



NTNU – Trondheim
Norwegian University of
Science and Technology

A Method for Rapid Localisation of Hydrocarbon Compounds on Surfaces Using Chemical Imaging and Back-Projection

Iselin Aakre

Chemistry

Submission date: May 2013

Supervisor: Bjørn Kåre Alsberg, IKJ

Norwegian University of Science and Technology
Department of Chemistry

Preface

The work in this thesis have been conducted at Department of Chemistry, Norwegian University of Science and Technology (NTNU) from August 2011 to May 2013.

Firstly, I want to thank my supervisor Professor Bjørn Kåre Alsberg, for helping me, believing in me, and telling me I needed to start writing. You were, of course, right.

Thanks to engineer Roger Aarvik (Department of Chemistry, NTNU) for help with supplying chemicals and laboratory equipment.

Thanks also to all my family and friends, and especially those of you who have brightened my days by sharing your lunch breaks with me.

Iselin Aakre
Trondheim, May 15, 2013

Sammendrag

Miljøforurensninger er et stort og sammensatt problem. Enkle, raske og pålitelige metoder for å oppdage slike forurensninger kan sees på som et første ledd mot en løsning.

Ved hjelp av et hyperspektralt kamera som tar bilder i området 923-1665 nm, kjemometriske metoder som diskriminant delvis minste kvadraters regresjon (discriminant partial least squares regression, DPLSR) og k -nærmeste naboer (k -NN) samt ulike former for preprosessering, er flere modeller for deteksjon av hydrokarboner på bakgrunner som jord, sand, stein, humus og vegetasjon utviklet. Disse bakgrunnene ble valgt som eksempler på naturlig forekommende overflater der det kan være interessant og nyttig å påvise eventuelle forekomster av hydrokarboner. Som hydrokarbon ble parafin brukt.

For enkelte av modellene viste det seg å ha stor betydning for feilen om spektrene ble normalisert eller ikke. Dette skyldes sannsynligvis at overflatene var ujevne og at gjennomtrengingen var forskjellig fra sted til sted. Da overflater i naturen vanligvis ikke er glatte, er det naturlig å anta at dette vil ha enda større betydning på naturlige bakgrunner enn i et laboratoriemiljø. Utglatting av dataene ga også lavere feil for mange modeller.

En enklere modell, der data fra kun 16 av de 148 tilgjengelige bølgelengdene ble benyttet, viste seg å gi overraskende gode resultater på de fleste bakgrunnene. Denne modellen var mindre selektiv enn DPLSR-modellene, og detekterte også andre hydrokarboner enn parafin. Dette antas å være fordi modellen gjenkjenner C-H-bindingen i hydrokarboner.

Den prosentvise feilen til modellene varierte mye mellom bakgrunnene. De ulike modellene hadde ulike sterke og svake sider. Vann forårsaket falske positive for noen modeller og bakgrunner.

En tidsserie av bilder av jord med og uten parafin tatt over et år, viste at en DPLSR modell basert på data fra den første uka hadde en lav prosentvis feil for alle bildene. Data fra andre bilder tydet likevel på at tiden kan ha en større effekt for andre bakgrunner enn jord.

DPLSR var i stand til å skille mellom n -heksan og n -heptan på en bakgrunn av jord. Dette sannsynliggjør et det er mulig å lage en selektiv modell som skiller mellom flere ulike hydrokarboner.

Abstract

Discharges of unwanted chemicals into nature is a large and complex problem. Easy, fast and reliable methods for detecting these contaminants can be viewed as a first step towards a solution.

Using a hyperspectral camera operating in the range 923-1665 nm, chemometric methods such as discriminant partial least squares regression (DPLSR) and k -nearest neighbours (k -NN) as well as various forms of preprocessing, models for detecting hydrocarbons on surfaces such as soil, sand, stone, humus and vegetation were developed. These surfaces were chosen as examples of naturally occurring surfaces where locating hydrocarbons can be interesting and useful. Paraffin were used as a hydrocarbon.

For some of the DPLSR models, it proved to be of great importance for the error whether or not the spectra were normalised. This is probably due to the fact that the surfaces were uneven, and that the penetration would differ from place to place. As surfaces in nature usually will not be smooth, it is natural to assume that this will have an even greater impact on natural backgrounds than in a laboratory environment. Smoothing of the data also improved many of the models.

A simpler model, where data from only 16 of the 148 available wavelengths were used, turned out to give surprisingly good results on most surfaces. This model was less selective than the DPLSR models, and will also detect hydrocarbons other than paraffin. This is probably due to this model recognising the C-H-bond.

The percentage error of the models varied widely between different surfaces. The different models have different strengths and weaknesses. Water caused false positives for some models and surfaces.

A time series of images of soil with and without paraffin taken over a year, showed that a DPLSR model based on data from the first week had a low percentage error for all the images. However, data from other images indicate that time could have a greater impact for other surfaces than soil.

DPLSR were able to distinguish between n -hexane and n -heptane on soil. This substantiates that it is possible to create a model that distinguishes between several different hydrocarbons.

Contents

List of abbreviations	vii
1 Introduction	1
2 Theory	5
2.1 Remote sensing	5
2.1.1 What is hyperspectral imaging?	5
2.1.2 Hyperspectral data	6
2.1.3 Hyperspectral cameras	8
2.1.4 Remote sensing of hydrocarbons	10
2.1.5 The PryJector	12
2.2 NIR-spectra of hydrocarbons	13
2.3 Chemometric principles	14
2.3.1 Classification and regression	14
2.3.2 Calibration and validation	15
2.3.3 Overfitting and underfitting	16
2.4 Preprocessing	17
2.4.1 Smoothing	17
2.4.2 Multiplicative scatter correction	21
2.4.3 Centering, scaling and autoscaling	22
2.4.4 Normalisation	23
2.4.5 Logarithm	23
2.4.6 Numerical differentiation	24
2.5 Principal component analysis	24
2.6 Partial least squares	27
2.7 k -nearest neighbours	29

3	Experimental	31
3.1	Hypotheses	31
3.2	The set up	32
3.3	Hyperspectral images	33
3.3.1	The effect of time	34
3.3.2	Different concentrations	35
3.3.3	Different surfaces	37
3.3.4	Time dependence, different backgrounds	37
3.3.5	Different hydrocarbons	40
3.4	Data sets	40
3.5	Calculations	42
4	Results	45
4.1	Detecting hydrocarbons	45
4.1.1	DPLSR models	45
4.1.2	k -NN models	48
4.1.3	Waveband models	50
4.2	Estimating the amount	55
4.3	Identifying hydrocarbons	56
4.3.1	DPLSR models	57
4.3.2	k -NN models	57
4.4	Time dependence	58
4.4.1	Can the time be estimated?	59
5	Discussion	61
5.1	Data sets	61
5.2	Comparing models	63
5.2.1	Hard and soft modelling	66
5.3	Different surfaces	66
5.4	Improving the k -NN models	69
5.5	A cheaper set up	71
5.6	Other methods	72
6	Conclusion	75
	Bibliography	77
A	Wavelengths for the hyperspectral camera	A-1
B	Surface specific error, data set 3b	A-3

List of abbreviations

Abbreviation	Meaning	Section
AOTF	Acousto-optic tunable filter	2.1.1
CCD	Charge-coupled device	2.1.1
DPLSR	Discriminant partial least squares regression	2.6
ETF	Electronically tunable filter	2.1.3
FPA	Focal plane array	2.1.3
FWHM	Full-width at half maximum	2.1.3
HI	Hydrocarbon Index	2.1.4
IR	Infrared	2.2
k -NN	k -nearest neighbours	2.7
LCTF	Liquid-crystal tunable filter	2.1.3
LDA	Linear discriminant analysis	5.6
LOO	Leave one out (cross-validation)	2.3.2
MA	Moving average	2.4.1.2
MFE	Morphological feature extraction	5.6
MLR	Multiple linear regression	2.3.1
MSC	Multiplicative scatter correction	2.4.2
NIPALS	Nonlinear Iterative Partial Least Squares	2.5
NIR	Near-infrared	2.2
NIRCI	Near-infrared chemical imaging	2.2
PC	Principal component	2.5
PCA	Principal component analysis	2.5
PCR	Principal component regression	2.5
PLS	Partial least squares	2.6
PLSR	Partial least squares regression	2.6
RF	Random forest	5.6
SNR	Signal-to-noise ratio	2.1.2.2
SVD	Singular value decomposition	2.5
SVM	Support vector machines	5.6
QWIP	Quantum well infrared photodetectors	2.1.3
TF	Tunable filter	2.1.3

Chapter 1

Introduction

Our society has made itself completely dependent upon a wide variety of chemicals. Of the tens of thousands of chemicals produced around the world every year, organic compounds represent the largest portion [29]. Petroleum and natural gas provide cheap starting materials for synthesising organic molecules [8]. Despite the best of human intentions, it seems inevitable that some of these chemicals eventually end up in nature. This will, and does, affect the environment.

Environmental systems are very complex, with a plethora of known and unknown variables to account for [27]. Large and multivariate data sets are needed to describe them, and interpreting these data sets are challenging.

Chemometrics can be used to analyse such complex data sets. There are several definitions of chemometrics. Varmuza [55] defines chemometrics as

Chemometrics concerns the extraction of relevant information from chemical data by mathematical and statistical tools.

In this thesis, chemometric methods and techniques are used to investigate data from hyperspectral images of alkane compounds applied to several different surfaces.

Long term leakage of hydrocarbons into nature will alter the soil composition. Changes such as formation of new minerals, bleaching, electrochemical and radiometric alterations of soil are documented long term effects of petroleum spills [43].

Most methods for detecting and monitoring organic chemicals in nature, involves grab sampling methods [29]. Grab sampling techniques are techniques which involves collecting an analyte and removing it from the surroundings before analysing the sample. This introduces several problems, such as chemical degradation [29]. It is also time consuming and potentially expensive.

To avoid these problems, non-invasive *in situ* techniques are needed. Remote sensing is a research field developed for investigating objects at some distance from the sensor. Research suggests that reduction of ferric iron, conversion of mixed-layer clay and feldspars to kaolinite and changes to the spectral reflection of the vegetation, are some of the long term effects of petroleum spills that can be detected by remote sensing and hyperspectral imaging [44]. However, other factors such as climate, soil composition and topography must also be taken into account, as the effect of these factors can be larger than the effect due to hydrocarbon spills.

With these concerns in mind, this thesis describes an alternative and non-invasive method for rapid localisation and *in situ* visualisation of hydrocarbon spills in nature. The method uses the PryJector [2], a device that combines a hyperspectral camera operating in the near-infrared area, a chemometric model, and a projector. Different models are developed, all of them with a focus on alkanes, but some of them seem to recognise also other types of hydrocarbons and organic compounds.

While remote sensing hyperspectral imaging generally is used for investigating and classifying large areas, the PryJector is more suitable for human-size problems. This makes it fitting for detecting small-scale spills of e.g. petroleum fuel, but it could also be useful for forensic uses such as arson. However, the models developed in this thesis are general and can also be used in other hyperspectral investigations, such as those conducted from aeroplanes and satellites.

The theory chapter starts with a section on hyperspectral imaging, a short overview of some of the earlier studies on remote sensing of hydrocarbons, and a brief presentation of the PryJector and a couple of its possible uses. This is followed by a section on some basic chemometric terms and principles.

Preprocessing can be a crucial part of data analysis. The focus in section 2.4 is on why and when the different kinds of preprocessing can be of use, without going into too much algorithmic detail. The same principle is used in the rest of the theory chapter, where different procedures such as

principal component analysis, partial least squares regression and k -nearest neighbours are described.

In the third chapter the underlying hypotheses, which the work is based on, are emphasised. The apparatus, software and methods used are described in more detail.

The fourth chapter summarises the results of some of the models developed in the course of this project, while the fifth chapter compares and to some extent appraise the models.

In the short concluding chapter, attention is brought to the most important results, and some of the challenges.

Chapter 2

Theory

2.1 Hyperspectral cameras and remote sensing

2.1.1 What is hyperspectral imaging?

Hyperspectral imaging is also known as spectroscopic imaging or chemical imaging, and is an imaging technique where both spectral and spatial information from an object is obtained [25]. Chemical information on both a macroscopic and microscopic level is available in a non-destructively way, without changing the composition of the sample [4].

A digital image is a matrix with intensity information recorded by a detector, often a charge-coupled device (CCD). It is often imagined as a 2-D function $f(x, y)$ where x and y are the coordinates and the amplitude of f is called the intensity of the image at that point [23]. An image point is also called a pixel [19].

A hyperspectral image is an image where each pixel contains not only one intensity value, but an entire spectrum [21]. This results in a 3-D data set (section 2.1.2). It is recorded by a hyperspectral camera (section 2.1.3). A hyperspectral camera can have hundreds of spectral bands, and differs from a multispectral camera which only have a handful of spectral bands [12].

2.1.2 Hyperspectral data

The 3-D data set recorded by a hyperspectral camera is called a data cube, hypercube or image cube. It has two spatial dimensions (x and y) and one spectral dimension (λ). This is illustrated in figure 2.1. The data cube is often large [17].

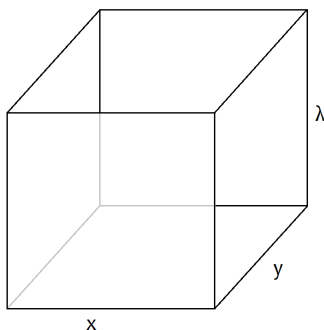


Figure 2.1: The data cube has two spatial dimensions (x and y) and one spectral dimension (λ).

By convention, independent variables are denoted X (not to be confused with the x coordinate above), while dependent variables, or response variables, are denoted Y [5]. For hyperspectral data, the X data is usually organised in a matrix with each object or pixel vertically and the different wavelengths or wavebands horizontally [13]. If there are only one dependent variable, the Y data will be a vector, whereas otherwise both X and Y will be matrices. This way of organising the data is ideal for many algorithms and data analysing tools [5], but destroys the image, that is, the X matrix normally does not include the spatial information from the data cube in figure 2.1.

2.1.2.1 Discretisation and digitalisation

An image is an attempt to represent the true world. While the real world (at least at most practical magnification levels) is continuous, a digital image is in its nature discrete. It is the discretisation of the image that makes it possible to store it as a computer file and carry out calculations on it [19].

Digitising the coordinate values is called sampling [23], and results in the pixels which the image consists of.

In a computer, all numbers (and everything else) is represented by a number of bits (binary digits). The higher the number of bits, the higher the resolution. The HySpex SWIR-320i camera used in this thesis, uses a 12-bit representation [39].

The discretisation limits the both the spatial resolution and the spectral resolution of the hyperspectral image [19]. This can in some cases cause unwanted effects.

Aliasing is an effect that causes different signals to become indistinguishable when sampled. This happens when the sampling rate is too low. In order to avoid aliasing, the sampling rate must be at least twice the highest frequency of the signal [46]. For hyperspectral imaging, this concerns both the spatial digitising of the coordinate values, or the pixels, and the distance between neighbouring spectral values.

2.1.2.2 Noise

The recorded spectra will always contain some noise. No instruments are perfect, and the observed signal x can thus be viewed as a sum of the true signal \tilde{x} and the noise e , as shown in equation 2.1 [5].

$$x = \tilde{x} + e \tag{2.1}$$

Noise can lead to loss of resolution, impede interpretation and complicate extraction of valuable information from the data [11]. Isolating the true signal from the noise is one of the main tasks of multivariate data analysis [13].

There are two main types of measurement noise: stationary noise and correlated noise [5]. For stationary noise, the noise at one point is independent of the noise at any other points. Stationary noise can be both *homoscedastic* and *heteroscedastic*. Homoscedastic noise have the same distribution, often the normal distribution, on the entire spectra. Heteroscedastic noise, on the other hand, is dependent on, and often proportional to, the intensity of the signal. This means that it will be largest around peaks.

When the noise level at a point depends on the noise level of the previous point, it is called correlated noise. For hyperspectral images, some

correlation between the noise from neighbouring pixels would be expected.

The signal-to-noise ratio (SNR) can be a useful parameter. It is, as the name proposes, defined as the signal divided with the noise, see equation 2.2. A large signal-to-noise ratio is, obviously, preferred.

$$SNR = \frac{\tilde{x}}{e} \quad (2.2)$$

2.1.2.3 Data compression

Data compression is a concentration of information [13]. Observed variables often contain common information. It is impractical and uneconomic to store this information twice. Algorithms that are able to operate directly on the compressed data will also save time [48].

Data compression can be either lossless or lossy. If the data compression are lossless, it is possible to reconstruct the original data from the compressed data. After lossy data compression, only an approximation to the original data can be recreated.

Hyperspectral images generally compress poorly using lossless techniques [17], although it can be achieved using techniques such as the integer Karhunen-Loève transform [38]. For lossy data compression, there are several possible techniques to use: wavelets [11] (section 2.4.1.5), splines, averaging (section 2.4.1.1) or latent variables such as principal components [48] (section 2.5). These lossy compression methods all aim to remove noise and increase the signal-to-noise ratio, which can give more accurate classification results [16]. However, if implemented poorly, lossy data compression can result in loss of important information.

2.1.3 Hyperspectral cameras

Different hyperspectral cameras, or imaging spectrometers, can be classified by how they record the hypercube. A whiskbroom scanner records the spectrum for only one pixel at a time, and can be categorised as a 0-D scanner [45]. A pushbroom scanner scans one line at a time, and is thus a 1-D scanner. Both whiskbroom and pushbroom scanners will traverse the scanning area to obtain an image, while a 2-D scanner is able to record an entire 2-D image simultaneously.

A whiskbroom or pushbroom scanner records the whole spectrum for each pixel or line at the same time, but because the detector arrays in image capture devices are, at most, two dimensional, they can only capture two dimensions at one time. Any additional spatial dimensions must thus be captured displaced in time [17]. This could be a source of error if the mechanical scanning operation is not entirely precise. An advantage with array devices, which detects light at several points simultaneously, is that they provide an uniformed background which give a high signal-to-noise ratio [52] (see section 2.1.2.2).

There are several different types of infrared focal plane arrays (IR FPAs). InGaAs, InSb, HgCdTe and quantum well infrared photodetectors (QWIPs) are some of the most widely used. InGaAs focal plane arrays have low dark current and noise, are able to operate at room temperature and are primarily used for applications requiring response in the 900-1700 nm near-infrared (NIR) region [52]. The cut-off wavelength can be extended to 2000 nm by varying the indium content.

Tunable filters (TFs) are common examples of filters used to separate out one wavelength or waveband at a time for recording in a 2-D detector. The time the apparatus needs for changing between wavelengths is called the tunability time. Ideally, this time should be as small as possible for faster data collection.

There are several different types of tunable filters [17]. Two of the most common are liquid-crystal tunable filters and acousto-optic tunable filters, both of which are variants of electronically tunable filters (ETFs). Electronically tunable filters are smaller, faster and have a larger spectral range than dispersive devices based on mechanical scanning [52].

Liquid crystal tunable filters (LCTFs) are build from a stack of polarisers and crystal quartz plates with a birefringent liquid. This makes the LCTF polarisation sensitive. LCTFs have usually a spectral resolution of several nanometres. The tunability time is limited by the relaxation time of the crystal, and is approximately 50 ms, although some devices can have a faster switching speed (≈ 5 ms) for a short sequence of wavelengths [17]. A NIR LCTF can typically be spectrally tuned from 1000 to 1700 nm, which coincides well with the InGaAs focal plane arrays cameras for the NIR region [52].

In an acousto-optic tunable filters (AOTFs), radio frequencies acoustic waves inside a crystal separates one single wavelength of light from a broadband source [17]. The wavelength of the filtered light can be controlled by

changing the frequency of the radio waves, which is applied to an array of LiNbO_3 piezoelectric transducers bonded inside the crystal [52]. For light frequencies from the near ultraviolet through the short-wave infrared region, crystals of tellurium dioxide (TeO_2) or mercury(I) chloride (Hg_2Cl_2) are used. The spectral resolution depends on the wavelength of the light, but can be as narrow as 1 nm full-width at half maximum (FWHM) [17]. The scanning speed is of the order of microseconds [52], which is significantly faster than for LCTFs.

An advantage with tunable filters is the ability to only record a subset of the available wavelengths. This can give faster data collection and a more compact data set [37]. A drawback with tunable filters is the fact that most of the light intensity is lost. At any given time, only one band of wavelengths passes through to the camera, while the rest of the emission spectrum is blocked by the filter [18].

2.1.4 Remote sensing of hydrocarbons

Remote sensing concerns, as the name suggests, investigating objects at some distance from the sensor without making physical contact with them [11]. Such techniques are much used in e.g. astronomy [15] and geology [54], where images from satellites or air crafts are used for mapping surfaces. Several studies have shown that it is possible to detect hydrocarbon oil seeps [12, 43, 53, 58] and materials containing hydrocarbons [28] by remote sensing.

Ellis et al. [12] successfully detected onshore oil seeps using an airborne sensor with a spatial resolution of 25 m^2 that recorded light from the visible through short-wave infrared part of the spectrum, divided into 128 wave bands. For calibration they build a spectral library including oil, tar, vegetation, soils and rocks, and the results were confirmed using a GPS and a portable hyperspectral sensor. Asphalt covered roads, plastic roofs and other hydrocarbon-based surfaces caused some confusion, and had to be identified and eliminated from the interpretation. This proves the importance of developing more selective models.

Hörig et al. [28] used an abandoned military training area in Berlin as a test field, with a parking lot, a lawn, trees and a gravel-paved area as reference objects. Both an airborne high resolution stereo camera (spectral range 385-2548 nm, 849 wave bands) and a HyMap scanner (spectral range 440-2543 nm, 128 wave bands) were used for recording the area. The hydro-

carbon bearing reference objects were characterised by absorption maxima at 1730 and 2310 nm. Plastic objects showed a sharp maximum at 1730 nm. This peak was less prominent in the spectra of oil bearing soils and rocks, but still significant enough for the materials to be detected by evaluating the HyMap spectra.

Based on previous studies showing that hydrocarbons have characteristic spectral signatures at 1730 and 2310 nm (see also section 2.2), Kühn et al. [34] developed an algorithm for airborne hyperspectral remote sensing of hydrocarbons. This *Hydrocarbon Index* (HI) uses one wavelength on each side of the 1730 nm absorption feature in addition to the intensity at 1729 nm. To demonstrate the HI, materials common for urban areas (concrete, grass, sand, plastic etcetera) were used as reference. The most accurate HI images were obtained for urban areas and bare ground, while the areas covered by vegetation appeared noisy. Dark-coloured hydrocarbons were not detected reliably.

Sanches et al. [43] aimed to use hyperspectral remote sensing for assessing the impact on vegetation of hydrocarbon leakages from the Brazilian pipeline system. Three planting slots were prepared and watered daily, while forced flows of gasoline and diesel were applied to the soil. Spectral measurements started a week prior to the first contamination, and were obtained using a portable high-resolution spectroradiometer detecting electromagnetic radiation in the spectral range 350-2500 nm. The results showed that contaminated brachiaria plants (*Brachiaria brizantha*) could be spectrally distinguished from healthy plants.

Detection of both off shore oil spills and naturally occurring petroleum hydrocarbons are of importance for the petroleum industry. Natural petroleum hydrocarbon seepage in the marine environment occurs where oil and gas leak from the seabed and rise to the surface, and are important for petroleum exploration. Both oil type and film thickness influence the reflectance spectra. Wettle et al. [58] used both a HyMap and a Quickbird sensor for detecting naturally occurring oil seeps. While the HyMap sensor was able to detect oil films with a thickness of $10^{-5} - 10^{-4}m$ depending on the oil type and its optical properties, the Quickbird sensor was unable to detect oil films in naturally occurring thicknesses.

2.1.5 The PryJector

The PryJector is a device for easy *in situ* visualisation of the chemical properties of a surface [2]. By combining a hyperspectral camera with a computer and an ordinary computer projector, the PryJector aims to make chemical information visible and easily accessible to humans.

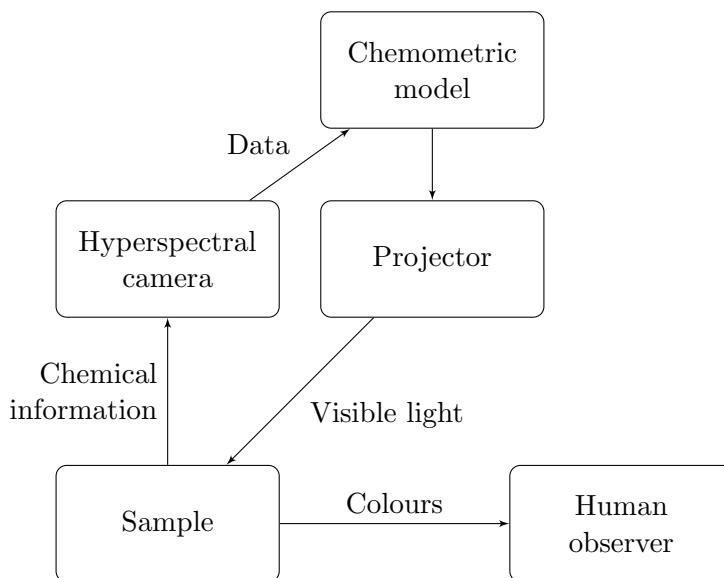


Figure 2.2: Schematic model of the idea behind the PryJector.

The key elements of the PryJector is shown in figure 2.2. Chemical information about a sample is recorded by a hyperspectral camera and sent to a chemometric model which analyses the data and aims to predict information about the composition of the sample. This information is coded to different colours for the different chemical compounds the PryJector recognises, and sent to a projector that projects the information back onto the sample as visible light, which can then be observed – and understood – by a human observer.

The current set up for the PryJector (see also section 3.2 or [2]), uses a pushbroom scanner (section 2.1.3) with an InGaAs focal plane array, and has a spectral range of 923-1665 nm.

The possible uses for the PryJector are many and varied. So far, only a couple of applications have been investigated: separating different kinds of pharmaceuticals used for pain relief [2] and detecting bones on surfaces

consisting of wood, sand and stone [3]. This thesis aims to develop a model for using the PryJector to detect hydrocarbons on surfaces such as soil, sand, stones and vegetation.

2.2 NIR-spectra of hydrocarbons

The Near-Infrared (NIR) region is the part of the infrared (IR) region of the electromagnetic spectrum that lies closest to visible light in energy. The exact definition in terms of wavelengths varies. In this thesis, the hyperspectral camera used operates in the area 923-1665 nm, and the focus will be on this part of the spectrum.

The NIR region is largely dominated by overtones of X-H stretching modes such as O-H, C-H, S-H and N-H, but there is also a large number of combination bands [6]. The fact that overtones of C-H bonds is so prominent in the NIR region, makes NIR analysis very useful for analysing hydrocarbons.

Cloutis [9] marks the C-H stretching overtones and combination bands near 1700 nm and the overlapping combination and overtone bands at 2200-2600 nm as the most promising regions in which to search for organic absorption bands. Both of these areas are outside the recording interval of the hyperspectral camera [39] used in this thesis (more in section 3.2), the first only barely. Kuhn et al. [34] focuses on the 1730 nm area when developing their *Hydrocarbon Index* (HI) for detection of hydrocarbons (see also section 2.1.4).

The C-H stretching vibration at 3500 nm has its first to fourth overtone at 1750, 1150, 880, and 700 nm [6]. The only true C-H stretching overtone in the area of interest is thus the second overtone at 1150 nm. The area has, however, also several combination bands that might be of interest.

Near-infrared chemical imaging (NIRCI) differs from near-infrared spectroscopy in that the sample is spatially heterogeneous. While it can be argued that – depending on the magnification – no sample is spatially homogeneous, in NIRCI this heterogeneity is exploited [37]. The chemical image is a direct consequence of the variation in chemical composition between the pixels. Instead of a single spectra, the chemical image consists of one spectrum in each pixel, and the total number of spectra can be very large.

2.3 Chemometric principles

2.3.1 Classification and regression

Classification and regression are two main types of problems. A regression problem relate each independent variable to one or several numbers, whereas a classification problem relate each independent variable to one of two or more categories or classes. If there are only two classes, it is called a Boolean classification problem [42].

Multiple linear regression (MLR) [57], is a linear regression model shown in equation 2.3. X is here the matrix with independent variables, Y is the matrix with dependent variables (a vector if there is only one dependent variable), β is a vector with parameters and ϵ is a vector of residuals.

$$Y = X\beta + \epsilon \quad (2.3)$$

The least squares solution b to this problem is shown in equation 2.4 [20], where X' is the transpose of the matrix X . This solution exists only when $X'X$ is assumed to be a nonsingular matrix [57], as it involves calculating the inverse of this matrix.

$$b = (X'X)^{-1}X'Y \quad (2.4)$$

However, for hyperspectral data, the X matrix tend to be close to singular, and therefore other regression methods are needed. Principal component regression (PCR, section 2.5) and partial least squares regression (PLSR, section 2.6) tend to perform well on hyperspectral data.

Classification methods creates a decision boundary which divides the variable space into regions belonging to each of the classes. Linear classification methods use linear decision boundaries, while other classifications methods, can handle more complex decision boundaries [1]. k -nearest neighbours (k -NN, section 2.7) is such a technique, where the decision boundary is not calculated explicitly.

Depending on the labelling of the classes, it will in many cases be possible to use a regression method for classification. Discriminant partial least squares regression (DPLSR, section 2.6) is an adjusted version of PLSR that can be used for classification.

2.3.2 Calibration and validation

Model development consists usually of two steps: the calibration step and the validation step [20]. Model parameters are estimated in the calibration step, while the validation step investigates how the model performs.

In the calibration step, one or more sets of variables are connected together [5]. Usually, the aim is to connect the independent variables to the dependent variables (section 2.1.2). To do this, a model is fitted to the data [13], and the parameters of the model is adjusted in order to describe the data as good as possible.

The data used in the calibration step is called the calibration data or the training data [5]. The composition of this data set is of high importance. If the model are to be used for prediction on new samples later, it is crucial that the calibration data set is representative of the data in such a way that it spans all the variations expected in the future samples [13]. There will always be a trade-off between time and money on one side, and the future predictability of the model on new data on the other side [5].

After calibration, the variation in the data can be expressed as a sum of a modelled part and a residual part, which hopefully is dominated by noise [13] (section 2.1.2.2).

Validation is a tool to estimate the prediction error, which is the expected error when the model is used for prediction on new samples. The optimal solution is to validate the model by testing its performance on an independent test, or validation set, with new samples not included in the training set used for development of the model. This is, however, not always possible due to time, economical or practical concerns.

If the data set is too small to be divided into two independent data set, cross-validation may be used. Cross-validation is, however, never as good as independent test set validation [13]. Cross-validation separates a data segment from the main data set. The main data set is used for creating a model, while the smaller data segment is used for validation. After the prediction error has been estimated, the data sets are combined before extracting a new data segment and repeating the process until all data segments have been used for validation once [14]. A special case of cross-validation is leave one out (LOO) cross-validation, where only one data point is extracted. Also other resampling methods, e.g. bootstrapping [1], can be used.

A hyperspectral data set is often large enough to be divided into two independent data sets: a calibration data set and an independent validation data set. It is important that both data sets is from the same population, and that they both span the space of variation where the model will be used in the future. To ensure that they are similar in this way, randomisation can be used to decide which data points should be in which data set.

For hyperspectral data, it is worth noting that a sample and a data point might not be the same thing. Hyperspectral data is a discretisation of presumed continuous data in two spatial directions (and one spectral direction). Because of this, it is somewhat random where one pixel end and another begin, and it is not fair to treat the neighbouring pixels as independent samples. Having one pixel in the calibration data set and its neighbour in the validation data set will make these data sets very similar to each other, and it might be better to get calibration and validation data sets from different images, or different regions in the images.

Regardless of the method used for validation, the percentage error in the validation step should be similar to the percentage error of the training set [5]. If the error for the training set is significantly lower than the validation error, the model is likely overfitting to the data.

2.3.3 Overfitting and underfitting

When creating a model from data, it is usually possible to create a more complex model with a smaller calibration error. While this might be tempting, the danger of overfitting is large. Overfitting means that the model explains too much of the variance in the calibration data set - it includes phenomena which are not relevant for the entire population, but are rather unique aspects of the calibration data set [1]. Because the overfitted model also describes noise present in the calibration data, it will usually fail to predict new objects with optimal accuracy [13].

Underfitting is the opposite of overfitting, and means that it is possible to create a better model that explains more of the systematic variance in the calibration data, and thus have a smaller error. There are interesting aspects to the data that is not included in the model, and a more sophisticated model would be preferred.

The validation error gives an indication of whether the model overfits or underfits to the data, and validation is a useful tool to prevent overfitting [13]. When choosing between different models, Ockham's razor provides a useful,

and in some ways intuitive, principle: the *simplest* hypothesis consistent with the data, is preferred [42]. Thus, a more complicated model is only desirable if it significantly lowers the validation error compared to a simpler model.

2.4 Preprocessing

Preprocessing is the term used for several different kinds of mathematical transforms of data. The goal in this part of the data analysis is to remove irrelevant sources of variation in the data [1], and ensure that all objects are described in a consistent way [7].

There are different types of preprocessing tools or transformations. Some are performed separately on each object/pixel, while others use information from all the objects (the whole population).

2.4.1 Smoothing

Smoothing aims to make the data more continuous by removing high frequency components in the signal, which are assumed to be random noise (see section 2.1.2.2).

2.4.1.1 Averaging

Averaging removes noise while reducing the total number of variables or objects in the data set. It is often used for exchanging replicates present in the data set with one mean value [13]. For sufficiently large populations, the central limit theorem [57] states that the mean will be normally distributed with variance equal to the variance of the objects divided by the sample size. This results in a higher signal-to-noise ratio.

For hyperspectral data, averaging in the spectral direction means that random noise will influence the data to a lesser degree, and the spectra will seem smoother, while the number of variables drastically decreases. Averaging two and two wavelengths will for instance give only half of the original data size, while averaging three and three data points will return a vector of a third the length of the original data vector.

Averaging in one or more spatial directions is also a possibility, and will increase the pixel size. As with averaging in the spectral direction, it might make the data look smoother up to a point, but from then on further averaging will make the individual pixels too large to be interesting.

In a way, hyperspectral data are already a result of averaging in both the spectral and spatial directions. For the hyperspectral images recorded in this work, each pixel represents an area of $1 \text{ mm} \times 1 \text{ mm}$ in the spatial direction, and each waveband represents 5 nm in the spectral direction. Each data point is thus an average value for all the light reflected from a given area with a wavelength in the given region.

2.4.1.2 Moving average

Moving average (MA) is a rather primitive form for smoothing, where the resultant smoothed data are a linear function of the raw data, and the only input given, except from the vector to be smoothed, is a window size. This window size must be an odd number. The window then moves over the vector, averaging the data it covers and assigning this value to the middle data point [5]. This smoothing technique will broaden peaks, which often is unwanted.

On hyperspectral data, it is possible to perform moving average smoothing both in the spectral and in one or more spatial directions. Smoothing in one or more spatial direction will make the picture more "blurry", and remove high contrasts and high frequencies. When smoothing the spectrum, "spiky" noise is in some ways restricted, but might still influence the spectrum (and will influence a broader part of the spectrum). A similar technique called *median filtering* uses the median instead of the mean, and is thus less influenced by "spiky" noise. This is illustrated in figure 2.3, which compares the effect of moving average and median smoothing. While the thin spike (at $x = 20$) influences a broad part of the moving average smoothed spectrum, it vanishes completely in the median smoothed spectrum.

Moving average is in many ways similar to averaging (section 2.4.1.1 above), but the size of the data is preserved. The loss of resolution is thus smaller for moving average. However, for hyperspectral data the amount of data is often large enough to make a decrease in the amount of data in exchange for a higher signal-to-noise ratio a good thing.

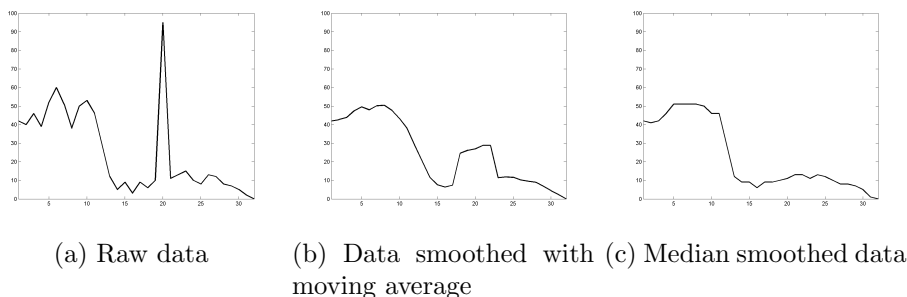


Figure 2.3: The effect of moving average and median smoothing with window size = 5.

2.4.1.3 Savitzky-Golay smoothing

Savitzky-Golay smoothing is a highly popular technique in chemistry [55]. The method is based on performing a convolution of the input signal by a suitable vector which controls the amount of smoothing. This vector can also be modified to perform differentiation by choosing other coefficients. Coefficients for smoothing and differentiation can either be calculated, or found in literature [5].

Equation 2.5 [55] gives the transformation of a vector component x_j where x_j^* is the new value, N is a normalisation constant, c_h are the coefficients and k is the number of neighbouring values at each side of j . The window size is thus $2k + 1$.

$$x_j^* = \frac{1}{N} \sum_{h=-k}^k c_h x_{j+h} \quad (2.5)$$

The Savitzky-Golay method smooths the data by locally fitting the data to a polynomial. This will in many cases give a better fit than averaging (section 2.4.1.1) or moving average (section 2.4.1.2), especially around peaks [5].

Performing regression for each moving window in Savitzky-Golay smoothing would be a computationally intense and time consuming operation [5]. Fortunately, the solution can be reformulated as a convolution [55], which considerably speeds up the process. This alternative and simplified strategy is one of the main advantages of the Savitzky-Golay method, and gives an exact solution to the problem.

As for moving average, an odd window size must be defined. A larger window size gives a smoother signal and removes more noise, but at a higher risk of blurring the signal [5]. Because the Savitzky-Golay method smooths the data by fitting the data to a polynomial, the degree of the polynomial must also be predefined. This degree must be lower than the window size.

2.4.1.4 Fourier transformation and low pass filtering

The Fourier transformation changes the domain of the data from the spectral domain to a frequency domain. This makes it possible to filter out unwanted frequencies. A low pass filter will keep only the low frequencies of the signal, while a high pass filter will keep the high frequencies.

For spectral data, the unsystematic noise is usually dominated by components with a higher frequency than the signal [11]. Because of this, a low pass filter will remove noise from the spectra, and the resulting spectrum will appear smoother. It is, however, important to keep in mind that if the signal has thin spikes, they will also be removed by this procedure.

Different kinds of low pass filters exist. The simplest, box filters, remove all frequencies above a cut-off frequency. This is also called hard thresholding. Other filters, such as gaussian filters, let different frequencies influence the result to different degrees. This is called soft thresholding [11]. Figure 2.4 illustrates the difference between these filters.

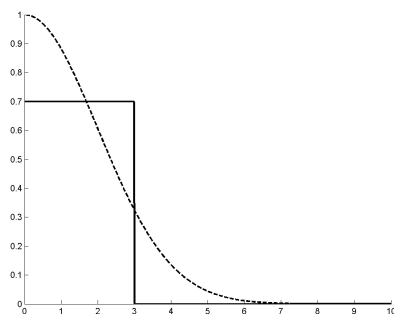


Figure 2.4: Examples of box filter (bold) and gaussian filter (dashed line).

2.4.1.5 Wavelets

Wavelet transforms involve fitting a spectrum to a series of functions based upon a basic shape called a wavelet [5]. In contrast to the smooth, periodic sinusoids of the Fourier transform, the wavelet functions are non-smooth and non-periodic [11]. This gives wavelets the advantage of being able to represent both smooth and locally bumpy functions with spikes in an efficient way [26].

A wavelet, $g(t)$, is usually a function of frequency and time, that add to zero [5]. First generation wavelets are translated or dilated versions of a single basic wavelet called the mother wavelet [11]. The choice of mother wavelet depend upon the problem at hand. Second generation wavelets, first introduced by Sweldens [49], uses the lifting scheme to create new custom designed wavelets. Second generation wavelets differ from first generation wavelets in that they not necessarily are translates or dilates of one fixed wavelet function [11].

There are several examples of wavelet transforms in literature. Wavelet transforms may make it possible to represent the entire spectra as a sum of only a few significant wavelets without losing too much information [5]. This can result in both data decompression (section 2.1.2.3) and denoising. Denoising can be achieved by using thresholding (see section 2.4.1.4 above) to remove the wavelet coefficients that corresponds to high frequencies [11].

2.4.2 Multiplicative scatter correction

Multiplicative scatter correction (MSC), also known as multiplicative signal correction, is a method developed to reduce the disturbing effect caused by light scattering for NIR data obtained by reflectance or transmission measurements of diffuse samples [55]. The difference in scattering between samples can be large, and represent a challenge for comparing different spectra [30].

MSC aims to simultaneously correct for two undesired scatter effects: multiplicative scatter effects (amplification) and additive scatter effects (offset) [13].

In the MSC method, a linear regression of the spectral variables against a reference spectrum is performed [33]. The mean spectrum is often used as a reference spectrum. The regression model is shown in equation 2.6 [30]

where the spectrum x_k is fitted to the reference or mean spectrum \bar{x}_k , and e_k is the error.

$$x_k = a + b\bar{x}_k + e_k \quad (2.6)$$

a represents the additive effect and b the multiplicative effect [55]. The corrected spectrum x_k^{MSC} is calculated from the original spectrum x_k and the regression parameters a and b , as shown in equation 2.7 [30].

$$x_k^{MSC} = \frac{x_k - a}{b} \quad (2.7)$$

It is also possible to adjust for only the multiplicative or the additive effect [55].

MSC traditionally uses the full spectral range for the regression [33]. Before any future prediction, all new samples must be corrected using the same reference spectrum as used for the calibration data set [13].

2.4.3 Centering, scaling and autoscaling

To center the data is to translate the coordinate system in such a way that the origin coincides with the mean of the data (also called the *mean center* [13]). This is done by subtracting the mean of each variable, \bar{x}_j , from the data x_{ij} , as shown in equation 2.8 [5].

$$x_{ij}^{centered} = x_{ij} - \bar{x}_j \quad (2.8)$$

It is common to center the data before further analysis. Principal component analysis (section 2.5) and partial least squares regression (section 2.6) almost always starts with automatically centering the data [5].

Scaling, or weighting, adjust the scale of the variables, and can be used to put the variables on a similar scale [5]. A particular common scaling factor is the inverse of the standard deviation of the variable. This is shown in equation 2.9 [13], where s_j is the standard deviation of variable j . Scaling by the standard deviation is also called standardisation or variance scaling [20]. Standardisation is most useful when the magnitude of the variables differ significantly [5].

$$x_{ij}^{standardised} = \frac{x_{ij}}{s_j} \quad (2.9)$$

Scaling can also be used to control to which extent different variables should be able to influence the model. Certain variables could be scaled up or down in order to make a larger or smaller contribution to the model [20].

Autoscaling gives all variables equal importance by both centering the data and dividing by the standard deviation [5], thus combining equations 2.8 and 2.9.

2.4.4 Normalisation

Normalisation is a much used preprocessing procedure which tries to reduce systematic differences between the observations [7]. Normalisation is performed separately on each object, and re-scales each object by dividing with a scaling factor. This scaling factor is often a common sum, usually 1 or 100 % [13]. To normalise the data, the sum of all the variables for a given object is calculated, and the variables is then divided by this sum. This is called total sum normalisation [7] to differentiate it from other normalisation techniques such as total area normalisation or normalising to the vector norm.

Some refer to normalisation to a vector sum, as described above, as *row scaling*, while normalisation is reserved for vectors whose sum of squares of the elements equal one [5].

Normalisation ensures that all observations in the data set is represented in an adequate and consistent way, which is a crucial step in preprocessing of data [7]. In hyperspectral imaging, normalisation can help adjust for uneven surfaces. An uneven surface will impact how much of the light is reflected, and normalisation can remove this difference between the individual pixels.

2.4.5 Logarithm

A logarithmic transformation may remove the skewness in the distribution of data [13]. For hyperspectral data, logarithmic transformations are often useful when there are large variations in intensities [5]. Logarithms with

different bases may be used, and both X and Y data may be logarithmically transformed. A logarithmic transformation is a commonly employed strategy for eliminating heteroscedastic noise (see also section 2.1.2.2), but the success of this procedure depends on certain characteristics of the heteroscedastic noise [35].

Due to the definition of the logarithm, it is impossible to take the logarithm of zeros or negative numbers. This can be dealt with by setting such numbers to a small, but positive number [5].

2.4.6 Numerical differentiation

Differentiation will often prove useful for removing systematic variance in the data. The first derivative of the data will be without any constant terms from the original data, while any linear terms will be removed from in the second derivative.

Differentiation might extract important information from the data, but will increase the noise [55]. As the noise increase for each differentiation, it is rare that derivatives above the second are used.

As mentioned in section 2.4.1.3 above, the Savitzky-Golay method can also be used for numerical differentiation.

2.5 Principal component analysis

Principal component analysis (PCA) is a method used for multivariate data. It is especially useful if there are more variables than samples, as multiple linear regression (section 2.3.1) cannot be used in those cases [20]. The principal components (PCs) are latent variables, which are linear combinations of the original variables [1].

In PCA, the independent data matrix X with rank a is written as a sum of a matrices, each with rank 1. These matrices can also be written as outer products of two vectors, a score vector t_i and a loadings vector p'_i . This is shown in equation 2.10, where the matrix X is written as a sum of a such vector products [20].

$$X = t_1 p'_1 + t_2 p'_2 + \dots + t_a p'_a \quad (2.10)$$

The NIPALS algorithm (Nonlinear Iterative Partial Least Squares [20]) for PCA often starts by centering the data, and then iteratively extracting one principal component at the time [13]. The principal components are chosen in such a way that the first principal component explains as much of the variance in the data set as possible, the second principal component explains as much of the variance left in the data set as possible while being orthogonal to the first principal component [20], and so on.

PCA can be viewed as creating a new coordinate system with the origin in the center of the data, and the coordinate axes in the direction of the maximum variation in the data set [13]. It is apparent that a crucial part of PCA is to determine the number of principal components, or coordinate axes, to investigate. To explain *all* the variance in the data set, it will be necessary to calculate a principal components where a is the rank of the X data matrix, which could be as many as there are original variables (provided there are at least as many samples as variables in the data set).

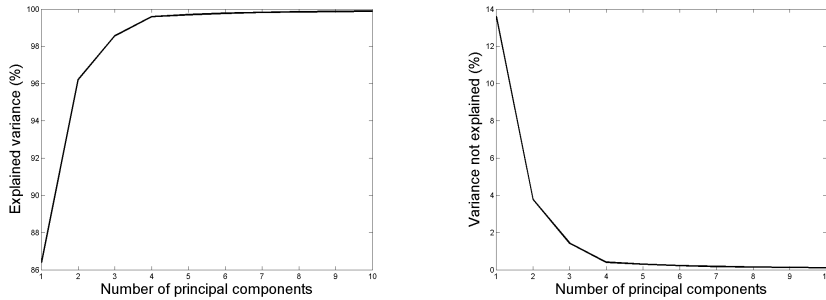
The rank of a matrix is a mathematical concept related to the number of independent sources of variation in the data set [13]. Ideally, this should be equal to the number of chemical compounds present in a mixture [5], but the noise present in the data set will usually cause the mathematical rank to appear larger than the chemical rank.

For this reason, the last principal components will mainly describe random error present in the data set, and it is neither necessary nor advantageous to calculate all the principal components. Figure 2.5 shows an example of how the explained variance can depend on the number of principal components. Four principal components explains 99.6 % of the variance in the data set, and would seem a sensible number of components to choose.

The variance left in the data set, not described by the principal components, is called the error of the model. This is shown in equation 2.11, where the data X is written as a product of a score matrix T and the transpose of a loadings matrix P [13]. These matrices are composed of the a first score and loadings vectors from equation 2.10. E is the error not explained by the a first principal components.

$$X = TP' + E \quad (2.11)$$

The score matrix T , gives the coordinates of the objects projected onto the new coordinate system. The loading matrix P contain information



(a) Explained variance (%) as a function of the number of principal components

(b) Variance not explained (%) as a function of the number of principal components

Figure 2.5: Choosing the number of principal components.

about relationships between the variables, and which variables are the most important for explaining the variance in the data set [13].

PCA can give valuable information about the underlying structure in the data, but in order to be useful for prediction, principal component regression (PCR) [13] is needed. Equation 2.12 shows how regression on the principal components can be done in an analogous way to equation 2.3 for multiple linear regression, but using the score matrix T instead of the data matrix X [20].

$$Y = TB + E \quad (2.12)$$

One of the main disadvantages with the MLR method (section 2.3.1), is the colinearity problem. If the matrix $X'X$ is singular, its inverse cannot be calculated, and no least squares solution (equation 2.4) may be found [57]. PCR avoids this problem by requiring the score vectors to be orthogonal to each other [20]. Equation 2.13 will thus always have a solution \hat{B} .

$$\hat{B} = (T'T)^{-1}T'Y \quad (2.13)$$

If the principal components are to be used for regression, the number of principal components chosen is crucial for the validity of the model, and must be chosen with care to avoid both over- and underfitting (see section 2.3.3). The optimal number of principal components for PCR will generally differ from the optimal number for PCA, because PCR lets the prediction

ability of the model determine the number of components [13]. Validation (section 2.3.2) is used to determine the number of principal components that maximises the prediction ability of the model. In some circumstances, the relevant variation is described by the last few, and not the first few, principal components [32]. This is, however, a rare case, and will not be discussed further.

PCA is a technique used in several areas under different names, such as Karhunen–Loève expansion, eigenvector analysis, the Hotelling transform or correspondence analysis [1]. It is closely related to singular value decomposition (SVD), which can be viewed as PCA where all the principal components are calculated.

2.6 Partial least squares regression

Partial least squares regression (PLS regression or PLSR) [20] is a much used technique in multivariate data analysis and chemometrics, and in particular for data from near-infrared spectroscopy [5]. It is in many ways similar to PCA (section 2.5), in that new latent variables are created as linear combinations of the original variables.

The main difference between PCA and PLSR lies in how the latent variables are chosen. PCA only uses information from the X matrix, and creates latent variables (or principal components) in such a way that they maximise the variance in the X matrix (independent variables) explained by the principal component. PLSR, on the other hand, also uses the information provided by the Y matrix (the variables dependent on X) and creates latent variables in such a way that it maximises the X - Y covariance and minimises the prediction error [13]. For this reason, PLSR usually needs less components than PCR for describing the structure in the X data relevant for predicting Y .

PLSR can be considered as consisting of two independent PCA decompositions, one for each of the X and Y block, and an inner relation linking the blocks together [20]. As shown in equation 2.14, the outer relation for the X block is similar as for PCA (see 2.11).

$$X = TP' + E \quad (2.14)$$

$$Y = UQ' + F \quad (2.15)$$

The Y block is decomposed in a similar way (equation 2.15), with U corresponding to T , Q corresponding to P and the resulting error after a factors denoted F [13]. However, these matrices must not be confused with the score and loading matrices from PCA. Methods such as MLR (section 2.3.1) and PCR assumes that the all error is in the X data, but PLSR takes into account that there can be errors in the Y data as well [5].

The key point of PLSR is link between the decompositions of X and Y . Instead of creating independent PCA models, the score from the decomposition of Y is used as a starting point for the decomposition of X – and vice versa. By letting U and T change places in this way, the decompositions gain information about and influence each other [20].

PLSR creates latent variables that both has a high variance and a high covariance with the response, in contrast to PCR which only maximises the variance in X [26]. Because the measured data never will be completely noise-free, only a subset of the principal components are used [20].

Two different PLSR routines exist: PLS1 and PLS2. The conceptual difference between them is minor, but PLS2 allows for more than one dependent variable [5], that is, for a Y matrix and not just a y vector (section 2.1.2).

In much the same way that choosing the number of principal components is a crucial step when performing PCR, the number of PLS components is a very important property of the PLSR model and must be chosen with great care [20]. The number must be large enough to explain most of the relevant variation in the data set, but low enough to avoid overfitting to the data (see section 2.3.3). This can e.g. be done by inspecting a plot of how the error changes for different number of components. Both visual inspection and more mathematically stringent methods, such as an F -test [57], can be employed.

One of the advantages with PLSR is the ability to calculate which of the objects and variables contribute mainly to the model, and which contribute mainly to the residual [20]. This can be visualised by plots of scores or loadings, and utilized to detect anomalies and outliers [13].

For classification problems, a version of PLSR called discriminant partial least squares regression (DPLSR) is much used in chemometrics [5]. In DPLSR, the Y -matrix have discrete values. For example, the elements can be coded as 1 if the object is a member of the class, and 0 otherwise. If there are k classes and n objects, this will result in a boolean $n \times k$ matrix for the calibration data. When performing prediction by DPLSR, an object is

more likely to belong to a class the closer the estimated y value is to 1 [5].

2.7 k -nearest neighbours

k -nearest neighbours (k -NN) classification [42] is a prototype method. In a prototype method, the training data are represented by a set of points which may not be examples from the training data set. The main challenge with the k -NN method is to determine how many and which prototypes to use [26].

Nearest neighbour methods is based on the assumption that an object is likely to be of the same class as the objects in the neighbourhood, that is, with similar x values [42]. In k -nearest neighbours, this neighbourhood consists of k points.

To decide a neighbourhood, some kind of distance metric is needed. The Euclidean distance d between points A and B in i dimensions is shown in equation 2.16 (generalised from [56]).

$$d(A, B) = \sqrt{\sum_i (x_i^B - x_i^A)^2} \quad (2.16)$$

The Euclidean distance is ill-suited when the variables have different scales. One solution to this problem is to standardise the variables (section 2.4.3) [42]. It is also possible to use another metric such as the Mahalanobis distance [5], which is similar to Euclidean distance but includes the covariance of the features as a scaling factor. If the variables are discrete features, the Hamming distance defines the distance to be the number of features on which the objects differ [42].

Each point is assigned to the class that the majority of its k nearest neighbours belong to. In the case of a tie, the class is usually chosen randomly between the tied classes [26]. When there are only two classes, ties can be avoided by choosing k to be an odd number. For large data sets an efficient way of finding the nearest neighbours are necessary, as calculating the distances between all points take far too long. Preprocessing of the training data may make this step more efficient [42].

Different values of k will give different results. Performing k -NN for a number of values of k can be used to spot anomalies or artefacts [5]. The

best value for k can be chosen using validation [42]. Cross-validation (section 2.3.2) is often used for this.

The k -nearest neighbours algorithm has been proved successful for a large number of classification problems. Because it is a prototype method, it handles irregular decision boundaries well [26]. The method is conceptually very simple [5], and easy to implement [42]. High-dimensional spaces may pose a problem, because the nearest neighbours sometimes are too far away for the method to be trusted.

The composition of the training set is essential for the performance of the model. If the number of objects in each class are not approximately equal, the model will be biased towards the class with the most representatives in the training set [5].

The k -NN method assumes that all variables are of the same importance. This is usually not true for spectroscopic data. Taking correlation between wavelengths into account by using the Mahalanobis distance metric will to some extent avert this problem [5].

Chapter 3

Experimental

3.1 Hypotheses

The main hypothesis of this thesis is that the near-infrared spectrum of hydrocarbons (e.g. paraffin) differs from the spectra of surfaces common in nature, e.g. soil, sand and vegetation. This difference is detectable by hyperspectral cameras, and as a consequence of this, chemometric methods can be used on hyperspectral images to create a model that discriminates between areas with hydrocarbons and areas without hydrocarbons.

Assuming that the main hypothesis holds, five additional hypotheses will be investigated:

1. Because water has a very strong absorption spectrum [59], presence of water will affect the model. For this reason, both wet and dry samples should be included in the data set.
2. Over time, chemical and biological reactions will occur [44]. This will cause the accuracy of the model to depend on how much time has passed after the hydrocarbons were added to the surface.
3. The amount of hydrocarbon present will influence the spectrum. On the same background, it should be possible to separate areas with different amounts of hydrocarbons.
4. Naturally occurring hydrocarbons or similar compounds (e.g. organic matter such as dead plants and animals) will give false positives. In a

more sophisticated model, this error should be minimised, but there is likely to be a trade-off between a high degree of detection and a high number of false positives.

5. Different hydrocarbons have different spectra. This means that it should be possible to create a selective model for target hydrocarbons. It should also be possible to create a model that discriminates between similar alkanes of different chain lengths, such as *n*-hexane and *n*-heptane, on some surfaces.

3.2 The set up

Figure 3.1 shows the current set up of the PryJector [2]. It consists of a sample table, a NIR source, a filter for blocking out visible light, a hyperspectral camera and a projector. The camera and the light source are mounted on a translator moving parallel to the sample table.

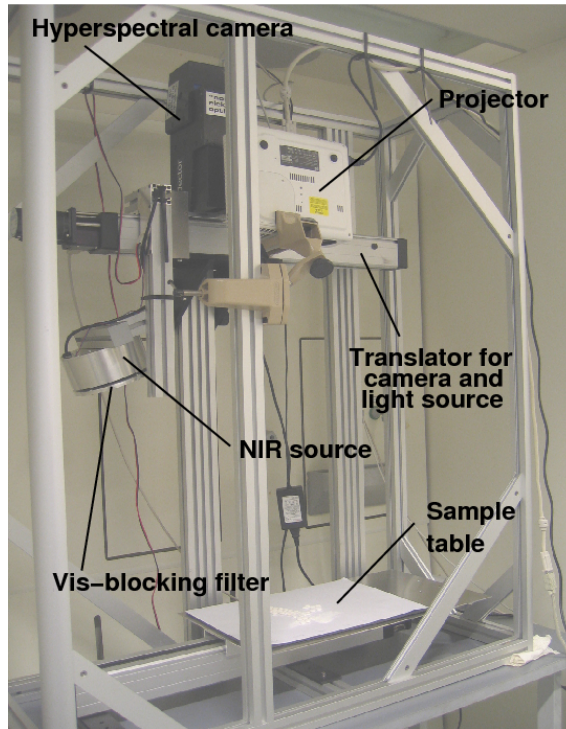


Figure 3.1: The PryJector. Illustration by Alsberg [2].

The light source used to illuminate the samples is a 150 W lamp which has a light intensity in both the visible and infrared region (400-2500 nm). To help enhance contrast when performing projection, the visible light is blocked out by a long wave pass filter with a cut-on at 850 nm, which almost completely blocks out all wavelengths below 800 nm [2].

The hyperspectral camera is a HySpex SWIR-320i from Norsk Elektro Optikk AS [39]. It is a pushbroom scanner with 320 pixels across and an InGaAs focal plane array [2]. The spectral range is 923-1665 nm with 148 different wavelengths approximately 5 nm apart. The wavelengths are given in appendix A. The camera is mounted at a distance of 100 cm from the sample table, giving a pixel size of 0.75 mm. The scan rate is 100 fps and the translation device moves at a speed of 7.5 cm/s.

The projector is an ordinary colour computer projector (Hewlett-Packard MP3222 with XGA, 1024×768 and 2000 lumens; Hewlett-Packard Co., Palo Alto, CA). When the PryJector is used in projection mode, the projector will project a chemical image with false colour onto the sample. This chemical image is updated continuously as the camera records new lines of the image [2].

In this thesis, the PryJector is used mainly for recording images. Although the projection feature were used to demonstrate and explore the limitations of an early model, this is not described further here.

3.3 Hyperspectral images recorded

Several hyperspectral images were recorded, in five phases. In the first phase (section 3.3.1), the main goal was to investigate the validity of the main hypothesis, that is, whether or not it was at all possible to create a model with the ability to distinguish paraffin and soil from soil without paraffin. Samples of water and soil were added to adjust for the effect of water (additional hypothesis 1). In addition to this, scans of the same samples were taken over the period of a year in order to test the time dependence of the model (additional hypothesis 2).

In the second phase (section 3.3.2), soil with different concentrations of paraffin were prepared to investigate additional hypothesis 3.

The third phase (section 3.3.3) includes a number of different surfaces, both inorganic and organic. This is both to make the model more robust,

and to test whether organic matter will result in false positives (additional hypothesis 4).

The images in the fourth phase (section 3.3.4) both includes different backgrounds (additional hypothesis 4) and explores the time dependence of the model further (additional hypothesis 2).

The last phase (section 3.3.5) introduces different hydrocarbons (additional hypothesis 5). Hexane and heptane were applied to soil, and attempts are made at distinguishing between them.

Two types of soil were used: Soil bought from *Plantasjen*, and compost soil from a private garden in Trondheim. In addition to this, leaves, plants, lichen, cones and moss from the area around Kyvannet, sand and humus from near Lianvannet and stones and grass from near Gløshaugen were used. These samples were gathered during the fall of 2012.

3.3.1 The effect of time

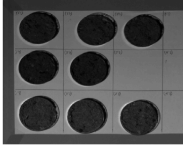
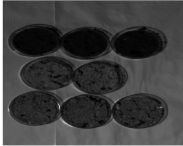
Eight Petri dishes were prepared in this first phase of the experimental work: three with pure soil (from *Plantasjen*, about 6 g soil/dish), two with a mix of soil and water (0.8 ml water/g soil, about 9 g soil/dish) and three with a mixture of soil and paraffin (0.8 ml paraffin/g soil, about 9 g soil/dish). Scans were taken over the period of a year, to investigate the possible effect of time (additional hypothesis 2). Table 3.1 gives an overview over the hyperspectral images recorded in this phase of the experiments. Water and soil were added in accordance with additional hypothesis 1 in section 3.1.

Soil and paraffin were mixed in a beaker by using a spoon. This resulted in a heterogeneous mixture. Soil and water were mixed in the same way.

The Petri dishes were placed in room temperature between the scans, and had lids on. In spite of this, the two dishes with soil and water (middle row in the mean images) would visibly dry out and crack over time. The samples with only soil (bottom row) would also crack, as can be seen in the last mean image.

These scans were used both to create models for locating hydrocarbons (in this case paraffin), and also to investigate the possible effect of time on the model (additional hypothesis 2).

Table 3.1: Hyperspectral images recorded in the first phase. Only the first and last mean image is shown.


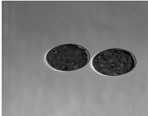
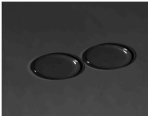
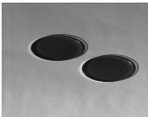
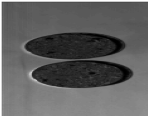
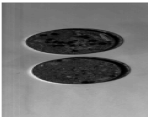
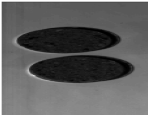
Number	Days after first scan	Mean image
1	0	
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	7	
9	14	
10	21	
11	28	
12	35	
13	42	
14	49	
15	56	
16	84	
17	112	
18	280	
19	308	
20	322	
21	336	
22	350	
23	365	

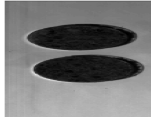
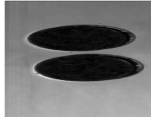
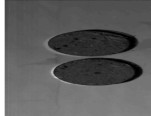
3.3.2 Different concentrations

To investigate whether or not additional hypothesis 3 is correct in assuming that the amount of hydrocarbon present might influence the validity of the

model, scans of soil and paraffin in various proportions were recorded, as shown in table 3.2. Soil and paraffin were mixed in a similar way as for the images in table 3.1.

Table 3.2: Hyperspectral images recorded in the second phase.

Soil [g]	Paraffin [ml]	Comment	Mean image
0	0	Empty plast petri dishes and lids	
12.4	0	Soil, uncrushed	
0	20	Pure paraffin, scans taken after time = 0 h, 4 h, 1 day, 2 days, 3 days, 1 week, 2 weeks	
0	0	Pure water (2 × 10ml)	
15.04	1.00	After time = 0 and 4 days	
15.08	2.50	After time = 0 and 4 days	
15.11	5.00	After time = 0 and 4 days	

Soil [g]	Paraffin [ml]	Comment	Mean image
15.12	10.00	After time = 0 and 4 days	
15.11	15.00	After time = 0 and 4 days	
14.98	0	Soil, crushed	

3.3.3 Different surfaces

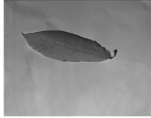


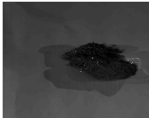
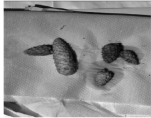
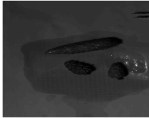
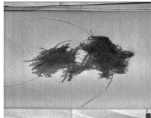
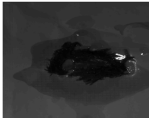

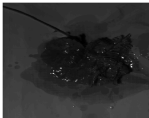

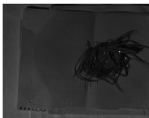

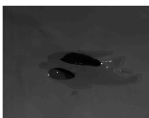
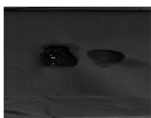

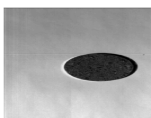
One of the main goals were to develop a model that could be used on several different surfaces. To achieve this, data from more backgrounds than soil were needed. Table 3.3 shows the hyperspectral images recorded in this phase. They span a range of surfaces, including both organic and inorganic matter. The paraffin were applied by immersing the samples in paraffin. This gave an uneven amount of paraffin on the samples.

To help correct for the effect of water (additional hypothesis 1 in section 3.1), water was added to some of the samples.

3.3.4 Time dependence, different backgrounds

Although the images in tables 3.1 and 3.2 above is recorded over time and thus can be used to detect changes over time (see additional hypothesis 2 in section 3.1), an additional time series of images were wanted for two reasons. Firstly, the recorded scans in tables 3.1 and 3.2 are all with a background of soil. Other backgrounds, like stones and humus, could give different results with respect to the time effect, as it is probable that different reactions would occur in a different environment.


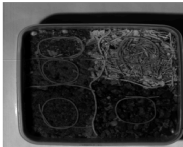
Table 3.3: Hyperspectral images recorded in the third phase.

Description	Mean image	Description	Mean image
Yellow-green leaf, a bit wet		Yellow-green leaf, with paraffin	
Two different kinds of lichen		Lichen with paraffin	
Cones from Scots Pine and Norway Spruce, some wet, some dry		Cones with paraffin	
Moss		Moss with paraffin	
Assorted leaves and plants, a bit wet		Assorted leaves and plants, with paraffin	
Grass, a bit wet		Grass with paraffin	
Stones, dry		Stones with paraffin	
Stones, wet		Wet sand and stones	
Sand, dry			

Secondly, and more importantly, the previously recorded samples were prepared by mixing paraffin and soil. It was reasoned that unwanted discharges of hydrocarbon into nature would result in a significantly different hydrocarbon concentration profile, where the hydrocarbon would migrate through the background over time. This would likely result in a high concentration at the surface at first, which would then decrease over time. The composition of the background would probably affect how fast and to what extent this would happen.

The images in table 3.4 contain five areas: humus, sand, stones and two different types of soil (from *Plantasjen* and compost soil). The height of the box caused some shadows. To control this effect, two pictures were taken at the same time, with the container rotated by 180° between the scans. Only one of these images is shown in the table.

Table 3.4: Hyperspectral images recorded in the fourth phase. Only the first and last mean image is shown.

Number	Time after added paraffin	Mean image
1	Taken before added paraffin	
2	As soon as possible	
3	5 min	
4	30 min	
5	4 hours	
6	23 hours	
7	2 days	
8	3 days	

3.3.5 Different hydrocarbons

Additional hypothesis 5 claims that different hydrocarbons will have different spectra, and that it should be possible to create a model that distinguishes between them. To investigate this, it was necessary to record images of different hydrocarbons.

n-hexane and *n*-heptane were chosen as two different alkanes with similar, but not identical, spectra. In addition, images of soil with paraffin were added.

Nine Petri dishes were used, and the soil in each Petri dish were weighted. To each dish, 5.00 ml of *n*-hexane, *n*-heptane or paraffin were added by slowly pouring the liquid over the soil. See table 3.5.

Table 3.5: Samples prepared in the 5th phase.

Number	Mass soil [g]	Hydrocarbon added
1	5.72	Paraffin
2	6.64	Hexane
3	6.56	Heptane
4	6.41	Heptane
5	6.58	Paraffin
6	6.44	Hexane
7	7.31	Hexane
8	6.19	Heptane
9	8.02	Paraffin

The hyperspectral image were recorded ten minutes after the last sample were prepared. The mean image of this hyperspectral image is shown in figure 3.2.

3.4 Data sets

Only a subset of the pixels from each hyperspectral image were included in the data sets used for developing models. Rectangular areas were chosen from the mean images, and the pixels from these areas were assembled in the data set.

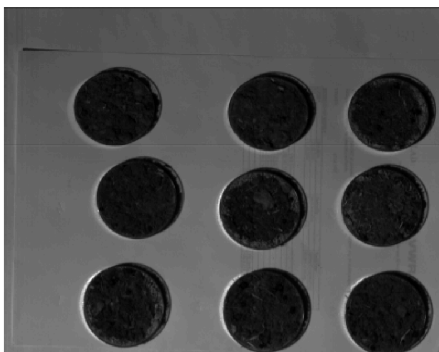


Figure 3.2: Mean image of samples with hexane, heptane and paraffin.

Table 3.6 gives an overview over the different data sets used and their composition.

Table 3.6: Data sets.

Id	Size	Images used
1a	109369	Images 1-8 from table 3.1
1b	376587	All images from table 3.1
2	126606	Images 2, 3, 5-10 from table 3.2
3a	268388	Images 1-8 from table 3.1, images 2-4 and 7-10 from table 3.2, and all images from table 3.3
3b	27536	A selection of the data from images 1 and 2 from table 3.1, images 8 and 9 from table 3.2, all images except dry sand from table 3.3, and images 1-4 from table 3.4
4	3263	The image shown in figure 3.2

Data set 3b were an attempt to create a smaller data set with a more equal number of positives and negatives, but with an even greater variation than data set 3a. This was deemed necessary because the size of data set 3a gave unwanted ramifications (see also section 5.1).

For data sets 1a-3b, the data were split into calibration and validation data sets randomly. For data set 4, the calibration and validation data sets were chosen from different Petri dishes (see figure 3.2).

3.5 Calculations

The calculations were performed using MATLAB R2011b [50] on an ASUS laptop with 6 GB ram and an Intel Core i5 processor. In addition to some functions from the Image Processing and Statistics toolboxes in MATLAB, in-house developed software were used for many procedures, including reading in data, data preparation, some preprocessing routines and PLSR/D-PLSR.

Figure 3.3 illustrates the process from the sample preparation to a finished model. After the samples were prepared and the hyperspectral images recorded by the PryJector, certain interesting regions in the images were selected for further analysis.

Several different preprocessing algorithms, as described in section 2.4, were performed on the data, which were split into a calibration data set and a validation data set.

After choosing which method to use (PLSR/DPLSR, k -NN or other), the model is created using the calibration data, and then validated using the validation data. For k -NN, the calibration data were used as a training data set, while the validation data were used as a test set.

The result is a model, which (hopefully) can be used by the PryJector for prediction on future data.

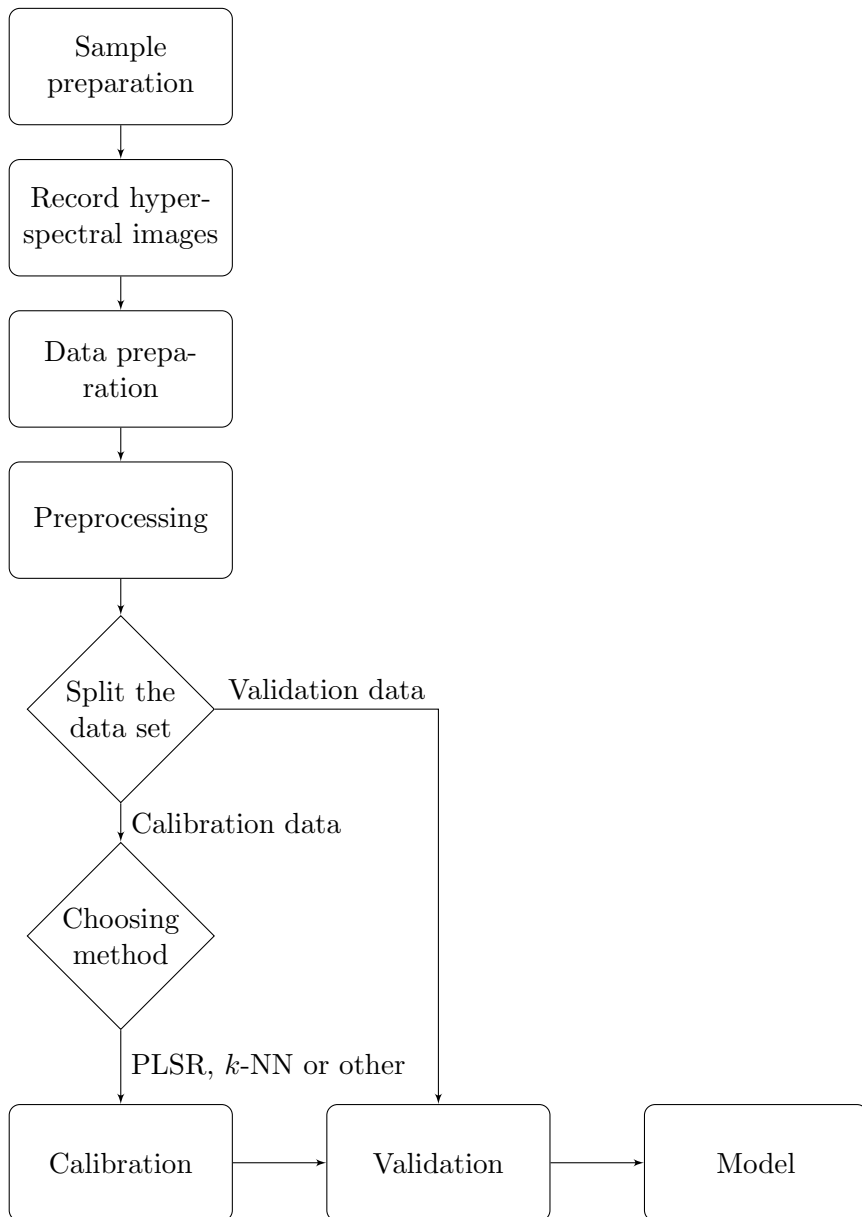


Figure 3.3: Creating a model

Chapter 4

Results

4.1 Detecting hydrocarbons

4.1.1 DPLSR models

Several DPLSR models were developed. Table 4.1 shows some of the DPLSR models and their percentage of error on data set 3a (see table 3.6). The DPLSR models were all generated in the same way, but the data were preprocessed in a number of different ways.

The preprocessing method which showed the most promise, proved to be normalisation (model number 3). This is assumed to be because normalisation removes some of the variance in intensity due to an uneven surface. Multiplicative scatter correction (section 2.4.2), which is a method developed to adjust for this scatter effect, also lowers the error significantly (model 13).

As shown in figure 4.1, the raw data have a saw tooth structure which is probably due to instrument error from the mechanical scanning operation (see also section 2.1.3 about the pushbroom scanner). This saw tooth structure is thought to be the reason median smoothing of the normalised data improves the models somewhat (compare models 3 and 5).

The plot of raw data suggests that there could be some heteroscedastic noise (section 2.1.2.2) that dominates the spectrum around its peaks. This is confirmed by figure 4.2, showing an approximately proportional relationship

Table 4.1: Overview over DPLSR models

Model	No. of comp.	Error (%)	Comment
1	12	9.3	Original data
2	3	23.6	Only selected wavelengths (1171-1246 nm)
3	7	4.3	Normalised data
4	7	5.0	1st derivatives of normalised data
5	7	4.1	Median smoothed normalised data (window = 7)
6	12	10.6	Savitzky-Golay smoothed data
7	12	10.7	Moving average-smoothed data (window = 3)
8	5	4.3	1st derivatives of moving average-smoothed, normalised data
9	15	9.4	1st derivatives of data
10	20	13.9	2nd derivatives of data
11	15	8.5	Centered data
12	17	9.0	Autoscaled data
13	6	4.6	Multiplicative scatter correction of data
14	7	9.2	Logarithm of data
15	6	8.5	Normalisation of log(data)

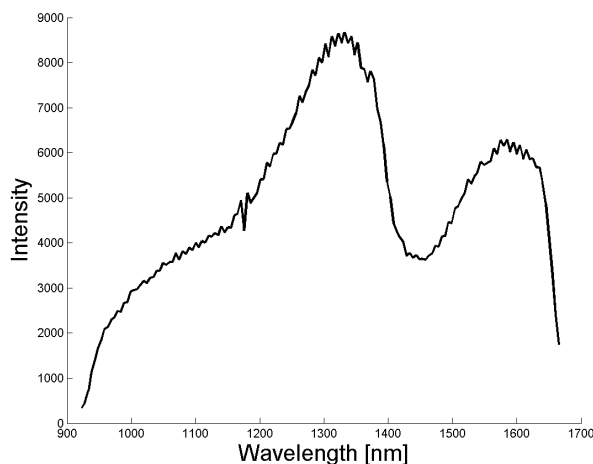


Figure 4.1: A typical raw data spectrum of soil.

between the mean values and the standard deviation of the recorded data. This plot is based on data from dry soil without paraffin from image 1 in table 3.1. If the noise were homoscedastic, no relationship between the mean and the standard deviation would be expected.

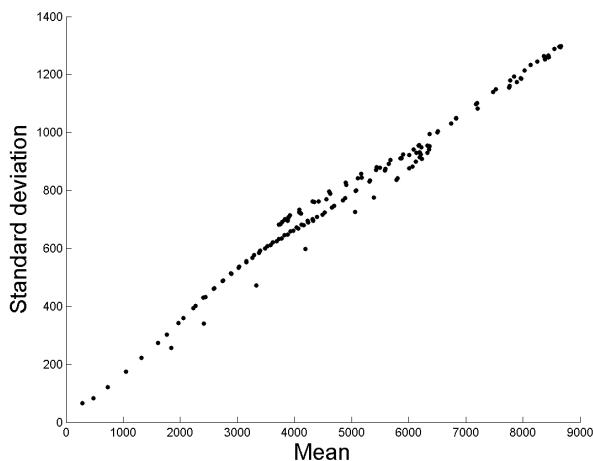


Figure 4.2: Plot of the mean versus the standard deviation for pure soil. Data taken from image 1, table 3.1.

Despite the logarithm transform (section 2.4.5) being commonly employed to remove heteroscedastic noise, the improvement in the error for the DPLSR model of the data after the logarithm transform (model 14) is negligible. Kvalheim et al. [35] recommends using a logarithmic transform before normalisation of data with heteroscedastic noise. This gave an only slightly better (model 15).

As described in section 2.6, the error will depend on the number of PLS components included in the model. The error profile for DPLSR of the original data (model 1), is shown in figure 4.3. As shown in table 4.1, 12 components are used in the model, giving an error of 9.3 %. Even though a higher number of components would give a smaller error, the addition of a 13th component is considered to not make enough of a difference in the error to justify adding the complexity of another component. This is in accordance with Ockham's razor (section 2.3.3).

Ockham's razor could also be employed when reasoning which kinds of preprocessing to use. For instance, with two digits after the decimal point the error of model 8 is 4.28 % while the error of model 3 is 4.33 %. Despite

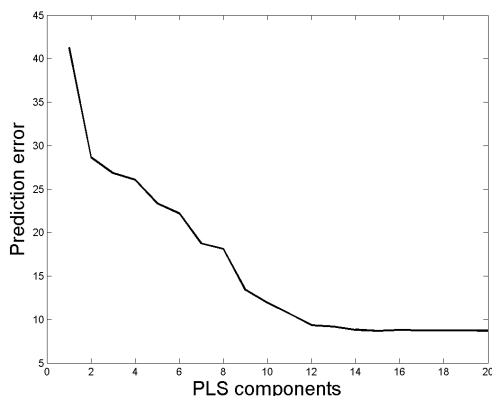


Figure 4.3: Percentage of error plotted against the number of PLS components.

this, model 3 is the model used for further analysis and comparison to other models. This is because this minor lowering of error is considered too small to justify both MA-smoothing and differentiating the spectra.

4.1.2 k -NN models

A procedure for k -NN (see section 2.7) were developed and used on data from the 8 first scans from table 3.1 (data set 1a from table 3.6). Two groups were used: with and without hydrocarbons (in this case paraffin), and each pixel in the validation data set were assigned to the group of which the majority of the k nearest neighbours in the calibration data set belonged.

Table 4.2 shows how the error changes for different (odd) values of k . No preprocessing were performed on the data. The percentage of false positives given in the table, is calculated with respect to the error, not the total number of spectra.

The lowest error is achieved by $k = 1$. This could imply a nice and clear separation between the classes. Figure 4.4 is a score plot from principal component analysis (section 2.5) of these data, and substantiate further that the classes are separated. Although there are some overlap between the groups, the data seems to form three clusters: dry soil (shown in black), soil with water (blue) and soil with paraffin (red). The latter is, to some extent, sandwiched between the other groups.

Table 4.2: k -NN models

k	Error (%)	False positives (%)
1	0.74	0.57
3	0.91	0.72
5	1.0	0.82
7	1.1	0.89

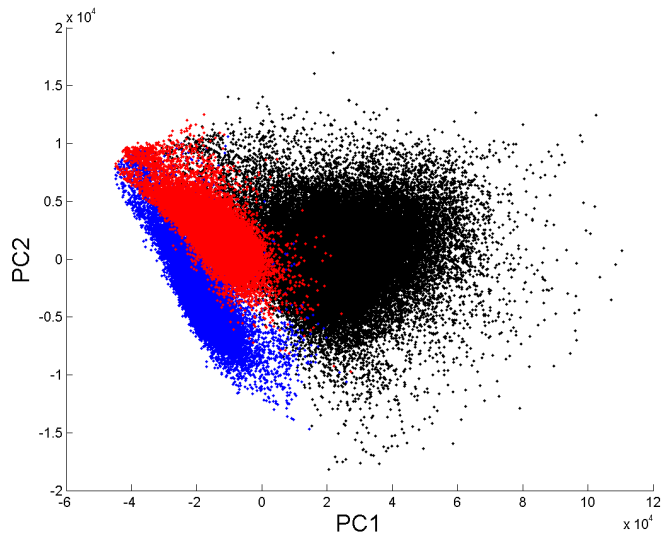


Figure 4.4: Score plot of the two first principal components from PCA of data set 1a from table 3.6. Dry soil is black, wet soil is blue and soil with paraffin is shown in red.

The low percentage of false positives means that a great majority of the errors were from the plastic dishes containing soil with paraffin. It is likely that the paraffin was unevenly mixed with the soil, which would cause some of the pixels to have more in common with soil without paraffin. Another reason for the low percentage of false positives, could be the fact that most of the data points are negatives, i.e. without hydrocarbons present. This could bias the model towards false negatives [5].

The corresponding DPLSR model based on raw data from the 8 first scans from table 3.1, had an error of 0.5 % with 4 PLS components. Attempts were made on performing k -NN on the data set used in the other models for detecting hydrocarbons (data set 3a from table 3.6), but the attempt was aborted when the script was still running after more than 46 hours. Attempts to use k -NN on data set 3b is described in section 5.1 in the discussion chapter.

4.1.3 Waveband models

A simpler model, using only some of the wavelengths, was also developed. This was inspired by the fact that the 1150 nm stretching overtone of the C-H bond was inside the 930-1670 nm range of the recorded wavelengths (see appendix A), and this could provide an area of wavelengths with characteristics specific for compounds with such bonds.

By inspecting plots of raw data, such as the one shown in figure 4.5, it was discovered that somewhere around 1200 nm the spectra of paraffin had a minimum, and that other samples with hydrocarbons present had a lower slope in this area than the corresponding sample *without* hydrocarbons present. This area was thought to correspond to the 1150 nm second overtone of the C-H stretching band (section 2.2), although it is, in fact, at a somewhat higher wavelength.

A couple of possible methods of creating a model were investigated. A smoothed spectrum was compared to the straight line through the 50th and 65th wavelength (1171-1246 nm), but because of the low slope of most spectra in this area, this strategy gave a high error with a large amount of false positives. Another and even simpler model normalised the spectra from the 50th to the 65th wavelength, and compared the intensity of the 56th or 57th wavelength of this normalised spectrum with a value chosen by trial and error. The best of these models used the 57th wavelength (1206 nm) and had an error of 6.4 %. Other attempts were made by using

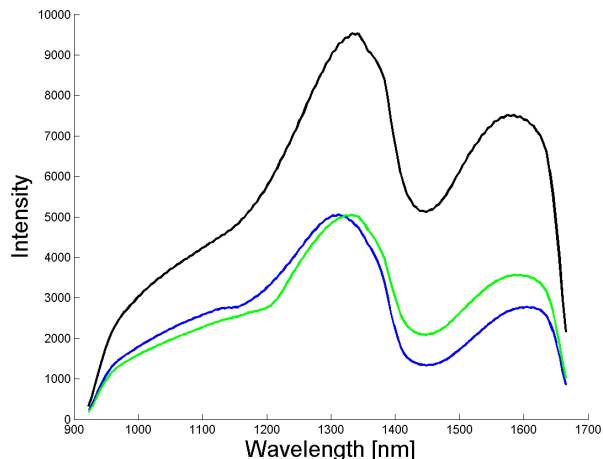


Figure 4.5: Means of raw data of soil (black), soil with water (blue) and soil with paraffin (green). The data is smoothed with a moving average filter, window size = 5.

the 2nd derivative or an algorithm similar to the HI-index described by Kuhn et al. [34].

The best results were obtained by a model that, using smoothed data, calculates the average of the 50th and 65th wavelength (see appendix A) and compares this average value with the intensity for each of the values in between. If the number of wavelengths with an intensity higher than the average value calculated is larger than 10, the pixel is marked as containing hydrocarbons. The MATLAB script is shown below.

```

first = 50;                                % Wavelength = 1171 nm
last = 65;                                  % Wavelength = 1246 nm
compare = 10;
n = size(X,1);
y_est = ones(n,1);                          % Class 1: no hydrocarbons
for i = 1:n
    m1 = X(i, first);
    m2 = X(i, last);
    m = (m1 + m2)/2;
    counter = 1;
    for j = first + 1 : last - 1
        if(X(i, j) < m)
            counter = counter + 1;
        end
    end
end

```

```

end
if(counter > compare)
    y_est(i,1) = 2;           % Class 2: hydrocarbons
end
end
end

```

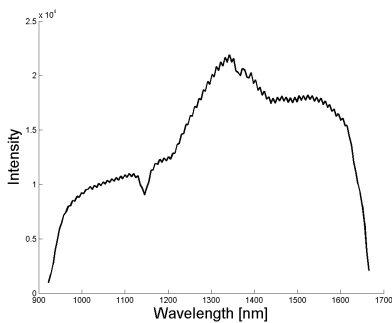
Different smoothing techniques (section 2.4.1) were investigated in an attempt to minimise the prediction error. An overview of some of these, and their error percentages, can be seen in table 4.3. All smoothing techniques improved the model compared to no smoothing (model 7). The procedure gave remarkably good results regardless of smoothing, with the percentage of error almost as low as for the best DPLSR models (see table 4.1). Normalising the data improved the model slightly. Scans of plastic and other hydrocarbons showed that these models did not discriminate between different kinds of hydrocarbons.

Table 4.3: Overview over models

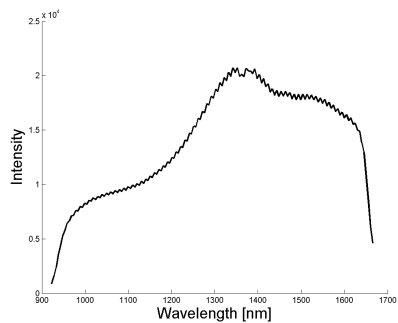
Model number	Wavelengths used	Smoothing technique	Error (%)	False positives (%)
1	50-65	Savitzky-Golay	7.4	57.8
2	50-65, 125	Savitzky-Golay	4.9	36.2
3	50-65, 125	Moving average	4.5	20.4
4	50-65, 125	Median smoothing	5.0	55.1
5	49-64	Averaging	6.1	20.4
6	50-65, 125	Fourier, low pass filter	5.1	47.4
7	50-65	No smoothing	10.2	52.3

Raw data plots of image 4 from table 3.2, figure 4.6c, shows that water has the same characteristic low slope as the hydrocarbons around 1200 nm. For this reason, water gave false positives in the first waveband model (model 1). To remove these false positives, model 2 (and subsequent models) were given an additional condition, and would not mark the pixel as containing hydrocarbons if the intensity of the 50th wavelength (1171 nm) were lower than three times the intensity of the 125th wavelength (1549 nm). This additional condition quantitatively removed the error resulting from water registering as false positives.

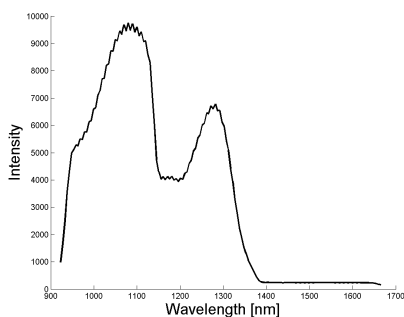
It should, however, be noted that the water spectra used above were from water in a plastic Petri dish. Raw data plots of water in a glass beaker shows another profile, as shown in figure 4.6d. As for the spectra of water



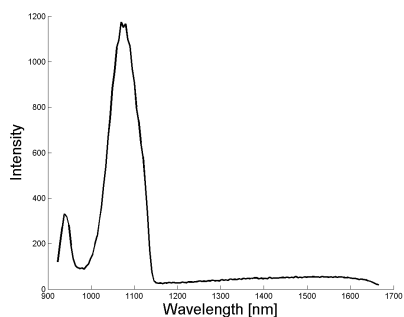
(a) Mean spectrum of a dry plastic Petri dish.



(b) Mean spectrum of a dry pyrex glass beaker.



(c) Mean spectrum of water in a plastic Petri dish.



(d) Mean spectrum of water in a pyrex glass beaker.

Figure 4.6: Raw data plots of water in different containers, and the different containers without water.

shown above, the spectra approaches zero for high wavelengths, but instead of this happening slightly before 1400 nm, it happens around 1150 nm. This made the previously described test for removal of false positives from water useless for these new data.

It also raises the question of whether the reason water in a plastic Petri dish gave false positives might stem from the plastic and not, in fact, from the water. If this is the case, it might not be desirable to treat these as false positives, as they actually should be treated as hydrocarbons. The mean spectrum of the plastic container (figure 4.6a) does have a minimum around the wavelength area used by the waveband model, which is not present in the mean spectrum of the class beaker (figure 4.6b).

The parameters used for the models described in table 4.3 were chosen by trial and error. A more throughout way of choosing parameters would be to find the calibration error for several models, and validate the models using a validation data set. Table 4.4 shows the 10 best models chosen by such a procedure. The percentage error is the validation error on data set 3a. All models use moving average-smoothed data, but with the window size varying from 3 to 9. The other variables were which wavelength area (see appendix A for wavelengths in nm) and what value to use for the compare parameter. Note that the additional test for removing false positives from water is not included here.

Table 4.4: Varying the parameters

Model	Error (%)	False positives (%)	Window size	Wavelength area	Compare
1	6.49	62.6	7	50-64	9
2	6.87	56.2	5	51-63	8
3	6.88	72.8	3	50-64	9
4	6.88	53.0	7	49-64	10
5	6.90	60.7	9	52-65	8
6	6.96	56.3	5	51-65	9
7	6.97	67.1	3	49-64	10
8	6.99	51.2	9	51-62	7
9	7.01	48.8	7	50-65	10
10	7.06	51.9	7	50-62	8

4.2 Estimating the amount of hydrocarbons

In accordance with additional hypothesis 3, attempts were made to estimate the amount of hydrocarbons.

As this is a regression problem, only PLSR have been used. Although k -NN, which is a classification method, can also be used for regression, this works best for low-dimensional problems [26], and has not been attempted here.

Figure 4.7 shows the result of an PLS regression on data set 2 from table 3.6. While there seems to be some correlation between the estimated amount of hydrocarbons and the experimental amount of hydrocarbons, it is clear that the predictability is not very good.

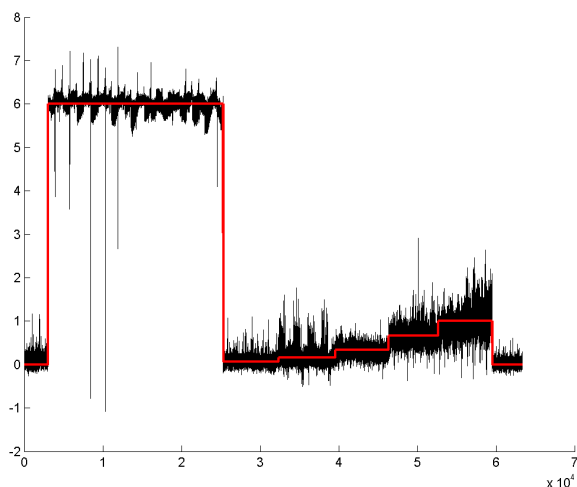


Figure 4.7: Estimated (black) and experimental (red) values. No paraffin = 0, pure paraffin = 6. Other values given as ml paraffin / g soil.

There are several possible ways for increasing the predictability of the model. So far, no preprocessing has been performed on the data. Normalisation, MSC and smoothing is likely to improve the model. If a model for estimating the amount of hydrocarbons were to be used with the back-projecting feature of the PryJector, it might be more interesting to use an average of several neighbouring pixels than to assess every pixel separately. This might very well give a much better estimate. An extreme case of how averaging can increase the predictability, is shown in table 4.5. The average

spectrum of all the data with the same experimental amount of hydrocarbons, is here used for prediction, and the error is very low. The model used is the same PLSR model as in figure 4.7.

Table 4.5: Estimated values, average spectra

Experimental value	Estimated value	Error	Error (%)
0	0.12	0.12	-
0.07	0.09	0.03	39.7
0.17	0.19	0.03	15.2
0.33	0.32	-0.01	-4.0
0.66	0.65	-0.01	-1.6
1.01	0.94	-0.06	-6.3
6	5.99	-0.01	-0.2

One of the challenges with creating this model, was how to choose the y data. For mixed soil and paraffin, the amount of paraffin in ml were divided with the mass of soil in g. Pure soil were thus given as 0. Pure paraffin were given as 6. This value were chosen because a similar PLSR model made without pure paraffin in the calibration data set, returned values around 6 for pure paraffin. The recorded spectrum for paraffin will have been affected by the background, which was a plastic dish.

It is also worth noting that the true amount of hydrocarbons is, due to the way samples were prepared, likely to vary between individual pixels from the same Petri dish. For this reason, the estimated amount of hydrocarbons might in some cases in fact be closer to the true amount for that pixel, than the experimental value is, but this is purely speculation and impossible to neither confirm nor falsify by the data. A model using the average spectra over four or nine neighbouring spectra, might for this reason give more stable results.

4.3 Identifying hydrocarbons

It was also investigated whether or not it was possible to identify alkanes of different chain-lengths. Scans of n -hexane and n -heptane, as described in section 3.3.5, were compared. The data set used is number 4 from table 3.6.

The models described below are not trained to detect hydrocarbons, but can be used in collaboration with one of the models presented in section 4.1.

For the same hyperspectral image, one model can detect areas with hydrocarbons present while another can identify the detected hydrocarbons.

4.3.1 DPLSR models

Table 4.6 shows the error for three DPLSR models for distinguishing hexane and soil from heptane and soil. In contrast to the results in section 4.1.1, the lowest error was achieved by DPLSR on raw data. Neither normalisation nor MSC improved the model.

Table 4.6: DPLSR models for distinguishing hexane from heptane

Model	No. of comp.	Error (%)	Comment
1	6	12.1	DPLSR of raw data
2	5	18.9	DPLSR of normalised data
3	4	16.1	DPLSR of data after MSC

The size of the error seems to suggest that it is possible to distinguish between different hydrocarbons by using DPLSR. A more optimal model with a lower error can probably be developed. *n*-hexane and *n*-heptane is, of course, very similar. It is reasonable to assume that hydrocarbons with more structural differences between them will be easier to distinguish.

4.3.2 *k*-NN models

Distinguishing between different hydrocarbons is a classification problem, which makes a classification method like *k*-NN an obvious choice. Despite this, the error for *k*-NN is large. This is shown in table 4.7, which gives the percentage error for different values of *k*. The results also shows that there are more heptane marked as hexane (about 20 % of the error), than hexane marked as heptane (about 80 % of the error).

The size of the error might be due to the fact that the calibration and validation data sets were chosen from different Petri dishes, and that no preprocessing was performed on the data prior to *k*-NN.

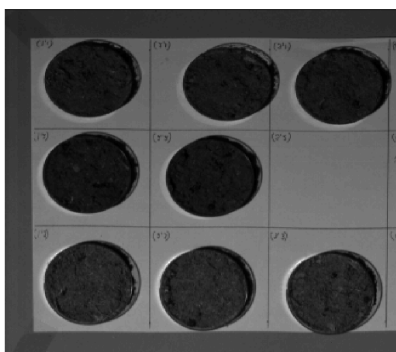
Table 4.7: k -NN models for distinguishing hexane from heptane

k	Error (%)	Hexane marked as heptane (%)
1	40.9	19.1
3	39.7	23.8
5	38.4	17.6
7	36.6	18.9
9	37.2	17.7
11	37.7	18.4

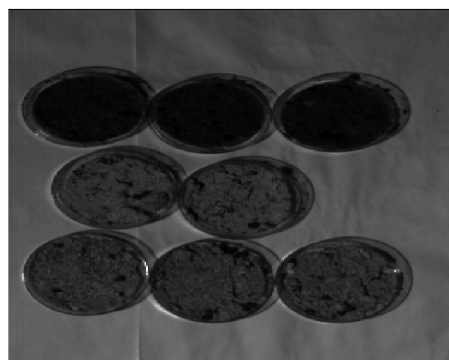
4.4 Time dependence of the models

One of the hypotheses in part 3.1, was that the accuracy of the model would depend on how much time had passed. To test this, scans of the same samples were performed during the course of one year (see table 3.1). This is data set 1b from table 3.6.

Mean images of the first and last scan in this time series is shown in figure 4.8. There are visible cracks in the Petri dishes with soil (bottom) and soil with water (middle). There are less cracks in the Petri dishes with soil and paraffin (top row).



(a) The first day



(b) After one year

Figure 4.8: Mean image of data, after 0 and 365 days.

Figure 4.9 shows how the error of detection changes over time. The DPLSR model is made from data set 1a in table 3.6, and uses only data recorded during the first week. Although the error seems to be somewhat higher after

a year, the error stays low during the whole year. Some of the increase in the error is probably due to the fact that, as can be seen from the mean images (figure 4.8), cracks in the soil have been introduced over time, causing the plastic to be visible, which might produce false positives.

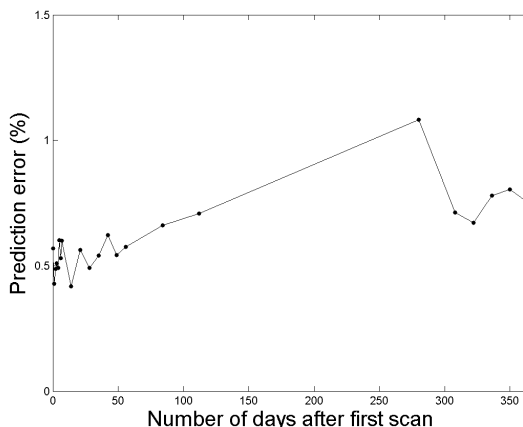


Figure 4.9: Plot showing how the percentage error of detection changes over time. The time is given in days.

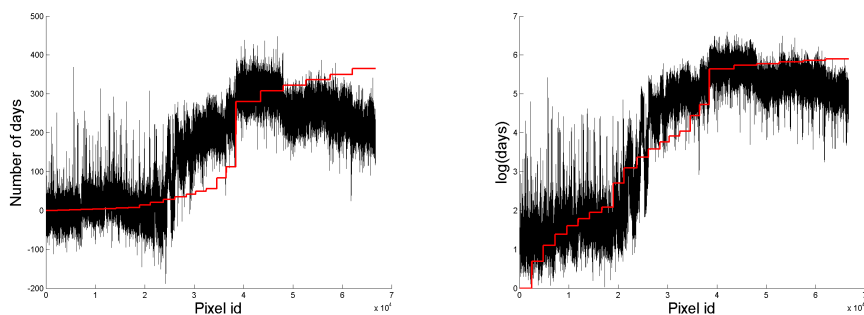
The results suggest that time is a less important variable than hypothesis 2 assumes. It is, however, possible that time will have a greater effect on other surfaces than soil. The samples used here are prepared by mixing paraffin and soil. Other ways of preparing the samples might have a larger error. The data from table 3.4, where paraffin is poured over the surfaces, seems to be more time dependent.

4.4.1 Can the time be estimated?

Figure 4.10 illustrates the results of a PLS regression of the data from data set 1b in table 3.6, using time after first scan as the y data. Only the pixels with with paraffin and soil are used.

There is some relationship between estimated and measured time, but the error is large. As with the concentration regression (section 4.2), it is likely that a higher accuracy might be achieved by spatially smoothing of the data. A smaller error is also achieved when the logarithm of the time is used.

It should be noted that these samples were prepared by mixing paraffin



(a) PLSR with y as the number of days after applying paraffin

(b) PLSR with y as the logarithm of the number of days after applying paraffin

Figure 4.10: Estimated (black) and experimental (red) values. PLSR of normalised data.

and soil. A higher time dependence is expected for data from samples where paraffin is poured over the soil, such as the images in table 3.4. For samples such as these, it is expected that the paraffin over time will migrate downwards through the sample, and to some extent disappear from the surface.

One problem with time series such as this, is that interesting long-term trends often are buried within short-term random fluctuations [5]. It is apparent from figure 4.10 that while there seems to be a certain trend in the data, using this model for prediction would not give good results for individual pixels. A routine using the mean or median value for a larger area might give better results.

Chapter 5

Discussion

5.1 Choosing calibration and validation data sets

How to choose the optimum size of the calibration and validation data sets, is an important problem with no easy answer [5]. It might be tempting to include as much data as possible in the calibration data set, but a larger data set will not always give better results. The data set must be large enough to span the relevant spectral variation for prediction on future data, but if the data set is too large, the relevant information might "disappear" in noise and irrelevant information.

It could be desirable to use the smallest data set possible, to save processing time (and money). For the PryJector, it is important that the processing time is short enough to make the PryJector able to operate "on the fly". However, the PryJector will generally use a pre-calibrated model, and the size of the data set used for calibration and validation will not necessarily influence the processing time of the final model, except for k -NN, where the number of prototypes in the training data set is essential.

The data set used for most of the calculations and testing/comparison of methods described in chapter 4, is data set 3a from table 3.6, which consists of almost 270 000 pixels, each with 148 wavelength bands. This turned out to be impractically large. Some procedures, like MSC (section 2.4.2) took hours to finish (although a more efficient implementation of MSC might be faster), while k -NN (sections 2.7 and 4.1.2) had to be aborted after nearly two days. To use this technique, a smaller data set would be necessary.

This is elaborated further in section 5.4.

It should be noted that the calculations were performed on an ordinary laptop (see section 3.5 for specifications). On a more powerful computer, the size of the data set would have been a lesser problem.

One of the other reasons that a larger data set might, in fact, be a bad thing, is the importance of the data set being "balanced". To illustrate this point, an extreme case would be a data set with 1000 data points, of which only 10 data point is of class 2, and the resulting 990 data points are all class 1. A model which automatically labels all data as class 1, would, of course, give a low error (1 %), but be completely useless for prediction. A data set with a smaller number of data points belonging to class 1, would in many cases result in a better model.

Data set 3a from table 3.6, which is the data set used for comparison between the models e.g. in table 4.1, has approximately 66 % of the pixels belonging to the background class (negatives), and only one third of the pixels containing hydrocarbons (positives). This must be taken into account when the different models are compared to each other. This was one of the motivations for calculating the percentage of false positives for the different models.

Even though the need for reducing the size of the data set were the catalyst for creating data set 3b, a more balanced data set were one of the determining factors for choosing how many pixels from each image to include in the data set. As a result of this 49.4 % of the pixels in data set 3b are negatives (no hydrocarbons present), and 50.6 % are positives (hydrocarbons present).

One way of reducing the size of the data set without removing any of the samples, is to use data compression (section 2.1.2.3). Data compression can give a more compact data set with a higher signal-to-noise ratio, and algorithms that can operate directly on the compressed data saves time [48].

There are several possible methods that could have been used, amongst them averaging (section 2.4.1.1), latent variables (section 2.5) and wavelets (section 2.4.1.5).

5.2 Comparing models

There are several things to keep in mind when comparing different models. One obvious way of comparing them, is to use the percentage of error, and then use the model with the lowest error. This will, however, not always give the best model.

To decide which model is the best model, one needs to answer the question: *what are the characteristics of a good model?* As explained in section 2.3.3, a good model neither overfits nor underfits to the data. In other words, the ideal model is sophisticated enough to describe all the relevant variation in the data set while neglecting the variation present due to noise. Most of the time, it is possible to create a more complicated model with a lower error, but such a model might describe the noise in the data instead of the relevant information. Ockham's razor should be kept in mind at all times, both while creating and comparing models.

For the models developed in this thesis, three aspects of the models are assumed to be especially important: the predictive ability, the robustness, and the simplicity of the model. The predictive ability is easily quantified as the percentage error for a given data set, and comparing models based on their error is the main principle for comparing models used in the results chapter. However, the ability to predict correctly also on new samples and previously unknown surfaces, is very important. A robust model can be applied to many different samples, and has a low and stable error on a variety of surfaces. Because of the spatial continuity of the data, the interpretation of the chemical image is not dependant upon every single pixel being classified correctly.

The models for detecting hydrocarbons can perform two types of errors: it can detect hydrocarbons where there in fact are none (type I error, also known as false positives), or it can fail to detect hydrocarbons (type II error, or false negatives) [57]. In most cases, type I errors are assumed to be worse than type II errors. This is not necessarily the case here. The severity of the different types of error will depend upon the purpose of the study. If the aim is to efficiently detect areas with hydrocarbon discharges into nature that will later be put through a more throughout investigation using different methods, some false positives may be a lesser problem.

When comparing the DPLSR models from table 4.1, it is important to compare both the error and the number of components. The ideal model would have a low error with a minimal number of components. Figure 5.1

shows the error plotted against the number of PLS components for both the original data (model 1) and the normalised data (model 3). It is apparent that the model using the normalised data is better, as it has a lower error for all number of components.

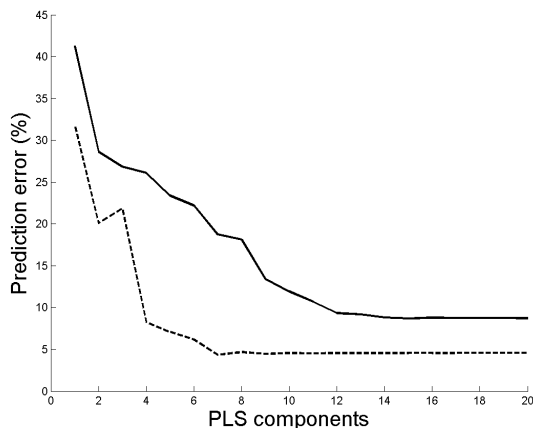


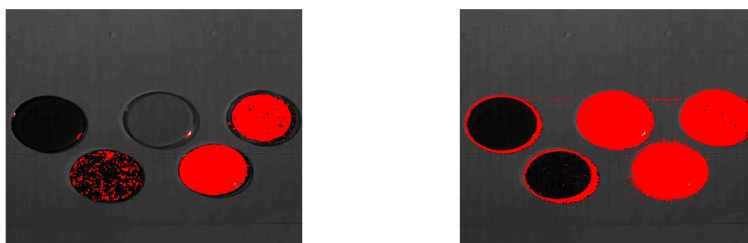
Figure 5.1: Error (%) as a function of the number of PLS components for original (solid) and normalised (dashed) data.

Another question that should be considered, is: *Why* is this model better than the others? The answer is, unfortunately, rarely self-evident. In this instance, however, an explanation may be given. Section 2.4.4 describes how normalisation will improve the prediction for hyperspectral data by removing the variation due to an uneven surface.

When comparing DPLSR models from table 4.1 with other models as the waveband models from table 4.3, it might be tempting to use only the total error. The information gained in such a way, is, however, rather restricted. Errors for different surfaces and images may be calculated, as well as the percentage of false positives versus false negatives, and all this information can give a better background for choosing which model to use. An example of this is given in section 5.3, where the percentage error of two models is given for each of the images from table 3.3.

There is also other techniques available. One of the main features of the PryJector is the ability to visually expose where the detected hydrocarbons are. This ability can easily be imitated by a computer. Figure 5.2 shows where two of the models with the overall lowest error, DPLSR of the normalised data set with 7 PLS components (model 3 from table 4.1)

and waveband model on MA-smoothed data with window size = 7 (model 3 from table 4.3), detects hydrocarbons. These models are applied on an image including water in a plastic dish, wet soil, an empty plastic dish, paraffin in a plastic dish, and paraffin and soil. The detected hydrocarbons are shown in red.



(a) Hydrocarbons detected by DPLSR model 3 (b) Hydrocarbons detected by waveband model 3

Figure 5.2: DPLSR model 3 and waveband model 3 applied on (left to right) water, soil with water, plastic, paraffin and soil with paraffin. Hydrocarbons detected by the models are shown in red.

The difference between the models are easy to grasp. The DPLSR model – which has only been trained to recognise paraffin – gives a negative for plastic, while the simple model – which probably recognises the C-H-bond in hydrocarbons – treats plastics and paraffin the same. This model seems to have a lower overall error on this image, while the DPLSR model has a number of false positives for the wet soil. Both models, however, recognises both paraffin and soil with paraffin quantitatively.

A visual inspection like the one shown above, gives a lot of information in a short time, but is of course not always viable. The percentage of error calculated for the different models, is calculated using data from more than thirty different scans. To compare all these images visually for all the models developed, would have been time consuming. For that reason, only the percentage of error is calculated for most of the models. For the models with the lowest percentage of error, the percentage of error on some specific backgrounds were calculated, and visual inspections were performed.

Table 5.1, compares the percentage error of DPLSR, k -NN and waveband model for a smaller data set (data set 3b from table 3.6). The parameters for the waveband model is adjusted to minimise the prediction error on this

Table 5.1: Comparing different models

Model	Error (%)	False positives (%)
DPLSR of raw data, 11 components	12.5	64.2
k -NN, $k = 5$	9.3	4.6
Waveband model	16.3	24.0

data set (see also appendix B). Data from table 3.4 is the largest part of this data set (19766 of 27536 pixels). This set of surfaces proved challenging to create a good model for (see also section 5.3). This is probably the reason the prediction error is larger than for the models developed from data set 3a.

5.2.1 Hard and soft modelling

Unlike hard modelling, where a mathematical model is developed based on first principles, soft modelling aims to describe the variation in the data by application of general criteria and constraints [10]. In other words, while hard modelling starts with some physical, chemical or mathematical law, soft modelling both begins and ends with the data.

The DPLSR, PLSR and k -NN models presented in chapter 4 are all examples of soft modelling. The waveband model (section 4.1.3), utilizes information about the NIR-spectrum and wavelengths of vibration overtone bands of hydrocarbons (section 2.2), and is thus an example of hard modelling, even though some parameters were chosen to fit the data through calibration.

5.3 Different surfaces

For a model to be useful, several criteria must be met (see section 5.2). One of the criteria is that the model should be applicable to several different surfaces. Ideally, it should be possible to use the same model on new surfaces not already investigated (not included in the calibration data). The surfaces of interest in this thesis are surfaces occurring in nature, such as soil, sand, stones, humus and vegetation.

Environmental systems are complex, and in order to create a model that

covers this complexity, a large number of samples are needed [27]. For this reason, a variety in surfaces were collected, both organic and inorganic. It proved challenging to create a model which gave good results on all surfaces. Some of the DPLSR models gave very good results on some surfaces, and poor results on others. The waveband model gave good or average results on most surfaces, while k -NN models typically gave good results on familiar surfaces, which does not necessary mean that it is useful for prediction on new samples.

Table 5.2 gives the image specific error for the DPLSR model 3 (from table 4.1) and waveband model 3 (table 4.3) on images from table 3.3.

Table 5.2: Percentage error on images from table 3.3

Description	DPLSR model	Waveband model
Leaf, a bit wet	0	0
Leaf, with paraffin	0.05	0
Lichen	7.1	0.4
Lichen with paraffin	33.5	12.5
Cones	1.5	0.6
Cones with paraffin	39.6	11.2
Moss	0	6.2
Moss with paraffin	1.7	1.1
Plants and leaves	0.01	0
Plants with paraffin	2.8	0.1
Grass, a bit wet	0	1.2
Grass with paraffin	18.1	0.3
Stones, dry	0	0.3
Stones with paraffin	96.1	28.7
Stones, wet	0.4	0.7
Wet sand and stones	2.4	5.5
Sand, dry	0	0.2

Both models have difficulties with the images of lichen with paraffin, cones with paraffin and stones with paraffin, but the waveband model is significantly better than the DPLSR model on all these images. The largest difference is for stones with paraffin, where the DPLSR model has an error of 96.1 % and the waveband model has an error of 28.7 %. This is also the image where both the models have the largest error.

In addition to these images where both models struggle, the DPLSR model

performs poorly for grass with paraffin (18.1 %) and to some extent lichen without paraffin (7.1 %). The waveband model performs significantly worse than the DPLSR model for moss without paraffin (6.2 % versus 0 % for the DPLSR model) and wet sand and stones without paraffin (5.5 % versus 2.4 %).

The other errors are small for both models. When the error is given as 0, all pixels in the data set were labelled correctly.

The percentage error varies greatly between the different surfaces. Additional hypothesis 4 expect false positives from vegetation. This seems not to be the case here, as the majority of the error stems from the model failing to detect paraffin, and the amount of false positives are small.

Stones seems to be a difficult surface for detecting paraffin, especially for DPLSR models. This was to some extent confirmed by applying DPLSR models from table 4.1 on the scans of humus, sand, stones and soil from table 3.4.

However, DPLSR models based on data set 3b from table 3.6, performed considerably better on stones. Appendix B shows the percentage error for a DPLSR model, a k -NN model and a waveband model for each of the different images and surfaces included this data set (this is the same models as in table 5.1). The error for detecting paraffin on stones varies for the different images and models (7.0-64.4 %). It is unsurprising that sand also have an error with about the same range (5.2-69.1 %). For sand and stone the error increases with time. This is unsurprising, as the paraffin is expected to flow downwards with time, and thus disappear from the surface. This effect is larger for sand than for stones.

What is surprising, however, is that it was the waveband model that performed the worst for both sand and stones with paraffin. While the DPLSR and k -NN models had errors from 5.2 % to 15.8 %, the waveband model failed to detect paraffin on 56.7 %-69.1 % of the pixels.

The error for detecting paraffin on humus, on the other hand, is lower for the waveband model (13.9-19.9 %) than for DPLSR (21.0-22.8 %) or k -NN (25.0-28.2 %). Table B.1 confirms that the waveband model generally has a low prediction error on vegetation also in this slightly modified form from the model used in table 5.3. Cones without paraffin (11.1 %), moss without paraffin (53.0 %) and grass without paraffin (18.2 %) are the only organic surface from table 3.3 with a prediction error higher than 10 % for this model.

From this, it would seem desirable to use the waveband model on vegetation, and the DPLSR or k -NN model on inorganic surfaces. For the PryJector, this could be chosen by the human operating it. However, it should also be possible to create a model that determines whether a surface is organic or inorganic. This model could then choose which of the models for detecting paraffin that should be used.

It is a challenge to construct realistic surfaces in a laboratory environment. Vegetation was particularly challenging in this regard, and the images in table 3.3 are not necessary good approximations to what similar samples would look like in nature. Ideally, the hyperspectral images should have been recorded outside in a more realistic and natural environment. For practical reasons, this was not possible, and instead the samples had to be collected and brought to the laboratory.

Different strategies for adding hydrocarbons were also used. Paraffin were mixed with the soil (images in tables 3.1 and 3.2), the samples were immersed in paraffin (images in table 3.3), or paraffin, hexane or heptane were poured over the surface (images in tables 3.4 and 3.5). The last of these strategies is considered the most realistic, at least with regard to using the models for detecting hydrocarbon fuel discharges.

Because the near-infrared radiation has a limited ability to penetrate matter [36], the background behind the sample will only be of importance for some of the samples. This is discussed briefly for water in section 4.1.3. Scans of different amounts of water showed that, as expected, the influence of the background on the spectra was diminished when the height of the water increased.

5.4 Improving the k -NN models

There are several problems with the k -NN approach currently used. As mentioned in section 5.1, the data set used for creating and validating most of the models, data set 3a from table 3.6, were too huge to perform k -NN, and k -NN were performed on some of the smaller data sets (1a and 3b) instead.

These data sets have a different composition from data set 3a, which makes direct comparison with models based on data set 3a difficult. It is also possible to create a smaller version of data set 3a. One possibility is to

only use some of the pixels from the original data sets. Attempts at using only pixel 1, 1001, 2001 and so forth were performed, but with poor results.

Another, and perhaps better, solution would have been to randomise which pixels to use. This is one of the principles used for assembling data set 3b. Randomisation would have been preferred because of the ordered nature of the data sets. Even though the calibration and validation data sets are chosen randomly, the data is collected sequentially, meaning that the 1001st data point from the calibration data set is likely to be spatially close to the 1001st data point from the validation data set. This means that it probably will also be spectrally more similar to the 1001st data point than e.g. the 1020th data point from the validation data set.

For k -NN, this means that it is a high probability that the 2nd data point in this new validation data set (the 1001st data point in the original validation data set), is a close neighbour to the 2nd data point in the new calibration data set. Thus, sequentially selecting pixels might cause the error for the validation data set to be lower than it realistically should be.

For large data sets, calculating the distance between all points is expensive [42]. Instead of performing k -NN in 148 dimensions, attempts were made at using only a small number of dimensions, where the dimensions chosen were the first principal components from PCA (section 2.5). Unfortunately, the error increased significantly, while the processing time stayed impractically long due to the number of objects being the same.

Another possibility for reducing the size of the data set is to perform spatially averaging (see section 2.4.1.1) over neighbouring pixels, and thus combining 4, 9, 16 or more data points. This has two advantages: the data set becomes smaller, and the signal-to-noise ratio will increase due to the fact that errors, statistically, will cancel and the variance will decrease as 1 divided by the number of pixels averaged (see also section 2.4.1.1). Removing noise and lowering the resolution in this way could also prove favourable for other methods than k -NN.

Images segmentation [23] is a field that concerns dividing an image into its constituent regions or objects, and is a way of extracting or isolating elements that belong together. It is a non-trivial problem, but image segmentation methods are in use in remote sensing. Segmentation methods could also be applied the problem here, to determine groups of neighbouring pixels with similar properties, which could be further processed together.

Cluster analysis [5] could have been used to create groups of similar pixels

not necessarily spatially connected. This could have provided useful for choosing prototypes for k -NN. Different kinds of cluster analysis exists.

Agglomerative clustering display the level of similarity between all objects by creating a tree of clusters (a dendrogram [5]) where the leaf nodes are individual objects [42], in this case pixels. Clusters of varying levels can be selected from the final result. Agglomerative clustering were attempted on data set 1a from table 3.6, but resulted in an out of memory error in MATLAB.

K-means clustering [42] divides the data into exactly k categories by iteratively assigning each object to the closest cluster. Closeness is calculated as the distance from the mean of the cluster. The result will depend upon the value for k , which can be chosen by cross-validation [1].

5.5 A cheaper set up

One of the key challenges with the PryJector, is to make it cheap enough to be commercially viable. The present set up is quite expensive, which clearly restricts its usefulness. Generally, hyperspectral cameras that operates in the NIR region are expensive.

It might, however, be possible to use a different, and cheaper, set up. The waveband model described in part 4.1.3 predicts the presence of hydrocarbons using only 16 or 17 preselected wavelengths. The idea for this procedure, came from the fact that the wavelengths of the stretching overtones of the C-H band is known, and that these wavelengths should have a certain characteristic profile for hydrocarbons.

As mentioned in part 2.2, the fourth stretching overtone of the C-H band is at 700 nm, and thus inside the visible part of the electromagnetic spectrum. If it is possible to create a model for this area similar to the waveband model described above, a regular camera might be used since the range of visible light is 400-780 nm [51]. The cost of producing the PryJector would in that case be considerably lower. The fourth stretching overtone might, however, be too weak for selective prediction. If that is the case, it would also be possible to use the third stretching overtone at 880 nm. This is outside the area of visible light, but might be inside the area a regular (or semi-regular) CCD in a digital camera might capture.

Water gave false positives in the first simple model, and an additional wavelength outside the region first selected was included in the model to

remove false positives from water. This might complicate things. When using a hyperspectral camera such as the one used in the work on this thesis, this is no problem, but if one were to use a simpler and cheaper setup, it might be a challenge to find a range of wavelengths that is selective for hydrocarbons.

While investigating scans of water on backgrounds other than soil/plastic, it became apparent that the spectrum were shifted, and that the waveband model thus gave false positives (see figure 4.6). For the PryJector, this might not be a problem, as the main goal is to identify hydrocarbons that are not detectable by the human eye. The model must thus be able to separate a mixture of hydrocarbons and soil from a mixture of water and soil, but if a lake gives a false positive, it is a smaller problem, as it would be quite obvious to the human using the instrument that this is, indeed, a false positive.

5.6 Other methods

There are a plethora of different methods that could be applied to the problem presented in this thesis. To present all of them would be impossible. This section aims merely to mention some possibilities that could result in a better and more robust model (see also sections 5.2 and 5.4).

Linear discriminant analysis (LDA) is a supervised classification method [24]. The method uses linear hyperplanes as decision boundaries [1], and maximises the between class/within class variation ratio [22]. Unlike k -NN, it can not handle complex decision boundaries, but it can be faster and more effective.

Human brains are remarkably good at recognising patterns. Neural networks are models for pattern recognition on complex data, that aims to simulate the human pattern recognition system [27]. The method uses complex non-linear functions with many parameters, which can be learned from noisy data [42]. Different kinds of neural networks exist, and they will often give better results than regression methods such as PLSR [47].

Support vector machines (SVMs) [40] is a kernel method, which similarly to neural networks can represent complex, non-linear functions. The main idea behind kernel methods is that data will always be linearly separable if they are mapped into a space of sufficiently high dimension [42]. SVMs use

an efficient training algorithm, and can reach a high accuracy with only a few training pixels [40].

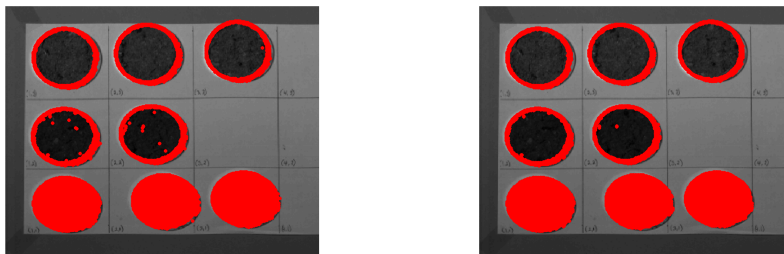
Different models will generally have different strengths and weaknesses (see also section 5.3). For this reason, it can be tempting to not choose only *one* model, but rather use a handful of them. This is the idea behind ensemble based systems [41], where several different models work together, and the final decision is based on the results from all the models. It is crucial that the individual models do not all err on the same objects, but as long as this prerequisite holds, ensemble methods may perform close to perfect even if the individual models each have a large error [42]. Ensemble based systems have been demonstrated to perform better than one single model for many different applications [41].

Random forests (RFs) [31] is one such ensemble method for classification or regression that uses several decision trees and base the decision on a majority voting of the trees. Decision trees is one of the simplest learning algorithms, and reaches its decision by performing a sequence of tests [42]. Which test to perform depend upon the result from the previous test. It has been shown that random forests in combination with morphological feature extraction (MFE) performs better than PCA and MFE for extracting and classifying regions in aerial images taken of urban areas [31].

None of the models developed in this thesis truly utilizes the fact that the data are spatially continuous. There are many ways to exploit this important property of hyperspectral images. One of the possibilities is image segmentation, as mentioned in section 5.4.

Another rather simple possibility is to use the spatial continuity to reduce the error by removing lonely pixels, that is, pixels that are classified as belonging to another class than its neighbours. This is demonstrated in figure 5.3 where waveband model 3 is applied on image 1 from table 3.1 (the images are upside down compared to the mean images in table 3.1). In figure 5.3b, a procedure marking pixels as negatives if none of its closest neighbours, using a four-connected neighbourhood, are classified as containing hydrocarbons is conducted. The model generally performs well on these data, but there are a few false positives in the Petri dishes in the middle row (containing soil and water). In image 5.3b, most of these lonely pixels are removed.

Because the procedure sketched above uses information about the four nearest neighbours of each pixel, the spectra of the neighbouring pixels must be known before prediction. For the PryJector, this would cause the projected



(a) Before removing lonely pixels.

(b) After removing lonely pixels. Most of the false positives are eliminated.

Figure 5.3: Hydrocarbons detected by waveband model 3 before and after removing lonely pixels. Hydrocarbons detected by the models are shown in red. The size of these pixels is exaggerated.

image to be lagging one line behind, as it will need to wait until the next line is recorded.

Chapter 6

Conclusion

Models for rapid localisation of hydrocarbon compounds on several surfaces including soil, sand, stones, humus and vegetation were developed with promising results. Discriminant partial least squares regression (DPLSR) proved a useful method, while the size of the data set proved challenging for the k -nearest neighbours method (k -NN). A fast and simple model using only a preselected area of the wavelengths available gave remarkably good results for most surfaces. Cones, humus and wet sand and stones proved challenging surfaces to model.

Selectivity were achieved through a DPLSR model able to discriminate between hexane and heptane on soil. This substantiates that it is possible to distinguish between different hydrocarbon compounds on the same surface. Introducing different surfaces might complicate this.

Bibliography

- [1] Bjørn K. Alsberg. Introduction to chemometrics (lecture notes), 2011.
- [2] Bjørn K. Alsberg, Trond Lø ke, and Ivar Baarstad. PryJector: a device for in situ visualization of chemical and physical property distributions on surfaces using projection and hyperspectral imaging. *Journal of forensic sciences*, 56(4):976–83, July 2011.
- [3] Bjørn K. Alsberg and Jø rgen Rosvold. Rapid localisation of bone fragments on surfaces using back-projection and hyperspectral imaging. *J. Forensic.Sci. (in press)*, 2013.
- [4] Vincent Baeten, Juan Antonio Fernández Pierna, and Pierre Dardenne. Hyperspectral Imaging Techniques: an Attractive Solution for the Analysis of Biological and Agricultural Materials. In Hans F. Grahn and Paul Geladi, editors, *Techniques and Applications of Hyperspectral Image Analysis*, pages 289–311. John Wiley & Sons Ltd, 2007.
- [5] Richard G. Brereton. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*. John Wiley & Sons Ltd, 2003.
- [6] Donald A. Burns and Emil W. Ciurczak, editors. *Handbook of Near-Infrared Analysis*. Taylor & Francis, 2nd edition, 2001.
- [7] Max Bylesjö, Olivier Cloarec, and Mattias Rantalainen. Normalization and Closure. In Stephen D. Brown, Romà Tauler, and Beata Walczak, editors, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, chapter 2.07, pages 109–127. Elsevier B.V., 2009.

- [8] Francis A. Carey. *Organic Chemistry*. McGraw-Hill Higher Education, 7th edition, 2008.
- [9] Edward A. Cloutis. Spectral reflectance properties of hydrocarbons: remote-sensing implications. *Science (New York, N.Y.)*, 245(4914):165–8, July 1989.
- [10] A. de Juan and Romà Tauler. Linear Soft-Modeling: Introduction. In Stephen D. Brown, Romà Tauler, and Beata Walczak, editors, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, chapter 2.12, pages 207–210. Elsevier B.V., 2009.
- [11] Ladan Ebadi, Helmi Z. M. Shafri, Shattri B. Mansor, and Ravshan Ashurov. A review of applying second-generation wavelets for noise removal from remote sensing data. *Environmental Earth Sciences*, March 2013.
- [12] J.M. Ellis, H.H. Davis, and J.A. Zamudio. Exploring for onshore oil seeps with hyperspectral imaging. *Oil and Gas Journal*, 99(37):49–58, 2001.
- [13] Kim H. Esbensen, Dominique Guyot, Frank Westad, and Lars P. Houmøller. *Multivariate Data Analysis in Practice*. Camo Process AS Paperback, 5th edition, 2002.
- [14] Kim H. Esbensen and Thorbjørn T. Lied. Principles of Image Cross-validation (ICV): Representative Segmentation of Image Data Structures. In Hans F. Grahn and Paul Geladi, editors, *Techniques and Applications of Hyperspectral Image Analysis*, pages 155–180. John Wiley & Sons, Ltd, 2007.
- [15] F. M. Flasar, V. G. Kunde, M. M. Abbas, R. K. Achterberg, P. Ade, A. Barucci, B. Bézard, G. L. Bjoraker, J. C. Brasunas, S. Calcutt, R. Carlson, C. J. Cesarsky, and B. J. Conrath. Exploring the Saturn system in the thermal infrared: The composite infrared spectrometer. *Space Science Reviews*, 115(1-4):169–297, 2005.
- [16] F. García-Vílchez, J. Muñoz Marí, M. Zortea, I. Blanes, V. González-Ruiz, G. Camps-Valls, A. Plaza, and J. Serra-Sagristà. On the impact of lossy compression on hyperspectral image classification and unmixing. *IEEE Geoscience and Remote Sensing Letters*, 8(2):253–257, 2011.
- [17] Nahum Gat. Imaging Spectroscopy Using Tunable Filters: A Review. *Proceedings of SPIE*, 4056:50–64, 2000.

- [18] Steven C. Gebhart. Instrumentation considerations in spectral imaging for tissue demarcation: comparing three methods of spectral resolution. *Proceedings of SPIE*, 5694:41–52, 2005.
- [19] Paul Geladi and Hans Grahn. *Multivariate Image Analysis*. John Wiley & Sons Ltd, 1996.
- [20] Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, 1986.
- [21] Paul L. M. Geladi, Hans F. Grahn, and James E. Burger. Multivariate Images, Hyperspectral Imaging: Background and Equipment. In Hans F. Grahn and Paul Geladi, editors, *Techniques and Applications of Hyperspectral Image Analysis*, pages 1–15. John Wiley & Sons, Ltd, 2007.
- [22] C Gendrin, Y Roggo, and C Collet. Pharmaceutical applications of vibrational chemical imaging and chemometrics: a review. *Journal of pharmaceutical and biomedical analysis*, 48(3):533–53, November 2008.
- [23] Rafael C. Gonzalez, Richard E. Woods, and Steven L. Eddins. *Digital Image Processing using MATLAB*. Pearson Prentice Hall, 2004.
- [24] A.A. Gowen, C P O’Donnell, P J Cullen, and S E J Bell. Recent applications of Chemical Imaging to pharmaceutical process monitoring and quality control. *European journal of pharmaceuticals and biopharmaceutics : official journal of Arbeitsgemeinschaft für Pharmazeutische Verfahrenstechnik e. V*, 69(1):10–22, May 2008.
- [25] A.A. Gowen, C.P. O’Donnell, P.J. Cullen, G. Downey, and J.M. Frias. Hyperspectral imaging – an emerging process analytical tool for food quality and safety control. *Trends in Food Science & Technology*, 18(12):590–598, December 2007.
- [26] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.
- [27] Philip K. Hopke. Environmental Chemometrics. In Stephen D. Brown, Romà Tauler, and Beata Walczak, editors, *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, chapter 4.03, pages 55–74. Elsevier B.V., 2009.

- [28] B. Hörig, F. Kühn, F. Oschütz, and F. Lehmann. HyMap hyperspectral remote sensing to detect hydrocarbons. *International Journal of Remote Sensing*, 22(8):1413–1422, May 2001.
- [29] James N. Huckins, Jimmie D. Petty, and Kees Booij. *Monitors of Organic Chemicals in the Environment: Semipermeable Membrane Devices*. Springer, 2006.
- [30] Tomas Isaksson and Tormod Næs. The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy. *Applied Spectroscopy*, 42(7):1273–1284, 1988.
- [31] Sveinn R. Joelsson, Jon Atli Benediktsson, and Johannes R. Sveinsson. Feature Selection for Morphological Feature Extraction using Random Forests. *Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG 2006*, pages 10–13, June 2006.
- [32] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.
- [33] Kirsten E. Kramer, Robert E. Morris, and Susan L. Rose-Pehrsson. Comparison of two multiplicative signal correction strategies for calibration transfer without standards. *Chemometrics and Intelligent Laboratory Systems*, 92(1):33–43, May 2008.
- [34] F. Kühn, K. Oppermann, and B. Hörig. Hydrocarbon Index – an algorithm for hyperspectral detection of hydrocarbons. *International Journal of Remote Sensing*, 25(12):2467–2473, June 2004.
- [35] Olav M. Kvalheim, Frode Brakstad, and Yi-zeng Liang. Preprocessing of analytical profiles in the presence of homoscedastic or heteroscedastic noise. *Analytical Chemistry*, 66(1):43–51, 1994.
- [36] Jeroen Lammertyn, Ann Peirs, Josse De Baerdemaeker, and Bart Nicolai. Light penetration properties of NIR radiation in fruit with respect to non-destructive quality assessment. *Postharvest Biology and Technology*, 18:121–132, 2000.
- [37] E. Neil Lewis, Janie Dubois, Linda H. Kidder, and Kenneth S. Haber. Near Infrared Chemical Imaging: Beyond the Pictures. In Hans F. Grahn and Paul Geladi, editors, *Techniques and Applications of Hyperspectral Image Analysis*, pages 335–361. John Wiley & Sons Ltd, 2007.

- [38] Nor Rizuan Mat Noor and Tanya Vladimirova. International Journal of Remote Investigation into lossless hyperspectral image compression for satellite remote sensing. *International Journal of Remote Sensing*, 34(14):5072–5104, 2013.
- [39] Norsk Elektro Optikk A/S. HySpex SWIR-320i.
- [40] Antonio Plaza, Jon Atli Benediktsson, Joseph W. Boardman, Jason Brazile, Lorenzo Bruzzone, Gustavo Camps-Valls, Jocelyn Chanussot, Mathieu Fauvel, Paolo Gamba, Anthony Gualtieri, Mattia Marconcini, James C. Tilton, and Giovanna Trianni. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment*, 113:S110–S122, September 2009.
- [41] R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–44, 2006.
- [42] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Pearson Education, Inc., New Jersey, 2nd edition, 2003.
- [43] I.D. Sanches, C.R. Souza Filho, L.a. Magalhães, G.C.M. Quitério, M.N. Alves, and W.J. Oliveira. Assessing the impact of hydrocarbon leakages on vegetation using reflectance spectroscopy. *ISPRS Journal of Photogrammetry and Remote Sensing*, 78:85–101, April 2013.
- [44] Dietmar Schumacher. Hydrocarbon-induced alteration of soils and sediments. In Dietmar Schumacher and Michael A. Abrams, editors, *Hydrocarbon migration and its near-surface expression*, chapter 6, pages 71–89. The American Association of Petroleum Geologists, 1996.
- [45] R. Glenn Sellar and Glenn D. Boreman. Classification of imaging spectrometers for remote sensing applications. *Optical Engineering*, 44(1), January 2005.
- [46] Claude. E. Shannon. Communication In The Presence Of Noise. *Proceedings of the IEEE*, 86(2):447–457, February 1998.
- [47] Xin-Hua Song and Philip K Hopke. Solving the Chemical Mass Balance Problem Using an Artificial Neural Network. *Environmental Science and Technology*, 30(2):531–535, 1996.
- [48] Suresh Subramanian, Nahum Gat, Alan Ratcliff, and Michael Eismann. Real-time hyperspectral data compression using principal com-

- ponents transformation. In *AVIRIS Earth Science and Applications Workshop*, Pasadena, CA, 2000.
- [49] Wim Sweldens. The Lifting Scheme: A Custom-Design Construction of Biorthogonal Wavelets. *Applied and Computational Harmonic Analysis*, 3(2):186–200, April 1996.
- [50] The MathWorks. MATLAB Release 2011b.
- [51] Paul A. Tipler and Gene Mosca. *Physics for Scientists and Engineers*. W. H. Freeman and Company, New York, 6th edition, 2008.
- [52] Chieu D. Tran. Principles , Instrumentation, and Applications of Infrared Multispectral Imaging, An Overview. *Analytical Letters*, 38:735–752, 2005.
- [53] Freek van der Meer and Paul van Dijk. Remote sensing and petroleum seepage: a review and case study. *Terra Nova*, 14:1–17, 2002.
- [54] Freek D. van der Meer, Harald M.a. van der Werff, Frank J.a. van Ruitenbeek, Chris a. Hecker, Wim H. Bakker, Marleen F. Noomen, Mark van der Meijde, E. John M. Carranza, J. Boudewijn De Smeth, and Tsehaie Woldai. Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):112–128, February 2012.
- [55] Kurt Varmuza and Peter Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press, 2009.
- [56] Gerard A. Venema. *The foundations of geometry*. Pearson Prentice Hall, Upper Saddle River, NJ, 2006.
- [57] Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, and Keying Ye. *Probability & Statistics for Engineers & Scientists*. Pearson Education, Inc., Upper Saddle River, NJ, 8th edition, 2007.
- [58] Magnus Wettle, Paul J. Daniel, Graham a. Logan, and Medhavy Thankappan. Assessing the effect of hydrocarbon oil type and thickness on a remote sensing signal: A sensitivity study based on the optical properties of two different oil types and the HYMAP and Quickbird sensors. *Remote Sensing of Environment*, 113(9):2000–2010, September 2009.
- [59] L G Weyer. Near-Infrared Spectroscopy of Organic Substances. *Applied Spectroscopy Reviews*, 21(1 & 2):1–43, 1985.

Appendix A

Wavelengths for the hyperspectral camera

Table A.1 shows the wavelengths for the hyperspectral camera used in this thesis. The wavelengths span 923-1665 nm.

Table A.1: Wavebands for the PryJector.

No.	Wavelength (nm)	No.	Wavelength (nm)	No.	Wavelength (nm)
1	923.071532	2	928.121589	3	933.171646
4	938.221703	5	943.271759	6	948.321816
7	953.371873	8	958.421930	9	963.471987
10	968.522043	11	973.572100	12	978.622157
13	983.672214	14	988.722271	15	993.772328
16	998.822384	17	1003.872441	18	1008.922498
19	1013.972555	20	1019.022612	21	1024.072668
22	1029.122725	23	1034.172782	24	1039.222839
25	1044.272896	26	1049.322953	27	1054.373009
28	1059.423066	29	1064.473123	30	1069.523180
31	1074.573237	32	1079.623293	33	1084.673350
34	1089.723407	35	1094.773464	36	1099.823521
37	1104.873577	38	1109.923634	39	1114.973691
40	1120.023748	41	1125.073805	42	1130.123862
43	1135.173918	44	1140.223975	45	1145.274032
46	1150.324089	47	1155.374146	48	1160.424202
49	1165.474259	50	1170.524316	51	1175.574373

APPENDIX A. WAVELENGTHS FOR THE HYPERSPSCTRAL
CAMERA

No.	Wavelength (nm)	No.	Wavelength (nm)	No.	Wavelength (nm)
52	1180.624430	53	1185.674487	54	1190.724543
55	1195.774600	56	1200.824657	57	1205.874714
58	1210.924771	59	1215.974827	60	1221.024884
61	1226.074941	62	1231.124998	63	1236.175055
64	1241.225112	65	1246.275168	66	1251.325225
67	1256.375282	68	1261.425339	69	1266.475396
70	1271.525452	71	1276.575509	72	1281.625566
73	1286.675623	74	1291.725680	75	1296.775736
76	1301.825793	77	1306.875850	78	1311.925907
79	1316.975964	80	1322.026021	81	1327.076077
82	1332.126134	83	1337.176191	84	1342.226248
85	1347.276305	86	1352.326361	87	1357.376418
88	1362.426475	89	1367.476532	90	1372.526589
91	1377.576646	92	1382.626702	93	1387.676759
94	1392.726816	95	1397.776873	96	1402.826930
97	1407.876986	98	1412.927043	99	1417.977100
100	1423.027157	101	1428.077214	102	1433.127270
103	1438.177327	104	1443.227384	105	1448.277441
106	1453.327498	107	1458.377555	108	1463.427611
109	1468.477668	110	1473.527725	111	1478.577782
112	1483.627839	113	1488.677895	114	1493.727952
115	1498.778009	116	1503.828066	117	1508.878123
118	1513.928180	119	1518.978236	120	1524.028293
121	1529.078350	122	1534.128407	123	1539.178464
124	1544.228520	125	1549.278577	126	1554.328634
127	1559.378691	128	1564.428748	129	1569.478805
130	1574.528861	131	1579.578918	132	1584.628975
133	1589.679032	134	1594.729089	135	1599.779145
136	1604.829202	137	1609.879259	138	1614.929316
139	1619.979373	140	1625.029429	141	1630.079486
142	1635.129543	143	1640.179600	144	1645.229657
145	1650.279714	146	1655.329770	147	1660.379827
148	1665.429884				

The waveband models from section 4.1.3 use wavelength number 50-65, that is 1170.5 nm through 1246.3 nm. Some models also use wavelength number 125, at 1549.3 nm.

Appendix B

Surface specific error, data set 3b

Table B.1 shows the percentage error for all images and surfaces included in data set 3b (table 3.6). The models are the same as in table 5.1.

The DPLSR model have 11 PLS components. The k -NN model uses $k = 5$. Both these models uses raw data. The waveband model uses data smoothed by moving average, window size = 9 and wavelengths from 1186 nm to 1246 nm (number 53-65 in appendix A).

Table B.1: Percentage error for the images and surfaces included in data set 3b (table 3.6).

Description	Table	DPLSR	k -NN	Waveband
Dry soil, image 1	3.1	0	6.3	0
Wet soil, image 1	3.1	19.0	3.2	7.1
Soil with paraffin, image 1	3.1	1.0	1.0	1.0
Dry soil, image 2	3.1	1.8	3.5	0.9
Wet soil, image 2	3.1	16.8	6.6	3.6
Soil with paraffin, image 2	3.1	0.5	4.7	1.0
10 ml paraffin / 15 g soil, after 0 days	3.2	1.6	8.1	5.6
10 ml paraffin / 15 g soil, after 4 days	3.2	0	0	6.2
15 ml paraffin / 15 g soil, after 0 days	3.2	0.7	4.4	0
15 ml paraffin / 15 g soil, after 4 days	3.2	0	0.7	1.5
Leaf, a bit wet	3.3	0	0	0.9
Leaf, with paraffin	3.3	0	2.8	0
Lichen	3.3	16.3	36.7	5.1
Lichen with paraffin	3.3	13.1	23.2	2.0

APPENDIX B. SURFACE SPECIFIC ERROR, DATA SET 3B

Description	Table	DPLSR	k -NN	Waveband
Cones	3.3	2.0	17.2	11.1
Cones with paraffin	3.3	57.6	37.4	3.0
Moss	3.3	1.0	1.0	53.0
Moss with paraffin	3.3	8.9	0	0
Plants and leaves	3.3	0	2.5	2.0
Plants with paraffin	3.3	2.0	0	0
Grass, a bit wet	3.3	0	2.5	18.2
Grass with paraffin	3.3	12.5	9.4	0
Stones, dry	3.3	3.6	0.5	1.5
Stones with paraffin	3.3	15.5	7.0	9.1
Stones, wet	3.3	30.5	6.1	5.6
Wet sand and stones	3.3	83.4	7.8	15.5
Humus, image 1	3.4	15.1	22.3	10.3
Stones, image 1	3.4	24.9	8.9	9.4
Sand, image 1	3.4	5.4	4.1	6.8
Soil from <i>Plantasjen</i> , image 1	3.4	12.6	7.1	4.5
Compost soil, image 1	3.4	1.8	3.5	1.0
Humus, image 2	3.4	21.0	27.6	13.9
Stones, image 2	3.4	9.8	7.0	59.2
Sand, image 2	3.4	5.2	5.8	66.5
Soil from <i>Plantasjen</i> , image 2	3.4	0.4	3.8	3.4
Compost soil, image 2	3.4	5.1	6.6	11.4
Humus, image 3	3.4	22.8	25.0	19.9
Stones, image 3	3.4	9.0	7.1	56.7
Sand, image 3	3.4	10.9	8.2	69.1
Soil from <i>Plantasjen</i> , image 3	3.4	0.8	3.0	3.4
Compost soil, image 3	3.4	2.9	5.8	13.8
Humus, image 4	3.4	22.2	28.2	16.2
Stones, image 4	3.4	9.8	12.8	64.4
Sand, image 4	3.4	15.8	12.0	65.8
Soil from <i>Plantasjen</i> , image 4	3.4	0.7	0.3	3.5
Compost soil, image 4	3.4	2.0	5.3	14.9
Total error		12.5	9.3	16.3