Classification of Chlorinated Phenols by Pattern Recognition Method on the Basis of Mass Spectra Losses

P. JURÁŠEK, M. SLIMÁK, R. BREŽNÝ, Š. VODNÝ, and M. KOŠÍK

Department of Wood, Pulp and Paper, Faculty of Chemical Technology, Slovak Technical University, CS-812 37 Bratislava

Received 17 February 1992

Chlorinated phenols were classified by the pattern recognition method PRIMA. For the purpose of the classification new characteristic features based on significant mass spectra losses were derived. Principal component analysis was used to present the separability of the selected classes. In the application of this approach there is no need for the spectral library and the spectra are interpreted much faster than by the library search methods. A reliability of the applied procedure for library spectra corresponded to 92 %. In the model mixture, 18 out of 21 substances were classified correctly.

Since the situation in the environmental area has been steadily deteriorating, attention has recently been paid to wastes produced by the pulp and paper industry. It has been shown that undesirable substances are primarily chlorinated organic compounds produced during bleaching the pulp by inorganic compounds of chlorine having a harmful effect on the health of man and on the environmental quality as well. By the chlorination of lignin, low-molecular, partially chlorinated compounds are formed representing only several kilograms per ton of lignin, but some of them having the toxic, mutagenic, and carcinogenic effects [1]. These organic compounds are not quite decomposable by the common plant procedures of effluent treatment and therefore the consequences of escaping of these substances into the environment are studied. Simultaneously, the search for procedures is made to reduce the formation of undesirable substances in the bleaching processes. An inevitable part of the solution of this problem is the ability of identifying and quantifying the chlorinated organic compounds. For the low-molecular chlorinated organic compounds the capillary gas chromatographic method associated with a flame ionization detector (FID) as well as with an electron capture detector (ECD) and a mass spectrometer detector (MSD) [2] proved to be suitable. The combination of GC—MS method provides a comprehensive information about the unknown sample and it belongs to the highly efficient analytical techniques.

The mass spectra gained during the common GC—MS analysis can be evaluated in several ways. The most reliable way of identification is the human spectral interpretation which requires, however, some information about the sample origin, some results from other spectral techniques and above all the profound knowledge of mass spectrometry. An important alternative of the spectral data interpretation are automatic computing approaches [3—6]. The library searching is a very frequently used method, when from a database of spectra the most similar to the unknown spectrum are selected. The integrated interpretation systems such as DENDRAL, STIRS, *etc.* and the chemometric detectors [7, 8] are less common. The maximum number of the pattern recognition method applications is available in the chemistry in the field of interpretation of the low-resolution mass spectrum [4, 5, 9]. Generally, these methods pursue the classification of objects into certain classes (groups) according to their features (properties, signs). The features by which the distribution is made can be derived from various properties of the objects (substances) and hence also from the mass spectrum [8]. The simple application of the pattern recognition method in mass spectrometry can be an inclusion of the unknown spectrum in one of the structural groups considered.

This paper is intended to classify the chlorinated phenols contained in bleaching pulp industry effluents by means of the pattern recognition method PRIMA [10] utilizing the features derived from the mass spectra losses.

EXPERIMENTAL

All evaluating programs, the library spectrum analyses and the model mixture measurements were performed by using the Hewlett—Packard GC—MS system (HP 59970 MS Chemstation) together with the HP-310 Series 3000 computer. The software was written in the program language HP Pascal 3.1. The specialized library called BLEACH was prepared for the analysis of bleaching effluents. The library consisted of 482 spectra taken from the commercial

EPA/NIH/NBS spectra collection, of spectra taken from the literature [11, 12], and of measured spectra of the model compounds.

The testing of the indicated method was done by applying the model mixture of chlorinated phenols. The mixture was analyzed using a mass spectrometer HP 5970B linked to a gas chromatograph HP 5890A equipped with a capillary column HP ULTRA-1 (25 m x 0.2 mm inner diameter, 0.33 μm film thickness). The injection was made in the split mode with a splitting ratio of 1 : 50 using the on-line dried helium as a carrier gas at the rate of 25 cm s^{-1}. The oven temperature was maintained at 50 °C for 1 min and programed at 10 °C min^{-1} to 270 °C and maintained for 10 min. The injector and transfer line temperatures were set at 250 °C and 280 °C, respectively. The mass spectrometer was operated in the TIC mode, mass scan range of m/z = 33—500 at the electron energy of 70 eV.

The four structural classes of chlorinated phenols, substances predominantly present in bleaching pulp industry waters, were defined as follows: *I*. chlorinated phenols, *II*. chlorinated 2-methoxyphenols (chlor. guaiacols), *III*. chlorinated 2,6-dimethoxyphenols (chlor. syringols), *IV*. chlorinated 1,2-dihydroxybenzenes (chlor. catechols). For classes *II*, *III*, and *IV* also acetylated compounds were considered. The pursued types of substances differ significantly in the interaction with enzymes under the acute and long-term toxic action, in chemical stability, and in other properties.

By applying the exploratory data analysis the separability of the selected classes was presented. A software package PEDAS/MS [13] was used for interpreting the investigated data.

The pattern recognition method PRIMA [10] was used for classification. It belongs to methods based on the concept of class distances being defined in the pattern space. For each structural group of substances under observation distances between the spectrum of unknown substance and the class centre of gravity are calculated on the basis of the centres of gravity and the dispersions of features and then used for classification. The PRIMA method is convenient for a minor training set or a computer, or for incomplete data.

Analysis of the Spectra

For each class an appropriate sublibrary was created of the mass spectra taken over from the specialized BLEACH library. For the purpose of the classification new characteristic features based on significant mass spectra losses were derived. Reference spectra of the individual structural groups were converted to the "loss spectra" and analyzed in detail. In computing the mass spectrum losses the following procedure was used: each peak in the mass spectrum was evaluated from the standpoint of its position and intensity in the mode derived from the results of the paper [14] concerned with the mass spectrum peaks significance. In such a way the original spectrum (Fig. 1*a*) was transformed (Fig. 1*b*), the significant peaks being those which are more intensive or found in the region of higher masses. In the next step, all possible losses occurring in the transformed spectrum were calculated. The loss $\Delta m/z$ is equal to the difference between m/z values of the two transformed peaks in the spectra. The intensity of loss (significance) was determined by the intensity product between two transformed peaks in the spectra. If a larger number of equal losses $\Delta m/z$ was calculated, it was the most intensive loss which was considered. Thus the loss spectrum (Fig. 1*c*) expressing an important part of the structural information contained in the original mass spectrum was calculated. Some of the losses were directly connected with the molecule fragmentation process. For example, Fig. 1*c* indicates that the most intensive losses $\Delta m/z$ for *x*,*y*-dichloro-2,6-dimethoxyphenol are: 2, 15, 28, 43, 46, 58, 61, *etc*. The selected losses can then be regarded as features characterizing the unknown spectrum. These features are quantitatively expressed by intensities of the peaks of the loss spectrum. The feature values were standardized at intervals of <0, 9999>. The seven highly important losses, interpretable from the aspect of fragmentation, were selected. For some of them it was useful to introduce the weight factors in order to improve the distinction between classes.

Classification

The seven of the value losses were calculated, in a manner indicated above, for each reference spectrum of individual classes. These values created a data matrix, the so-called training set. From all the BLEACH library spectra and also from the mass spectra of the GC—MS analysis of model mixture the data matrices, *i.e.* the so-called recognition sets were created in the same way. Both the training set and the recognition set were then used as input data for the own pattern recognition process. The classification was proceeding in two steps. In the first step the distances of the point corresponding to the unknown spectrum and the centre of gravities of the classes were calculated. In the second step, the class to which the unknown spectrum belongs was determined.

The program system enabling the libraries to be created and the spectra analysis and classification to be performed consists of five programs (Scheme 1):

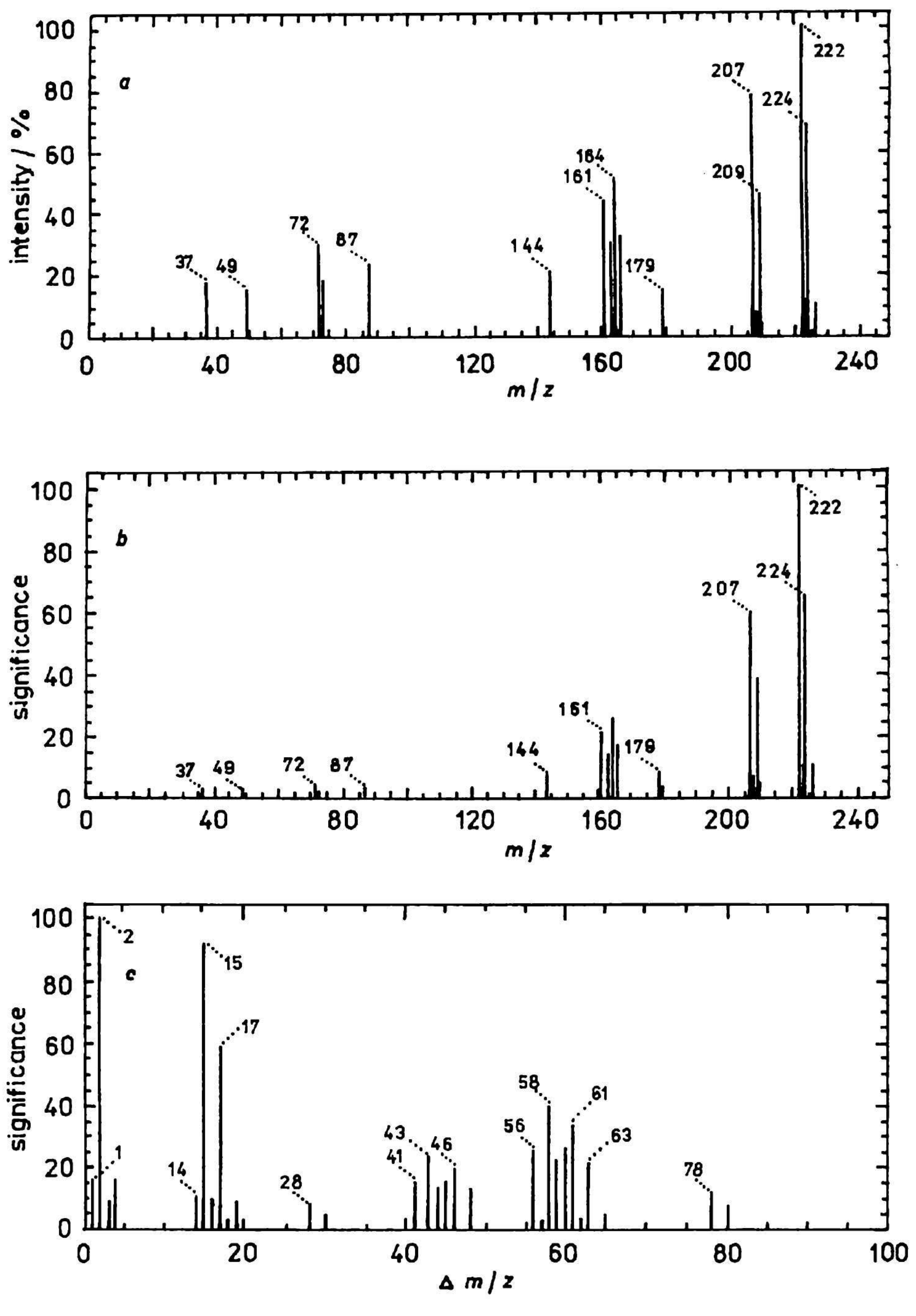


Fig. 1. The creation of losses in the mass spectrum of *x*,*y*-dichloro-2,6-dimethoxyphenol. *a*) Original mass spectrum; *b*) transformed spectrum according to the significance of peaks; *c*) loss spectrum.

MSLIB – it permits to create the specialized libraries by importing the mass spectra from commercial libraries, own measurements, literature, *etc.*;

MSANAL – the analysis of the library reference spectra (the study directed at the occurrence of fragments, losses, and "intelligent features"), the choice of the most significant features;

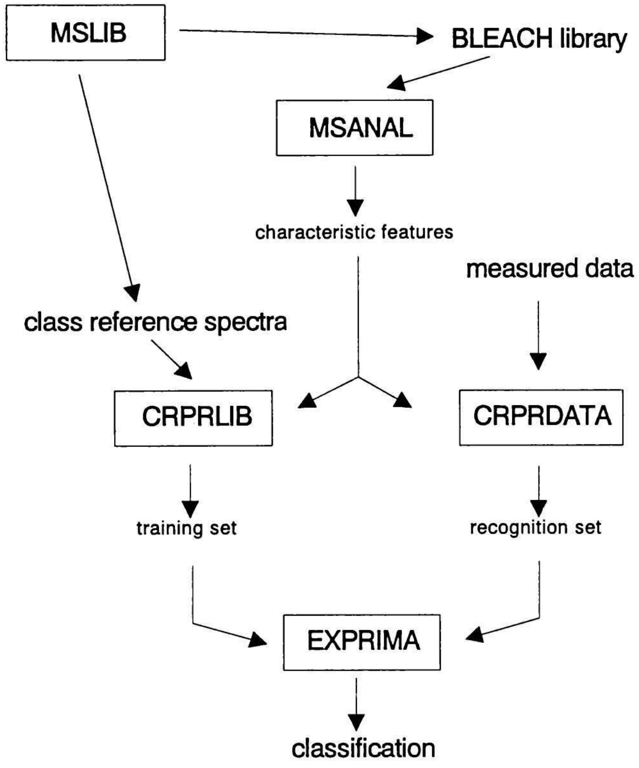
CRPRLIB – the formation of a training set from reference spectra;

CRPRDATA – it creates a recognition set from the measured data;

EXPRIMA – the classification process according to the PRIMA method [10].

RESULTS AND DISCUSSION

The goal of this work was to classify the chlorinated phenols contained in bleaching pulp industry



Scheme 1

A program system performing the analysis of spectra and the classification of measured data.

effluents into four classes according to mass spectra losses. In Table 1 are shown the most importabt real losses $\Delta m/z$ for individual structural groups along with the losses which showed to be the most characteristic in their loss spectra. The most intensive losses $\Delta m/z$ found in the reference loss spectra were: 2, 15, 28, 29, 35, 36, 43, 46, 58, 61, 64, 71, 72, and 99. The losses occurring due to isotopic peaks of the spectra were not considered, which means 4, 6, 13, 17, 26, 33, 34, 37, 41, 45, 48, 63, 65, 66, 67, 73, 74, *etc*. For the purpose of the feature selection more attention was paid to the losses interpretable from the aspect of fragmentation of the substances considered. At last, the following losses $\Delta m/z$ were chosen: 15 ($CH_3^\bullet$), 29 ($CHO^\bullet$), 36 (HCl), 43 ($CH_3^\bullet$ and CO), 46 (H_2O and CO), 58 (CH_2O and CO), and 64 (HCOCl; HCl and CO). The intensive loss $\Delta m/z = 2$ was not used because it is characteristic of each of the classes considered and it does not discriminate them. In order to make better distinction among the classes the weight factors 3 and 6 were estimated for the losses 29 and 46, respectively. In calculating the loss spectra the following limitations proved to be useful: the losses were computed from the transformed mass spectrum peaks being > 1 % and their $m/z > 62$; only thirty of the most intensive losses were applied (in the case of acetylated spectra forty); the maximum loss considered was $(\Delta m/z)_{\mathrm{max}} = 65$. For acetylated spectra it is suitable to reduce several times, before calculating the loss spectrum, the intensive fragment $m/z = 43$ found in the measured spectra.

The calculated training set data were investigated using exploratory data analysis. By applying the principal component analysis these multivariate data were projected onto a suitable plane. Thus, the sets of reference mass spectra of selected structural classes were represented graphically as a scatter plot (Fig. 2). Each point in the scatter plot corresponds to a spectrum. The distances and positions of the points are determined by the similarities of the spectra and therefore similar mass spectra form clusters in the scatter plot. Fig. 2 shows a clear separability of the classes considered. The outlying spectra (*a*, *b*, *c*, *f*, *g*, *h*, and *i*) belong to acetylated compounds, with the much more intensive fragment $m/z = 43$ than the other characteristic fragment intensities, and therefore their separation from the clusters is reasonable. The library spectra *d* and *e* (two mass spectra of 2-chloro-6-methoxy phenol) do not provide the sufficiently characteristic losses occurring in the other chlorinated guaiacols [15]. By a more detailed analysis (loading-loading plots) the most relevant losses $\Delta m/z$ were determined: 15, 46, 64, and 43 which correspond to real losses of $CH_3^\bullet$; H_2O and CO; HCOCl or HCl and CO; $CH_3^\bullet$ and CO, respectively.

A reliability of the method developed was tested in several ways. The recognition ability of this tech-

Table 1. Characteristic Real Losses $\Delta m/z$ in the Mass Spectra of Substances for Classes *I*—*IV* and the Most Intensive Losses in Their Loss Spectra

Class	Real losses $\Delta m/z$ (structural type)	The most intensive losses $\Delta m/z$ in loss spectra
I	36 (HCl), 28 (CO)	2, 28, 35, 36, 64, 65, 71, 99
II	15 ($CH_3^\bullet$), 28 (CO), 36 (HCl)	2, 15, 28, 29, 36, 43, 64
III	15 ($CH_3^\bullet$), 28 (CO), 43 ($CH_3^\bullet$ and CO) 46 (H_2O and CO), 58 (CH_2O and CO)	2, 15, 28, 43, 46, 58, 61, 65
IV	36 (HCl), 28 (CO), 46 (HCOOH), 29 ($CHO^\bullet$) 72 (HCl and HCl), 63 (COCl), 64 (HCl and CO; HCOCl)	2, 29, 36, 43, 46, 63, 64, 65, 72, 99

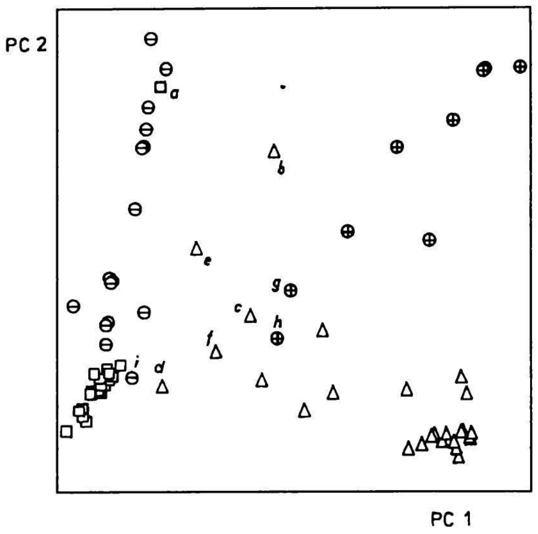


Fig. 2. Principal component plot of reference mass spectra sets (training set). Features (losses $\Delta m/z$): 15, 29, 36, 43, 46, 58, and 64. Axes: first and second principal component; 56.5 % and 27.8 % of total variance, respectively. The outliers are marked by characters. □ Chlorinated phenols, △ chlorinated guaiacols, ⊕ chlorinated syringols, ⊖ chlorinated catechols.

nique is determined by the classification of all training set spectra. In this case the total recognition ability is 84 % (Table 2). Besides the recognition ability, the classification efficiency was tested on all spectra of the specialized BLEACH library. The results are shown in Table 3. The total successfulness was 92 %. In most cases of the incorrect determination there were anticipated the structures which were "close" to structural types of the given classes (an absence or excess of some of the substituents). The worse results were obtained for the first class in which some chlorinated benzenes and methylphenols were incorrectly included and for the second class in which some of the aliphatic, core-free benzene substances (hydrocarbons, alcohols, aldehydes, and carboxylic acids) and also various silanized compounds were incorrectly included. Consequently, the applied method is suitable for mixtures containing mainly the substances from structural classes considered (*e.g.* mixtures from the pulp industry effluents). In order to differentiate these substances more properly, the sequence of losses in the mass spectrum or the more detailed evaluation of distances of the unknown spectra to the classes can be considered, or additional features (*e.g.* characterizing the region of low masses, the fundamental peak, the spectral centre of gravity, selected fragments, *etc.*) may be introduced.

The satisfactory results were achieved on applying the method to the model mixture components (Fig. 3). The incorrect classification was accomplished in three cases (Table 4) (the successfulness of 86 %). The peak No. *15* was slightly intensive and its mass spectra were incomplete or deformed. During the classification, several scans of the peak were evaluated.

The loss spectra employed for the formation of features of classification methods seem to be suitable for the structural types of substances the fragmentation of which proceeds, from the aspect of the losses, in a similar way and in the case when the important mass spectra peaks are more intensive than the peaks nearby and when the former peaks are not essentially lower (the minimum intensity of 5—10 %) than the base peak. It is clear that the more intensive molecular peak and good chromatographic separation of the mixture components are relevant. The influence of a background in the field of higher masses is unfavourable since it is just the "heavy" fragment that has an important role in evaluating the significance of the peaks.

Table 2. Reliability of the Method for the Training Set Reference Spectra (Recognition Ability)

Class	Number of spectra	Determined correctly (%)	Nondetermined	Determined incorrectly (false positive)
I	45	39 (87)	6	0
II	28	26 (93)	2	0
III	9	8 (89)	1	0
IV	21	14 (67)	4	3

Table 3. Classification Results for 482 BLEACH Library Spectra

Class	Determined correctly	Nondetermined	Determined incorrectly (false positive)	Unclassified*	Correct results/%
I	39	6	61	376	86
II	26	2	75	379	84
III	8	1	6	467	99
IV	14	4	10	454	97

* Other structural type.

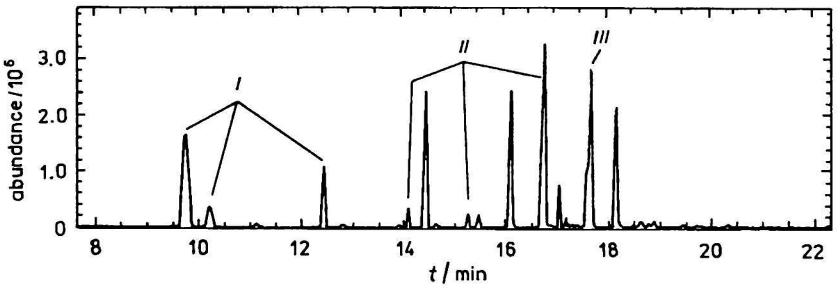


Fig. 3. Chromatogram for the model mixture of chlorinated phenols and the results of classification achieved by the pattern recognition method. The recognized classes: *I*. chlorinated phenols, *II*. chlorinated guaiacols, *III*. chlorinated syringols.

Table 4. Results of Classification for the Model Mixture of Chlorinated Phenols Achieved by the Pattern Recognition Method

Peak	Retention time/min	Substituent	Class	Classification
1	9.72	*x*,*y*-dichloro	*I*	+
2	10.20	2,4-dichloro	*I*	+
3	11.13	*x*,*y*-dichloro, 4-methyl		–
4	12.44	2,4,5-trichloro	*I*	+
5	12.80	4-hydroxy-3-methoxybenzaldehyde		N
6	13.89	*x*,*y*-dichloro, 1,2-dimethoxy		N
7	14.08	4,5-dichloro, 2-methoxy	*II*	+
8	14.42	*x*,*y*-dichloro, 1,2-dimethoxy		–
9	14.62	Nonidentified compound		N
10	15.25	*x*,*y*,*z*-trichloro, 2-methoxy	*II*	+
11	15.45	*x*-chloro-4-hydroxy-2-methoxybenzaldehyde		N
12	16.09	*x*,*y*,*z*-trichloro, 1,2-dimethoxy		N
13	16.74	*x*,*y*,*z*-trichloro, 2-methoxy	*II*	+
14	17.02	*x*,*y*,*z*,*u*-tetrachloro, 1,2-dimethoxy		N
15	17.16	*x*,*y*-dichloro, 2,6-dimethoxy	*III*	–
16	17.64	*x*,*y*,*z*-trichloro, 2,6-dimethoxy	*III*	+
17	18.13	*x*,*y*,*z*-trichloro, 2-methoxy, 4-propyl		N
18	18.63	Nonidentified compound		N
19	18.75	Nonidentified compound		N
20	18.88	Nonidentified compound		N
21	19.45	Nonidentified compound		N

+ Determined correctly; – determined incorrectly or undetermined (missing); N nonclassified (other structural type).

The main advantage of the described technique over the library search methods is that the spectra can be interpreted much faster and that the spectral library is not needed during the application. The approach applied can facilitate the evaluation of analyses and the determination of toxic effects of such substances present in mixtures (bleaching waters, effluents, plant extracts). The system can be used efficiently by means of a chemometric detector which is able to search selectively in the complicated mixture the desirable types of substances in a short time.

Acknowledgements. We are grateful to Professor K. Varmuza (Technical University, Vienna) for valuable advice and comments.

REFERENCES

1. Voss, R. H., Wearing, J. T., Mortimer, R. D., Kovacs, T., and Wong, A., *Paperi ja Puu 12*, 809 (1980).
2. Lindström, K. and Nordin, J., *J. Chromatogr. 128*, 13 (1976).
3. Small, G. W., *Anal. Chem. 59*, 535A (1987).
4. Martinsen, D., *Mass Spectrom. Rev. 4*, 461 (1985).
5. Chapman, J. R., *Computers in Mass Spectrometry.* Academic Press, London, 1978.
6. Jurášek, P., Brežný, R., and Košík, M., *Chem. Papers 46*, 184 (1992).
7. Varmuza, K., *Trends Anal. Chem. 7*, 50 (1988).
8. Lohninger, H. and Varmuza, K., *Anal. Chem. 59*, 236 (1987).
9. Varmuza, K., *Pattern Recognition in Chemistry.* Springer-Verlag, Berlin, 1980.
10. Juricskay, I. and Veress, G. E., *Anal. Chim. Acta 171*, 61 (1985).
11. Knuutinen, J., *Academic Dissertation.* University of Jyväskylä, Jyväskylä, 1984.
12. Knuutinen, J. and Korhonen, I. O. O., *Org. Mass Spectrom. 22*, 70 (1987).
13. Varmuza, K., *PEDAS/MS: PC Version of Exploratory Data Analysis of Spectra (Mass Spectrometry).* Technical University, Vienna, 1990.
14. Pesyna, G. M., McLafferty, F. W., Venkataraghavan, R., and Dayringer, H. E., *Anal. Chem. 47*, 1161 (1975).
15. Knuutinen, J. and Korhonen, I. O. O., *Org. Mass Spectrom. 19*, 96 (1984).

Translated by L. Hadrbulcová