Oil Shale, 2019, Vol. 36, No. 3, pp. 410–430
doi: https://doi.org/10.3176/oil.2019.3.04
ISSN 0208-189X

The chemometric approach to identification of residual oil contamination at former primitive asphalt pavement plants

Jelena Jurjeva[(a)*], Mihkel Koel

[(a)] Department of Materials and Environmental Technology, Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia; Eurofins Environment Testing Estonia OÜ, Paavli 5–3, 10412 Tallinn, Estonia
[(b)] Department of Chemistry and Biotechnology, Tallinn University of Technology, Akadeemia tee 15, 12618 Tallinn, Estonia

Abstract. *This study investigated polycyclic aromatic hydrocarbons (PAHs) and hydrocarbon oil index (HOI) pollution in the soil on the territories of two former primitive asphalt pavement plants (APPs) in Estonia. The standard quantitative methods for the chemical characterisation of the oils consisted of an initial screening, by using a gas chromatography-flame ionization detector (GC-FID), and, for a more detailed analysis, of gas chromatography-mass spectrometry (GC-MS). A combination of chemometric and analytical methods was used to identify the sources of PAHs, which could be attributed to the soil pollution at the plants. The identification and classification of oil spills were performed using chemometric techniques, such as the principal component analysis (PCA) and the clustering analysis (CA), which is based on Jaccard similarity. The application of the chemometric techniques enabled the clustering and discrimination of polluted soils into four groups, according to oil type. Several different methods of CA, such as single, complete and average linkages, were tested and the results were compared.*

Keywords: *residual pollution, oil spills, chemometrics, principal component analysis, cluster analysis.*

1. Introduction

Residual pollution is still a crucial environmental problem in Estonia and its clean-up has been too slow. Any assessment of pollution includes a subsequent and constant monitoring of carefully selected parameters, which give information about the contamination size and risks to the groundwater, surface water, air and soil. According to the Estonian Ministry of the Environment, the most serious past contamination was detected in the soil at former primitive asphalt pavement plants (APPs). The main sources of contamination were old oils, which were spilt into the soil. At the moment, there are approximately 30

* Corresponding author: e-mail *jelenajurjeva@eurofins.com*

unclean asphalt pavement plants on the territory of Estonia [1–3].

For the identification of oil pollution the most commonly analysed compounds found in the soil are polycyclic aromatic hydrocarbons (PAHs), volatile organic compounds (benzene, toluene, ethylbenzene and three xylene isomers), polychlorinated biphenyls (PCBs), heavy metals and hydrocarbon oil index (HOI). The U.S. Environmental Protection Agency (EPA) has reported 16 PAH compounds (Fig. 1) as being priority pollutants, including naphthalene (Nap), acenaphthylene (Acy), acenaphthene (Acp), fluorene (Flu), phenanthrene (Phe), anthracene (Ant), fluoranthene (Fla), pyrene (Pyr), chrysene (Chr), benz(*a*)anthracene (B(*a*)Ant), benzo(*b*)fluoranthene (B(*b*)F), benzo(*k*)fluoranthene (B(*k*)F), benzo(*a*)pyrene (B(*a*)P), benzo(*g*,*h*,*i*)perylene (B(*g*,*h*,*i*)P), dibenz(*a*,*h*)anthracene (DiAnt) and indeno(*1*,*2*,*3-cd*)pyrene (Ind) [4, 5].

PAHs are important to be determined because of their possible toxic, carcinogenic and mutagenic properties. These compounds are released into the atmosphere during the incomplete combustion of organic materials (e.g. coal, oil, petrol, wood) and they are then precipitated onto the soil. In the soil, or in sediments, they tend to adsorb tightly onto suspended particulate matter

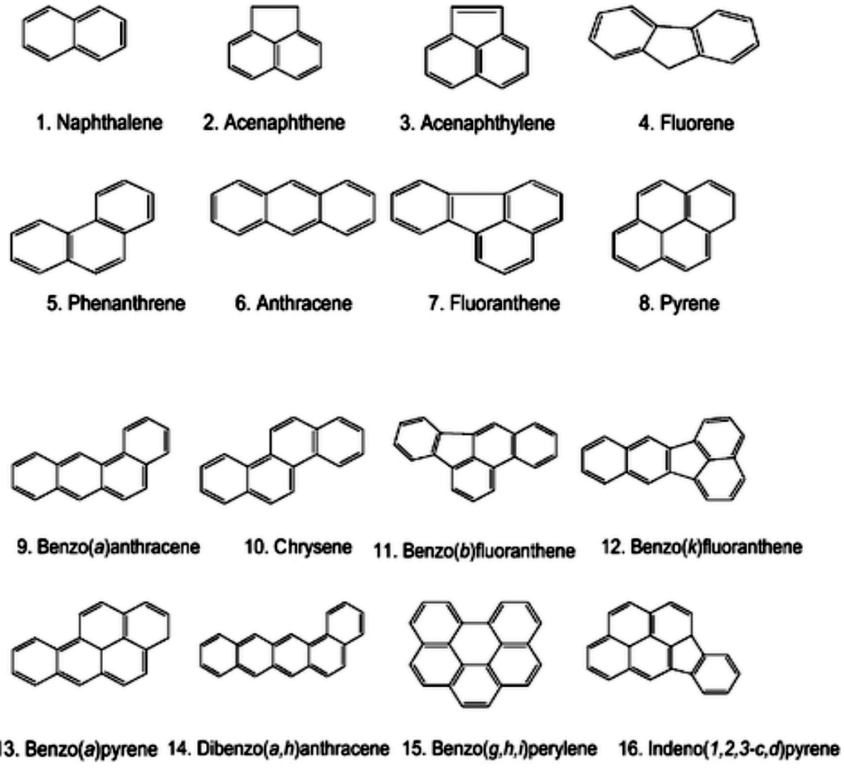

Fig. 1. The chemical structure of 16 EPA-reported PAHs [6].

[7–9]. The source of PAHs contamination in the soil can be pyrogenic, being caused or produced by combustion of heat fuels, or petrogenic, relating to the origin or formation of rocks. PAH diagnostic (binary) ratios can be applied to tracing out the sources of contamination in polluted areas [10, 11].

The hydrocarbon oil index is defined as the total amount of hydrocarbons which can be extracted from the sample by a non-polar solvent and then eluted between *n*-decane ($C_{10}H_{24}$) and *n*-tetracontane ($C_{40}H_{82}$) on an apolar capillary gas chromatographic (GC) column. The analysis of HOI is performed using a gas chromatography-flame ionization detector (GC-FID). The general term "mineral oil" comprises petroleum products, with a complex mixture of hydrocarbons, which can be found in diesel, kerosene, home heating oils, heavy fuel oils, transformer oil and lubricants. Due to the widespread use of mineral oils, these petroleum hydrocarbons are the most common organic contaminants in the soil and sediments, especially at former industrial and military sites. The said hydrocarbons can contaminate water or be consumed by organisms that can enter the human food chain [4, 12–14].

For oil hydrocarbon fingerprinting, the pollution source identification is based on chromatographic methods, which are the most effective for this purpose [15]. Gas chromatography-flame ionization detection (GC-FID) is a technique which is used for an initial screening. Gas chromatography-mass spectrometry (GC-MS) gives information about the content of oil and it is able to separate compounds in complex mixtures [15, 16].

The analytical methods generate a large amount of data, which is sometimes difficult to interpret. The multivariate statistical methods, like the principal component analysis (PCA) and the cluster analysis (CA), give a better resolution and a more adept separation quality of the samples. The chemometric methods extract the hidden information, to individualise and classify the samples into groups. The chemometric approach, together with the chromatographic methods, can help to identify and classify the soil, based on the type of contamination. Various statistical and numerical software programs (SPSS, R, MATLAB, Minitab and Excel Stat) are used to simplify these processes [17–19].

The principal component analysis decomposes the matrix into products of the scores matrix, the transposed loadings matrix plus the residuals matrix. This reduction of the data allows presenting the initial data in new coordinates or principal components (PCs). The newly generated PCs explain most of the information from the dataset. The loadings plot shows the importance of the different variables that are responsible for the clustering in the scores plot. The scores plot provides information about the relationships between individual objects, showing the groups, outliers, etc. [16, 20, 21].

The cluster analysis is used for the grouping of samples, according to the type of similarity. Its main task is to recalculate the numerical values of similarity between the new group and the rest of the objects. The next step consists in the further grouping of the data until all the objects have been

merged into one large group. The output of the hierarchical cluster analysis (HCA) is a dendrogram which visualises the grouping of samples in a two-dimensional space [16, 22, 23].

The study by Mali et al. [19] demonstrated that the chemometric approach (PCA/CA and factorial analysis of variance (ANOVA)) was advantageous for assessing and modelling the contamination patterns of highly polluted areas, and thus, it could contribute to the effective monitoring of their quality. In the study by Miki et al. [24], CA was used to identify the sources of the parent and alkylated PAHs in the sediments. The contaminated sites were categorised on the basis of PAHs composition, in order to find their primary sources within the site groups.

In spite of the numerous studies that have focused on the analysis of PAHs, there are only a few investigations on their distribution and contamination identification on the territories of past plants. In this work, analytical and chemometric tools were used to identify soil contamination at two former primitive asphalt pavement plants, Jänesselja asfaltbetoonitehas in Pärnu County, Southeast Estonia and Maadevahe asfaltbetoonitehas on Saaremaa Island, West Estonia. First of all, GC-FID chromatograms were recorded for an initial screening, in order to determine soil types and estimate the extent of weathering. For a detailed fingerprinting, the content and distribution of PAHs were determined by GC-MS. The binary ratios of PAHs were calculated. The samples from both APPs were compared and, using oil standards (diesel, light fuel oil (LFO), used motor oil, shale oil, heavy fuel oil (HFO)) were classified. PCA and CA were used to cluster the samples. The binary ratios of PAHs were used to distinguish their potential sources in the environmental samples.

2. Material and methods

2.1. Soil samples, reagents and equipment

The chemical fingerprints of 20 spilt oil-containing soil samples from Jänesselja APP and 36 samples from Maadevahe APP were analysed. The samples were collected from various locations and from different depths. Prior to the analyses, the samples were registered and equipped with numbers, and stored at 4 °C. Hexane was used as a solvent for extraction of oils and PAHs. In this study, for PAHs calculation, the specific internal standards (ISTDs) Naphthalene D8, Acenaphthene D10, Phenanthrene D10, Anthracene D10, Pyrene D10, Benz(*a*)anthracene D12, Benzo(*a*)pyrene D12 and Dibenz(*ah*)-anthracene D14 were chosen. The mixtures of the above-mentioned standards were prepared and spiked for all the samples, together with the blank sample and control samples. For the HOI internal standard, *n*-tetracontane with a final concentration of 20 mg/l in an extraction solution (hexane) was used. Analyses for HOI were performed on an Agilent 7890B equipped with a flame ionization detector (FID), and an Agilent 7693 autosampler. The column used

for the analyses was an Agilent J&W GC Column DB-1 (15 m × 320 μm × 0.10 μm). The analyses of the target EPA PAH compounds were performed on an Agilent 6890N GC system equipped with an Agilent Technologies 5973 mass selective detector, and an Agilent 7683 Series Autosampler. The column used was the Zebron 20 m × 180 μm × 0.18 μm ZB-5MS. In case of both instruments, all the samples were injected in a splitless injection mode, with helium as the carrier gas. Further, the PAH quantitation was performed in the selected ion monitoring (SIM) mode. The validations were established based on six calibration points for PAHs and on eight calibration points for HOI, with the correlation coefficient (R^2) greater than 0.995 for each component.

2.2. Samples preparation

The samples were prepared according to a modified method, ISO 16703:2011 [25]. Approximately 10 g of each soil sample was weighed into a tube. The PAH mix internal standards were spiked into the soil samples from APPs, the blank sample and control samples. 10 ml of acetone, 5 ml of the extraction solution with the HOI internal standard and 5 g of NaCl were then added to the sample tube. The samples were shaken for 16 hours. 1 ml of each extract was used for PAHs analysis by GC-MS and approximately 4 ml of the rest of the solutions was eluted through a Florisil column for the HOI analysis by GC-FID. The polar substances were removed by a clean-up with Florisil and the non-polar compounds were eluted through the column with the extraction solutions. One blank sample was processed together with the soil samples. The limit of quantification (LOQ) of each PAH was 0.003 mg/kg and 20 mg/kg for HOI. The identification of the compounds was based on the retention times of the calibration standards. The quantification of PAHs was performed using the internal standard quantification method, by comparing the area of the quantification ion to that of the corresponding deuterated quantification standard.

2.3. Principal component analysis

The initial data matrix consisted of 54 samples (rows) and concentrations of 16 individual PAHs (columns). Two samples were excluded from the chemometric analysis because the concentrations of most of the PAHs contained were below LOQ. The PAH concentrations were calculated as follows:

$$C_{\mathrm{PAH}} = \frac{(C - C_0) \times V}{M}, \tag{1}$$

where C_{PAH} is the PAH concentration in the soil sample, mg/kg (on a dry matter basis); C is the PAH concentration obtained from the calibration curve, mg/l; C_0 is the PAH concentration in the blank sample obtained from the calibration curve, mg/l; V is the extraction solution volume, l; and M is the weight of the

soil sample, kg (on a dry matter basis).

To identify soil pollution on a dry matter basis, the binary ratios of PAH isomers, such as Ant/Phe, Fla/Pyr, Ant/(Ant + Phe), Fla/(Fla + Pyr), Ind/(Ind + B(*g*,*h*,*i*)P) and B(*a*)Ant/(B(*a*)Ant + Chr), and the ratios of the four-to-six-ring parent PAHs to the sum of the two- and three-ring parent PAHs (HMW/LMW) were calculated. In PCA, a matrix consisting of 54 samples and 7 PAH binary ratios was used.

Prior to PCA, the data was standardised using two methods, Min-Max scaling and autoscaling. In the Min-Max scaling (normalisation), the data was scaled to a fixed range, usually to 0 to 1, and this was typically calculated via the following equation:

$$X\text{norm} = \frac{X - X\text{min}}{X\text{max} - X\text{min}}, \tag{2}$$

where *X*norm stands for the the normalised *X* value, *X*min is the minimum *X* value and *X*max is the maximum *X* value.

In order to avoid the problem of incompatibility between different scales, the data was often centred and all the values were divided by the standard deviation for each variable. The autoscaling was performed via the following equation:

$$X\text{aut} = \frac{X - X\text{avg}}{X\text{std}}, \tag{3}$$

where *X*aut means the *X* value (recalculated value) after using the autoscaling method, *X*avg stands for the average *X* value and *X*std denotes the standardised *X* value.

The Min-Max scaled and autoscaled data were then subjected to PCA. In this study, the Microsoft Excel Macro using the nonlinear iterative partial least squares (NIPALS) algorithm was applied to finding the eigenvectors of the most important principal components [26–28].

2.4. Cluster analysis

The hierarchical clustering methods grouped the objects according to the similarity between them. Based on the Jaccard similarity index, two sites that were most similar to each other were classified into one group. The Jaccard similarity index was calculated by the following equation:

$$Sj = \frac{c}{a + b - c}, \tag{4}$$

where *a* is the number of PAH species in sample A, *b* is the number of PAH species in sample B and *c* is the number of similar PAH species in sample A and sample B [29, 30].

To quantify the distance between the two clusters, single, complete and average linkages were used.

2.4.1. Single linkage clustering

As in all agglomerative cluster analyses, the single linkage began with a matrix of similarity (or dissimilarity) coefficients. Firstly, the most similar pair(s) of the samples, or the first clusters, was(were) found. The second most similar pair(s) of the samples, or the highest similarity between a sample and the first cluster, whichever was greater, was(were) then found. A disadvantage of the single linkage clustering was that it tended to produce long clusters.

2.4.2. Complete linkage clustering

Unlike the single linkage clustering, the complete linkage clustering was often inclined to generate opposite extremes by producing highly compact clusters. The method calculated the similarity measures after the new groups were formed and two groups with the highest similarity were always merged first.

2.4.3. Average linkage clustering

The average linkage clustering, differently from the single and complete linkage clusterings, produced no extremes. In order to compute the average similarity between a sample and the existing cluster, the type of "average" had to be precisely defined by using the unweighted or weighted technique. If group A consisted of *Na* objects and group B of *Nb* objects, then in the unweighted technique, the new similarity (*Sab*) was calculated by the following equation:

$$Sab = \frac{NaSa+NbSb}{Na+Nb}, \quad (5)$$

where *Na* is the number of objects in group A and *Nb* is the number of objects in group B.

The weighted mean value was calculated as follows [23, 30]:

$$Sab = \frac{Sa+Sb}{2}. \quad (6)$$

For defining the similarity between the samples, R software was used. R is a programming language for statistical analysis, graphical representation and reporting [31].

3. Results and discussion

3.1. Initial screening

In this study, GC-FID was used for an initial screening of oils. The respective chromatograms of soil samples were compared with those of control oils and the levels of oils weathering were estimated. Figure 2 shows the chromatograms of control oils used in the study for identification of the following oil spills: diesel, LFO, shale oil grade C, used motor oil, fresh motor

oil and HFO. The retention times between 3.5 and 6 min were attributable to HOI from decane (C_{10}) to heneicosane (C_{21}), and the retention times between 6 and 8.5 min were assignable to fractions C_{21}–C_{40}. From Figure 2a it can be seen that the major components in diesel were alkanes C_{10}–C_{24}, while LFO (Fig. 2b) contained mostly alkanes C_{10}–C_{32}. Motor oils (Figs. 2d and 2e) were found to consist mainly of unresolved complex mixtures (UCMs), while no UCMs were identified in diesel (Fig. 2a) or LFO (Fig. 2b). Figure 3 depicts the GC-FID chromatograms of soil samples from Jänesselja and Maadevahe APPs. All the chromatograms indicate the presence of the mixture of different oils. This study found no samples that would have been polluted with one type of oil which would have been similar to control oils. The samples from Jänesselja APP reveal different patterns of chromatograms, depending on the source background (Figs. 3b, 3c, 3d). These chromatograms display pollution with different oils, like diesel, LFO, lubricating oil, HFO, waste oil and shale oil. Hydrocarbons are represented by a wide range of species, from *n*-decane

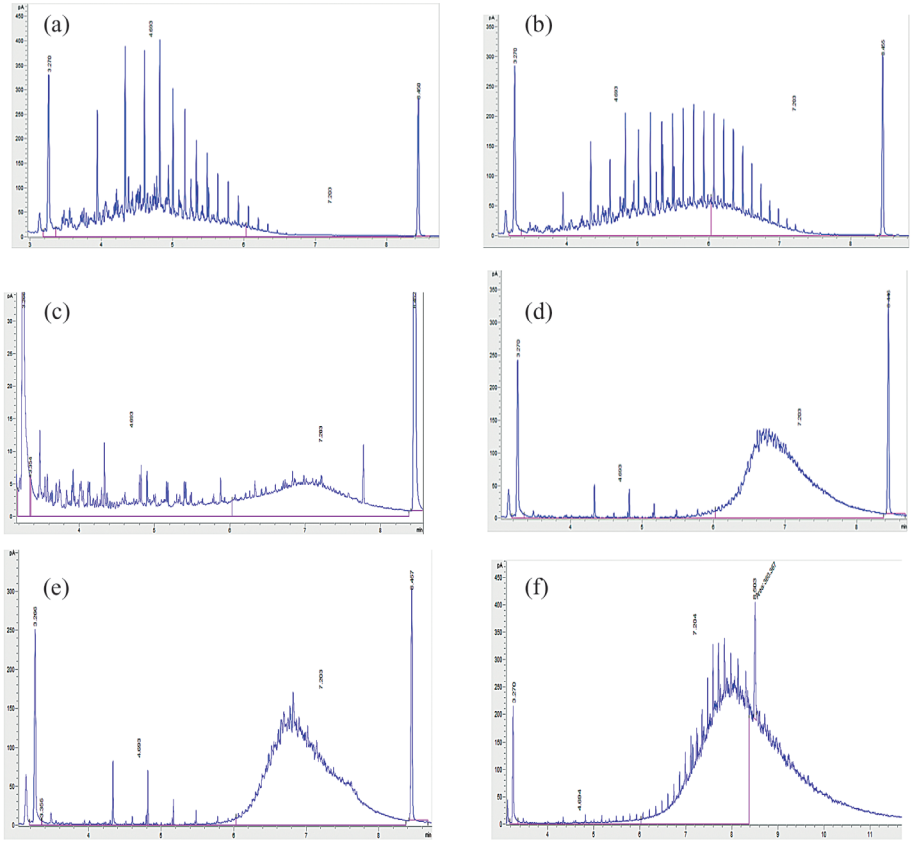

Fig. 2. GC-FID chromatograms of control oils: (a) diesel, (b) LFO, (c) shale oil grade C, (d) used motor oil, (e) fresh motor oil, (f) HFO.

to high molecular weight paraffins (> C_{40}). The GC-FID chromatograms of selected oil samples from Maadevahe APP (Fig. 3a) exhibit a similar pattern of chromatograms (hydrocarbons C_{10}–C_{40}), which suggests soil pollution sourced from one type of oil or a mixture of oils. Based on this information, mainly shale oil was used by Maadevahe APP [32].

The soil samples from Maadevahe APP (Fig. 3a) and sample 16 from Jänesselja APP (Fig. 3b) were probably polluted with a mixture of diesel, sample 14 (Fig. 3c) was polluted with a mixture of LFO, waste oil and HFO, and sample 4 (Fig. 3d), with a mixture of LFO and shale oil. When petroleum products are released into the environment, they tend to weather by evaporation, water solubilisation and oxidation. Oxygen from the air and biological organisms transform the petroleum products, increasing the degradation of *n*-alkanes. After *n*-alkanes have been removed, the remaining constituents appear as a hump on the GC-FID chromatograms, with a few discernible peaks. The chromatograms of oil samples from both APPs show much lower *n*-alkane concentrations (small peaks) and higher phytane concentrations in them, in addition to UCMs, than in fresh oils. The loss of *n*-alkanes and increase in UCMs are indicative of oil weathering processes [33, 34].

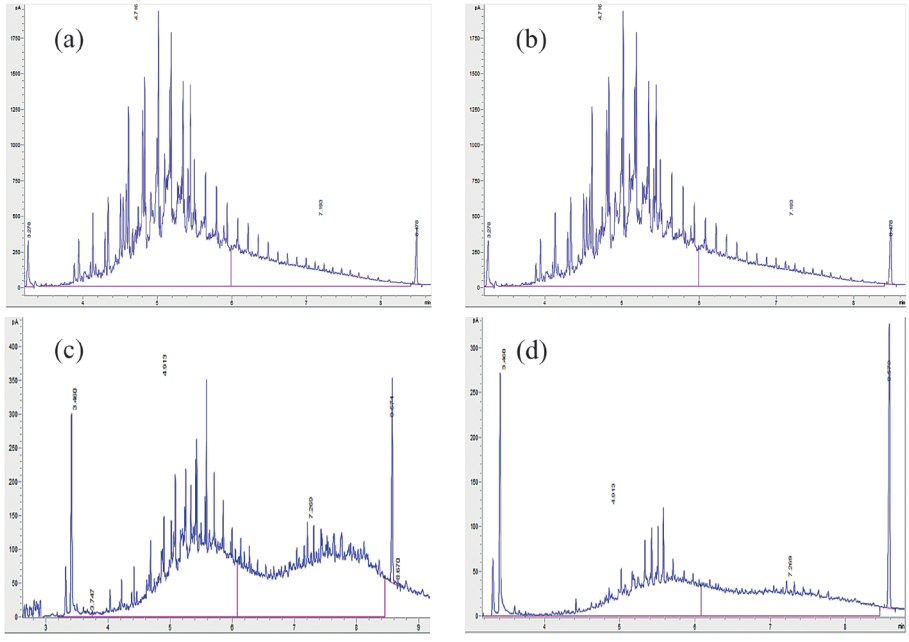

Fig. 3. GC-FID chromatograms of samples from: (a) Maadevahe APP, (b)–(d) Jänesselja APP.

3.2. Determination of PAHs in soils

PAHs in soil were identified using a modified method, ISO 18287:2006 [35]. For the identification of the compounds, one appropriate analyte retention time was chosen and two or three different fragmentation ions (quantifier and qualifier) were used (Table 1). The qualifier to quantifier ion area ratio, m/z, is unique and should not differ more than ± 20% from the same ratio in calibration standards.

The most sensitive ions were used to calculate the contents of individual PAHs by using the corresponding calibration curve. The analyte concentrations were calculated from the ratio of the samples to the internal standard peak area.

Table 1. Ions used for the qualification of PAHs and ISTDs used for the quantification of the final concentrations of PAHs

PAH	Mass fragment, m/z	Internal standard (ISTD) and m/z
Naphthalene	128, 127	Naphthalene D8; 136,137
Acenaphthylene	152, 151	Acenaphthene D10; 162, 164
Acenaphthene	154, 152	Acenaphthene D10; 162, 164
Fluorene	166, 165	Acenaphthene D10; 162, 164
Phenanthrene	178, 176	Phenanthrene D10; 188,184
Anthracene	178, 176	Anthracene D10; 188, 184
Fluoranthene	202, 200	Pyrene D10; 212, 210
Pyrene	202, 200	Pyrene D10; 212, 210
Benz(*a*)anthracene	228, 226	Benz(*a*)anthracene; 240, 236
Chrysene	228, 226	Benz(*a*)anthracene; 240, 236
Benzo(*b*)fluoranthene	252, 250	Benzo(*a*)pyrene D12; 264, 260
Benzo(*k*)fluoranthene	252, 250	Benzo(*a*)pyrene D12; 264, 260
Benzo(*a*)pyrene	252, 250	Benzo(*a*)pyrene D12; 264, 260
Dibenz(*ah*)anthracene	278, 277, 276	Dibenz(*ah*)anthracene D14;292, 291
Benzo(*g*,*h*,*i*)perylene	276, 277	Dibenz(*ah*)anthracene D14;292, 291
Indeno(*1*,*2*,*3-cd*)pyrene	276, 277	Dibenz(*ah*)anthracene D14;292, 291

3.3. Detailed fingerprinting

3.3.1. Soil samples contamination identification using PAH binary ratios and PCA

Seven PAH diagnostic ratios, namely Ant/Phe, Fla/Pyr, Ant/(Ant + Phe), Fla/(Fla + Pyr), Ind/(Ind + B(*g*,*h*,*i*)P), B(*a*)Ant/(B(*a*)Ant + Chr) and HMW/LMW, as well as PCA were used to identify the sources of oil in the soil samples from Jänesselja and Maadevahe APPs. The same PAH diagnostic ratios were also calculated for control oils, like diesel, LFO, used motor oil, shale oil and HFO, and compared with those for the soil samples. Figure 4 shows the average PAH binary ratios for the samples from both APPs, with standard deviations. All the above ratios of samples from Jänesselja APP had higher standard deviations, which indicated that the samples differed from each other to a greater degree than did those from Maadevahe APP. The high dispersion of Ant/Phe and HMW/LMW ratios just indicates a wider distribution in the grouping of the samples in further PCA, but this does not influence the grouping itself.

The average binary ratio of Ind/(Ind + B(*g*,*h*,*i*)P) for all the 54 samples was 0.48. One sample had a value > 0.5, which might be due to biomass combustion. The other samples had values between 0.42 and 0.5, indicating petroleum combustion as the possible source of PAHs. The variation in PAH sources revealed by these indices could be due to the different sampling sites, with different hydrological and spatial conditions. The Ant/(Ant + Phe) for

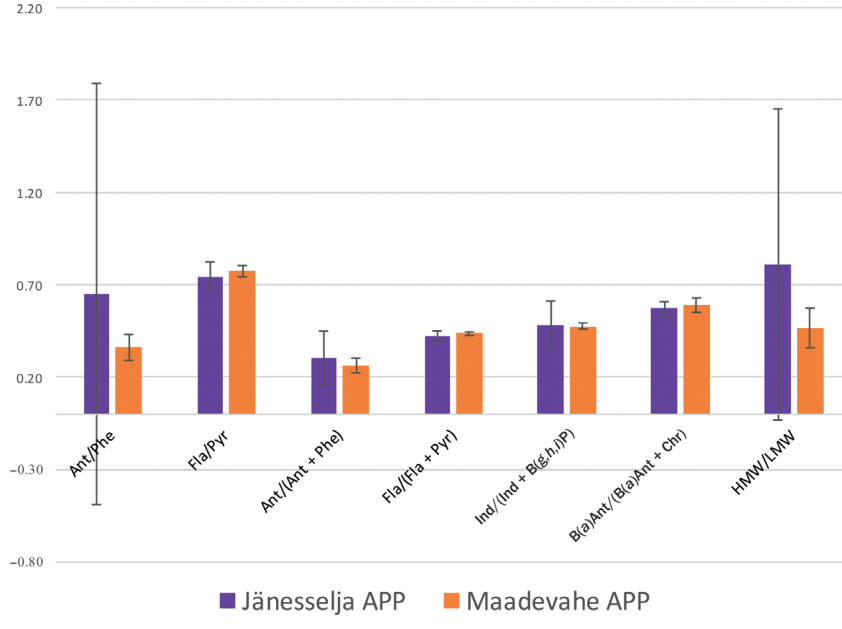

Fig. 4. The average PAH binary ratios for the samples from Jänesselja and Maadevahe APPs, with standard deviations.

Table 2. PAH binary ratios for samples from Jänesselja and Maadevahe APPs

PAH binary ratio / Sample	Ant/Phe	Fla/Pyr	Ant/(Ant + Phe)	Fla/(Fla + Pyr)	Ind/(Ind + B(*g*,*h*,*i*)P)	B(*a*)Ant/(B(*a*)Ant + Chr)	HMW/LMW
Jänesselja APP	0.21–5.19	0.63–0.95	0.18–0.84	0.38–0.49	0.42–0.71	0.51–0.65	0.02–2.63
Maadevahe APP	0.18–0.50	0.70–0.87	0.15–0.33	0.41–0.46	0.44–0.50	0.38–0.62	0.06–0.70

the samples from Jänesselja APP was in the range of 0.18–0.84 and for those from Maadevahe APP, between 0.15 and 0.33 (Table 2). Generally, Ant/(Ant + Phe) < 0.1 indicates that PAHs in the soil sourced from petroleum combustion, whereas Ant/(Ant + Phe) > 0.1 might mean contamination from wood and coal combustion [10].

PAHs in the soils with Fla/(Fla + Pyr) < 0.4, B(*a*)Ant/(B(*a*)Ant +Chr) < 0.2 and Ind/(Ind + B(*g*,*h*,*i*)P) < 0.2 were mainly from petroleum contamination. PAHs with ratios of 0.4 < Fla/(Fla + Pyr) < 0.5, 0.2 < B(*a*)Ant/(B(*a*)Ant + Chr) < 0.4 and 0.2 < Ind/(Ind + B(*g*,*h*,*i*)P) < 0.35 originated predominantly from the combustion of petroleum. PAHs with Fla/(Fla + Pyr) > 0.5, B(*a*)Ant/(B(*a*)Ant + Chr) > 0.4 and Ind/(Ind + B(*g*,*h*,*i*)P) > 0.35 came chiefly from the combustion of coal, shale oil and biomass. Most soil samples from APPs had Fla/(Fla+Pyr) < 0.5, B(*a*)Ant/(B(*a*)Ant + Chr > 0.4 and Ind/(Ind + B(*g*,*h*,*i*)P) > 0.35. The above values suggested that contamination at both APPs originated from the combustion of oil shale and petroleum [10].

The scores plot in Figure 5 shows the PAH binary ratios of 54 samples from APPs, which contained diesel, LFO, used motor oil, shale oil and HFO as the first and second principal components (PC1 and PC2 respectively). The first two PCs comprised 70.69% of the total components of the complex mixture of oil samples. The first principal components were responsible for 45.46% of the total oil spill compounds, while the second principal components were responsible for 25.23% of the total variance. The scores plot reveals that the soil samples mostly contained used motor oil. In this study, the researchers used for analysis fresh motor oil in the mixture of different used motor oils, from diesel to engine petroleum. Diesel, shale oil, HFO and LFO were not identified in the soil samples. The chemical composition of weathered oils differed from that of fresh oils. The most noticeable changes in PAHs composition in the weathered soils were the depletion of naphthalenes, degradation of alkylated PAHs and increase of chrysenes [33, 36]. Being situated distant from the other

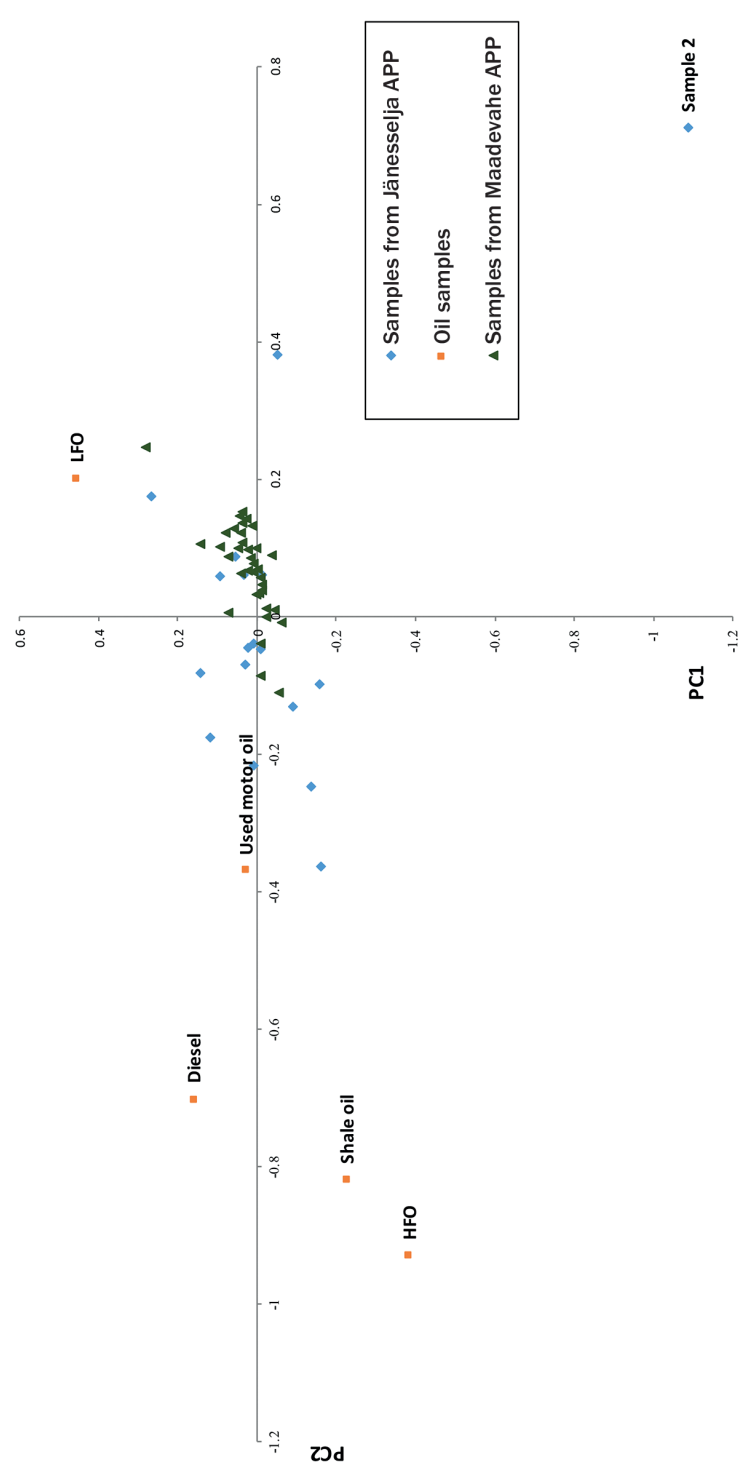

Fig. 5. The scores plot of PAH binary ratios of the samples from Jänesselja and Maadevahe APPs and oil samples.

soil samples and control oils in Figure 2, sample 2 from Jänesselja APP was identified as an outlier. The Ant/Phe ratio of the sample was 5, while for all the other samples it was < 0.66. The source of anthracene in sample 2 could not be identified.

3.3.2. The classification of soil samples from Jänesselja APP

To group the soil samples from Jänesselja APP, another PCA approach was applied, leaving sample 2 out as an outlier from further analysis (Fig. 6).

Sample 4 was classified into an individual cluster. Unlike all the other samples, this sample had a high Fla/Pyr ratio. With Ind/(Ind + B(*g*,*h*,*i*)P) > 0.5, sample 13 was distinguished from the other samples whose respective ratio was considerably lower. This suggested contamination of the sample with PAHs from combustion of biomass. The HMW/LMW of sample 15 was less than 0.1, in samples 10 and 1 more than 2 and in samples 14 and 7, approximately 1. HMW/LMW ≥ 1.0 was indicative of a pyrogenic source and HMW/ LMW < 1.0 suggested a petrogenic source. Most samples from Jänesselja and Maadevahe APPs had similar PAH binary ratios and could therefore be classified into one group. Contamination in the soil samples from both plants originated most likely from shale oil combustion. Samples 4 and 13 could be clearly differentiated from the other samples by PAH binary ratios. Samples 1, 3, 7, 10, 12 and 15 could also be distinguished from the other samples by PAH binary ratios, but this difference was not as great as in case of samples 3 and 14. Based on PCA, several samples were found to be similar in PAH binary ratios. Sample pairs 5 and 6, 9 and 11, and 16 and 17 were established to be similar in composition, based on origin, sampling time and location. However, in the case of sample pairs 3 and 15, 1 and 10, and 14 and 7, the similarity could not be traced, based on the information available. The oil samples that were closely located in the scores plot (central cluster) had a similar chemical composition, based on PAH binary ratios (Ant/Phe, Fla/Pyr, Ant/(Ant + Phe), Fla/(Fla + Pyr), Ind/(Ind + B(*g*,*h*,*i*)P, B(*a*)Ant/(B(*a*)Ant + Chr), and HMW/LMW). Conversely, the oil samples that were located far apart in the scores plot differed in chemical composition, and this dissimilarity increased as the distance between the samples increased. Figure 7 shows the scores plot and the loadings plot of the samples. The grouping of the samples was mainly based on the ratio HMW/LMW. In grouping sample 2, mainly Ant/Phe and Ant/(Ant + Phe) were taken into account, while sample 4 was grouped chiefly on the basis of Fla/Pyr and Fla/(Fla + Pyr), and sample 13, on the basis of Ind/(Ind + B(*g*,*h*,*i*)P).

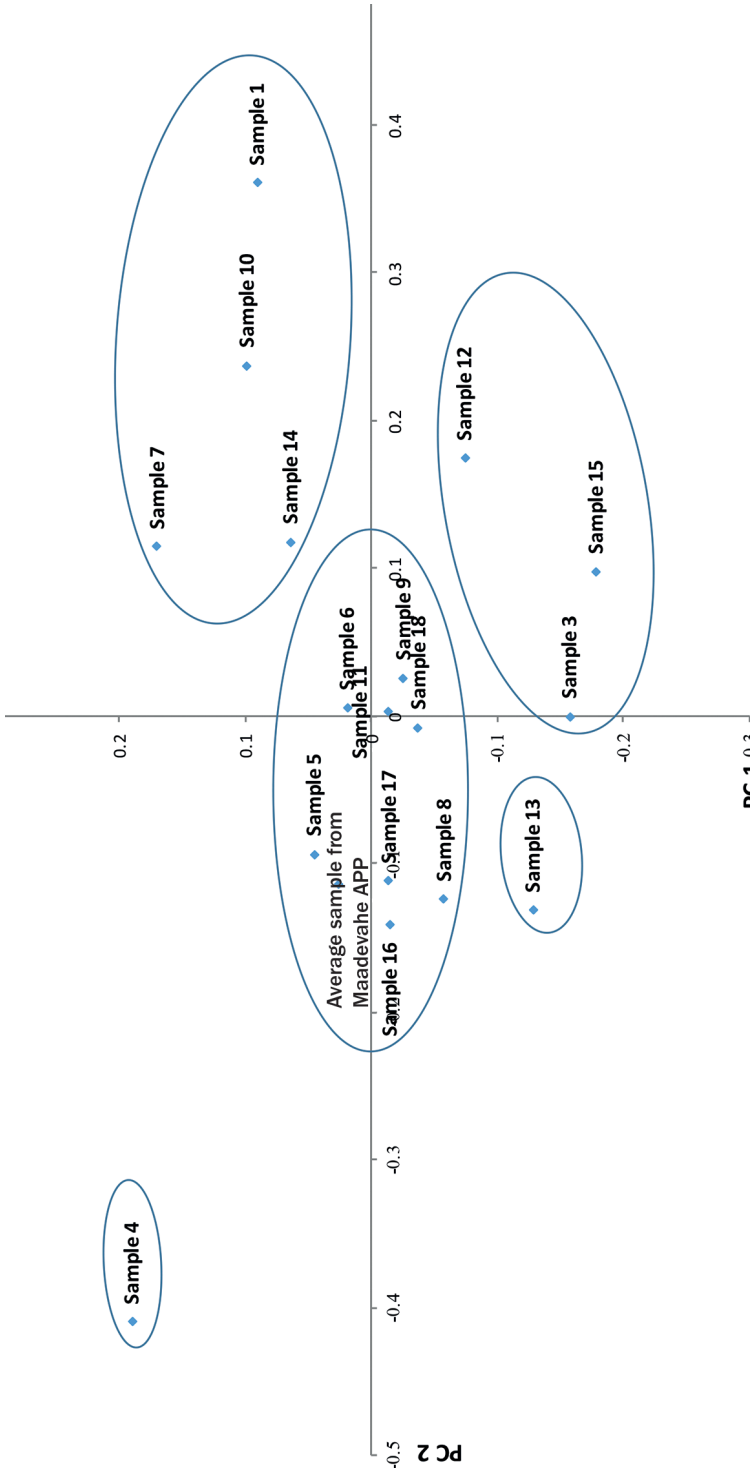

Fig. 6. The grouping of samples from Jänesselja APP on the basis of PAH binary ratios.

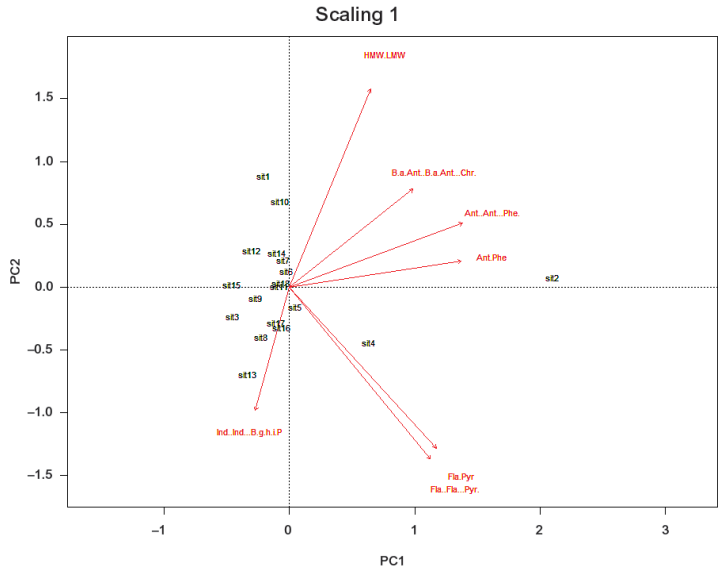

Fig. 7. The scores plot and the loadings plot of samples.

3.3.3. Cluster analysis

Another chemometric technique, cluster analysis, was used for clustering the samples. The data consisted of binary ratios of seven PAH isomers (Ant/Phe, Fla/Pyr, Ant/(Ant + Phe), Fla/(Fla + Pyr), Ind/(Ind + B(*g*,*h*,*i*)P), B(*a*)Ant/(B(*a*)Ant + Chr), and HMW/LMW) contained in 18 samples from Jänesselja APP. Distance was used as a measure of similarity/dissimilarity between the samples. Before the distances were computed, the PAH binary ratios were first normalised by the maximum-minimum range. Three kinds of linkage methods (single, average, complete) were tested in order to find out the most suitable linkage method for determination of these ratios. All the linkage methods merged the samples almost into the same groups, but the average distance between the groups better reflected the relationship between PAH binary ratios and oil types.

Figure 8 depicts the output of the average linkage method using the Jaccard distance as a similarity measure. From the figure it can be seen that the sampling sites fall into two major groups. The dendrogram obtained from HCA revealed that sampling sites with similar PAH concentrations were clustered into the same group. Based on the clusters obtained from CA using chromatogram patterns, as described in Section 3.1, the sampling sites in this study were grouped according to oil mixture type. Cluster 1 included sample 2 (the outlier). Sample 13 in cluster 2 was polluted mostly with LFO and sample 4 in cluster 3 with a mixture of shale oil and diesel. Cluster 4 consisted of samples which were primarily polluted with HFO (probably mazut), while cluster 5 comprised samples containing the mixture of shale oil and LFO.

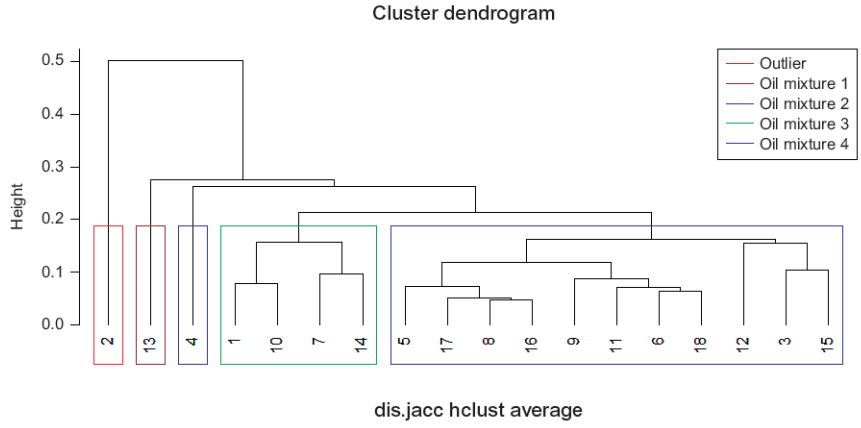

Fig. 8. The dendrogram of the groups of samples from Jänesselja APP.

Some sample pairs, namely 1 and 10, 7 and 14, 8 and 16, and 3 and 15, were located next to each other on the dendrogram (Fig. 8) and in the scores plot (Fig. 7). Oil mixtures 3 and 4 on the dendrogram included the same samples as in the scores plot (Fig. 6). PCA and CA classified the samples into the same groups. The clusters generated were highly reliable as the samples in the groups had similar PAH binary ratios and similar sampling backgrounds.

4. Conclusions

The application of two chemometric techniques, cluster analysis and principal component analysis, and two analytical methods, gas chromatography-mass spectrometry and gas chromatography-flame ionization detection, were used to determine the distribution of polycyclic aromatic hydrocarbons in the soils on the territories of past asphalt pavement plants, as well as to identify possible sources of old oils. The initial oil screening revealed that the soil samples from Jänesselja APP were contaminated with different mixtures of oils, such as diesel, light fuel oil, heavy fuel oil, lubricating oil, waste oil and shale oil. The GC-FID chromatograms of the selected oil samples from Maadevahe APP showed a similar pattern, which suggested that the soil at this plant was polluted with one type of oil, namely shale oil. The composition and physical properties of spilt oils in the soil samples changed during the weathering processes (evaporation, water solubilisation and oxidation), and differed from those of fresh oils.

The binary ratios of 16 PAHs revealed that the contamination in the samples from Maadevahe and Jänesselja APPs originated from petroleum and shale oil combustion. Based on the results of PCA and CA, oils or oil mixtures, with

which the soil samples from Jänesselja APP were mostly polluted, could be divided into four groups: oil mixture 1 (mostly LFO), oil mixture 2 (shale oil and diesel), oil mixture 3 (HFO, waste oils and diesel), oil mixture 4 (shale oil, LFO and diesel). As a result of degradation and migration of PAHs in soil, it was difficult to identify the sources of mixed oils and the relative contribution of each possible source to pollution according to PAH binary ratios only. It can be concluded that using the combination of analytical and chemometric methods in oil spill fingerprinting considerably contributed to interpreting the data, based on PAH binary ratios, while in determining possible sources of old oils, it saved a great amount of time and avoided the high costs of oils classification. PCA and CA can be useful tools to examine various relationships among different samples. These methods are able to identify similar sample groups or sample pairs in the scores plots or on the dendrograms.

Acknowledgements

The authors would like to thank Sander Sannik of Eurofins Environment Testing Estonia OÜ for his help in writing the manuscript, Mati Salu of Maves AS for assisting in sampling at Jänesselja APP, and the Estonian Ministry of Education and Research for financial support (Grant No. IUT 33-20).

REFERENCES

1. Eesti Päevaleht. *In the road construction there occurred a 6680-tonne mazut spill*. Report. http://epl.delfi.ee/news/eesti/tee-ehitusele-sattusette-6680-tonnine-masuudireostus?id=51193631. Last visited September 2018 (in Estonian).
2. AS Maves. *The control and monitoring of hazardous residual contamination*. Report. Tallinn, 2004 (in Estonian).
3. Mas, S., de Juan, A., Tauler, R., Olivieri, A. C., Escandar, G. M. Application of chemometric methods to environmental analysis of organic pollutants: A review. *Talanta*, 2010, **80**(3), 1052–1067.
4. Ramirez, M. I., Arevalo, A. P., Sotomayor, S., Bailon-Moscoso, N. Contamination by oil crude extraction – Refinement and their effects on human health. *Environ. Pollut.*, 2017, **231**(1), 415–425.
5. Eesti keskkonnauuringute keskus. *The development of the methodology for reducing residual pollution at former military and industrial objects, Phase I*. Report. 2013 (in Estonian).
6. Bruzzoniti, M. C., Fungi, M., Sarzanini, C. Determination of EPA's priority pollutant polycyclic aromatic hydrocarbons in drinking waters by solid phase extraction-HPLC. *Anal. Methods*, 2010, **2**, 739–745.
7. Jefimova, J., Irha, N., Reinik, J., Kirso, U., Steinnes, E. Leaching of poly-

cyclic aromatic hydrocarbons from oil shale processing waste deposit: A long-term field study. *Sci. Total Environ.*, 2014, **481**, 605–610.
8. Retnam, A., Zakaria, M. P., Juahir, H., Aris, A. Z., Zali, M. A., Kasim, M. F. Chemometric techniques in distribution, characterisation and source apportionment of polycyclic aromatic hydrocarbons (PAHs) in aquaculture sediments in Malaysia. *Mar. Pollut. Bull.*, 2013, **69**(1–2), 55–66.
9. Lau, E. V., Gan, S., Ng, H. K. Extraction Techniques for Polycyclic Aromatic Hydrocarbons in Soils. *Int. J. Anal. Chem.*, **2010**, Article ID 398381, 2010.
10. Pongpiachan, S., Hattayanone, M., Tipmanee, D., Suttinun, O., Khumsup, C., Kittikoon, I., Hirunyatrakul, P. Chemical characterization of polycyclic aromatic hydrocarbons (PAHs) in 2013 Rayong oil spill-affected coastal areas of Thailand. *Environ. Pollut.*, 2018, **233**, 992–1002.
11. Dudhagara, D. R., Rajpara, R. K., Bhatt, J. K., Gosai, H. B., Sachaniya, B. K., Dave, B. P. Distribution, sources and ecological risk assessment of PAHs in historically contaminated surface sediments at Bhavnagar coast, Gujarat, India. *Environ. Pollut.*, 2016, **213**, 338–346.
12. Koch, M., Liebich, A., Win, T., Nehls, I. Certified Reference Materials for the determination of mineral oil hydrocarbons in water, soil and waste. *Forschungsbericht 272*, Berlin, 2005.
13. Konečný, F., Boháček, Z., Müller, P., Kovářová, M., Sedláčková, I. Contamination of soils and groundwater by petroleum hydrocarbons and volatile organic compounds – Case study: ELSLAV BRNO. *Bull. Geosci.*, 2003, **78**(3), 225–239.
14. Wüst, B. *Measuring Hydrocarbon Oil Index according to ISO 9377-2 (DIN H53). Environmental Application. Ga Chromatography*. https://www.agilent.com/cs/library/applications/app_note.pdf. Last visited October 2018.
15. Christensen, J. H., Tomasi, G. Practical aspects of chemometrics for oil spill fingerprinting. *J. Chromatogr. A*, 2007, **1169**(1–2), 1–22.
16. Lubes, G., Goodarzi, M. Analysis of volatile compounds by advanced analytical techniques and multivariate chemometrics. *Chem. Rev.*, 2017, **117**(9), 6399−6422.
17. Hopke, P. K. Chemometrics applied to environmental systems. *Chemom. Intell. Lab. Syst.*, 2015, **149**(B), 205–214.
18. Singh, I., Juneja, P., Kaur, B., Kumar, P. Pharmaceutical applications of chemometric techniques. *ISRN Anal. Chem.*, **2013**, Article ID 795178, 2013.
19. Mali, M., Dell'Anna, M. M., Notarnicola, M., Damiani, L., Mastrorilli, P. Combining chemometric tools for assessing hazard sources and factors acting simultaneously in contaminated areas. Case study: "Mar Piccolo" Taranto (South Italy). *Chemosphere*, 2017, **184**, 784–794.
20. Panchuk, V., Yaroshenko, I., Legin, A., Semenov, V., Kirsanov, D. Application of chemometric methods to XRF-data - A tutorial review. *Anal.*

Chim. Acta, 2018, **1040**, 19–32.
21. Gad, H. A., El-Ahmady, S. H., Abou-Shoer, M. I., Al-Azizi, M. M. Application of chemometrics in authentication of herbal medicines: a review. *Phytochem. Analysis*, 2013, **24**(1), 1–24.
22. Novák, M., Palya, D., Bodai, Z., Nyiri, Z., Magyar, N., Kovács, J., Eke, Z. Combined cluster and discriminant analysis: An efficient chemometric approach in diesel fuel characterization. *Forensic Sci. Int.*, 2017, **270**, 61–69.
23. Brereton, R. G. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. Wiley, Chichester, 2003.
24. Miki, S., Uno, S., Ito, K., Koyama, J., Tanaka, H. Distributions of polycyclic aromatic hydrocarbons and alkylated polycyclic aromatic hydrocarbons in Osaka Bay, Japan. *Mar. Pollut. Bull.*, 2014, **85**(2), 558–565.
25. ISO 16703:2011. *Soil quality - Determination of content of hydrocarbon in the range C10 to C40 by gas chromatography*, 2011.
26. Hupp, A. M., Marshall, L. J., Campbell, D. I., Smith, R. W., McGuffin, V. L. Chemometric analysis of diesel fuel for forensic and environmental applications. *Anal. Chim. Acta*, 2008, **606**(2), 159–171.
27. Raschka, S. *About Feature Scaling and Normalization – and the effect of standardization for machine learning algorithms*. https://sebastianraschka.com/Articles/2014_about_feature_scaling.html. Last visited November 2018.
28. Risvik, H. *Principal Component Analysis (PCA) & NIPALS algorithm*, 2007. https://folk.uio.no/henninri/pca_module/pca_nipals.pdf. Last visited November 2018.
29. Oksanen, J. *R package version. Multivariate Analysis of Ecological Communities in R: vegan tutorial*, 2011.
30. Romesburg, H. C. *Cluster Analysis for Researchers*. Lulu Press, 2004.
31. Zuur, A., Ieno, E., Meesters, E. *A Beginner's Guide to R*. Springer Publishing Company, 2009.
32. AS Maves. http://www.maves.ee/Projektid/2003/OJRK_pingerida.htm. Last visited January 2019 (in Estonian).
33. Barakat, A. O., Mostafa, A. R., Qian, Y., Kennicutt II, M. C. Application of petroleum hydrocarbon chemical fingerprinting in oil spill investigations – Gulf of Suez, Egypt. *Spill Sci. Technol. B.*, 2002, **7**(5–6), 229–239.
34. Zemo, D. A., Bruya, J. E., Graf, T. E. The application of petroleum hydrocarbon fingerprint characterization in site investigation and remediation. *Ground Water Monit. R.*, 1995, **15**(2), 147–156.
35. ISO 18287:2006. *Soil quality - Determination of polycyclic aromatic hydrocarbons (PAH) - Gas chromatographic method with mass spectrometric detection (GC-MS)*, 2006.
36. Malmquist, L. M., Olsen, R. R., Hansen, A. B., Andersen, O.,

Christensen, J. H. Assessment of oil weathering by gas chromatography – mass spectrometry, time warping and principal component analysis. *J. Chromatogr. A*, 2007, **1164**(1–2), 262–270.

Presented by A. Siirde
Received April 10, 2019