UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Department of Physics and Technology

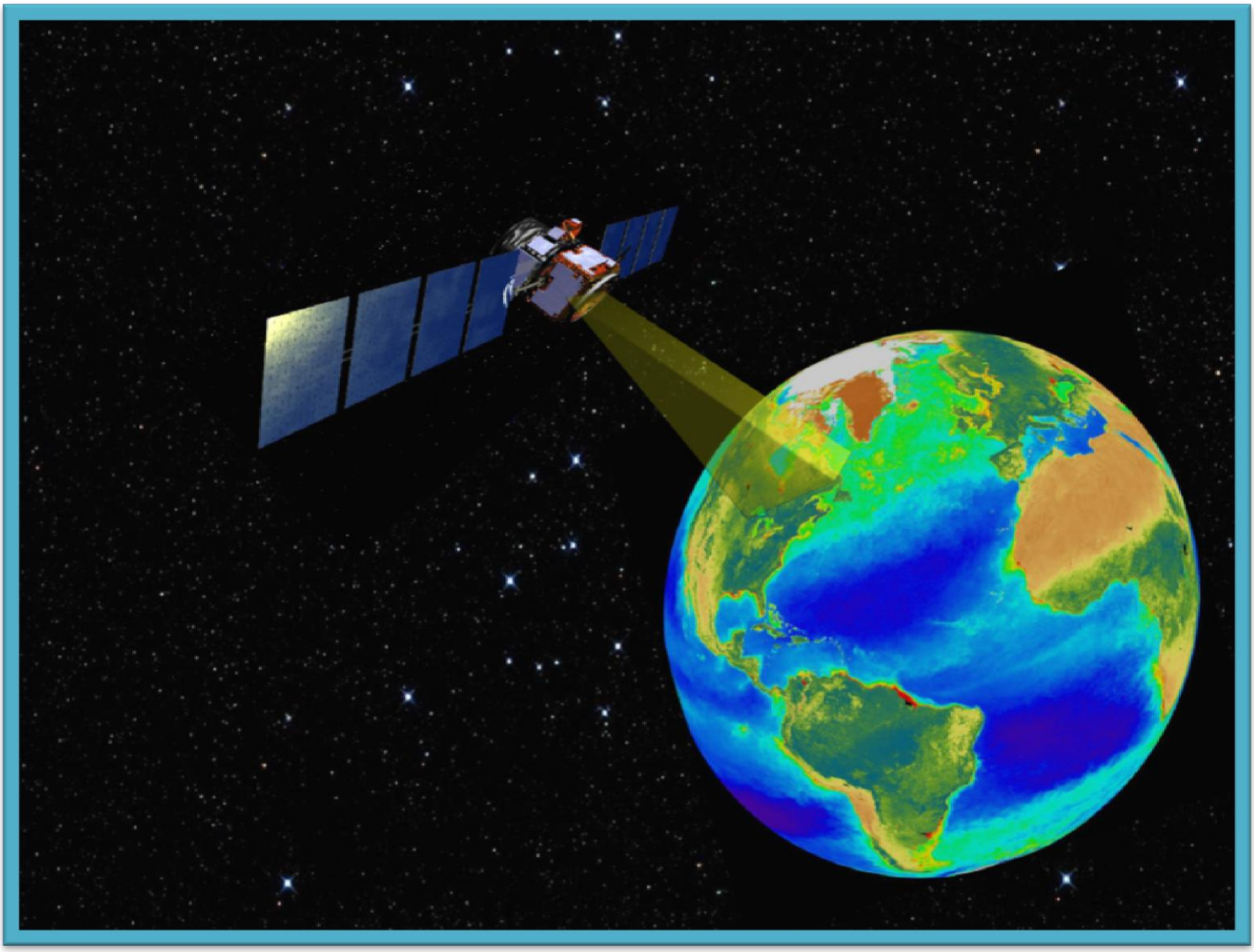
Sensitivity Analysis of Gaussian Process Machine Learning for Chlorophyll Prediction from Optical Remote Sensing

—

Katalin Blix

FYS-3941 Master's Thesis in Applied Physics and Mathematics

May 2014



UNIVERSITY OF TROMSØ

Contents

Abstract

The machine learning method, Gaussian Process Regression (GPR), has lately been introduced for chlorophyll content mapping from remotely sensed data. It has been shown that GPR has outperformed other machine learning and empirical methods in accuracy, speed and stability. Moreover, GPR not only estimates the chlorophyll content, it also provides the certainty level of the prediction, allowing the assessment of additional certainty maps. However, since GPR is a non-linear kernel based regression method, the relevance of the features are not accessible directly from the weights. The main contribution of this thesis is to develop a procedure for feature sensitivity analysis in order to assign relative importance to the features. The sensitivity analysis was introduced for the predictive mean function and for the predictive variance function of the Gaussian process. Then the empirical estimates for the derived sensitivity functions were applied to a land chlorophyll dataset and to two ocean chlorophyll datasets. The sensitivity analysis revealed the most important spectral bands for land chlorophyll and for ocean chlorophyll prediction. Applying the proposed methodology to the land chlorophyll dataset discovered that bands outside the chlorophyll absorption spectrum also contribute to the prediction of chlorophyll. The results of the sensitivity analysis of the ocean chlorophyll datasets open the possibility of discriminating between Case-1 water and Case-2 water condition. The method also provides additional information through the sensitivity of the predictive variance. Thus, not only the most relevant spectral bands can be revealed, but also the stability of the variance for the feature in interest can be accessed.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor Associate Professor Robert Jenssen for his excellent supervision, guidance and help throughout the writing of this thesis. The door to his office was always open whenever I had a question. Thank you for all the encouragement, inspiration, comments, discussions and kindness during my work. It was a pleasure to be able to work with this very exciting development in methodology and applying it to the chlorophyll datasets. I have enjoyed every moment of working with my thesis. You have motivated me all the way during my studies, shared your knowledge with me and stimulated me to pursue a career in research.

I would like to acknowledge Associate Professor Gustavo Camps-Valls for providing me with the chlorophyll datasets, allowing me to study the astonishing world of the primary producers.

I would like to acknowledge all my teachers at the University of Tromsø for their patience and guidance during my years in Tromsø. Thank you all for teaching me to be strong and committed.

Foremost, I would like to thank my mum, my sister Boglárka and my niece Kincső Katalin, for their endless support and love. Thank you for accepting that I study and live in Norway. Thank you for all the skype conversations, which gave me the power to hold out and manage to build up my life by myself, far away from my home and my beloved. Thank you for visiting me whenever I missed you so much that I could not go on, regardless of time and expense. I would especially like to express my deepest appreciation to my mum for the long discussions in connection with my studies. You have shared with me your bright ideas, which always helped me to solve my tasks.

In particular, I would like to thank my boyfriend Erik. There are no worlds to describe how grateful I am for meeting you. This thesis would never have been born without you. Thank you for your excellent ideas, advice, inspiration and support. You kept me motivated all the way during my years at the University of Tromsø. You are the best thing that ever happened to me.

List of Figures

List of Tables

Symbols and nomenclature

Part I (a)		
Symbol	**Property**	**Units**
a	absorption factor	
A_i	model coefficient	
b	bias term	
$\hat{b}$	estimated bias term	
B	regularization constant (SVM and KRR)	
c	speed of light	ms^{-1}
C	chlorophyll content	$\mu\mathrm{gl}^{-1}$
d	distance between the sun's surface and the Earth's surface	m
dA	aperture size	
E	energy collected by the sensor	Joule
$f(\cdot)$	activation function of the NN	
h	Planck constant	
h_{st}	sensor transfer function	
J	objective function of the SVM	
J_{KRR}	modified regularized loss function (KRR)	
$K(\cdot,\cdot)$	covariance matrix	
$\mathbf{L_b}$	radiance that reaches the bottom of the ocean	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
$\mathbf{L_p}$	path radiance	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
$\mathbf{L_s}$	radiance reflected by the water surface	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
$\mathbf{L_t}$	total collected radiance	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
$\mathbf{L_v}$	radiance that penetrates the air-water interface	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
L_w	water-leaving radiance	$\mathrm{Wsr}^{-1}\mathrm{m}^{-2}$
L_ϵ	epsilon insensitive lost function of the SVM	
P	power	W
r	distance between the aperture of the sensor and the Earth's surface	m
R	reflectance ratio	
R_s	radius of the sun	m
S	energy radiated from the source	Joule
$S_\mathbf{i}$	incoming radiant flux	W
$S_\mathbf{r}$	radiant flux at the sensor aperture	W
T_s	temperature of the sun	K
u	output of the linear combiner of the NN	

Part I (b)		
Symbol	**Property**	**Units**
w	synaptic weight of the neuron of the NN	
$\mathbf{w}$	weight (SVM)	
$\hat{\mathbf{w}}$	estimated weight (SVM)	
x	input	
$\mathbf{X}$	training data	
$\mathbf{X}_\star$	test data	
y	output	
$\hat{y}$	estimated output	
α_i	Lagrange multiplier	
α	vector containing the Lagrange multipliers	
β	constant	
ϵ	penalty term of the SVM	
λ	wavelength	m
ν	frequency	Hz
τ	dwell time	s
ρ	reflectance or surface albedo	
ρ_{NIR}	reflectance in the near-infrared region	
ρ_R	reflectance in the red region	
ζ	slack variable	
CI	chlorophyll vegetation index	
DVI	difference vegetation index	
$NDVI$	normalized difference vegetation index	
RVI	ratio vegetation index	

Part II	
Symbol	**Property**
$\mathbf{I}$	identity matrix
K	covariance function
$\mathbf{K} = \mathbf{K}(\mathbf{X}, \mathbf{X})$	covariance matrix between $\mathbf{X}$ and $\mathbf{X}$ (GP)
$\mathbf{K}_\star = \mathbf{K}(\mathbf{X}, \mathbf{X}_\star)$	covariance matrix between $\mathbf{X}$ and $\mathbf{X}_\star$ (GP)
$\mathbf{K}_{\star\star} = \mathbf{K}(\mathbf{X}_\star, \mathbf{X}_\star)$	covariance matrix between $\mathbf{X}_\star$ and $\mathbf{X}_\star$ (GP)
$\mathbf{K}_y = \mathbf{K} + \sigma_n \mathbf{I}$	noisy covariance matrix (GP)
$\mathbf{f} = f(\mathbf{x})$	function of the input $\mathbf{x}$
$\mathbf{f}_\star = f(\mathbf{x}_\star)$	function of the test input $\mathbf{x}_\star$
$p(\cdot)$	probability function
$\mathbf{x}$	input vector
$\mathbf{X}$	input matrix
$\mathbf{x}_\star$	test input vector
$\mathbf{X}_\star$	test input matrix
$x_\star$	test input point
$\mathbf{y}$	output vector
y	output point
R^2	coefficient of determination
$RMSE$	root-mean-squared error
SS_{res}	residual sum of squares
SS_{tot}	total sum of squares
$\boldsymbol{\mu}$	mean function
$\boldsymbol{\mu}_{\mathbf{f}_\star \mid \mathbf{f}}$	predictive mean function of the noiseless GP
$\boldsymbol{\mu}_{\mathbf{f}_\star \mid y}$	predictive mean function of the noisy GP
$\boldsymbol{\Sigma}$	covariance matrix of the Multivariate Gaussian distribution
$\mathbf{Var}_{\mathbf{f}_\star \mid \mathbf{f}}$	predictive variance function of the noiseless GP
$\mathbf{Var}_{\mathbf{f}_\star \mid y}$	predictive variance function of the noisy GP
σ_n^2	noise variance
ϵ	Gaussian noise
ν	scaling factor
λ	characteristic length-scale
$\boldsymbol{\Theta}$	vector collecting the hyperparameters

Part III	
Symbol	**Property**
b	bias term (linear regression)
$\mathbf{K}_\star = \mathbf{K}(\mathbf{X}, \mathbf{X}_\star)$	covariance matrix between $\mathbf{X}$ and $\mathbf{X}_\star$ (GP)
$\mathbf{K}_{\star\star} = \mathbf{K}(\mathbf{X}_\star, \mathbf{X}_\star)$	covariance matrix between $\mathbf{X}_\star$ and $\mathbf{X}_\star$ (GP)
$\mathbf{K}_y = \mathbf{K} + \sigma_n \mathbf{I}$	noisy covariance matrix (GP)
$p(\cdot)$	probability function
s	sensitivity
$\hat{s}$	estimated sensitivity
$\hat{s}_{\boldsymbol{\mu}}$	estimated sensitivity of the predictive mean function of the GP
$\hat{s}_{\mathbf{Var}}$	estimated sensitivity of the predictive variance function of the GP
$\mathbf{w}$	weight vector (linear regression)
$\mathbf{x}$	input vector
$\mathbf{x}_\star$	test input vector
$\mathbf{y}$	output vector
y	output point
$\hat{y}$	estimated output
MSE	mean squared error
$\boldsymbol{\mu}_{\mathbf{f}_\star \mid y}$	predictive mean function of the noisy GP
$\mathbf{Var}_{\mathbf{f}_\star \mid y}$	predictive variance function of the noisy GP
σ_n^2	noise variance
ν	scaling factor
λ	characteristic length-scale
$\phi(\mathbf{x})$	function of $\mathbf{x}$

Introduction

Remote sensing of chlorophyll from space provides the possibility to detect changes in chlorophyll content. Changes in chlorophyll concentration indicate changes in the photosynthetic activity. Both terrestrial and marine primary producers use photosynthesis to live and grow. However, following the photosynthetic activity of primary producers by chlorophyll content mapping has different applications, depending on whether the the environment is terrestrial or marine.

Monitoring the photosynthetic activity of terrestrial plants provides information about the vegetation productivity, vegetation stress and the land cover. It has been shown, that in a crop the concentration of chlorophyll is directly related to yield, allowing the possibility of estimating vegetation productivity from remotely sensed data [Cracknell et al., 2009]. Vegetation stress can be revealed by following the reflectance spectrum over time. Shifts in the position of the red-edge might indicate stressed vegetation due to water deficiency, pollutant stress, disease or stress due to heavy metals. Gaining knowledge about the vegetation productivity and health status of vegetation provides useful information for agriculture applications [Joiner et al., 2013]. Land cover mapping is mostly used for forest applications. In addition, remote sensing of chlorophyll content offers the possibility to the assessment of the terrestrial carbon budget [Joiner et al., 2013].

Remote sensing of ocean chlorophyll content allows to access information about the marine primary producers, the phytoplankton. Phytoplankton are the beginning of the marine food chain, thus their presence indicate the occurrence of fish. Ocean color images are widely used in the fishing industry for locating areas rich in fish [1]. In addition to the ecological importance of phytoplankton, they also act like biological pumps due to their capability of removing carbon dioxide (CO_2) from the atmosphere. Since phytoplankton are photosynthetic organisms, they take up carbon dioxide [Reynolds, 2006]. Dying phytoplankton sink to the bottom of the oceans, where they accumulate in the sediments. Thus, they remove CO_2 from the atmosphere. Therefore, monitoring ocean chlorophyll content provides a tool to achieve deeper understanding of the contribution of CO_2 to the climate [2].

All these applications require an accurate and fast chlorophyll content mapping. There have been developed several parametric and machine learning methods for chlorophyll content estimation from multi- and hyperspectral data [Verrelst et al., 2012b]. Parametric methods, such as vegetation indices and spectral band ratios, have been widely used for chlorophyll content estimation purposes due to their fast performance. Some of the disadvantages of these models are that they might be affected by confounders and they cannot adapt to changes in the variation of biophysical parameters and illumination, which might lead to inaccurate predictions. In addition, they make use of only a few spectral bands for chlorophyll estimation, thus resulting uncertainty whether the right combination of spectral bands are being used [Verrelst et al., 2012b]. In order to overcome the drawbacks of methods based on vegetation indices and spectral band ratios, machine learning methods have been introduced. One of the advantages of the ma-

[1] http://oceancolor.gsfc.nasa.gov/

[2] earthobservatory.nasa.gov/Features/Phytoplankton/page2.php

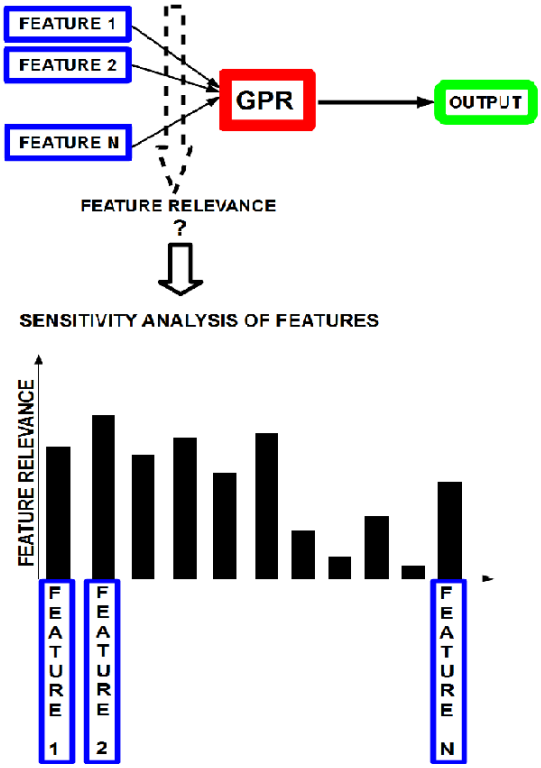


Figure 1: Sensitivity analysis of GPR.

chine learning methods is that they are not based on the relationship between the variables [Verrelst et al., 2012a]. Machine learning models have shown an improved performance, specially the recently introduced Gaussian Process Regression (GPR) [Verrelst et al., 2012a]. GPR differs from other machine learning methods not only in its predictive power, but also in its principles. The methodology of the GPR is based on Bayes' theorem, which can be followed back to basic axioms of probability. The other advantageous property of GPR is that it provides an additional information about the prediction, the predictive variance. Thus the output of the regression is not only the estimated chlorophyll content, but also the estimated variance, which reveals the strength of the prediction.

Although GPR has shown an excellent predictive performance, the information about the relative relevance of the features being used for regression is lost, since GPR is a non-linear kernel method. It hasn't been proved yet what controls the regression. Therefore, the main contribution of this thesis was to focus on the derivation of a new methodology for feature selection, the sensitivity analysis of features, in order to discover the driving mechanisms of the GPR. Figure 1 illustrates how the sensitivity analysis of features can be used to reveal feature relevance in the GPR. The sensitivity analysis aimed to reveal the importance of the spectral bands being used for chlorophyll content mapping, thus retrieving information about characteristics of the study areas. Another advantage of applying sensitivity analysis for GPR is the possibility of accessing information about the variance of the most relevant bands. This is in contrast with other methods, since only the GPR provides the predictive variance in addition to the predictive mean. The sensitivity analysis of features discovered that not only bands around the red-edge are the most important for land chlorophyll estimation, but also bands that fall outside the chlorophyll absorption region. In the case of the sensitivity analysis of ocean chlorophyll, the method revealed that the most relevant bands for ocean chlorophyll content

estimation differ whether Case-1 water or Case-2 water conditions are presence.

Although this thesis focused on the prediction of chlorophyll, the methodology of the sensitivity analysis translates to other fields as well[3].

Structure of the thesis

This thesis contains three parts.

Part I

The first part describes the background theory. It gives an overview of the importance of the chlorophyll molecule. Since this thesis focused on chlorophyll content estimation, it is important to gain deeper understanding of the output, the estimated chlorophyll. Then it describes the remote sensing sensors and the methods being used for land chlorophyll and ocean chlorophyll content mapping, because the input data originates from passive imaging systems. Finally, examples for machine learning methods for chlorophyll content prediction are discussed. The examples of the machine learning methods are described for the benefit of the reader, thus allowing the reader to understand the advantages of the GPR in comparison of other machine learning methods, which have been used for chlorophyll content mapping from remotely sensed data.

Part II

The second part of the thesis gives a detailed description of the Gaussian Process Regression (GPR). In order to understand why GPR differs from other machine learning methods, in addition to introducing the principles of the method, the derivation from the Multivariate Gaussian distribution is presented. Thus it can be realized, that GPR is a Bayesian statistical model, in contrast to the other machine learning models. Then GPR is applied to a land chlorophyll dataset and to two ocean chlorophyll datasets. The results of the regressions are presented and discussed, and the predictive performance of the GPR is evaluated by using widely used model criteria.

Part III

The third part is the main focus of the thesis, namely the sensitivity analysis of features. First the principles of the sensitivity analysis are described. Then the sensitivity analysis is derived for the GPR and the performance of it is tested on two controlled datasets. Finally, the sensitivity analysis is applied to the land chlorophyll dataset and to the ocean chlorophyll datasets. Then the resulting most relevant features are interpreted by using the gained knowledge from Part I and from Part II.

[3]The illustration of the translational nature of the methodology can be seen in Appendix A

Part I

Background theory

Chapter 1

The importance of chlorophyll in photosynthesis

Plants, phytoplankton, algae and certain bacteria are capable of harvesting light energy in order to build up their own molecules [Cotterill, 2002]. The process, which uses light energy to fuel metabolism, is called photosynthesis. The process, photosynthesis, in phytoplankton does not differ from land plants [Reynolds, 2006], thus the description of the photosynthesis is valid both for marine phytoplankton and for land plants. During photosynthesis, carbon-dioxide and water is converted to glucose and oxygen in the presence of light energy [Cotterill, 2002]. This can be expressed by the following formula

$$6CO_2 + 6H_2O \quad \xrightarrow{light \quad energy} \quad C_6H_{12}O_6 + 6O_2, \tag{1.1}$$

where the term CO_2 is the carbon-dioxide, H_2O is water, $C_6H_{12}O_6$ is glucose and O_2 is oxygen. Photosynthesis in leafs and phytoplankton takes place in specialized organelles called chloroplasts. Chloroplasts contain an inner membrane called thylakoid membrane which forms long folds in the stroma of the chloroplast. A green pigment called chlorophyll is located in the folded thylakoid membrane[1]. Figure 1.1 shows the structure of a chloroplast and the thylakoid membrane.

[1] http://www.nature.com/scitable/topicpage/photosynthetic-cells-14025371

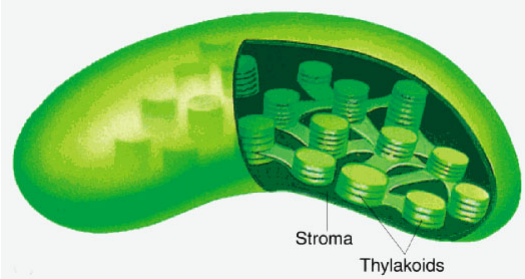


Figure 1.1: Chloroplast (from www.uic.edu).

1.1 Chlorophyll, the green pigment

The reason that certain organisms can perform photosynthesis by capturing photons is that they contain chlorophyll which can absorb incident light. The electromagnetic spectrum spreads from short wavelengths (10^{-14} m) to long wavelengths (10^{2} m). However, chlorophyll molecules can absorb only certain wavelengths of light from the visible part of the electromagnetic spectrum. Figure 1.2 shows the electromagnetic spectrum. The visible light region of the spectrum is a short interval of the entire spectrum.

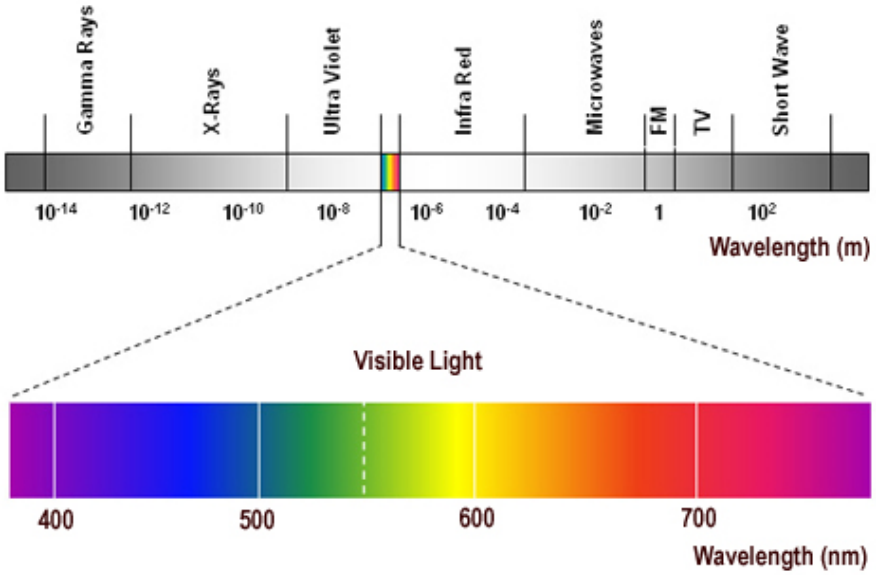


Figure 1.2: The electromagnetic spectrum (from www.pro-lite.co.uk).

There are several types of chlorophyll, namely chlorophyll a, b, c and d. Green leafs (higher plants) contain both chlorophyll a and b in order to increase the efficiency of the photosynthesis [Cotterill, 2002]. Chlorophyll a and b absorb red and blue light, and reflect green light. Thus chlorophyll can absorb photons only with certain wavelengths. Chlorophyll a and b have characteristic light absorption maxima, shown in Fig. 1.3.

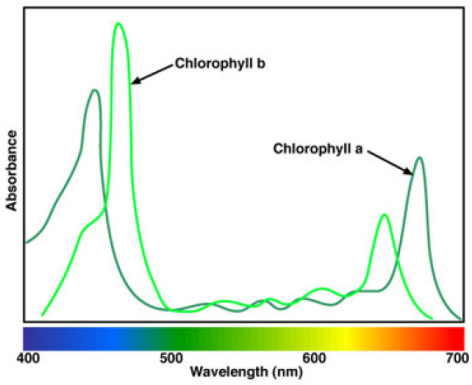


Figure 1.3: Absorption spectrum of chlorophyll a and b. (Adopted from [Kemp et al., 2012].)

1.1.1 Trapping of light

When light illuminates plants and phytoplankton, photons get absorbed by both chlorophyll a and b molecules (and other accessory pigments). This can be described by

$$Chlorophyll + h\nu = Chlorophyll_{excited}, \tag{1.2}$$

where h is the Planck constant, ν is the frequency of the incoming photon before absorption and $Chlorophyll$ represents any chlorophyll pigment before the absorption of the photon. $Chlorophyll_{excited}$ is the chlorophyll, after the absorption of the photon. The subscript $excited$ indicates the new energy level of the Chlorophyll molecule. This is due to the fact that the absorption of the photon gave access energy to the Chlorophyll molecule, thus allowing it to reach a higher energy level. The frequency can be expressed as the fraction of the speed of light c and the wavelength λ, $\nu = \frac{c}{\lambda}$. Thus only certain photons with certain wavelengths can be absorbed by the chlorophyll molecules due to the quantization of energy levels [Govindjee, 1975]. This explains why only certain photons with certain wavelengths can be absorbed. Absorption of photons leads to excitations in these molecules [Roháček and Barták, 1999]. The access energy causes changes in the distribution of the electron structures of the chlorophyll molecules. Chlorophyll molecules are closely spaced thus allowing a rapid transfer of the excitation energy to each other in order to return to their ground state. However, due to the molecular difference between chlorophyll a and b, the efficiency of transferring excitation energy also differs slightly [Govindjee, 1975]. The energy transfer efficiency from chlorophyll b to a is believed to be 100 %. Excitation energy migrates from one pigment to another until it reaches an electron trap called *reaction center* where the energy gets trapped in order to drive photochemical reactions [Govindjee, 1975]. Figure 1.4 shows the pathway of energy migration.

This excitation energy drives photochemical reactions which makes photosynthetic processes possible. However, only a part of the excitation energy contributes to photochemical reactions. A part of this energy is converted to heat, which is a non-radiative de-excitation pathway, and a part of it is re-emitted in a form of *chlorophyll fluorescence*, which is a radiative de-excitation pathway [Roháček and Barták, 1999]. These three processes, photochemistry, heat loss and chlorophyll fluorescence compete for excitation energy [Baker, 2008]. Figure 1.5 shows the schematic presentation of the three competitive pathways.

Figure 1.5 shows that light energy is absorbed by chlorophylls which is associated with photosystem 2. There are two photosystems in plants and phytoplankton, photosystem 1 and 2. These photosystems are the collections of several hundreds of chlorophyll molecules (and

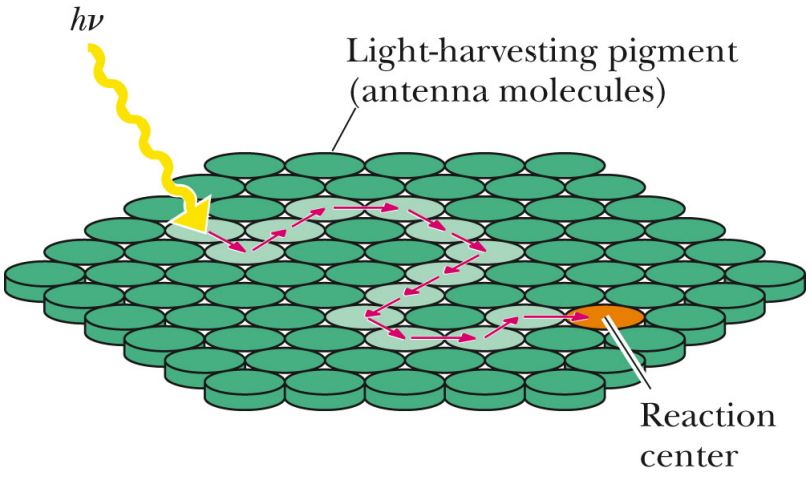


Figure 1.4: Energy transfer between the pigments (from www.studyblue.com).

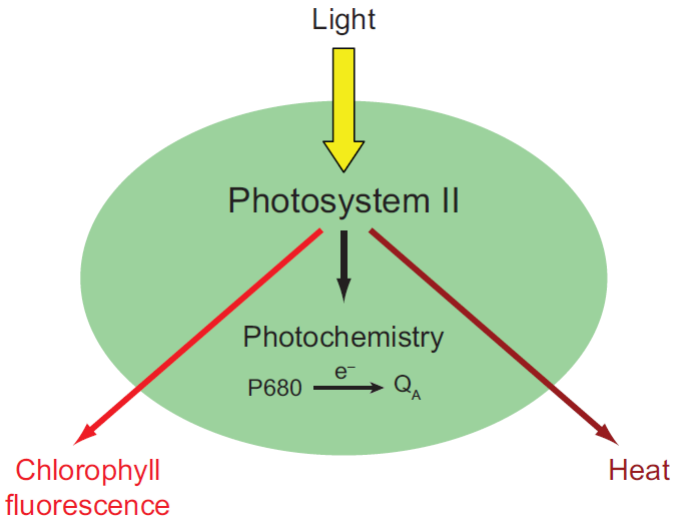


Figure 1.5: Schematic presentation of the three pathways [Baker, 2008].

other components) [Govindjee, 1975]. This energy can lead to photochemical reactions. During photochemical reactions an electron is transferred from the reaction center of the chlorophyll molecule (P680) to the primary acceptor (Q_A) [Baker, 2008]. Figure 1.5 also shows that the excitation energy can be lost in a form of heat or chlorophyll fluorescence. This can be described by

$$Clorophyll_{excited} \quad \rightarrow \quad Chlorophyll + h\nu^{'}, \tag{1.3}$$

where $Clorophyll_{excited}$ is the Chlorophyll molecule with the access energy arising form the absorbed photon. $Chlorophyll$ is the Chlorophyll molecule after emitting the access energy in a form of chlorophyll fluorescence $h\nu^{'}$, and returning into its ground state. The term $\nu^{'}$ corresponds to the frequency of the re-emitted photon. It is worth to mention that although both chlorophyll a and b are present in intact leaves of higher plants and in some phytoplankton, only chlorophyll a distributes to chlorophyll fluorescence due to the high efficiency of transferring excitation energy from chlorophyll b to chlorophyll a [Roháček and Barták, 1999]. Chlorophyll fluorescence is believed to be a predominant emission at room temperature, it represents only 3-5 % of the total absorbed energy [Roháček and Barták, 1999]. However measuring the time varying chlorophyll fluorescence emission allows to detect changes in the distribution of the absorbed energy. This is due to the fact that chlorophyll fluorescence pathway is in competition with the heat loss pathway and the photochemical pathway for the absorbed energy. Thus changes in the chlorophyll fluorescence emission indicates changes in the photochemical reactions and/or in heat conversion (due to the law of conservation of energy) [Baker, 2008].

Figure 1.6 summarizes the relationship between the absorption and the fluorescence of the chlorophyll molecules. Figure 1.6 shows the distinct energy levels of a chlorophyll molecule and the corresponding absorbance spectrum. Figure 1.6 also shows that fluorescence takes place from the lowest excited state to the ground state. The other states (higher states) do not contribute to fluorescence due to rapid relaxation which doesn't result radiation. The fluorescence intensity shows the fluorescence spectrum. It can be observed that the fluorescence maximum is slightly shifted to longer wavelengths (red line) comparing to the absorption maximum (black line). This is due to thermal relaxations [Govindjee, 1975].

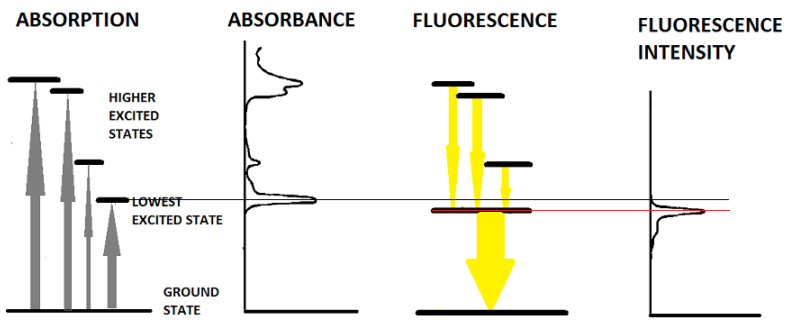


Figure 1.6: Absorption and fluorescence (after [Govindjee, 1975]).

1.2 Chlorophyll content prediction

Measuring chlorophyll fluorescence allows to gain knowledge about the health status of the vegetation [Verrelst et al., 2011]. Chlorophyll fluorescence provides information about the rate of photosynthesis and the amount of chlorophyll [Abbott and Letelier]. Fluorescence emission increases under stress conditions. Fluorescence emission shows an inverse correlation with photosynthetic activity [Chaerle et al., 2007]. This is due to the fact that the electron transport is partially blocked in the photosynthetic process [Edner et al., 1995]. Chlorophyll fluorescence signal can be retrieved from satellite measurements which allows to estimate the chlorophyll content [Verrelst et al., 2012a].

Chapter 2

Passive remote sensing of vegetation

This chapter is mainly based on the books by [van Zyl, 2006]
and [Jones & Vaughan, 2010]. Passive remote sensing detects reflected sun radiation without using artificially generated radiation. Passive remote sensing sensors capture radiation most commonly in the visible and near-infrared part of the electromagnetic spectrum. These visible and infrared imaging sensors monitor the Earth's surface. Thus passive imaging systems can give informations about the vegetation properties by collecting the reflected radiation and studying the spectral characteristic of it. Therefore, understanding the principles of the way these systems operate, is necessary.

2.1 Principles of passive imaging systems

In order to understand how satellites can collect information about vegetation, the basic principles of passive imaging systems are presented. Figure 2.1 shows how information about the Earth's surface can be collected. Figure 2.1 is discussed in detail in this section.

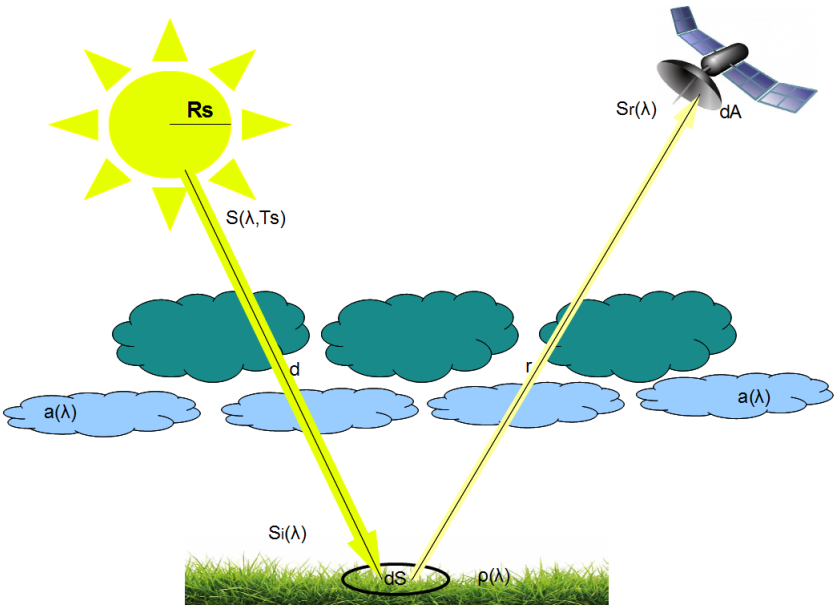


Figure 2.1: Principles of remote sensing

The source of illumination is the sun. The sun radiates energy with a given temperature T_s and wavelength λ. Denote this energy $S(\lambda, T_s)$. This energy is radiated spherically from the sun's surface. Denote the radius of the sun R_s. $S(\lambda, T_s)$ decreases with increasing distance from the sun's surface. Denote the distance between the sun's surface and the Earth's surface d. When radiation propagates through the Earth's atmosphere a part of it is absorbed by the particles of the atmospheric components. Denote this absorption factor $a(\lambda)$. The spectrum of the incoming radiant flux at the surface of the Earth can be written by

$$S_{\mathbf{i}}(\lambda) = S(\lambda, T_s) \left(\frac{R_s}{d}\right)^2 a(\lambda), \tag{2.1}$$

where the index **i** indicates that the radiant flux is **i**ncoming. A part of $S_{\mathbf{i}}(\lambda)$ is reflected by the Earth's surface. This ability of the Earth's surface is called reflectance or surface albedo. The surface albedo depends on the characteristics of the surface and it is denoted $\rho(\lambda)$. The reflected radiation propagates through the atmosphere $a(\lambda)$ once again until it reaches the aperture of the sensor. The energy, carried by the reflected radiation decreases with increasing distance from the Earth's surface. Denote the distance between the aperture of the sensor and the Earth's surface r. For Lambertian surfaces[1] this attenuation can be expressed by $\frac{1}{2\pi r^2}$. Thus the radiant flux at the aperture of the sensor $S_{\mathbf{r}}(\lambda)$ can be written by

$$S_{\mathbf{r}}(\lambda) = S_{\mathbf{i}}(\lambda)\rho(\lambda)dSa(\lambda)\frac{1}{2\pi r^2} = S(\lambda, T_s) \left(\frac{R_s}{d}\right)^2 a^2(\lambda)\rho(\lambda)dS\frac{1}{2\pi r^2}, \tag{2.2}$$

where dS is the surface element, which reflects the incoming radiant flux. The index **r** in $S_{\mathbf{r}}(\lambda)$ refers to **r**eflected. The power $P(\lambda)$ collected by the sensor with an aperture size dA is

$$P(\lambda) = S_{\mathbf{r}}(\lambda)dA = S(\lambda, T_s) \left(\frac{R_s}{d}\right)^2 a^2(\lambda)\rho(\lambda)dS\frac{1}{2\pi r^2}dA. \tag{2.3}$$

Usually the sensor collects $P(\lambda)$ over a given bandwidth (from λ_a to λ_b) for a short time, called dwell time τ, with a given efficiency. The efficiency of the sensor can be described by the sensor transfer function $h_{st}(\lambda)$. Thus the energy collected by the sensor (E) can be expressed by

$$E = \int_{\lambda_a}^{\lambda_b} S(\lambda, T_s) \left(\frac{R_s}{d}\right)^2 a^2(\lambda)\rho(\lambda)dS\frac{1}{2\pi r^2}dAh_{st}(\lambda)\tau d\lambda. \tag{2.4}$$

This integral can be approximated for relatively short bandwidths by

$$E \approx S(\lambda_0, T_s) \left(\frac{R_s}{d}\right)^2 a^2(\lambda_0)\rho(\lambda_0)dS\frac{1}{2\pi r^2}dAh_{st}(\lambda_0)\tau\Delta\lambda, \tag{2.5}$$

where λ_0 denotes the wavelength at the center of the measurement bandwidth and $\Delta\lambda$ is the bandwidth.

2.2 Passive imaging systems for monitoring vegetation

Figure 2.2 shows that incident solar radiation on green leaves is partially absorbed, in order to drive photosynthetic processes, and partially reflected due to the characteristics of the plant. A small fraction of the absorbed radiation is re-emitted as fluorescence. Thus the radiation measured by the sensor is the composition of the reflected radiation and the fluorescence. Figure 2.2 also shows the chloroplast where reflectance, absorption and fluorescence takes place. This section gives an overview about passive imaging sensors which can measure the composition of the reflected radiation and fluorescence.

[1] Lambertian surface: reflected light is scattered uniformly over the hemisphere.

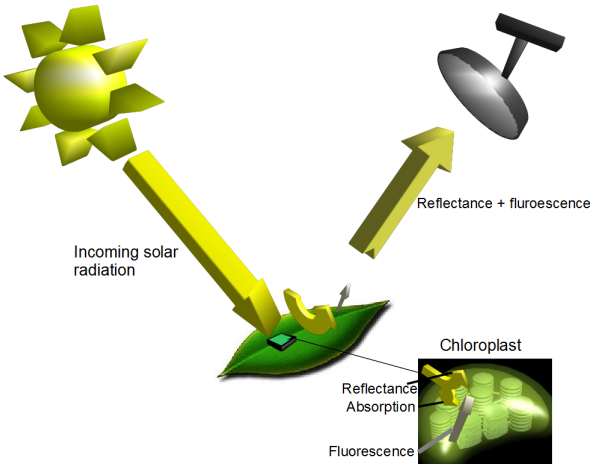


Figure 2.2: Monitoring vegetation.

2.2.1 Resolution

Spatial resolution

A surface element can be represented by a pixel. Spatial resolution refers to the size of the pixel. Spatial resolution depends on the optical system of the sensor and the height above the ground. Passive imaging systems can be grouped into low-resolution systems, medium-resolution systems and high-resolution systems [Jones & Vaughan, 2010]. Low-resolution systems have a spatial resolution of approximately 1 km, medium-resolution systems from 100 m to 1 km and high-resolution systems can have a spatial resolution from 5 m to 100 m [Wojtaszek, 2010]. Low-resolution systems can image a large area with coarse details, while high-resolution systems can image a smaller area with fine details[2].

Spectral resolution

Sensors can detect over broad spectral bands or over narrow spectral bands. The spectral resolution of sensors operating on many narrow spectral bands is high. These imaging system are referred to as multi-, super-, and hyperspectral imagery depending on the number of spectral resolution channels they operate on. Multispectral imagers use a few wavebands, superspectral imagers record between 10 and 50 spectral bands and hyperspectral imagers can record between 50 and 200 wavebands. Hyperspectral imagers are often referred to as imaging spectrometers [Jones & Vaughan, 2010]. Images taken by imaging spectrometers are called multiple images, where each image corresponds to one waveband. Such multiple images can be stacked in order to form a three-dimensional cube, where x and y corresponds to the spatial coordinates and the third direction λ corresponds to the spectral dimension. Figure 2.3 shows a hyperspectral image cube. The advantage of hyperspectral imagers over multi-, and superspectral imagers is that they contain more information. Thus finer details can be retrieved [Jones & Vaughan, 2010]. Figure 2.4 shows the difference between an multispectral and an hyperspectral image cube.

[2]from: http://www.nrcan.gc.ca/earth-sciences/geography-boundary/remote-sensing/fundamentals/985

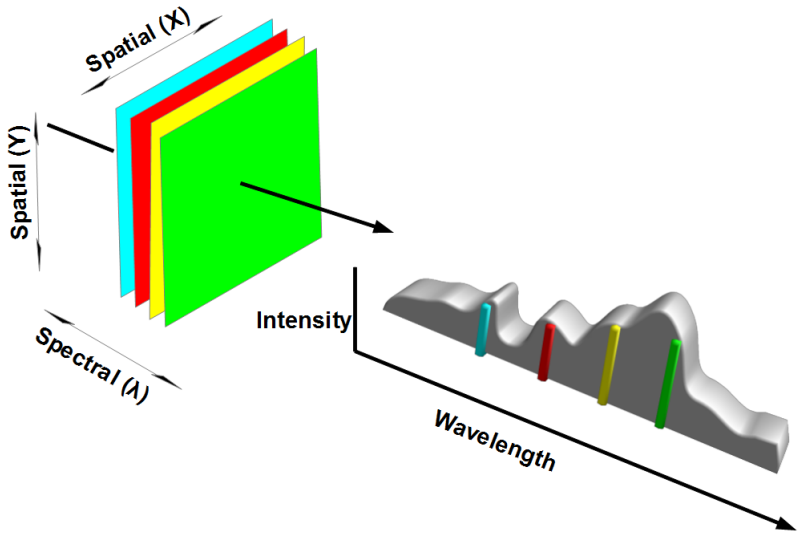


Figure 2.3: Hyperspectral image cube (after www.ece.gatech.edu).

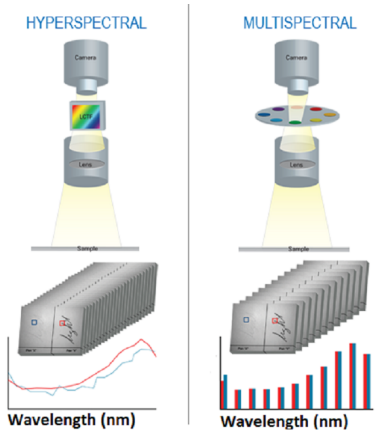


Figure 2.4: Multispectral and hyperspectral image cube (from www.chemimage.com).

2.2.2 The structure of the imaging systems

Figure 2.5 shows the main components of the imaging systems.

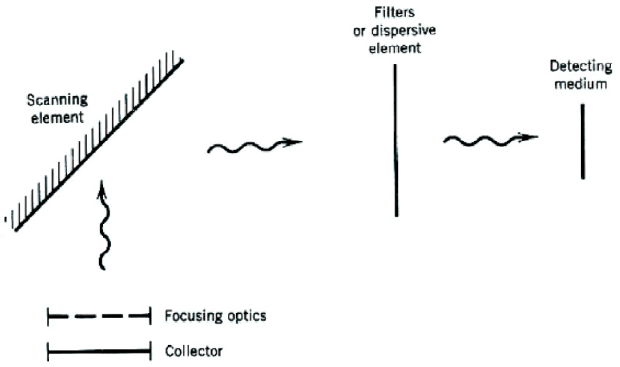


Figure 2.5: Structure of the imaging systems (from [van Zyl, 2006]).

The size of the collecting aperture determines the available energy for the sensor. The collector (lens or curved mirror) collects the radiation. Then the focusing optics focuses the radiation on the detecting medium through the scanning element (if the imaging systems has only a few detecting elements) and the dispersive element, which splits the incident radiation into several spectral components. Finally the detecting medium transforms the collected energy into a recordable information. In the visible and near-infrared part of the spectrum CCD (charge-coupled device) detectors are used due to their sensitivity to wavelength between 400 and 1100 nm.

2.2.3 The types of the imaging systems

There are three main types of imaging systems, referred to as framing cameras, scanning systems and pushbroom imagers [van Zyl, 2006]. Figure 2.6 shows the three type of imaging systems. The swath width refers to the length of the scan line of the detector. (Swath is the strip of the surface under the platform where data is collected from.) Along-track direction refers to the parallel direction of the line of flight, while cross-track direction is the direction perpendicular to the line of flight [Jones & Vaughan, 2010]. The framing camera takes a snapshot of the surface area. Scanning systems and pushbroom imagers build up an image line by line as the platform moves. Scanning systems project the image of one pixel on a single detector. It uses across-track scanning, thus lines are scanned as the platform propagates. Pushbroom imagers use along track scanning. It uses a line array of detectors in order to cover all pixels in cross-track direction allowing a longer dwell time on each pixel.

Passive imaging systems can operate both spaceborne and airborne. In the following section an example of a spaceborne operating imaging spectrometer and an example of an airborne imaging spectrometer are described.

2.2.4 The spaceborne imaging spectrometer: CHRIS

CHRIS (Compact High Resolution Imaging Spectrometer) is a high-resolution imaging spectrometer [Jones & Vaughan, 2010]. CHRIS acquires images over the Earth surface in the spectral range between 415 nm and 1050 nm [Cutter et al., 2000]. It can operate in various modes

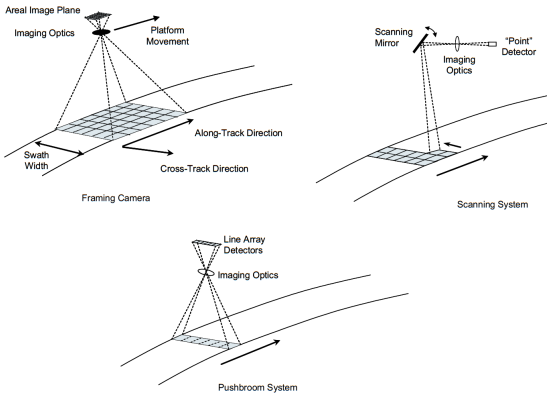


Figure 2.6: Imaging systems (from [van Zyl, 2006]).

[Verrelst et al., 2012a]. For instance, CHRIS can operate with 19 spectral bands with a spectral sampling interval between 1.25 nm and 11 nm at 25 m spatial resolution at nadir[3]. It can is also operate in a mode when 62 spectral bands are available at a 50 m spatial resolution [Jones & Vaughan, 2010]. CHRIS was mounted on a small satellite, called PROBA (Project for On-Board Autonomy). PROBA has operated at 830 km altitude in pushbroom mode [Cutter et al., 2000]. Operating in pushbroom mode has the advantage that the dwell time is longer, hence the spatial resolution and the width of the bandwidth is improved [van Zyl, 2006]. PROBA can acquire 5 images of one target at 5 different view of zenith angles[4] in one satellite overpass [Jones & Vaughan, 2010].

2.2.5 The airborne imaging spectrometer: CASI

CASI (Compact Airborne Spectrographic Imager) is also an imaging spectrometer operating in small aircraft. The spectral range of CASI is between 423 and 946 nm. It has 288 spectral bands [Borstad et al., 1989], and 1.9 nm sampling intervals [Jones & Vaughan, 2010]. CASI provides 512 spatial pixels per scan line [Jones & Vaughan, 2010].

2.3 Remote sensing of terrestrial chlorophyll

Chapter 2 described the principles of passive remote sensing and the importance of passive imaging systems in monitoring vegetation. One of the purposes of monitoring vegetation by passive imaging systems is to map the chlorophyll content of plants. Figure 2.2 shows the connection between the measured signal and plant activity (absorption, reflectance and chlorophyll fluorescence). Chapter 1 gave on overview about the importance of chlorophyll fluorescence. It should be emphasized that the measured signal gives information about the reflected radiation and the re-emitted radiation (chlorophyll fluorescence). Chlorophyll fluorescence is related to chlorophyll content. It was also mentioned that the chlorophyll fluorescence signal is tiny, only 1 % or 2% of the absorbed light [Verrelst et al., 2012a]. In order to use passive imaging systems to monitor chlorophyll content based on the measured chlorophyll fluorescence signal, various approaches have been developed. The following sections give an overview of the most commonly used chlorophyll retrieval methods.

[3]Nadir: looking vertically downwards [Jones & Vaughan, 2010].
[4]Zenith: vertically upwards [Jones & Vaughan, 2010].

2.3.1 Vegetation indices

Vegetation indices are dimensionless quantities. They indicate the presence of green vegetation. Vegetation indices are based on the characteristics of chlorophyll absorption. The presence of chlorophyll indicates a strong absorption at wavelengths shorter than 700 nm, strong reflectance between 700 and 1300 nm. The spectral reflectance between 1300 and 2500 nm is almost the same as for pure water [van Zyl, 2006]. The absorption decreases rapidly around 700 nm. In the observed spectrum this is manifested as a sharp increase in reflectance (around 700 nm). The sudden change at the red-edge (700 nm) indicates the presence of green vegetation. There haven't been found evidence that other natural surfaces could show this rapid change [Jones & Vaughan, 2010]. Measuring the variation in spectral signature of vegetation allows the estimation of the health status of vegetation. Thus changes in spectral signature indicates changes in chlorophyll concentration. Figure 2.7 shows the characteristic change at the red-edge for green vegetation. It also shows how the reflectance changes for unhealthy vegetation. There are several types of vegetation indices. In the following a few commonly used

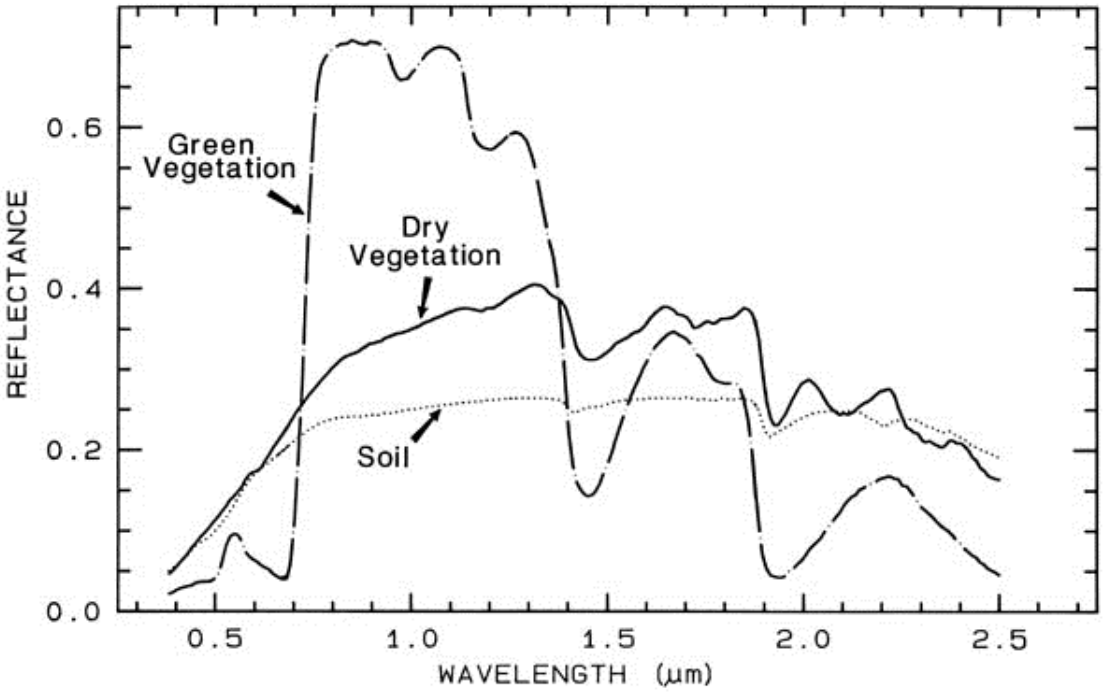


Figure 2.7: Spectral reflectance of various natural surfaces (from http://bluemarble.ch/).

vegetation indices are presented. The presented vegetation indices are based on the book from [Jones & Vaughan, 2010].

Difference vegetation index

The difference vegetation index (DVI) can be expressed by

$$DVI = \rho_{NIR} - \rho_R, \tag{2.6}$$

where ρ_{NIR} is the reflectance in the near-infrared and ρ_R is the reflectance in the red region of the spectrum. For green vegetation the DVI results in a greater value than for bare soil surface. This is due to the fact that healthy vegetation has the sharp change between the near-infrared and red region (Fig. 2.7), while bare soil doesn't have this characteristic.

Ratio vegetation index

The ratio vegetation index (RVI) is the ratio of ρ_{NIR} and ρ_R,

$$RVI = \frac{\rho_{NIR}}{\rho_R}. \tag{2.7}$$

Normalized difference vegetation index

The normalized difference vegetation index ($NDVI$) is the fraction of DVI and the sum of the ρ_{NIR} and ρ_R,

$$NDVI = \frac{\rho_{NIR} - \rho_R}{\rho_{NIR} + \rho_R}. \tag{2.8}$$

This model has the advantage that it ranges between 0 and 1 in the absence of clouds, snow and water surfaces.

Chlorophyll vegetation index

$$CI_{590} = \frac{\rho_{880} - \rho_{590}}{\rho_{590}}, \tag{2.9}$$

where CI_{590} is chlorophyll index. The index 590 and 880 refers to the wavelength (nm). Using the observed reflectance at 590 nm (ρ_{590}) and 880 nm (ρ_{880}) in order to calculate CI_{590} gives the best sensitivity to crop chlorophyll content.

Chapter 3

Ocean colour remote sensing

Chapter 2 introduced the principles of passive imaging systems and the applications of passive imaging systems for monitoring vegetation. Passive imaging systems are also used for monitoring ocean chlorophyll content. However, the composition of the collected radiance by the sensor is different for oceans since the optical properties of oceans differ from the terrestrial ones. This chapter gives an overview of the collected water leaving radiance by the sensor and describes two examples for ocean colour remote sensing. The total collected radiance $\mathbf{L_t}$ by the sensor can be written by

$$\mathbf{L_t} = \mathbf{L_p} + \mathbf{L_s} + \mathbf{L_v} + \mathbf{L_b}, \tag{3.1}$$

where $\mathbf{L_p}$ is the path radiance, also referred to atmospheric noise, $\mathbf{L_s}$ is the radiance, which is reflected by the water surface, $\mathbf{L_v}$ is the radiance that penetrates the air-water interface and interacts with the constituents of the water and $\mathbf{L_b}$ is the radiance that reaches the bottom of the ocean and gets reflected by the surface of the ocean bottom [Jensen, 2007]. Figure 3.1 illustrates the paths of the radiances which reach the sensor. In order to retrieve information about the ocean chlorophyll content, $\mathbf{L_v}$ has to be isolated from the other three radiances which contribute to the total measured radiance. $\mathbf{L_v}$ can be expressed from Eq. (3.1) by

$$\mathbf{L_v} = \mathbf{L_t} - (\mathbf{L_p} + \mathbf{L_s} + \mathbf{L_b}). \tag{3.2}$$

In practise the isolation of $\mathbf{L_v}$ is carried out by radiometric correction of the collected data, thus $\mathbf{L_p}$, $\mathbf{L_s}$ and $\mathbf{L_b}$ can be removed [Jensen, 2007]. Furthermore, the component $\mathbf{L_v}$ does not necessary provide exclusive information about the ocean chlorophyll content. This is due to the

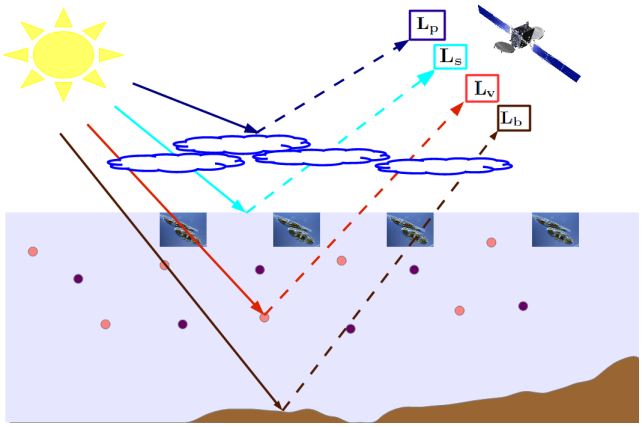


Figure 3.1: The received water leaving radiance by the sensor.

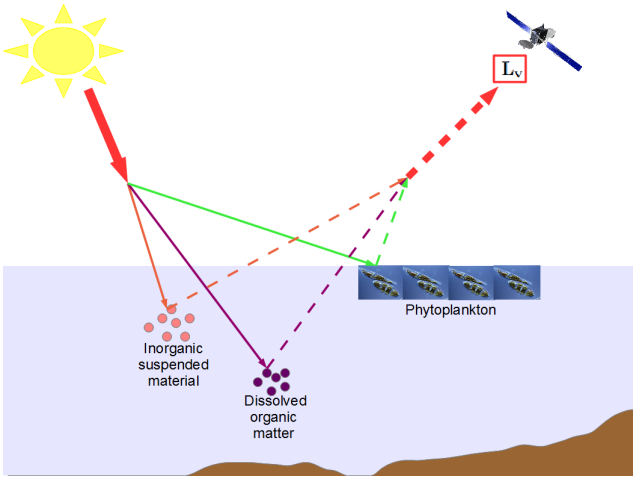


Figure 3.2: The composition of $\mathbf{L_v}$.

fact, that ocean water might consist of the following constituents in addition to pure water: inorganic suspended material, dissolved organic matter (DOM) and phytoplankton [Li et al., 2009]. These constituents have different optical properties, thus they result the characteristic colour of the sea. Thus $\mathbf{L_v}$ can be expressed as a function of the pure seawater w, inorganic suspended material ISM, dissolved organic matter DOM and the chlorophyll content Chl. This can be written by

$$\mathbf{L_v} = f\left(w_\lambda, ISM_\lambda, DOM_\lambda, Chl_\lambda\right), \tag{3.3}$$

where λ indicates the wavelength [Jensen, 2007]. Figure 3.2 illustrates the composition of $\mathbf{L_v}$. In order to understand how the chlorophyll content of the oceans can be recovered from remotely sensed data, the constituents which contribute to the spectral characteristics of the ocean water are described.

Pure seawater

Pure seawater can absorb and scatter the incoming solar radiation. Absorption by seawater increases with increasing wavelength, thus most red light from the visible spectrum is absorbed before it could be scattered towards the sea surface and then to the sensor. Whereas scattering takes place mostly in the blue part of the spectrum and it decreases with increasing wavelength. This results in the blue appearance of the pure seawater [Robinson, 2004].

Inorganic suspended material

Inorganic suspended materials refer to suspended sediments which might originate from bottom sediments, river-borne particles, eroded coastal and beach deposits and long- and short-range of atmospheric particulates [Li et al., 2009]. The composition and size distribution of the suspended sediments have a great variety [Robinson, 2004]. These various properties of the inorganic suspended materials contribute to the colour of the sea. Due to the great diversity of suspended sediments, a universal model for the absorption and scattering spectrum has not yet been achieved. However, since these inorganic suspended material appears usually in coastal waters, empirical models for specific locations have been developed [Robinson, 2004]. It has been observed that increasing suspended sediment content results an increase across the reflectance spectrum [Robinson, 2004].

Dissolved organic matter

Dissolved organic matter originates from decaying vegetation [Jensen, 2007]. Decaying vegetation might come from marine organism and/or input from terrestrial matter [Li et al., 2009]. Dissolved organic matter consists of humic acids and fluvic acids [Robinson, 2004]. Dissolved organic matter also referred to yellow substances or *gelbstoff* due to their optical properties. They have a strong absorption in the blue part of the visible spectrum. The absorption decreases with increasing wavelength, resulting a least absorption in the yellow, middle part of the spectrum [Robinson, 2004].

Phytoplankton

The description of the contribution to the water-leaving radiance from phytoplankton is based on the book by [Robinson, 2004]. Phytoplankton are small plant like organisms. Just like terrestrial plants they use photosynthesis in order to build up their own molecules. Their main photosynthetic pigment is chlorophyll-a. Some species also contain accessory pigments as well, such as chlorophyll-b, chlorophyll-c and carotenoids. These various pigments have their own characteristic absorption spectra, each with a peak at different wavelengths. There are two certain characteristics absorption peaks, one at about 443 nm, and another one at about 675 nm. However, due to "packing" effect of the pigments in the cells and additional pigments arriving from decaying phytoplankton and primary production wastes, the peaks of the absorption spectrum are broadened. Scattering by cells also appears, although it is small and it is not uniform. Thus the reflectance spectrum due to absorption and scattering from the phytoplankton population tends to decrease at wavelength below around 540 nm, and increases at longer wavelength. This effect is enhanced when chlorophyll concentration increases, with a minimum at around 440 nm due to chlorophyll absorption. The other minimum occurs at around 660 nm. This minimum might be masked by chlorophyll fluorescence, which has its reflectance peak at around 685 nm. The range between 550 nm and 600 nm shows a quite constant reflectance spectrum independent of the amount of the chlorophyll content.

It can be concluded that all these constituents contribute to the colour of the sea dependently of their amount in the oceans. Therefore, the estimation of ocean chlorophyll content considers the presence and the amount of these constituents. A general approach, which considers the contribution of the constituents to the measured reflectance spectra, is the grouping of oceans.

3.1 Case-1 and Case-2 waters

Global waters can be divided into Case-1 and Case-2 waters based on their optical properties [Robinson, 2004]. Case-1 water refers to waters whose inherent optical properties are dominated by phytoplankton and phytoplankton degradation products. Most open ocean waters are Case-1 waters. Case-2 waters apply to all other waters, such as coastal and inland waters. Case-2 waters can contain coloured dissolved organic matter and/or inorganic particles instead of (or in addition to) phytoplankton [Matsushita et al., 2012]. Due to the different content of Case-1 and Case-2 waters the optical properties are also different depending on the type of water. Therefore, care should be taken when chlorophyll content estimation from satellite data is carried out. Chlorophyll content estimation is more straight forward for Case-1 waters, due to the characteristic reflectance spectra of ocean chlorophyll. However, estimating chlorophyll content from Case-2 waters is more challenging due the additional reflectance spectra of other materials besides (or instead of) chlorophyll [Robinson, 2004].

3.2 SeaWiFS

SeaWiFS (Sea-viewing Wide-Field-of-view Sensor) is an advanced scanning system, developed specifically for ocean monitoring. It was carried by the satellite called SeaStar on an orbit 705 km above the Earth [Jensen, 2007]. It had the capability for tilting the sensor scan axis by 20° forwards and backwards thus sun glint could be reduced or avoided [Robinson, 2004]. SeaWiFS operated with eight spectral channels. The channels ranged from 402 nm to 885 nm, where the bandwidths were 20 nm or 40 nm. The spatial resolution was 1.13 km × 1.13 km at nadir, and the swath width was 2801 km [Hooker et al., 1992].

3.3 MERIS

The description of the MERIS instrument is based on the article by [Rast et al., 1999]. MERIS (Medium-resolution imaging spectrometer) was launched on the European environmental satellite, Envisat-1. MERIS is an imaging spectrometer, and it operates in push-broom mode. The pixel size at nadir is 260 m across-track and 300 m along-track. MERIS can operate with 15 programmable spectral bands between 390 nm and 1040 nm. The spectral bands can range from 1.25 nm to 30 nm. The swath width is 1150 km. The acquired data can have two spatial resolutions, 300 m and 1200 m. The main purpose of MERIS is to provide information about the biophysical parameters of the oceans, with particular focus on the ocean chlorophyll content.

3.4 Remote sensing of marine chlorophyll

3.4.1 Spectral band ratios

Retrieval of ocean chlorophyll from water-leaving radiance spectrum analytically has not yet been derived. The most common approach of estimating ocean chlorophyll from satellite ocean color data is based on empirical algorithms. These algorithms attempt to fit regression models to real and simulated data [Robinson, 2004]. In the following the most commonly used estimation methods from spectral band ratios are presented. The description of the methods are based on the book by [Robinson, 2004].

CZCS algorithms

The CZCS (Coastal Zone Color Scanner) algorithms were derived after water-leaving radiance L_w become available by launching the CZCS on the Nimbus-7 satellite in 1978. The water-leaving radiance, L_w, refers here to the radiance that has penetrated the ocean and a part of it was reflected by the sea bottom, and a part of it was absorbed and/or scattered by the sea water and the constituents of the sea water before it left the sea surface. Thus L_w corresponds to the composition of $\mathbf{L_b}$ and $\mathbf{L_v}$. The CZCS algorithms are also referred to simple blue-green band ratios. The blue-green band ratio algorithms are based on the characteristic spectra of phytoplankton populations. Phytoplankton populations tend to decrease the reflectance around 540 nm due to absorption and backscattering. This effect is enhanced by increasing chlorophyll content, resulting a minimum at 440 nm. Therefore, the most important spectral bands for ocean chlorophyll estimation are centred at 443 nm, 520 nm and 550 nm, where the bandwidths are 20 nm. Thus the estimated chlorophyll content C in $\left[\frac{\mu g}{l}\right]$ can be expressed by

$$C = 1.130 \left(\frac{L_w^{443}}{L_w^{550}} \right)^{-1.71} \qquad \text{for} \qquad C < 1.5 \tag{3.4}$$

and

$$C = 3.326 \left(\frac{L_w^{520}}{L_w^{550}} \right)^{-2.44} \qquad \text{for} \qquad C < 1.5, \tag{3.5}$$

where L_w is the water-leaving radiance and the superscripts indicate the wavelength in nm. Equation (3.4) is the most sensitive, due to the absorption maximum of the chlorophyll, which appears around 443 nm. However, when chlorophyll content increases, the absorption of blue light increases as well, which results a small value for the water-leaving radiance L_w^{443}, therefore Eq. (3.5) provides a better estimate for chlorophyll content. Even-though ocean chlorophyll content estimates for satellite data were derived, it was desired to improve their accuracies. In 1997 the SeaWiFS (Sea-viewing Wide Field of view Sensor) sensor was launched in order to provide further satellite data. In addition, an ocean colour mission, called SeaBam (SeaWiFS Bio-optical Algorithm Mini-workshop), was held. SeaBam gathered a large dataset of matched radiance-chlorophyll data from 919 stations, thus allowing the improvement of the CZCS algorithms.

SeaBam algorithm

The SeaBam dataset allowed the derivation of new empirical algorithms. The first such algorithm was the OC2 algorithm, which was initially developed for processing SeaWiFS data. The OC2 algorithm is based on the ratio R of the remote sensing reflectance R_{rs} on 490 nm and 555 nm, $R = R_{rs}^{490}/R_{rs}^{555}$. The OC2 algorithm can be written by

$$C = 10^{A_0 + A_1 R + A_2 R^2 + A_3 R^3} + A_4, \tag{3.6}$$

where C is the chlorophyll content, R is the reflectance ratio and $A_i, i = 1, 2, 3, 4$ are the model coefficients. The OC2 algorithm provided a better fit than the CZCS algorithms. However, the OC2 algorithm was replaced by an improved model, called the OC4 algorithm. The OC4 model is an updated version of the OC2 model, the only difference is that the band ratio R is now the one with the largest value selected from $R = R_{rs}^{443}/R_{rs}^{555}$, $R = R_{rs}^{490}/R_{rs}^{555}$ and $R = R_{rs}^{510}/R_{rs}^{555}$. The model coefficients for the OC2 and OC4 algorithm are presented in Tabel 3.1. The two

Table 3.1: The model coefficients for the OC2 and OC4 algorithm

Model	A_0	A_1	A_2	A_3	A_4
OC2	0.3410	-0.30010	2.8110	-2.0410	-0.0400
OC4	0.4708	-0.38469	4.5338	-2.4434	-0.0414

models can be compared by computing model statistics, such as the root mean squared error, RMSE. The resulting RMSE for the OC2 model was $RMSE_{(OC2)} = 0.172$, while the RMSE for the OC4 model was slightly lower, $RMSE_{(OC4)} = 0.156$. Thus the OC4 algorithm provided a better fit. Therefore, the OC4 algorithm has been used in order to estimate ocean chlorophyll content from SeaWiFS data. In order to achieve further improvement of the algorithm, more data is being collected, and added to the database. However, since the further development of an empirical model is limited to the adjustments of the model coefficients, machine learning methods have been investigated in order to provide ocean chlorophyll estimates from satellite data.

Chapter 4

Examples of machine learning methods

This chapter gives an overview how machine learning methods can compete with vegetation indices and spectral band ratios in the prediction of chlorophyll from remotely sensed data. Since this thesis has focused on studying the relatively new machine learning method, the GPR, it is important to understand the contribution of this new method compared to the already existing machine learning models.

Machine learning methods learn the relationship between the input variable (reflectance and chlorophyll fluorescence) and the training data (actual measured chlorophyll content). The established relationship can be used in order to predict chlorophyll content. Figure 4.1 illustrates the learning mechanism of machine learning system. The input is the training data, which is fed to the learning system. The produced output of the system is then compared to the target output (desired output). Based on some criterion an error function/ error term is defined. After minimising the error, the system is updated. The process continues until there are no changes in the error term. In [Verrelst et al., 2011] several machine learning algorithms were tested

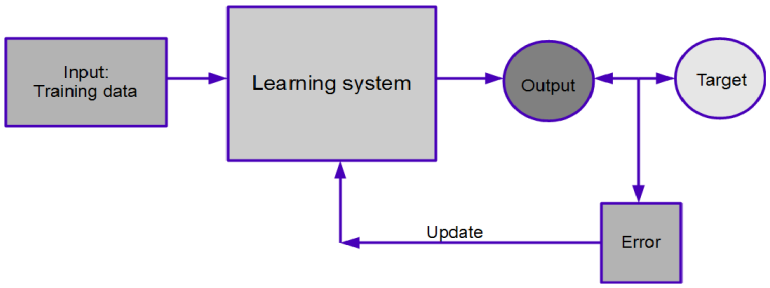


Figure 4.1: Machine learning flowchart.

in order to find the best method to predict chlorophyll content. The tested algorithms were the following: neural network, support vector regression, kernel ridge regression and Gaussian process regression [Verrelst et al., 2011]. In the following these machine learning methods are briefly discussed.

4.1 Neural Network

Neural networks have been used for predicting ocean-chlorophyll content and land-chlorophyll content from remote sensing data. Neural networks are nonlinear models with the ability of learning the relationship between the input (satellite-received radiance) and the output (chlorophyll content) [Canziani et al., 2008]. The description of the neural networks is based on the book

by [Haykin, 1999]. A neural network consist of neurons. Neurons are information-processing units, which are connected to each other and organized in layers. These information-processing units are the basis of the neural networks. Neurons consists of a set of connecting links, each characterized by a weight, a linear combiner and an activation function. Figure 4.2 shows the structure of a neuron. Neuron k can be expressed by

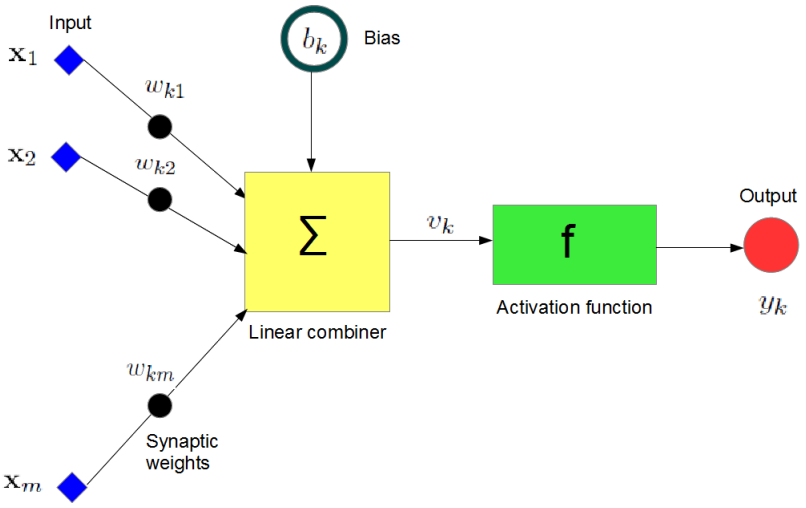


Figure 4.2: Model of a neuron.

$$u_k = \sum_{j=1}^{m} w_{kj} x_j, \tag{4.1}$$

where $x_j | j = 1, .., m$ are the inputs, $w_{kj} | j = 1, .., m$ are the synaptic weights for neuron k and u_k is the output of the linear combiner. The output y_k of the neuron k is given by

$$y_k = f(u_k + b_k), \tag{4.2}$$

where $f(\cdot)$ is the activation function and b_k is a bias term. The activation function limits the amplitude of the output of a neuron. The activation function is often referred to as squashing functions, because it squashes the amplitude range of the output to some finite value. There exists a great variety of activation functions. One of the most commonly used activation function is the sigmoid function due to its advantageous properties. The sigmoid function only produces outputs between 0 and 1 and it is easily differentiable, allowing computational simplicity. In order to learn the connection between the inputs and outputs by using neural network, the output y_k of neuron k is compared to the target output t_k. This comparison is done by defining an error term, $\epsilon_k = t_k - y_k$. Thus the learning process of the neural network consists of minimizing the error terms, then updating the weights iteratively, until convergence occurs. Minimizing the error terms is done by defining a cost function $\varepsilon = \frac{1}{2}\epsilon_k^2$, and differentiating the cost function with respect to the weights. The learning process is illustrated on Fig. 4.3. Figure 4.3 shows that the output y_k of the neuron is compared to the target t_k through the error term. After minimizing the difference between the output and the target, the weights are updated. This process continues until convergence occurs. Usually neural networks consist of several neurons, which are connected to each other. This type of neural network are referred to multilayer neural networks. Neural networks have been applied for chlorophyll prediction both for marine data and terrestrial data.

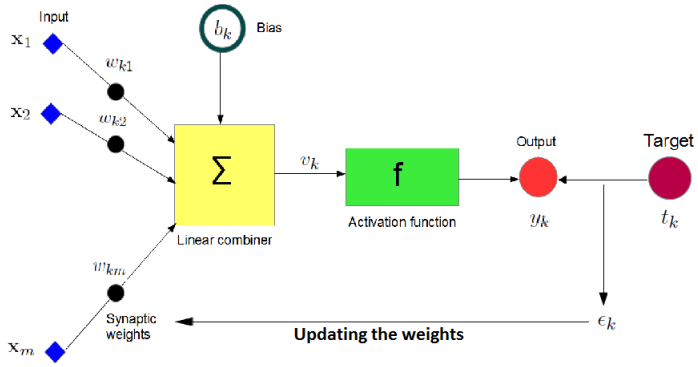


Figure 4.3: Learning process of a neuron.

Neural networks for ocean chlorophyll prediction

Neural networks showed a good performance in the prediction of ocean-chlorophyll content [Canziani et al., 2008]. However, neural networks have limitations in modelling processes, since only generalized process statements can be made [Whitehead et al., 1997]. In order to apply neural networks for high dimensional input data, a great number of parameters need to be obtained, such as the number of hidden layers, the number of hidden nodes, activation functions, epochs, weights initialization methods and parameters of the training algorithm. This might lead to computational difficulties [Zhan et al., 2003].

Neural networks for land chlorophyll prediction

The performance of neural networks have also been tested for land-chlorophyll prediction. Despite the fact that neural networks can learn the relationship between the measured radiance and the canopy parameter, it showed instability when it was applied to validation data. In addition, neural networks have been computational expensive comparing to other machine learning methods, such as support vector regression, kernel ridge regression and Gaussian process regression [Verrelst et al., 2011].

4.2 Support Vector Regression

An alternative for neural networks is Support Vector Regression (SVR) [Verrelst et al., 2011]. SVR is a useful function estimation method for high-dimensional input space. SVR has the advantage that the number of free parameters depend only on the the number of support vectors, but they are independent of the dimensionality of the input space [Vapnik et al., 1997]. The following paragraph is based on the book by [Murphy, 2012]. In order to make the solution vector $\mathbf{w}$ independent of the training inputs, the epsilon insensitive loss function was proposed. The epsilon insensitive loss function is defined by

$$L_\epsilon = \begin{cases} 0 & \text{if } |y - \hat{y}| < \epsilon \\ |y - \hat{y}| - \epsilon & \text{otherwise.} \end{cases} \tag{4.3}$$

Where y is the output, $\hat{y}$ is the predicted output and ϵ is the penalty term. Equation (4.3) suggests that only points outside an ϵ-tube are penalized. Furthermore, an objective function

has to be defined in order to perform regression. The objective function can be written by

$$J = B\sum_{i=1}^{N} L_\epsilon(y_i, \hat{y}_i) + \frac{1}{2}||\mathbf{w}||^2, \tag{4.4}$$

where the predicted output $\hat{y}_i = f(\mathbf{x}_i) = \mathbf{w}^T\mathbf{x}_i + b$ and the constant $B = \frac{1}{\beta}$ is a regularization constant. The term $\mathbf{x}_i$ is the input, $\mathbf{w}$ is the weight and b is the bias. This objective function is not differentiable. An approach to overcome this problem is to introduce slack variables ζ_i^+, ζ_i^-. The slack variables are informative about the degree to which each point lies outside the ϵ-tube. The slack variables can be expressed by

$$y_i \leq f(\mathbf{x}_i) + \epsilon + \zeta_i^+ \tag{4.5}$$

$$y_i \geq f(\mathbf{x}_i) + \epsilon + \zeta_i^-. \tag{4.6}$$

Thus Eq. (4.4) can be written by

$$J = B\sum_{i=1}^{N}(\zeta_i^+ + \zeta_i^-) + \frac{1}{2}||\mathbf{w}||^2. \tag{4.7}$$

Minimizing Eq. (4.7) with respect to the linear constraints in Eq. (4.5) and Eq. (4.6), and also with respect to the positive constraints $\zeta_i^+ \geq 0$ and $\zeta_i^- \geq 0$ yields that the estimated weights $\hat{\mathbf{w}}$ has a solution

$$\hat{\mathbf{w}} = \sum_i \alpha_i \mathbf{x}_i, \tag{4.8}$$

where $\alpha_i \geq 0$ is the Lagrange multiplier. Thus the predicted output can be computed by using Eq. (4.8)

$$\hat{y}(\mathbf{x}) = f(\mathbf{x}_i) = \hat{\mathbf{w}}^T\mathbf{x}_i + \hat{b} = \hat{\mathbf{w}}^T\mathbf{x} + \hat{b} = \left(\sum_i \alpha_i \mathbf{x}_i^T\mathbf{x}\right) + \hat{b}. \tag{4.9}$$

In order to get a kernelized solution introduce a kernel function κ for the term $\mathbf{x}_i^T\mathbf{x}$. Finally the predicted output expressed by the kernel function is

$$\hat{y}(\mathbf{x}) = \left(\sum_i \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})\right) + \hat{b}. \tag{4.10}$$

SVR for predicting ocean chlorophyll

In the work by [Zhan et al., 2003] SVR was applied in order to retrieve ocean chlorophyll. They showed that SVR has several advantageous properties, such as accuracy, fewer parameters as by neural networks, unique, global minimum solution and high generalization ability. Furthermore, the generalization performance of SVR doesn't depend on the dimensionality of the input space. [Kwiatkowska and Fargion, 2003] also showed that SVR is capable of learning the complex relationship between radiance and actual ocean chlorophyll data, and it can adopt to the variations of the sensor, such as seasonal trends, scan angles and spatial variations.

SVR for predicting land chlorophyll

The application of SVR to land chlorophyll data showed a fast training performance. In addition, SVR was quite accurate in order to predict land chlorophyll content [Verrelst et al., 2011]. However, SVR gives rise to a quadratic programming problem, due to the cost function and the linear constrains [Verrelst et al., 2011].One way to overcome the quadratic programming problem is to introduce Kernel Ridge Regression for chlorophyll prediction. In the following the KRR model is described.

4.3 Kernel Ridge Regression

The description of Kernel Ridge Regression is based on the book by [Schölkopf et al., 2013]. Kernel Ridge Regression is a simplified special case of Support Vector Regression. In section 4.2 it was showed that the predicted output can be computed by estimating the weights. The estimation of the weights was carried out by minimizing Eq. (4.7) with respect to $\zeta_i^+ \geq 0$ and $\zeta_i^- \geq 0$ and the linear constraints in Eq. (4.5) and Eq. (4.6). KRR simplifies the problem by ignoring the bias terms and setting $\epsilon = 0$. In addition, KRR seeks to minimize the squared of the simplified loss function. Thus, the modified regularized loss function can be written by

$$J_{KRR} = \frac{B}{2}\sum_{i=1}^{N}\zeta^2 + \frac{1}{2}||\mathbf{w}||^2 = \frac{B}{2}\sum_{i=1}^{N}\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 + \frac{1}{2}||\mathbf{w}||^2. \tag{4.11}$$

Minimizing and solving Eq. (4.11) yields the solution of the predictive output. It can be shown that the predictive output can be expressed by [Verrelst et al., 2011]

$$\hat{y}_i = K(\mathbf{X}_\star, \mathbf{X})\left(K(\mathbf{X}, \mathbf{X}) + \frac{1}{B}\right)^{-1}\mathbf{y} = K(\mathbf{X}_\star, \mathbf{X})\boldsymbol{\alpha}, \tag{4.12}$$

where $K(\cdot, \cdot)$ is the covariance matrix, $\mathbf{X}$ is the training data, $\mathbf{X}_\star$ is the test data and $\boldsymbol{\alpha}$ is a vector containing the Lagrange multipliers α_i.

KRR for predicting land chlorophyll

Applying KRR instead of SVR in order to predict land chlorophyll has an advantage that the solution can be expressed in a closed form, thus quadratic programming can be avoided [Verrelst et al., 2011]. However, KRR is not sparse, thus all the training samples in the final solution includes the weight α_i [Verrelst et al., 2011].

Part II

Gaussian Processes

Chapter 5

Gaussian Process Regression

5.1 Background on Gaussian Processes

Gaussian Processes dates back to the 1940s, when the Wiener-Kolmogorov prediction theory was introduced in order to predict trajectories of military objects [MacKay, 2003]. In the 1950s, a South African mining engineer, D. G. Krige developed empirical methods in order to predict true ore-grade distributions. The method was based on data from already existing gold-mines [Minnitt and Assibey-Bonsu, 2003]. In 1963, a geostatistical, optimal spatial linear prediction method was introduced by a French mathematician Georges Matheron. The method is called kriging, after Krige's work. Kriging is a minimum-mean-squared-error method, used for spatial prediction [Cressie, 1993]. Kriging is similar to Gaussian Process regression, although the derivation and the interpretation differs from it. Since kriging was developed for geostatistical applications, it is mainly used for low-dimensional problems [Boyle, 2007]. Furthermore, kriging also tends to ignore any probabilistic interpretation of the model [MacKay, 2003].

In 1978 A. O'Hagan introduced a Bayesian approach for general regression, where a prior over functions was defined by using Gaussian processes [Sacks et al., 1989]. The use of Gaussian processes for machine learning appeared first in the 1990s. In 1994 Radford M. Neal showed that Bayesian neural networks converge to Gaussian processes under certain conditions. He investigated the properties of priors for network weights and biases. These priors were independent Gaussian distributions. He showed that as the number of hidden units goes to infinity, the prior over functions converges to a Gaussian process [Neal, 1995]. This led to the idea of investigating the replacement of supervised neural networks by Gaussian processes. In 1996 Christopher K. I. Williams and Carl Edward Rasmussen carried out experiments in order to approximate neural networks by Gaussian process. Furthermore they introduced a new machine learning method for regression with Gaussian processes and tested the performance of it on actual datasets. Gaussian Process regression was found to show a good performance on real-world problems [Williams and Rasmussen, 1996]. The introduction of Gaussian processes for regression resulted in interest for further investigations of the method and the application of the model. In addition to regression, Gaussian processes have also been developed for classification problems. However, using Gaussian processes for classification problems is more demanding. This is due to the fact, that Gaussian process regression assumes a Gaussian likelihood, and the Gaussian likelihood combined with a Gaussian process prior results in a Gaussian process posterior over functions. On the other hand, classification problems deal with discrete targets, thus Gaussian likelihood assumption is not an appropriate choice [Rasmussen and Williams, 2006]. However, it has been shown that it is possible to use Gaussian processes for classification tasks as well, with promising results [Rasmussen and Williams, 2006].

This thesis is mainly concerned with Gaussian processes for regression, therefore the classifi-

cation approach with Gaussian processes is not presented here. More details about the classification approach can be read in the book from [Rasmussen and Williams, 2006]. In the following the principles of Gaussian Process Regression is discussed and the derivation of the method is presented.

5.2 Principles of Gaussian Process Regression

This chapter introduces the principles of Gaussian Process Regression. The following paragraph is based on the book by [Murphy, 2012] and [Rasmussen and Williams, 2006]. Gaussian Process Regression (GPR) assumes that the output (chlorophyll content) is the function of the input (radiance). The regression part of the GPR is to learn the function, the connection between the actual chlorophyll content and the measured radiance, while the prediction part uses the regression to predict unseen function values for test inputs. Thus in order to be able to perform prediction, first the function between the output and input variables has to be established. To be able to establish the connection between the output and the input variables a training dataset has to be available.

Assume that the training dataset consists of a set of observations. Denote the training dataset to $D = \{(\mathbf{x}_i, y_i)|i = 1, 2, ..., N\}$, where the output y_i is the actual chlorophyll content and the input $\mathbf{x}_i$ is the measured radiance. Furthermore assume that under GPR y_i is a function of $\mathbf{x}_i$. This can be written by $y_i = f(\mathbf{x}_i) = \mathbf{f}_i$.

The approach is to infer distributions over these functions given D, and use this to make prediction for new inputs. The approach is based on Gaussian Processes (GP).

A GP defines a prior over functions which can be converted to a posterior over functions given D. In order to represent a distribution over a function, it is enough to define a distribution over the function's values at a finite, but arbitrary set of points, $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$. GP assumes that $p(f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_N))$ is jointly Gaussian with some mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$. It can be observed that the GPR differs from other machine learning methods in that way, that under GPR distributions over functions are inferred and a prior over functions is defined. Figure 5.1 illustrates the prior and the posterior of the GP for three random functions. Panel (A) shows the three random functions drawn from a GP prior. Panel (B) indicates the posterior of the three random functions, which is the prior conditioned on the observed data. The observed data is presented by the + signs. Figure 5.1 also shows the confidence region, which is the mean plus and minus twice the standard deviation. The certainty level is presented by the grey shaded area.

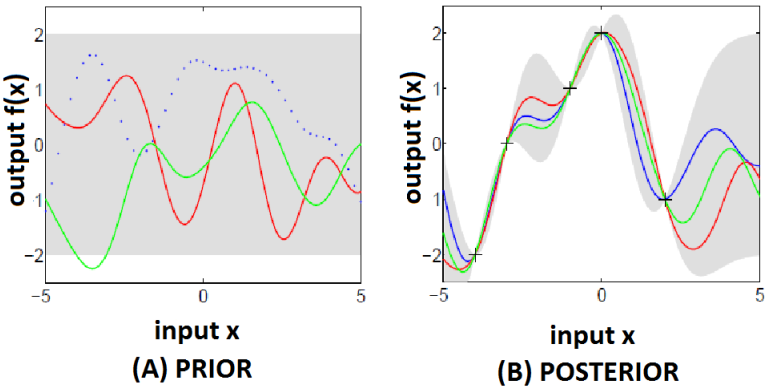


Figure 5.1: Fig.(A) shows three functions drawn from a GP prior. Fig.(B) represents the posterior of the three random functions. (Figure is adopted from [Rasmussen and Williams, 2006].)

A GP is completely specified in terms of a mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$. This can be expressed by

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ . \\ . \\ . \\ \mathbf{f}_N \end{bmatrix} \sim N\left(\boldsymbol{\mu}, \quad \mathbf{K}\right). \tag{5.1}$$

The mean is often assumed to be zero. This assumption is due to the flexibility of the GP [Murphy, 2012]. Thus Eq. (5.1) modifies to

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ . \\ . \\ . \\ \mathbf{f}_N \end{bmatrix} \sim N\left(\mathbf{0}, \quad \mathbf{K}\right), \tag{5.2}$$

where $\mathbf{K}$ is given by

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}. \tag{5.3}$$

The covariance $\mathbf{K}$ is composed by computing the covariance function $k(\cdot, \cdot)$ for each entry. Equation (5.2) can be learned when training data $D = \{(\mathbf{x}_i, y_i) | i = 1, 2, ..., N\}$ is available, since it is assumed that $\mathbf{f}_i = y_i = f(\mathbf{x}_i)$. For simplicity denote the joint distribution of the $\mathbf{f}_i$'s to $\mathbf{f}$, and express Eq. (5.2) by

$$\mathbf{f} \sim N\left(\mathbf{0}, \quad \mathbf{K}\right). \tag{5.4}$$

The regression part of the GPR is learning Eq. (5.4). Then use this to preform prediction for a new input. Assume a new input $\mathbf{x}_\star$. By using Eq. (5.4) the output $\mathbf{f}_\star$ of $\mathbf{x}_\star$ can be predicted. This is the prediction part of the GPR. The approach is to form the joint distribution of $\mathbf{f}$ with $\mathbf{f}_\star$, $p(\mathbf{f}_\star, \mathbf{f})$. Under GP this joint distribution is also jointly Gaussian distributed with zero mean

and a new covariance matrix $\mathbf{K}_{new}$. This can be expressed by

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_\star \end{bmatrix} \sim N\left(\mathbf{0}, \quad \mathbf{K}_{new}\right) = p(\mathbf{f}_\star, \mathbf{f}), \tag{5.5}$$

where the new covariance matrix is

$$\mathbf{K}_{new} = \begin{bmatrix} \mathbf{K} & \mathbf{K}_\star \\ \mathbf{K}_\star{}^T & \mathbf{K}_{\star\star} \end{bmatrix}, \tag{5.6}$$

where $\mathbf{K}_\star = \mathbf{K}(\mathbf{X}, \mathbf{X}_\star)$ is the covariance matrix between the training data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i]$ and the test data $\mathbf{X}_\star$, $\mathbf{K}_\star{}^T = \mathbf{K}(\mathbf{X}_\star, \mathbf{X})$ is the transposed of the covariance matrix between the training data and test data and $\mathbf{K}_{\star\star} = \mathbf{K}(\mathbf{X}_\star, \mathbf{X}_\star)$ is the covariance between the test data with itself. Equation (5.5) can be factorized as the posterior distribution time the marginal by applying Bayes' rule,

$$p(\mathbf{f}_\star, \mathbf{f}) = p(\mathbf{f}_\star|\mathbf{f})p(\mathbf{f}). \tag{5.7}$$

Here the term $p(\mathbf{f}_\star|\mathbf{f})$ is the posterior distribution and $p(\mathbf{f})$ is the marginal. It can be shown that

$$p(\mathbf{f}_\star|\mathbf{f}) \sim N(\boldsymbol{\mu}_{f_\star|f}, \mathbf{Var}_{f_\star|f}) \tag{5.8}$$

and

$$p(\mathbf{f}) \sim N(\mathbf{0}, \mathbf{K}). \tag{5.9}$$

Thus the expected value of predicted output $\mathbf{f}_\star$ is the mean of the posterior distribution and it is given by

$$\boldsymbol{\mu}_{f_\star|f} = \mathbf{K}_\star{}^T\mathbf{K}^{-1}\mathbf{f}, \tag{5.10}$$

and the variance is

$$\mathbf{Var}_{f_\star|f} = \mathbf{K}_{\star\star} - \mathbf{K}_\star{}^T\mathbf{K}^{-1}\mathbf{K}_\star. \tag{5.11}$$

The term $\mathbf{f}$ in Eq. (5.10) denotes the collection of the observed function values, the noiseless outputs. One of the advantages of the GPR is, that in addition to prediction it also provides the variance of the predicted value. Knowledge about the variance of the prediction reveals the certainty of the prediction.
However, usually the output is corrupted by noise. In this case the outputs of the training dataset modify to $\mathbf{y} = \mathbf{f} + \epsilon$, where ϵ is additively, independently and identically distributed Gaussian noise $\epsilon \sim N(0, \sigma_n^2\mathbf{I})$. Then Eq. (5.5) modifies to

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_\star \end{bmatrix} \sim N\left(\mathbf{0}, \quad \mathbf{K}_{noisy}\right), \tag{5.12}$$

where the noisy covariance $\mathbf{K}_{noisy}$ is given by

$$\mathbf{K}_{noisy} = \begin{bmatrix} \mathbf{K}_y & \mathbf{K}_\star \\ \mathbf{K}_\star{}^T & \mathbf{K}_{\star\star} \end{bmatrix} \tag{5.13}$$

and the submatrix $\mathbf{K}_y$ includes the noise term

$$\mathbf{K}_y = \mathbf{K} + \sigma_n^2\mathbf{I}. \tag{5.14}$$

Furthermore the joint distribution of the noisy outputs and the predictive output is given by

$$p(\mathbf{f}_\star, \mathbf{y}) = p(\mathbf{f}_\star|\mathbf{y})p(\mathbf{y}), \tag{5.15}$$

where $p(\mathbf{f}_\star|\mathbf{y}) \sim N(\boldsymbol{\mu}_{f_\star|y}, \mathbf{Var}_{f_\star|y})$ and $p(\mathbf{y}) \sim N(\boldsymbol{\mu}_y, \mathbf{Var}_y)$. Thus the noisy predictive mean is

$$\boldsymbol{\mu}_{f_\star|y} = \mathbf{K}_\star{}^T \mathbf{K}_y^{-1} \mathbf{y}, \tag{5.16}$$

and the noisy variance is

$$\mathbf{Var}_{f_\star|y} = \mathbf{K}_{\star\star} - \mathbf{K}_\star{}^T \mathbf{K}_y^{-1} \mathbf{K}_\star. \tag{5.17}$$

The next section shows how these results can be derived by factorizing the multivariate joint Gaussian distribution.

5.3 Factorization of the multivariate joint Gaussian distribution

In order to understand how the multivariate joint Gaussian distribution can be used for prediction, the factorization of the multivariate joint Gaussian distribution is performed. This section gives a general derivation of the factorization of the multivariate joint Gaussian distribution. The results of this derivation is applied for Gaussian Process regression. The starting point of the derivation is the multivariate Gaussian distribution.

Multivariate Gaussian

The multivariate Gaussian distribution of an l-dimensional random vector $\mathbf{x}$ is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{l/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \tag{5.18}$$

where $\boldsymbol{\mu}$ is the expectation, $\boldsymbol{\Sigma}$ is the $l \times l$ covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix [Theodoridis and Koutroumbas, 2009].

The notations in this section differ from Section 5.2 in order to indicate a general multivariate Gaussian distribution, which is a special case of the Gaussian Process. The notational changes are the following: Before $\mathbf{f}$ or $\mathbf{y}$ indicated a set of random variables of a stochastic process, here $\mathbf{x}$ denotes a vector, and $\mathbf{X}$ is the matrix collecting all the vectors $\mathbf{x}$ in its columns. In addition, the covariance matrix of the prior was earlier denoted to $\mathbf{K}$, here the covariance matrix is indicated by $\boldsymbol{\Sigma}$.

Conditional multivariate Gaussian distribution

Assume that $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is jointly Gaussian with the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Figure 5.2 shows a jointly Gaussian distribution for an arbitrary $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, with the mean $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and the covariance $\boldsymbol{\Sigma} = \begin{pmatrix} 0.25 & 0.2 \\ 0.2 & 1 \end{pmatrix}$. The left-hand panel in Fig. 5.2 is the probability density $p(\mathbf{x}_1, \mathbf{x}_2)$ and the right-hand panel is the underlying contour plot of the probability density. The joint distribution can be factorized by [Murphy, 2012]

$$p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1|\mathbf{x}_2) p(\mathbf{x}_2), \tag{5.19}$$

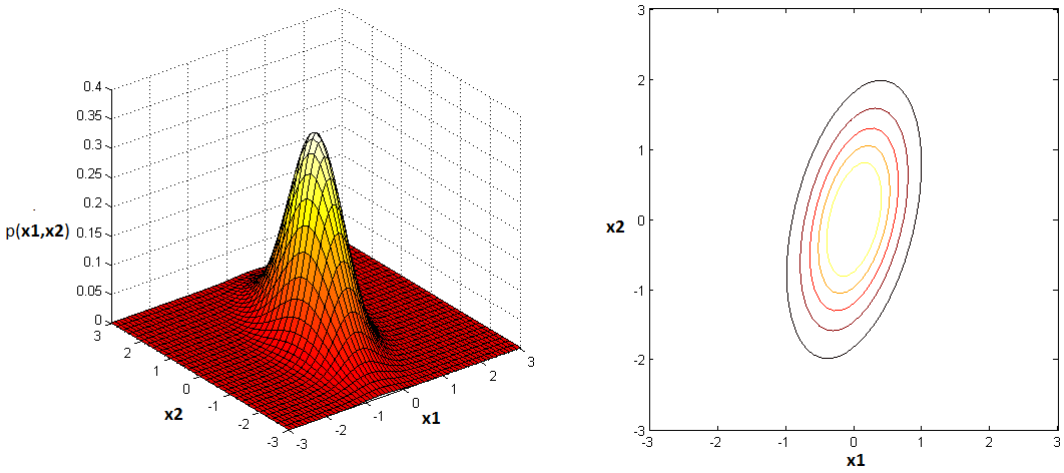


Figure 5.2: Multivariate jointly Gaussian distribution.

where $p(\mathbf{x}_1|\mathbf{x}_2)$ is the conditional distribution of $\mathbf{x}_1$ conditioned on $\mathbf{x}_2$ and $p(\mathbf{x}_2)$ is the marginal of $\mathbf{x}_2$. The conditional distribution can be applied for Gaussian Process regression. Therefore the main focus is the derivation of the conditional distribution. In order to find the conditional distribution of a multivariate Gaussian distributed variable, the joint distribution has to be computed. The joint distribution of $\mathbf{x}$ is given by [de Freitas, 2011]

$$p(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{(2\pi)^{(l_1+l_2)/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}\right), \tag{5.20}$$

where l_1 is the dimension of $\mathbf{x}_1$ and l_2 is the dimension of $\mathbf{x}_2$. Factorizing Eq. (5.20) allows the derivation of the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ and the marginal distribution $p(\mathbf{x}_2)$ [de Freitas, 2011]. In order to simplify the computations, first the exponential term was expanded.

The exponential term

Denote

$$P = \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T (\boldsymbol{\Sigma})^{-1} \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}, \tag{5.21}$$

where

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1}. \tag{5.22}$$

$$P = \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}^{-1} \begin{pmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{pmatrix}. \tag{5.23}$$

In order to factorize Eq. (5.23), $\boldsymbol{\Sigma}$ can be expressed by the Schur complement of the matrix. The Schur complement of the matrix $\boldsymbol{\Sigma}$ is [Petersen & Pedersen, 2008]

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{5.24}$$

The inverse of the matrix $\boldsymbol{\Sigma}$ can be expressed by using the Schur complement

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}. \tag{5.25}$$

In the following denote the Schur complement to [Murphy, 2012]

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}. \tag{5.26}$$

Thus Eq. (5.23) can be expressed by

$$P = \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}^T \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}. \tag{5.27}$$

Equation (5.27) can be factorized by applying matrix algebra. The steps of the computations are the following

$$\begin{aligned} P = &\begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \end{bmatrix} \times \\ &\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \end{aligned} \tag{5.28}$$

$$\begin{aligned} P = &\begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \end{bmatrix} \times \\ &\begin{pmatrix} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \end{aligned} \tag{5.29}$$

$$\begin{aligned} P = &\begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \times \\ &\begin{pmatrix} \mathbf{I} & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}. \end{aligned} \tag{5.30}$$

Then using the associative property of matrix multiplications Eq. (5.30) becomes

$$\begin{aligned} P = &\begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \times \\ &\begin{bmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1) - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ (\mathbf{x}_2 - \boldsymbol{\mu}_2) \end{bmatrix}. \end{aligned} \tag{5.31}$$

After completing the last step of matrix multiplications, the resulting equation is

$$\begin{aligned} P = &\left((\mathbf{x}_1 - \boldsymbol{\mu}_1)^T - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} \right) \left((\mathbf{x}_1 - \boldsymbol{\mu}_1) - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \right) \\ &+ (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2). \end{aligned} \tag{5.32}$$

Finally transposing Eq. (5.32) by using that [Petersen & Pedersen, 2008]

$$(A + B)^T = A^T + B^T,$$

$$(ABC...)^T = ...C^T B^T A^T$$

and

$$\left(A^{-1}\right)^T = \left(A^T\right)^{-1},$$

the resulting equation is

$$\begin{aligned} P = &\left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right)^T \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right) \\ &+ \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^T \boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right). \end{aligned} \tag{5.33}$$

Substituting Eq. (5.33) into the exponential

$$E = \exp\left(-\frac{1}{2}P\right) \tag{5.34}$$

and using that

$$\exp\left(A + B\right) = \exp\left(A\right) \times \exp\left(B\right),$$

the exponential term (E) of the joint distribution is

$$\begin{aligned} E = &\exp\left(-\frac{1}{2}\left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right)^T \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right)^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right)\right) \\ &\times \exp\left(-\frac{1}{2}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^T \boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)\right). \end{aligned} \tag{5.35}$$

The normalization constants

In order to normalize the constants,

$$(2\pi)^{(l_1+l_2)/2}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}, \tag{5.36}$$

the following relationship for matrix determinant was used. The determinant of $\boldsymbol{\Sigma}$ can be expressed by using the Schur complement ($\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}$) by

$$\left|\boldsymbol{\Sigma}\right| = \left|\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix}\right| = \left|\boldsymbol{\Sigma}_{22}\right| \cdot \left|\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right|. \tag{5.37}$$

Thus Eq. (5.36) can be expressed by

$$(2\pi)^{(l_1+l_2)/2}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}} = (2\pi)^{(l_1+l_2)/2}\left(\left|\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right|\left|\boldsymbol{\Sigma}_{22}\right|\right)^{\frac{1}{2}}, \tag{5.38}$$

and Eq. (5.38) can be written in the following form [de Freitas, 2011]

$$(2\pi)^{l_1/2}\left|\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}\right|^{\frac{1}{2}}(2\pi)^{l_2/2}\left|\boldsymbol{\Sigma}_{22}\right|^{\frac{1}{2}}. \tag{5.39}$$

The factorized multivariate joint Gaussian distribution

Finally the resulting multivariate joint Gaussian distribution can be expressed by combining Eq. (5.35) and Eq. (5.39)

$$\begin{aligned} p(\mathbf{x}_1, \mathbf{x}_2) = & \frac{1}{(2\pi)^{l_1/2} \left| \frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}} \right|^{\frac{1}{2}}} \times \\ & \exp\left(-\frac{1}{2} \left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right) \right)^T \left(\frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}} \right)^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right) \right) \right) \\ & \frac{1}{(2\pi)^{l_2/2} |\boldsymbol{\Sigma}_{22}|^{\frac{1}{2}}} \times \exp\left(-\frac{1}{2} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right)^T \boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right) \right). \end{aligned} \tag{5.40}$$

Equation (5.40) can be recognized as the product of the conditional distribution and the marginal $p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1|\mathbf{x}_2)p(\mathbf{x}_2)$, thus the factorization is successfully fulfilled. Both the conditional distribution and the marginal distribution are Gaussian distributed [de Freitas, 2011]

$$p(\mathbf{x}_1|\mathbf{x}_2) \sim N(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \tag{5.41}$$

$$p(\mathbf{x}_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}), \tag{5.42}$$

where

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right) \tag{5.43}$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \frac{\boldsymbol{\Sigma}}{\boldsymbol{\Sigma}_{22}}. \tag{5.44}$$

Figure 5.3 shows the factorized joint Gaussian distribution. The figure on the top is the contour plot of the same joint Gaussian distribution as in Fig. 5.2. The figure in the middle is the conditional distribution $p(\mathbf{x}_1|\mathbf{x}_2)$ and the bottom figure is the marginal distribution $p(\mathbf{x}_2)$. Figure 5.3 shows how the variance decreases for the conditional distribution compared with the marginal distribution. This decrease of the variance is due to the increasing certainty when information about the data (here $\mathbf{x}_2$) is available. The conditional MVG (multivariate Gaussian) distribution can be used for prediction under a Gaussian Process. It has been shown that for an $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$, with the mean $\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}, \end{pmatrix}$, the conditional distribution of $\mathbf{x}_1$ given $\mathbf{x}_2$ is normally distributed with the mean $\boldsymbol{\mu}_{1|2}$ and variance $\boldsymbol{\Sigma}_{1|2}$ is

$$p(\mathbf{x}_1|\mathbf{x}_2) \sim N(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}), \tag{5.45}$$

where

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{x}_2 - \boldsymbol{\mu}_2 \right) \tag{5.46}$$

and

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{5.47}$$

These results also yield for

$$p(\mathbf{x}_2|\mathbf{x}_1) \sim N(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}), \tag{5.48}$$

where [Petersen & Pedersen, 2008]

$$\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} \left(\mathbf{x}_1 - \boldsymbol{\mu}_1 \right) \tag{5.49}$$

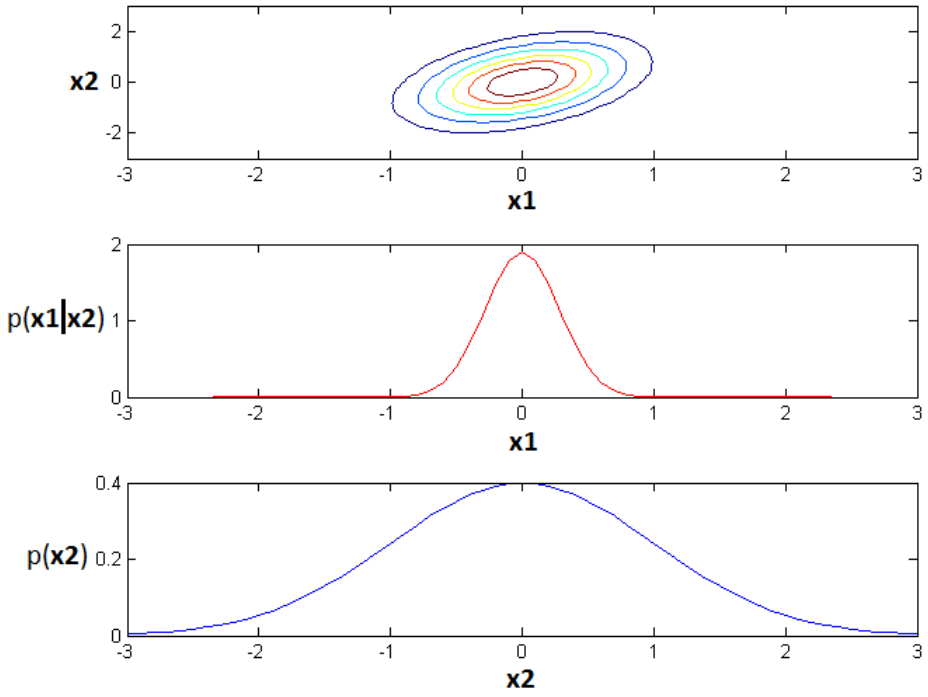


Figure 5.3: The joint Gaussian distribution, the conditional distribution and the marginal distribution respectively.

and

$$\mathbf{\Sigma}_{2|1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}. \tag{5.50}$$

Equation (5.48) can be recognized as the conditional distribution of the noise-free Gaussian Process in Eq. (5.8), Eq. (5.49) corresponds to the predictive mean in Eq. (5.10) and Eq. (5.50) corresponds to the predictive variance in Eq. (5.11).

It can be observed that the predictive mean, Eq. (5.10), and the predictive variance, Eq. (5.11), only depend on the observed data and the choice of the covariance function. Therefore the choice of the covariance function is crucial.

5.4 The covariance function

This section is based on the book by [Rasmussen and Williams, 2006]. Input values situated closely to each other tend to have similar output values. Therefore points that are near to a test point can be informative in prediction. When Gaussian Process is assumed the covariance function is the measure of this nearness. There are a great variety of covariance functions. However, there is a certain covariance function which is most frequently used in Gaussian Process regression due to its advantageous properties. This most commonly used covariance function is the squared exponential covariance function.

The squared exponential covariance function

The squared exponential covariance function can be defined by

$$K_{nn'} = \nu^2 \exp\left\{ -\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_n^d - x_{n'}^d}{\lambda^d}\right)^2 \right\}, \tag{5.51}$$

where ν is the scaling factor, λ is the characteristic length-scale, which controls the smoothness of the covariance function and d is the dimension of x. The squared exponential covariance function is stationary, isotropic, symmetric and positive semidefinite. Stationarity means that the function is invariant to translations in the input space. Thus a stationary covariance function is a function of $\mathbf{x} - \mathbf{x}'$. An isotropic function is only a function of $|\mathbf{x} - \mathbf{x}'|$, therefore it is invariant to all rigid motions. The squared exponential covariance function is symmetric, since $K_{nn'} = K_{n'n}$. The covariance function can be used in order to build a covariance matrix, where the entries of the covariance matrix are defined by the covariance function. If a real covariance matrix is positive semidefinite, then it has to satisfy that for all vectors $\mathbf{v}$, $Q(\mathbf{v}) = \mathbf{v}^T\mathbf{K}\mathbf{v} \geq 0$. A symmetric matrix is positive semidefinite if and only if all of its eigenvalues are non-negative. The squared exponential covariance function is infinitely differentiable, thus it is very smooth. Under Gaussian Process regression the most widely used covariance function is the squared exponential covariance function, due to its advantageous properties. However, the squared exponential covariance function contains free parameters (ν and λ^d). These free parameters are also referred to hyperparameters. The hyperparameters of $K_{nn'}$ have to be estimated in order to get the best regression.

5.5 Estimation of the hyperparameters

It was stated in Eq. (5.12) that the joint distribution of the noisy observations $\mathbf{y}$ and the test function values $\mathbf{f}_\star$ is

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_\star\end{bmatrix} \sim N\left(\mathbf{0}, \quad \mathbf{K}_{noisy}\right). \tag{5.52}$$

It can be observed from Eq. (5.12) that $\mathbf{y} \sim N(\mathbf{0},\ \mathbf{K}_y)$ [Rasmussen and Williams, 2006], where the covariance matrix is computed by the squared exponential covariance function, thus it includes the hyperparameters. Denote the collection of the hyperparameters to $\mathbf{\Theta} = \{\nu, \lambda, \sigma_n\}$, where σ_n is the noise level. The hyperparameters can be estimated by applying the Maximum Likelihood Estimation (MLE) method. MLE refers to the maximization of the likelihood functions. This can be done by differentiating the likelihood with respect to the hyperparameters, setting the derivatives to zero, and finally solving the resulting equation with respect to the hyperparameters [Walpole et al., 2007]. For a Gaussian process, the log marginal likelihood of $\mathbf{y}$ conditioned on the collection of the observed input vectors $\mathbf{X}$ and the hyperparameters $\mathbf{\Theta}$, can be written by

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}) = -\frac{l}{2}\log 2\pi - \frac{1}{2}\mathbf{y}^T\mathbf{K}_y\mathbf{y} - \frac{1}{2}\log|\mathbf{K}_y|. \tag{5.53}$$

Thus the hyperparameters can be computed by the MLE method to Eq. (5.53). The optimization is performed by differentiating Eq. (5.53) with respect to $\mathbf{\Theta}$. This can be written by

$$\frac{\delta}{\delta\Theta_i}\log p(\mathbf{y}|X, \mathbf{\Theta}) = \frac{1}{2}\mathbf{y}^T\mathbf{K}^{-1}\frac{\delta\mathbf{K}}{\delta\Theta_i}\mathbf{K}^{-1}\mathbf{y} - \frac{1}{2}tr\left(\mathbf{K}^{-1}\frac{\delta\mathbf{K}}{\delta\Theta_i}\right). \tag{5.54}$$

Thus the hyperparameters of the covariance function from Eq. (5.51) can be computed by solving Eq. (5.54). The importance of the hyperparameters is presented on Fig. 5.4. The parameters in Fig. 5.4 (a) are $\mathbf{\Theta} = \{\nu, \sigma_n, \lambda\} = \{1, 0.1, 1\}$, on Fig. 5.4 (b) $\mathbf{\Theta} = \{\nu, \sigma_n, \lambda\} = \{1.08, 0.00005, 0.3\}$ and on Fig. 5.4 (c) $\mathbf{\Theta} = \{\nu, \sigma_n, \lambda\} = \{1.16, 0.89, 3\}$. The + signs indicate the observed data, the blue line is the underlying function and the grey, shaded area corresponds to the confidence region. The hyperparameters on Fig. 5.4 (a) was obtained by applying Eq. (5.54). Fig. 5.4 (b) and (c) were computed by setting $\lambda = 0.3$ and $\lambda = 3$ respectively, then optimizing the rest of the hyperparameteres by using Eq. (5.54) [Rasmussen and Williams, 2006]. It can be observed that the resulting, best prediction was obtained, when all the hyperparameters were computed by optimizing the log marginal likelihood (Fig. 5.4) (a)).

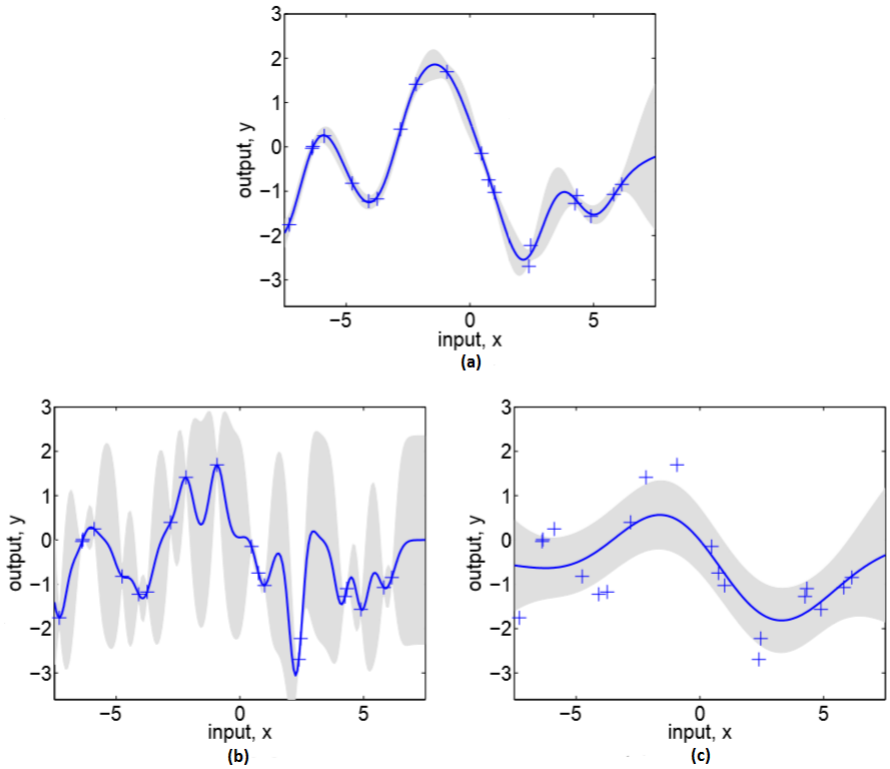


Figure 5.4: The importance of the hyperparameters (from [Rasmussen and Williams, 2006]).

Chapter 6

Results of the GPR

GPR has been proven to be a powerful machine learning method for chlorophyll content prediction [Verrelst et al., 2011], [Verrelst et al., 2012a], [Verrelst et al., 2012b], [Pasolli et al., 2010]. However, the driving mechanism of the GPR in the prediction of chlorophyll, hasn't been yet understood. Part III is going to introduce a method, the sensitivity analysis of features, which might uncover the most relevant features in the regression process. For the benefit of the reader this chapter shows the performance of the GPR when all the features are being used for chlorophyll content prediction. Thus the importance of the introduction of the sensitivity analysis (in Part III) can be understood. Furthermore, carrying out GPR with all the available features allows the visual illustration of the performance of the GPR. In order to show the flexibility of the GPR, in addition to its predictive performance, chlorophyll content estimation was performed both on a land chlorophyll dataset and on two ocean chlorophyll datasets. This chapter presents the results of the regressions. The implementations of the GPRs were carried out in Matlab. The chlorophyll datasets were already preprocessed, thus direct implementation could be performed.

6.1 Land chlorophyll

The estimation of land chlorophyll content from remotely sensed data was performed by using all the available bands for GPR. The performance of the GPR was tested on a test image. In the following section the description of the data is presented, so that the resulting chlorophyll content estimates can be better understood in the context of the underlying biophysical structures.

6.1.1 Description of the data

The training data were collected during the SPARC-2003 and SPARC-2004 campaigns, in Barrax, La Mancha in Spain [Verrelst et al., 2012b]. The output training data is the actual chlorophyll content. The chlorophyll content were measured for certain crops (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard and wheat) in Barrax. The test area has a dimension of 5 km $\times$ 10 km, and it consists of dry land (65 %) and irrigated land (35%) [Verrelst et al., 2012b]. There had been 18 Elementary Sampling Units (ESU) visited among the 9 different kind of crops. One ESU has the dimension of 20 m^2. There had been taken 50 chlorophyll measurements at each ESU [SPARC Report, 2004]. Leaf chlorophyll content was measured in order to gain knowledge about the distribution of the chlorophyll in certain plant species. The instrument which was used is the Chlorophyll Content Meter CCM-200. The measurements were performed independently [SPARC Report, 2004]. The instrument was properly calibrated and GPS coordinates were taken thus the data set could be used for model train-

ing. At the same time CHRIS images and CASI images were collected from the corresponding area. CHRIS images provided the input data for training the GPR. CHRIS was mounted on the spacecraft, called PROBA. PROBA has the advantage that it can acquire 5 images from 5 zenith angles. The CHRIS data was acquired in Mode 1. CHRIS Mode 1 corresponds to 62 bands, where the bandwidth is between 5.6 nm and 33 nm. The dimension of the bandwidth depends on the wavelength. The spatial resolution in Mode 1 is 34 m at nadir. In order to minimize angular and atmospheric effects only near-nadir images were used from the two campaigns (12.07.2003 and 15.07.2004) [Verrelst et al., 2012b]. The CASI images were collected due to another campaign, Sentinel-3 Experiment (SEN3EXP) campaign. However, during the SEN3EXP campaign CHRIS images were also acquired. These CASI and CHRIS images from the SEN3EXP campaign were used in the article by [Verrelst et al., 2011] to evaluate the portability of the already trained GPR model. This project focuses on the prediction performance of the GPR, therefore the portability of the GPR model is not described here. (For the interested readers the portability of the GPR model can be found in the article from [Verrelst et al., 2012a] and [Verrelst et al., 2011].) Figure 6.1, 6.2 and 6.3 show the position of Barrax, a satellite image of the test area and the area where chlorophyll measurements were taken, respectively.

Figure 6.1: The position of the test area (from [SPARC Report, 2004]).

It can be observed in Fig. 6.2 that vegetated areas appear as green circles. This is due to the irrigation culture of the area [SPARC Report, 2004]. Figure 6.3 shows the locations where the chlorophyll content measurements were taken and the type of crop they correspond to.

6.1.2 GPR

The dataset $D = \{\mathbf{X}, \mathbf{y}\}$, where $\mathbf{X}$ corresponds to the input of the Gaussian Process $\mathbf{X}$ and $\mathbf{y}$ corresponds to the output of the Gaussian Process $\mathbf{y}$. The input dataset, $\mathbf{X}$, has a dimension of 135x62, where 135 corresponds to the 135 locations, and 62 corresponds to the number of bands. Each of the 135 locations were measured on 62 wavelengths. The resulting spectra were collected in matrix the matrix $\mathbf{X}$. The output dataset, $\mathbf{y}$ corresponds to the measured chlorophyll contents of the 135 locations. $\mathbf{y}$ is a vector, with the dimension of 135x1. In order to test the performance of the GPR, training data and test data were required. Both $\mathbf{X}$ and $\mathbf{y}$ were divided to training and test data. Leave-one-out cross validation were used for prediction. $\mathbf{X}$ was divided to $k = 135$ vectors with a dimension of 1x62, and $\mathbf{y}$ was divided to $k = 135$ points (1x1). In order to perform LOO (Leave-one-out), one vector from $\mathbf{X}$ and the corresponding point form $\mathbf{y}$ were held out from the dataset for testing purposes. The rest of $\mathbf{X}$ and $\mathbf{y}$ were used for training. Thus the test data had a dimension of 1x62 from $\mathbf{X}$, and 1x1 from $\mathbf{y}$, while the training data had a dimension of k-1 both for $\mathbf{X}$ and $\mathbf{y}$. The training and the testing process was performed

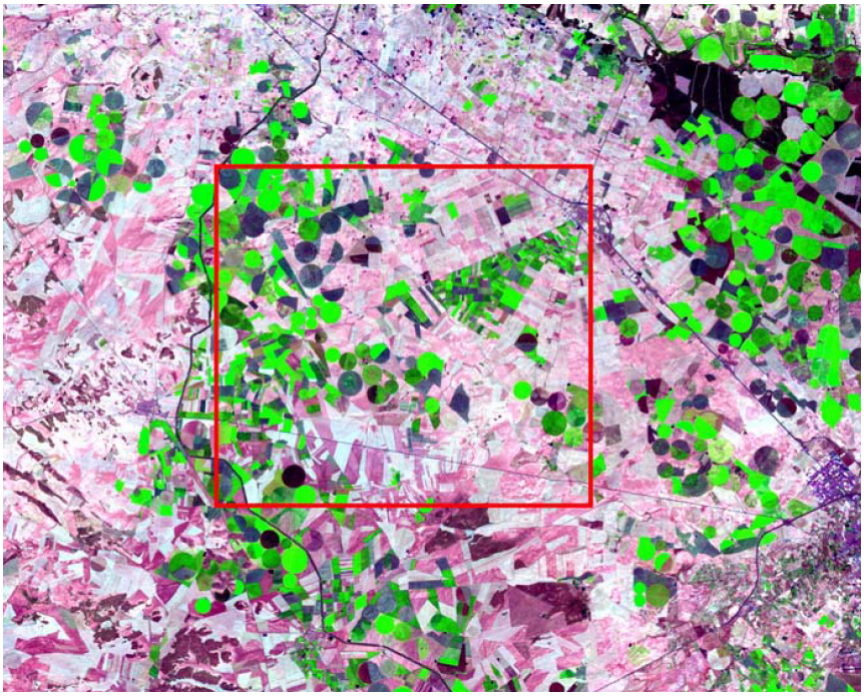


Figure 6.2: Landsat TM satellite image. The test area is indicated by the red square (from [SPARC Report, 2004]).

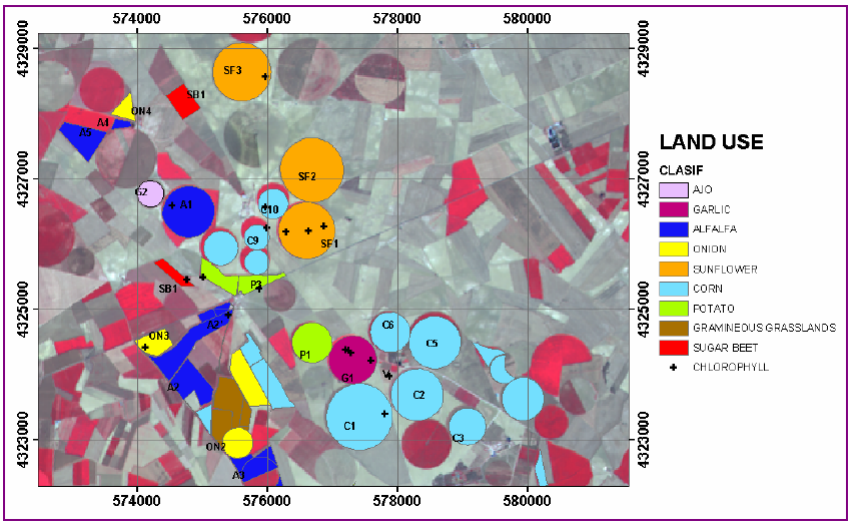


Figure 6.3: Chlorophyll content measurements (from [SPARC Report, 2004]).

for k-times [Rasmussen and Williams, 2006]. The optimization of the hyperparameters were carried out by using the Gaussian Processes for Machine Learning (GPML) toolbox, which is a freely available toolbox for Matlab users, provided by [Rasmussen and Williams, 2006]. In order to estimate the hyperparameters by the Maximum Likelihood Estimation (MLE) method with the GPML toolbox, the hyperparameters need to be initialized. The hyperparameter ν and σ_n was determined empirically, whereas the characteristic length-scales were computed by using Silverman's rule. Further details about the computations of the initial length-scales by applying the Silverman's rule can be found in the article by [Jenssen et al., 2006]. The hyperparameters were estimated only once in order to overcome the computational costs of the initializations. (All regressions in Part II and Part III with the land chlorophyll dataset were carried out with the same hyperparameters.) After the optimization of the hyperparameters was performed, GPR was applied by implementing the LOO method. The covariance matrices were computed by the squared exponential covariance function. The resulting predictive chlorophyll content and the actual chlorophyll content was plotted, so that the performance of the GPR with all the available bands is illustrated. Figure 6.4 shows the predicted chlorophyll contents and the measured chlorophyll values. It can be observed in Fig. 6.4 the predictive mean function values (predicted chlorophyll) of the GPR follow well the actual chlorophyll content values. In order to illustrate the additional advantageous property of the GPR, the predictive

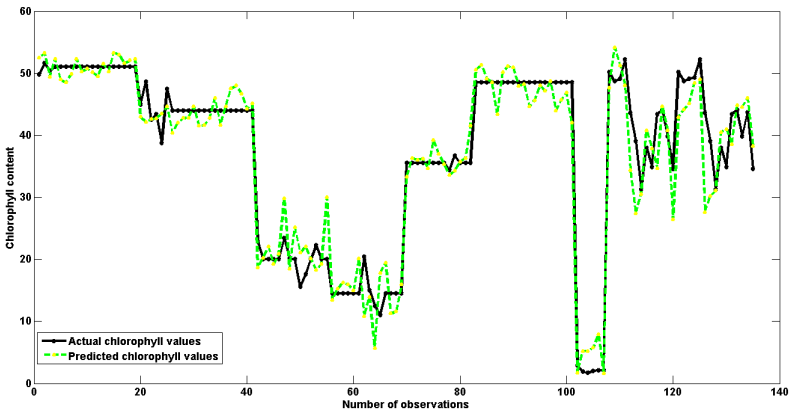


Figure 6.4: Predicted chlorophyll (62 bands).

variance, the computed predictive variances were also plotted. Figure 6.5 shows the resulting predictive variances together with the predictive means and measured chlorophyll values. The confidence region on Fig. 6.5 was computed by adding and subtracting twice the square-root of the predicted variance $\mathbf{Var}_{f_\star|y}$ to the predicted mean $\boldsymbol{\mu}_{f_\star|y}$. Thus the confidence regions can be mathematically expressed by

$$\boldsymbol{\mu}_{f_\star|y} \pm 2\sqrt{\mathbf{Var}_{f_\star|y}}, \tag{6.1}$$

where $\boldsymbol{\mu}_{f_\star|y}$ and $\mathbf{Var}_{f_\star|y}$ are vectors, collecting the predicted means and the predicted variances, respectively. To illustrate the meaning of the confidence region, the red squared part of Fig. 6.5 was enlarged. It can be shown how the confidence region increases when the test point deviates from the training points. Thus the confidence region allows the access of additional informations about the prediction. The predictive variance also reveals the certainty level of the estimated values. Fig. 6.5 suggests quite certain estimated chlorophyll contents, since the confidence region is relative narrow.

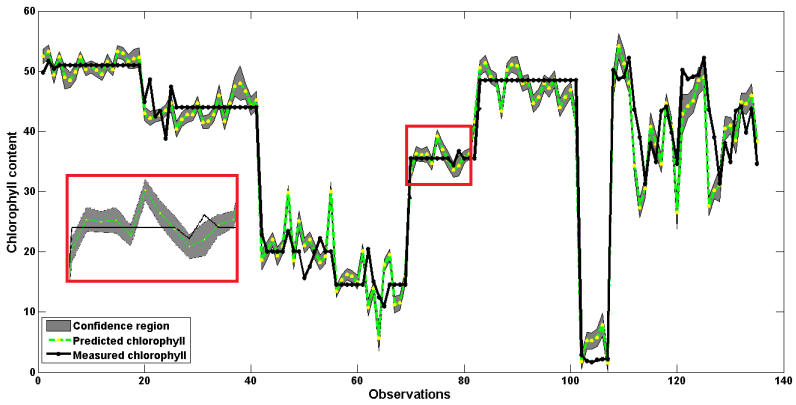


Figure 6.5: Predicted chlorophyll and corresponding confidence region (62 bands).

Evaluation of the performance of the GPR

The model performances were evaluated by computing the root-mean-squared error $RMSE$ and the coefficient of determination R^2. The $RMSE$ is informative about the accuracy of the model, and it can be written by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\mu^i_{f_\star|y} - y^i\right)^2}{N}}, \qquad \text{for} \qquad i = 1, ..., N, \tag{6.2}$$

where $\mu^i_{f_\star|y}$ is the estimated chlorophyll content, y^i is the actual (measured) chlorophyll content and N is the size of the test data set. The coefficient of determination R^2 is the measure of the goodness-of-fit, and it can be computed by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{6.3}$$

where $SS_{res} = \sum_{i=1}^{N}\left(\mu^i_{f_\star|y} - y^i\right)^2$ is the residual sum of squares and $SS_{tot} = \sum_{i=1}^{N}\left(y^i - \bar{y}\right)^2$ is the total sum of squares. The term $\bar{y}$ indicates the mean value of the measured chlorophyll contents of the test set, and it is given by $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y^i$. The $RMSE$ of the GPR with 62 bands was 3.8160 and the R^2 was 0.9293, which indicates good model performance. Computing the $RMSE$ and R^2 values allows the comparison of the GPR with all the available bands with the results of the sensitivity analysis of features which is going to be introduced in Part III.

Testing the GPR on a test image

The GPR was applied to an image, so that the chlorophyll predictive power of the model could be tested. However, ten bands had to be removed from the test image, because the provided data had this flaw. The removed bands were the following: band-1, band-2, band-3, band-21, band-22, band-23, band-24, band-25, band-26 and band-27. Therefore, the training of the GPR was performed once more, but now only with 52 bands, so that a reasonable prediction on the test image could be carried out. Removing these bands from the training set did not effected significantly the model performance, the model criteria haven't changed. Finally, the chlorophyll content prediction on the test image was performed. Figure 6.6 shows the resulting chlorophyll content image. Figure 6.6 shows that the model could discriminate between the areas with and

without chlorophyll. The irrigated parcels are clearly identified and indicated with an orange colour. The blue areas most probably corresponds to dry vegetation or bare soil. To gain deeper

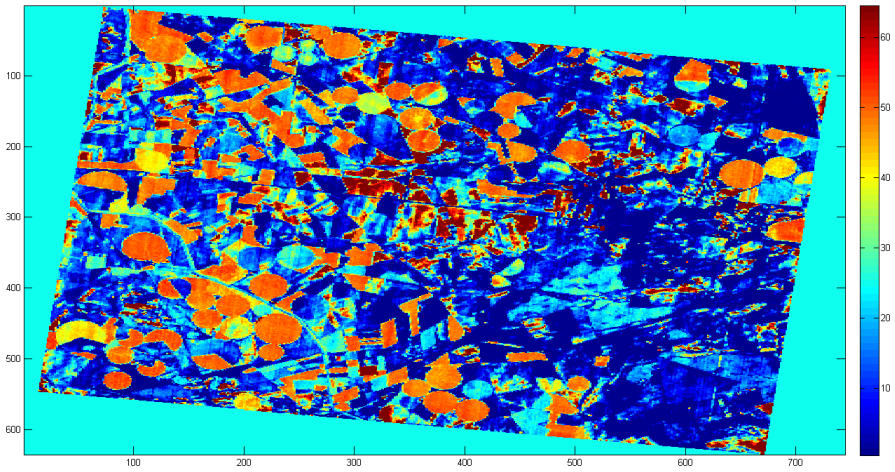


Figure 6.6: Predicted chlorophyll content map.

understanding about the strength of the prediction, the variance map of the test image was also implemented. The resulting predictive variance map can be seen in Fig. 6.7. The variance map in Fig. 6.7 shows the areas in dark blue, where the certainty level of the prediction is highest. The red areas correspond to greater uncertainty about the estimated chlorophyll content. The higher uncertainty of the predicted values might occur due to fewer or no chlorophyll content measurements from those areas.

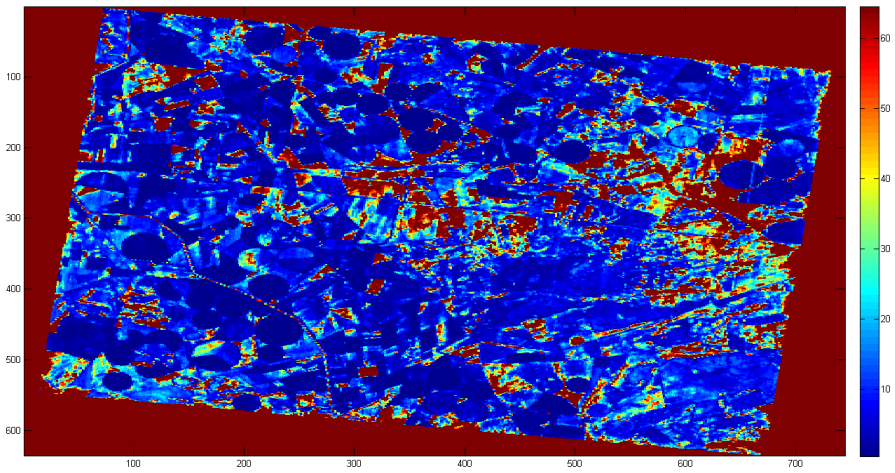


Figure 6.7: Predicted variance map.

It can be summarized that the GPR showed a good performance of estimating land chlorophyll content. The regression function in Fig. 6.4 follows the actual chlorophyll content function. The confidence region in Fig. 6.5 is quite narrow, indicating high certainty for the prediction. The model statistics also confirmed a good regression. Applying GPR for pixelwise chlorophyll

content prediction for a test image resulted the identification of the irrigated parcels, dry vegetation and bare soil (Fig. 6.6). The predicted variance map in Fig. 6.7 revealed the strength of the prediction.

6.2 Ocean chlorophyll

In order to illustrate the flexibility of the GPR, the model was tested for ocean chlorophyll estimation as well. The estimation of ocean chlorophyll content from remotely sensed data can be challenging due to the complex composition of ocean waters. Various dissolved matter might influence the reflectance spectra, which can lead to difficulties in the identification of the chlorophyll signal. Since global ocean waters occur both as Case-1 and Case-2 waters (for details see Part I Section 3.1), it is important that the method that estimates ocean chlorophyll content from remotely sensed data can adopt to both of the two conditions. GPR was applied to two ocean chlorophyll datasets: the SeaBam dataset and the MERIS dataset. The SeaBam dataset is mostly representative for Case-1 waters, whereas the MERIS dataset contains Case-2 waters as well [Camps-Valls et al., 2006].

Chlorophyll content prediction was performed by applying GPR on the datasets with all the available channels. The model performances were evaluated by computing the root-mean-squared error $RMSE$ and the coefficient of determination R^2 by using Eq. (6.2) and Eq. (6.3), respectively. The estimated chlorophyll values and the actual chlorophyll values were plotted for both of the datasets in order to illustrate the performance of the regressions. In addition, the estimated variances were also computed and plotted as confidence regions, thus the certainty of the predictions could be visualized. The confidence regions were computed by using Eq. (6.1). The following sections describe the datasets and the results of the predictions.

6.2.1 SeaBam

Description of the data

The SeaBam dataset gathers 919 ocean chlorophyll measurements around the United States and Europe [Camps-Valls et al., 2006]. The dataset consists of coincident in situ remote sensing reflectance on five channels, which corresponds to some of the SeaWiFS channels and are presented in Table 6.1, and chlorophyll-a concentration measurements. The chlorophyll-a concentrations range between 0.019-32.79 mgm^{-3}. The dataset is mostly representative for Case-1 waters [O'Reilly et al., 2000]. The chlorophyll data, the in situ chlorophyll contents and the reflectances, were converted to the logarithmic domain, based on the assumption that these bio-optical quantities follow a log-normal distribution [Pasolli et al., 2010]. The dataset was previously sorted in an increasing order of the chlorophyll concentration [Zhan et al., 2003]. GPR for estimating ocean chlorophyll content was implemented by two methods. One of the

Table 6.1: The SeaBam channels (nm).

1	2	3	4	5
402-422	433-453	480-500	500-520	545-565

methods (Method A) was to split the SeaBam dataset into a training dataset and a test dataset [Zhan et al., 2003]. The training dataset consist of 460 measurements, while the test dataset contains 459 measurements. The division of the SeaBam dataset was done by first sorting the entire dataset in an increasing order of the chlorophyll concentrations, then the odd ordered

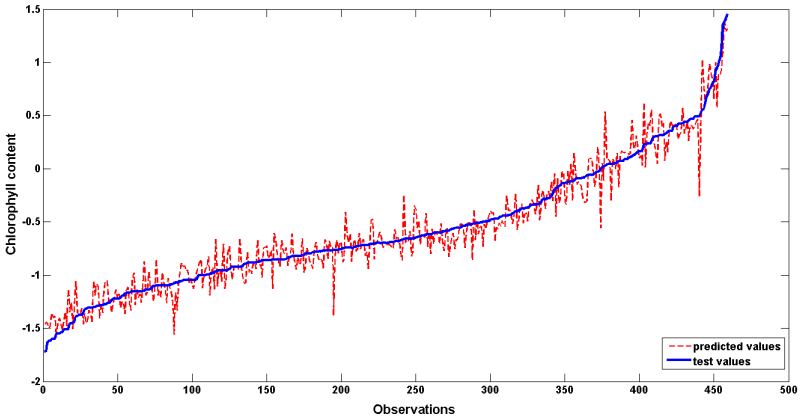


Figure 6.8: Predicted chlorophyll and measured chlorophyll.

samples were picked out to form the training set ($N_{tr} = 460$), while the rest of the samples was chosen to form the test set ($N_{test} = 459$) [Zhan et al., 2003]. The other method (Method B) was the Leave-One-out method, when only one sample was hold out for testing ($N_{test} = 1$), while the rest of the dataset ($N_{-1} = 918$) was used for training in a loop which run for 919 times. It should be noted that Method B leads to a better model performance due to a greater number of training samples. In the following the results of Method A and Method B are presented.

GPR

Method A

GPR was performed on the training dataset ($N_{tr} = 460$) with all the available bands. In order to do regression, the hyperparameters $\boldsymbol{\Theta} = \{\nu, \boldsymbol{\lambda}, \sigma_{noise}\}$ had to be estimated. The optimization of the hyperparameters was carried out by maximizing the log-marginal likelihood with the GPML toolbox. The initial hyperparameters were chosen to have the same value as in the article by [Pasolli et al., 2010]. After computing $\boldsymbol{\Theta}$, the training of the GPR was implemented on the training set. Finally, the trained GPR was tested on the test set. The results of the predictions were plotted together with the measured chlorophyll content in Fig. 6.8, so that the function learning ability of the GPR could be visualized. It can be observed in Fig. 6.8, that although the prediction function is wobbly, it follows the trend of the actual chlorophyll content function. The confidence region of the prediction was also computed by using Eq. (6.1). The results were plotted in Fig. 6.9 with the prediction function and the actual chlorophyll content function. The red square indicates the enlarged functions. This zoomed in area shows the behaviour of the confidence region. The increased confidence region indicates deviation from the training points. However, the confidence region in Fig. 6.9 is quite narrow, besides few predictions, indicating a relative high certainty of the estimated chlorophyll content values. In order to evaluate the prediction mathematically the $RMSE$ and R^2 values were computed by using Eq. (6.2) and Eq. (6.3). The $RMSE$ and R^2 resulted a value of 0.1437 and 0.9408, respectively, which indicates an excellent fit, supporting the visual results in Fig. 6.8 and in Fig. 6.9. (It should be noted, that the increasing trend on Fig. 6.8 and Fig. 6.9 is due to the fact, that the data was sorted during preprocessing.) Both Fig. 6.8 and Fig. 6.9 indicate that the GPR model follows the measured chlorophyll content.

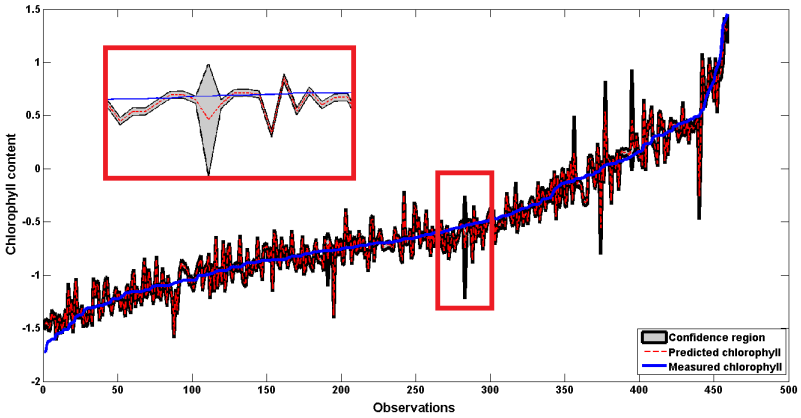


Figure 6.9: Predicted chlorophyll and the corresponding confidence region.

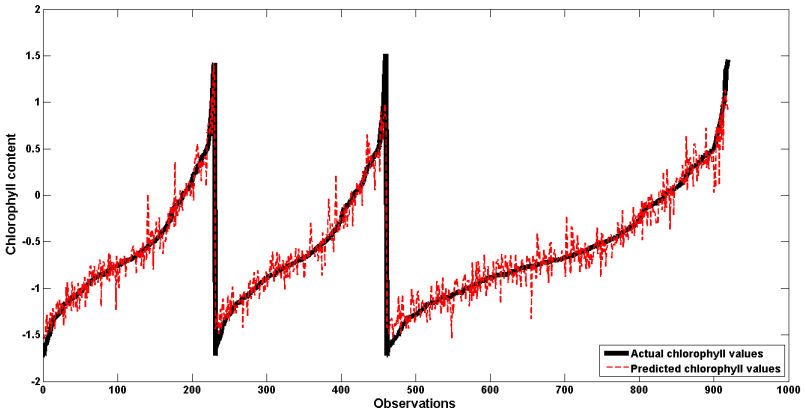


Figure 6.10: Predicted chlorophyll and measured chlorophyll.

Method B

Gaussian Process Regression was also performed by applying Leave-One-Out method in order to further improve the regression. The results of the predicted chlorophyll contents and the measured chlorophyll values are shown in Fig. 6.10. The pattern in Fig. 6.10 occurs due to the preprocessing of the data. The GP regression function could follows the trend of the measured chlorophyll content function. The confidence region was also computed and plotted in Fig. 6.11. It can be observed how the confidence region blows up when the test point differs from the training values. Applying Method B to the SeaBam dataset resulted some improvements in the $RMSE$ and R^2, namely $RMSE = 0.1414$ and $R^2 = 0.9432$, indicating that the regression has a strong predictive performance.

6.2.2 MERIS

Description of the data

The MERIS dataset is a synthetic dataset, where 5000 coincident chlorophyll-a concentrations and reflectance were simulated [Camps-Valls et al., 2009]. The chlorophyll-a concentrations range between 0.021 mgm^{-3} and 53.4429 mgm^{-3}. The remote sensing reflectance were sim-

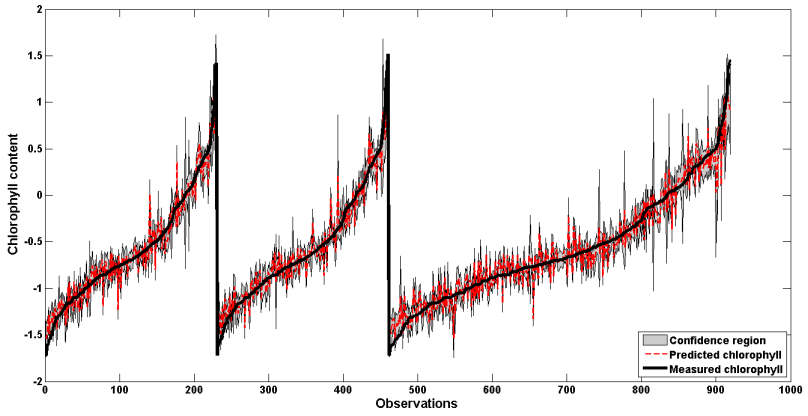


Figure 6.11: Predicted chlorophyll and the corresponding confidence region.

ulated on eight channels. The channels are presented in Table 6.2. The MERIS dataset was also converted to the logarithmic domain based on the same assumption as by the SeaBam dataset. The MERIS dataset is representative for both Case-1 and Case-2 subsurface waters. In order to perform chlorophyll content prediction, the MERIS dataset was randomly divided to a training set ($N_{tr} = 1000$) and to a test set ($N_{test} = 4000$) [Pasolli et al., 2010].

Table 6.2: The MERIS channels (nm).

1	2	3	4	5	6	7	8
407.5-417.5	437.5-447.5	485-495	505-515	555-565	615-625	660-670	677.5-685

GPR

Chlorophyll content estimation was first performed by using all the eight channels of the MERIS dataset. The first 100 estimated chlorophyll contents and the corresponding actual chlorophyll contents are presented on Fig. 6.12. The reason that only 100 values are plotted is that the dataset is large, thus plotting all the 4000 estimates would not allow the visualization of the performance of the regression. However, the shown estimated chlorophyll contents in Fig. 6.12 represents well the predictions. Figure 6.12 shows that the regression function follows almost perfectly the actual chlorophyll content function, suggesting a strong predictive performance. The confidence region was computed for the MERIS dataset as well, by using Eq. (6.1). The results for the first 100 observations were plotted in Fig. 6.13. It can be observed that the confidence region is very narrow with some few exceptions, indicating a quite high certainty level. Figure 6.13 suggests that the prediction of the MERIS dataset is reliable. Both Fig. 6.12 and Fig. 6.13 shows that using all the available channels for chlorophyll content prediction with GPR results accurate and reliable estimates. However, in order to confirm these results, the model statistics for the MERIS dataset were also computed. The $RMSE$ value was 0.0060 and the R^2 was 0.9999, indicating an almost perfect prediction.

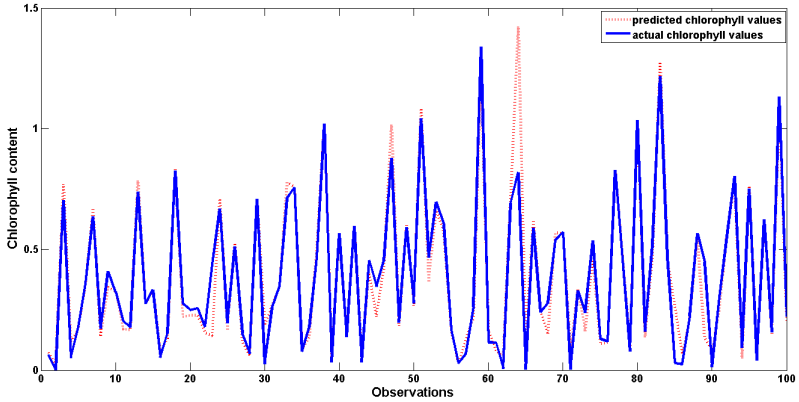


Figure 6.12: Predicted chlorophyll and actual chlorophyll.

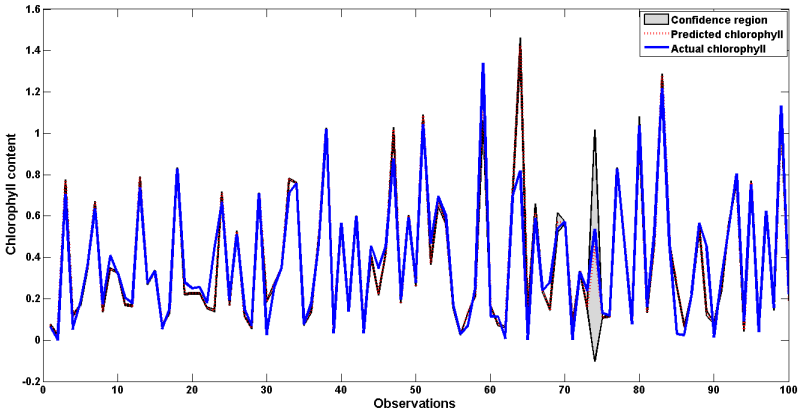


Figure 6.13: Predicted chlorophyll and the corresponding confidence region.

Conclusions

Performing GPR on the land chlorophyll dataset and on the ocean chlorophyll datasets with all the available bands showed the predictive power of this relatively new machine learning method. Figure 6.4, Fig. 6.8, Fig. 6.10 and Fig. 6.12 illustrated the predicted chlorophyll contents and the actual chlorophyll contents as functions, allowing visual comparison of the results. The additional confidence region figures, Fig. 6.5, Fig. 6.9, Fig. 6.11 and Fig. 6.13, showed how the advantageous property of the GP, the predictive variance, can be interpreted. Care should be taken in the areas, where the confidence region blows up, since these predictions indicate uncertainty. Note that although GPR performs excellently for chlorophyll content prediction, the model doesn't provide information about the relevance of the bands due to its non-linearity. The driving mechanism of the GPR hasn't been understood yet. Knowledge about the importance of the bands being used during regression, could reveal the relevance of the underlying biophysical parameters.

Part III

Sensitivity analysis of features

Gaussian Processes (GPs) have been a very influential machine learning method in the last few years. The method has been shown to have an accurate and fast predictive capability. In addition, through the predictive variance the certainty level of the prediction can also be accessed. GPs have been performed very good for the important task for chlorophyll content prediction [Verrelst et al., 2012b]. Part II showed the excellent performance of the method for chlorophyll content prediction and the advantage of having information about the predictive variance. However, it is unclear which features are the most important with respect to the spectral bands.

This thesis will for the first time develop a method for sensitivity analysis of the features for GPs. The methodology focuses on chlorophyll content prediction due to the great importance of global chlorophyll content mapping. Revealing the most relevant spectral regions of the chlorophyll absorption spectrum has been frequently studied in various scientific field, resulting some preliminary knowledge about the expected spectral band importance. The results of the sensitivity analysis of features for chlorophyll content prediction can uncover new informations and/or it can immerse the already existing knowledge.

The GPs sensitivity analysis is developed for both the predictive mean $\boldsymbol{\mu}_{f_\star|y}$ and for the predictive variance $\mathbf{Var}_{f_\star|y}$. Although this thesis applies the model for chlorophyll content prediction, the method translates to numerous fields.

Chapter 7

Sensitivity analysis

7.1 Motivation

There have been developed several parametric and machine learning models for chlorophyll content prediction. One of the most widely used parametric methods for chlorophyll content mapping from hyperspectral data is vegetation indices (VI) for land chlorophyll and spectral band ratios for ocean chlorophyll, due to the simplicity of the models [Verrelst et al., 2012b]. The description of vegetation indices and spectral band ratios can be found in Part I. Both vegetation indices and spectral band ratios use only a few spectral bands for chlorophyll content estimation, which allows them to provide a fast performance. However, there remains an uncertainty whether the most relevant combination of these bands are being used [Verrelst et al., 2012b]. In addition these models can be affected by confounding factors, thus resulting under- or overestimates of the chlorophyll content. In order to overcome the drawbacks of the parametric models, machine learning methods for chlorophyll content prediction have lately been introduced. Neural networks, support vector machines and relevance vector machines have been successfully applied for retrieving chlorophyll content from remotely sensed data. However, it has been demonstrated in the article by [Verrelst et al., 2011] that the recently introduced machine learning method, Gaussian Process Regression, has outperformed both these machine learning methods and the parametric models. Therefore the sensitivity analysis of features, which was introduced by this thesis, was developed for the GPR. Discovering the most important spectral bands being used for chlorophyll content prediction by applying GPR, would provide a powerful tool for understanding global chlorophyll content mapping.

7.2 Principles of the sensitivity analysis

Introduction

Although machine learning models have been shown to have better predictive performances than vegetation indices [Verrelst et al., 2011] and spectral band ratios [Pasolli et al., 2010], they are non-linear models, thus the importance of the features cannot directly be extracted. In contrast, the linear regression models provide easily accessible feature importance. The linear regression model can be expressed by

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b, \tag{7.1}$$

where $\hat{y}$ is the predicted output, $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input vector and b is the bias term. Equation (7.1) can be further expanded by

$$\hat{y} = w_1 x_1 + w_2 x_2 + ... + w_i x_i + b, \qquad \text{for} \qquad i = 1, ..., N. \tag{7.2}$$

Equation (7.2) suggests that the relevance of the features can be directly retrieved from the weights. In order to illustrate how the linear regression model can assign feature relevance to the inputs through the weights, consider a model with five features. Thus Eq.(7.2) can be written by

$$\hat{y} = w_1 x_1 + w_2 x_2 + ... + w_i x_i + b, \qquad \text{for} \qquad i = 1, 2, 3, 4, 5. \tag{7.3}$$

Furthermore assume, that the linear regression model resulted the following solution for Eq. (7.3)

$$\hat{y} = 0.005 x_1 + 3.2 x_2 + 2.5 x_3 + 0.000001 x_4 + 6.23 x_5 + 1.3. \tag{7.4}$$

The weights in Eq. (7.4) assign importance to the features. The example in Eq. (7.4) shows that the relevance of the features are the following in a descending order: x_5, x_2, x_3, x_1 and x_4. Thus, the higher the value of the weight, the more the feature contributes to the prediction of the output $\hat{y}$. However, when non-linear machine learning methods are used for regression, the weights cannot be assigned to the features in the same fashion as by the linear regression models. In order to access feature relevance, sensitivity maps have been introduced by [Zurada et al., 1994], [Kjems et al., 2002] and [Rasmussen et al., 2011]. Feature selection for neural networks has been discussed in the article by [Zurada et al., 1994], while the sensitivity analysis for Support Vector Machines (SVM) was first introduced by [Rasmussen et al., 2011]. However, to the best of the author's knowledge, sensitivity analysis of features for GPR (and Kernel Ridge Regression (KRR)[1]) has not yet been derived. Introducing sensitivity maps for SVM inspired the derivation of the sensitivity analysis of features for GPs. In the article by [Rasmussen et al., 2011] sensitivity analysis for neuroimaging was applied in order to produce sensitivity maps of the brain. They showed that SVM, can be used for analysing neuroimaging data. Their aim was to identify those regions in the brain, which activate for visual stimuli. This was achieved by the introduction of the sensitivity maps. These sensitivity maps reflect the importance of the so called voxels (volume units of the brain) for various visual stimuli.

The probabilistic approach

For neuroimage applications of the sensitivity map, a probabilistic approach was used for determining voxel importance [Kjems et al., 2002], [Rasmussen et al., 2011], [Strother et al., 2002]. For the first time this thesis introduces this probabilistic approach for the sensitivity analysis of features in the context of GPR for chlorophyll content prediction. The sensitivity for the jth feature/spectral band can be interpreted as the expected value of the squared derivative of the function with respect to its arguments [Rasmussen et al., 2011]. This can be expressed by

$$s_j = \int \left(\frac{\partial \phi(\mathbf{x})}{\partial x_j} \right)^2 p(\mathbf{x}) d\mathbf{x}, \tag{7.5}$$

where $\phi(\mathbf{x})$ is the function, $p(\mathbf{x})$ is the probability density function over the inputs $\mathbf{x}$ and s_j is the sensitivity of the feature j. The objective of the sensitivity map defined by Eq. (7.5) is to measure the changes of the derivative of the function $\phi(\mathbf{x})$ in the jth direction. (Here the jth direction corresponds to the jth spectral band.) The derivatives can have both positive and negative sign, which can result in both positive and negative sensitivities. This might lead to the cancellation of the terms [Rasmussen et al., 2011]. In order to avoid the possibility of cancellation of the terms, the derivatives are squared. Therefore, the resulting sensitivity map

[1]Note that both GPR and KRR have some common properties, namely, that both of the two methods can be expressed in a linear combination. Thus, the new methodology for feature selection for GPR can easily be derived for KRR. However, this thesis aimed to study the unexploited properties of the GPR in the analysis of chlorophyll datasets, therefore the derivation of sensitivity maps for KRR hasn't been carried out here.

will be positive $s_j \geq 0$ for all bands [Kjems et al., 2002]. Performing the integral in Eq. (7.5) might lead to computational complexities. Thus an empirical estimate of the integral was derived [Kjems et al., 2002]. The estimated sensitivity for the jth feature can be written by

$$\hat{s}_j = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\phi(\mathbf{x}^{(n)})}{\partial x_j} \right)^2, \tag{7.6}$$

where N denotes the number of training samples. Equation (7.6) shows that the value of $\hat{s}_j$ depends on the derivative of the function in the given direction.

In order to illustrate the interpretation of Eq. (7.6), assume an arbitrary function with two variables. Figure 7.1 represents this arbitrary function. Furthermore, assume that this function changes in direction x_1, while it is constant in direction x_2 (Fig. 7.1). Thus, summing the squared derivatives in the direction x_1 would result a positive contribution to the sensitivity map, whereas the direction x_2 would appear as an unimportant feature in the sensitivity map. The sensitivity analysis of the function presented on Fig. 7.1 would yield that direction x_1 is the most sensitive. The dimensionality of the function can be extended to multiple directions, and

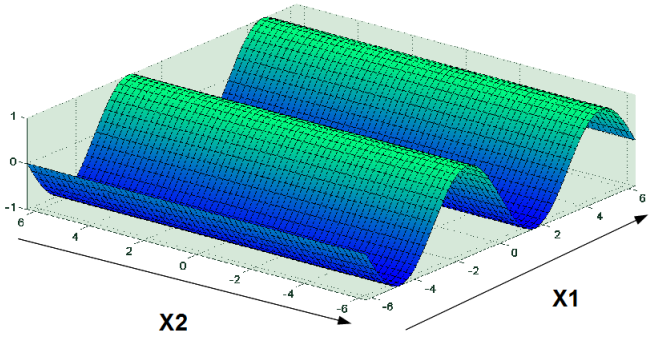


Figure 7.1: Illustrating the interpretation of the sensitivity map.

the sensitivities for each directions can be computed. The most sensitive directions would reveal the direction where the function fluctuates most. The intuitive interpretation of the sensitivity map depends on the choice of the function. For neuroimage applications, the sensitivity map was derived for the discriminant function of the SVM classifier. Hereafter the function $\phi(\mathbf{x})$ denotes either the predictive mean function of the GPR, $\phi(\mathbf{x}) = \boldsymbol{\mu}_{f_\star|y}$, or the predictive variance function of the GPR, $\phi(\mathbf{x}) = \mathbf{Var}_{f_\star|y}$. When the function is the predictive mean function $\phi(\mathbf{x}) = \boldsymbol{\mu}_{f_\star|y}$, the sensitivities can provide information about the relevance of the bands, which have been used for chlorophyll content estimation. The following section describes the derivation of the sensitivity map for the predictive mean function of the GPR.

7.3 Sensitivity of the predictive mean

Consider the case when the function is the the predictive mean $\phi(\mathbf{x}) = \boldsymbol{\mu}_{f_\star|y}$. The predictive mean can be expressed by

$$\boldsymbol{\mu}_{f_\star|y} = \mathbf{K}_\star{}^T \mathbf{K}_y^{-1} \mathbf{y}, \tag{7.7}$$

where the terms are the same as in Part II, Section 5.2. Furthermore, the predictive mean can be expressed as the linear combination of N kernel functions, where each kernel function is centred

on a training point [de Freitas, 2013]. This can be written by

$$\boldsymbol{\mu}_{f_\star|y} = \mathbf{K}_\star{}^T \mathbf{K}_y^{-1}\mathbf{y} = \sum_{p=1}^{N} \alpha_p \mathbf{K}(\mathbf{x}_p, \mathbf{x}_q), \tag{7.8}$$

where $\boldsymbol{\alpha} = \mathbf{K}_y^{-1}\mathbf{y}$ and $\mathbf{K}(\cdot,\cdot)$ is the squared exponential covariance function between the training vectors $\mathbf{x}_p$ and the test points x_q or test vectors $\mathbf{x}_q$, and it's given by $\mathbf{K}(\mathbf{x}_p, \mathbf{x}_q) = \nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2\right)$, where $d = 1, ..., D$ corresponds to the dimension (feature). Thus the sensitivity of the predictive mean can be written by

$$s_{\boldsymbol{\mu}_j} = \int \left(\frac{\partial \boldsymbol{\mu}_{f_\star|y}}{\partial x_j}\right)^2 p(\mathbf{x}) d\mathbf{x}. \tag{7.9}$$

The resulting empirical estimate of the predictive mean sensitivity is

$$\hat{s}_{\boldsymbol{\mu}_j} = \frac{1}{N}\sum_{q=1}^{N}\left(\frac{\partial \sum_{p=1}^{N} \alpha_p \mathbf{K}(\mathbf{x}_p, \mathbf{x}_q)}{\partial x_j}\right)^2. \tag{7.10}$$

The partial derivative of $\boldsymbol{\alpha}\mathbf{K}(\mathbf{x}_p, \mathbf{x}_q)$ with respect to element x_j can be written by

$$\frac{\partial \boldsymbol{\alpha}\mathbf{K}(\mathbf{x}_p, \mathbf{x}_q)}{\partial x_j} = \frac{\partial \sum_{p=1}^{N} \alpha_p \nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2\right)}{\partial x_j}. \tag{7.11}$$

Thus the derivative is given by

$$\frac{\partial \boldsymbol{\alpha}\mathbf{K}(\mathbf{x}_p, \mathbf{x}_q)}{\partial x_j} = \sum_{p=1}^{N} \alpha_p \frac{(x_{p,j} - x_{q,j})}{\lambda_j^2} \nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2\right). \tag{7.12}$$

The sensitivity of the predictive mean can be computed by combining Eq. (7.10) and Eq. (7.12)

$$\hat{s}_{\boldsymbol{\mu}_j} = \frac{1}{N}\sum_{q=1}^{N}\left(\sum_{p=1}^{N} \alpha_p \frac{(x_{p,j} - x_{q,j})}{\lambda_j^2} \nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2\right)\right)^2. \tag{7.13}$$

In order to illustrate the performance of the sensitivity analysis for the predictive mean, a controlled experiment was carried out. The illustration was motivated by the example of [Rasmussen et al., 2011]. The first part of the illustration aims to perform GPR, and compare the performance of it with linear regression. The second part computes the sensitivity map for the predictive mean of the GPR and also for the linear regression, thus allowing the comparison of the results.

7.3.1 Illustrating the concept of the sensitivity of the predictive mean

The generated dataset for the regression $D = \{\mathbf{X}, \mathbf{y}\}$ holds the input matrix $\mathbf{X}$ and the the corresponding output vector $\mathbf{y}$. The input matrix $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \mathbf{x}_4]$ contains four vectors, $\mathbf{x}_i | i = 1, 2, 3, 4$, where each vector consist of 200 observations. Each of the four dimensions were independently generated from the standard normal distribution. The output $\mathbf{y}$ is the product of $\mathbf{x}_1$ and $\mathbf{x}_2$. Furthermore, in order to produce a sensitivity map, the input $\mathbf{X}$ was embedded in a high dimensional matrix $\mathbf{I}$, with a size of 1600×200, where now each $\mathbf{x}_i$ corresponds to equally sized submatrices in $\mathbf{I}$. Figure 7.2 illustrates the setup of the matrix $\mathbf{I}$ and the four submatrices. Region 1. and 4. corresponds to $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively, while region 2. and 3. is for $\mathbf{x}_3$ and $\mathbf{x}_4$. Thus, only region 1. and 4. are informative, since $\mathbf{x}_1$ and $\mathbf{x}_2$ forms the output $\mathbf{y}$. The sensitivity map is expected to assign special importance to region 1. and 4.

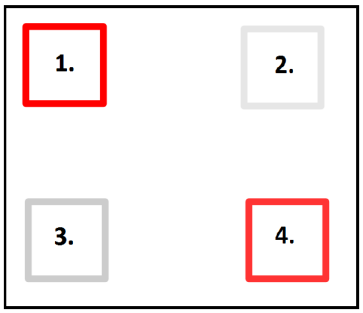


Figure 7.2: The four submatrices in matrix $\mathbf{I}$.

Comparison of the GPR with the linear regression

In order to train and test the GPR and the linear regression, the dataset D was split into two parts. 75 % of D, $\mathbf{X}_{train} = 150 \times 4$ and $\mathbf{y}_{train} = 150 \times 1$, was used for training, while 25%, $\mathbf{X}_{test} = 50 \times 4$ and $\mathbf{y}_{test} = 50 \times 1$, was used for testing. The results of the testing of the GPR and the linear regression are presented on Fig. 7.3. Figure 7.3 indicates an excellent performance of

Figure 7.3: Illustration of the performance of the GPR and the linear regression.

the GPR, while the linear regression resulted in worse prediction. In order to access numerical information of the performances of the predictions, model statistics were computed. The chosen criterion was the mean squared error, MSE, which can be written by

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2\,, \tag{7.14}$$

where N is the number of observations, $\hat{y}_i$ is the predicted value and y_i is the target value (output). The mean squared error of the GPR was $MSE = 7.9864 \cdot 10^{-11}$, while the MSE for the linear regression was $MSE = 0.7759$, which is in good correspondence with Fig. 7.3.

Sensitivity maps

The sensitivity maps for the predictive mean of the GPR and the linear regression were computed for matrix $\mathbf{I}$. The resulting sensitivity maps are shown on Fig. 7.4. Figure 7.4 indicates that the

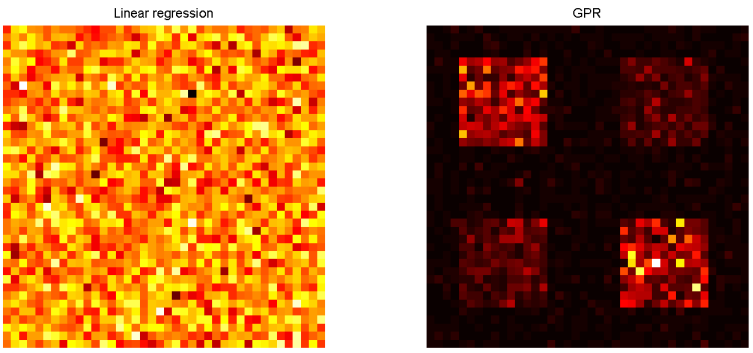


Figure 7.4: Sensitivity maps for the GPR and the linear regression.

sensitivity map of the GPR found the four embedded regions, and assigned special importance to region 1. and 4., while the sensitivity map for the linear regression could not distinguish between these regions. These results gave inspiration to perform sensitivity analysis on chlorophyll data in order to gain information about the most relevant bands for chlorophyll content prediction.

7.4 Sensitivity of the predictive variance

GPR, as opposed to other machine learning methods (NN, SVM, KRR), provides the predictive variance function in addition to the predictive mean function. Therefore, for the first time, the sensitivity analysis of the variance of the GPR is derived. The predictive variance of the Gaussian process can be expressed by

$$\mathbf{Var}_{f_\star|y} = \mathbf{K}_{\star\star} - \mathbf{K}_\star{}^T \mathbf{K}_y^{-1} \mathbf{K}_\star = \mathbf{K}_{\star\star} - \sum_{p=1}^{N}\sum_{p'=1}^{N} A_{pp'} \mathbf{K}(\mathbf{x}_p, \mathbf{x}_q)\mathbf{K}(\mathbf{x}_{p'}, \mathbf{x}_{q'}), \tag{7.15}$$

where $\mathbf{A} = \mathbf{K}_y^{-1}$ [Obádovics, 2010]. Furthermore, Eq. (7.15) can be expanded in terms of the covariances by

$$\mathbf{Var}_{f_\star|y} = \mathbf{K}_{\star\star} - \sum_{p=1}^{N}\sum_{p'=1}^{N} A_{pp'}\nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2\right)\nu^2 \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\left(\frac{x_{p'}^d - x_{q'}^d}{\lambda^d}\right)^2\right). \tag{7.16}$$

After performing the multiplication of the exponential term, the resulting equation for the predictive variance can be expressed by

$$\mathbf{Var}_{f_\star|y} = \mathbf{K}_{\star\star} - \sum_{p=1}^{N}\sum_{p'=1}^{N} A_{pp'}\nu^4 \exp\left(-\frac{1}{2}\left\{\sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d}\right)^2 + \sum_{d=1}^{D}\left(\frac{x_{p'}^d - x_{q'}^d}{\lambda^d}\right)^2\right\}\right). \tag{7.17}$$

In order to find the sensitivity of the predictive variance, the partial derivative of Eq. (7.17) with respect to element x_j needs to be computed. This can be written by

$$\frac{\partial \mathbf{Var}_{f_\star|y}}{\partial x_j} = -\sum_{p=1}^{N}\sum_{p'=1}^{N} A_{pp'}\nu^4 \left\{ \frac{(x_{p,j} - x_{q,j}) + (x_{p',j} - x_{q',j})}{\lambda_j^2} \right\} \times \exp\left(-\frac{1}{2}\left\{ \sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d} \right)^2 + \sum_{d=1}^{D}\left(\frac{x_{p'}^d - x_{q'}^d}{\lambda^d} \right)^2 \right\} \right). \tag{7.18}$$

Thus the estimate for the sensitivity of the predictive variance can be expressed by applying Eq. (7.18) and Eq. (7.6)

$$\hat{s}_{\mathbf{Var}_j} = \frac{1}{N}\sum_{q,q'=1}^{N} \left(-\sum_{p=1}^{N}\sum_{p'=1}^{N} A_{pp'}\nu^4 \left\{ \frac{(x_{p,j} - x_{q,j}) + (x_{p',j} - x_{q',j})}{\lambda_j^2} \right\} \exp\left(-\frac{1}{2}\left\{ \sum_{d=1}^{D}\left(\frac{x_p^d - x_q^d}{\lambda^d} \right)^2 + \sum_{d=1}^{D}\left(\frac{x_{p'}^d - x_{q'}^d}{\lambda^d} \right)^2 \right\} \right) \right)^2. \tag{7.19}$$

The performance of the sensitivity of the predictive variance was also tested on a controlled dataset. However, testing the sensitivity of the predictive variance would have been challenging to carry out on the same dataset as it was used in Section 7.3.1. This is due to the fact that the function of the predictive variance is independent of the output. In addition, all the four signals which were embedded in matrix $\mathbf{I}$ (in Section 7.3.1) were independently generated from standard normal distribution, therefore the variance functions should not differ significantly. Thus the sensitivity map of the predictive variance is not expected to be able to distinguish between the regions. In order to produce a sensitivity map for the predictive variance, a more illustrative dataset was chosen. The following section describes the illustration of the concept of the sensitivity of the predictive variance.

7.4.1 Illustrating the concept of the sensitivity of the predictive variance

The choice of the dataset was inspired by the article from [Jenssen et al., 2012]. The chosen dataset was extracted from the USPS Handwritten Digits dataset [2]. Figure 7.5 shows the input data, which consists of six matrices. Each matrix represents the handwritten digit 0. The dimension of the matrices is 16×16, which was converted to a 256×1 vector. Then the six vectors were stacked together in a matrix and transposed, in order to form to input data matrix $\mathbf{X}$. Thus $\mathbf{X}$ has the dimension of 6×256. The output data $\mathbf{y}$ holds the labels of the digits. Since all the digits represent the same number, digit 0, the labels assigned to them were identical, namely, 1. Thus $\mathbf{y}$ is a 6×1 vector, with the elements 1.

Sensitivity maps

In order to gain more information about the sensitivity of the predictive variance, the comparison of the sensitivity of the predictive mean with the sensitivity of the predictive variance was carried out. The sensitivity of the predictive mean and the predictive variance were computed for each

[2] The data is freely available on the website: *cs.nyu.edu/ roweis/data.html*

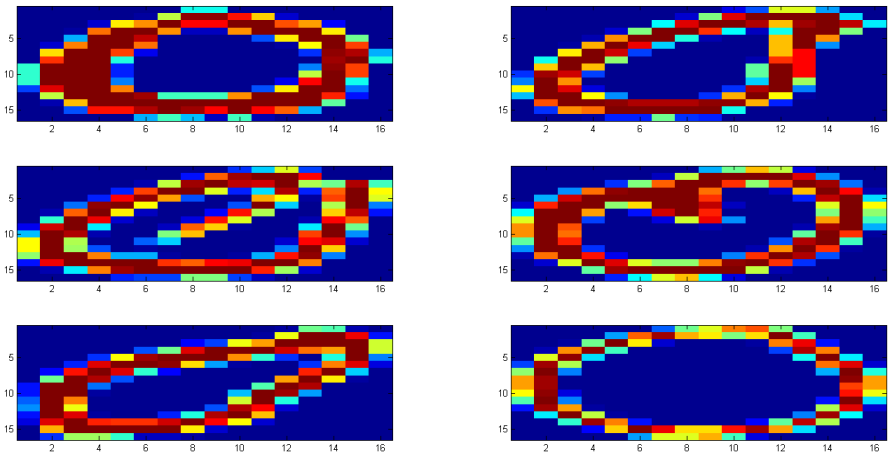


Figure 7.5: The input data.

pixels by applying Eq. (7.13) and Eq. (7.19), respectively. The resulting sensitivities were converted back to a 16×16 matrix. The resulting sensitivity maps are presented on Fig. 7.6. Figure 7.6 suggests that both the sensitivity of the predictive mean function and the sensitivity of the predictive variance function can distinguish between pixels containing information and pixels without information. The sensitivity map of the predictive variance indicates the pixels where the predictive variance function takes the greatest shifts. For better visualization, the

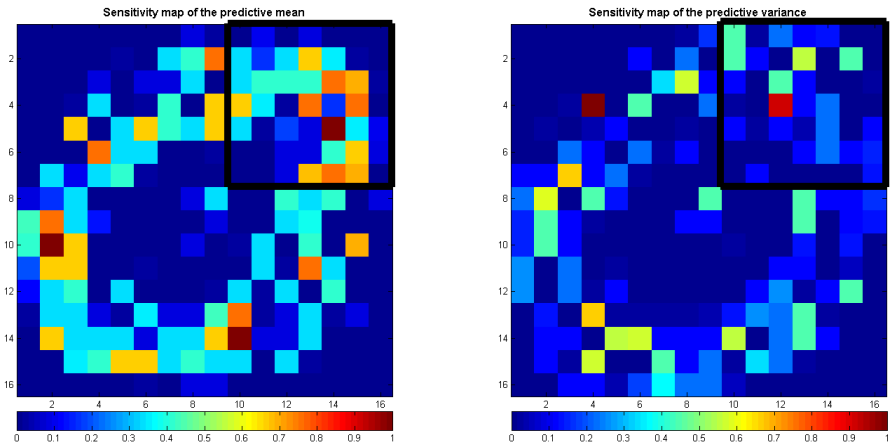


Figure 7.6: The sensitivity map of the predictive mean and the predictive variance.

black squared parts of the sensitivity maps in Fig. 7.6 were enlarged. These enlarged areas are presented in Fig. 7.7. It can be seen that the the most sensitive pixels for the predictive mean corresponds to the red and orange areas (left panel). This means that these areas are the most important in the prediction of the digit 0. The right panel shows the corresponding sensitivity map of the predictive variance. The red, yellow and green pixels indicate the areas where the variance fluctuates most, whereas the blue pixels corresponds to a more stable variance. Generally, areas with more stable variance are preferable for prediction. Thus, computing the

sensitivity map of the predictive variance, in addition to the sensitivity map of the predictive mean, allows to choose the features which contributes most to the prediction and have the most stable variance. In the case of the digit 0, Fig. 7.6 and Fig. 7.7 suggest to choose features where the pixels of the predictive mean have red/orange and yellowish colours, furthermore, from the chosen pixels, consider the ones, where the sensitivity of the predictive variance resulted blue coloured pixels. In addition to the sensitivity maps, the histogram of the sensitivities were also

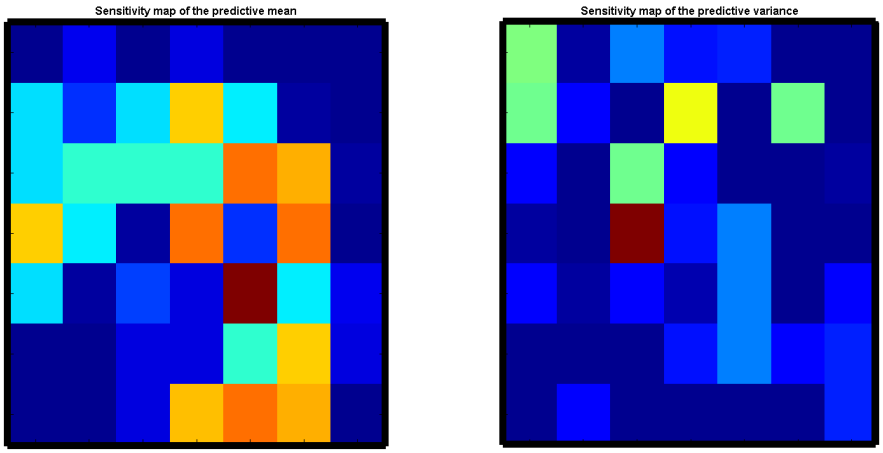


Figure 7.7: The enlarged area of the sensitivity map of the predictive mean and the predictive variance.

computed. The histogram is shown in Fig. 7.8. The histograms presented in Fig. 7.8 might help the selection of the pixels that contribute most to the prediction of the digit 0 and have the most stable variances. For example, the sensitivities, marked with the red square, indicate pixels, where the sensitivity of the predictive mean is relative high and the sensitivity of the predictive variance is quite low.

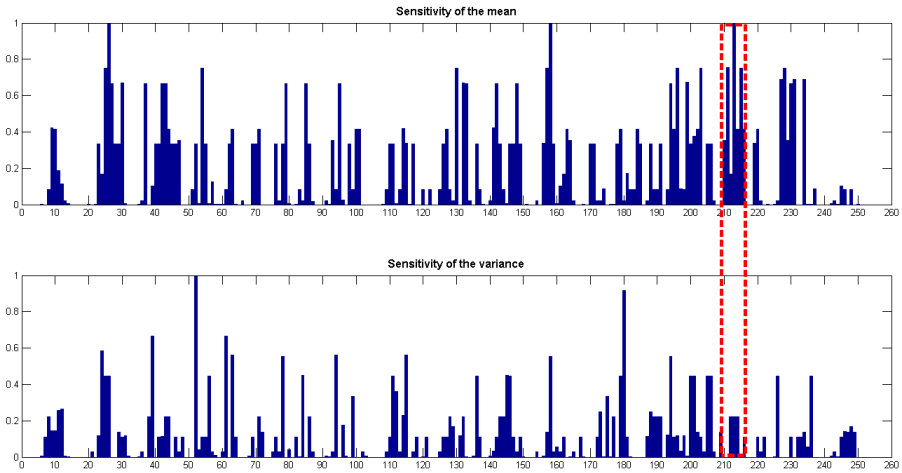


Figure 7.8: The histogram of the sensitivity of the predictive mean and the predictive variance.

Computing the sensitivity map of the predictive variance function allows to access additional

information about the features used for prediction. The controlled experiment with the digit dataset resulted promising performance of the sensitivity analysis. Therefore, the sensitivity analysis of features was carried out on the chlorophyll datasets. The following chapter presents the result of the sensitivity analysis of the chlorophyll datasets.

Chapter 8

Results of the sensitivity analysis

Sensitivity analysis of features was carried out for the land chlorophyll dataset and for the ocean chlorophyll datasets. In order to find the most sensitive bands the sensitivity of the predictive mean function was implemented. The sensitivity of the predictive mean function, defined by Eq. (7.13), was programmed as a function called "SensitivityMean", in the function environment in Matlab. (All implementations were performed in Matlab for the sensitivity analysis as well.) "SensitivityMean" needs the dataset $D = \{\mathbf{X}, \mathbf{y}\}$ and the hyperparameters of the GPR $\mathbf{\Theta} = \left\{\nu, \boldsymbol{\lambda}, \sigma^2_{noise}\right\}$ for input variables to compute the output variable, which contains the sensitivities of the predictive mean $\hat{s}_{\boldsymbol{\mu}_j}$. "SensitivityMean" was applied to the chlorophyll datasets to determine the most sensitive spectral bands. After these bands were found, the sensitivity analysis of the predictive variance function was implemented, so that bands with the most stable variance from the most relevant bands could be chosen. The sensitivity of the predictive variance was defined by Eq. (7.19). The implementation of Eq. (7.19) was also carried out in the function environment. The function that calculates the sensitivities of the predictive variance $\hat{s}_{\mathbf{Var}_j}$ is called "SensitivityVariance", which also takes the dataset $D = \{\mathbf{X}, \mathbf{y}\}$ and the hyperparameters of the GPR $\mathbf{\Theta} = \left\{\nu, \boldsymbol{\lambda}, \sigma^2_{noise}\right\}$ for input variables in order to produce the output variable $\hat{s}_{\mathbf{Var}_j}$. The function "SensitivityVariance" was also applied to the chlorophyll datasets. The resulting most sensitive bands with the most stable variances were chosen in order to carry out GPR with the selected features. Then the results of the GPR was evaluated by using model criteria so that comparison with the performance of the GPR with all the available features could be preformed. In order to carry out a sanity check, GPR with the least sensitive bands with unstable variances, was carried out. Finally, the results were compared and analysed.

8.1 Sensitivity analysis of the land chlorophyll dataset

Sensitivity analysis of the land chlorophyll dataset was carried out on the same dataset as it was described in Part II in Chapter 6. The sensitivities of the predictive mean and the predictive variance were computed by applying function "SensitivityMean" and "SensitivityVar", respectively. The input variables to the functions were the land chlorophyll dataset and the hyperparameters, which had been calculated previously in Part II in Chapter 6. In order to apply the most sensitive bands to the available test image for chlorophyll content mapping ten bands had to be removed, since these bands were ruined, and would have affected the predictions (see Part II, Section 6.1.2). Therefore, the sensitivity analysis of the spectral bands was performed twice. The first time with all the 62 bands, so that the most important features with the most stable variance could be discovered. The second time the ten ruined bands were removed, then sensitivity analysis was carried out for the 52 bands in order to be able to perform chlorophyll content mapping on the CHRIS image with the most relevant spectral bands. The

resulting sensitivities of the predictive mean and predictive variance for 62 bands and for 52 bands were plotted as histograms. Figure 8.1 shows the histograms. The black squares in Fig.

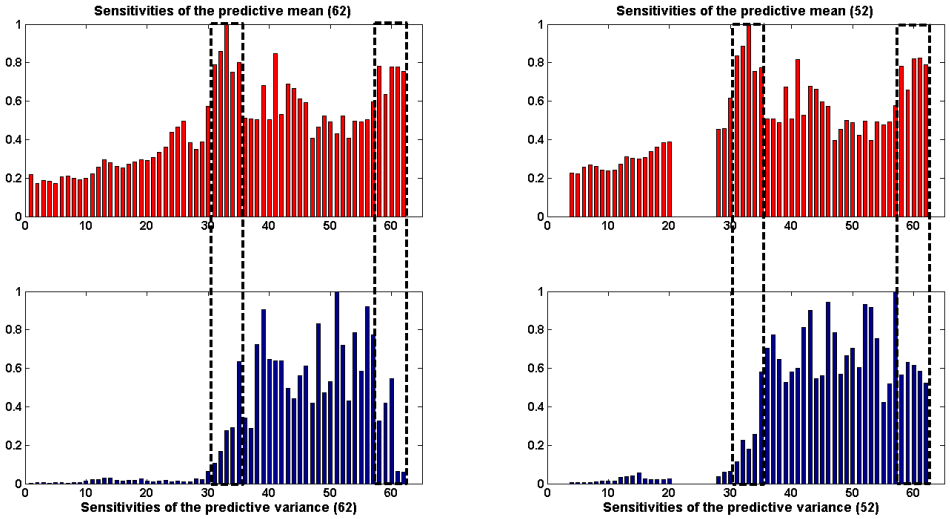


Figure 8.1: Sensitivities of the predictive mean.

8.1 show the most relevant bands with the corresponding variances. It can be observed that the sensitivity analysis of the predictive mean resulted that bands between band-30 and band-35, and bands above band-58 are the most important features for land chlorophyll content prediction from the test site in both cases [1]. When sensitivity analysis was performed with 62 bands, the most important spectral bands had relative stable variances. Interestingly, when the ten bands (corresponding to the ruined bands in the test image) were removed, the sensitivity analysis of the predictive variance assigned greater values to the highest bands (band-61 and band-62). This might be due to the fact, that removing the ten bands changed the ratio of the variances relative to each other. Figure 8.1 also shows the features which resulted in low sensitivities both for the predictive mean and for the predictive variance. These features are the bands below band-30, with specially low values for the ones below band-10.

It can be summarized that the resulting 10 most sensitive spectral bands with relative stable variances were found to be band-31, band-32, band-33, band-34, band-35, band-41, band-58, band-60, band-61 and band-62. These bands partially corresponds to the expected most relevant bands for chlorophyll content prediction. Bands between band-31 and band-35 correspond to centred wavelengths between 712.17 nm and 737.76 nm. This part of the visible part of the electromagnetic spectrum is called the red-edge. The presence of vegetation indicates a sharp change around the red-edge in the measured reflectance spectrum [Jones & Vaughan, 2010]. Thus finding bands around the red-edge to be most important for land chlorophyll prediction is in good correspondence with the expectations. In contrast, band-41 and the highest bands (band-58, band-60, band-61 and band-62) fall outside the chlorophyll absorption spectrum. However, the results of the sensitivity analysis suggest that these bands also contribute to the prediction of land chlorophyll. (For further details see Appendix B.)

In order to thoroughly analyse the results of the sensitivity analysis of features, GPR was performed with the 10 most important spectral bands. The GPR with the 10 most relevant bands were carried out on the land chlorophyll dataset by using LOO method. The results are presented in Fig. 8.2. The left panel in Fig. 8.2 shows the predicted chlorophyll content and

[1] Both for 62 bands and 52 for bands.

the measured chlorophyll content, while the right panel presents the confidence region of the predicted values. The red squares indicates enlarged areas in order to provide better illustrations of the GPR. It can be observed in the enlarged parts that the GPR could learn the relationship between the chlorophyll content and measured reflectance spectrum with the 10 most sensitive bands quite well. The confidence region is narrow, indicating a relative certain prediction. In

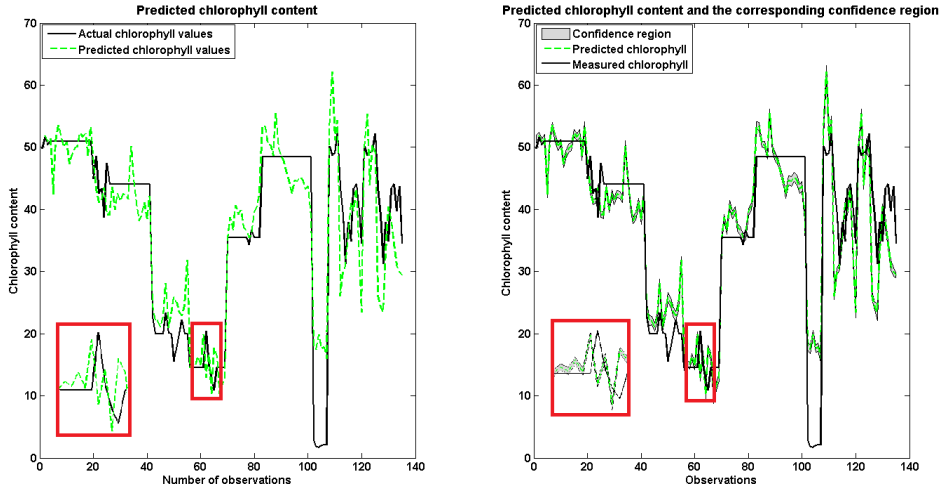


Figure 8.2: Predicted chlorophyll content and confidence region with the 10 most sensitive bands.

order to retrieve deeper understanding about the performance of the GPR with the 10 most sensitive bands, model statistics were computed. The $RMSE$ and the R^2 (Part II: Eq. (6.2) and Eq. (6.3)) was 6.0931 and 0.8197, respectively. These values still indicate a relative good regression. However, the computed model statistics for the GPR with the 10 most sensitive bands show that the regression worsened in comparison with the GPR with all the available bands. Therefore, sanity check was performed with the 10 least sensitive bands.

The resulting 10 least sensitive bands were the lowest bands, between band-1 and band-10. However, band-1, band-2 and band-3 are corrupted in the test image. In order to be able to perform a sanity check for the test image, instead of using the resulting 10 least important bands of the sensitivity analysis for the 62 bands, the 10 least relevant bands of the sensitivity analysis for 52 bands were chosen [2]. Thus the 10 least important bands used for GPR are the following: from band-4 to band-12 and band-16. GPR was carried out by using the LOO method. In order to be able to compare the results visually with the results of the GPR with the 10 most important bands (Fig. 8.2), the predicted chlorophyll contents together with the actual chlorophyll contents, and the corresponding confidence region were plotted in Fig. 8.3. The left panel in Fig. 8.3 shows the predicted chlorophyll content of the GPR by using the 10 least important bands and the actual chlorophyll content, while the right panel indicates the corresponding confidence region as well. The red squares present the same enlarged areas as in Fig. 8.2. It can be observed that the predicted chlorophyll content function deviates significantly from the actual chlorophyll content function. Comparing Fig. 8.3 with Fig. 8.2 suggests that there is a significant difference whether the 10 least or the 10 most sensitive bands are used for chlorophyll content prediction. It should be mentioned that the confidence region in Fig. 8.3 looks surprisingly still narrow, indicating high certainty level for the prediction. In order to access numerical information about the prediction with the 10 least sensitive bands, the model

[2]The sensitivity analysis for both 62 bands and 52 bands resulted the same bands to be the most important in the prediction of chlorophyll.

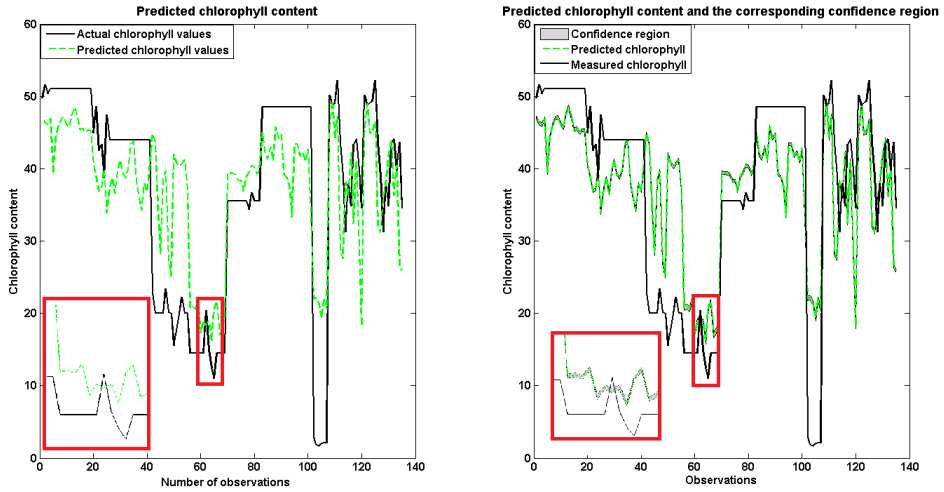


Figure 8.3: Predicted chlorophyll content and confidence region with the 10 least sensitive bands.

statistics were computed. The resulted $RMSE$ value was 9.4558 and the R^2 was 0.5657. Both the $RMSE$ and R^2 value confirmed that the regression has considerably worsened when the 10 least sensitive bands are used for chlorophyll content prediction.

The regression figures (Fig. 8.2 and Fig. 8.3) and the model statistics ($RMSE$ and R^2) show that the results of the sensitivity analysis of features for the land chlorophyll dataset are reasonable. Therefore, chlorophyll content mapping by using the 10 most important spectral bands was performed on the CHRIS image. The predicted chlorophyll content map with the 10 most sensitive bands is presented in Fig. 8.4. Figure 8.4 shows that using GPR with only the 10

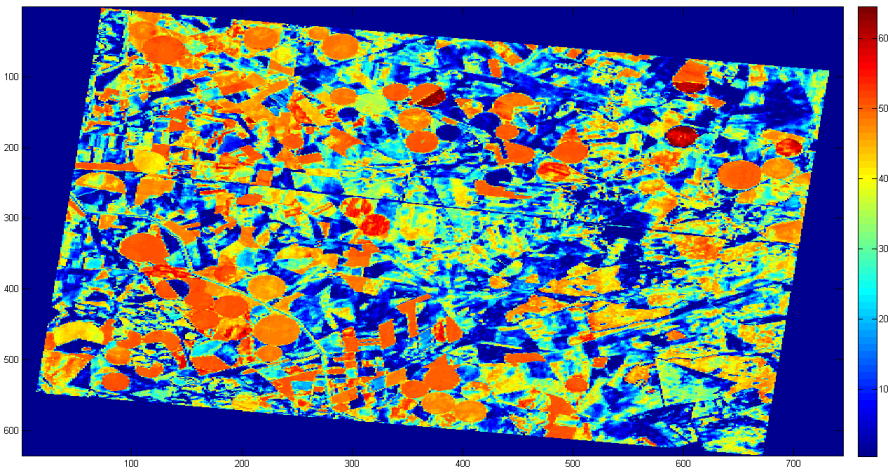


Figure 8.4: Predicted chlorophyll content map with the 10 most sensitive bands.

most important bands for chlorophyll content mapping could still identify the circular parcels. Areas where the predicted chlorophyll content is the greatest are indicated with red and orange colours, whereas region with little or no chlorophyll content are shown in blue colour. In order to illustrate the advantageous property of the GPR, the predictive variance, the certainty map of

the prediction with the 10 most relevant bands were also implemented. The resulting certainty map can be shown in Fig. 8.5. The certainty map presents the computed predictive variances for each pixel. Figure 8.5 shows that the certainty level of the prediction for the circular parcels are quite high (blue/ dark blue), while the regions in red colour indicate uncertain predictions. A possible interpretation of Fig. 8.4 and Fig. 8.5 is as it follows. Using the 10 most sensitive

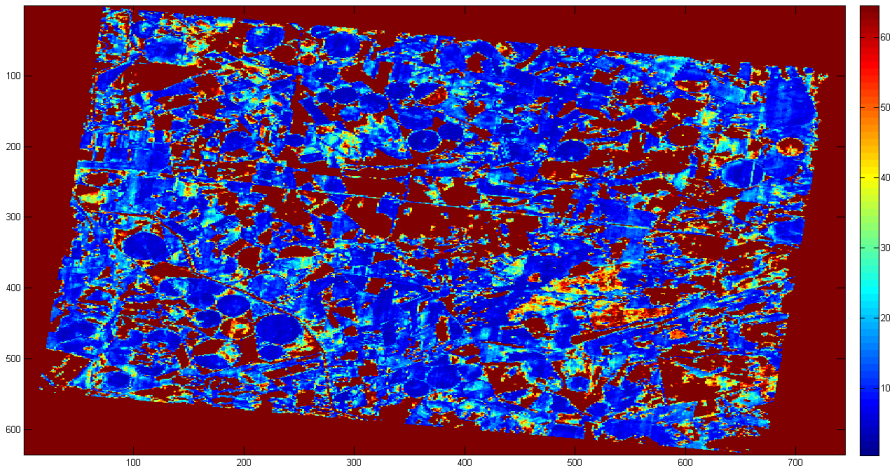


Figure 8.5: Predicted variance map with the 10 most sensitive bands.

bands for chlorophyll content mapping can find the areas which contain chlorophyll (irrigated parcels), with a relative high certainty. Areas (dry vegetation and/ or bare soil) which contain little or no chlorophyll are also identified by using the 10 most sensitive bands, however the certainty level for the estimated chlorophyll content for these region is relative low.

GPR for chlorophyll content mapping with the 10 least sensitive bands were also carried out in order to perform sanity check. Figure 8.7 shows the result of the estimated chlorophyll content for each pixel on the test image. It can be observed that the chlorophyll content assigned to the pixels is significantly less as it was when the 10 most sensitive bands were used for chlorophyll content mapping. Figure 8.6 also shows that not all the circular parcels were identified. Using the 10 least important bands for chlorophyll content prediction resulted little chlorophyll content. Figure 8.6 indicates an area which is dominated with dry vegetation or no vegetation at all. In order to gain information about the certainty level of the prediction with the 10 least sensitive bands, the predictive variances were computed for each pixel on the test image. The resulting variance map can be seen in Fig. 8.7. It can be observed that using the 10 least sensitive bands for chlorophyll content mapping resulted a remarkably increased uncertainty. Most areas in Fig. 8.7 are indicated with red colour, which corresponds to high uncertainty. In addition, those circular parcels, which were identified in Fig. 8.6, resulted decreased certainty level in Fig. 8.7. The light blue, greenish colour assigned to these regions corresponds to decreased certainty for the predicted chlorophyll content.

8.1.1 Summary

The sensitivity analysis of the land chlorophyll dataset resulted that band-31, band-32, band-33, band-34, band-35 and band-41 are the most sensitive bands. These bands correspond to wavelengths centred at 712.17 nm, 718.37 nm, 724.74 nm, 731.23 nm, 737.76 nm and 780.01 nm respectively. These results indicate that bands around the red-edge are most sensitive in chloro-

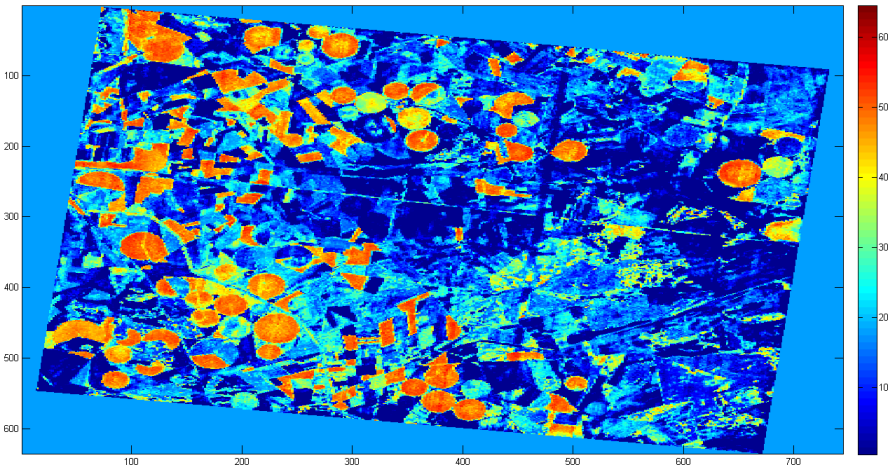


Figure 8.6: Predicted chlorophyll content map with 10 least sensitive bands.

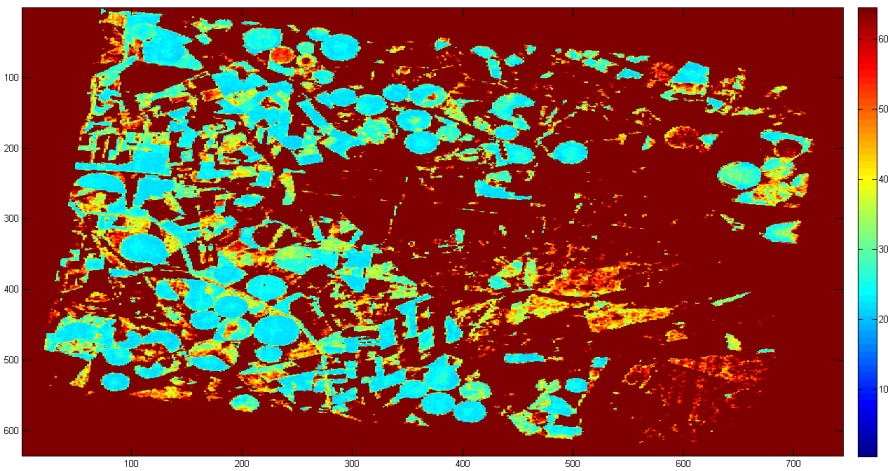


Figure 8.7: Predicted variance map with the 10 least sensitive bands.

phyll content prediction with GPR. However, bands corresponding to wavelengths 957.79 nm and wavelengths between 978.47 nm and 999.94 nm were also included in the ten most sensitive bands. This could be due to the fact that dry vegetation results different reflectance spectrum. This might indicate that bands corresponding to the dry vegetation reflectance spectra can also contribute to chlorophyll content mapping. In order to perform sanity check the least important bands were also analysed. The least important bands correspond to the lowest wavelengths of the measured spectrum, between 410.56 nm and 583.14 nm. Using the same amount of bands for carrying out GPR with the most sensitive bands and with the least sensitive bands allowed to gain information about the differences of the predictions. Both the regression figures (Fig. 8.2 and Fig. 8.3) and the computed model statistics confirmed that using the 10 most relevant bands for chlorophyll content prediction results a better model. The model statistics for the predictions are summarized in Table 8.1. The computed $RMSE$ and R^2 values in Table 8.1 shows that using the 10 least sensitive bands for GPR results worse regression.

Table 8.1: Evaluation of the land chlorophyll dataset.

Band	**RMSE**	$\mathbf{R^2}$
31, 32, 33, 34, 35, 41, 58, 60, 61, 62	6.0931	0.8197
4, 5, 6, 7, 8, 9, 10, 11, 12, 16	9.4558	0.5657

In order to reveal information about the biophysical structures that might contribute to chlorophyll content prediction, the CHRIS image which was used for testing the GPR with the 10 most sensitive and 10 least sensitive bands, was approximately localized in Fig. 6.2 (Part II). Furthermore, the locations of the chlorophyll content measurements was attempted to be identified, based on Fig. 6.3 (Part II). Unfortunately the exact coordinates of the chlorophyll content measurements were unknown, therefore the identification of the regions and crop types were based on visual judgement ability. Figure 8.8 shows the approximate location of the CHRIS image in the blue square, whereas the red square indicates the area of the test site of Barrax. Figure 8.8 also shows the possible type of crops that fell in the CHRIS image. It can be noticed that most of the measurements which are included in the CHRIS image are chlorophyll content measurements originating from potato, alfalfa, onion, garlic and corn. The absorption spectrum of the various type of crops might differ slightly, which might manifest in the measured reflectance spectrum. Using sensitivity analysis to reveal the most sensitive spectral bands could contribute to the deeper understanding of the various absorption spectra of vegetation. Figure 8.9 shows the resulting predicted chlorophyll content maps and the corresponding predicted variance maps of the GPR by using the 10 most sensitive and 10 least sensitive bands. Figure 8.9 indicates the differences between the amount of predicted chlorophyll contents. This might be due to the fact that the most sensitive bands, specially the ones around the red-edge, are characteristics for those type of crops which are in the CHRIS image, whereas using the least sensitive bands for chlorophyll content mapping in the test image by GPR, could not discriminate between the various type of vegetation. Note the blue circular parcels in the predicted chlorophyll content map by using the 10 least relevant bands. The predicted variance maps in Fig. 8.9 also indicates the differences between the predicted chlorophyll contents. The variance map produced by the 10 most sensitive bands shows high certainty of the estimated values, specially for the vegetated regions, while the variance map predicted with the 10 least bands assigns low certainty to the predicted values. It can be observed that the highest uncertainty occurs in the areas where there are little or no chlorophyll. These uncertain region are specially apparent when the 10 least sensitive bands were used for prediction. However, the sensitivity analysis resulted that bands corresponding to wavelengths above 950 nm are also important in the prediction, thus the

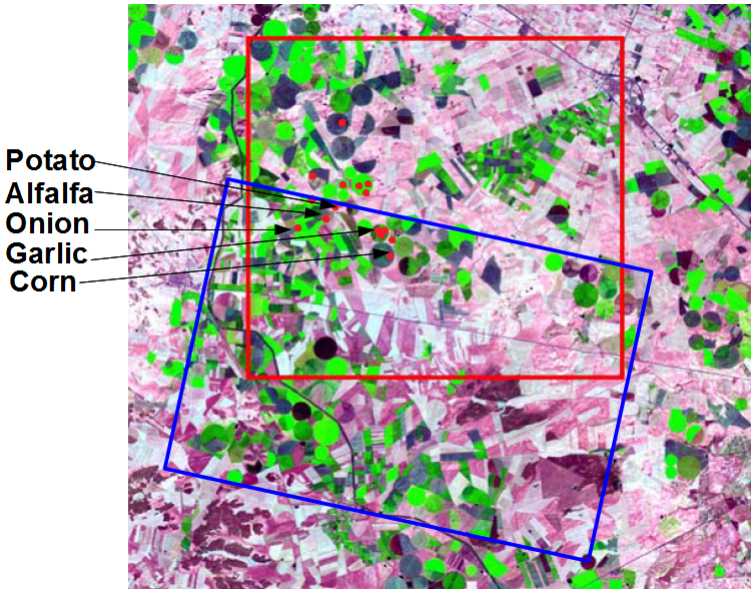


Figure 8.8: GPR with the 10 most important bands and with the 10 least important bands.

exclusion of those bands, might have resulted this great uncertainty. It can be concluded the the

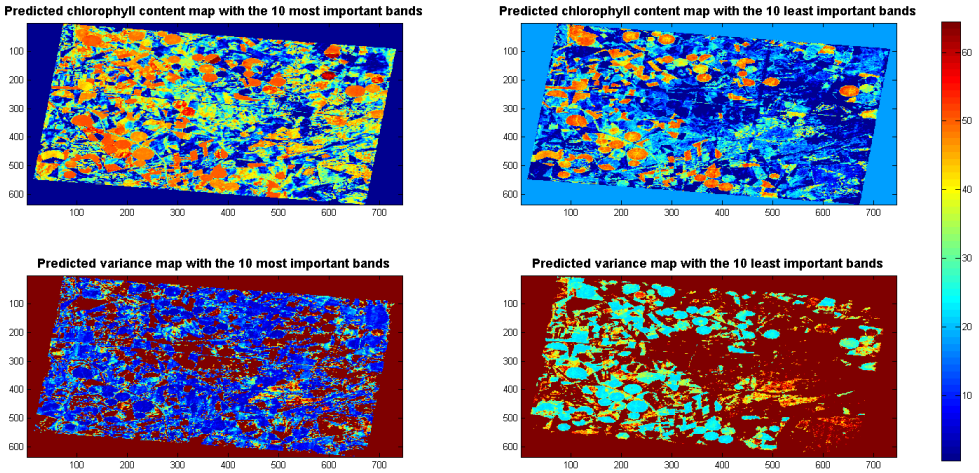


Figure 8.9: GPR with the 10 most important bands and with the 10 least important bands.

sensitivity analysis could most probably reveal the most important bands for land chlorophyll prediction for the test site. GPR with the most relevant bands resulted reasonable predictions. Thus the introduced methodology for sensitivity analysis of features can be recommended for further studying of terrestrial vegetation. The methodology might provide the possibility to uncover unknown information about the absorption/reflectance spectrum of vegetation.

8.2 Sensitivity analysis of the ocean chlorophyll datasets

Ocean chlorophyll content prediction from multispectral data can be challenging when Case-2 water conditions are present [Robinson, 2004]. One of the available ocean chlorophyll datasets

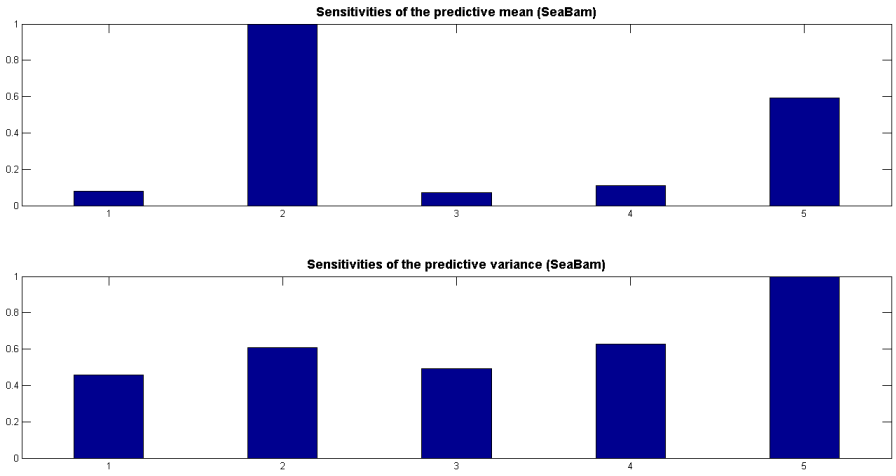


Figure 8.10: Sensitivity maps of the SeaBam dataset.

is mostly representative for Case-1 water conditions (SeaBam dataset), while the other dataset also includes measurements from Case-2 water conditions (MERIS dataset). Applying sensitivity analysis of features to these datasets might provide the possibility to be able to discriminate between Case-1 and Case-2 waters. Therefore the sensitivity analysis of features was carried out for the ocean chlorophyll datasets as well. The sensitivities of the predictive mean function and the predictive variance function were computed by using the implemented sensitivity functions, "SensitivityMean" and "SensitivityVariance", respectively. The resulting most sensitive bands were picked out in order to perform GPR. Sanity check was carried out by using the least sensitive bands for GPR. The results were analysed and interpreted.

8.2.1 The SeaBam dataset

The sensitivity analysis of the predictive mean function of the GPR for the Seabam dataset was performed by applying the function "SensitivityMean" to the dataset. In order to gain information about the variance of the most sensitive bands, the sensitivities of the predictive variances were also computed. The resulting sensitivity maps can be seen in Fig. 8.10. It can be observed in Fig. 8.10 that band-2 is the most desirable band for ocean chlorophyll content prediction when Case-1 water conditions are dominated, since band-2 has the highest sensitivity of the predicted mean and a relative low sensitivity of the predictive variance, indicating a quite stable variance. The highest sensitivities of the predictive mean are band-2 and band-5 respectively. In addition the sensitivity analysis of the predictive variance resulted that band-2 has more stable variance in comparison with band-5. The results of the sensitivity analysis suggest that band-2 and band-5, which correspond to 433-453 nm and 545-565 nm respectively, represents the most relevant bands for predicting chlorophyll content from satellite data. A possible interpretation of these results is the following. Ocean chlorophyll occurs in phytoplankton. Phytoplankton are usually represented in populations in the ocean. Dying phytoplankton cause scattering. Absorption from phytoplankton populations and backscattering occur around 540 nm [Robinson, 2004]. However, slightly below 540 nm there is a decrease in the measured reflectance, while slightly above 540 nm there is an increase in the measured reflectance. Band-5 starts at 545 nm, which might cover the increasing reflectance of the measured spectrum. In addition, increasing ocean chlorophyll content results an absorption maximum at 440 nm [Robinson, 2004], which corre-

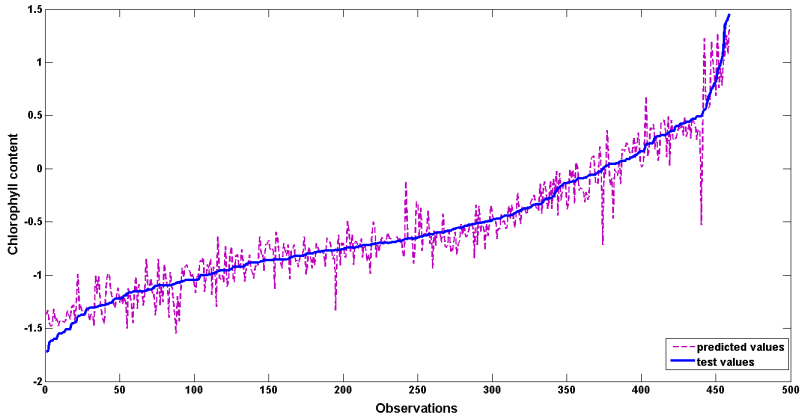


Figure 8.11: Predicted chlorophyll content by using only band-2 and band-5 (Method A).

sponds to a minimum in the reflectance. The wavelength, 440 nm falls in the range of band-2. Figure 8.24 shows the Seabam bands (and MERIS bands) and the expected changes in the reflectance due to the presence of chlorophyll.

GPR was carried out with the two most important bands (band-2 and band-5) by both Method A and Method B.

Method A

GPR was performed by using only band-2 and band-5. The procedure of the prediction was the same as it was described in Part II in Section 6.2.1. Figure 8.11 shows the predicted chlorophyll content function and the actual chlorophyll content function. It can be observed in Fig. 8.11 that even though the predicted chlorophyll content function is fluctuating, it still can learn the trend of the actual chlorophyll content function. In order to benefit from the advantageous property of the GPR, namely the predictive variance, the confidence region was also computed by applying Eq. (6.1). Figure 8.12 shows the confidence region of the prediction, the predicted chlorophyll content function and the actual chlorophyll content function. The black square indicates an enlarged area of the functions. It can be observed that the confidence region is narrow, indicating high certainty for the prediction. In order to evaluate the model performance of the GPR with the two most sensitive bands, model statistics were computed. The resulting $RMSE$ was 0.1600 and the R^2 value was 0.9266. These values indicate a good model performance. In order to carry out a reasonable comparison of the sensitivities of the bands, sanity check was performed with the two least sensitive bands (band-1 and band-3). Fig. 8.13 shows the resulting estimated chlorophyll content values and the measured chlorophyll contents. It can be noticed that the predicted chlorophyll content function shows an increased wobbly behaviour, with values that greatly deviate from the actual chlorophyll content values. The confidence region was also computed, so that the certainty level of the prediction could be revealed. Figure 8.14. shows the confidence region of the predictions. Comparing the enlarged area in Fig. 8.14 with the enlarged region in Fig. 8.12 reveals that the confidence region increased, when the two least sensitive bands were used for regression, indicating increased uncertainty of the estimated values. The computed $RMSE = 0.2651$ and $R^2 = 0.7988$ values also confirm that the model performance is worse when band-1 and band-3 are used for chlorophyll content prediction. These results show that the sensitivity analysis performed well. The methodology could successfully assign feature relevance to the SeaBam spectral bands.

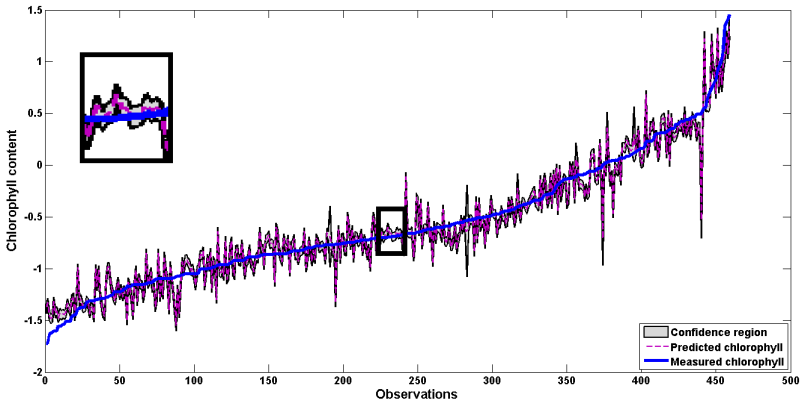


Figure 8.12: Predicted chlorophyll content and the corresponding confidence region by using only band-2 and band-5 (Method A).

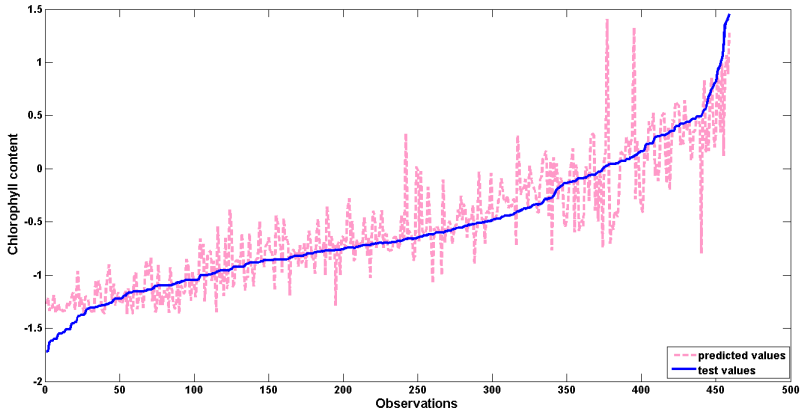


Figure 8.13: Predicted chlorophyll content by using only band-1 and band-3 (Method A).

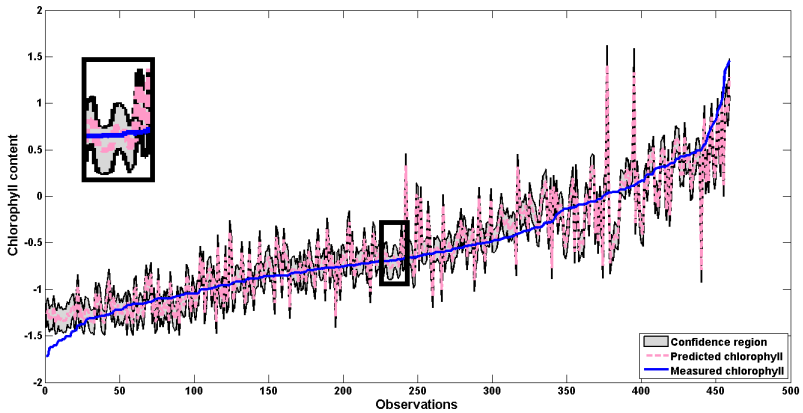


Figure 8.14: Predicted chlorophyll content and the corresponding confidence region by using only band-1 and band-3 (Method A).

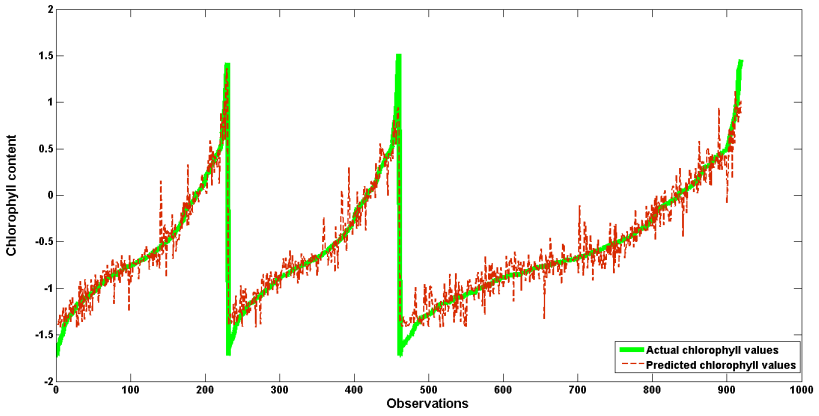


Figure 8.15: Predicted chlorophyll content by using only band-2 and band-5 (Method B).

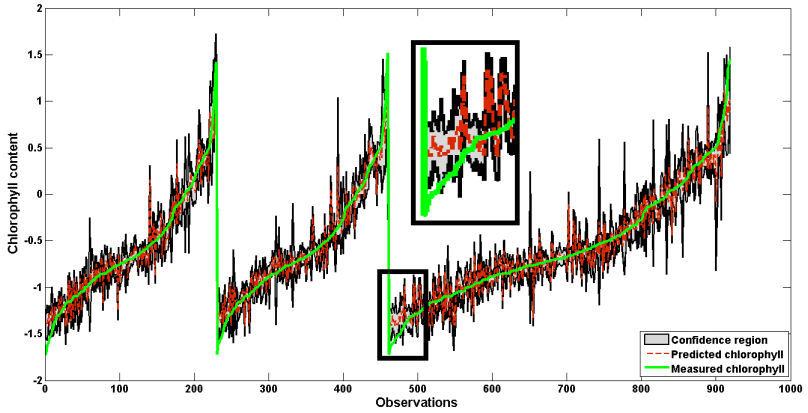


Figure 8.16: Predicted chlorophyll content and the corresponding confidence region by using only band-2 and band-5 (Method B).

Method B

GPR with the two most important bands was carried out by using Method B (Part II Section 6.2.1) as well, so that the performance of the model with the two most important bands could be evaluated on a larger dataset. The predicted chlorophyll contents and the measured chlorophyll contents are presented in Fig. 8.15. It can be observed that the predicted chlorophyll content values follow the trend of the measured chlorophyll content values similarly to Method A in Fig. 8.11. The confidence region was computed for the case of Method B as well. Figure 8.16 shows the confidence region, the predicted chlorophyll content function and the actual chlorophyll content function. The black squared area presents an enlarged part of the function, in order to provide better visibility of the confidence region. The confidence region is narrow, showing a high certainty level of the prediction. The model performance improved slightly when Method B was used. This improvement occurs due to the increased training set. The $RMSE$ decreased to 0.1526, while the R^2 increased to 0.9338. GPR with the least sensitive bands (band-1 and band-3) was also performed by using Method B. The predicted chlorophyll contents and the measured chlorophyll contents are presented in Fig. 8.17. The predicted chlorophyll content values show greater deviations from the actual chlorophyll content values. The corresponding confidence

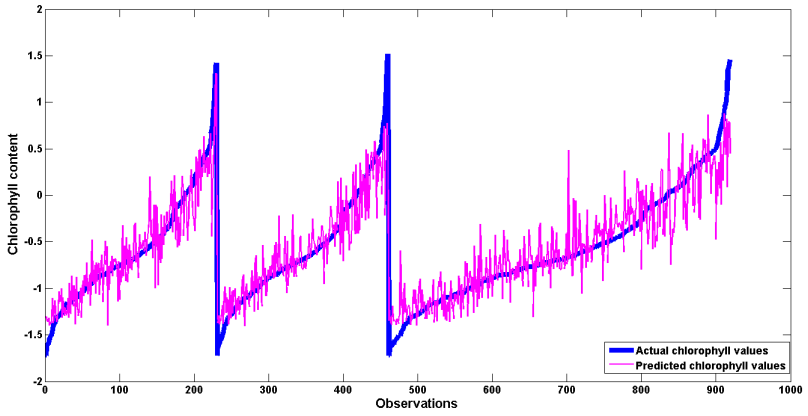


Figure 8.17: Predicted chlorophyll content by using only band-1 and band-3 (Method B).

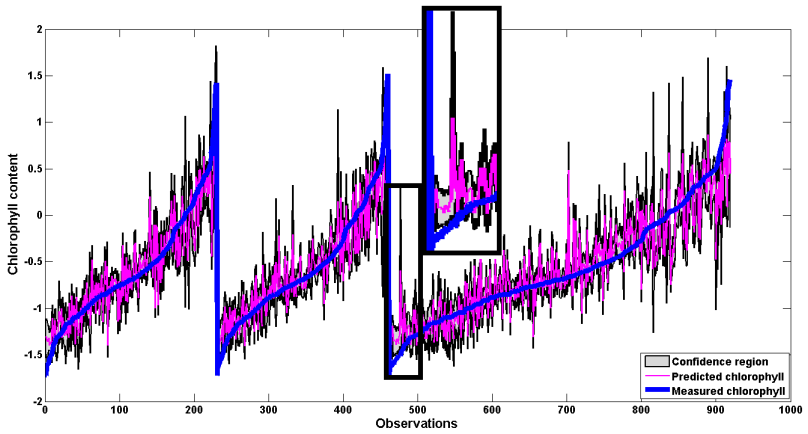


Figure 8.18: Predicted chlorophyll content and the corresponding confidence region by using only band-1 and band-3 (Method B).

region can be seen in Fig. 8.18. It can be observed that by using the least sensitive bands resulted increased confidence region. The black square in Fig. 8.18 shows that in addition to the greater uncertainty, the distance between the predictions and the measured values increased as well. The chlorophyll content function is more wobbly. The $RMSE$ value for the GPR with band-1 and band-3 was 0.2527, while the R^2 value was 0.8185, confirming that the model performance worsened, when the two least sensitive bands are used for ocean chlorophyll content prediction.

Summary

Applying the sensitivity analysis of features to the SeaBam dataset resulted that band-2 and band-5 are the most important spectral region for ocean chlorophyll content prediction, when Case-1 water condition dominates. Figure 8.10 also revealed that not only the most important bands can be discovered by using the feature selection methodology, but due to the additional property of the GPR, the predictive variance, the sensitivity of the variance can be accessed as well. The resulting most relevant spectral bands are in good correspondence with the expectations. Case-1 water condition refers to phytoplankton dominated waters. Since phytoplank-

ton contain chlorophyll, the expected most important spectral bands should correspond to the chlorophyll absorption spectrum. The resulting most important bands of the sensitivity analysis reflects this expected absorption spectrum of the ocean chlorophyll. In addition, performing GPR with these two most relevant bands, also corroborated the results of the sensitivity analysis. Table 8.2 summarizes the computed model statistics for the GPRs. It can be observed that the model performance worsened only slightly when band-2 and band-5 were used for chlorophyll content prediction in comparison to the prediction when all the five channels were used. However, using band-1 and band-3 for prediction resulted significantly decreased model performance. Thus, it can be concluded, that the sensitivity analysis has shown a reasonable performance,

Table 8.2: Evaluation of the SeaBam dataset.

Method	**Band**	**RMSE**	$\mathbf{R^2}$
A	1, 2, 3, 4 and 5	0.1437	0.9408
A	2 and 5	0.16	0.9266
A	1 and 3	0.2651	0.7988
B	1, 2, 3, 4 and 5	0.1414	0.9432
B	2 and 5	0.1526	0.9338
B	1 and 3	0.2527	0.8185

when it was applied to the SeaBam dataset.

8.2.2 The MERIS dataset

The MERIS dataset representative for both Case-1 and Case-2 waters. Finding the most important spectral regions for the MERIS dataset can contribute to the understanding of ocean chlorophyll content prediction when Case-2 waters are also present in the mapped area. The sensitivity analysis of features was performed by using the function "SensitivityMean" and "SensitivityVariance" in order to compute the sensitivity map of the predictive mean function and the predictive variance function of the GPR, respectively. The resulting sensitivity maps are presented in Fig. 8.19. The most sensitive bands for the predictive mean function are band-1, band-6, band-7 and band-8. The sensitivity map of the predictive variance shows that band-1 and band-6 have a slightly more stable variance than band-7 and band-8. However, the sensitivity map of the predictive mean assigns a relative high importance to band-8. It can be observed on Fig. 8.19 that although band-8 has the most unstable variance, the relative difference between the sensitivity of the predictive variance for band-8 and the sensitivity of the variance for band-1, band-6 and band-7 is not that significant. Therefore, in the case of the MERIS channels, it can be concluded, that the most relevant band for Case-1 and/or Case-2 water conditions is band-8. The interpretation of the results is based on the book by [Robinson, 2004]. Chlorophyll fluorescence occurs at about 685 nm, which corresponds to band-8. Chlorophyll-a molecules have an absorption maximum at 660 nm in addition to the absorption maximum at 440 nm. Band-7 ranges from 660 nm to 670 nm, which corresponds to the second absorption maximum. The sensitivity analysis resulted that also band-6 and band-1 were sensitive. This might be due to the fact, that the MERIS dataset contains Case-2 waters as well, thus other dissolved matters, such as dissolved organic matter (DOM), also contribute to the measured reflectance spectrum. Increasing DOM content leads to an increased reflectance spectrum around 620 nm, which corresponds to band 6. In addition, absorption by DOM increases from 440 nm to 400 nm. Band-1 offers the possibility for distinguishing between chlorophyll content and

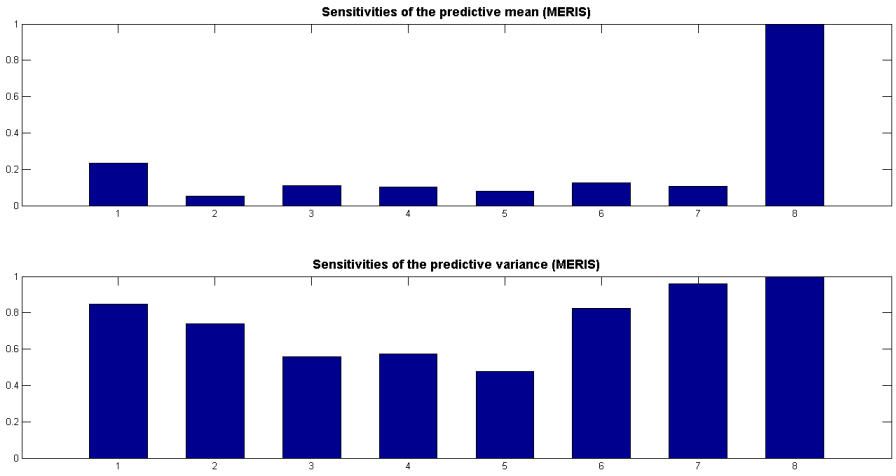


Figure 8.19: Sensitivity maps of the MERIS dataset.

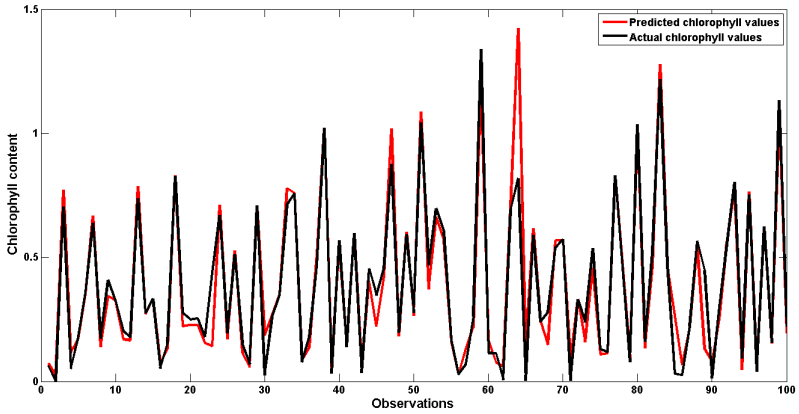


Figure 8.20: Predicted chlorophyll content and actual chlorophyll content (using the four most sensitive bands).

DOM. Thus, the computed sensitivities suggest also the Case-2 waters conditions. Figure 8.24 illustrates the MERIS bands (and SeaBam).

GPR was performed with the four most important bands in order to gain information about the performance of the regression when only four bands are used for prediction. GPR was carried out by the same method as it was done in Part II in Section 6.2.2. The resulting predicted chlorophyll contents and the actual chlorophyll contents for the first 100 observation can be seen in Fig. 8.20. The predicted chlorophyll content function shows a very good fit, even though only half of the available bands were used for prediction. The confidence region was also computed by using Eq. (6.1) (Part II). The resulting confidence region, the predicted chlorophyll content function and the actual chlorophyll content function for the first 100 observations are presented in Fig. 6.13. It can be observed that the confidence region is very narrow for most of the predictions. However, there are some exceptions, when the confidence region increases. This might be due to the fact, that the value differs from the majority of the observations. For example, observation 75 in Fig. 8.21 indicates an observation, where the confidence region shows great deviation. The confidence region for most of the observations in Fig. 8.21 shows

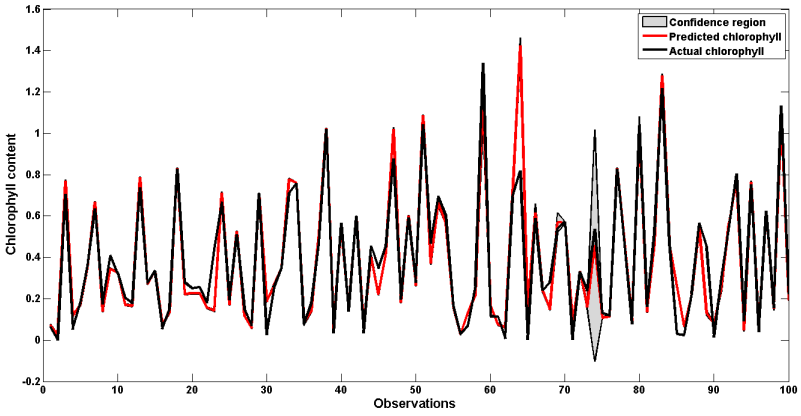


Figure 8.21: Predicted chlorophyll content and the corresponding confidence region (using the four most sensitive bands).

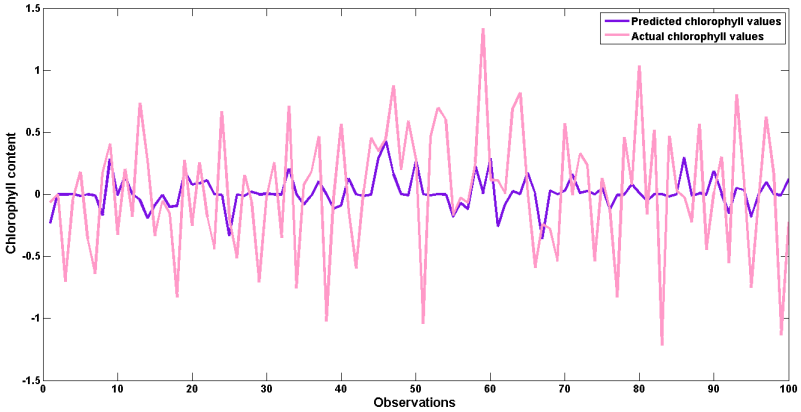


Figure 8.22: Predicted chlorophyll content and actual chlorophyll content (using the four least sensitive bands).

high certainty level of the predicted values. Presenting the entire predicted chlorophyll content function wouldn't have been informative, due to the large number of observations. Therefore, the computed model statistics reveal deeper understanding about the model performance. The resulting $RMSE$ and R^2 were 0.1352 and 0.7908, respectively. These values indicate a quite good regression.

In order to perform sanity check, GPR was carried out by using the four least sensitive bands as well. The resulting predicted chlorophyll contents and the actual chlorophyll content can be seen in Fig. 8.22. Figure 8.22 shows that using the four least important spectral bands for chlorophyll content prediction resulted worsening in the model fit. The deviation between the predicted values and the measured values are quite large. In order to gain information about the reliability of the prediction, the confidence region was computed. The resulting confidence region for the first 100 observations is presented in Fig. 6.13. It can be seen that the confidence region increased significantly. Figure 8.23 also shows how the pattern of the confidence region can be interpreted in order to access more information about the prediction. Although the confidence region increased, there are some observations when it decreases. These narrowed parts indicate, that the certainty level of the prediction is high in those predicted values. It can be observed

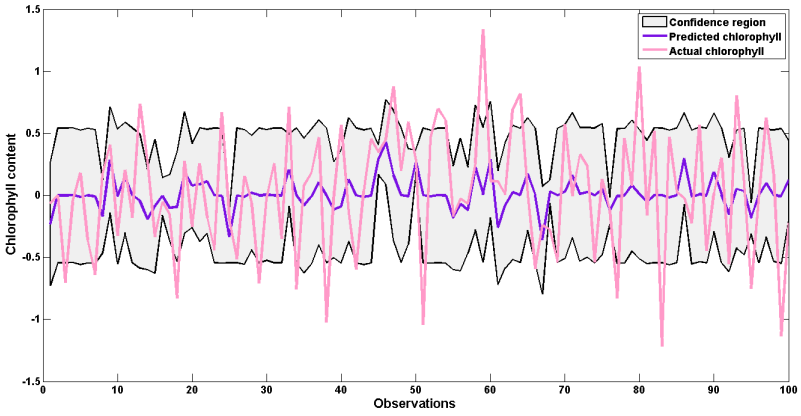


Figure 8.23: Predicted chlorophyll content and the corresponding confidence region (using the four least sensitive bands).

that the regions where the confidence region rapidly decreases, corresponds to those areas where the predicted chlorophyll content approaches to the actual chlorophyll content. Thus, using the additional property, the predictive variance, of the GPR can reveal information about the prediction.

The performance of the GPR with the four least sensitive bands was evaluated by using model criteria. The computed $RMSE$ value was 0.4582 and the R^2 value was 0.1443. These model statistics indicate a remarkable worsened prediction, particularly the resulting R^2 value, which is very low, signifying a bad model fit.

Summary

The sensitivity analysis of the predictive mean function for the MERIS dataset resulted that bands, which fall outside the chlorophyll absorption spectrum, are also relevant for prediction. Since the MERIS dataset includes measurements from Case-2 waters as well, other structures also contribute to the reflectance spectrum (such as DOM and/or sediments). Band-8 resulted the highest sensitivity relative to the rest of the bands. This indicates most probably the chlorophyll fluorescence (Fig. 8.24), thus showing the presence of the primary producers. In addition, band-6 was also found to be an important bands, which also shows chlorophyll content. However, the sensitivity analysis assigned importance to band-6 and band-1, allowing the discrimination between Case-1 water conditions and Case-2 water conditions. Figure 8.24 shows the position of band-1 and band-6. It can be observed that these bands cover the characteristic reflectance spectrum of the DOM.

The sensitivity analysis of the predictive variance hasn't revealed significant difference in the stability of the variances for the four most sensitive bands. This suggests that the variance of the four most sensitive bands do not differ significantly from each other.

Performing GPR with most and least sensitive bands also confirmed that the sensitivity analysis resulted reasonable results. Table 8.3 shows the computed model statistics. Note the difference in the R^2 value between the model with the four most important and with the four least relevant bands.

It can be concluded the using the methodology of the introduced sensitivity analysis of features for ocean chlorophyll content prediction, not only can reveal the different type of water conditions, but it can also provide information about the type of constituents of the ocean.

Table 8.3: Evaluation of the MERIS dataset.

Band	**RMSE**	$\mathbf{R}^2$
1, 2, 3, 4, 5, 6, 7 and 8	0.0060	0.9999
1, 6, 7 and 8	0.1352	0.7908
2, 3, 4 and 5	0.4582	0.1443

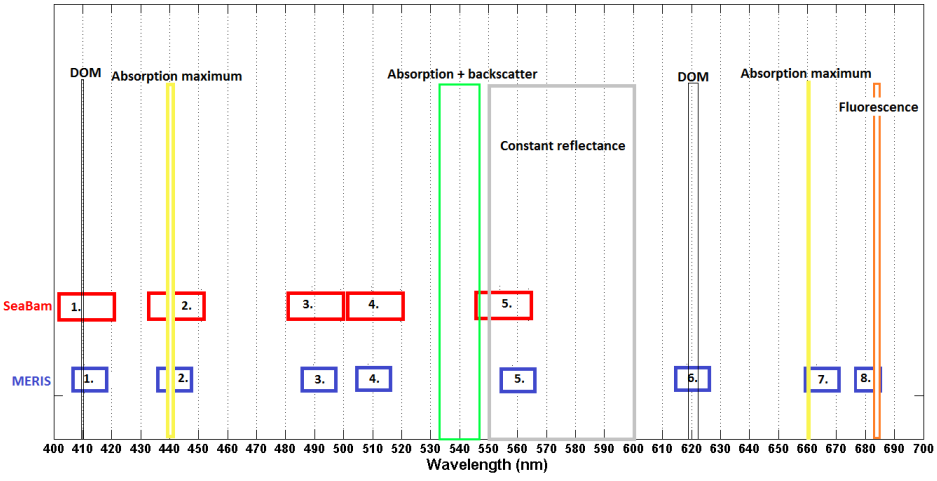


Figure 8.24: SeaBam and MERIS bands.

Part IV

Chapter 9

Conclusion

Gaussian process regression showed an excellent performance for predicting both land chlorophyll content and ocean chlorophyll content from remotely sensed multi- and hyperspectral data. The trained GPR can provide pixelwise prediction, thus allowing chlorophyll content mapping. This thesis examined the flexibility of the this nonparametric machine learning method. GPR could easily learn the relationship between spectral reflectance and measured chlorophyll content regardless whether the data originates from a terrestrial environment or a marine environment. One of the advantages of applying GPR for chlorophyll content mapping (besides its excellent predictive performance) is that the approach provides an additional information about the prediction, namely the variance. The predictive variance shows the certainty of the prediction. Thus using the predictive variance for producing certainty maps in addition to chlorophyll content maps, can retrieve the power of the predictions, allowing the identification of regions where the certainty of the predictions are high or eventually low. This can be a powerful tool, when chlorophyll content maps provide the base for other scientific fields. After examining the predictive capabilities of GPR, the main contribution of this thesis, the sensitivity analysis of features, was introduced. Empirical estimates for the sensitivity of the predictive Gaussian process mean function and variance function were derived. Then these Gaussian process sensitivity models were applied to the chlorophyll datasets in order to determine the relevance of the spectral bands, and to gain information about the variance of the most important spectral bands. The sensitivity analysis of the predictive mean for the land chlorophyll dataset showed that bands around the red-edge are the most important for chlorophyll content prediction. In addition, bands outside the chlorophyll absorption spectrum also seems to contribute to the estimation of the chlorophyll. The sensitivity analysis of the predictive variance revealed that bands centred around 720 nm and bands centred around 990 nm have the most stable variance, thus choosing these bands for chlorophyll content mapping by GPR seems to be a reasonable choice. The sensitivity analysis of features for ocean chlorophyll content prediction was performed on two datasets. One of the datasets, the SeaBam data, is mostly representative for Case-1 water conditions, whereas the other dataset, the MERIS dataset, includes both Case-1 and Case-2 water measurements. The sensitivity analysis of the SeaBam dataset resulted that band-2 (433-453 nm) is the most relevant band with the most stable variance for Case-1 water chlorophyll content mapping. This is in good correspondence with the largest absorption peak (440 nm) of the chlorophyll-a molecule. The sensitivity analysis of the MERIS dataset resulted that band-1, band-6, band-7 and band-8 are the most relevant bands for predicting ocean chlorophyll content, when Case-2 water conditions are also present. The sensitivity analysis of the variance of these bands hasn't revealed a remarkable difference of the stability of the variance. This suggests that all these four bands contribute significantly to the prediction of ocean chlorophyll, when both Case-1 and Case-2 water conditions are present.

It can be concluded that the sensitivity analysis of features for the GP can reveal relevance of the spectral bands and also the stability of the variance of these channels. Using these most relevant and stable bands as an input for the GPR results a fast, easy and accurate prediction of chlorophyll content.

Chapter 10

Future work

Sensitivity analysis can provide a powerful tool for feature selection when the dataset has high dimension. Therefore further study of the method on hyperdimensional data is recommended. The GPR and the sensitivity analysis of features were carried out on chlorophyll data. However, the methods are not limited to chlorophyll content prediction, thus allowing applications for various scientific fields. In order to illustrate the translational potential of the introduced methodology, some examples for the possible applications of the sensitivity analysis of features are presented in Appendix A.

For future applications, a web side is going to be created in order to provide easy accessibility of the method. It is planned that the implemented sensitivity functions and also GPR functions going to be available for Matlab users, where the methods would only require a dataset for input. The web side would help the further development of the methodology in addition to the application purposes.

Appendix A

Examples for applications of the sensitivity analysis

Out of the scope of the thesis some examples for the possible applications of the sensitivity analysis of features are presented. These examples show that the sensitivity analysis of features, introduced in this thesis, has the untapped potential to reveal the importance of features, in other fields as well.

A.1 Sensitivity analysis of the Madelon dataset

The Madelon dataset is an artificial dataset[1]. The dataset has 500 features, where only 20 features are informative. The training dataset consists of 2000 observations. In order to illustrate the sensitivity analysis of features, the first 100 observations and the corresponding labels were chosen. Then the functions "SensitivityMean" and "SensitivityVariance" were applied to the chosen training data. The resulting sensitivity maps can be seen in Fig. A.1. Figure A.1 shows the computed sensitivities of the predictive mean and the predictive variance. It can be observed that both the sensitivity analysis of the predictive mean and the sensitivity analysis of the predictive variance could discriminate between the features. Unfortunately, there were no information provided about the position of the 20 informative features. However, the resulting 20 most sensitive features of the sensitivity analysis were located. Figure A.2 presents the sensitivity map of the predictive mean for all the features, the located 20 most important features and the corresponding sensitivities of the predictive variance. Figure A.2 shows that the 20 most important features are spread evenly. The corresponding sensitivities of the predictive variances are quite low, suggesting stable variances, with no outstanding values. Applying the sensitivity analysis of features to the Madelon dataset showed that the model can assign feature relevance to other datasets besides the chlorophyll data.

A.2 Sensitivity analysis of the digit dataset

In Section 7.4.1 the sensitivity analysis of features was carried out on a digit dataset. The focus in Section 7.4.1 was to illustrate the performance of the sensitivity of the predictive variance. Furthermore, the digit dataset can also be used for identifying the most important pixels for discriminating digits from their surroundings. Therefore, the sensitivity analysis was tested on the same dataset as it was described in Section 7.4.1. The training data consists of an input matrix $\mathbf{X}$ and the corresponding output vector $\mathbf{y}$. The size of $\mathbf{X}$ is 6×256, where 6 corresponds to

[1] The dataset can be found on the web side: http://archive.ics.uci.edu/ml/datasets/Madelon.

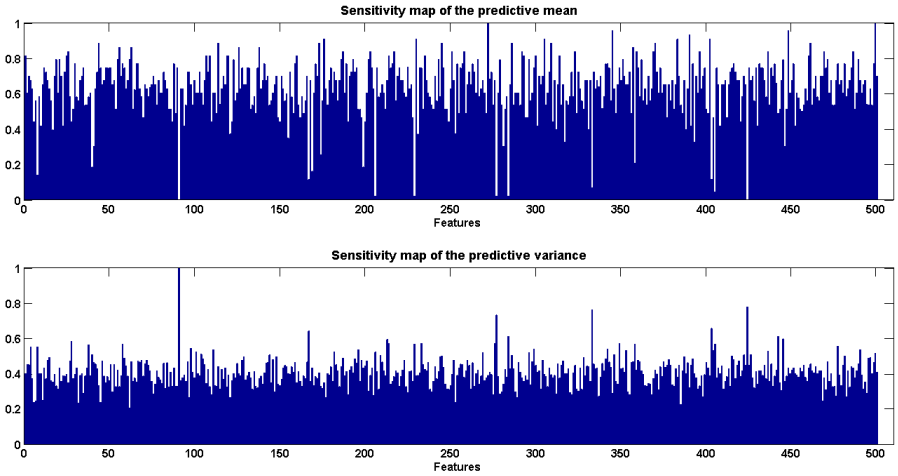


Figure A.1: Sensitivity analysis of the Madelon dataset.

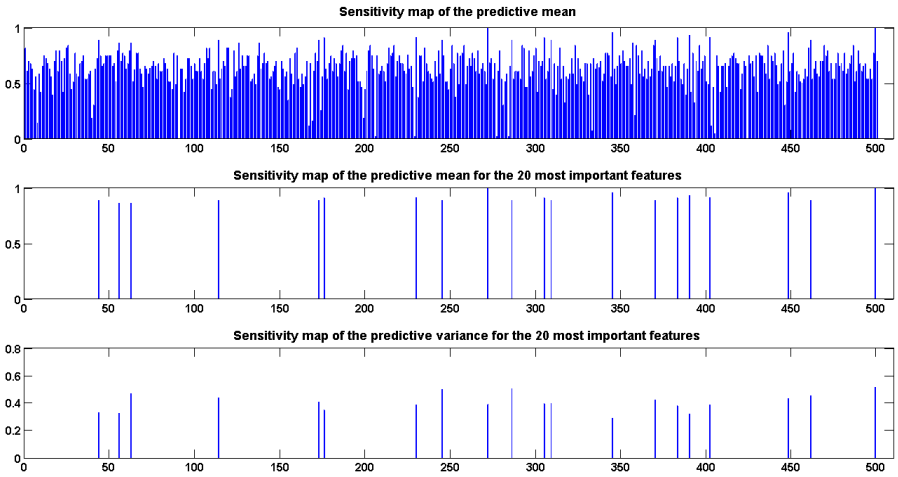


Figure A.2: The position of the 20 most sensitive features and corresponding sensitivities of the variances of the Madelon dataset.

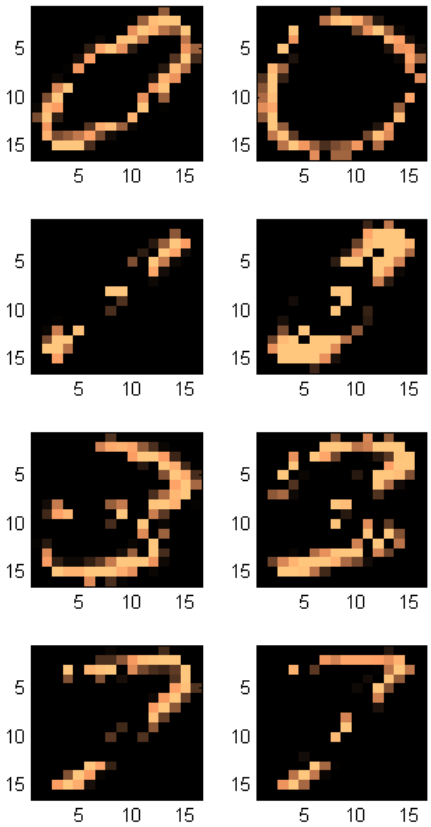


Figure A.3: Sensitivity analysis for digit identification.

the observations and 256 is the dimension of each observation. All the 6 observations correspond to the handwritten digit 0. The output vector $\mathbf{y} = [1\ 1\ 1\ 1\ 1\ 1]^T$ holds the labels. The function "SensitivityMean" (and the additional "SensitivityVariance") was applied to the training data (Section 7.4.1). The resulting sensitivities of the pixels were ranked in a descending order. Thereafter, the position and the value of the first 128 most sensitive pixels were picked out. In order to test whether these most sensitive pixels can be used for identifying digits, a test set was chosen. The test set contains eight digits. The chosen digits were: 0, 0, 1, 1, 3, 3, 7 and 7. Then the pixels of each test digit was set to the value zero. Finally, the resulting most sensitive 128 pixels were located in each test digit, and the corresponding pixels were set to the value of the resulting sensitivities in order to test whether the 128 most sensitive pixels can be used for identifying digits. Figure A.3 shows the results. It can be observed that using the most sensitive pixels to the test set, could identify the test digits. All numbers are clearly distinguishable. Note that although the pixel sensitivities were computed on a training set, which consists of only the digit 0, the resulting 128 most sensitive pixels could be used for distinguishing other digits (1, 3 and 7 in addition to 0) from their environment. Applying the sensitivity analysis of features to the digit dataset for finding the most important pixels for digit distinguishing, showed a good performance. This example also suggests that the introduced methodology for feature selection translates to other fields.

Appendix B

Determining feature relevance by using λ_j

The sensitivity analysis of features for the land chlorophyll dataset showed that bands outside the chlorophyll absorption spectrum also contribute to the prediction of chlorophyll (Section 8.1). Finding spectral bands outside the chlorophyll absorption spectrum to be important in the prediction of chlorophyll content has been done by [Verrelst et al., 2012b]. Although sensitivity analysis of features in the context of GPR for land chlorophyll content prediction by using the methodology introduced in Part III in Chapter 7 was first carried out in this thesis, the identification of the most important spectral bands in the same context have been studied by [Verrelst et al., 2012b]. Their approach for finding the most relevant bands by using GPR for the same chlorophyll dataset was based on assigning feature relevance to the characteristic length-scales hyperparameters λ_j.

It has been pointed out in the book by [Rasmussen and Williams, 2006] that the relative value of λ_j can be interpreted as feature relevance. The greater the value of λ_j, the lower the relevance of the feature. Inversely, relative low values of the characteristic length-scales, indicate important features. Using this approach for determining the importance of the spectral bands for the land chlorophyll dataset, [Verrelst et al., 2012b] found that the four most important bands were band-18, band-25, band-57 and band-60. Their results partially overlaps with the resulting most important spectral bands using the methodology introduced by this thesis. They also found that bands (band-57 and band-60) outside the chlorophyll absorption spectrum contributes to the prediction of the land chlorophyll content. However, the other two relevant bands, band-18 and band-25, are in contrast of the resulting most relevant bands (bands between band-31 and band-35) by using the sensitivity analysis of features.

Bibliography

[Abbott and Letelier] M. R. Abbott and R. M. Letelier *Algorithm Theoretical Basis Document Chlorophyll Fluorescence (MODIS Product Number 20)* (College of Oceanic and Atmospheric Sciences, Oregon State University)

[Baker, 2008] N. R. Baker *Chlorophyll Fluorescence: A Probe of Photosynthesis In Vivo* (The Annual Review of Plant Biology 59: 89-113, 2008)

[Borstad et al., 1989] G. A. Borstad, D. A. Hill and R. C. Kerr *Use of the Compact Airborne Spectrographic Imager (CASI): Laboratory examples* (Geoscience and Remote Sensing Symposium, G. A. Borstad Associates Ltd., 100-Marine Technology Centre, 1989, Sidney)

[Boyle, 2007] P. Boyle *Gaussian Process for Regression and Optimisation* (Doctor of Philosophy in Computer Science, Victoria University of Wellington, 2007)

[Camps-Valls et al., 2006] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López and J. Calpe-Maravilla *Retrieval of oceanic chlorophyll concentration with relevance vector machines* (Remote Sensing of Environment 105 23-33, 2006)

[Camps-Valls et al., 2009] G. Camps-Valls, J. Muñoz-Marí, L. Gómez-Chova, K. Richter and J. Calpe-Maravilla *Biophysical Parameter Estimation With a Semisupervised Support Vector Machine* (IEEE Geoscience and Remote Sensing Letters, Volume 6, NO. 2, 2009)

[Canziani et al., 2008] G. Canziani, R. Ferrati, C. Marinelli and F. Dukatz *Artificial Neural Networks and Remote Sensing In The Analysis of Highly Variable Pampean Shallow Lakes* (Mathematical Bioscienes and Engineering, Volume 5, Number 4, 2008)

[Chaerle et al., 2007] L. Chaerle, I. Leinonen, H. G. Jones and D. Van Der Straeten *Monitoring and screening plant populations with combined thermal and chlorophyll fluorescence imaging* (Journal of Experimental Botany, Vol. 58, No. 4, pp. 773-784, 2007)

[Cotterill, 2002] R. M. J. Cotterill *Biophysics An Introduction* (John Wileyn and Sons, Chichester, England, pp. 187-199, 2002)

[Cracknell et al., 2009] A. P. Cracknell, V. Krapivin and C. A. Varotsos *Global Climatology and Ecodynamics Anthropogenic changes to Planet Earth* (Praxis Publishing Ltd, Germany, 2009)

[Cressie, 1993] N. A. C. Cressie *Statistics for Spatial data* (John Wileyn and Sons, USA, 1993)

[Cutter et al., 2000] M. A. Cutter, D. R. Lobb and R. A. Cockshott *Compact High Resolution Imaging Spectrometer (CHRIS)* (Acta Astronautica Vol. 46, Nos. 2-6, pp. 263-268,Kent, 2000)

[de Freitas, 2011] N. de Freitas *Multivariate Gaussian Models* (Lecture notes, CPSC540, KPM Book Sections: 5 and 31.2, 2011)

[de Freitas, 2013] N. de Freitas *Gaussian Processes* (Lecture notes, CPSC540, KPM Book Sections: 4.3 and 15.2, 2013)

[Do, 2007] C. B. Do *Gaussian process* (Lecture notes, Stanford Engineering Everywhere, 2007)

[Edner et al., 1995] H. Edner, J. Johansson, P. Ragnarson, S. Svanberg and E. Wallinder *Remote Monitoring of Vegetation using a Fluorescence Lidar System in Spectrally Resolving and Multi-Spectral Imaging Modes* (EARSeL Advances in Remote Sensing, Vol.3, No. 3-VII, 1995)

[Govindjee, 1975] Govindjee *Bioenergetics of Photosynthesis* (Academic press, 1975)

[Haykin, 1999] S. Haykin *Neural Networks - A Comprehensive Foundation (Second Edition)* (Prentice Hall, 1999)

[Hooker et al., 1992] S. B. Hooker, W. E. Esaias, G. C. Feldman, W. W. Gregg and C. R. McClain *SeaWiFS Technical Report Series Volume 1, An Overview of SeaWiFS: and Ocean Color* (NASA Technical Memorandum 104566, Vol. 1, 1992)

[Jensen, 2007] J. R. Jensen *Remote Sensing of the Environment An Earth Resource Perspective* (Pearson Prentice Hall, 2007)

[Jenssen et al., 2006] R. Jenssen, D. Erdogmus, K. E. Hild II , J. C. Principe and T. Eltoft *Information cut for clustering using a gradient descent approach* (Pattern Recognition Society, doi:10.1016/j.patcog.2006.06.028, Elsevier, 2006)

[Jenssen et al., 2012] R. Jenssen, M. Kloft, A. Zien, S. Sonnenburg and K.-R. Müller *A Scatter-Based Prototype Framework and Multi-Class Extension of Support Vector Machines* (PLOS ONE, Vol. 7, issue 10, e42947, 2012)

[Jones & Vaughan, 2010] H. G. Jones & R. A. Vaughan *Remote sensing of vegetation* (Oxford University Press, Oxford, 2010)

[Joiner et al., 2013] J. Joiner, L. Guanter, R. Lindstrot, M. Voigt, A. P. Vasilkov, E. M. Middleton, K. F. Huemmrich, Y. Yoshida and C. Frankenberg *Global monitoring of terrestrial chlorophyll fluorescence moderate-spectral-resolution near-infrared satellite measurements: methodology, simulations, and applications to GOME-2* (Atmospheric Measurement Techniques, 6, 2803-2823, 2013)

[Kemp et al., 2012] N. Kemp, C. McMullen and A. Shih *Mixotrophic Lipid Enrichment* (Brown Industries,Inc., 2012)

[Kjems et al., 2002] U. Kjems, L. K. Hansen, J. Anderson, S. Frutiger, S. Muley, J. Sidtis, D. Rottenberg and S. C. Strother *The Quantitative Evaluation of Functional Neuroimaging Experiments: Mutual Information Learning Curves* (NeuroImage 15, 772-786, doi:10.1006/nimg.2001.1033, Elsevier Science, USA, 2002)

[Kwiatkowska and Fargion, 2003] E. J. Kwiatkowska and G. S. Fargion *Application of Machine-Learning Techniques Toward the Creation of a Consistent and Calibrated Global Chlorophyll Concentration Baseline Dataset Using Remotely Sensed Ocean Color Data* (IEEE Transactions on Geascience and Remote SEnsing, Vol. 41, NO, 12, 2003)

[Li et al., 2009] D. Li, J. Shan and J. Gong *Geospatial Technology for Earth Observation* (Springer Science+Business Media, LLC 2009)

[Matsushita et al., 2012] B. Matusushita, W. Yang, P. Chang, F. Yang and T. Fukushima *A simple method for distinguishing global Case-1 and Case-2 waters using SeaWiFS measurements* (ISPRS Journal of Photogrammetry and Remote Sensing 69 74-87, 2012)

[Minnitt and Assibey-Bonsu, 2003] R. C. A. Minnitt and W. Assibey-Bonsu *Keynote Address: A tribute to Prof. D. G. Krige for his contributions over a period of more than half a century* (South African Institute of Mining and Metallurgy, 2003)

[MacKay, 2003] D. J. C. MacKay *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, 2003)

[Murphy, 2012] K. P. Murphy *Machine Learning: A Probabilistic Perspective* (Massachusetts Institute of Technology, ISBN 978-0-262-01802-9, 2012)

[Neal, 1995] R. M. Neal *Bayesian Learning for Neural Networks* (Doctor of Philosophy in Computer Sciences, University of Toronto, 1995)

[Obádovics, 2010] Gy. J. Obádovics *Lineáris algebra* (Scolar kiadó, Budapest, 2010)

[O'Reilly et al., 2000] J. E. O'Reilly and 24 Coauthors *SeaWiFS Postlaunch Calibration and Validation Ananlysis* (Part 3. NASA Tech. Memo. 2000-206892, Vol. 11, S. B. Hooker and E. R. Firestone, Eds., NASA Goddard Space Flight Center, 2000)

[Pasolli et al., 2010] L. Pasolli, F. Melgani and E. Blanzieri *Gaussian Process Regression for Estimating Chlorophyll Concentration in Subsurface Waters From Remote Sensing Data* (IEEE Geoscience and Remote Sensing letters, Vol. 7, NO. 3, 2010)

[Petersen & Pedersen, 2008] K. B. Petersen & M. S. Pedersen *The Matrix Cookbook* (Petersen & Pedersen, 2008)

[Rasmussen et al., 2011] P. M. Rasmussen, K. H. Madsen, T. E. Lund and L. K. Hansen *Visualization of nonlinear kernel models in neuroimaging by sensitivity maps* (Neuroimage 55 (2011) 1120-1131)

[Rasmussen and Williams, 2006] C. E. Rasmussen and C. K. I. Williams *Gaussian Process for Machine Learning* (MIT Press, Massachusetts Institute of Technology, 2006)

[Rast et al., 1999] M. Rast, J. L. Bézy and S. Bruzzi *The ESA Medium Resolution Imaging Spectrometer MERIS-a review of the instrument and its mission* (INT. J. Remote Sensing, Vol. 20, NO. 9, 1681-1702, 1999)

[Reynolds, 2006] C. S. Reynolds *The Ecology of Phytoplankton* (Cambridge University Press, 2006)

[Robinson, 2004] I. S. Robinson *Measuring the Oceans from Space: The principles and methods of satellite oceanography* (Praxis Publishing Ltd, Germany, 2004)

[Roháček and Barták, 1999] K. Roháček and M. Barták *Technique of the modulated chlorophyll fluorescence: basic concepts, useful parameters, and some applications* (Photosynthetica 37 (3): 339-363, 1999)

[Sacks et al., 1989] J. Sacks, W. J. Welch, T. J. Mitchell and H. P. Wynn *Design and Analysis of Computer Experiments* (Statistical Science, Vol. 4, No. 4., pp. 409-423, 1989)

[Schölkopf et al., 2013] B. Schölkopf, Z. Luo and V. Vovk *Empirical Inference Festschrift in Honor of Vladimir N. Vapnik* (Springer-Verlag, Berlin Heidelberg, 2013)

[SPARC Report, 2004] ESA, SPARC Data Acquisition Report (Contract number: 18307/04/NL/FF, University of Valencia,2004)

[Strother et al., 2002] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte and D. Rottenberg *The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework* (NeuroImage 15, 747-771, doi:10.1006/nimg.2001.1034, Elsevier Science, USA, 2002)

[Theodoridis and Koutroumbas, 2009] S. Theodoridis and K. Koutroumbas *Pattern Recognition* (Elsevier Inc.,Oxford, 2009)

[van Zyl, 2006] C. E. J. van Zyl *Introduction to the Physics and Techniques of Remote Sensing* (John Wiley & Sons, New Jersey, 2006)

[Vapnik et al., 1997] V. Vapnik, S. E. Golowich and A. Smola *Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing* (NIPS, 1997)

[Verrelst et al., 2011] J. Verrelst, J. Muñoz, L. Alonso, J. Delegido, J. P. Rivera, G. Camps-Valls and J. Moreno *Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3* (Remote Sensing of Environment 118 127-139, 2011)

[Verrelst et al., 2012a] J. Verrelst, L. Alonso, J. P. R. Caicedo, J. Moreno and G. Camps-Vall *Gaussian Process Retrieval of Chlorophyll Content From Imaging Spectroscopy Data* (IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2012)

[Verrelst et al., 2012b] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido and J. Moreno *Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques* (IEEE Transactions on Geoscience and Remote Sensing, Volume 50, NO. 5, 2012)

[Yitayew, 2012] T. G. Yitayew *Multi-sensor Data Fusion and Feature Extraction for Forest Applications* (University of TromsÃ¸, FYS-3900 Master's Thesis in Physics, 2012)

[Walpole et al., 2007] R. E. Walpole, R. H. Myers, S. L. Myers and K. Ye *Probability and Statistics for Engineers and Scientists* (Pearson International Edition, 2007)

[Whitehead et al., 1997] P. G. Whitehead, A. Howard and C. Arulmani *Modelling algal growth and transport in rivers: a comparison of time series analysis, dynamic mass balance and neural network techniques* (Hydrobiologia 349: 39-46, 1997)

[Williams and Rasmussen, 1996] C. K. I. Williams and C. E. Rasmussen *Gaussian Processes for Regression* (MIT Press, 1996)

[Wojtaszek, 2010] M. V. Wojtaszek *Data acquisition and integration 6.: Remote Sensing* (University of West Hungary Faculty of Geoinformatics, 2010)

[Zhan et al., 2003] H. Zhan, P. Shi and C. Chen *Retrieval of Oceanic Chlorophyll Concentration Using Support Vector Machines* (IEEE Transactions on Geoscience and Remote Sensing, Vol. 41, NO. 12, 2003)

[Zurada et al., 1994] J. M. Zurada, A. Malinowski and I. Cloete *Sensitivity Analysis for Minimization of Input Data Dimension for Feedforward Neural Network* (IEEE International Symposium on Circuits and Systems, Vol. 6, pp. 447-450, 1994)