Pergamon

Ann. occup Hyg., Vol. 40, No. 2, pp. 145–169, 1996
Elsevier Science Ltd
Copyright © 1996 British Occupational Hygiene Society
Printed in Great Britain. All rights reserved
0003–4878/96 $15 00+0 00

0003–4878(95)00070–4

CHEMOMETRICS IN OCCUPATIONAL HYGIENE—HOW AND WHY! A PICTURE CAN TELL MORE THAN A THOUSAND WORDS AND FIGURES!

Erik Bye

National Institute of Occupational Health, P.O. Box 8149 Dep., 0033 Oslo, Norway

(*Received in final form* 5 *June* 1995)

Abstract—In this introductory article the author argues for an increased use of a multivariate analytical approach to the complex problems encountered in occupational hygiene. Relations between exposure at the work place and reported health effects are mostly so complicated and depend on so many factors that methods other than the traditional statistical techniques should be applied. *Chemometrics* is a field within chemistry where mathematics, statistics and modern computer technology are used to perform multidimensional data analysis. Graphical plots are extensively used to extract the most relevant information from the measurements. The possibility of performing *soft modelling* through pattern recognition and multifactorial regression analysis will simplify the management of large data sets. A 'metric' philosophy is introduced to describe similarity and dissimilarity among many objects characterized with many variables. This article emphasizes the use of *principal component analysis* and *partial least-squares* regression for such purposes. Application of the *SIMCA* method for classification of objects is also described. These methods are not dependent upon a priori formulated hypotheses, as in the classical modelling techniques. Instead of being restricted to accepting or rejecting previously formulated hypotheses, these methods may lead to new insights and unperceived features of a complex problem. The application of such exploratory methods may produce new hypotheses and further investigations are necessary to confirm or discard any 'new' chemometric findings. Copyright © 1996 British Occupational Hygiene Society. Published by Elsevier Science Ltd.

INTRODUCTION

The occupational hygienist is faced with complex problems. As efforts are intensified towards a description and analysis of the total work environment there is a need for a multivariate approach. This is particularly so in studies where occupational health problems are related to the complex hygienic quality at the work place. There are several analytical methods that can handle large, complex data. Some of these methods apply graphical presentations and plots extensively to simplify interpretation.

In an earlier paper (Schneider *et al.*, 1993) a favourite plot of *chemometrics* appeared in this journal for the first time. In that article Schneider *et al.* (1993) used a *variable loading plot* to interpret correlations between the respirable fibre concentrations and various properties of vitreous fibre types, including the content of oil from the production. Such a variable loading plot may be used to find the most relevant or important variables in an investigation. The recognition of variables with low or no importance at all is quite helpful in cases with many variables. In addition the correlations among the variables can be depicted. A variable loading plot is schematically shown in Fig. 1. The variable loadings (X1 ... X6) are given with co-ordinates in the range ± 1 and related to two axes, t_1 and t_2. These are calculated so

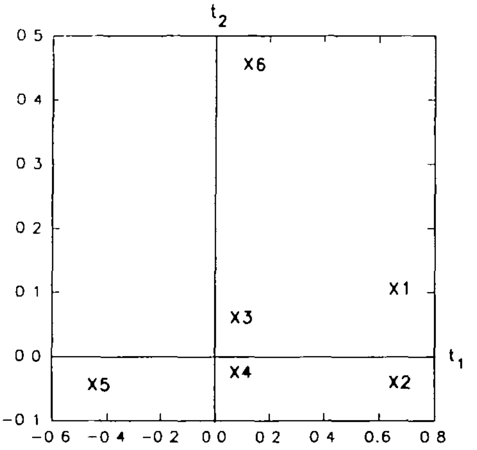

Fig. 1. Variable loading plot for a data set with six vanables (X1 . . . X6). The loadings are referred to axes (principal components) which account for most of the variance. (See text for further details on the calculations.)

that they account for most of the variance of the data set. (These axes are the so-called *principal components*. A more detailed description of this will be given below.) Variables far out along the axes, that is variables with large loadings, are important (X1, X2, X5, X6). Less important variables are located near the origin (X3, X4). The axes are calculated to be orthogonal and variables lying close together (X1 and X2) are highly correlated whereas the variables X1 and X6 are uncorrelated. Variable X5 is negatively correlated with X1 and X2 [Quite recently another chemometric article was presented in this journal, showing a multivariate calibration technique for aerosol analysis (Bye, 1994). This application is discussed in *Example* 3 in this paper.]

Chemometrics is a field within chemistry where mathematics and statistics are used to extract the most important information from large amounts of chemical data (Sharaf *et al.*, 1986; Massart *et al.*, 1988). The branch of chemometrics includes: (a) signal processing (Massart *et al.*, 1988); (b) experimental design (Deming and Morgan, 1987); and (c) multivariate data analysis, for example *classification* and *prediction* (Wold *et al.*, 1983). I will focus on the latter applications in this introductory article and describe why and how *principal component* and *multivariate regression* methods could be used in occupational hygiene.

Nature is a complex system of many interrelated factors. Thus, in order to understand this it is only reasonable that we have to study and consider more than one variable at the time. Analogously the interplay between work and human health is multidimensional and should be examined with multivariate methods. In its broadest sense, the fundamentals of chemometrics are indeed based on multidimensional philosophy and strategy. Chemometrics is most commonly associated with *principal component analysis* (*PCA*) for pattern recognition purposes (Kowalski and Wold, 1982; Wold *et al.*, 1987) and *partial least squares* (*PLS*) for *multivariate regression* or *multivariate calibration* purposes (Martens and Næs, 1989). These methods represent a *soft modelling* approach to the complex questions encountered in occupational hygiene and health. They may contribute with a quantitative description, a systematic organization and a deeper understanding of relevant information in a data set. Extensive use of graphical plots for interpretation and presentation makes chemometric methods a powerful analytical instrument in research fields with large amounts of data, especially if complex and disorganized.

PRINCIPAL COMPONENT ANALYSIS (PCA)

The main concept of chemometrics is that similarity or dissimilarity between *objects* (samples) can be measured by the 'distance' between the objects in a multidimensional measurement space. An object characterized by *p variables* is looked upon as a *p*-dimensional vector in the space represented by the *p* variable axes. Numeric values of the variables, that is the *parameters*, determine the position of the objects in this space (Wold *et al.*, 1987). This 'metric' representation is easily visualized in two dimensions as shown in Fig. 2.

Example 1*: hygienic quality of a workplace*

The hygienic quality at the same work operation is compared in two different foundries A and B. This example has been constructed to illustrate the applicability of the basic chemometric methods for occupational hygiene studies. Respirable dust (RD), the quartz content (Q), the Fe concentration (Fe), the CO level and the temperature (T) have been measured. Personal sampling for 10 workers are reported in Table 1.

Each worker (sample) is taken as an object characterized by the five variables in Table 1. Using respirable dust and quartz as the co-ordinate axes, the data in Table 1 may be plotted as in Fig. 2. All the workers are represented as points in this two-dimensional data space. Inclusion of the CO exposure is illustrated by the three-dimensional diagram in Fig. 3. Such two- and three-dimensional pictures may be very useful in looking for a pattern in the data structure, that is any systematic

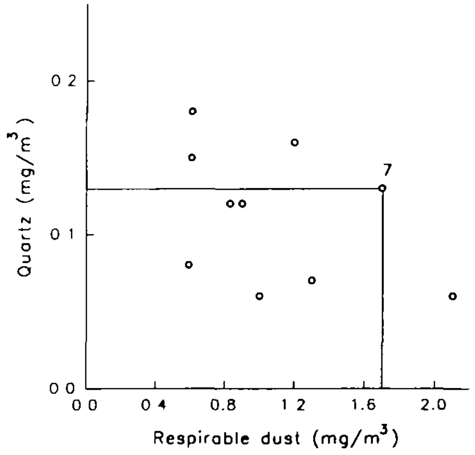

Fig. 2. Two-dimensional representation of the 10 foundry workers at two different plants (see Table 1) exposed to respirable dust (RD) and quartz (Q). Each worker (object) is characterized by the two variables RD and Q.

Table 1. Time-weighted average exposure parameters for 10 workers at the same work operation in two different foundries, A and B. Workers 1, 2, 3, 7 and 10 belong to plant A

Operator	Respirable dust (mg m^{-3})	Quartz (mg m^{-3})	Fe (mg m^{-3})	CO (ppm)	Temperature (°C)
1	0 83	0.12	0.20	7	26
2	0.61	0.15	0 10	15	29
3	0.61	0.18	0 15	28	34
4	0.59	0.08	0.05	42	29
5	1.30	0.07	0.20	7	25
6	2.10	0.06	0.40	19	22
7	1.70	0.13	0.25	2	20
8	1.00	0.06	0.27	23	27
9	0.90	0.12	0.15	30	30
10	1.20	0.16	0.30	25	24

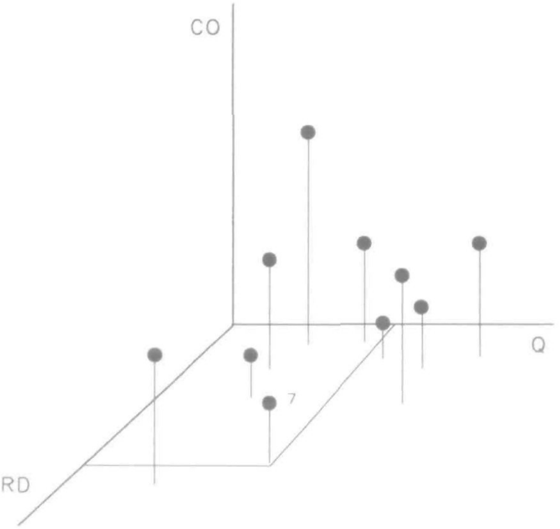

Fig. 3. Three-dimensional representation of the 10 foundry workers from two different plants (see Table 1) exposed to respirable dust (RD), quartz (Q) and CO. Each worker (object) is characterized by the three variables: RD, Q and CO.

variation among the data points. However, at this stage you have to decide on the viewpoint and the view direction to have the optimized orientation.

The principles of a 'metric' representation of samples may also be extended and applied for higher dimensions. However, with more than say, four variables characterizing the samples, there are difficulties with perceiving the relations and covariance between the objects and variables. A fourth dimension can be introduced in the three-dimensional plot in Fig. 3. This can be done with a marker to show the level of this variable for each point. In Fig. 4 the Fe parameter is plotted in increasing order, according to the concentrations. The interpretation is, however, not straightforward!

With more than four variables describing the samples, it is a complex task to extract the data structure and most important information of the investigated

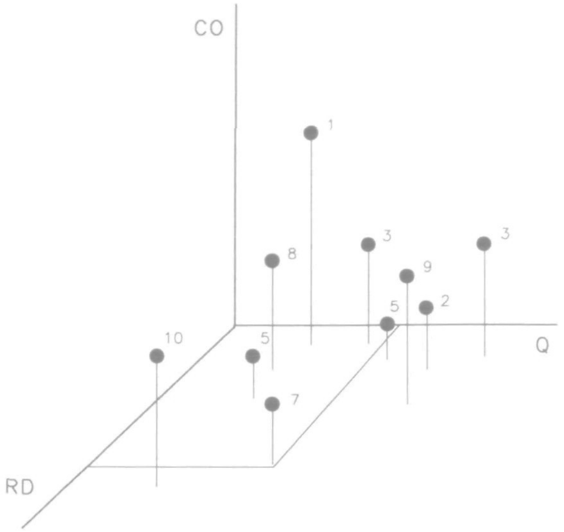

Fig. 4. A three-dimensional scatter plot of the 10 foundry workers with an indicator representing the level of Fe exposure in increasing order as the fourth dimension

system. Questions such as: What are the relations between the objects? Are there more than one class of objects? Which variables are most important? Which variables are most appropriate for discrimination between the classes? Are there variables of no importance? Which variables are correlated?—are frequently encountered when we examine correlations between health effects and exposure at work.

PCA is one possible way to study such questions. A large number of objects (samples) described with many variables (properties) can be handled simultaneously with this analytical method. The observations are organized as an **X** matrix (table), with m rows, one for each object and p columns, one for each variable, as illustrated in Fig. 5.

The main purpose of the principal component analysis is to calculate a smaller number of new variables (A). They should describe the phenomenon being studied quite well. These $A<p$ variables are the *principal components* (PCs) and are linear combinations of the original variables. They are calculated by a least-squares method (Wold, 1982) as orthogonal components, successively describing decreasing amounts of the variance of the data set. A principal component may be presented geometrically as shown in Fig. 6. Objects are seen as a point swarm (here two dimensional) and the first principal component is calculated as the best possible linear fit to the data points. This may be compared to the straight line in linear regression, except that we in the PC case have errors in both the original x and y variables. The projection of the ith object onto the principal component defines the position of this object t_{1i} along the component, that is the *object score*. The variance accounted for by the first principal component is subtracted from the total variance of the data matrix. Principal component 2 (PC2) is the direction in space that is orthogonal to PC1 and is associated with the maximum residual variance of the data

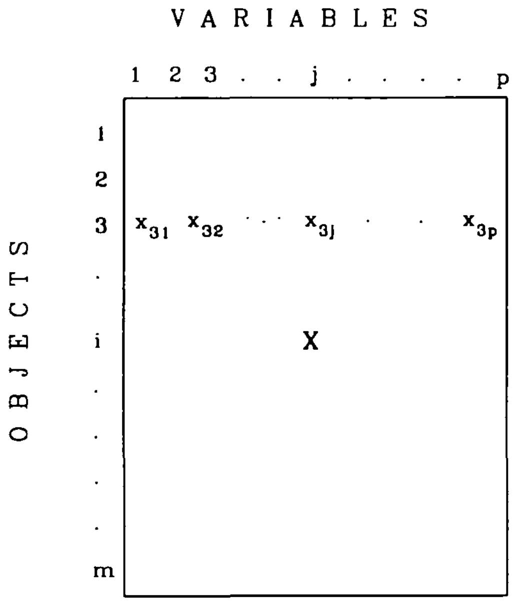

Fig. 5. A data table X organized on the basis of *m* objects (the rows) characterized by *p* variables (the columns).

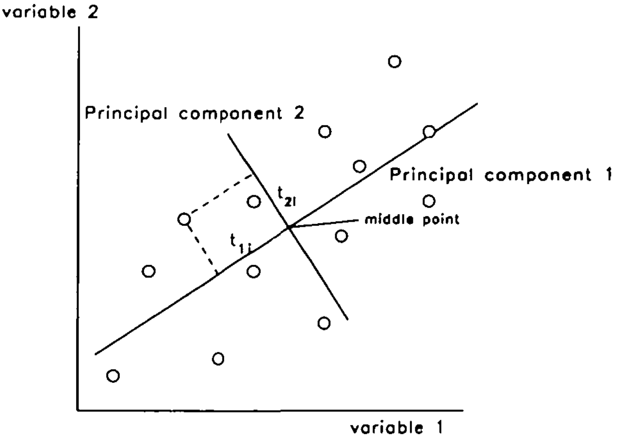

Fig. 6. Principal component analysis. The principal component is a linear combination of the two variables 1 and 2 such that it describes the maximum of the systematic variation in the data set. The first two orthogonal principal components (PC1 and PC2) are shown All calculations are based on the middle point of the data set Object scores t_{1i} and t_{2i} are the projections of point *i* (object *i*) onto the respective components.

set. For practical reasons during the calculations (and plotting), the mean value of all the objects is used as origin. This is the average of each column in the table in Fig. 5. Therefore the two orthogonal principal components (PC1 and PC2) represent nearly 'a rotation' of the original co-ordinate axis.

Object scores, given as t_{1i}, t_{2i} in Fig. 6, give the position of the *i*th object in the 'new' co-ordinate system defined by the principal components. *Variable loadings* define the directions of the principal components relative to the original axis. Mathematically they are the direction cosines for the principal components relative to the original variables. They represent the contribution from each variable to the principal component. In two and three dimensions the PCs may be looked upon as a 'rotation' of the original axis. For higher dimensions we will have a 'new data space' with orthogonal axis, describing most of the variance of the data set. Very often a few components describe most of the systematic variability in the data. These are new variables and they thus reduce the dimension of the problem.

Through the calculation of principal components we have a 'new' one-, two-, three- or *A*-dimensional model space, depending upon the number of significant components (*A*). The projections of the objects onto the principal components represent the positions in the 'new' co-ordinate space. A plane defined by the first two principal components will thus function like a two-dimensional window into the *p*-dimensional data space. These two components hold most of the variance of the data set. We can see the positions of the objects at their score values. Similar objects tend to cluster together whereas different objects lie further apart. This is shown in Fig. 7 with the workers from the two similar workplaces in foundries A and B. Information is taken from Table 1, each worker being characterized by five exposure parameters. The first two principal components account for 82% of the variance. The groups of workers are seen, although the variation within group II is quite large. Groups I and II are actually plants A and B. However, no direct information about the work place origin has been introduced in the data analysis. This separation is less obvious in Figs 2 and 3. In this way, the object score plot for the first two principal components has provided us with a two-dimensional window into the five-dimensional data space. Maximum separation and perhaps also the maximum discrimination among the objects has been obtained by PCA.

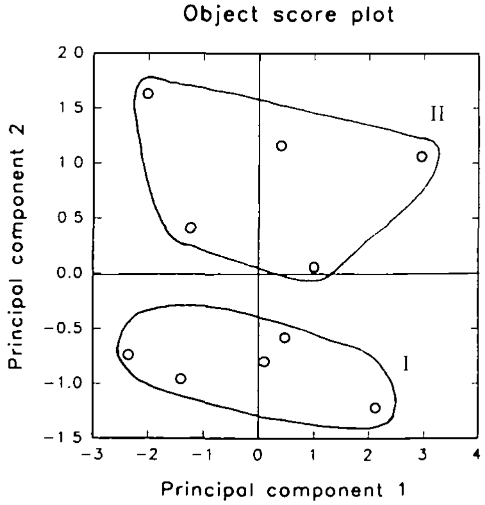

Fig. 7. Object score plot for the first two principal components, describing 61.5 and 20.5% of the variance, respectively. Group I and II are identical to plant A and B. No information of the sample origin has been introduced during the data analysis.

Projecting the loadings onto the same principal components gives the *variable loading plot*, shown in Fig. 8. These loadings describe the contribution of the original variables to the respective principal component. Variables far out along the axis are important for the variance along that component, whereas variables near the origin are less important. Variables lying close together in such a plot are correlated, whereas variables separately situated in the plot are uncorrelated. Such variables contribute independently to the two different (orthogonal) components.

All five variables in Fig. 8 are important. The respirable dust (RD) and the Fe measurements are highly correlated since they lie close together. The temperature (T) and the respirable dust concentration are inversely correlated since they lie close to a line through and on opposite sides of the origin. Quartz concentration (Q) and the respirable dust are not correlated since the two variables Q and RD mainly contribute separately to the two orthogonal principal components. A simultaneous inspection of the object scores (Fig. 7) and the variable loadings (Fig. 8) will give information about the distribution of the objects and the corresponding influence of the variables. This can be combined in *a biplot*, see Fig. 9, where you can see the data structure or *pattern* of the objects and the variable influence. Variables far out along one axis are important for the distribution of objects along the respective axis.

The variables RD, Fe and T are most important for principal component 1. Workers with high exposure to respirable dust (RD) and Fe are located to the right in Fig. 9. Furthermore, the variables Q and CO are important for principal component 2, and workers with the highest exposure to quartz are found in the lower part of Fig. 9. All the variables are important for the two principal components because none of them are located close to the origin, that is they all have high loadings. By contrast, this is the case in the schematic illustration in Fig. 1, with a low importance of variables X3 and X4.

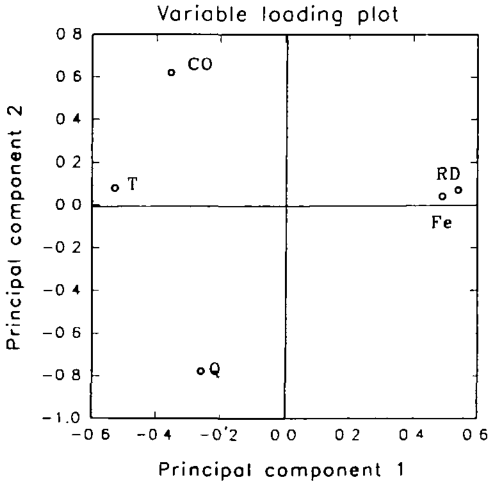

Fig. 8. Variable loading plot for the first two principal components. See Table 1 for explanation of the variable names. Respirable dust (RD) and Fe concentration are correlated and they are important for component 1. Quartz (Q) and the CO level is important for component 2

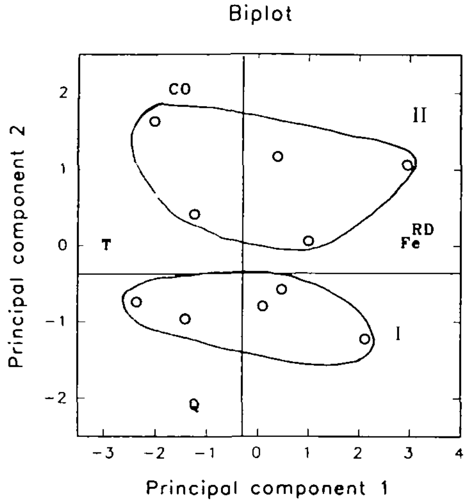

Fig. 9. A biplot of the object scores (see Fig. 7) and the variable loadings (see Fig. 8) for the first two principal components. This plot simplifies the interpretation of the variable influence to the object pattern.

Through the calculation of the two principal components, as much as 82% of the systematic variability in the exposure data is modelled. The two PCs are linear combinations of the original five variables. They are two new variables and the dimension of the problem (work place quality) has thus been reduced. The object score plot in Fig. 7 shows that there is a systematic difference between plant A and B along the second component. This is seen by the position of the CO and quartz variables. And remember, no information about the work place origin of the workers was introduced during the principal component analysis. An extensive use of object score plots and variable loading plots represents the possibility to display and interpret the most important information in the data matrix. This is the relation between the objects and the variables. Models with more than two principal components need 'windows' with higher order components. This is a most powerful advantage with this analytical method of chemometrics and provides the analyst with graphical tools that simplify the interpretation of highly complex data sets. Associations between health outcomes and exposure pattern and occupational hygiene standard as studied by multivariate methods will be discussed in Example 4.

Sampling errors in occupational hygiene are by far the largest problem when taking measurements. Variations by factors of 2–3 can be observed for personal samples on different lapels of the same worker. Larger variations may be seen in exposure levels, even for time-weighted averages of 8-h sampling. Such variation may be due to spatial, locational and job type changes. These experimental variations could well exceed any differences found between factories (Kromhout *et al.*, 1993). Useful and important correlations may not be derived, but chemometrics should help in seeing this variability. It is beyond the scope of this introductory article to discuss this in any detail. However, many samples from each workplace should at least indicate the dominant characteristics.

THE SIMCA METHOD

A principal component model of a class of objects may be used to examine whether an unknown object belongs to the class or not. The model is constructed by *a training* or *calibration set*. Any *unknown objects* included in the calculation and in the object score plot will show if the *class membership* is correct. The exploratory part of the data analysis to establish the training set may be called *unsupervised learning*. Application of the model for testing of unknown objects is called *supervised learning*. *Class boundaries* and a modified *F*-test (based on the residual variances) can be used to quantify class membership. Furthermore, separate clusters may be described by disjoint models, and the methodology includes criteria to evaluate the *discrimination power* of the variables for the separation between classes. These ideas were developed as the *SIMCA* method in the 1970s (*S*oft *I*ndependent *M*odelling of *C*lass *A*nalogy) by Wold and his group (Wold and Sjöström, 1977). It will be beyond the scope of this introductory article to describe all these features of the SIMCA method in detail. Classification of human skeletal muscle fibres provides a suitable illustration of the method and some of the features (Bye *et al.*, 1989). This example has been selected because no simple and relevant investigation has been reported on a SIMCA application within occupational hygiene.

Example 2: human skeletal muscle fibres

Human skeletal muscle fibres are normally classified by visual evaluation or by measurements of the optical density of histochemically stained muscle sections which is closely correlated to the enzyme activities of the fibres, and is typical for the fibre types. A data set comprising several stained fibre types has been reanalysed by the SIMCA method. We would like to consider the following questions: (a) How many distinct fibre types are present? (b) Which staining techniques are the best? (c) How many staining techniques are necessary to separate the three fibre types? The data set contained the same 32 fibres in 12 different thin muscle sections, each section stained by a different colour technique. Thus altogether 12%-transmission light measurements were used as the variables, 12 for each object (fibre). A principal component analysis gave a model with two components, accounting for 95% of the variance. The object score plot is shown in Fig. 10. Three distinct fibre types are easily recognized (Type I, Type IIa and Type IIb) in accordance with the visual inspection. However, three fibres (Nos 1, 2 and 3 in Fig. 10) deviate somewhat from the main cluster regions.

Figure 11 shows the corresponding variable loading plot. Except for variables 4, 10 and 11 all the others are important for the two PCs since they are located far out along the axes. Considering the pattern in Figs 10 and 11 together, we can conclude that Type I fibres have high values for the variables to the right in the loading plot (along component 1). Similarly, Type IIa fibres have high values for variables 3, 9 and 8 whereas the opposite is the case for Type IIb fibres. Variables 6 and 8 are most important for component 1 and 2, respectively and the two-dimensional scatter plot in Fig. 12 illustrates the separation ability of the two variables. This scatter plot confirms the impact of the variables to the object structure, since the patterns in Figs 10 and 12 are quite similar.

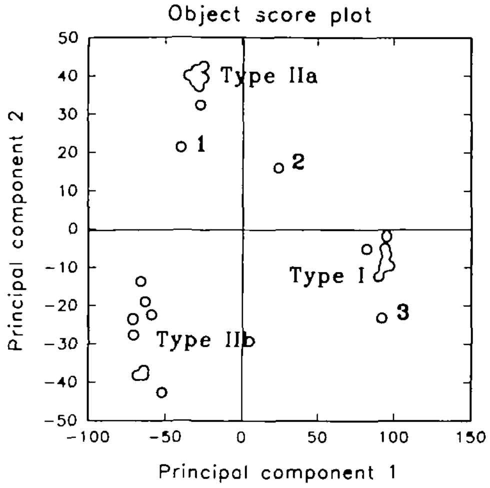

Fig. 10. Object score plot for the first two principal components, describing 81 and 14% of the variance, respectively. Three muscles fibre types are recognized, in addition to three separate objects (Nos 1, 2 and 3).

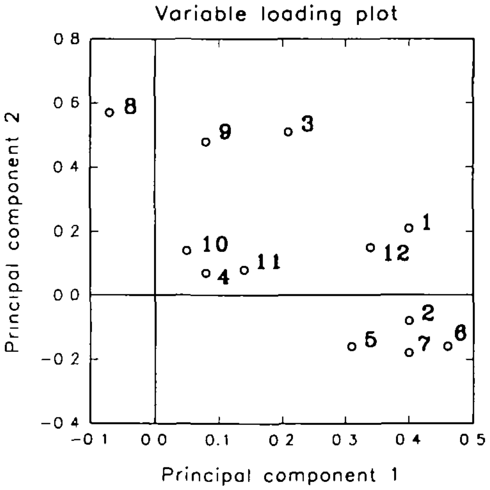

Fig. 11. Variable loading plot for the first two principal components. The variables are different colouring techniques for the staining of the muscle fibres.

When we know that object No. 3 (see Fig. 10) belongs to Type I, we can see from Fig. 12 that the value of variable 8 is somewhat low for this object. This is in accordance with object No. 3 lying opposite to variable 8 in Fig. 11. From Fig. 11 we can see that variables 1 and 12 are correlated. So are variables 2, 5, 6 and 7, since they are located closely to each other in the loading plot. The correlation between variable 1 and 12 is shown in Fig. 13.

With separate (disjoint) PC models for each fibre type, and class boundaries associated with each group, outliers may be identified and quantified (Wold *et al*,

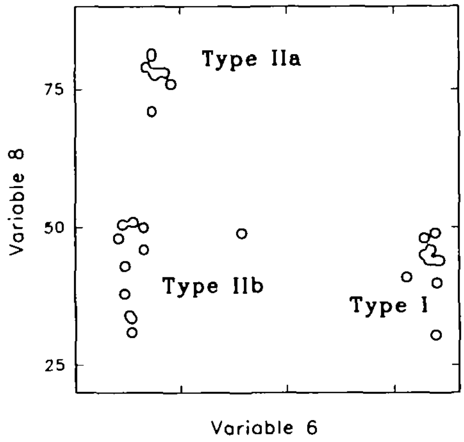

Fig. 12. Scatter plot of the muscle fibres to illustrate the separation ability of the two colouring techniques, variables 6 and 8. The variables were selected from the loading plot (Fig 11), as most important for components 1 and 2, respectively.

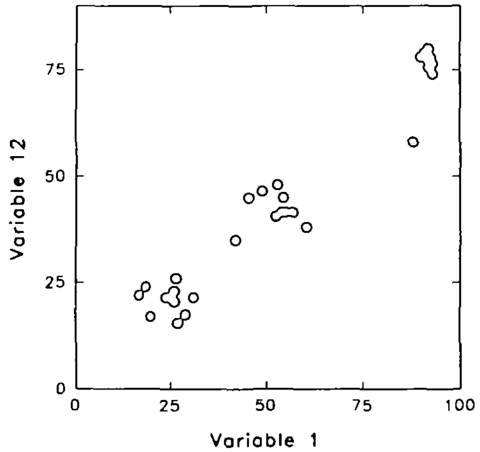

Fig. 13. Scatter plot of variables 1 and 12, to show the correlation as observed in the loading plot in Fig 11.

1981). Class boundaries are calculated from the residual variance and can be used to determine whether clusters are 'really separate groups'. Table 2 gives the class distances (CD) for the three clusters, along with the normalized distances between the clusters and the three separate objects, see Fig. 10. A class distanced (CD) of 3.0 means that the distances between the classes are three times the mean standard deviation of the middle point of the classes.

The fibre located between Type IIa and IIb (object No. 1 in Fig. 10) was originally classified as a Type IIa fibre. However, a location between IIa and IIb and outside the class boundaries for the two classes indicates a 'new' fibre type. Actually a Type IIab fibre has been suggested by others. The fibre located between Type I and Type IIa (object No. 2) is actually a fourth type, Type IIc. It is included in this example to illustrate the presence of an 'outlier'.

Table 2. Class distances (CD) and rejection criteria for membership determination. A class distance higher than 3.0 is considered to be a significant separation. The rejection criteria for 'outlier detection' and the object distances to the class is based on a simplified F-test. A significance level of $P = 0.01$ is used for the class boundaries

	Class distance		Object distance			
Fibre type	IIa	IIb	1	2	3	Rejection criteria of the class (for objects)
I	10.6	10.8	39.4	23.3	6 3	5 9
IIa	—	6.3	6 6	18.8	—	3.6
IIb	—	—	13 6	31.7	—	4.3

Table 3. Discrimination power (DP) of the variables for discrimination between the classes (fibre types). A numerical value larger than 3.0 is considered to be appropriate for separation between two classes

	IIa		IIb	
Fibre type	Variable	DP	Variable	DP
I	7	30.5	6	33.3
	6	29.1	7	31 1
	2	20.9	12	21.1
IIa			3	8.9
			12	6.9
			1	6.5

Table 3 refers the discrimination powers (DP) of the variables, which give a quantitative measure of the separation power between two classes. According to Wold *et al.* (1981) a DP of 3.0 is significant to consider the classes as distinct, that is the variables have significant separation ability. This means that the class distance is three times the intraclass variation (estimated standard deviation), calculated from the residual variance of the class. According to Table 3 variables 7, 6 and 3 are the most appropriate variables to discriminate among the three fibre types. According to the variable loading plot in Fig. 11, variables 7 and 6 are correlated. As shown in Fig. 15 the two variables 6 and 3 give a good separation between the clusters. In practice the most important variables may be combined according to loadings and the discrimination powers to conclude on the most appropriate variables for discrimination. (The small differences in the loadings between the variable pairs 6 and 8, and 7 and 3 are of no significance here; compare Fig. 12 and Fig. 14.)

PARTIAL LEAST SQUARES (PLS)

Signal processing, experimental design and principal component analysis are included in chemometrics, although these methods have been well known for decades. However, along with the increased use of principal component analysis (*PCA*) and the development of the SIMCA classification technique, the *PLS* method, that is the *partial least-squares* regression, was introduced (Wold *et al.*, 1983). This concept was based on the early work of Wold on 'systems under indirect observation' (1966, 1975, 1982). The technique provides the analyst with a two-block

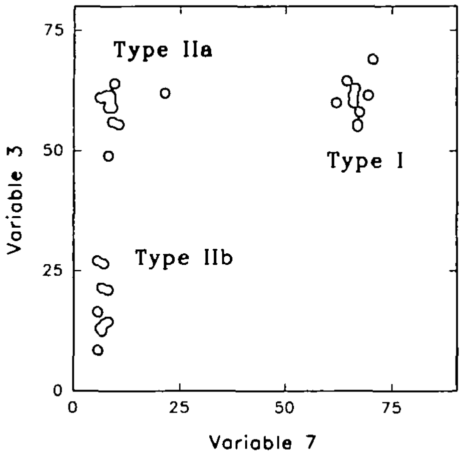

Fig. 14. Scatter plot to illustrate the two variables with the largest Discrimination Powers along the first two principal components.

regression method that is also based on principal component models. However, in contrast to *multiple regression* (MR) information of the *response* variable (to be predicted) is also taken into account in the construction of the model. If we call the *independent* variables the **X** matrix, similarly to the previous paragraph, the *dependent* or response variables are called the **Y** matrix. In practice, the data set is organized as two tables, the *x* and *y* parameters. Each object is again looked upon as a *p*-dimensional vector, but now with one or more (*k*) intrinsic properties (the *y* parameters), as illustrated in Fig. 15. These may be properties that are difficult, expensive or time-consuming to measure.

An interdependent principal component model is constructed based on the known objects, that is the *training* or the *calibration set*. The model is optimized to account for the variance in **Y**. If there is systematic variation between the variables in the **X** and **Y** matrix, these correlations are built into the PLS model. These components are 'quite similar' to PCs and may be called *PLS components*. They are also linear combinations of the original variables. The predictive ability of the model is evaluated with a test set and the model is then used to predict the *y* values of the unknown objects. Test sets with measured *y* values are used for comparison with the predicted *y* parameters only and not included in the model.

For calibration purposes this technique is called *multivariate calibration* (Martens and Næs, 1989). Spectroscopic data from a large spectral range (or the whole spectrum) are used in the **X** matrix instead of only one single wavelength. In this approach PLS may be used to determine the concentrations of one or several chemical compounds in complex mixtures. Here the traditional univariate calibration often fails. The training set contains known concentrations, that is the *y* values, of the compounds of interest. With one response variable the method is called *PLS*1, whereas *PLS*2 is the acronym for regression on more than one effect variable.

Another approach of PLS is the QSAR application, that is the *Quantitative Structure–Activity Relationships*. A frequent problem may be to correlate biological

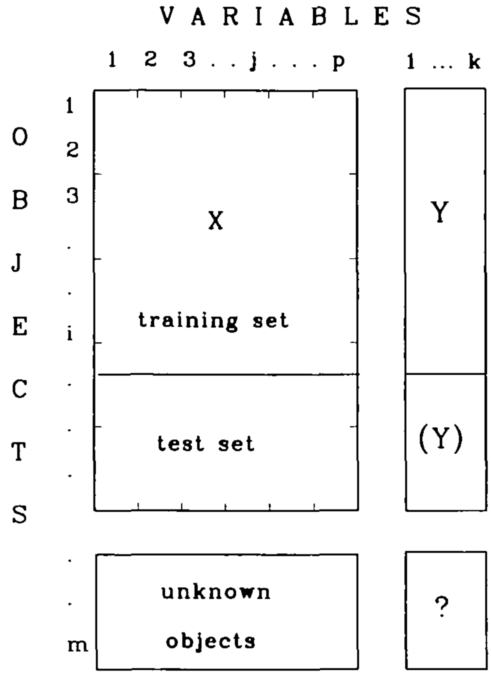

Fig. 15. The general organization of data matrices in PLS analysis. The **X** matrix (the measured independent parameters) and the **Y** matrix (the dependent parameters to be predicted) are set up in two blocks. The data set consists of three subgroups of objects: the training set, the test or validation set and the unknown objects. *m* objects (the *rows*) are characterized by *p* independent variables (the *columns*), whereas the effect table (Y) may consist of one or more additional dependent variables (*k*). The *y* parameters of the test set are used for evaluation of the constructed PLS model. *y* parameters of the unknown objects (given as ?) are predicted.

activity to chemical properties or to the chemical structure. Physico-chemical properties of the compounds are introduced as the **X** matrix whereas the *y* variables contain the corresponding biological effect parameters. QSAR studies with PLS have most frequently been used on drug design and studies of toxic effects in the environment (Dunn, 1989; Lundstedt, 1991). However, laboratory experiments on selected assays to elucidate toxic effects in humans have also been investigated by PLS models (Nordén *et al.*, 1978; Wold *et al.*, 1985). A study of health effects relevant for industrial handling of chemicals was recently published (Eriksson *et al.*, 1994) and the two applications of PLS, *multivariate calibration* and *QSAR*, will be described in the examples below.

Example 3: quantitative determination of silica mixtures

When we are performing a traditional calibration for quantitative analysis, we normally apply a so-called univariate regression technique. *One* instrumental signal (one variable) from a chemical substance is associated with, for example, its concentration. This works well with distinct and 'pure' peaks without interferences. Less distinct spectra, interferences or mixtures ask for more advanced calibration methods. With the PLS method a large spectral range or the whole spectrum is the 'signal' that is related to the amount of an unknown compound. This is accomplished by using the signal strength, for example, the absorbance, at several wavelengths as x variables in the calibration procedure while the concentrations are the y variables. Quantitative determination of silica mixtures with infrared spectroscopy will be used to display the application of a multivariate calibration strategy (Bjørsvik and Bye, 1991; Bye, 1994).

The dust exposure in foundries and refractories is quite complex. The dust may contain crystalline as well as amorphous modifications of silica and the quantitative determination can be done with X-ray diffraction (XRD) for quartz and cristobalite (Altree-Williams, 1977; Bye, 1983) or infrared spectroscopy (i.r.) for crystalline and amorphous silica (Tuddenham and Lyon, 1960). Mixtures of these silica modifications may be analysed with the combined XRD–i.r. method (Bye *et al.*, 1980). However, the XRD instrumentation is expensive and needs highly qualified laboratory personnel. Thus the application of only i.r. spectroscopy for quantitative determination of silica mixtures might be advantageous, the method being quick, simple and performed on a standard laboratory instrument. However, the primary absorption band of the various silica modifications interferes severely in i.r. as seen in Fig. 16.

Our standard method with i.r. uses only the absorption band at 800 cm^{-1} (Bye *et al.*, 1980), while for the PLS application, the spectral profile was recorded at 10 cm^{-1} intervals in the spectral range 900–600 cm^{-1}. A calibration set of 18 binary mixtures of quartz, cristobalite and amorphous SiO_2 (0–100%) was prepared by weighing. One central ternary sample (1/3, 1/3, 1/3) was also included in the calibration set. The observed *%-transmission* values were used as the **X** matrix, whereas the relative concentrations made up the dependent y parameters. Prior to the PLS calculations a PCA was performed to inspect the structure of the data set. Two principal

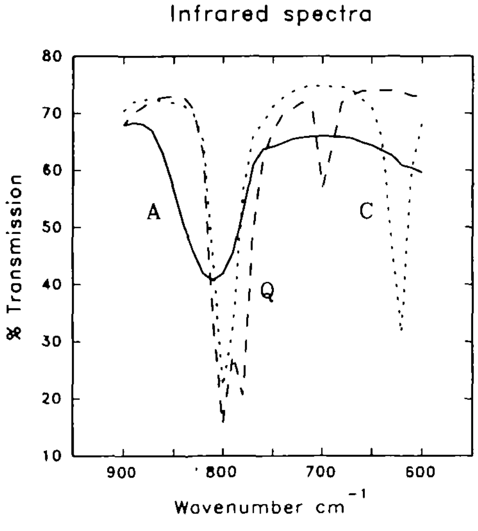

Fig. 16. Infrared spectra of quartz (Q– – – – –), cristobalite (C··················) and amorphous silica (A————). The transmission values are only given for the most interesting region (900–600 cm^{-1}) for the silica modifications, revealing the severe interference for the primary absorption bands.

components explained 97% of the variance of the **X** matrix and the object score plot is shown in Fig. 17.

The traditional three-component mixture triangle is easily recognized in the figure, and remember that only the transmission values were used as input parameters. No direct information about the silica concentrations was introduced in the PCA calculation! The corresponding variable loading plots are shown in Fig. 18. Without going into details on the interpretation, we can see that the largest loadings are located at the wavelengths for the primary silica absorption bands at 750–850 cm^{-1}. In addition there are large loadings around the secondary bands for quartz (700 cm^{-1}) and cristobalite (620 cm^{-1}). This implies that these variables have the largest influence of the spectral variation between the samples, and that they are most important for the characterization of the samples.

A PLS model was constructed with the same spectroscopic data as the x parameters and with the silica concentrations as the response variables. Two components explained 99.5% of the variance in Y and a set of 'unknown' samples were then used to test the PLS model, see Table 4.

Figure 19 illustrates the quality of the model, showing the measured (nominal) vs the predicted quartz content. The results were similar for the other two components. Therefore, the multivariate approach with PLS applied to i.r. spectroscopy made it possible to simultaneously determine the concentrations of three interfering substances with acceptable accuracy.

Example 4: quantitative models for skin corrosion by carboxylic acids

In a recent paper (Eriksson *et al.*, 1994) a search was made for a QSAR-model to predict the dermal effects of corrosive carboxylic acids. Corrosive chemicals may induce two types of biological effects: (1) irritant contact dermatitis (skin irritation); and (2) allergic contact dermatitis (skin sensitization). Furthermore, skin irritation may be categorized in acute primary irritation, cumulative irritation or corrosion. According to legislation the compounds are classified in three

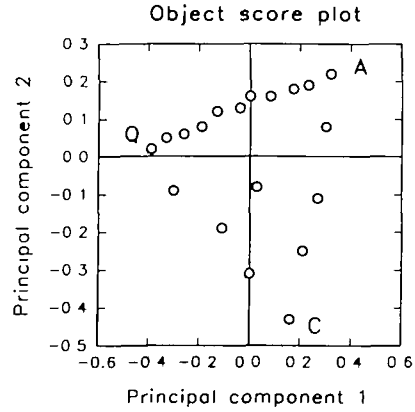

Fig. 17. Object score plot for the first two principal components (PC1 and PC2) for the training set. The design pattern for a three-component mixture (0–100%) is recognized with the additional centre mixture (1/3, 1/3, 1/3) of the three silica components (A = amorphous silica; Q = quartz; C = cristobalite). See Bye (1994) for further details.

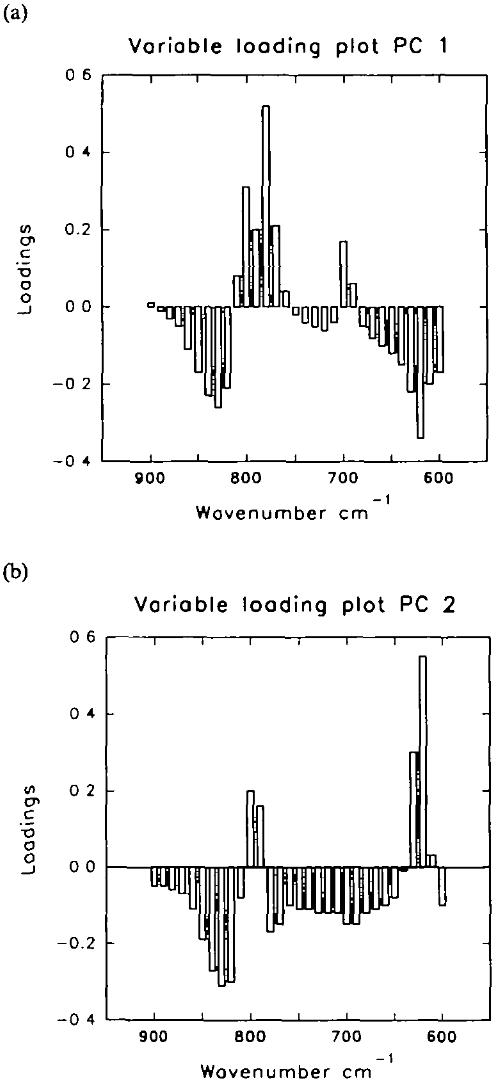

Fig. 18. Variable loading plot for (a) the first and (b) the second principal components of the training set The characteristic (and largest) loadings are located at 800, 700 and 620 cm^{-1}, that is at the most characteristic silica absorption bands; compare with Fig. 16

categories: strongly, moderately and weakly corrosive, depending upon their corrosive strength toward epithelium tissue. However, in practical work the occupational hygienist will observe that classification is difficult owing to lack of corrosion data. Corrosive properties of diluted solutions of the chemicals are even less investigated. The aim of the investigation was thus to establish a systematic way to classify the corrosive chemicals and to predict corrosion properties of uninvestigated compounds.

The nine physico-chemical properties: molecular weight, melting point, density, refractive index, pK_a (acid constant), log P (octanol–water partition coefficient),

Table 4. Nominal and predicted compositions of the three-component silica mixtures for the test samples. A = amorphous silica, Q = quartz, C = cristobalite (Bye, 1994)

Sample	Determined by weighing			Predicted by PLS		
	A	Q (wt %)	C	A	Q (wt %)	C
T1	20.1	20.5	59.4	17.3	21.4	61.3
T2	50.2	22.3	27 5	48.5	23.1	28.4
T3	59.7	20.4	19 9	63.2	19.4	17.4
T4	19.9	35.0	45.1	15 1	36.1	49.8
T5	20.0	60.1	19.9	19 8	57.2	22 7

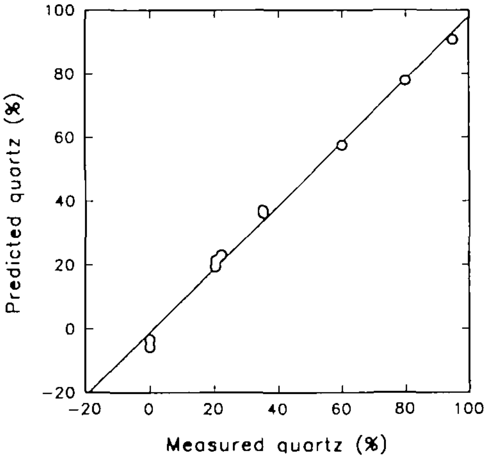

Fig. 19. Measured quartz concentration (%) vs predicted quartz concentration (%) for the test set with the three silica modifications. The prediction is based on a PLS2 model with two components, describing 99.5% of the variance in the composition matrix (Y). The absolute prediction errors, that is the standard deviations of predictions, are in the range of 1.5–3.0% for the three silica components. This is comparable with the standard X-ray diffraction and i.r. spectroscopic methods (Bye, 1994).

electronegativity and the energy of the highest occupied and the lowest unoccupied molecular orbitals were introduced as the **X** matrix for 45 aliphatic carboxylic acids. Nine acids were selected as the training set together with a validation set (test set) of six compounds. One additional test compound outside the domain of the training set was included to show the ability for the model to operate (predict) outside the 'trained' region. Selection of the other acids in the training and test set was based on experimental design and PCA, to be representative for the molecular family. These 15 carboxylic acids were tested for cutaneous corrosion on adult rabbits. Each substance was categorized as corrosive or not, and assigned the lowest observed effect concentration (LOEC). A three-component PLS model described 83% of the biological effect variance. The consecutive application of the model on the test set represents a necessary evaluation of the model. Figure 20 shows the measured LOEC values vs the predicted, together with the corresponding values for the training set. A satisfactory predictive ability of the model is evident from the high correlation in Fig. 20. The final stage in application of the model was to predict the corrosive properties

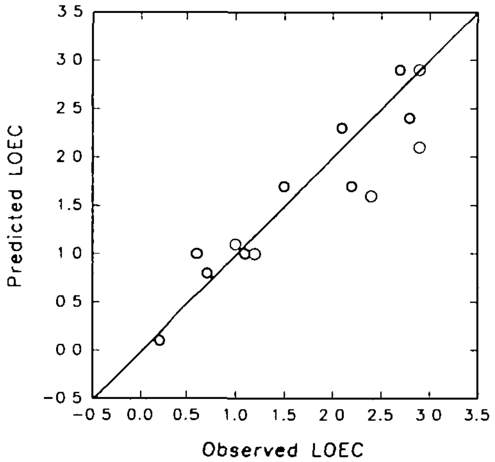

Fig. 20. Observed vs predicted LOEC (*lowest observable effect concentration*) of the training (o) and test set (o) for skin corrosion, exerted on rabbits by carboxylic acids The prediction is based on a PLS1 model with two components (Eriksson *et al.*, 1994).

of the 30, until now untested, carboxylic acids. These results can be used in further work on corrosive chemicals.

One question is: 'What is the requirement for a new chemical to be included for valid corrosivity predictions?' The basic idea is that your calibration and test set are representative for future carboxylic acids. Selection of these sets are therefore very important. Object scores for new compounds have to be in a certain range—not too different from the object scores of the calibration set. This is controlled through PCA. Furthermore, some of the SIMCA techniques (for example, class boundaries and object residuals) can be used to verify the class membership. In addition, most chemometric software has outlier warnings during the prediction of new objects.

In general, such quantitative models have several prospects. First of all such models provide us with a systematic way of classifying chemicals, for example, with respect to material properties or biological activities. Mathematical and statistical models used in investigations on biological effects of chemicals may reduce the need for animal experiments. Furthermore, the application of such models, constructed through a combined use of experimental design, PCA and PLS, represents an efficient method for ranking toxic chemicals in a screening phase. Such ranking can be of great help in deciding which compounds should undergo extensive, expensive and time-consuming experiments. The work of Jonsson *et al.* (1989), on a strategy for ranking toxic chemicals in the environment illustrates how these multivariate chemometric methods can be combined and used as a systematic approach in the complex work of hazard reduction.

Associations between occupational hygiene and health data

The QSAR example described above has been selected to illustrate how PLS can be used to study associations between several x variables and one y variable. This would be the situation if the foundry data (**X** matrix) in Example 1 should be correlated to one health parameter. For foundry workers this could be a respiratory effect and the most usual PLS1 technique would be used. Also several y variables (health effects) can be studied by the same model simultaneously, through the use of the PLS2 technique. However, no adequate investigations have been reported on the association between exposure or occupational hygiene data and health effects using PLS so far.

Nevertheless, multivariate methods other than PLS have been used in various studies to describe health endpoints in the industry. A comprehensive evaluation of the health status among workers in the Japanese ceramic industries of different size was reported by Huang *et al.* (1993). In a cross-sectional study it was found that the occupational health status was significantly higher in the large factories than in the smaller ones. Prevalences of silicosis and tuberculosis were used as the indices for health status. In addition, several medical screening tests were subjected to principal component analysis. The high morbidity of silicosis and pulmonary tuberculosis in smaller companies contributed most to the decline in the overall health level. These findings were associated with earlier observations on the exposure level for quartz. High correlations between the most important factors derived from PCA and the two pulmonary outcomes gave thus indirectly information about the hygienic level.

In a recent study by Tielemans *et al.* (1994) a canonical correlation analysis was carried out to study the relationship between a set of organic dust exposure measurements and a set of ventilatory function variables. The investigation comprised 390 male workers within grain processing and animal feeding. Increased organic dust exposure was closely correlated with a decrease in ventilatory functions. However, an independent effect of overall organic dust exposure and the number exposure-years was observed on the MEVF curve (maximum expiratory ventilatory function). The regression method used is somewhat similar although not exactly identical to PLS.

FOUR LEVELS OF PATTERN RECOGNITION (PARC)

It is very convenient to describe PCA, SIMCA and PLS together in a presentation like this. The methods are closely related in several ways: (i) the samples (objects) are looked upon as vectors in the multidimensional variable or measurement space; (ii) the data set is organized as a data table, including training set, test set and unknown samples; (iii) principal component models (or closely related in the case of PLS) are built for extraction of the systematic and correlated information among the samples; (iv) graphical presentations, through the object score and the variable loading plots are used for interpretation; and (v) the philosophy for all three techniques is based on multivariate approaches, handling all the experimental data simultaneously. The inclusion of one or more dependent variables in the data treatment, completes the concept of *Four Levels of Pattern Recognition (PARC)*: Level (1): classify an unknown compound within one of the present established classes. Level (2): classify an unknown object inside one or outside all present classes, i.e. an outlier. Level (3): prediction of one dependent response variable of an unknown object. Level (4): prediction of more than one response variable (Albano *et al.*, 1978; Wold *et al.*, 1983).

An integrated use of these techniques represents a powerful set of tools for pattern recognition. In particular, the introduction of various quantitative measures for the correlations among objects and variables is convenient. This includes class distances, discrimination power, class membership determination and prediction power (Wold, 1976; Wold and Sjöström, 1977; Sharaf *et al.*, 1986). Such a combined approach is outlined here because questions within the area of occupational hygiene and health can be partially classification and partially prediction problems. If the prediction of a dependent variable should be done in a system where there are several classes of objects, the combined application of SIMCA and PLS can assist the occupational hygienist substantially. One may first utilize the opportunity to establish the basic characteristics and structural pattern of the groups of items. Second, one can develop quantitative relationships between these basic properties and, for example, the biological effects. These results can finally be connected to the work place operations.

Models, causality and validation

So far this introduction to chemometric methods has emphasized the advantage of studying more than one variable at the time. In fact, as many variables as possible, having a suggested although not definite connection to the problem can be considered. When combining the formulation of new hypothesis with exploratory data analysis and mathematical model formulation, there are two important topics to discuss:

—the difference between a mathematical and mechanistic model understanding;
—the separation (distinction) between causal relations and incidental correlations.

Many scientists stick to the philosophy of studying phenomena based on a mechanistic and theoretical (mental) understanding of the ongoing processes. This is possible with 'uni- or oligo-variate' problems or minor fragments on topics that are well studied and described for a long time. However, for complex problems like: (i) indoor air quality and health effects; or (ii) mineral fiber properties and lung cancer, the number of suggested and important variables is quite large. Fundamental knowledge about the processes is still so scarce that good and comprehensive mechanistic models are not yet achievable (Austin *et al.*, 1992; Dement, 1990). Mostly these two complex topics are studied in a confined or limited variable space. This is probably because it is mentally difficult to handle more than one or a few variables simultaneously. The final goal of obtaining comprehensive mechanistic knowledge enforces researchers to look for an explanation based on few variables, irrespective of how complex and multifactorial the problem is.

Mathematical models from a chemometric approach through experimental measurements of such complex problems, may however lead to empirical and practical models. They may not provide us with an easily accessible understanding of the underlying mechanism. However, we may obtain both important and relevant qualitative and quantitative information from such models. For many cases this may be satisfactory, at least in an initial stage, such as the exploratory phase. Such models might give 'new information' through practical functioning by solving the initial problem. This in turn may unravel connections that can guide us to a deeper causal understanding of complex problems. A complex exposure situation may be described by a mathematical model, involving many work operations, several workroom characteristics and process factors. Such a model may help the occupational hygienist in improving the work environment. This can be done without giving a mechanistic interpretation of the interplay between all the various factors, the exposure pattern and the health effects.

However, there are reasons to be cautious in handling many variables during an 'exploratory' phase. If by chance, some variables incidentally should be correlated, there are possibilities for misinterpretation. Such 'new' exploratory chemometric findings should only be recognized and accepted as significant evidence after repeated and properly designed investigations have been performed. Since the main application is classification or unraveling of general relational trends between the objects and variables there should be only limited possibilities for serious misinterpretations. When the techniques are applied to organize and graphically display the measurements through projections, the inference from statistical artifacts should be of minor concern. Even occasional and intermediate misinterpretation of causality in an observed connection should not be a fundamental hindrance for more exploratory approaches in occupational hygiene studies. However, the quality of the data is critical under such circumstances.

A critical scientific attitude toward new and unexpected or 'strange' findings are necessary to warrant that any conclusions have, for example, biological credibility. On the other hand, a properly functioning quantitative empirical model (such as the QSAR models), with uncertain causal relations, may still help in the prevention of future work-related problems. From an empirical model, describing air quality and health problems, we might predict that a certain class of indoor air characteristics would impose problems. With critical scientific considerations to these questions, such work environmental problems may be reduced in the future, considering even 'non-causal models with only empirical functioning'.

However, in order to work with mathematical models for complex exposure pattern and occupational health effects we have to rely on the quality of the models. This quality depends on several factors, among which the quality of the *training set* is essential. Earlier I emphasized the importance of the *test set* in model validation. The use of the test set is the best and only way to validate your model thoroughly. To avoid biased models and ensure plausible results, quality control has to be included during the model construction as well. Such validation can be done with cross-validation (Wold, 1978) or leverage correction (Martens and Næs, 1989).

It is outside the scope of this introduction to go into details about the procedures for cross-validation and leverage correction. This brief description is given just to point to the importance of valid models and how the credibility of the models may be achieved through standard validation techniques. As already stated, the quality of the training and test objects is crucial. Whenever possible, experimental design should be used to obtain representative samples. This means that the samples span the variation in properties (the measured variables) that can be anticipated for future samples. In studies where the exposure pattern is correlated to health effects, it is not always possible to design an exposure situation with large enough variation. However, one should look for experimental conditions where the actual variables vary as much as possible and include these 'situations' in the model. A practical guide for most of the aspects of chemometrics may be found in a recent textbook published by Esbensen *et al.* (1994).

CONCLUDING REMARKS

Since the early 1970s when Svante Wold (University of Umeå, Sweden) introduced the name *chemometrics* for the first time, the field has developed like a cascade. Special software, journals, congresses and an increasing amount of textbooks are indicators of this development. Within the field of occupational hygiene and health, we should definitely welcome and promote an increased use of multidimensional philosophy and analytical methods. This is due to the complexity of the work environment and the correlations between exposure and health effects. All the measured variables for a large number of samples may be analysed simultaneously. Data matrices with up to 32 500 × 32 500 elements can easily be handled by a modern desk-top computer. The dimension of a complex problem can be reduced to a few principal components. Graphical presentation of the results offers user friendly interpretation facilities. Issues like indoor air quality and health effects, fiber properties and fiber toxicities and correlations between chemical structures and toxic reactions, all connected to exposure at work, should definitely benefit from a multivariate approach.

Chemometrics being nearly a way of life rather than purely a collection of powerful methods, offers a bridge for occupational hygienists into a sphere for multidimensional problem solving. Applying the framework of chemometrics may increase the exploration of 'The total working environment'. This may be the future key to solving the general multidimensional problem: what are the most important variables that relate to the reported health effects?

It is definitely not a question of why multivariate methods should be applied within the field of occupational hygiene—it is merely a question of how! Hopefully, chemometrics will increase the consciousness around experimental design and qualified measurements. Complex systems can be investigated and even latent or hidden information may be revealed.

Acknowledgements—Hans-R. Bjørsvik, Dr Roald Bjørklund and Dr Kim Esbensen are highly acknowledged for all inspiring discussions and valuable criticism during the preparation of this article. I am also indebted to Dr John McK Ellison, former editor of *The Annals of Occupational Hygiene*, who encouraged me to present this review article.

REFERENCES

Albano, C., Dunn, W., III, Edlund, U., Johansson, E., Nordén, B., Sjöström, M. and Wold, S. (1978) Four levels of pattern recognition. *Analyt. chim. Acta.* **103**, 429–443.

Altree-Williams, S. (1977) Quantitative X-ray diffractometry on respirable dust collected on Nucleopore filters. *Ann. occup. Hyg.* **20**, 109–126.

Austin, B. S., Greenfield, S. M., Weir, B. R., Anderson, G. E. and Behar, J. V. (1992) Modeling the indoor environment. *Environ. Sci. Technol.* **26**, 851–858.

Bjørsvik, H. R. and Bye, E. (1991) Multivariate calibration applied to infrared spectroscopy for quantiative determination of crystalline and amorphous silica. *Appl. Spectrosc.* **45**, 771–778.

Bye, E. (1983) Quantitative microanalysis of cristobalite by X-ray diffraction. *J. appl. Crystall.* **16**, 21–23

Bye, E. (1994) Chemometrics in aerosol analysis—quantitative analysis of silica dust mixtures by multivariate calibration applied to infrared spectroscopy. In *Inhaled Particles VII* (Edited by Dodgson, J. and McCallum, R. I.), pp. 519–525. Elsevier Science, Oxford.

Bye, E., Edholm, G., Nicholson, D. G. and Gylseth, B. (1980) On the determination of crystalline silica in the presence of amorphous silica. *Ann. occup. Hyg.* **23**, 329–334.

Bye, E., Grønnerød, O. and Vogt, N. B. (1989) Multivariate classification of histochemically stained human muscle skeletal fibres by the SIMCA method. *Histochem. J.* **21**, 15–22.

Dement, J. M. (1990) Overview: Workshop on fiber toxicology research needs. *Environ. Hlth Perspect.* **88**, 261–268.

Deming, S. N. and Morgan, S. L. (1987) *Experimental Design, Data Handling in Science and Technology*, Volume 3. Elsevier, Amsterdam.

Dunn, W. J., III (1989) Quantiative structure-activity relationships (QSAR) *Chemometr. & Intell. Lab. Syst.* **6**, 181–190.

Eriksson, E., Berglind, R. and Sjöström, M. (1994) A multivariate quantitative structure–activity relationship for corrosive carboxylic acids. *Chemometr & Intell. Lab. Syst.* **23**, 235–245.

Esbensen, K., Midtgaard, T. and Schönkopf, S. (1994) *Multivariate Analysis—In Practice. A Training Package*. Camo AS, Trondheim.

Huang, J., Shibata, E , Takeuchi, Y. and Okutani, H. (1993) Comprehensive health evaluation of workers in the ceramics industry. *Br. J. Ind. Med.* **50**, 112–116.

Jonsson, J., Eriksson, L., Sjöström, M. and Wold, S. (1989) A strategy for ranking environmentally occurring chemicals. *Chemometr. & Intell. Lab. Syst.* **5**, 169–186.

Kowalski, B. R. and Wold, S. (1982) Pattern recognition in chemistry. In *Handbook of Statistics* (Edited by Krishnaia, P. R. and Kanal, L. N.), Volume 2, pp. 673–697. North-Holland, Amsterdam.

Kromhout, H., Symanski, E. and Rappaport, S. M. (1993) A comprehensive evaluation of within- and between-worker components of occupational exposure to chemical agents. *Ann. occup. Hyg.* **37**, 253–270.

Lundstedt, T. (1991) A QSAR strategy for screening of drugs—and predicting their clinical activity. *Drug news & Perspect.* **4**, 468–475.

Nordén, B., Edlund and Wold, S. (1978) Carcinogenicity of polycyclic aromatic hydrocarbons studied by SIMCA pattern recognition. *Acta Chem. Scand.* **B32**, 602–608.

Martens, H. and Næs, T (1989) *Multivariate Calibration*. John Wiley & Sons, Chichester, UK.

Massart, D. L., Vandeginste, B. G. M., Deming, S. N , Michotte, Y and Kaufman, L. (1988) *Chemometrics: A Textbook. Data Handling in Science and Technology*, Volume 2. Elsevier, Amsterdam.

Sharaf, M. A., Illman, D. L. and Kowalski, B. R. (1986) Chemometrics. In *Chemical Analysis*, Volume 82. Wiley Interscience, New York.

Schneider, T., Husemoen, T., Olsen, E., Christensen, V. and Kamstrup, O. (1993) Airborne fibre concentration during standardized building insulation with bonded man-made vitreous fibre insulation material having different nominal diameters and oil content. *Ann. occup. Hyg.* **37**, 631–644.

Tielemans, E., Heederick, D. and Van Pelt, W. (1994) Changes in ventilatory function in grain processing and animal feed workers in relation to exposure to organic dust. *Scand. J Wk. Environ Hlth.* **20**, 435–443.

Tuddenham, W. M and Lyon, R. I. D. (1960) Infrared techniques in the identification and measurement of amorphous silica. *Analyt. Chem.* **32**, 1630–1634

Wold, H. (1966) Estimation of principal components and related models by iterative least-squares. In *Multivariate Analysis* (Edited by Krishnaia, P. R.). Academic Press, New York.

Wold, H (1975) *Perspectives in Probability and Statistics* (Edited by Gani, J.). Academic Press, New York.

Wold, H. (1982) Soft modelling: The basic design and some extensions. In *Systems Under Indirect Observations* (Edited by Jøreskog, K. G. and Wold, H.), Volume 2, pp. 1–54. North-Holland, Amsterdam.

Wold, S. (1976) Pattern recognition by means of disjoint principal components models. *J Pattern Recogn.* **8**, 127–139.

Wold, S. (1978) Cross-validatory estimation of the number of components in factor and principal component models. *Tecnometrics* **20**, 397–405.

Wold, S., Albano, C., Blomquist, G., Coomans, D., Dunn, W. J., III, Edlund, B., Eliasson, S., Hellberg, S., Johansson, E , Nordén, B., Sjöström, M., Söderström, B. and Wold, H. (1981) Pattern recognition by means of disjoint principal component models. Philosophy and methods. In *Symposium i Anvendt Statistik* (Edited by Höskuldson, A., Conradson, K., Sloth Jensen, B and Esbensen, K.). Arranged by NEUCC, RECAU, RECKU.

Wold, S., Albano, C., Dunn, W. J. III, Esbensen, K., Hellberg, S., Johansson, E. and Sjöström, M (1983) Pattern recognition: finding and using regularities in multivariate data In *Food Research and Data Analysis* (Edited by Martens, H. and Russwurm, H, Jr.), pp 147–188. Applied Science, London.

Wold, S., Dunn, W. J. and Hellberg, S. (1985) Toxicity modeling and prediction with pattern recognition *Environ. Hlth Perspect.* **61**, 257–268.

Wold, S., Esbensen, K. and Geladi, P. (1987) Principal component analysis. *Chemometr. & Intell. Lab Syst.* **2**, 37–52.

Wold, S. and Sjöström, M. (1977) SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In *Chemometrics: Theory and Application* (Edited by Kowalski, B. R.), ACS Symposium Series 52, pp. 243–282. American Chemical Society, Washington DC.