Norwegian University of Life Sciences
Faculty of Science and Technology

Philosophiae Doctor (PhD)
Thesis 2019:77

Hyperspectral imaging: algorithmic advances in variable selection and applications to wood science

Hyperspektral avbildning: algoritmiske fremskritt innen variabelt utvalg og anvendelser til trevitenskap

Petter Stefansson

Hyperspectral imaging: algorithmic advances in variable selection and applications to wood science

Hyperspektral avbildning: algoritmiske fremskritt innen variabelt utvalg og anvendelser til trevitenskap

Philosophiae Doctor (PhD) Thesis

Petter Stefansson

Norwegian University of Life Sciences
Faculty of Science and Technology

Ås (2019)

Thesis number 2019:77
ISSN 1894-6402
ISBN 978-82-575-1637-6

Supervisors

Dr. Ingunn Burud

Faculty of Science and Technology, Norwegian University Of Life Sciences, Ås, Norway

Prof. Thomas Kringlebotn Thiis

Faculty of Science and Technology, Norwegian University Of Life Sciences, Ås, Norway

Dr. Kristian Hovde Liland

Faculty of Science and Technology, Norwegian University Of Life Sciences, Ås, Norway

Dr. Lone Ross Gobakken

Norwegian Institute of Bioeconomy Research, Ås, Norway

Evaluation committee

Prof. Olle Hagman

Luleå University of Technology, Skellefteå, Sweden

Prof. Federico Marini

Dept. of Chemistry, University of Rome “La Sapienza”, Rome, Italy

Prof. Knut Kvaal

Faculty of Science and Technology, Norwegian University Of Life Sciences, Ås, Norway

Summary

According to Beer's Law there is a linear dependence between the absorbance of a material and the concentration of an absorbing species in the material. Thus, if one is interested in modeling the concentration of an absorbing species, it should be possible to do so by utilizing a linear model to describe the concentration of the species from a measurement of the absorbance of the material. This thesis is concerned with developing such models from hyperspectral measurements taken in the visible (vis) and near infrared (NIR) region of the electromagnetic spectrum.

When developing such models, it is frequently the case that a majority of the wavelengths within a measured spectrum are not absorbed by the species of interest - and should therefore preferably be excluded from the developed model in order to optimize its performance. The process of identifying unnecessary wavelengths is often driven by trial and error, as such it tends to be time consuming and computationally demanding. During the work leading up to **Paper I** we discovered a conceptually very simple technique which allows calculations to be recycled when developing *partial least squares* (PLS) models from different combinations of wavelengths. The technique can greatly reduce the computational cost of fitting multiple regression models with various combinations of included/excluded wavelengths to a dataset. In **Paper II** we incorporate the findings of **Paper I** into a *genetic algorithm* (GA) and demonstrate that the technique also can be used to simultaneously evaluate—in a computationally efficient manner—combinations of wavelengths which are preprocessed using different techniques.

In **Paper III** and **IV** we develop models which solve wood science related issues.

In **Paper III** samples of spruce (*Picea abies*) treated with a phosphorus-based flame retardant compound were scanned using a NIR hyperspectral camera. The resulting data was subsequently used to develop a PLS model which estimated the phosphorous content from the spectral signal.

In **Paper IV** samples of thermally modified pine (*Pinus sylvestris*) were repeatedly scanned over time as they dried. The resulting time series sequences of hyperspectral NIR data was used to develop a regression model capable of estimating the moisture content of the pine from the spectra.

In **Paper V**[1] a generic method is developed for studying and summarizing hyperspectral time series sequences in terms of known and unknown variations. The main idea of the presented method is that spectral variations of known origin are removed from the data. The remaining residual data, containing variation of unknown origin, is then subjected to dimensionality reduction in order to identify new previously unknown variations in the data; variations which in the case of hyperspectral time series data may exhibit temporal as well as spatial patterns of interest. The developed concept was experimentally evaluated in **Paper V** on a piece of unmodified spruce (*Picea abies*) which was monitored using a vis-NIR hyperspectral camera as it dried over the course of 21 hours.

[1] Although referred to as a paper in this thesis for convenience, **Paper V** is really a book chapter and not a conventional paper.

Sammendrag

I følge Beer's Law er det en lineær avhengighet mellom absorbansen av et materiale og konsentrasjonen av en absorberende art i materialet. Således, hvis man er interessert i å modellere konsentrasjonen av en absorberende art, bør det være mulig å gjøre det ved å benytte en lineær modell for å beskrive konsentrasjonen av arten fra en måling av absorbans av materialet. Denne avhandlingen er opptatt av å utvikle slike modeller fra hyperspektrale målinger som er tatt i det synlige (vis) og nær infrarøde området (NIR) i det elektromagnetiske spekteret.

Når man utvikler slike modeller, er det ofte slik at et flertall av bølgelengdene innenfor et målt spektrum ikke blir absorbert av artene som er av interesse - og derfor bør de helst utelukkes fra den utviklede modellen for å optimalisere ytelsen. Prosessen med å identifisere unødvendige bølgelengder er ofte drevet av prøving og feiling, og har en tendens til å være tidkrevende og vanskelig. Under arbeidet frem til **Paper I** oppdaget vi en konseptuelt veldig enkel teknikk som gjør det mulig å resirkulere beregninger når vi utvikler delvis minste kvadrater (PLS) -modeller fra forskjellige kombinasjoner av bølgelengder. Teknikken kan redusere beregningskostnadene for å montere flere regresjonsmodeller med forskjellige kombinasjoner av inkluderte/ekskluderte bølgelengder til et datasett. I **Paper II** innlemmer vi funnene fra **Paper I** i en *genetisk algoritme* (GA) og demonstrerer at teknikken også kan brukes til å evaluere—på en beregningseffektiv måte—kombinasjoner av bølgelengder som er forbehandlet med forskjellige teknikker.

I **Paper III** og **IV** utvikler vi modeller som løser trevitenskapelige problemer. I **Paper III** ble prøver av gran (*Picea abies*) behandlet med en fosforbasert

flammehemmende forbindelse skannet ved bruk av et NIR hyperspektralkamera. De resulterende data ble deretter brukt til å utvikle en PLS-modell som estimerte fosforinnholdet fra spektralsignalet.

I **Paper IV** ble prøver av termisk modifisert furu (*Pinus sylvestris*) gjentatte ganger skannet over tid mens de tørket. De resulterende tidsseriesekvensene med hyperspektrale NIR-data ble brukt til å utvikle en regresjonsmodell som var i stand til å estimere fuktighetsinnholdet i furu fra spektrene.

I **Paper V**[1] utvikles en generisk metode for å studere og oppsummere hyperspektrale tidsseriesekvenser i form av kjente og ukjente variasjoner. Hovedideen med den presenterte metoden er at spektrale variasjoner av kjent opprinnelse blir fjernet fra dataene. De resterende restdataene, som inneholder variasjon av ukjent opprinnelse, blir deretter utsatt for dimensjonalitetsreduksjon for å identifisere nye tidligere ukjente variasjoner i dataene; variasjoner som i tilfelle av data fra hyperspektrale tidsserier kan utvise tidsmessige så vel som romlige mønstre av interesse. Det utviklede konseptet ble eksperimentelt evaluert i **Paper V** på et stykke umodifisert gran (*Picea abies*) som ble overvåket ved bruk av et vis-NIR hyperspektralkamera da det tørket i løpet av 21 timer.

[1]Selv om det er referert til som en artikkel i denne oppgaven for enkelhets skyld, er **Paper V** virkelig et bokkapittel og ikke en vanlig journal-artikkel.

Acknowledgments

I would like to thank all my co-authors, supervisors and co-supervisors. Especially, I would like to thank Ingunn Burud and Thomas Thiis for deciding to hire me as a PhD student in the first place. Given how astonishingly mediocre/pathetic my previous university grades were, and how I completely lacked any type of experience in the field of hyperspectral imaging, it still puzzles me greatly to this day why you chose to hire me. But I'm happy you did, because I have learned a lot during my years in Norway.

Petter Stefansson

Malmö, August 2019

List of papers

Paper I

Stefansson, Petter; Indahl, Ulf; Liland, Kristian; Burud, Ingunn. *Orders of magnitude speed increase in Partial Least Squares feature selection with new simple indexing technique for very tall datasets.* Journal of Chemometrics 2019 (in press).

Paper II

Stefansson, Petter; Liland, Kristian; Thiis, Thomas; Burud, Ingunn. *Fast method for GA-PLS with simultaneous feature selection and identification of optimal preprocess technique for datasets with many observations.* Submitted to Journal of Chemometrics.

Paper III

Stefansson, Petter; Burud, Ingunn; Thiis, Thomas; Gobakken, Lone; Larnøy, Erik. *Estimation of phosphorus-based flame retardant in wood by hyperspectral imaging—a new method.* Journal of Spectral Imaging 2017; Volume 7.

Paper IV

Stefansson, Petter; Thiis, Thomas; Gobakken, Lone; Burud, Ingunn. *Hyperspectral NIR time series imaging used as a new method for estimating the moisture content dynamics of thermally modified scots pine.* Submitted to Wood Material Science Engineering.

Paper V

Stefansson, Petter; Fortuna, João; Rahmati, Hodjat; Burud, Ingunn; Konevskikh, Tatiana; Martens, Harald. *Hyperspectral time series analysis: Hyperspectral image data streams interpreted by modeling known and unknown variations.* Hyperspectral Imaging, Volume 32, 1st edition, 2019.

Contents

CHAPTER 1

Introduction & aims

The main theme of this thesis is making models for predicting material properties based on the way the material in question interacts with light. Models which take a multivariate signal as input, collected in a nondestructive way by shining a light at the material and measuring how it reflects/absorbs light at different wavelengths, and outputs a prediction about the material. Such as its content of a particular chemical. The instrument used throughout this thesis to measure how a material interacts with light has been a hyperspectral camera; which is an instrument capable of measuring how an object reflects light as a function both of wavelength and spatial location in the object. In contrast to traditional spectroscopic instruments which only measure the spectral reflectance/absorbance in a single point, hyperspectral cameras can be used to study heterogeneous materials where the reflectance/absorbance spatially varies within the material. An example of such a material is wood, which is studied using hyperspectral imaging in several of the articles in this thesis. Because hyperspectral measurements are spatially resolved, the models developed from such data can be used to estimate the spatial distribution of a chemical's content in the material—generating what is known as a *chemical map* of the material—which opens the door to new ways of analyzing the material and allows rapid and nondestructive surveying of the material.

In two of the papers included in this thesis, models utilizing hyperspectral data are developed from experimentally collected measurements and used to address

wood-related research questions. Namely, to estimate the phosphorus content in spruce surfaces and to model the moisture content of thermally modified pine. Why such estimates are needed will be outlined in section 1.3. In two other papers, hyperspectral model development is explored from a more theoretical angle, where techniques are introduced which allow models to be developed faster than previously possible by introducing new, computationally efficient, techniques for conducting variable selection. I.e., identifying which of the wavelengths in a measured hyperspectral signal hold significant predictive capacity over the chemical/material property one wishes to model. Identification of such wavelengths is important both because it can increase the performance of the developed model and because it can aid in understanding the underlying data and its relation to the analyte of interest. In another paper, hyperspectral time series data of a drying pine sample is studied, and a generic new methodology is developed and evaluated for analyzing spatiotemporal spectral data. The developed technique allows large amounts of spectral data to be quantified and summarized in terms of wavelength-dependent variations of either beforehand known and unknown origin.

> Fundamentally, this thesis can be said to be concerned with two things: algorithmic advances in wavelength selection and investigating wood-related research questions by developing models where the input consists of hyperspectral data.

1.1 Objective of this research

In the work leading up to this thesis three hyperspectral datasets were experimentally collected, details of these will be given in chapter 3. These datasets are central to the thesis as all the included articles and presented

findings relates to these datasets in one form or another: either a dataset was used to practically address a wood-related research question, or the dataset was used to evaluate theoretical concepts related to wavelength selection and/or spectral preprocessing. Wavelength selection and spectral preprocessing are both concepts which will be introduced and motivated in chapter 2. Suffice to say at this point is that they are both integral parts of developing functional models from hyperspectral data.

The topic of this thesis—studying wood using hyperspectral imaging and developing novel computational techniques for hyperspectral wavelength selection—is investigated from different angles in five scientific papers. The individual research objective of each paper in this thesis can be summarized as follows:

1. the aim of **Paper I** is to put forth a new technique for performing fast variable selection with partial least squares regression, suitable for hyperspectral data containing many observations in relation to the number of variables;
2. the aim of **Paper II** is to develop a method which competently identifies both a wavelength section and a suitable choice of preprocessing technique for a given hyperspectral dataset. This will be realized by generalizing the research outcome of **Paper I** such that it can be applied to several differently preprocessed versions of a dataset simultaneously, thus making it easier to review the performance of different preprocessing techniques after they have undergone wavelength selection;
3. the aim of **Paper III** is to evaluate the possibility of using hyperspectral NIR imaging to nondestructively estimate the amount of phosphorous-based flame retardant present in spruce surfaces;

④ the aim of **Paper IV** is to investigate if hyperspectral imaging in NIR region can be used as a means of estimating the moisture content within thermally modified pine and if the technique can be used to gain insight to how thermally modified wood dries differently in various regions of a board by studying samples over time as they dry;

⑤ the aim of **Paper V** is to introduce a new explorative method for analyzing spatiotemporal spectral data, i.e. hyperspectral time series data. The technique will require the user to provide information about spectral variations of known origin believed to be present in the dataset and will attempt to output a summary of detected systematic spectral variations whose origins are unknown to the user. This will then enable the user to study the unknown spectral variations and ideally bring them into the domain of known variations whilst additionally allowing the data to be greatly compressed.

Although the objective varies between the papers in the thesis, the papers are interconnected on either a conceptual level, or in terms of the dataset used in the research, or both. In addition to all articles utilizing hyperspectral data depicting wood in one way or another, some of the papers are more directly linked to each other: **Paper I** and **Paper II** are both about computationally efficient variable selection performed using a novel indexing technique; **Paper IV** and **Paper V** are both concerned with analyzing hyperspectral time series data of pine; **Paper III** and **Paper IV** are both about developing regression models for predicting the spatial distribution of chemicals in wood. Figure 1.1 illustrates how the different papers of the thesis are connected to each other and to the datasets.

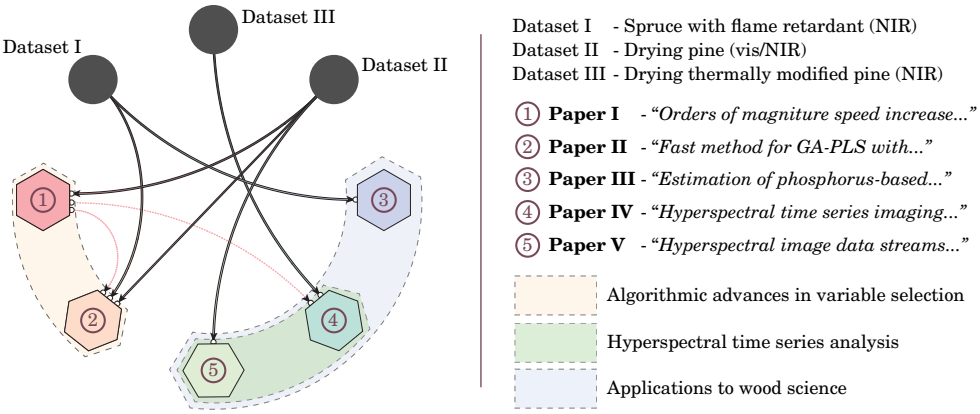


FIGURE 1.1 Illustration of how the five papers in the thesis are related to each other and to the three datasets.

1.2 Layout of the thesis

In section 1.3 a motivation is given for why it is worthwhile to study wood, and why hyperspectral measurements are particularly well suited for certain wood material studies. The section will also motivate the wood-related research goals of **Paper III** and **IV**. The second main topic of this thesis—variable selection—will be introduced and motivated separately in section 2.4. Chapter 2 contains an explanation of all relevant concepts used in the papers which are attached in the appendix. Chapter 3 provides a summary of the experimentally collected datasets. Chapter 4 presents and discusses the key findings of the thesis. List form conclusions and research outcomes of the thesis are given in chapter 5.

1.3 Why study wood?

As a construction material wood is experiencing something of a renaissance at the moment—increased demand for environmentally sustainable building materials has led to an increased use of wood as it is a material which is both renewable and stores carbon [1]. Life-cycle assessment studies of carbon

emissions of various construction materials consistently indicate that increasing the use of wood in the construction sector, by means of substituting for instance concrete, lowers the long term carbon footprint of a building [2, 3]. Despite its environmental advantages, there are aspects of wood as a construction material which are suboptimal.

Historically, fire safety has long been the Achilles' heel of wood as it is a notoriously combustible material. An illustrative local example of this is that Oslo, the capital of Norway, was largely annihilated by city fires on four separate occasions between the years 1523 and 1624 alone—roughly once every 25 years—after which the king of Norway forbade the use of wood as a construction material within the city walls [4]. This example is far from unique since many European cities are plagued by a history of periodic fires followed by a ban/limitation imposed on the use of wood as a construction material. Today most European cities have however lifted their bans and it is permitted again to build with wood, often even multistorey buildings [5]. Provided, of course, that the buildings fulfill the regulations regarding fire safety. In order to meet such regulations, wood is sometimes impregnated with flame retardant chemicals before being used in the construction phase. The purpose of such treatments is to delay the structural failure and the flashover-point of the wood such that the occupants are given ample time to evacuate the building in case of a fire. However, a problem associated with such treatments is that it is known that substances impregnated into wood such as flame retardants leach out from the wood over time [6]. The rate at which the leaching occurs can be difficult to predict. In order to verify that the fire safety requirements of a wood-based building are still met some arbitrary time after the construction of the building, surveying methods are therefore needed which can quantify the current chemical composition of wooden structures. Using well-established chemical analysis methods such as inductively coupled plasma (ICP) analysis, this can be achieved by extracting

samples from the wooden structure and determining the chemical composition of the samples in a laboratory. The disadvantages of this is that: (1) the technique is destructive since material needs to be physically removed from the structure before it can be analyzed; (2) it is point based, meaning that only small regions of the structure is surveyed. If there has been a large spatial variability in the leaching rate throughout the structures lifetime, for instance due to harsh microclimate conditions in certain parts of the building such as around window sills etc., a potentially important spatial variation in chemical content within the structure could be missed. The development of new, preferably nondestructive, surveying methods are therefore warranted—methods capable of surveying large areas of a wooden structure.

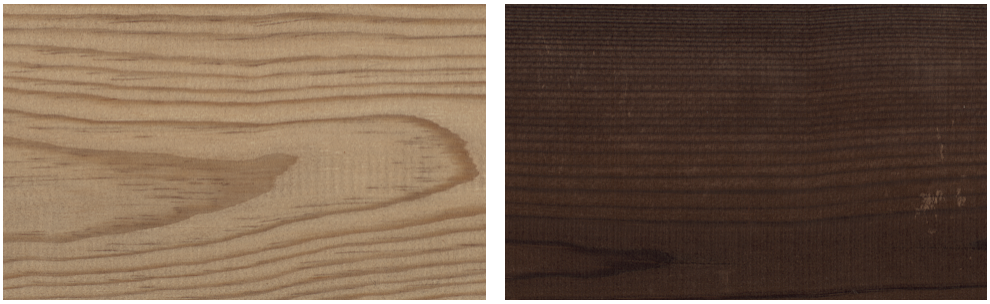
Another weakness of wood besides its combustibility is its susceptibility to moisture-related problems. These problems can take many forms: risk of mould growth and decay are elevated in moist conditions [7], differential swelling and contraction of different regions of a wood board due to spatial variations in moisture content can cause the wood or coatings applied to the wood to crack [8], the wood can bend and deform as a result of transient wetting and drying, its color can significantly change when exposed to moisture for longer periods [9] which can be aesthetically displeasing, etc. One treatment method for combating moisture-related inconveniences is *thermal modification* [10, 11]. Simply put, thermal modification of wood is a process where wood is heated in an oven with a reduced supply of oxygen. Visually the effect of the treatment is that the wood gets notably darker [10, 12], as can be seen in Fig. 1.2 showing two pieces of pine in both unmodified and thermally modified form. Property-wise there are multiple effects of the treatment. For instance the dimensional stability of the wood can be improved [13] and its resistance against fungi and mold growth increases [14, 15]. The thermal modification process does however decrease the strength of the wood [12], which is why thermally treated wood is generally not

used in load-bearing structures, but rather in facades, flooring, decking etc. [15]. Because nothing is added to the wood and organic compounds are burnt away during the treatment, the wood also weighs less after the thermal modification compared to before [10, 12]. It is recognized that the moisture dynamics of wood is altered by thermal treatments and that wood's equilibrium moisture content (EMC) decreases when thermally modified [10, 11]. Precisely how the moisture uptake in a piece of wood spatially changes after thermal modification is largely unknown. Some wood technologists speculate that the moisture uptake of thermally modified wood ought be more evenly distributed throughout the wood compared to unmodified wood—since some of the chemical causes for wood's heterogeneousness are burnt away in the thermal modification process. Previous studies have been able to use magnetic resonance imaging (MRI) to study and visualize the distribution of moisture in thermally modified pine [16, 17], but the resolution offered by such techniques is relatively low. New techniques are therefore needed which can predict the distribution of moisture at a high spatial resolution in order to broaden the current scientific understanding of the moisture dynamics of thermally modified timber.

> In this thesis, hyperspectral imaging is evaluated as a tool for estimating the content of a flame retardant chemical in spruce (**Paper III**) and as a tool for studying the drying dynamics of thermally modified pine (**Paper IV**).

1.3.1 Why study wood with hyperspectral imaging?

Detailed knowledge of how a material reflects/absorbs radiation can be often be used to approximate several interesting things about the material. Since the reflectance/absorbance of an object can be measured nondestructively, techniques based on the utilization of spectroscopic data have become popular in the



a) b)

FIGURE 1.2 Visual appearance of two pieces of Scots pine in a) unmodified form; b) thermally modified form.

field of wood science, where the near-infrared (NIR) region of the spectrum is of particular interest due to the multitude of useful properties which can be estimated from such measurements. Studies have shown that information contained in the NIR spectrum can be used to predict several useful properties of wood. The mechanical stress of wood [18], the density [19] and the moisture content [20] are a few examples of properties which can be nondestructively estimated from the NIR spectrum.

Wood is also a heterogeneous material in several ways. During a tree's lifetime, its growth rate varies with seasonal changes in temperature and precipitation which gives rise to lower density *earlywood* regions and higher density *latewood* regions of the wood [21]. This is clearly visible in most wood boards as lighter respectively darker regions—as seen for example in Fig. 1.2. Wood from the dead inner part of the tree, the *heartwood*, is also different from living wood taken from the outermost parts of the trunk, the *sapwood* [22]. This heterogeneity makes hyperspectral imaging a particularly well-suited tool for studying wood. The proven techniques for estimating material properties from a spectrum, which traditionally have been point-based, are still applicable when using hyperspectral measurements, but hyperspectral imaging additionally

opens the door to studying spatial variations within the heterogeneous wood surface. In other words: if a property of wood can be estimated using point-based spectroscopic measurements, hyperspectral imaging can be used to estimate the spatial distribution of the same property within the wood. Despite the many advantages, it should be mentioned that spectroscopic methods are associated with a few drawbacks, the main one being that the penetration depth of visible/near-infrared radiation into the wood is very limited; effectively only the surface of the wood can be measured. Models developed from such surface measurements will therefore not be able to estimate property variations taking place deeper into the wood. This is not an issue when the surface-level content of a chemical or property is of interest. But when the average chemical content/material property throughout the thickness of the wood is of interest, this is a severe disadvantage. In such cases it is generally necessary to assume that the wood sample has a uniform distribution of the studied analyte throughout the thickness of the wood such that the spectra observed on the surface is representative of the wood as a whole. An assumption which to a varying degree will always be inaccurate.

CHAPTER 2

Methods & theoretical background

This chapter will serve as a foundation for understanding the papers included in the appendix. First, a general introduction will be given explaining why it is in theory possible to deduce something about a material simply by shining a light at it and measuring the reflected signal. This chapter will then transition into an increasingly practical explanation of all the key concepts and methodologies used throughout the thesis.

2.1 Why is it even possible to nondestructively say something about a material by studying the way it interacts with light?

The electrons of an atom are spatially confined around the center (the nucleus) of their atom according to a probabilistic distribution of locations [23]. By directing a source of electromagnetic radiation, i.e. light, towards an atom, the atom can absorb the photon energy. In doing so the energy level of the atom is elevated and the probability distribution associated with the location of the electrons changes. This photon absorption can only occur if the radiation shun

on the atom contains photons carrying an energy exactly equivalent to any of the energy levels of the atom [24]. The energy carried by a photon depends on the wavelength/frequency of the photon as there is a direct and well-established relation between energy and wavelength [25]:

$$E_{\text{phot}} = \frac{h \times c}{\lambda}. \tag{2.1}$$

In Eq. (2.1) E_{phot} is the energy of the photon, h is the Planck constant, c is the speed of light in vacuum and λ is the wavelength of the photon. Energy levels and electron orbitals differ between various atom types, and it is the electron orbitals of an atoms which usually dictate the chemical behavior of the atom [23]. By exposing an atom to radiation and measuring which wavelengths it absorbs, it is therefore possible to deduce information about the atom type—and by extension chemical information. In most material science however, materials made up from many different types of atoms, arranged into different types of molecules, are generally of interest. The previously mentioned energy levels associated with electrons orbiting the nucleus within atoms is one out of multiple energy types needed to describe the total energy state of a molecule. Generally, the energy state of a molecule E_{mol} can be approximated as the sum of four types of energies [26]:

$$E_{\text{mol}} = E_{\text{elect}} + E_{\text{vib}} + E_{rot} + E_{\text{trans}}. \tag{2.2}$$

In Eq. (2.2), E_{elect} is the energy associated with the electrons of the atoms within the molecule, E_{vib} is the vibrational energy of the molecule, E_{rot} is the rotational energy of the molecule and E_{trans} is the translational energy of the molecule. When atoms combine into molecules, they do so through various chemical bonds. The nature of these bonds causes the atoms within the molecules to be in constant motion; the atoms constantly vibrate along the chemical bonds like oscillating springs, the entire molecule rotates about the molecular axes and in liquids and

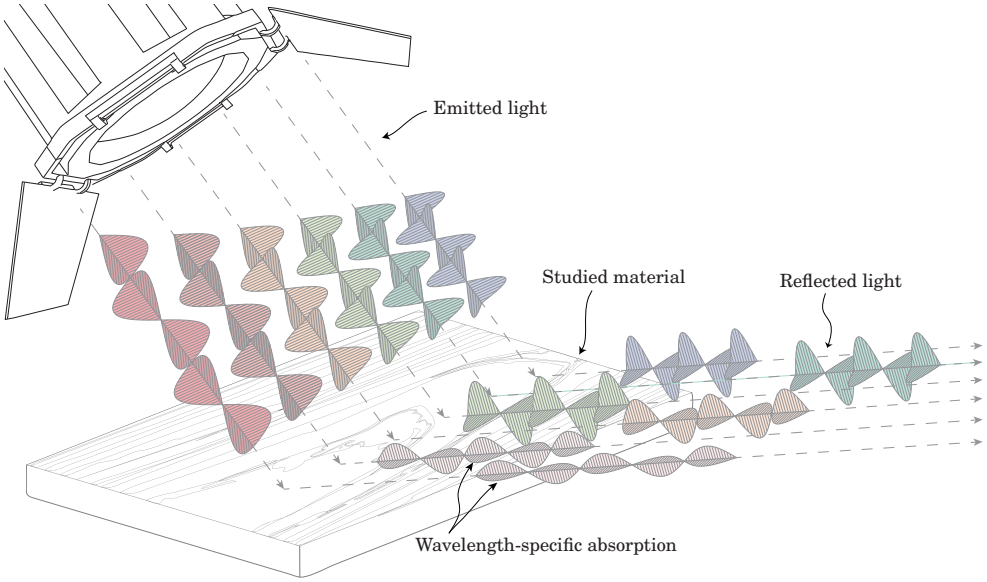


FIGURE 2.1 Illustration of fundamental concept used in this thesis to nondestructively study materials: light from an artificial source is emitted towards a material. The material reflects much of the light again but depending on the composition of the material certain frequencies of the light are greatly diminished in the reflected signal due to absorption. By measuring the reflected signal, and knowing the wavelength distribution of the emitted light, the absorbance of the material can be inferred and used to deduce properties about the studied material.

gases the molecule moves within the material [26]. The energy associated with these motions are notated as E_{vib}, E_{rot} and E_{trans} respectively in Eq. (2.2). For practical spectroscopic purposes, the translational energy of a molecule can be ignored. The remaining three energies (E_{elect}, E_{vib}, E_{rot}) however, all have specific energy levels associated with them which can cause photons from an incident source of radiation to be absorbed by the molecule. By exposing a material to a controlled source of radiation and measuring which photon energies, i.e. which wavelengths, the material absorbs, it becomes possible to deduce chemical information about the material. This concept is illustrated in Fig. 2.1.

Because photon absorption only occurs when the energy of an incident photon precisely matches the discrete energy amount associated with one of the molecule's energy levels, radiation shun on a material can intuitively be expected

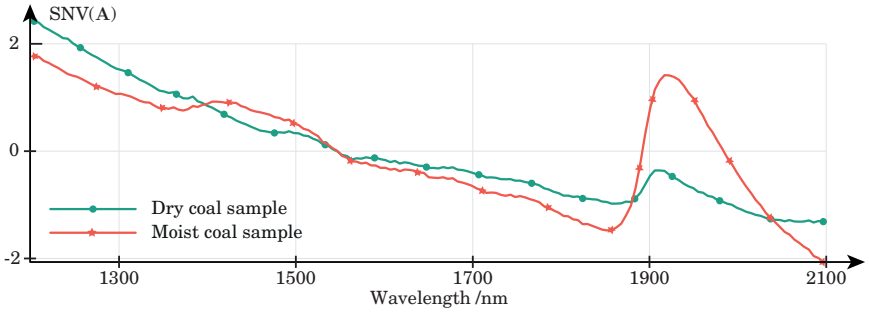


FIGURE 2.2 Measured absorbance spectrum of a coal sample in moist condition (red) and after drying for 6 hours (green). Displayed absorbance values are preprocessed using standard normal variate.

to be absorbed at specific, sharply defined, wavelength locations in an absorbance spectrum. When measuring the absorbance spectrum of a material in practice however, chemically induced absorption regions are rarely sharp, but rather smooth and distributed across a number of wavelengths. One reason for this is that materials are composed of many molecules, with varying vibrational and rotational states. These variations cause absorption to occur probabilistically across a number of wavelengths, with the peak of the absorption located around the energy level most probable to cause absorption [27]. Figure 2.2 shows an example of a measured absorbance spectrum depicting a piece of coal, measured in both wet and dry condition. As seen in the figure, the absorption peak around 1900-2000 nm which is caused by water in the sample, appears across a distribution of wavelengths.

How much the absorbance is altered by an absorbing species of a given concentration is mathematically described in Beer's law, which states that the absorbance $\mathbf{A}$ at a given wavelength λ follows the relation [28]:

$$\mathbf{A}(\lambda) = \varepsilon(\lambda) \times \ell \times c. \tag{2.3}$$

In Eq. (2.3) ε is the molar extinction coefficient, ℓ is the path length of the light, and c is the concentration of an absorbing species. The molar extinction

coefficient ε is an intrinsic material property. Provided that the path length ℓ is reasonably similar between different samples of a material, the absorbance $\mathbf{A}$ therefore has a very clear connection to the concentration c of an absorbing species. Simply put, higher concentration of an absorbing species causes higher absorbance. By experimentally measuring the absorbance of several samples of a material containing different concentrations of a chemical species one wishes to study, it therefore becomes possible to deduce and quantify the amount by which the absorbance should be weighted at different wavelengths to produce a good estimation of the concentration. It should be noted however that Beer's law is said to be valid only when a set of conditions are met, among which is that the attenuating medium must not scatter the radiation [29]. In practice, as will be discussed in the next section, this is an example of a condition that is difficult to fulfill. Some deviations from Beer's law are therefore virtually inevitable in practice. Despite this, highly useful approximations of an analyte's concentration are still perfectly possible to model. One way of numerically approximating the concentration of an analyte from the absorbance of a material, which is widely used in this thesis, is to assign a coefficient β (estimated from measurement data) to each measured wavelength which is multiplied with the absorbance of that wavelength:

$$c \approx \beta_1 \times \mathbf{A}(\lambda_1) + \beta_2 \times \mathbf{A}(\lambda_2) + ... + \beta_n \times \mathbf{A}(\lambda_n) \tag{2.4}$$

2.1.1 Why is it generally nontrivial to say something about a material by studying the way it interacts with light?

Absorption taking place due to the energy of a photon precisely matching any of the molecule's characteristic energy levels is known as *fundamental*

absorption. The frequencies at which this type of absorption occurs are known as the *fundamental frequencies* of the molecule [30]. However, in addition to fundamental absorption, photons possessing an energy which is a multiple of any of the energies of the fundamental frequencies of a molecule—for instance twice or three times the required energy amount—can also be absorbed by the molecule and cause elevation to higher energy levels. Such energy levels are known as *overtone frequencies*. The first multiple of a fundamental frequency is called the first overtone, the second multiple is the second overtone, etc. Because overtone absorption requires more energy (i.e. a higher frequency of light) than fundamental absorption, the overtone absorption will take place at shorter wavelengths in the electromagnetic spectrum relative to the fundamental absorption. Generally, fundamental absorption is much stronger in intensity compared to the overtones, and with each increasing overtone the magnitude of the absorption becomes lower [24]. Thus, the characteristic information of a molecule can be repeated multiple times across the wavelength range of an absorbance spectrum, similarly to an echoing sound becoming weaker and weaker with each echo. If the energy of a photon equals the sum of two or more vibrational energy levels within a molecule, it is also possible for a photon to be absorbed by the molecule, and in doing so triggering multiple vibrational excitations simultaneously [24]. Due to the multitude of ways in which photons can be absorbed by molecules, particularly molecules containing a large number of atoms, a single type of molecule can cause absorption at many different wavelengths. When studying materials composed of many different molecules, the absorption triggered by different molecules within the material may be located at overlapping wavelength regions, thus making it difficult to identify a wavelength region related only to the chemical substance of interest. Another layer of complexity is added to the analysis of the spectra by undesired interference in the signal.

2.1.1.1 Signal interference

Inadvertently, a measured spectrum is in practice likely to contain undesired interferences in the signal in addition to the absorbance patterns of interest. These interferences can have multiple sources of origin. For instance, if the studied material contains water, a change in temperature will cause the absorption peaks associated with the water to shift to slightly different wavelengths [31]. Thus, the exact same sample measured on two occasions may yield different spectra if the temperature of the sample has changed. In addition to changes in the sample, temperature changes in the instrument used to make the measurement can also induce interference in the signal.

Provided that measurements are taken in a controlled environment however, the main source of undesired signal interference is light scatter interference. Light scatter is when light changes trajectory and deviates from its originally straight path [32]. Micro-structure variations in the surface geometry of a material, presence of bubbles, droplets, fibers and density variations are examples of things which can scatter electromagnetic radiation [33]. The consequence of light scatter is that even thought the distance between a light source, a sample and a detector are held constant, the distance the light has to traverse before reaching the detector—the optical path length—can vary greatly when different samples are measured. Even when the measured samples are of a similar chemical composition [34]. Light scatter is particularly prevalent in near-infrared spectra where up to 99 % of the variation can be caused by light scatter in severe cases [31]. In the measured spectra this can manifest itself in several ways, it can for instance cause offsets and slope changes between the spectra of measured samples [32] or cause shifts in the position of peaks in the spectra [35]. Figure 2.3 illustrates the problematic nature of light scatter interference. The figure shows the measured absorbance at 19 different wavelengths in the

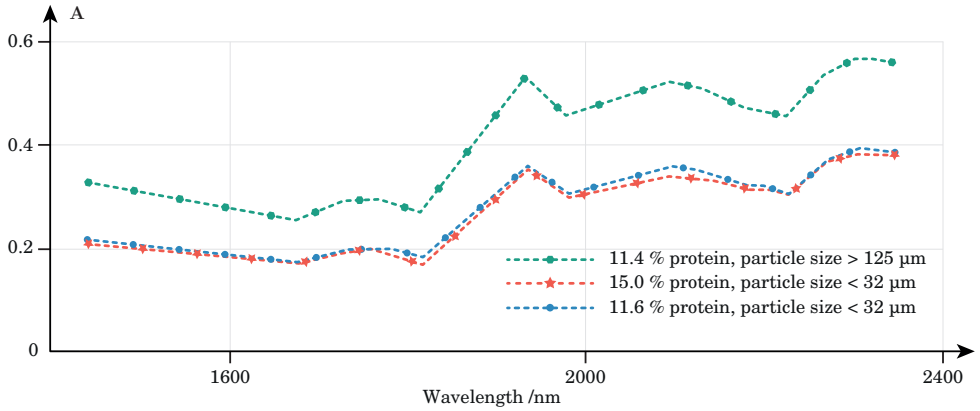


FIGURE 2.3 Measured absorbance spectra of three types of sieved wheat flours. Two of the wheat flours have a nearly identical chemical composition with a protein content of 11.4 and 11.6 % respectively, yet—due to differences in particle size—their spectra are substantially different. The third flour sample is chemically different from the others with 15 % protein content, yet its spectrum is very similar to the sample of similar particle size. Figure adapted from [31].

NIR region for three types of wheat flour. Two of the samples in the figure have a nearly identical protein content, whereas the third has a higher protein content. Despite the fact that two of the samples are nearly identical in chemical composition, there is a substantial difference in their measured absorbance spectra. The cause of this is that light scatters differently in the flours due to their different particle sizes, resulting in a seemingly uniform absorbance offset between the samples. Needless to say, this scatter-induced signal interference makes it problematic to directly model the relationship between the absorbance spectrum of a sample and a chemical constituent within the sample—since the scatter interference masks, and often overshadows, the spectral changes caused by chemical absorption. There exists however, as will be the subject of section 2.5, mathematical transformations which can be applied to the measured spectra to reduce the issues related to interference and to enhance spectral features of interest.

2.2 Hyperspectral imaging

2.2.1 Obtaining a hypercube

When an image is taken with a conventional digital camera, the data structure of the image consists of three separate two-dimensional matrices, where each matrix describes the spatial distribution of the intensity of red, green and blue color respectively as shown in Fig. 2.4 a). Hyperspectral line-scanning cameras on the other hand, such as the ones used to collect the datasets in this thesis, only take images consisting of a single spatial dimension, i.e. a line of pixels. In return for the reduced spatial detection capacity, each collected line contains substantially more colors than three, as shown in Fig. 2.4 b). When the number of color channels is large enough, each pixel can be viewed not merely as a composite of three discrete colors, but as a spectrum describing the measured light intensity as a function of wavelength. Furthermore, the wavelengths which the hyperspectral camera measure do not necessarily need to be confined to the visible part of the spectrum, but can be of frequencies either higher or lower than visible to humans.

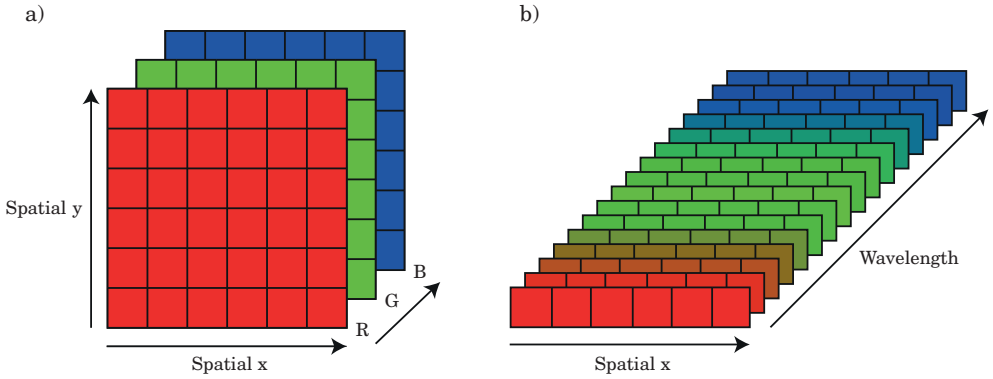


FIGURE 2.4 Illustration of a) data structure of an image taken with a conventional digital camera containing two spatial dimensions for each of the three color channels (red, green and blue); b) data structure of a single line image taken with a hyperspectral line-scanning camera.

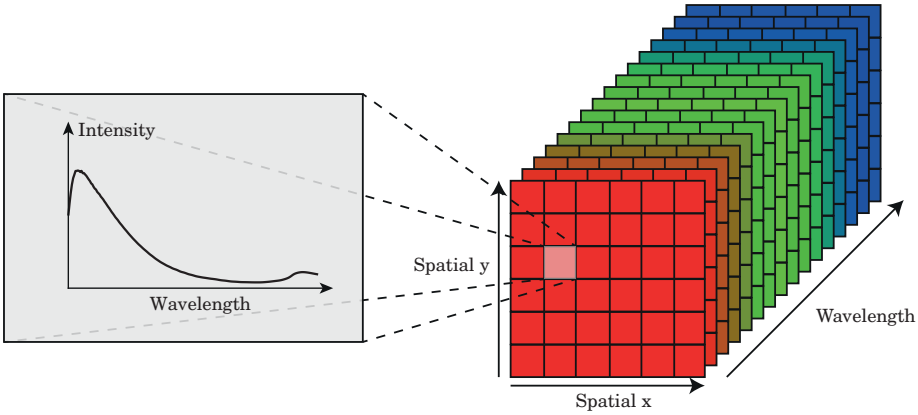


FIGURE 2.5 Illustration of the data structure of a hypercube. Multiple line images—such as the one shown in Fig. 2.4 b)—captured by a line-scanning camera are stacked on top of each other to form a three-dimensional image containing a spectrum of light intensity as a function of wavelength in every spatial position of the image.

To obtain images with a line-scanning camera containing two spatial dimensions, it is necessary to either drag the camera above the object one wishes to study, in a direction perpendicular to the original line orientation, while repeatedly taking pixel rows at each new location. Or to keep the camera stationary whilst moving the object in front of the camera whilst taking line images. The resulting sequence of images, where each image contains a single line of pixels, can then be stacked on top of each other to form a data structure consisting of two spatial dimensions. Each pixel of the resulting image will then contain a seemingly continuous spectrum of colors as shown in Fig. 2.5. Such three-dimensional images are commonly referred to as hyperspectral images or *hypercubes*. The linear motion needed to take such images is typically achieved by fastening either the camera or the object being studied to a *translation stage*, which is a motorized device for translating rotational movement from a motor into linear movement of the sample/camera.

2.2.2 Converting the values of a hypercube into reflectance

The numerical values of an obtained hypercube greatly depends on the illumination of scene being photographed, as well as camera settings such as integration time. As such, attempting to use the raw measured spectral signal within a hypercube to deduce anything about chemical or physical properties of the photographed object is greatly suboptimal; since images taken of the exact same object at different points in time can, and most likely will, result in different hyperspectral images due to changes in ambient light conditions etc. Furthermore, the raw measured spectra will appearance-wise often closely resemble the spectra emitted by the light source used to illuminate the object. Which for instance can be a halogen light or sunlight. Figure 2.6 shows the measured spectrum of a Norway Spruce (*Picea abies*) wood sample in comparison to the spectrum of a halogen light source which was used to illuminate the spruce sample. To make the measured spectral signal invariant to changes in illumination and to assign more easily interpreted and meaningful numerical values to the signal within the hyperspectral images, an additional object—a reference object—is often included in the same image together with the object being studied. The reflectance properties of the reference object should ideally be well known for the entire wavelength region detected by the camera. The purpose of including a reference object with a known reflectance in the image is that it can be used to infer what the reflectance of any other object in the image is.

In all of the image acquisitions performed in the projects included in this thesis, a reference rod made of *spectralon* was included in the images. Spectralon is a material which is often used as a reference material due to its exceptionally high and uniform diffuse reflectance across a broad wavelength range. For all visible light Spectralon's reflectance is typically over 99 %, and it also has excellent

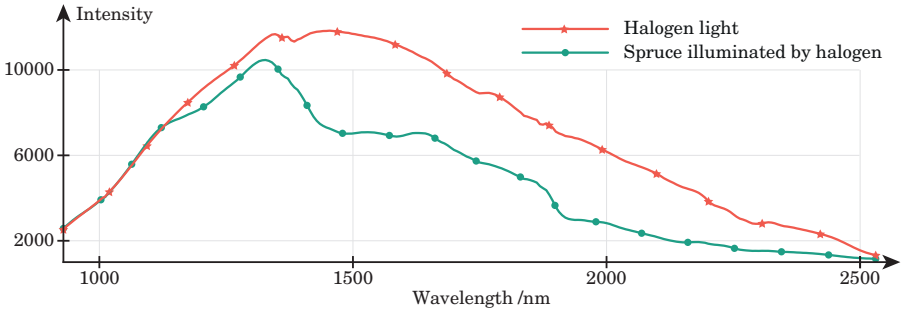


FIGURE 2.6 Spectral signal intensity measured of a halogen light source (red line) and a piece of spruce (green line) which is being illuminated by the halogen light source shown in red.

reflectance in the near-infrared region up to 2500 nm [36]. By assuming that the signal intensity of pixels depicting the spectralon rod correspond to 100 % reflectance it becomes possible to approximate what the reflectance of every other pixel within a hypercube corresponds to—simply by dividing every pixel of the image with the intensity value measured on the spectralon. By doing so the values within the hypercube are assigned an easily interpreted physical property, namely reflectance. After this transformation, images taken of an object at different points in time under different light conditions will be, within some margin of error, identical to each other. Since variations in illumination will also affect the reference material. This technique—dividing the value of pixels with the value of a reference pixel to obtain the reflectance—would have been adequate if the camera would have offered perfect fidelity in its photon count. In practice however, one also has to account for a phenomena known as *dark current*. Just like conventional cameras, hyperspectral cameras use CCD sensors (or CMOS/MTC) to detect incoming light. The purpose of such sensors is to translate impinging photons into electric charges, which can then be quantified and used to measure the amount of light reaching the camera. Dark current is the name of thermally generated electric charges which are continuously misinterpreted by such detectors as impinging photons [37]. The charge generated by dark current—which has multiple known sources [38]—is indistinguishable from the

charge generated by actual photons reaching the sensor. Thus, the practical consequence of dark current is that the signal obtained using a CCD/CMOS/MTC sensor is a superposition of desirable photon-induced signal and undesirable dark current-induced signal. Therefore, all hyperspectral images, even ones taken in complete absence of impinging light, will still contain a measured signal intensity greater than 0. To quantify the amplitude of the noise caused by dark current, each hyperspectral image acquisition typically ends with taking a number of line images with the shutter of the hyperspectral camera closed. The resulting signal from these pixels then represents the signal measured by the sensor despite the absence of impinging photons, i.e. an approximation of the dark current signal. The amount of dark current noise received by a CCD depends on the temperature of the detector—the higher the temperature the higher the dark current noise. To minimize the effect of dark current, hyperspectral cameras are therefore actively cooled. When the dark current signal has been approximated, it can be removed from the hyperspectral image by subtracting the measured dark current signal from every other pixel of the image. Thus, the final conversion from raw measured signal intensity into reflectance can be calculated as:

$$\mathbf{R} = \frac{\boldsymbol{I} - \boldsymbol{I}_{\mathrm{d}}}{\boldsymbol{I}_0 - \boldsymbol{I}_{\mathrm{d}}}. \tag{2.5}$$

Where $\boldsymbol{I}_0$ represents the signal measured on the reference object with roughly 100 % reflectance, $\boldsymbol{I}_{\mathrm{d}}$ is the signal measured with the camera shutter closed and $\boldsymbol{I}$ is the signal of all other pixel of the image. Because an image will likely contain spatial variations in sample illumination, dark current and CCD sensor sensitivity along each line of pixels, the correction in Eq. (2.5) is performed separately for each position along the acquired lines, i.e. column by column in the hypercube. It should also be mentioned that if the object being scanned is stationary together with a light source and the camera is moving above the sample during the scan, there will also likely be illumination variations in the

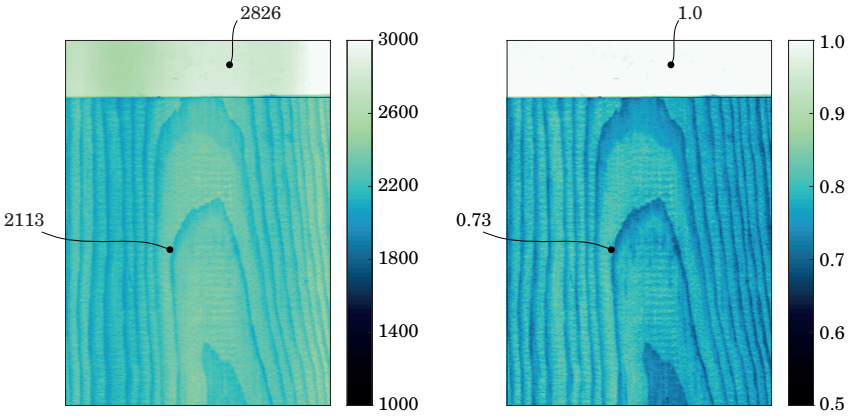


FIGURE 2.7 Slice of a hyperspectral image depicting a pine sample at 761 nm before and after white reference correction. Left image shows the raw measured pixel intensity data. Right image shows the same data but after it has been passed through Eq. (2.5) and converted into reflectance relative to the white reference object located at the top of the image. Signal captured with the shutter closed (I_d) has been omitted from the illustration but is typically located at the bottom of the image.

vertical (row) direction of an image, unless the sample is illuminated in a perfectly uniform way. If this is the case Eq. (2.5) becomes inaccurate. It is therefore generally preferable to keep the camera stationary together with the light source and move the sample in front of the camera during the acquisition. In this way each scanned line experiences the same illumination, regardless of how long the sample being scanned is. Figure 2.7 shows a hyperspectral image of a pine sample before correction with Eq. (2.5) (left) with integer values, and after the correction (right) where the data has been normalized such that values within the white reference object correspond to 1.0 (100 % reflectance) and all values within the sample lie somewhere between 0 and 1.

2.2.3 Converting a hypercube into absorbance

After the measured hyperspectral data has been passed through Eq. (2.5), it describes a property inherent to the object being monitored, the reflectance, and can thus be used to study different aspects of the object. However, when the ambition is to use the spectral data in a statistical model to correlate it

to a chemical concentration of some sort, absorbance, rather than reflectance, is often preferred. As previously mentioned in section 2.1, the reason for this is that provided Beer's law is upheld, there is a linear relationship between the absorbance at a specific wavelength and the concentration of the absorbing species [24]. Beer's law states that the absorbance, $\mathbf{A}$, can be calculated as the logarithm of the inverted transmittance, $\mathbf{T}$:

$$\mathbf{A} = \log_{10}(1/\mathbf{T}). \tag{2.6}$$

In hyperspectral imaging, radiation reflected off of objects is often measured rather than radiation transmitted through the object. In such cases it is common to simply treat the reflectance as an analog to transmittance, thereby making the equation:

$$\mathbf{A} = \log_{10}(1/\mathbf{R}). \tag{2.7}$$

Although Eq. (2.7) is used extensively in spectroscopy it should be mentioned that the expression may not be entirely correct, as no functioning theory has ever been developed proving that it is valid to substitute the transmittance $\mathbf{T}$ with reflectance $\mathbf{R}$ when using Beer's law to obtain absorbance [39]. As such, the resulting values of Eq. (2.7) are by some authors not referred to as absorbance values but rather as *apparent absorbance* or simply as "$\log 1/\mathbf{R}$ values". The $\log_{10}(1/\mathbf{R})$ values are however related to the absorbance and it has been shown that there is an *almost linear* relation between the concentration of an absorbing compound and its contribution to the $\log_{10}(1/\mathbf{R})$ value [40], which nevertheless makes it practically useful to use Eq. (2.7). Henceforth in this thesis, any spectral data that has gone through the $\log_{10}(1/\mathbf{R})$ transformation will for convenience just be referred to as absorbance data ($\mathbf{A}$) with an understanding that the values are produced with Eq. (2.7) and may not correspond exactly to the true

absorbance.

2.2.4 Unfolding a hypercube

A final step before using the hyperspectral images as input to a statistical model and correlating the data to a response is to rearrange the structure of the data. Most statistical modeling techniques are designed to operate on datasets structured into two-dimensional matrices with observations along the rows of the matrix and variables along the columns. In the case of hyperspectral images, each pixel is an observation and each measured wavelength is a variable. Before developing models with the hyperspectral data, the data therefore needs to conform to this structure. This is achieved by spatially collapsing each hypercube and stacking the columns on top of each other as shown in Fig. 2.8, thereby combining the original two-dimensional spatial information of the images into one dimension. This procedure is known as *unfolding* a hypercube. When hyperspectral datasets consist of multiple hypercubes, each cube is unfolded and stacked vertically on top of the others to form a tall two-dimensional matrix containing the entire dataset. Such a data matrix is often denoted $\boldsymbol{X}$. The response value to be modeled from the spectral data, for instance the chemical content associated with each hyperspectral image, is rearranged into a vector (or matrix in the case of multiple responses), $\boldsymbol{y}$, of the same length as $\boldsymbol{X}$—such that each row of $\boldsymbol{X}$ maps to a response in $\boldsymbol{y}$.

2.3 Regression modeling

In order to predict something useful from the spectra contained within a hyperspectral image, such as the spatial distribution of a particular chemical, a mathematical transformation is needed. This transformation is generally realized with regression modeling. The aim of regression modeling is to find a

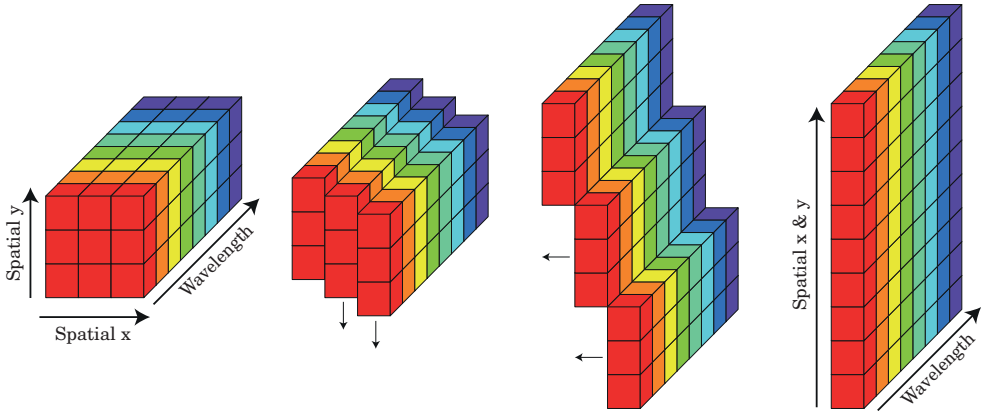


FIGURE 2.8 Unfolding of a three-dimensional hypercube into a two-dimensional matrix of structure observation $\times$ variables suitable for regression modeling.

function f, generally parameterized by some coefficients $\boldsymbol{\beta}$, which can translate independent data $\boldsymbol{X}$ to one or multiple response variables $\boldsymbol{y}$:

$$\boldsymbol{y} = f(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{\epsilon}. \tag{2.8}$$

Preferably with a minimal translation error $\boldsymbol{\epsilon}$. The motivation for seeking such a function rather than measuring $\boldsymbol{y}$ directly is commonly that $\boldsymbol{X}$ is easier and/or cheaper to obtain than $\boldsymbol{y}$ or that $\boldsymbol{y}$ is simply not accessible to measure. In *linear regression* the response variable $\boldsymbol{y}$ is assumed to be a linear function of the parameters in the model [41]. The simplest and probably most well known example of a linear function is that of a straight line: $y = mx + b$. When using the equation of a line as a basis for a regression model describing the linear relationship between the model's input and output, all parameters whose values need to be estimated—the intercept b and slope m in the case of a straight line—are often all denoted with the symbol β; with different subscripts distinguishing the different parameters. The regression modeling analogue of $y = mx + b$ is therefore often expressed as:

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{2.9}$$

Where the error term ϵ accounts for deviations from the linear relation. In cases where the value of y is suspected to be influenced by more than one independent variable, as is often the case in spectroscopy, Eq. (2.9) can easily be extended to accommodate such additional predictor variables by adding more terms to the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon. \tag{2.10}$$

Equation (2.10) is still linear, but as the number of terms in the equation increase it becomes increasingly difficult to picture what the modeled geometrical connection between input and output looks like. When only one independent variable is involved, the model uses a straight line to estimate y from $\boldsymbol{x}$; when two independent variables are involved, the model uses a plane to estimate y from $\boldsymbol{x}$; when three or more independent variables are used the model describes a *hyperplane*, i.e. a plane with higher dimensionality than two. A traditional and well-established way of estimating the unknown regression coefficients in Eq. (2.10) is by using ordinary least squares (OLS) fitting. The OLS solution to Eq. (2.10) can for instance easily be obtained by [41]:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.11}$$

Where $\hat{\boldsymbol{\beta}}$ is a vector containing all the estimated coefficients. The OLS solution works well when the regression problem being solved only contains a few independent variables [42]. Additionally, these few independent variables need to be uncorrelated to each other. If they are not, the regression coefficient estimation becomes unstable [43]. Throughout the papers in this thesis, the independent $\boldsymbol{X}$ data consists of hyperspectral data where the $\boldsymbol{X}_1, \boldsymbol{X}_2, ...\boldsymbol{X}_n$ variables represents the absorbance at individual wavelengths in close proximity to each other. Under such circumstances neither of the mentioned prerequisites

for robust and successful OLS estimation are fulfilled—as there are a lot of variables in spectral data, many of which are heavily correlated. To successfully estimate the coefficients of Eq. (2.10) when the variables in $\boldsymbol{X}$ are large in number and heavily correlated, an alternative fitting procedure to OLS is therefore necessary.

One way of estimating the regression coefficients of Eq. (2.10) which is robust against noisy, numerous and correlated $\boldsymbol{X}$ variables [42] is *partial least squares regression* (PLS). PLS is extensively used in areas related to spectral modeling and is used throughout this thesis (**Papers I**, **II**, **III** and **IV**) for estimating the concentration of various analytes from absorbance data. The reason for PLS's robustness is that the method has a dimensionality reduction built into to the parameter estimation. If each of the n variables in $\boldsymbol{X}$ ($\boldsymbol{X}_1, \boldsymbol{X}_2, ...\boldsymbol{X}_n$), i.e. each wavelength of the spectral data, is thought of as defining its own coordinate axis in an n-dimensional space, PLS defines a new A-dimensional subspace (where $A \leq n$). The original $\boldsymbol{X}$ data is then projected down to this A-dimensional space forming a lower dimensional representation of the data which is used to estimate the coefficients needed to model $\boldsymbol{y}$. This dimensionality reduction concept can be seen exemplified in Fig. 2.9 using three wavelengths from a set of NIR spectra. Unlike other commonly used dimensionality reduction methods such as *principal component analysis* (PCA) which only considers the $\boldsymbol{X}$ data when forming the lower dimensional subspace, PLS takes both the $\boldsymbol{X}$ and $\boldsymbol{y}$ data into consideration when orienting the A-dimensional subspace and in the orientation process attempts to maximize the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ [44]. This ensures that information relevant for the modeling of $\boldsymbol{y}$ from $\boldsymbol{X}$ is caught in the first few subspace dimensions. If most of the variables in $\boldsymbol{X}$ are uncorrelated to each other and essential for the prediction of $\boldsymbol{y}$, A will need to be high in order to not lose any critical information in the dimensionality reduction. If many of the variables in $\boldsymbol{X}$ are highly correlated to each other or irrelevant for the prediction

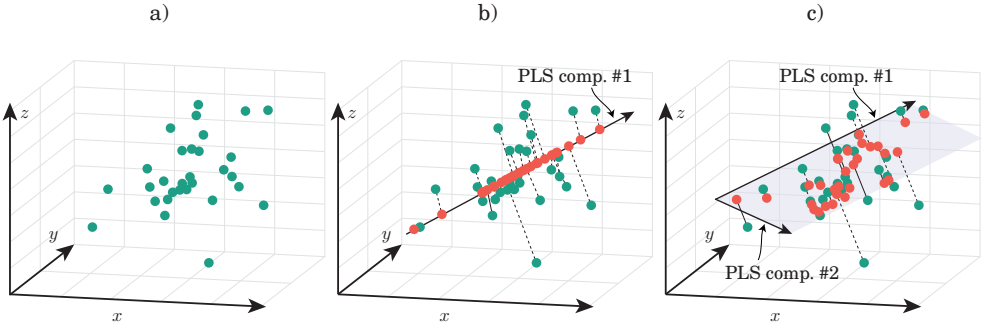


FIGURE 2.9 Illustration of dimensionality reduction performed during PLS on a dataset containing three variables ($n = 3$). Scatter plot a) shows original data as green dots. Each independent variable (x, y, z) defines a separate axis. Scatter plot b) shows the original data points (green) together with orange points projected down on a line defined by the first component/dimension identified by PLS ($A = 1$). Scatter plot c) shows the original data points (green) together with orange points projected down to a two-dimensional plane ($A = 2$) defined by the first two components identified by PLS. PLS components were estimated using the SIMPLS algorithm [45].

of $\boldsymbol{y}$, the relation between $\boldsymbol{X}$ and $\boldsymbol{y}$ can be adequately captured even when A is substantially lower than n.

To establish which value of A is optimal for a dataset, a technique called *k-fold cross-validation* is often used. In k-fold cross-validation the observations of a dataset (both $\boldsymbol{X}$ and $\boldsymbol{y}$) are partitioned into k segments called folds. A regression model is then trained using $k - 1$ of the folds; the last k^{th} fold is withheld from the model. Afterwards, the model is given the withheld $\boldsymbol{X}$ data as input and attempts to predict the withheld response values associated with the fold. The mean squared error between the models prediction of the response and the actual response is then calculated and stored. The process is then repeated k times with a different fold withheld from the model during its training until all folds have been withheld from the model once. The resulting k mean squared errors obtained from this are then used to form a *cross-validated Root Mean Squared Error* (RMSE_{cv}) which is an estimate of how well the model is expected to perform on average on new, previously unseen, data. Figure 2.10 illustrates the concept

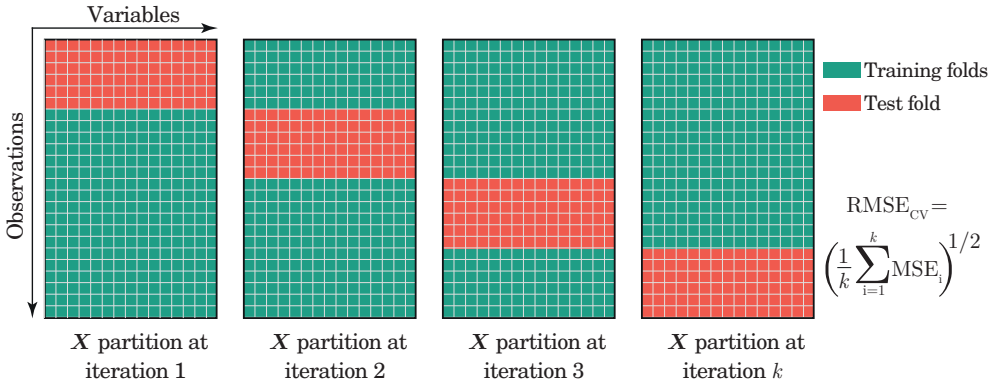


FIGURE 2.10 Illustration of concept behind k-fold cross-validation. k different models are fitted iteratively, in each iteration a different partition of the data is used to fit/train the model and to test its performance. The final $\mathrm{RMSE}_{\mathrm{cv}}$ value is afterwards calculated using the mean squared error of all iterations to obtain an estimate of the model's average performance when confronted with new data.

behind k-fold cross-validation and the equation used to determine the $\mathrm{RMSE}_{\mathrm{cv}}$. In order to use k-fold cross-validation to determine the optimal subspace dimension A in a PLS model, the cross-validation is performed with different values of A. The A value found to produce the lowest $\mathrm{RMSE}_{\mathrm{cv}}$ is then used in the final model. Alternatively, if the $\mathrm{RMSE}_{\mathrm{cv}}$ continues to decrease as A increases, but does so insignificantly, the A value can be set manually to favour a low complexity model over high complexity models with nearly identical performance as this may be advantageous in terms of robustness.

2.4 Variable selection

Hyperspectral data contains absorbance measurements at typically a few hundred different wavelengths, which are often equally spaced apart. When attempting to correlate the spectral information of such data to a response—such as the concentration of a particular chemical in the surface of a sample—it is generally the case that only a subset of the measured wavelengths is relevant for mapping the signal to the response. Inadvertently, the collected spectra are

thus likely to contain absorbance measurements at numerous wavelengths which are irrelevant when it comes to modeling the dependent variable. Even thought the dimensionality reduction of PLS will alleviate some of the issues related to irrelevant variables, including such redundant wavelengths in the regression model can render it less accurate and robust than it would have been had these regions been excluded, even when using PLS [46]. Wavelength selection, which is more generally referred to as variable selection, feature selection or subset selection, is the process by which the relevant wavelength regions of a dataset are separated from the irrelevant or noisy regions. More concretely, if an independent dataset is stored in a two-dimensional matrix $\boldsymbol{X}$ with observations along the rows and variables along the columns, the aim of variable selection is to identify which columns of $\boldsymbol{X}$ should be included in the regression model and which ought to be excluded.

What makes variable selection challenging is that the problem has no known analytical solution which can be used to quickly identify the optimal variable subset. To find the single best performing variable subset the only option is therefore to exhaustively evaluate every possible combination of variables and compare how they perform. However, the total number of possible variable combinations is $(2^n - 1)$ where n is the number of available variables to choose from, i.e. the number of columns in $\boldsymbol{X}$. As soon as the number of variables in the dataset is more than just a handful, a rapid combinatorial explosion takes place which makes the number of possible subsets astronomically large. If the number of wavelengths in a measured spectrum is 256 for instance, the number of possible subsets is more than 10^{77}, which is well beyond what is computationally feasible to evaluate. Instead of evaluating every possible subset, variable selection techniques are often heuristic and strives to identify subsets which are as good as practically possible, rather than globally optimal.

Variable selection techniques are often categorized as either filter-, wrapper-

or embedded methods depending on how they attempt to identify and quantify the value of variables in the data [47]. The category of techniques exclusively used throughout the papers of this thesis are the wrapper methods, as such, the concepts behind filter and embedded methods will not be expanded upon here. Wrapper methods work by generating candidate subsets which are then evaluated by fitting regression models using the generated variable subsets and predicting the response of a dataset. The resulting prediction error associated with each variable selection, which for instance can be a $RMSE_{cv}$ value, is then fed back into the wrapper algorithm which is guided by the provided performance measure when generating the next set of candidate subsets. Thus, as the name suggests, wrapper methods can be thought of as wrapping around a regression model in order to create a feedback loop which drives the search for new subsets. The iterative concept behind wrapper methods is illustrated in Fig. 2.11. One of the main criticisms of wrapper methods is that they are generally very slow [48]. This is because wrapper methods are iterative and driven by trial and error. As such, they tend to require a large number of subsets to be evaluated before a satisfying subset is found, which is significantly computationally costly and thus time consuming when a large number of variables are involved [49].

> The need for novel, computationally efficient, techniques that can speed up the feature selection process is apparent, particularly in fields concerned with large datasets such as hyperspectral data. Developing such techniques will be the research topic of **Paper I** and **Paper II**.

In the next two sections the concept behind two variable selection methods which are commonly used in spectroscopy are outlined, both of which are used in the papers of the appendix.

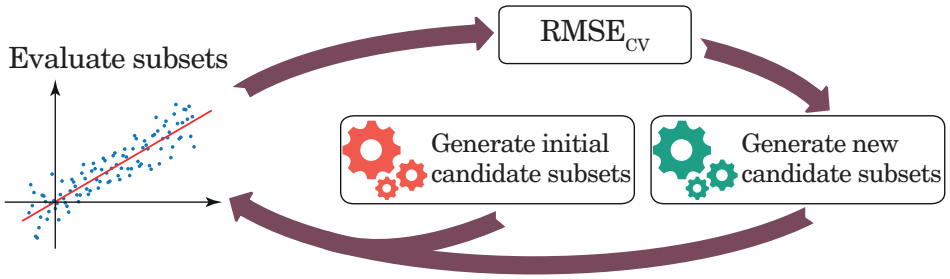


FIGURE 2.11 High-level illustration of cyclical concept behind wrapper-based feature selection methods. An initial pool of candidate variable subsets is generated and cross-validated, the $RMSE_{cv}$ of the subsets is then used to decide which subsets should be generated and evaluated next. How new feature subsets are produced varies greatly between different feature selection algorithms.

2.4.1 Step-wise methods

Among the simplest forms of variable selection techniques are the step-wise methods. In these methods, a single variable is either added or removed from a candidate variable selection in an attempt to improve the solution. The most basic forms of step-wise methods are *forward selection* and *backwards elimination*. In forwards selection an initial subset is generated containing no active variables. Each of the inactive variables are then one by one activated and a regression model is trained and evaluated using the newly activated variable. Whichever variable yields the lowest prediction error from this becomes permanently active and becomes the start-case for the next cycle of the selection procedure. In the next cycle, each of the reminding inactive variables are one by one activated and evaluated, the best performing of which again becomes permanently activated. This pattern continues until no activation can be made without increasing the prediction error of the model. In backwards elimination, the concept is similar but executed in reverse: an initial subset is generated containing all the possible variables in the dataset. From the initial subset each single variable is inactivated one by one and evaluated. The best performing inactivation becomes permanent. Figure 2.12 summarizes the concept behind both forward selection

- Initialize variable selection vector v with all elements inactive (set to 0)

while Model error decreases

  for Each inactive variable in v

    - Temporarily activate variable in v and solve the regression problem using the column subset from X indicated by v. Store RMSE_{cv}

  end

  - Permanently activate variable in v which gave lowest RMSE_{cv}

end

- Initialize variable selection vector v with all elements active (set to 1)

while Model error decreases

  for Each active variable in v

    - Temporarily inactivate variable in v and solve the regression problem using the column subset from X indicated by v. Store RMSE_{cv}

  end

  - Permanently inactivate variable in v which gave lowest RMSE_{cv}

end

FIGURE 2.12 Pseudocode of variable selection with forward selection (left) and backwards elimination (right). Conceptually the two methods are very similar, the difference being that one seeks to add variables to a model whereas the other seeks to remove variables from a model. Differences between the two algorithms are highlighted in green.

and backwards elimination in pseudocode.

Step-wise methods are deterministic in their outcome and easy to implement. However, one of the drawbacks of the step-wise methods' simplicity is that they terminate the search as soon as no single variable can individually improve upon the current solution. It may well still be that the activation/inactivation of two or more variables in a solution would have improved upon it when a step-wise method terminates its search. The successfulness of step-wise methods can therefore vary greatly between datasets.

2.4.2 Genetic algorithms

Genetic algorithms (GA) belong to a category of algorithms known as *evolutionary algorithms* (EA). As the name suggests, these algorithms attempt to solve optimization tasks by using mechanisms inspired by concepts observed in natural evolution. A genetic algorithm starts off by generating a group of different randomly created solutions to an optimization problem. When using GA to perform variable selection, randomly initialized solutions refers to subset vectors which have had their elements randomly assigned either zero (inactive variable) or one (active variable). In GA lingo the group of generated solutions is known as the *population* of the algorithm, each solution of the population is in

turn called an *individual* or a *chromosome*.

After the initialization, the next step of the algorithm is to evaluate the performance of each available solution, which is done by cross-validating a separate regression model using each of the individuals of the population. Since the solutions are randomly generated without any consideration for the specific dataset at hand, they are all likely to perform poorly. Inevitably, some of the solutions will however still perform slightly better than the others purely by chance. The goal of the GA is to suppress the type of solutions which give rise to models with large prediction errors and to favor and refine solutions which give rise to low prediction errors. This is accomplished by the use of a set of genetic operators: *parent selection*, *crossover*, *mutation* and *survivor selection*; all of which are inspired by processes taking place in natural evolution. Parent selection consists is selecting a subset of the individuals from the population which are in the algorithmic steps to come going to be used to generate new solutions. This is commonly done probabilistically such that the performance of each solution is converted into a probability of being selected—individuals resulting in lower model errors are assigned higher probability of being selected, and vice versa. From the probability distribution a number of solutions are then drawn, a majority of which are likely to be among the top performing variable selections in the population. The selected solutions are referred to as *parents*.

The parents are then used to create new solutions via the crossover operator. During crossover, parents are pair-wise selected and combined to form *children*. A common way of achieving this is to randomly pick two locations along the parents and then let the resulting children alternately inherit parts of their two parents. This concept can be seen illustrated in Fig 2.13. The new solutions—the children—generated by crossover are then added to the original population. To introduce additional exploration into the algorithm and to enable solutions to be formed which are not represented in the current population, the entire

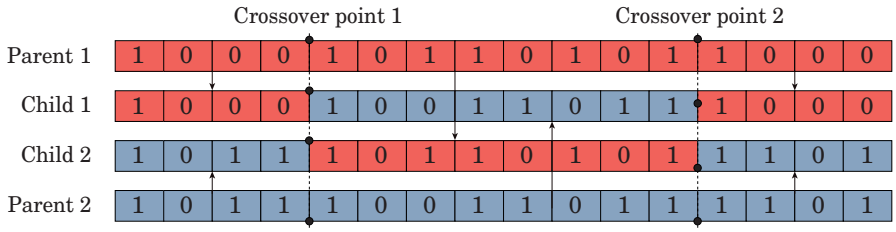


FIGURE 2.13 Illustration of a crossover operator (2-point crossover): two points are randomly picked along two parents, two new child solutions are then formed from the parents by alternately copying segments in between the crossover points from the parents into the children.

population then undergoes mutation. During mutation, each individual has a small chance to have one or more of its elements flipped—such that a value of one becomes a value of zero and vice versa. After the crossover and mutation stages, the population is once again evaluated by using each variable selection to generate a regression model which is cross-validated to get a measure of its performance. At this stage of the algorithm, the population has grown to consist of both the original population and the new crossover-generated children. The final step of the GA is therefore to discard a number of solutions such that the population returns to its original size. This is realized with survivor selection, which in essence can be performed using the same method of selection as during parent selection.

After the survivor selection, one generational cycle is complete and the algorithm starts over again. The next iteration however, the algorithm does not randomly initialize its population but instead starts with the population from the previous generation. The population then goes through parent selection, crossover, mutation, survivor selection again and again until a specified number of iterations (generations) has been reached. This concept is summarized in pseudocode in Fig. 2.14. As the iterations go by, the population experiences an evolution where fit individuals are favored, survives over time and gets to spread their genes (vector elements) throughout the population via crossover. Poorly

- Initialize population of variable selections randomly
- for Specified number of generations
  - Select parents
  - Create new solutions with crossover from parents
  - Mutate population
  - Evaluate performance of all solutions
  - Select survivors for the next generation
- end

FIGURE 2.14 Pseudocode of concept behind variable selection with a genetic algorithm.

performing solutions are less likely to be picked as parents and are eventually killed off by the algorithm during the survivor selection.

Compared to for instance step-wise methods, genetic algorithms are arguably substantially more complicated. However, they do not suffer from some of the drawbacks of step-wise methods such as early termination upon encountering a local optima. Genetic algorithms will continue to evolve new solutions until the user chooses to terminate, which may result in better variable selections. A disadvantage of genetic algorithms is that they can be immensely computationally costly, especially if a large population size is used. Whereas a step-wise method can converge within a few hundred evaluations, genetic algorithms have no upper limit on the number of variable selection evaluations needed, since this depends entirely on the user-specified number of generations before termination and the size of the population used. The concepts introduced in **Paper I** and **Paper II** does however considerably lessen some of the computational burden associated with evaluating a large number of variable subsets.

2.5 Preprocessing

The initial transformation from raw data to absorbance outlined in section 2.2.2-2.2.3 is an attempt to make the spectral data respond linearly to changes in chemical concentration in the sample [50]. As mentioned in section 2.1.1,

NIR spectra are often heavily influenced by undesired interferences such as light scattering. When developing regression models this can cause significant problems, since the linear relationship between the measured absorbance and the concentration of the sought analyte becomes spoiled [34]. Additional mathematical transformations may therefore be applied to the spectra in an attempt to correct for undesired phenomena in the signal and/or to enhance the spectra in some way in order to either improve the regression performance or to increase the interpretability of the data. These mathematical transformation procedures are referred to as *preprocessing* methods. This section describes some of the main methods used to preprocess spectra, all of which are used at some point in the papers included in this thesis. The effect each of the described preprocessing methods has on a spectral dataset is shown in Fig. 2.15.

2.5.1 Multiplicative Scatter Correction (MSC)

The goal of *Multiplicative Scatter Correction* (MSC) is to isolate scatter-related variations in a spectral dataset from chemical-related absorption variations and then ”correct” the spectra such that the scatter variations in all the spectra are as similar to each other as possible [51]. Ideally one is after the correction then left with a set of spectra where the variations caused by chemical absorption are intact, whilst variations in the spectra caused by physical interferences such as light scattering are suppressed. The fundamental concept which MSC relies on is that the wavelength dependency of scattered light is often different from that of the wavelength dependency of chemically-induced light absorption [31, 52].

MSC works by comparing a set of spectra to a reference spectrum and then attempting to correct the set of spectra such that they have the same scatter level as the reference. Ideally, the reference spectrum used in MSC should be a pure reflectance/absorbance spectrum of the sample being studied, without any scatter interference present; since this would cause the corrected spectra to also

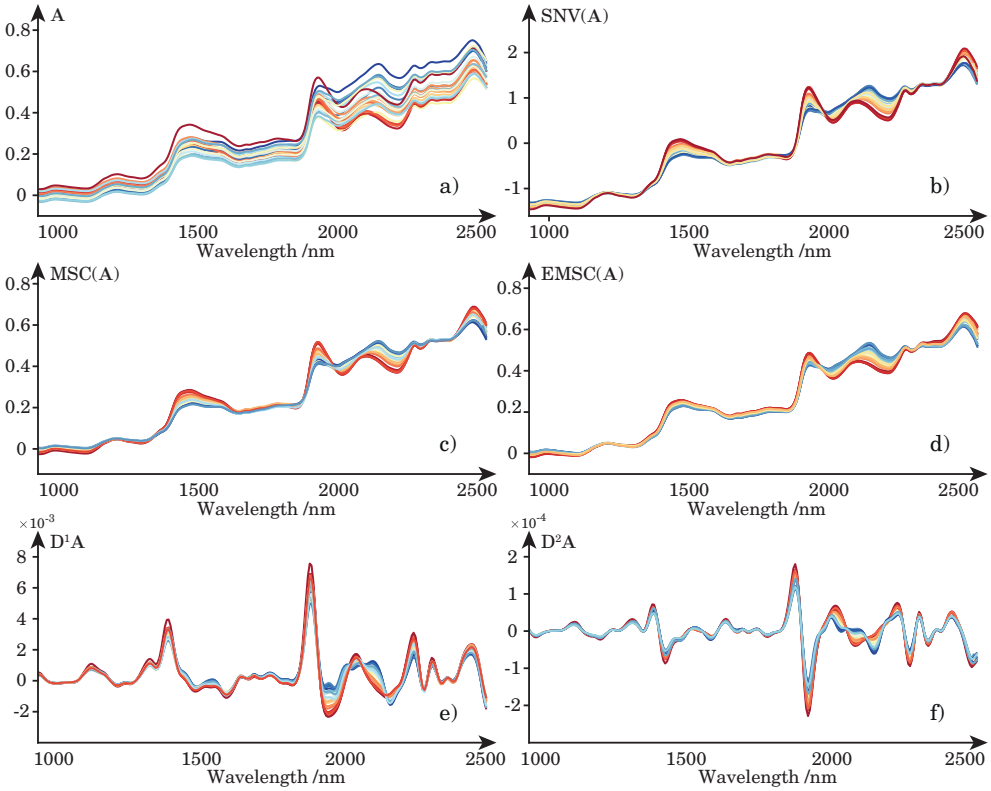


FIGURE 2.15 Average absorbance spectra of each sample in dataset I (spruce treated with a varying concentration of flame retardant) preprocessed with different methods: a) absorbance spectra without preprocessing; b) Standard Normal Variate; c) Multiplicative scatter correction; d) Extended Multiplicative Scatter Correction; e) First order savitzky-golay derivative; f) Second order savitzky-golay derivative. All spectra are color-coded on a scale from red to blue where redder color indicates a higher concentration of phosphorus in the surface of the wood and vice versa.

be free of scatter effects. Such an ideal reference spectrum is however typically not available. Instead a common practice is to use the average spectrum of a collected dataset as a reference. When the correction is performed all the spectra in the dataset will then be adjusted to contain a similar scatter profile as the reference, which is often adequate to improve the regression performance.

Figure 2.16 illustrates how MSC isolates scatter-related variations from chemical absorbance variations. Subplot a) of Fig. 2.16 shows the measured reflectance spectra of two pieces of spruce from dataset I. Both pieces are of the same species

and cut from the same board, the main difference between the samples is that one has been treated with a high concentration of a flame retardant chemical whereas the other has been treated with a low concentration. As can be seen in the figure, there is a large vertical offset between the two spectra across large parts of the wavelength range. The dotted black line in between the two spectra in Fig. 2.16 a) is the average spectrum of dataset I, which is here used as a reference spectrum. Figure 2.16 b) shows the same spectra, but this time plotted versus the reference spectrum of the dataset on the abscissa (x-axis), rather than against wavelengths. By visualizing the spectra in this way it becomes clear that in this domain each of the two individual spectra falls roughly along a straight line. The assumption in MSC is that chemical light absorption is the cause of the smaller deviations around the straight lines [52] shown in Fig. 2.16 b), whereas the rotational offset from the reference spectrum is caused by physical effects such as light scatter. By fitting linear regression lines through each spectrum in Fig. 2.16 b), an intercept β_0 and a slope β_1 can be estimated which describes the additive and multiplicative deviations from the reference spectrum. With these two parameters estimated the spectral data can be vertically shifted to have an intercept of 0 and rotated to the same slope as the reference spectrum. Since this transformation is done in an identical way for all the measured points within a spectrum, the smaller deviations around the regression lines seen in Fig. 2.16 b) will still be preserved after the transformation. Figure 2.16 c) shows the same spectra after the MSC correction. The intercept and slope required for MSC are typically estimated using ordinary least squares fitting, for example using Eq. (2.11). Once the β_0 and β_1 parameters are estimated the MSC correction of a spectrum $\boldsymbol{s}$ is performed using:

$$\boldsymbol{s}_{\text{MSC}} = \frac{\boldsymbol{s} - \beta_0}{\beta_1}. \tag{2.12}$$

Where $\boldsymbol{s}_{\text{MSC}}$ is the corrected spectrum, β_0 is the estimated intercept and β_1 is the

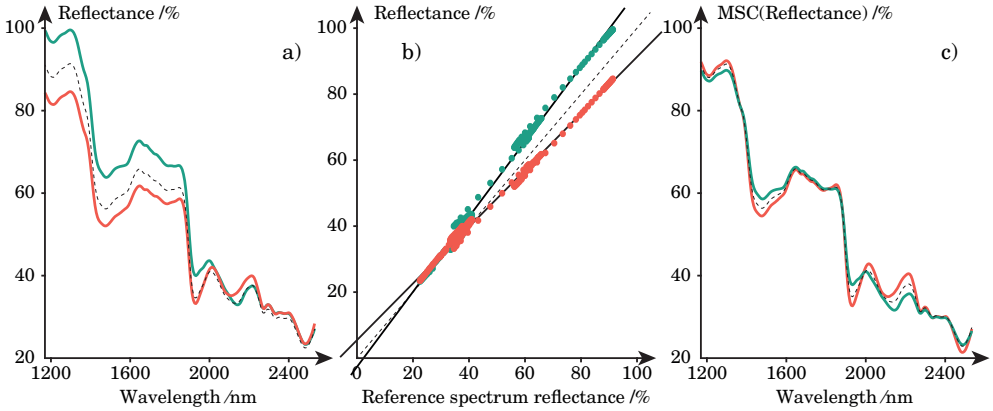


FIGURE 2.16 Concept behind MSC. Subplot a) shows the average reflectance spectrum of two spruce samples, one treated with a high concentration of phosphorous (green line) and one treated with a low concentration (orange line) together with the average spectrum of the entire dataset (dotted black line). Subplot b) shows the same spectra as in subplot a) but plotted against the average spectrum on the x-axis. Each point in subplot b) corresponds to a different wavelength of the spectra shown in subplot a). Subplot c) shows the same spectra after the MSC correction, i.e. after the spectra have been corrected with Eq. (2.12).

estimated slope.

2.5.2 Extended Multiplicative Scatter Correction (EMSC)

Extended Multiplicative Scatter Correction [53] (EMSC) is as the name suggests an extension to MSC. The difference between EMSC and MSC is that in EMSC new β terms are added to Eq. (2.12). In effect this means that instead of fitting straight lines through the spectra as shown in Figure 2.16 b) when estimating β_0 and β_1, a more complex regression line is fitted. One motivation for this is simply that in some datasets it is not adequate to estimate the scatter variations using straight lines (MSC). One common way of extending the MSC model, which is referred to as the *basic EMSC model* [54], is to fit a polynomial regression curve through the data using a linear term and a quadratic term in addition to the intercept and slope. Once the regression coefficients for the

polynomial curve have been estimated, the EMSC correction is performed by extending Eq. (2.12) with more subtractions in the numerator. In the case of basic EMSC this would mean:

$$\boldsymbol{s}_{\text{EMSC}} = \frac{\boldsymbol{s} - \beta_0 - \beta_2 \times \boldsymbol{v} - \beta_3 \times \boldsymbol{v}^2}{\beta_1}. \tag{2.13}$$

In Eq. (2.13) $\boldsymbol{v}$ is a vector of linearly increasing values for each wavelength, which estimates the linear trend, $\boldsymbol{v}^2$ is the squared version of the linear vector accounting for quadratic trends. β_2 and β_3 are the estimated regression coefficients associated with $\boldsymbol{v}$ and $\boldsymbol{v}^2$. As in Eq. (2.12), β_0 and β_1 are the intercept and the slope of the data in relation to the reference spectrum. Using a polynomial expression to correct undesired variations in the measured spectra often work well, and was used to correct the datasets in **Paper III** and **Paper IV** with great success.

Another approach to EMSC however, is to extend the MSC model with constituent spectra describing a specific type of variation within the measured data instead of using artificially constructed polynomial features. For instance, the polynomial vector $\boldsymbol{v}^2$ in Eq. (2.13) can be replaced with a vector containing an absorbance spectrum of pure water covering the same wavelength region as the measured spectra. During the parameter estimation the coefficients of the model will then attempt to accredit some of the variation in the measured spectra to the provided constituent spectrum—effectively quantifying the influence the constituent has on the spectra. The spectral variation caused by the constituent can then, once approximated, be suppressed from the measured spectra by subtracting its effect. Using constituent spectra in the EMSC modeling instead of polynomial features with little to no physical interpretation has at least two advantages:

1) once the regression coefficients associated with a particular constituent spectrum have been estimated for all the spectra in a hyperspectral image,

the regression coefficients can be viewed spatially as a 2D image, thus giving insight to how the constituent varies within the material. This is similar to a chemical map produced with regression modeling, but different in the sense that the numerical values of the EMSC coefficients are unitless whereas a regression model can be developed to make predictions in the same unit as the provided response data. Is it also different in the sense that the EMSC estimation of a constituent is unsupervised, since it does not involve a known target constituent concentration.

2) basing an EMSC model on constituent spectra can be used as a means of testing ones understanding of the spectral variation present in the measured data. If the variation in the measured spectra are adequately modeled by a set of user-provided constituent spectra, this can be seen as an indication that the underlying causes of the spectral variations are well understood.

Once the regression coefficient for a given constituent spectra has been estimated, the spectral variations caused by a particular constituent can also be excluded from the EMSC correction—thus ensuring that the spectral variation caused by the constituent is preserved after the EMSC correction. It should be noted however, that quantifying the presence of constituents with EMSC based on their pure constituent spectra is not an exact science and the estimation will inevitably contain imperfections. Partly because constituent spectra often change when interacting with a material; for instance, bound water in the cell walls of a wood sample is likely to have a different spectral signature from that of free water measured separately in a Petri dish. The measured spectra of a material can therefore typically not be perfectly described as a linear combination of its individually measured pure constituent spectra.

2.5.3 Standard Normal Variate (SNV)

Standard Normal Variate (SNV) [55] is another commonly used scatter-correction method. The SNV transformation of a spectrum s is defined as:

$$s_{\text{SNV}} = \frac{s - \bar{s}}{\text{std}(s)}. \tag{2.14}$$

Where $\bar{s}$ denotes the mean value of the s spectrum across all its wavelengths and $\text{std}(s)$ is the standard deviation of the spectrum. By comparing Eq. (2.14) to the MSC correction of Eq. (2.12) it becomes evident that there are great similarities between the two preprocessing techniques. The main difference is that SNV does not require a reference spectrum to be defined as MSC does; each spectrum is individually transformed and is not affected by the other observations/spectra of the dataset. Outcome-wise the corrected spectra obtained from SNV are also often similar in shape to the MSC corrected ones, as shown in Fig. 2.15 b) and c). Avoiding the need of a reference spectrum can be seen as the main advantage SNV has over MSC. A potential disadvantage of SNV is that the values of the corrected spectra no longer can be interpreted as absorbance values—as the transformed signal will always be centered around zero and have unit variance.

2.5.4 Spectral derivatives

The linear relationship between spectra and response can sometimes be improved by using the derivative of the spectra instead of the original spectra. The derivative of an absorbance spectrum describes the rate of change in absorbance at various wavelengths. Therefore, two absorbance spectra which have a constant offset between them, due to for instance undesired signal interference, but are otherwise of the same shape will after derivation be identical—because their rate of change in absorbance is identical regardless of

the absolute absorbance values. As previously mentioned in section 2.1.1.1, taking multiple absorbance measurements of samples with similar chemical composition can result in vastly different absorbance spectra due to optical path length variations, instrumental drifts, etc. Spectral derivatives can aid in suppressing such undesired variations by filtering away low-frequency trends in the data, reducing scatter effects, bringing the spectra together around a common baseline and thus enhancing meaningful variations [56]. Figure 2.17 illustrates the effect of applying first order spectral derivation to an absorbance spectrum. The upper part of Fig. 2.17 contains a measured absorbance spectrum along with two artificially constructed signal interferences applied to the original spectrum. The first interference is purely additive, and constructed by adding a constant c to the original spectrum. The second interference is both additive and multiplicative: in addition to the added constant c the spectrum is scaled with a multiplier which increases with wavelength. The lower part of Fig. 2.17 contains the first order derivative of each of the three spectra. As can be seen, the original spectrum and the spectrum with an additive constant interference are identical after the derivation. The spectrum with a multiplicative wavelength dependent interference however is still different from the other two spectra after the derivation, but its baseline is removed. Depending on the type of interferences present in a dataset, derivations higher than the first order could be beneficial. Applying first order derivation removes the baseline of any signal, applying second order derivation removes the baseline and any linear trend in the signal, which is why second order derivation is particularly popular in spectroscopy [57].

Since absorbance spectra are measured at a discrete number of wavelengths, and the underlying continuous function of the spectra is unknown, the true derivative cannot be analytically established but must instead be numerically approximated. The most straightforward way of numerically estimating the derivative of a signal is by using *finite difference* derivation. The derivative of

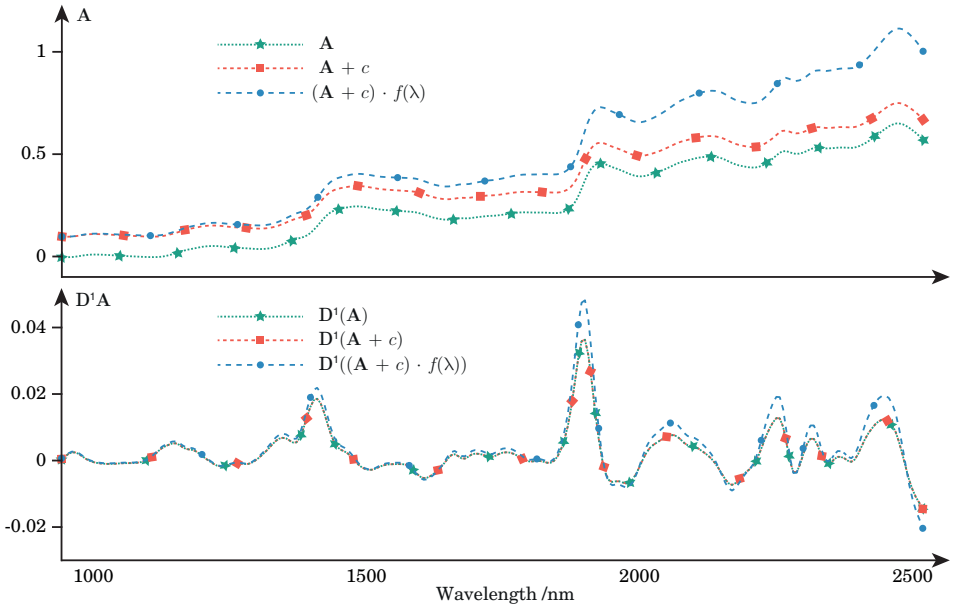


FIGURE 2.17 Effect of spectral derivation. Upper subplot contains the average absorbance spectrum from dataset I (green), along with two artificially altered versions of the same spectrum. The red curve represents the mean absorbance spectrum with an added constant c to it, in this example 0.1. The blue curve represents the mean absorbance spectrum with the added constant and a multiplicative disturbance which becomes increasingly amplified at longer wavelengths. In this example the wavelength-dependent multiplicative factor is a linearly increasing multiplier between 1.0 and 1.5. The lower subplot shows the first derivative of the same three spectra. The green and red curves—which in the upper subplot are separated—are identical after the derivation. The blue spectrum however, is still different.

a function f is defined as [58]:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}. \tag{2.15}$$

If the function f is a spectrum, determining the derivative of the spectrum would thus require the value of $f(x + h)$ to be known for a small value of h, ideally infinitesimally small. But since hyperspectral cameras measure spectra at a set of discrete points, the value of $f(x + h)$ is unknown for infinitesimally small values of h. In practice, the wavelengths of a spectrum measured with a hyperspectral camera are typically 5 or so nanometers apart. Approximating the derivative of a spectrum using finite difference consists in numerically inserting values into Eq. (2.15) using the smallest available value of h. Since dividing each point along the derivative with the constant h merely scales the signal uniformly and does not affect the shape of the derivative (provided that the spacing between measured wavelengths is constant), the division part of Eq. (2.15) is sometimes omitted when performing finite difference. If the difference between the measured wavelengths of a spectrum is 5 nm, the shape of the first order derivative of the spectrum $\mathbf{A}$ can therefore be approximated using finite difference as $\mathbf{A}(\lambda + 5\text{ nm}) - \mathbf{A}(\lambda)$ for each available wavelength λ. The obvious benefit to finite difference is its extreme simplicity, in high-level programming languages it can be efficiently implemented in a single line of code. The main drawback of the method is that it inflates high-frequency noise within spectra, which becomes increasingly severe when approximating higher order derivatives. Because of the noise inflation issue, spectral derivation is instead generally performed using techniques which incorporate some type of signal smoothing into the derivation process. One such derivation technique is *Savitzky-Golay filtering / derivation* [59].

Savitzky-Golay filtering is in essence a signal smoothing technique, but the

way the smoothing is mathematically formulated makes it advantageous to also incorporate derivation into the smoothing process. The technique works by iteratively sliding a window of some arbitrary width w across a spectrum. As the window is traversing the signal, at each new location a regression line is fitted with least squares regression using only the data points from the signal which are within the bounds of the window. The middle point of the regression line is then extracted and used in the new smoothed spectrum as a replacement for the data point at the same wavelength in the original signal. When fitting the regression line at each location the basic equation of a straight line, Eq. (2.9), can be used. Alternatively, a polynomial equation can be used, for instance of the form:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{2.16}$$

It is in this polynomial formulation that the opportunity for derivation emerges. Because the expression in Eq. (2.16) can analytically be derived into:

$$\hat{y'} = \beta_1 + 2\beta_2 x. \tag{2.17}$$

And consequently, once the coefficients β_0, β_1 and β_2 from Eq. (2.16) have been estimated with least squares fitting, the β_1 and β_2 coefficients can be inserted into Eq. (2.17) instead of Eq. (2.16) to obtain a smoothed first derivative version of the signal, rather than just a smoothed version. Figure 2.18 illustrates the main concept behind Savitzky-Golay derivation—i.e. sliding a window of width w across a spectrum and iteratively fitting polynomial regression lines which are used to approximate the derivative. By analytically deriving the expression of the first derivative a second time, the second derivative of the signal can be obtained. The maximum derivative order possible during Savitzky-Golay filtering is determined by the polynomial degree used in the polynomial regression line.

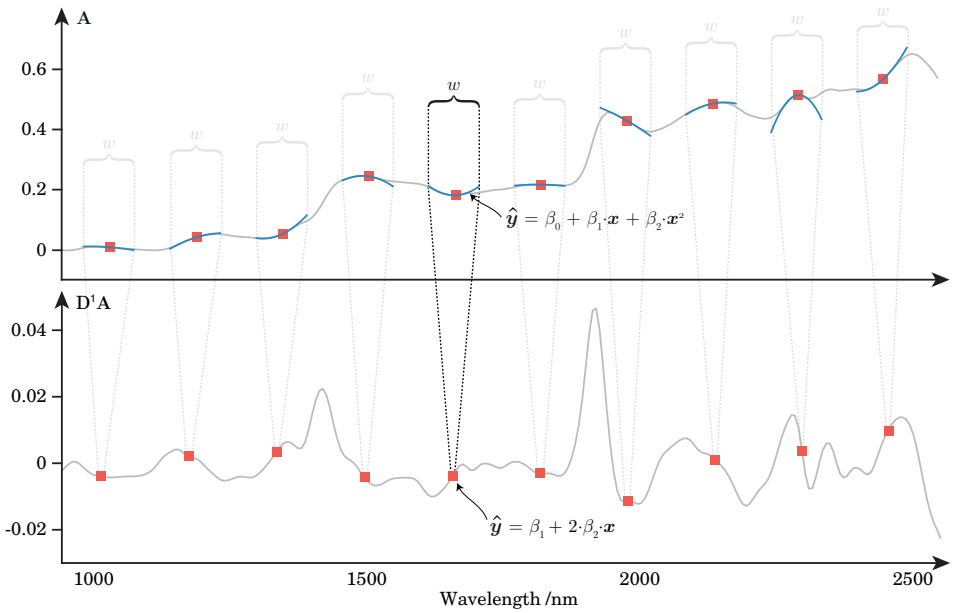


FIGURE 2.18 Concept behind Savitzky-Golay derivation. At each point along the spectrum a polynomial regression line is fitted to data within a window of width w. The coefficients of the regression line are then inserted into an equation describing the derivative of the polynomial line.

CHAPTER 3

Summary of datasets

During the work leading up to this thesis, three hyperspectral datasets were generated. All three datasets depict wood samples and were made in collaboration with the *Norwegian Institute of Bioeconomy Research* (NIBIO). Since these datasets are central to understanding the papers in the appendix, a summary of what they contain and how they were collected will be given here.

3.1 Dataset I - Hyperspectral NIR data of spruce samples treated with flame retardant

40 samples of Norway spruce (*Pincea abies*) were cut into dimension 50×50×10 mm. 35 of the square samples were originally untreated whilst 5 were impregnated upon delivery with a phosphorous-based flame retardant compound from Akzo Nobel called *Preventor AntiFlame*. The samples were conditioned in a climate chamber at 20 °C and 65 % relative humidity until all samples had reached their equilibrium moisture content. Using a pure undiluted version of the Preventor AntiFlame compound, given to us by Akzo Nobel, seven different concentrations of the solution were made by diluting it with water. These solutions were composed of 0 %, 17 %, 33 %, 50 %, 67 %, 83 % and 100 % Preventor AntiFlame. The different concentrations were then poured into

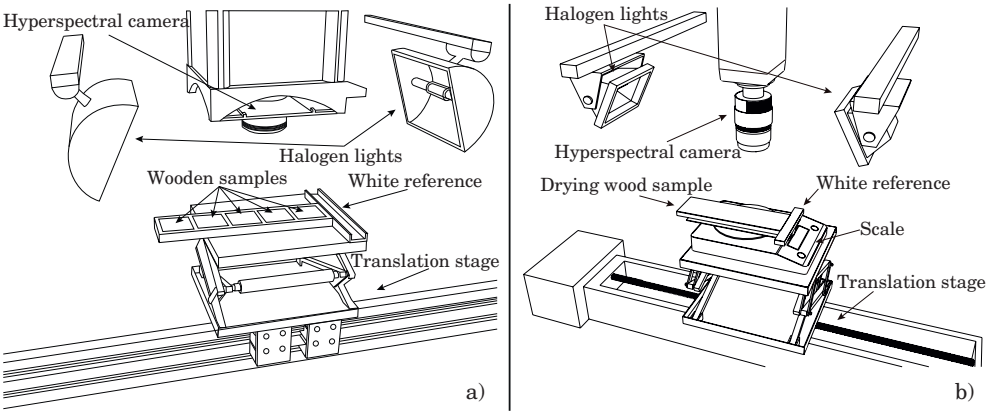


FIGURE 3.1 Illustration of experimental setup used during hyperspectral signal acquisition of a) dataset I; b) dataset II and III.

separate Petri dishes whereupon the 35 untreated spruce samples were immersed in the solutions for 30 seconds such that each of the concentrations were used to treat five samples each.

A sample mount capable of holding five samples in a straight vertical line together with a white reference was designed and 3D-printed. All 40 samples were then, five at a time, placed in the sample mount together with a spectralon white reference rod and scanned using a Specim SWIR hyperspectral camera in the wavelength range 929-2531 nm. The experimental setup of of the image acquisition can be seen illustrated in Fig. 3.1 a). After the NIR reflectance signal of each sample was measured, a 0.4-0.5 mm layer was shaved off from the surface of each of the 40 samples using a planer. An inductively coupled plasma (ICP) spectroscopy analysis was then performed by NIBO on the removed surface layers to determine (destructively) the average phosphorus content ($g{\cdot}kg^{-1}$) of each surface layer.

This dataset was generated primarily to investigate the possibility of predicting the surface concentration of phosphorous in a nondestructive way by correlating it to the collected NIR signal, i.e. calibrating a regression model using the hyperspectral NIR signal as $\boldsymbol{X}$ and the phosphorus content of each sample as $\boldsymbol{y}$,

which was done in **Paper III**. This dataset was also used in **Paper II** to evaluate the performance of a variable selection concept.

3.2 Dataset II - Hyperspectral vis/NIR time series data of pine during drying

A pine (*Pinus sylvestris*) sample from a forest in Hobøl, Norway, was cut into dimension 280×100×18 mm. Afterwards it was placed in a drying oven at 103 °C where it was kept during 48 hours. Over the course of a few hours the sample was then repeatedly weighted to ensure that its weight had become stable, at which time the sample's dry weight was measured. The sample was then fully submerged under water and left to soak for 24 hours. After the soaking period the sample was placed on top of a digital scale, which in turn was attached to a translation stage underneath a hyperspectral camera as shown in Fig. 3.1 b).

Using an in-house written MATLAB script, the hyperspectral camera was automated to take images of the pine sample at eight minute intervals whilst also recording the weight registered by the scale at the time of each image acquisition. The sample was then monitored for approximately 21 hours, resulting in a time series containing both a hyperspectral image and a sample weight at each time step. The hyperspectral camera (Specim, Oulu, Finland) measured 200 bands in the 392-1022 nm region.

The primary purpose of the dataset was to use it to experimentally evaluate the method for studying spatiotemporal spectral data presented in **Paper V**. The dataset was also used to evaluate variable selection concepts in **Paper I** and **Paper II** by correlating the measured spectral data to the moisture content of the wood.

3.3 Dataset III - Hyperspectral NIR time series data of thermally modified pine during drying

The sample preparation and experimental setup of dataset III is similar to dataset II, the major differences being: (1) the type and number of samples used; (2) the type of hyperspectral camera used; (3) the temporal resolution of the time series. Eight samples of thermally modified pine (*Pinus sylvestris*) were cut into the dimension of 280×100×18 mm. The samples were then dried in an oven at 103 °C, this time for 4 days. Afterwards the dry weight of all eight samples was measured. The samples were then submerged under tap water were they were kept for approximately one and a half months to allow the samples to become completely saturated. When taken from the water after the soaking period the samples were one by one scanned by an automated hyperspectral camera whilst situated on top of a digital scale as shown in Fig. 3.1 b). For this dataset a NIR camera (HySpex SWIR-384) was used covering the wavelength range 953–2516 nm. The camera was programmed to take one image per minute and was monitoring each sample for approximately 21.5 hours, resulting in a time series sequence of roughly 1290 hyperspectral images per sample. For each of the time steps the average moisture content (MC) of the samples was calculated using the following expression:

$$MC(t) = \frac{w(t) - w_{\text{Dry}}}{w_{\text{Dry}}} \times 100. \tag{3.1}$$

In Eq. (3.1) w_{Dry} represents the pre-established dry weight of each sample and $w(t)$ represents the measured weight of a wood sample at time t as it dries.

The purpose of the dataset was to allow a model to be generated from the spectral

data capable of estimating the spatial distribution of moisture in each wood sample as it dried by using the average moisture content at each time step as a response value ($\boldsymbol{y}$). This concept was explored in **Paper IV**.

CHAPTER 4

Results & discussion

4.1 Method for fast hyperspectral wavelength selection

The aim of **Paper I** was to develop a computationally efficient method for conducting variable selection on hyperspectral data with partial least squares (PLS) regression. To that end, several published PLS algorithms were initially implemented and evaluated in terms of computational efficiency. During these evaluations it was noted that in certain types of PLS algorithms—*kernel PLS* algorithms—the first computational step in the algorithm is to calculate the products $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$. In the remaining part of these algorithms the original $\boldsymbol{X}$ and $\boldsymbol{y}$ data are never referenced again. I.e., only $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ are needed to estimate regression coefficients using these algorithms. In the work leading up to **Paper I** we discovered that the ability of fitting regression coefficients from only $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ is immensely beneficial when conducting feature selection on tall datasets, for reasons which will be outlined below.

If $\boldsymbol{X}_i$ represents a matrix containing only a subset of the columns from the full $\boldsymbol{X}$ matrix, $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ (and $\boldsymbol{X}_i^\top \boldsymbol{y}$) are needed to estimate regression coefficients for the subset using kernel PLS. The conventional, obvious, approach used for determining $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ is to multiply $\boldsymbol{X}_i^\top$ with $\boldsymbol{X}_i$. In **Paper I** however, we developed a new technique which allows $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ to be established without matrix

multiplication for any column subset i, provided that the full $\boldsymbol{X}^\top \boldsymbol{X}$ is calculated beforehand. The technique is most intuitively understood by visually examining what calculations are involved in an $\boldsymbol{X}^\top \boldsymbol{X}$ matrix multiplication. Given an $\boldsymbol{X}$ matrix and its transpose $\boldsymbol{X}^\top$, the calculations involved in establishing the $\boldsymbol{X}^\top \boldsymbol{X}$ product with matrix multiplication are shown in Fig. 4.1. The columns of $\boldsymbol{X}$ are colored in Fig. 4.1 to emphasize a pattern: each color/column from the original $\boldsymbol{X}$ matrix only influences one row and one column of $\boldsymbol{X}^\top \boldsymbol{X}$. More concretely, the j^{th} column from $\boldsymbol{X}$ only affects the j^{th} row and column of the $\boldsymbol{X}^\top \boldsymbol{X}$ product. Once calculated, the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix contains all possible combinations of dot products between the column vectors in $\boldsymbol{X}$. Obtaining $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ from $\boldsymbol{X}^\top \boldsymbol{X}$ is therefore simply a matter of filtering out vector products which are not relevant for the column subset i. The technique introduced in **Paper I** for obtaining $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ via filtering consists in representing a column subset i as a vector of ones and zeros which is aligned along both the rows and columns of $\boldsymbol{X}^\top \boldsymbol{X}$. A value of zero in the vector represents a column being excluded from the column subset i and a value of one represents inclusion. $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ can be obtained simply by extracting elements from $\boldsymbol{X}^\top \boldsymbol{X}$ which intersect with one in both the horizontal and vertical direction. This concept is visualized in Fig. 4.2. Although not shown in Fig. 4.2, the indexing technique works in an identical fashion when obtaining $\boldsymbol{X}_i^\top \boldsymbol{y}$ from the full $\boldsymbol{X}^\top \boldsymbol{y}$.

Calculating $\boldsymbol{X}^\top \boldsymbol{X}$ only to extract a single $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ matrix via indexing would clearly be a nonsensical and inefficient way of establishing $\boldsymbol{X}_i^\top \boldsymbol{X}_i$, since this would involve calculating dot products which will later be filtered away. The benefit of the technique emerges when multiple different subsets are used; in such cases calculations which would have been identical between various column subsets no longer need to be recomputed.

The practical implications of the described indexing technique is that variable selection with kernel PLS can be performed in a new way. Figure 4.3 illustrates

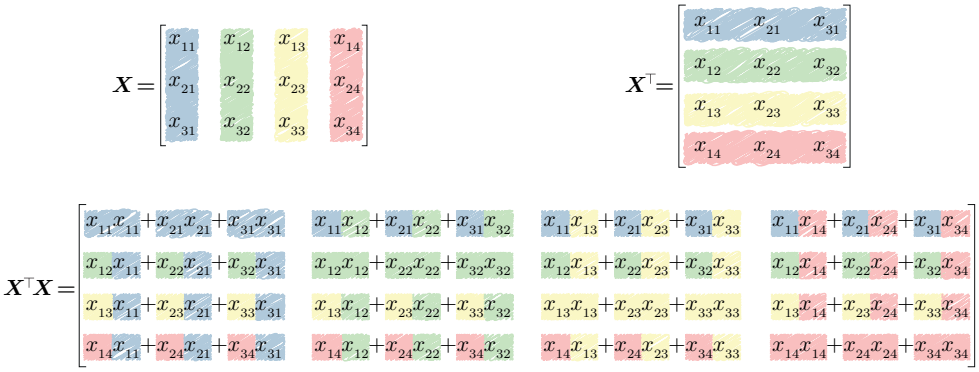


FIGURE 4.1 Illustration of calculations involved when forming the matrix product $\boldsymbol{X}^\top \boldsymbol{X}$. An $\boldsymbol{X}$ matrix is shown in the top left with each column assigned a unique color. $\boldsymbol{X}^\top$, the transposed version of $\boldsymbol{X}$, is shown in the top right. The lower part of the figure illustrates how the elements from $\boldsymbol{X}^\top$ and $\boldsymbol{X}$ blend together when forming $\boldsymbol{X}^\top \boldsymbol{X}$. In the resulting matrix product, the influence of the j^{th} column from $\boldsymbol{X}$ is confined to the j^{th} row and column of $\boldsymbol{X}^\top \boldsymbol{X}$.

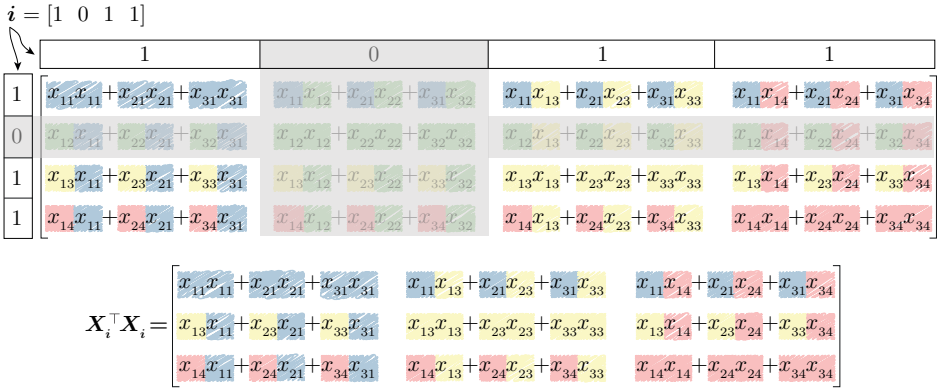


FIGURE 4.2 Illustration of filtering technique introduced in **Paper I** for obtaining an $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ matrix for a column subset $\boldsymbol{i}$ given $\boldsymbol{X}^\top \boldsymbol{X}$. The upper part of the figure contains the $\boldsymbol{X}^\top \boldsymbol{X}$ from Fig. 4.1 along with an example subset vector $\boldsymbol{i} = [1\ 0\ 1\ 1]$ which has been aligned along both the height and width of $\boldsymbol{X}^\top \boldsymbol{X}$. The lower part of the figure contains the $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ matrix, obtained by extracting rows and columns from $\boldsymbol{X}^\top \boldsymbol{X}$ which are intersecting with a value of one in the $\boldsymbol{i}$ vector along both rows and columns.

the conventional algorithmic approach to performing variable selection with kernel PLS (left), and the new way enabled by the findings of **Paper I**. As shown in the right part of Fig. 4.3, the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix multiplication, which is often the most computationally demanding operation in a kernel PLS algorithm, is only performed once regardless of the number of different subsets being evaluated. How much of an improvement in computational efficiency the algorithm to the right in Fig. 4.3 offers in relation to the left algorithm depends on the size of the dataset used and the number of variable selections being evaluated. The taller the $\boldsymbol{X}$ and $\boldsymbol{y}$ data are and the more column subsets are being evaluated, the greater the performance gains. When a dataset is not strongly overdetermined, or when only a handful of subsets are evaluated, the performance benefits can be negligible. When evaluating a large number of variable subsets on a very tall dataset however, it is not uncommon for the method proposed in **Paper I** to be hundreds or thousands of times faster compared to the conventional approach. This is shown in Fig. 4.4 which contains the results of a benchmark where kernel PLS was used on a dataset of size $10^7 \times 100$ to fit regression coefficients for batches of randomly generated subset vectors. The x-axis in Fig. 4.4 describes the number of random subset vectors evaluated with PLS and the y-axis describes the time in seconds it took to estimate regression coefficients for all subsets. Note that the y-axis of Fig. 4.4 is logarithmic—each major y-tick corresponds to an order of magnitude increase in the required calculation time.

Although it was not experimentally demonstrated as a part of **Paper I**, it is easy to see that the same indexing technique can also be applied to Eq. (2.11) to efficiently obtain the OLS solution for any column subset of $\boldsymbol{X}$. Eq. (2.11) is repeated here for convenience:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2.11}$$

```
for Each variable selection
    • Extract relevant columns from X
    • Calculate X^T X and X^T y using extracted columns
    • Solve the regression problem using X^T X and X^T y
end
```

```
• Calculate X^T X and X^T y using all available columns
for Each variable selection
    • Index into the full X^T X and X^T y and extract regions relevant for the current variable subset
    • Solve the regression problem using extracted quantities
end
```

FIGURE 4.3 Side-by-side pseudocode illustrating the conventional way of performing variable selection with kernel PLS (left), and the method enabled by the indexing technique introduced in **Paper I** (right).

As seen in Eq. (2.11), OSL estimated regression coefficients can, as in the case of kernel PLS, be obtained directly from the $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ quantities. The indexing technique from **Paper I** is therefore directly applicable. Unpublished benchmark results indicate that the technique from **Paper I** offers very similar performance benefits in the OLS case as in the kernel PLS case.

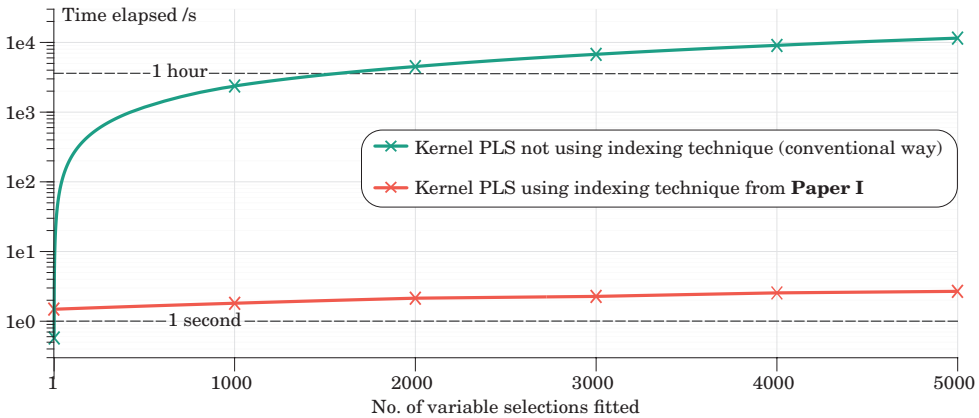


FIGURE 4.4 Benchmark of computational time required for estimating regression coefficients using kernel PLS for batches of randomly generated variable selections. The $\boldsymbol{X}$ matrix used was of size $10^7 \times 100$. The *Modified kernel algorithm #2* [60] was used to estimate regression coefficients without indexing technique (green) and with indexing technique (red). Calculations were performed using an Intel i7-7700K @ 4.2 GHz.

4.2 Method for faster wavelength selection with multiple preprocessing techniques

In addition to identifying relevant wavelengths to include in a model, another challenge which needs to be overcome during hyperspectral model development is choosing which spectral preprocessing method to use. The aim of **Paper II** was to leverage the technique introduced in **Paper I** in such a way that it could be used to perform wavelength selection on multiple preprocessing techniques simultaneously. Which would allow several combinations of preprocessing techniques and wavelength selections to be quickly evaluated together within a single algorithm, making it easier to identify which combination of pretreatment method and wavelength subset is appropriate for a given dataset.

One of the main advantages of kernel PLS algorithms which operate on $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ as opposed to $\boldsymbol{X}$ and $\boldsymbol{y}$ is that they require very little working memory during the fitting process. The reason for this is that the $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$ products are considerably smaller in size compared to $\boldsymbol{X}$ and $\boldsymbol{y}$, provided that the $\boldsymbol{X}$ and $\boldsymbol{y}$ data contains substantially more rows than columns. Since, if $\boldsymbol{X}$ is an $m \times n$ matrix, $\boldsymbol{X}^\top \boldsymbol{X}$ will always be an $n \times n$ matrix. If m for instance is 1 000 000 and n is 256, in practical terms this means that $\boldsymbol{X}$ requires roughly one gigabyte of memory to store (if stored as 32-bit floating point numbers). Storing $\boldsymbol{X}^\top \boldsymbol{X}$ on the other hand, only requires 0.25 megabyte. Holding multiple $\boldsymbol{X}^\top \boldsymbol{X}$ matrices in memory and operating on them is therefore not an issue. The fact that multiple $\boldsymbol{X}^\top \boldsymbol{X}$ matrices can easily be held in memory was utilized in **Paper II**. The developed algorithmic concept is summarized below.

A spectral dataset $\boldsymbol{X}$ is preprocessed with different methods which are

horizontally concatenated into a wide matrix $\boldsymbol{X}_{\text{Tot}}$:

$$\boldsymbol{X}_{\text{Tot}} = [\boldsymbol{X} \;\; f_1(\boldsymbol{X}) \;\; f_2(\boldsymbol{X}) \;\; \ldots \;\; f_p(\boldsymbol{X})]. \tag{4.1}$$

In Eq. (4.1) f_1, f_2, f_p refers to different spectral preprocessing functions such as the ones described in section 2.5. The large $\boldsymbol{X}_{\text{Tot}}$ matrix is then condensed in size into a considerably smaller $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix with conventional matrix multiplication. An example of what $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ can look like is shown in Fig. 4.5, which depicts an $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix generated by applying six different preprocessing methods to dataset I. As can be seen in Fig. 4.5, $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ exhibits a clear chequered pattern. Each submatrix along the diagonal of the chequered structure contains the intact $\boldsymbol{X}^{\top}\boldsymbol{X}$ product for each of the differently preprocessed versions of $\boldsymbol{X}$. Thus, a single $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix contains all the necessary information needed to estimate PLS coefficients for each of the involved f_1, f_2, f_p preprocessing methods, and every imaginable wavelength selection within the preprocessed dataset. The data needed for PLS with any given preprocessing method can easily be obtained by indexing into the matrix and extracting relevant elements. Moreover, the $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix can easily be used to create models utilizing wavelengths preprocessed in different ways by extracting elements from $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ which are not part of the any of the submatrices along the diagonal and using these in a model.

When extracting data from $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ corresponding to a specific variable subset of a specific preprocessing method, the indexing technique introduced in **Paper I** can still be used, provided that a small modification is applied to the subset vector. The only modification needed is that the subset vector $\boldsymbol{i}$ used during the extraction needs to be padded with zeros in all locations which are outside of the preprocessing-specific region of interest. This concept of padding solutions with zeros can be seen illustrated in Fig. 4.6.

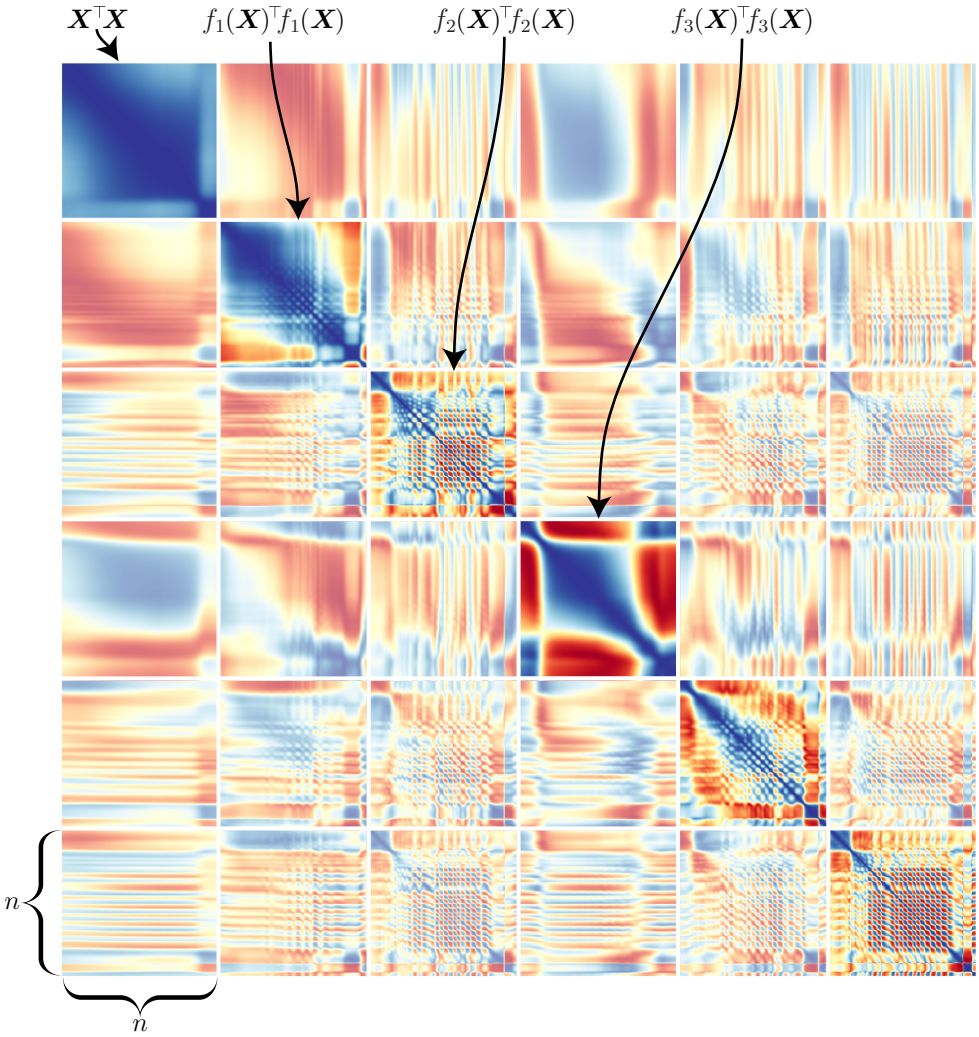


FIGURE 4.5 Illustration of $\boldsymbol{X}_{\text{Tot}}^\top \boldsymbol{X}_{\text{Tot}}$ matrix product obtained by preprocessing the NIR data of dataset I with six different methods and horizontally placing them next to each other in $\boldsymbol{X}_{\text{Tot}}$ prior to performing the matrix multiplication. In addition to being immensely aesthetically pleasing, the data contained in the matrix can be used to fit PLS models for any variable subset of any of the involved preprocessing techniques in $\boldsymbol{X}_{\text{Tot}}$ by indexing into the matrix using a modified version of the indexing concept introduced in **Paper I**.

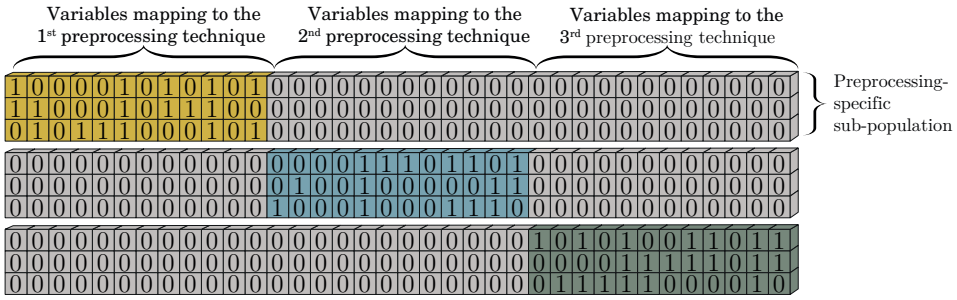


FIGURE 4.6 Illustration of padding variable selections with zeros to make them align to their preprocessing-specific area of an $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix.

The concept of extracting regions of an $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix relating to different preprocessing methods and using them to efficiently fit PLS models is general enough that it can be used to perform feature selection on multiple preprocessing techniques with any wrapper-based technique. In **Paper II**, we demonstrate the technique specifically for use with genetic algorithms and use it to evolve wavelength selections for multiple preprocessing techniques simultaneously. The motivation behind this choice of wrapper method is that genetic algorithms are notoriously slow since they operate on whole populations of solutions and thus require a lot of subsets to be evaluated. For this reason variable selection with genetic algorithms stand to benefit substantially from a computational speedup in the PLS coefficient estimation. To practically incorporate the indexing technique into a GA and evolve feature selections on multiple preprocessing techniques simultaneously a few modifications to a conventional GA were necessary: (1) the population was divided into different preprocessing-specific sub-populations, i.e. each individual of the population was restricted to only operate in one square along the diagonal of the $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix relating to its assigned preprocessing technique; (2) the crossover operator had to be restricted in such a way that it did not mix together individuals which were assigned different preprocessing methods; (3) the parent selection and survivor selection had to be carried out separately for each preprocessing technique, otherwise the best performing preprocessing techniques would quickly start to dominate the population and kill

off techniques with lower performance[1].

In **Paper II** we implement the described technique and evolved variable selections for 12 different preprocessing techniques within a single algorithm. The run time of our developed technique was experimentally measured and compared to four GA implementations using conventional PLS fitting procedures. Our results indicate that the method is faster than any of the traditional approaches we compared it against. Compared to traditional PLS methods, the technique introduced in **Paper II** lends itself especially well to parallel implementations on specialized hardware such as GPUs, which makes the computational efficiency of the suggested method even more compelling and substantially faster than the conventional alternatives. The reason for this is that most PLS algorithms operate on the $\boldsymbol{X}$ data, which is often large in size when working with hyperspectral data. This makes it difficult to operate on multiple subsets of the quantity simultaneously in parallel without running in to memory issues. This is especially true when performing the computations on GPUs where the available working memory is typically substantially lower compared to the system's RAM. Since a variable subset of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ is often vastly smaller in size compared to the corresponding subset of $\boldsymbol{X}_{\text{Tot}}$, thousands of such subsets can be extracted and operated upon in parallel at a low memory cost when using kernel PLS with our introduced indexing technique. With the developed method from **Paper II**, hundreds or thousands of kernel PLS algorithms can therefore be executed in parallel on a GPU—all operating on different variable selections of various preprocessing techniques—which allows for very high parameter estimation throughput.

[1] In terms of finding a model which minimizes the RMSE_{cv}, it may be advantageous to let poorly performing preprocessing techniques die off, but this was not the desired outcome of our demonstration in **Paper II**.

4.3 Estimating phosphorus in spruce

The research objective of paper **Paper III** was to develop a model for nondestructively estimating the concentration of a flame retardant treatment in spruce surfaces. To that end, 40 spruce samples were treated with a phosphorus-based flame retardant and scanned with a NIR hyperspectral camera as described in section 3.1. The surface layer of each sample was removed and analyzed using ICP to establish a ground truth value of the average phosphorus content in each sample. Phosphorus is the active substance in the studied flame retardant and can be seen as proportional to the flame retardant's ability of protecting against fire. Using EMSC to alleviate scatter interferences in the collected hyperspectral data, we demonstrated that it is possible to use parts of the near-infrared region of an absorbance spectrum to estimate the concentration of phosphorous in the surface of Norway Spruce (*Pincea abies*) boards using PLS regression. This was accomplished by modeling the ICP-established phosphorus content from the measured absorbance spectrum. Backwards elimination was used to determine which of our measured wavelengths in the 929-2531 nm region held significant predictive abilities over the phosphorous content. The backwards elimination identified a sequence of 17 wavelengths, all clustered around a single absorbance peak in the 2400-2531 nm region of the spectra. Furthermore, when using these 17 wavelengths the lowest RMSE_{cv} was obtained with a single PLS component ($A = 1$), resulting in a model of very low complexity.

By using hyperspectral imaging, as opposed to simpler point-based measurements, we demonstrated that the spatial distribution of phosphorous in the spruce surface could be estimated. Which revealed two things: (1) greater deposits of phosphorus were generally located in the earlywood regions of the spruce surface compared to the latewood regions; (2) resin shares some of the same spectral characteristics as wood treated with a very high phosphorus

content, which leads to resin being misclassified by our model as a region of very high phosphorous concentration. With regards to the first point it should be noted that although it intuitively seems reasonable that the chemical uptake is higher in the lower density earlywood areas as suggested by our model, we cannot know for sure that the modeled spatial distribution of phosphorus is accurate. This is simply because we only have the laboratory-established phosphorus content expressed as an average value for each spruce sample—the spatial distribution behind the average value is unknown to us. Because we are not aware of any alternative method which can provide the spatial distribution of phosphorus in the wood, we have no means of validating the distribution estimated by our developed model.

Figure 4.7 shows the spatially resolved predictions of phosphorus for 14 of the spruce samples in **Paper III**. The developed model had an R^2 of 0.87 when exposed to 15 validation samples which were previously unseen by the model. Potential use-cases of the model include: (1) verifying that an adequate concentration of flame retardant treatment is present on spruce boards before use in a construction; (2) surveying existing constructions, potentially on a large scale such as whole building facades etc. which will allow nondestructive inspection of the treatment an arbitrary length of time after the initial construction.

Although we demonstrated in **Paper III** that hyperspectral imaging can be used in a laboratory environment to successfully estimate the phosphorous content of spruce boards, before such a model is deployed in practice further studies will have to be performed to verify that the method is robust enough to be reliably used outdoors where a number of disturbances are present which may influence the spectra. Especially since the model was found to be tricked by naturally occurring phenomena such as resin.

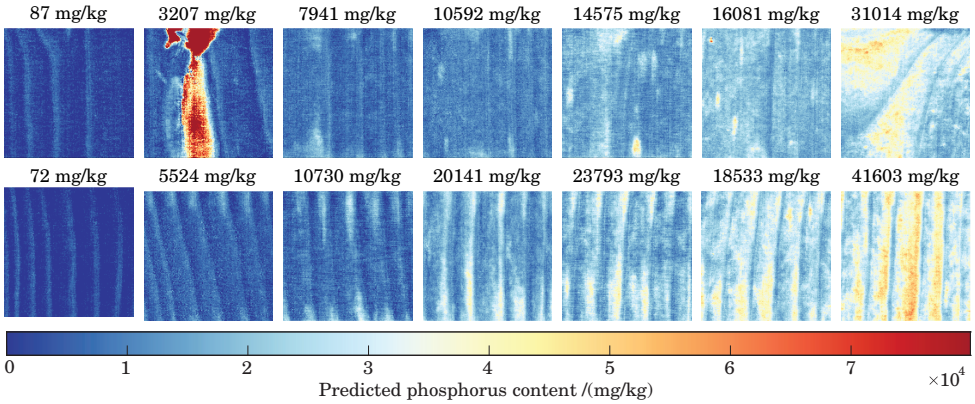


FIGURE 4.7 Modeled spatial distribution of phosphorus content in 14 of the spruce samples from **Paper III**. Value above each sample represents the ICP-established average phosphorus content of the sample. The second sample from the left in the upper row contains resin defects which is misclassified as phosphorus.

4.4 Estimating moisture in thermally modified pine

A previous study by Kobori et al. [61] showed that vis-NIR hyperspectral imaging combined with PLS regression can be used to nondestructively estimate the moisture content of pine. The primary aim of **Paper IV** was to investigate if similar techniques can be used to estimate the moisture content of thermally modified pine samples. To experimentally evaluate this, eight thermally modified pine samples were soaked in tap water for one and a half months and then monitored with a hyperspectral NIR camera on a minute-by-minute basis as they dried, as explained in section 3.3. Each sample was monitored roughly one day, resulting in a massive time series dataset containing over 9500 hyperspectral images. For each hyperspectral image taken, the sample weight was recorded. The weight was then used to calculate a sample-average moisture content associated with each image using Eq. (3.1). PLS regression was then used to map the collected spectral signal to the measured average moisture content of the samples at each time step. To enhance the performance of

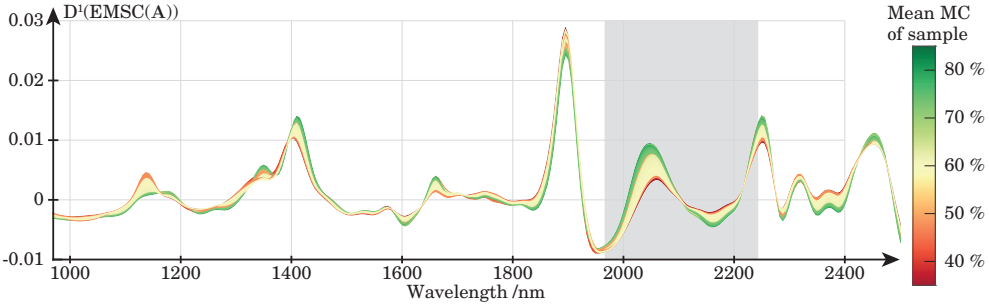


FIGURE 4.8 Mean absorbance spectrum of every hyperspectral image in the eight time series acquisitions from **Paper IV**. The spectra are preprocessed with EMSC followed by first order Savitzky-Golay derivation (window size 7, polynomial degree 1). Gray region in the figure indicates wavelength selection identified by the moving window feature selection algorithm. All spectra in the figure are colored according to the average moisture content of the sample they originate from; green color means the pine sample is wet, red means it is dry.

the developed model, different combinations of preprocessing techniques and feature selection algorithms were evaluated on the data. Extended multiplicative scatter correction (EMSC) followed by first order Savitzky-Golay derivation was found to produce the lowest model prediction error of all evaluated pretreatment techniques. The best performing wavelength selection was identified using *moving window* variable selection, and consisted in 52 wavelengths in the 1966-2244 nm region. The moving window variable selection algorithm consists in simply moving a window of a specified width across the spectra and evaluating the performance, i.e. $RMSE_{cv}$, of a model based only on the variables within the window at each step. Figure 4.8 shows the mean preprocessed spectrum of every image in the dataset along with the wavelength subset found to be highly correlated to the moisture content of the pine. The final PLS model was calibrated on six of the hyperspectral time series, and then applied to the two unseen samples' time series (approximately 2400 images) in order to verify the models performance on new observations. On average the calibrated model was capable of estimating the sample-average moisture content of the new pine samples with 2.7 % prediction error.

In order to obtain separate estimates of the average moisture content within

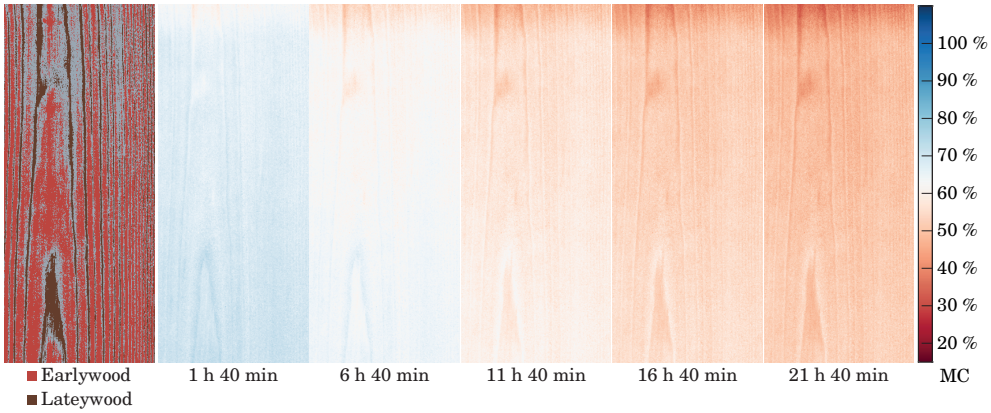


FIGURE 4.9 Generated early-/latewood segmentation mask for one of the thermally modified pine samples from **Paper IV** (left) along with the spatial distribution of PLS modeled moisture content for the same sample at five points in time throughout the drying process.

early- and latewood regions of each board, the collected hyperspectral images were spatially segmented using an early-/latewood segmentation algorithm developed by Smeland et al. [62]. The segmentation algorithm consists in first performing a principal component analysis on a hyperspectral image and then forming a histogram of the scores from one of the principal components. Two thresholds are then placed in the histogram, and pixels with a score value below the lower threshold are classified as earlywood while pixels with a score value greater than the higher threshold are classified as latewood. Score values between the two thresholds are not considered earlywood nor latewood. Once a spatial segmentation mask had been developed for each of the eight samples, the PLS model was used to estimate the spatial distribution of moisture content for each pine board at each time step during the drying sequences. The segmentation masks were then superimposed onto the spatial predictions, thus allowing estimates of early- and latewood moisture content to be separately extracted. Figure 4.9 (left) shows the result of the spatial early-/latewood segmentation of one of the samples in the study along with chemical maps of estimated moisture content for the same sample at five points in time during the time series.

Our results indicate that—albeit not by a particularly large margin—all eight pine samples consistently had a higher PLS modeled average moisture content in the earlywood than in the latewood. Averaged across all samples and time steps the modeled difference between early- and latewood moisture content was 1.7 %. However, as in the case of the chemical maps produced in **Paper III**, one of the main limitations of **Paper IV** is that we have no means of validating the spatial predictions produced by the PLS model since only the sample-average true moisture content is known to us. As such, we cannot know for sure that the PLS modeled contrast between early- and latewood is valid. If the modeled contrast between early- and latewood moisture content is indeed accurate, then the results from **Paper IV** suggests that there is typically a rather small (only about 1.7 %) difference in moisture content between early- and latewood regions of thermally modified pine—which is a desired property for wood to possess since large moisture differentials may lead to cracks and deformations. Further studies should however be conducted where pine samples are studied both before and after undergoing thermal modification in order to conclude that the small moisture differential can be accredited to the thermal modification and not any other factor.

4.5 Method for studying known and unknown variations in hyperspectral time series

In **Paper V** the aim was to develop a generic method for modeling and summarizing signal variations in spatiotemporal hyperspectral data, i.e. hyperspectral time series data, in such a way that variations of known origin are separated from variations of unknown origin. The purpose of our developed method can be seen as threefold: (1) it allows a dataset to be substantially reduced in size by describing the data in terms of a few interpretable components, thereby

eliminating the need to store the entire spectral data; (2) ones understanding of the collected spectral dataset in question can be tested by seeing how much of the data can be modeled using the user-provided variation vectors of known origin; (3) the unmodeled and therefore unknown part of the signal can be studied separately from the known variations which may allow it to be interpreted as something meaningful, thus broadening ones understanding of the dataset.

Because hyperspectral time series data is four-dimensional, it can easily be massive in size and unfeasible to store unpartitioned in RAM. Many traditional methods of analyzing the data are therefore not applicable. To circumvent the issues associated with data size, our developed method operates in an online fashion—analyzing data incrementally as it arrives in a stream over time. As explained in section 2.5.2, the variations in a measured spectrum can be accredited to different constituent spectra using the extended multiplicative scatter correction (EMSC) model. Such constituent spectra can describe spectral variations of either chemical or physical origin. The method introduced in **Paper V** utilizes this concept to quantify and create a highly compressed representation of the variations of known origin in the data. The unmodeled part of the data not captured by the EMSC model is then further scrutinized in an attempt to identify systematic but previously unknown variations present in the data which can both enhance the understanding of the dataset and make it more compressible. The concept which was developed in **Paper V** can be summarized in four steps:

1. A hyperspectral image arrives and is transformed into absorbance using the concepts explained in section 2.2.2 - 2.2.3.
2. An attempt is made to model the known part of the signal using an EMSC model equipped with as many relevant user-defined constituent spectra $(\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_j)$ as possible. With least squares fitting the EMSC model quantifies the concentration of each provided constituent spectra into a set of regression coefficients $(\beta_{i,1}, \beta_{i,2}, ..., \beta_{i,j})$ for each of the i pixels in the

image. Each spectrum from the hypercube is then modeled as an additive combination of the constituents:

$$\hat{\boldsymbol{z}}_i = \boldsymbol{c}_1 \times \beta_{i,1} + \boldsymbol{c}_2 \times \beta_{i,2} + ... + \boldsymbol{c}_j \times \beta_{i,j}. \tag{4.2}$$

Inevitably, there will be a discrepancy between the EMSC modeled spectra $\hat{\boldsymbol{z}}_i$ and the measured spectra $\boldsymbol{z}_i$, this step will thus produce a residual signal containing variations of unknown origin which are not well described by the user-defined spectra.

③ The residuals from the EMSC model are then subjected to dimensionality reduction. Since the method is developed to handle incrementally arriving data, where only a fraction of the dataset is observable at a given time, the dimensionality reduction technique has to accommodate this. Our suggested method uses the *On-The-Fly Processing* (OTFP) [63] algorithm to compress the hyperspectral residual data into a low-dimensional set of loading and score vectors. At this point the original hyperspectral image can be removed from memory and a new one can be inserted into step ①.

④ Once enough data has passed through steps ①, ② and ③ the loading vectors from OTFP can then be analyzed manually or via other modeling procedures in an attempt to understand the systematic spectral variation unexplained by the provided constituent spectra. Using the stored regression coefficients from step ②, the constituent spectra and the scores and loading vectors from step ③, the discarded hyperspectral signal from any previous timestep can be recreated with a high degree of fidelity.

In **Paper V** we experimentally evaluate the concept outlined above on a time series containing 150 vis/NIR hyperspectral images depicting a pine board as it dried during 21 hours (dataset II). A set of five user-defined constituent spectra was used in the EMSC model to quantify known spectral variations. Three of

which were intended to account for physical light scattering effects in the signal and two of which were intended to account for chemical absorption expected to influence the measured signal.

Apart from the large data compression enabled by the described method, the data exploration aspect of the method—where systematic and previously unknown spectral variations are presented to the user via a set of interpretable loading vectors—is arguably one of the most compelling things about the approach. The exploration side of the methodology was however held back in our experiments by the fact that our five-component EMSC model managed to describe nearly all the variation in the data. Figure 4.10 shows 10 measured absorbance spectra from the pine sample in its wet state (left) compared to the EMSC corrected versions of the same 10 spectra (right). The EMSC corrected spectra shown in the figure are spectra where the quantified known variations are removed. Since the EMSC model is based on a reference spectrum, deviations from the reference spectrum after the EMSC correction corresponds to unmodeled variations. In total, the five-component EMSC model was able to account for 98.94 % of the variance in our experimental time series dataset, leaving just 1.06 % of the signal unexplained.

Within the unexplained 1.06 % residual data the OTFP algorithm was able to identify a number of systematic variations. The most dominant variation found by OTFP is shown in Fig. 4.11. The left part of Fig. 4.11 shows the spectral signature of the dominant variation identified in the EMSC residuals, the right part shows its temporal development during the 21 hours of drying. As can be seen in the figure, the identified variation has a very clear temporal development, and lessens in intensity as time progresses, i.e. as the pine sample gets drier, indicating that whatever the identified variation represents it is likely related to water.

From the results of the experiments performed during the work leading up to

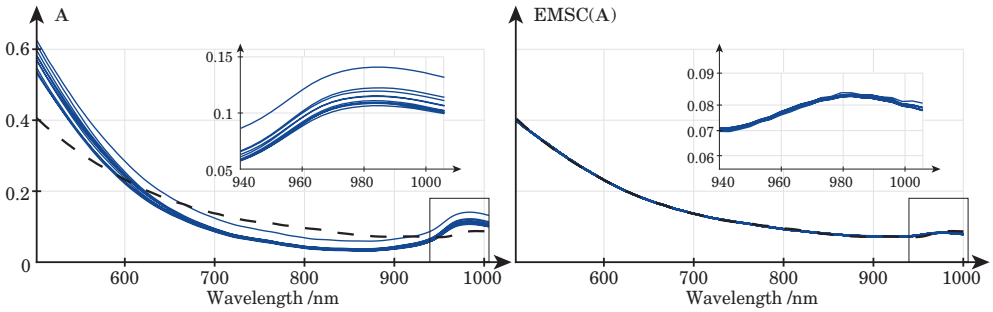


FIGURE 4.10 Ten absorbance spectra of from a wet pine board before EMSC correction (left) and after EMSC correction (right). The black dotted line in the figure represents the reference spectrum used during EMSC. Deviations from the reference spectrum in the right figure represents the residual signal which was sent to OTFP for dimensionality reduction. The 940-1005 nm region which is heavily associated with water absorption is magnified in each figure.

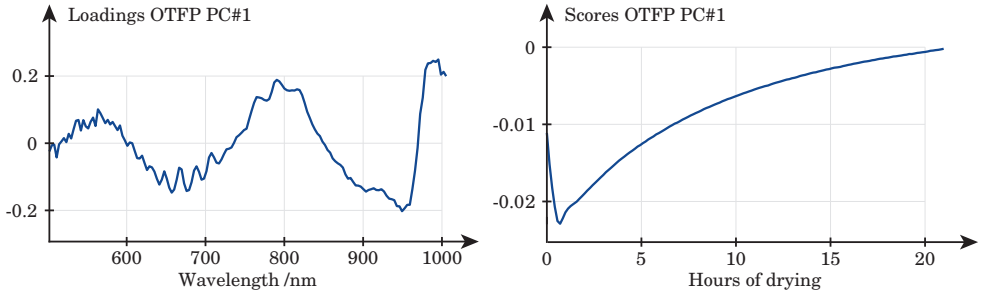


FIGURE 4.11 Dominant variation (first principal component) identified by the OTFP algorithm after processing the residuals from the EMSC model. Left plot shows the loading vector, describing the wavelength dependence of the identified variation. Right plot shows the score value as a function of time, averaged across all pixels in an image, describing how the variation developed during the 21 hour long signal acquisition.

Paper V, one can only conclude that although the suggested method appears to work as intended, the dataset we used to showcase the method was admittedly not overwhelmingly interesting in the sense that it did not contain a lot of ”mysterious unexplained variations” from which major revelations about the dataset could be derived.

CHAPTER 5

Conclusions

Several research topics related to hyperspectral imaging, model development and applications thereof to wood science were addressed in the papers of this thesis. The outcome of the performed work is listed below:

- a new way of doing variable selection with PLS was introduced, which is especially well suited for hyperspectral data with many more observations than variables;
- a new way of performing wavelength selection with GA-PLS was introduced, allowing the selection to efficiently be performed on multiple preprocessing methods simultaneously;
- it was practically demonstrated that hyperspectral imaging in the NIR region can be used to predict the phosphorus content in spruce surfaces;
- it was practically shown that hyperspectral imaging in the NIR region can be used to estimate the moisture content of thermally modified pine;
- a new method was introduced for analyzing hyperspectral time series data, capable of compressing streaming measurement data, quantifying variations of known and unknown origin in the data and presenting the user with a summary of poorly understood systematic variations.

5.1 Suggestions for future research

For tall datasets the indexing technique introduced in **Paper I** allows PLS models to be fitted for a massive amount of variable subsets in a small fraction of the time it traditionally would take to calibrate the models. This opens the door to new possible ways of doing wavelength selection which previously would have been considered unfeasible due to the computational cost. For instance, conventional step-wise methods such as forward selection can be modified such that instead of evaluating every possible activation of one single variable each step of the algorithm, every possible activation of k variables are evaluated each step of the algorithm. The number of variable selection evaluations needed at the first step of such an algorithm would then be $(n + 1)!/(k! \times (n - k + 1)!)$ instead of just n. If n is 256 and k is 3 for instance, nearly three million variable selections need to be evaluated at the first step, instead of 256. Algorithm modifications such as this are conceptually easy to invent and implement in code, but would likely be considered impractical prior to the introduction of **Paper I** due to their computational cost and are therefore currently most likely unexplored. When using the indexing technique from **Paper I** and performing the PLS computations in parallel on graphics card(s), calibrating a few million PLS models is no longer an issue. Modifying and extending simple existing variable selection algorithms such as step-wise methods to incorporate a more rigorous exploration of subsets, for instance in the manner mentioned above, could therefore help such algorithms overcome local optima—at least ones of variable size up to k—which may lead to the identification of better subsets than what a traditional step-wise method would find. How much such modifications to existing variable selection algorithms potentially could improve existing feature selection algorithms should therefore be studied further.

As alluded to in section 4.3, despite hyperspectral imaging having been shown

in **Paper III** to be capable of estimating phosphorous in spruce in a laboratory environment, further studies should be carried out to investigate how well the technique works outdoors in the presence of various interferences. Additionally, it should be verified that the model works as intended when the wood has aged and weathered. Another open question is whether or not the same technique can be used to estimate the phosphorous content of wood boards which have a coating, such as paint, applied to the wood surface in addition to the flame retardant.

A previous study has shown that hyperspectral imaging can be used to estimate the moisture content of pine [61], **Paper IV** demonstrated that it is also a viable technique for estimating moisture in thermally modified pine. This opens the door to new methods of studying how the thermal modification influences the wood: if a board (non-thermally modified) is monitored with hyperspectral imaging as it dries and then the same board is thermally modified and monitored again during similar drying conditions, this would enable a direct insight to how the moisture dynamics is altered by the thermal modification.

Bibliography

[1] J. Hildebrandt, N. Hagemann, and D. Thrän, "The contribution of wood-based construction materials for leveraging a low carbon building sector in europe," *Sustainable Cities and Society*, vol. 34, pp. 405–418, 2017.

[2] A. H. Buchanan and S. Levine, "Wood-based building materials and atmospheric carbon emissions," *Environmental Science Policy*, vol. 2, no. 6, pp. 427–437, 1999.

[3] U. Y. Ayikoe Tettey, A. Dodoo, and L. Gustavsson, "Carbon balances for a low energy apartment building with different structural frame materials," *Energy Procedia*, vol. 158, pp. 4254–4261, 2019.

[4] J. Sjåvik, *Historical dictionary of Norway*. Scarecrow Press, 2008.

[5] RISE Research Institutes of Sweden, "Technical guideline for fire safety in timber buildings," 2017. [Online]. Available: https://www.sp.se/FSITB

[6] W. Ellis and R. Rowell, "Flame-retardant treatment of wood with a diiscyanate and an oligomer phosphonate," *Wood and Fiber Science*, vol. 4, pp. 367–375, 1989.

[7] K. Sandin, *Praktisk byggnadsfysik*, 1st ed. Studentlitteratur, 2010.

[8] P. A. Schweitzer, *Atmospheric degradation and corrosion control*, 1st ed. Marcel Dekker, 1999.

[9] S. Karlsen Lie, "Surface mould growth as a contributor to visual changes of exterior wooden claddings," Ph.D. dissertation, Norwegian University of Life Sciences, 2019.

[10] C. A. S. Hill, *Wood modification*. Wiley, 2006.

[11] B. Esteves and H. Pereira, "Wood modification by heat treatment: a review," *BioResources*, vol. 4, no. 1, 2009.

[12] D. Johansson, "Strenght and colour response of solid wood to heat treatment," Ph.D. dissertation, Luleå University of Technology, 2005.

[13] SP Technical Research Institute of Sweden, *Benchmarking and State of the art for Modified wood*, 2013.

[14] E. Dunningham and R. Sargent, *Review of new and emerging international wood modification technologies*, 2015.

[15] D. Cirule, A. Meija-Feldmane, E. Kuka, B. Andersons, N. Kurnosova, A. Antons, and H. Tuherm, "Spectral sensitivity of thermally modified and unmodified wood," *BioResources*, vol. 11, no. 1, 2015.

[16] M. A. Javed, P. M. Kekkonen, S. Ahola, and V.-V. Telkki, "Magnetic resonance imaging study of water absorption in thermally modified pine wood," *Holzforschung*, vol. 69, no. 7, pp. 899–907, 2015.

[17] P. M. Kekkonen, A. Ylisassi, and V.-V. Telkki, "Absorption of water in thermally modified pine wood as studied by nuclear magnetic resonance," *The Journal of Physical Chemistry C*, vol. 118, no. 4, pp. 2146–2153, 2014.

[18] J. Sandak, A. Sandak, D. Pauliny, V. Krasnoshlyk, and O. Hagman, "Near infrared spectroscopy as a tool for estimation of mechanical stresses in wood," *Advanced Materials Research*, vol. 778, pp. 448–453, 2013.

[19] T. Fujimoto, H. Kobori, and S. Tsuchikawa, "Prediction of wood density independently of moisture conditions using near infrared spectroscopy," *Journal of Near Infrared Spectroscopy*, vol. 20, no. 3, pp. 353–359, 2012.

[20] K. Watanabe, S. D. Mansfield, and S. Avramidis, "Application of near-infrared spectroscopy for moisture-based sorting of green hem-fir timber," *Journal of Wood Science*, vol. 57, no. 4, pp. 288–294, 2011.

[21] T. Higuchi, "Formation of earlywood, latewood,and heartwood," in *Biochemistry and Molecular Biology of Wood*, T. E. Timell, Ed. Berlin: Springer, 1997, ch. 6.

[22] R. K. Bamber, *Sapwood and Heartwood*. Forestry Commission of New South Wales, 1987.

[23] M. Orchin, R. S. Macomber, A. R. Pinhas, and R. Marshall Wilson, *The vocabulary of organic chemistry*, 2nd ed. Wiley-Interscience, 2005.

[24] I. Murray and P. Williams, "Chemical principles of near-infrared technology," in *Near-infrared technology in the agricultural and food industries*, P. Williams and K. Norris, Eds. St. Paul, Minnesota, USA: American Association of cereal chemists, 1987, ch. 2, pp. 17–22.

[25] A. P. French and E. F. Taylor, *An introduction to quantum physics*. W.W. Norton, 1978.

[26] M. R. Roussel, *A Life Scientist's Guide to Physical Chemistry*. Cambridge University Press, 2012.

[27] G. Wypych, *Handbook of material weathering*, 5th ed. ChemTec, 2013.

[28] D. W. Ball, *Field guide to spectroscopy*. SPIE, 2006.

[29] D. A. Skoog, D. M. West, F. J. Holler, and S. R. Crouch, *Fundamentals of analytical chemistry*, 8th ed. Brooks/Cole Cengage Learning, 2014.

[30] J. Workman Junior, *Practical guide to interpretive near-infrared spectroscopy*. CRC Press, 2008.

[31] H. A. Martens, "Quantitative interpretation of non-selective chemical data," Ph.D. dissertation, Technical University of Norway, Trondheim, 1985.

[32] M. Mancini, G. Toscano, and Å. Rinnan, "Study of the scattering effects on nir data for the prediction of ash content using emsc correction factors," *Journal of Chemometrics*, 2019.

[33] Å. Rinnan, F. v. d. Berg, and S. B. Engelsen, "Review of the most common preprocessing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[34] M. N. Leger, "Alleviating the effects of light scattering in multivariate calibration of near-infrared spectra by path length distribution correction," *Applied Spectroscopy*, vol. 64, no. 3, pp. 245–254, 2010.

[35] T. Konevskikh, "Modeling scattering and absorption in the infrared spectroscopy of cells and thin films," Ph.D. dissertation, Norwegian University of Life Sciences, 2017.

[36] G. T. Georgiev and J. J. Butler, "Long-term calibration monitoring of spectralon diffusers brdf in the air-ultraviolet," *Appl. Opt.*, vol. 46, no. 32, pp. 7892–7899, Nov 2007. [Online]. Available: http://ao.osa.org/abstract.cfm?URI=ao-46-32-7892

[37] J. Mavor and P. Denyer, "The dependence of ccd dark current upon power dissipation," *Microelectronics Reliability*, vol. 17, no. 3, p. 403–404, 1978.

[38] A. Tasch, R. Brodersen, D. Buss, and R. Bate, *Dark-current and storage-time considerations in charged-coupled devices*. Texas Instruments Incorporated, 1973, pp. 179–180.

[39] G. Birth S. and H. G. Hecht, "The physics of near-infrared reflectance," in *Near-infrared technology in the agricultural and food industries*, P. Williams and K. Norris, Eds. St. Paul, Minnesota, USA: American Association of cereal chemists, 1987, ch. 1, p. 9.

[40] R. H. William, "Data analysis: Wavelength selection methods," in *Near-infrared technology in the agricultural and food industries*, P. Williams and K. Norris, Eds. St. Paul, Minnesota, USA: American Association of cereal chemists, 1987, ch. 3, p. 33.

[41] S. Chatterjee and A. S. Hadi, *Regression Analysis by Example*. Wiley, 2012.

[42] S. Wold, M. Sjostrom, and L. Eriksson, "Pls-regression: a basic tool of chemometrics," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, p. 109–130, 2001.

[43] M. Williams, C. Grajales, and D. Kurkiewicz, "Assumptions of multiple regression: Correcting two misconceptions," *Practical Assessment, Research Evaluation*, vol. 18, no. 11, p. 1–14, 2013.

[44] H. Abdi, "Partial least square regression pls-regression," in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Thousand Oaks, Calif: SAGE Publications, 2007, p. 3.

[45] S. de Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[46] H. Kubinyi, "Evolutionary variable selection in regression and pls analyses," *Journal of Chemometrics*, vol. 10, no. 2, pp. 119–133, 1996.

[47] T. Mehmood, K. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, Aug. 2012.

[48] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[49] T. Phuong, Z. Lin, and R. Altman, "2005 ieee computational systems bioinformatics conference (csb'05) - choosing snps using feature selection," 2005, pp. 301–309.

[50] H. Martens, S. Å. Jensen, and P. Geladi, "Multivariate linearity transformations for near-infrared reflectance spectrometry," in *Nordic Symposium on Applied Statistics*, O. H. J. Christie, Ed. Stavanger, Norway: Stokkand Forlag, 1983, pp. 205–234.

[51] M. Maleki, A. Mouazen, H. Ramon, and J. De Baerdemaeker, "Multiplicative scatter correction during on-line measurement with near infrared spectroscopy," *Biosystems Engineering*, vol. 96, no. 3, pp. 427–433, 2007.

[52] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[53] H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 9, no. 8, pp. 625–635, 1991.

[54] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 92–99, 2012.

[55] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[56] H. Dehghani, F. Leblond, B. W. Pogue, and F. Chauchard, "Application of spectral derivative data in visible and near-infrared spectroscopy," *Physics in Medicine and Biology*, vol. 55, no. 12, pp. 3381–3399, 2010.

[57] H. W. Siesler, Y. Ozaki, S. Kawata, and H. M. Heise, *Near-infrared spectroscopy*, 1st ed. Wiley-VCH, 2002.

[58] W. Karush, *The crescent dictionary of mathematics*, 1st ed. MacMillan Publishing Company, 1962.

[59] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[60] B. S. Dayal and J. F. MacGregor, "Improved pls algorithms," *Journal of Chemometrics*, vol. 11, no. 1, pp. 73–85, 1997.

[61] H. Kobori, N. Gorretta, G. Rabatel, V. Bellon-Maurel, G. Chaix, J.-M. Roger, and S. Tsuchikawa, "Applicability of vis-nir hyperspectral imaging for monitoring wood moisture content (mc)," *Holzforschung*, vol. 67, no. 3, 2013.

[62] K. A. Smeland, K. H. Liland, J. Sandak, A. Sandak, L. R. Gobakken, T. K. Thiis, and I. Burud, "Near infrared hyperspectral imaging in transmission mode: Assessing the weathering of thin wood samples," *Journal of Near Infrared Spectroscopy*, vol. 24, no. 6, pp. 595–604, 2016.

[63] R. Vitale, A. Zhyrova, J. F. Fortuna, O. E. de Noord, A. Ferrer, and H. Martens, "On-the-fly processing of continuous high-dimensional data streams," *Chemometrics and Intelligent Laboratory Systems*, vol. 161, pp. 118–129, 2017.

Errata

The following formal errors were corrected after the committee had evaluated the thesis.

Side	Line	Original text	Corrected text
ii	16	Dr. Federic Marini	Prof. Federico Marini
4	17-19	…for a given hyperspectral spectral dataset.	…for a given hyperspectral dataset.
9	23-24	Detailed knowledge of how a material reflects/absorbs radiation can be often be used…	Detailed knowledge of how a material reflects/absorbs radiation can often be used…
11	16-17	An assumption which is to a varying degree will always be inaccurate.	An assumption which to a varying degree will always be inaccurate.
17	10-12	…the overtone absorption will take place at shorter wavelength in the electromagnetic spectrum…	…the overtone absorption will take place at shorter wavelengths in the electromagnetic spectrum…
22	8-9	…result in different hyperspectral images due to changes ambient light conditions etc.	…result in different hyperspectral images due to changes in ambient light conditions etc.
30	23	…in the orientation process attempts to maximize the covariance between $\boldsymbol{X}$ and y [44].	…in the orientation process attempts to maximize the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ [44].
34	23	…the concept behind two variable selections methods…	…the concept behind two variable selection methods …
38	19-20	…gets to spread their genes (vector elements) to throughout the population…	…gets to spread their genes (vector elements) throughout the population…
41	1	Average absorbance spectra of each sample in the dataset I…	Average absorbance spectra of each sample in dataset I…
62	6	Time elapsed time /s	Time elapsed /s
63	7	…wavelength selections to be quickly be evaluated…	…wavelength selections to be quickly evaluated…
64	25	This is concept of padding solutions with zeros can be seen…	This concept of padding solutions with zeros can be seen…

Paper I

Orders of magnitude speed increase in Partial Least Squares feature selection with new simple indexing technique for very tall data sets

P. Stefansson, U. G. Indahl, K. H. Liland, and I. Burud
Faculty of Science and Technology, Norwegian University Of Life Sciences.

Feature selection is a challenging combinatorial optimization problem that tends to require a large number of candidate feature subsets to be evaluated before a satisfying solution is obtained. Due to the computational cost associated with estimating the regression coefficients for each subset, feature selection can be an immensely time consuming process and is often left inadequately explored. Here we propose a simple modification to the conventional sequence of calculations involved when fitting a number of feature subsets to the same response data with partial least squares model fitting. The modification consists in establishing the covariance matrix for the full set of features by an initial calculation and then deriving the covariance of all subsequent feature subsets solely by indexing into the original covariance matrix. By choosing this approach, which is primarily suitable for tall design matrices with significantly more rows than columns, we avoid redundant (identical) recalculations in the evaluation of different feature subsets. By benchmarking the time required to solve regression problems of various sizes, we demonstrate that the introduced technique outperforms traditional approaches by several orders of magnitude when used in conjunction with Partial Least Squares (PLS) modeling. In the supplementary material, we provide code for implementing the concept with kernel partial least squares regression.

Keywords: PLS, feature selection, variable selection, subset selection, Kernel PLS

1. INTRODUCTION

For centuries linear least squares fitting has been one of the most important statistical tools for mapping a set of independent variables, $\mathbf{X}$, to a dependent response variable, y. Its use has today spread to nearly every quantitative field of science, and in areas such as bioinformatics and chemometrics multiple linear regression is employed extensively. A serious challenge in the data driven fields consists in identifying which column vectors contained within a potentially megavariate data matrix $\mathbf{X}$ are significantly correlated to the response vector y and which are not. The process of eliminating non-informative variables from $\mathbf{X}$ is typically referred to as either *feature selection*, *variable selection* or *subset selection*. Only an exhaustive search is guaranteed to identify the globally best feature subset, which becomes computationally infeasible as soon as the number of independent variables in the data are more than just a few. Some heuristic method is therefore often required in practice to identify a combination of variables that is considered 'good enough'. Examples of feature selection methods frequently used in bioinformatics and chemometrics include: *forward selection*, *backward selection*, *genetic algorithms* (GA), *simulated annealing* and *interval PLS* (iPLS) [1, 2]. Each feature selection technique has its own advantages and disadvantages; stepwise methods such as forward/backward selection are for instance relativley fast to use but are prone to getting stuck in local optimas [1]. Population based feature selection methods such as genetic algorithms can overcome such local optimas, but are often in comparison extremely slow, which limits the circumstances under which they may be applied sucessfully [3]. A characteristic trait of most wrapper-based feature selection methods is that they are driven by trial and error. As such, most methods generally require a large number of candidate subsets to be evaluated before a useful solution is obtained. Due to the computational cost of calibrating regression models for each of the subsets this process can be very time consuming for big data sets.

Here, we present a seemingly trivial insight to matrix multiplications which, when utilized for feature selection purposes in a specific way, leads to nontrivial speedups in the execution time required to explore the performance of a large number of variable subsets subject to linear modeling. The introduced technique is suitable to use together with any feature selection strategy that requires a large number of candidate subsets to be evaluated, such as any of the feature selection methods mentioned above. Our approach involves an extensive reuse of the results from identical calculations that occur across the evaluation of

different variable subsets. The calculations available for reusing can easily be applied together with a particular version of the Partial Least Squares (PLS) method in order to obtain substantial improvements in computational performance.

2. THE CONCEPT

2.1. Relevant background information

In ordinary least squares (OLS) regression problems we assume that our $(\mathbf{X}, y)$–data are well described by a linear model of the form

$$y = \mathbf{X}\beta + \epsilon, \tag{1}$$

where the error term ϵ should be 'small'. The least squares solution of this equation is obtained by the estimated vector of regression coefficients, $\hat{\beta}$, minimizing the sum of squared errors $\|\epsilon\|^2 = \|y - \hat{y}\|^2$, where the vector of fitted response values $\hat{y}$ is calculated by the matrix-vector multiplication $\hat{y} = \mathbf{X}\hat{\beta}$.

When the $m \times n$ matrix $\mathbf{X}$ is composed of a large number of observations $m \gg n$ where n denotes the number variables, finding the OLS solution of the linear system with respect to β in the numerically most accurate way may become slow and computationally costly. The *normal equations* for the OLS solution of Eq. 1 is the $n \times n$-system

$$\mathbf{X}^\top y = \mathbf{X}^\top \mathbf{X}\beta. \tag{2}$$

Solving the normal equations directly is usually not recommended due to the potentially unfavourable condition number of $\mathbf{X}^\top\mathbf{X}$ with associated numerical issues. On the other hand, provided that $m \gg n$, Eq. 2 yields a much smaller system of equations which can be solved considerably faster also when accounting for the computational costs of forming the products $\mathbf{X}^\top y$ and $\mathbf{X}^\top\mathbf{X}$. In addition to the OLS use case, these products are also exploited in some partial least squares algorithms, such as kernel PLS, to reduce the dimensions of the regression problem under consideration and speed up the $\hat{\beta}$-estimation.

The feature selection problem requires comparison of a large number of competing models. The computational advantages of solving normal equations rather than the original systems may therefore justify a priority over numerical precision in the explorative phase of a feature selection study where most candidate models will be discarded anyway. For the most attractive feature combination candidates found in the explorative phase, the final feature evaluation and -selection should be based on repeated modeling using numerically stable algorithms for solving the associated regression problems. In the following section we investigate a simple, but seemingly overlooked, aspect of the $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ calculations and demonstrate that these quantities, which may be the most computationally costly calculations involved in solving Eq. 2, only need to be calculated once regardless of how may column subsets of $\mathbf{X}$ one decides to evaluate.

2.2. Technique for reusing $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ calculations with indexing

Consider a simple design matrix, $\mathbf{X}$, containing three observations of four independent variables, each indicated with its own color:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{bmatrix}.$$

And its transpose

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{bmatrix}$$

When multiplying $\mathbf{X}^\top$ with $\mathbf{X}$ to form the quantity $\mathbf{X}^\top\mathbf{X}$ needed to solve Eq. 2, the colors—i.e. the independent variables—will 'blend' with each other according to the rules of linear algebra as shown in Fig. 1.

As can be seen in Fig. 1, the influence of the first column of $\mathbf{X}$—the blue one—can be traced along the first column and the first row of $\mathbf{X}^\top\mathbf{X}$. The second column of the original $\mathbf{X}$ matrix—the green one—only influences the second column and the second row of $\mathbf{X}^\top\mathbf{X}$. This pattern continues throughout the resulting matrix product regardless of what dimension the original $\mathbf{X}$ matrix is of. In summary; the influence of the j^{th} column of $\mathbf{X}$ is confined to the j^{th} row and column of $\mathbf{X}^\top\mathbf{X}$.

This means that if one already has determined the $\mathbf{X}^\top\mathbf{X}$ matrix using all available columns of $\mathbf{X}$ and then, for instance, wishes to obtain what the matrix product would have been had the last column of $\mathbf{X}$—the red one—not been included, it is not necessary to redo the matrix multiplication with one less column. Instead the last row and column of the full $\mathbf{X}^\top\mathbf{X}$ matrix product can simply be discarded.

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} 1\cdot1+2\cdot2+3\cdot3 & 1\cdot4+2\cdot5+3\cdot6 & 1\cdot7+2\cdot8+3\cdot9 & 1\cdot10+2\cdot11+3\cdot12 \\ 4\cdot1+5\cdot2+6\cdot3 & 4\cdot4+5\cdot5+6\cdot6 & 4\cdot7+5\cdot8+6\cdot9 & 4\cdot10+5\cdot11+6\cdot12 \\ 7\cdot1+8\cdot2+9\cdot3 & 7\cdot4+8\cdot5+9\cdot6 & 7\cdot7+8\cdot8+9\cdot9 & 7\cdot10+8\cdot11+9\cdot12 \\ 10\cdot1+11\cdot2+12\cdot3 & 10\cdot4+11\cdot5+12\cdot6 & 10\cdot7+11\cdot8+12\cdot9 & 10\cdot10+11\cdot11+12\cdot12 \end{bmatrix}$$

FIG. 1: Illustration of how the elements of $\mathbf{X}^\top$and $\mathbf{X}$ would blend during the calculation of the matrix product $\mathbf{X}^\top\mathbf{X}$. Numbers are colored according to their column origin in the example $\mathbf{X}$ given in section 2.2.

In fact, when $\mathbf{X}^\top\mathbf{X}$ is calculated including the complete set of n variables in $\mathbf{X}$, each of the matrix products $\mathbf{X}_i^\top\mathbf{X}_i$ obtained by defining $\mathbf{X}_i$ as one of the (2^n-1) possible column subset matrices of $\mathbf{X}$ can be derived without additional calculations—simply by indexing into the already available full product $\mathbf{X}^\top\mathbf{X}$. The same idea holds true for finding $\mathbf{X}_i^\top y$ by indexing into $\mathbf{X}^\top y$. Thus for any variable subset matrix $\mathbf{X}_i$, the associated normal equations $\mathbf{X}_i^\top y = \mathbf{X}_i^\top\mathbf{X}_i\beta$ can be obtained directly by an appropriate indexing into the full normal equations (2). To see in mathematical notation why this is true, assume that the vector $y \in \mathbb{R}^m$ and the matrix $\mathbf{X} = [x_1\ x_2\ ...\ x_n]$ has dimension $m \times n$ where the column vectors $x_1,\ ...,\ x_n \in \mathbb{R}^m$. Note that except for the diagonal values of $\mathbf{X}^\top\mathbf{X}$, exactly two vectors are involved in the calculation of each entry for both

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} x_1^\top x_1 & x_1^\top x_2 & ... & x_1^\top x_n \\ x_2^\top x_1 & x_2^\top x_2 & ... & x_2^\top x_n \\ : & ... & x_i^\top x_j & : \\ x_n^\top x_1 & x_2^\top x_n & ... & x_n^\top x_n \end{bmatrix} \text{ and } \mathbf{X}^\top y = \begin{bmatrix} x_1^\top y \\ x_2^\top y \\ : \\ x_i^\top y \\ : \\ x_n^\top y \end{bmatrix}.$$

Each matrix– and vector entry is calculated by taking dot products of the form $x_i^\top x_j$ and $x_i^\top y$, respectively. Elimination of all entries involving any specific vector x_i will therefore remove its contribution to the final product—clearly without influencing any of the remaining dot products. As will be demonstrated below, the performance implications of this observation are profound when using a kernel PLS algorithm for evaluating

multiple variable combinations.

3. APPLICATION TO FEATURE SELECTION WITH PARTIAL LEAST SQUARES, PLS

In fields such as chemometrics where multicolinearity amongst the variables in $\mathbf{X}$ is common, a partial least squares approach is often used instead of ordinary least squares due to the robustness benefits that comes with using latent rather than actual variables under such conditions [4, 7]. For large data sets where the $\mathbf{X}$ matrix is either very tall or very wide *Kernel PLS* algorithms have been developed as faster alternatives to the conventional NIPALS PLS fitting procedure [5, 8, 9]. The primary focus in this paper lies on situations where $\mathbf{X}$ is tall, i.e. has substantially more rows than columns ($m \gg n$). In such cases, the original Kernel PLS algorithm [5] or any of the existing derivatives/improvements of which [6] generally constitutes good algorithm choices in terms of computational efficiency. One property of these kernel algorithms—that turns out to be greatly beneficial when conducting feature selection—is that they base their entire parameter estimation process around the information content of the covariance matrices $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$. Because they operate on the covariance matrices, the indexing strategy introduced in section 2.2 can easily be incorporated into the fitting process in order to further increase the computational efficiency of evaluating multiple feature subsets. Furthermore, calculating the quantities $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ is generally amongst the very first steps involved in kernel based PLS algorithms, which makes the necessary alteration required to incorporate the indexing technique from section 2.2 trivial to implement and limited to a small part of the calculation sequence. To speed up the coefficient fitting of a batch of feature selections with a kernel PLS algorithm, the only modification required compared to a conventional naive approach is to loop through the subsets one by one *after* the initial covariance matrices have been calculated, rather than placing the same loop around the entire PLS algorithm. Essentially this means that the variable selection procedure is placed inside the PLS algorithm rather than wrapped around it. The fundamental differences between the two approaches are respectively made clear in algorithms 1 and 2 which depict the concepts behind conventional variable selection using kernel PLS and the suggested modified implementation. Practically implementing the matrix indexing technique described in section 2.2 is exceptionally simple in most high-level programming languages: first the quantities $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ are calculated with the full set of variables. Then, the relevant covariance elements for any feature subset can be extracted from these quantities by applying the same indexing logic across all dimensions. The most straightforward approach to achieving this is to represent a particular feature selection as an n-dimensional Boolean vector and then applying the vector as a means of indexing into $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$, i.e. only including the dot products of intersecting $true$-elements. The process can then be repeated for all feature subsets in a loop as shown in algorithm 2.

Algorithm 1 PLS(X, y, SubSets)

```
/* Loop over candidate subsets */
for i ← 0 to k do
  /* XX and Xy for relevant subset */
  X_i = X[:,SubSets[i,:]]
  XX_i = X_i^⊤ × X_i
  Xy_i = X_i^⊤ × y

  /* Kernel PLS using XX_i and Xy_i */
  β_i = KernelPLS(XX_i,Xy_i)
end for
```

Algorithm 1: Pseudocode explaining how a batch of feature selections conventionally would be estimated using kernel PLS. Input variables are assumed to be an $m \times n$ design matrix $\mathbf{X}$, an $m \times 1$ response vector y and a $k \times n$ Boolean matrix $SubSets$ containing k different subsets represented as $1 \times n$ vectors.

Algorithm 2 FastPLS(X, y, SubSets)

```
1: /* Calculate full covariance matrices */
2: FullXX = X^T×X
3: FullXy = X^T×y
4:
5: /* Loop over candidate subsets */
6: for i ← 0 to k do
7:   /* XX and Xy for relevant subset */
8:   XX_i = FullXX[SubSets[i,:],SubSets[i,:]]
9:   Xy_i = FullXy[SubSets[i,:]]
10:
11:  /* Kernel PLS using XX_i and Xy_i */
12:  β_i = KernelPLS(XX_i,Xy_i)
13: end for
```

Algorithm 2: Pseudocode illustrating modifications to algorithm 2 necessary to reuse covariance calculations between variable subsets to speed up the fitting of a batch of feature selections using a kernel PLS algorithm.

3.1. BENCHMARK OF FEATURE SELECTION WITH RANDOM SEARCH

To experimentally validate the supposed performance benefits that comes with reusing the full $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ calculations for all feature subsets rather than individually determining them for each variable subset, both methods (algorithm 1 & 2) were benchmarked in terms of execution time over various problem sizes. Four different $\mathbf{X}$ matrices with 100 columns and 10^4, 10^5, 10^6 and 10^7 rows, respectively, were generated and populated with pseudorandom data together with five y vectors with the same numbers of rows. Batches containing 1, 1000, 2000, 3000, 4000 and 5000 feature selections were randomly generated with a uniform distribution of active and inactive variables. Using the same input data the PLS regression between $\mathbf{X}$ and y was then performed using both the method that reuses covariance calculations (algorithm 2) and the conventional kernel PLS method (algorithm 1). For both algorithms the maximum number of PLS components was set to 15. The kernel algorithm used to perform the parameter estimation during the benchmark was the *Modified kernel algorithm #2* [6]. In appendix 1, a MATLAB implementation of this algorithm is provided with the indexing technique from algorithm 2 incorporated. The implementation of the Modified kernel algorithm #2 provided in the appendix differs slightly from Dayal and MacGregor's original algorithm [6] in the sense that we have included a stabilizing reorthogonalization (line 42, appendix 1), eliminated some redundant intermediate calculations (the lines 44-49, appendix 1) and simplified the regression coefficient calculations (line 50, appendix 1).

The runtimes for the subset sizes k ($1 \leq k \leq 5000$) that were not directly evaluated in our benchmark experiments were linearly interpolated to provide a more coherent trend line. Figure 2 shows the result of the benchmark and indicates that the performance benefits of reusing the full $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ for each feature subset grows as the number of evaluated subsets increases. When only evaluating one feature subset the modification described in section 2.2 naturally offers no performance benefits at all and is consistently slightly slower than the conventional approach. When evaluating a large number of feature subsets for a regression problem with many observations however, the technique described in section 2.2 is several orders of magnitude faster than the conventional method and peaks in our tests at a runtime decrease of roughly 5920x ($m = 10^7$, $k = 5000$).

A drawback with PLS algorithms in terms of computational efficiency is that they are inherently serial in their execution since each fitted component builds upon the previous one. Because $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ which a kernel PLS algorithm operates on are typically very small in size compared to the full $\mathbf{X}$

and y, kernel PLS algorithms require very little working memory as they run. An advantage of this is that it allows multiple kernel PLS instances to be executed in parallel across several threads such that several feature subsets are evaluated simultaneously—even though each individual algorithm runs in serial. In the supplementary material found online [LINK?], a GPU implementation of algorithm 2 written in CUDA C is available which assigns one thread—of potentially thousands available on modern GPUs—to the fitting of each of the k feature selections. The GPU implementation can also be called from MATLAB through the MEX interface. The benchmark results of this implementation are shown in green in Fig. 2. When including the CUDA implementation in the comparison, the speedup increases even more and peaks at around 7316x at $m = 10^7$ $k = 5000$ compared to algorithm 1.

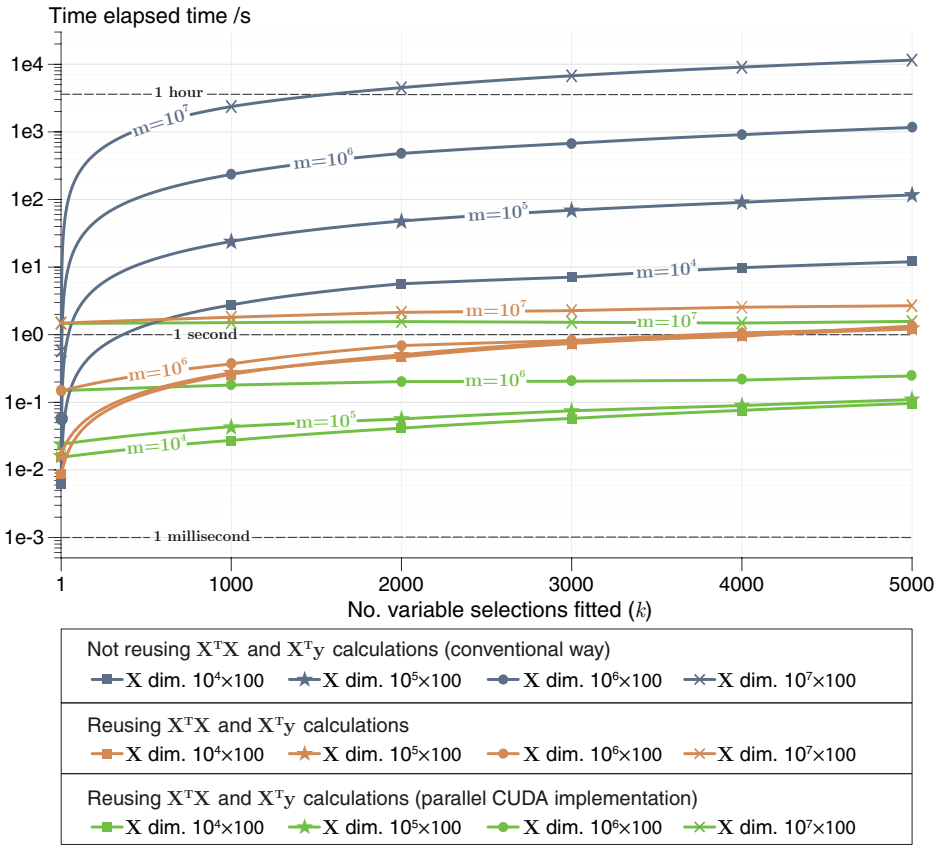


FIG. 2: Benchmark results from fitting batches of feature selections of varying sizes to random data using an improved version of the modified kernel #2 PLS algorithm with and without reusing calculations. Blue lines represents calculations performed according to algorithm 2, orange lines according to algorithm 3. CPU benchmarks were performed on an Intel i7-7700K @ 4.2 GHz. Green lines represents the parallel CUDA implementation of algorithm 3 executed on an Nvidia GTX1080ti @ 1.6 GHz. The maximum number of PLS components was set to 15 in all cases.

3.2. BENCHMARK OF COMMONLY USED FEATURE SELECTION METHODS

The results from the random search benchmark in section 3.1 provides a good overview of how the PLS calculation time scales with problem size. It does not, however, clearly convey what speedup one could expect in practice when implementing the indexing technique from section 2.2 together with commonly used non-random feature selection methods along with real data. In this section, three commonly used feature selection methods: forward selection, backward selection and a genetic algorithm, are therefore benchmarked together with a hyperspectral data set to fill this void. The data set used in the benchmark consists of six vis-NIR hyperspectral time series sequences where each sequence depicts a separate wood sample of the species Scots pine (*Pinus sylvestris*). Initially, each wood sample was submerged entirely under water and left to soak for 24 hours. After the soaking period the wood samples were taken from the water and placed one by one on a digital scale, which in turn was positioned underneath a hyperspectral camera. Over the course of roughly 21 hours, the absorbance of each wood sample was then monitored by the camera using 190 bands in the $500 - 1005$ nm region as the wood dried. In total, 843 hyperspectral images were taken and the absorbance spectra from all images are used as $\mathbf{X}$ in the data set. The digital scale placed underneath the wood samples was used to measure the weight of the wood samples as it decreased over time due to moisture evaporation. The sample weight was then recalculated into an average moisture content of each wood sample for each point in time. The time dependent moisture content is used as the response (y) in the data set. For more information on the data set the reader is referred to [11]. When performing regression on hyperspectral data, each pixel of the involved hyperspectral images can be viewed as a unique observation. When arranging such a data set into a two-dimensional design matrix, the number of rows (m) corresponds to the total number of pixels in all images—which can easily add up to several million or billion in number. To make the data set easier to work with, the spatial resolution of the original hyperspectral images can be lowered by averaging together neighboring pixels, resulting in an $\mathbf{X}$ matrix with any desired number of rows. During the benchmark in this section the spectral resolution (n) of the data set was kept constant at 190 bands, while the spatial resolution (m) of the design matrix was down-sampled to $0.5\text{e}3$, $1.0\text{e}3$, $0.5\text{e}4$, $1.0\text{e}4$, $0.5\text{e}5$, $1.0\text{e}5$, $0.5\text{e}6$ and $1.0\text{e}6$ respectively. In each feature selection algorithm a 10-fold cross-validation was performed and the cross-validated root mean square error, $RMSE_{cv}$, was used to drive the search. In the forward and backward selection benchmark, the selection process was terminated as soon as an iteration caused the $RMSE_{cv}$ to increase. The genetic algorithm used a population size of 200 and ran for 200 generations before terminating. In all benchmarks the maximum number of considered PLS components was set to 15. The results from the three benchmarks are shown in Fig.3. As can be seen Fig.3, the calculation time required to perform backward selection and genetic algorithm was greatly decreased by the use of the indexing technique from section 2.2, while forward selection benefited substantially less from the indexing technique. Table 1 summarizes the average and maximum observed speedup of algorithm 2 compared to algorithm 1 for each of the feature selection methods in the benchmark. As demonstrated by the random search benchmark, the speedup is directly related to the number of subsets being evaluated, which differs greatly between the feature selection methods. In the case of the genetic algorithm, which terminated after a fixed number of generations, the number of subsets to evaluate is deterministic and often high compared to the other two methods, which is why it is natural for the GA to benefit a lot from the suggested method. In the case of forward- and backward selection, the number of subsets evaluated throughout the feature selection process depends on when the termination criterion is triggered, which in turn is data set specific. In the present example, backwards selection was sped up a great deal more by the suggested indexing technique than forward selection, it should be noted however, that this pattern could well be reversed for data sets that converges on a large number of active features. Furthermore, when performing calculations on a GPU there is always an overhead accosted with transferring data onto and from the device. When the computational workload is low, such as during the initial stages of a forward selection algorithm, the cost of transferring data to and from the GPU is too high to be completley amortized away by the offered parallelization of the PLS computation. This is why the CPU implementation of algorithm 2 can be seen to outperform the GPU implementation of algorithm 2 under some circumstances in the benchmark.

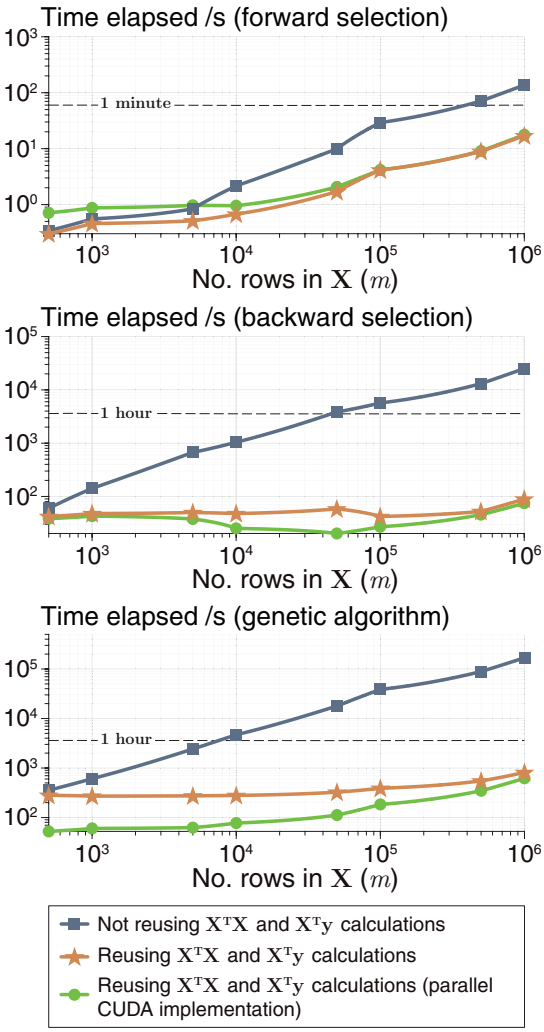


FIG. 3: Benchmark results from performing feature selection with kernel PLS on $\mathbf{X}$ matrices of a varying number of rows (m) using three commonly used feature selection algorithms; forward selection (upper), backward selection (middle) and genetic algorithm (lower). Blue lines represents feature selection performed using algorithm 2, orange lines according to algorithm 3. CPU benchmarks were performed on an Intel i5-6300HQ @ 2.3 GHz. Green lines represents the parallel CUDA implementation of algorithm 3 excecuted on an Nvidia GTX1080ti @ 1.6 GHz. The benchmark results only include the time elapsed when fitting regression coefficients $\hat{\beta}$ with kernel PLS, the additional time required to compute $RMSE_{cv}$ is not included since it is unrelated to the choice of PLS algorithm and identical in the three cases.

TABLE I: Average and maximum observed speedup of algorithm 2 compared to algorithm 1 for the three benchmarked feature selection methods.

Feature selection method	**Avg. speedup**	**Max. speedup**
Forward selection (CPU)	5x	8x
Forward selection (GPU)	4x	8x
Backward selection (CPU)	96x	281x
Backward selection (GPU)	136x	335x
Genetic algorithm (CPU)	69x	207x
Genetic algorithm (GPU)	127x	271x

4. CONCLUSIONS AND DISCUSSION

Many heuristic feature selection algorithms are largely driven by trial and error, because of this they tend to be rather time consuming - which creates a demand for computationally fast PLS fitting procedures such as kernel PLS. By taking advantage of a new simple indexing technique, the computational cost of fitting multiple variable subsets of the same data with PLS regression is substantially reduced. In cases were the design matrix $\mathbf{X}$ consists of a far greater number of observations than variables, we have demonstrated that the proposed technique offers a speedup of several orders of magnitude compared to the conventional approach when evaluating regression models for a large number of different variable subsets. The speedup is achieved by performing the computationally expensive covariance matrix calculations $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top y$ only once using the complete set of variables within $\mathbf{X}$ and then reusing the already calculated results for all subsequent feature subsets, rather can recalculating the covariances for each individual subset. It should be emphasized that the performance benefits of this technique becomes greater the larger the number of evaluated feature subsets becomes. In the special case of considering only one feature subset, the method offers no improvements at all since there are no calculations to reuse.

Because the success of many heuristic variable selection algorithms depend on the ability to explore a search space by evaluating the performance of a large number of subsets, the technique introduced here has the potential of improving essentially all wrapper-based feature selection methods by enabling more feature subsets to be evaluated per unit of time than previously possible. The kernel PLS algorithms are, however, not among the most numerically stable PLS alternatives. It is therefore recommended to recalibrate the most promising feature combination(s) by using a numerically more stable PLS algorithm, such as bidiag2 [10], before carefully evaluating, choosing and deploying the final model. It should also be mentioned that there are other flavors of PLS, such as sparse partial least squares regression (SPLS) which circumvents the need for conventional feature selection by producing sparse linear combinations of the original features within the algorithm [12]. Although SPLS requires the optimization of additional built-in parameters, it may in some cases—such as when the computational cost is not of critical importance or when the data set is small—be worthwhile to consider as a viable alternative approach.

Lastly it should also be mentioned that the indexing technique introduced in this paper is not limited to partial least squares regression. Indeed, in cases where solving an ordinary least squares regression problem through the normal equations is numerically acceptable, the indexing technique introduced in this paper is trivial to implement together with OLS and offers speedups on par with the ones demonstrated for PLS in section 3.

REFERENCES

[1] Y. Saeys, I. Inza and P. Larranaga, *A review of feature selection techniques in bioinformatics*. Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[2] T. Mehmood, K. Liland, L. Snipen and S. Sæbø, *A review of variable selection methods in Partial Least Squares Regression*. Chemometrics and Intelligent Laboratory Systems, vol. 118, pp. 62-69, 2012.

[3] Z. Xiaobo, Z. Jiewen, M. Povey, M. Holmes and M. Hanpin, *Variables selection methods in near-infrared spectroscopy*. Analytica Chimica Acta, vol. 667, no. 1-2, pp. 14-32, 2010.

[4] M. Anzanello and F. Fogliatto, *A review of recent variable selection methods in industrial and chemometrics applications*. European J. of Industrial Engineering, vol. 8, no. 5, p. 619, 2014.

[5] F. Lindgren, P. Geladi and S. Wold, *The kernel algorithm for PLS*. Journal of Chemometrics, vol. 7, no. 1, pp. 45-59, 1993.

[6] B. Dayal and J. MacGregor, *Improved PLS algorithms*. Journal of Chemometrics, vol. 11, no. 1, pp. 73-85, 1997.

[7] D. Kepplinger, P. Filzmoser, K. Varmuza, *Variable selection with genetic algorithms using repeated cross-validation of PLS regression models as fitness measure*. https://arxiv.org/pdf/1711.06695.pdf.

[8] N. Kettaneh, A. Berglund and S. Wold, *PCA and PLS with very large data sets*. Computational Statistics & Data Analysis, vol. 48, no. 1, pp. 69-85, 2005.

[9] S. Rännar, F. Lindgren, P. Geladi and S. Wold, *A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm*. Journal of Chemometrics, vol. 8, no. 2, pp. 111-125, 1994.

[10] Å. Björck and U. Indahl, *Fast and stable partial least squares modelling: A benchmark study with theoretical comments*. Journal of Chemometrics, vol. 31, no. 8, p. e2898, 2017.

[11] P. Stefansson, J. Fortuna, H. Rahmati, I. Burud, T. Konevskikh and H. Martens, *Hyperspectral time series analysis: Hyperspectral image data streams interpreted by modeling known and unknown variations*. Hyperspectral imaging. Analysis and applications. , 1st ed., 2019. [IN PRESS]

[12] H. Chun and S. Keleş, *Sparse partial least squares regression for simultaneous dimension reduction and variable selection*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 72, no. 1, p. 3-25, 2010.

Appendix 1: PLS coefficient estimation of a batch of feature selections with reused covariance

```
function [ Beta ] = PLSvarsel(X, y, A, VarSels)
% Filename: PLSvarsel.m
% Description: Matlab function which estimates regression coefficients for a batch of variable
% selections using the PLS algorithm 'Modified kernel algorithm #2'. For reference regarding
% the fitting sequence see: Improved PLS algorithms, Journal of chemometrics Vol. 11 p 73–85.
% Inputs:
%        1. a [m–by–n] double–precision design matrix X.
%        2. a [m–by–1] double–precision response vector y.
%        3. a [1–by–1] double–precision scalar, A, specifying maximum PLS components.
%        4. a [k–by–n] logical matrix with k variable selections.
% Outputs:
%        1. a [n–by–A–by–k] array, Beta, with fitted coefficients for all feature selections.
%                 Inactive variables are given a coefficient value of 0.
% Syntax:
%        Beta = PLSvarsel(X,y,A,VarSels);
%
% Written  2017–10–04 by Petter Stefansson.
% Modified 2018–03–01 by Ulf Indahl.
%% ––––––––––––––––––––––– Calculate full covariance matrices X'X and X'y ––––––––––––––––––––
XX = X'*X;
Xy = X'*y;
% Memory allocation for Beta.
k = size(VarSels,1); n = size(X,2);
Beta = zeros(n,A,k);

% Loop over all variable selections.
for v = 1 : k
    %% ––––– Index into XX and Xy using a variable selection to acquire new covariances –––––
    smallXX = XX(VarSels(v,:),VarSels(v,:));
    smallXy = Xy(VarSels(v,:));
    smalln  = size(smallXX,1);
    % Ensure number of PLS components <= number of variables.
    if A > smalln;  MaxComps = smalln;  else;  MaxComps = A;  end

    %% ––––––––––– PLS on extracted covariances using Modified Kernel#2 algorithm ––––––––––
    % Memory allocation for matrices W, P, R and vector b.
    W = nan(smalln,MaxComps);    P = nan(smalln,MaxComps);
    R = nan(smalln,MaxComps);    b = zeros(smalln,1);

    % PLS Component loop.
    for i = 1 : MaxComps
        w = smallXy – W(:,1:i–1)*(W(:,1:i–1)'*smallXy);
        w = w/sqrt(w'*w);
        r = w – R(:,1:i–1)*(P(:,1:i–1)'*w);
        smallXXr = smallXX*r;
```

```
46          tt = r'*smallXXr;
47          p = smallXXr/tt;
48          q = (r'*smallXy)/tt;
49          smallXy = smallXy - smallXXr*q;
50          b = b + r*q;
51          W(:,i) = w;
52          R(:,i) = r;
53          P(:,i) = p;
54          Beta(VarSels(v,:),i,v) = b;
55      end
56  end
```

Paper II

FAST METHOD FOR GA-PLS WITH SIMULTANEOUS FEATURE SELECTION AND IDENTIFICATION OF OPTIMAL PREPROCESSING TECHNIQUE FOR DATASETS WITH MANY OBSERVATIONS

UNDER REVIEW

Petter Stefansson
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Kristian H. Liland
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Thomas K. Thiis
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Ingunn Burud
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

ABSTRACT

A fast and memory-efficient new method for performing genetic algorithm partial least squares (GA-PLS) on spectroscopic data preprocessed in multiple different ways is presented. The method, which is primarily intended for datasets containing many observations, involves preprocessing a spectral dataset with several different techniques and concatenating the different versions of the data horizontally into a design matrix $\boldsymbol{X}$ which is both tall and wide. The large matrix is then condensed into a substantially smaller covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}$ whose resulting size is unrelated to the number of observations in the dataset, i.e. the height of $\boldsymbol{X}$. It is demonstrated that the smaller covariance matrix can be used to efficiently calibrate partial least squares (PLS) models containing feature selections from any of the involved preprocessing techniques. The method is incorporated into GA-PLS and used to evolve variable selections for a set of different preprocessing techniques concurrently within a single algorithm. This allows a single instance of GA-PLS

to determine which preprocessing technique, within the set of considered methods, is best suited for the spectroscopic dataset. Additionally, the method allows feature selections to be evolved containing variables from a mixture of different preprocessing techniques. The benefits of the introduced GA-PLS technique can be summarized as threefold: (1) for datasets with many observations, the proposed method is substantially faster compared to conventional GA-PLS implementations based on NIPALS, SIMPLS, etc. (2) using a single GA-PLS automatically reveals which of the considered preprocessing techniques results in the lowest model error. (3) it allows the exploration of highly complex solutions composed of features preprocessed using various techniques.

Keywords: GA-PLS, Genetic algorithms, Partial least squares, Feature selection, Preprocessing

1 Introduction

Multivariate calibration on spectroscopic near-infrared (NIR) data typically involves overcoming two main challenges: choice of spectral preprocessing method and selection of relevant wavelengths to include in the model. The goal of spectral preprocessing is to remove unwanted noise from the measured spectral signal (e.g. light scattering due to physical phenomena), to emphasize features related to the specific chemometric problem at hand and to improve the comparability between samples [1]. The purpose of wavelength selection is to limit the spectral region used by the model to only areas relevant for modeling the analyte of interest; thus enabling faster and more accurate predictions whilst in addition providing a better understanding of the underlying process that generated the data [2]. In many real-world circumstances, a priori information regarding what constitutes a suitable wavelength selection and/or preprocessing technique is not available. To address the challenge of wavelength selection, a myriad of feature selection algorithms can be found suggested in chemometric literature such as *simulated annealing*, *step-wise methods*, *interval PLS* (iPLS) and *genetic algorithms* (GA) [3, 4, 5]. In the realm of NIR spectroscopy, genetic algorithms have been extensively used for several decades and proven to be a well suited strategy for the selection of relevant wavelengths when used together with partial least squares (PLS) regression [6, 7, 8, 9] to form a particular niche of GA commonly abbreviated as *GA-PLS*. Genetic algorithms operate on a *population* of potential solutions to an optimization task; a population which is improved upon over consecutive generations in an attempt to mimic natural evolution. Using a population of candidate solutions to an optimization task has several advantages. For instance, it reduces the risk of the search getting stuck in a local optimum and allows for a better exploration of the solution space [10]. One of the main drawbacks of GA-PLS is that the computational cost of evaluating the performance of each candidate solution in the population is often high—making the algorithm highly time consuming. In the interest of saving time it is therefore tempting to use a small population size when running a GA, it has however been demonstrated [11] that using a small population size can greatly lower the success rate of genetic algorithms. If one wishes to explore how well various different preprocessing methods perform together with GA-PLS, the computational cost of doing so further increases linearly with every preprocessing technique added to the search.

Recently, a new method for calibrating multiple PLS models using different variable subsets of the same dataset was discovered [12] which greatly reduces the computational cost of performing PLS feature selection on strongly overdetermined, i.e. 'long and thin', regression problems. The computational speedup of the method introduced by Stefansson et al. [12] is accomplished by deriving the covariance matrix of any variable subset directly from a covariance matrix calculated using the full set of variables in the dataset; thereby bypassing the need to calculate the covariance for each individual subset. The derived covariance matrices for the various feature subsets are then used to estimate PLS regression coefficients using kernel PLS at a low computational cost. In this study, the

method introduced in [12] will be used in the context of GA-PLS and expanded upon to demonstrate that the method not only is capable of reducing the run-time of a conventional GA-PLS, but can also be used to efficiently incorporate multiple different preprocessing techniques within a single population of a GA-PLS. The candidate solutions of the population can then either be restricted to evolve solely within their own preprocessing-specific region of the solution space, which produces the equivalent effect of running multiple genetic algorithms in parallel. Or, the search space region in which candidate solutions are permitted to operate in can be unrestricted, thereby enabling models to evolve containing features from a mixture of different preprocessing techniques.

To experimentally demonstrate that our suggested methodology works in practice, the proposed concept will be applied to two hyperspectral datasets containing NIR and vis-NIR spectra and used to perform simultaneous feature selection and identification of optimal preprocessing method. To quantify the computational efficiency gains offered by the method, the introduced technique will be benchmarked against other GA-PLS implementations using well-established PLS algorithms (SIMPLS, NIPALS, Bidiag2 and modified kernel algorithm #2).

The remaining part of this paper is organized as follows: in section 2.1 a brief overview of how conventional GA-PLS algorithms work is given, which will serve as a primer leading up to the new modified/extended GA-PLS concept outlined in section 2.2. In this section it is described how a GA-PLS can be implemented in order to efficiently operate on multiple preprocessing methods simultaneously using the technique from [12]. In section 3 a summary of the preprocessing methods and datasets used to evaluate the proposed algorithm is given as well as a description of the computational efficiency benchmarks performed. The results of which are presented in section 4.

2 Method

2.1 Overview of GA-PLS

The conventional way of representing a candidate solution to a feature selection problem using GA-PLS is by a Boolean vector of the same length as the total number of variables in the dataset. Such vectors are generally referred to as *chromosomes*. Each Boolean value of the vector corresponds to a variable being either excluded from the feature selection, 0, or included, 1. Many different chromosomes placed together in a Boolean matrix constitutes a population. The Boolean values of each solution are typically initialized randomly at the start of the algorithm. The feature subset specified by each chromosome of the population is then used to calibrate one or more PLS regression models on the dataset. The resulting regression coefficients are used to model the target response of the dataset, and the error associated with the model's predictions are reduced into a statistical measure of the feature subset's performance, such as the cross-validated root-mean-squared error, RMSE_{cv}. The cross-validated performance measure of each chromosome is then converted into a *fitness* value, which is a scalar value that communicates the quality of an encoded solution to the GA and defines what improvement means [13]. How well a chromosome translates into a fitness value, i.e. how well it solves the optimization task compared to all the other candidate solutions of the population, determines how it gets treated by the GA-PLS in the steps to come, where fitter individuals are favored compared to individuals with a lower fitness. The bias within the generational cycle of a genetic algorithm towards favoring solutions with high fitness is typically realized using two stages of selection: *parent selection* and *survivor selection* [13]. In the parent selection stage it is decided which individuals are used to generate new offspring solutions which will drive evolution forward. In the survivor selection stage, all individuals of the population compete for survival into the next generation. If the fitness value of a chromosome is low compared to the rest of the population, it is more likely to be discarded from the population. In the interest of conciseness the mechanics

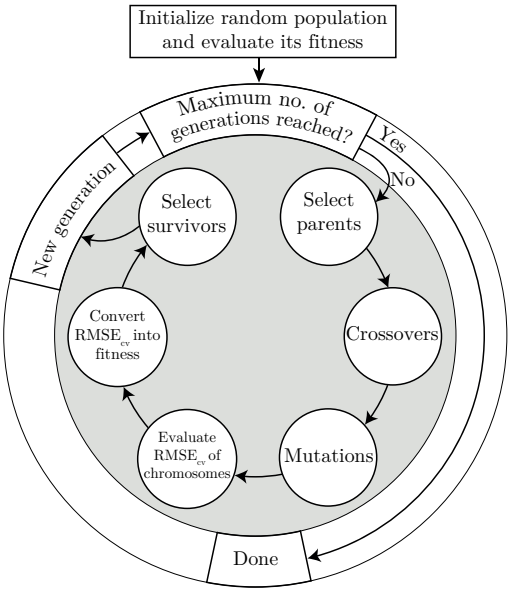


Figure 1: Flowchart of typical GA-PLS.

of the different selection procedures will not be explained here, for an excellent overview of some commonly used selection operators the reader is referred to [13].

In addition to favoring individuals already known to have a high fitness through the selection operators, genetic algorithms have the ability to generate new–potentially improved–solutions to an optimization task by combining and/or refining already successful chromosomes to form new ones. This is achieved using a set of stochastic variation operators: *crossover* and *mutation*. The crossover operator takes a pair of parent chromosomes as input and produces one or several offspring chromosomes as output, analogously to how genes from two parents are combined to generate children in nature. An example of a frequently used crossover operator in GA-PLS is *2-point crossover,* which is performed by randomly selecting two points along two parent chromosomes and letting the resulting offspring chromosomes alternately inherit sections from their parents between the crossover points. The second variation operator, the mutation, is a unary process that takes a single chromosome as input and outputs a slightly altered version of the same chromosome which replaces the original in the population. In GA-PLS, *bitwise mutation* is commonly used. In bitwise mutation the Boolean value of each element within a chromosome is flipped–thereby turning a value of 0 into 1 and vice versa–according to a mutation probability p_m. Typically p_m is conservatively set to a small number since a too high value could have a destructive impact on the evolutionary progress [14]. Placing all the mentioned components together in the order illustrated by the flowchart in Fig. 1 forms a conventional PLS-GA.

2.2 Extending/modifying GA-PLS to efficiently operate on multiple versions of the same data with kernel PLS

Certain kernel PLS algorithms, such as the modified kernel algorithm #2 [15], base their regression coefficient calculations on the covariance matrices of the dataset, $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}^\top \boldsymbol{y}$, rather than the $\boldsymbol{X}$ and $\boldsymbol{y}$ data itself. For strongly overdetermined systems, kernel algorithms can vastly reduce the computational time required to do PLS compared to traditional well-established algorithms such

as NIPALS or SIMPLS [16]. Calculating the covariance is however still a costly operation for large $\boldsymbol{X}$ matrices. In a previous paper, Stefansson et al. [12] showed that when performing subset selection, the computationally costly $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{y}$ covariance calculations required for kernel PLS only needs to be established once using the full set of variables within $\boldsymbol{X}$. The covariance matrix that any combinatorial column subset of $\boldsymbol{X}$ would have resulted in, had it been separately calculated, can then be retrieved from the full covariance matrix without further computations simply by indexing into it. The greater the number of evaluated feature subsets become, the larger the efficiency gains of this technique becomes. When evaluating the performance of an entire population in GA-PLS, this technique enables all the individuals of a population to share the same $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{y}$ covariance matrices as a basis for their regression coefficient estimations. Sharing covariance matrices between multiple feature subsets has shown to greatly decrease the computational time required to estimate PLS regression coefficients under many circumstances [12]. Additionally, the full covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}$ will be of dimension $n \times n$, where n denotes the number of columns in $\mathbf{X}$. For strongly overdetermined systems where the number of rows (m) in $\boldsymbol{X}$ is substantially greater than the number of columns ($m \gg n$), this means that the memory required to store $\boldsymbol{X}^\top\boldsymbol{X}$ is greatly reduced compared to storing $\boldsymbol{X}$ itself. This in turn makes the kernel PLS-based indexing technique highly amendable to parallel GPU implementations, where the available memory is often relatively low. In order to derive the covariance matrix of a feature subset from a full $\boldsymbol{X}^\top\boldsymbol{X}$ matrix, a Boolean chromosome of length n can simply be aligned along both the rows and columns of the full $n \times n$ $\boldsymbol{X}^\top\boldsymbol{X}$ matrix, whereupon only elements intersecting with 1 in both dimensions are extracted. The submatrix one receives from this extraction will then be identical to the covariance matrix that would be obtained had the covariance matrix been calculated using only the column subset in $\boldsymbol{X}$ specified by the same chromosome. The same concept also holds true for $\boldsymbol{X}^\top\boldsymbol{y}$, except only one dimension is involved when extracting elements intersecting with 1. This indexing concept is visually exemplified using two small chromosomes and a dummy $\boldsymbol{X}^\top\boldsymbol{X}$ matrix in Fig. 2. For a more thorough description of the indexing technique along with a MATLAB implementation of it, the reader is referred to [12].

Considering that evaluating the fitness of a GA-PLS population tends to be the overwhelmingly most time-consuming step within the evolutionary cycle of the algorithm, and that forming the covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}$ in turn tends to be the most computationally costly part of a kernel PLS algorithm, the indexing technique illustrated in Fig. 2 can alone substantially increase the efficiency of a GA-PLS [12]. However, since it has been shown that an $\boldsymbol{X}^\top\boldsymbol{X}$ matrix can be indexed into using the simple technique illustrated in Fig. 2 in order to obtain the covariance matrix related to any subset of columns in $\boldsymbol{X}$, it naturally follows that if $\boldsymbol{X}$ is widened by adding additional columns to it, any covariance matrix containing these added features can also be retroactively retrieved from the resulting $\boldsymbol{X}^\top\boldsymbol{X}$ matrix. By treating a spectroscopic data matrix $\boldsymbol{X}$ with multiple different preprocessing techniques and horizontally concatenating these matrices into one wide matrix:

$$\boldsymbol{X}_{\text{Tot}} = [\boldsymbol{X} \;\; f_1(\boldsymbol{X}) \;\; f_2(\boldsymbol{X}) \;\; f_3(\boldsymbol{X})...],$$

it therefore also follows that the covariance matrix for any of the individual constituent matrices within $\boldsymbol{X}_{\text{Tot}}$, i.e. $\boldsymbol{X}^\top\boldsymbol{X}$, $f_1(\boldsymbol{X})^\top f_1(\boldsymbol{X})$, $f_2(\boldsymbol{X})^\top f_2(\boldsymbol{X})$, $f_3(\boldsymbol{X})^\top f_3(\boldsymbol{X})$, ..., can be retroactively retrieved from the matrix multiplication $\boldsymbol{X}_{\text{Tot}}^\top\boldsymbol{X}_{\text{Tot}}$. After all, horizontally concatenating the different preprocessed versions of the data is effectively a widening of $\boldsymbol{X}$ by adding additional columns to it. In fact, the matrix multiplication $\boldsymbol{X}_{\text{Tot}}^\top\boldsymbol{X}_{\text{Tot}}$ will result in a patchwork of all the possible covariance matrix combinations that can be generated from all the concatenated matrices within $\boldsymbol{X}_{\text{Tot}}$. Fig. 3 shows the resulting $\boldsymbol{X}_{\text{Tot}}^\top\boldsymbol{X}_{\text{Tot}}$ matrix calculated for the first dataset in the study (a description of the dataset is given in section 3.1) using the preprocessing techniques: $\text{D}^1\boldsymbol{X}$, $\text{D}^2\boldsymbol{X}$, $\text{MSC}(\boldsymbol{X})$, $\text{MSC}(\text{D}^1\boldsymbol{X})$ and $\text{MSC}(\text{D}^2\boldsymbol{X})$ concatenated together when forming $\boldsymbol{X}_{\text{Tot}}$. The notation used to abbreviate preprocessing techniques is clarified in section 3.2. As indicated by the arrows in Fig. 3, the pure covariance matrices for all of the differently preprocessed versions of $\boldsymbol{X}$ can be found as intact submatrices along the diagonal of $\boldsymbol{X}_{\text{Tot}}^\top\boldsymbol{X}_{\text{Tot}}$. Because of this emerging structure, a

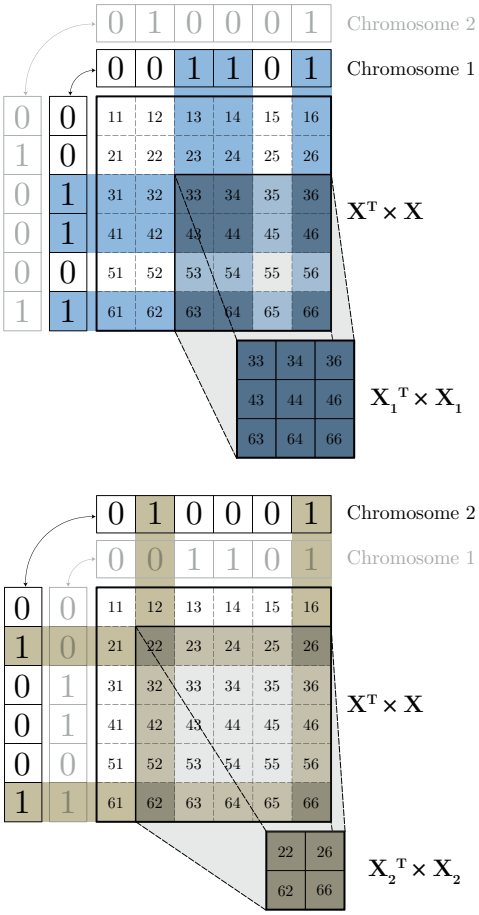


Figure 2: Example of the concept behind the indexing technique needed to extract chromosome-specific subset covariance matrices ($\boldsymbol{X}_1^\top \boldsymbol{X}_1$, $\boldsymbol{X}_2^\top \boldsymbol{X}_2$) from a covariance matrix formed using the full set of available features in $\boldsymbol{X}$. All chromosomes in the population share an underlying $\boldsymbol{X}^\top \boldsymbol{X}$ matrix from which they extract their own subset covariance matrix by retrieving elements intersecting with 1 along both the rows and columns. The extracted subset covariance matrices can then be used to calibrate PLS models needed for GA-PLS using kernel PLS.

single $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix can be used in combination with the indexing technique shown in Fig. 2 to calibrate PLS models containing any combinatorial feature subset of any of the involved preprocessing techniques—or indeed even models containing a mixture of features preprocessed with different methods. As a consequence of this, multiple subpopulations can be positioned within a larger GA-PLS population and be efficiently calibrated in parallel with each other since the data used during PLS can be shared between all chromosomes, regardless of spectral pretreatment method. In order to extract a feature selection corresponding to a specific preprocessing technique from the large $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ matrix, a small alteration in the chromosome encoding is required. For the genes of a chromosome of length n to map onto the i^{th} $n \times n$ covariance matrix along the diagonal of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$, the chromosome has to be offset by $(i-1) \times n$ elements, where n denotes the length of the spectra in $\boldsymbol{X}$. Equivalently, one could also pad the chromosome with zeros in all locations corresponding to features that are outside of the preprocessing-specific region the chromosome should operate in. The concept of padding chromosomes with zeros in order to form preprocessing-specific subpopulations is illustrated in Fig. 4.

During the selection, mutation and crossover stages of the GA-PLS, one has the choice of either restricting the subpopulations to evolve exclusively within a preprocessing-specific region, or to allow the individuals to combine with each other free of restrictions across the entire solution space to form complex feature selections containing a mixture of preprocessing methods. When only restricted, preprocessing-specific, subpopulations are used, only the submatrices along the diagonal of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ will be referenced during the kernel PLS. In such cases, all off-diagonal submatrices in $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ do not need to be calculated. When including subpopulations permitted to evolve feature subsets containing variables from a mixture of preprocessing techniques however, all regions of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ will potentially be referenced.

By including or excluding observations (rows) of $\boldsymbol{X}_{\text{Tot}}$ according to a desired cross-validation scheme, several different versions of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ can be precomputed and stored in a three-dimensional cache, as can be seen illustrated in Fig. 5, prior to commencing the GA-PLS. During the RMSE_{cv} evaluation stage of the GA-PLS, different versions of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ can then be retrieved from the cache and used for calibrating and cross-validating the individuals of a population. This eliminates the need of performing any costly covariance calculations at all within the GA-PLS generational loop.

The full set of steps required to implement the GA-PLS technique described in this section is given as pseudocode in Algorithm 1. In the case where subpopulations are restricted to be confined within their own preprocessing-specific region of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$, the selection and mutation stages of the GA-PLS are carried out in serial, whilst the RMSE_{cv} evaluations can be performed simultaneously in parallel for all subpopulations by sharing a common $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$.

3 Experimental

In order to experimentally verify the viability of the GA-PLS technique described in section 2.2 (Algorithm 1), an implementation of the algorithm was used to perform simultaneous feature selection and identification of optimal preprocessing technique on two hyperspectral datasets. Hyperspectral imaging data is a particularly suitable type of spectroscopic data to use with the kernel PLS based modeling procedure proposed in [12] since each pixel of the hyperspectral images are treated as a new observation—which results in inherently overdetermined datasets with substantially more rows than columns. To evaluate the computational efficiency of the proposed GA-PLS method, a benchmark was conducted where the time required for performing GA-PLS on one of the hyperspectral datasets using Algorithm 1 was compared to GA-PLS implementations utilizing more well-known PLS fitting procedures.

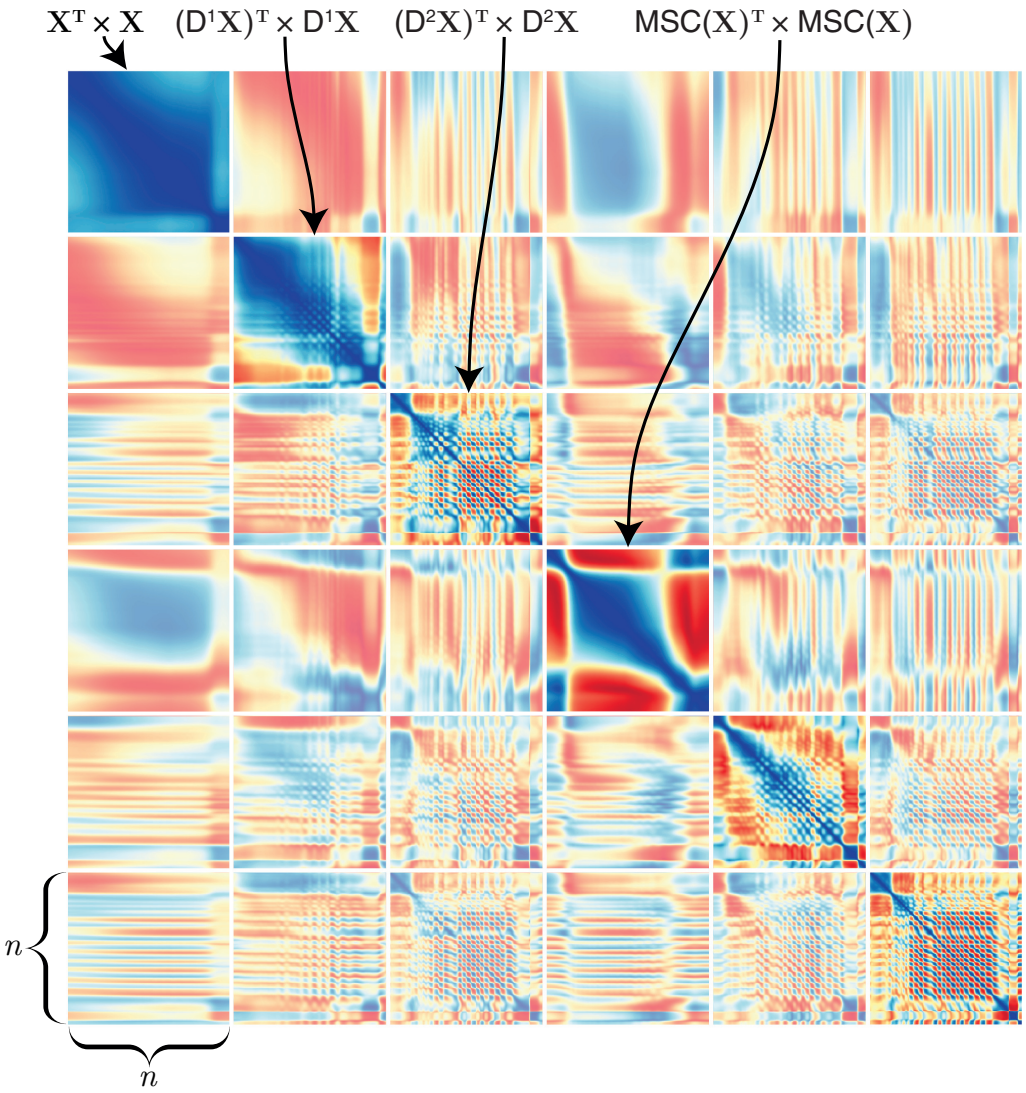


Figure 3: Illustration of the resulting grid-like landscape of covariance matrices formed when multiplying $\boldsymbol{X}_{\text{Tot}}^{\top}$ with $\boldsymbol{X}_{\text{Tot}}$. The pure covariance matrices of each individual preprocessing method contained in $\boldsymbol{X}_{\text{Tot}}$ can be found as submatrices along the diagonal of the data structure. By applying the indexing technique introduced by Stefansson et al. [12] the covariance matrix required to solve the PLS regression for any of the involved preprocessing techniques can easily be derived from the full grid of covariances–as well as any combinatorial subset of features within the preprocessing-specific covariance matrices. Small white spaces have been added between the submatrices in the figure to emphazise the grid structure.

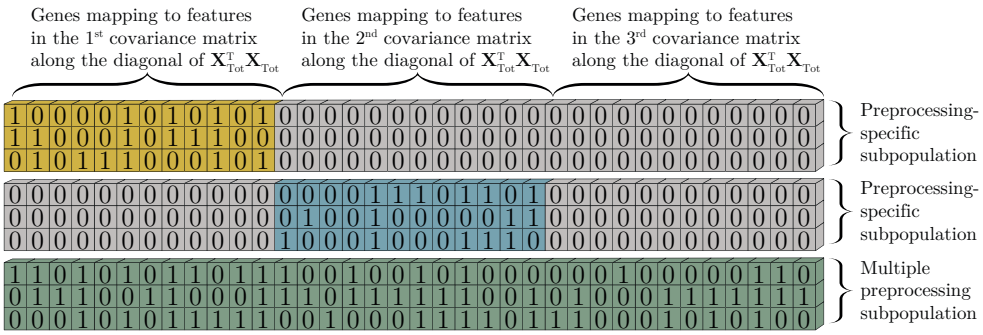


Figure 4: Illustration of padding needed to align a preprocessing-specific subpopulation to a designated area of a large covariance matrix containing multiple preprocessed versions of $\boldsymbol{X}$, such as the one shown in Fig. 3. Row 1-6 illustrate chromosomes of length n padded with zeros in order to span the full width of $\boldsymbol{X}_{\text{Tot}}^{\top}\boldsymbol{X}_{\text{Tot}}$ (here $3 \times n$). Row 7-9 illustrates a subpopulation operating on multiple preprocessing methods.

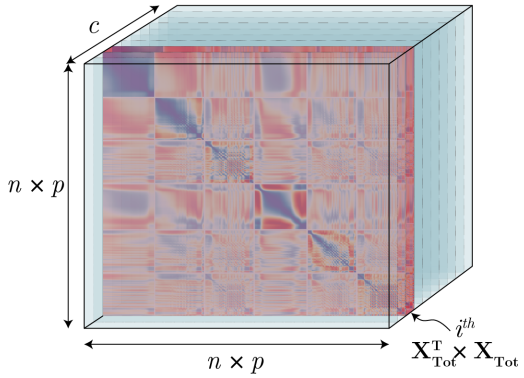


Figure 5: Illustration of data structure used for storing a cache of precomputed covariance matrices needed for calibrating and cross-validating PLS models for any feature subset of any of the involved preprocessing methods. Each two-dimensional slice contains a patchwork of $p \times p$ different submatrices of dimension $n \times n$, where p represents the number of involved preprocessing techniques and n represents the length of a spectrum. For simplicity all spectra are assumed to be of length n regardless of preprocessing method in the illustration, in practice however, this may not necessarily be the case.

Algorithm 1 Fast GA-PLS for multiple simultaneous preprocessing methods

Preprocess $\mathbf{X}$ with p different methods and horizontally concatenate into $\mathbf{X}_{\text{Tot}}$
Iteratively include/exclude rows of $\mathbf{X}_{\text{Tot}}$ and y according to a cross-validation scheme and calculate caches of c different $\mathbf{X}_{\text{Tot}}^{\top}\mathbf{X}_{\text{Tot}}$ and $\mathbf{X}_{\text{Tot}}^{\top}y$ matrices
Initialize a population containing p preprocessing-specific Boolean subpopulations
Sample k $\mathbf{X}_{\text{Tot}}^{\top}\mathbf{X}_{\text{Tot}}$ and $\mathbf{X}_{\text{Tot}}^{\top}y$ matrices from the caches and use to batch-evaluate the RMSE_{cv} of the entire population using kernel PLS with the indexing technique from section 2.2. Convert RMSE_{cv} into fitness
for $gen \leftarrow 1$ to $maxgen$ **do**
    for $s \leftarrow 1$ to p **do**
        Select parents in subpopulation $\#s$
        Generate offspring using selected parents
        Merge the offspring with subpopulation $\#s$
        Mutate subpopulation $\#s$
    end for
    Sample k $\mathbf{X}_{\text{Tot}}^{\top}\mathbf{X}_{\text{Tot}}$ and $\mathbf{X}_{\text{Tot}}^{\top}y$ matrices from the caches and use to batch-evaluate the RMSE_{cv} of the entire population using kernel PLS with the indexing technique from section 2.2. Convert RMSE_{cv} into fitness
    for $s \leftarrow 1$ to p **do**
        Select survivors for subpopulation $\#s$
    end for
end for

Algorithm 1: Pseudocode of steps needed to efficiently perform GA-PLS on multiple preprocessed version of a dataset. Input data is assumed to be a design matrix $\boldsymbol{X}$, a response vector $\boldsymbol{y}$, a scalar parameter p corresponding to the number of preprocessing methods included in the algorithm, a scalar parameter c indicating the cache size to use, a scalar parameter k ($k \leq c$) describing the number of folds to use during k-fold cross-validation and a scalar parameter $maxgen$ controlling the number of generations to run the GA-PLS.

3.1 Description of the two datasets used

1) The first hyperspectral dataset contains absorbance spectra from 40 $50 \times 50 \times 10$ mm sized wood samples of the species Norway spruce (*Pincea abies*) treated with various concentrations of a phosphorous-based flame retardant compound. The hyperspectral images were acquired using a push broom camera (Specim, Oulu, Finland) with 256 bands in the 929-2531 nm wavelength range. After the signal acquisition, a 0.4-0.5 mm layer was removed from the surface of each wood sample and sent to an inductively coupled plasma (ICP) analysis where the average phosphorus content of each sample was determined. The laboratory established average phosphorus content ($g \cdot kg^{-1}$) associated with each wood sample is the response variable $\boldsymbol{y}$ of the dataset. The size of the design matrix $\boldsymbol{X}$ containing all the absorbance data is $10^5 \times 256$. For a more comprehensive description of this dataset the reader is referred to [17].

2) The second dataset consists of six hyperspectral time series sequences. Each sequence contains 150 images depicting a separate $280 \times 100 \times 18$ mm wood sample of the species Scots pine (*Pinus sylvestris*); resulting in a dataset of 6×150 images. Initially, all six wood samples were submerged under water and left to soak for 24 hours. When taken from the water after the soaking period the samples were placed on a digital scale, which in turn was positioned underneath a hyperspectral push broom camera (Specim, Oulu, Finland) measuring 200 bands in the 392-1022 nm wavelength region. During the course of 21 hours, hyperspectral images were repeatedly taken of each sample every eight minute as they dried. The digital scale which the wood samples were positioned on top of simultaneously provided an estimate of the average moisture content of the sample for each of the 150 images taken during the 21 hours. The weight-estimated average moisture content (%) in each wood sample at each time step is the response variable of the dataset. The size of the design matrix

Abbreviation	Category
$\mathbf{X}$	No preprocessing
$\mathrm{D}^1\mathbf{X}$	Derivative
$\mathrm{D}^2\mathbf{X}$	Derivative
$\mathrm{D}^3\mathbf{X}$	Derivative
$\mathrm{MSC}(\mathbf{X})$	Scatter correction
$\mathrm{SNV}(\mathbf{X})$	Scatter correction
$\mathrm{MSC}(\mathrm{D}^1\mathbf{X})$	Combination
$\mathrm{SNV}(\mathrm{D}^1\mathbf{X})$	Combination
$\mathrm{MSC}(\mathrm{D}^2\mathbf{X})$	Combination
$\mathrm{SNV}(\mathrm{D}^2\mathbf{X})$	Combination
$\mathrm{MSC}(\mathrm{D}^3\mathbf{X})$	Combination
$\mathrm{SNV}(\mathrm{D}^3\mathbf{X})$	Combination

Table 1: Summary of all preprocessing methods used in the experiments.

$\boldsymbol{X}$ containing all the absorbance data is $10^4 \times 200$. A more thorough description of this dataset can be found in [18].

3.2 Description of the preprocessing techniques used

Preprocessing is often considered a crucial part of the data exploration/model development phase, particularly in the field of hyperspectral imaging [19]. There are many ways in which spectroscopic data can be preprocessed, most methods however fall into one out of two categories: derivation methods or scatter correction methods. The purpose of derivation is to remove baseline effects in the spectra and/or to enhance spectral features. Scatter correction methods on the other hand seek to remove scatter effects from the spectra, thereby making samples more comparable to each other. Scatter correction techniques are considered especially relevant for near-infrared data since the NIR region is susceptible to disturbances caused by light scattering [1]. Combinations of preprocessing techniques belonging to these two categories can also be used together when treating a spectral dataset in order to obtain benefits associated with both categories. In our experiments, first (D^1), second (D^2) and third order derivation (D^3) were used as derivation methods. All spectral derivations were carried out using Savitzky-Golay derivation [20] with a window size of 11 and a polynomial order of 3. Multiplicative Scatter Correction (MSC) [21] and Standard Normal Variate (SNV) [22] were used as scatter correction methods. With the restriction that derivation is carried out prior to scatter correction, six additional scatter corrected derivation methods were formed and used in the experiments by combining the above-mentioned techniques, namely $\mathrm{MSC}(\mathrm{D}^1)$, $\mathrm{SNV}(\mathrm{D}^1)$, $\mathrm{MSC}(\mathrm{D}^2)$, $\mathrm{SNV}(\mathrm{D}^2)$, $\mathrm{MSC}(\mathrm{D}^3)$ and $\mathrm{SNV}(\mathrm{D}^3)$. All of the 12 preprocessing techniques used in the experiments are summarized in Table 1.

3.3 Algorithm settings

In order to perform feature selection on all the 12 preprocessing techniques listed in Table 1 with a single GA-PLS, a population containing 12 preprocessing-specific subpopulations was used. All elements outside of the relevant region for each subpopulation was padded with zeros as conceptually illustrated in row 1-6 of Fig. 4. In addition to the 12 preprocessing-specific subpopulations, a 13^{th} subpopulation was also included in the population, which was not associated with any particular preprocessing technique but instead was permitted to operate unrestrictedly across the whole search space (as illustrated in the chromosomes of row 7-9 in Fig. 4). Each subpopulation contained 512

Setting	Value
Population size	512×13
Number of parents	512×13
Number of offspring	512×13
Maximum generations	200
Maximum PLS components	15
Parent selection	Roulette wheel selection
Survivor selection	Tournament selection
Crossover operator	2-point crossover
Mutation operator	Bitwise
Mutation probability p_m	$1/n$
Cross-validation procedure	k-fold ($k = 10$)

Table 2: Summary of settings used during GA-PLS experiments. For a detailed description of roulette wheel and tournament selection the reader is referred to [13].

chromosomes, resulting in a total population size of 512×13. The number of generated offspring in each generation was set to 512×13, such that 1024×13 feature selections were evaluated for each generation of the algorithm.

The fitness of each chromosome was defined as its minimum RMSE_{cv}, across the number of considered latent variables, multiplied with minus one such that a high fitness corresponds to a low RMSE_{cv} and a low fitness corresponds to a high RMSE_{cv}. All relevant GA-PLS settings used in our experiments are summarized in Table 2. Because parts of GA-PLS algorithms are stochastic and different runs produce different outcomes, the feature selection was repeated ten times for both datasets to increase the robustness of the results. The results from these ten runs were then averaged.

3.4 Benchmark of the computational performance of the suggested GA-PLS

The computational efficiency of the GA-PLS technique introduced in section 2.2 (Algorithm 1) was compared against GA-PLS implementations containing four well-established PLS fitting procedures: NIPALS [23] , SIMPLS [24], Bidiag2 [25] and modified kernel algorithm #2 (without indexing technique from section 2.2) [15]. Dataset 1 (spruce treated with a flame retardant) was used in the benchmark together with the 12 preprocessing techniques listed in Table 1. Each algorithm was used to perform feature selection on the 12 differently preprocessed versions of the dataset and the runtime of each algorithm was measured. For simplicity, the 13^{th} unrestricted subpopulation mentioned in section 3.3 was not included in the benchmark. To measure the effect in calculation time of an increasing population size, the benchmark was performed using subpopulation sizes containing 16, 32, 64, 128, 256, 512 and 1024 individuals. The number of parents and the number of offspring generated were set to the same size as the subpopulation. All benchmarks were performed using MATLAB (R2018a). MATLAB's built-in *plsregress* function was used to fit regression coefficients when benchmarking the SIMPLS based GA-PLS. The NIPALS, Bidiag2 and modified kernel algorithm #2 based GA-PLS implementations were written in MATLAB with [23], [25] and [15] respectively as reference. The MATLAB code used for estimating regression coefficients with kernel PLS and the indexing technique described in section 2.2 can be found in the supplementary material of [12]. In addition to the CPU implementation of Algorithm 1, a GPU implementation of the algorithm written in CUDA C was also included in the comparison (executed from MATLAB through the MEX interface) which estimated the regression coefficients of all feature subsets within a population in parallel. All computations were performed using 32-bit floating-point numbers. The motivation

for using 32-bit floating-point numbers rather than 64-bit is that most consumer-grade GPUs have a substantially higher throughput when performing arithmetic operations on 32-bit numbers [26]. So much so that it can be considered wasteful from a hardware utilization perspective to perform calculations in 64-bit precision when the use of 32-bit numbers is numerically acceptable.

In the benchmark, the algorithm settings were chosen according to Table 2. However, because of the tremendous time required to calculate the regression coefficients using NIPALS, SIMPLS, Bidiag2 and modified kernel algorithm #2 at large population sizes, the GA-PLS's in the benchmark using these PLS methods only ran a single generational cycle, instead of the 200 suggested by Table 2. The reason for this is that benchmarking these algorithms on a single CPU using 200 generations would have taken several years. The measured runtime of one generation for these algorithms was scaled by a constant factor in order to approximate the time required for running 200 generations of the algorithms. Hence, the benchmark results presented in this paper for NIPALS, SIMPLS, Bidiag2 and modified kernel algorithm #2 are only an empirical estimate of the time required to perform a full 200 generation GA-PLS, an estimate based on the assumption that all generations take an equal amount of time to complete. Both the CPU and GPU based implementations of Algorithm 1 however, were measured using the full 200 generations.

4 Results

4.1 Variable selection results

Table 3 contains a summary of the average number of active variables, average number of optimal PLS components and the average RMSE_{cv} of each subpopulation after 200 generations when applying Algorithm 1 to dataset 1. Figure 6 a) shows how the RMSE_{cv} developed for the individuals of the population over the course of 200 generations. Each chromosome of the population is colored according to its subpopulation belonging. Interestingly, in this dataset none of the included preprocessing techniques were alone able to surpass the performance of the subpopulation operating on the unpreprocessed absorbance data. The subpopulation operating across the whole search domain unrestrictedly, however, was able to find feature selections which resulted in a slightly lower RMSE_{cv} than that of those using regular absorbance data. The unrestricted subpopulation was found to contain active features from a broad range of preprocessing methods after 200 generations. Each included preprocessing method except $\boldsymbol{X}$ and $\text{MSC}(\text{D}^1\boldsymbol{X})$ was responsible for 8-14 % of the active genes in the average unrestricted chromosome. $\boldsymbol{X}$ and $\text{MSC}(\text{D}^1\boldsymbol{X})$ were only responsible for roughly 1 % each of the active genes within the unrestricted subpopulation, which can be seen as somewhat surprising considering that the unpreprocessed version of the data $\boldsymbol{X}$ individually performs very well.

Table 4 and Fig. 6 b) shows the corresponding set of results from applying Algorithm 1 to dataset 2. As seen in Table 4 and Fig. 6 b), in this dataset all the included preprocessing techniques managed to improve the RMSE_{cv} compared to the unpreprocessed absorbance data. After 200 generations, the unrestricted subpopulation operating on features from a mixture of preprocessing methods again resulted in the lowest RMSE_{cv}. Although the unrestricted subpopulation did contain approximately twice the number of active variables compared to any of the preprocessing-specific populations, its average optimal number of latent variables was still relatively low. This could indicate that the subpopulations permitted to operate on a mixture of preprocessing methods are evolving a favorable redundancy into their regression models by including multiple features which describe a similar correlation—resulting in models containing many variables which are still characterizable using a low complexity latent space. For dataset 2, the active genes within the unrestricted subpopulation after 200 generations were less uniformly distributed across the different preprocessing methods compared to dataset 1. Moreover, the correlation between the performance of an individual preprocessing method and the extent to which the preprocessing method contributed genes to the unrestricted subpopulation

Preprocessing method	Num. var.[a]	Num. lat.[b]	$RMSE_{CV}$
$\mathbf{X}$	44	6	7394
$D^1\mathbf{X}$	37	5	7499
$D^2\mathbf{X}$	42	15	7825
$D^3\mathbf{X}$	85	14	8247
MSC($\mathbf{X}$)	38	4	7514
SNV($\mathbf{X}$)	38	4	7473
MSC($D^1\mathbf{X}$)	47	5	8238
SNV($D^1\mathbf{X}$)	35	2	8070
MSC($D^2\mathbf{X}$)	45	1	10687
SNV($D^2\mathbf{X}$)	58	8	8380
MSC($D^3\mathbf{X}$)	86	14	8908
SNV($D^3\mathbf{X}$)	92	14	8687
Multiple	62	9	6966

[a]*Average number of active variables in the subpopulations after 200 generations.* [b]*Average number of optimal latent variables in the subpopulations after 200 generations.*

Table 3: Summary of results after 200 generations of applying the proposed GA-PLS method (Algorithm 1) to dataset 1. Results are averaged across ten trials.

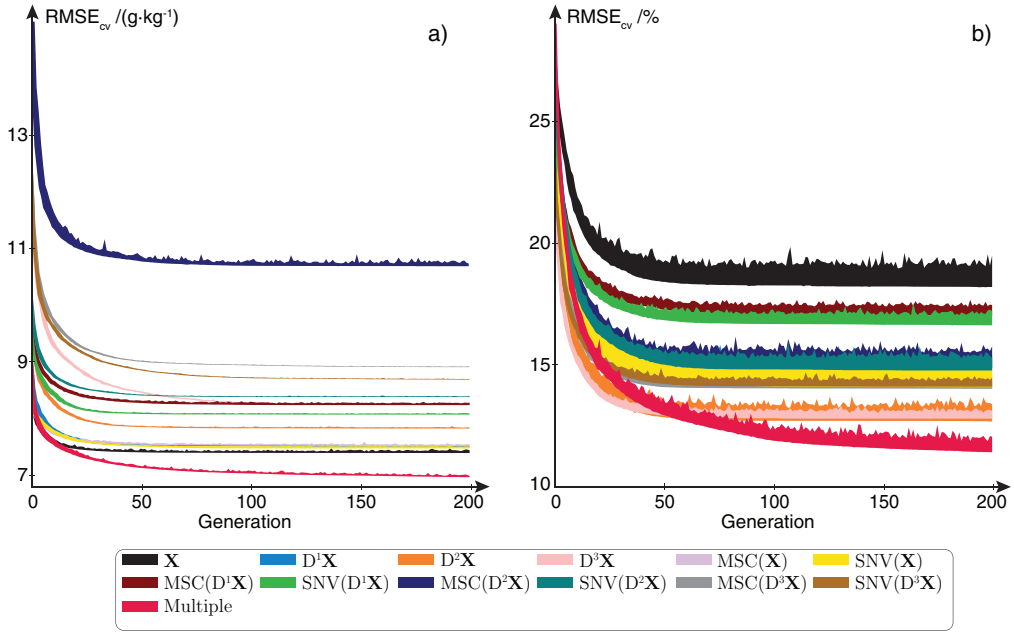


Figure 6: Cross-validated root-mean-squared error, $RMSE_{cv}$, as a function of generation for every member of the population when applying Algorithm 1 to a) dataset 1 and b) dataset 2. Results are averaged across ten trials.

Preprocessing method	Num. var.[a]	Num. lat.[b]	RMSE_{CV}
$\mathbf{X}$	**25**	**8**	**18.20**
$\text{D}^1\mathbf{X}$	**24**	**6**	**14.87**
$\text{D}^2\mathbf{X}$	**23**	**6**	**12.68**
$\text{D}^3\mathbf{X}$	**24**	**10**	**12.80**
$\text{MSC}(\mathbf{X})$	**19**	**7**	**14.06**
$\text{SNV}(\mathbf{X})$	**20**	**7**	**14.09**
$\text{MSC}(\text{D}^1\mathbf{X})$	**23**	**9**	**17.02**
$\text{SNV}(\text{D}^1\mathbf{X})$	**24**	**8**	**16.62**
$\text{MSC}(\text{D}^2\mathbf{X})$	**22**	**11**	**14.85**
$\text{SNV}(\text{D}^2\mathbf{X})$	**22**	**10**	**14.77**
$\text{MSC}(\text{D}^3\mathbf{X})$	**23**	**11**	**14.02**
$\text{SNV}(\text{D}^3\mathbf{X})$	**24**	**8**	**14.11**
Multiple	**43**	**7**	**11.42**

[a]*Average number of active variables in the subpopulations after 200 generations.* [b]*Average number of optimal latent variables in the subpopulations after 200 generations.*

Table 4: Summary of results after 200 generations of applying the proposed GA-PLS method (Algorithm 1) to dataset 2. Results are averaged across ten trials.

was somewhat more intuitive, i.e. methods performing well on their own in general contributed a larger number of variables to the unrestricted subpopulation. The lowest number of active genes in the unrestricted subpopulation, 0.7 %, originated from the unpreprocessed $\boldsymbol{X}$, whilst the highest number of active genes, 14 %, originated from $\text{D}^3\boldsymbol{X}$.

4.2 Benchmark results

Figure 7 shows the results from the benchmark of measuring the runtime of GA-PLS algorithms using different PLS fitting procedures at various subpopulation sizes. Subpopulation sizes whose runtimes' were not explicitly measured are linearly interpolated in the figure to provide a clearer view of the trend. As can be seen in the figure, there is a substantial difference in computational efficiency between the GA-PLS implementations using any of the conventional set of PLS fitting procedures—NIPALS, SIMPLS, Bidiag2 and modified kernel algorithm #2—and the GA-PLS implementation described in section 2.2 (Algorithm 1). Furthermore, the parallel GPU-based version of Algorithm 1 offers an additional massive leap in speedup compared to the CPU version of the same algorithm.

5 Discussion & conclusions

A fast new method for performing GA-PLS subset selection on multiple preprocessed versions of a dataset within a single algorithm was presented. The main benefit of the introduced method is that it enables the performance of multiple spectral pretreatment methods to be explored in parallel with each other in a manner which is substantially less time consuming than previously possible. Indeed, evaluating the GA-PLS performance of multiple pretreatment methods for tall datasets could previously require weeks, months or even years of calculation time; using the technique introduced in this paper the same exploratory search can now be done in an afternoon. The technique we introduced in this paper is primarily intended for strongly overdetermined systems due to two reasons: (1) the method relies on kernel PLS (specifically Modied kernel algorithm #2 [15]), which is susceptible

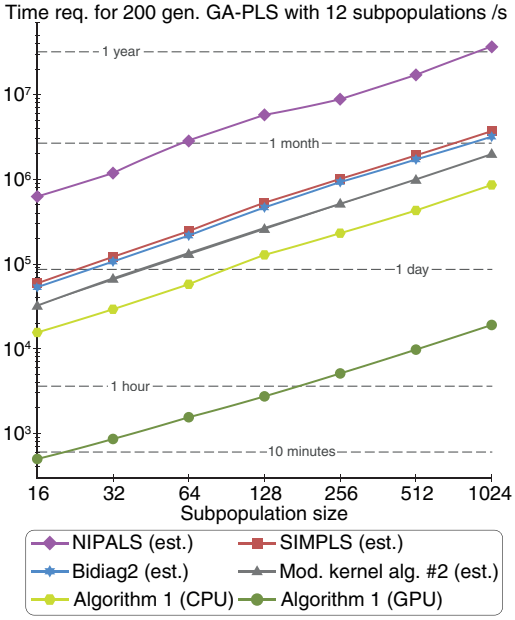


Figure 7: Benchmark results of measuring the calculation time (in seconds) required for performing GA-PLS with the 12 preprocessing techniques listed in Table 1 at various population sizes using six different algorithms. NIPALS, SIMPLS, Bidiag2 and modified kernel algorithm #2 were only timed for 1 generation, the measured time was then scaled to approximate the time required for 200 generations. "Subpopulation size" in the figure refers to the number of chromosomes related to each preprocessing technique. Since twelve preprocessing methods were included in the GA-PLS, the total population size is twelve times that of the subpopulation size. All CPU-based calculations were performed on an i7-7700K @ 4.2 GHz. The GPU-based implementation of Algorithm 1 was calculated on a GTX1080ti @ 1.6 GHz. All calculations were perfomed using 32-bit precision floating points numbers. The size of the $\boldsymbol{X}$ matrix used in the benchmark was $10^5 \times 256$.

to numerical issues for systems that are not overdetermined. (2) the computational cost savings offered by the covariance indexing technique used in the method increase the more overdetermined the system is [12].

The use of feature selections containing variables from a mixture of preprocessing methods was included in this paper primarily to showcase the possibility of doing so and the ease by which it can be done with the introduced method. Although this type of feature selection did result in the lowest RMSE_{cv} for both of our evaluated datasets, in many cases it may still not be advisable to use this approach for a few reasons. For instance, it increases the complexity of the resulting model and arguably lowers its interpretability. With the increased combinatorial complexity available to the GA-PLS, the potential of evolving overfitted solutions—which is widely recognized as an ever-present risk when dealing with GA-PLS—also rises. Additional care should therefore be taken when designing the cross-validation procedure to use in the GA-PLS if such complex models are permitted. In addition to enabling a quick evaluation of the performance of a set of different preprocessing methods, as we have demonstrated here, the introduced method could also be suitable for finding an optimal set of parameters to use together with one particular preprocessing method. For instance, if one has decided to use Savitzky-Golay derivation or smoothing on a dataset, the set of preprocessing methods

used in Algorithm 1 can easily be changed into a set containing replicates of the same preprocessing method but with different parameter settings, i.e. different window widths and polynomial degrees in the case of Savitzky-Golay filtering. Different subpopulations can then be appointed to different parameter configurations of the preprocessing method to provide an indication of which parameter settings are best suited for the dataset. Lastly, it should be noted that because kernel PLS is not the most numerically stable PLS procedure, it could be beneficial to use the technique described in this paper as a means of quickly exploring a spectral dataset and how it responds to various pretreatment methods. Then, once a satisfying combination of preprocessing technique and feature subset has been identified, the final PLS model can be recalibrated using a numerically more precise fitting procedure, such as Bidiag2.

References

[1] A. Rinnan, F. Berg, and S. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[2] H. Zhao, K.-W. Huan, X.-G. Shi, F. Zheng, L.-Y. Liu, W. Liu, and C.-Y. Zhao, "A variable selection method of near infrared spectroscopy based on automatic weighting variable combination population analysis," *Chinese Journal of Analytical Chemistry*, vol. 46, no. 1, pp. 136–142, 2018.

[3] T. Mehmood, K. H. Liland, L. Snipen, and S. S. Saebo, "A review of variable selection methods in partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 118, pp. 62–69, 2012.

[4] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[5] L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, "Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy," *Applied Spectroscopy*, vol. 54, no. 3, pp. 413–419, 2000.

[6] K. Song, L. Li, L. Tedesco, S. Li, N. Clercin, B. Hall, Z. Li, and K. Shi, "Hyperspectral determination of eutrophication for a water supply source via genetic algorithm-partial least squares (ga-pls) modeling," *Science of The Total Environment*, vol. 426, pp. 220–232, 2012.

[7] R. Leardi and A. L. González, "Genetic algorithms applied to feature selection in pls regression: how and when to use them," *Chemometrics and Intelligent Laboratory Systems*, vol. 41, no. 2, pp. 195–207, 1998.

[8] D. Jie, L. Xie, X. Fu, X. Rao, and Y. Ying, "Variable selection for partial least squares analysis of soluble solids content in watermelon using near-infrared diffuse transmission technique," *Journal of Food Engineering*, vol. 118, no. 4, pp. 387–392, 2013.

[9] A. Durand, O. Devos, C. Ruckebusch, and J. Huvenne, "Genetic algorithm optimisation combined with partial least squares regression and mutual information variable selection procedures in near-infrared quantitative analysis of cotton viscose textiles," *Analytica Chimica Acta*, vol. 595, no. 1-2, pp. 72–79, 2007.

[10] R. S. G. F. C. Reynès, S. d. Souza and B. Vidal, "Selection of discriminant wavelength intervals in nir spectrometry with genetic algorithms," *Journal of Chemometrics*, vol. 20, no. 3-4, pp. 136–145, 2006.

[11] M. S. Y.a. Zhang, Q. Ma and H. Furutani, "Effects of population size on the performance of genetic algorithms and the role of crossover," *Artificial Life and Robotics*, vol. 2, no. 15, pp. 239–243, 2010.

[12] P. Stefansson, U. G. Indahl, K. H. Liland, and I. Burud, "Orders of magnitude speed increase in partial least squared feature selection with new simple indexing technique for very tall data sets," *Journal of Chemometrics*, 2019.

[13] A. E. Eiben and J. E. Smith, *Introduction to evolutionary computing, Berlin: Springer Berlin Heidelberg*, 2015.

[14] R. Leardi, "Genetic algorithm-pls as a tool for wavelength selection in spectral data sets," *Data Handling in Science and Technology*, pp. 169–196, 2003.

[15] D. B. S. and M. J. F., "Improved pls algorithms," *Journal of Chemometrics*, vol. 11, no. 1, pp. 73–85, 1997.

[16] F. Lindgren, P. Geladi, and S. Wold, "The kernel algorithm for pls," *Journal of Chemometrics*, vol. 7, no. 1, pp. 45–59, 1993.

[17] P. Stefansson, T. Thiis, L. Gobakken, and E. Larnøy, "Estimation of phosphorus-based flame retardant in wood by hyperspectral imaging—a new method," *Journal of Spectral Imaging*, vol. 7, 2018.

[18] P. Stefansson, J. Fortuna, H. Rahmati, I. Burud, T. Konevskikha, and H. Martens, *Hyperspectral time series analysis: Hyperspectral image data streams interpreted by modeling known and unknown variations (in print)*, 2019.

[19] M. Vidal and J. M. Amigo, "Pre-processing of hyperspectral images. essential steps before image analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 138–148, 2012.

[20] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[21] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 3, no. 39, pp. 491–500, 1985.

[22] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[23] D. L. Massart, *Handbook of Chemometrics and Qualimetrics*. Burlington: Elsevier p. 336, 1998.

[24] S. de Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[25] A. Björck and U. G. Indahl, "Fast and stable partial least squares modelling: A benchmark study with theoretical comments," *Journal of Chemometrics*, vol. 31, p. 8, 2017.

[26] N. Ploskas and N. Samaras, *GPU programming in Matlab*. 2016.

Paper III

Peer Reviewed Paper openaccess Special Issue on Chemometrics in Hyperspectral Imaging

Estimation of phosphorus-based flame retardant in wood by hyperspectral imaging—a new method

Petter Stefansson,[a,*] Ingunn Burud,[b] Thomas Thiis,[c] Lone R. Gobakken[d] and Erik Larnøy[e]

[a]Faculty of Science and Technology, Norwegian University of Life Sciences, Drøbakveien 31, 1430 Ås, Norway.
E-mail: petter.stefansson@nmbu.no
[b]Faculty of Science and Technology, Norwegian University of Life Sciences, Drøbakveien 31, 1430 Ås, Norway.
ORCID: https://orcid.org/0000-0003-0637-4073
[c]Faculty of Science and Technology, Norwegian University of Life Sciences, Drøbakveien 31, 1430 Ås, Norway
[d]Norwegian Institute of Bioeconomy Research, PO Box 115, 1431 Ås, Norway
[e]Norwegian Institute of Bioeconomy Research, PO Box 115, 1431 Ås, Norway. ORCID: https://orcid.org/0000-0002-8724-4010

It is recognised that flame retardant chemicals degrade and leach out of flame-protected wood claddings when exposed to natural weathering. However, the ability to survey the current state of a flame retardant treatment applied to a wood cladding, an arbitrary length of time after the initial application, is limited today. In this study, hyperspectral imaging in the near infrared to short-wavelength infrared region is used to quantify the amount of flame retardant present on wooden surfaces. Several sets of samples were treated with various concentrations of a flame retardant chemical and scanned with a push broom hyperspectral camera. An inductively coupled plasma (ICP) spectroscopy analysis of the outermost layer of the treated samples was then carried out in order to determine each sample's phosphorus content, the active ingredient in the flame retardant. Spectra from the hyperspectral images were pre-processed with extended multiplicative scatter correction, and the phosphorus content was modelled using a partial least squares (PLS) regression model. The PLS regression yielded robust predictions of surface phosphorus content with a coefficient of determination, R^2, between 0.8 and 0.9 on validation data regardless of whether the flame retardant chemical had been applied to the surface of the wood or pressure-impregnated into it. The result from the study indicates that spectral imaging around the 2400–2531 nm wavelength region is favourable for quantifying the amount of phosphorus-based flame retardant contained in the outermost layer of non-coated wooden claddings. The results also reveal that the uptake of phosphorus-based flame retardant does not occur uniformly throughout the wood surface, but is to a larger extent concentrated in the earlywood regions than in the latewood.

Keywords: hyperspectral imaging, NIR, flame retardant treated wood

Introduction

Throughout the European building sector wood is strengthening its position as a preferred building material in many applications.[1] This trend is incentivised by the European Union's *Europe 2020* targets, which aim to substantially decrease the greenhouse gas emissions and increase energy-efficiency in the EU by the year 2020, partly by promoting the sustainable use of wood in construction.[2] Wood materials used in buildings already benefit from

Correspondence
Petter Stefansson (petter.stefansson@nmbu.no)

Received: 15 September 2017
Revised: 30 November 2017
Accepted: 31 January 2018
Publication: 15 February 2018
doi: 10.1255/jsi.2018.a3
ISSN: 2040-4565

Citation
P. Stefansson *et al.*, "Estimation of phosphorus-based flame retardant in wood by hyperspectral imaging—a new method", *J. Spectral Imaging* **7,** a3 (2018). https://doi.org/10.1255/jsi.2018.a3

having a low carbon footprint,[3] and through increased focus on material selection which is fit-for-purpose and by facilitating predictable long-term performance, their footprint can be reduced even further. However, the combustibility of wood can be a challenge when wood is used as a construction material. Due to the risk of fire, many European countries adopted an outright ban on constructing timber buildings taller than two stories for a long time. More developed building regulations and an increased understanding of fire safety has, however, in the last few decades caused many of these bans to be lifted. It is now both permitted and often encouraged to build multi-story timber buildings in most European countries.[4] One effective strategy for increasing the fire safety of wood-based buildings is to use wood materials which prior to assembly in the construction phase, have been treated with a phosphorus-based flame retardant chemical. In an event of fire, such treatments react by producing a char surface which prevents the degradation of the wood and delays the point of flashover.

An important fire safety aspect, which currently is largely unknown, is how wood materials treated with flame retardant chemicals are affected by natural weathering and how the treatment withstands long-term exposure to the external climate. It is known that phosphorus, a crucial component in many flame retardant treatments, tends to leach out over time from the wood it has been impregnated into.[5] By weighing a piece of lumber before and after it undergoes a chemical treatment to increase its flame resistance, it is possible to determine the total uptake of flame retardant compound in the wood. To determine the amount of chemical which resides in the outermost layer of a wood cladding treated with a flame retardant, an arbitrary length of time after the initial treatment, is considerably more difficult. Especially if it is to be done in a non-destructive way. For this reason, the ability to survey and inspect the treatment status of existing wooden constructions or stored lumber is severely limited.

In this paper, a new method for quantifying the current concentration of flame retardant in the outer layer of wood samples using near infrared (NIR) hyperspectral imaging is developed and evaluated. NIR spectroscopy is widely used for quality control in numerous fields, e.g., the food industry,[6] and the potential for using NIR spectroscopy in the field of wood science has successfully been demonstrated in previous studies.[7–9] NIR hyperspectral imaging has also been applied to wood surfaces to determine wood moisture content,[10] map chemical composition,[11] determine wood extractive content,[12] map weathering by UV radiation,[13] for identification of compression wood[14] and for detecting show-through resin defects on painted lumber.[15]

The method proposed in this paper involves mapping the NIR absorbance spectrum of a wood sample, by means of a linear regression model, to its corresponding surface concentration of phosphorus. Phosphorus is the active ingredient in many flame retardant chemicals and can be seen as proportional to its ability to protect the wood against fire. By measuring the NIR spectra using hyperspectral imaging, it also becomes possible to spatially resolve the absorbance properties of the wood, which enables completely non-destructive and rapid surveying of large areas of wood such as facades or wooden decks.

Method

Sample preparation and signal acquisition

Five boards of Norway spruce (*Picea abies*), originating from different logs, were each cut into seven samples of dimension 50 × 50 × 10 mm, giving a total of 35 samples. The samples were conditioned in a climate chamber at 20 °C and 65 % relative humidity until the samples had reached equilibrium moisture content. The samples were treated with *Preventor AntiFlame* from Akzo Nobel, which is a flame retardant wood coating commonly used in the construction industry. According to the Akzo Nobel, the undiluted chemical is typically diluted with distilled water to reach any desired solution concentration prior to applying it to wood, and a concentration of 66 % is common when pressure-impregnating the chemical into wood. In our study, the Preventor AntiFlame was diluted with distilled water into seven different concentrations: 0, 17, 33, 50, 67, 83 and 100 %. Each of the seven concentrations was then used to treat 5 of the 35 samples in a manner such that each concentration was represented once in each of the original wood logs. This was done to ensure that any model developed from the data set had the ability to reliably estimate the concentration of flame retardant chemical in the wood despite natural variations in the physical appearance of samples originating from different trees, which could affect their spectrum.

To ensure even and comparable uptake of each solution, the solutions were poured into petri dishes and each sample was immersed in the solution for 30 seconds. In

addition to the immersed samples, five samples that were industrially pressure-impregnated using the same flame retardant chemical at a concentration of 66 % were also included in the study to verify that the spectral fingerprint of the flame retardant substance is the same independent of treatment method.

A custom plastic sample mount was designed and 3D-printed, which could accommodate five samples at a time in addition to a Spectralon white reference plate during image acquisition. The samples were then placed inside the plastic mount, illuminated by two halogen lights and scanned using a push broom hyperspectral camera (Specim, Oulu, Finland) with 256 bands in the 929–2531 nm range, as can be seen illustrated in Figure 1.

After signal acquisition, a three-dimensional representation of the reflective properties of each sample was acquired, with spatial width along one dimension, spatial height along one dimension and reflected light along the third as can be seen illustrated in Figure 2. The spatial resolution of each image was 151 × 151 pixels, resulting in a pixel dimension of 0.33 mm^2.

When the reflective properties of each sample had been measured, a 0.4–0.5 mm layer was removed from the surface of each sample using a planer, which was cleaned with alcohol between every sample to prevent cross contamination. The phosphorus content in the removed layers was analysed according to the ICP21100 Thermo Jarell Ash ICP-IRIS HR Duo method.[16] The laboratory established surface concentration of phosphorus for each sample were used as response values when developing the regression model.

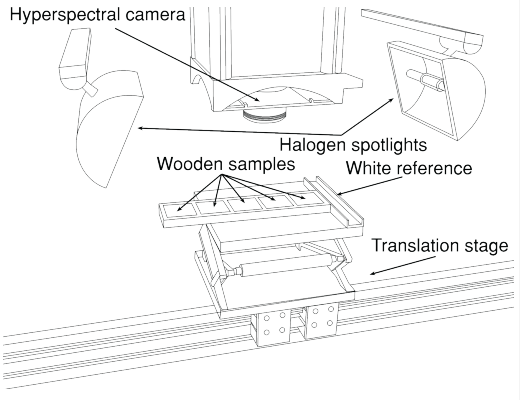


Figure 1. Illustration of experimental setup.

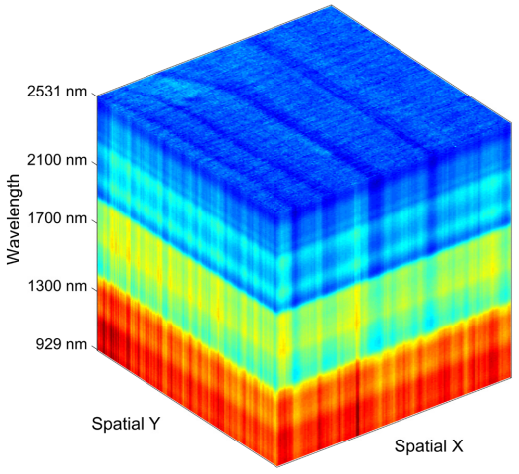


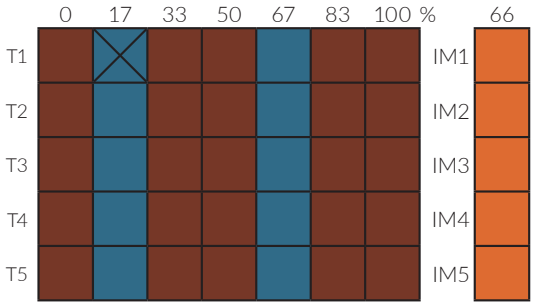
Figure 2. Structure of 3-D hypercube of a wood sample with NIR reflectance as a function of wavelength and spatial position.

The hyperspectral data were divided into a training set (5/8 of the data) and a validation set (3/8 of the data). The samples treated with concentrations 0, 33, 50, 83 and 100 % were chosen as training data to be used when developing the regression model. Samples treated with concentrations of 17 % and 67 %, together with the pressure-impregnated samples, were consequently not used in the calibration of the regression model. These samples could therefore be used to validate the final model's ability to generalise to new unseen data and estimate the phosphorus content of new wood samples. One of the five samples treated with 17 % chemical concentration was, however, later excluded from the validation data set since it was found to be covered in resin, which severely distorted its spectral properties. Table 1 shows a summary of all the samples in the data set and their treatment.

Data pre-processing

The signal from each hyperspectral image was first recalculated into reflectance, relative to a white reference, by subtracting the mean value of each pixel row in the spatial direction of a dark region (taken with the camera shutter closed), and dividing by the mean intensity value of a white calibration plate included in the image, also for each pixel row. The reflectance spectra, R, were transformed into absorbance, A, using the rela-

Table 1. Overview of samples included in the study. Maroon-coloured cells represent samples used as calibration data, whilst blue and orange cells indicate immersed, T, and pressure-impregnated, IM, samples used to validate the PLS model's performance. The crossed-out cell represents a resin covered sample which was excluded from the validation data set.



tion $A = \log_{10}(1/R)$. All spectra were pre-processed using extended multiplicative scatter correction (EMSC), which is a pre-processing technique that corrects for undesired physical variations between the samples, such as surface geometry causing light scattering effects,[17] while attempting to retain spectral variations caused by chemical differences between the samples. This is achieved by fitting each raw absorbance spectrum using ordinary least squares (OLS) fitting to a design matrix containing a constant term, a linear term consisting of a vector with all wavelengths included in the spectrum, a quadratic term containing a squared version of the wavelengths and a reference spectrum which is assumed to contain less scatter effects than each individual raw spectrum. The spectral correction is then carried out using

$$S_{Corr} = \frac{S_{Raw} - b_1 - b_2 \cdot w - b_3 \cdot w^2}{b_4} \quad (1)$$

where b_{1-4} are the regression coefficients, acquired with OLS, corresponding to the consecutive terms in the design matrix. w is the vector of wavelengths involved, S_{Raw} is the uncorrected spectrum and S_{Corr} is the corrected spectrum.

The reference spectrum used during our EMSC was set to the average spectrum of all samples within the training data set, thereby avoiding information leakage between the training and validation data which were both processed using the same reference spectrum. For a more in-depth description of EMSC, the reader is referred to References 17 and 18.

Wavelength selection and PLS regression

Noisy and irrelevant parts of the spectra which were not correlated to the phosphorus content in any statistically meaningful way were removed by a *backwards elimination* algorithm. In backwards elimination, an initial regression model is created using all available variables, which in this case corresponds to wavelengths. The model is then created again N times, where N is the number of available wavelengths, with one of the available wavelengths inactivated each iteration until all wavelengths have been excluded once. The wavelength selection with the lowest root mean square error of cross-validation ($RMSE_{CV}$) of all the evaluated models then becomes the new start-case and the process starts over again recursively until no single wavelength inactivation can be made that improves the $RMSE_{CV}$. The backwards elimination algorithm utilised in this paper is summarised in pseudocode in Figure 3. All fitting of regression coefficients was done using a C implementation of the *Bidiag2* partial least squares algorithm proposed in Reference 19, which was chosen due to its proven numerical stability and computational efficiency. The partial least squares (PLS) regression was configured to consider at most 25 components. Since backwards elimination can require several thousand regression models to be fitted before terminating, the spatial resolution of the hyperspectral data was down-sampled with regional pixel averaging to allow for

Algorithm 1 Pseudocode for Backwards elimination

```
Continue = true
while Continue == true do
    /* Evaluate full X */
    RMSE[0] = PLS(X, y)

    /* Evaluate every inactivation */
    for n ← 1 to N do
        RMSE[n] = PLS(X[:, ! = n], y)
    end for

    /* Find index of lowest RMSEcv */
    n = argmin(RMSE[0 : N])

    /* Stop if full X was best */
    if n == 0
        Continue = false
    /* Else make inactivation permanent */
    else
        X = X[:, ! = n]
    end if
end while
```

Figure 3. Pseudocode for the backwards elimination algorithm used to select relevant wavelengths. Cross-validation loop omitted for clarity. The variable *N* symbolises the dynamically changing width of the design matrix X.

faster model development into 625 spectra (25 × 25) per hypercube.

Results and discussion

Pre-processing

Figure 4 shows the mean spectrum from all samples in the data set before (left) and after (right) being processed using EMSC. As can be seen to the right in Figure 4, the most significant change induced in the spectrum when varying the concentration of flame retardant compound appears to be in the 1900–2531 nm region. It is also noteworthy that the pressure-impregnated samples, which are displayed as dashed black lines in Figure 4, exhibit slightly different absorbance properties from the immersed samples at wavelengths shorter than 1500 nm, which suggests that treatment method influences the absorbance properties of the sample in certain spectral regions. The absorbance difference between immersed and pressure-impregnated samples does, however, appear to become much less significant in the far-end of the spectrum.

Wavelength selection

The $RMSE_{cv}$ kept decreasing as the backwards elimination algorithm ran until only 17 wavelengths remained active. As can be seen in Figure 5, illustrating the location of these wavelengths, all 17 active wavelengths were found in the 2400–2531 nm range. In terms of applying the model to the pressure-impregnated samples, the wavelengths selected by the backwards elimination algorithm seem to be among the most favourable, since the absorbance of all wood samples is similar in this wavelength region, regardless of treatment method. The number of PLS components deemed optimal in the final variable selection with respect to $RMSE_{CV}$ was one, resulting in a model of low complexity. This identified wavelength range for determining phosphorus agrees with wavelength regions which have previously been reported in studies of phosphorus content in dried woody plant species[20] and in soil.[21] The latter study in soil phosphorus concentration also found large regression coefficient values at other wavelengths, including the visible region. Gillon *et al.*[22] studied phosphorus content in plant material with spectra in the visible (Vis) and NIR regions and found that most of the spectral information was found in the NIR part of the spectrum. Although optical properties in the Vis region of the spectrum were not measured for samples in this study, the supposed correlation between absorbance in the NIR region and phosphorous content agrees with the present findings.

Model performance

Figure 6 shows the phosphorus content from the ICP analysis versus the final regression model's prediction of the phosphorus content based on the samples absorbance for both the samples within the training data set (left) and the validation data set (right). The regression model's RMSE of the phosphorus content of the training data was 6055 mg kg^{-1} with a coefficient of determination of 0.90. For the validation samples immersed in 17 % and 67 % chemical solution, the RMSE was 4297 mg kg^{-1} with a coefficient of determination of 0.87. Similar good regression results have been obtained when predicting phosphorus in soil and ground plant material,[21,22] whereas no previous study has been found using this technology on phosphorus in wooden surfaces.

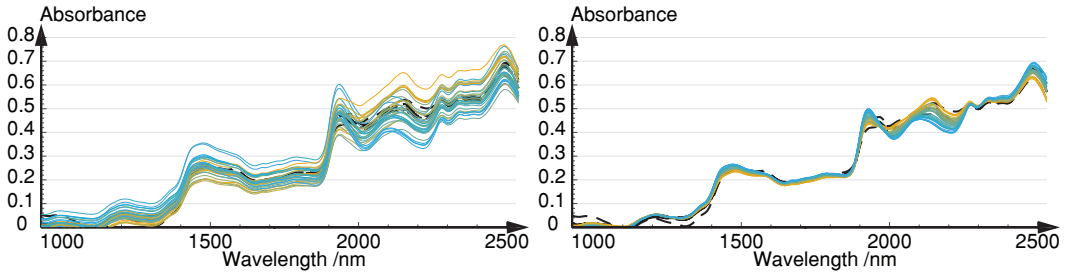


Figure 4. Mean spectrum of each sample in the dataset before (left) and after (right) being pre-processed using EMSC. Line colour indicates phosphorus content in the sample with yellow representing lower concentration and blue higher. Black dotted line indicates the spectrum from the pressure-impregnated samples.

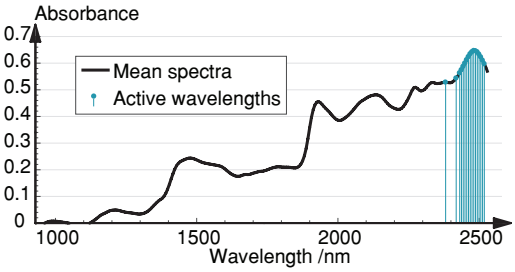


Figure 5. Wavelengths selected by backwards elimination. Blue lines represent wavelengths which were active after backwards elimination. The black curve represents the mean absorbance spectrum of the training dataset.

When using the model to predict the phosphorus content in the pressure-impregnated samples, the RMSE degraded to 24,394 mg kg^{-1} with a coefficient of determination of 0.87. However, as can be seen both in the right part of Figure 6 and in Table 2 which provides a summary of the ICP results for all samples, this apparent collapse in RMSE can almost entirely be accredited to just one of the five impregnated samples, a sample which according to the ICP results supposedly has several times the phosphorus content compared to the other impregnated samples. It is unclear why sample IM1 has such an abnormally high phosphorous content considering that all impregnated samples were treated with the same solution. We can only speculate that it could be due to an exceptional deviation in sample density, an aggregation effect caused by an irregular surface curvature or a result of too high moisture conditions during storage. If the sample causing the RMSE collapse is disregarded as an outlier the RMSE of the four remaining impregnated samples becomes 8650 mg kg^{-1}, which is more comparable to the immersed samples.

When reviewing the results in Figure 6 it is important to realise that, as is often the case in multivariate calibration with hyperspectral data, only the mean response value of each sample is known. Because multiple spectra share the same target value, it is therefore inevitable for the calibrated model to produce a distribution of errors around the target. In the absence of an alternative measurement technique which can be used to validate the chemical variations within each sample, the convention is to arrange the distribution of predicted values into two-dimensional images and visually determine if the in-sample variations display a credible pattern or not. Figure 7 shows the regression model's phosphorous prediction of every measured spectra for the immersed samples as a function of pixel position, arranged into chemical maps, with applied concentration of flame retardant increasing in the horizontal direction and the different wooden logs in the vertical direction. The model generally estimates a higher concentration of phosphorus in the earlywood regions than in the

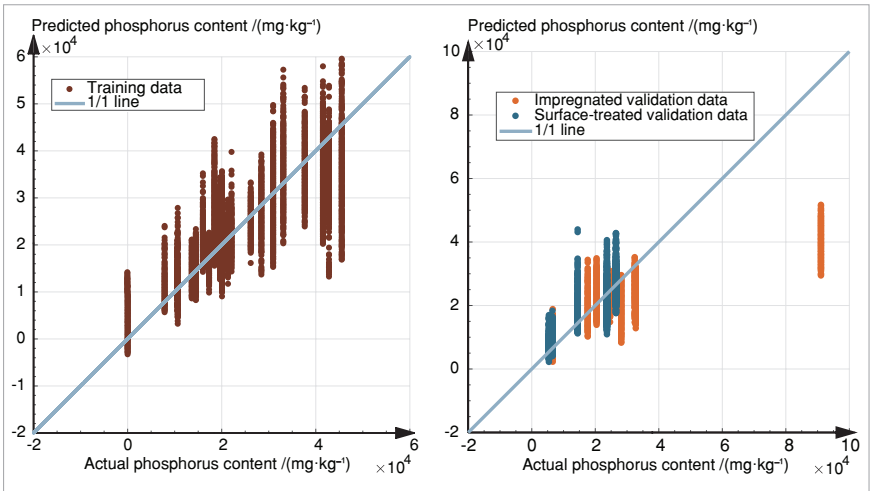


Figure 6. Regression plot of actual vs predicted phosphorus content for the training data (left) and the validation data (right). Each dot corresponds to one of the 25 × 25 spectra from the down-sampled hypercubes.

Table 2. Phosphorous concentration established by ICP for all samples together with mean, μ, and standard deviation, σ, across replicates. All values are in the unit mg kg^{-1}.

Sample \ Conc.	0 %	17 %	33 %	50 %	67 %	83 %	100 %
T1	87	3207	7941	10,592	14,575	16,081	31,014
T2	72	5524	10,730	20,141	23,793	18,533	41,603
T3	90	6632	14,586	19,715	26,643	33,177	45,600
T4	20	6436	13,714	21,414	20,532	28,516	42,888
T5	80	6724	17,392	26,249	17,749	37,730	22,169
μ	70	5705	12,872	19,622	20,658	26,807	36,655
σ	29	1475	3639	5678	4773	9305	9811

Sample \ Conc.	66 %
IM1	91,188
IM2	24,769
IM3	32,871
IM4	32,542
IM5	28,314
μ	41,937
σ	27,733

latewood regions. This makes intuitive sense, since the earlywood regions of Norway Spruce have lower density and are more susceptible to absorbing liquids, but this is a phenomenon which, to the authors' knowledge, has never been demonstrated before with flame retardant treatments. Thumm *et al.* have previously demonstrated that resin affects the spectral signature of wood around the 1180 nm and 1370 nm regions.[15] As can be seen on sample T1 17 % in Figure 7, which depicts the modelled phosphorous of the resin covered sample, it is clear that

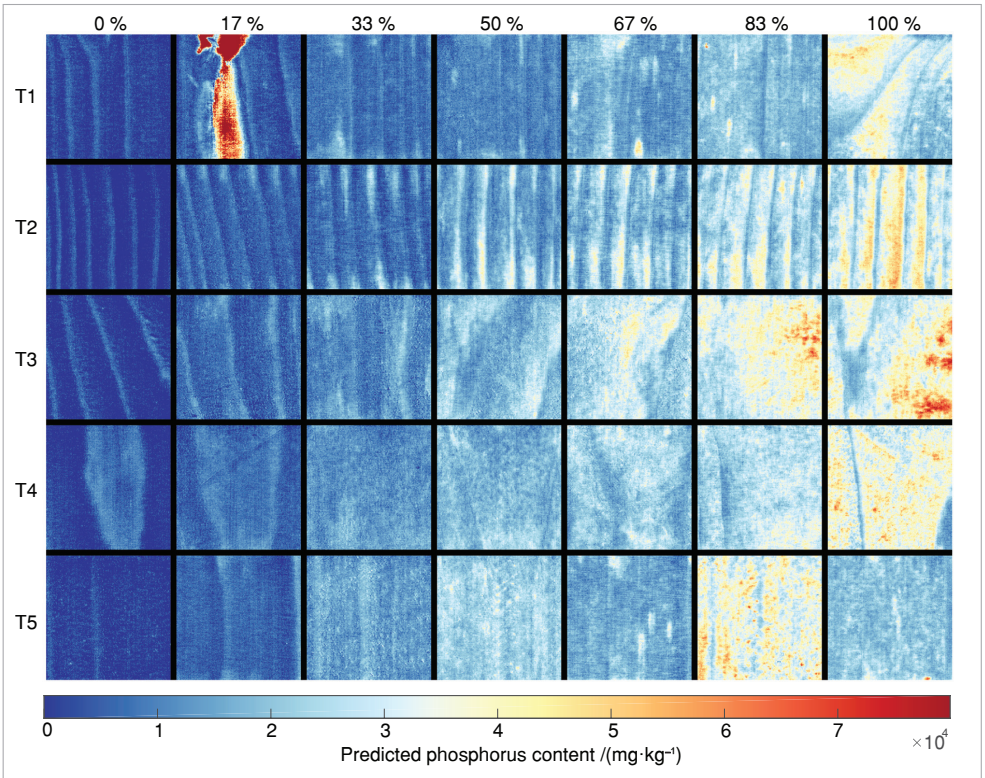


Figure 7. Prediction of phosphorus content for every immersed sample in the data set. Sample T1 17 % illustrates the outlier sample covered in resin which was not considered when evaluating the performace of the model.

resin also alters the spectra in the 2400–2531 nm range, which in this case causes a local misclassification of phosphorous. If the spectral signature of resin was studied further it is possible that a different wavelength subset could be identified which would allow a model to accurately predict phosphorous whilst being unaffected by the presence of resin.

Figure 8 shows the chemical maps for the pressure-impregnated samples. As can clearly be seen in Figure 8, the model does estimate, in accordance with the ICP analysis, that one of the pressure-impregnated samples has a substantially higher surface concentration of phosphorus than the others. The precise quantity of phosphorus reported by the ICP analysis and the regression model does however differ substantially as shown in Figure 6. If indeed the phosphorus concentration reported by the ICP analysis for this sample is valid, it is not surprising that the regression model struggles in its estimation since it has a response value far higher than anything used in the calibration of the model.

Conclusions

We demonstrated that NIR hyperspectral imaging together with PLS regression can be used as a novel non-destructive tool for surveying the current condition of phosphorus-based flame retardant chemical compounds, both surface-applied and pressure-impregnated into, samples of Norway spruce. In most cases our model was able to predict the phosphorus content in wood surfaces with a high degree of accuracy with an R^2 of 0.87 on independent validation samples. However, since the method works by measuring how light is reflected off the surface, it is vulnerable to surface defects, such as resin stains, which locally conceals the true physical properties of the wood surface and can cause misleading phosphorus estimates. When using the model to estimate the spatial distribution of phosphorus in our samples, the chemical uptake does not occur entirely evenly throughout the wood. Instead, the highest concentration of phosphorus was generally found in the earlywood regions of the surfaces.

We identified that 2400–2531 nm appears to be a key wavelength region when it comes to estimating phosphorus. Since this region was at the limit of what our instrument could measure, further studies should investigate the possibility of estimating phosphorus using an instrument capable of detecting longer wavelengths into the infrared region.

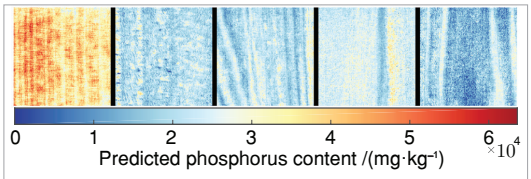


Figure 8. Prediction of phosphorus content of the pressure-impregnated samples. Left-most square represents the sample which, according to the ICP analysis, had a substantially higher phosphorus content compared to the other pressure-impregnated samples.

Acknowledgement

The authors wish to thank Akzo Nobel Coating AS for supplying the flame retardant chemical along with the pressure-impregnated samples used in the study. This work has been funded by the Norwegian Research Council in the project "WOOD/BE/BETTER" code 225345.

References

1. J. Hildebrandt, N. Hagemann and D. Thrän, "The contribution of wood-based construction materials for leveraging a low carbon building sector in Europe", *Sustain. Cities Soc.* **34,** 405–418 (2017). doi: https://doi.org/10.1016/j.scs.2017.06.013
2. M. Herczeg, D. McKinnon, L. Milios, I. Bakas, E. Klaassens, K. Svatikova and O. Widerberg, *Resource Efficiency in the Building Sector.* ECORYS Nederland BV, Rotterdam (2014).
3. "Woodworking - Growth - European Commission", *Growth* (2017). https://ec.europa.eu/growth/sectors/raw-materials/industries/forest-based/woodworking_en [Accessed: 14 September 2017].
4. *Technical Guideline for Fire Safety in Timber Buildings.* RISE Research Institutes of Sweden (2017). https://www.sp.se/FSITB [Accessed: 14 September 2017].
5. W.D. Ellis and R.M. Rowell, "Flame-retardant treatment of wood with a diisocyanate and an oligomer phosphonate", *Wood Fiber Sci.* **21(4),** 367–375 (1988). https://wfs.swst.org/index.php/wfs/article/view/546

6. P. Gou, E. Santos-Garcés, M. Høy, J. Wold, K. Liland and E. Fulladosa, "Feasibility of NIR interactance hyperspectral imaging for on-line measurement of crude composition in vacuum packed dry-cured ham slices", *Meat Sci.* **95(2),** 250–255 (2013). doi: https://doi.org/10.1016/j.meatsci.2013.05.013

7. A. Sandak, J. Sandak and R. Meder, "Tutorial: assessing trees, wood and derived products with near infrared spectroscopy: hints and tips", *J. Near Infrared Spectrosc.* **24(6),** 485–505 (2016). doi: https://doi.org/10.1255/jnirs.1255

8. A. Sandak, J. Sandak and M. Riggio, "Assessment of wood structural members degradation by means of infrared spectroscopy: an overview", *Struct. Contr. Health Monitor.* **23(3),** 396–408 (2016). doi: https://doi.org/10.1002/stc.1777

9. M. Schwanninger, J. Rodrigues and K. Fackler, "A review of band assignments in near infrared spectra of wood and wood components", *J. Near Infrared Spectrosc.* **19(5),** 287–308 (2011). doi: https://doi.org/10.1255/jnirs.955

10. H. Kobori, N. Gorretta, G. Rabatel, V. Bellon-Maurel, G. Chaix, J. Roger and S. Tsuchikawa, "Applicability of vis-NIR hyperspectral imaging for monitoring wood moisture content (MC)", *Holzforschung* **67(3),** 307–314 (2013). doi: https://doi.org/10.1515/hf-2012-0054

11. A. Thumm, M. Riddell, B. Nanayakkara, J. Harrington and R. Meder, "Near infrared hyperspectral imaging applied to mapping chemical composition in wood samples", *J. Near Infrared Spectrosc.* **18(6),** 507–515 (2010). doi: https://doi.org/10.1255/jnirs.909

12. T. Lestander, M. Finell, R. Samuelsson, M. Arshadi and M. Thyrel, "Industrial scale biofuel pellet production from blends of unbarked softwood and hardwood stems—the effects of raw material composition and moisture content on pellet quality", *Fuel Process. Technol.* **95,** 73–77 (2012). doi: https://doi.org/10.1016/j.fuproc.2011.11.024

13. I. Burud, K. Smeland, K. Liland, J. Sandak, A. Sandak, L. Gobakken and T. Thiis, "Near infrared hyperspectral imaging in transmission mode: assessing the weathering of thin wood samples", *J. Near Infrared Spectrosc.* **24(6),** 595–604 (2016). doi: https://doi.org/10.1255/jnirs.1253

14. R. Meder, J. Brawner, G. Downes and N. Ebdon, "Towards the in-forest assessment of Kraft pulp yield: comparing the performance of laboratory and hand-held instruments and their value in screening breeding trials", *J. Near Infrared Spectrosc.* **19(5),** 421–429 (2011). doi: https://doi.org/10.1255/jnirs.954

15. A. Thumm and M. Riddell, "Resin defect detection in appearance lumber using 2D NIR spectroscopy", *Eur. J. Wood Wood Prod.* **75(6),** 995–1002 (2017). doi: https://doi.org/10.1007/s00107-017-1188-5

16. G. Ogner, T. Wickstrøm, G. Remedios, S. Gjelsvik, G.R. Hensel, J.E. Jacobsen, M. Olsen, E. Skretting and B. Sørlie, *The Chemical Analysis Program of the Norwegian Forest Research Institute 2000.* Norwegian Forest Research Institute, Ås, Norway, p. 23 (1999).

17. H. Martens and E. Stark, "Extended multiplicative signal correction and spectral interference subtraction: new preprocessing methods for near infrared spectroscopy", *J. Pharmaceut. Biomed. Anal.* **9(8),** 625–635 (1991). doi: https://doi.org/10.1016/0731-7085(91)80188-F

18. N. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial", *Chemometr. Intell. Lab. Sys.* **117,** 92–99 (2012). doi: https://doi.org/10.1016/j.chemolab.2012.03.004

19. Å. Björck and U. Indahl, "Fast and stable partial least squares modelling: a benchmark study with theoretical comments", *J. Chemometr.* **31(8),** e2898 (2017). doi: https://doi.org/10.1002/cem.2898

20. C. Petisco, B. García-Criado, B. Vázquez de Aldana, I. Zabalgogeazcoa, S. Mediavilla and A. García-Ciudad, "Use of near-infrared reflectance spectroscopy in predicting nitrogen, phosphorus and calcium contents in heterogeneous woody plant species", *Anal. Bioanal. Chem.* **382(2),** 458–465 (2005). doi: https://doi.org/10.1007/s00216-004-3046-7

21. N. Dhawale, V. Adamchuk, R.A.V. Rossel, S. Prahser, A.A. Ismail and J.K. Whalen, "Predicting extractable soil phosphorus using visible/near-infrared hyperspectral soil reflectance measurements", *The Canadian Society for Bioengineering 2013 Annual Meeting*, Saskatoon, Saskatchewan, 7–10 July 2013, Paper No. CSBE13-047 (2013).

22. D. Gillon, C. Houssard and R. Joffre, "Using near-infrared reflectance spectroscopy to predict carbon, nitrogen and phosphorus content in heterogeneous plant material", *Oecologia* **118(2),** 173–182 (1999). doi: https://doi.org/10.1007/s004420050716

Paper IV

HYPERSPECTRAL NIR TIME SERIES IMAGING USED AS A NEW METHOD FOR ESTIMATING THE MOISTURE CONTENT DYNAMICS OF THERMALLY MODIFIED SCOTS PINE

Petter Stefansson
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Thomas K. Thiis
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Lone R. Gobakken
Norwegian Institute of Bioeconomy Research
PO Box 115, 1431 Ås

Ingunn Burud
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

ABSTRACT

The purpose of this research is to develop a method for estimating the spatially and temporally resolved moisture content of thermally modified Scots pine (*Pinus sylvestris*) using remote sensing. Hyperspectral time series imaging in the NIR wavelength region (953-2516 nm) was used to gather information about the absorbance of eight thermally modified pine samples each minute as they dried during a period of approximately 20 hours. After preprocessing the collected spectral data and identifying an appropriate wavelength selection, partial least squares regression (PLS) was used to map the absorbance data of each pine sample to a distribution of moisture contents within the samples at different time steps during the drying process. To enable separate studying and comparison of the drying dynamics taking place within the early- and latewood regions of the pine samples, the collected images were spatially segmented to separate between early- and latewood pixels. The results of the study indicate that the 1966-2244 nm region of a NIR spectrum, when preprocessed with extended multiplicative scatter correction and first order derivation, can be used to model the average moisture content of thermally modified pine using PLS. The methods presented in this paper allows for estimation and visualization of the intrasample spatial distribution of moisture in thermally modified pine wood.

Keywords: Hyperspectral imaging, thermally modified pine, TMT, hyperspectral time series, moisture content, PLS

1 Introduction

Altering the properties of timber using heat is a practice that dates back thousands of years [1]. It is well known that exposing timber to high temperatures in an oxygen deficient environment—i.e. thermally modifying it—can increase the dimensional stability of the wood and improve its resistance towards moisture-related inconveniences such as fungi and mold growth [2, 3, 4]. More recently, along with an increased demand for environmentally friendly construction materials, *thermally modified timber* (TMT) has rapidly gained popularity in applications such as claddings, decks and floors partly due to the nontoxic and eco-friendly nature of the treatment [3, 4]. The existing body of literature is however still limited with respect to how thermal modifications affect certain aspects of the moisture characteristics of wood. It is known that the equilibrium moisture content (EMC) of wood decreases when undergoing thermal modification [5, 6]. It is also known that as wood dries, gradients of varying moisture content are formed in the wood structure in both the radial, tangential and longitudinal direction [7]. Internal differences in the moisture content of a wood board cause swelling and contraction to occur at different rates within the board, which in turn leads to tensile stresses in the wood which may cause several undesired consequences: for instance, it may cause the wood, or coatings applied to the wood surface, to crack [8], or it may cause the wood to deform, sometimes permanently [7]. Methods enabling the spatial distribution of moisture within a wood board to be quantified is therefore of interest within wood sciences. Previous studies have used magnetic resonance imaging (MRI) to study the distribution of moisture within both thermally modified [9, 10] and unmodified [11] pine. The spatial resolution of MRI is however still relatively low, and such instruments are large and difficult to use outdoors to survey existing structures. Near infrared (NIR) spectroscopy has become a popular tool in wood sciences due to its ability of nondestructively allowing several useful wood properties to be characterized. Previous studies have shown that the density of wood [12], the mechanical stress of wood [13], the geographical growth region of wood [14] and its moisture content [15] can all be approximated from nondestructive NIR measurements of the sample. Using hyperspectral imaging, as opposed to traditional point based NIR measurements, allows the NIR data to be used to nondestructively approximate the spatial distribution of wood properties. Kobori et al. demonstrated that hyperspectral imaging in the vis-NIR wavelength region can be used together with multivariate regression as a viable means of determining the spatial distribution of moisture content in unmodified pine [16]. Myronycheva et al. used hyperspectral NIR imaging to exploratively study the chemical composition of thermally modified pine using principal component analysis [17].

In the present study, the feasibility of using hyperspectral imaging in the near infrared region (953-2516 nm) to estimate the moisture content distribution of thermally modified pine is evaluated. The ambition of the study is to develop a multivariate regression model capable of nondestructively estimating the spatial distribution of moisture in thermally modified pine samples based on the individual spectra of each pixel of a sample. *Partial least squares regression* (PLS) will be used to calibrate a regression vector which maps the spectra of the pine samples to a corresponding moisture content. The predicted distributions will then be spatially segmented such that separate estimates are obtained of the moisture content within the early- and latewood regions of each samples. To ensure that the developed model can accurately predict the moisture content of samples at a wide variety of different moisture contents, hyperspectral time series imaging will be used to gather spectral measurements of each sample on a minute-by-minute basis as the samples dry over the course of a day. Whilst hyperspectral cameras cannot be used to measure radial moisture variations in a sample (due to the limited surface penetration depth of the radiation), the relatively high spatial resolution

offered by such cameras allow for a detailed view of the tangential and longitudinal variations at the surface.

2 Materials & methods

2.1 Sample preparation & image acquisition

Eight thermally modified boards of Scots pine (*Pinus sylvestris*) were cut into samples of dimension $18 \times 100 \times 280$ mm. The boards were bought at a local (20 km south of Oslo, Norway) lumberyard and were manufactured by Moelven (thermally treated at 210-215 °C). After being cut, the eight samples were dried at 103 °C for 120 hours to ensure that only chemically bound moisture resided in the samples. The dryness of the samples was verified by repeatedly measuring the weight of the samples and confirming that their weight had stabilized before removing them from the oven, at which point the oven-dry weight of each sample was recorded. We assume that the moisture content at the oven-dried state corresponds to 0 %.

In addition to the dry weight, the average annual ring distance $\text{Å}_{\bar{x}}$ of each sample was measured in the radial direction (calculated according section 8 of the SKANORM 2 method [18]) and a dry density ρ_0 was calculated for every sample. When establishing the dry density of our samples, the original volume of $18 \times 100 \times 280$ mm was used as dry volume since any shrinkage which may have occurred during drying was too small for us to reliably measure.

After the dry weights were established, the samples were fully submerged in tap water for a period of approximately one and a half months. After the soaking period, the samples were one at a time taken from the water and placed on a digital scale which in turn was placed on a translation stage situated underneath a hyperspectral camera as can be seen depicted in figure 1.

The hyperspectral line scan camera (HySpex SWIR-384) situated above the sample was automated to scan each sample every minute for a period of roughly 21.5 hours as the pine dried. During the first hour or so a film of free water was still present on the surface of some samples which distorted the measured spectra and partly concealed the spectra of the pine sample. The first one hundred images

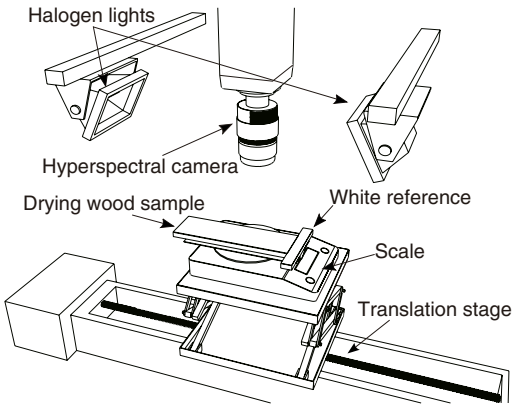


Figure 1: Illustration of experimental setup. A drying wood sample is positioned on a digital scale which in turn is situated under a hyperspectral push-broom camera and illuminated with halogen spotlights. Figure from [19].

Table 1: Dry weights, average annual ring distance ($\mathring{A}_{\bar{x}}$), dry density (ρ_0), initial moisture content (MC_{High}), final moisture content (MC_{Low}), difference between highest and lowest moisture content (MC_{Range}), average moisture content (MC_{μ}) and standard deviation of moisture content (MC_{σ}) of all samples in the study.

Sample	Dry weight	$\mathring{A}_{\bar{x}}$	ρ_0	MC_{High}	MC_{Low}	MC_{Range}	MC_{μ}	MC_{σ}
S1	188.9 g	2.8 mm	375 kg/m^3	73.7 %	44.1 %	29.5 %	56.0 %	8.2 %
S2	187.0 g	3.1 mm	371 kg/m^3	66.9 %	38.9 %	28.1 %	49.4 %	7.6 %
S3	202.2 g	2.2 mm	401 kg/m^3	80.6 %	53.2 %	27.4 %	64.2 %	7.2 %
S4	211.8 g	2.2 mm	420 kg/m^3	63.3 %	33.9 %	29.4 %	44.1 %	7.8 %
S5	185.8 g	3.7 mm	369 kg/m^3	85.4 %	47.1 %	38.3 %	61.0 %	10.3 %
S6	193.5 g	2.7 mm	384 kg/m^3	70.0 %	40.2 %	29.8 %	51.6 %	8.2 %
S7	190.7 g	3.0 mm	378 kg/m^3	77.5 %	45.3 %	32.3 %	57.2 %	8.7 %
S8	205.5 g	3.0 mm	408 kg/m^3	65.1 %	35.6 %	29.5 %	46.6 %	8.0 %

(i.e. the data from the first 100 minutes of drying) from each sample's time series was therefore removed from the dataset. The hyperspectral time series data used in the study consists in 1196 images per pine sample, depicting the samples between 1.5 and 21.5 hours of drying. The room the image acquisition/drying took place in was conditioned to be approximately 21 °C. For every image taken, the corresponding sample weight was also registered. Using the preestablished dry weight, the average moisture content (MC) of each pine sample was calculated for each time point during the drying process based on the instantaneous scale reading using the relation [16]:

$$\mathrm{MC} = \frac{w_{\text{wetwood}} - w_{\text{dryweight}}}{w_{\text{dryweight}}} \times 100 \ \ (\%) \tag{1}$$

where w_{wetwood} represents the weight of the drying pine sample and $w_{\text{dryweight}}$ represents the predetermined dry weight of the same sample. Table 1 provides a summary of each sample's recorded dry weight, average annual ring distance, dry density, together with highest, lowest, range, average and standard deviation of the calculated moisture content during the drying process.

The hyperspectral camera registered 288 equally spaced bands in the 953–2516 nm range. The spatial resolution of the region of interest of each sample was 801 × 335 pixels. The complete dimensions of the collected hyperspectral dataset is therefore 801 × 335 × 288 × 1196 × 8 (rows × columns × spectral bands × time × sample). Which equates to roughly 5.3 terabytes of spectral intensity data when stored in double-precision format. The region of interest extracted out from each hyperspectral image included roughly 87 mm of the width of each board and 209 mm of the length, centered around the middle of the board. The spatial size of each pixel corresponds to 0.227 × 0.227 mm.

The resulting structure of the experimentally collected data for each sample can be seen conceptually illustrated in figure 2: each sample is associated with its own four-dimensional hyperspectral time series as well as a one-dimensional time series of its average moisture content.

2.2 Normalization & linearization of raw spectral signal

The spectrally resolved light intensity images $\mathbf{I}(\lambda)$ registered by the hyperspectral camera were initially converted into reflectance units $\mathbf{R}(\lambda)$, relative to a *Spectralon* white reference plate included

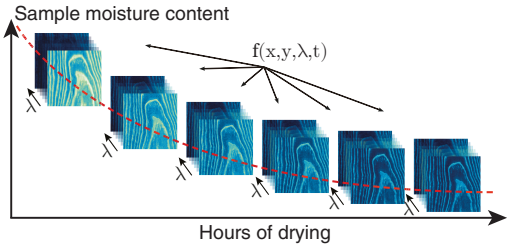


Figure 2: Illustration of hyperspectral time series data of a drying thermally modified pine sample. The spectral signal of the pine sample is resolved through both time and space. At each time step the average moisture content of the sample is known. Figure from [19].

in each image, according to

$$\mathbf{R}(\lambda) = \frac{\mathbf{I}(\lambda) - \mathbf{I}_d(\lambda)}{\mathbf{I}_0(\lambda) - \mathbf{I}_d(\lambda)}. \tag{2}$$

In Eq. 2, $\mathbf{I}_0(\lambda)$ represents the measured light intensity of the Spectralon white reference and $\mathbf{I}_d(\lambda)$ represents the dark signal of each image (signal captured with the camera shutter closed). The reflectance images were then transformed into apparent absorbance $\mathbf{A}(\lambda)$ in accordance with Lambert Beer's law [20]:

$$\mathbf{A}(\lambda) = \log_{10}(1/\mathbf{R}). \tag{3}$$

2.3 Regression & data division

A partial least squares regression model was calibrated which took the hyperspectral time series data as input ($\mathbf{X}$) and mapped each spectrum to the corresponding scalar moisture content of the sample (y), i.e. the sample-average moisture content at a specific time step. The average moisture content of each sample for each time step—the response values of the dataset—are shown in figure 3. During the PLS calibration the average spectrum of each hyperspectral image was used as input.

To test the developed model's ability to generalize to new unseen data, the hyperspectral time series of two samples, S5 and S2, were randomly chosen and withheld from the calibration procedure. After calibrating the model on the remaining six samples' time series the model was applied to the data from the two withheld time series (2×1196 images) to validate its performance on new data. In total, 7176 hyperspectral images were included in the training set and 2392 images were included in the validation set.

To enhance the performance of the PLS model, spectral preprocessing and wavelength selection was applied to the measured absorbance data. In order to identify a spectral preprocessing technique that would yield a low prediction error in the PLS regression, a grid search over different common NIR preprocess methods was therefore conducted. In addition to the unprocessed absorbance spectra, the methods included in this search were: *Savitzky-Golay derivation* [21] of first, second and third order, *Multiplicative Scatter Correction* [22] (MSC), *Standard Normal Variate* [23] (SNV) and *Extended Multiplicative Scatter Correction* [24] (EMSC) as well as pairwise combinations of these methods. For each of the preprocessing methods a PLS model was calibrated and cross-validated with 10-fold cross-validation. The preprocessing technique resulting in the lowest *cross-validated root-mean-squared-error* (RMSEcv) was chosen for the final model.

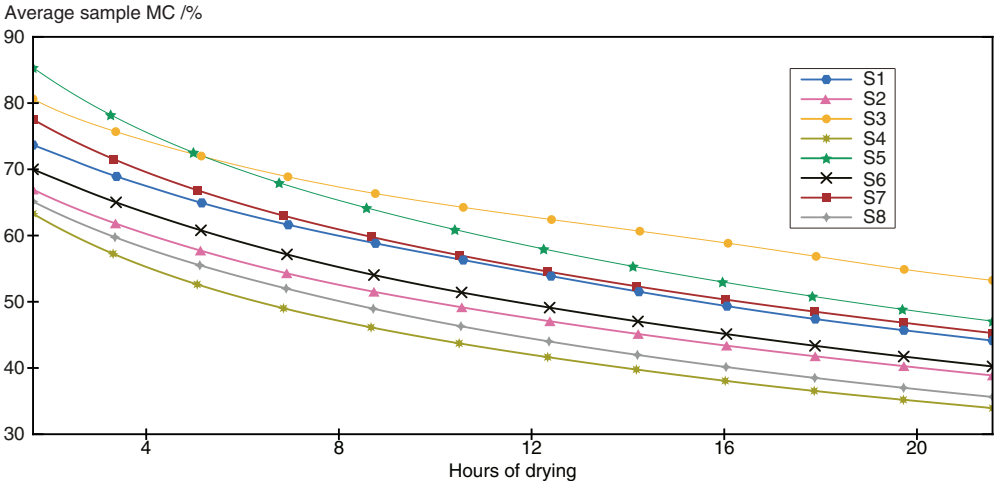


Figure 3: Calculated average moisture content of all thermally modified pine samples in the study during the drying period.

Once a suitable preprocessing technique was identified, the preprocessed data was subjected to variable selection in order to further enhance the model's performance by eliminating irrelevant or noisy wavelengths. *Forward selection*, *backwards elimination*, *interval PLS* [25] (iPLS) in backwards mode and *moving window* variable selection [26, 27] (MW) were applied to the data. When applying interval PLS and moving window variable selection every interval/window width between 1 and n was tested, where n denotes the total number of variables in the spectra. Because this wavelength selection search required many thousands of PLS models to be calibrated and evaluated, the feature selection calibrations were performed using the kernel PLS feature selection technique introduced by Stefansson et al. [28] in order to speed up the feature selection process.

Once a combination of preprocessing technique and feature selection had been identified the final PLS model was calibrated using the *bidiag2* [29] algorithm. All modeling was performed in MATLAB 2019a [30].

2.4 Early-/latewood image segmentation

To allow for separate estimates to be obtained of the moisture content within early- and latewood regions of a sample during the drying process, the pixels of each hyperspectral image was categorized according to wood type belonging. Wood is a notoriously inhomogeneous material and the seasonal growth patterns takes many, sometimes relatively complex, forms which makes manual segmentation tedious and nontrivial. To enable semi-automatic segmentation of the dataset we employed the principal component analysis-based segmentation technique introduced by Smeland et al. [31] for discriminating between early- and latewood pixels within our hyperspectral images. The method consists in performing *principal component analysis* (PCA) on a hyperspectral image and then forming a histogram from the resulting scores associated with one of the principal components. Two thresholds are then placed in the histogram, and datapoints below the lower threshold are classified as earlywood whereas datapoints above the higher threshold are classified as latewood.

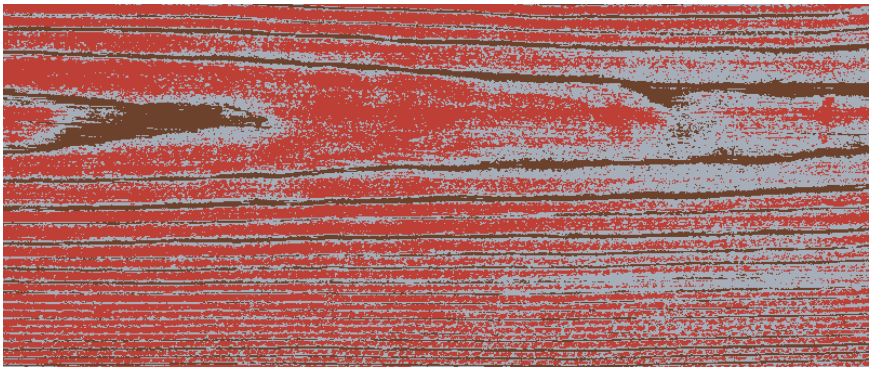


Figure 4: Spatial early-/latewood segmentation of one of the samples in the study (sample S7). Brown color indicates pixels classified as latewood and red color indicates pixels classified as earlywood. Gray color indicates intermediate wood which was not treated as either early- or latewood.

3 Results

3.1 Segmentation

Smeland et al. [31] found that the wood segmentation algorithm they developed worked best using the second principal component from PCA and suggested that thresholds be positioned at the 25th and 65th percentile in the scores histogram. For our dataset however, we found that the first principal component worked better than the second and that the percentile thresholds needed to be tweaked manually for each sample in order to adequately approximate the early- and latewood distribution observed by studying the samples visually. An example of a generated image segmentation mask can be seen in figure 4. In the figure, red color indicates earlywood, brown color indicates latewood, the intermediary region between the two classes which was not considered as either early- or latewood is shown in gray. Since our samples were kept stationary during the time series acquisition and only negligible contraction of the samples was found to take place during the drying process, the segmentation was performed only once per sample and then applied to all images within the time series. The segmentation was performed using the last image of each series, i.e. the image corresponding to the driest sample state.

3.2 PLS regression modeling

The grid search over spectral preprocessing techniques indicated that a combination of extended multiplicative scatter correction followed by first order Savitzky-Golay derivation yielded the lowest cross-validated prediction error. During EMSC the *basic EMSC model* [32] was used, which entails a model containing an intercept term, slope term, linear term and a quadratic term. The average spectrum from the training dataset was used as a reference spectrum in the EMSC correction. The Savitzky-Golay derivation was carried out with a window size of seven and a polynomial degree of one.

The best performing variable selection was identified using the moving window algorithm. The region found by the algorithm consisted in 52 wavelengths between 1966 nm and 2244 nm. Figure 5 shows the average spectrum of every hyperspectral image in the dataset after preprocessing along the region identified during wavelength selection.

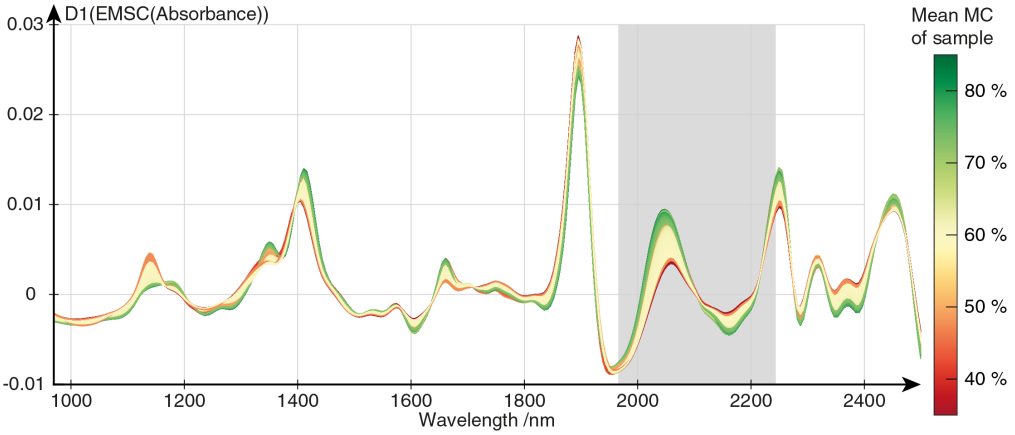


Figure 5: Mean absorbance spectrum of every time series image in the collected dataset preprocessed with basic EMSC followed by first order Savitzky-Golay derivation. Gray region indicates wavelength region identified by the moving window feature selection algorithm. All spectra in the figure are colored according to the average of moisture content of the sample they originate from.

Using the combination of identified preprocessing and wavelength selection, the lowest $RMSE_{cv}$ (and first local minima) after 10-fold cross-validation was obtained using nine PLS components, which was subsequently used when calibrating the final PLS model. Figure 6 shows a regression plot of PLS-modeled vs. average measured moisture content of each image in the eight time series sequences. Blue dots indicate data originating from any of the six training time series, red squares indicate data from the two validation time series. The model's root-mean-squared-error, RMSE, on the training data was 2.1 %, with a coefficient of determination, R^2, of 0.98. Applying the model to the validation data resulted in a RMSE of 2.7 % and a R^2 of 0.97. For most samples in the study, the discrepancy between modeled MC and sample-average MC was larger during the first few hours of drying compared to the rest of the drying sequence.

3.3 PLS modeled spatial distribution & temporal development of moisture content

Figure 7 shows chemical maps of the PLS-estimated spatial distribution of moisture content for every sample in the study obtained by applying the PLS model to the full resolution hyperspectral data. The upper row depicts the samples at the start of their time series, i.e., after 100 minutes of drying. The lower row depicts the same samples at the end of their time series, i.e., after 21.5 hours of drying. Some samples, such as S4, S6, S7 and S8, can be seen in the figure to have a locally higher moisture content at the top of the sample, indicating a slower rate of drying at the top. A possible explanation for this could be that in our experimental setup the Spectralon white reference plate is located at the top edge of the sample—potentially hindering drying to freely take place there.

Figure 8 shows the average PLS-estimated early- and latewood moisture content for every image in the dataset displayed as eight individual time series/drying curves. These curves were obtained by applying the developed early-/latewood segmentation masks of each sample onto the modeled spatial distribution of moisture content for every time step in the series. The moisture content estimates for all early- and latewood pixels were then separately averaged to form two drying curves per sample; one containing the estimated drying curve for latewood and the other for earlywood.

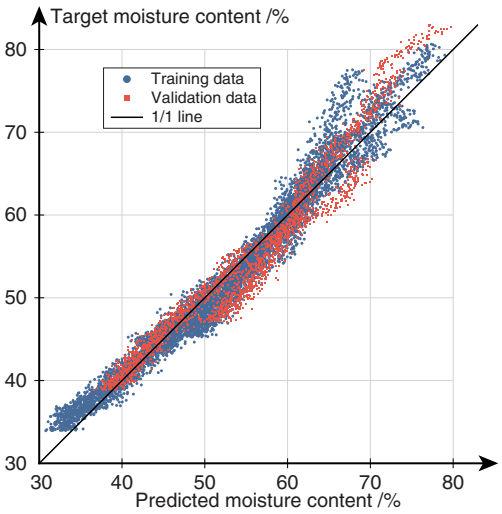


Figure 6: Regression plot of PLS modeled vs. measured mean moisture content for every image of the dataset. Blue dots represent images belonging to the training data, red squares represent images belonging to the validation data.

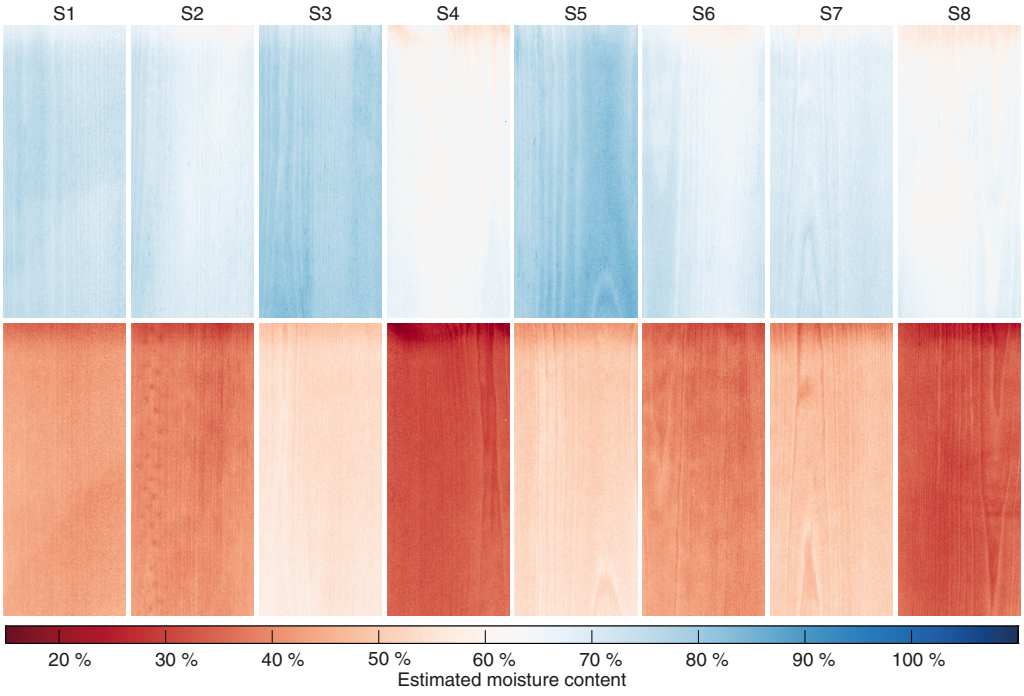


Figure 7: Spatial distribution of PLS-estimated moisture content for every sample in the study. Upper row depicts the samples at the initial stages of drying, lower row depicts the same samples approximately 20 hours of drying later.

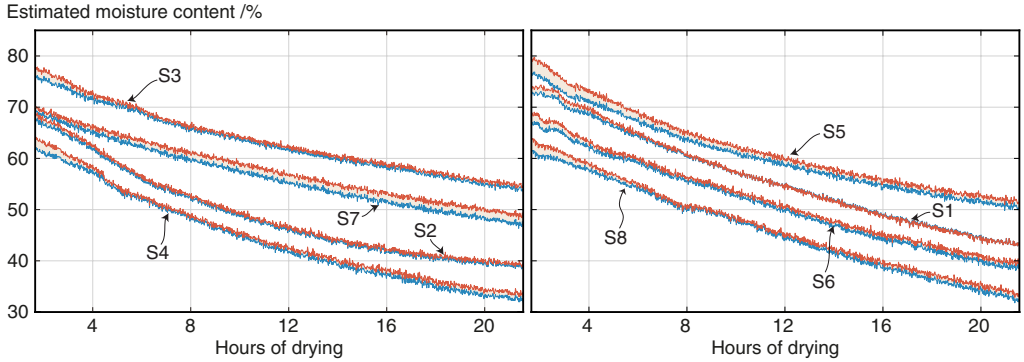


Figure 8: Temporal development of predicted moisture content of all samples (S1-S8). In each subplot the orange lines indicate the average modeled moisture content in all earlywood pixels of a sample, blue lines indicate the average moisture content in the latewood pixels of a sample.

During the first few hours of drying the estimated moisture content of the earlywood was noticeably higher than the latewood estimates for all samples. Averaged across all samples the estimated moisture content was 1.7 % higher in the earlywood than the latewood at the beginning of the time series and decreased over time down to a 0.8 % difference at the end of the drying process. The samples with the lowest estimated moisture difference between early- and latewood, S2, S1 and S3, all have a fine grained early- and latewood structure with densely packed growth regions as can be seen in figure 7. It is also interesting to note that the sample with the greatest radial annual ring distance of the sample set, S5, also had the largest estimated moisture differential between early- and latewood (2.9 % during the initial stages of drying).

4 Discussion & conclusions

Our developed PLS model, calibrated on six time series consisting of 7176 hyperspectral images in total, proved capable of estimating the average moisture content of thermally modified pine with a high degree of accuracy. Extended multiplicative scatter correction combined with first order derivation proved useful in relieving light scatter from the spectra and enhancing the correlation between spectra and moisture content. The wavelength region 1966-2244 nm was found to hold strong predictive capacity over the moisture content within the pine. This region includes several wavelengths which are known to cause absorption in free water at room temperature, although it slightly misses the absorption peak which occurs around 1930-1950 nm [33]. Fujimoto et al. [34] found 1980 nm to be a region representative of water absorption in larch wood, which is included in our identified region. Our model produced the best results at nine principal components, which is rather high. Kobori et al. [16] found six PLS components to be optimal for estimating MC in unmodified pine using vis-NIR hyperspectral spectral data—thus, both experiments indicate that a surprisingly high number of latent variables is beneficial when estimating the moisture content of pine using PLS regardless of thermal treatment.

By superimposing a segmentation mask onto the PLS-estimated spatially and temporally resolved distributions of moisture content, our presented method allows separate estimates to be obtained of the MC of early- and latewood regions within a board during drying. However, as is often the case with studies such as this one—where a model is trained using a measured sample-average response value and later used to estimate the spatial distribution of the response throughout the sample—a major limitation is that the chemical maps generated by the regression model cannot easily be validated;

since the true pixel-by-pixel distribution of moisture is unknown to us. We therefore cannot conclude that the spatial predictions are accurate, only that the spatial estimates appear realistic upon visual inspection and that the PLS model is capable of estimating the average moisture content of a thermally modified pine sample based on the average spectra of the sample. Further studies should therefore investigate the possibility of training a regression model with hyperspectral input data together with spatially resolved response values—obtained for instance by using magnetic resonance imaging on the same samples as are scanned with a hyperspectral camera.

Despite the successfulness of using surface reflected visible and near-infrared radiation to model the moisture content of wood samples, demonstrated in this study for thermally modified pine as well as in [16] for unmodified pine, it is important to note that such models rest on the assumption that the moisture content is consistent throughout the thickness of the sample. This assumption is of course never entirely valid, and it likely increases in invalidity when thicker wood samples are used with a more inhomogeneous internal annual ring structure. To circumvent this issue entirely remote sensing technologies with a greater penetration depth, such as MRI, are necessary. To lessen the effects of this limitation in our own experimental setup, thinner samples could have been used in our experiment—such that there is less room for unobserved radial moisture variations in the sample. In some preliminary experiments we conducted prior to the experiment presented here however, we monitored thinner samples with hyperspectral imaging and found that although it certainly seems possible to estimate the moisture content of such samples, the drying dynamics of thin samples ($\approx$ 2-3 mm thickness) appears substantially different from that of thicker boards ($\approx$ 2 cm thickness). When using thin samples the moisture evaporated rapidly around the edges of the sample when exposed to the warm halogen light of the experimental setup. Our motivation for choosing larger samples in this experiment is that we wanted a slower, more controlled, drying process.

References

[1] SP Technical Research Institute of Sweden, *Benchmarking and State of the art for Modified wood*.

[2] D. Sandberg and A. Kutnar, "Thermally modified timber: Recent developments in europe and north america," *Wood and Fiber Science*, vol. 44, pp. 28–39, 2016.

[3] D. Cirule, A. Meija-Feldmane, E. Kuka, B. Andersons, N. Kurnosova, A. Antons, and H. Tuherm, "Spectral sensitivity of thermally modified and unmodified wood," *BioResources*, vol. 11, no. 1, 2015.

[4] E. Dunningham and R. Sargent, *Review of new and emerging international wood modification technologies*, 2015.

[5] C. A. S. Hill, *Wood modification*. Wiley, 2006.

[6] B. Esteves and H. Pereira, "Wood modification by heat treatment: a review," *BioResources*, vol. 4, no. 1, 2009.

[7] C. P. Edward, *How wood shrinks and swells*, 1957.

[8] P. A. Schweitzer, *Atmospheric degradation and corrosion control*, 1st ed. Marcel Dekker, 1999.

[9] P. M. Kekkonen, A. Ylisassi, and V.-V. Telkki, "Absorption of water in thermally modified pine wood as studied by nuclear magnetic resonance," *The Journal of Physical Chemistry C*, vol. 118, no. 4, pp. 2146–2153, 2014.

[10] M. A. Javed, P. M. Kekkonen, S. Ahola, and V.-V. Telkki, "Magnetic resonance imaging study of water absorption in thermally modified pine wood," *Holzforschung*, vol. 69, no. 7, pp. 899–907, 2015.

[11] S. Hameury and M. Sterley, "Magnetic resonance imaging of moisture distribution in pinus sylvestris l. exposed to daily indoor relative humidity fluctuations," *Wood Material Science and Engineering*, vol. 1, no. 3-4, pp. 116–126, 2006.

[12] T. Fujimoto, H. Kobori, and S. Tsuchikawa, "Prediction of wood density independently of moisture conditions using near infrared spectroscopy," *Journal of Near Infrared Spectroscopy*, vol. 20, no. 3, pp. 353–359, 2012.

[13] J. Sandak, A. Sandak, D. Pauliny, V. Krasnoshlyk, and O. Hagman, "Near infrared spectroscopy as a tool for estimation of mechanical stresses in wood," *Advanced Materials Research*, vol. 778, pp. 448–453, 2013.

[14] A. Sandak, J. Sandak, and M. Negri, "Relationship between near-infrared (nir) spectra and the geographical provenance of timber," *Wood Science and Technology*, vol. 45, no. 1, pp. 35–48, 2010.

[15] K. Watanabe, S. D. Mansfield, and S. Avramidis, "Application of near-infrared spectroscopy for moisture-based sorting of green hem-fir timber," *Journal of Wood Science*, vol. 57, no. 4, pp. 288–294, 2011.

[16] H. Kobori, N. Gorretta, G. Rabatel, V. Bellon-Maurel, G. Chaix, J.-M. Roger, and S. Tsuchikawa, "Applicability of vis-nir hyperspectral imaging for monitoring wood moisture content (mc)," *Holzforschung*, vol. 67, no. 3, 2013.

[17] O. Myronycheva, E. Sidorova, O. Hagman, M. Sehlstedt-Persson, O. Karlsson, and D. Sandberg, "Hyperspectral imaging surface analysis for dried and thermally modified wood: An exploratory study," *Journal of Spectroscopy*, vol. 2018, pp. 1–10, 2018.

[18] K. Bohumil, *Skandinaviske normer for testing av små feilfrie prøver av heltre.*, 1992.

[19] P. Stefansson, J. Fortuna, H. Rahmati, I. Burud, T. Konevskikha, and H. Martens, *Hyperspectral time series analysis: Hyperspectral image data streams interpreted by modeling known and unknown variations. In Hyperspectral Imaging, volume 32.*, 1st ed. Elsevier, 2019.

[20] Å. Rinnan, F. v. d. Berg, and S. B. Engelsen, "Review of the most common pre-processing techniques for near-infrared spectra," *TrAC Trends in Analytical Chemistry*, vol. 28, no. 10, pp. 1201–1222, 2009.

[21] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[22] P. Geladi, D. MacDougall, and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat," *Applied Spectroscopy*, vol. 39, no. 3, pp. 491–500, 1985.

[23] R. J. Barnes, M. S. Dhanoa, and S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Applied Spectroscopy*, vol. 43, no. 5, pp. 772–777, 1989.

[24] H. Martens and Stark, "Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 9, no. 8, pp. 625–635, 1991.

[25] L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen, "Interval partial least-squares regression (ipls): A comparative chemometric study with an example from near-infrared spectroscopy," *Applied Spectroscopy*, vol. 54, no. 3, pp. 413–419, 2000.

[26] S. Fang, M.-Q. Zhu, and C.-H. He, "Moving window as a variable selection method in potentiometric titration multivariate calibration and its application to the simultaneous determination of ions in raschig synthesis mixtures," *Journal of Chemometrics*, vol. 23, no. 3, pp. 117–123, 2009.

[27] J.-H. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, "Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data," *Analytical Chemistry*, vol. 74, no. 14, pp. 3555–3565, 2002.

[28] P. Stefansson, U. G. Indahl, K. Liland, and I. Burud, "Orders of magnitude speed increase in partial least squares feature selection with new simple indexing technique for very tall datasets," *Journal of Chemometrics (accepted, in print)*, 2019.

[29] Å. Björck and U. G. Indahl, "Fast and stable partial least squares modelling: A benchmark study with theoretical comments," *Journal of Chemometrics*, vol. 31, no. 8, p. e2898, 2017.

[30] MATLAB, *version 9.6.0 (R2019a)*. The MathWorks Inc., 2019.

[31] K. A. Smeland, K. H. Liland, J. Sandak, A. Sandak, L. R. Gobakken, T. K. Thiis, and I. Burud, "Near infrared hyperspectral imaging in transmission mode: Assessing the weathering of thin wood samples," *Journal of Near Infrared Spectroscopy*, vol. 24, no. 6, pp. 595–604, 2016.

[32] N. K. Afseth and A. Kohler, "Extended multiplicative signal correction in vibrational spectroscopy, a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 117, pp. 92–99, 2012.

[33] J. A. Curcio and C. C. Petty, "The near infrared absorption spectrum of liquid water," *Journal of the Optical Society of America*, vol. 41, no. 5, p. 302, 1951.

[34] T. Fujimoto, Y. Kurata, K. Matsumoto, and S. Tsuchikawa, "Application of near infrared spectroscopy for estimating wood mechanical properties of small clear and full length lumber specimens," *Journal of Near Infrared Spectroscopy*, vol. 16, no. 6, pp. 529–537, 2008.

Paper V

Hyperspectral time series analysis: Hyperspectral image data streams interpreted by modeling known and unknown variations

In print

Petter Stefansson
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

João Fortuna
Department of Engineering Cybernetics
Norwegian University of Science and Technology NTNU
7034 Trondheim Norway

Hodjat Rahmati
Idletechs AS
7010 Trondheim Norway

Ingunn Burud
Faculty of Science and Technology
Norwegian University of Life Sciences NMBU
Drøbakveien 31, 1430 Ås

Tatiana Konevskikha
Department of Fundamental Mathematics
Perm State University PSU
Bukirev street 15, 614990 Perm

Harald Martens
Department of Engineering Cybernetics
Norwegian University of Science and Technology NTNU
7034 Trondheim Norway

Abstract

In the present chapter, we experimentally demonstrate a generic method for compact quantification and interpretation of multichannel spatiotemporal data ("hyperspectral video") in terms of known and unknown variation types. The process of drying a wet wood sample was characterized by a series of high-resolution hyperspectral images taken at 150 consecutive time steps over a period of 21 hours. Each pixel intensity was measured at 159 wavelength channels in the vis-NIR region, which resulted in a time series of approximately half a terabyte of raw spectral data. Passing the massive stream of data through a four-stage data modeling procedure resulted in

a substantially compressed 10 component bilinear model comprised of five a priori known and five newly discovered spectral components. From this compressed subspace model, a filtered version of the original data stream could be reconstructed. First, the measured intensity spectra were normalized by transformation into reflectance and linearized into apparent absorbance units. These absorbance spectra were secondly submitted to simplified causal modeling of known phenomena by *Extended Multiplicative Signal Correction* (EMSC) - to identify variations in the main light scattering and light absorption variation types. Thirdly, the high-dimensional stream of lack-of-fit residual spectra from EMSC were analyzed for possible remaining systematic structures, e.g. due to unknown and hence unmodeled variation types by an adaptive bilinear modeling method (ABLM). In the final modeling stage, the dynamics of the various known and unknown physical and chemical variations of the drying process were assessed from their temporal developments.

Keywords Hyperspectral time series · EMSC · OTFP · Light scattering · Light absorbance · wood · drying · image analysis · multivariate

1 Introduction

Hyperspectral time series and other multichannel spatiotemporal spectral measurement processes give overwhelming streams of Quantitative Big Data. The raw data are non-selective, in the sense that they are affected by many different sources of variation - variations in sample physics (e.g. light scattering) and sample chemistry (composition) as well as variations in the light source and the camera (systematic errors). To identify, separate, quantify and interpret the various sources of information in such data streams is thus a challenge. It is important to quantify phenomena already known for the given type of measurements, as well as also discovering, quantifying and displaying unexpected, but systematic variation patterns in the data. Here we demonstrate how a combination of known and unknown physical, chemical and instrumental variation types can be discovered and summarized in a combination of mechanistic and empirical multivariate data models.

Diffuse multi-channel reflectance spectroscopy in e.g. the visible (vis) and near-infrared (NIR) wavelength range is a fast and informative methodology for simultaneous measurement of a range of chemical and physical properties in complex biological samples [1, 2]. However, in order to resolve selectivity problems of chemical and physical variations with similar spectral effects, multivariate calibration is required [3]. In hyperspectral *imaging*, where each pixel in an image is represented by a spectrum of reflected light, the spatial distribution of these properties can be quantified in a heterogeneous sample such as a piece of wood, using mathematical modeling and Multivariate Image Analysis (MIA) [4].

When a given sample is measured repeatedly *over time* by hyperspectral imaging, the resulting "hyperspectral video" also provides information about temporal developments. The multichannel spatiotemporal measurements in hyperspectral video generate really Big Data. In this chapter a piece of Scots pine was monitored using hyperspectral imaging in the vis-NIR wavelength region for a period of 21 hours as the wood underwent desorption from a moisture saturated state to a dry state. We use statistical data driven models to estimate the temporal development of properties within the wood as it dries and demonstrate how a seemingly overwhelming stream of hyperspectral video data can be converted into relatively simple quantitative spatiotemporal information by a combination of various pragmatic mathematical modeling techniques.

Light's interaction with complex materials like wood is dominated by two phenomena - chemical light absorption and physical light scattering - both of which are fairly well understood: Due to electronic or molecular resonances in chemical compounds, light at a given wavelength is absorbed – i.e. converted into heat, or converted into light emitted at other wavelengths. Changes in the degree of absorbed light between samples are useful for quantification of their chemical composition. Variations in the physical properties of samples (e.g. the amount, size, shape and refractive index of particles) cause variations in several light scattering phenomena (angular distribution of reflected or transmitted light, effective optical path length, specular surface reflectance etc.); measuring these is useful for quantification of the samples' physical properties.

The present chapter analyzes a stream of closely related hyperspectral images, representing a hyperspectral time series analyzed with respect to the common underlying spectral variation patterns. In addition, since the wood sample is monitored over time in a fixed position, the spatial and temporal structure of these variation patterns can be studied qualitatively and quantitatively.

In theory, causal mathematical modeling of how these phenomena affect light measurements is therefore possible: For instance, light absorption and light scattering can be modeled via the complex refractive index [5]. However, such causal modeling requires detailed information about chemical and physical structures in the samples measured, and this information may not be available. Moreover, it may require a measurement set-up and/or extensive measurements involving simultaneous transmittance/reflectance measurements in an integrating sphere etc. For routine analysis, such quantification is usually too cumbersome, too slow or too expensive.

Instead, this chapter employs a combination of four simple mathematical approximation model stages:

1. Normalization and linearization of the raw data to facilitate subsequent linear modeling according to Beer-Lambert's Law.

2. Modeling known structures: Linear additive/multiplicative model of an extended version of Beer-Lambert's Law (*Extended Multiplicative Signal Correction*, EMSC [6, 7]) to quantify and extract known chemical and physical variations.

3. Modeling unknown structures: Bilinear search for unknown and hence unmodeled variation types in the spectral residuals after the EMSC based on Adaptive Bi-Linear Modeling (ABLM) using *On-The-Fly-Processing* (OTFP) [8].

4. Temporal kinetics modeling of the known and unknown state variables (component time series) obtained from the EMSC and ABLM based data-models of the hyperspectral video measurements.

The outline of the chapter is as follows: First we describe the actual experiment and the mathematical modeling methods employed to analyze the results. Then we summarize the results: First a preliminary, over-all assessment of the complexity of the drying process is given, based on the reduction of the wood sample's weight. Then the entirety of the hyperspectral measurements are analyzed with respect to known and unknown phenomena, based on the EMSC and OTFP methods, respectively. Finally, the kinetics of the EMSC and OTFP temporal parameters are assessed.

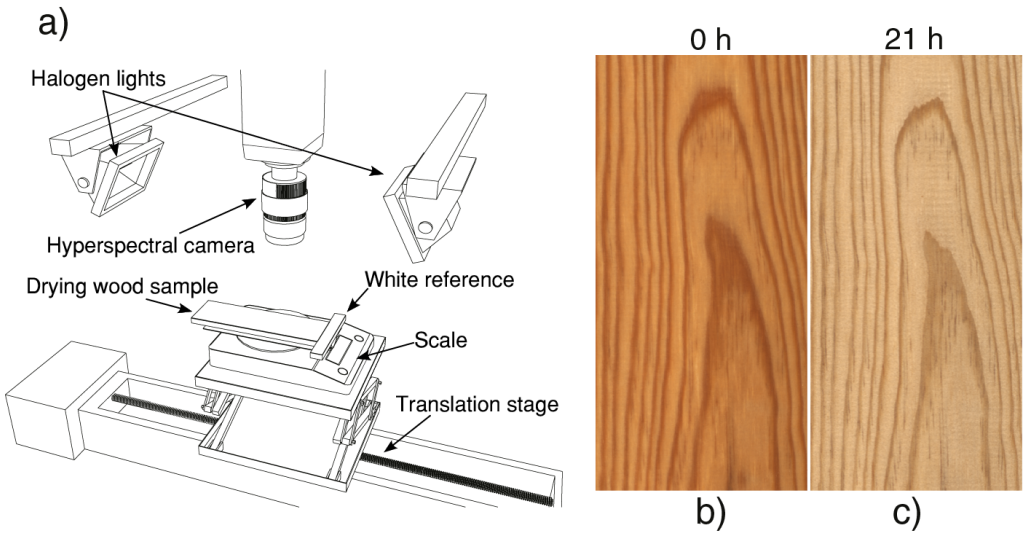


Figure 1: The experiment. a) Illustration of experimental setup used to measure the spectral reflectance and weight of a drying wood sample. b) RGB rendering of wood sample in wet state (drying time = 0 hours). c) RGB rendering of wood sample in dry state (drying time = 21 hours).

2 The experiment

2.1 Drying of a wood sample

A wood sample of the species Scots pine (*Pinus sylvestris*) originating from a forest in Hobøl, roughly 30 km south of Oslo, Norway, was cut into dimensions $18 \times 100 \times 280$ mm. The sample was then placed in a drying oven where it was dried at a temperature of 103 $^\circ$ C for 48 hours until it was ensured, through repeated weight measurements, that as much as possible of the water in the material had been evaporated. The sample was then taken from the oven and immediately weighed in order to have its dry weight determined to be 245.46 g. Once the sample's dry weight was established, the sample was submerged in water and left to soak for approximately 24 hours. After the soaking period, the sample was removed from the water and placed on a digital scale. Its initial wet weight was 336.51 g. The digital scale with the wood sample and a *Spectralon* white reference, was attached to a translation stage as can be seen illustrated in Fig. 1 a).

The wood sample was artificially illuminated by two halogen spotlights positioned on either side of a hyperspectral camera which monitored the sample for a period of about 21 hours. From the weight loss of the wood sample registered by the scale, the relative moisture content could be calculated at different time points and used for a preliminary assessment of the kinetic complexity of the drying process.

2.2 Monitoring the drying process by hyperspectral time-lapse camera ("hyperspectral video")

The hyperspectral line scan camera (Specim, Oulu, Finland) captured images with a spatial resolution of 4480×1312 pixels (pixel dimension 69.2×68.6μm), each characterized by 200 wavelength bands in the 392-1022 nm range. Because bands at the edge of the cameras detection limit were found to suffer from a low signal-to-noise ratio, the wavelength range was cropped after the signal acquisition down to 159 bands covering the 500-1005 nm region. The software controlling the experimental setup was programmed to automatically repeat the image acquisition of the sample every eight minutes over a period of about 21 hours as the sample and the setup remained completely untouched, resulting in 150 hyperspectral frames of the sample at various moisture contents. The spatial resolution of the region of interest within each frame (area excluding white reference plate, wood edges, etc.) was 2200×1070 pixels, resulting in a four-dimensional dataset of size 2200×1070×159×150 (rows×columns×spectral bands×time). Altogether, this corresponds to roughly 418 GB of data when stored in double-precision floating-point format. The sample was not moved during the 21 hours, therefore the pixels of each hyperspectral image are expected to correspond to the same location of the sample, apart from a minor offset caused by the contraction of the wood sample as it dries out. Thus the resulting data set can be seen as four-dimensional, with each one-dimensional pixel spectra being a function of both space and time as can be seen illustrated in Fig. 2 a).

The recorded photon count ("light intensity" $\boldsymbol{I}$) in each of the 150×159 images has 2D spatial information about the physical and chemical structure of the wood sample; the distinction between early- and latewood growth zones (lighter respectively darker regions of the wood) are clearly visible. These zones reflect seasonal fluctuations in the growth rate of the tree during its lifetime due to variations in temperature and precipitation, resulting in chemical and physical differences in wood structure. This sample heterogeneity is symbolized by the two wood sub-images inserted in Fig. 1 b-c for illustration.

In order to use the massive stream of high-resolution hyperspectral images to study the physics and chemistry of the wood drying process, these measured intensity data $\boldsymbol{I}$ were passed through several mathematical modeling stages: 1. Response linearization, 2. Semi-mechanistic multivariate modeling of known effects, 3. Data-driven multivariate modeling of remaining unknown effects, and finally, 4. Statistical summary and kinetic modeling of the dynamics of the known and unknown effects.

3 Mathematical modeling stages

3.1 Weight-based assessment of the kinetic complexity of the drying process

The weight-based average moisture content ($\mathbf{water}_{\%}$) in the wood sample was calculated as

$$\mathbf{water}_{\%} = \frac{\mathbf{w}_{\text{wood}} - 245.46\ g}{245.46\ g} \times 100\ (\%). \tag{1}$$

Where 245.46 represents the weight of the wood sample dried at for 21 hours and $\mathbf{w}_{\text{wood}}$ represents the varying wood weight during the drying period. In order to assess the

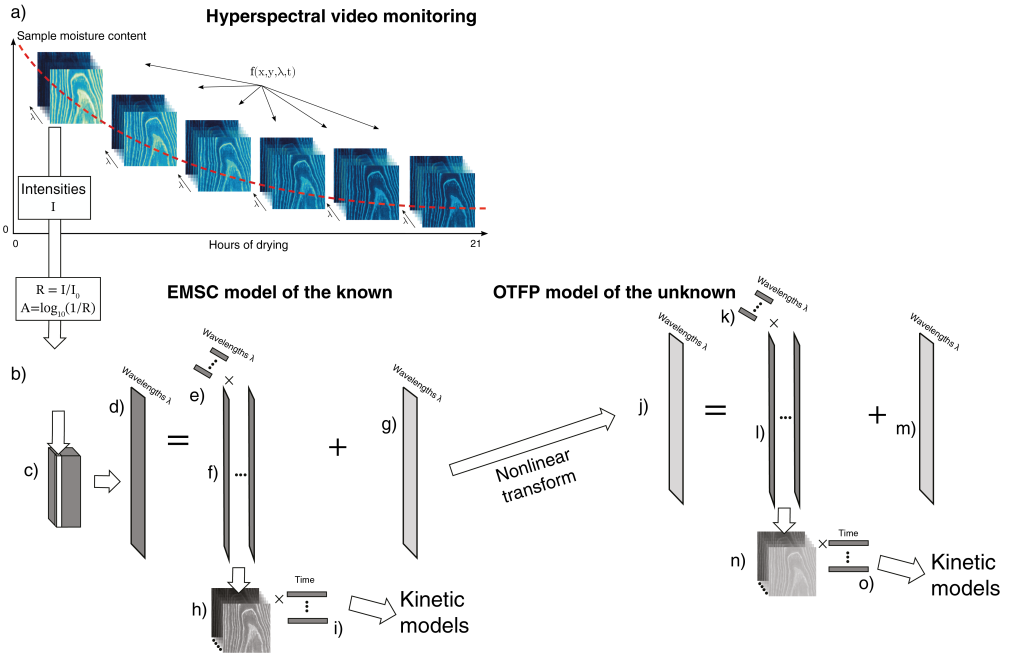


Figure 2: Overview of experimental data acquisition and modeling of hyperspectral video. a) Input data: (2200 × 1070) pixels × 159 wavelength channels × 150 time points. b) - i) Model what is known about input data: EMSC modeling of 2-way input data for 159 wavelength channels at 353 100 000 pixels (2200 × 1070 × 150) × 159 wavelengths, and spatiotemporal averaging. j) - o) Model what is unknown: Adaptive bilinear modeling of 2-way residual data for 22 353 100 000 pixels × 159 wavelengths.

over-all complexity of the drying process, the temporal derivative of **water**$_{\%}$ was calculated: For $y_t = water_{\%,t}$ at time t, its temporal derivative dy_t/dt at time #t was calculated as

$$\frac{dy_t}{dt} = \frac{y_{t+1} - y_{t-1}}{h_{t+1} - h_{t-1}} \tag{2}$$

where h is the actual time at time point #t (in hours). The derivative dy_t/dt was then plotted against $\boldsymbol{y}$ itself, in order to assess the kinetics of the weight-based drying process.

Since the drying process appeared to be quite complex, a simpler alternative was also employed, namely a graphical assessment of how well the drying loss dynamics followed the simple first order reaction:

$$\frac{dy_t}{dt} = -k \times y_t \tag{3}$$

where k is the rate constant. Integrating this equation over time gives the following expression in natural logarithms (ln):

$$\ln(y_t) = -k \times t + \ln(y_{t=0}) \tag{4}$$

Hence, a plot of $\ln(y_t)$ vs time t (here: in hours) should give a straight line if the process follows first order kinetics. Observed deviations from a single straight line indicate a more complex drying process.

3.2 Hyperspectral modeling

A hyperspectral video is a 3-way data set (wavelength×channels×pixels×time points) that, in principle, could have been modeled by a tensor-algebraic 3-way model, e.g. a PARAFAC-model [9] of the type

$$A = B \otimes C \otimes D + E \tag{5}$$

where A is the 3-way video input data, B, C and D represent a low-dimensional model with vectors in the wavelength-, pixel- and time-domains, respectively, and E represents measurement noise.

However, the sheer amount of input data makes this N-way modeling too computer intensive. Moreover, the light scattering varies with the drying time (conf. Fig. 1 b-c), and this is likely to involve changes in the effective optical path length. Hence, since path length variations give non-additive effects, this purely additive 3-way modeling was discarded.

Instead, the following sequence of theory-driven and data-driven modeling steps (Fig. 2) were chosen here.

3.2.1 Response linearization

The light intensities $\boldsymbol{I}$ from the wood sample, measured at each of the wavelengths for each of the pixels at each of the points in time Fig. 2 a) were converted to reflectance units by $R = (I - I_d)/(I_0 - I_d)$, where I_0 represents the intensity of a spectralon white reference plate and I_d is the dark signal (image taken with the shutter closed). The reflectance data were in turn linearized with respect to chemical response by the conventional transformation to apparent absorbance $A = \log_{10}(1/R)$. After this response normalization and linearization of each of the $> 5 \times 10^{10}$ individual light intensity readings, the 2D image of apparent absorbance at each wavelength channel at each point in time Fig. 2 b) is unfolded into a 1D column vector. This resulted in a virtual 3-way array of absorbance Fig. 2 c) with 2,354,000 pixels per image and 159 wavelengths × 150 time points.

This array was unfolded into a virtual 2-way absorbance matrix Fig. 2 d) of 353,100,000 pixels × 159 wavelengths which was submitted to three more stages of modeling:

1. **Modeling the known:** The 159-dimensional spectrum of each of the 353,100,000 pixels was submitted to a semi-causal modeling of what is known about how light interacts with matter in complex samples such as wood, based on the Extended Multiplicative Signal Correction (EMSC) model [6, 7].

   In total five different phenomena with known spectra were modeled, in order to estimate their unknown spatial and temporal distributions. The purpose of this stage is to model the linearized absorbance spectra in terms of a sum of variation types whose spectral profiles Fig. 2 e) are known but actual levels in each pixel Fig. 2 f) is unknown. After subtracting these five estimated effects, the residual spectra Fig. 2 g) should only contain random measurement noise – if the chosen mechanistic model had been perfect.

   The chosen EMSC model spectra Fig. 2 e) represent three known physical and two known chemical variation patterns that are expected to affect the apparent

absorbance spectrum $\mathbf{z}_i$ of each pixel $i = 1, 2, ...nPixels$ relative to a chosen, "typical" reference pixel $\mathbf{m}$ (in our case chosen to be the mean of all pixels in the image taken after 21 h of drying), according to the model

$$\mathbf{z}_i = b_i \times (\mathbf{m}' + \Delta c_{i,\text{Water}} \times \mathbf{s}'_{\text{Water}} + \Delta c_{i,\text{WoodPigment}} \times \mathbf{s}'_{\text{WoodPigment}}) + a_i \times \mathbf{1}' + d_i \times \mathbf{f}' + \boldsymbol{\varepsilon}_i \quad (6)$$

In this equation $\mathbf{m}, \mathbf{s}_{\text{Water}}, \mathbf{s}_{\text{WoodPigment}}, \mathbf{1}, \mathbf{f}, \mathbf{z}_i, \boldsymbol{\varepsilon}_i$ are column-wise vectors with the same length as the number of wavelength channels. $b_i, \Delta c_{i,\text{Water}}, \Delta c_{i,\text{WoodPigment}}, a_i$ and d_i are scalars. The reference spectrum $\mathbf{m}$ (Fig. 3 left, top spectrum) was chosen in the modeling for estimating the effective relative path length parameter b_i, in each pixel i and it is corrected for by division. The reason why it is important to estimate the relative optical path length is that according to Beer-Lambert's law, the absorbance effects of path length variations can be very big, and are multiplicative, while e.g. chemical pigment variations can be very small and give additive absorbance effects.

This parameter, the relative optical path length, is intended as a pragmatic measure of the "diffuse thickness" [10] of the wood sample at different states of drying, relative to that of the dry reference sample which defined the reference spectrum $\mathbf{m}$. Popularly speaking, b_i should thus show how far, on "average", the photons travel inside the wood after hitting the wood surface before they emerge again at the surface to be detected by the camera. The longer a photon travels in the chemically absorbing environment, the higher the probability is that it will be lost by chemical absorption and converted into heat or lost by other means. This relative optical path length estimate b_i is expected to vary more or less inversely proportional to the scattering coefficient s in sample #i: $b_i = \mathbf{1}/s$, but with the simplifying assumption that s is the same at all measured wavelengths.

In order to attain robust estimates of the relative path length variations by projection of input spectra $\mathbf{Z}_i$ on reference spectrum $\mathbf{m}$, the other variation types that also affect the input spectra must be modeled so as to avoid alias errors: A flat baseline of length 159 Fig. 3 left, spectrum #2) and a straight line $\mathbf{f}$ with monotonically increasing values evenly spaced between -1 and 1 Fig. 3 left, spectrum #3) are chosen for estimating the pixel's spectral baseline-offset a_i and baseline slope d_i, respectively, again due to physical light scattering variations, and are corrected for by subtraction.

A wood color spectrum $\mathbf{s}_{\text{WoodPigment}}$ Fig. 3 left, spectrum #4) was chosen for estimating and subtracting spatially visible wood structure variations $\Delta c_{i,\text{WoodPigment}}$. It was defined as the average scatter-corrected difference between the absorbance spectra of early- and latewood pixels in the last of the images (the driest state).

The absorbance spectrum of water, $\mathbf{s}_{\text{Water}}$ (Fig. 3 left, spectrum #5) was chosen in order to quantify how the concentration of water differs from the presumed concentration of water in the reference pixel, $\Delta c_{i,\text{Water}}$, and subtract this effect. The water spectrum, $\mathbf{s}_{\text{Water}}$, was defined as the wavelength dependent specific absorption coefficient of water within the 500-1005 nm range (data from [11]). Residual spectrum $\boldsymbol{\varepsilon}_i$ represents any variation in the input absorbance spectrum $\mathbf{z}_i$ that is not described by this EMSC model, once the unknown parameter values [b_i, $\Delta c_{i,\text{WoodPigment}}$, $\Delta c_{i,\text{Water}}$, a_i, d_i] in each pixel #i of these known spectral variation types have been estimated. In order to estimate these parameters simultaneously in a linear regression model, the products $b_i \times \Delta c_{i,\text{Water}}$ and

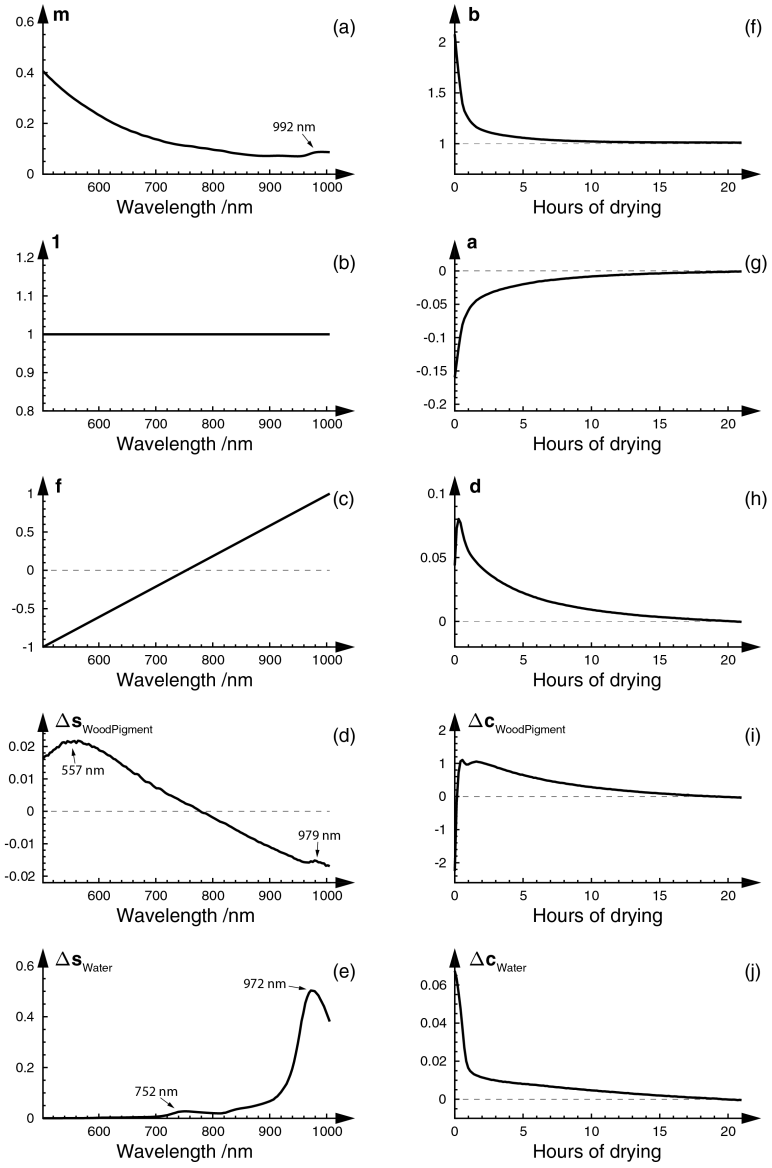


Figure 3: Modeling the known: Spectral and temporal structure of the parameters from Extended Multiplicative Signal Correction (EMSC). Left column shows EMSC model spectra chosen for modeling apparent absorbance; (a) Reference spectrum $\mathbf{m}$ for estimating optical path length, calculated as the average spectra of the last (driest) image in the series. (b) Constant "spectrum" for estimating baseline offset. (c) Linear "spectrum" for estimating baseline slope. (d) Dominant pigment spectrum $\Delta\mathbf{s}_{\text{WoodPigment}}$, defined as the average difference between early- and latewood pixels in the last (driest) image in the series. (e) Water spectrum $\Delta\mathbf{s}_{\text{Water}}$. Right column shows the temporal development of all EMSC parameters (estimated at each point in time by averaging over all image pixels).

$b_i \times \Delta c_{i,\text{WoodPigment}}$ were redefined as $g_{i,\text{Water}}$ and $h_{i,\text{WoodPigment}}$ respectively. Hence, the EMSC model is rewritten:

$$\mathbf{z}_i = b_i \times \mathbf{m}' + g_{i,\text{Water}} \times \mathbf{s}'_{\text{Water}} + h_{i,\text{WoodPigment}} \times \mathbf{s}'_{\text{WoodPigment}} + a_i \times \mathbf{1}' + d_i \times \mathbf{f}' + \boldsymbol{\varepsilon}_i \tag{7}$$

Defining the set of known model spectra:

$$\mathbf{M} = [\mathbf{m},\ \mathbf{s}_{\text{Water}},\ \mathbf{s}_{\text{WoodPigment}},\ \mathbf{1},\ \mathbf{f}]' \tag{8}$$

And the corresponding set of unknown parameters (pixel properties)

$$\mathbf{p}_i = [b_i,\ g_{i,\text{Water}},\ h_{i,\text{WoodPigment}},\ a_i,\ d_i] \tag{9}$$

the description for each pixel #i becomes a simple linear model:

$$\mathbf{z}_i = \mathbf{p}_i \times \mathbf{M} + \boldsymbol{\varepsilon}_i \tag{10}$$

This allows the unknown parameter values in $\mathbf{p}_i$ to be estimated by ordinary least squares (OLS) regression of each pixel spectrum $\mathbf{z}_i$ on $\mathbf{M}$:

$$\mathbf{p}_{i,\text{est}} = \mathbf{z}_i \times \mathbf{M}' \times (\mathbf{M} \times \mathbf{M}')^{-1}. \tag{11}$$

The multivariate modeling methods used here, EMSC and OTFP, are both based on weighted least squares. Therefore, it is important to balance the presumed relevance of the 159 wavelength channels against their estimated noise levels. Uncertainties in the measured signal from the hyperspectral camera vary across different wavelengths due to e.g. the spectral response of camera detector. To account for this in the modeling stage, wavelengths were weighted with a vector of weights $\mathbf{v}_\lambda$, $\lambda = 1, 2, ..., \Lambda$ according to their signal-to-noise ratio (SNR) so that the signal of wavelengths associated with greater uncertainties were down weighted, whilst signal originating from wavelengths with low noise level were to a greater extent preserved. The signal-to-noise ratio of an image was approximated by dividing the average raw signal intensity in the white reference region of an image with the average signal intensity in the dark reference region of the same image:

$$SRN = \frac{P_{\text{signal}}}{P_{\text{noise}}} = \frac{P_{\text{white}}}{P_{\text{dark}}}. \tag{12}$$

As the detector and the halogen lights used to artificially illuminate the wood sample become warmer over time, slight changes in their characteristics occur which causes minor changes in the signal-to-noise ratio between images taken at different points in time during an extended image acquisition period such as 21 hours. The SNR was in the present study therefore calculated for nine different images, evenly sampled from the first to the last image of the acquisition period, and then averaged into one SNR curve. The average SNR curve was then smoothed and normalized to lie in the 0.4-0.9 range before used as a weight vector in both EMSC and OTFP. Figure 4 shows the final, smoothed, weight vector $\mathbf{v}$ together with the SNR curve.

The ordinary least squares solution therefore is replaced by a weighted least squares (WLS) solution:

$$\mathbf{p}_{i,\text{est}} = \mathbf{z}_i \times \mathbf{V} \times \mathbf{M}' \times (\mathbf{M} \times \mathbf{V} \times \mathbf{M}')^{-1}. \tag{13}$$

where weights $\mathbf{V} = \text{diag}(\mathbf{v})$ where $\mathbf{v} = [v_\lambda,\ \lambda = 1, 2, ..., \Lambda]$ balances the sum-of-squares contributions from the different wavelength channels with respect to their

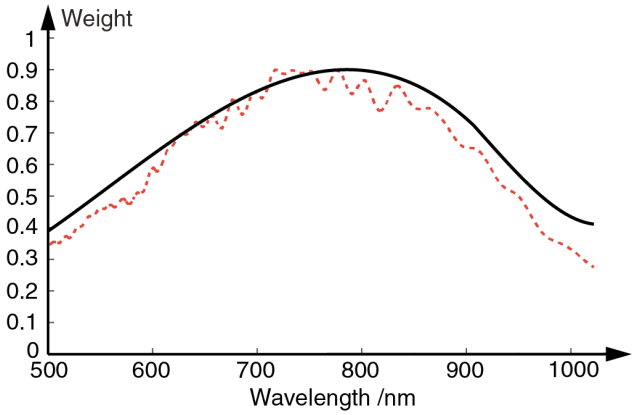


Figure 4: Weight vector used to assign weights to different wavelength regions during EMSC and OTFP. Red dotted line represents measured signal-to-noise ratio. Dark solid line represents smoothed S/N curve used as weight vector $\mathbf{v}$ in both EMSC and OTFP.

relevance and noise levels. Once the relative effective optical path length b_i had been estimated for every pixel, the chemical parameters were estimated by division:

$$\Delta c_{i,\text{Water}} = b_i^{-1} \times g_{i,\text{Water}} \tag{14}$$

$$\Delta c_{i,\text{WoodPigment}} = b_i^{-1} \times h_{i,\text{WoodPigment}} \tag{15}$$

Then, for each of these five EMSC model elements, the many spatiotemporal pixel parameters were summarized in terms of spatial structure Fig. 2 h), comparison of the first and the last image in the drying sequence) and their temporal developments Fig. 2 i). The obtained time series were finally analyzed with respect to their kinetic properties.

To correct for the physical light scattering variations while retaining the chemical light absorbance variations, the EMSC post-processing of the input spectra is defined as

$$\mathbf{z}_{i,\text{corrected}} = b_i^{-1} \times (\mathbf{z}_i - a_i \times \mathbf{1}' - d_i \times \mathbf{f}') \tag{16}$$

which corresponds to

$$\mathbf{z}_{i,\text{corrected}} = \Delta c_{i,\text{Water}} \times \mathbf{s}'_{\text{Water}} + \Delta c_{i,\text{WoodPigment}} \times \mathbf{s}'_{\text{WoodPigment}} + b_i^{-1} \times \boldsymbol{\varepsilon}_i \tag{17}$$

The estimated residual spectrum of each pixel $\boldsymbol{\varepsilon}_i$ (Fig. 2 g), is obtained by:

$$\boldsymbol{\varepsilon}_i = \mathbf{z}_i - b_i \times \mathbf{m} - g_{i,\text{Water}} \times \mathbf{s}'_{\text{Water}} - h_{i,\text{WoodPigment}} \times \mathbf{s}'_{\text{WoodPigment}} - a_i \times \mathbf{1}' + a_i \times \mathbf{f}' \tag{18}$$

This residual spectrum is expected to contain random measurement noise as well as possible unmodeled spectral structures remaining after the EMSC modeling. Before the residual spectra $\boldsymbol{\varepsilon}_i$ were submitted to further scrutiny to search for unknown features, they were scaled in order to remove the non-additive effects of varying path length:

$$\boldsymbol{\varepsilon}_{i,\text{scaled}} = b_i^{-1} \times \boldsymbol{\varepsilon}_i. \tag{19}$$

This EMSC modeling yielded rescaled and weighted 159-dimensional residual spectra $\boldsymbol{\varepsilon}_{i,\text{scaled}}$ Fig. 2 j) for more than 350 million 159-dimensional spectra:

$$\mathbf{E}_{\text{scaled}} = [\boldsymbol{\varepsilon}_{i,\text{scaled}}, 1, 2, ..., 353\ 100\ 000]. \quad (20)$$

2. **Modeling the unknown:** In order to look for unknown, and hence unmodeled patterns of variations in the hyperspectral video data, the residuals $\mathbf{E}_{\text{scaled}}$ were passed to the next stage in the modeling: A joint analysis to discover, quantify and display unknown, unmodeled spectral variation patterns and to separate these from the background of (presumably random) measurement noise.

   The estimated residual spectra $\mathbf{E}_{\text{scaled}}$ are expected to contain not only "random" measurement noise, but also systematic—but unknown—variation structures. These may be due to errors in the shape of the mechanistic model (here: normalization, linearization, linear EMSC model), errors in the employed model elements (here: estimated spectra of $\mathbf{s}_{\text{Water}}$ and $\mathbf{s}_{\text{WoodPigment}}$) or unmodeled phenomena (here: e.g. physical specular reflection from wood surface, or chemical variation in the cellulose/lignin ratio in different parts of the wood). Bilinear data modeling by *principal component analysis* (PCA) and PCA-like methods are useful for discovering, quantifying and graphically displaying non-random covariation structures in data. For "path length"-scaled, mean-centered variables and weighted residuals the model may thus be described as:

$$\mathbf{E}_{\text{ABLM,scaled,weighted}} = \mathbf{t}_1 \times \mathbf{p}'_1 + ... + \mathbf{t}_a \times \mathbf{p}'_a + ...\mathbf{t}_A \times \mathbf{p}'_A + \mathbf{E}_{\text{ABLM,final,weighted}} \quad (21)$$

   where the bilinear contribution from each principal component $\#a = 1, 2, ..., A$ in this case consists of the product of the spatiotemporal scores $\mathbf{t}_a = [t_{i,a}, i = 1, 2, ...nPixels]$ and the loading spectra $\mathbf{p}_a = [p_{k,a}, k = 1, 2, ...nWavelengths]$. Ideally, the first PCs are included in the final model because they represent non-random covariation structures, while $\mathbf{E}_{\text{ABLM,final,weighted}}$ represents the remaining random noise.

   In the present case, the number of spectra in $\mathbf{E}_{\text{scaled}}$ is very high - for different 2D pixel positions at different times. To simplify the explorative adaptive bilinear modeling, the On-The-Fly-Processing (OTFP) software from Idletechs AS (www.idletechs.com) was employed, in order to be able to handle this stream of Quantitative Big Data on a regular PC within a reasonable computation time. The OTFP yields results very similar to a weighted principal component analysis (PCA) [8], but develops this bilinear model sequentially and therefore allows PCA-like modeling of more or less continuous streams of high-dimensional data such as spectra from hyperspectral imaging. The same statistical weight vector of the wavelength channels Fig. 4 was used in the OTFP as in the EMSC.

   The OTFP analysis will gradually, but automatically, discover, extract and quantify any clear systematic covariation pattern remaining in the residuals. The components in the bilinear model consist of the product of unknown spectral loading profiles Fig. 2 k) and unknown spatiotemporal score vectors Fig. 2 l), and represent a succession of "unexpected patterns". In the present case, four patterns from the residuals were found to have smooth spectral loadings. To be on the safe side, a fifth principal component was also included in the final model.

   Like for the spatiotemporal EMSC scores, the estimated spatiotemporal OTFP score patterns of each of these five PCA components was refolded with respect to its 2D

wood picture Fig. 2 n) and its temporal development Fig. 2 o) during the drying process.

The residuals Fig. 2 m) after this two-stage modeling were briefly interpreted graphically. Since no important patters were found, they were just summarized statistically and then discarded.

3. **Temporal kinetics modeling:** Each of the average temporal score vectors from the EMSC modeling of known phenomena Fig. 2 i) and OTFP modeling of unknown phenomena Fig. 2 o) were finally fitted to the simplified first-order kinetic model. The purpose of this was to shed some light on the temporal development of the hyperspectral video measurements, and hopefully also on the mechanisms governing how the optical properties of the wood sample change during the drying process.

   For each of the EMSC and OTFP components the average score was computed at each of the 150 time points $t = 1, 2, ..., 150$ by averaging over all pixels at time point t. After defining the component's score vector as y_t^*, $t = 1, 2, ..., 150$, a linearly transformed score vector $y_t = y_t^* \times a + b$ was defined so as to ensure that y_t is positive and falling gradually towards zero. The scaling factor a was defined to ensure that one unit of change in y_t corresponded to one unit change in apparent absorbance of the input data. Offset b was then defined so that the minimum value $y_{t=21\text{hrs}}$ was 0.001 absorbance units or more. Then the temporal derivative dy_t/d_t was computed using Eq. 2.

   The dynamic complexity of the component time series y_t^* could then be assessed by plotting dy_t/d_t vs time t: If the processes affecting component y_t^* had followed simple first-order kinetics, then the 150 data points would fall along a straight line (conf. Eq. 4). Hence, deviations from the straight lines indicate kinetics that is more complex.

4 Results

4.1 Average moisture in the drying process

The total weight of the wood sample ($\mathbf{w}_{wood}$) is shown in Fig. 5 a) for the drying period 0-21 h, in total consisting of 150 time points. The weight-based average moisture content ($\mathbf{water}_{\%}$) in the wood sample is shown in Fig. 5 b). The figure shows that the wood sample has 37.09 % moisture at time $t = 0$, and after $t = 21$ h of low-temperature drying, the wood sample still contains 6.20 %" moisture, compared to its oven dried state.

In order to study the over-all kinetics of the drying process, the average moisture content $\mathbf{water}_{\%}$, its temporal derivative $d(\mathbf{water}_{\%})/dt$ Eq. 2 is plotted against drying time Fig. 5 c) and against $\mathbf{water}_{\%}$ itself Fig. 5 d). In theory, if $d(\mathbf{water}_{\%})/dt$ vs $\mathbf{water}_{\%}$ in Fig. 5 d) had displayed a simple linear relationship, this would have meant that the drying process follows simple first-order kinetics. But at least four different near-linear relationships at different drying stages may be observed: 0-0.5 h, 0.5-1.5 h, 1.5-10 h and 10-21 h. This indicates that the drying of wood represents a rather complex dynamic process, where especially the initial hour or two follow very different kinetics.

A more noise robust but also less sensitive way to study the observed dynamics is illustrated by plotting development of $\mathbf{y} = \mathbf{w}_{\text{wood}}$ Fig. 5 e) and $\mathbf{y} = \mathbf{w}_{\text{water}}$ Fig. 5 f) on ln-scales against drying time. While a simple first-order kinetics process would have yielded a single

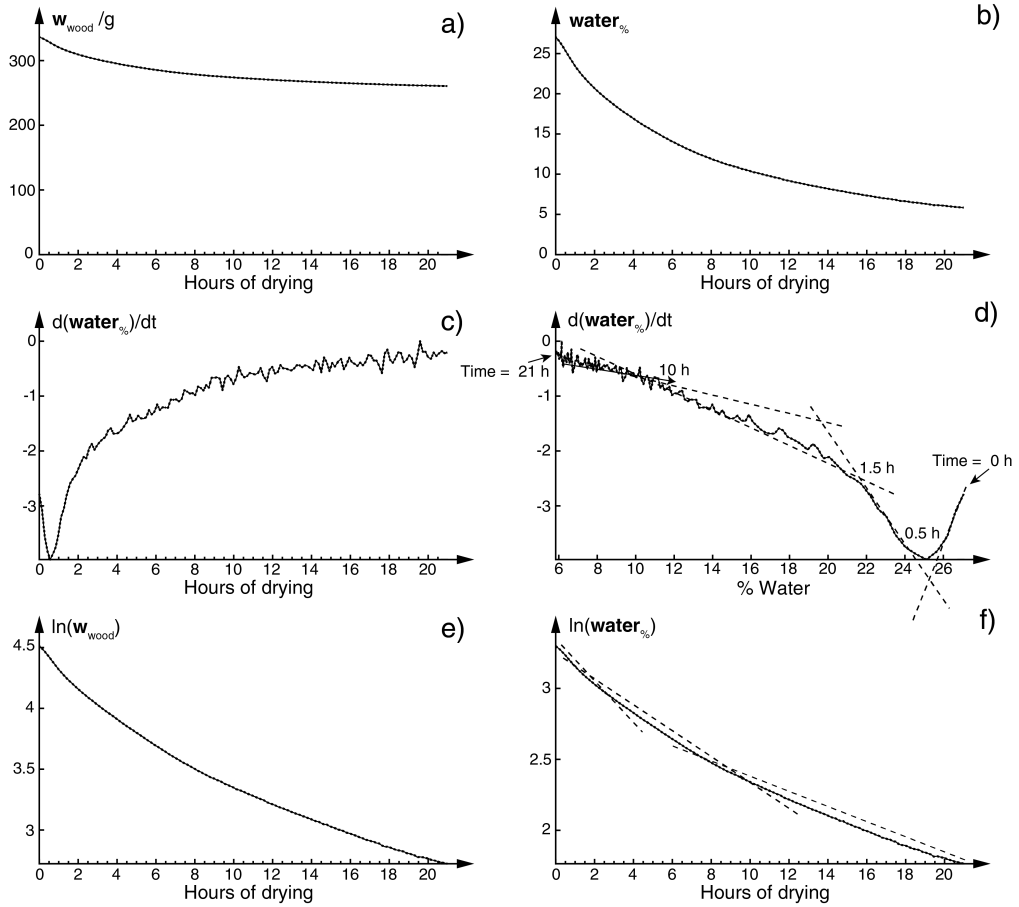


Figure 5: a) Weight of wood sample as function of drying time. b) % Water in wood sample as function of drying time, calculated as $\mathbf{water}_{\%} = 100 \times (\mathbf{w}_{\text{wood}} - 245.46g)/245.46$. c) Rate of change $d(\mathbf{water}_{\%})/dt$ as function of drying time. d) Rate of change $d(\mathbf{water}_{\%})/dt$ as function of $\mathbf{water}_{\%}$, with four local approximation lines. e) $\ln(\mathbf{w}_{\text{wood}} - 245.46)$ as function of drying time. f) $\ln(\mathbf{water}_{\%})$ as function of drying time, with three local approximation lines.

straight line $ln(y) \approx -k \times t + ln(y_{t=0})$, clear curvatures are observed. In Fig. 5 f) three local approximation line segments are drawn, to illustrate the complex drying dynamics.

Hence, while the weight loss of the wood sample of course gives important insight into the over-all drying process, it does not shed light on the different mechanisms involved in the process. Better methods are in the end needed in order to monitor and study the different mechanisms leading to water evaporation in the wood sample.

4.2 Hyperspectral imaging input spectra

Figure 6 illustrates the modeling of the vis-NIR spectra in terms of the physical and chemical phenomena expected to dominate the data, for ten typical pixels. The top left and right subplots represent the absorbance spectra of pixels at $t = 0$ h (i.e. very wet wood sample) and $t = 21$ h (i.e. dry wood sample). The inserted windows magnifies the wavelength region were water absorbance is seen most clearly.

As expected, the absorbance spectra of wet wood (Fig. 6, top left) shows a clear peak in the 940-1005 nm region which the specific absorption coefficient spectra (Fig. 3 e) suggests is heavily associated with water absorption. After drying for 21 hours it has decreased, but not disappeared.

Over-all, the apparent absorbance spectra are seen to be dominated by brown wood pigments showing their strongest light absorbance in the blue and green wavelength regions (< 550 nm). Moreover, the shorter wavelengths in the visible region appear to have a significantly higher absorbance in a wet wood sample than in a dry one. There is also considerable absorbance variations in the water absorbance range. This is not counterintuitive considering that a wet piece of wood is perceived as darker than a dry one even to the human eye, and probably represent variations in physical light scattering properties. Less intuitively however, is the fact that the absorbance decreases, i.e. the wood is perceived as lighter, in the 650-900 nm range when saturating the sample with water. This difference in absorbance is likely an effect of different scatter properties between the two states of the sample. Introducing water into the wood pores reduces lateral light scattering [12], which in turn allows more light to be reflected back to the camera.

Hence, the water-related physical properties of the wood, as seen from its light scattering properties (as well as its drying kinetics in Fig. 5 d) and Fig. 5 f) have considerable complexities.

4.3 Modeling known structures by EMSC

The purpose of the EMSC is to quantify and remove effects from the spectra that we think we understand. The middle row of curves in Fig. 6 shows the spectra of the same subset of pixels from the wood sample in both wet ($t = 0$ h, left) and dry condition ($t = 21$ h, right) after the EMSC correction:

$$\mathbf{z}_{i,\text{corrected}} = \mathbf{m} + \boldsymbol{\varepsilon}_{i,\text{scaled}}. \tag{22}$$

As can be seen in the middle row of Fig. 6, clearly the EMSC model largely succeeds in modeling the "known" types of spectral variations, and then removing them, both by subtractions (baseline variations, water and wood pigment variations) and by division (optical path length variations). The EMSC-corrected spectra $\mathbf{z}_{i,\text{corrected}}$ are brought together around the chosen reference spectrum $\mathbf{m}$. By comparing the spectra of the sample in dry vs wet state in the 940-1005 nm region—a region heavily associated with water absorption—it

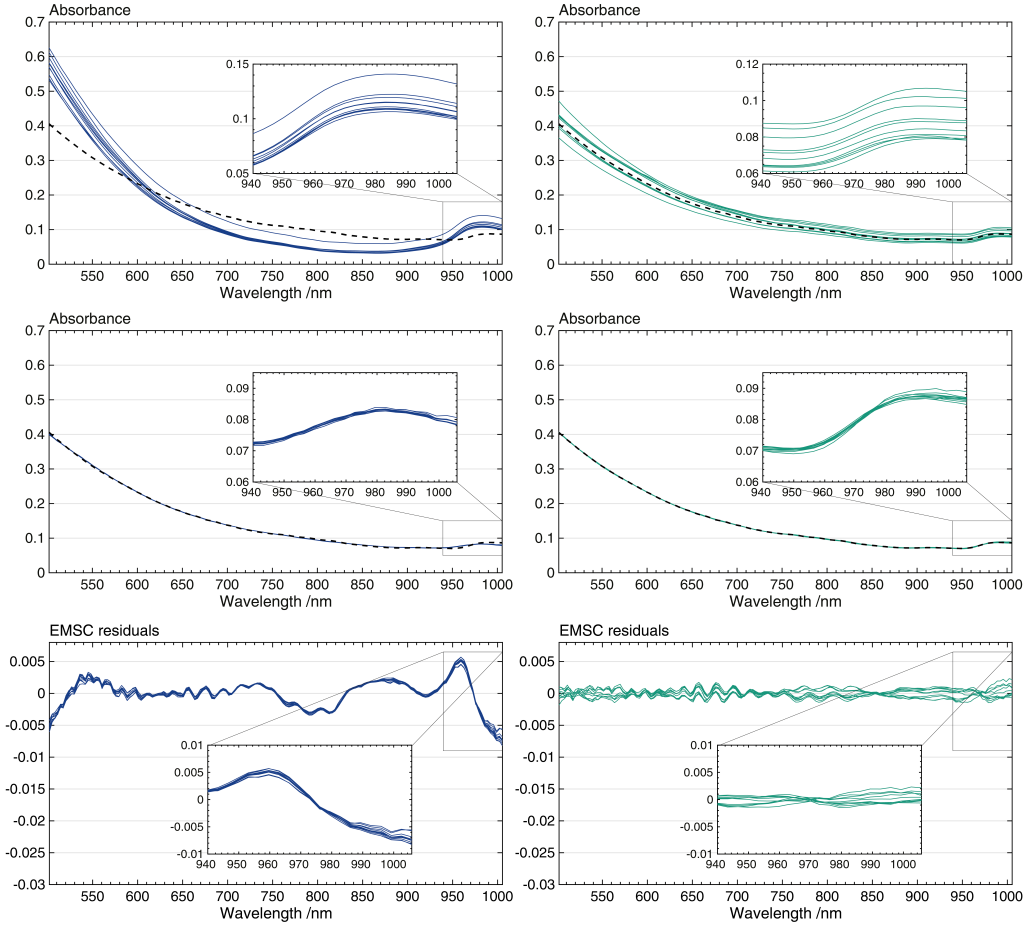


Figure 6: Apparent absorbance spectra from 10 typical pixels of the wood sample in wet condition (left) at $t = 0$ h and dry condition (right) at $t = 21$ h. The black dotted line represents the chosen reference spectrum m, which is the average of all pixels in the image taken after 21 hours of drying. Top figures show spectra before EMSC pre-processing. Middle figures show spectra after EMSC pre-processing. Bottom figures show unmodeled spectral residuals after the EMSC modeling. Windows within the figures show a magnification of the 940-1005 nm region strongly associated with water absorption.

is clear that the characteristics of the spectra are different in the two states even after the spectral correction. In this region, the dry sample exhibits an S-shaped absorbance while the wet version of the same sample has a more blunt and flat absorbance spectra. Because the sample underwent desorption "drying" in a laboratory with a room temperature of about 20 °C, the sample will certainly still contain chemically bound water after the 21 hours of drying. It is possible that the differences in absorbance characteristic seen in the 940-1005 nm range can be accredited to different absorbance properties of free vs chemically bound water in the wood.

For more graphical resolution, the bottom curves of Fig. 6 shows the estimated residual spectra $\boldsymbol{\varepsilon}_{i,\text{scaled}}$ (without reference spectrum $\mathbf{m}$) for the same pixels in sample in the wet ($t = 0$ h) and dry ($t = 21$ h) condition. Because most of the residuals are non-zero, there are undoubtedly unknown spectral phenomena taking place within the sample which are not completely captured by the chosen EMSC model spectra shown in Fig. 3 (left side). The EMSC accounted for more than 99 % of the total variance in the 159 weighted wavelength channels. But it is clear that the residuals are not just random measurement noise. Unmodeled spectral structures are clearly visible, in particular in the water absorbance region and at the shortest wavelengths. Before moving on to analyze these unknown variations, the spatiotemporal properties of the EMSC parameter estimates will be discussed.

4.4 Temporal development of fitted EMSC parameters

Figure 3 (right side) shows the temporal development of the fitted EMSC parameters associated with the EMSC model vectors (left side), averaged across all pixels at each of the 150 points in time $t = 0, ..., 21$ h.

Within the first hour of drying, all the modeled properties of the wood sample appear to change rapidly. It is important to note that the hyperspectral monitoring process of the wood sample started immediately after the sample had been taken from a state of full submersion in water. As a consequence of this, there was an amount of liquid water on top of the wood during the initial period of drying, effectively forming a film of water covering the wood and partially cloaking the spectral properties of the wood itself. This probably caused the particularly complex drying behavior at the beginning of the drying process, as also seen in Fig. 5 d).

Figure 6 (bottom left) showed a lot of unmodeled water absorption at $t = 0$ h. It is obvious that the chosen EMSC model spectra alone do not allow an adequate description for the spectra of wood with the highest moisture content. Since absorbance spectra of bound water probably overlap with the EMSC model of free water (Fig. 3 c), so-called alias errors [13] are expected in the EMSC estimates. For this reason, some of the initial changes seen in the modeled properties of the sample could be misleading during the initial stages of drying, as they have large uncertainties associated with them.

After one to two hours, the average of the multiplicative optical path length score b_i (Fig. 3 f) seems to approach the level of the reference. The average additive baseline scores a_i and d_i (Fig. 3 g and h) likewise level off. Hence, at first glance the different physical light scattering parameters show somewhat similar behavior.

Concerning the chemical absorbance effects, the average scores for $\Delta c_{i,\text{WoodPigment}}$ should ideally have been constant over time, to the extent the structure of the wood sample itself is constant. As can be seen in subplot (e) of Fig. 3, this is not the case. While $\Delta c_{i,\text{WoodPigment}}$

does not vary as much as the other EMSC model parameters, some variation is evident. This could indicate that the in-situ estimated spectra describing the pigment differences within the wood is too crude to adequately model the inhomogeneity of the wood. The origin of this is not clear. But it could be related to the small spatial contraction expected when the wood dries.

Because less energy is required to evaporate free water within a wood sample compared to bound water [12], the drying-out of a water saturated piece of wood occurs in at least two phases. First a rapid evaporation of free water takes place, which is then followed by a slower evaporation of bound water. As such, the temporal development of average sample water content suggested by the EMSC model in Fig. 3 j) appears realistic; the curve quickly drops during the first hour, suggesting the evaporation of free water and surface water, which is then followed by a significantly slower drying process for the remaining 20 hours of drying.

The EMSC-based estimate of $\Delta c_{i,\mathrm{Water}}$, is based on a projection on the spectrum of free water, $\mathbf{s}_{\mathrm{Water}}$, so it is intended to be proportional to the concentration of free water. This behavior of the estimated free water is what might be expected if the free water were primarily situated on or close to the wood surface. It is distinctly different from that of the weight-based over-all moisture content (Fig. 5 b), which represents loss of both free and bound water.

4.5 Spatial development of fitted EMSC parameters

Figure 7 shows the corresponding spatial distribution of the EMSC parameters, for two of the 150 time points: $t = 0$ h (wet; top) and $t = 21$ h (dry, corresponding to the reference time point for spectrum $\mathbf{m}$; bottom). In the figure, light rendering means high while dark rendering means low values. All the EMSC parameters show clear spatial patterns of early- vs latewood, in addition to the differences between wet and dry wood.

In summary, the effective optical path length is lower in dry wood than in wet wood. This is probably due to an increase in light scattering as air replaces water in the wood pores and therefore increases the variability in the refractive index inside the wood material. On the other hand, the absorbance baseline offset increases upon drying. This may also be due to the increasing light scattering, e.g. by increasing the angular distribution of reflected light and thus reducing the fraction of light reaching the narrow angle of the camera's light detector. The consequence of the path length reduction is apparently stronger than the consequence of the increased spectral absorbance level, since the over-all effect is to render dry wood visually lighter than wet wood. A more detailed spatial comparison of early- and latewood pixels confirm the opposite trends of these two effects of light scattering.

The slightly decreasing baseline slope indicates a slightly increased wavelength dependency of this absorbance baseline. The pigment concentration estimate $\Delta c_{\mathrm{WoodPigment}}$ increases slightly; this is probably due to insufficient model detail. Finally, the estimate of free water concentration, $\Delta c_{\mathrm{Water}}$, falls distinctly upon drying, as expected. The EMSC modeling accounted for 98.04 % of the weighted variance.

4.6 Modeling unknown structures by OTFP

Each of the residual spectra remaining after the "theory-driven" EMSC modeling of known phenomena, illustrated for 10 of the pixels at the bottom of Fig. 6, were divided by the

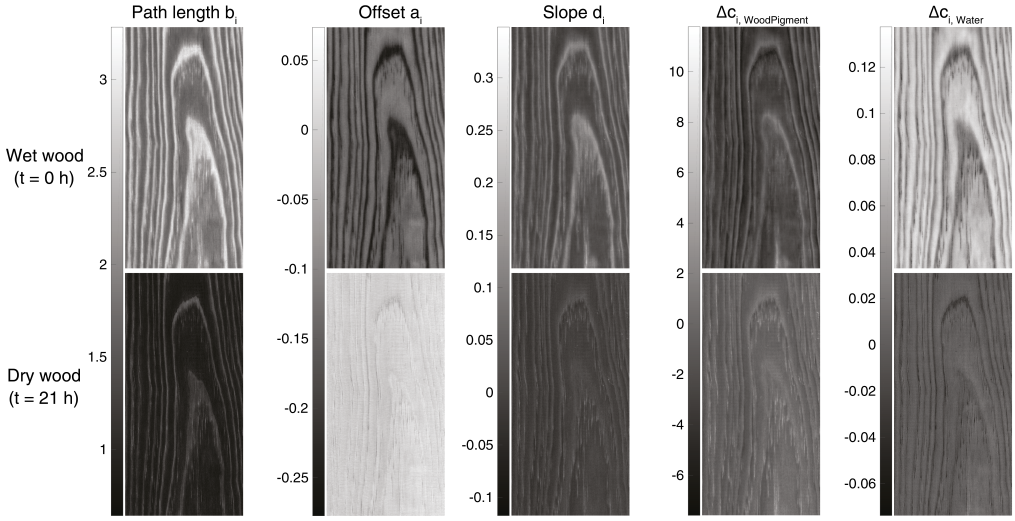


Figure 7: Modeling the known: Spatial structure of EMSC parameters. 2D visualization of fitted EMSC parameters in wet wood sample, i.e. $t = 0$ h, (upper row) and dry wood sample, i.e. $t = 21$ h, (lower row) for all parameters used in the EMSC model.

estimates of their relative effective optical path length b_i. These more than 350 million residual spectra, accounting for 1.06 % of the total variance of the input spectra, were then submitted to the OTFP software system, one small batch at a time.

Since the OTFP is a weighted least squares procedure that describe the stream of incoming spectra with as few principal components as possible, the first step in the OTFP is to multiply each residual spectrum by the same relevance-vs-reliability weights Fig. 4 that were used in the EMSC modeling. Then the bilinear "PCA-like" model was gradually developed, to describe as much of the variation in the incoming stream of residual data with as few principal components as possible.

The sequence of five first OTFP components representing unknown, but systematic spectral structures, decreased this further, from 1.06 % to 0.52 %, 0.15 %, 0.05 %, 0.02 % and 0.01 %, respectively.

4.7 Spectral and temporal development of estimated OTFP parameters

Figure 8 shows the results for the first five PCs, for the deweighted spectral loadings for the 159 wavelength channels (left), and the temporal scores, averaged over all the pixels in each of the 150 points in time $t = 0, 1, 2, ..., 21$ h (right). The first four PCs display clear, smooth, loadings and scores. And while the fifth PC shows rather noisy loadings, its score vector is smooth. Comparing their relative sums-of-squares contributions based on weighted loadings, PCs #1 and #2 accounted for 95 % and 4.8 %, while PCs #3, #4 and #5 together accounted for only 0.2 %. Hence, only the first two PCs dominate the weighted spectral residuals.

All five OTFP PCs show very different behavior for the first hour of drying, compared to the rest of the 21-hour drying period. Moreover, PCs #1, #2 and #3 show clear signatures

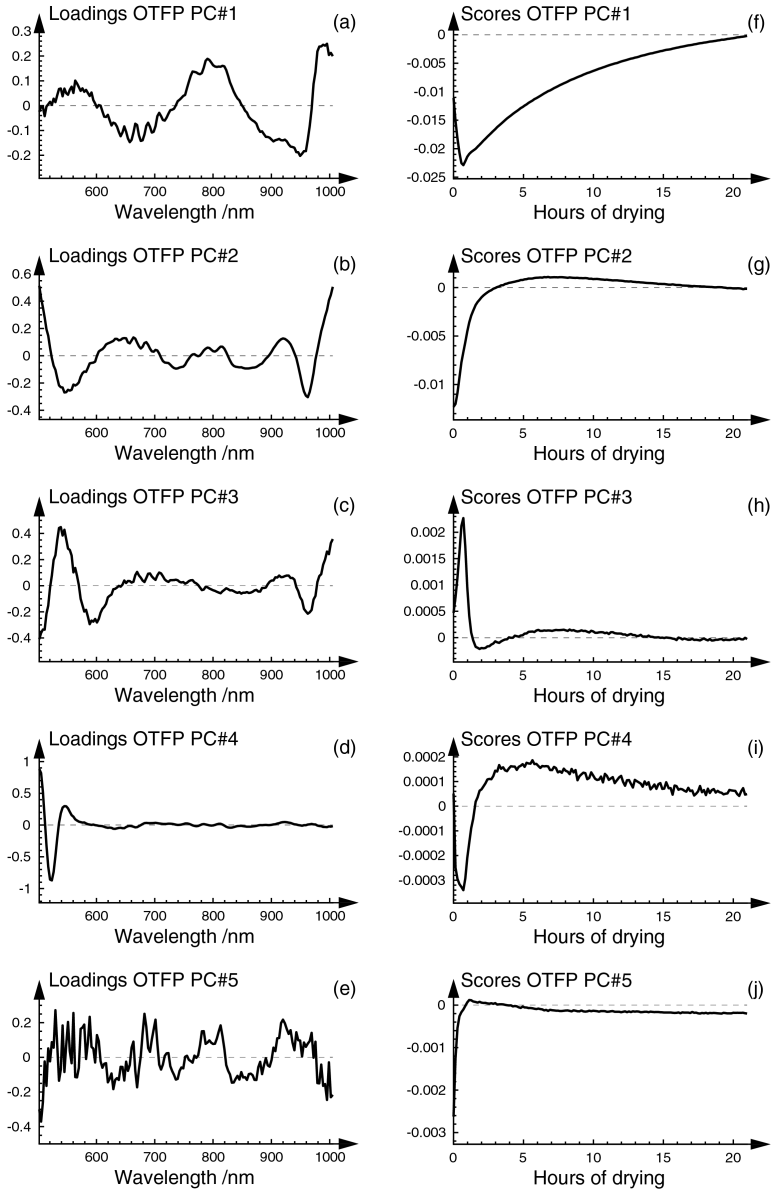


Figure 8: Modeling the unknown: Spectral and temporal structure of the parameters from adaptive bi-linear modelling in the On-The-Fly-Processing (OTFP) implementation. Left column shows the OTFP model spectra estimated for modeling of apparent absorbance. De-weighted loadings for component 1-5. Right column shows the temporal development of the ABLM parameters (estimated at each point in time by averaging over all image pixels).

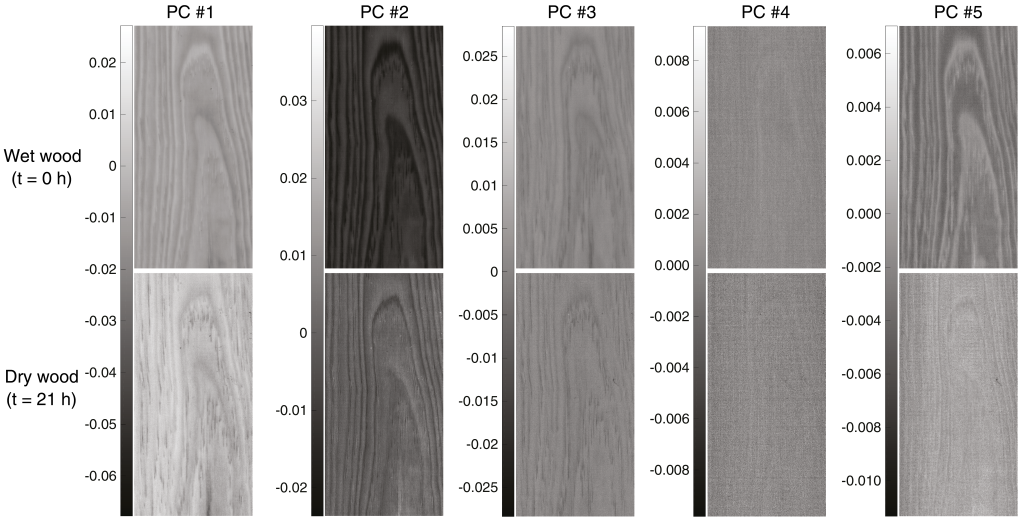


Figure 9: Modeling the unknown: Spatial structure of OTFP parameters. 2D visualization of reconstructed OTFP scores of wet the wood sample, i.e. $t = 0$ h, (upper row) and dry wood sample, i.e. $t = 21$ h, (lower row) for the five first PCs.

in the water absorbance region above 900 nm. PCs #2, #3, #4 likewise show clear spectral features at the lowest wavelengths. PCs #1 and #2 also show some structure at intermediate wavelengths.

4.8 Spatial structure of estimated OTFP parameters

Figure 9 shows the corresponding spatial distribution of the OTFP parameters, for two of the 150 time points: $t = 0$ h (wet; top) and $t = 21$ h (dry; bottom). In the figure, light rendering means high while dark rendering means low values. In particular, the first, second and fifth component show spatial differences between wet and dry wood.

For instance, PCs #1 and #2 reveal a local spatial structure at the bottom of the wood sample that are not visible in the EMSC parameters. However, in general, their temporal and spatial structures of the OTFP components were not as clear as for the EMSC parameters in Fig. 3 and Fig. 7.

A subsequent axis rotation, from the basic, orthogonal PCA score-and loading- representation of the OTFP to a simpler bilinear structure (by e.g. varimax or independent component analysis ICA) may be expected to give easier wood-related interpretation. However, that is beyond the scope of the present chapter, which is primarily intended to demonstrate how massive streams of hyperspectral video data can be efficiently modeled in terms of known and unknown features.

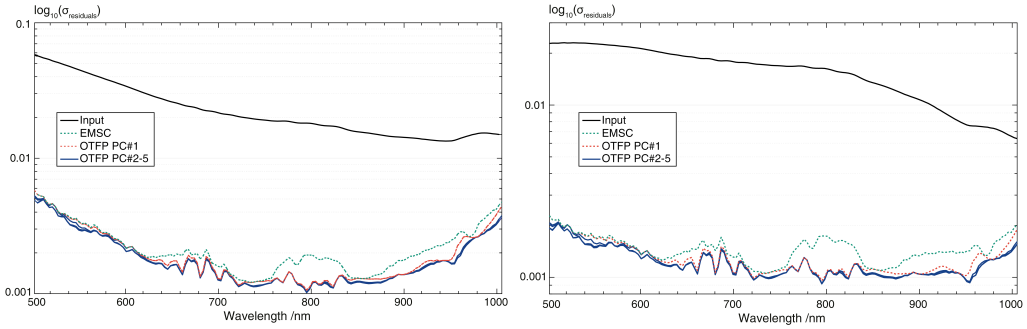


Figure 10: How the variation at the different wavelengths was explained by the sequence of modeling steps: In the input data, after EMSC and after OTFP PCs #1, #2, #3, #4 and #5. Top: Residual standard deviations, statistically weighted. Bottom: Residual standard deviations, de-weighted.

4.9 Over-all model assessments

4.9.1 Statistical summaries

Figure 10 (left) shows how the standard deviation of each of the 159 wavelength channels is reduced by the succession of modeling stages, for the weighted wavelength channels used in the weighted least squares modeling in EMSC and OTFP. It is clear that the EMSC modeling of known spectral phenomena (dashed green curve) has removed a very large part of the variation. However, substantial residuals remain in visible range (610-720 nm) and in two NIR ranges (750-840 nm and 870-1010 nm). Interestingly the water absorption region 920-970 nm shows clear, unmodeled spectral features.

Most of the residuals after the EMSC, including the remaining water absorptions, are accounted for by the first OTFP component. The effects of OTFP components 2, 3, 4 and 5 are hardly visible in the plot. Figure 10 (right) show the same residual standard deviations after having removed statistical weights Fig. 4.

For statistically optimal EMSC and OTFP modeling from a weighted least squares point of view, the inverse of the residual standard deviation after five OTFP components might have been used as wavelength weights. However, we have chosen to retain the original weights in shown Fig. 4, in order to emphasize the chemical relevance in the longer wavelengths (water absorption) and the shortest wavelengths (wood pigment effects).

The following statistics summarize the weighted residual standard deviations (left, Fig. 10) over the 159 weighted wavelengths: The EMSC modeling of known spectral phenomena accounted for as much as 98.93 % of the total initial variance, leaving 1.06 % unexplained. The sequence of five OTFP components representing unknown, but systematic spectral structures, decreased this further to 0.52 %, 0.015 %, 0.05 %, 0.02 % and 0.01 % of the total initial variance, respectively. Hence, in total, about 99.99 % of the input variance was explained by the modeling of both known and unknown spectral phenomena.

It should be noted that due to partial overlap between the known and unknown spectral variation phenomena in the drying wood, alias errors are to be expected in the EMSC scores, which may affect the subsequent OTFP modeling due to the multiplicative (rather

than purely additive) residual definition. Moreover, the estimated OTFP loading spectra have the unnatural property of being orthogonal to the chosen EMSC spectra [13]. Fortuna et al. [14] presented a method for resolving alias problems in purely additive systems. This approach might also have been useful here. But that would require special attention to the nonadditive estimation and correction for the optical path length, so that is not pursued here.

4.9.2 Temporal kinetics modeling

The kinetic analysis of the over-all drying process Fig. 5 showed that the weight-based estimates of the moisture content dynamics were rather complex. The phase space of the moisture percentage displayed least four types of kinetics. The EMSC and OTFP models have shown that several different variation phenomena affect the spectroscopic properties of the wood sample during the drying process.

The final modeling step is an attempt at assessing the complexity of the processes going on during the drying process of the wood sample. Each of the individual temporal score averages from the EMSC model Fig. 3, right) and the OTFP model (Fig. 8, right) were modified and fit to the decay model corresponding to a simple first-order reaction. In each case, the fit included a rescaling, a sign change if needed and an offset correction, followed by a logarithmic transformation.

Figure 11 shows that for the first hour of drying, the simple first-order kinetic model did not fit. But in the time period 1-20 h drying, the log transforms for four of the five EMSC parameters as well as the first (and dominant) OTFP parameter fitted well to the linear model expected for first-order process dynamics, with different rate constants.

4.9.3 Over-all model summary

A time series of hyperspectral images, comprising around 300 million spectra, was logarithmically linearized, weighted statistically and then decomposed by mathematical subspace modeling into "known" components and "unknown" components.

The "known" components represented three physical light scattering effects and two chemical absorbance effects expected to affect the measured spectra. Together, these five "known" components accounted for 98.94 % of the variance of all the statistically weighted spectra, leaving 1.06 % variance unexplained.

Then a sequence of "unknown" components, i.e. unexpected, but systematic variation patterns, were discovered and extracted from the spectral residuals. The five first components reduced the remaining unexplained variance to 0.52 %, 0.015 %, 0.05 %, 0.02 % and 0.01 %, respectively. Hence, by compressing the 159 wavelength channels into these 10 modeled components, more than 99.99 % of the variance of all the absorbance data was thereby accounted for.

Linear and bilinear modeling was used in the component estimation, and compensated for by subtraction, in analogy to Beer's law. Moreover, one of the estimated physical components, the relative effective optical path length, was corrected for by division, also in analogy to Lambert's law.

The dynamics of the wood drying process was further assessed by tentatively fitting each of the component averages, after a suitable linear transformation, to the loglinear kinetic

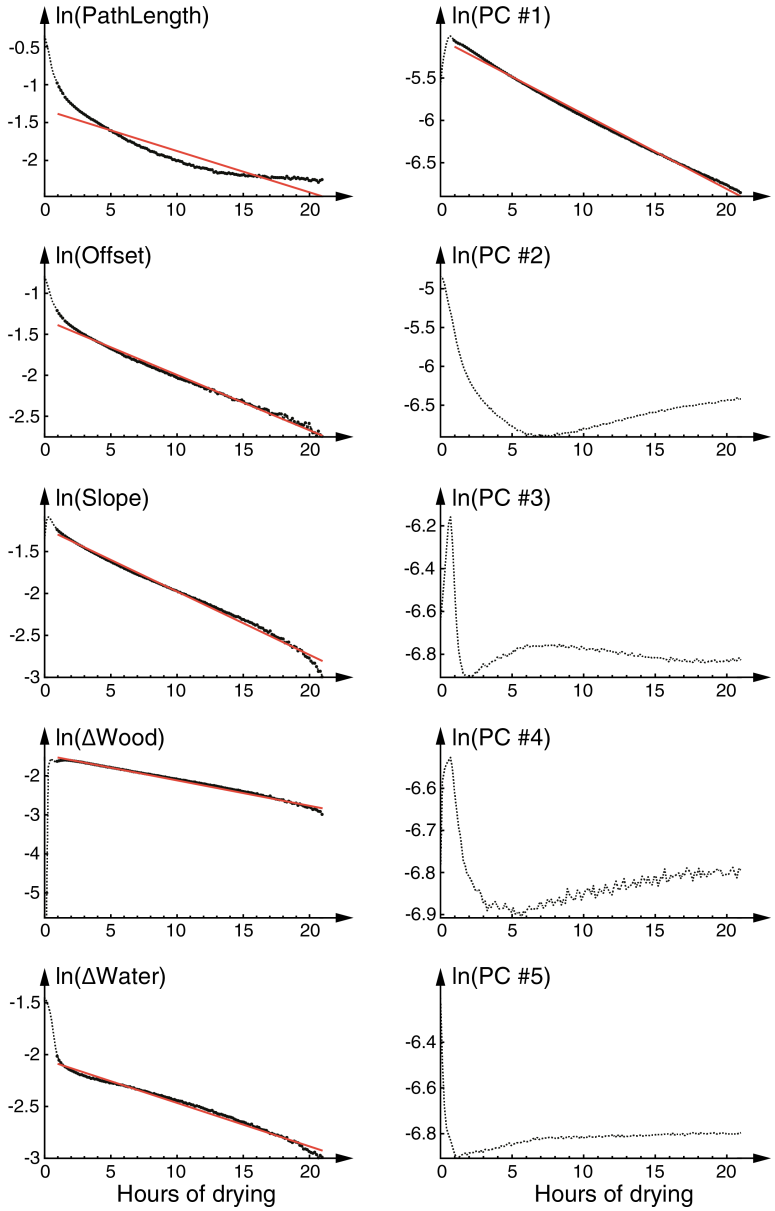


Figure 11: Kinetic modeling of the hyperspectral video developments: Analysis of the known and unknown temporal developments of the parameters from EMSC (left) and OTFP (right), averaged over all pixels, as first order dynamic processes. Dotted: ln(normalized parameters); densely dotted represent the data points used in the least squares estimation of the kinetic parameters. Straight red lines: Model fitted.

model. Some of the known and unknown components fitted quite well to this model of a first-order reaction, but with different reaction rates, while other components displayed more complex time developments.

Compared to the initial assessment of the over-all weight loss during the wood drying process, the hyperspectral video appeared to give much more detailed, quantitative account of the complex water diffusion- and evaporation processes along with their kinetics. It also revealed how their spatial distribution in the wood changed from wet to dry wood.

5 Conclusions

By monitoring a drying wood sample we experimentally demonstrated a generic way in which a stream of hyperspectral time series data can be modeled in terms of a priori known and unknown constituent spectra to enable large dimensionality reduction of the data, essentially without any loss of information. An additional benefit of the described methodology, apart from enabling substantial compression of the data, is that it autonomously highlights unidentified systematic spectral variations within the sample being studied, which aids in further exploration and understanding of the underlying chemical and physical processes causing the variations. The kinetic analysis of the weight loss curve of our drying wood sample indicated that the over-all drying process of the wood sample was rather complex. This complexity was addressed, by resolving the hyperspectral time series data in terms of a multivariate, mixed multiplicative – additive EMSC modeling, involving three known physical variation phenomena related to varying light scattering and two known chemical variation phenomena related to changing wood composition. After removal of these known effects, the hyperspectral imaging data had clear unmodeled spectral variations, particularly in images taken during the first hour of drying.

Most of these residual variations were picked up by the subsequent data-driven OTFP modeling, which revealed two major and a couple of minor unexpected variation patterns. Four of the known variation phenomena and one of the unexpected variation patterns seemed to follow relatively simple first-order kinetics. Most notably, the effective optical path length did clearly not follow first-order kinetics.

The RGB images of the wood sample in Fig. 1 b) and c) showed the wood to be darker and more yellow-brown when in a wet state compared to when dry. This corresponds well to how the drying seemed to affect the light scattering, causing several types of variations, the most dominant one being a strong reduction in the effective optical path length.

6 Acknowledgement

References

[1] Phil Williams. *Near-infrared technology in the agricultural and food industries*. American Association of Cereal Chemists, 2004.

[2] Marena Manley. Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. *Chem. Soc. Rev.*, 43(24):8200–8214, 2014.

[3] Harald Martens and Tormod Naes. *Multivariate calibration*. Wiley, 2002.

[4] Hans Grahn and Paul Geladi. *Techniques and applications of hyperspectral image analysis*. J. Wiley, 2007.

[5] A. Kohler, J. Sulé-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, and D. G. et al. van Pittius. Estimating and correcting mie scattering in synchrotron-based microscopic fourier transform infrared spectra by extended multiplicative signal correction. *Applied Spectroscopy*, 62(3):259–266, 2008.

[6] Harald Martens and Edward Stark. Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 9(8):625–635, 1991.

[7] Harald Martens, Jesper Pram Nielsen, and Søren Balling Engelsen. Light scattering and light absorbance separated by extended multiplicative signal correction. application to near-infrared transmission analysis of powder mixtures. *Analytical Chemistry*, 75(3):394–404, 2003.

[8] Raffaele Vitale, Anna Zhyrova, João F. Fortuna, Onno E. de Noord, Alberto Ferrer, and Harald Martens. On-the-fly processing of continuous high-dimensional data streams. *Chemometrics and Intelligent Laboratory Systems*, 161:118–129, 2017.

[9] Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2):149–171, 1997.

[10] Gerald S. Birth. Diffuse thickness as a measure of light scattering. *Applied Spectroscopy*, 36(6):675–682, 1982.

[11] Water and ice data [internet]. npsg.uwaterloo.ca. [cited 13 february 2018]. available from: http://www.npsg.uwaterloo.ca/data/water.php, 2018.

[12] Mats Andersson, Linda Persson, Mikael Sjöholm, and Sune Svanberg. Spectroscopic studies of wood-drying processes. *Optics Express*, 14(8):3641, 2006.

[13] Harald Martens. The informative converse paradox: Windows into the unknown. *Chemometrics and Intelligent Laboratory Systems*, 107(1):124–138, 2011.

[14] João Fortuna and Harald Martens. Multivariate data modelling for de-shadowing of airborne hyperspectral imaging. *Journal of Spectral Imaging*, 6, 2017.

ISBN: 978-82-575-1637-6
ISSN: 1894-6402

Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no