EUROPEAN NETWORK OF FORENSIC SCIENCE INSTITUTES
DRUGS WORKING GROUP

Guideline for the use of Chemometrics in Forensic Chemistry

Ref. Code: DWG-CFC-001
Issue No: 001
April 2020

INDEX

1 AIMS

There is a clear increasing trend in the use of chemometrics in forensic laboratories. This can be seen in forensic literature covering different disciplines such as drug profiling, arson debris analysis, spectral imaging, glass analysis, age determination, and more. In particular, current chemometric applications cover spectral (i.e. FT-IR) and chromatographic (i.e. GC-MS) data. All this has created a need for reliable and structured interpretation and assessment of both analytical and chemometric results.

From a literature survey the recently used data pretreatments and chemometric methods in forensic chemistry were identified. The common practices of chemometrics are collected in order to help forensic scientists understand and utilize chemometrics in their everyday work tasks. Also, an easy to use software tool (ChemoRe) is created for this purpose

This Guideline and the software tool (ChemoRe) aims to provide an easy starting point for a forensic chemist to apply chemometrics and sharpen the understanding of the critical points in the forensic workflow from an incident to reporting of results. Together, the guideline and software tool will support routine forensic work and help create high-quality measures and processes that authorities can rely on.

This work is part of EU project: Steps towards European Forensic Area (STEFA) - Work package G02: Chemometrics: easy to use tools to process and interpret chemical data of illicit drug samples.

2 SCOPE

The initiative to develop a guideline accompanied with a tailored software tool for forensic chemists enabling them to apply chemometric methods was taken by the European Network of Forensic Science Institutes (ENFSI) Drugs Working Group (DWG) during the Annual Meeting in 2015. An initial team of three senior forensic chemists and three forensic statisticians formed the DWG subcommittee Chemometrics. This combination of expert knowledge quickly proved fundamental. Thorough discussions were required to achieve a mutual understanding of knows and not-knows performing chemometrics in forensic chemistry in routine casework. Familiar standard terms in each discipline like identification, discrete variable, continuous (spectral) data, classification, interpretation etc., allegedly understood by everyone in the same way, had to be discussed in depth and agreed until the objectives of the project were unequivocally understood.

The ENFSI DWG identified a deficiency in knowledge as well as in development and application of chemometric methods to solve questions regarding material comparison or classification either between seized samples or seized samples against a database. While the Working Group has organized trainings and workshops over several years, the number of laboratories exploiting chemometrics did not increase although the need was expressed. It turned out that some laboratories developing and applying chemometric methods had employed statisticians among their personnel. Even with an understanding of the possibilities that chemometric methods can significantly support the forensic casework, the laboratories without a statistician on board hardly developed in house methods. A further obstacle was that no easy-to-use software tool, applicable by a forensic chemist without a broader statistical background, was available.

The DWG subcommittee Chemometrics was therefore requested to fill this gap and to compose a tailored guideline and an easy-to-use software tool (called ChemoRe) for development and application of chemometric methods. The software will help the laboratory to apply chemometrics starting from the data of the analytical instrument over validation of the method until assessment and reporting of the chemometric results. It was found essential to have a clear understanding of the forensic workflow and what are the questions assigned to the forensic laboratory as well as how the chemometric workflow proceeds. These are illustrated in Figures 4.1 and 4.2.

It is to be noted here, that in addition to chemometric methods this guideline focuses on the assessment and interpretation of chemometrics results with respect to chemical data. This approach applies no matter of the legal systems of different countries.

When final reporting is considered, national regulations, standards and practices need to be taken into account. Therefore, the final reporting phase, potentially also covering further steps like evaluation (hypothesis-based determination of likelihood ratios), is not described in depth (see Chapter 10).

The project found great interest regarding quick realisation, exceeding the fields of drug analysis, as the principles can be applied in any other forensic discipline where similar data or databases occur. Probably due to this broad applicability, the work received funding through the European Union’s Internal Security Fund — Police: 779485 — STEFA — ISFP-2016-AG-IBA-ENFSI during a period of 29 months from January 2018 to May 2020.

The content of this Guideline represents the views of the authors only and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

However, according to the concept of the project, the core elements in this guideline have been published in a peer reviewed journal [1, 2, 3]. The software tool named ChemoRe is based on R, a free software environment for statistical computing and graphics (https://www.r-project.org/) and depends heavily on the shiny-package (Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application Framework for R. https://CRAN.R-project.org/package=shiny).

This guideline is available on the ENFSI website (http://enfsi.eu/documents/forensic-guidelines/). The software ChemoRe with its validation report and the user manual of ChemoRe are made available via the ENFSI webpage on EPE (Europol Platform for Experts). Further information on access can be obtained from the acting chairman of the ENFSI Drugs Working Group (http://enfsi.eu/about-enfsi/structure/working-groups/drugs/).

3 DEFINITIONS AND TERMS

The abbreviations given refer to the analytical and chemometric methods that are used throughout the guideline.

Table 3.1: Abbreviations and short descriptions of analytical and chemometric methods

Analytical methods	
Abbreviation	**Short description**
GC-MS	Gas chromatography mass spectrometry
LC-TOFMS, LC-MS/MS, LC-MS	Liquid chromatography time-of-flight mass spectrometry, Liquid chromatography tandem mass spectrometry, Liquid chromatography mass spectrometry
GC-FID	Gas chromatography flame ionization detector Fast gas chromatography flame ionization detector
GC-IRMS	Gas chromatography isotopic ratio mass spectrometry
ICP-MS	Inductively coupled plasma mass spectrometry
ICP-AES	Inductively coupled plasma atomic emission spectrometry
XRF / EDXRD	X-ray diffraction spectrometry / Energy-dispersive X-ray diffraction
FT-IR	Fourier-transform infrared spectroscopy
MIR	Mid infrared spectroscopy
NIR	Near infrared spectroscopy
Raman	Raman spectroscopy

Chemometric methods	
PCA	*Principal component analysis*
OLS-R	*Ordinary least squares-regression*
PLS-DA	*Partial least squares-discriminant analysis*
HCA	*Hierarchical cluster analysis*
LDA	*Linear discriminant analysis*
PLS-R	*Partial least squares-regression*
SVM	*Support vector machines*
QDA	*Quadratic discriminant analysis*
PCC	*Pearson's correlation coefficient*
SIMCA	*Soft independent modelling by class analogy*
LogReg	*Logistic regression*
CCSWA	*Common components and specific weights analysis*
k-NN	*k-nearest neighbours*
NN	*Neural networks*
OPLS-DA	*Orthogonal projections to latent structures-discriminant analysis*
PA	*Predictive agreement*
PCR	*Principal component regression*
RF	*Random forest modelling*
DA	*Discriminant analysis*

4 INTRODUCTION

4.1 Chemometrics in Forensic Chemistry

Forensic literature shows a clear trend towards increasing use of chemometrics (i.e. multivariate analysis and other statistical methods) when forensic data is gathered. This can be seen in different disciplines such as drug profiling, arson debris analysis, spectral imaging, glass analysis, age determination, and more. In particular, current chemometric applications cover low-dimensional (in this guideline also described as **Type 1** data, e.g. drug impurity profiles, see also Chapter 5.3) [4-5] as well as high-dimensional data (**Type 2** data, e.g. FT-IR spectra) [6-8] and are therefore useful in many forensic disciplines. Because of this, there is an increasing need in forensic chemistry for guidance on how to perform reliable and structured processing, analysis and interpretation of analytical data.

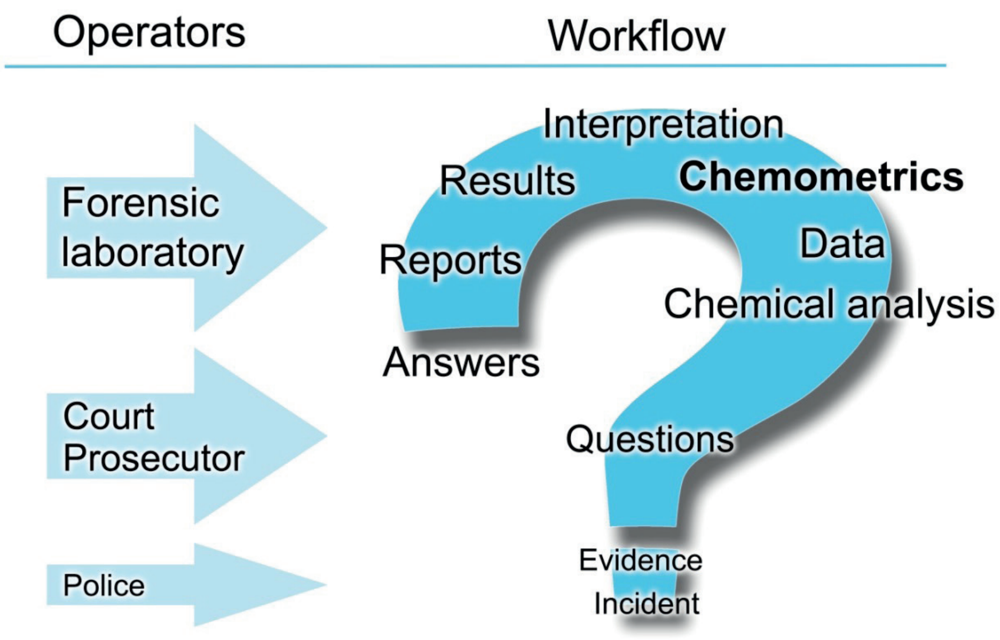
The forensic workflow in routine casework (Figure 4.1) usually starts at the police investigation site and ends in the courtroom. Physical evidence from the site collected by forensic investigators or police officers is analysed in a forensic laboratory according to the request of the police, prosecutor or the court of law. Traditionally, forensic samples collected or seized are subjected to physical and chemical analyses. The results of these analyses are typically used for identification and quantification of effective substances in order to support the judicial process [6, 9].

Additional information to the forensic process (e.g. for illicit drug profiling) may be provided when the data is further analysed using statistical methods [10-13]. This application of multivariate analysis and other statistical methods, also called chemometrics, includes processing the data from the forensic chemical analysis in different ways, e.g. through data selection, data pre-processing or calculation of similarity scores between samples.

Chemometrics can provide *additional* information in complex crime cases and enhance productivity by improving the processes of data handling and interpretation in various applications. Beside this use in routine chemical casework, chemometrics can be used to process large sets of case data for police tactical or intelligence tasks, as well as crime analysis and prevention purposes by enhancing the usability of database information [14-17].

The results and conclusions of forensic analyses need to be communicated in a comprehensible form and explained with sufficient clarity to investigative units and to the court of law in order to be used effectively. Ultimately, the forensic analysis must answer the original request presented by the investigative unit.

Figure 4.1 Illustration of a case workflow from incident to expert report, different operators involved and where in the process chemometrics would be applied



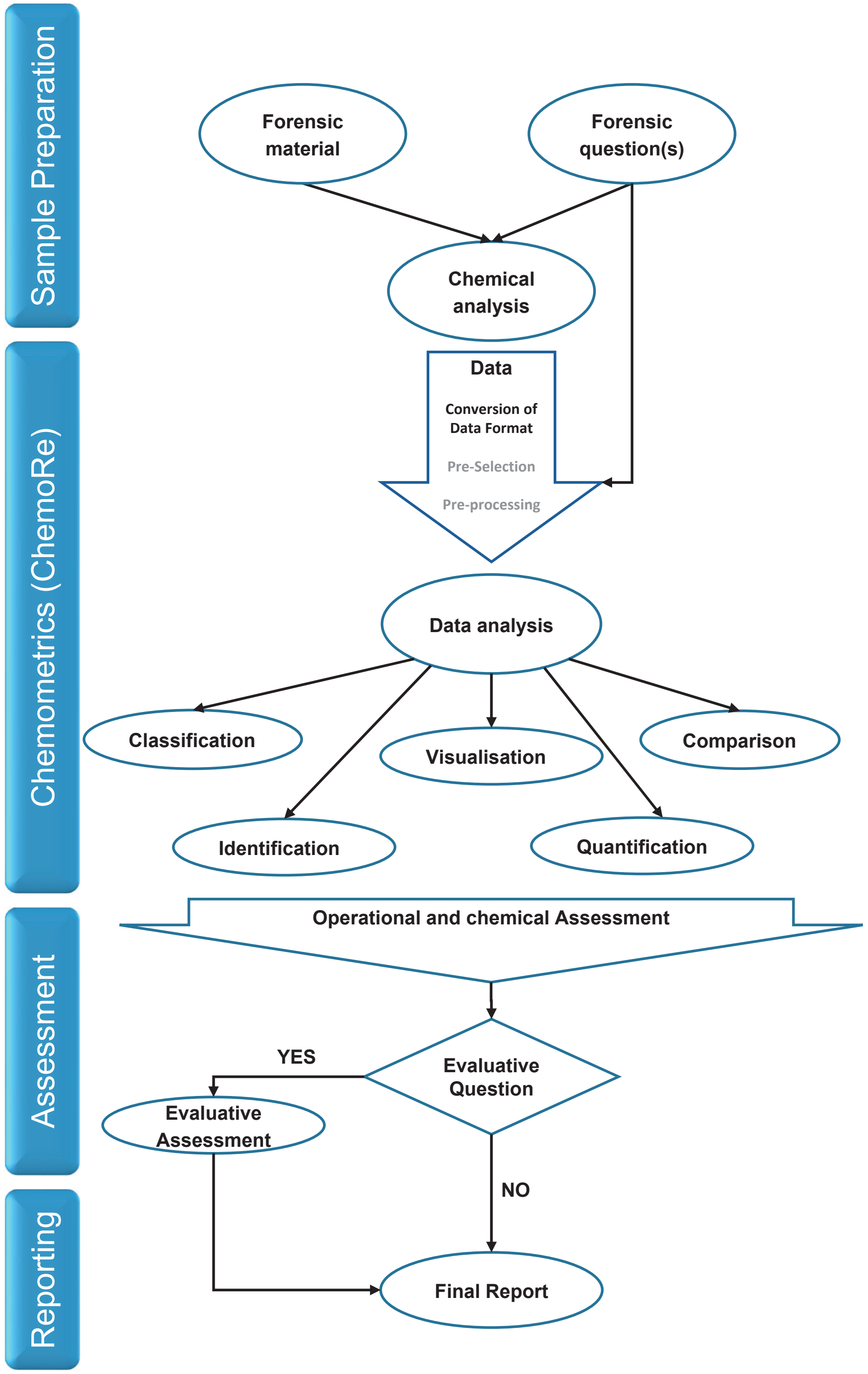


Figure 4.2 Steps before, during and after applying chemometrics (ChemoRe)

According to Matthias Otto's textbook Chemometrics [18] can be defined as:

"the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analysing chemical data"

As presented in article 1 [1], a literature survey was performed to summarize currently used chemometric methods in the field of forensic chemistry. Based on that survey a selection of common practices has been collected to this guideline and the software tool ChemoRe, aimed at helping forensic chemists to utilize chemometrics in their everyday work tasks.

Analytical chemists are using chemometrics in order to extract information of multivariate chemical data. Modern analytical instruments or a combination of instruments provide a tremendous amount of data. Usually a high number of descriptive variables but a comparable low number of samples are available and need to be correlated answering complex questions like [19]:

- Visualisation of multivariate data sets,
- Relationships between data sets,
- Recognition of internal structures,
- Classification or identification
- Comparison
- (Quantification)

Figure 4.3 summarizes the specific types of analytical methods used to produce the analytical data. And Figure 4.4 illustrates which chemometric methods were used to treat Type 1 and Type 2 data.

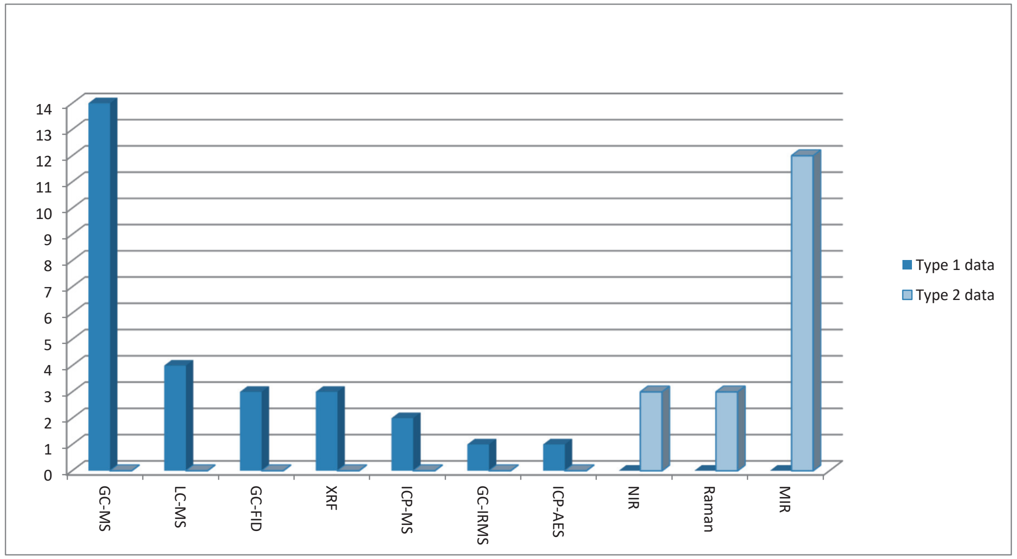


Figure 4.3 Analytical methods used to produce Type 1 and Type 2 data, as found in the literature survey (41 articles from 2011 to 2018) [1

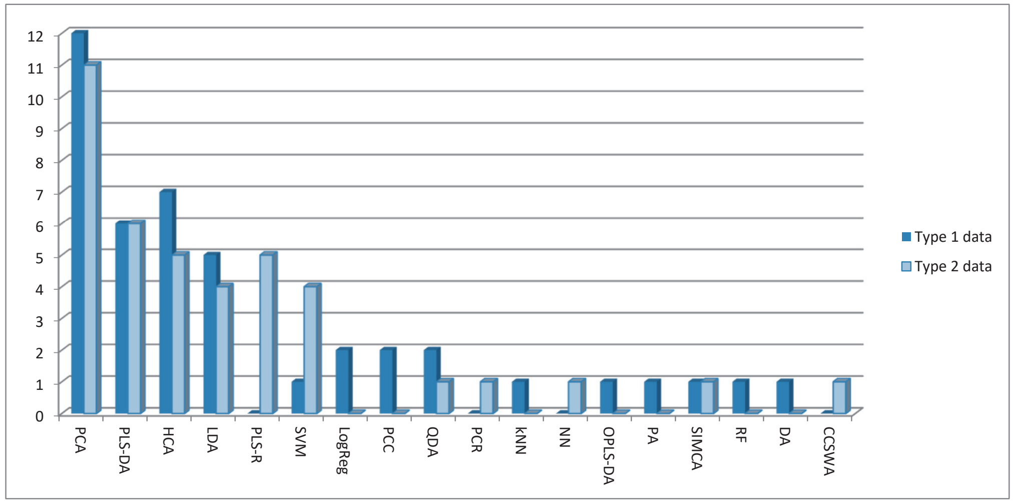


Figure 4.4 Chemometric methods used to produce Type 1 and Type 2 data, as found in the literature survey (41 articles from 2011 to 2018) [1].

It can be noted that the analytical method used depends strongly on whether the data is of Type 1 or Type 2. It is natural that spectroscopic methods produce Type 2 data. On the other hand, GC-MS and LC-MS methods prevail with Type 1 data.

As for chemometric methods, it can be seen from Figure 4.4 that PCA, PLS-DA, HCA and LDA are all quite frequently used, for both types of data. There is no clear dependence between the type of data and chemometric method. Additionally Figure 4.4 shows that there is a wide variety of chemometric methods applied to both Type 1 and 2 data.

4.2 Chemometrics for intelligence purposes and potential evidence in court

Forensic science (like chemical measurement) is considered as a mean to generate forensic evidence in a probative process. The utilisation of the chemical or physical trace (e.g. analytical results from physical remnant of an activity) has for long focused only on its primary role as evidence [1-3, 16-17, 20]. However, in the past years numerous authors have emphasized the contribution of forensic science and the data provided by the latter to crime analysis and to investigation [17]. In the field of illicit drugs, the importance and benefit of forensic data (i.e. drug profiling) as a law enforcement tool has been widely recognized and discussed in literature. The analytical methods do not only allow to identify and quantify illicit compounds of questioned material to support the judicial process but by further processing (e.g. with the help of chemometric methods) provide additional information for intelligence purposes [5]. Until today, only few countries have put their efforts to develop, incorporate and implement the routinely use of illicit drug profiling during the investigation process. However, many authors have highlighted (the importance and benefits of illicit drug profiling as an important law enforcement tool principally during the investigation for tactical and strategic purposes at national and international level [21-23].

When illicit drug profiling is used as piece of evidence in court, usually investigative units or prosecutors have asked for case to case comparisons [5, 21-23]. The sought information in this case is to know whether seized substances share a common origin in order to prove, for instance, that cases under investigation are related or drug trafficking. In that case, the comparison results are added as piece of evidence to be presented in court. The chemical or physical profiling may relate the substances or the materials to each other but not the individuals that are behind these seizures. Therefore, usually more information is needed to assess the accused’s conviction in court.

4.3 Description of the basic questions in Forensic Chemistry

Of course, the range of questions Forensic Chemistry is requested is much broader than given in this subchapter. Beside the need to report which compounds are present in a sample (factual drug identification) or which amount of an identified compound is present (quantitative analysis), more complex questions may concern attributing a sample to a class (classification or grouping) or elucidating the similarity of one sample to another sample (comparison). The terms classification and comparison are described in more details hereinafter.

Classification has several aspects to consider and to differentiate for a good handling and communication.

Characteristics of data or in a more practical way – our samples to analyse – can be such distinct that an overlapping is not possible. Metaphorically speaking the sample can only belong to a predefined ‘box’ and there is no overlapping of this box to any other. Analytical results appoint to just one of these boxes. In such cases – also known as factual analysis – there is no reasonable uncertainty to consider. The assessment of the analytical result(s) may only respect the validity of the applied method (quality assurance aspects) to the kind of question asked. If the predefined ‘boxes’ are fully separated, also chemometric methods are capable to appoint a sample to just *one* of these boxes. Furthermore, if the characteristics of a class are discriminative enough, a class can be reduced to 1 representative, e.g. a substance like cocaine. In such a case the classification is considered as identification.

Examples are:

- Pharmaceutical tablets, original or fake
- Synthetic routes if specific substances are present

Figure 4.5 Illustration for a classification by pre-defined, non-overlapping classes

However, in practice, boxes can overlap to a certain extent, i.e. the characteristics of a sample can point only to a certain extent to one box, but to a lesser extent also to an adjacent one. If certain classes (boxes) are pre-defined, the chemometric application can still support a classification.

Examples are:

- Distinction of drug-hemp and fibre-hemp
- Distinction of South-West Asia or South-East-Asia heroin

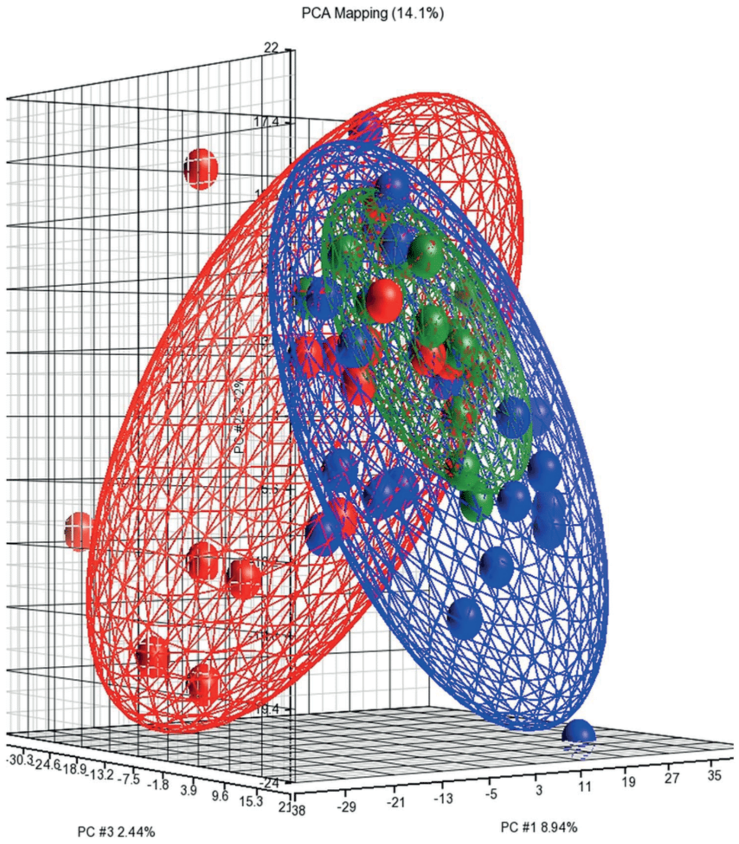


Figure 4.6 Classification by pre-defined, overlapping classes [24]

When classification' is wished out of a dataset, a 'but there are no or only vague impressions of the kind and number of 'classes', a grouping can be calculated by means of chemometric applications. Then the term of *grouping* should be used instead of (pre-defined) classification as the groups are 'post-created' by the chemometric application. By changing (increasing) the dataset, the number and size of groups and the criteria to appoint a group can change, while by classification, the criteria of a class are fixed.

Examples are:

- Drugs profiling by the assumption that different batches can be distinguished. Each batch represents a group and it cannot be pre-determined how many batches are in a database. Batches are added or deleted (because they are supposed to be consumed) from the database.

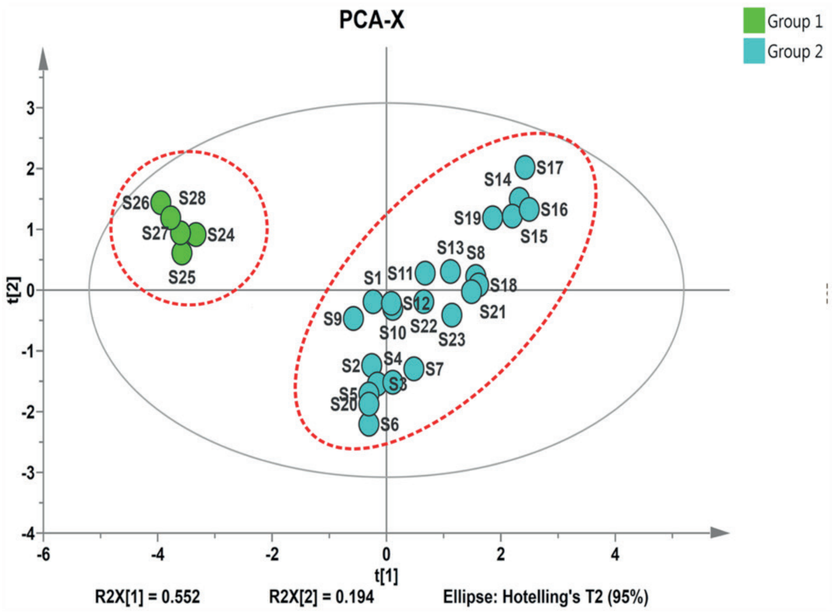


Figure 4.7 Grouping as a result of a chemometric application [25]

In forensic drug chemistry a focus on the elucidation of strategic information out of chemical investigations on seized illegal drugs can lead to such a grouping, i.e. ‘drug classification’. It needs a systematic chemical investigation over a longer period of time applying the same analytical scheme on a certain amount of individual drug samples. Prominent information is the determination of origin, common methods of clandestine drug manufacturing, identifying key precursor chemicals and dismantling distribution networks [4, 14, 26-27]. Also, estimations of how long a laboratory has been operating and what is the scale and output volumes of a certain drug production site [28]. The aim is to support drug intelligence programs of law enforcement with either general information or deeper intelligence information.

Comparison supports the investigative side of law enforcement on a case-file based level. Statements of similarity provide links between seized drug samples, so comparison constitutes an evidential part of forensic expertise. It may provide information on the relation of drug dealers (i.e. seized material attributed to them) and users or the relation between different drug dealers for prosecution purposes.

The chemical profiling of heroin can be taken as an example. Two chromatograms of pairwise (peak-wise) comparisons on the basis of impurities formed during manufacturing are presented in Figures 4.8 and 4.9.

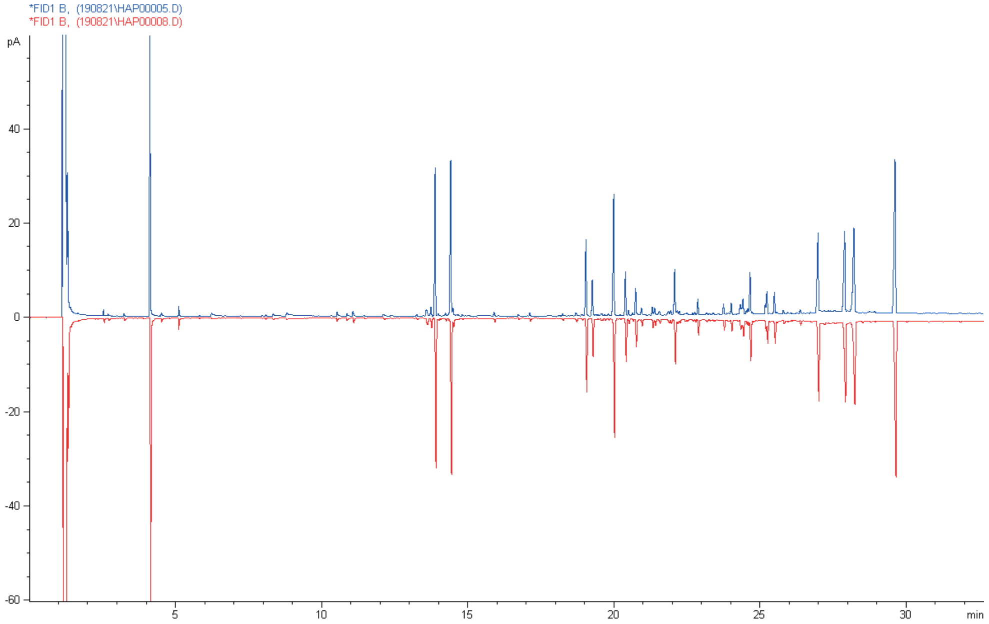


Figure 4.8 Very high similarities of two heroin samples by comparison (impurity profiling)

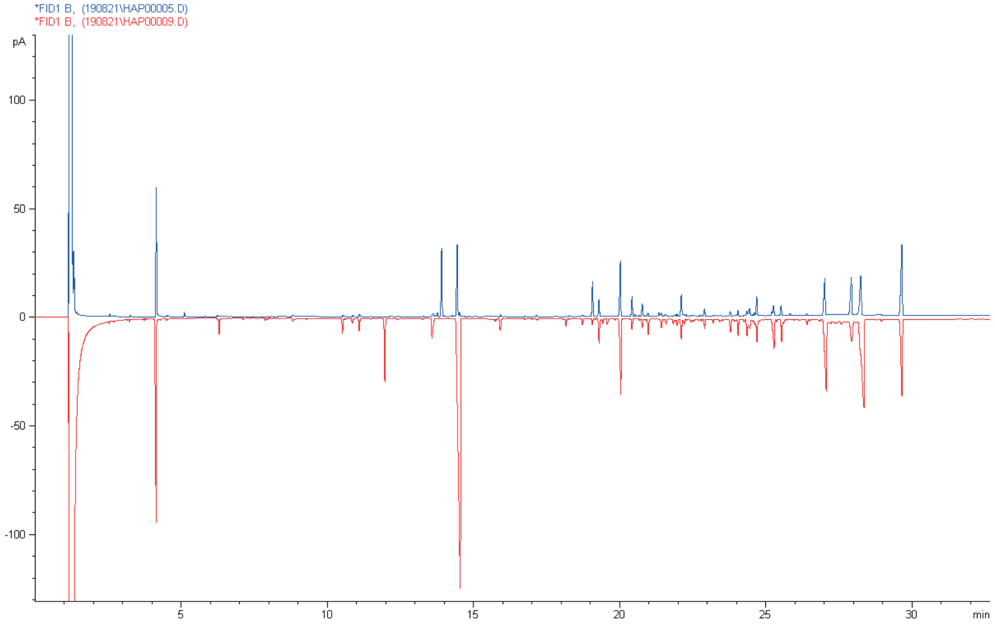


Figure 4.9 Quite a large difference of two heroin samples by comparison (impurity profiling)

Using the chromatograms, the forensic chemist could only describe verbally the visual similarity of the two samples and making maybe some inferences by knowing which peak represents which substance (impurity).

The chemist could make some conclusions like if these peaks come from the origin of a plant or synthetic route or hydrolysis/decomposition or being an adulterant or diluent. However, a number or percentage of similarity cannot be expressed by just a visual comparison of the chromatogram.

Here the strength of chemometrics comes into play giving a number or degree of similarity (expressed like as a distance of two samples in an n-dimensional space).

After validation of the applied chemometric method a threshold number / distance (see also the given examples in Figures 11.4, 11.6 or 11.11) gives the degree of similarity and will enable a conclusion.

The chapters 7 to 11 of this guideline will deepen the description and understanding of that process.

5 DATA HANDLING

5.1 Verification of the quality of the data

In forensic casework of illicit drug analysis, the question to be answered first is: ‘Is an illicit drug present?’ In forensic laboratories, the requirements for the detection and identification or quantification of drugs are predefined, tested and validated for the methods in use. However, each technique has its advantages and limitations. The quality control measures in use (e.g. internal standard, calibration samples, QC samples and blank samples) has to be acceptable and the correctness of the data (e.g. retentions time, correct identification and integration of the compound) needs to be checked before starting the data processing (by chemometric methods).

ENFSI DWG Best practise manual (BPM) for controlled drug analysis (DWG-CDA-001) gives recommendations covering analytical methods, procedures, quality principles, training processes and approaches to the forensic analysis of illicit drugs[29].

5.2 Exploration of data

Sometimes a signal from an analytical system might have a very low peak height or area, and is therefore below the integration threshold. This is due to a low concentration, small sample size (weight), dilution during sample pretreatment, injection volume, split factor etc. The output value of such a small (not integrated) signal will be zero, although visible as a signal. However, if a small signal is considered as relevant variable to be included in the chemometric application, an artificial correction is necessary to the numerical output. The zero value is not valid for further calculations and therefore a very small value above zero needs to be included as a variable in the chemometric application.

5.3 Selection of the ‘correct‘ parameters (variables)

Type 1 data typically consists of a series of signals with peak height or an integrable peak area determining the variable. Each peak usually stands for a single characteristic information (i.e. a molecule, an ion, an element, etc.). The relevance of each signal for answering the question asked is to be evaluated at the earliest possible stage. This evaluation needs expert knowledge regarding the overall composition of the item like the contribution of impurities from the biosynthesis or lab synthesis, including processing and purification steps. Irrelevant signals might be included in the data imported to chemometric method. This irrelevant data may or may not have influence to the result regarding similarity of two items. Irrelevant variables could be ‘filtered’ out by regression methods like PCA (see chapter 7.4). However, it is logic and preferable that only relevant signals are chosen as variables for chemometrics.

For example, in illegal cocaine seizures, truxillines and cis- and trans-cinnamoylcocaine are part of the biosynthesis remaining in the final product as impurity. Norcocaine is built through the chemical purification step and ‘man made’. Both contribute to a very specific total composition assumed to change in each production batch which is used to determine the similarity or dissimilarity of seizures regarding the source attribution to a certain batch [30].

Another fact to consider the significance of each variable and its reasonable selection for the chemometric application is its correlation to another variable. Illustrated by cocaine seizures again, benzoylecgonine is the first hydrolysis product of cocaine and found in every sample. The more benzoylecgonine is built, the less cocaine is in the sample while the other impurities are not affected by this hydrolysis and do not change the profile. Benzoylecgonine and cocaine are therefore correlated variables and its use as variable should be carefully considered and tested through the method validation (see chapter 7). This does not mean that correlated compounds shall be excluded. One possible way is to take the sum of both variables (its areas) as a new variable.

Type 2 data, typically spectral data, are commonly used as such, i.e. the whole spectral range. However, it might be sufficient, that for answering the question asked only one or more specific spectral bands or ranges cover the information relevant for comparison or classification. Type 2 data usually comprises hundreds of variables, each representing a measurement point in a continuous domain such as time or wavelength. The variables are thus obtained by sampling from a continuous function (curve), but serve to represent the entire function. These variables are therefore often highly correlated, and as opposed to type 1 data, do not represent any clearly defined features of the measured material by themselves. Instead, the information is often contained in the correlations between these variables.

Spectroscopic techniques such as FT-IR, NIR and Raman encompass information on the composition of the sample material, usually a mixture of components. Spectroscopic techniques don´t isolate the response of individual compounds, unlike chromatographic (Type 1) techniques. This integral information on the material can also be used for comparison or classification in chemometric applications.

5.4 Importing analytical data to ChemoRe

The data provided by the analytical methods can roughly be categorized into two types based on their properties:

Type 1 data consists of measurements with relatively few dimensions or variables, usually much less than a hundred. This includes, for example, data from calculating the areas of peaks of interest appearing at specific retention times in chromatographic signals. Such data often represents amounts of distinct compounds or other clearly separate characteristics of an item that describe its properties.

Type 2 data consists of measurements over a continuous range. This includes spectroscopic data where the spectrum associated with the material is measured over, e.g., wave lengths or raw chromatographic signals where measurements are made over retention time. Type 2 data usually comprises hundreds of variables, each representing a measurement point in a continuous domain such as time or wavelength. The variables are thus obtained by sampling from a continuous function (curve), but serve to represent the entire function. These variables are therefore often highly correlated, and as opposed to Type 1 data, do not represent any clearly defined features of the measured material by themselves. Instead, the information is often contained in the correlations between these variables.

The data must be provided in a specific format in order to be imported to the ChemoRe software. ChemoRe currently supports CSV (*.csv) and Excel (*.xls, *.xlsx) file formats for data input. These files must further be arranged so that the columns correspond to the variables (e.g. chemical compounds, classes or IDs) and the rows correspond to the samples.

In the data, the names of the variables should always be at the first row. When it comes to CSV-files, the user should be mindful of the delimiter and decimal separator used in their respective locale. In English speaking countries the default is typically to use a comma as a delimiter and a point as decimal separator but in some countries the comma is used as the decimal separator while a semicolon is used as the delimiter.

ChemoRe allows selecting these according to the settings used to write the data file. When using Excel files, this is not an issue. Table 5.1 exemplifies the formatting of the data using the 'iris' dataset provided with many statistical software. Here, Item #, Sepal.Length, Petal.Length, Petal.Width and Species are variables and each row under the first one corresponds to a single measured sample.

Table 5.1: Illustration of formatting data for ChemoRe

Item #	Sepal. Length	Sepal. Width	Petal. Length	Petal.Width	Species
1	5,1	3,5	1,4	0,2	setosa
2	4,9	3	1,4	0,2	setosa
3	4,7	3,2	1,3	0,2	setosa
4	4,6	3,1	1,5	0,2	setosa
51	7	3,2	4,7	1,4	versicolor
52	6,4	3,2	4,5	1,5	versicolor
53	6,9	3,1	4,9	1,5	versicolor
54	5,5	2,3	4	1,3	versicolor
55	6,5	2,8	4,6	1,5	versicolor
103	7,1	3	5,9	2,1	virginica
104	6,3	2,9	5,6	1,8	virginica
105	6,5	3	5,8	2,2	virginica
106	7,6	3	6,6	2,1	virginica
107	4,9	2,5	4,5	1,7	virginica

For example, Agilent instruments by ChemStation software allow data export to CSV –file via Custom Report function. Obtained CSV- file can be converted to Excel format and organized as described before importing the data to ChemoRe software.

In the case of Type 2 data, additional steps might be needed. If the data consists of spectra, each wavelength should be considered a variable. This means each spectrum included in the data should have a measured intensity at the same wavelength number.

6 DATA PRE-PROCESSING

As explained in Chapter 5, the *data acquisition* from an analyzed sample is done via the analytical instrument and may be steered by the analyzing chemist. The data processing like integration of peak areas is fully defined by the analytical method setup, and by choosing the data reporting method the analyzing chemist moreover can decide on a certain selection of information to be exported.

These steps can be automatized using macros. Templates (e.g. validated macros) for the most used analytical instruments (GC-MS and FT-IR) are provided by the ENFSI DWG upon request.

As a third step before data pre-processing, the forensic chemist may use chemical knowledge in the manual selection of data with respect to the question asked (see also Chapter 5.2). Diluents or decomposition products might e.g. be of non-interest or are very strongly correlated to other variables and do not need to be considered in the final dataset. It may be beneficial to eliminate variables contributing less information from the dataset immediately, as also explained in Chapter 5.2. On the other hand, computer performance nowadays allows fast calculations with large number of potentially correlated or not obviously discriminative variables, so potential discriminative information of variables at the stage of the untreated .csv file should not be thrown overboard without a good reason.

The starting point for the *ChemoRe* software is then the dataset obtained through the steps described above, in *.csv* file format. The analytical data is categorized into Type 1 or Type 2 data, based on whether they are low- or high-dimensional in the number of variables (see Chapter 5.3), *ChemoRe* is able to handle both types.

Now the first step of the chemometric method development is to consider the need of *data pre-processing*. Data pre-processing can consist of e.g. data normalization, data transformation or dimension reduction. For this a variety of methods exist, such as standard normal variate transformation, square or fourth root or logarithmic transformations or principal component analysis. Visualization of the data may lead the way as to what data pre-processing is to be used and is discussed in Chapter 6.1. The purpose and advantages and disadvantages of the various types of data pre-processing are explained in more detail in Chapter 6.2.

6.1 Visualization of the data

Different visualization methods like box-plots and histograms can be helpful in the selection of data pre-processing methods. These methods indicate the distribution and range of each variable. It is important to note that box-plots are not necessarily a suitable visualization tool to select the most powerful discriminative variables of the dataset. This is because a variable with low inter-variability could still be a well discriminative variable. Box-plots might help however in the setup of the method, to visualize the effect of data pre-processing to get weighed contributions of each selected variable and as such is integrated in *ChemoRe*. An Excel-based macro file for data visualization is provided by the ENFSI Drugs Working Group upon request.

6.2 Methods commonly used for data pre-processing

There are several methods that can be applied for data pre-processing. These can be used alone or in combination. The overall purpose of data pre-processing is to meet requirements on the data for the specific chemometric method used, more specifically the underlying statistical inference procedures. If one uses the raw analytical data in the chemometric method, the results of the methods will often not be reliable.

As an example, assume we have 20 samples of seized amphetamine powder and the analytical data are chromatograms of these samples. Assume further that these samples are considered to represent some source of interest (e.g. production of amphetamine powder at a particular illegal laboratory). There may be striking differences between the chromatograms of a number of samples, and these differences may be assessed and understood from a chemical point of view. However, the differences may lead to a chemometric output that is not in concordance with such a chemical understanding. The reason for this is that the chemometric method may be such that the data it uses should show what is referred to as “normal variation”, i.e. varying symmetrically around a centre point. If this is not the case, then the striking differences may be interpreted by the chemometric method as anomalies which do not fit with the chemist’s view of the situation.

There is of course statistical theory behind recommendations of data pre-processing, but in general we can say that there should not be too much imbalance in the data. Data pre-processing aims at scaling or transforming the original analytical data, without losing much relevant information, so that the pre-processed data behave according to the statistical models that are assumed in the chemometric methods. Most of the conclusions that are expected to be drawn from the chemometric analysis can also be drawn using transformed data. The most important exception is when the chemometric method is designed to predict the actual value of a sample, for instance quantifying the purity. The predictions must then be back-transformed to the original scale. This procedure is usually already implemented in the software used, and the chemist should normally not need to bother about that.

The most commonly used data pre-processing procedures include normalization, arithmetic transformations and principal component analysis (PCA). They will be described below.

6.2.1 Normalization

Normalization is used to adjust values measured on different scales to a notionally common scale. E.g. each target variable in a sample (like a chromatographic peak area of a target compound) is divided by a certain number, such as the concentration of an effective substance in that sample, or with the sum of all target peaks. The result of this will be that transformed peak areas are closer to each other in absolute value, whereas relative differences between them are kept. After transformation the values have less chemical meaning, although ratios of peak areas may have an element of chemical interpretation. The advantage of the transformation is that the chemometric method will not base conclusions on the fact that absolute areas of some peaks are dominating much smaller ones. In this way the smaller peaks are interpreted with equal importance as the peaks with large areas, and not only as noise.

As an illustrating example we consider peak areas of the substance N-Benzylpyrimidine appearing as an impurity compound in amphetamine samples. For 744 casework samples of amphetamine powder, in Figure 6.1 to the left a histogram of the original peak areas of N-Benzylpyrimidine are shown, and to the right a histogram of the peak areas divided (normalized) by the internal standard used in the chromatographic analysis of the sample. Comparing the two graphs it can be seen that in the left graph the *x*-axis ranges from 0 to 70,000,000 which means there are enormous differences in peak areas of N-Benzylpyrimidine between the 744 samples, while in the right graph the range is from 0 to 35. The shapes of the two histograms are very similar though, which means that the relative sizes of the peak areas are to a large extent kept, while they are much closer in absolute value.

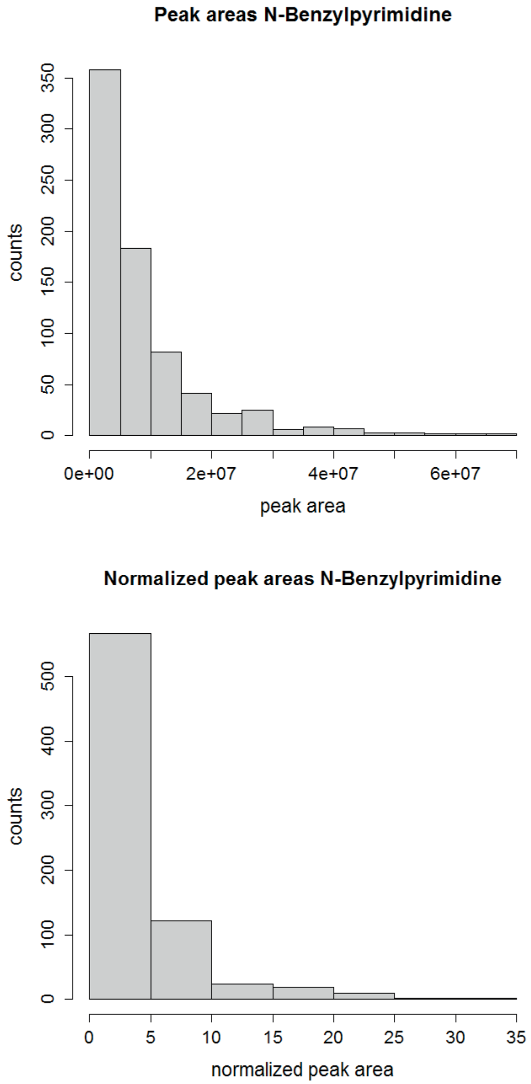


Figure 6.1 *Histograms of original peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (to the left) and normalized peak areas by dividing the peak area with the peak area of the internal standard used (to the right)*

6.2.2 Standardization

By *standardization* it is usually meant that normalization of individual variables takes place based on their standard deviations, i.e. replacing each value y with y/s, where s is the calculated standard deviation based on all values. The formula for calculating *s* is

$$s = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

where *n* is the number of data points (y_1, y_2, … , y_n) and

$$\bar{y} = \frac{1}{n}\cdot\sum_{i-1}^{n} y_i$$

that is the mean (or average) of all values.

For the 744 samples of amphetamine powder, the standard deviation of the peak areas of N-Benzylpyrimidine is 10,489,823. Dividing all 744 peak areas by 10,489,823 produces standardized peak areas. In Figure 6.2 a histogram is shown of these values. As we can see (and expected) the standardized data behave like normalized data, i.e. the range of values is small so that the values from different peak areas are closer in absolute value, while their relative sizes are (to a large extent) kept.

Whether data should be normalized or standardized is a question related to the chemometric method applied. Some methods will require standardized values because the underlying statistical model is designed for that. That is due to the fact that the standard deviation of standardized values is always equal to one, which might be a requirement. For most chemometric methods it will not matter whether normalization or standardization is used as pre-processing. They both perform scaling of the original data, and the choice to be made is that of a suitable scaling constant.

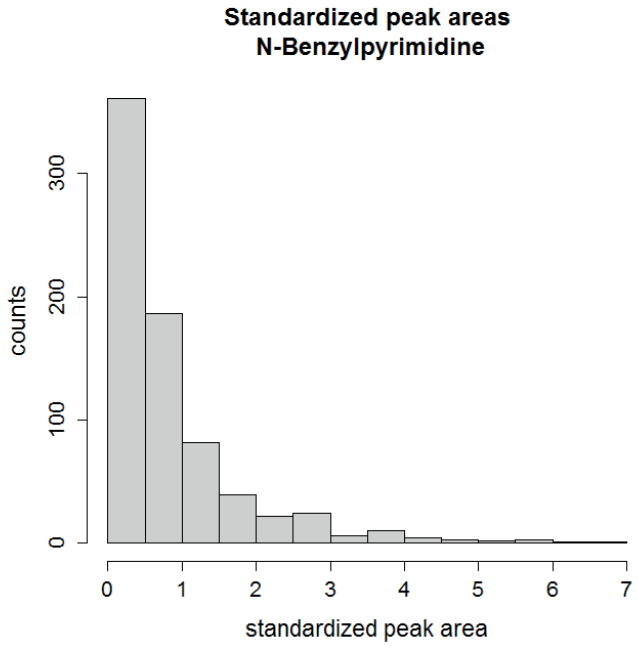


Figure 6.2 Histograms of standardized peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1). The original peak areas have been divided by their standard deviation 10,489,823

6.2.3 Z-scores

A third normalizing procedure is by means of so-called *z*-scores. Z-scores are calculated by replacing each value y by the value

$$z = \frac{y-\bar{y}}{s}$$

whereas before $s = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}$ and $\bar{y} = \frac{1}{n}\cdot\sum_{i-1}^{n} y_i$,

that is the z-score is the difference between the current value and the mean of all values divided by the standard deviation

After performing this transformation the resulting values (z_1, z_2, … , z_n) will have a mean value (average) of zero and a standard deviation of 1. In Figure 6.3 a histogram is shown of the z-scores of the 744 peak areas for N-Benzylpyrimidine used to produce Figures 6.1 and 6.2. Here the mean is 8,568,532 and the standard deviation (as before) 10,489,823. Hence the z-scores are calculated by replacing each peak area y with z=(y – 8,568,532) / 10,489,823.

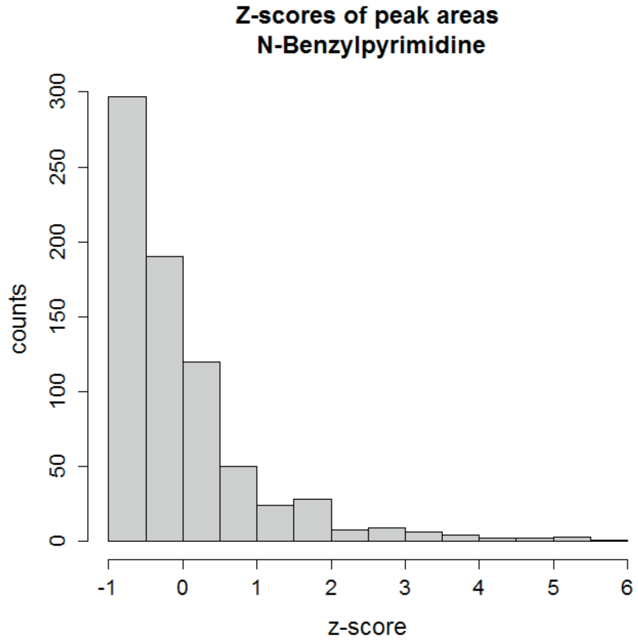


Figure 6.3 Histograms of z-scores of the peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1). The mean 8,568,532 of the peak areas have been subtracted from each peak area and these differences have been divided by the standard deviation 10,489,823 of the peak areas

The shape of the histogram in Figure 6.3 is (as expected) the same as of the histogram in Figure 6.2, while the location of the histogram in Figure 6.3 is translated to the left.

The purpose of computing z-scores is not so much to produce values with zero mean and standard deviation 1, but to make the transformed values closer to a so-called standard normal distribution. A normal distribution is always symmetric around its mean and a standard normal distribution has zero mean and standard

deviation 1. The letter ‘z’ stems from that in the statistical literature a random variable with a standard normal distribution is often denoted as ‘*Z*’. Several of commonly used chemometric methods, e.g. linear discriminant analysis, have as a requirement that the variables involved follow normal distributions. If they deviate significantly from normal distributions, the underlying statistical model will provide erroneous results and hence the output from such an analysis will not be very reliable.

6.2.4 Data transformation

Data transformation is the application of a deterministic mathematical function to each point in a data set so that each data point y_i is replaced with the transformed value $w_i = f(y_i)$, where f is the function applied. Transforms are usually applied so that the data appear to better meet the assumptions of the underlying statistical inference procedure of the chemometric method that is to be applied, or to improve the interpretability or appearance of graphs. It is important though to remember that no chemical understanding of transformed data should be expected.

There are a number of standard transformations used in chemometrics of which we will present three.

Square root transformation is as it reads to replace each data point y with its square root $y^{0.5}$. The effect is to some extent the same as with normalization, i.e. that the transformed values are closer to each other in absolute value. However, a more important effect is that data points that are heavily deviating in absolute value from the majority of the data points will after taking square roots be much closer. This transformation thus affects large values more than it affects small values, and in turn the entire transformed data set looks more symmetric than the original data set, i.e. has more the shape of values randomly spread around a centre point. This also implies that the transformed data are closer to being normally distributed, which (as said before) is an important requirement for some chemometric methods (like linear discriminant analysis). Feeding such a method with asymmetrical data will render an output that is not reliable.

In Figure 6.4 a histogram is shown of square root transformed peak areas of N-Benzyl-pyrimidine in the 744 samples of seized amphetamine powder illustrated in Figure 6.1. Comparing this graph with the left graph of Figure 6.1 (i.e. the histogram of original peak areas) we can see that the distribution (variation) is less skewed with the square root transformed peak areas. However, we cannot say that it has turned to be symmetric.

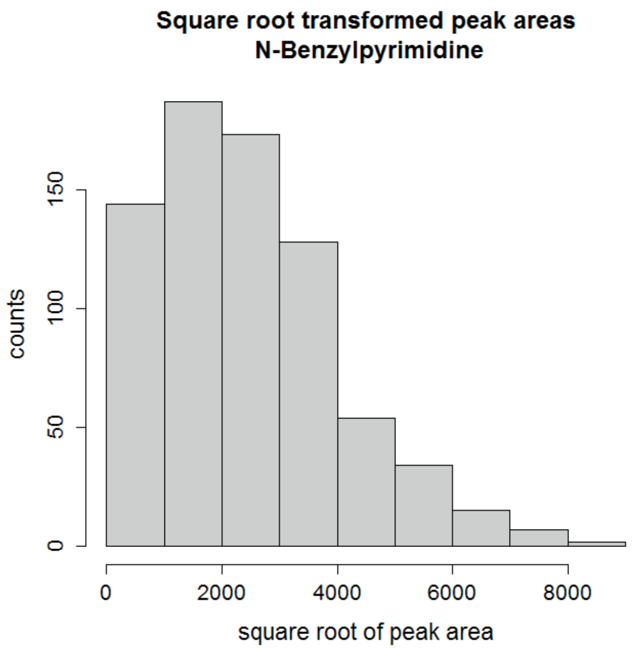


Figure 6.4 *Histograms of square root transformed peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1)*

Square root transformation is a first choice when one wants to reduce the excess influence of large values and make the data more symmetric. However, some data sets show such strong deviations from symmetry that the square root transformation is not sufficient. The next choice is then often to use the fourth root transformation, i.e. replacing each data point y with its fourth root $y^{0.25}$. This transformation affects large values even more than the square root, still not affecting small values that much. In Figure 6.5 a histogram is shown of fourth root transformed peak areas of N-Benzylpyrimidine in the 744 samples of seized amphetamine powder illustrated in Figure 6.1. The shape of this histogram is now almost symmetric and definitely less skewed than the histogram in Figure 6.4. This means that the fourth root transform is better than the square

root transform if the aim is to make data (more) symmetric. We might even conclude that the fourth root transformation have excessed the aim in that the shape is almost left-skewed (compared to the original right-skewed shape).

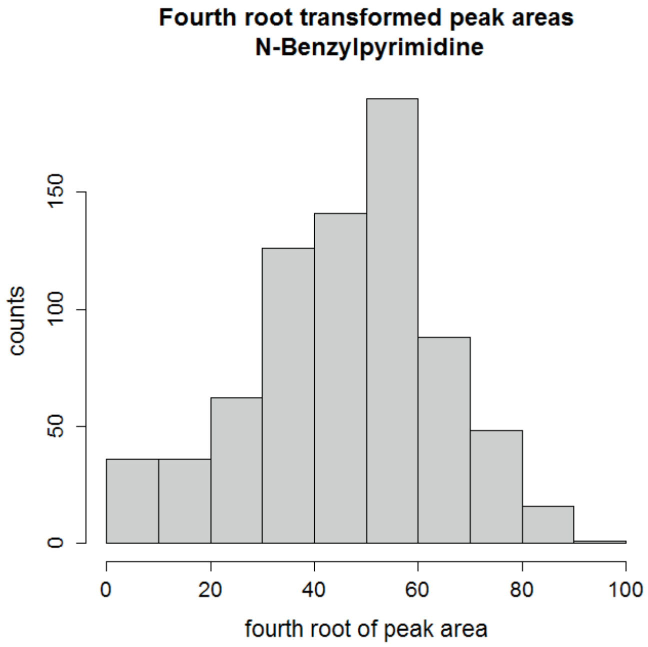


Figure 6.5 *Histograms of fourth root transformed peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1)*

Common for square root and fourth root transformations is that they cannot be used with negative values. Such values are of course not natural to obtain as analytical output, but if the analytical data has been pre-processed with standardization or z-scores (see above) it is not possible to compute square roots or fourth roots (or any roots).

Should the data set show even stronger deviations from symmetry that cannot be alleviated with fourth root transformation, the next choice is to use the logarithm (or simply log) transformation. That is, each data point y is replaced by its natural logarithm $\ln(y) = \log_e(y)$. It is also the logarithm that is implemented as the default logarithm in many statistical software. In Figure 6.6 a histogram is shown of logarithm transformed peak areas of N-Benzylpyrimidine in the 744 samples of seized amphetamine powder illustrated in Figure 6.1. The shape of this histogram is now clearly left-skewed, which means that the logarithm transformation is too strong for these data. The aim was to make data more symmetric and this way we have rather made them asymmetric in the other direction.

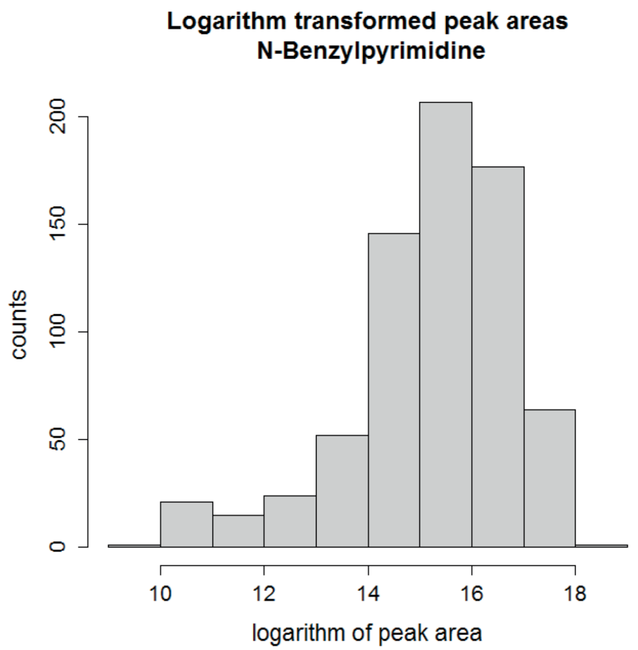


Figure 6.6 Histograms of logarithm transformed peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1)

Logarithms cannot be used with zeros (nor with negative values), and while negative values are not expected in analytical outputs zeros may very well be. This problem can be circumvented by replacing the zeroes with a small value, small in that sense that it would not be of the same magnitude as generally small values in the analytical output. For instance, in a chromatogram there may be very small peaks, sometimes even hard to discern, but they have chemical significance. One way to find a suitable “small value” for replacement of zeros would be to use the lowest peak area divided by 100 as a general replacement value. Another way could be to use the limit of detection.

It is always of great importance to assess the presence of zeros in analytical output. Some zeros may be structural, and if so, these variables should sometimes be omitted from chemometric analysis. For instance, a sample material in which a particular substance is completely missing (no peak in the chromatogram) cannot easily be chemometrically compared with a sample material in which this substance is present. There must be a manual comparison of the analytical outputs before any chemometric method is considered. Making a distance calculation between samples like this may give very misleading results. However, sometimes the zeros are there simply because the concentration of the corresponding substances is so small that they couldn’t be detected in the chromatographic analysis. In that case the replacement of these zeros with a small value as described above makes sense in the subsequent chemometric analysis.

If we would like to make the original peak areas of N-Benzylpyrimidine in the set of 744 samples of amphetamine powder looking more normally distributed, we should first transform the peak areas so that the transformed values are more symmetric. As could be seen above, this can be achieved by applying the fourth root transform. Thus, calculating z-scores on fourth root transformed data would possibly give us what we are looking for. In Figure 6.7 is shown a histogram of z-scores of fourth root transformed peak areas of N-Benzylpyrimidine in the 744 samples of seized amphetamine powder. Note that for calculating the z-scores we have to calculate the mean and the standard deviations of the fourth root transformed peak areas, but we do not explicitly show these numbers here since they have no chemical meaning.

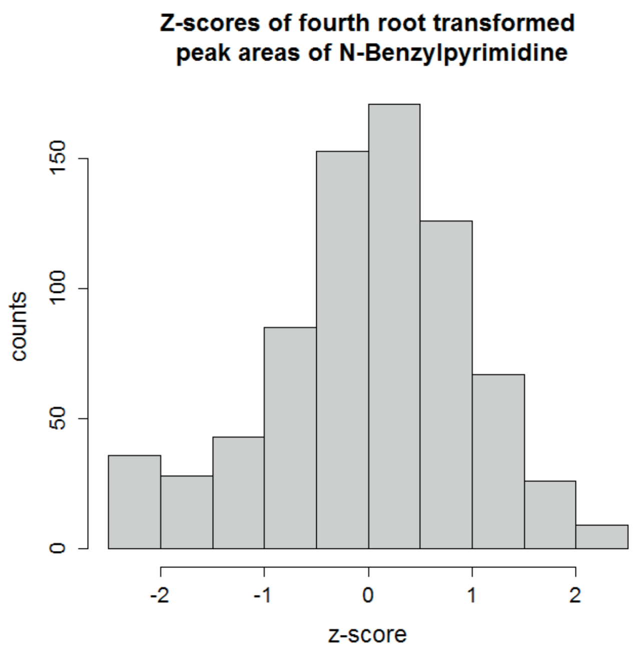


Figure 6.7 Histograms of z-scores of fourth root transformed peak areas of N-Benzylpyrimidine in 744 samples of seized amphetamine powder (same original data as those used for producing Figure 6.1)

The shape of the histogram in Figure 6.7 is close to symmetric around the value 0. We can also see that the range of values is between roughly –2.5 and 2.5. This corresponds quite well with a standard normal distribution, since for such a distribution about 95% of the values are between –2 and 2 and about 99% of the values are between –2.6 and 2.6. Hence, z-scores of fourth root transformed peak areas of N-Benzylpyrimidine would fit well with the use of e.g. linear discriminant analysis.

6.2.5 Principal component analysis

Principal component analysis or PCA is another common way of performing data pre-processing. PCA is a multivariate method that is used to simplify interpretation of large data matrices. The procedure involves replacement of original variables with a reduced number of ‘principal components’, which explain the variation in the outcomes of the raw data as efficiently as possible. Since PCA is used both as a tool for data pre-processing as for data analysis it will be discussed in depth in Chapter 7.

7 DATA ANALYSIS

In the following a description is given of different types of methods used to further analyse pre-processed data. Statistical analysis of chemical data depends largely on the forensic question. If the question is on identification or classification, various classification methods can be used to answer questions where it is necessary to assign a sample to one of a number of predefined classes or when one wishes to identify a compound. Regression can be used to quantify the amount of a substance in a sample based on predictor variables. Dissimilarity measures can be used to compare samples to each other, with high values of dissimilarity indicating a different origin and low values of similarity indicating a same origin. As an extension of this idea, clustering can be used to group samples based on their similarity in order to obtain groups of samples with possibly the same origin. For all of these questions, one may use dimensionality reduction methods to simplify the data and extract relevant variables for further analysis.

7.1 Classification techniques

As stated, if the forensic question is on identification or classification, various classification methods can be used. Three important examples are linear discriminant analysis, logistic regression and partial least squares discriminant analysis.

Linear discriminant analysis

In Linear Discriminant Analysis (LDA), a linear combination of features is sought for that yields an optimal division into two classes of the data under consideration. The procedure may be generalized to more than two classes.

LDA is a ‘supervised’ learning method, which means that all data points in the training data set must contain information concerning the class from which that data originates. Moreover, the method is constructed to be used with normally distributed features (variables), and if they show severe deviations from such distributions, the output may not be reliable. However, as was shown in subchapter 5, it may be possible to transform the original data to obtain more normal-like distributions. In many cases this will suffice.

As an example, consider chemical profiling of amphetamine in which peak areas of 28 pre-selected impurity compounds are monitored in chromatographic output. The monitoring of these impurity compounds was agreed on in EC financed projects [31, 32], and the data is of Type 1. In subchapter 6.2.4 a data set with peak areas in 744 samples of seized amphetamine powder was used to illustrate different methods of data pre-processing on one of the analytes. This data set will be used in the description of other methods for data analysis.

In the following (and in the description of other methods for data analysis), this data set will be used, however with some slight modifications. Six of the impurity compounds are Ketoxime 1, Ketoxime 2, DPIA 1, DPIA 2, DPIMA1 and DPIMA2. These constitute three pairs (Ketoxime 1 & Ketoxime 2; DPIA 1 & DPIA 2; DPIMA 1 & DPIMA 2) and within each pair the correlation between the two compounds is very high. Therefore, the two compounds in each of these three pairs are summed to a new variable, which is named after the common name, i.e. the three variables Ketoxime, DPIA and DPIMA are created. Nevertheless, they are still referred to as impurity compounds in the descriptions below. Hence, we will refer to 25 impurity compounds.

14 of the 25 impurity compound peak areas show variation that is either close to normal in original scale or becomes close to normal upon having applied square root or fourth root transforms (see subchapter 6.2.4). These compounds and which transformations were used are presented in Table 7.1.

Now, besides these compounds assume there is also information about the country of origin for each of the 744 samples. The countries are The Netherlands, Sweden, Lithuania, Poland and 'Other' which means another country than those four, i.e. in total five. Note that the attribution of countries is completely artificial and has no correspondence with the original data. It has been made to provide a simple example of how LDA works.

We would now investigate whether a linear discriminant analysis can be carried out with which we could classify new samples to their countries of origin. A linear discriminant is a linear combination of the used variables (compounds). The number of linear discriminants that can be found are at most the number of classes (here country categories) minus 1, i.e. in total four.

Table 7.1: 14 of the impurity compounds that in original scale or upon data transformation shows normal variation

Impurity compound	**Transformation of peak areas**
N-Acetylamphetamine	
N-Formylamphetamine	
Benzylamphetamine	
DPPA[a]	fourth root
DPIA[b]	
DPIMA[c]	fourth root
Naphthalene 1	
Naphthalene 2	fourth root
N-Benzoylamphetamine	fourth root
2-Oxo[d]	square root
2,6-Dimethyl-3,5-diphenylpyridine	fourth root
2,4-Dimetyl-3,5-diphenylpyridine	fourth root
Pyridine 7 and 14	
2,6-Diphenyl-3,4-dimethylpyridine	square root

[a] 1,3-Diphenyl-2-propylamine

[b] N,N-di-(β-phenylisopropyl)amine

[c] N,N-di-(β -phenylisopropyl)methylamine

[d] 2-Oxo-1-phenyl-2-(β-phenyl-isopropylamino)ethane

Hence, in this example, there could be at most four linear discriminants and each is of the mathematical form

$$\begin{aligned} LD_i = \; & a_{1i} \times (\text{peak area of N-Acetylamphetamine}) \\ + \; & a_{2i} \times (\text{peak area of N-Formylamphetamine}) \\ + \; & a_{3i} \times (\text{peak area of Benzylamphetamine}) \\ + \; & a_{4i} \times (\text{fourth root of peak area of DPPA}) \\ + \; & \ldots \\ + \; & a_{14i} \times (\text{square root of peak area of 2,6-Diphenyl-3,4-di-} \\ & \quad \text{methylpyridine}) \end{aligned}$$

where the coefficients a_{1i} , a_{2i} , … , a_{14i} for *i = 1, 2, 3, 4* are determined such that the set of linear discriminants (their values) discriminates as well as possible between the groups defined by the five country categories.

Using linear discriminants to classify new observations is done by so-called 'Bayesian updating'. When carrying out LDA so-called prior probabilities are attached for the different classes. The default setting of these probabilities is using the proportions of each class in the training data set. For a new sample the so-called posterior probabilities of the classes are calculated by updating the prior probabilities with the information contained in the calculated discriminants. The sample may then be classified to the class that has the highest 'posterior probability', or other decision rules may be applied.

As an illustration, five new samples are to be classified with respect to their country of origin by using the LDA analysis. In the data set of 744 samples (the training data) the proportions of the five country categories are: The Netherlands (36%), Sweden (24%), Lithuania (18%), Poland (20%), Other (2%). Using the default settings of the LDA, the prior probability for a sample to have The Netherlands as origin is 0.36, the prior probability for a sample to have Sweden as origin is 0.24 etc. In Table 7.2, computed posterior probabilities of the country categories for each of the five new samples are shown.

Table 7.2: Computed posterior probabilities of each of the five country categories (The Netherlands, Sweden, Lithuania, Poland and Other) for each of five new samples that need to be classified

New sample	**The Netherlands**	**Sweden**	**Lithuania**	**Poland**	**Other**
1	0.737	7×10^{-7}	0.243	0.020	7×10^{-7}
2	2×10^{-7}	0.9999	5×10^{-12}	8×10^{-5}	3×10^{-6}
3	0.448	2×10^{-8}	0.542	0.009	8×10^{-7}
4	0.455	0.009	0.007	0.529	2×10^{-5}
5	2×10^{-4}	7×10^{-6}	5×10^{-6}	1×10^{-4}	0.9997

For sample 2 and 5 it can be seen from Table 7.2 that the posterior probabilities are very high for Sweden (sample 2) and Other (sample 5). For these two samples the classification to country category can be said to be unambiguous.

In contrast, for samples 3 and 4 the two highest posterior probabilities are not that far from each other in value and hence using the maximum posterior probability as a decision rule may cause misclassification. Note that the sum of probabilities of each sample is always 1.

Logistic regression

Logistic regression is used to predict the value of a binary response variable, values of which are typically expressed as 0 or 1. For given values of the explanatory variables or predictors, logistic regression is used to establish the probability of the response variable taking the value of 1. This is achieved by taking a linear combination of the predictors, with an additional transformation of this combination to 'force' it to the 0-1 range. This method is useful for producing a statistical classifier for purposes of e.g. identifying compounds.

Logistic regression can be said to be a more generally usable tool for classification than LDA. The reason for this is that the predictors need not be normally distributed, or even close, whereas for LDA they need to be. Besides, if normally distributed features are used as predictors in a classification problem with two classes, it is usually possible to obtain similar results with logistic regression as with LDA. Logistic regression can also be extended for use with more than two categories. The method is then often referred to as 'multinomial' or 'polytomic' (logistic) regression. The word 'logistic' stems from the fact that the transformation of the linear combination to obtain a value that is in the 0-1 range in most cases is the logistic function.

As an example, we return to the data with impurity profiling of samples of seized amphetamine powder used in the previous example of LDA. In this data we artificially attributed countries of origin to the sample. Assume now that we are only interested in whether the sample originates from The Netherlands or not. This means that we have a response variable with value 1 if the sample originates from The Netherlands and 0 otherwise. We can now use the peak areas of the impurity compounds listed in Table 7.1 as predictors. As stated, with logistic regression it is not necessary to transform any of these peak areas since there is no requirement of normal distributed predictors. However, we will use the transforms as given in Table 7.1 since these transforms also reduce a possible imbalance in the data with values largely deviating from the core of the data set (cf. subchapter 6).

Using the logistic function as transformation of the linear combination, we are fitting the model

$$Prob(\text{The Netherlands}) = \frac{e^{\text{linear combination}}}{1 + e^{\text{linear combination}}}$$

where stands for the probability that a sample originates from The Netherlands and 'linear combination' is of the form

c_0

+ c_1× (peak area of N-Acetylamphetamine)

+ c_2 × (peak area of N-Formylamphetamine)

+ c_3 × (peak area of Benzylamphetamine)

+ c_4 × (fourth root of peak area of DPPA)

+…..

+ c_{14} × (square root of peak area of 2,6-Diphenyl-3,4-dimethylpyridine)

where the coefficients c_0, c_1, c_2, …, c_{14} are constants that will be estimated when carrying out the logistic regression analysis. Note that in contrast with the linear combinations in LDA there is a baseline constant c_0 which is common for regression models in general. This prevents forcing a model that would state that if every predictor is zero (here that none of the compounds are present in the sample) then the probability must be ½ (which may not be the case).

With these coefficients estimates it is possible to predict the probability that a new sample originates from The Netherlands, and if that probability exceeds a pre-defined threshold the sample is classified as such. A threshold is not trivially set, and it is not customary to set it to 0.5. Rather a clearly higher value is used (e.g. 0.7) since the risk of misclassification might otherwise be too large. See further subchapter 8 about method validation.

Now carrying out the logistic regression analysis on the 744 sample of seized amphetamine powder leads to the following estimated linear combination:

–2.54

+ 0.0086 × (peak area of N-Acetylamphetamine)

– 0.0419 × (peak area of N-Formylamphetamine)

+ 0.0237 × (peak area of Benzylamphetamine)

+ 0.0002 × (fourth root of peak area of DPPA)

+ 0.0378 × (peak area of DPIA)

– 0.9833 × (fourth root of peak area of DPIMA)

– 0.0377 × (peak area of Naphthalene 1)

+ 0.1747 × (fourth root of peak area of Naphthalene 2)

– 0.2956 × (fourth root of peak area of Benzoylamphetamine)

+ 0.2827 × (square root of peak area of 2-Oxo)

– 0.1863 × (fourth root of peak area of 2,6-Dimethyl-3,5-diphenylpyridine)

+ 1.2977 × (fourth root of peak area of 2,4-Dimethyl-3,5-diphenylpyridine)

+ 0.1871 × (fourth root of peak area of Pyridine 7 and 14)

– 1.7529 × (fourth root of peak area of 2,6-Dimethyl-3,5-diphenylpyridine)

A difference compared to LDA is that the estimated coefficients of this linear combination have a straightforward qualitative interpretation. Since the logistic transformation is monotonic, an interpretation of a coefficient with a positive value is that the probability of the sample originating from The Netherlands increases when the corresponding predictor increases. Similarly, an interpretation of a coefficient with a negative value is that the probability of the sample originating from The Netherland decreases when the corresponding predictor increases.

Hence, if any (or several) of the peak areas (transformed or not) of N-Acetyl-amphetamine, Benzyl-amphetamine, DPPA, DPIA, Naphthalene 2, 2-Oxo, 2,4-Dimethyl-3,5-diphenylpyridine and Pyridine 7 and 14 is larger in one sample compared to another sample, while the rest of the peak areas are about the same, the former sample has a *higher* probability of originating from The Netherlands than the latter sample.

If any (or several) of the peak areas (transformed or not) of N-Formylamphetamine, DPIMA, Naphthalene 1, Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,6-Dimethyl-3,5-diphenylpyridine is larger in one sample compared to another sample, while the rest of the peak areas are about the same, the former sample has a *lower* probability of originating from The Netherlands than the latter sample.

Now, consider again the five new samples that were classified with LDA. There, one sample (no. 1) was attributed to The Netherlands as its origin while the rest were attributed to other countries. We can now use the formula for *Prob (The Netherlands)* with the estimated linear combination and predict the probability for each sample to originate from The Netherlands. In Table 7.3 these predicted probabilities for all five samples are given.

Table 7.3: Predicted probabilities of five new samples to originate from The Netherlands

New sample	**Predicted probability of sample originating from The Netherlands**
1	0.625
2	0.057
3	0.604
4	0.335
5	0.187

If we define a decision rule as ‘the predicted probability must exceed 0.7’, then none of the samples will be attributed to The Netherlands as their origin. Nevertheless, we can see in Table 7.3 that the sample with the highest predicted probability of originating from The Netherlands is sample 1. This sample was also attributed to The Netherlands in the previous LDA analysis.

Partial least squares

Partial least squares regression (PLS-R) is a multivariate method that models the relation between two blocks of variables commonly referred to as *X* (predictors) and *Y* (responses). Simplified, the procedure can be described as a simultaneous calculation of PCs of each matrix under certain constraints. The calculated PCs for *X* and *Y* are adjusted to maximize the covariance between them. Therefore, the directions of the obtained PCs are somewhat different to those in ordinary PCA. Given values for *X*, the *Y* values are obtained by first transforming the *X* values to the corresponding PCs, these are transformed to the PCs of *Y* by ordinary least squares regression and finally values of *Y* are obtained by transforming back from the PC space. PLS is quite general and is especially useful for when the predictors are highly collinear or when there are more predictors than observations. Partial Least Squares - Discriminant Analysis (PLS-DA) is a special case of this where the *Y* matrix is categorical, representing discrete classes. Now, the PCs for the *X* matrix are adjusted to maximize the covariance between the components and the classes indicated by the *Y* matrix.

7.2 Quantification: regression techniques

Ordinary least squares regression (OLS-R), also: simple linear regression, is a traditional method for using numerical variables called predictors to predict the value of another numerical variable called the response. It is appropriate for purposes of quantification and it accomplishes this by finding an optimal linear transformation of the predictors to predict the value of the response variable. OLS-R has many applications in chemical analysis, and is not (like other methods) connected to a typical forensic question.

An example is the following:

Khat is a flowering plant (Catha edulis in Latin) that is native in East Africa and on the Arabian Peninsula. It contains the stimulant cathinone, and is chewed to get an inebriation effect. It is classified by WHO as a drug of abuse and as an illicit drug in several Western World countries. Khat is gathered in bundles of sprigs where a bunch is considered to be of “chewing size”. A number of bundles historically comprising a daily dose are then covered with a banana leaf *(partly to conserve some of their freshness), and called a 'marduuf'* [1]. *Consignments of such marduufs are then smuggled to users living in the Western World but practicing drug culture from their home countries.*

When a consignment for delivery is discovered and confiscated by the customs there are issues with estimating the quantity of the sprigs (since the banana leaves are not classified). The most accurate estimate would be to remove all banana leaves and weigh all bunches of sprigs together, but evidently this takes a lot of both time and effort. One idea is therefore to investigate whether a deduction factor can be determined and used on the weight of the whole consignment (i.e. including the banana leaves).

In a study 260 marduufs were selected, and their gross weights and wrapping weights were registered. In Figure 7.1 a scatter plot of wrapping weights against gross weights are shown.

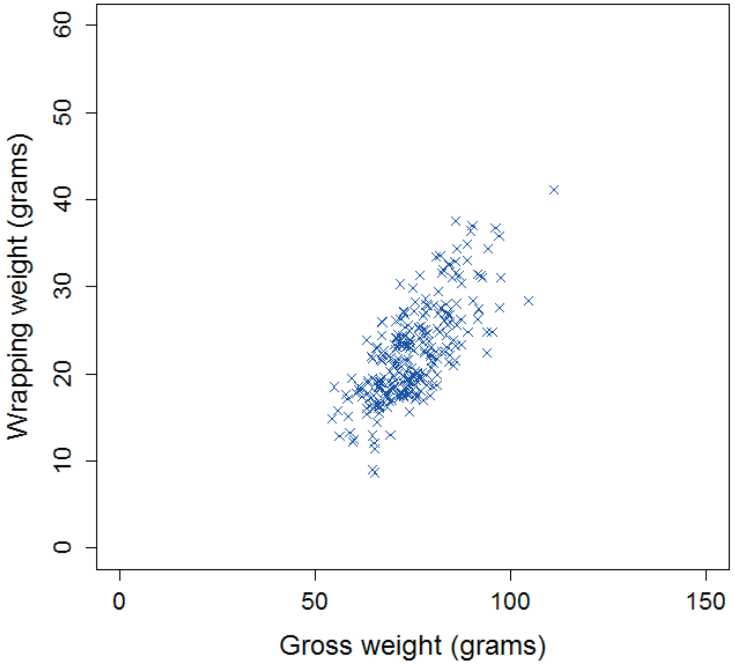


Figure 7.1 Scatter plot of wrapping weights against gross weights for 260 selected marduufs with khat

[1] Marduuf is a street market unit for Khat; consisting two bunches of Khat leaves

Although the plot in Figure 7.1 is rather noisy, there is clearly a correlation between the wrapping weight and the gross weight. If a factor for deduction of the wrapping weight should be found, this would correspond with the slope of a fitted line to the scatter plot. However, the data is scattered around a point that is far from the origin of the coordinate system and should we force such a line to go through the point (0;0) it would not fit well with the points. Hence, to fit a line we must include some kind of offset parameter. The statistical model for these data is:

Wrapping weight = *intercept* + *slope* × Gross weight + deviation (grams)

where the parameter *intercept* is the offset parameter = the value on the y-axis where the line will cross, the parameter *slope* is the slope of the line and deviation is the vertical deviation from a point to the line (there are both positive and negative deviations). This is referred to as a (simple) linear regression model where Wrapping weight is the 'response variable' and Gross weight is the 'explanatory variable' or 'predictor' (cf. the description of logistic regression above).

Now, using the method of ordinary least squares a (regression) line is fitted with *intercept* estimated by – 9.7 and *slope* by 0.43. Hence the fitted line can be written as:

Wrapping weight = –9.7 + 0.43 × Gross weight (grams).

Note that the term 'deviation' in the model is there to explain the position of a single point, on average this deviation should be zero. The regression line represents the expected (or mean) relationship between Wrapping weight and Gross weight, but is also at the same time the best prediction of the Wrapping weight.

In Figure 7.2 the fitted line is added to the previous scatter plot.

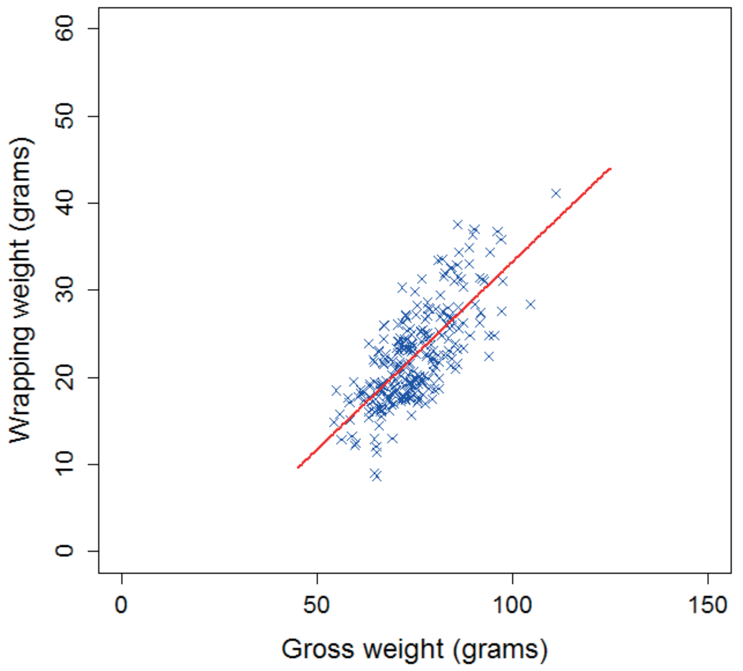


Figure 7.2 Scatter plot of wrapping weights against gross weights for 260 selected marduufs with khat, together with a fitted regression line with intercept –9.7 and slope 0.43

Now, for a new consignment of marduufs the predicted weight of the sprigs (the Net weight) can be predicted using the following formula:

predicted Net weight = Gross weight – predicted Wrapping weight =

= Gross weight – (–9.7 + 0.43 × Gross weight) =

= 0.57 × Gross weight + 9.7 (grams).

This formula is not the same as just applying a deduction factor to the gross weight, but for consignments with large weights (several kilograms) the addition of 9.7 grams is negligible and might as well be omitted.

It should be said that even if the fitted regression line captures the assumed background relationship between Wrapping weight and Gross weight, many points are quite far from the line. This might mean that not all variation in the data has been captured by the model. When applying OLS to these kinds of data, a measure of the fit is the so-called '*coefficient of determination*'. This is a measure between 0 and 1, where 0 means that no variation has been explained by the regression model, and 1 means that all variation has been explained. Values above 0.8 are often considered as sufficiently high, but it depends on the kind of application the fitted model shall be used for. For this fitted model the coefficient of variation is 0.73, which may be considered too low when the purpose is to get an accurate prediction of the Net weight of a consignment.

7.3 Comparison by dissimilarity-based methods

In comparison problems, where it is of interest to discover whether samples share a common source, different measures of dissimilarity, calculated between pairs of samples from numerical variables, can be used to indicate linked samples. The prerequisite for this is that the selected dissimilarity measure adequately summarises relevant differences between data items. To ensure this, careful consideration of the properties of different measures is necessary.

The Euclidean and Manhattan distances are typical examples of what could be called 'classical' dissimilarity measures. The first of these is the usual 'straight line' distance between two points in space. The second one, on the other hand, corresponds to the distance when travelling along each coordinate axis separately. Indeed, the Manhattan distance is sometimes called the 'taxicab' metric as a reference to the way taxis traverse the grid layout in Manhattan, New York. In some literature it is also called 'city block distance'.

Alternatively, one may consider dissimilarity measures such as the Pearson correlation distance and cosine distance. Both of these are, in a sense, measures of shape similarity and are quite closely related. The former is based on the Pearson correlation coefficient and it measures how well one can reconstruct one of the vectors using a linear transformation of the other one. In particular, it ignores any absolute differences in the samples and is only affected by differences in relation to the mean and standard deviation calculated over the indices of the vectors. In contrast, cosine dissimilarity measures the difference in the angles of the two vectors, again ignoring absolute differences. The two measures are connected by the fact that Pearson correlation is the cosine similarity of z-score transformed vectors.

In subchapter 7.5 it will be shown how different measures of dissimilarities may affect hierarchical clustering analysis (HCA).

7.4 Principal component analysis

Principal component analysis (PCA) is a multivariate method that is used to simplify interpretation of large data matrices. The procedure involves replacement of original variables with so-called orthogonal principal components (PCs), which explain the variation in the different dimensions as efficiently as possible. PCA provides a graphical overview of both the samples (by means of 'scores plots') and variables (by means of 'loadings plots') in the database, using only a few dimensions. The scores plot makes it possible to identify classes of samples, i.e. samples that are similar and dissimilar. Loading plots can be used to find variables that are positively or negatively correlated to each other. The position of the variables in the loading plot can also be used to explain the position of the samples in the scores plot. An advantage of PCA is that it can handle co-variation of variables.

As an example, consider again chemical profiling of amphetamine in which peak areas of 25 pre-selected impurity compounds are monitored in chromatographic output. For the set of 744 samples of seized amphetamine powder (see chapter 6), we illustrate potential correlations between the peak areas of six of these compounds (Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine). In Figure 7.3 pairwise scatter plots of the peak areas of these six substances are shown.

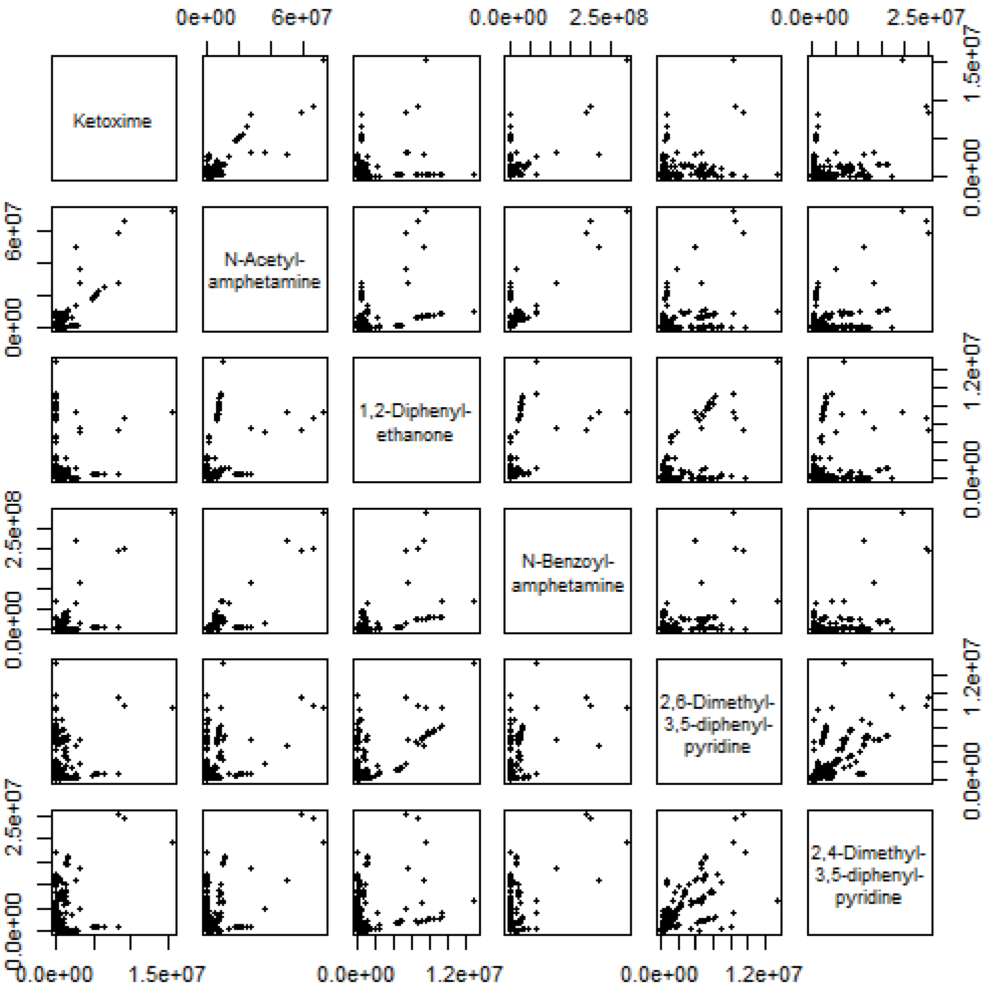


Figure 7.3 Pairwise scatter plots of peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine

The scatter plots in Figure 7.3 reveal mostly fairly weak pairwise relationships between the peak areas of the six compounds. The matrix of pairwise correlation coefficients is shown in Table 7.4. The pairwise correlation coefficient used is Pearson's correlation coefficient which is computed as

$$Correlation\left(\mathbf{y}^{(1)},\mathbf{y}^{(2)}\right)=\frac{\sum_{i=1}^{n}\left(y_i^{(1)}-\bar{y}^{(1)}\right)\cdot\left(y_i^{(2)}-\bar{y}^{(2)}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i^{(1)}-\bar{y}^{(1)}\right)^2\cdot\sum_{i=1}^{n}\left(y_i^{(2)}-\bar{y}^{(2)}\right)^2}}$$

where $y_i^{(1)}$ is the observation for compound 1 in sample i, $y_i^{(2)}$ is the observation for compound 2 in sample i, and the sample means of these two compound values are denoted by and respectively.

Table 7.4 Pairwise correlation coefficients of peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine

	Ketoxime	N-Acetyl-amphetamine	1,2-Diphenyl-ethanone	N-Benzoyl-amphetamine	2,6-Dimethyl-3,5-diphenyl-pyridine	2,4-Dimethyl-3,5-diphenyl-pyridine
Ketoxime	1	0.86	0.26	0.67	0.31	0.41
N-Acetyl-amphetamine	0.86	1	0.50	0.87	0.43	0.51
1,2-Diphenyl-ethanone	0.26	0.50	1	0.58	0.59	0.28
N-Benzoyl-amphetamine	0.67	0.87	0.58	1	0.48	0.51
2,6-Dimethyl-3,5-diphenyl-pyridine	0.31	0.43	0.59	0.48	1	0.75
2,4-Dimethyl-3,5-diphenyl-pyridine	0.41	0.51	0.28	0.51	0.75	1

Note that even if the scatter plots in Figure 7.3 do not indicate strong relationships, the pairwise correlation coefficients between peak areas of Ketoxime and N-Acetylamphetamine and between peak areas of N-Acetylamphetamine and N-Benzoylamphetamine are both over 0.8.

Principal component analysis (PCA) of the peak areas of the six compounds will give as output so-called principal components (PC1, PC2, …), which are six linear combinations of the compounds' peak areas that are independent (orthogonal) of each other. The total number of principal components is always the same as the total number of variables, but as we shall see later, they decrease in importance from the first to the last. In this example the first principal component is:

PC1 = 0×(peak area of Ketoxime) + 0.256×(peak area of N-Acetylamphetamine) + 0×(peak area of 1,2-Diphenylethanonen + 0.961×(peak area of N-Benzoylamphetamine) + 0×(peak area of 2,6-Dimethyl-3,5-diphenylpyridine) + 0×(peak area of 2,4-Dimethyl-3,5-diphenylpyridine) =

= 0.256×(peak area of N-Acetylamphetamine) + 0.961×(peak area of N-Benzoylamphetamine)

The coefficients in front of the compounds' peak areas are called *loadings* and give some understanding about how the peak area of a specific compound affects the principal component. If no scaling pre-processing has been made to the data, the loadings cannot be compared with each other. As an example: in this first principal component, we can see that it depends only the peak areas of N-Acetylamphetamine and N-Benzoylamphetamine, but since no scaling pre-processing was done it is not meaningful to compare the two loadings (0.256 and 0.961 respectively).

All six principal components are presented in Table 7.5, in which the rows are the principal components, the columns are the 'different compounds' peak areas, and the entry in each cell is the loading on the column compound peak area for the principal component of the row.

Table7.5: Loadings for each principal component in a PCA of the peak areas of the compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine in 744 samples of seized amphetamine powder

Principal Component	Peak area of...					
	Ketoxime	N-Acetylamphetamine	1,2--Diphenylethanone	N-Benzoylamphetamine	2,6-Dimethyl-3,5-diphenylpyridine	2,4-Dimethyl-3,5-diphenylpyridine
PC1		0.256		0.961		
PC2	0.155	0.725		–0.255	0.193	0.590
PC3	0.135	0.603			–0.318	–0.712
PC4	0.106		0.800		–0.532	0.243
PC5	0.399		–0.501		0.719	–0.265
PC6	0.887	–0.198	0.325		–0.246	

To be able to compare the loadings on different compound peak areas the peak areas need to be standardized or transformed to *z*-scores before the principal component analysis is carried out. In Table 7.6 loadings are shown for the six principal components from a PCA of *standardized* peak areas from the analyzed 744 samples of seized amphetamine powder.

Table 7.6: Loadings for each principal component in a PCA of standardized peak areas of the compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine in 744 samples of seized amphetamine powder

Principal Component	Standardized peak area of...					
	Ketoxime	N-Acetylamphetamine	1,2--Diphenylethanone	N-Benzoylamphetamine	2,6-Dimethyl-3,5-diphenylpyridine	2,4-Dimethyl-3,5-diphenylpyridine
PC1	0.397	0.470	0.347	0.461	0.383	0.376
PC2	0.512	0.361	–0.300	0.203	–0.583	–0.370
PC3	0.208		–0.748	–0.167	0.112	0.597
PC4	0.583		0.150	–0.661	0.358	–0.269
PC5		0.172	0.418	–0.434	–0.583	0.518
PC6	0.444	–0.786	0.183	0.306	–0.184	0.157

Note that the loadings in Table 7.6 are completely different from the loadings of Table 7.5, and in particular, that peak areas with no loadings in Table 7.4 can have loadings in Table 7.6. Consider for example the loadings for the first principal component (PC1), i.e. the first row of Table 7.6. These can now be interpreted as follows.

Assume two different materials for which PC1 values have been computed.

- If the standardized peak area of Ketoxime is K units higher in material 1 compared to material 2, while all other standardized peak areas are about the same for the two materials, then PC1 will be 0.397×K units higher for material 1 than for material 2.

- If the standardized peak area of N-Acetylamphetamine is K units higher in material 1 compared to material 2, while all other standardized peak areas are about the same for the two materials, then PC1 will be 0.470×K units higher for material 1 than for material 2.

- Similar for the remaining 4 coefficients.

Note that a negative loading will have the reverse interpretation. For instance, if the standardized peak area of 1,2-Diphenylethanone is K units higher in material 1 compared to material 2, while all other standardized peak areas are about the same, then PC2 will be 0.300×K units *lower* than for material 2.

A difference in K units between two standardized peak areas is equivalent to a difference in K×s units between the peak areas in original scale, where “s” is the standard deviation computed from the set of peak areas for the current compound.

Now, the main purpose with PCA is to reduce the number of dimensions for the interpretation of the data. Since any of the six compounds can have loadings in a principal component, each principal component captures some part of the total variation in data. The first principal component always captures the largest part of the variation, the second principal component captures the second largest part of the variation etc. The idea is that a subset of the principal components (the first, the first two, the first three, …) may together capture a sufficient amount of the total variation for drawing conclusions, e.g. about groupings in the data. In software for principal component analysis there is an option to produce a so-called ‘scree plot’. This is a graph that depicts the amount of variation (statistical variance) that is captured by each principal component. In Figure 7.4 a scree plot is shown for the PCA done on the original peak areas of the six compounds of the 744 samples of seized amphetamine powder.

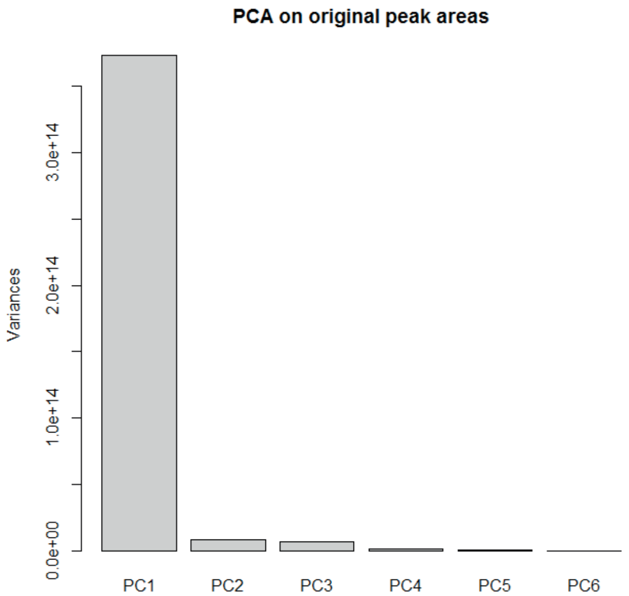


Figure 7.4 Screen plot of the six principal components obtained from PCA done on original peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine. The bar of each principal component shows the amount of variance in the data set captured by that component

We can see in Figure 7.4 that the first principal component captures a very large part of the variance compared to the other components. This can be interpreted as it would suffice to retain the first principal component from the PCA to draw conclusions about the data set. In other words, the dimension may be reduced from six (as is the number of original variables) to one. However, the scree plot based on PCA on non-standardized data may not always give rise to such a clear interpretation. In Figure 7.5 is shown the scree plot for the PCA done on standardized peak areas (i.e. the loadings of which were given in Table 7.6.

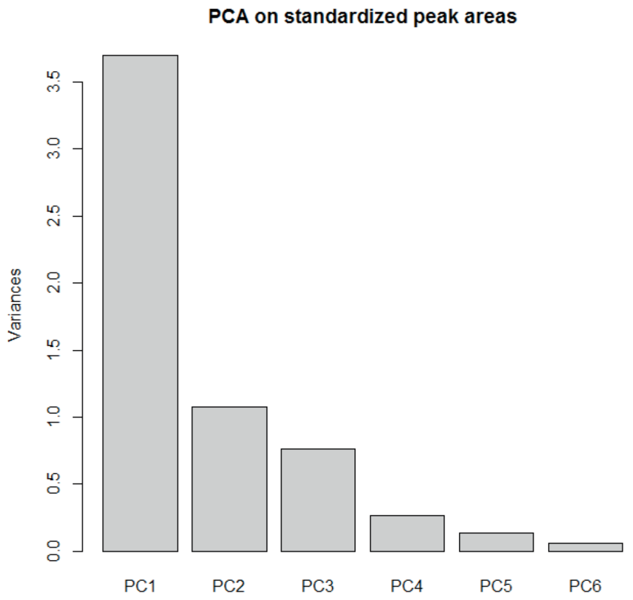


Figure 7.5 *Screen plot of the six principal components obtained from PCA done on standardized peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine. The bar of each principal component shows the amount of variance in the standardized data set captured by that component*

Compared to the graph in Figure 7.5 we observe that the variances are of a much smaller in magnitude. This is so since the data has been standardized before PCA was applied, i.e. al variables have standard deviation 1. We can also see that the bars for PC2, PC3, …, PC6 are much higher relative to the bar for PC1 than was the case in Figure 7.4. From a scree plot for PCA done on standardized variables there is an easy to use rule for deciding how many principal components should be retained to capture a sufficient amount of the total variation for conclusions to be drawn. All principal components capturing each a variance at least equal to one should be retained, while the other components can be left out from further interpretation. In the graph of Figure 7.5 it can be seen that the first two principal components have each a captured variance greater than one, while the rest of the components have not. Hence, we should retain the first two principal components for further conclusions to be drawn. Note that this choice was not obvious from inspection of the scree plot in Figure 7.4. That plot rather pointed towards retaining only the first principal component.

Once we have decided upon the principal components to be retained, we can illustrate with a score plot and a loadings plot. A score plot means a scatter plot of the values of the (retained) principal components - named *scores*. Two-dimension scatter plots in which the scores of one principal component is plotted against the scores of another principal component are easiest to interpret but with today's graphical tools in software it is also fairly easy to interpret a three-dimensional plot, especially if it can be rotated freely. However, it is only meaningful to produce score plots for the principal components retained, but should these be more than two then several plots may be needed for drawing useful conclusions.

A scatter plot may reveal previously not known subgroups of the data set shown as swarm of points more or less separated from each other. If the data points are already grouped (before the PCA is done) this can be indicated in the scatter plot by choosing different colors of the points depending on to which group they belong.

In Figure 7.6 is shown a score plot with the scores of principal component 2 (PC2) against the scores of principal component 1 (PC1) from the PCA of standardized peak areas of the six compounds of the 744 samples of seized amphetamine powder

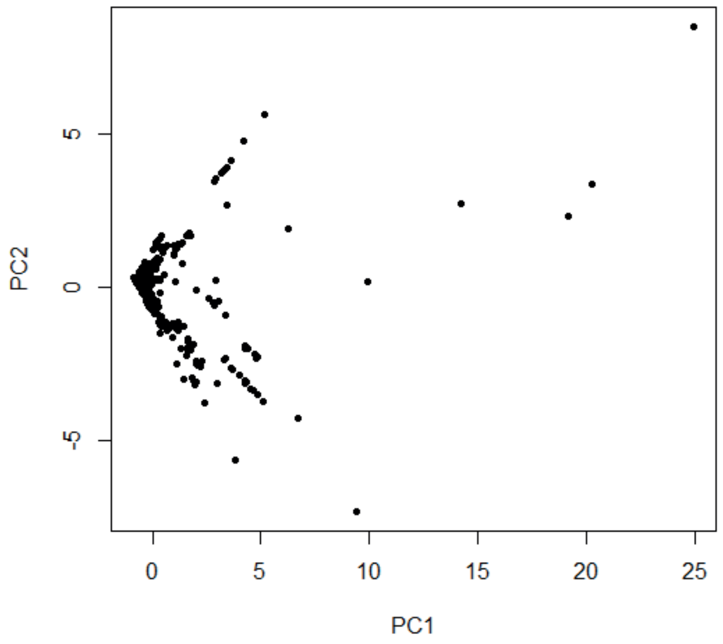


Figure 7.6 *Score plot of principal component 2 (PC2) scores against principal component 1 (PC1) scores obtained from PCA done on standardized peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine*

The score plot of Figure 7.6 does not reveal any clear swarms of points indicating groups in the data set. Now, we might be especially interested in the few points that deviate from the funnel-shaped pattern of the majority of the points in the graph. For each sample in the data set we have information about the dry concentration of amphetamine in the powder. If we classify this into the classes “below 30%”, “from 30% to 60%” and “above 60%” we can color the points according to this classification. In Figure 7.7 such a graph is shown, i.e. the same points as plotted in Figure 7.6 are plotted in Figure 7.7, but with colors depending on the amphetamine dry concentration class (black: below 30%, red: between 30% and 60%; and green: above 60%).

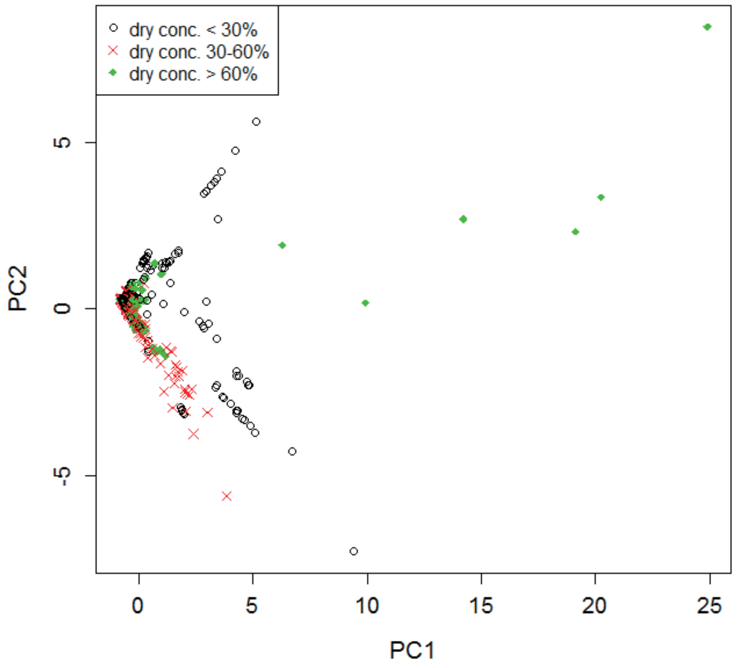


Figure 7.7 *Score plot of principal component 2 (PC2) scores against principal component 1 (PC1) scores obtained from PCA done on standardized peak areas of the six compounds Ketoxime, N-Acetylamphetamine, 1,2-Diphenylethanone, N-Benzoylamphetamine, 2,6-Dimethyl-3,5-diphenylpyridine and 2,4-Dimethyl-3,5-diphenylpyridine obtained from chemical profiling of 744 samples of seized amphetamine. The colors and symbols of the points are due to the amphetamine dry concentration class to which they belong (black circles: below 30%; red crosses: from 30% to 60%; green diamonds: above 60%)*

In the plot of Figure 7.7 it can be seen that all points deviating from the general funnel-shaped patters are samples with dry concentration of amphetamine above 60%.It can also been seen that samples with dry concentration of amphetamine between 30 % and 60 % tend to have negative scores on PC2 (a majority of the red crosses are below PC2 =0).

7.5. Hierarchical cluster analysis (unsupervised classification)

Hierarchical cluster analysis (HCA) is a method for analysing group structure in a dataset based on pairwise dissimilarities. HCA starts by assigning each sample to its own group or cluster. After this, at each stage based on a user-specified rule, two closest groups are combined to form a new, bigger cluster. This continues until only one cluster remains. This process produces a hierarchy of possible clustering solutions that is group assignments for the dataset. This hierarchy can be represented as a tree – a so-called *dendrogram* – that can be examined as is to explore the structure of the dataset or cut at certain height to provide a specific clustering solution for the dataset. The application of HCA requires the user to choose the dissimilarity measure for determining which samples are similar as well as so called 'linkage' method to determine which groups of samples are similar. The first of these depends on what is a useful metric for determining similarity between any two samples, the details of which were described in section on comparison methods. For the second choice, it is important to realise that even if pairwise dissimilarities have been determined between samples, this does not uniquely determine how dissimilar any two groups of samples are with respect to the used dissimilarity measure. There are several methods for assessing group dissimilarity, or linkage, such as single, average, complete and Ward's method linkages. Single linkage means that for any group of samples, the distance between the groups is assigned as the minimum of the pairwise dissimilarities between the groups. Average linkage involves the arithmetic mean of the pairwise dissimilarities and complete linkage refers to using the maximum of these dissimilarities. Ward's method is in a sense more complex and is theoretically appropriate only when using Euclidean distances. Its basic idea is to minimize the variability within each formed group. That is, at each stage the two groups for which merging produces the minimal increase in internal variability are combined.

We can illustrate how HCA is applied by again using impurity profiling data from the 744 samples of seized amphetamine powder that has been used in several previous examples. This time we will use all 25 impurity compounds with no data transformation since no general requirements are imposed on the distributional properties of the data (as was the case for LDA). What is necessary is that a dissimilarity measure exists that can be applied on the data. In this example all variables (impurity compounds) are continuously varying, which means that a lot of dissimilarity measures can be used. We will compare the clustering outputs for the two measures Euclidean distance and Pearson correlation distance (see also subchapter 7.3) together with single linkage and average linkage.

The purpose is to obtain knowledge about how many laboratories have produced amphetamine for the market during the last 6 months. The dendrogram showing the output of an HCA is a tree diagram with its outermost twigs at the bottom and the root at the top. However, since we have 744 samples, each of which constitutes one of the outermost twigs, there is not enough space for a figure in this document to show such a dendrogram in which all these twigs are discernible. Therefore, for illustration purposes, we randomly select 50 of the 744 samples and apply HCA to these.

The dendrograms for all four combinations of Euclidean distance/Pearson correlation distance and simple linkage/average linkage are shown in Figures 7.8 - 7.11.

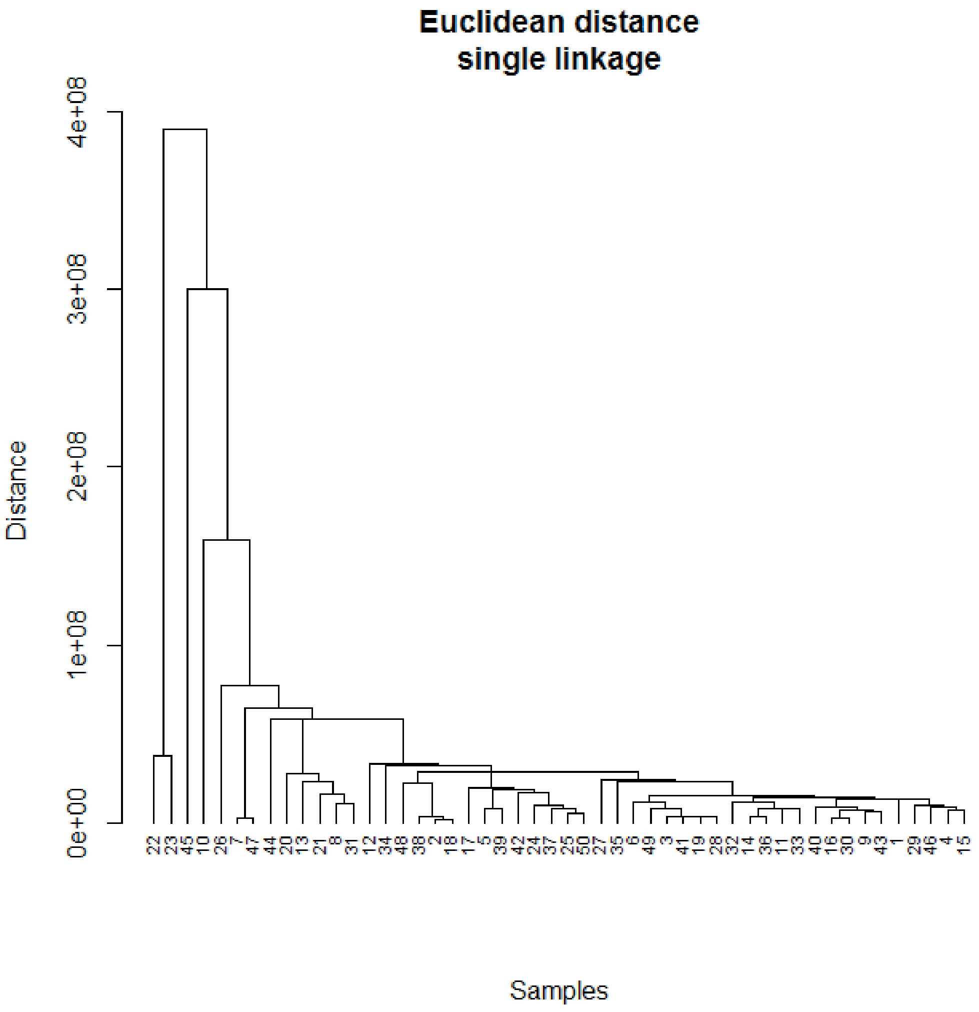


Figure 7.8 *Dendrogram for the hierarchical clustering of a selection of 50 samples from 744 samples of seized amphetamine powder using Euclidean distances and single linkage*

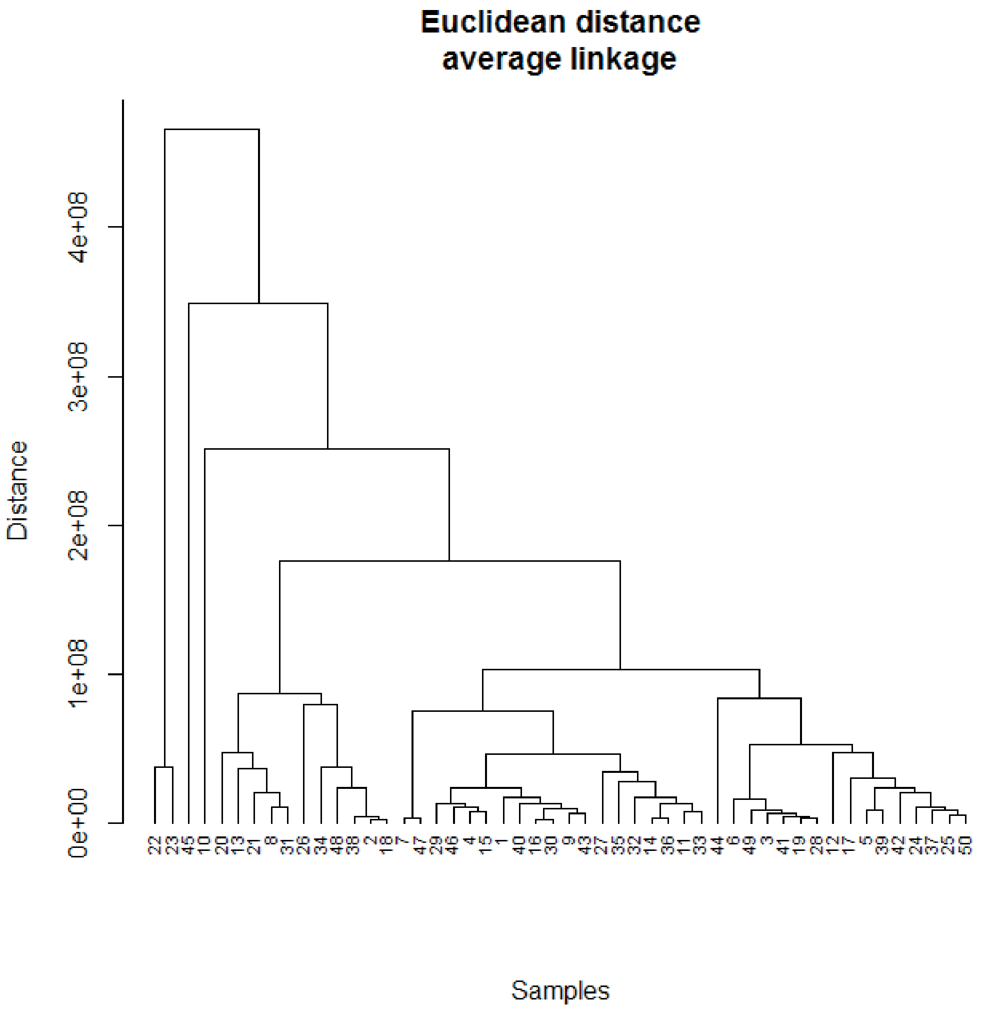


Figure 7.9 *Dendrogram for the hierarchical clustering of the same selection of 50 samples as of Figure 7.8 using Euclidean distances and average linkage*

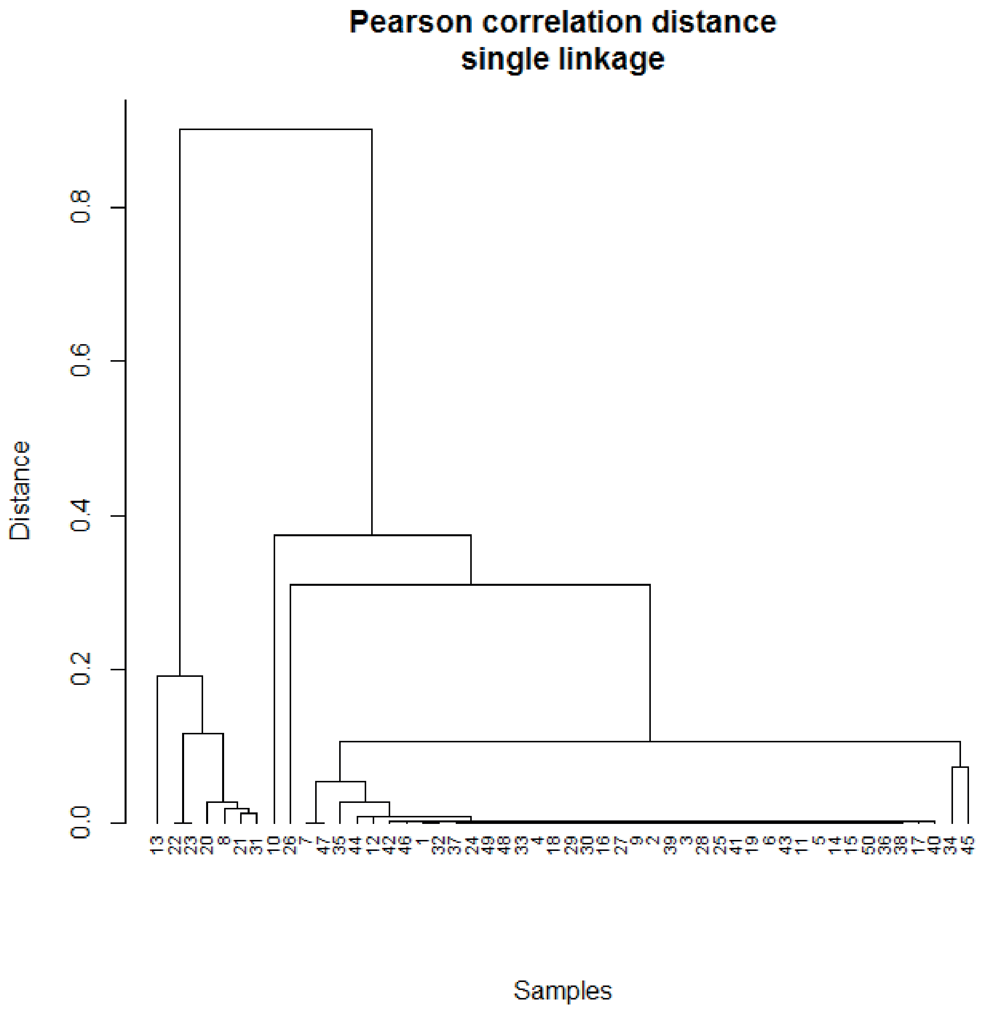


Figure 7.10 *Dendrogram for the hierarchical clustering of the same selection of 50 samples as of Figures 7.8 and 7.9 using Pearson correlation distances and single linkage*

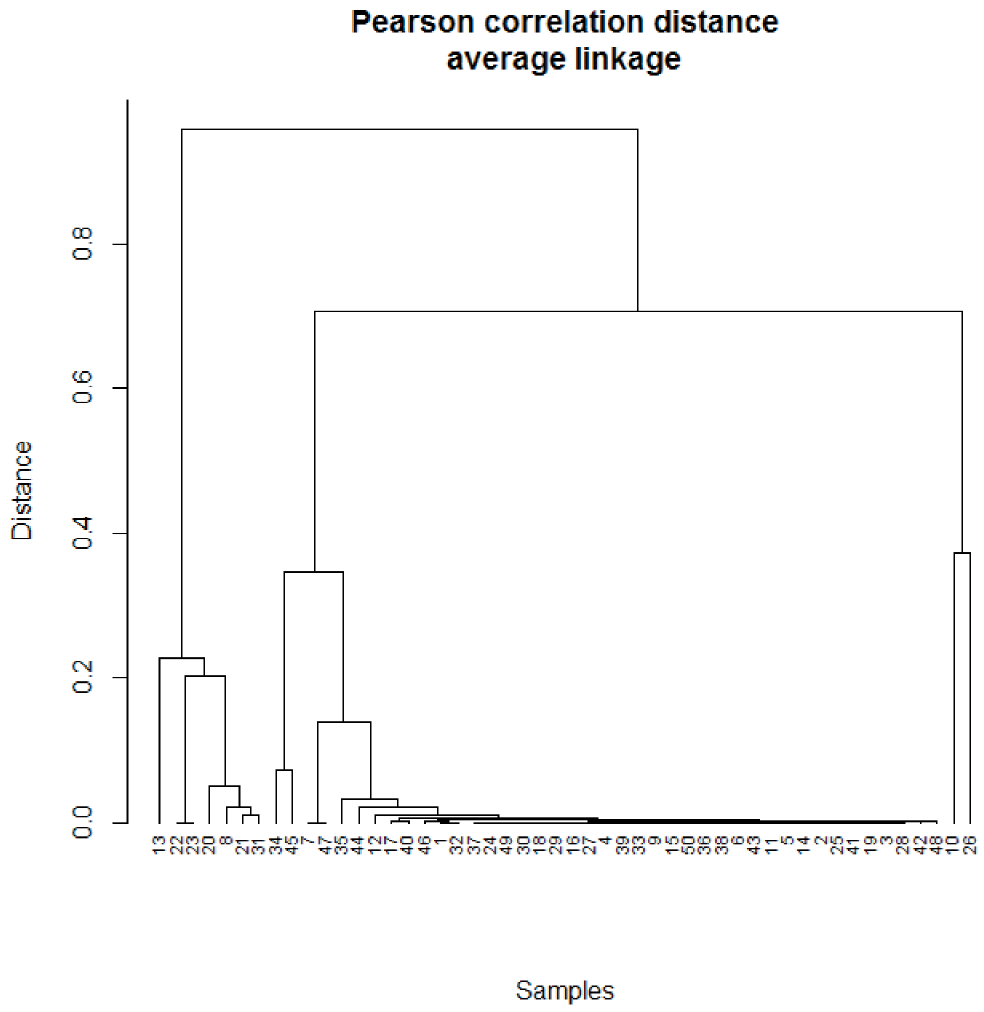


Figure 7.11 *Dendrogram for the hierarchical clustering of the same selection of 50 samples as of Figures 7.8, 7.9 and 7.10 using Pearson correlation distances and average linkage*

The dendrograms in Figures 7.8 - 7.11 look quite different even if a kind of right-skewed tree can be seen in all graphs. In particular, it should be noted that the order of the samples on the x-axis is different between the plots. This is so since for each method the first pairs that are joined into a cluster depend on the dissimilarity measure used and the samples are sorted so that no joining lines are crossing each other.

Now, to deem upon the number of producing laboratories we should in each dendrogram start from the top and follow the splitting into clusters as we move down vertically (decreasing the distance). At the beginning, the splitting into smaller clusters is not frequent with respect to the decrease of the distance, but at some point, it suddenly becomes much more frequent. In Figure 7.8 this point is about a distance of 4×10^7. This is illustrated by the red dashed line in Figure 7.12 below. Above this line there are at most 8 clusters in the dendrogram, and hence 8 is a reasonable estimate of the number of laboratories producing amphetamine.

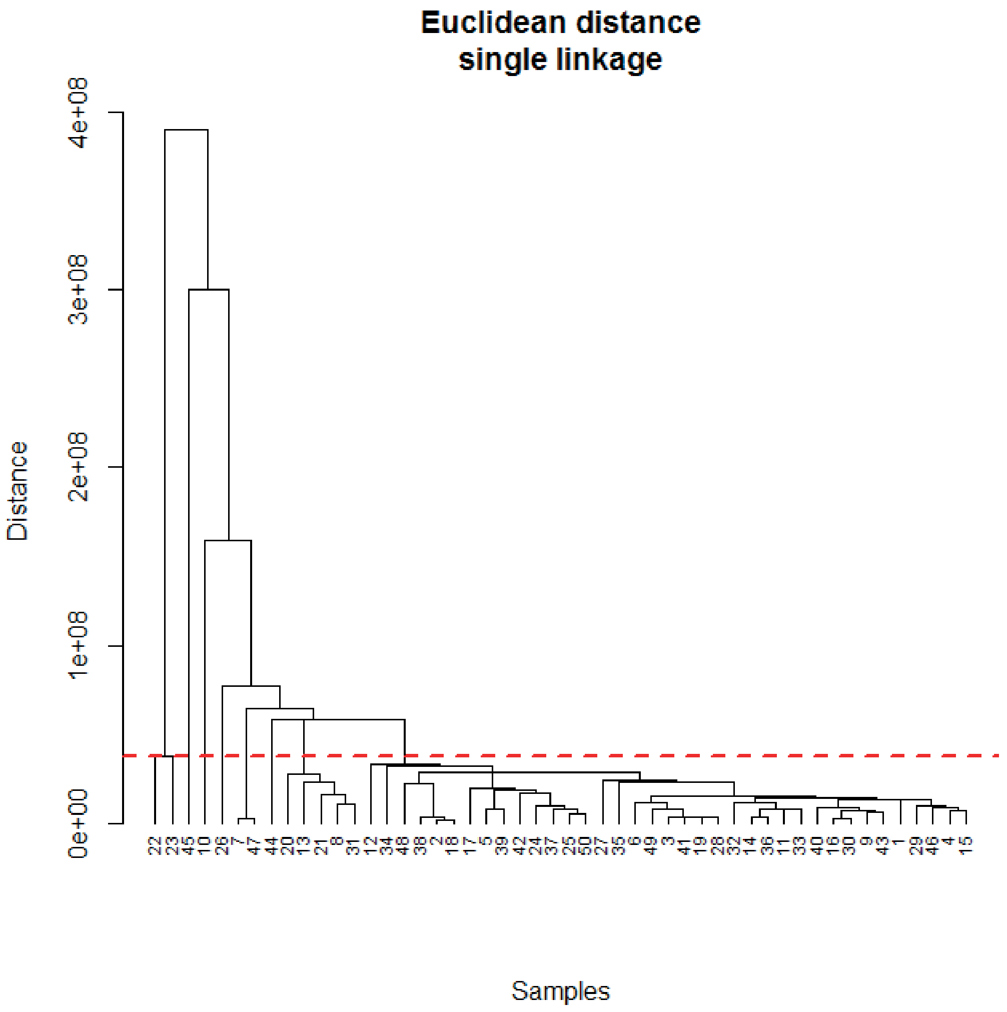


Figure 7.12 Dendrogram shown in Figure 7.8 with an added line at the distance where the splitting into smaller clusters become frequent

Reasoning in the same way for the dendrograms in Figures 7.9-7.11 the estimated number of laboratories producing amphetamine is from 8 to 10.

When setting the decision limit to HCA it needs to be validated, i.e. which distance (illustrated as red line in Figure 7.12) gives the picture closest to the truth of different batches.

8 WHICH DATA PRE-PROCESSING AND ANALYSIS TO USE FOR DIFFERENT KIND OF QUESTIONS?

The question may be asked which data pre-processing and data analysis methods are fit to answer our specific forensic questions, namely for identification, quantification, classification and comparison. This topic is not easily covered but some pointers can be given.

With respect to question on identification and classification, usually discriminant analyses methods will be used, such as LDA, Logistic Regression or PLS-DA as described above. Pre-processing by normalization or standardization usually is a sound idea.

With respect to the question about quantification usually regression-based methods such as OLS-R can be used.

With respect to comparison-based questions, normalization and/or weighing is often beneficial in order to get values on the same scale. Comparison may be based on the analysis of dissimilarities between samples.

9 METHOD VALIDATION

9.1 Validation of the applied method before going into practice or casework

In order to apply any of the chemometric methods described in casework, they should first be validated. Chemometric methods will produce different types of results depending on the type of question posed and the results need to be evaluated appropriately. For the purpose of identification, classification and comparison, the performance of the method can be expressed in terms of *error rates*, such as false positive and false negative rate. These rates express how often the method arrives at an erroneous conclusion and they can be assessed based on applying the method on test data.

For identification, a false positive means that the method falsely states that a substance is present, while false negative means that the method fails in identifying a present substance. For classification, a false positive corresponds to falsely concluding an item belonging to a class and a false negative is a failure to conclude class membership. For comparison, a false positive means erroneously concluding that two items are connected, while a false negative means failure to connect two items. Performance of identification and comparison system typically take place through Receiver Operating Characteristic (ROC) curves, see chapter 9.2. Performance of classification systems are often based on error rates such as will be described in chapter 9.3.

For quantification, like the determination of the concentration of the effective substance in a sample, statistical models can be produced to predict quantities from given input variables. A simple way to evaluate such models is by assessing a combined measurement uncertainty that covers the inhomogeneity of the sample as well as variation coming from the sample preparation and measurement process. This can be done by calculating the bias and standard deviation of the residuals from the prediction given by the statistical model.

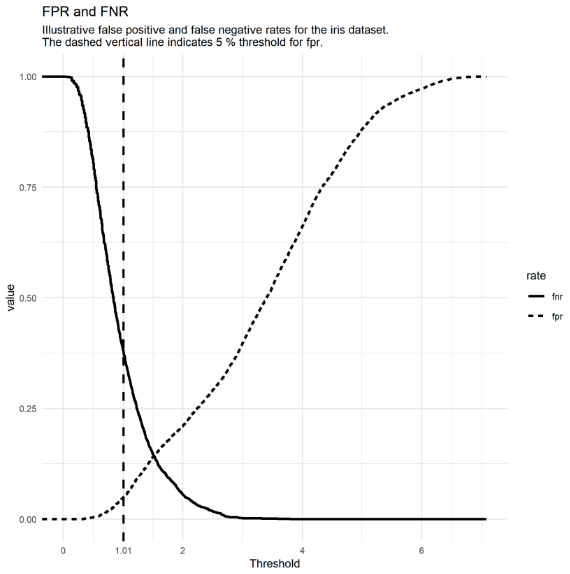
In cases of e.g. comparison or quantification, the selected chemometric method needs a prior training stage. In practice, the dataset available may be divided into a training and a testing datasets. As a rule of thumb, 80% of the data can be used as training data and 20% for testing data. The choice of the training set should encompass the whole variety of the data (samples). It is important that data covers the normally observed variation and that the data is from samples with prior knowledge. This kind of data is needed for training and testing datasets.

9.2 ROC curves

When a chemometric method depends on the selection of a decision threshold, Receiver Operating Characteristic (ROC) curves can be used to evaluate the overall performance of the method. A specific decision threshold can be determined by examining false positive and false negative rate curves plotted against possible thresholds in order to choose a threshold that offers a desired level of performance for that method.

9.2.1 How to create ROC curves?

An illustration of a false positive and a false negative rate curve and the corresponding ROC curve is given in Figure 9.1. Suppose that there is a classification system where a higher outcome makes it more plausible that the answer to the question (e.g. two tablets are made with the same type of tableting machine) is positive. Then for each possible outcome *s* of the system it is determined what the fraction of outcomes in the test set for which the real answer is positive and for which the outcome of the system is <*s,* this is the *false negative rate*. The same can be done for the part of the test set that in reality is negative, but has an outcome of the system that is >*s*. This is the *false positive rate*. In the ROC curve, the true positive rate (=1-false negative rate) is plotted against the false positive rate. The Area Under Curve (AUC) can be used as a metric for method performance.



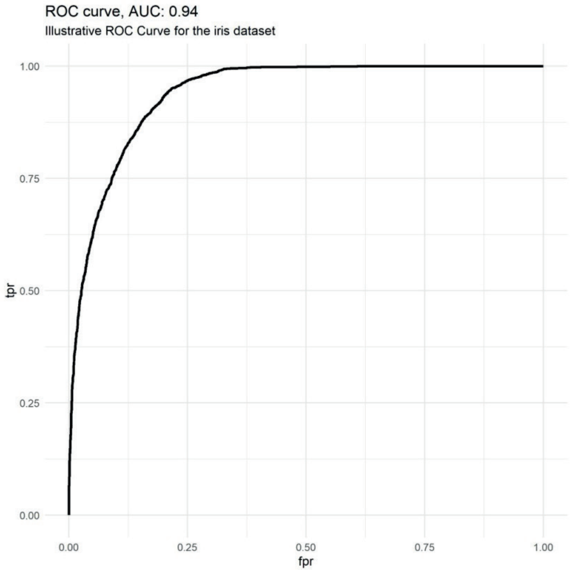


Figure 9.1. An illustration of a false positive and a false negative rate curve (with the threshold of 1.01 corresponds to 5% false positive rate for identifying pairs belonging to the same class), and the corresponding ROC curve. These are based on pairwise Euclidean distances over the so-called 'iris' dataset available with e.g. the base package of the R programming language.

9.2.2 What information can be taken from ROC curves?

ROC curves depict the performance of systems in which a choice is made between two possible scenarios, for example whether a substance is identified, or whether two samples are from the same source or not. Systems like these are usually based on the determination of similarity or distance measures, and in the case of comparisons, the test data set is used to gather these similarities/distances between samples of the same source and samples of different sources. On the basis of these distances, performance of this system is expressed by (FP/FN & TP/TN) a ROC curve, as shown in Figure 9.1. This curve presents the proportion of true positives as a function of the proportion of false positives and illustrates the performance of the system. The closer the curve is to the upper left corner, the better the performance of chemometric method is at predicting correctly whether two samples come from the same source or not. A diagonal curve would imply that the chemometric method adds no information and predictions based on it would be as good as random guessing. The selection of a threshold always involves a tradeoff between false positive and false negative rates.

9.3 Performance of classification systems: confusion matrices

When the chemometric method involves classification into more than 2 classes, ROC curves are typically not applicable as such for performance evaluation of the method. In this case the method can be readily evaluated using confusion matrices, overall accuracy or so-called one-versus-all error rates.

A confusion matrix is a cross-tabulation of the predictions of the classification method against true classes. In Table 9.1, an example is given using the iris dataset and classes given by linear discriminant analysis. The rows correspond to the predictions of the algorithm while the columns correspond to the true classes. The diagonal of the matrix tells how many times the predicted class matched the true class while off diagonal values give different errors. For example, the first row tells how many times 'setosa', 'versicolor' and 'virginica' classes, respectively, were predicted as belonging to the 'setosa' class. Here, no mistakes were made, suggesting good separation between 'setosa' and the other classes. On the second row it can be seen that one 'virginica' class sample was predicted to belong to 'versicolor' class and, on the third row, that two 'versicolor' class samples were predicted as belonging to the 'virginica' class. This suggests that the 'versicolor' and 'virginica' classes are not completely separable by the algorithm. A confusion matrix can easily be extended to any number of classes.

Table 9.1: Confusion matrix based on classification results on the iris-dataset using linear discriminant analysis

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	1
virginica	0	2	49

When there are more than two classes, there is no obvious meaning to terms such as 'positive' or 'negative' cases and as such it makes no sense to speak of false positives and false negatives. However, it is still possible to speak of accuracy in terms of how many times the prediction made by the classification method was correct. Continuing the example using iris dataset with linear discriminant analysis, the accuracy of the algorithm could be calculated as the sum of the diagonal element (50 + 48 + 49 = 147) divided by the number of total cases (150) resulting in the accuracy of 147 / 150 = 0.98. That is, the algorithm has 98% accuracy and, on this dataset, made the wrong classification 2% of the time.

It is possible to use similar methods as with binary classification with multiple classes by considering one class at the time. In this approach each class in turn is considered to be the 'positive' class while all the other classes are considered 'negative', allowing calculation of, for example, false positive and negative rates from the perspective of each class separately. The results for the iris dataset are given in Table 9.2. Here each row corresponds to the class that is considered to represent the 'positive' case in that case. For the 'setosa' class, no errors are made. For the 'versicolor' class, the false positive rate (FPR) is 0.01, telling that 1% of all non-versicolor samples were falsely predicted to be 'versicolor', and the false negative rate (FNR) is 0.04 indicating that 4% of all versicolor samples were wrongly classified into some other class. For the class 'virginica', we see that the corresponding error rates are 2% for both FPR and FNR.

Table 9.2: One-vs-all false positive and false negative rates for each class in the iris dataset using linear discriminant analysis

Class	FPR	FNR
setosa	0	0
versicolor	0.01	0.04
virginica	0.02	0.02

10 ASSESSMENT / INTERPRETATION OF RESULTS

Whenever the methods that have been developed are applied in practice / casework, the question may be asked: “How should the assessment and interpretation of the chemometric results take place?”

The assessment may take place on two levels, as described below,
as well as in article 3 [3].

The first level is an operational assessment that evaluates the performance of the chemometric method (operational level). With the given relevant performance criteria, identification, classification or comparison may take place, in order to make a decision about samples being from a certain category or not. These predefined values may be taken from various points of view, e.g. they are fixed by the law, derived by scientific arguments, described in the literature or agreed upon with the ‘customer’. In the latter case they may change according to the needs of the customer (e.g. police intelligence may ask for less restrictive values than the court).

The second level is a chemical assessment of the chemometric result. Given the original chemical data, the forensic chemist evaluates whether the chemometric results make sense according to the chemical properties, and as such functions as a ‘safety net’.

In the case that a decision criteria is used that is based on a threshold value – accepting a certain false positive and false negative rate - reporting might take place in terms of “based on the similarity value between the samples, and the chemical assessment of the forensic chemist, it is concluded that they are from the same batch”, or similar alternatives. The communication can be given straight, comparable to a technical report, where a substance is identified according to given criteria.

Depending on the kind of question asked, for example when answering the question requires evaluative reporting, the interpretation may need further steps like expressing a likelihood ratio. The ChemoRe software is not designed to directly express results for evaluative reporting. There are other tools like SaiLR [33] that have been designed for this, so further details on expression an evaluative interpretation and communication of these types of results are out of the scope of this guideline.

11 EXAMPLES OF CHEMOMETRIC METHODS USED IN FORENSIC CHEMISTRY RELATED TO CASEWORK OF ILLICIT DRUGS

In Chapter 3 the forensic workflow and the relation between different phases in forensic casework involving illicit drugs is explained. Herein it is described where chemometric methods are located. This chapter describes by using three examples how chemometrics may be applied. The examples cover Type 1 and Type 2 data (low- and high dimensional), which are used for classification, comparison, identification, and quantification. For all examples a stepwise description will be given of

- The forensic question under consideration
- The available background data for training and testing purposes
- The data pre-processing taking place
- The chemometric analysis involved
- The performance of the procedure given the test data

The data pre-processing and analysis has been performed by applying the ChemRe software from that the relevant results are visualised and exported.

11.1 Example 1. XTC tablets: case-to-case comparison by external characteristics

Case history

In police operation some XTC tablets were seized during a police investigation in addition to a tableting machine. The police were interested in whether the seized tablets and tablets seized earlier were made using this machine.

Tablets, as final products, can be characterized and linked to a certain type of tableting machine by their physical dimensions [34].

Forensic Question

The following forensic question has to be answered: “Are the seized tablets from the same (or similar) tabletting machine?”

Selection of Parameters

The forensic laboratory identified MDMA in the seized tablets by routine analytical methods and decided to use several physical parameters, namely diameter, thickness and weight of the XTC tablets for chemometric evaluation. This data is relevant when the forensic laboratory is answering to the question whether the seized tablets were stamped with the seized or a similar tableting machine or not. A similar tableting machine would mean a machine that has comparable molds and punches as the seized machine.

Classification of Parameters

Example 1 contains Type 1 data with a training set. Data pre-processing (transformation) and data analysis are performed to answer a case to case comparison question.

Training and Test Data

The origin of the data used in this example is the CHAMP-project database [13, 32] where XTC- tablets are grouped in batches (160 batches in total). From this database, measurements corresponding to 2 batches were removed to serve as the origin of the seized items while the remaining 158 batches in the database were split into training and test datasets with 80% and 20% of the data included in each set respectively. The two removed batches were further divided in four separate groups that represent individual batches to serve as simulated case data.

The physical measurements of the tablets are presented in Table 11.1. For illustration purposes, information about color and logo of the tablets is included additionally (Figure 11.1. Smiley tablet).

Table 11.1: Physical measurements (in millimeters and milligrams) and additional information on 4 batches of ecstasy tablets. Here Sample ID is a code used to identify each tablet. Prefix ‘S’ refers to tablets that are from the seized machine while prefix ‘Q’ refers to tablets found later, that is ‘questioned’ tablets

Sample ID	Batch	Diameter	Thickness	Weight	Logo	Color
S1	1	8,1	3,5	198	Smiley	Yellow
S2	1	8	3,5	199	Smiley	Yellow
Q1	2	8	3,4	198	Smiley	Yellow
Q2	2	8,1	3,5	196	Smiley	Yellow
Q3	3	8	3,5	198	Smiley	Red
Q4	3	8	3,5	196	Smiley	Red
Q5	4	9,1	3,6	301	Smiley	Yellow
Q6	4	9	3,5	291	Smiley	Yellow

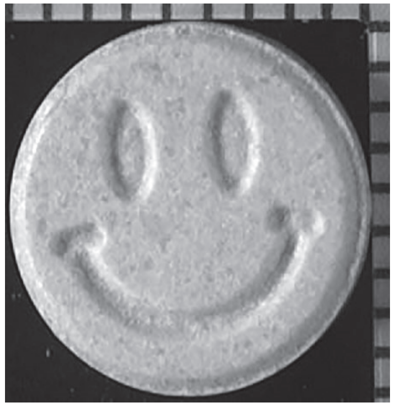
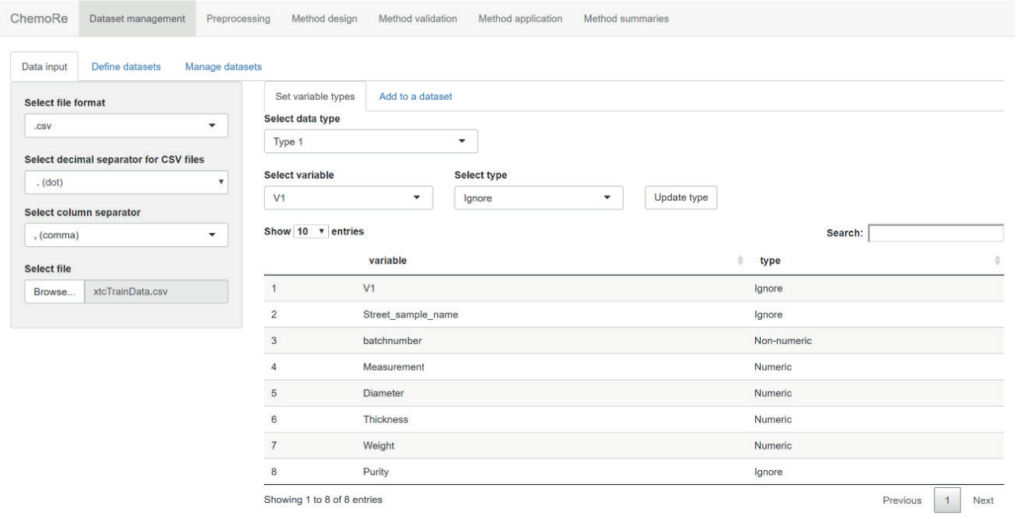


Figure 11.1 Example logo (smiley) on a seized ecstasy tablet

Data import

After removing the two batches determined as origins of the requested tablets, the training dataset composed of 80% of the remaining 158 batches was imported into the ChemoRe software in form of a *.csv file (Figure 11.2).



Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select file: xtcTrainData.csv

Select data type: Type 1

Select variable: V1

Select type: Ignore

Show 10 entries

Search:

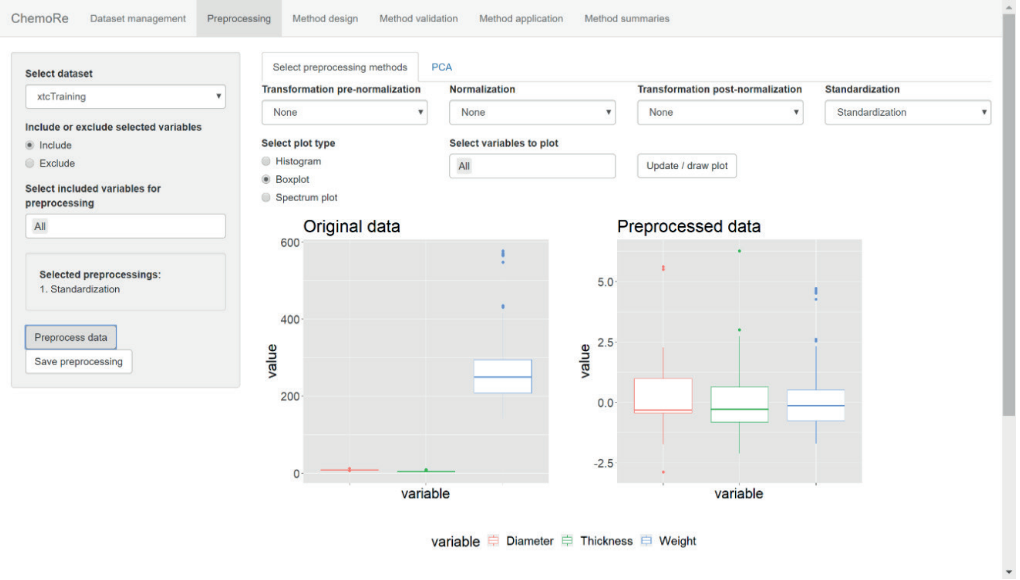
	variable	type
1	V1	Ignore
2	Street_sample_name	Ignore
3	batchnumber	Non-numeric
4	Measurement	Numeric
5	Diameter	Numeric
6	Thickness	Numeric
7	Weight	Numeric
8	Purity	Ignore

Showing 1 to 8 of 8 entries

Figure 11.2 Data import of physical parameters of the seized tablets (type 1, numeric)

Data pre-processing

As the variables of the data were measured on different scales it was necessary to transform them to z-scores (Figure 11.3). This required calculating the means and standard deviations of each variable. The z-scores are obtained by subtraction of the mean and division by the standard deviation for each variable in the data.



Select dataset: xtcTraining

Include or exclude selected variables: ● Include ○ Exclude

Select included variables for preprocessing: All

Selected preprocessings:
1. Standardization

Transformation pre-normalization: None

Normalization: None

Transformation post-normalization: None

Standardization: Standardization

Select plot type: ○ Histogram ● Boxplot ○ Spectrum plot

Select variables to plot: All

Figure 11.3 Effect of data pre-processing by standardization (z-score transformation)

Chemometrics - Method development

In order to compare the transformed physical parameters of XTC-tablets pairwise to each other, a measure of dissimilarity was selected, together with a threshold for concluding whether two tablets originate from the same batch. The Euclidean distance was chosen for this purpose, though other measures could have been used likewise. Euclidean distance as a simple distance measure seems to be reasonable for a first attempt. To determine a threshold, the measure was evaluated on a test dataset.

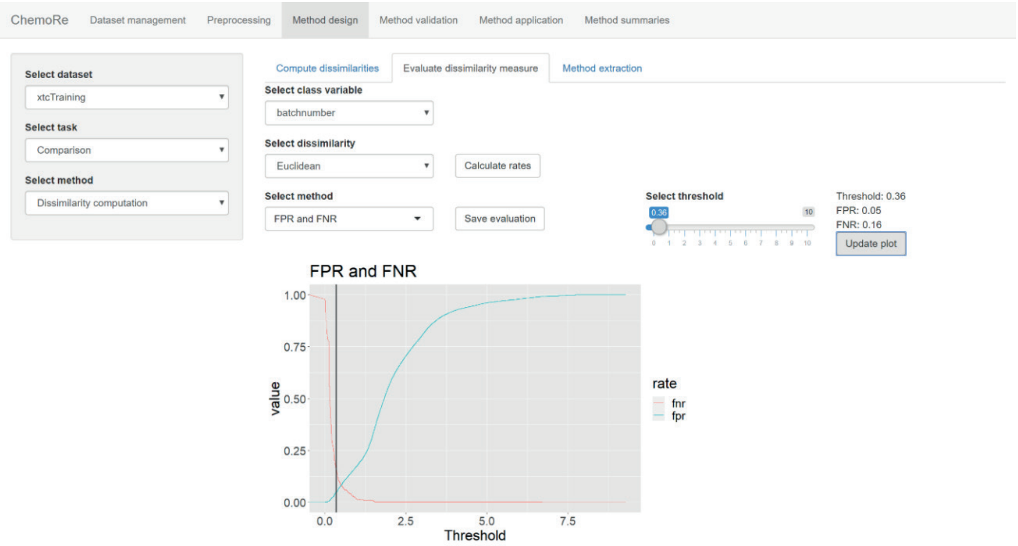
The calculated means and standard deviations depend on the observed data and therefore are determined on the training dataset, as explained above. For this the database reserved for method development (486 tablets) was split randomly into a training dataset (80% of the batches; 391 tablets) and a test dataset (20% of the batches; 95 tablets).

After transformation of the data, the Euclidean distance was calculated between all samples in the test set. The obtained distance values were used to determine whether a pair originates from the same source, referred to as a positive case, or from different sources, referred to as a negative case. This determination is done by comparing the distance value against a chosen threshold (same source if the value is below the threshold and different sources if the value is above the threshold).

Based on these distances a graph is given showing the false positive rate versus the false negative rate. Further the performance of this system is expressed by a ROC curve. A decision threshold was determined to achieve an acceptable rate of error assisted by both graphs.

Result:

In Figure 11.4 it is shown how the false positive and false negative rates on the test database develop as a function of all possible threshold values for the dissimilarity of pre-processed variables. The general performance of this method for the purpose of comparison can be seen in the ROC curve (Figure 11.5). In order to choose a decision threshold, an acceptable error rate needs to be chosen. Assisted by Figure 11.4 (fpr and fnr) a 5% false positive rate is deemed acceptable, based on the test dataset, the threshold should be set at 0.31 for the Euclidean distance. This corresponds to a false negative rate of 17%.



Select dataset: xtcTraining

Select task: Comparison

Select method: Dissimilarity computation

Select class variable: batchnumber

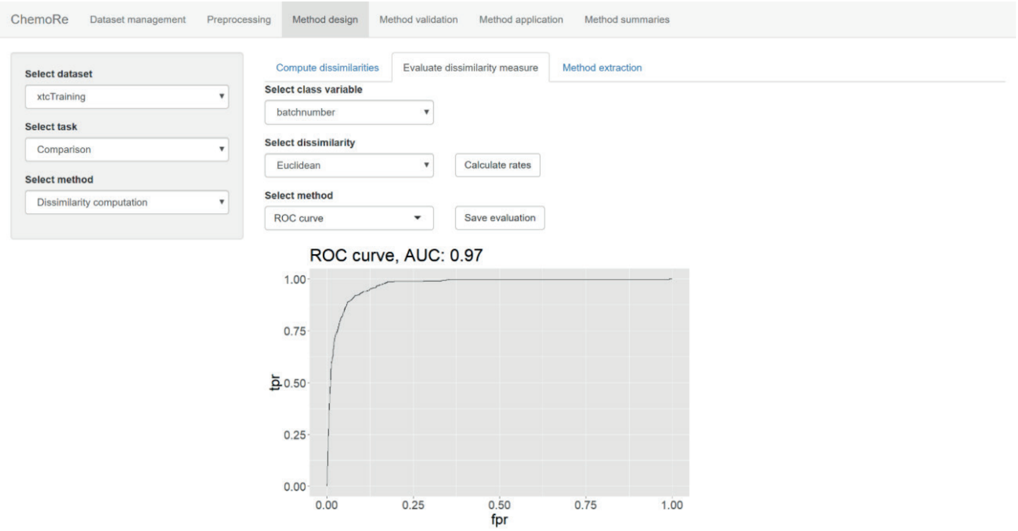
Select dissimilarity: Euclidean

Select method: FPR and FNR

Select threshold: 0.36 (10)

Threshold: 0.36
FPR: 0.05
FNR: 0.16

Figure 11.4 *Illustration of the false positive and false negative rates on the test database*



Select dataset: xtcTraining

Select task: Comparison

Select method: Dissimilarity computation

Select class variable: batchnumber

Select dissimilarity: Euclidean

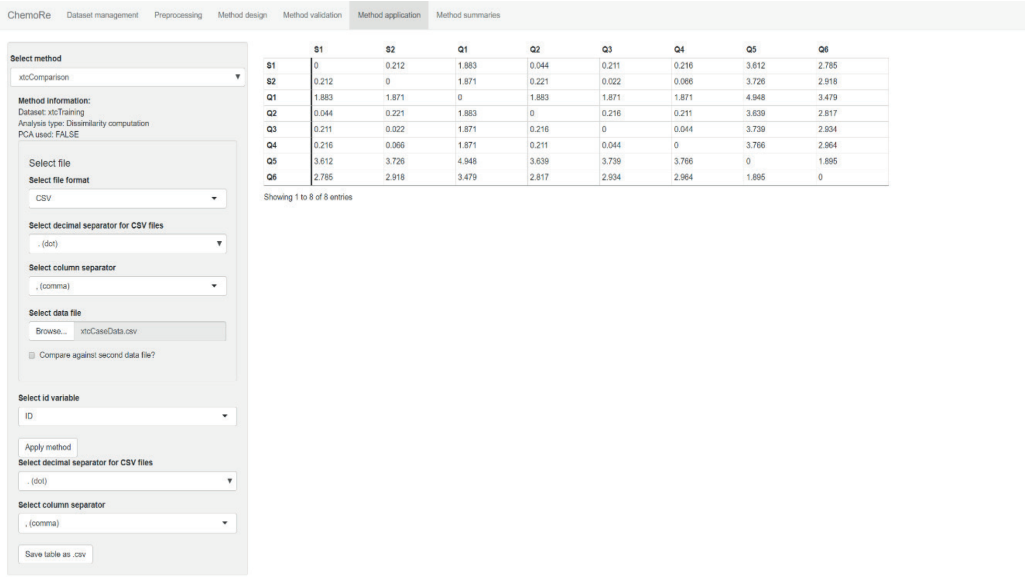
Select method: ROC curve

Figure 11.5 *ROC curve of the method, depicting the true positive rate as a function of the false positive rate*

We turn to the results in the actual case-to-case comparison. In Figure 11.6 dissimilarity values are shown for all pair-wise comparisons of the tablet batch under consideration.

Euclidean distance is below the 0.31 threshold indicate a connection between tablets from the seized machine to the questioned tablets seized later on. Whether a Euclidean distance above the 0.31 threshold does not indicate such a link.

From here we can see that tablets from the seized machine, titled S1 and S2 in the Figure, are indicated to have the same source as questioned tablets Q1-4, whereas according to the Euclidean distance the questioned tablets Q5 and Q6 are indicated to have a different source. As the seized tablets are colored yellow but tablets Q3 and Q4 are colored red, it seems likely that the production batch of these tablets was different, even if the tableting machine used was the same. Additionally, tablets Q5 and Q6 share the color with tablets S1-2, but have different dimensions. It is possible they are made from same raw material, but using different tableting machines.



Select method: xtcComparison

Method information:
Dataset: xtcTraining
Analysis type: Dissimilarity computation
PCA used: FALSE

Select file

Select file format: CSV

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select data file: xtcCaseData.csv

□ Compare against second data file?

Select id variable: ID

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

	S1	S2	Q1	Q2	Q3	Q4	Q5	Q6
S1	0	0.212	1.883	0.044	0.211	0.216	3.612	2.785
S2	0.212	0	1.871	0.221	0.022	0.066	3.726	2.918
Q1	1.883	1.871	0	1.883	1.871	1.871	4.948	3.479
Q2	0.044	0.221	1.883	0	0.216	0.211	3.639	2.817
Q3	0.211	0.022	1.871	0.216	0	0.044	3.739	2.934
Q4	0.216	0.066	1.871	0.211	0.044	0	3.766	2.964
Q5	3.612	3.726	4.948	3.639	3.739	3.766	0	1.895
Q6	2.785	2.918	3.479	2.817	2.934	2.964	1.895	0

Showing 1 to 8 of 8 entries

Figure 11.6 Dissimilarity values for all pair-wise comparisons of the tablet batches under consideration considering a threshold value of 0.31

Note that there is a possibility of misclassification when two similar tableting machines are used independently. This can be taken into account when interpreting the results in a forensic report by a statement such as: “the tablets are made with the same or a similar machine”. Provided there are prescribed limits for the rates of misclassification, the report can be considered as factual. The meaning of the statement is that the Euclidean distance between the tablets is below the chosen threshold. If no limits are prescribed it may be more advisable to write an evaluative report.

11.2 Example 2. Amphetamine: case-to-case and database comparison

Case history

An investigation of an amphetamine distribution network was performed in a medium size town where an individual was suspected to act as a local sales person and 4 amphetamine samples were recovered from his apartment. Following this, several seizures were made within a short period of time from various users resulting in 7 amphetamine samples. Additionally, during the same time frame, a package containing a further amphetamine sample was found by a hiker in a nearby forest.

Forensic Question

Now, the investigative police had two questions. Firstly, the police wanted to determine whether the samples found from users had the same source as the samples confiscated from the suspect's apartment, amounting to a *case-to-case comparison*. Secondly, the police wished to know if the sample found by the hiker was related to amphetamine seen earlier in the area, including what was found at the suspect's home, constituting a *database comparison*.

Selection of Parameters

Profiling analysis was done according to the questions presented by the investigative police. Amphetamine can be produced by several synthetic routes [4]. The most common routes used are the Leuckart method, Reductive amination and the Nitrostyrene method. The different synthetic routes produce different chemical impurities to the final product and need to be identified accordingly. A chemical profiling analysis was performed using the harmonized profiling method (European Harmonized Method for the Profiling of Amphetamine, EHMPA) that is based on the analysis of chemical impurities originating from synthetic process [31]. The chemical analysis was done from homogenized samples by a GC-MS instrument after sample pre-processing. An example chromatogram of a GC-MS measurement is presented in Figure 11.7, the exact analytical method is described in literature [35-37]. A segment of the raw data of the measured impurities is presented in Table 11.2 and contains the integrated peak areas of 26 target compounds (impurities).

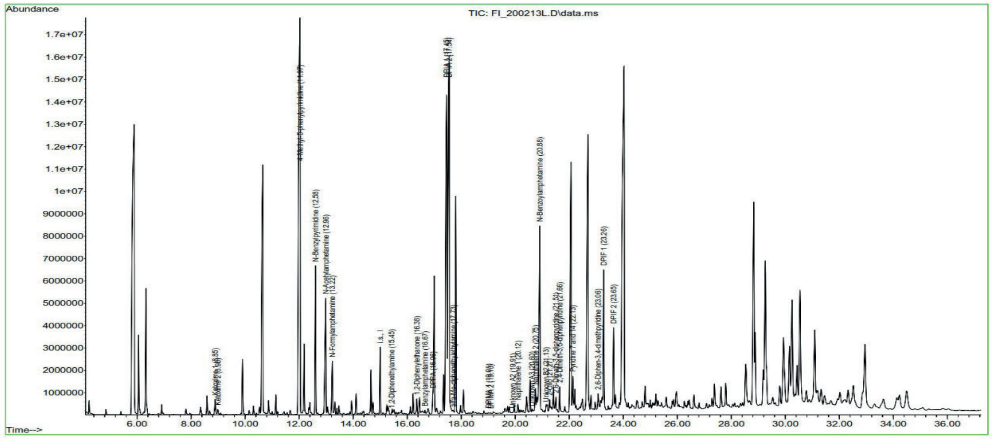


Figure 11.7 Chromatogram of amphetamine data analyzed by GC-MS.

Table 11.2: Example of the raw data (peak areas) of amphetamine profiling analysis. The data of the example is available upon request

	Sample 1	Sample 2	...	Sample 13
V1[b]	4092	186233		400 *
V2	200 [a]	200 [a]		200 [a]
V3	77351	171272		4704
V4	1368282	1126129		18314
V5	200 [a]	55490		200 [a]
V6	2671	10950		200 [a]
V7	19103	94993		200 [a]
V8	57915	428840		4549
V9	81117	331794		200
...	...	...		...
V26	259602	2155433		33555

a. *Value of 200 is added to replace a missing value or a value which is under 1% of the measured peak area of internal standard.*

b. *Variables V1 and V10 are sum values of two isomers*

Classification of Parameters

Example 2 contains Type 1 data without a training set; the data transformation (pre-processing) and data analysis are performed to answer comparison and clustering questions.

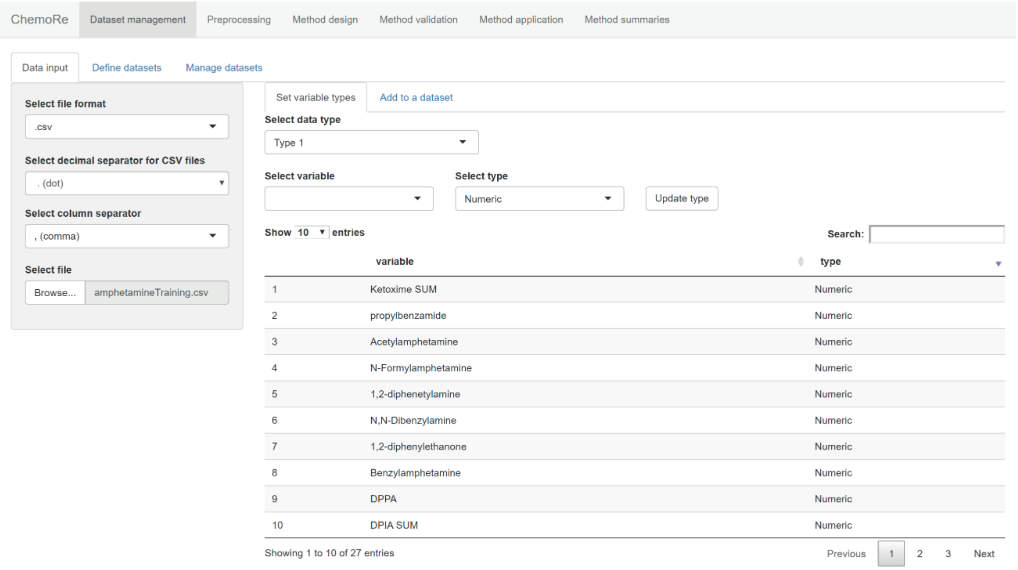
Training and Testing Data

The raw data presented is an example of Type 1 data, consisting of a limited number of peaks. In this example data was extracted from the amphetamine profiling database of the Finnish National Bureau of Investigation - Forensic Laboratory, for the purpose of training and testing a comparison method. The data was collected with the criterion that the origin of samples was known. It should be noted that the origin is 'known' based on forensic analysis and police information. While this is not optimal for the development and validation of a chemometric method, it is sufficient for the purposes of this example and represents a typical forensic situation. The data consists of impurity profiles of amphetamine samples, comprising (as mentioned above) 26 chromatographic peak areas of selected target compounds that correspond to impurities resulting from the manufacturing process. Answering the question whether an amphetamine seen earlier in the area, a comparison using a database have to be performed.

In order to find linkages between all seizures made before or seizures made in a certain time interval, the questioned data sets can be compared to data of the whole database or the time interval of interest by applying the same method.

Data import

The data sets have been imported into the ChemoRe software in form of a *.csv-file (Figure 11.8).



Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select file: amphetamineTraining.csv

Select data type: Type 1

Select variable:

Select type: Numeric

Show 10 entries

Search:

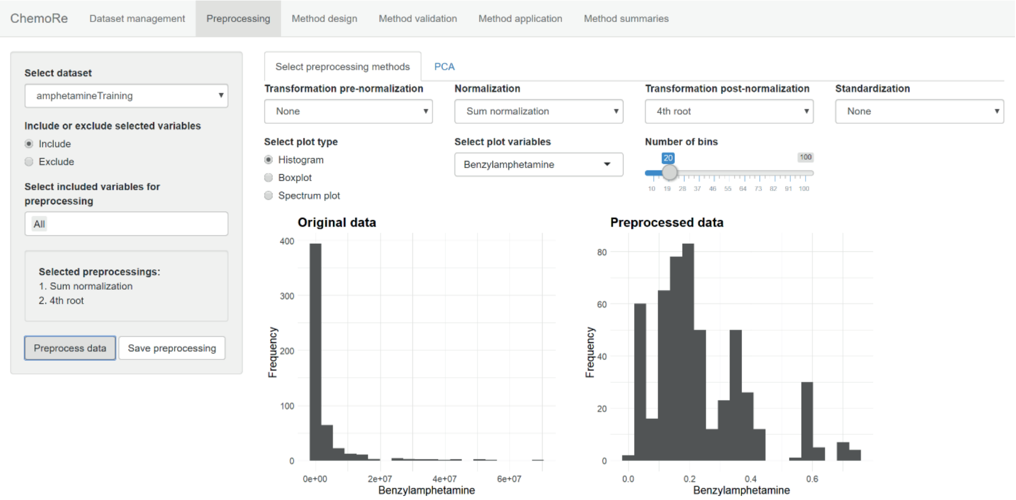
	variable	type
1	Ketoxime SUM	Numeric
2	propylbenzamide	Numeric
3	Acetylamphetamine	Numeric
4	N-Formylamphetamine	Numeric
5	1,2-diphenetylamine	Numeric
6	N,N-Dibenzylamine	Numeric
7	1,2-diphenylethanone	Numeric
8	Benzylamphetamine	Numeric
9	DPPA	Numeric
10	DPIA SUM	Numeric

Showing 1 to 10 of 27 entries

Figure 11.8 Import of impurity profiles into ChemoRe. Each Dataset consists of 26 chromatographic peak areas (type 1 data; numeric values)

Data pre-processing

Due to differing concentrations of amphetamine in samples, the measured peak areas can be on different scales between samples. This makes it sensible to transform the data in each sample to the same scale. Additionally, certain impurities may typically be present in higher quantities in one sample than in others. Therefore, it is a good idea to further transform the normalized values to make them more comparable in terms of importance. This normalization is achieved by dividing each profile by the sum of the peak areas in that profile which forces all measurements to the range from 0 to 1. This ensures that all impurity profiles are on the same scale between samples. Further data pre-processing is then performed by calculation of the fourth root of each normalized peak area [31, 37]. Calculating logarithms or square roots are alternative methods for transforming the data (also referred to as 'weighing' in the forensic literature). The effect of data pre-processing on original peak area of one target compound and on an original chromatogram represented by all 26 components is shown in Figures 11.9 and 11.10.



Select dataset: amphetamineTraining

Include or exclude selected variables
● Include
○ Exclude

Select included variables for preprocessing: All

Selected preprocessings:
1. Sum normalization
2. 4th root

Transformation pre-normalization: None
Normalization: Sum normalization
Transformation post-normalization: 4th root
Standardization: None

Select plot type
● Histogram
○ Boxplot
○ Spectrum plot

Select plot variables: Benzylamphetamine

Number of bins: 20 (100)

Figure 11.9 Effect of data pre-processing applied to one component (Benzylamphetamine) by Sum normalization and 4th root

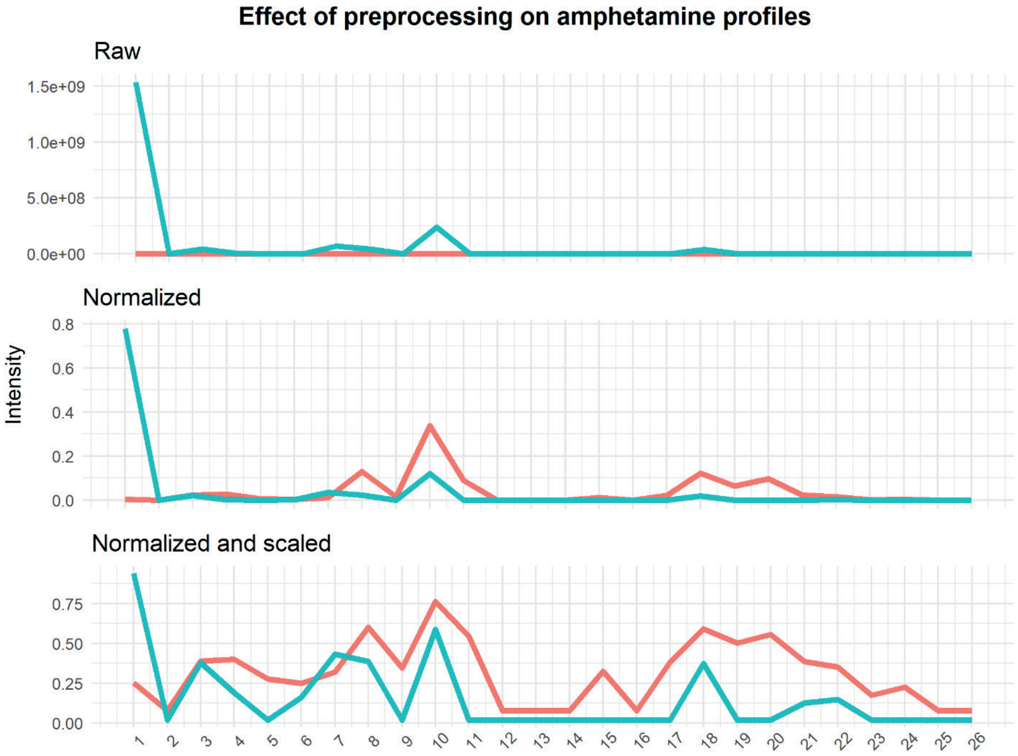


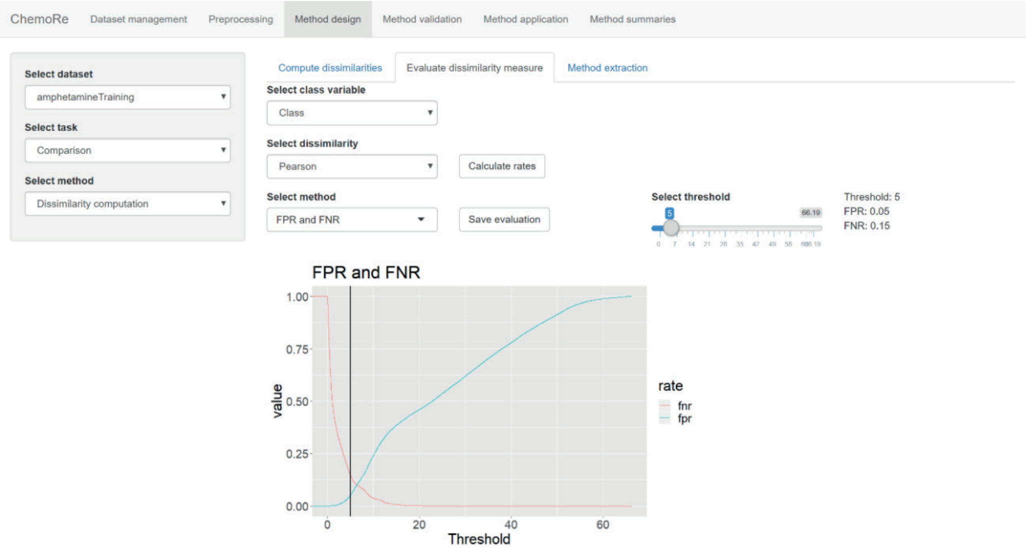
Figure 11.10 Illustration of the effect on the variables (x-axis) of data pre-processing (normalization and fourth root)

Chemometrics - Method development

To answer the presented questions of case-to-case and database comparison, comparisons were performed on sets of amphetamine samples. To this end, a measure of dissimilarity between the pre-processed profiles was used, describing to what extent two impurity profiles differ. Pearson distance is chosen here, following recommendations in the EHMPA method [31]. Intuitively, this distance measures the similarity between the shapes of the profiles rather than absolute differences between profiles which has previously been considered advantageous. As an alternative other dissimilarity measures could be used, including the Euclidean and Manhattan distance. Based on the Pearson distances observed between samples of the same source and of different sources, the comparison method is evaluated similarly to example 1 by producing a ROC curve and corresponding false positive and false negative rate curves. Based on these results, a decision threshold is determined. The threshold involves a certain (fixed) error rate that is evaluated using the training dataset. Pairwise comparison of the confiscated amphetamine samples is performed according to this threshold.

Results:

The effects of normalization and scaling on the amphetamine impurity profiles are visualized by a graph with the pre-processed peak area on the ordinate and the component number on the horizontal axis. Examples of two profiles and the effect of the pre-processing on them are visualized in Figure 11.10. After pre-processing, the Pearson distances were calculated for each pair in the test dataset. For readability the distances were scaled to be in range from 0 to 100. Based on these distances, the false positive and false negative rates were calculated and these are shown in Figure 11.11. Based on these results the decision threshold was set at 5.07[1] as this produced 5% false positive rate (0.05 on the y-axis, intersection with green line) on the test dataset and 14% false negative rate (0.14 on the y-axis, intersection with red line). The overall performance of the method can be seen in the ROC curve in Figure 11.12. Given the threshold that was chosen, the method was applied to the case data with results for the samples seized from the users and from the suspect.

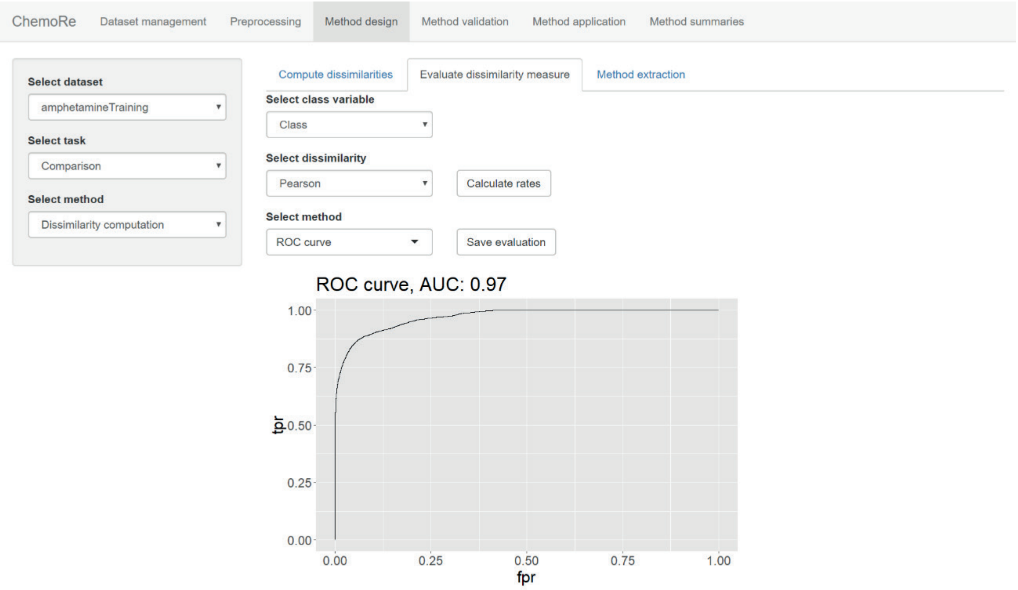


Select dataset: amphetamineTraining
Select task: Comparison
Select method: Dissimilarity computation
Select class variable: Class
Select dissimilarity: Pearson
Select method: FPR and FNR
Select threshold: 5 (66.19)

Threshold: 5
FPR: 0.05
FNR: 0.15

Figure 11.11 Illustration in ChemoRe of the false positive and false negative rates on the test database

[1] Pearson distances were multiplied by factor 100 due to a better visualization and readability.

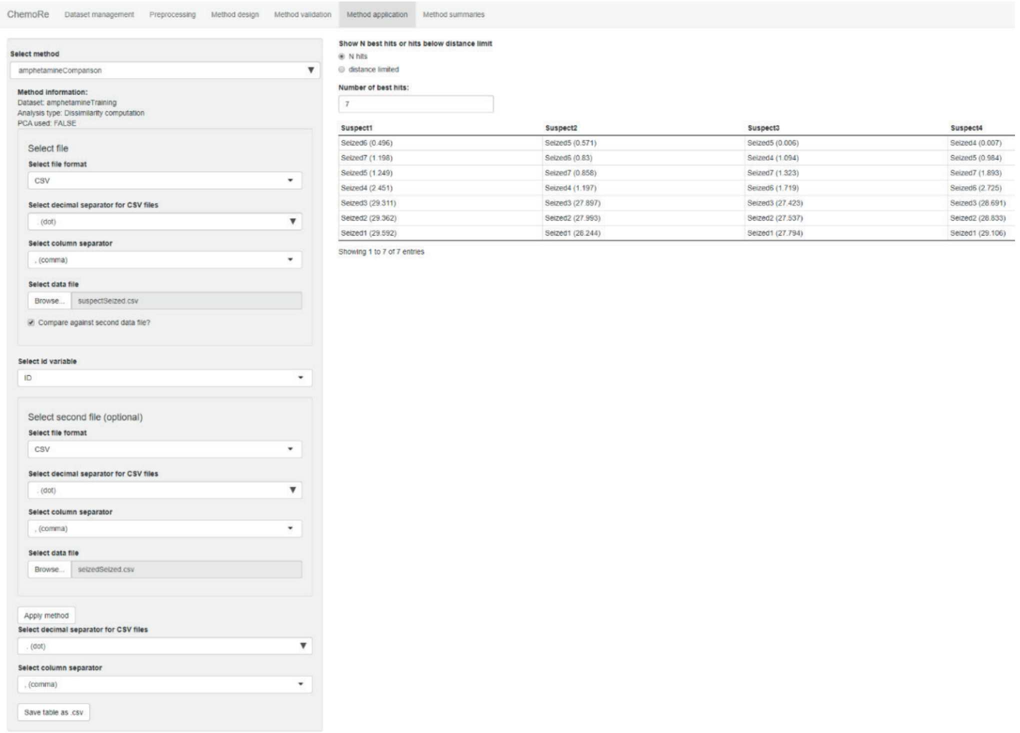


Select dataset: amphetamineTraining
Select task: Comparison
Select method: Dissimilarity computation
Select class variable: Class
Select dissimilarity: Pearson
Select method: ROC curve

Figure 11.12 The ROC curve for the amphetamine comparison method based on Pearson correlation distance. This curve together with the Area Under Curve (AUC) statistic provide information on the general performance of the method

The findings of the profiling study were as follows:

Four of the seven seized amphetamine samples were found linked to the 4 samples from the suspected sales person (Figure 11.13). Thus, the result of profiling analysis supports the investigative theory of a local distributor (case-to-case). The other 3 seized samples were linked to each other but showed a different type of impurity profile than the remaining 4 seized samples (Figure 11.14).



Select method: amphetamineComparison

Method information:
Dataset: amphetamineTraining
Analysis type: Dissimilarity computation
PCA used: FALSE

Select file format: CSV
Select decimal separator for CSV files: . (dot)
Select column separator: , (comma)
Select data file: suspectSeized.csv
☑ Compare against second data file?

Select id variable: ID

Select second file (optional)
Select file format: CSV
Select decimal separator for CSV files: . (dot)
Select column separator: , (comma)
Select data file: seizedSeized.csv

Select decimal separator for CSV files: . (dot)
Select column separator: , (comma)

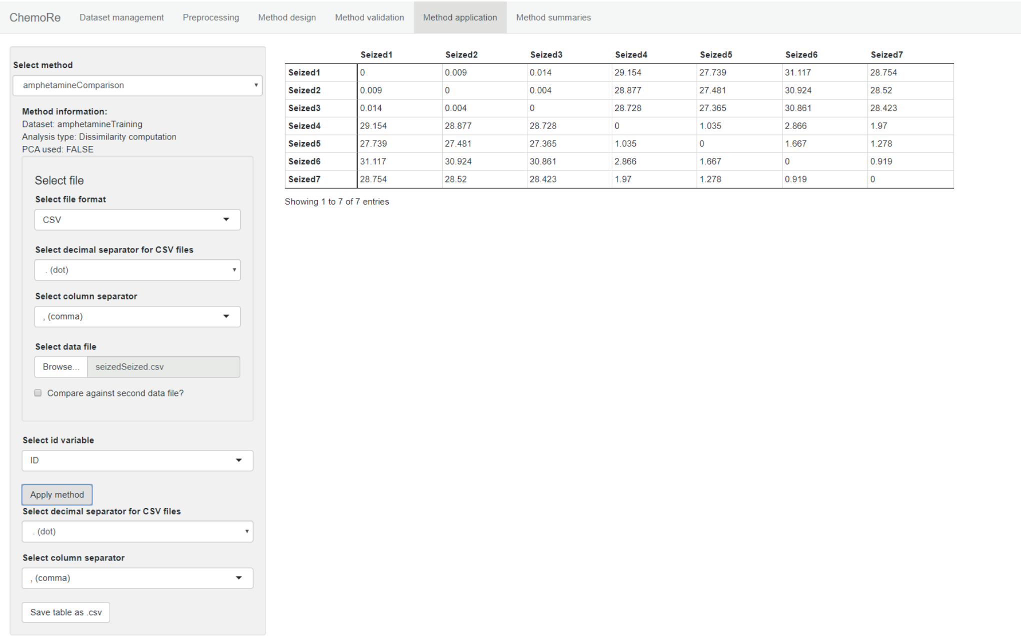
Show N best hits or hits below distance limit
● N hits
○ distance limited

Number of best hits: 7

Suspect1	Suspect2	Suspect3	Suspect4
Seized6 (0.496)	Seized5 (0.571)	Seized5 (0.006)	Seized4 (0.007)
Seized7 (1.198)	Seized6 (0.83)	Seized4 (1.094)	Seized5 (0.984)
Seized5 (1.249)	Seized7 (0.858)	Seized7 (1.323)	Seized7 (1.893)
Seized4 (2.451)	Seized4 (1.197)	Seized6 (1.719)	Seized6 (2.725)
Seized3 (29.311)	Seized3 (27.897)	Seized3 (27.423)	Seized3 (28.691)
Seized2 (29.362)	Seized2 (27.993)	Seized2 (27.537)	Seized2 (28.833)
Seized1 (29.592)	Seized1 (28.244)	Seized1 (27.794)	Seized1 (29.106)

Showing 1 to 7 of 7 entries

Figure 11.13 Amphetamine seized from the suspect compared to samples seized from users



Select method: amphetamineComparison

Method information:
Dataset: amphetamineTraining
Analysis type: Dissimilarity computation
PCA used: FALSE

Select file format: CSV
Select decimal separator for CSV files: . (dot)
Select column separator: , (comma)
Select data file: seizedSeized.csv
□ Compare against second data file?

Select id variable: ID

Select decimal separator for CSV files: . (dot)
Select column separator: , (comma)

	Seized1	Seized2	Seized3	Seized4	Seized5	Seized6	Seized7
Seized1	0	0.009	0.014	29.154	27.739	31.117	28.754
Seized2	0.009	0	0.004	28.877	27.481	30.924	28.52
Seized3	0.014	0.004	0	28.728	27.365	30.861	28.423
Seized4	29.154	28.877	28.728	0	1.035	2.866	1.97
Seized5	27.739	27.481	27.365	1.035	0	1.667	1.278
Seized6	31.117	30.924	30.861	2.866	1.667	0	0.919
Seized7	28.754	28.52	28.423	1.97	1.278	0.919	0

Showing 1 to 7 of 7 entries

Figure 11.14 Pairwise comparisons of amphetamine seized from users

Later on, the forensic laboratory was requested to evaluate these 3 samples in order to find relations to seizures made during the previous year on illegal street markets. After database comparison database links to 4 persons (suspect 1-4) in another city were observed. No previous information that would have connected these cases was available. This additional information could be provided to the investigating police unit.

As an outcome of such a profiling analysis a distribution network was discovered in Finland. The results of profiling analysis are presented in a graphical chart in Figure 11.15. Also, the sample found from the forest was linked to the 4 samples from the suspect. This additional information could be provided to the investigating police unit.

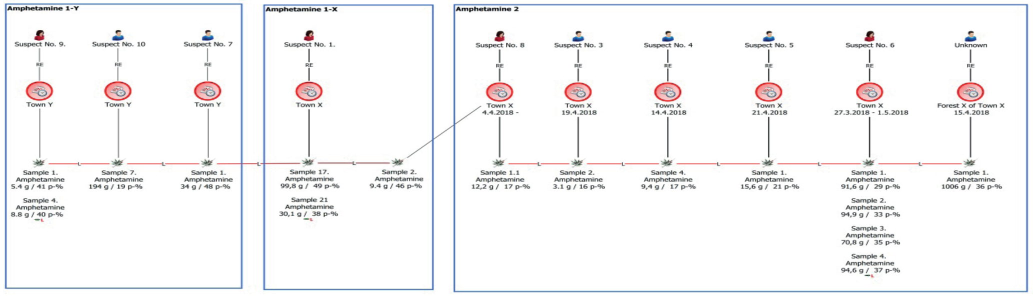

Amphetamine 1-Y
Suspect No. 9.
Suspect No. 10
Suspect No. 7
RE
Town Y
Sample 1. Amphetamine 5.4 g / 41 p-%
Sample 4. Amphetamine 8.8 g / 40 p-%
Sample 7. Amphetamine 194 g / 19 p-%
Sample 1. Amphetamine 34 g / 48 p-%
L
Amphetamine 1-X
Suspect No. 1.
Town X
Sample 17. Amphetamine 99,8 g / 49 p-%
Sample 21 Amphetamine 30,1 g / 38 p-%
Sample 2. Amphetamine 9.4 g / 46 p-%
Amphetamine 2
Suspect No. 8
Suspect No. 3
Suspect No. 4
Suspect No. 5
Suspect No. 6
Unknown
Town X 4.4.2018 -
Town X 19.4.2018
Town X 14.4.2018
Town X 21.4.2018
Town X 27.3.2018 - 1.5.2018
Forest X of Town X 15.4.2018
Sample 1.1 Amphetamine 12,2 g / 17 p-%
Sample 2. Amphetamine 3.1 g / 16 p-%
Sample 4. Amphetamine 9,4 g / 17 p-%
Sample 1. Amphetamine 15,6 g / 21 p-%
Sample 1. Amphetamine 91,6 g / 29 p-%
Sample 2. Amphetamine 94,9 g / 33 p-%
Sample 3. Amphetamine 70,8 g / 35 p-%
Sample 4. Amphetamine 94,6 g / 37 p-%
Sample 1. Amphetamine 1006 g / 36 p-%

Figure 11.15 Graphical visualization of profiling analysis results

In cases of amphetamine profiling the identity or content of adulterants or diluents is not part of the chemical profiling and has therefore not been included in this example because the production of amphetamine is in focus. Analysis of adulterants or diluents could be performed if the answering to the original question from the investigative police unit would require such analysis e.g. if the trafficking route is of importance.

Note that the procedure described for amphetamine could in principle be applied to the comparison of items of any kind of material for which type 1 data have been obtained, e.g. other drugs, oil, metal or glass. For such comparisons, there may already be preferred ways to pre-process the data and measure dissimilarity but the general procedure should still be the similar.

11.3 Example 3. Cocaine: identification and quantitation

Case history and Forensic Question

In a police investigation, the question arose whether certain material contained cocaine or not. Besides the identification, it was also of interest to determine the purity of the cocaine.

For this purpose, analytical techniques like GC-MS and GC-FID are commonly used [38]. However, in Eliaerts et al. [6] the implementation of a combined identification and fast quantification method for cocaine samples based on a FT-IR instrument is described. The FT-IR instrument is suitable for this type of analysis because the measurement is fast, a sample preparation is not required and it has a high identification power. The quantitative result obtained by the FT-IR provides a 'rough estimate' of the cocaine concentration. If a more precise result is needed, the analysis done by GC-FID or GC–MS required.

Selection of Parameters

The authors of "Rapid classification and quantification of cocaine in seized powders with ATR-FTIR and chemometrics, Drug Test" [6] supplied the FT- IR data of cocaine that was subsequently used for training and testing purposes. Typical FT-IR spectra of cocaine preparations are presented in Figure 11.16.

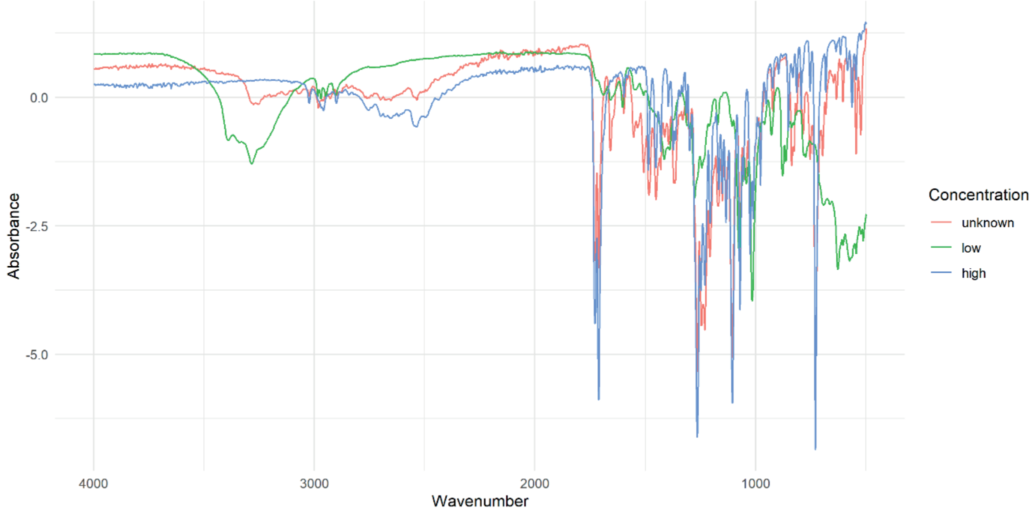


Figure 11.16 FTIR-spectrum of unknown sample (red) between cocaine samples of low (green) and high purity (blue)

Classification of Parameters

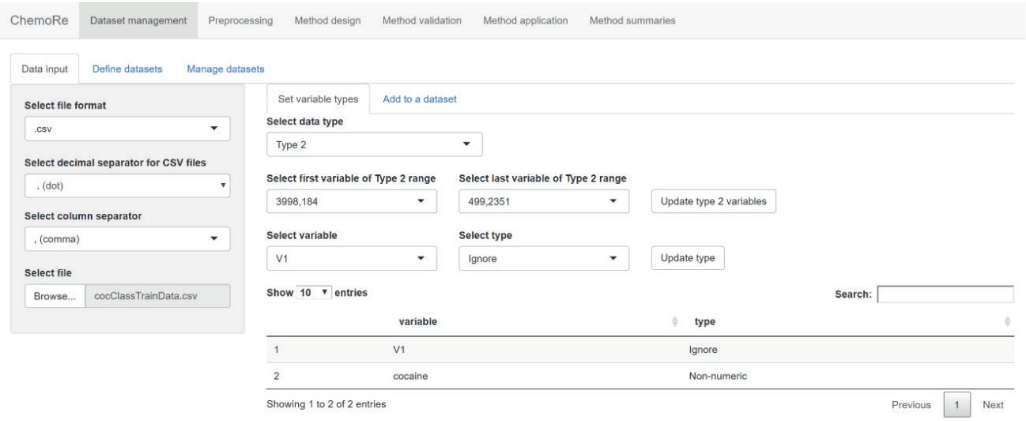
Example 3 contains Type 2 data with a training dataset; the data pre-processing and data analysis are performed to answer classification (identification) and quantification (determination of concentration) questions.

Training- and Testing Data

For both identification and quantification, the corresponding datasets were divided into training and testing datasets by randomly sampling 80% of the data to the former, and using the rest for testing. In the case of identification this resulted in 412 and 103 samples for training and testing respectively, and in the case of quantification, this resulted in 302 and 76 samples for training and testing respectively. Note that the datasets for identification and quantification were different, since in order to do quantification the cocaine must already be identified.

Data import

The spectral data sets have been imported into the ChemoRe software in form of a *.csv-file for classification (Figure 11.17) and quantification (Figure 11.18).



Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select file: cocClassTrainData.csv

Select data type: Type 2

Select first variable of Type 2 range: 3998,184

Select last variable of Type 2 range: 499,2351

Select variable: V1

Select type: Ignore

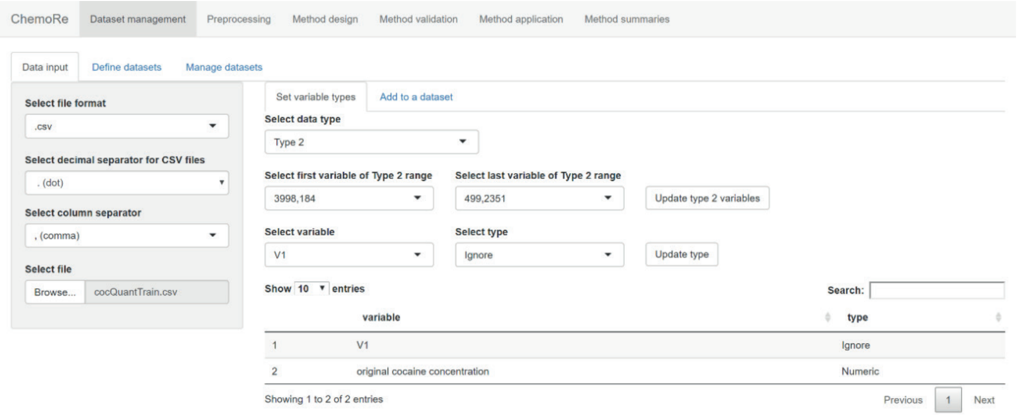
Show 10 entries

Search:

	variable	type
1	V1	Ignore
2	cocaine	Non-numeric

Showing 1 to 2 of 2 entries

Figure 11.17 Import of spectral data for classification. Each Dataset consists of an FT-IR spectrum in the range of 500 – 4000 wavenumbers (type 2 data; non-numeric)



Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select file: cocQuantTrain.csv

Select data type: Type 2

Select first variable of Type 2 range: 3998,184

Select last variable of Type 2 range: 499,2351

Select variable: V1

Select type: Ignore

Show 10 entries

Search:

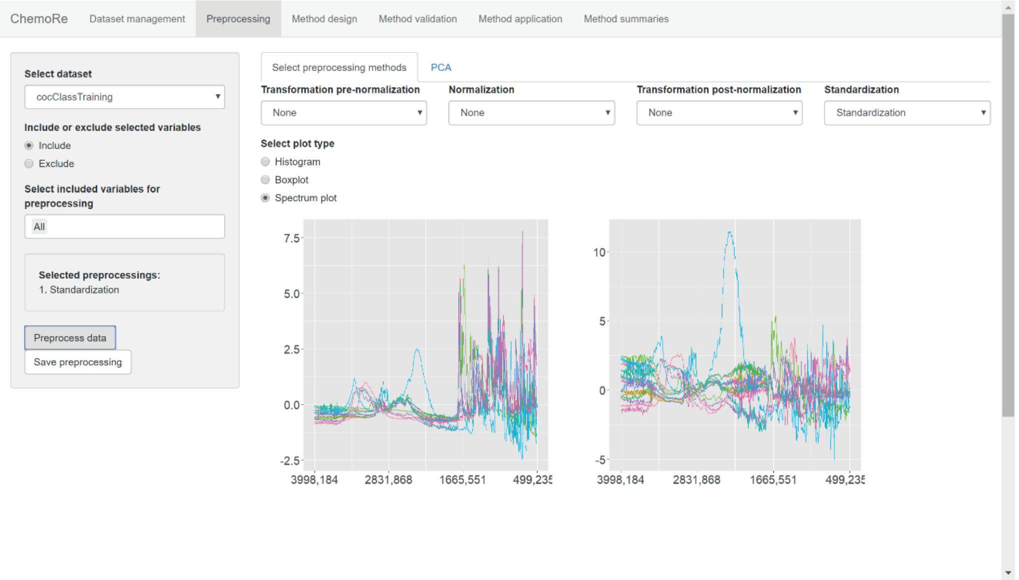
	variable	type
1	V1	Ignore
2	original cocaine concentration	Numeric

Showing 1 to 2 of 2 entries

Figure 11.18 Import of spectral data for quantification. Each Dataset consists of an FT-IR spectrum in the range of 500 – 4000 wavenumbers (type 2 data; numeric)

Data pre-processing

Although, Eliaerts et.al used standard normal variate transformation (SNV) as the pre-processing method for FT-IR spectral data. In this example, as a pre-processing step, each FT-IR spectrum was transformed to a z-score (as explained above) in order to limit the effect of baseline differences between measurements (Figure 11.19 classification, Figure 11.20 quantification). For this, sample means and variances were calculated from training data for each of the 2440 variables, corresponding to signal intensities of measured wavelengths ranging from 500 to 4000 nm, and these were used for standardization of both training and test data. PCA was then applied in both cases to the standardized data (classification Figure 11.21 and quantification Figure 11.22). The principal components explaining 99% of the total variance in the data were retained in both cases which resulted in 44 and 24 variables retained, for identification and quantification respectively.



Select dataset: cocClassTraining

Include or exclude selected variables
● Include
○ Exclude

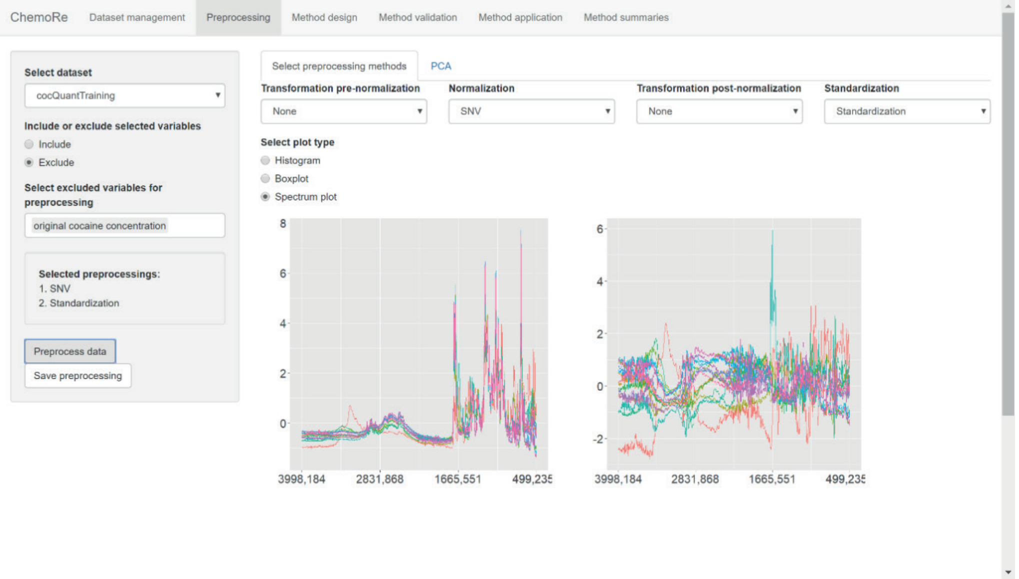
Select included variables for preprocessing: All

Selected preprocessings:
1. Standardization

Transformation pre-normalization: None
Normalization: None
Transformation post-normalization: None
Standardization: Standardization

Select plot type
○ Histogram
○ Boxplot
● Spectrum plot

Figure 11.19 Data pre-processing (classification) - effect of standardization (z-score transformation) applied to spectral data of cocaine preparations



Select dataset: cocQuantTraining

Include or exclude selected variables
○ Include
● Exclude

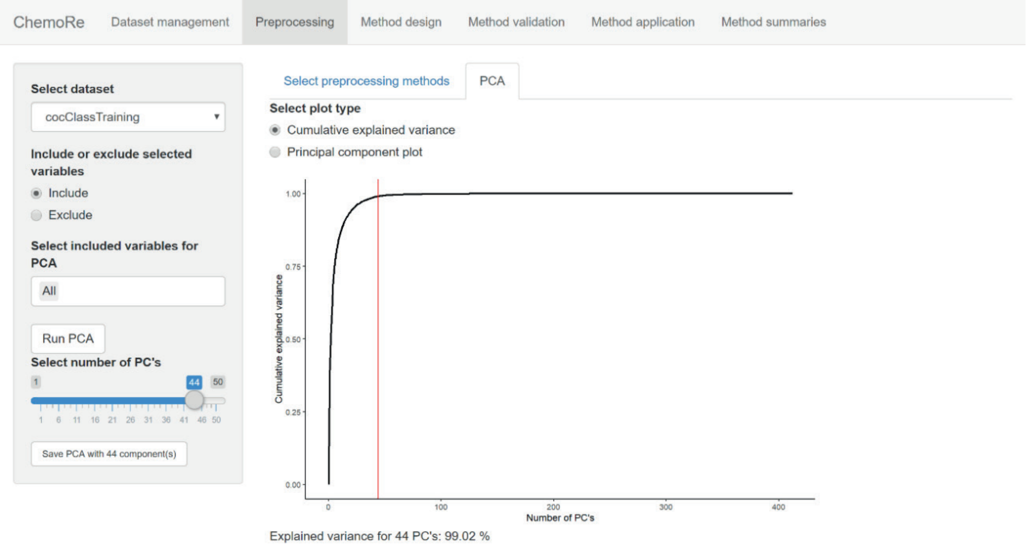
Select excluded variables for preprocessing: original cocaine concentration

Selected preprocessings:
1. SNV
2. Standardization

Transformation pre-normalization: None
Normalization: SNV
Transformation post-normalization: None
Standardization: Standardization

Select plot type
○ Histogram
○ Boxplot
● Spectrum plot

Figure 11.20 Data pre-processing (quantification) - effect of SNV transformation followed by standardization applied to spectral data of cocaine preparations



Select dataset: cocClassTraining

Include or exclude selected variables
● Include
○ Exclude

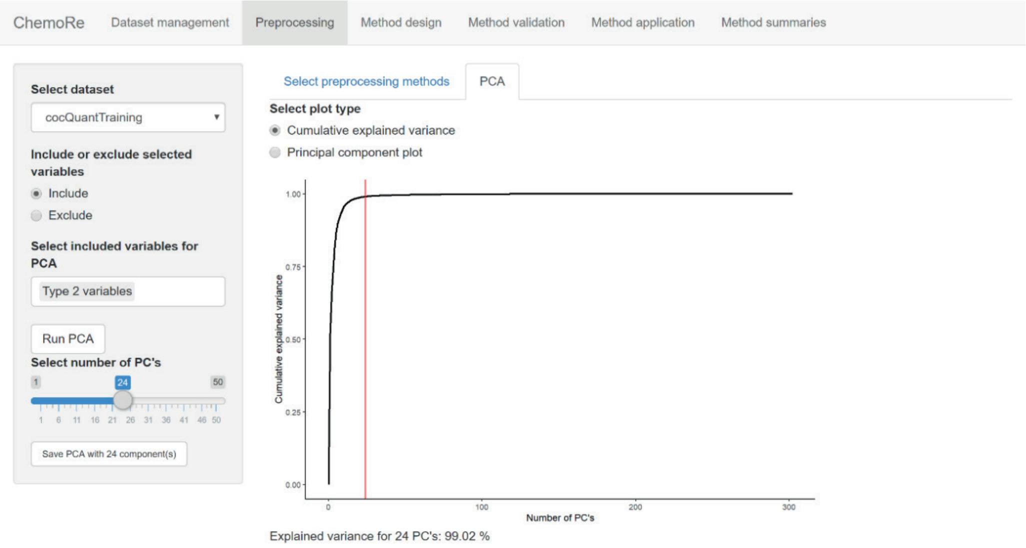
Select included variables for PCA: All

Select number of PC's: 44

Select plot type
● Cumulative explained variance
○ Principal component plot

Explained variance for 44 PC's: 99.02 %

Figure 11.21 Data pre-processing (classification) – reducing dimensionality by applying PCA after standardization to spectral data of cocaine preparations



Select dataset: cocQuantTraining

Include or exclude selected variables
● Include
○ Exclude

Select included variables for PCA: Type 2 variables

Select number of PC's: 24

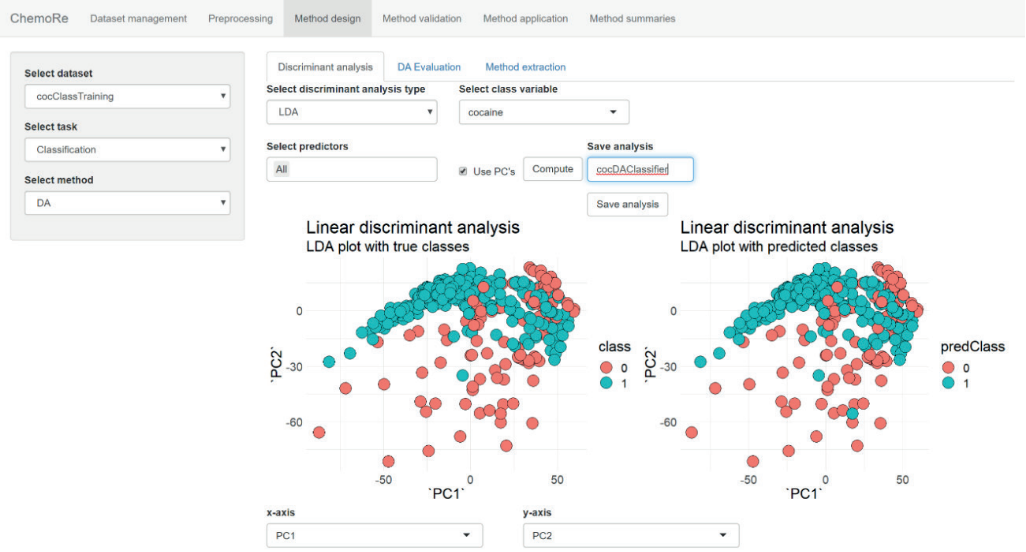
Select plot type
● Cumulative explained variance
○ Principal component plot

Explained variance for 24 PC's: 99.02 %

Figure 11.22 Data pre-processing (quantification) – reducing dimensionality by applying PCA after standardization to spectral data of cocaine preparations

Chemometrics - Method development

The first question of identification is framed as a statistical classification and the latter, (quantification) as a regression problem. To classify and quantify cocaine from the pre-processed data, Eliaerts et al. applied Support vector machines for discriminant analysis and regression (SVM-DA and SVM-R). In the current example, a simpler chemometric method is applied for the same purpose. To this end, LDA model is trained to determine whether a sample contains cocaine or not (Figure 11.23). In addition to identification, linear regression is used for quantification of cocaine shown in Figure 11.24. Again, PCA pre-processing is applied to training data. The regression model is then fitted to the reduced data using the measured concentration of cocaine as response variable. For the classification by LDA and the quantification by linear regression, separate training and test datasets are used for training and evaluation of the performance of the methods.



Select dataset: cocClassTraining

Select task: Classification

Select method: DA

Select discriminant analysis type: LDA

Select class variable: cocaine

Select predictors: All

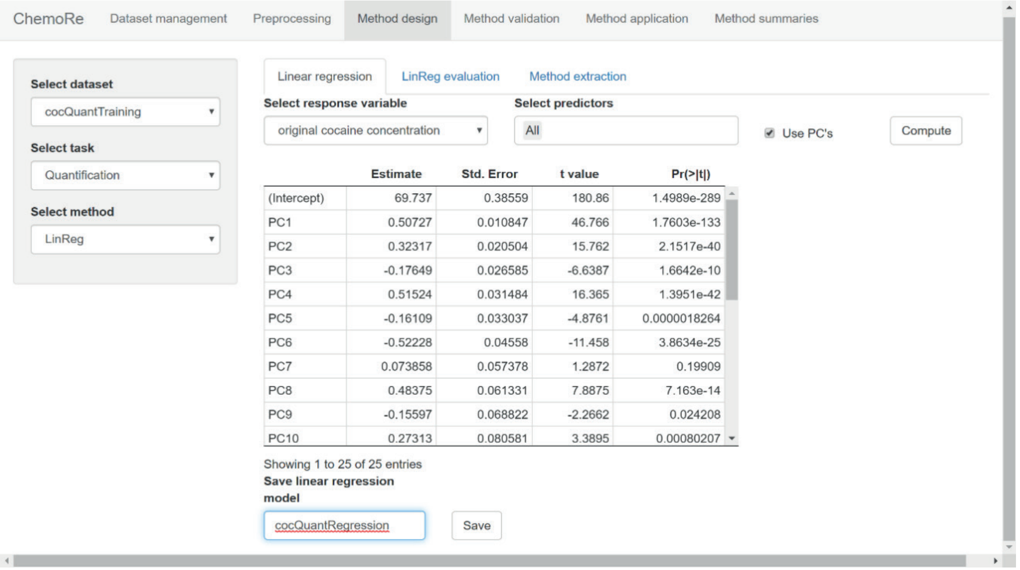
☑ Use PC's

Save analysis: cocDAClassifier

x-axis: PC1

y-axis: PC2

Figure 11.23 Applying LDA for classification whether a seized material contain cocaine or not



Select dataset: cocQuantTraining

Select task: Quantification

Select method: LinReg

Select response variable: original cocaine concentration

Select predictors: All

☑ Use PC's

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	69.737	0.38559	180.86	1.4989e-289
PC1	0.50727	0.010847	46.766	1.7603e-133
PC2	0.32317	0.020504	15.762	2.1517e-40
PC3	-0.17649	0.026585	-6.6387	1.6642e-10
PC4	0.51524	0.031484	16.365	1.3951e-42
PC5	-0.16109	0.033037	-4.8761	0.0000018264
PC6	-0.52228	0.04558	-11.458	3.8634e-25
PC7	0.073858	0.057378	1.2872	0.19909
PC8	0.48375	0.061331	7.8875	7.163e-14
PC9	-0.15597	0.068822	-2.2662	0.024208
PC10	0.27313	0.080581	3.3895	0.00080207

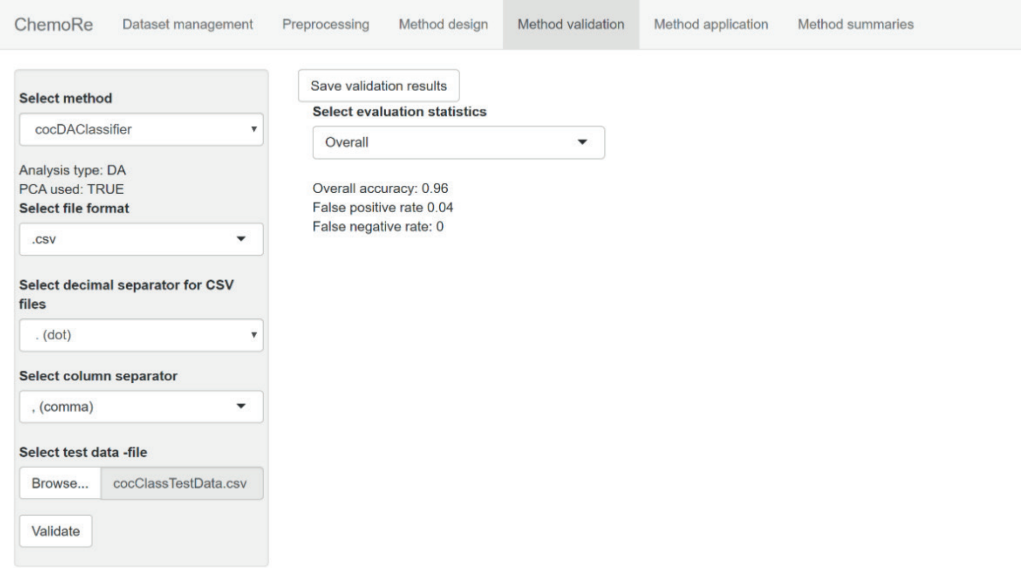
Showing 1 to 25 of 25 entries

Save linear regression model: cocQuantRegression

Figure 11.24 Linear regression model based on training dataset of seized cocaine preparations

Results:

For identification purposes, an LDA model was fitted to the pre-processed training data as described earlier. This can be framed as a classification problem by considering the property of a sample containing cocaine as defining a class while missing of cocaine defines the other class. The resulting model made correct predictions on the test data in 96% of the time, with approximately 4% false positive results and 0% false negatives (Figure 11.25).



Select method: cocDAClassifier

Analysis type: DA
PCA used: TRUE

Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

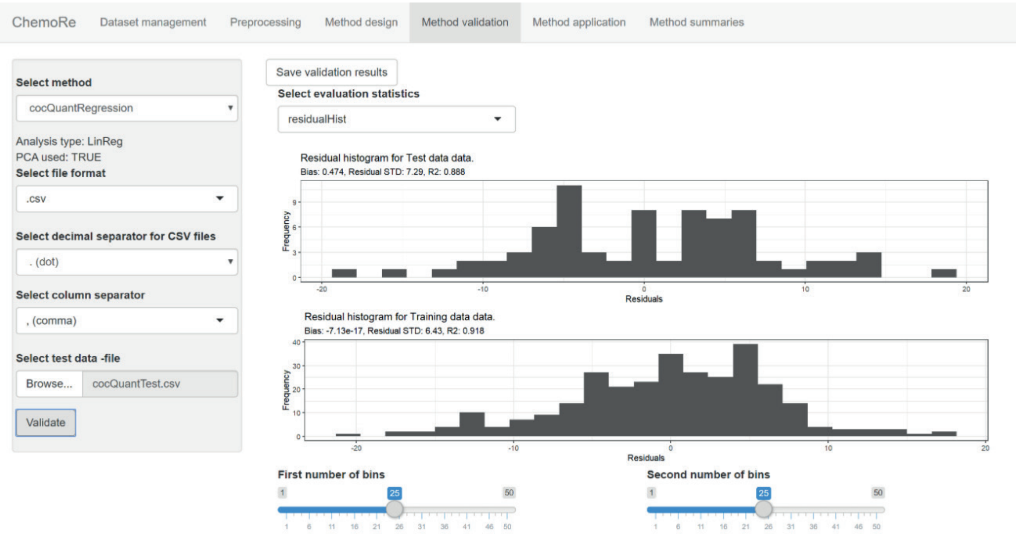
Select test data -file: cocClassTestData.csv

Select evaluation statistics: Overall

Overall accuracy: 0.96
False positive rate 0.04
False negative rate: 0

Figure 11.25 Validation parameter of classification model

It should be noted that, unlike in the previous examples, no threshold for classification is determined by user and as such the error rates do not depend on user choices. For quantification, a linear regression was fit using the 24 retained principal components as predictors. Applying the linear regression on the test data resulted in coefficient of determination of 0.89 with bias of approximately 0.47 percent points and residual standard deviation of approximately 7.30 percent points (Figure 11.26).



Select method: cocQuantRegression

Analysis type: LinReg
PCA used: TRUE

Select file format: .csv

Select decimal separator for CSV files: . (dot)

Select column separator: , (comma)

Select test data -file: cocQuantTest.csv

Select evaluation statistics: residualHist

Residual histogram for Test data data.
Bias: 0.474, Residual STD: 7.29, R2: 0.888

Residual histogram for Training data data.
Bias: -7.13e-17, Residual STD: 6.43, R2: 0.918

Figure 11.26 Validation parameter of quantification model

The performance of the identification and quantification method is somewhat reduced compared to [6] where a more advanced method was applied. Here simpler methods were preferred in order to illustrate how standard chemometrics may be applied on spectral data.

As to the final example 3 on simultaneous identification and quantification of cocaine, here it is illustrated that with the application of chemometric methods, the spectral data of FT-IR can be used for identification as well as quantification. It should be noted that the number of principal components that are retained is critical for the performance of the model and here the choice of retaining components explaining 99% of the total variance is rather ad hoc. One could also optimize this number by so-called cross validation or by selecting the variance limit based on the performance measured on the test set, but in this case, it would be necessary to use additional data to test the end result in order to avoid overfitting. While here statistical classification methods were used for the purposes of identification, the same process could be extended for a forensic classification problem where one needs to classify samples according to multiple known sources.

12 REFERENCES

[1] M. Bovens, B. Ahrens, I. Alberink, A. Nordgaard, T. Salonen, S. Huhtala, Chemometrics in forensic chemistry-Part I: Implications to the forensic workflow, Forensic Science International, 301 (2019) 82-90

[2] T. Salonen, B. Ahrens, M. Bovens, J. Eliaerts, S. Huhtala, A. Nordgaard, I. Alberink, Chemometrics in forensic chemistry — Part II: Standardized applications – Three examples involving illicit drugs, Forensic Science International, 307 (2020) 110138

[3] A. Nordgaard, B. Ahrens, I. Alberink, T. Salonen, S. Huhtala, M. Bovens, Chemometrics in Forensic Chemistry – Part III: Assessment and interpretation of results, Forensic Science International, submitted March 6th 2020

[4] N. Stojanovska, S. Fu, M. Tahtouh, T. Kelly, A. Beavis, K. Kirkbride, A review of impurity profiling and synthetic route of manufacture of methylamphetamine, 3,4-methylenedioxymethylamphetamine, amphetamine, dimethylamphetamine and p-methoxyamphetamine, Forensic Sci. Int. 224 (2013) 8-26.

[5] J. Broseus, S. Huhtala, P. Esseiva, First systematic chemical profiling of cocaine police seizures in Finland in the framework of an intelligence-led approach, Forensic Sci. Int. 251 (2015) 87-94, doi:10.1016/j.forsciint.2015.03.026.

[6] J. Eliaerts, P. Dardenne, N. Meert, F. van Durme, N. Samyn, K. Janssens, K. De Wael, Rapid classification and quantification of cocaine in seized powders with ATR-FTIR and chemometrics, Drug Test. Analysis 9 (2017) 1480-1489, doi:10.1002/dta.2149.

[7] T.S. Groberio, J.J. Zacca, E.D. Botelho, M. Talhavini, J.W.B. Braga, Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR, Forensic Sci. Int., 257 (2015) 297–306.

[8] M. Monfreda, F. Varani, F. Cattaruzza, S. Ciambrone, A. Proposito, Fast profiling of cocaine seizures by FTIR spectroscopy and GC-MS analysis of minor alkaloids and residual solvents, Science and Justice 55 (2015) 456–466.

[9] A.C. Moffat, M.D. Osselton, B. Widdop, J. Watts, Clarke's Analysis of Drugs and Poisons 4th Ed., Pharmaceutical Press, 2011.

[10] E. Lock, Development of a harmonized method for the profiling of amphetamine, PhD. Thesis, Institute de police Scientifique, University of Lausanne, Switzerland, 2005.

[11] C.S.L. Jonson, L. Strömberg, Two-level classification of Leuckart amphetamine, Forensic Sci. Int. 69 (1994) 31-44.

[12] C.S.L. Jonson, Amphetamine profiling – improvements of data processing, Forensic Sci. Int. 69 (1994), 45-54.

[13] L. Dujourdy, V. Dufey, F. Besacier, N. Miano, R. Marquis, E. Lock, L. Aalberg, S. Dieckmann, F. Zrcek, J.S. Bozenko jr. Drug intelligence based on organic impurities in illicit MA samples, Forensic Sci. Int. 177 (2008) 153-161.

[14] J. Broseus, S. Baechler, N. Gentile, P. Esseiva, Chemical profiling: A tool to decipher the structure and organization of illicit drug markets: An 8-year study in Western Switzerland, Forensic Sci. Int. 266 (2016) 18–28.

[15] P. Esseiva, L. Dujourdy, F. Anglada, F. Taroni, P. Margot A methodology for illicit heroin seizures comparison in a drug intelligence perspective using large databases, Forensic Sci. Int. 132 (2003) 139-152.

[16] O. Ribaux, A. Girod, S.J. Walsh, P. Margot P, S. Mizrahi, V. Clivaz, Forensic intelligence and crime analysis. Law, Probability and Risk 2 (2003) 47-60.

[17] O. Ribaux, S.J. Walsh, P. Margot, The contribution of forensic science to crime analysis and investigation: Forensic intelligence. Forensic Sci. Int. 156 (2006) 171-181.

[18] M. Otto, Chemometrics: Statistics and Computer-Applications in Analytical Chemistry, Third Edition, Wiley-VCH Verlag GmbH, Weinheim, 2017.

[19] J. Mazerski, Introduction in Chemometrics, in Chemometrics – Methods and Applications, edited by Dariusz Zuba and Andrzej Parczewski, Institut of Forensic Research Publishers, pp 11-15, 2006

[20] S. Meola, M.Sc. Thesis, Institute de police Scientifique, University of Lausanne, Switzerland, 2013.

[21] Ribaux O, Baylon A, Roux C, Delémont O, Lock E, Zingg C, Margot P. Intelligence-led crime scene processing. Part I: Forensic intelligence. Forensic Science International 2010; 195 (1–3): 10-16.

[22] Morelato M, Beavis A, Tahtouh M, Ribaux O, Kirkbride P, Roux C. The use of forensic case data in intelligence-led policing: The example of drug profiling. Forensic Science International 2013; 226 (1–3): 1-9.

[23] Esseiva P, Ioset S, Anglada F, Gasté L, Ribaux O, Margot P, Gallusser A, Biedermann A, Specht Y, Ottinger E. Forensic drug Intelligence: An important tool in law enforcement. Forensic Science International 2007; 167 (2–3): 247-254.

[24] CORRIGENDUM: Centenarians, but not octogenarians, up-regulate the expression of microRNAs - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Principal-component-analysis-PCA-of-the-small-non-coding-RNA-profiles-in-mononuclear_fig5_233902313, [accessed 11 Mar, 2020]

[25] https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0157601.g004; doi: https://doi.org/10.1371/journal.pone.0157601.g004, [accessed 17 Oct, 2019]

[26] F. Been, P. Esseiva, O. Delemont, Analysis of illicit drugs in wastewater – Is there an added value for law enforcement?, Forensic Sci. Int.I 266 (2016) 215-221, http://dx.doi.org/10.1016/j.forsciint.2016.05.032

[27] S. Schneider, F. Meys, Analysis of illicit cocaine and heroin samples seized in Luxembourg from 2005–2010, Forensic Sci. Int. 212 (2011) 242–246, doi:10.1016/j.forsciint.2011.06.027

[28] L. Stride Nielsen, P. Villesen, C. Lindholst, Variation in chemical profiles within large seizures of cocaine bricks, Forensic Sci. Int. 280 (2017) 194-199, https://doi.org/10.1016/j.forsciint.2017.10.007

[29] ENFSI Best Practice Manual (BPM) for controlled drug analysis, DWG-CDA-001, http://enfsi.eu/documents/best-practice-manuals/, [accessed 11 March, 2020]

[30] S. Lociciro, P. Hayoz, P. Esseiva, L. Dujourdy, F. Besacier, P. Margot, Cocaine profiling for strategic intelligence purposes, a cross-border project between France and Switzerland Part I. Optimisation and harmonisation of the profiling method, Forensic Sci. Int., 167 (2007) 220–228

[31] Development of a Harmonized Method for the Profiling of Amphetamines, Project – SMT-CT98-2277, DG Research of the EU Commission, 2003.

[32] Citizens and Governance in A Knowledge-Based Society - Providing health, security and opportunity to the people of Europe - Collaborative Harmonisation of Methods for Profilling of Amphetamine Type Stimulants (CHAMP), Project - CIS8-CT-2004-502126, 2004.

[33] DG Research of the EU Commission, ENFSI Monopoly MP2013-T6, SaiLR - Software for Likelihood Ratio Calculation, Project - Home/2013/ISEC/MO/ENFSI/4000005962

[34] R. Marquis, C. Weyermann, C. Delaporte, P. Esseiva, L. Aalberg, F. Besacier, J.S. Bozenko jr, R. Dahlenburg, C. Kopper, F. Zrcek, Drug intelligence based on MDMA tablets data 2. Physical characteristics profiling, Forensic Sci. Int., 178 (1) (2008), 34-39, https://doi.org/10.1016/j.forsciint.2008.01.014

[35] K. Andersson, K. Jalava, E. Lock, H. Huizer, E. Kaa, A. Lopes, A. Poortman-van der Meer, M.D. Cole, J. Dahlén, E. Sippola, Development of a harmonised method for the profiling of amphetamines: IV. Optimisation of sample preparation, Forensic Sci. Int. 169 (2007) 64–76, doi:10.1016/j.forsciint.2006.10.017

[36] K. Andersson, K. Jalava, E. Lock, Y. Finnon, H. Huizer, E. Kaa, A. Lopes, A. Poortman-van der Meer, M.D. Cole, J. Dahlén, E. Sippola, Development of a harmonised method for the profiling of amphetamines: III. Development of the gas chromatographic method, Forensic Sci. Int. 169 (2007) 50–63, doi: http://dx.doi.org/10.1016/j.forsciint.2006.10.018.

[37] K. Andersson, E. Lock, K. Jalava, H. Huizer, S. Jonson, E. Kaa, A. Lopes, A. Poortman-van der Meer, E. Sippola, L. Dujourdy, J. Dahlén, Development of a harmonised method for the profiling of amphetamines VI: evaluation of methods for comparison of amphetamine, Forensic Sci. Int. 169 (2007) 86-99, doi:http://dx.doi.org/10.1016/j.forsciint.2006.10.020.

[38] J.F. Casale, R.W. Waggoner Jr., A chromatographic impurity signature profile analysis for cocaine using capillary gas chromatography, J. For. Sci. 36 (5) (1991) 1312–1330, doi:http://dx.doi.org/10.1520/JFS13154J.

13 APPENDIX

This guideline can be found from the ENFSI website:
http://enfsi.eu/documents/forensic-guidelines/

ChemoRe software and following additional material are available upon request from ENFSI Drugs Working Group and from the EPE web site (Europol Platform for Experts):

1. ChemoRe user manual
2. ChemoRe validation report
3. Data of the examples 1 to 3
4. Tips and Tricks of Data Export

14 AUTHORS IN ALPHABETICAL ORDER

Dr. Björn Ahrens
Bundeskriminalamt Wiesbaden, KT-45, Äppelallee 45, D-65203 Wiesbaden, Germany
bjoern.ahrens@bka.bund.de

Dr. Ivo Alberink
Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB, Den Haag, The Netherlands
i.alberink@nfi.nl

Dr. Michael Bovens (Chairman ENFSI Drugs Working Group Subcommittee Chemometrics)
Zurich Forensic Science Institute, P.O. Box, 8021 Zurich, Switzerland
michael.bovens@for-zh.ch

Sami Huhtala (Work Package Leader of the EU Direct Grant STEFA-G02)
National Bureau of Investigation, Jokiniemenkuja 4, P.O. Box 285, FI-01301 Vantaa, Finland
sami.huhtala@poliisi.fi

Dr. Anders Nordgaard
National Forensic Centre, Swedish Police Authority, 58194 Linköping, Sweden
anders.nordgaard@polisen.se

Tuomas Salonen
University of Helsinki, Department of Mathematics and Statistics, P.O. Box 68, FI-00014 Helsinki Finland
tuomas.x.salonen@helsinki.fi

15 ACKNOWLEDGEMENTS

We acknowledge Juhana Rautavirta for substantial contribution to the development of the ChemoRe software.

We acknowledge Joy Eliaerts for the valuable contributions to this guideline and to the software tool ChemoRe.

EU commission for funding through the European Union's Internal Security Fund — Police: 779485 — STEFA — ISFP-2016-AG-IBA-ENFSI during a period of 29 months from January 2018 to May 2020.

Special thanks to the ENFSI Drugs Working Group for supporting the Subcommittee Chemometrics.

****