ELSEVIER

Chemometrics and Intelligent Laboratory Systems 44 (1998) 31–60

Chemometrics and intelligent laboratory systems

Chemometrics in food science—a demonstration of the feasibility of a highly exploratory, inductive evaluation strategy of fundamental scientific significance

L. Munck *, L. Nørgaard, S.B. Engelsen, R. Bro, C.A. Andersson

Chemometrics Group, Food Technology, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

Abstract

At the roots of science lies observation and data collection from the world as is and from which conclusions can be induced after classification. This is far from the present theory-driven, deductive, normative stage of science which depends heavily on modelling discrete functional factors in laboratory experiments and suppresses the aspect of interaction. In spite of its successes, science today has great difficulty in adapting to the changes which technology has created to cope with registering and evaluating real data from the world, such as in food production chains. This paper demonstrates that it is possible and profitable with the help of new technology to reintroduce an explorative, inductive strategy to investigate the chemistry of a complex food process as is with a minimum of a priori assumptions. The food process investigated is a sugar plant and the tools necessary in this strategy include a multivariate screening method (fluorescence spectroscopy), an arsenal of chemometric models (PCA, PLS, principal variables), including multiway models (PARAFAC, Tucker), and a computer. Not only can chemical criteria and process parameters throughout the process be validly predicted by the screening method, but process irregularities as well as chemical species can also be detected and validated by multiway chemometric techniques. Inspired by examples from the food area, the paper further discusses the nature of the exploration method in the selection of tools and data. The aim is to study complex processes as a whole in order to model interaction of the underlying latent functional factors which may later be defined more precisely by deductive methods. These methods in combination with an appropriate multivariate screening method allow for unique identification of objects—a significant prerequisite for a viable, exploratory, inductive data strategy which is needed as a fundamental complement to prevalent normative research in order to obtain a science on the interdisciplinary level.

Keywords: Chemometrics; Food science; Multiway models

″. . . mathematics is bound to become an increasingly experimental science with less of a claim to absolute truth″

Gregory Chaitin [1]

1. The need for a new multivariate approach in interdisciplinary evaluation

The food and health area receives special attention from the public in the present accelerated change driven by technology. Chemistry and chemical data play decisive roles here. Classical basic research based on laboratory experimentation has made ap-

* Corresponding author.

0169-7439/98/$ - see front matter
PII: S0169-7439(98)00074-4

parent a wide range of natural and manmade chemical species which appear as functional and antifunctional factors in food science and nutrition. Food science is thus, today, in the very centre of the scientific cyclone, drawing on a wealth of disciplines from chemistry and physics [2,3], mathematics and statistics [4], to biology, genetics, medicine, microbiology [5], agriculture, technology and environmental science, and even further to the cognitive sciences like sensory [6] and consumer analysis and psychology as well as to other social disciplines like economy. Such an elaborate web of contacts increases the need for the establishment of basic principles for intercontextual multivariate data communication which are necessary tools to create a real science on the interdisciplinary level. Chemometrics might help here.

The present rapid change is supported but not primarily driven by science. Instead, inventors mainly outside the universities develop technology to advance to the forefront with a much more flexible operational strategy than science. The technologists are focusing on finding a surprising technological fix that is visible and attractive to the consumer and which thus can secure a market. Science often comes long afterwards and explains why technology works and what side effects it has by studying interferences to present hypotheses.

During the Second World War, the organisation of technological product development and the supporting science became much more effective, as vividly described in the classic OECD report by Erich Jantsch in 1967 [7]. The aim of the development outlined by Jantsch is essentially to 'invent the future' by technological forecasting, which Jantsch describes as a management discipline systematically exploiting goal-oriented science in order to realize technology or, in other words, to achieve technological transfer with a high degree of probability.

Exploratory technological forecasting starts by pragmatically evaluating the present knowledge base and is directed towards the future, while normative technological forecasting first defines a future goal or model by evaluating needs, wishes and possibilities and works backwards toward the present in order to realize it. In classical science, these two outlooks are related to inductive and deductive problem-solving, respectively [8]. When technology and science were young, they worked in an inductive, exploratory way, for example to describe, classify and utilize the chemical compounds which were isolated by distillation, precipitation and crystallization and analyzed by their colour, smell, taste, solubility and reactivity. The patterns of relationships which could be induced from the information from these early studies inspired a theoretical model thinking in formulating general hypotheses from which new, specific and detailed principles and new, confirmative experiments could be deduced [9]. Thus, in food science and related industry, data evaluation today is primarily performed by classical statistical [10] and hard engineering methods [11] based on distributional assumptions and solution of complex differential equations, which were necessary before the advent of the computer. These methods are, however, only relevant for a part of real life where the sufficient causal understanding is already available and underlying assumptions fit, such as in representative sampling techniques, and on the molecular level when, e.g., modelling heat-transfer in food processes.

Before the advent of the computer, the necessary strategy to cope with issues in the multivariate complex world was through problem reduction. The different functional factors in the laboratory were isolated one by one at the expense of control of covariance and overview. Data are still evaluated by a mathematical language based on axioms which are more tuned to the logic of the mathematical machinery than to that of chemistry and the world outside the laboratory. Therefore, the present crisis in today's science is rooted in a lack of an accepted strategy in interdisciplinary science, despite the political quest for such a cooperation. We maintain that in the science of the future new strategies and data, analytical algorithms and procedures will play a fundamental role in creating a dialogue on equal terms between the normative, deductive and the exploratory, inductive principles. We will now focus on an example of how the computer, a specific screening method and a range of chemometric tools mostly funded on vector algebra adapted from mathematical methods of social science [12] may be used by the human brain to upgrade the exploratory, inductive research method which is greatly neglected today. Hempel [8] explains the current attitude: ''Scientific knowledge is not obtained by the method of induction based on earlier collection of data but rather by 'The hypothe-

sis method': that is, to invent alternative hypotheses deduced from earlier known knowledge as preliminary answers of the problem under study and thereafter testing these hypotheses empirically''.

2. Exploring the beet sugar manufacturing process by spectrofluorometry and chemometrics—an example of a highly exploratory, inductive research strategy

We will begin by presenting the sequence of chemometric results of the exploratory investigation expressed as a graphic interface which is easily cognitively accessible for any person. In Appendices A–C, we will comment in more detail on how we use the chemometric machinery involved, with emphasis on the new multiway techniques.

Sugar or sucrose [13] is the most abundant disaccharide in nature and has been a world leading commodity for centuries mainly due to its sweet taste properties. Originally, it was extracted from sugar canes but today more than half of the world production comes from sugar beets. Sugar is probably the most chemically pure food component produced with a typical purity of 99.999%. Colour and purity play a great role when evaluating sugar quality.

In 1992, we heard from a sugar production expert that UV-lamps and filters were used in Denmark during the war for visual classification of sugar according to purity. There was a typical blue fluorescence for less pure sugars. With our background in fluorescence analysis in foods [14], but without any in-depth knowledge of sugar production, we contacted and established a dialogue with the Danish company Danisco Sugar. We started by analyzing samples which we knew nothing about in our Perkin Elmer LS50B spectrofluorometer. After presenting the results to the sugar technologists, we obtained successively more information about process conditions and about chemical analyses of the products for interpretation which we included in our chemometric models. The measurement conditions are described or referred to in the text of the figures and tables.

In Fig. 1A, we see the complex fluorescence spectra, each with 1023 data points from 34 different sugar samples from the year 1993. In order to get an overview of this complex information, we performed a data reduction by principal component analysis (PCA) to reduce the data to a few (three) principal components (PCs).

The PCA score plot in Fig. 1B (PC#1 vs. PC#3) reveals 3 clusters which the sugar technologists identified as average weekly samples from the sugar campaign (production period) from week 1 to 14 for 3 sugar factories called A, B and C. The different raw material and processing conditions of the different factories in 1993 obviously had a unique fluorescence signature.

We then obtained 10 kinds of univariate chemical analyses for each of the 34 samples which are presented as spectra after scaling in Fig. 1C. We performed a separate PCA score analysis of the chemical data which also revealed 3 clusters (Fig. 1D) corresponding to 3 factories and similar to the spectrofluorometric investigation (Fig. 1B). When combining loadings and scores for the chemical analyses in a bi-plot (Fig. 1E), we could see that ash, colour and amino–N analyses are situated in the same area as samples from factory C which indicates that these have especially high values. Because the independent classification based on fluorescence data (Fig. 1B) indicates that factory C is especially high in fluorescence, we induced the hypothesis that fluorescence might be directly or indirectly related to some of the chemical analyses. In order to test this, we performed a partial least squares (PLS) regression analysis on the 34 samples correlating whole fluorescence spectra with ash. The result reveals a significant correlation coefficient of -0.92, which indicates that fluorescence analysis could be a candidate as a screening method for quality in sugar production.

This indication is further verified in a PLS study [15] with 81 whole fluorescence spectra from 6 different factories showing especially high correlations with amino–N, ash and colour (Table 1). Five wavelengths were selected by the principal variables method (see Appendix C) which altogether gave reasonable prediction models with amino–N, colour and ash, indicating that an 'on-line' screening method could be devised based on a simple filter instrument.

When a PCA was performed on fluorescence information of mean weekly sugar samples during the campaign for one factory, a horseshoe-formed time

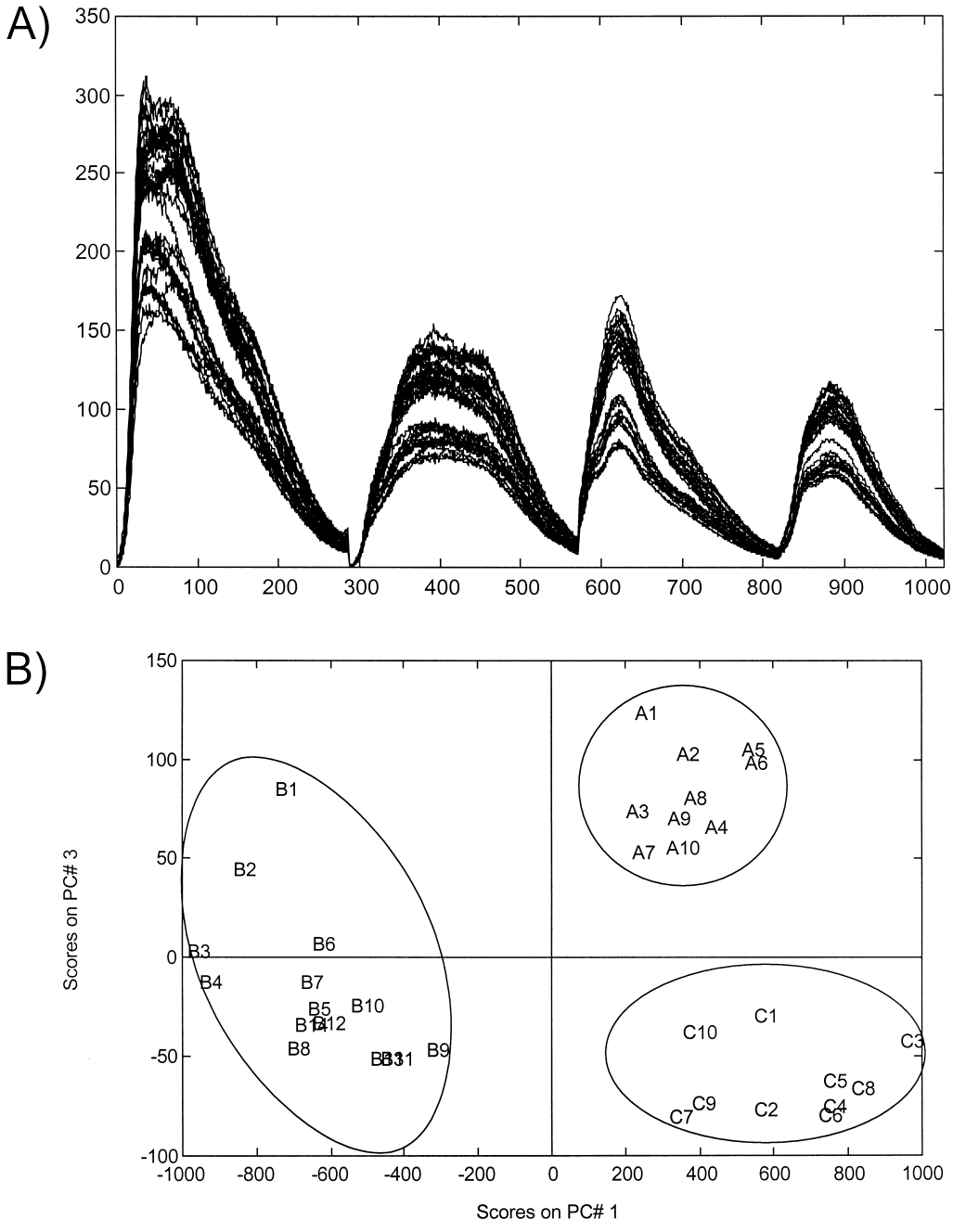
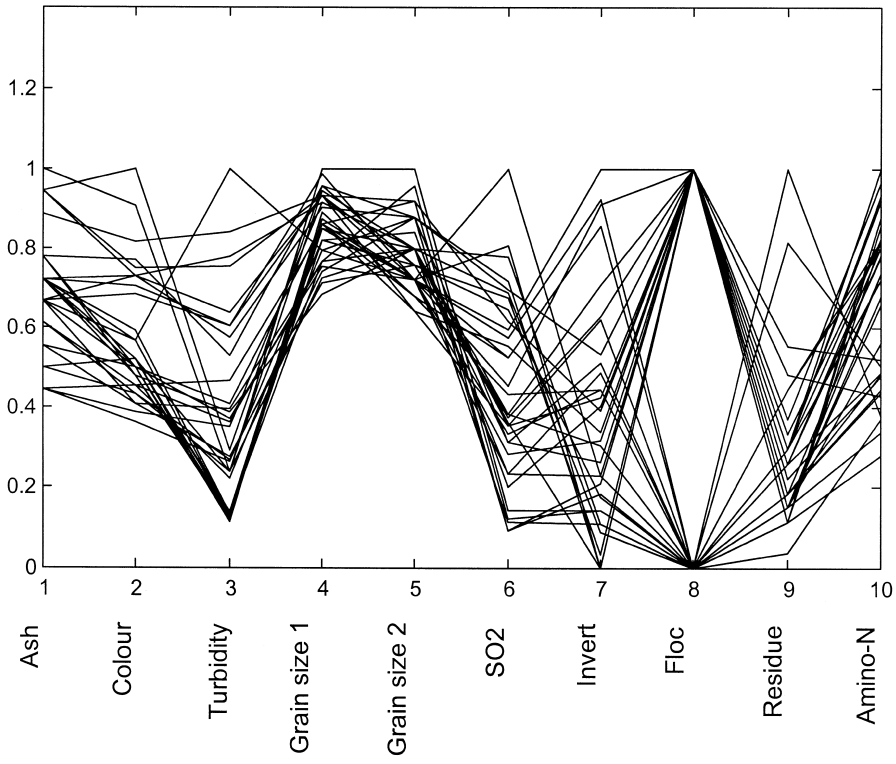


Fig. 1. (A) Uncorrected fluorescence emission spectra of 34 sugar samples. The spectra are recorded from a solution of sugar in water at excitation wavelengths 230 nm, 240 nm, 290 nm and 340 nm. The emission ranges sampled with 1 nm intervals are 275–560 nm, 275–560 nm, 311–560 nm, and 361–560 nm, respectively (in total 1023 data points). See Ref. [15] for further details. (B) A score plot from a PCA on the spectra; three clusters are seen corresponding to samples from three different factories (A, B, and C). (C) Chemical data on the same 34 samples (scaled to a maximum value of 1). (D) Score plot from a PCA on the chemical data; again three clusters are seen corresponding to samples from three different factories. (E) Bi-plot based on chemical data.

C)



D)

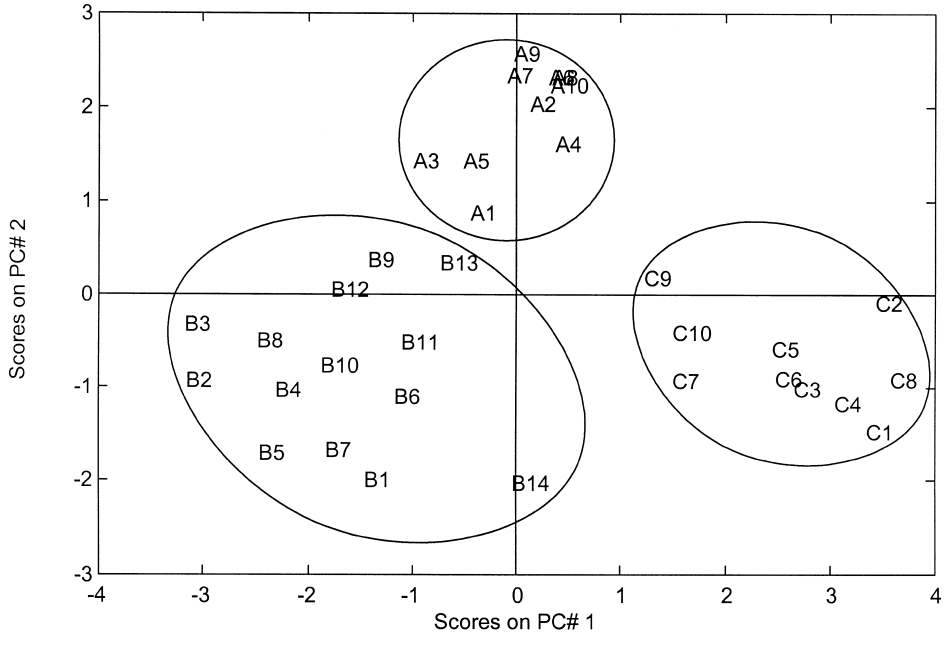


Fig. 1 (continued).

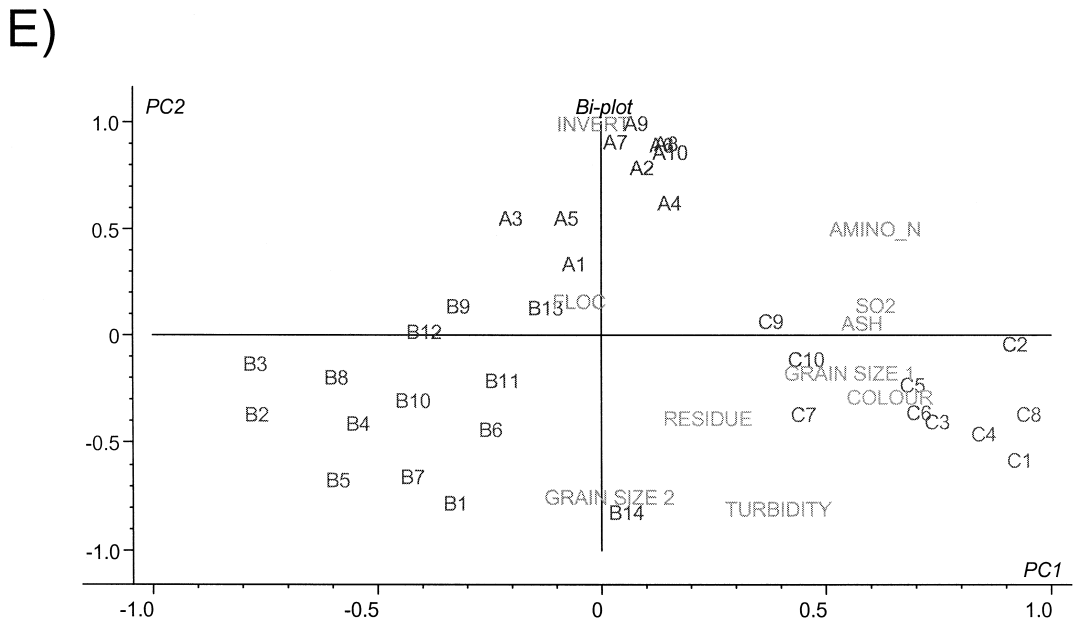


Fig. 1 (continued).

trend could be envisaged for some factories (Fig. 2A), but not for others (Fig. 2B) which where rather chaotic. These two extreme PCA score plots selected from six factories with data from 1993 were described by the sugar engineers as their best and worst functioning factories. The trend in the PCA analysis of spectra in Fig. 2A tentatively represents changes in beet raw material chemistry due to growing conditions, age, climate and storage and the resulting adjustments in process technology.

A similar PCA score plot of sugar fluorescence information from the campaign start of the best functioning sugar factory in 1994 is displayed in Fig. 2C. A total of 106 sugar samples were taken during the first three days of the sugar campaign. The PCA score plot representing these sugar spectra starts at the bottom with samples 3, 2, 5, 6, 4, 9, 8, 7, moving upwards to the right, then straight to the left and ending up in an area of balance from score -50 to score $+50$ of PC2. At the same time, the number of significant principal components diminishes from 4–5 to 1–2, indicating normal operating conditions. However, in the area of relative balance we can still envisage in a local PCA (Fig. 2D) a segregation in two sample clusters 40–74 and 75–106, indicating a fundamental change in the process conditions after sample 74. This change could be identified in the factory records as a process breakdown. Sample 88 is an unexplained outlier. We conclude that it would be worthwhile to investigate whether the fluorescence information could be used to assist the process engineer in indicating the balance of the process in the form of PCA graphics.

We now move upwards in the process chain from sugar to analyze thick juice—an important unpure intermediate product in sugar production. In an earlier preliminary study on thick juice [15], we obtained results similar to those as with sugar with regard to fluorescence analysis, however less clear cut, in the classification of factories and correlation to chemical analyses. We then employed a more advanced analysis than two-way PCA, namely 4-way Tucker [16,17], which is explained in more detail in Appendix A. Undiluted thick juice does not display fluorescence due to concentration quenching. It is possible to 'develop' fluorescence information by dilution. By simultaneously using fluorescence landscapes for partially quenched (1:15 Fig. 3A) and unquenched dilutions (1:150 Fig. 3B) we obtain four external parameters with 47 samples, two levels of dilution, 20 excitation wavelengths and 311 emission wavelengths constituting a 4-way data array of order

Table 1

(A) Full spectrum prediction errors for sugar samples (dissolved in water)[a]

	Mean	Range	# PC's	RMSEP[b]	*r*
Amino–N (ppm)	2.631	0.28–4.91	1	0.314	0.96
Colour	21.8	11–44	5	2.4	0.94
Ash %	0.0110	0.004–0.017	3	0.0012	0.91
SO_2 (ppm)	4.16	0.8–8.2	3	1.08	0.85
Invert (ppm)	36.8	0–92	3	17.6	0.74
Turbidity	0.498	0.19–1.30	4	0.204	0.72

(B) Prediction results for colour, ash, and amino–N based on five excitation–emission wavelengths pairs selected by the principal variables algorithm[cd]

	Mean	Range	# PC's	RMSEP[b]	*r*
Amino–N	2.631	0.28–4.91	1	0.280	0.96
Colour	20.9	11–34	5	2.6	0.90
Ash%	0.011	0.004–0.017	3	0.0013	0.91

[a]All models are PLS1 models [15].
[b]Root mean square error of prediction.
[c]The excitation (nm)/emission (nm) wavelengths used for prediction were 230/361, 230/310, 230/333, 230/454 and 340/419.
[d]See Appendix C.

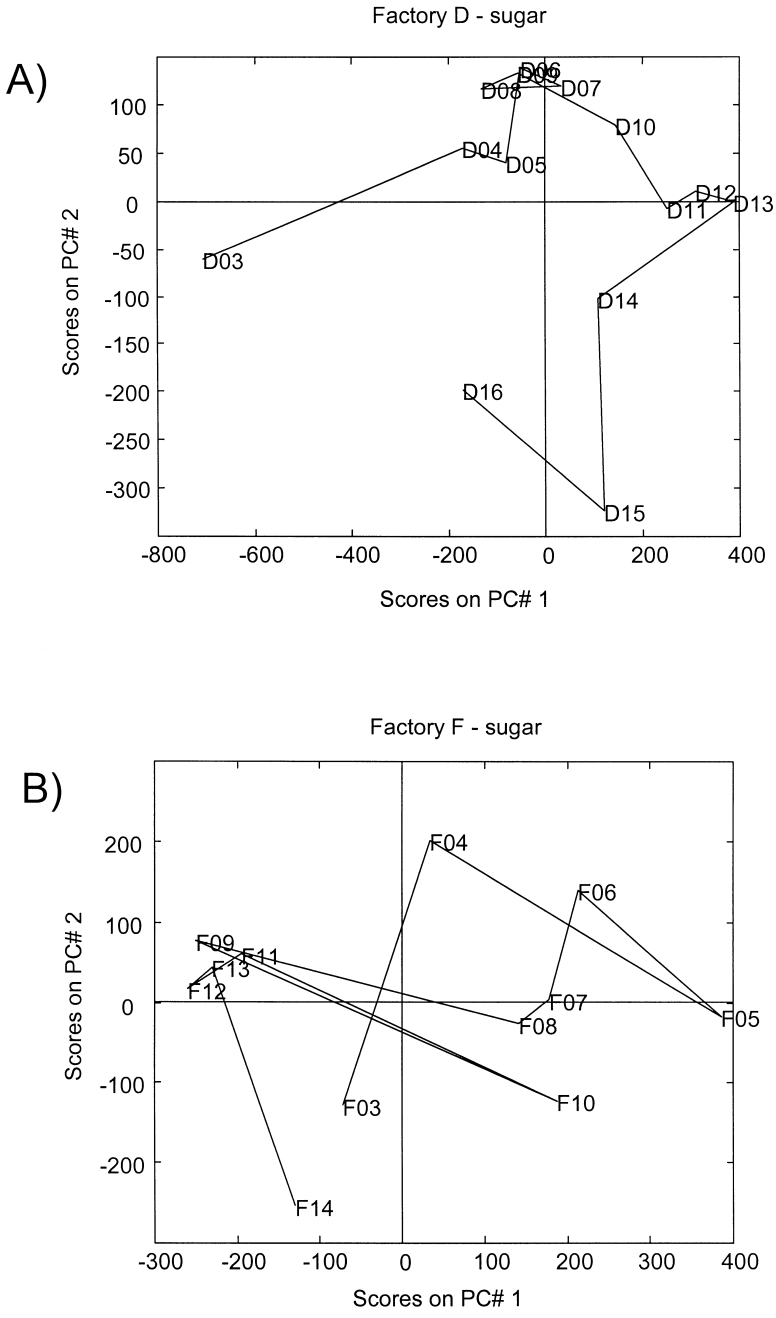


Fig. 2. (A) and (B) Score plots from a PCA of fluorescence spectra recorded on weekly collected samples from two factories. Factory D (A) was known to be the best functioning factory, while factory F (B) was known to be the worst functioning factory. (C) Score plot from a PCA of fluorescence spectra recorded on 106 sugar samples from the first three days of operation in a given sugar factory. (D) Score plot of a PCA on the last 87 samples. The numbering is chronological.

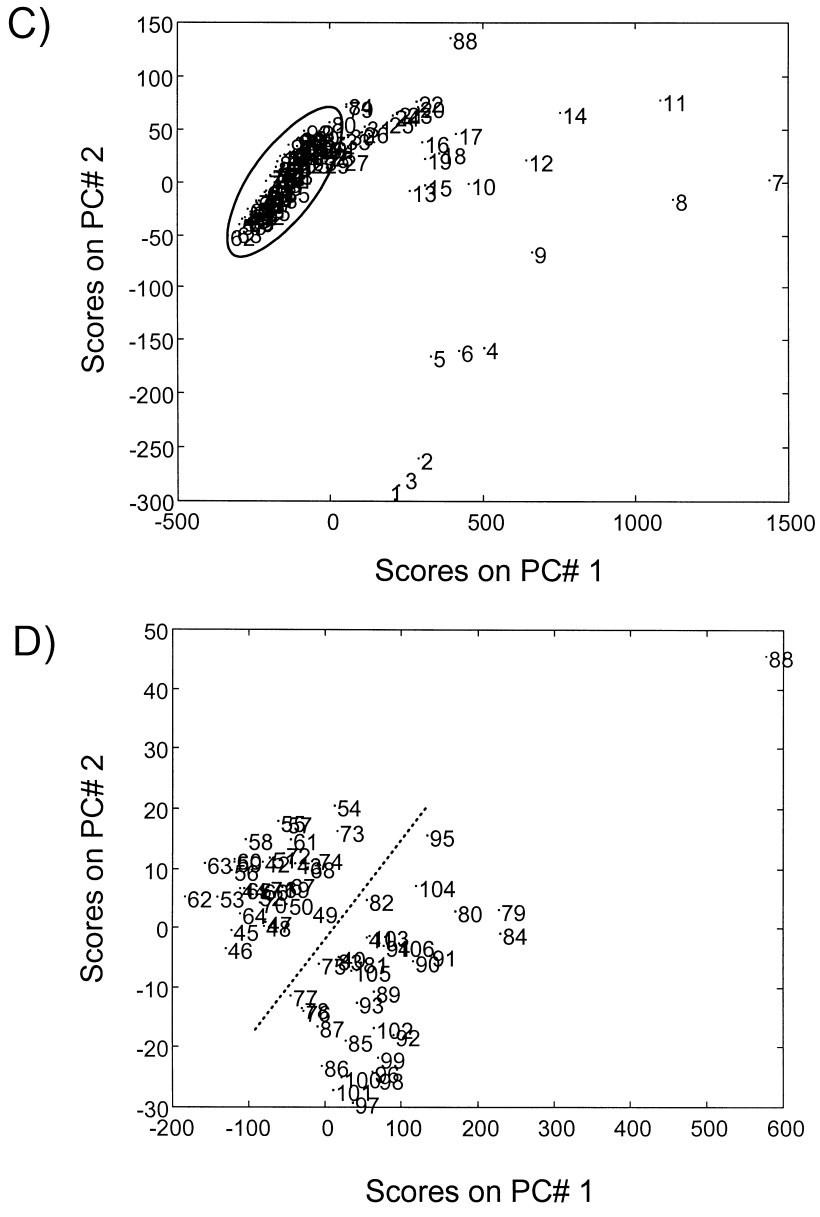


Fig. 2 (continued).

$(47 \times 2 \times 20 \times 311)$. The plot of the PC scores 2 and 3 is displayed in Fig. 3C showing a clear-cut classification into 5 factories (a, b, d, e and f) and with a clear tendency of timing within each cluster from below to above, ranging from the early to the late samples. This classification is much more clear-cut than that obtained from the PCA score plots in the thick juice material from different factories investigated by Nørgaard [15] where factories were overlapping and where the time aspect of the samples could not be modelled in the same plot. This underlines the advantages of respecting and exploiting the structure of

the data and selecting chemometric algorithms accordingly, which are further discussed in Appendices A–C.

We will now proceed further upstream in the sugar process to beet production in agriculture. The price paid to the farmer for the beets is regulated by the

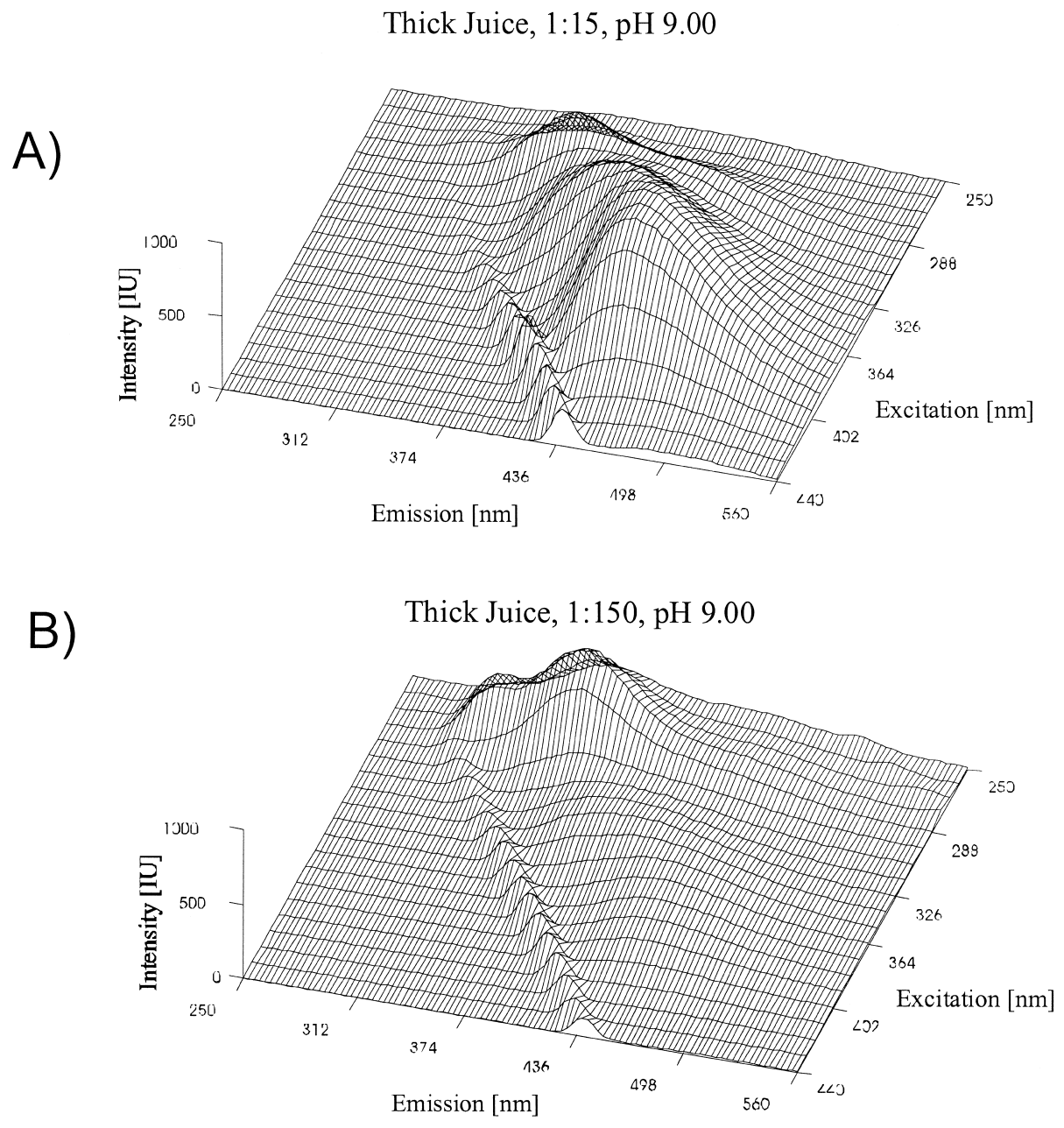


Fig. 3. (A) and (B) Fluorescence landscapes of one thick juice sample in two concentrations. Note how the fluorescence signal in the UV region is quenched in the 1:15 dilution (A) and becomes dominant in the lower concentration (B). (C) A Tucker score plot showing the pattern of principal components two and three of the sample mode from 4-way PCA. Two principles are illustrated by this plot: samples from the same factories (a, b, d, e, and f) are clustered nicely together and simultaneously the shift of the samples according to week number (e.g., d1 to d10) reveals that temporal information is present in the fluorescence landscapes.

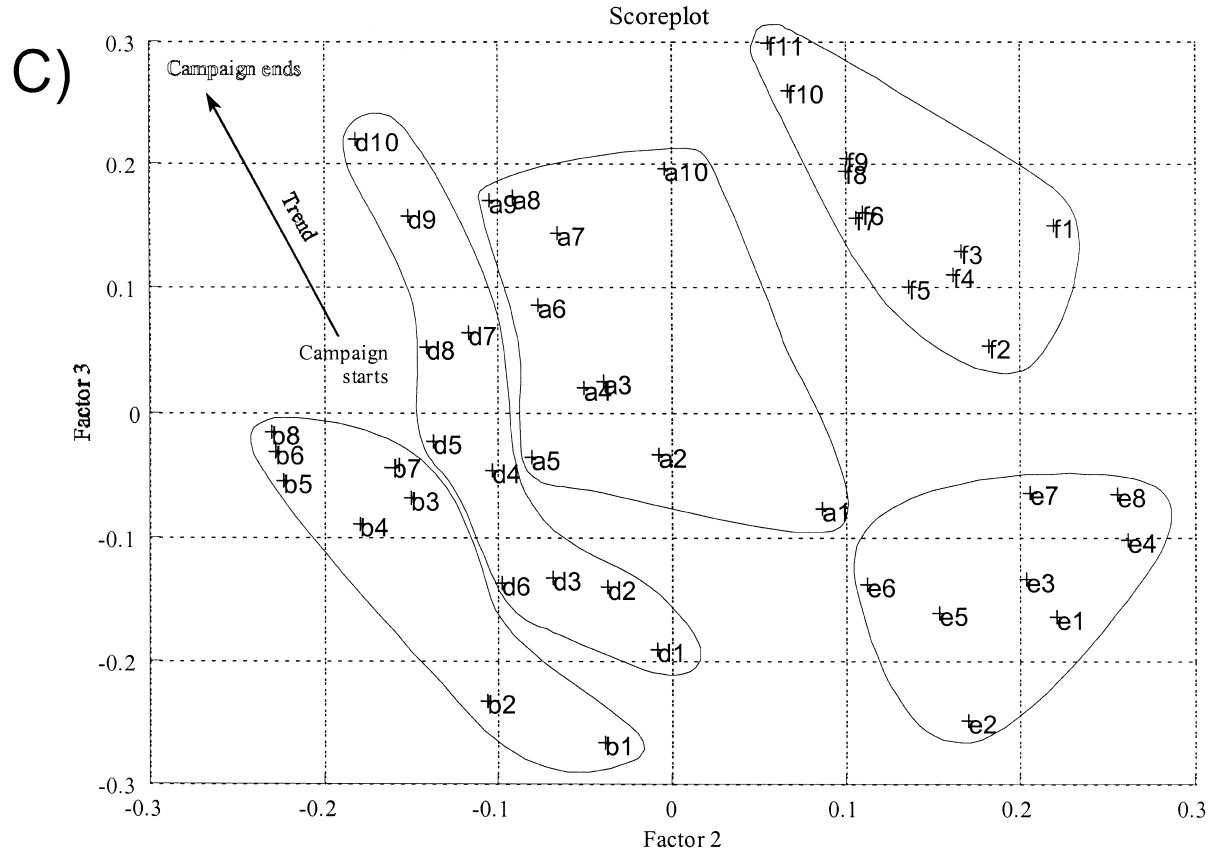


Fig. 3 (continued).

sugar and amino–nitrogen content of the beet juice, the latter indicative of potentially colour-forming molecules which could intervene with purification of white sugar by crystallization.

Fig. 4A displays fluorescence information from 24 sugar beet mash samples taken from the receiving station of a sugar factory. As seen in Fig. 4B, there is an excellent correlation between whole fluorescence spectra and amino–N in these samples. In order to preliminarily investigate the variation in fluorescence between sugar beets from different farms, three 15-kg sugar beet samples were taken from nine farms. Fig. 4D shows the PCA clustering analysis of the corresponding fluorescence spectra of the beet juices from Fig. 4C. There is a clear clustering effect of the fluorescence information related to farm site which not only depends on amino–N, but which also indicates differences in the complex underlying chemistry due to beet variety, sowing time, soil, fertilizer and weather which has to be understood by further systematic trials with laboratory verification and by correlation to technological quality. The fluorescence method could thus be a candidate for a screening analysis for beet quality to be used by the plant breeding companies and farmers to optimize the plant growing conditions and the beet varieties.

We will now investigate the evaluation possibilities of another multi-way generalization of PCA, namely PARAFAC [18,19] (the mechanism of which is discussed in more detail in Appendix B), to study 268 sugar samples, each averaging 8 h of processing (equal to one shift) by fluorescence from a three-month campaign in 1995 from a well-controlled sugar factory. Contrary to the unconstrained Tucker model, the three-way PARAFAC model (268 samples, 571 emission wavelengths (Fig. 5A) and 7 excitation

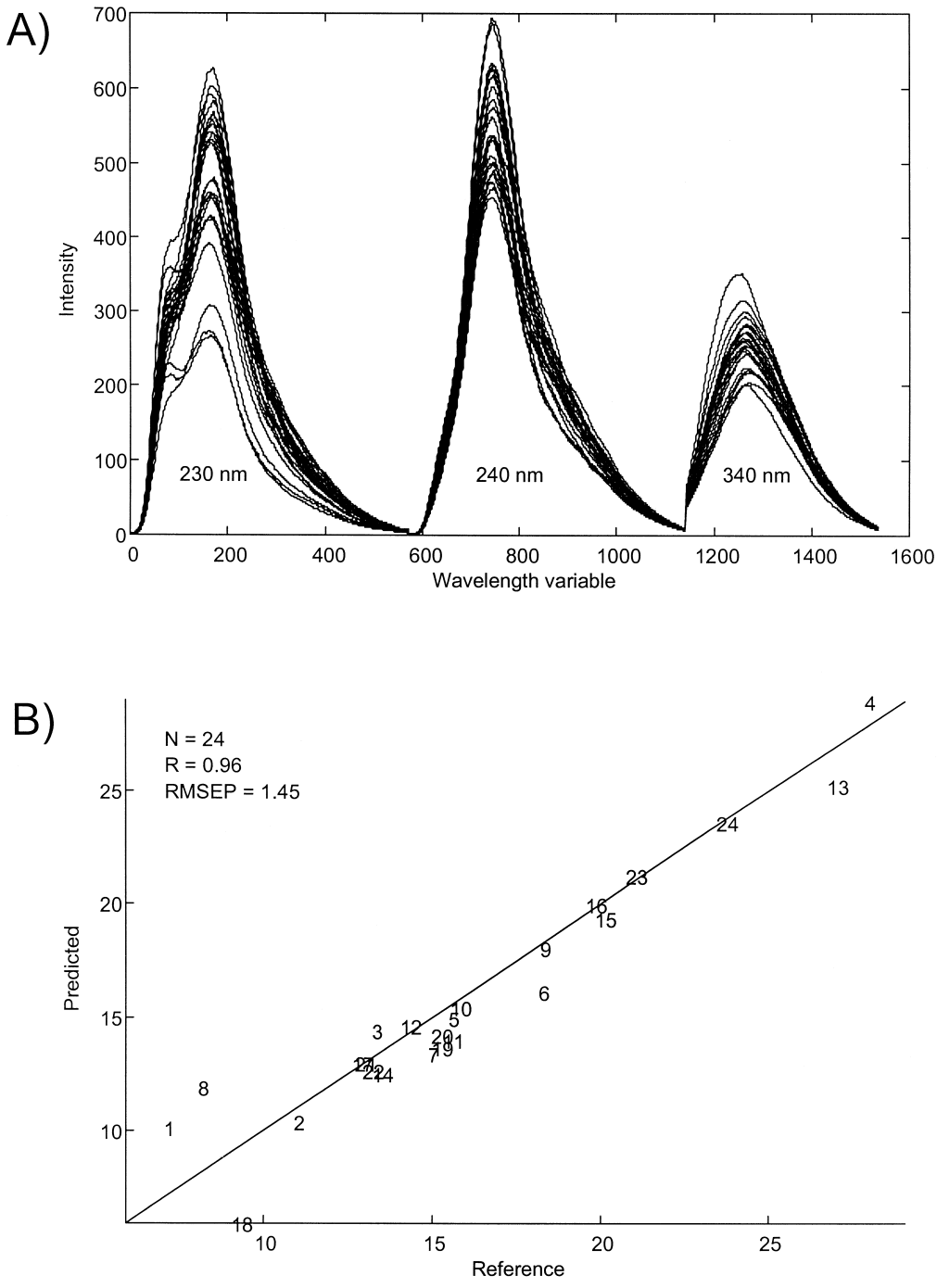
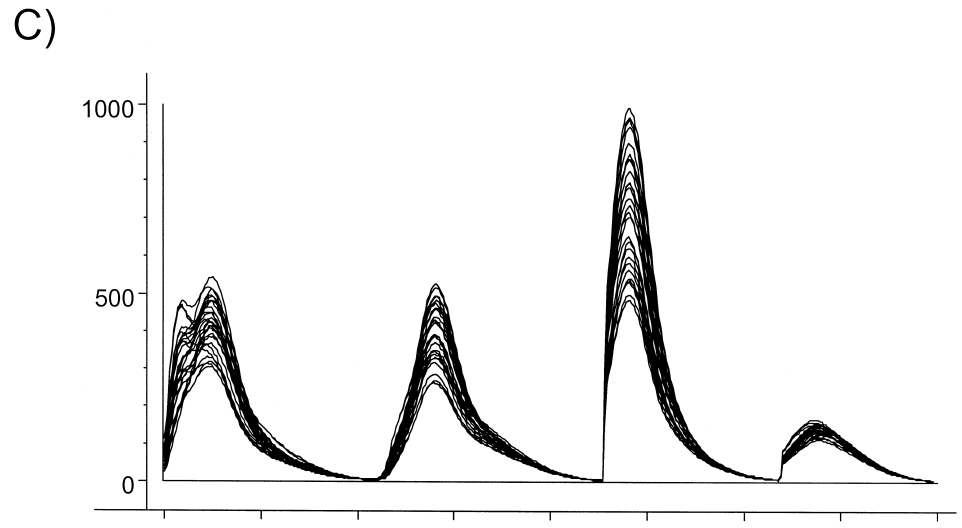


Fig. 4. (A) Fluorescence raw emission spectra of 24 sugar mash samples. Excitation 230 nm, 240 nm, and 340 nm (emission ranges 275–560 nm). (B) Predicted versus measured plot of amino–N values. Based on a three-factor PLS-model with fluorescence spectra as independent variables and amino–N as the dependent variable. (C) Raw fluorescence spectra recorded on sugar beet mash samples from nine different farms (three sample from each farm, i.e., in total 27 samples). The excitation/emission wavelengths are the same as those displayed in Fig. 1. (D) A score plot showing that the beets from the same farm no. 4, 7, 9, 10, 12, 15, and 19, in the fluorescent fingerprint seen in the mash samples.



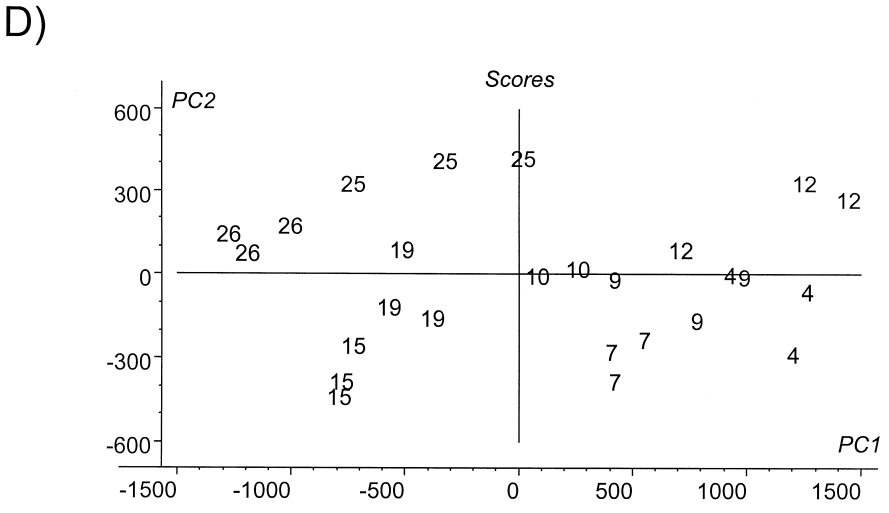


Fig. 4 (continued).

wavelengths) may allow direct recovery of some of the pure spectra from the underlying chemical substances.

In this study, four loadings called pseudospectra could be resolved, two of which were identified by comparing emission and excitation pseudospectra with the true spectra of tyrosine (Fig. 5B) and tryptophane (Fig. 5C, see also discussion in Appendix B). Fig. 5D shows the four emission pseudospectra and their correlations to the process parameters colour and ash. In this preliminary study, it is observed that the four component candidates have different patterns of correlation, pointing at the possibility that they may be used as indicator substances, e.g., for colour or ash alone or in combination. Compound 4 is obviously the best indicator for colour.

In Fig. 5E, the scores for the four pseudospectra during the campaign are shown. The components

show a high degree of covariation, especially in the beginning of the campaign, revealing a tendency toward higher peaks during weekends. The variation levels off during the season when outdoor temperature is decreasing. Around shift 200, on about the 15th of November, compound 4 scores steadily rise,

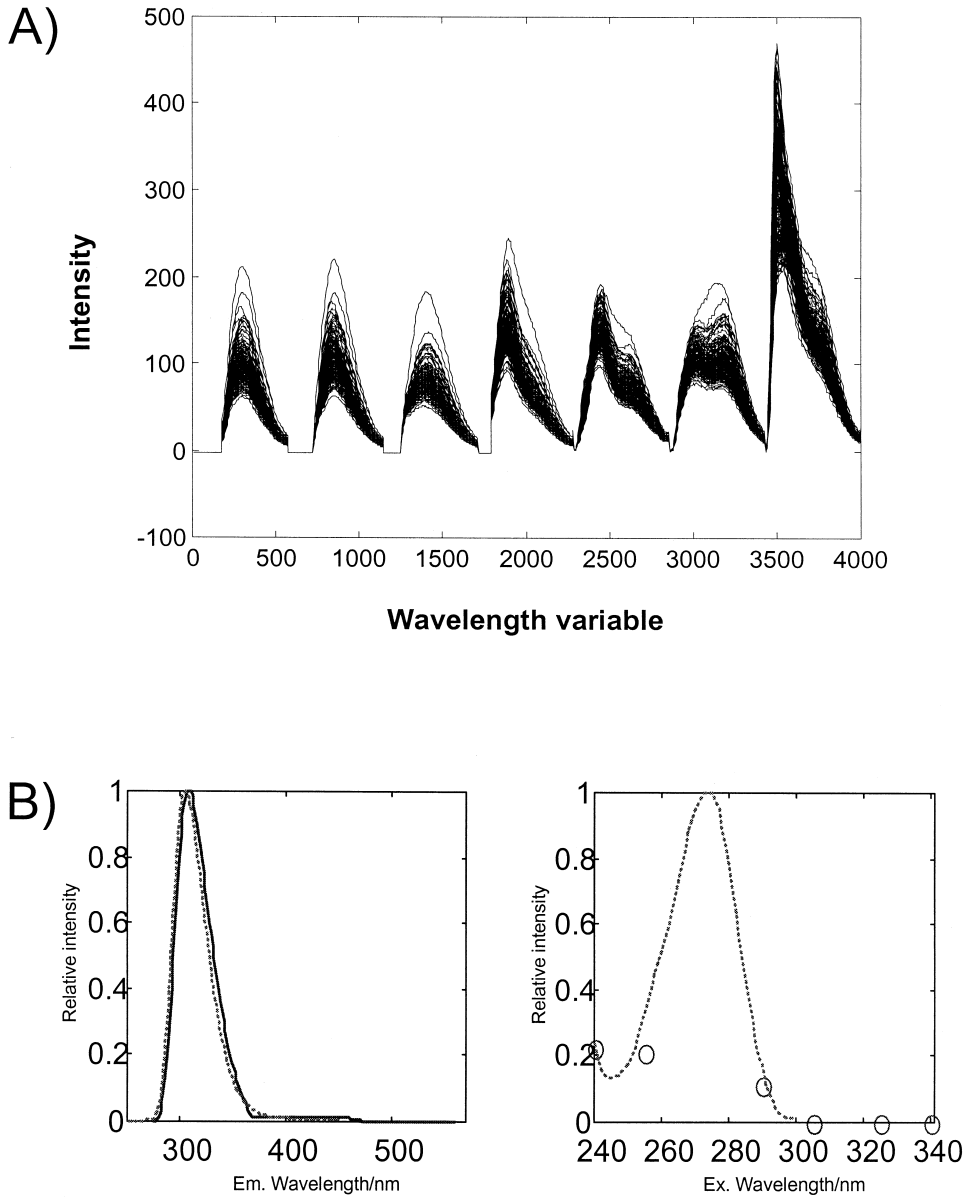
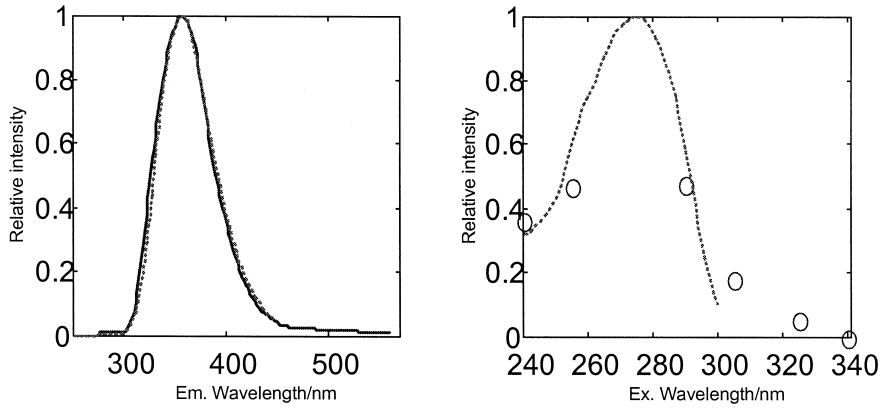


Fig. 5. (A) Raw fluorescence emission spectra of 268 sugar samples sampled as a mean spanning eight h equal to one shift during a three-month campaign (1995). The samples were measured at excitation wavelengths 230, 240, 255, 290, 305, 325, and 340 nm (emission ranges were all 275–560 nm). (B) Pseudo-emission and excitation spectra for compound 2 compared with pure tyrosine (dashed). To the left the emission parameters are shown and to the right the excitation parameters are shown. (C) Pseudo-emission and excitation spectra for compound 3 compared with pure tryptophane (dashed). To the left the emission parameters are shown and to the right the excitation parameters are shown. (D) PARAFAC emission loadings 1–4 and their correlations to ash and colour. (E) Concentrations (scores) of the four pseudocomponents.

C)



D)

	Colour (r)	Ash (r)	Pseudoemission spectra
Comp. 1	0.60	0.69	
Comp. 2 "Tyrosine"	-	0.42	
Comp. 3 "Trypthophane"	0.43	0.64	
Comp. 4	0.81	-	

Fig. 5 (continued).

while scores for the other three components are more constant.

Factory records as well as interviews with the process engineers revealed that beets are stored longer during weekends which may produce heat due to microbiological activity which is reflected in higher fluorescence scores for all four components as well as an increase in colour. The change in the level of compound 4 and the increased colour development could be explained by frozen beets due to the coming winter and the resulting process adjustments. Compound 4 could thus be an indicator for colour as well as for frozen beets. These observations has to be verified and generalized in more detailed studies with other factories and other production years.

The variation of the fluorescence pseudocomponents during the production campaign clearly indicates temperature effects covariant with colour of sugar. We may therefore induce a hypothesis from real life data that temperature in the receiving beet stores may have a major impact on the precursors of sugar colour which should be checked by monitoring temperature in the store.

We have demonstrated that with a minimum of prior knowledge of sugar technology and chemistry we are able to establish a constructive, exploratory

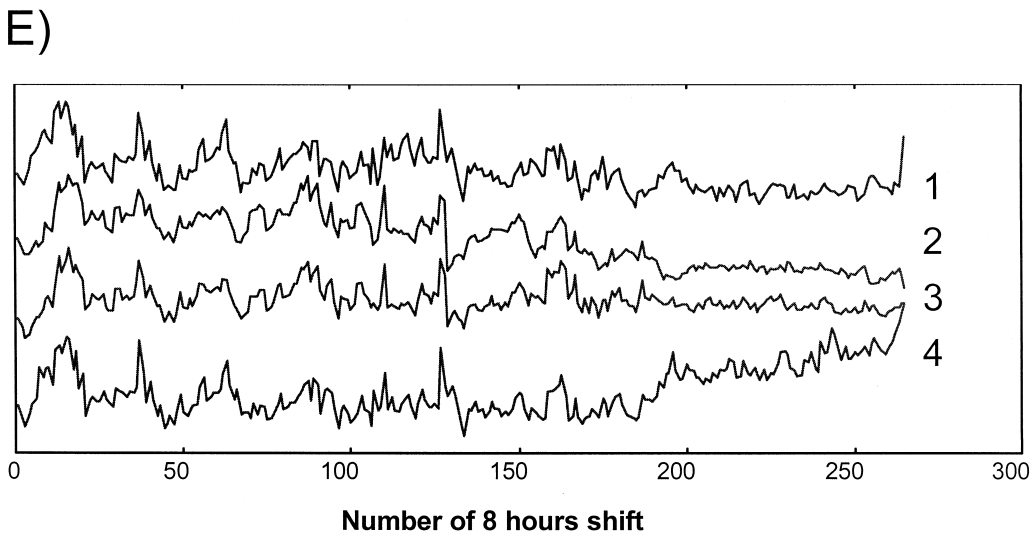


Fig. 5 (continued).

dialogue with the sugar technologists throughout the whole production chain using the tools of a fluorescence screening analysis, chemometric software and the computer. Together we have been able to identify a range of process events which the fluorescence analysis had picked up. At the same time we have shown that the fluorescence screening method has the potential for providing a holistic fingerprint of the state of chemistry in the process in the form of 4 fluorophores which correlates with a range of important quality parameters throughout the beet sugar manufacturing process and which may be used as indicator substances which is further demonstrated in Appendix B.

3. What chemometrics and food science can learn from each other

In his outline on the roots of mathematics in human culture, Barrow [4] emphasizes the inherent weakness of the human brain in multivariate analysis and the fundamental role of written symbols and basic assumptions axioms—the fundamental on which the mathematical machinery is built. It should be acknowledged that 'axioms' are also a fundamental part of human cognition—a method to keep a working platform of consistency in bookkeeping in a complex universe. This is often practised without thinking too much, for example by the chemist in the laboratory as well as by the food consumer in daily life. However, when trying to exploit mathematics in real life, such as in food production, it becomes as crucial to define 'the axioms' of chemistry and food production as those of the mathematical models which are used to describe and predict events in data from food processes.

Food production is dependent on the demand of markets in thousands of complex production chains regulated by the monetary principle and governmental and international regulations. The functional unit is 'man as selector' [20] in different roles as consumer, distributor, manufacturer, as well as raw material and secondary material supplier.

This exploratory selection process with the individual consumer in the centre may be elucidated by a model for learning—'the selection cycle' (Fig. 6) [20] related to the concept of the perceptual cycle in psychology [21] (p. 37)—comprising different steps starting with a primary selection hypothesis inspired from the global area (0) proceeding with an inventory/screening analysis (I) and selection of material and methods (II), followed by testing/evaluating (III) which results in a secondary (IV) selection hypothesis are valid for the local area.

After an introductory round the individual selector proceeds in increasingly more focused and limited

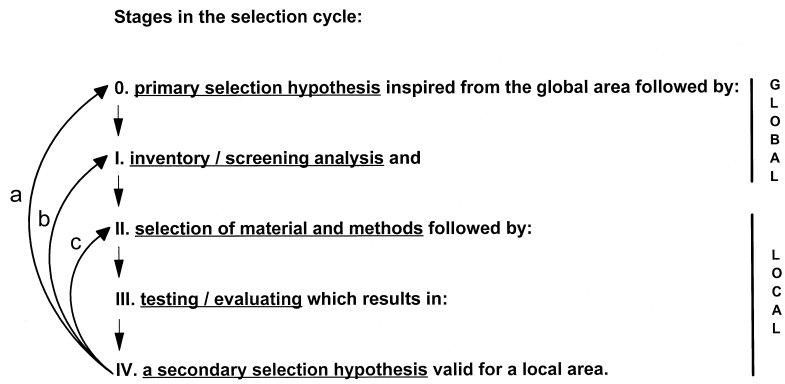


Fig. 6. The selection cycle [20].

rounds (Fig. 6a–c) (e.g., omitting point 0 (b) or even 0 and 1 (c)) in the selection cycle. Thus, in each cycle, the dynamically adapted secondary selection hypothesis (IV) is validated experimentally (III) in numerous revolutions. It is a common phenomenon that in the progress of time the secondary hypothesis (IV:*n*) and its derived propositions from the local area have often overshadowed the more global primary selection hypothesis derived from. e.g., society. It now lives its own life in the context of society in the mind of the selector in spite of its local limitations. In this way serious bias could be introduced unintentionally.

The food experience of the consumer tells that the selection cycle contains both global and local as well as visible and hidden domains. When buying food in the store, the selector starts with a primary selection hypothesis (0) implying acquisition of defined foods with expectations regarding culinary quality, health and economy in a long-range perspective. He/she then evaluates foods in the local area with regard to visible (screening) characteristics (I) like colour, packaging and price. After selection (II) the individual 'develops' hidden qualities such as smell, taste and tenderness by cooking the food at home (III). This may generate a reaction in the form of a new purchase policy (IV) which will then be checked in later cycles. The exploratory behaviour of the consumer creates information about foods in the local area which again may reinforce or weaken a specific behaviour of selection.

The global aspect of food selection [20] includes the part of the accumulated feedback on the physiology of the consumer which is caused by his/hers own food selection and consumption. It also contains the hidden feedback effects [20] of nature which Darwin called 'natural selection', now also including the indirect influence of the selection force on the food production environment of the activity of a large population of human selectors exploiting resources and the resulting effects on their food quality and health.

Returning to our example on exploratory analysis by fluorescence screening, we find that indeliberately we worked exploratively according to the selection cycle model: we attempted an analysis in the 'global area'—the beet sugar production chain—by using chemometrics.

Without extensive knowledge of sugar manufacturing we used the fluorescence screening method to pose a question to the process as follows: ''Is fluorescence analysis chemically and technologically relevant as a screening method for control and prediction of parameters of industrial interest?'' This is the primary hypothesis (0) in the selection cycle. After analyzing (1:1) sets of sugar products with fluorescence spectroscopy, we could select (II:1) and evaluate (III:1) sugar samples belonging to defined factories and processes as well as identify time effects due to date of delivery throughout the season. We could also identify process balance in a start-up test (III:1) by analyzing the sugar product as well as indicating a minor breakdown in the balance point.

From these results, we could induce a preliminary secondary selection hypothesis (IV:I) that a sugar sample could be looked upon as 'a datalogger' which

integrates information from the production chain upstream that could be read by a fluorescence spectrophotometer and evaluated by PCA.

In our second selection cycle, we proceed on our data selection adventure in the local area—the laboratory—by comparing the classification of traditional sugar analyses (I:2) with fluorescence analyses in two separate PCAs. Due to the fact that samples with high fluorescence have high ash, colour and amino–nitrogen values, we selected (II:2) the PLS algorithm which gave good correlation in an evaluation (III:2) between fluorescence spectroscopy at 5 specific wavelengths and sugar quality, indicating direct or indirect relationships. This fact was used to formulate a new and more specific secondary selection hypothesis (IV:2) that fluorescence could be used as a preliminary screening method for direct analysis of purity in sugar. In a third selection cycle this hypothesis was expanded to the whole production chain. In a fourth selection cycle, we enlarged our third secondary hypothesis by suggesting that behind the fluorescence spectra lies information from discrete chemical compounds which may be used as 'indicator substances'. These substances reflect chemical composition of sugar and intermediate products as well as process parameters. To solve this problem we selected multiway exploratory algorithms such as Tucker and PARAFAC. From a complete material of 8 h average sugar samples from an entire sugar campaign PARAFAC displayed 4 different pseudospectra (loadings) corresponding to 4 discrete compounds (fluorophores), two of which could be preliminarily identified. The four pseudospectra were shown to be able to model process observations, such as frozen beets and quality criteria like ash and colour, as well as other important process parameters as discussed in Appendix B.

Finally, in the fifth turn in the selection cycle we aim at more precisely identifying the underlying chemical compounds by high pressure liquid chromatography in the local area, the research laboratory, which is outside the scope of this paper. Thus, we do not forget to check the results from the exploratory screening with our chemical interpretation of the problem.

In the longer perspective, we aim to feed back the integrated experience of the multivariate fluorescence perspective from the five selection cycles into the primary area (0), the beet sugar industry, in the form of an established 'global' control method covering the production chain from beet production to sugar.

In our sugar process example, with our sensitive spectrofluorometric method we are not measuring sugar, which is non-fluorescent, but rather a selection of impurities such as fluorescent amino acids, phenols and their reaction products with reducing sugars: the high molecular coloured melanoidines and melanines. The sugar processing engineer tries hard to avoid the formation of colour by adjusting pH with CaO and adding reducing agents in the form of SO_2.

In traditional chemical analysis, one starts by defining the hundreds of chemical substances involved in a process, as was done for the sugar industry by Madsen et al. [22] in order to understand color formation. If the target hypothesis is to find easily identifiable indicator substances by which to model quality and process characteristics, we suggest that our exploratory, inductive method by introducing a multivariate screening method in the global area of the sugar factory would be more economical than a normative, deductive strategy based only on a priori chemical knowledge, chromatography and classical statistics as studied in the local area—the research laboratory.

We can thus conclude that the strategy of exploratory chemometric analysis in the example is closer to the behaviour of 'man as selector' performing in the food production chain than to how statisticians operate today. While statistics is mainly directed toward probabilistic methods in modelling noise, identifying the object as a void in the space of noise, exploratory data analysis and chemometrics is more deterministic [23]. It instead tries to model the contours of data objects by data experimentation in the computer.

In our example, statistical validation is completed with two other alternatives: calibration/test set validation (data experimentation) and interviews with the processing engineers, including confirmation from process data banks. It must be pointed out that exploratory data analysis, which contains an important inductive, empirical element of validation through enumeration [8], does have a more humble profile [24] in a restricted context than classical mathematics and statistics. It places less demand on finding the abso-

lute (generalized) truth (see citation by Chaitin in the introduction), but instead aims at finding an adequate and more precise local truth of equal or higher importance which is time- and context-dependent. It is basically a provisional detective work [25], trying to explore the partly unknown territory of the world outside the laboratory where hard hypotheses are likely to neglect covariance and synergy and therefore are insufficient and inefficient. An endeavour of reversed logics might be fruitless in a classical situation relying on univariate analyses where each object has just a few characteristics, a multivariate analytical situation with many informative data points attached to each object increases the uniqueness of the description. In classification it allows safe detection of outliers, thus greatly increasing the validity of the results.

From our platform of data technology in chemometrics, we can clearly see how it was necessary before the computer to develop a very special form of deep, rigorous and general thinking [26] aimed at identifying the laws of nature. The goal is to obtain consensus in the form of a global rational opinion as a 'science map of reality' through organized, intersubjective communication [26]. Such an inflexible outlook is rather strange for model creation in the normal human mind which is characterized by pragmatism and cognitive flexibility, although with a short memory.

In fact, as the physicist and historian Thomas Kuhn points out [27]: ''The investigations of classical science have few quantitative points of contact with nature, because investigations of those contact points usually demand such laborious instrumentation and approximation and because nature itself needs to be forced to yield the appropriate result, the route from theory of law to measurement can almost never be travelled backward. Numbers gathered without some knowledge of the regularities to be expected almost never speak for themselves. Almost certainly they remain just numbers''.

We have here applied our global (with regard to fluorescence) screening method and exploratory data analysis and gone from measurements of sugar samples to a theory of selected indicator substances for process control. Is it not this fairly straight forward travel from the measurement of phenomena from real life to construction of a theory which Kuhn calls 'backwards', which we have just humbly attempted and to a large extent succeeded in?

Obviously, new multivariate screening methods and data evaluation methods based on induction using the computer, which Kuhn [27] and Hempel [8] were unaware of (and still the vast majority of scientists are today), open up new possibilities for connecting data from the world as it is with science—if one can obtain a common platform for 'the axioms' and contexts of mathematics and those of the world under study. This issue is further exemplified in Appendices A–C.

We may thus conclude that there is a major conceptual distance between the aspiration of science of global understanding of natural phenomena in its generalized sense and global evaluation of measurements as is from the real world for prediction and control. This discrepancy has to be further understood and bridged by a new strategy combining screening methods, mathematics and information technology. We can thus look upon the flow of information in our sugar process example as a dialogue between two connected selection cycles—one global (sugar production) and one local (the laboratory).

Attempts by leading physicists to introduce a new paradigm change in science, such as in the now classic book by Prigogine and Stenger [9] (since 1979) 'Order out of chaos—Man's new dialogue with science' are only slowly being acknowledged. They see the world as an open self-organizing system which develops while consuming energy. The world is heterogeneous. It contains simple as well as complex, reversible as well as irreversible and probabilistic (e.g., due to thermal movement of molecules) and deterministic (e.g., due to DNA in organisms) including chaotic moments. This new outlook on the world, combined with exploratory data analysis, is much more relevant for describing the dynamic situation in food science than classical hard modelled science with its mathematics and statistics which, however, is still relevant in special cases. One should thus be cautious in introducing a priori biased statistical evaluation techniques in such a world without defining context in an inventory in the start of the selection cycle.

As food technologists we, of course, gratefully acknowledge the laws of nature as defined by science in our food technology research. But our pri-

mary task is not to produce the eternal and general. We do not aim to make a factory which produces the same product from the same raw materials by the same technology forever. Instead, we are interested in controlling the timely, transient and specific traits of the production, so that the company may withstand competition for another year. The generally acknowledged mathematical language which should be used in the future to model such data should be more compatible with this context and to the new science of Prigogine and Stenger [9]. Today it is not.

We now see the great opportunity to directly study order out of chaos in Prigogine's and Stengers' sense by applying multivariate screening methods in real life (e.g., in a sugar factory) as evaluated by the computer and exploratory data analysis. It is therefore of great wonder to us that most scientists, including Prigogine, investigating self-organizing systems are still apparently working with hard modelling, deductive methods alone and have not yet found their way to supplement with the new multivariate methods. Science is indeed conservative. It has not yet discovered all the new kinds of freedom which the computer may introduce. It is possible within the limits of the screening analysis and the mathematical algorithm with the exploratory method to discover unknown phenomena directly. It is only possible for classic science to obtain new knowledge outside its traditional deductive system of hypotheses indirectly through unexpected interference, e.g., in discovering environmental problems.

The classical, positivistic science presumptions [9,26,27] of the world are still dominant in the present normative-deductive culture and severely restrict chemometrics. They focus on deduction from a priori hypotheses based on fully transparent factors which can be seen directly or revealed after experimentation. As long as the present consensus in statistical hard modelling and validation rules, the more flexible, soft exploratory data models which introduce latent factors and empirical validation, such as PLS regression, will not be accepted as a science. This is due to the incomplete transparency of these algorithms which for the mathematicians are undecidable by lack of mathematical proofs, in spite of their better robustness and ability to adjust to a changing context by experimental validation reflecting human behaviour in the selection cycle.

In fact, the operation of the PLSR algorithm makes a dialogue possible between screening data from the world as it is and laboratory data. This is expressed in finding common latent factors in a cyclic adaptation process which embodies a dialogue between the global and local principle, between the real world and sciences, just as in the selection cycle discussed previously.

It is obvious that chemometrics can contribute to food science with new more flexible data programs which display the exploratory results in cognitively accessible graphical data interfaces. Food science and chemistry on the other hand stimulates the chemometrician to take new contexts into consideration in the development of models suitable for real world data which is exemplified in the Appendices A–C.

In practical life, respect for the 'axioms' of the world in the form of contexts is more important than transparency. In science it seems to be the reverse. Transparency is preferred based on the axioms of the mathematical machinery, far from the contexts of the world which was supposed to be studied. Because of its lack of complete transparency we could thus for the moment look upon chemometrics more as a technology than as a scientific discipline—a very vital technology which already has proven its potential in chemistry and in other related technologies [23,28] —an invention the results of which science should explore and incorporate in its basic principles.

As early as 1941, Emil Post, one of the co-discoverers with Turing [4] (p. 292) of non-computable operations, wrote [29] the following comment regarding the divide between meaning and formalism in mathematics: ''mathematical thinking is, and must be, essentially creative. It is to the writer's continuing amazement that ten years after Gödel's remarkable achievement current views on the nature of mathematics are thereby affected only to the point of seeing the need of many formal systems, instead of a universal one. Rather has it seemed to us inevitable that these developments will result in a reversal of the entire axiomatic trend of the late nineteenth and early twentieth centuries with a return to meaning and truth. Postulation thinking will then remain as but one phase of mathematical thinking''.

It should thus be possible to assemble a mathematical algorithm to describe and predict complex conditions in the real world inspired by finding order

in observational measurements of nature by consulting the computer. Such an endeavour must respect the mechanisms how humans best senses complex information.

While we wait for the breakthrough of the new interdisciplinary science [9] where exploratory, inductive chemometrics is an integrated part as an established option, we could with the support of the relatively recently discovered computer contribute to the basic mathematical language of the new science by balancing the normative and exploratory principles in a dialogue, as described in our example. In this work food technology is an excellent Trojan horse in the conservative scientific city of Troy, harbouring research teams prepared to fight for the revolutionary new science and its new mathematics while awaiting the right moment and better times.

Acknowledgements

The inspiring cooperation with Ole Hansen, Lars Bo Jørgensen and John Jensen, Danisco Sugar Development Center Nakskov, Denmark in exploring beet sugar production is gratefully acknowledged. We also thank John Hørlyck and Peter Henriksen for professional assistance in maintaining equipment and in measuring the samples as well as Gilda Kischinovsky for valuable help in preparing the manuscript. The four years of research summarized in this paper was supported by Nordic Industrial Fund project No. P93149 (salary to Rasmus Bro), EU-AIR2 project No. CT94-1416 'AFFLUENCE' (equipment and salary to Claus Andersson), Føtek-2 project 'On-line/at-line screening methods - spectrometric structural analysis for process and quality control in food production' (salary to John Hørlyck and Gilda Kischinovsky), and by the Danish Veterinary and Agricultural Research Council (equipment and salary to Lars Nørgaard).

Appendix A. Selecting and adjusting chemometric models to represent different contexts of the world

Chemometrics has arisen as a hybrid with contributions from various sciences like econometrics, psychometrics, classical statistics and physics. The mixed background is reflected in the way the chemometrician actually conducts the data analysis. Central aspects in data analysis are the selection of data as well as the selection of suitable models, combined with adaptation of the models to a given problem. Classification, for example through PCA, is a fundamental first step in an exploratory data investigation of a given data set (e.g., fluorescence spectra), employing data reduction into latent variables in this way revealing resemblances and outliers.

In the framework used throughout this paper we see the alternation between the selection of models and the selection of data which again influences the selection of material for analysis and the technological focus of the project. The data analyst might follow different chemical roads depending on the goal of the investigation. However, the exploratory approach starting with an inventory with a data classification from a multivariate screening method is to be preferred in the beginning of an investigation in order to minimize bias. After revealing the data structure, both surprising and expected elements can be identified from which more specific correlation models may be created using a range of new chemometric methods. These include the new multi-way methods employed in our example with sugar process fluorescence analyses.

There are various models for analysing multi-way data sets, see Kroonenberg [A1]. In Figs. 7 and 8 we shall focus on the *N*-way principal component analysis (*N*-way PCA) which is a generalization of the 3-way Tucker3 model [A2] to *N*-way data arrays as well as the PARAFAC model [A3]. The authors would like to draw the reader's attention to the fact that the generalization of bilinear PLSR to multilinear PLSR (*N*-PLS) was given by Bro [A4].

A.1. Tucker model

As with conventional two-way PCA, the model uses a projection technique whereby the systematic variation in data is reduced to a few representative factors. Due to some mathematical features (i.e., factors are non-unique and can be rotated) of the model and its solutions, the term *N*-way PCA is often used to describe the Tucker 3 model. Fig. 7 provides a basis for presenting the *N*-way PCA. The 3-way PCA model of a 3-way data array $\mathbf{X}$ of order (r_1, r_2, r_3) is depicted in the figure. The array is decomposed into

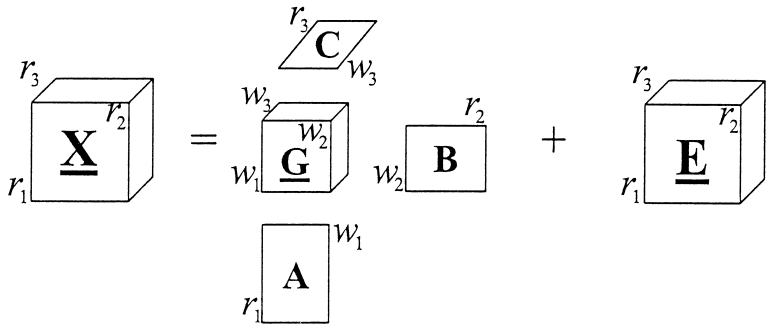


Fig. 7. Tucker.

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 **A** (r_1, w_1), **B** (r_2, w_2) and **C** (r_3, w_3). The mathematical representation is as follows

$$x_{ijk} = \sum_{f=1}^{F} \sum_{g=1}^{G} \sum_{h=1}^{H} a_{if} b_{jg} c_{kh} g_{fgh} + e_{ijk} \qquad (1)$$

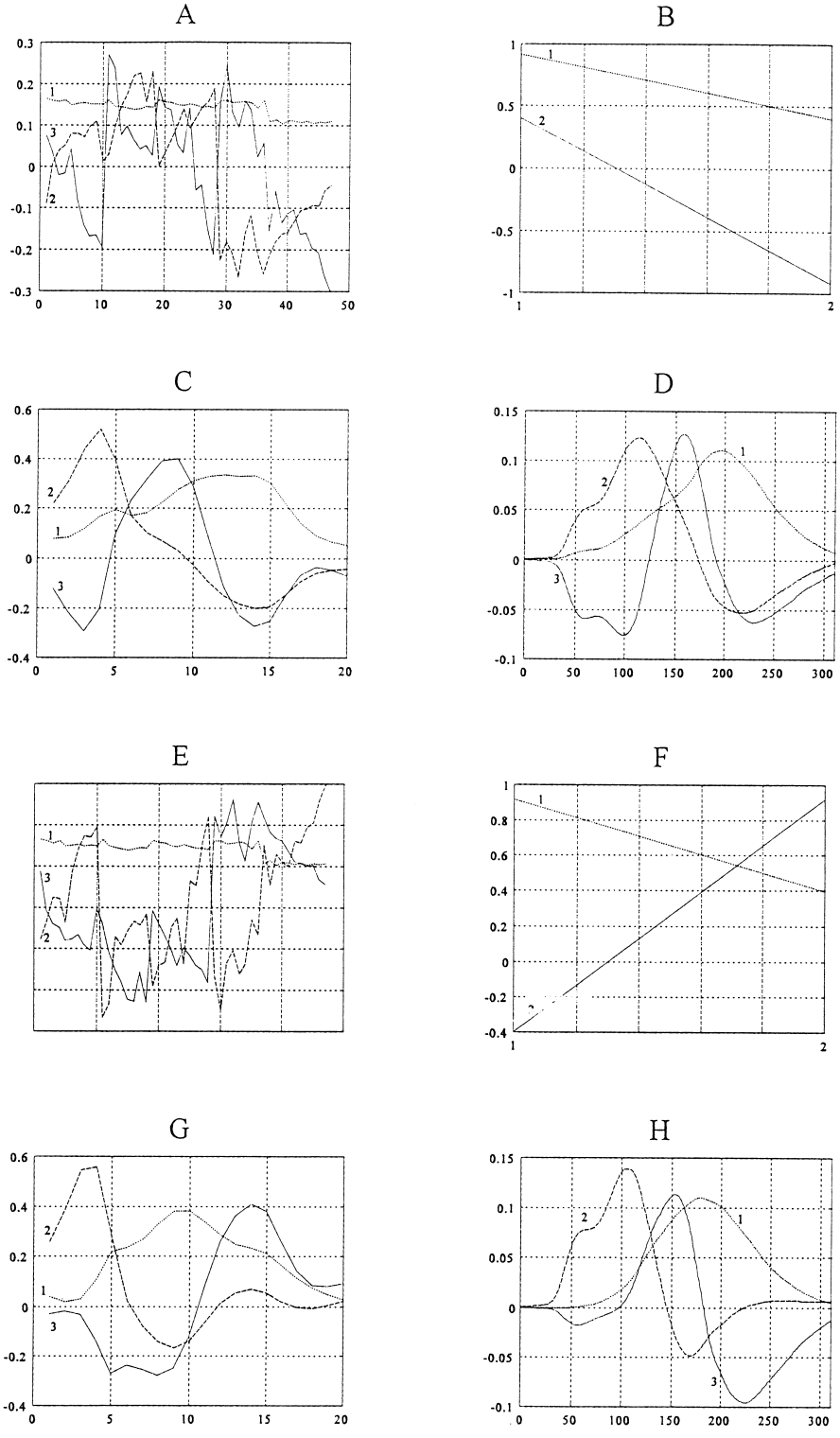
The number of factors in each of the three ways, i.e., w_1, w_2 and w_3, must be determined by the analyst from a priori knowledge about **X** or by evaluating models with different combinations of w_1, w_2 and w_3, choosing the order that gives the most accurate model of **X**. The correct number of factors is found as a compromise between having a good fit and as few factors as possible. 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**. Upon calculation of the model, the factors in the three component matrices **A**, **B**, and **C** and the core **G** must be interpreted. Since the factors are orthogonal, hence linearly independent, the squared core elements are proportional to the variation explained by the combination of factors in question. Thus, if $g_{i,j,k}$ is the largest squared element in **G**, the combination of factor i in the first mode, factor j in the second mode and factor k in the third mode explains most of the variation in **X** and the analyst should give these factors special attention when interpreting the model.

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 having the exact same fit. Returning to the exploratory power of the squared elements of the core, one can perform *selective* transformations of a solution to give a core where only a few squared entries are significant [A5]. Having only a limited number of significant core entries allows the analyst to focus on a few combinations of more significant/general factors. Hence, we use an unsupervised algorithm to select a solution from this infinity of solutions to yield a model for interpretation which is simple as possible.

A.2. Data (an in-depth treatment of this data set was given in Andersson et al. [A6])

Fluorescence intensity landscapes, or excitation-emission matrices, were measured on 47 thick juice samples from the 1994 sugar campaign. Five factories contributed thick juice samples. Two typical landscapes from one sample are shown in Fig. 3A–B in the main text. Note that the peaks in the ultraviolet area do not decrease from A) to B) with dilution. This is caused by concentration quenching, or inner-ab-

Fig. 8. (A) Factors in the first way representing variation in the thick juice samples. (B) Factors in the second way describing concentration effects. (C) Factors explaining the excitation profiles. (D) Factors explaining the emission profiles. (E) Rotated sample factors. (F) Rotated concentration profiles. (G) Rotated excitation factors. (H) Rotated emission factors.


A
B
C
D
E
F
G
H

sorption effect [A7]. Each sample has been diluted volumetrically 1:15 and 1:150 with pH 9.00 NH_4Cl in double ion exchanged and Si-free water. Both of these dilutions were measured using 20 excitation wavelengths (250–440 nm, 10-nm intervals) and 311 emission wavelengths (250–560 nm, 1-nm intervals). At the excitation and emission sites 10 nm slit widths were used. The instrument was the Perkin Elmer LS50B spectrofluorometer. As indicated by Fig. 3 (main text), the combination of a narrow emission slit width and generally low turbidity allows for neglecting the Rayleigh scattering. Since each intensity measurement in the collected data depends on four external parameters, the sample number (47 samples), the concentration (two levels of dilution, 1:15 and 1:150), the emission wavelength and the excitation wavelength, the measured intensities constitute a 4-way data table of order (47, 2, 311, 20). We will apply a 4-way PCA model for analysis of these data. The 4-way PCA used in this application can be conceived as an extension of the decomposition illustrated in Fig. 7 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.

In order to find the optimal numbers of factors for the 4-way PCA model, several models of different orders were investigated. Table 2 shows the relative increase in explained sum-of-squares (SS) as the order of the models increase. The total number of parameters is shown in the far right column of Table 2. The findings shown in Table 2 suggest that a model of order (3, 2, 3, 3) should be chosen. For the factors to be representative a good fit to **X** is paramount, hence 96.25% of SS explained seems appropriate in comparison with the models of higher orders. The number of parameters should be kept as low as possible in accordance with the principle of parsimony. Parsimonious models reduce the risk for fitting non-systematic trends (i.e., noise). Note that the model does not improve in fit when using more than two factors in the second mode. This is in concordance with the number of observations in the second mode: one cannot derive three or more orthogonal solutions in a mode that is only spanned by two variables. When moving from analysis of two-way data to multi-way data, we expect increased stability towards outliers. This is due to the increase in selectivity. Measuring many independent characteristics of samples will offer more scales on which to evaluate the goodness or suitability of the sample for modelling by the model in question. This is the so-called second-order advantage. The *N*-way PCA and the two-way PCA have the non-uniqueness in common, since factors from these two classes of models may be rotated by orthogonal transformations without affecting the fit.

Table 2
The explained sum-of-squares of the data as a function of the number of factors in the 4-way PCA model of sugar fluorescence measurements from the material in Fig. 3A–C

Model order	Expl. SS (%)	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

The sample-to-sample variation among the 47 samples is condensed in the factors in the first way. The three factors in the first way are depicted in Fig. 8. The factor denoted 1 describes a significant change of level in the samples. Factors marked 2 and 3 also reveal systematic behaviour. The factors describing the concentration levels are shown in Fig. 8. Fig. 8 reveals the behaviour of the intensities as a function of the excitation wavelength. However, it should be remembered that the factors are orthogonal. This makes interpretation with regard to chemical properties difficult. Fig. 8 shows the principal components describing the variation in the fourth way which relates to the emission wavelength. In the 54 ($= 3 \cdot 2 \cdot 3 \cdot 3$) element large core array the five most significant squared entries and their factor combinations are $2.04 \cdot 10^{10}$ (1,1,1,1), $2.27 \cdot 10^{9}$ (1,1,3,1), $1.20 \cdot 10^{9}$ (1,1,1,3), $9.92 \cdot 10^{8}$ (1,2,1,3) and $5.46 \cdot 10^{8}$ (1,1,2,2). From these values we see that no clear-cut factor combinations can be used for further data exploration. If the factors are properly rotated and the core correspondingly counter-rotated, a more simple structure of the core may be selected.

Thus, to improve the interpretability of the core array, the solution was rotated to yield maximum variance-of-squares of the core [A5]. After transfor-

mation, the variance-of-squares of the core array changed from $4.11 \cdot 10^{20}$ to $5.46 \cdot 10^{20}$, i.e. an increase of 32%. The variance-of-squares of the optimised core elements were $2.36 \cdot 10^{10}$ (1,1,1,1), $1.73 \cdot 10^{9}$ (1,1,2,2), $9.50 \cdot 10^{8}$ (1,2,1,3), $1.49 \cdot 10^{8}$ (1,2,2,3) and $1.03 \cdot 10^{8}$ (1,2,1,2). Note how the largest elements of the rotated core have absorbed variation described by the minor ones. Upon rotation the factors were as plotted in Fig. 8E–H. The variation expressed by the factors in Fig. 8 can be plotted in a more convenient way as in Fig. 3C (main text) where factor 2 and factor 3 are plotted against each other (corresponding to a PCA score plot). The conclusions drawn from this plot are presented in the main text.

Appendix B. parafac

B.1. Model

Consider a fluorescence data set with typical elements, x_{ijk}, where x_{ijk} is the intensity of the ith sample excited by light at the jth excitation wavelength and measured at the k'th emission wavelength. Theoretically, such data can be approximated as

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk} \qquad (2)$$

where a_{if} is the concentration of the fth analyte in the ith sample, b_{jf} is the relative emission emitted at wavelength j of analyte f, and c_{kf} is the relative amount of light absorbed at the excitation wavelength k of analyte f. This relation holds for diluted solutions, and if b_{jf} is (approximately) independent of c_{kf} [A8].

The fluorescence model is equivalent to the PARAFAC (parallel factor analysis) model initially proposed by R.A. Harshman [A9] and Carroll and Chang [A10]. Leurgans and Ross [A11], Leurgans et al. [A12], Ross and Leurgans [A13], and Nørgaard [A14] describe in detail the rationale for using PARAFAC models for modelling fluorescence data. The PARAFAC model is very closely related to ordinary two-way PCA, as exemplified graphically in Fig. 9.

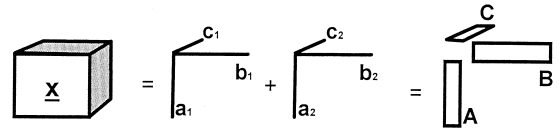


Fig. 9. A two-component PARAFAC model of the three-way array **X** (residuals omitted for brevity). The vector and matrix products to the right of the equal sign are equivalent to ordinary outer products, i.e. the first component represented by a1, b1, and c1 gives a rank-one part of the model of the same size as **X**, each element being a triple product $a_{i1} b_{j1} c_{k1}$.

Where two-way PCA gives one score and one loading matrix, one gets one score matrix and two loading matrices in a PARAFAC model of a three-way data set; one for each variable mode in the data. Therefore, a PCA model is a bilinear model, while PARAFAC is a trilinear model. The PARAFAC model is unique [A3,A15]. This means that if the model is appropriate for the data one need not impose orthogonality or other mathematical constraints to identify the model. Furthermore, instead of abstract latent variables, the true underlying phenomena are found. In this case it means that it is possible to estimate the underlying emission and excitation spectra and concentration profiles simply by decomposing the fluorescence data by a PARAFAC model.

B.2. Data

Sugar was sampled every 8 h during a campaign (approximately three months) at a sugar plant in Scandinavia, providing a total of 268 samples three of which were discarded in this study. Each sugar sample was dissolved in un-buffered water (2.25 g/15 mL) and the solution was measured spectrofluorometrically (Perkin Elmer LS50B). For every sample the emission spectra from 275–560 nm was measured in 0.5 nm intervals (571 wavelengths) at seven excitation wavelengths (230, 240, 255, 290, 305, 325, 340 nm). Laboratory determinations of the quality of the produced sugar were also available. These quality measures are ash content and colour. In addition, several automatically sampled process variables were available, including temperature, flow, and pH determinations at different points in the process. Typically these variables are very noisy and sampled at quite different rates.

A four-component PARAFAC model of the fluorescence data is appropriate in this case. However, for an unconstrained model a large portion of the loadings have negative areas at lower wavelengths. The reason for this is that 60% of the data are missing in this area, due to Rayleigh scattering. Therefore, the model is based on only one to four excitations below 360 nm. This causes some of the estimated emission loadings to be uncertain.

As the parameters of the PARAFAC model reflect concentrations and emission and excitation spectra, non-negativity seems a valid constraint to use in order to remedy this problem. One may infer that non-negativity should not be necessary, since the model should be identifiable even without using non-negativity. The adequacy of the unconstrained model, however, only holds to the extent that the PARAFAC model is correct for the data. There is a portion of the data that is missing due to Rayleigh scatter. Also, very likely a portion of the data that has not been set to missing values may be influenced by Rayleigh scatter to a slight degree, and therefore the data do not necessarily behave according to a trilinear systematic variation plus random noise. Furthermore, heteroscedasticity, quenching and other deviations from the model can cause the estimated parameters to deviate from strict non-negativity.

Very similar results are obtained by an unconstrained and a non-negativity constrained model. In the sample and excitation modes the loadings of the two models are highly correlated ($r = 0.99$). Further, the problems arising in the unconstrained model can be explained by the amount of missing values and model mis-specification.A four-component non-negativity constrained PARAFAC model results in the emission loading vectors displayed in Fig. 10a. The spectra seem mainly reasonable, but for one spectrum, the bump slightly above 300 nm seems to be more of a numerical artefact than real (Fig. 10b). This is plausible because many variables are missing in this area. One important aspect indicates that the spectrum should really be unimodal namely, that the most likely fluorophores in sugar (amino acids, simple phenols, and derivatives) have unimodal emission spectra due to the Kasha rule [A7,A8].

The above reasoning led to specifying a new model where all emission spectra were estimated under unimodality constraints and remaining parameters under non-negativity constraints. The estimated model was stable (Fig. 10c) and the estimated excitation spectra and relative concentrations did not vary considerably from that of the non-negativity constrained model. This strongly confirms the assumption that the cause of the artefact is mainly due to the amount of missing data in the specific region. It means that the unimodality is probably a valid constraint, and it also implies that unimodality is mainly necessary for improving the visual appearance of the emission loadings, hence enabling better identification of the underlying analytes.

Fig. 5B,C (main text) show selected estimated emission spectra, which fit well with the emission

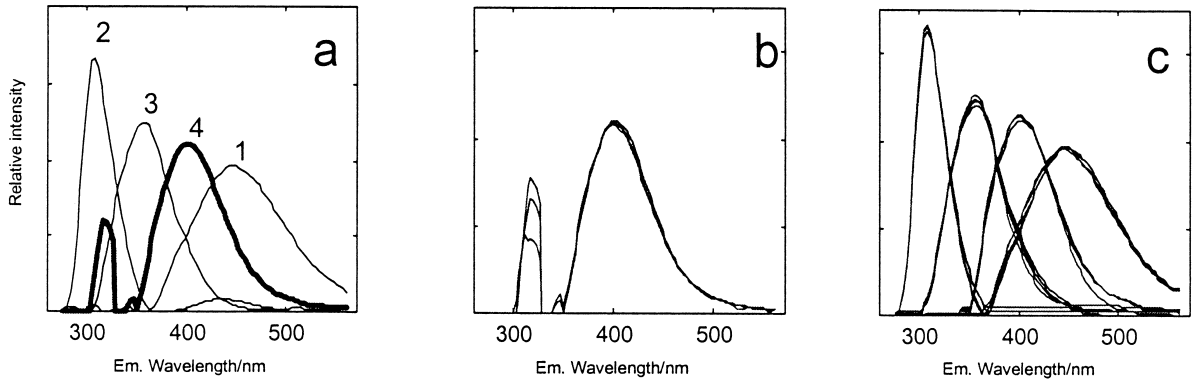


Fig. 10. Estimated emission spectra from fluorescence data. (a) Four spectra estimated using non-negativity. The 'suspicious' spectrum, 1, is marked with a thicker line. (b) Suspicious spectrum estimated from four different subsets using non-negativity. (c) Estimated spectra from different subsets using unimodality.

spectra of pure tyrosine and tryptophane respectively, two substances of known technological importance. The excitation spectra of tyrosine and tryptophane crudely agreed with those of the pure chemicals due to the limited number of seven excitation wavelengths employed with a gap between 255 nm and 290 nm. The spectra of tyrosine and tryptophane were acquired under quite dissimilar circumstances (pH 9, whereas the solutions used here was unbuffered) in experiments unrelated to this study. Still, the striking similarity with regard to the emission spectra confirms that the PARAFAC model is capturing chemical information. In order to verify with more confidence the identity of the underlying analytes we have confirmed the fluorescence signatures of the pseudospectra in column chromatography fractions of thick juice.

B.3. Using PARAFAC scores for modelling quality

The scores (**A**) of the model of the fluorescence data are estimates of concentrations. Initially, the correlation between the PARAFAC scores and the process variables was investigated. For some process variables there were almost no correlations, but for a large number excellent correlations were obtained. Examples of can be seen in Fig. 11.

A calibration model was made for predicting ash and colour from PARAFAC scores, The models for predicting ash content and colour of the sugar were excellent. The predicted values and the reference values are shown in Fig. 12. Note that, disregarding the fact that no cross- or test set validation has been performed, the prediction models are only based on four regression coefficients each, hence quite impressive. The above model substantiates, that it is possible to use fluorescence for on-line or at-line monitoring of sugar quality. This is important, as these parameters are currently only determined every 8 h and with a certain lag, as the laboratory analysis takes time.

The models described in this application based on fluorescence data are quite extraordinary. They give a direct connection between the raw material, process parameters and the final sugar quality (as defined by laboratory measurements defining the internal as well as the external consumer quality). As such,

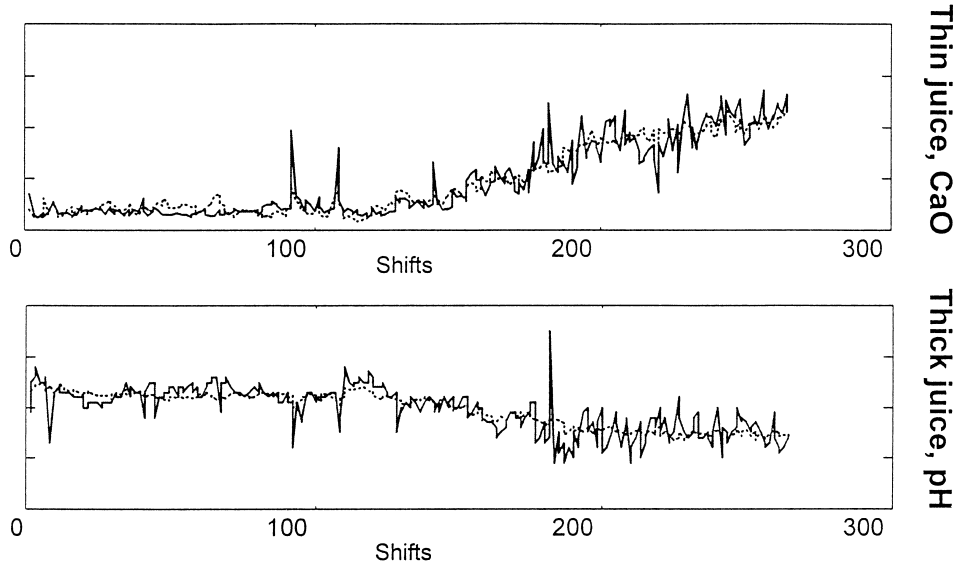


Fig. 11. Predictions of two important process variables. Unbroken lines are reference values. Notice the smoothing effect of the predictions based on fluorescence analysis of 8 h mean sugar samples representing one shift. The fitted values obtained using multiple linear regression (MLR) are shown. MLR was chosen, because the condition of the independent variables (265×4) is excellent, hence no problems arising from collinearity are expected.

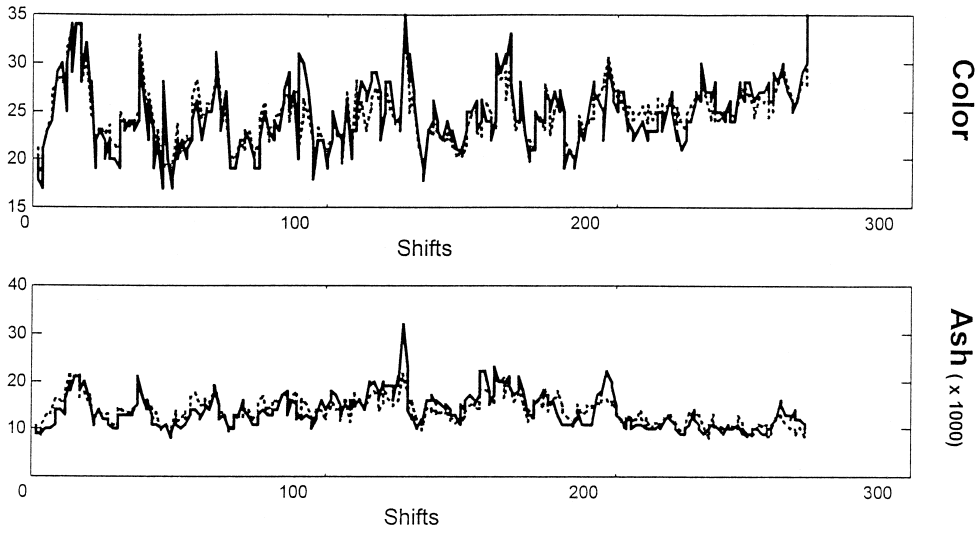


Fig. 12. Predictions of colour and ash from PARAFAC scores. Thick lines are reference values and thin lines the predicted values.

the conceptual idea behind the results reach far beyond the specific data treated here.

Appendix C. Principal variables (PV)

The PV model is based on exactly the same principles as is PCA and PLS. In PCA the first loading vector is the eigenvector corresponding to the largest eigenvalue of $(\mathbf{X}'\mathbf{X})^2$, while in PLS we look for the weight vector which is the eigenvector corresponding to the largest eigenvalue of $(\mathbf{X}'\mathbf{Y})^2$ [A16]. In PV we investigate exactly the same matrix products, but since we are interested in finding manifest variables and not latent factors we seek the largest diagonal elements of the matrices $(\mathbf{X}'\mathbf{X})^2$ (in the 'PCA'-case) or $(\mathbf{X}'\mathbf{Y})^2$ (in the regression case) corresponding to first principal variables. In PCA and PLS $\mathbf{X}$ is orthogonalised with the information described by the first latent factor. This also holds in the PV algorithm, where the $\mathbf{X}$ matrix is orthogonalised with the manifest variable: $\mathbf{X}_{\text{new}} = \mathbf{X} - \boldsymbol{v} \cdot \boldsymbol{k}$, where $\boldsymbol{v}$ is the column corresponding to the first principal variable and $\boldsymbol{k}$ is the loading.

Next the variables selected by the PV-algorithm are used in an ordinary multiple linear regression (MLR) with y as the dependent variable in order to develop a predictive model based only on the selected variables. We see here the synergistic combination of classical statistics (MLR) and new chemometric methods (principal variables). See the main text for applications (Table 1).

C.1. References to Appendices

[A1] P.M. Kroonenberg, Three-mode component models, A survey of the literature, Statistica Applicata, 4 (1992) 619–633.

[A2] L. Tucker, Some notes on three-mode factor analysis, Psychometrika, 31 (1966) 279–311.

[A3] R. Bro, PARAFAC: Tutorial and applications, Chemometrics and Intelligent Laboratory Systems, 38 (1997) 149–171.

[A4] R. Bro, Multi-way calibration. Multilinear PLS, Journal of Chemometrics, 10 (1996) 47–61.

[A5] R. Henrion and C.A. Andersson, Diagonality versus variance of squares for simple-structure transformations of N-way core arrays, Journal of Chemometrics, submitted.

[A6] C.A. Andersson, L. Munck, R. Henrion and G. Henrion, Analysis of N-dimensional data arrays from fluorescence spectroscopy of an intermediara sugar product, Fresenius Journal of Analytical Chemistry, 359 (1997) 138–142.

[A7] S.G. Schulman, Fluorescence and Phosphorescence Spectroscopy: Physicochemical Principles and Practice, Pergamon Press, Oxford, 1977.

[A8] G.W. Ewing, Instrumental Methods of Chemical Analysis, McGraw-Hill Int. Ed., New York, 1985.

[A9] R.A. Harshman, Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-mode factor analysis, UCLA Working Papers in phonetics 16 (1970).

[A10] J.D. Carroll and J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition, Psychometrika, 35 (1970) 283–319.

[A11] S. Leurgans, R.T. Ross, Multilinear models in spectroscopy, Statistical Science, 7 (1992) 289–319.

[A12] S. Leurgans, R.T. Ross and R.B. Abel, A decomposition for three-way arrays. SIAM Journal of Matrix Analysis and Applications, 14 (1993) 1064–1083.

[A13] R.T. Ross and S. Leurgans, Component resolution using multilinear models, Methods in Enzymology, 246 (1995) 679–700.

[A14] L. Nørgaard, A multivariate chemometric approach to fluorescence spectroscopy, Talanta, 42 (1995) 1305–1324.

[A15] R.A. Harshman and M.E. Lundy, PARAFAC: Parallel factor analysis, Computational Statistics and Data Analysis, 18 (1994) 39–72.

[A16] A. Höskuldsson, Prediction Methods in Science and Technology, Thor Publishing, Copenhagen, 1996.

References

[1] G. Chaitin, in: J. Horgan (Ed.), The End of Science — Facing the Limits of Knowledge in the Twilight of the Scientific Age, Helix Books, Addison-Wesley, Reading, MA, 1996, p. 230.

[2] P.C. Williams, K. Norris (Eds.), Near Infrared Technology in Agricultural and Food Industries, Am. Assoc. Cereal Chem., St. Paul, MN, 1987.

[3] H. Martens, M. Martens, in: K.I. Hildrum, T. Isaksson, T. Næs, A. Tandberg (Eds.), Near Infrared Spectroscopy—Bridging the Gap between Data Analysis and NIR Applications, Ellis Horwood, 1992, 1–10.

[4] J.D. Barrow, Pi in the Sky—Counting, Thinking and Being, Penguin Books, London, 1993.

[5] J.C. Frisvad, M. Norsker, Use of correspondence analysis partial least squares on linear and unimodal data, J. Chemom. 10 (1996) 677–685.

[6] M. Martens, E. Risvik, H. Martens, Matching sensory and instrumental analyses, in: J.R. Piggott, A. Paterson (Eds.), Understanding Natural Flavors, Blackie Academic and Professional, Chapman & Hall, London, 1994, 61–76.

[7] E. Jantsch, Technological Forecasting in Perspective, Organisation for Economic Cooperation (OECD), Paris, 1967.

[8] C. Hempel, Philosophy of Natural Science, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[9] I. Prigogine, I. Stengers, Order out of Chaos — Man's New Dialogue with Nature, Flamingo Paperbacks, London, 1984.

[10] Analytical Methods Committee, Uncertainty of measurement: implications of its use in analytical science, R. Soc. Chem. London, Analyst 120 (1995) 2303–2308.

[11] R.P. Singh, J.R. Banga, Recent advances in food process optimization, Chem. Industry, London 13 (1994) 511–514.

[12] C.W. Snyder, H.G. Law, J.A. Hattie, Overview of multimode analytic methods, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald, Research Methods for Multimode Data Analysis, Praeger Scientific, New York, 1984, 2–35.

[13] S.B. Engelsen, S. Pérez, Internal motions and hydration of sucrose in a diluted water solution, J. Mol. Graph. Modelling 15 (1997) 122–131.

[14] L. Munck (Ed.), Fluorescence Analysis in Foods, Longman Scientific and Technical, Harlow, England, 1989.

[15] L. Nørgaard, Classification and prediction of quality and process parameters of beet sugar and thick juice by fluorescence spectroscopy and chemometrics, Zuckerindustrie 120 (1995) 970–981.

[16] L. Tucker, Some notes on three-mode factor analysis, Psychometrika 31 (1966) 279–311.

[17] C.A. Andersson, L. Munck, R. Henrion, G. Henrion, Analysis of *N*-dimensional data arrays from fluorescence spectroscopy of an intermediary sugar product, Fresenius J. Anal. Chem. 359 (1997) 138–142.

[18] R. Bro, PARAFAC: Tutorial and applications, Chemom. Intell. Lab. Systems 38 (1997) 149–171.

[19] R.A. Harshman, M.E. Lundy, PARAFAC: parallel factor analysis, Computational Stat. Data Anal. 18 (1994) 39–72.

[20] L. Munck, Man as selector—a Darwinian boomerang striking through natural selection, in: J. Aa. Hansen (Ed.), Environmental Concerns, Elsevier, London, 1991, 211–227.

[21] M.W. Eysenck, Principles of Cognitive Psychology, Lawrence Erlbaum Associates Hove., UK, 1993.

[22] R.F. Madsen, W. Kofod Nielsen, B. Winstrom-Olsen, T.E. Nielsen, Formation of colour compounds in production of sugar from sugar beet, Sugar Technology Reviews, Elsevier, Amsterdam, 6 (1978/79) 49–115.

[23] S. Wold, Chemometrics, what do we mean with it, and what do we want from it? InCINC'94 (1994), http://www.emsl.pnl.gov:2080/docs/incinc/chem_phd/SWdoc.html.

[24] P. Gould, Destructive retrieval in the realm of three-mode thinking, in: H.G. Law, C.W. Snyder, J.A. Hattie, R.P. McDonald, Research Methods for Multimode Data Analysis, Praeger Scientific, New York, 1984, 536–591.

[25] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley Publishing, 1977.

[26] J. Ziman, Reliable Knowledge - An Exploration of the

Grounds for Belief in Science, Cambridge Univ. Press, Cambridge, 1978.
[27] T.S. Kuhn, The Essential Tension - Selected Studies in Scientific Tradition and Change, The Univ. of Chicago Press, Chicago and London, 1977.
[28] H. Martens, Multivariate Calibration: Combining harmonies from an orchestra of instruments into reliable predictions of chemical composition, Proceedings, International Statistical Institute, Tokyo, 1987, 1–18.
[29] E.L. Post, Absolutely unsolvable problems and relatively undecidable propositions: Account of an anticipation (1941), printed in the book The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions, Raven Press, New York, 1965.