrease with dilution from 1:15 to 1:150. The spectral excitation and emission profiles marked 1 in Fig. 3C–D give indications to the profiles of the fluorophores being common to the samples.

Bearing in mind that the samples are ordered factory-wise after increasing week number, we return to the sample profiles in Fig. 3A. Sample profile number 2 appears to reflect time-dependent events since the level generally increases as the week of sampling increases. There is a shift in this temporal development going from sample 28 to 29, corresponding to going from factory d to e. Also sample profile number 3 appears to reflect intensities that are inversely related to the week number, albeit, this trend is not as obvious as in the case of profile 2. Additionally, the profiles not only reflect time dependences but also give rough indications of different levels for the factories. We have chosen to extend Fig. 3A with a scoreplot where sample profiles 2 and 3 are plotted against each other, as shown in Fig. 5. This plot fully exploits the information in the two profiles as discussed above by combining the trends from two independent factors in one plot. The relationship among the samples becomes clear since samples from the same factories are grouped almost without overlaps. Furthermore, these two factors reveal a development in time, that is, there is a trend in the plot that the samples are dispersed within the clusters according to the time of sampling (as indicated by the inserted arrow). Hence, sample profiles 2 and 3 contain fluorometric information that describes the temporal behaviour of the thick juices as the campaign runs. Also, plots of sample profiles 1 vs. 2 and 3 have been investigated, but as indi-

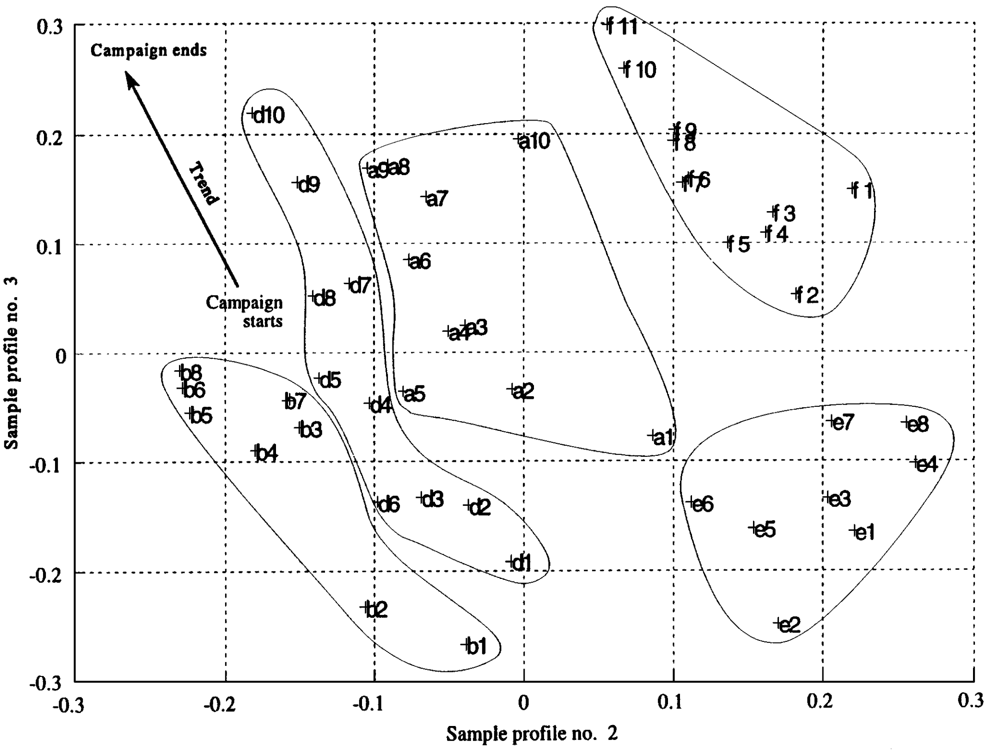


Fig. 5 A scoreplot combining the information in sample profiles no. 2 and 3. The letters a, b, d, e and f each relate to a factory and the numbers designate the week of sampling. This plot reveals two important trends in the fluorescence data: Grouping according to factory and a development in time

cated by sample profile 1 in Fig. 3A, this factor contains only very general information that cannot reveal detailed differences between neither time nor factory among the samples.

5 Results

Explorative soft modelling, *in casu* 4-way PCA, has substantiated the use of spectrofluorometry as a screening method. By showing that the collected 4-way data array cannot only classify samples according to factories, but also give an indication of temporal conditions, fluorometry gives promise as a very relevant source of information that is related to variations in the raw beets and the state of the factory as well. Without explicit modelling of the many uncontrollable parameters (some being difficult to asses or quantify, e.g. growing conditions and weather conditions) causing the differences between samples, the results from the 4-way PCA has proven that spectrofluorometric measurements give promise as an important screening method for process control. By temporal characterizing of the thick juice, the process control will be able to adjust conditions accordingly. On the basis of the presented results a project has been initiated aiming at developing a spectrofluorometer for in-line screening. This will improve our understanding of the relation between measured fluorescence signals and the extent of colouration. The data analytical part of the project will include extensive use of chemometric multi-way models, as the one presented.

Acknowledgments We wish to thank Danisco Sugar for providing the thick juice samples. C.A. and L.M. are grateful to the Affluence AIR2-CT94-1416 project fund and the Føtek-2 grant “Online”. The present work is a result of cooperation between the universities HUB, Berlin and KVL, Copenhagen.

References

1. Andersson CA (1995) Flow injection analysis, fluorometry and chemometrics. Master thesis. Royal Veterinary and Agricultural University, Food Technology, Copenhagen
2. Nørgaard L (1995) Zuckerindustrie 120:970–981
3. Schulman SG (1977) Fluorescence and Phosphorescence Spectroscopy: Physicochemical Principles and Practice. Pergamon Press, Oxford
4. Kroonenberg PM (1992) Stat App 4:619–633
5. Tucker L (1966) Psychometrika 31:279–311
6. Magnus JR, Neudecker N (1988) Matrix differential calculus with applications in statistics and econometrics. Wiley, Chichester
7. Henrion R (1994) Chemom Intell Lab Syst 25:1–23
8. Kroonenberg PM, De Leeuw J (1980) Psychometrika 45: 69–97
9. Henrion R, Andersson CA (1997) J Chemometrics (submitted)