SEQUENTIAL AND NON-SEQUENTIAL HYPERTEMPORAL CLASSIFICATION AND CHANGE DETECTION OF MODIS TIME-SERIES

by

Trienko Lups Grobler

Submitted in partial fulfilment of the requirements for the degree

Philosophiae Doctor (Engineering)

in the

Department of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Built Environment and Information Technology
UNIVERSITY OF PRETORIA

September 2012

SUMMARY

SEQUENTIAL AND NON-SEQUENTIAL HYPERTEMPORAL CLASSIFICATION AND CHANGE DETECTION OF MODIS TIME-SERIES

by

Trienko Lups Grobler

Promoter(s):	Prof. J.C. Olivier, Dr. W. Kleynhans and Dr. A.J. van Zyl
Department:	Electrical, Electronic and Computer Engineering
University:	University of Pretoria
Degree:	Philosophiae Doctor (Engineering)
Keywords:	noise-harmonic features, Coloured Simple Harmonic Oscillator, Support Vector Machine, Moderate Resolution Imaging Spectroradiometer, Cumulative Sum, Ornstein-Uhlenbeck process, inductive simulator, hypertemporal classification, hypertemporal change detection, sequential analysis

Satellites provide humanity with data to infer properties of the earth that were impossible a century ago. Humanity can now easily monitor the amount of ice found on the polar caps, the size of forests and deserts, the earth's atmosphere, the seasonal variation on land and in the oceans and the surface temperature of the earth. In this thesis, new hypertemporal techniques are proposed for the settlement detection problem in South Africa. The hypertemporal techniques are applied to study areas in the Gauteng and Limpopo provinces of South Africa. To be more specific, new sequential (windowless) and non-sequential hypertemporal techniques are implemented. The time-series employed by the new hypertemporal techniques are obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) sensor, which is on board the earth observations satellites Aqua and Terra. One MODIS dataset is constructed for each province.

A Support Vector Machine (SVM) [1] that uses a novel noise-harmonic feature set is implemented to detect existing human settlements. The noise-harmonic feature set is a non-sequential hypertemporal feature set and is constructed by using the Coloured Simple Harmonic Oscillator (CSHO) [2]. The CSHO consists of a Simple Harmonic Oscillator (SHO) [3], which is superimposed on the Ornstein-

Uhlenbeck process [4]. The noise-harmonic feature set is an extension of the classic harmonic feature set [5]. The classic harmonic feature set consists of a mean and a seasonal component. For the case studies in this thesis, it is observed that the noise-harmonic feature set not only extends the harmonic feature set, but also improves on its classification capability.

The Cumulative Sum (CUSUM) algorithm was developed by Page in 1954 [6]. In its original form it is a sequential (windowless) hypertemporal change detection technique. Windowed versions of the algorithm have been applied in a remote sensing context. In this thesis CUSUM is used in its original form to detect settlement expansion in South Africa and is benchmarked against the classic band differencing change detection approach of Lunetta et al., which was developed in 2006 [7]. In the case of the Gauteng study area, the CUSUM algorithm outperformed the band differencing technique. The exact opposite behaviour was seen in the case of the Limpopo dataset.

Sequential hypertemporal techniques are data-intensive and an inductive MODIS simulator was therefore also developed (to augment datasets). The proposed simulator is also based on the CSHO. Two case studies showed that the proposed inductive simulator accurately replicates the temporal dynamics and spectral dependencies found in MODIS data.

OPSOMMING

SEKWENSIËLE EN NIE-SEKWENSIËLE HIPERTEMPORALE KLASSIFIKASIE EN VERANDERING-OPSPORING VAN MABS-TYDSREEKSE

deur

Trienko Lups Grobler

Promotor(s):	Prof. J.C. Olivier, Dr. W. Kleynhans en Dr. A.J. van Zyl
Departement:	Elektriese, Elektroniese en Rekenaar-Ingenieurswese
Universiteit:	Universiteit van Pretoria
Graad:	Philosophiae Doctor (Ingenieurswese)
Sleutelwoorde:	harmoniesegeraas-kenmerkstel, Gekleurde Eenvoudige Harmoniese Ossillator, Steunvektormasjien, Matigeresolusie- Beeldskeppende Spektrale Radio-ontvanger, Kumulatiewesom, Ornstein-Uhlenbeck proses, induktiewe simulator, hipertemporale klassifikasie, hipertemporale verandering-opsporing, sekwensiële analiese

Satelliete gee die mensdom die geleentheid om dinge van die aarde te leer wat nie ’n eeu gelede moontlik was nie. Die mensdom kan nou maklik die hoeveelheid ys op die pole, die grootte van woude en woestyne, die aarde se atmosfeer, die seisoenale veranderinge op land en in die oseane, asook die temperatuur op die aarde se oppervlak monitor. In hierdie proefskrif word nuwe hipertemporale tegnieke vir die sogenaamde nedersettingsopsporingsprobleem in Suid-Afrika beskryf. Die nuwe hipertemporale tegnieke word toegepas op studie-areas in die Gauteng- en Limpopoprovinsies van Suid-Afrika. Om meer spesifiek te wees, nuwe sekwensiële (vensterlose) en nie-sekwensiële hipertemporale tegnieke word bespreek. Die tydsreekse wat deur die hipertemporale tegnieke benodig word, word deur die Matigeresolusie- Beeldskeppende Spektrale Radio-ontvanger (MABS) sensor verskaf, wat gemonteer is op die aardobservasiesatelliete Aqua en Terra. Een MABS-datastel is saamgestel vir elke provinsie.

’n Steunvektormasjien (SVM) [1] wat ’n nuwe harmoniesegeraas-kenmerkstel gebruik om bestaande nedersettings op te spoor, is geïmplementeer. Die nuwe harmoniesegeraas-kenmerkstel is ’n nie-sekwensiële hipertemporale-kenmerkstel en is saamgestel deur die Gekleurde Eenvou-

dige Harmoniese Ossillator (GEHO) te gebruik [2]. Die GEHO bestaan uit 'n Eenvoudige Harmoniese Ossillator (EHO) [3] wat gesuperponeer is op die Ornstein-Uhlenbeck-proses [4]. Die harmoniesegeraas-kenmerkstel is 'n uitbreiding van die klassieke harmoniese-kenmerkstel [5]. Die klassieke harmoniese-kenmerkstel bestaan uit 'n gemiddelde en 'n seisoenale komponent. Aan die hand van die gevallestudies in hierdie proefskrif is daar gevind dat die harmoniesegeraas-kenmerkstel nie net 'n uitbreiding van die klassieke harmoniese-kenmerkstel is nie, maar dat die harmoniesegeraas-kenmerkstel ook die klassifikasie-vermoë van die klassieke harmoniese-kenmerkstel verbeter.

Die Kumulatiewesom- (KUMSOM) algoritme is in 1954 deur Page ontwikkel [6]. In sy oorspronklike vorm is dit 'n sekwensiële hipertemporale veranderingopsporingstegniek. Afgeknotte weergawes van die algoritme is al vantevore in 'n afstandswaarneming-konteks gebruik. In hierdie proefskrif word KUMSOM in sy oorspronklike vensterlose vorm gebruik om die vorming van nuwe nedersettings in Suid-Afrika op te spoor. Die KUMSOM-algoritme word ook met die bandaftrekkingmetode vergelyk, wat in 2006 deur Lunetta et al. ontwikkel is [7]. In die geval van die Gauteng-gevallestudie lewer KUMSOM beter resultate as die bandaftrekkingmetode. Die presiese teenoorgestelde gedrag is waargeneem in die geval van Limpopo.

Sekwensiële hipertemporale tegnieke is baie data-intensief, gevolglik is 'n induktiewe MABS-simulator ontwerp wat datastelle kan vergroot. Die nuwe simulator is ook gebaseer op die GEHO. Twee gevallestudies het gewys dat die induktiewe simulator wel die temporale dinamika en spektrale afhanklikheid kan dupliseer wat voorkom in MABS-data.

ACKNOWLEDGEMENTS

In the first place I would like to thank God for the many blessings that I have received out of his gracious hand. Second, I would like to thank my supervisors, Prof. J.C. Olivier, Dr. W. Kleynhans and Dr. A.J. van Zyl, and my colleagues, Dr. B.P. Salmon and Mr. E.R. Ackermann, for their technical contributions. I would also like to thank my wife, Mrs. C.I. Grobler, friends and family (especially my parents and brother Mr. P.N. Grobler) for their support. Lastly, I would like to thank my sister, Mrs. J.B. Louw, who helped me obtain some of the reference material used in this thesis. The Council for Scientific and Industrial Research (CSIR) (especially Mr. A. van der Merwe) should also be thanked for their financial contribution. The thesis was proofread by Mrs. M.B. Bradley.

In die begin het God die hemel en die aarde geskep. Die aarde was heeltemal onbewoonbaar, dit was donker op die diep waters, maar die Gees van God het oor die waters gesweef. Toe het God gesê: "Laat daar lig wees!" En daar was lig.

Genesis 1:1-3, Bybel Nuwe Vertaling, 1983

This dissertation is dedicated to my wife and parents.

LIST OF ABBREVIATIONS

AG	Asymmetric Gaussian
AIRS	Atmospheric Infrared Sounder
AMSR-E	Advanced Microwave Scanning Radiometer-EOS
AMSU	Advanced Microwave Sounding Unit
ANN	Artificial Neural Network
AR	Autoregressive
ARL	Average Run Length
ASN	Average Sample Number
ASTER	Advanced Spaceborne Thermal Emission and Reflection Radiometer
ATMS	Advanced Technology Microwave Sounder
AVHRR	Advanced Very High Resolution Radiometer
B	Blue
BB	Blackbody
BFAST	Breaks for Additive Seasonal and Trend
BOB	Bureau of Budget
BRDF	Bidirectional Reflectance Distribution Function
CERES	Clouds and the Earth's Radiant Energy System
CSIR	Council for Scientific and Industrial Research

CrIS	Cross-track Infrared Sounder
CSHO	Coloured Simple Harmonic Oscillator
CUSUM	Cumulative Sum
CVA	Change Vector Analysis
CZCS	Coastal Zone Color Scanner
DAAC	Distributed Active Archive Center
DL	Double Logistic
DN	Digital Number
DOD	Department of Defense
DOI	Department of the Interior
EDOS	EOS Data and Operations System
EHO	Eenvoudige Harmoniese Ossillator
EOS	Earth Observing System
EOSMRWG	EOS Science and Mission Requirements Working Group
ERTS	Earth Resources Technology Satellite
ESE	Earth Science Enterprise
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
EVI	Enhanced Vegetation Index
FFT	Fast Fourier Transform
FP	False Positive Rate

FPAR	Fraction of Photosynthetically Active Radiation
FPAs	Focal Plane Assemblies
G	Green
GEHO	Gekleurde Eenvoudige Harmoniese Ossillator
GES DAAC	GSFC Earth Sciences DAAC
GIS	Geographic Information System
GMS	Geostationary Meteorological Satellite
GOES	Geostationary Operational Environmental Satellite
GSFC	Goddard Space Flight Center
HSB	Humidity Sounder for Brazil
HyspIRI	Hyperspectral Infrared Imager
i.i.d.	independent and identically distributed
ICA	Independent Component Analysis
IFOV	Instantaneous Field of View
IFOV	Instantaneous Field of View
INSAT	Indian National Satellite System
IPO	Integrated Program Office
IR	Infrared
IRS	Indian Remote Sensing Satellite
ISRO	Indian Space Research Organisation

ITOS	Improved TIROS Operational System
JAXA	Japan Aerospace Exploration Agency
JERS	Japanese Earth Resource Satellite
KDE	Kernel Density Estimation
KUMSOM	Kumulatiewesom
LAADS	L1 and Atmosphere Archive and Distribution System
LAI	Leaf Area Index
LDCM	Landsat Data Continuity Mission
LP DAAC	Land Processes DAAC
LWIR	Long-wave Infrared
MABS	Matigeresolusie- Beeldskeppende Spektrale Radio-ontvanger
MISR	Multi-angle Imaging SpectroRadiometer
MODAPS	MODIS Adaptive Processing System
MODIS	Moderate Resolution Imaging Spectroradiometer
MOPITT	Measurements of Pollution in the Troposphere
MSS	Multispectral Scanner
MTPE	Mission to Planet Earth
MWIR	Mid-wave Infrared
NASA	National American Aeronautics and Space Administration
NDVI	Normalised Difference Vegetation Index

NESDIS	National Environmental Satellite Data and Information Service
NIR	Near-Infrared
NOAA	National Oceanic and Atmospheric Administration
NPOESS	National Polar Orbiting Environmental Satellite Series
NPP	NPOESS Preparatory Project
NSIDC DAAC	National Snow and Ice Data Center DAAC
OA	Overall Accuracy
OC	Operating Characteristic
OCDPS	Ocean Color Data Processing System
OLS	Ordinary Least Squares
OMPS	Ozone Mapping and Profiler Suite
PCA	Principal Component Analysis
POES	Polar-orbiting Operational Environmental Satellite
PROSAIL	PROSPECT + Scattering by Arbitrary Inclined Leaves
R	Red
RBV	Return Beam Vidicon
ROC	Receiver Operating Curve
SAR	Synthetic Aperture Radar
SBRC	Hughes/Santa Barbara Research Center
SD	Solar Diffuser

SDSM	Solar Diffuser Stability Monitor
SEASAT	Sea Satellite
SHO	Simple Harmonic Oscillator
SMA	Spectral Mixture Analysis
SMMR	Scanning Multichannel Microwave Radiometer
SNR	Signal-to-Noise Ratio
SPOT	Système Probatoire d'Observation de la Terre
SPRT	Sequential Probability Ratio Test
SRCA	Spectral Radiometric Calibration Assembly
SSE	Sum of Squared Error
SV	Space View
SVM	Support Vector Machine
SWIR	Short-wave Infrared
TDRSS	Tracking and Data Relay Satellite System
TIROS	Television Infrared Observation Satellite
TM	Thematic Mapper
TOMS	Total Ozone Mapping Spectrometer
TP	True Positive Rate
UAV	Unmanned Aerial Vehicle
UME	Uniformly Most Efficient

UN	Unitited Nations
USA	United States of America
USGS	United States Geological Survey
USSR	Union of Soviet Socialist Republics
VIIRS	Visible Infrared Imager Radiometer Suite

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Satellites provide humanity with data to infer properties of the earth that were impossible a century ago. Humanity can now easily monitor the amount of ice found on the polar caps, the size of forests and deserts, the earth's atmosphere, the seasonal variation on land and in the oceans and the surface temperature of the earth.

In this thesis satellite data are used to detect and estimate human settlement expansion. Anthropogenic changes to the environment are driven by the need to provide food, water and housing to more than 7 billion people. Unfortunately humanity's need to survive has a negative effect on the environment [8]. For example, human settlement expansion on the outskirts of Xalapa city, the capital city of the state of Veracruz in Mexico, is causing severe environmental damage in the region [9]. Monitoring the growth of settlements around the world is important as it could enable multiple governments to enforce sustainable development, which would decrease humanity's negative impact on the environment. Monitoring settlement expansion in South Africa is the primary focus of the thesis. Monitoring settlement expansion is especially important in South Africa as it is one of the most pervasive forms of land cover change found in southern Africa [10].

The chapter starts by explaining the importance of monitoring settlement expansion in South Africa (in greater detail than in the previous paragraph) and then continues by briefly introducing the techniques and the data that are used to detect settlement expansion in Section 1.2, Section 1.3 and Section 1.4. The main problem statement is given in Section 1.6, which also discusses the main contributions made by this thesis. The chapter ends with a list of publications written during the course of the study and a brief overview of the remaining chapters.

1.1 SETTLEMENT DETECTION

Human settlement expansion in South Africa is often unplanned and informal in nature, meaning that the settlements form in randomly selected places, without any provision for electricity, running water, refuse removal or water-borne sewage. These informal settlements usually develop as people move closer to employment opportunities [10].

According to a report from the nineteenth special session of the general assembly of the Unitited Nations (UN), the South African government needs to be empowered to plan, implement, develop and manage human settlements [11]. As mentioned before, predicting where human settlements will form is rather difficult. For this reason, the main focus of this thesis is to develop affordable techniques that will aid the South African government in monitoring the expansion of human settlements so that efficient decisions can be made regarding infrastructure development and resource allocation. Remote sensing provides an attractive solution to this problem, since the data of certain remote sensing sensors are free and provide large-scale monitoring capabilities. In this thesis remote sensing data will be the primary tool used to monitor settlement expansion. Two provinces of South Africa were selected as study regions, namely Gauteng and Limpopo. Gauteng was selected as it has a higher population growth rate than the remaining provinces of South Africa, which makes it an attractive study area. Limpopo was chosen due to the fact that it is a very poor province of South Africa [12]. Poverty usually leads to the formation of informal settlements.

1.2 HYPERTEMPORAL APPROACHES

Most of the remote sensing classification and change detection techniques available in literature are multi-temporal (usually bi-temporal or single date) techniques [13, 14]. In contrast, the approaches that are investigated in this thesis are all hypertemporal techniques. Hypertemporal techniques fully exploit the information located in hypertemporal time-series to classify or detect changes. A hypertemporal time-series is defined as a time-series that consists of frequent equal-spaced observations [10]. The benefit of using the temporal dimension effectively is that the date selection problem is circumvented [15]. When working with multi-temporal algorithms, selecting optimal dates is important, as class separability may be different during different seasons. Another possible benefit of hypertemporal techniques is that a time-series can provide phenological metrics, which are used for discerning between different vegetation types. There already exists a few hypertemporal classification

and change detection techniques in literature [5, 7, 10, 15–21]. The approaches cited are definitely not an exhaustive list.

1.3 DATA SELECTION

The MCD43A4 MODIS product consists of Bidirectional Reflectance Distribution Function (BRDF) corrected land surface reflectance (eight-day composite, 500 m resolution) time-series. The product was chosen to investigate the hypertemporal techniques discussed in this thesis, because the MCD43A4 product provides a long, reliable high temporal remote sensing time-series. Another benefit of the adjusted land spectral reflectance product is that it significantly reduces the anisotropic scattering effects of surfaces under different illumination and observation conditions [22]. Furthermore, MODIS data, when compared to Advanced Very High Resolution Radiometer (AVHRR) data, exhibit enhanced spectral and radiometric resolution, wide geographical coverage and improved atmospheric corrections, while preserving the same temporal resolution [15].

1.4 SEQUENTIAL APPROACHES

Sequential hypertemporal approaches are a relatively new subset of current hypertemporal remote sensing techniques that are available in literature. Up to now the focus in remote sensing has mainly been on sequential classification approaches [23]. A good literature review on the field of sequential analysis can be found in [24, 25]. Sequential approaches are threshold-based techniques. Sequential approaches keep sampling observations until an on-line statistic crosses a predefined threshold. The main advantage of sequential approaches is that on average, sequential approaches usually require fewer observations than fixed-sample-size approaches. The reason for the speed increase is that sequential approaches terminate uniquely for each observable sequence. Sequential approaches try to jointly optimise the accuracy and detection delay of the classifier or change detector. It should be clear that the accuracy of a classifier or change detector (which is the primary focus of remote sensing literature) is not the only design criterion to consider when designing classifiers and change detectors. The detection delay of a classifier or change detector is also an important design criterion [23]. Sequential classification and its application in remote sensing are studied in detail in [23]. One of the objectives of this thesis is to verify the preliminary sequential results of [23] and extend sequential analysis to the remote sensing change detection realm. Interest in sequential techniques is expressed in this thesis because, the South African government not only needs to detect settlement expansion,

but also needs to do so as quickly as possible.

1.5 INDUCTIVE SIMULATION

Sequential hypertemporal approaches usually rely on densities. As sequential approaches employ densities, they require large amounts of training data. For the current thesis, large amounts of training data are not available and therefore an efficient simulator that can augment datasets is required. Most remote sensing simulators in the literature are deductive simulators, which means that they employ the biophysical laws that govern the reflection of light [26, 27]. In contrast to deductive simulators, an inductive simulator uses a mathematical (inductive) model that is fitted directly on an existing dataset. The aim of an inductive model is to model the statistical characteristics of the original dataset and therefore it can be used for dataset augmentation.

1.6 PROBLEM STATEMENT

At this stage enough background has been discussed to formulate the fundamental problem statement of this thesis:

Problem Statement: Develop new sequential or non-sequential hypertemporal remote sensing techniques to detect settlement expansion in South Africa.

The existing techniques investigated and the contributions made by this thesis in trying to solve the above problem statement are discussed in the following three sections.

1.6.1 Existing hypertemporal techniques

The following existing hypertemporal techniques were investigated in this thesis:

1. Ackermann [23] applied sequential analysis to the remote sensing field and in doing so developed the time-varying maximum likelihood classifier. Ackermann also developed the minimum distance classifier [16].

2. Lhermitte et al. [5] showed that an efficient classifier could be created by using only the mean and seasonal harmonic components of a remotely sensed time-series. For the remainder of this thesis, the harmonic feature group proposed by Lhermitte et al. is denoted by $\boldsymbol{\iota}$.

3. Carrão et al. [15] showed that temporal features can provide good separability, which inspired the development of the temporal feature group $\boldsymbol{\zeta}$ (defined formally in Section 4.2.4.2).

4. Lunetta et al. [7] developed the band differencing algorithm (a hypertemporal change detection approach).

1.6.2 Contribution to hypertemporal classification

The following contribution was made in the hypertemporal remote sensing classification field:

1. A non-sequential hypertemporal SVM classifier was implemented. The SVM classifier uses a novel (an outcome of this thesis) noise-harmonic feature group $\boldsymbol{\theta}$ (where the symbol $\boldsymbol{\theta}$ is used to represent this noise-harmonic feature group), which is an extension (in size but also in classification capability) of the classic harmonic feature group $\boldsymbol{\iota}$ proposed in [5]. The feature group $\boldsymbol{\theta}$ is constructed from the CSHO [2]. The SVM using $\boldsymbol{\theta}$ is benchmarked against the minimum-distance classifier, the time-varying maximum likelihood classifier, an SVM classifier using the harmonic feature group $\boldsymbol{\iota}$ and an SVM using the temporal feature group $\boldsymbol{\zeta}$ (see Chapter 4 and Chapter 5 for more detail) [5, 15, 16, 23]. Generally the SVM classifier using $\boldsymbol{\theta}$ outperformed the minimum-distance classifier, the time-varying maximum likelihood classifier, the SVM classifier using the harmonic feature group $\boldsymbol{\iota}$ and the SVM using the temporal feature group $\boldsymbol{\zeta}$. The performance results of the new hypertemporal technique were published in [2].

1.6.3 Contribution to hypertemporal change detection

The following contributions were made to the hypertemporal remote sensing change detection field:

1. The sequential change detection algorithm called CUSUM (windowless version) [6] was introduced into the remote sensing field and benchmarked against the popular hypertemporal approach developed by Lunetta et al. (see Chapter 4 for more detail) [7, 10, 28]. This thesis therefore builds on and extends the work done by Ackermann [23], which mainly focused on sequential detection (had a smaller scope). Windowed versions of the CUSUM algorithm have been used in a remote sensing context [28, 29]. The problem of windowed approaches, is that it

becomes important to select an optimal window length. If the window is chosen either too small or too large then the change detection capability of the approach deteriorates. The windowless CUSUM approach presented here, is more flexible as it circumvents the optimal window length issue. The results of the windowless CUSUM approach were published in [30].

2. To implement the CUSUM algorithm effectively, an inductive simulator was developed (see Section 1.5 and Chapter 4 for more detail). In selective cases, inductive statistical models similar to the one used in this thesis have been used to simulate a single time-series [31]. The complex issue of replicating multispectral correlation and dependence was not undertaken in [31]. The inductive simulator developed in this thesis accurately enforces multispectral correlation and dependence. The details of this simulator were published in [32].

1.7 PUBLICATIONS AND RELATED WORK

The following conference papers (where C# implies a conference paper) were produced during the course of the PhD study:

[C1] E.R. Ackermann, T.L. Grobler, A.J. van Zyl, K.C. Steenkamp and J.C. Olivier, "Minimum error land cover separability analysis and classification of MODIS time series data", *IEEE International Geoscience and Remote Sensing Symposium*, Vancouver, Canada, July 2011, pp. 2999–3002.

[C2] T.L. Grobler, E.R. Ackermann, J.C. Olivier and A.J. van Zyl, "Systematic Luby Transform codes as incremental redundancy scheme", *IEEE AFRICON*, Livingston, Zambia, September 2011, pp. 1–5.

[C3] E.R. Ackermann, T.L. Grobler, A.J. van Zyl and J.C. Olivier, "Belief propagation for nonlinear block codes", *IEEE AFRICON*, Livingston, Zambia, September 2011, pp. 1–6.

[C4] T.L. Grobler, E.R. Ackermann, A.J. van Zyl, W. Kleynhans, B.P. Salmon and J.C. Olivier, "Sequential classification of MODIS time series", *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 6236–6239.

[C5] B.P. Salmon, W. Kleynhans , F. van den Bergh, J.C. Olivier, W.J. Marais, T.L. Grobler, K.J. Wessels, "A search algorithm to meta-optimize the parameters for an extended Kalman filter to

improve classification on hyper-temporal images", *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 4974–4977.

[C6] W. Kleynhans, B.P. Salmon, J.C. Olivier, F. van den Bergh, K.J. Wessels and T.L. Grobler, "Detecting land-cover change using a sliding window temporal autocorrelation approach", *IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, July 2012, pp. 6765–6768.

The following journal papers (where J# implies a journal paper) were published during the course of the PhD study:

[J1] T.L. Grobler, A.J. van Zyl, J.C. Olivier, W. Kleynhans, B.P. Salmon and W.T. Penzhorn, "Wu's algorithm and its possible application in Cryptanalysis", *African Journal of Mathematics and Computer Science Research*, vol. 5, no. 1, pp. 1–8, January 2012.

[J2] T.L. Grobler, E.R. Ackermann, J.C. Olivier, A.J. van Zyl and W. Kleynhans, "Land-Cover separability analysis of MODIS time-series data using a combined Simple Harmonic Oscillator and a Mean Reverting Stochastic Process", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 3, pp. 857–866, June 2012.

[J3] W. Kleynhans, B.P. Salmon, J.C. Olivier, F. van den Bergh, K.J. Wessels, T.L. Grobler, K.C. Steenkamp, "Land cover change detection using autocorrelation analysis on MODIS time series data: detection of new human settlements in the Gauteng province of South Africa", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 3, pp. 777–783, June 2012.

[J4] T.L. Grobler, E.R. Ackermann, A.J. van Zyl, J.C. Olivier, W. Kleynhans and B.P. Salmon, "Using Page's Cumulative Sum Test on MODIS time-series to detect land-cover changes", *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 2, pp. 332–336, March 2013.

[J5] T.L. Grobler, E.R. Ackermann, A.J. van Zyl, J.C. Olivier, W. Kleynhans and B.P. Salmon, "An inductive approach to simulating multispectral MODIS surface reflectance time series", *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 3, pp. 446–450, May 2013.

[J6] E.R. Ackermann, T.L. Grobler, W. Kleynhans, J.C. Olivier, B.P. Salmon and A.J. van Zyl, "Cavalieri Integration", *Quaestiones Mathematicae*, vol. 35, no. 3, pp. 265–296, September

2012.

1.8 LAYOUT OF THESIS

The outline of the thesis is as follows:

Chapter 2: The chapter provides a broad overview of the remote sensing field, which includes a brief history of remote sensing, an introduction to the physical principles behind remote sensing, an overview of remote sensing platforms and an introduction to the MODIS sensor. The MODIS data used by the classification and change detection algorithms investigated in this thesis are also presented in this chapter.

Chapter 3: The chapter provides a broad overview of the sequential analysis field. It starts with the Neyman-Pearson optimal classification result, which is the predecessor of modern sequential analysis. The chapter then continues to the field of sequential classification, which is discussed by using two frameworks, namely Wald's framework and the Bayesian framework. From sequential classification the chapter progresses to a group of statistical change detection algorithms grouped under the collective name of quickest detection. The quickest detection techniques discussed in the chapter are divided into Bayesian and non-Bayesian approaches. Two main non-Bayesian approaches are discussed, namely the CUSUM stopping time and the Shiryaev-Roberts stopping time (as well as its variants). The main reason for including this chapter is to provide the theoretical background knowledge required to implement CUSUM as a sequential hypertemporal remote sensing change detection algorithm.

Chapter 4: The chapter provides the technical details of the newly proposed algorithms as well as the benchmarking sequential and non-sequential hypertemporal classification and change detection algorithms investigated in the thesis. The details of the inductive simulator developed for the CUSUM algorithm are also found in this chapter. Furthermore, the chapter contains literature reviews of remote sensing classification and change detection.

Chapter 5: The chapter starts with preliminary data analysis results obtained from the test datasets. These results can be used to predict the performance of the different classification and change detection approaches. The chapter then gives the classification and change detection accuracies and rankings of the different sequential and non-sequential hypertemporal classification and change de-

tection algorithms investigated in the thesis.

Chapter 6: In this chapter the main conclusions from Chapter 5 are summarised.

CHAPTER 2

REMOTE SENSING

The main aim of this chapter is to introduce two datasets. All the hypertemporal techniques presented in Chapter 4 are applied to these two datasets. The two datasets are discussed in Section 2.8. The first few sections of this chapter however introduce the basic principles of remote sensing, as well as the MODIS sensor. These introductory sections are needed to understand the content of the datasets in Section 2.8.

Remote sensing is the science of converting data about the earth's surface, recorded with remote (distant) sensor platforms, into usable information. The remote sensors archive how the earth's surface reflects or transmits electromagnetic energy at different wavelengths and thus record an electromagnetic spectral signature of the earth's surface [33].

2.1 HISTORY OF REMOTE SENSING

It can be argued that the moment in time which gave birth to photography was in fact also the starting point of spaceborne remote sensing. Photography was invented in 1839. Those early photographs were created by the photographic processes of Nicephore Niepce, William Henry Fox Talbot, and Louis Jacques Daguerre [33]. The first aerial photograph was taken (of Bievre, France) by a Parisian photographer named Gaspard Felix Tournachon (from a balloon). The earliest *existing* aerial photograph was taken from a balloon over Boston in 1860 by James Wallace Black and immortalised by Oliver Wendell Holmes [33]. The First and Second World War sparked the widespread use of aerial photography as a surveillance tool. The use of aerial photography for environmental purposes only became popular after the Second World War [10]. The term "remote sensing" was first coined by Evelyn Pruit after recognising that "aerial photography" no longer accurately described the different

images recorded, as some images (at that point) were recorded by using wavelengths outside the visible spectrum [34]. The next step in the evolution (starting in the 1960s) of remote sensing was when humans started using spaceborne platforms to house remote sensing sensors. The space era, which is still continuing, can be discussed under four headings, namely *military reconnaissance satellites*, *manned space flight*, *meteorological satellites* and *earth resource satellites* [35].

2.1.1 Military reconnaissance satellites

Before 1960, the United States of America (USA) and the former Union of Soviet Socialist Republics (USSR) used aerial photographs to keep track of each other's military capability. However, at the Surprise Attack Conference in Geneva (in 1958) it was proposed for the first time to use satellites to gather military information [35]. CORONA was one of the first military programmes under which satellites were launched into space to perform military reconnaissance (active during the 1960s). These missions were usually very short in duration, typically no longer than one or two weeks. These early systems were constrained, since they could only carry a limited amount of film. The film canister was ejected and picked up as it descended to earth [35]. Later systems could store images in digital format and transported the data to the earth via telemetry.

2.1.2 Manned space flight

On April 12, 1961 Yuri Gagarin became the first person to orbit the earth. Although no photos were taken during this mission, it became apparent that spaceborne earth observation had great potential. The USA also started manned space missions in the 1960s, culminating in the moon landing in 1969 during the Apollo programme [35]. The Mercury (1961-1963), Gemini (1965-1966) and Skylab (1973-1974) programmes were some of the manned American programmes that captured pictures of the earth. The Russians also conducted their own manned missions, which included the Vostok and Voskhod programmes, which were analogous to the Mercury and Gemini missions [35].

2.1.3 Meteorological satellites

Meteorological satellites (weather satellites) paved the way for the modern earth resource satellites. Meteorological satellite, TIROS-1, was the first satellite that was used for earth observation and was launched by the USA on April 1, 1960 [35]. Both *polar orbiting* and *geostationary satellites* are used for weather prediction. A geostationary satellite completes its orbit every 24 hours, so that it can

always monitor one specific place on earth and is usually found at a higher altitude than polar orbiting satellites. Polar orbiting satellites do not complete their orbit in 24 hours and can therefore survey the entire surface of the earth [36].

There are a few polar orbiting satellite programmes worth highlighting [36]:

1. ITOS/NOAA or POES: ITOS-1 was launched on January 23, 1970, while NOAA-1 was launched on December 11, 1970. It is noteworthy to mention that NOAA-6 (launched on June 27, 1979) contained the first in a series of AVHRRs, the predecessor of MODIS. The ITOS/NOAA programme is administrated by the National Oceanic and Atmospheric Administration (NOAA). The most recent Polar-orbiting Operational Environmental Satellite (POES) launched is NOAA-19, launched on June 2, 2009.

2. Nimbus: Nimbus-1 was launched on August 28, 1964. Nimbus-7 (launched on June 27, 1979) carries the Coastal Zone Color Scanner (CZCS), the Total Ozone Mapping Spectrometer (TOMS) and the Scanning Multichannel Microwave Radiometer (SMMR). The Nimbus satellites were put into space by the National American Aeronautics and Space Administration (NASA).

There are four important geostationary satellite programmes, that together provide complete coverage (weather) of the globe, namely [36]:

1. Meteosat: Meteosat-1 was launched on November 23, 1977. The Meteosat programme is administrated by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) and covers Europe and Africa.

2. GOES: GOES-1 was launched on October 16, 1975. The Geostationary Operational Environmental Satellites (GOESs) are operated by the National Environmental Satellite Data and Information Service (NESDIS) and have been developed by NOAA. There are two main GOES satellites in use, GOES-W, which services the western Americas and the Atlantic Ocean, and GOES-E, which covers the eastern Americas and the Pacific.

3. Indian INSAT: INSAT-1B was launched on August 30, 1983. The Indian National Satellite System (INSAT) satellites where launched by the Indian Space Research Organisation (ISRO) and provides coverage of India and the Indian Ocean.

4. Japanese GMS: GMS-1 was launched on July 14, 1977. The Geostationary Meteorological Satellite (GMS) programme is driven by the Japan Meteorological Agency and covers South-East Asia and Japan.

The coverage areas for each of the geostationary weather satellites are displayed in Figure 2.1.

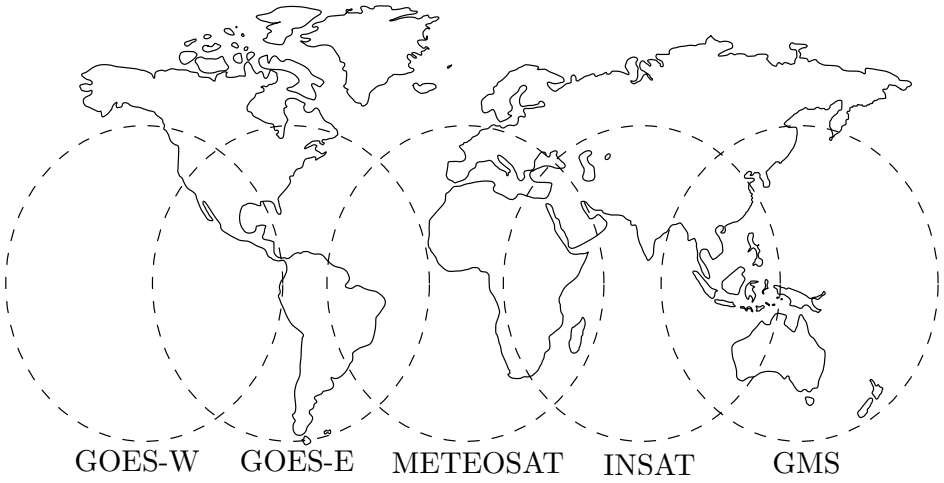


Figure 2.1: Worldwide coverage by international geostationary weather satellites.

2.1.4 Earth resource satellites

While weather satellites have been monitoring the earth's atmosphere since 1960 and have largely been considered useful, there was no real appreciation of land data from space before the development of earth resource satellites. The development of earth resource satellites can be divided into three generations. The first generation is mainly characterised by the fact that basic remote sensing sensors were used. The Landsat and Système Probatoire d'Observation de la Terre (SPOT) satellites are prime examples of first phase earth resource programmes. It can be argued that a second generation began with the inception of the Earth Observing System (EOS) programme, which hailed in the era of sophisticated remote sensing sensors that could survey the earth and allow humans to track climate change. The EOS was part of NASA's Mission to Planet Earth (MTPE), now called the Earth Science Enterprise (ESE). It is important to note that these generations are not mutually exclusive, and that some intersection does occur. At the moment earth resource satellite development is moving into a third generation that started with the launch of NPP. The third phase will be characterised by the use of remote sensing sensors that are far more advanced (giant leap) than the sensors housed in, for instance, Terra. These sensors will have the explicit goal of achieving the original vision of EOS,

which is to collect a 15-year global data set to address questions on climate change.

2.1.4.1 First generation

The idea of a civilian satellite that could be used for scientific earth surveillance was proposed in 1965 by William Pecora, director of the United States Geological Survey (USGS), and was inspired by the photographs taken on the Mercury, Gemini and Apollo missions in the 1960s. Unfortunately this idea was met with heavy criticism from the Bureau of Budget (BOB) and the Department of Defense (DOD), since the BOB thought high-altitude aircraft would be better suited to the task and the DOD was concerned that earth resource satellites would jeopardise its military reconnaissance missions.

In 1966 NASA felt pressure from the Department of the Interior (DOI), after USGS convinced the DOI to announce that the DOI would be starting its own earth resource satellite programme. This forced NASA to accelerate the building of an earth resource satellite. Unfortunately, a limited budget and sensor disagreements between application agencies again delayed the satellite construction process. Finally, by 1970 NASA received authorisation to build a satellite. The first earth resource satellite, Landsat-1, was launched on July 23, 1972 by NASA; at that time the satellite was known as the Earth Resources Technology Satellite (ERTS) [33]. Landsat-1 carried the Return Beam Vidicon (RBV) and Multispectral Scanner (MSS) systems. Seven satellites were launched in the Landsat series, some of which are still functioning today. Landsat-4, launched on July 16, 1982, carried the Thematic Mapper (TM), another predecessor of MODIS [33].

A few other earth resource satellite programmes (started as part of the first generation) worth mentioning are [35]:

1. SEASAT: SEASAT was managed by NASA's Jet Propulsion Laboratory and was launched on June 27, 1978. SEASAT had on board the first spaceborne Synthetic Aperture Radar (SAR), but unfortunately only functioned for three months.

2. SPOT: SPOT is a high-resolution, optical imaging earth observation satellite programme managed by Spot Image based in Toulouse, France. SPOT-1 was launched on February 22, 1986.

3. IRS: IRS are a series of earth observation satellites, built, launched and maintained by ISRO. IRS-1A was launched on March 17, 1988.

4. JERS: JERS-1 was launched on February 11, 1992 and was administrated by the Japan Aerospace Exploration Agency (JAXA).

2.1.4.2 Second generation

In the early 1980s, there was a merger between the human spaceflight missions and the earth science missions of NASA, which was termed System Z. System Z fostered the idea of having one giant spacecraft carrying a variety of sophisticated earth observation sensors, including radar. System Z changed into the EOS in 1983, after scientists realised that multiple small missions would lead to better results [37, 38]. In the beginning the System Z earth observation system was to consist of two large (15-ton) platforms called EOS-A and EOS-B. After a reduction in size, the original sun-synchronous system EOS-A was renamed to EOS-Terra to emphasise its main function, namely to make land observations. EOS-Terra was launched on December 18, 1999. EOS-B was renamed to EOS-Aqua, as EOS-B would focus on ocean observation, and was launched on May 4, 2002. EOS-Terra carries the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER), the Multi-angle Imaging SpectroRadiometer (MISR), MODIS, the Measurements of Pollution in the Troposphere (MOPITT) and the the Clouds and the Earth's Radiant Energy System (CERES). EOS-Aqua carries the Advanced Microwave Scanning Radiometer-EOS (AMSR-E), the Atmospheric Infrared Sounder (AIRS), the Advanced Microwave Sounding Unit (AMSU), CERES, the Humidity Sounder for Brazil (HSB) and MODIS. Many EOS missions have been launched over the last decade [37, 38].

2.1.4.3 Third generation

During the mid 1990s a series of decisions were taken that NASA's EOS programme would only be regarded as a proof of concept and that NOAA would eventually be responsible for developing systems to study climate change. At about the same time the USA government decided to combine the low earth orbiting satellite programmes of NOAA and the DOD into the National Polar Orbiting Environmental Satellite Series (NPOESS). The responsibility of administrating NPOESS was assigned to the newly created Integrated Program Office (IPO) consisting of NASA/NOAA/DOD [37, 38]. The NPP satellite was originally developed by the IPO, until DOD participation in the project was dissolved. The NPOESS Preparatory Project (NPP) satellite is intended to bridge the gap between old (Terra) and new systems (still to be launched) and was launched on October 28, 2011. The NPP sa-

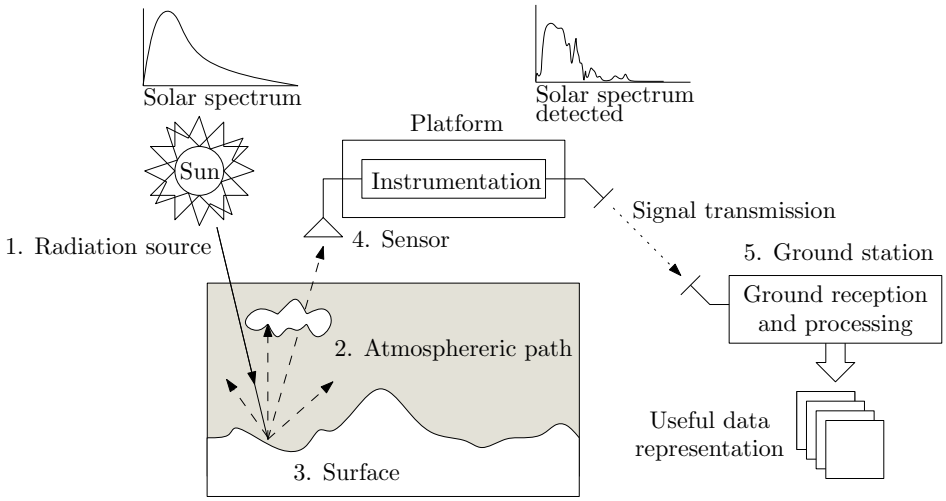


Figure 2.2: Signal and data flow in a typical remote sensing system (from [23]).

tellite carries five instruments, including the Advanced Technology Microwave Sounder (ATMS), the Cross-track Infrared Sounder (CrIS), CERES, the Visible Infrared Imager Radiometer Suite (VIIRS) and the Ozone Mapping and Profiler Suite (OMPS). MODIS is a predecessor of VIIRS. The Landsat Data Continuity Mission (LDCM) and Hyperspectral Infrared Imager (HyspIRI) are two new developments that will also form part of the third generation [37, 38].

2.2 A TYPICAL REMOTE SENSING SYSTEM

A general remote sensing platform is depicted in Figure 2.2 and consists of five main parts, namely the radiation source, the atmospheric path, the surface, the remote sensor and the ground reception station [39].

The sun is arguably the best known and most widely used source of electromagnetic energy and its energy is distributed throughout the electromagnetic spectrum. The electromagnetic energy from the sun travels through the atmosphere towards the surface of the earth. When the electromagnetic energy travels through the atmosphere, some of the energy is absorbed or scattered. The remaining energy arrives at the surface of the earth, where the energy is absorbed, reflected or transmitted. The absorbed energy can be re-emitted at a different wavelength. The reflected and emitted wavelengths travel back through the atmosphere towards the remote sensing sensor, where it is finally recorded. The atmosphere again absorbs and scatters some of the reflected and emitted energy. In the last step the recorded data are sent to a ground station where the data are processed to create useful

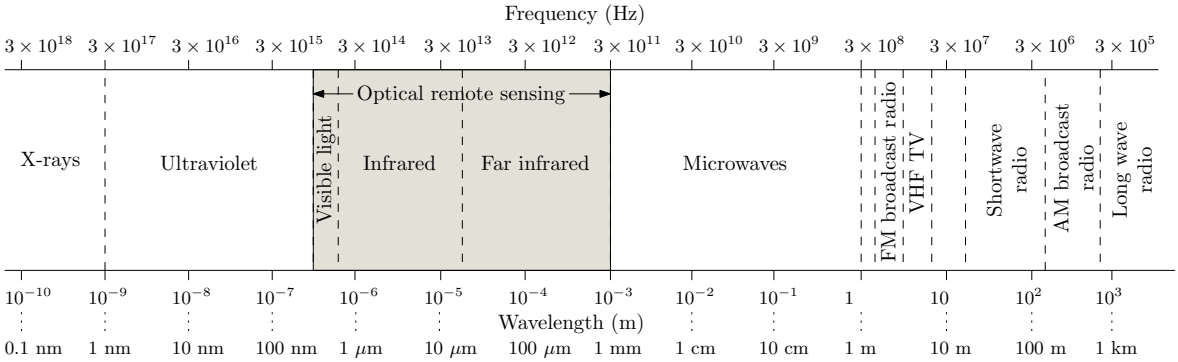


Figure 2.3: The electromagnetic spectrum, showing the region of interest for optical remote sensing (from [23]).

information [39].

2.3 ELECTROMAGNETIC RADIATION

The most important principles of electromagnetic radiation are discussed in this section. The section starts by introducing the electromagnetic spectrum and is followed by a section that explains how electromagnetic radiation propagates. Radiation units are discussed in Section 2.3.3, while Section 2.3.4 explains what blackbody radiation is.

2.3.1 Electromagnetic spectrum

In Figure 2.3 a segment of the electromagnetic spectrum is shown. Electromagnetic spectrum divisions were created for convenience and by tradition for each discipline and is therefore defined differently in other sources [34]. The ultraviolet, visible, infrared and microwave regions are usually used for remote sensing purposes.

Near ultraviolet radiation is known for its ability to induce fluorescence, emission of visible radiation, in some materials. Unfortunately ultraviolet radiation is severely scattered by the atmosphere and therefore not used very often in a remote sensing context [34].

The visible and infrared region together form the optical region [23]. The optical region is usually divided further into smaller regions. However more than one division are, used in literature, as shown in Table 2.1. The near infrared and mid-infrared regions are close to the visible region and have similar characteristics to visible light, and for this reason can be recorded via films, filters and cameras. The

far infrared region is reasonably far removed from the visible region. In everyday terminology this region is known as the thermal region, consisting of "heat" [34].

Table 2.1: Some common optical regions of the electromagnetic spectrum.

	Region	Wavelength (μm)
Visible	Blue	0.4 – 0.5
	Green	0.5–0.6
	Red	0.6–0.7
Infrared [23]	Near IR	0.7–1.4
	Short-wave IR	1.4–3.0
	Mid-wave IR	3.0–8.0
	Long-wave IR	8.0–15.0
	Far IR	15.0–1000
Infrared [35]	Photographic IR	0.7–0.9
	Very near IR	0.7–1.0
	Reflected IR	0.7–3.0
	Near IR	0.7–3.0
	Thermal IR	3.0–1000

The microwave region is usually used in an active remote sensing system.

2.3.2 Propagation of electromagnetic radiation

Electromagnetic radiation is produced through several means, including changes in the energy levels of electrons, acceleration of electrical charges, decay of radioactive substances and the thermal motion of atoms and molecules [34]. Electromagnetic radiation propagates by means of a transverse wave. Electromagnetic radiation consists of a perpendicular electric (E) and a magnetic field (H) that increase and decrease in phase with each other [35]. Transverse waves have a few important properties, namely [34]:

1. Wavelength (λ) is the distance between two successive peaks. Wavelength can be measured in everyday units of length, but the wavelengths of the electromagnetic radiation that is relevant to most remote sensing sensors are so short that less known units are usually employed, which

include the micrometre (μm: 10^{-6}) and the nanometre (nm: 10^{-9}).

2. Amplitude (A) is equal to the height of each peak. Amplitude is often measured as spectral irradiance (W.m^{-2}.μm^{-1}), expressed as watts per square metre per micrometre (as energy levels per wavelength interval).

3. Frequency (f) is measured in Hz and is defined as the number of crests passing a fixed point in a second.

All matter above 0 K produces electromagnetic radiation, and all electromagnetic radiation travels at the speed of light, $c = 2.99893 \times 10^8$ ms^{-1}. Because all electromagnetic radiation travels at the same speed, an inverse relation exists, between the wavelength and frequency of electromagnetic radiation, which is expressed mathematically as [35]

$$c = \lambda f. \tag{2.1}$$

In Equation 2.1, λ is measured in m and f is measured in Hz. Equation 2.1 explains why a transverse wave with a longer wavelength has a lower frequency when compared to a transverse wave with a shorter wavelength wave.

2.3.3 Radiation units

Although many electromagnetic radiation characteristics can be explained eloquently through wave theory, another theory offers useful insights when describing how electromagnetic energy reacts with matter. This theory, called the particle theory, states that electromagnetic radiation is absorbed and emitted in units called photons or quanta. The energy of a quantum is given by [33]

$$Q = hf,$$

where Q represents the energy of a quantum in joules (J), h is Planck's constant equal to 6.626×10^{-34} J.s and f represents frequency measured in Hz.

The rate $\frac{dQ}{dt}$ at which photons strike a surface is called radiant flux Φ measured in watts (W). Radiant exitance (M) and irradiance (E) are defined as $\frac{d\Phi}{dA}$, where A denotes area measured in m^2. The difference between radiant exitance and irradiance is that radiant exitance refers to the rate at which photons are emitted from a unit area, while irradiance refers to the rate at which photons strike a unit area. Spectral radiant exitance (M_λ) and spectral irradiance (E_λ) differ from M and E in that they des-

cribe how the energy is distributed with respect to wavelength across the electromagnetic spectrum and is therefore defined as $\frac{dM}{d\lambda}$ and $\frac{dE}{d\lambda}$ respectively and measured in W.m^{-2}.μm^{-1} [39].

The radiometric units introduced up to this stage, take into account energy, time, wavelength and area. A variable that is still unaccounted for is the viewing angle and radiance L thus takes into account the viewing angle and is defined as

$$L = \frac{d^2\Phi}{dAd\Omega\cos\theta},$$

where L is the observed or measured radiance (W.m^{-2}.sr^{-1}) in the direction θ and Ω is the solid angle (sr) subtended by the observation or measurement. As in the case of radiant exitance and irradiance, radiance also has a spectral counterpart called spectral radiance $L_\lambda = \frac{dL}{d\lambda}$ [39].

2.3.4 Blackbody radiation

A blackbody is an ideal body which, if it existed, would be a perfect absorber and a perfect radiator, absorbing all incident radiation, reflecting none, and emitting radiation at all wavelengths [39]. In remote sensing, the exitance curves of blackbodies at various temperatures can be used to model naturally occurring phenomena such as solar radiation and terrestrial emittance. The spectral radiant exitance M_λ (measured in W.m^{-2}.μm^{-1}) of a blackbody for different temperatures is described through Planck's law [39]

$$M_\lambda = \frac{\varepsilon c_1}{\lambda^5(e^{c_2/\lambda T} - 1)}, \tag{2.2}$$

where ε is emittance (dimensionless), c_1 is the first radiation constant and is equal to 3.7413×10^8 W.μm^4.m^{-2}, λ is radiation wavelength with units μm, c_2 is the second radiation constant, which is equal to 1.4388×10^4 μm.K, and T is the absolute radiant temperature in K.

Emittance (emissivity) is the ratio of the radiation given off by a surface to the radiation given off by a blackbody at the same temperature; a blackbody has an emissivity of 1, while a whitebody (perfect reflector) has an emissivity of 0. All other objects (greybodies) have an emissivity between 0 and 1 [39].

Alternatively, Planck's law can be described in terms of the radiation frequency f by using the follo-

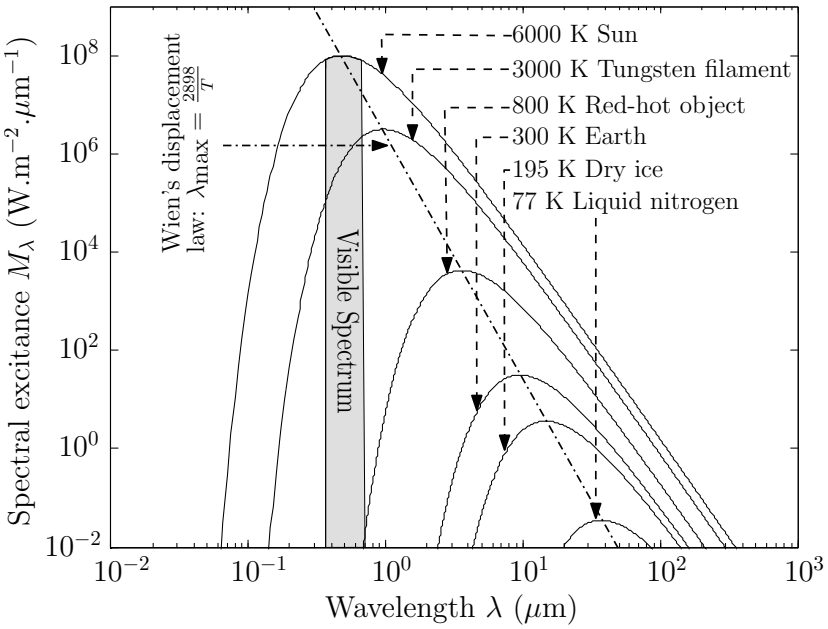


Figure 2.4: Blackbody radiation at various temperatures

wing substitutions

$$M_\lambda d\lambda = -M_f df$$
$$\lambda = \frac{c}{f}$$
$$\frac{d\lambda}{df} = -\frac{c}{f^2}.$$

After the above substitutions have been made, Equation 2.2 changes to

$$M_f = \frac{\varepsilon c_1 f^3}{c^4(e^{c_2 f/cT} - 1)}. \tag{2.3}$$

If Equation 2.3 is integrated over all frequencies, the radiant exitance M (measured in W.m^{-2}) will be obtained for a blackbody. That is

$$M = \int_0^\infty M_f df = \int_0^\infty \frac{\varepsilon c_1 f^3}{c^4(e^{c_2 f/cT} - 1)} df. \tag{2.4}$$

If $x = \frac{c_2 f}{cT}$ and $dx = \frac{c_2}{cT} df$ are substituted into Equation 2.4 the following is obtained

$$M = \frac{\varepsilon c_1 T^4}{c_2^4} \int_0^\infty \frac{x^3}{e^x - 1} dx \tag{2.5}$$

$$= \frac{\varepsilon c_1 T^4 \zeta(4)\Gamma(4)}{c_2^4} \tag{2.6}$$

$$= \frac{\varepsilon c_1 \pi^4}{15 c_2^4} T^4$$

$$= \varepsilon \sigma T^4, \tag{2.7}$$

where σ is the Stefan-Boltzmann radiation constant, which is equal to 5.6693×10^{-8} W.m^{-2}.K^{-4} and T is absolute temperature measured in K [39].

Equation 2.5 and Equation 2.6 are equal, since it is a well-known fact that $\int_0^\infty \frac{x^{n-1}}{e^x-1} dx$ is equal to the product $\zeta(n)\Gamma(n)$ for all $n \in \mathbb{N}$, where $\zeta(n) = \sum_i^\infty \frac{1}{i^n}$ is the well-known Riemann zeta function and $\Gamma(n) = (n-1)!$ is the gamma function [39, 40]. Equation 2.7 is known as the Stefan-Boltzmann radiation law, which states that the total radiation emitted from a blackbody is proportional to the fourth power of its absolute temperature [34].

If Equation 2.2 is differentiated with respect to wavelength, the derivative is set to 0, and the resulting equation solved $\lambda_{\max}$ (measured in μm) is obtained, which is the wavelength at which maximum emittance occurs for a given absolute temperature [39]. The result of this procedure is Wien's displacement law [34]

$$\lambda_{\max} = \frac{2898}{T}. \tag{2.8}$$

In Equation 2.8 the constant 2898 is measured in μm.K, while T represents absolute temperature, which is measured in K. Wien's displacement law states that the wavelength of maximum emittance for a blackbody is inversely proportional to absolute temperature [34].

2.4 ATMOSPHERIC INTERACTIONS

The atmospheric path is a critical component of any remote sensing system. When electromagnetic radiation travels through the atmosphere a lot of scattering and absorption takes place. Scattering alters the direction in which the electromagnetic radiation propagates, while absorption leads to attenuation in signal strength. When designing remote sensing systems it is of great importance to keep the effect of scattering and absorption in mind, as it would serve no purpose to record electromagnetic radiation in an atmospheric absorption window. Absorption and scattering are caused by particles and

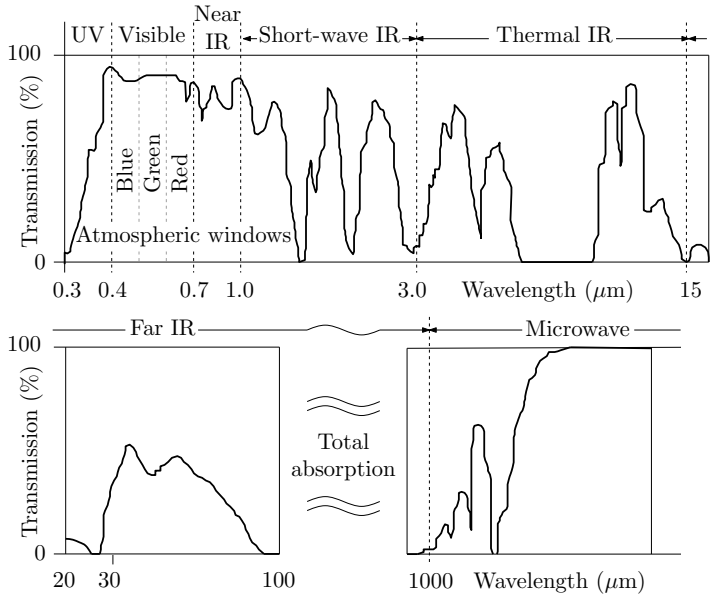


Figure 2.5: Atmospheric electromagnetic transmission windows (from [23]).

gases contained in the atmosphere.

2.4.1 Atmospheric absorption

As photons collide with atmospheric molecules, some of the radiation is absorbed through electron orbital transitions and induced vibrations, heating up the atmosphere. Nitrogen, oxygen, carbon dioxide, ozone and water vapour all absorb electromagnetic radiation at different wavelengths. The set of frequencies that a gaseous mixture can absorb consists of the union of all the the frequencies that the constituent gases of the gaseous mixture can absorb. The atmosphere contains nitrogen, oxygen, carbon dioxide, ozone and water vapour, therefore the net effect of this gaseous mixture in the atmosphere is atmospheric absorption windows. The parts of the spectrum that are not affected heavily by absorption (in which the transmission of electromagnetic radiation is high) are known as atmospheric transmission windows [23]. The atmospheric transmission windows are displayed in Figure 2.5.

2.4.2 Atmospheric scattering

Scattering is mainly caused when electromagnetic energy is redirected from its original propagation path via particulates or large gas molecules [23]. There are three basic types of scattering taking place in the atmosphere that affect electromagnetic radiation, namely *Rayleigh*, *Mie* and *nonselective*

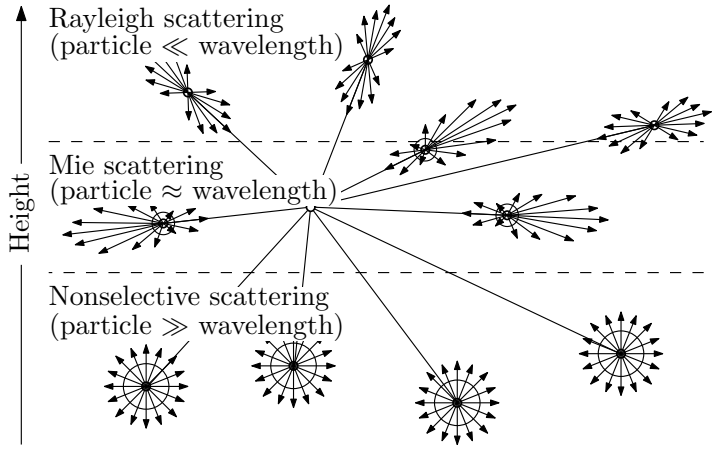


Figure 2.6: Atmospheric scattering (from [23]).

scattering, as shown in Figure 2.6.

2.4.2.1 Rayleigh scattering

Rayleigh scattering takes place at high altitudes, where the radiation wavelengths are much larger than the size of the particulates. In Rayleigh scattering, the volume-scattering coefficient σ_λ (with units cm^{-1}) is given by

$$\sigma_\lambda = \frac{4\pi^2 NV^2}{\lambda^4}\frac{(n^2-n_0^2)^2}{(n^2+n_0^2)^2}, \tag{2.9}$$

where N is the number of particles per cm^3, V is the volume of scattering particles (cm^3), λ is the radiation wavelength measured in cm, n is the refractive index of particles and n_0 is the refractive index of the medium. From Equation 2.9 it is clear that the scattering coefficient is proportional to the inverse fourth power of wavelength and this causes the shorter blue wavelengths to be scattered toward the ground much better than the longer red wavelengths, which makes the sky appear blue. As the sun approaches the horizon, the rays of the sun follow a longer path through the atmosphere, which in turn leads to an increase in blue wavelength scattering, leaving only the red wavelengths to reach the human eye (making a sunset appear orange/red) [39]. The primary components responsible for scattering at these altitudes include atmospheric gases such as oxygen and nitrogen or tiny specks of dust. Rayleigh scattering is symmetrical, with equal amounts of forwardscatter and backscatter [23].

2.4.2.2 Mie scattering

Mie scattering occurs closer to the ground than Rayleigh scattering, where the diameter of the particulates is about the same as the wavelength of radiation. For the most universal situation, in which there is a continuous particle-size distribution, the Mie scattering coefficient σ_λ (with units km^{-1}) is given by the following relationship

$$\sigma_\lambda = 10^5 \pi \int_{a_1}^{a_2} N(a) K(a,n) a^2 da,$$

where $N(a)$ is the number of particles in the interval a to $a+da$ (cm^{-3}), $K(a,n)$ is the scattering coefficient (cross-section measured in cm^{-1}), a represents the radius of the spherical particles (in cm) and n is the index of the refraction of particles [39]. Aerosols, dust particles, pollen, smoke and water vapour are the main causes of Mie scattering. Mie scattering is not as dependent on wavelength as Rayleigh scattering and mainly affects the visible spectrum. As can be seen from Figure 2.6, Mie scattering mainly causes forward scattering [23].

2.4.2.3 Nonselective scattering

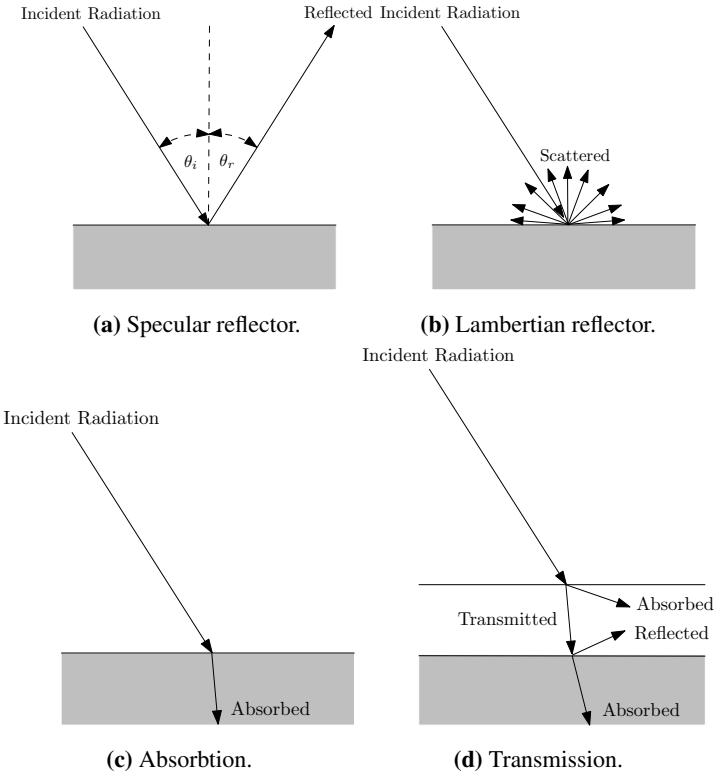
Nonselective scattering occurs at low altitudes, where the particles are usually much larger than the wavelength of radiation. Nonselective scattering scatters electromagnetic radiation uniformly and is not really dependent on wavelength. This kind of scattering is caused by large particulates such as dust, water droplets, ice crystals and hail. Since nonselective scattering scatters electromagnetic radiation uniformly, it is responsible for the fact that clouds appear white [23].

2.5 SURFACE INTERACTION

The previous section focused on the scattering and absorption effects of the atmosphere. In this section a closer look is taken at what happens to electromagnetic radiation when it reaches the earth's surface.

2.5.1 Reflection, absorption and transmission

When electromagnetic energy reaches the earth's surface, the incident radiation may be absorbed, reflected or transmitted [35]. The three phenomena introduced above are displayed in Figure. 2.7.



(a) Specular reflector. **(b)** Lambertian reflector.

(c) Absorbtion. **(d)** Transmission.

Figure 2.7: Interaction of electromagnetic radiation with a surface (from [10]).

Atoms and molecules contain electrons. To explain the concept of absorption, it is useful to imagine that the electrons are attached to the atoms via springs. The electrons vibrate continuously at a specific frequency, called the natural frequency. When an electromagnetic wave impedes on an atom that is vibrating at the same frequency as the frequency of the incident wave, the energy is absorbed by the atom via the resonance principle. Absorption can also occur due to electron orbital transitions and is not limited to vibration inducement. The absorbed radiation can then be re-emitted at a different wavelength. Reflection and transmission of electromagnetic waves occur because the frequencies of the incident waves do not match the natural frequencies of the objects. During reflection the atoms start vibrating for a short while, after which the energy is simply re-emitted at the same wavelength as the incident wave. In the case of transmission the radiation is passed on through the bulk of the material and emitted on the other side of the material at the same wavelength as the incident wave.

The roughness of the surface determines the type of reflection that will occur. A very smooth surface acts as a specular reflector, for which the reflection angle, θ_r, equals the incidence angle, θ_i. A very rough surface acts like a Lambertian reflector (diffuse reflector), which scatters the incident radiation

uniformly in all directions [35]. It should be clear that for remote sensing purposes a Lambertian reflector is preferred, since specular reflectors usually appear dark from most angles, as the incident radiation is not reflected uniformly [23].

2.5.2 Albedo

Absorption, reflectance and transmission are related via

$$E_I(\lambda) = E_A(\lambda) + E_R(\lambda) + E_T(\lambda), \tag{2.10}$$

due to the principle of conservation of energy. Where E_I denotes the incident radiation, E_A denotes the absorbed radiation, E_R denotes the reflected radiation and E_T denotes the transmitted radiation, with all energy components being a function of wavelength. Equation 2.10 states that all incident radiation is absorbed, reflected or transmitted. The albedo (spectral reflectance) of a surface is given by the ratio of the electromagnetic radiation reflected from a surface to the total electromagnetic radiation incident on the surface and is expressed mathematically as [33]

$$\rho(\lambda) = \frac{E_R(\lambda)}{E_I(\lambda)}.$$

2.5.3 Bidirectional Reflectance Distribution Function

There is a function that can describe the scattering characteristics of a surface much better than albedo can, namely the BRDF.

BRDF is usually denoted by the symbol f with units sr^{-1} and defined as

$$f(\theta, \phi, \theta', \phi') = \frac{dL'(\theta', \phi')}{dE(\theta, \phi)},$$

where dE is the irradiance (units W.m^{-2}), dL' is the reflected radiance (units $\text{W.m}^{-2}\text{.sr}^{-1}$), θ is the zenith angle of the radiation source, ϕ is the azimuthal angle of the radiation source and the primed angles refer to the location of the sensor. The relationship between irradiance dE and incident radiance dL is expressed mathematically as

$$\begin{aligned} dE(\theta, \phi) &= L(\theta, \phi)\cos(\theta)d\omega \\ &= L(\theta, \phi)\cos(\theta)\sin(\theta)d\theta d\phi \end{aligned}$$

where $d\omega$ is the solid angle defined as $\sin(\theta)d\theta d\phi$ [39].

2.5.4 Spectral signature of vegetation, soil and water

A graph of the spectral reflectance (albedo) of an object as a function of wavelength is termed a spectral reflectance curve (spectral signature) of the object [33]. Different types of objects have different spectral signatures and spectral signatures can therefore be used for classification. The spectral signature for three diverse types of objects are displayed in Figure 2.8.

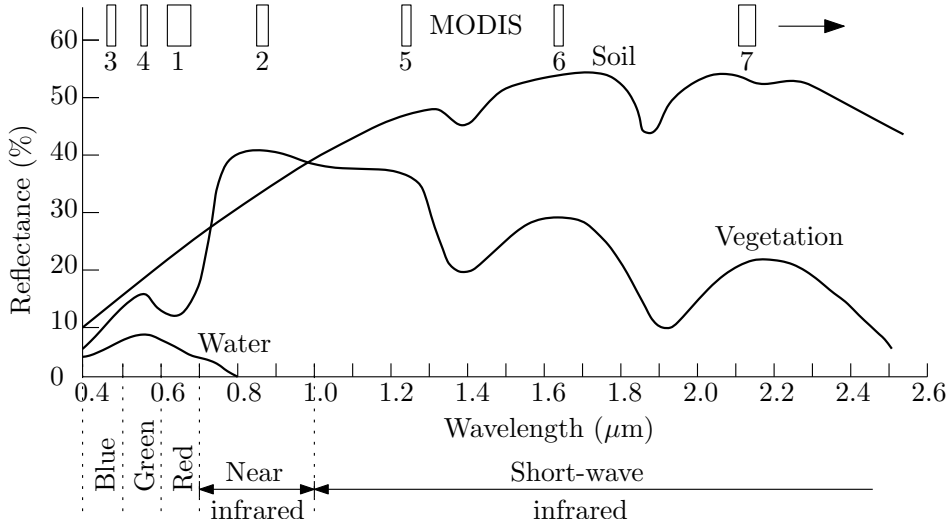


Figure 2.8: Spectral reflectance charachteristics of common earth surface materials in the visible and near-to-mid-infrared range. The positions of the MODIS spectral bands are also indicated (from [23]).

2.5.4.1 Vegetation

The spectral signature of lush green vegetation is characterised by a "peak-valley" configuration [33]. Vegetation appears green, since chlorophyll strongly absorbs radiation in the blue and red bands, while heavily reflecting radiation from the green band. As depicted in Figure 2.9, chlorophyll is a green pigment, which is contained in sacs called chloroplasts [35]. The peak in the spectral signature of the green band is clearly visible in Figure 2.8. Most of the Near-Infrared (NIR) radiation reaches the leaf's spongy mesophyll tissue, where 40 to 50 percent of the NIR radiation is reflected. The reflection caused by the spongy mesophyll tissue is responsible for the fairly flat spectral signature found in Figure 2.8 between 0.7 μm and 1.3 μm [33, 35]. The cell structures of different vegetation types vary a lot, which leads to discernible NIR reflection patterns [39]. The remaining NIR radiation is transmitted. Multiple layers of leaves in a plant canopy provide the opportunity for hierarchical layers of transmittance and reflectance. Hence the infrared reflectance increases with the number

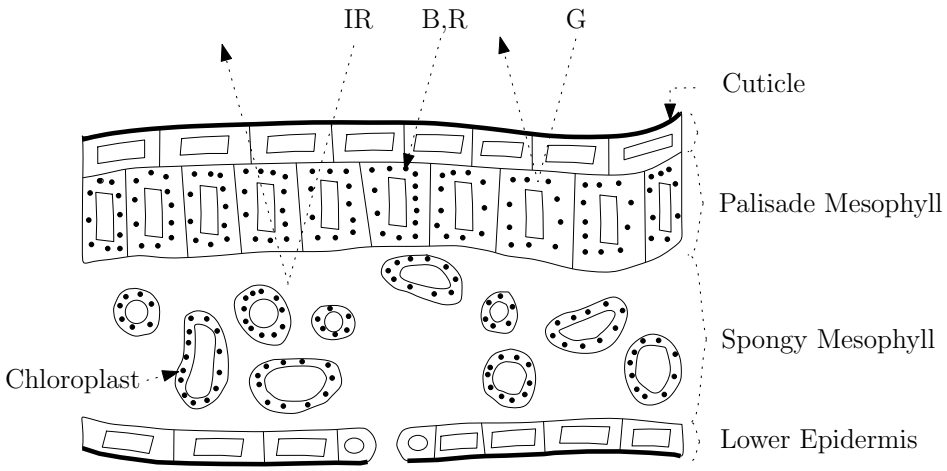


Figure 2.9: Generalised cross-section showing the cell structure of a green leaf.

of layers of leaves in a canopy [33]. Healthy vegetation consists mainly of water and the spectral signature of vegetation therefore also contains water absorption bands at 1.4μm, 1.9 μm and 2.7 μm which, are located in the Short-wave Infrared (SWIR) region.

Spectral ratioing (or image differencing) is a popular transform thst is used on remotely sensed data to enhance the interpretability of the data. Normalised Difference Vegetation Index (NDVI) is an example of such a spectral ratioing technique and for the MODIS sensor is calculated with

$$\text{NDVI} = \frac{(\text{Band 2}) - (\text{Band 1})}{(\text{Band 2}) + (\text{Band 1})}. \tag{2.11}$$

It can clearly be seen from Figure 2.8, that vegetation reflects much less radiation in the red band (MODIS 1) than in the NIR band (MODIS band 2) and vegetation will consequently have a large NDVI value. Soil will have a lower NDVI value, since there is not much difference between the amounts of radiation that are reflected in the red and NIR bands (in the case of soil). NDVI can be used to identify vegetative areas [23].

2.5.4.2 Soil

The soil curve in Figure 2.8 does not have a "peak-valley" appearance. Some of the factors affecting soil reflectance include moisture content, soil texture (proportion of sand, silt or clay), surface roughness, presence of iron oxide and organic matter content [33]. Moist soil has a lower reflectance if compared to dry soil. As with vegetation, the effect of moistness (water content) on reflectance is amplified in the water absorption bands. The soil texture influences the soil's capability of retaining water. Clay particles are smaller than those of silt, which in turn are smaller than those of sand. Sand

is thus more porous when compared to clay. Clay can retain water the best, while sand has the lowest retention capability due to its porous nature (which is caused by the size of the sand granules) [39]. The surface roughness of a soil and the presence of iron oxide and organic matter in a soil will also significantly decrease its reflectance capability [33].

2.5.4.3 Water

As can be seen from Figure 2.8, most of the radiation incident upon water is either absorbed or transmitted. The longer the wavelength of incident radiation, the better it is absorbed by water, and therefore water appears blue-green in the visible spectrum, and dark in the infrared range. Suspended sediments or shallow water bodies may cause increased reflection. The increased reflectance may even be observable in the NIR region [23].

2.6 REMOTE SENSING PLATFORMS

Different remote sensing platforms and systems are discussed in this section. The section concludes with the resolution of remote sensing sensors.

2.6.1 Ground-based, airborne and spaceborne platforms

There are three main types of remote sensing platforms, namely *ground-based*, *airborne* and *space-borne* platforms.

Ground-based sensors can usually only work on a small scale and are thus normally used for generating ground truth data. Balloons, aircraft and more recently Unmanned Aerial Vehicles (UAVs) are all airborne platforms. Satellites are the primary platforms used to host spaceborne sensors [23]. Active and passive remote sensing systems can be found on both airborne and spaceborne platforms.

2.6.2 Passive and active remote sensing systems

Remote sensing systems can be grouped into two mayor system categories, named *active* and *passive* systems. In a passive remote sensing system the sun is used as the source of electromagnetic radiation, while in an active system, such as radar, the system produces its own radiation. The difference between a passive and an active system is illustrated in Figure 2.10. Most active radiation systems

(radars) produce radiation in the spectrum bands where the sun does not radiate with high intensity, such as the microwave region. A passive sensor can also be designed to measure the thermal radiation produced by the earth, so that the earth becomes the radiation source of the passive system instead of the sun.

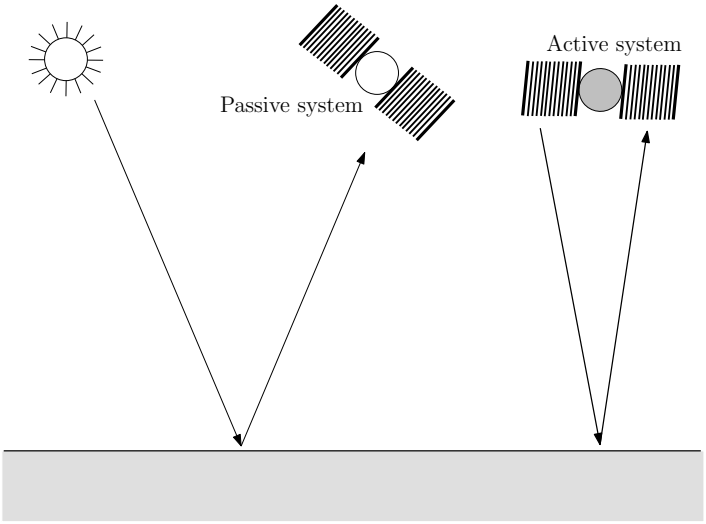


Figure 2.10: Active and passive remote sensing system.

Two main acquisition scanners are employed by passive remote sensing systems, namely the transverse scanner and the *pushbroom* scanner. A transverse (also termed across-track or *whiskbroom*) scanner is an electro-mechanical device that obtains data from narrow swaths of terrain (by using a scanning mirror), which are at right angles to the direction of movement. The scanning mirror sweeps across the satellite's ground track. The scanning mirror directs reflected (or emitted) radiation towards the on-board detectors. A pushbroom scanning system does not rely on a scanning mirror to direct radiation onto a detector, but instead employs a linear array of detectors. Each detector in the array measures the radiation reflected from a small area on the ground, which is known as a ground resolution cell [35].

2.6.3 Resolution of remote sensing sensors

One way of comparing different remote sensing sensors with one another is to compare their resolution. The resolution of a remote sensing sensor can be divided into four categories, namely its *spectral*, *spatial*, *temporal* and *radiometric* resolution [23].

2.6.3.1 Spectral resolution

Most remote sensing systems record data from different spectral bands. The width of these spectral bands is known as the spectral resolution of the sensor. If these spectral bands are not small (low resolution) and there are large gaps between them, the sensor is called a multispectral sensor; on the other hand, if the resolution is high and there are almost no gaps between bands, the sensor is known as a hyperspectral sensor [23]. The difference between multispectral and hyperspectral sensors is illustrated in Figure 2.11.

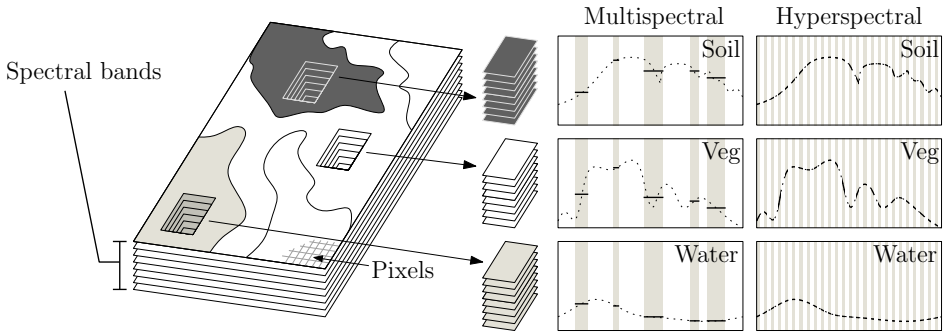


Figure 2.11: Interaction of reflected light with surface materials, showing multispectral and hyperspectral signatures (from [23]).

2.6.3.2 Spatial resolution

The spatial resolution depends on the Instantaneous Field of View (IFOV), which is the angular cone of visibility that describes the surface area from which radiation is recorded by the sensor at any instant in time.

Increasing the spatial resolution of a sensor can also decrease its spectral resolution. If the sensor's IFOV is made narrower (which increases its spatial resolution) the area which is monitored by the sensor is decreased. The smaller surveillance area means less energy reaches the sensor, which means that the Signal-to-Noise Ratio (SNR) of the sensor will also be lower. To compensate for this the scanning bandwidth must be increased, which in turn lowers the spectral resolution of the sensor [23].

2.6.3.3 Temporal resolution

The temporal resolution of a sensor refers to the time interval it needs to make successive measurements of the same physical location. The temporal resolution is actually dependent on the remote sensing platform carrying the sensor. There is normally a trade-off between the temporal and spatial resolution of a sensor. The higher the spatial resolution the lower the temporal resolution. This trade-off is due to the sensor's swath width. If the sensor has a large swath width, then the revisit time to a specific location will be shorter. The revisit time will be shorter, because the entire earth can be surveyed faster [23].

2.6.3.4 Radiometric resolution

The radiometric resolution of a sensor is measured in bits. The number of bits that a sensor has available to record values determines the number of unique levels of radiation it can measure. For example, MODIS, has a radiometric resolution of 12 bits, thus MODIS can detect $2^{12} = 4096$ unique levels of radiation.

Remote sensing data are commonly expressed with Digital Numbers (DNs) ranging from 0 to $2^b - 1$, where b is the radiometric resolution of the sensor, in bits [23].

2.6.4 Signal-to-noise ratio

Another important characteristic of a remote sensing sensor is its SNR. The SNR is defined as the the energy contained in the received signal divided by the energy in the noise that is generated by aberrations in the electronics. It is desirable to have a sensor with a high SNR [35].

2.7 MODERATE RESOLUTION IMAGING SPECTRORADIOMETER

In this section the MODIS sensor is discussed in detail. A literature review of land cover mapping applications with MODIS data can be found in [41]. The MODIS product selected for this thesis is discussed in Section 2.7.5.

2.7.1 History of MODIS

In 1983 NASA created the EOS Science and Mission Requirements Working Group (EOSMRWG) with the explicit purpose of developing a global concept for EOS. The EOSMRWG report delivered in 1984 called for several instruments to survey the earth. In 1984 NASA formed an instrument panel for each facility sensor proposed by the EOSMRWG report. The MODIS instrument panel suggested the development of two sensors, the MODIS-T and the MODIS-N. The original MODIS-N was a conventional imaging filtered radiometer capable of surveying 35 spectral bands, while MODIS-T was supposed to be a 64-band imagining spectroradiometer with the ability to tilt to avoid sun-glint from the oceans. The management and development of the MODIS sensor was assigned to Goddard Space Flight Center (GSFC), where it was decided to develop the MODIS-T sensor in-house and to outsource the MODIS-N sensor. In 1991 the Hughes/Santa Barbara Research Center (SBRC) was assigned the contract to build the MODIS-N sensor. Soon after the SBRC contract started, major restructuring of EOS occurred, which led to a decision to keep the MODIS-N design and to scrap the MODIS-T design. Over the next few years the SBRC developed and fabricated two MODIS flight models. The first of these models was installed in EOS-Terra, which was subsequently launched on December 18, 1999. The second model was installed in EOS-Aqua, which was launched on May 4, 2002 [37, 38].

2.7.2 MODIS sensor characteristics

Both EOS-Terra and EOS-Aqua are polar-orbiting sun-synchronous platforms. The orbital height of the EOS platforms is 705 km at the equator. Each MODIS instrument has a two-sided scan mirror with a maximum scan angle of 55° at either side of nadir, providing a nominal swath width of 2330 km. Because of to the large swath width, the MODIS sensor surveys the earth every one to two days [42, 43]. The predecessors of MODIS are NOAA's AVHRR and Landsat's TM. MODIS was compared to AVHRR in Section 1.3. Although the TM sensor provides a higher spatial resolution than MODIS, the TM sensor is charachterised by incomplete spatial coverage, low temporal resolution and cloud contamination [23]. A total of 36 spectral bands are surveyed by MODIS inside the spectral region 0.412-14.235 μm. The first two bands are located in the Red (R) (0.648 μm) and NIR (0.858 μm) regions and have a spatial resolution of 250 m. The next five MODIS bands (bands 3-7: 0.470 μm, 0.555 μm, 1.240 μm, 1.640 μm and 2.13 μm) have a spatial resolution of 500 m and are located in the visible to the SWIR regions. The remaining 29 bands (bands 8–36) have a spatial resolution

of 1000 m and are located in the the Mid-wave Infrared (MWIR) and Long-wave Infrared (LWIR) regions. The MODIS instrument has a 12-bit radiometric resolution [42]. The characteristics of the 36 spectral bands of MODIS are displayed in Table 2.2 and Table 2.3. Additional MODIS characteristics can be found in Table 2.4.

2.7.3 MODIS products

The data captured by the MODIS sensor can be subjected to a few levels of processing [42]:

- Level 0: The initial data set, which is automatically derived from the instrumental raw data.
- Level 1A: Contains geodetic information.
- Level 1B: Calibrated radiances for all bands and surface reflectance values for selective bands.
- Level 2: Derived geophysical variables at the same resolution and location as level 1 source data (swath products).
- Level 2G: Level 2 data mapped on a uniform space-time grid scale (sinusoidal).
- Level 3: Gridded variables in derived spatial and/or temporal resolutions.
- Level 4: Model output or results from analyses of lower-level data.

The raw MODIS data are transferred to ground stations in White Sands, New Mexico, via the Tracking and Data Relay Satellite System (TDRSS). The raw data are then forwarded to the EOS Data and Operations System (EDOS) at GSFC, where level 0 processing takes place. Level 1A and level 1B data are generated by GSFC Earth Sciences DAAC (GES DAAC). Higher-level MODIS land and atmosphere products are produced by the MODIS Adaptive Processing System (MODAPS), and distributed by three Distributed Active Archive Centers (DAACs), namely the L1 and Atmosphere Archive and Distribution System (LAADS), the Land Processes DAAC (LP DAAC) and the National Snow and Ice Data Center DAAC (NSIDC DAAC). Ocean colour products are produced and distributed by the Ocean Color Data Processing System (OCDPS) [37, 38].

There are close to 40 MODIS products available. Most MODIS product names start with three specific letters, which may be MOD, MYD or MCD. MOD indicates that the product was derived using only

Table 2.2: A summary of MODIS spectral bands 1-27.

Band	Wavelength (μm)	IFOV (m) [at nadir]	Primary use	Spectral region
Band 1	0.62–0.67	250×250	Land/Cloud/Aerosols Boundaries	R
Band 2	0.841–0.876	250×250	Land/Cloud/Aerosols Boundaries	NIR
Band 3	0.459–0.479	500×500	Land/Cloud/Aerosols Properties	B
Band 4	0.545–0.565	500×500	Land/Cloud/Aerosols Properties	G
Band 5	1.230–1.250	500×500	Land/Cloud/Aerosols Properties	SWIR
Band 6	1.628–1.652	500×500	Land/Cloud/Aerosols Properties	SWIR
Band 7	2.105–2.155	500×500	Land/Cloud/Aerosols Properties	SWIR
Band 8	0.405–0.420	1000×1000	Ocean Colour/Phytoplankton/ Biogeochemistry	B
Band 9	0.438–0.448	1000×1000		B
Band 10	0.483–0.493	1000×1000		B
Band 11	0.526–0.536	1000×1000		G
Band 12	0.546–0.556	1000×1000		G
Band 13	0.662–0.672	1000×1000		R
Band 14	0.673–0.683	1000×1000		R
Band 15	0.743–0.753	1000×1000		NIR
Band 16	0.862–0.877	1000×1000		NIR
Band 17	0.890–0.920	1000×1000	Atmospheric Water Vapour	NIR
Band 18	0.931–0.941	1000×1000	Atmospheric Water Vapour	NIR
Band 19	0.915–0.965	1000×1000	Atmospheric Water Vapour	NIR
Band 20	3.660–3.840	1000×1000	Surface/Cloud Temperature	MWIR
Band 21	3.929–3.989	1000×1000	Surface/Cloud Temperature	MWIR
Band 22	3.929–3.989	1000×1000	Surface/Cloud Temperature	MWIR
Band 23	4.020–4.080	1000×1000	Surface/Cloud Temperature	MWIR
Band 24	4.433–4.498	1000×1000	Atmospheric Temperature	MWIR
Band 25	4.482–4.549	1000×1000	Atmospheric Temperature	MWIR
Band 26	1.360–1.390	1000×1000	Cirrus Clouds Water Vapour	NIR
Band 27	6.535–6.895	1000×1000	Cirrus Clouds Water Vapour	MWIR

Table 2.3: A summary of MODIS spectral bands 28-36.

Band	Wavelength (μm)	IFOV (m) [at nadir]	Primary use	Spectral region
Band 28	7.175–7.475	1000×1000	Cirrus Clouds Water Vapour	LWIR
Band 29	8.400–8.700	1000×1000	Cloud Properties	LWIR
Band 30	9.580–9.880	1000×1000	Ozone	LWIR
Band 31	10.780–11.280	1000×1000	Surface/Cloud Temperature	LWIR
Band 32	11.770–12.270	1000×1000	Surface/Cloud Temperature	LWIR
Band 33	13.185–13.485	1000×1000	Cloud Top	LWIR
Band 34	13.485–13.785	1000×1000	Cloud Top	LWIR
Band 35	13.785–14.085	1000×1000	Cloud Top	LWIR
Band 36	14.085–14.385	1000×1000	Cloud Top	LWIR

Table 2.4: MODIS Design Specifications.

Orbit	705 km, 10:30 AM descending node or 1:30 PM ascending node, sun-synchronous, near polar, circular
Scan rate	20.3 rpm, cross track
Swath dimension	2330 km (cross track) by 10 km (along track at nadir)
Telescope	17.78 cm off-axis, a focal (collimated), with intermediately held stop
Size	$1.0 \times 1.6 \times 1.0$ m^3.
Weight	250 kg
Power	225 W (orbital average)
Data rate	11 Mbps (peak daytime)
Quantisation	12 bits
Spatial resolution (at nadir)	250m (bands 1–2) 500m (bands 3–7), 1000 m (bands 8–36)
Design life	5 years

data from EOS-Terra, MYD indicates that data from EOS-Aqua was used and MCD indicates that the product was generated using data from EOS-Terra and EOS-Aqua. Normally two numbers follow the three letters and indicate the intended application of the product. A list of these numbers with their appropriate descriptions can be found in Table 2.5 and Table 2.6.

Most standard MODIS land products use a sinusoidal grid tiling system. Tiles are 10 degrees by 10 degrees at the equator. The tile coordinate system starts at (0,0) (horizontal tile number, vertical tile number) in the upper left corner and proceeds right (horizontal) and downward (vertical). The tile in the bottom right corner is (35,17).

2.7.4 MODIS design

MODIS is a whiskbroom scanning radiometer with a double-sided paddle wheel scan mirror, which operates at 20.3 rpm. The incident radiation (earth view) is reflected from the scan mirror to a fold mirror. From the fold mirror the radiation is reflected onto the primary mirror of the Afocal Gregorian Telescope. The radiation is reflected by the primary mirror and then passes through a field stop. After passing through the field stop the radiation falls upon the secondary mirror of the telescope. The secondary mirror then reflects the incident radiation to three dichroic beamsplitters. The beamsplitters split the radiation into four regions: NIR, visible, SWIR and MWIR, and LWIR, after which the radiation ends up on four Focal Plane Assemblies (FPAs), one for each region. The MODIS sensor also houses four on-board calibrators. The name of each calibrator is given below [37, 38]:

1. Solar Diffuser (SD) and Solar Diffuser Stability Monitor (SDSM),
2. Spectral Radiometric Calibration Assembly (SRCA),
3. Blackbody (BB),
4. Space View (SV).

2.7.5 The MCD43A4 product

The MCD43A4 MODIS product consists of seven BRDF corrected land surface reflectance (eight-day composite, 500 m resolution) time-series [22]. BRDF is discussed in Section 2.5.3. The reason for selecting this product was discussed in Section 1.3. The product is built from a 16-day rolling window

Table 2.5: The MODIS Product Codes 01–28

Product Code#	Description
01	Level-1A Radiance Counts
02	Level-1B Calibrated, Geolocated Radiances
03	Geolocation Data Set
04	Aerosol Product
05	Total Precipitable Water
06	Cloud Product
07	Atmospheric Profiles
08	Gridded Atmosphere Products (Level 3)
09	Atmospherically Corrected Surface Reflectance
10	Snow Cover
11	Land Surface Temperature and Emissivity
12	Land Cover/ Land Change
13	Vegetation Indices
14	Thermal Anomalies, Fires and Biomass Burning
15	Leaf Area Index and FPAR
16	Surface Resistance and Evapotranspiration
17	Vegetation Production, Net Primary Productivity
18	Normalised Water Leaving Radiance
19	Pigment Concentration
20	Chlorophyll II, Fluorescene
21	Chlorophyll and Pigment Concentration
22	Photosynthetically Active Radiation
23	Suspended Solids Concentration in Ocean Water
24	Organic Matter Concentration
25	Coccolith Concentration
26	Ocean Water Attenuation Coefficient
27	Ocean Primary Productivity
28	Sea Surface Temperature

Table 2.6: The MODIS Product Codes 29–MODISALB.

Product Code#	Description
29	Sea Ice Cover
31	Phycoerythrin Concentration
35	Cloud Mask
36	Total Absorption Coefficient
36	Total Absorption Coefficient
37	Ocean Aerosol Properties
39	Clear Water Epsilon
43	Albedo-16 day (Level 3)
44	Vegetation Cover Conversion and Continuous Fields
MODISALB	Snow and Sea Ice Albedo

of acquisitions obtained from the Terra and Aqua satellites, which explains the use of "MCD" in the product name. An MCD43A4 pixel value consists of seven reflection ratios (at 500 m resolution). The seven reflection ratios are located in the seven MODIS land bands. The raw MCD43A4 data are DNs (16-bit unsigned integer values). The raw 16-bit data of MCD43A4 should not be confused with the raw radiation value (which is a 12-bit value) recorded by the MODIS sensor. The raw MCD43A4 data should be multiplied by 0.0001 to obtain reflection ratios. The temporal period of MODIS MCD43A4 (if an observation is produced every eight-days) roughly translates to 45 observations per year. NDVI is calculated from MCD43A4 by using Equation 2.11. In the remainder of this thesis the phrase "MODIS pixel" refers to the seven time-series at 500 m resolution that are associated with the MCD43A4 product.

2.8 DATASET DESCRIPTION

The datasets used in this thesis are constructed from the eight-day composite MODIS MCD43A4 BRDF corrected 500 m land surface reflectance product. The study areas associated with the MODIS MCD43A4 datasets span a total area of approximately 230 km^2 in Gauteng and 800 km^2 in Limpopo, South Africa. Gauteng and Limpopo are provinces in South Africa and their physical location is shown in Figure 2.12.

Gauteng is the smallest province in South Africa. The name "Gauteng" is derived from the Sesotho word meaning "Place of Gold". The name chosen for Gauteng is appropriate as it is the economic heart of South Africa. The capital of Gauteng is Johannesburg [12].

Limpopo is the northernmost province of South Africa. It is named after the Limpopo River. "Limpopo" is the Zulu word for "waterfalls". The Limpopo province houses the largest hunting industry in the country. The capital of Limpopo is Polokwane and was formerly known as Pietersburg [12].

The reasons for selecting Gauteng and Limpopo as study regions were discussed in Section 1.1.

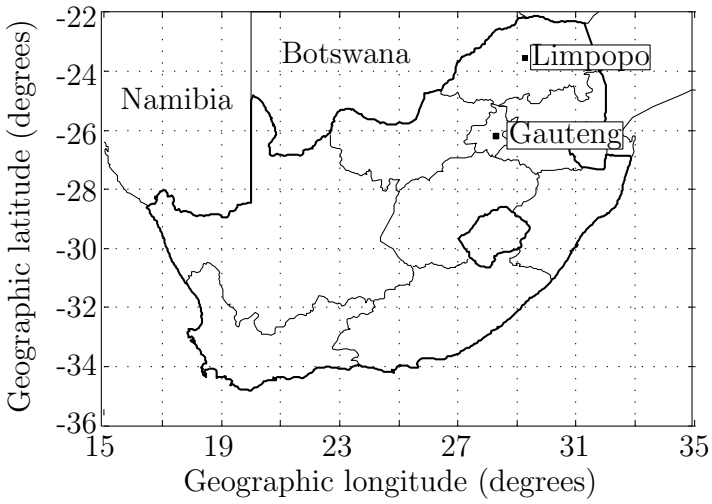


Figure 2.12: The physical location of the Gauteng province and the Limpopo province [2] © IEEE 2012.

Two land cover classes are considered: *settlements* and *natural vegetation*, denoted by s and v respectively. In this thesis the settlements class contains pixels that contain more than 50% buildings (construction), whereas the vegetation class contains pixels with more than 90% vegetation.

The above class classification rule is illustrated below with an example. Figure 2.13 is a Google Earth^{TM} image of a populated area in Gauteng. Four red parallelograms are visible in Figure 2.13. Each red parallelogram represents a pixel that is actually surveyed by the MODIS sensor and is 500 m$\times$500 m in size. If the vegetation settlement classification rule mentioned above is applied to Figure 2.13, only the bottom right pixel would be classified as a vegetation pixel, while the remaining three would be classified as settlement pixels.

Two MODIS MCD43A4 datasets are used to investigate the hypertemporal techniques discussed in

Figure 2.13: A Google Earth^{TM} image of a populated area in Gauteng (courtesy of Google Earth^{TM}).

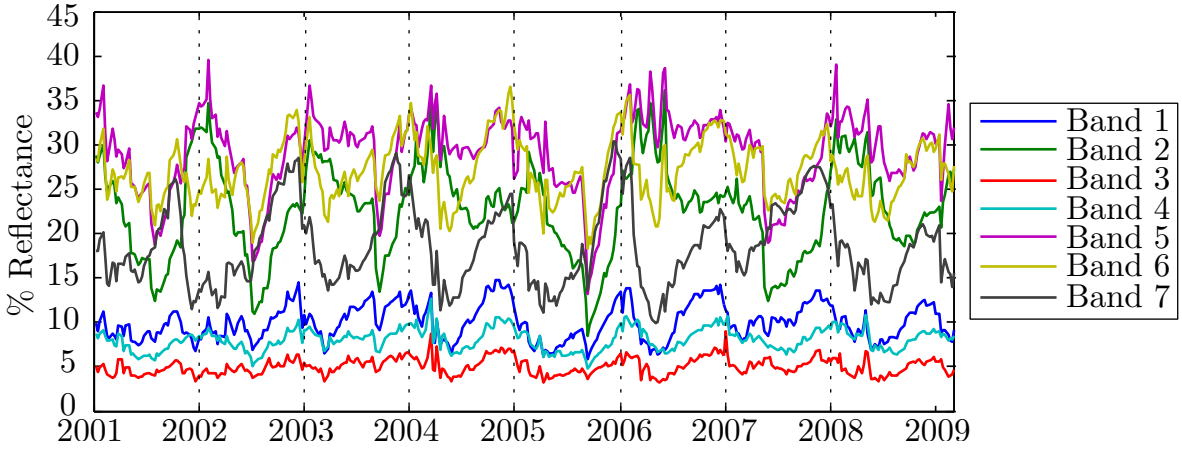


Figure 2.14: A random vegetation MODIS pixel in Gauteng.

this thesis. The Gauteng dataset consists of 1106 MODIS pixels, while the Limpopo dataset contains 3349 MODIS pixels and was selected with visual (human) interpretation of two high-resolution SPOT images from 2001 and 2009. MODIS pixels that, according to the SPOT images, either did not change or changed from vegetation to settlement were selected. Each MCD43A4 MODIS pixel contains seven time-series with $I = 368$ observations (extracted between January 2001 and March 2009). An NDVI time-series can be added to a MODIS pixel and is computed using the first two spectral land bands. The Gauteng and Limpopo datasets are respectively divided into the three classes: natural vegetation (592 Gauteng pixels and 1497 Limpopo pixels), settlements (333 Gauteng pixels and 1735 Limpopo pixels) and real land cover change from vegetation to settlement (181 Gauteng pixels and 117 Limpopo pixels). A random vegetation MODIS pixel in Gauteng is displayed in Figure 2.14.

2.9 CONCLUSION

The chapter provided a broad overview of the remote sensing field, which included a brief history of remote sensing, an introduction to the physical principles behind remote sensing, an overview of remote sensing platforms and an introduction to the MODIS sensor. The MODIS data used by the classification and change detection algorithms investigated in this thesis were also presented in this chapter.

CHAPTER 3

SEQUENTIAL ANALYSIS

The main reason for including this chapter in the thesis is to provide the theoretical background knowledge required to implement CUSUM as a sequential hypertemporal remote sensing change detection algorithm. It is recommended to first study Section A.1 if the reader is unfamiliar with stochastic calculus. Stochastic calculus provides the mathematical framework needed to understand and study sequential analysis.

In this chapter, different statistical techniques are investigated either to classify observations or to detect changes in the underlying distribution of observation. All the techniques investigated have no pre-determined sample size and are thus purely *sequential* or *on-line*. The study of statistical sequential classification and change detection techniques is known collectively as *sequential analysis*. Good literature reviews can be found in [24, 25] on the subject of sequential analysis. The main advantage of a sequential approach is that on average sequential approaches require fewer observations than a fixed-sample-size approach while maintaining the same probability of error. The reason for this is that sequential algorithms terminate uniquely for each observable sequence. In an ambiguous case the algorithm will take longer to terminate than in an unambiguous case [23]. The chapter starts with Neyman and Pearson's 1933 seminal result [44], which provides an optimal fixed-sample-size classification strategy. Neyman and Pearson's result inspired Wald [45, 46] to develop a sequential solution to the classification problem during the 1940s. Optimality was subsequently proven by Wald and Wolfowitz in 1948 [47]. The sequential classification problem, also known as *sequential detection*, is discussed in two frameworks, namely in Wald's framework (frequentest) (Section 3.3) [46] and in a Bayesian framework (Section 3.4) [48]. From sequential detection the chapter progresses to a group of change detection algorithms grouped under the collective name of *quickest detection*. Quickest detection techniques are statistical techniques capable of detecting a change as quickly as possible

after it occurs (using different measures for the delay). Statistical change detection has its roots in the seminal papers of Shewhart [49] and Page [6]. The quickest detection techniques discussed in this chapter are divided into Bayesian (see Section 3.5) and non-Bayesian (see Section 3.6) approaches. The problem of quickest detection was first cast into a Bayesian framework in 1952 [50], and was subsequently solved in 1963 by Shiryaev [51]. The most famous non-Bayesian change detection algorithm is arguably the CUSUM stopping time, first developed by Page (see Section 3.6.1) [6]. It has been shown that the CUSUM stopping time is asymptotically optimal [52] (when employing the *worst case expected delay* as a performance measure). The asymptotically optimal result was later extended by showing that the CUSUM stopping time is in fact exactly optimal [53, 54]. An alternative to the CUSUM stopping rule was proposed in 1966 and is known as the *Shiryaev-Roberts* stopping time (see Section 3.6.2) [51, 55]. An extension to the Shiryaev-Roberts stopping time known as the *Shiryaev-Roberts-Pollak* stopping time was developed in 1985 by Pollak [56]. The Shiryaev-Roberts-Pollak method is a third-order asymptotically optimal sequential procedure when employing *Pollak's performance measure* (which is less restrictive than the worst case expected delay) [56]. More recently the Shiryaev-Roberts-Pollak stopping time was extended to the *deterministic Shiryaev-Roberts* stopping time [57, 58], which can uniformly outperform both Shiryaev-Roberts and Shiryaev-Roberts-Pollak for appropriately chosen starting conditions [57,58]. For a good theoretical introduction to sequential analysis the reader is referred to [48], while the reader is referred to [59] for an overview that focuses more strongly on implementation specifics.

3.1 NEYMAN-PEARSON

The following section closely follows the notation of [60, 61]. Let $\mathbf{z}^n = \{z_k\}_{\{k=1,2,\cdots,n\}}$ be an independent and identically distributed (i.i.d.) sequence of real observation of size n following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots,n$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots,n;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively. The *problem* is to determine which hypothesis is true by only looking at the observations. Furthermore, let $q_0(\mathbf{z}^n)$ and $q_1(\mathbf{z}^n)$ denote n-dimensional density functions. Let T be a function of the observations, known as the *test statistic* and let R be the image of T. The image R can be divided into

a *critical region* R_0 and a *region of acceptance* R_1, such that $R_0 \cup R_1 = R$. If $T(\mathbf{z}^n)$ fall into R_0, $\mathscr{H}_0$ is rejected. Constructing a hypothesis test thus requires selecting a test statistic and a critical region. The probability α of rejecting $\mathscr{H}_0$ when it is in fact true is known as the *level of significance* or the *size of the test* and is equal to

$$\alpha = \int_{\{\mathbf{z}^n : T(\mathbf{z}^n) \in R_0\}} q_0(\mathbf{z}^n) d\mathbf{z}^n.$$

The probability α is also known as the probability of a false alarm P_{FA} or the type I error. The *power of the test* $1-\beta$ is defined as the probability of accepting $\mathscr{H}_1$ if it is in fact true, also known as the probability of detection P_{D}, and is equal to

$$1-\beta = \int_{\{\mathbf{z}^n : T(\mathbf{z}^n) \in R_0\}} q_1(\mathbf{z}^n) d\mathbf{z}^n.$$

In other words, β is the probability of accepting $\mathscr{H}_0$ when it is in fact false (type II error).

Neyman and Pearson [44, 61] derived the following theorem, which states that the likelihood ratio Λ is the best possible choice of T. The likelihood ratio maximises the P_{D} given a specific false alarm rate P_{FA}.

Theorem 1 (Neyman-Pearson) *To maximise the P_D for a given P_{FA} decide $\mathscr{H}_1$ if*

$$\Lambda(\mathbf{z}^n) = \frac{q_1(\mathbf{z}^n)}{q_0(\mathbf{z}^n)} > \gamma,$$

where the threshold γ is found from

$$P_{FA} = \int_{\{\mathbf{z}^n : \Lambda(\mathbf{z}^n) > \gamma\}} q_0(\mathbf{z}^n) d\mathbf{z}^n.$$

3.2 KULLBACK-LEIBLER DIVERGENCE

Theorem 1 stipulates that Λ is the optimal test statistic and some of the unique properties of Λ that make it a useful tool when building a classifier or a change detector should therefore be highlighted. Instead of calculating the likelihood ratio

$$\Lambda_k = \prod_{i=1}^{k} \frac{q_1(z_i)}{q_0(z_i)}, \tag{3.1}$$

the log-likelihood ratio

$$S_k = \sum_{i=1}^{k} s_i, \tag{3.2}$$

where

$$s_i = \ln \frac{q_1(z_i)}{q_0(z_i)}, \tag{3.3}$$

could be calculated. The sum S_k is derived from $\ln\Lambda_k$.

In probability theory and information theory, the Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions Q_0 and Q_1 and is defined as

$$\begin{aligned} D_{\mathrm{KL}}(Q_1\|Q_0) &= \int_{-\infty}^{\infty} q_1(z_1) ln \frac{q_1(z_1)}{q_0(z_1)} dz_1 \\ &= \mathbb{E}_1\left[ln\frac{q_1(z_1)}{q_0(z_1)}\right] \\ &= \mathbb{E}_1[s_1]. \end{aligned}$$

The reason why z_1 can be used is that the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ is a sample path of a discrete, stationary stochastic process. In other words, the density of the first observation of multiple sample paths equals the density from which the sequence $\mathbf{z}$ is drawn. Kullback-Leibler divergence is always positive, implying that S_k will have positive drift under $\mathscr{H}_1$, because $\mathbb{E}_1[s_1] > 0$. Under $\mathscr{H}_0$, S_k will have negative drift, since $\mathbb{E}_0[s_1] = -\int_{-\infty}^{\infty} q_0(z) ln\frac{q_0(z)}{q_1(z)} dz = -D_{\mathrm{KL}}(Q_0\|Q_1) < 0$. It should be perfectly clear that $\ln\Lambda$ is a good statistic to use when building a classifier, since under $\mathscr{H}_1$, S_k experiences positive drift, while under $\mathscr{H}_0$, S_k experiences negative drift [59].

3.3 HYPOTHESIS TESTING: WALD'S FORMULATION

The following section closely follows the notation of [59]. The problem with Neyman-Pearson is that the sample size has to be chosen before the threshold can be computed and therefore the algorithm is non-sequential. In contrast with Neyman-Pearson, Wald's formulation, the Sequential Probability Ratio Test (SPRT), is a sequential approach and is the Uniformly Most Efficient (UME) test among all sequential tests. In general the art of classifying as quickly (no predetermined sample size) and accurately as possible is known as *sequential detection*.

The problem introduced in Section 3.1 is now restated without limiting the sample size. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observation following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively.

Furthermore, let the sequence $\mathbf{z}$ be adapted to the filtration $\mathscr{F}_k = \sigma(\{z_k\}_{\{k=1,2,\cdots\}})$. The *problem* is to determine which hypothesis is true by only looking at the observations.

The above problem can be solved by using a *sequential statistical test*. A sequential statistical test for testing between hypotheses $\mathscr{H}_0$ and $\mathscr{H}_1$ is defined by a *sequential decision rule*, which is the pair (δ, T), where T is a *stopping time* and δ a *decision function*. In the case of the SPRT the stopping time is equal to

$$T = T^{\text{SPRT}}_{\{A,B\}} = \inf\{k \geq 0 | \Lambda_k \notin (A,B)\}, \tag{3.4}$$

where Λ_k was defined in Equation 3.1, and the decision function (after stopping) is

$$\delta_T = \begin{cases} 0 & \text{when } \Lambda_T \leq A \\ 1 & \text{when } \Lambda_T \geq B. \end{cases}$$

In other words, Wald's test keeps on sampling until the likelihood ratio crosses the *exit thresholds* A or B, at which time a decision is made. If Λ_T is less or equal to A, hypothesis $\mathscr{H}_0$ is accepted, if Λ_T is greater or equal to B, hypothesis $\mathscr{H}_1$ is accepted. The type I error α is equal to the probability $P_0(\delta_T = 1)$, where the subscript refers to the fact that $\mathscr{H}_0$ is assumed to be true. The probability of a type II error β is equal to the probability $P_1(\delta_T = 0)$.

The log-likelihood ratio S_k (Equation 3.2) can also be used instead of Λ_k to derive the sequential decision rule, (δ, T). In the log-likelihood domain the SPRT stopping time is equal to

$$T = T^{\text{SPRT}}_{\{-a,h\}} = \inf\{k \geq 0 | S_k \notin (-a,h)\}, \tag{3.5}$$

where $\ln A = -a$ and $\ln B = h$. The decision rule now becomes

$$\delta_T = \begin{cases} 0 & \text{when } S_T \leq -a \\ 1 & \text{when } S_T \geq h. \end{cases}$$

Overshoot is an important concept that is often used to analyse the performance of the SPRT algorithm and should therefore be defined formally. Let

$$\mathscr{O}(T, S_T, -a, h, \delta_T) = \begin{cases} |S_T + a| & \text{when } \delta_T = 0 \\ |S_T - h| & \text{when } \delta_T = 1. \end{cases} \tag{3.6}$$

When inspecting Equation 3.6 it should be clear to the reader that $\mathscr{O}(T, S_T, -a, h, \delta_T)$ is a random variable. If $\mathscr{O}(T, S_T, -a, h, \delta_T) \equiv 0$ then there is no overshoot (S_T always equals one of the boundaries $\{a, h\}$), however when $\mathscr{O}(T, S_T, -a, h, \delta_T) \not\equiv 0$ then overshoot does occur.

As mentioned before, Wald's SPRT is the UME test among all sequential tests. This fact is formally stated by the Wald-Wolfowitz theorem, given below without proof [47,48].

Theorem 2 (Wald-Wolfowitz) *Suppose* (T,δ) *is the Sequential Probability Ratio Test, SPRT*(A,B) *with* $0 < A \leq 1 \leq B < \infty$, *and let* (T',δ') *denote any other sequential decision rule with* $\max\{\mathbb{E}_0[T'],\mathbb{E}_1[T']\} < \infty$, *and satisfying*

$$\alpha' = P_0(\delta'_{T'} = 1) \leq P_0(\delta_T = 1) = \alpha \text{ and } \beta' = P_1(\delta'_{T'} = 0) \leq P_1(\delta_T = 0) = \beta,$$

with

$$P_0(\delta_T = 1) + P_1(\delta_T = 0) < 1.$$

Then

$$\mathbb{E}_0[T'] \geq \mathbb{E}_0[T] \text{ and } \mathbb{E}_1[T'] \geq \mathbb{E}_1[T].$$

At this point the natural question arises, "*How can the thresholds A and B be selected to achieve a certain probability of error?*" It turns out that it is quite complex to find the exact thresholds A and B, but quite simple to find approximations of A and B that typically work well in practice. These practical estimates are known as *Wald's approximations of A and B*. When $\Lambda_k \geq B$, sampling stops and hypothesis $\mathscr{H}_1$ is chosen. Clearly the decision rule leads to [23]

$$\prod_{i=1}^{k} q_1(z_k) \geq B \cdot \prod_{i=1}^{k} q_0(z_k) \implies P_1(\delta = 1) \geq B \cdot P_0(\delta = 1), \tag{3.7}$$

which can be interpreted as the probability of observing z_k under $\mathscr{H}_1$ is at least B times bigger than under $\mathscr{H}_0$. Furthermore, since $\mathscr{H}_1$ was chosen, the type I error is equal to $P_0(\delta = 1)$. Recognising the type II error β to be equal to $P_1(\delta = 0)$, an upper limit for B can be derived by using Equation 3.7 and is equal to

$$B \leq \frac{1-\beta}{\alpha}. \tag{3.8}$$

Similarly a lower limit for A can be derived and is equal to

$$A \geq \frac{\beta}{1-\alpha}. \tag{3.9}$$

Wald's approximations for A and B are derived by replacing the inequalities with equalities in Equation 3.9 and Equation 3.8 and are thus equal to

$$\tilde{A} = \frac{\beta}{1-\alpha} \text{ and } \tilde{B} = \frac{1-\beta}{\alpha}.$$

Wald's approximations for the log-likelihood domain can be found similarly and are equal to

$$-\tilde{a} = \ln\frac{\beta}{1-\alpha} \text{ and } \tilde{h} = \ln\frac{1-\beta}{\alpha}.$$

Alternatively *Wald's approximate error probabilities* can be calculated with the correct exit boundaries A and B, resulting in

$$\tilde{\alpha} = \frac{1-A}{B-A} \text{ and } \tilde{\beta} = A\frac{B-1}{B-A}. \tag{3.10}$$

3.3.1 The OC and ASN functions of the SPRT

The problem stated at the beginning of Section 3.3 is restated with additional information to help in defining the Operating Characteristic (OC) and the Average Sample Number (ASN) functions properly. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$), which obeys one of the two hypotheses,

$$\mathscr{H}_0 : \theta = \theta_0$$

versus

$$\mathscr{H}_1 : \theta = \theta_1;$$

best, where $\theta \in \Theta$ is a unique property of the random variable generating the sequence $\mathbf{z}$, for example θ could represent the mean of a Gaussian random variable and $\Theta \subset \mathbb{R}$ is the possible range of θ. The problem could also be stated for a parameter list $\boldsymbol{\theta}$, but for simplicity it is not done here. If θ is equal to θ_0, the random variable generating the sequence $\mathbf{z}$ will obey density q_0, similarly when θ equals θ_1 the sequence $\mathbf{z}$ will be distributed according to density q_1. Assume now that the sequence $\mathbf{s}$ is derived from $\mathbf{z}$ by using Equation 3.3 and is also i.i.d.. Under $\mathscr{H}_0$ the random variable generating $\mathbf{s}$ will have density f_0 and under $\mathscr{H}_1$ the random variable will have density f_1, which is shorthand for f_{θ_0} and f_{θ_1}, respectively. In general, when the unique property of $\mathbf{z}$ is equal to θ, the sequence $\mathbf{s}$ will be distributed according to density f_θ. In effect the parameter θ determines the density of the random variable generating $\mathbf{s}$.

The probability $\mathscr{Q}(\theta)$ of accepting hypothesis $\mathscr{H}_0$, treated as a function of $\theta \in \Theta$, when the exit thresholds $-a$ and h are fixed, is called the *operating characteristic* function. In other words the type I error α is equal to $1-\mathscr{Q}(\theta_0)$ and the type II error β is equal to $\mathscr{Q}(\theta_1)$ [59].

The *average sample number* $\mathbb{E}_\theta[T]$ is the mean number of sample points required to make a decision when performing a hypothesis test, with fixed exit thresholds $-a$ and h, as a function of $\theta \in \Theta$. In

the cases where θ equals θ_0 or θ_1 the shorthand notations $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ are used. If the sequence $\mathbf{z}$ follows $\mathscr{H}_0$, the expected number of samples required to make a decision is equal to $\mathbb{E}_0[T]$, while $\mathbb{E}_1[T]$ is defined similarly [59].

3.3.2 Wald's approximations

The OC function $\mathscr{Q}(\theta)$ can be approximated by $\tilde{\mathscr{Q}}(\theta)$ with equation

$$\tilde{\mathscr{Q}}(\theta) = \begin{cases} \frac{e^{-\omega_0(\theta)h}-1}{e^{-\omega_0(\theta)h}-e^{\omega_0(\theta)a}} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{h}{h+a} & \text{when } \mathbb{E}_\theta[s_1] = 0, \end{cases} \tag{3.11}$$

where $\omega_0(\theta)$ is the unique non-zero real number which satisfies

$$\begin{aligned} \mathbb{E}_\theta[e^{-\omega_0(\theta)s_1}] &= \int_{-\infty}^{\infty} e^{-\omega_0(\theta)s_1} f_\theta(s_1) ds_1 \\ &= 1, \end{aligned} \tag{3.12}$$

if a non-zero solution exists, otherwise $\omega_0(\theta) = 0$, and $\mathbb{E}_\theta[s_1]$ is defined as

$$\mathbb{E}_\theta[s_1] = \int_{-\infty}^{\infty} s_1 f_\theta(s_1) ds_1. \tag{3.13}$$

Wald's approximation of $\mathscr{Q}(\theta)$ is derived from the following well-known *identity of Wald* (see Theorem 6) [48]

$$\mathbb{E}_\theta\left[e^{-\omega S_T}\left(\mathbb{E}_\theta\left[e^{-\omega s_1}\right]\right)^{-T}\right] = 1, \tag{3.14}$$

where T is the SPRT stopping time defined in Equation 3.5. Wald's identity is valid for all $\omega \in \{\omega | \mathbb{E}_\theta[e^{-\omega s_1}] < \infty\}$. Equation 3.14 can be transformed into

$$\mathbb{E}_\theta\left[e^{-\omega S_T - T\ln\mathbb{E}_\theta[e^{-\omega s_1}]}\right] = 1, \tag{3.15}$$

trivially. If ω is substituted with $\omega_0(\theta)$, Equation 3.15 reduces to

$$\mathbb{E}_\theta\left[e^{-\omega_0(\theta)S_T}\right] = 1. \tag{3.16}$$

If the excess over the boundaries $-a$ and h is ignored, S_T is approximately equal to either $-a$ or h, and Equation 3.16 becomes

$$e^{-\omega_0(\theta)h}[1 - P_\theta(S_T \leq -a)] + e^{\omega_0(\theta)a} P_\theta(S_T \leq -a) \approx 1, \tag{3.17}$$

where $P_\theta(S_T \leq -a)$ is equal to $\mathcal{Q}(\theta)$. By making $\mathcal{Q}(\theta)$ the subject of Equation 3.17, Equation 3.11 is obtained.

The ASN function $\mathbb{E}_\theta(T)$ can be approximated via $\tilde{\mathbb{E}}_\theta(T)$ with equation

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{-a\hat{\mathcal{Q}}(\theta)+h(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1]} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{a^2\hat{\mathcal{Q}}(\theta)+h^2(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1^2]} & \text{when } \mathbb{E}_\theta[s_1] = 0. \end{cases} \tag{3.18}$$

The approximation of the ASN function was also derived through another *identity of Wald*, namely

$$\mathbb{E}_\theta[T] = \begin{cases} \frac{\mathbb{E}_\theta[S_T]}{\mathbb{E}_\theta[s_1]} & \text{if } \mathbb{E}[s_1] \neq 0 \\ \frac{\mathbb{E}_\theta[S_T^2]}{\mathbb{E}_\theta[s_1^2]} & \text{if } \mathbb{E}[s_1] = 0, \end{cases} \tag{3.19}$$

where $\mathbb{E}_\theta[S_T]$ equals

$$-a\tilde{\mathcal{Q}}(\theta) + h[1-\tilde{\mathcal{Q}}(\theta)], \tag{3.20}$$

and $\mathbb{E}_\theta[S_T^2]$ equals

$$a^2\tilde{\mathcal{Q}}(\theta) + h^2[1-\tilde{\mathcal{Q}}(\theta)], \tag{3.21}$$

if the excess over the boundaries is ignored. The stopping time T used in Wald's identity is again the SPRT stopping time in Equation 3.5. The result of substituting Equation 3.20 and Equation 3.21 into Equation 3.19 is Equation 3.18.

Wald's approximations can be restated in the likelihood domain in which case $\tilde{\mathcal{Q}}(\theta)$ and $\tilde{\mathbb{E}}_\theta[T]$ can be expressed as

$$\tilde{\mathcal{Q}}(\theta) = \begin{cases} \frac{B^{-\omega_0(\theta)}-1}{B^{-\omega_0(\theta)}-A^{-\omega_0(\theta)}} & \text{when } \mathbb{E}_\theta(s_1) \neq 0 \\ \frac{\ln B}{\ln(BA^{-1})} & \text{when } \mathbb{E}_\theta(s_1) = 0, \end{cases}$$

and

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{\ln A\hat{\mathcal{Q}}(\theta)+\ln B(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1]} & \text{when } \mathbb{E}_\theta[s_1] \neq 0 \\ \frac{(\ln A)^2\hat{\mathcal{Q}}(\theta)+(\ln B)^2(1-\hat{\mathcal{Q}}(\theta))}{\mathbb{E}_\theta[s_1^2]} & \text{when } \mathbb{E}_\theta[s_1] = 0. \end{cases}$$

In the special case when $\theta = \theta_0$ or $\theta = \theta_1$ then $\tilde{\mathcal{Q}}(\theta_0)$ and $\tilde{\mathcal{Q}}(\theta_1)$ reduces to

$$\tilde{\mathcal{Q}}(\theta_0) = \frac{B-1}{B-A} \text{ and } \tilde{\mathcal{Q}}(\theta_1) = A\frac{B-1}{B-A},$$

respectively, which is nothing more than the approximations already stated in Equation 3.10. The function $\tilde{\mathbb{E}}_\theta(T)$ also simplifies in the two special cases $\theta = \theta_0$ and $\theta = \theta_1$ to

$$\tilde{\mathbb{E}}_0[T] = (\mathbb{E}_0[s_1])^{-1}\left[\tilde{\alpha}\ln\left(\frac{1-\tilde{\beta}}{\tilde{\alpha}}\right) + (1-\tilde{\alpha})\ln\left(\frac{\tilde{\beta}}{1-\tilde{\alpha}}\right)\right], \tag{3.22}$$

$$\tilde{\mathbb{E}}_1[T] = (\mathbb{E}_1[s_1])^{-1}\left[(1-\tilde{\beta})\ln\left(\frac{1-\tilde{\beta}}{\tilde{\alpha}}\right) + \tilde{\beta}\ln\left(\frac{\tilde{\beta}}{1-\tilde{\alpha}}\right)\right], \tag{3.23}$$

respectively, where $\tilde{\alpha}$ and $\tilde{\beta}$ are Wald's probability of error approximations [59].

3.3.3 Exact computation

For a given θ, let $P_\theta(-a|y) = P_\theta(y)$ be the probability that S_k (Equation 3.2), starting from y, reaches the lower bound $-a$, and let $\mathbb{E}_\theta[T|y] = N_\theta(y)$ be the expected number of sample points required by the SPRT algorithm to terminate when S_k starts at y, i.e. $S_k = \sum_{i=1}^{k} s_i + y$ [59]. It should now be clear that $\mathscr{Q}(\theta)$ is equal to $P_\theta(0)$ and $\mathbb{E}_\theta[T]$ is equal to $N_\theta(0)$. It is widely known that $P_\theta(y)$ and $N_\theta(y)$ respectively satisfy the following two Fredholm integral equations of the second kind [62],

$$P_\theta(y) = \int_{-\infty}^{-a-y} f_\theta(s_1)ds_1 + \int_{-a}^{h} P_\theta(s_1)f_\theta(s_1-y)ds_1,\ -a \le y \le h, \tag{3.24}$$

$$N_\theta(y) = 1 + \int_{-a}^{h} N_\theta(s_1)f_\theta(s_1-y)ds_1,\ -a \le y \le h, \tag{3.25}$$

which can be solved through a system of linear equations that approximate Equation 3.24 and Equation 3.25 [63]. The derivation of $P_\theta(y)$ and $N_\theta(y)$ is based upon the theory of a random walk with absorbing and reflecting boundaries (barriers) [59]. See Section 3.3.5.2 for further details. Another exact approach is the Markov method of Brook [64].

3.3.4 Simulation

The easiest way to compute $\mathscr{Q}(\theta)$ and $\mathbb{E}_\theta[T]$ is through simulation. The pseudo-code for computing the OC and ASN functions is given in Listing 3.1 (listed at the end of the chapter). The functions obtained via simulation become more accurate as N becomes larger.

3.3.5 Example: Gaussian random variable

Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) generated by a Gaussian random variable with density $\mathscr{N}(\theta,1)$. The problem is to choose one of two

fixed hypotheses so that the data sequence fits that hypothesis best. The following two hypotheses are under consideration:

$$\mathscr{H}_0 : \theta = \theta_0 = 0$$

versus

$$\mathscr{H}_1 : \theta = \theta_1 = 1.$$

For this example the increment sequence $\mathbf{s}$ becomes $\{s_k\}_{\{k=1,2,\cdots\}} = \{z_k - \frac{1}{2}\}_{\{k=1,2,\cdots\}}$, since

$$\begin{aligned} s_k &= \ln \frac{q_1(z_k)}{q_0(z_k)} \\ &= \ln \frac{e^{-(z_k-1)^2}}{e^{-z_k^2}} \\ &= z_k - \frac{1}{2}, \end{aligned}$$

and is thus also an independent Gaussian sequence with density $f_\theta(s_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(s_1-(\theta-\frac{1}{2}))^2}{2}}$. Example realisations of z_k and S_k when $\theta = \theta_0$ or $\theta = \theta_1$ are given in Figure 3.1, while the probability density functions of the random variable generating z_k and s_k under the same assumption of θ are given in Figure 3.2. Note that the example sequences are classified correctly for the exit thresholds equal to $h = 3$ and $-a = -3$ and that the stopping times are equal to $T = 15$ and $T = 11$ in Figure 3.1c and Figure 3.1d, respectively.

The OC and ASN functions offer insight into the problem given above and will be calculated by using Wald's approximation as well as the exact computational approach presented in Section 3.3.3.

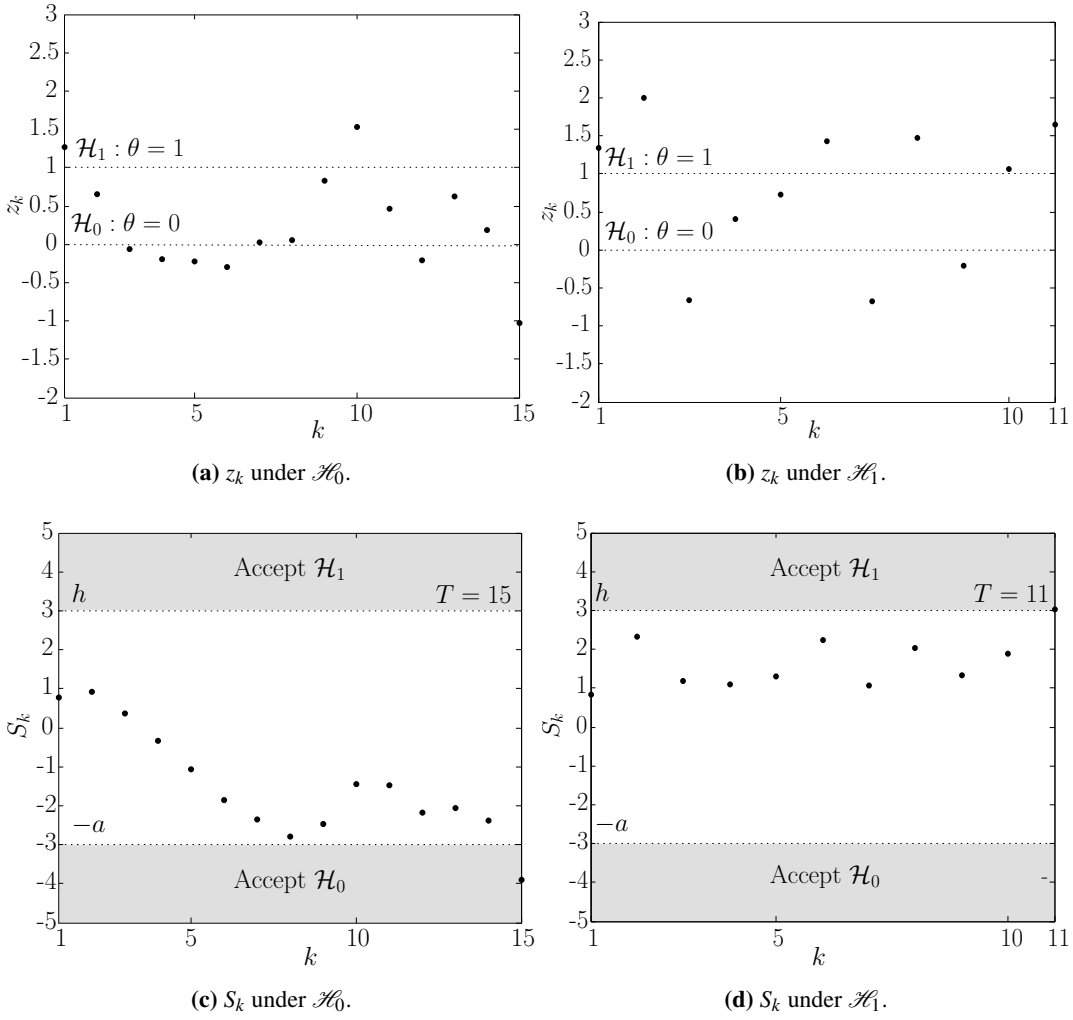
3.3.5.1 Wald's approximation

The first step in calculating the OC and ASN functions is to determine $\omega_0(\theta)$ by using Equation 3.12 and for this example is equal to

$$\begin{aligned} \mathbb{E}[e^{-\omega_0(\theta)s_1}] &= \int_{-\infty}^{\infty} e^{-\omega_0(\theta)s_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{(s_1-(\theta-\frac{1}{2}))^2}{2}} ds_1 \\ &= 1 \\ e^{-\omega_0(\theta)(\theta-\frac{1}{2})+\frac{1}{2}\omega_0^2(\theta)} &= e^0 \\ \omega_0(\theta) &= 2\theta - 1. \end{aligned} \tag{3.26}$$

The next step is to calculate $\mathbb{E}_\theta[s_1]$ by using Equation 3.13, in order to attain

$$\mathbb{E}_\theta[s_1] = \theta - \frac{1}{2}.$$



(a) z_k under $\mathscr{H}_0$.

(b) z_k under $\mathscr{H}_1$.

(c) S_k under $\mathscr{H}_0$.

(d) S_k under $\mathscr{H}_1$.

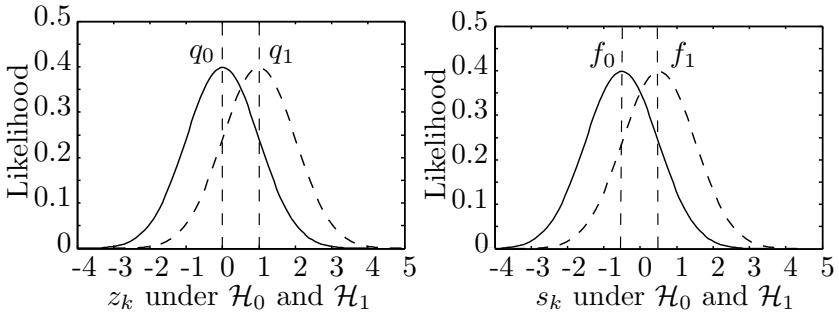
Figure 3.1: Example sequences of z_k and S_k for the unit variance Gaussian example having a mean of $\theta_0 = 0$ and $\theta_1 = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively. The exit thresholds are equal to 3 and -3.

The quantity $\mathbb{E}_\theta[s_1^2]$ is also required and is equal to 1 since $s_1^2 \sim \chi_1^2$. The approximate OC function is determined by substituting Equation 3.26 into Equation 3.11 to obtain

$$\tilde{\mathscr{Q}}(\theta) = \begin{cases} \frac{e^{-(2\theta-1)h}-1}{e^{-(2\theta-1)h}-e^{(2\theta-1)a}} & \text{when } \theta \neq \frac{1}{2} \\ \frac{h}{h+a} & \text{when } \theta = \frac{1}{2}. \end{cases} \tag{3.27}$$

Wald's approximated OC function for the Gaussian example is presented in Figure 3.3.

The approximate ASN function is calculated by substituting $\mathbb{E}_\theta[s_1]$, $\mathbb{E}_\theta[s_1^2]$ and Equation 3.27 into



(a) Probability density function of z_k. **(b)** Probability density function of s_k.

Figure 3.2: Probability density functions of the random variable generating z_k and s_k under $\mathscr{H}_0$ and $\mathscr{H}_1$.

Equation 3.18, which gives

$$\tilde{\mathbb{E}}_\theta[T] = \begin{cases} \frac{1}{\theta-\frac{1}{2}}\left[\frac{1-e^{a(2\theta-1)}}{e^{-h(2\theta-1)}-e^{a(2\theta-1)}}h - \frac{e^{-h(2\theta-1)}-1}{e^{-h(2\theta-1)}-e^{a(2\theta-1)}}a\right] & \text{when } \theta \neq \frac{1}{2} \\ ah & \text{when } \theta = \frac{1}{2}. \end{cases}$$

Wald's approximated ASN function for the Gaussian example is presented in Figure 3.4.

3.3.5.2 Exact computation

By substituting $f_\theta(s_1)$ into Equation 3.24 and Equation 3.25 and applying the method of Gaussian quadrature (Section A.2) [63, 65], Equation 3.24 and Equation 3.25 can be reduced to

$$\tilde{P}(y) = \Phi\left(-a-y-\left(\theta-\frac{1}{2}\right)\right) + \sum_{k=1}^{m} A_k \cdot (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_k-y-(\theta-\frac{1}{2}))^2} \cdot \tilde{P}(y_k), \tag{3.28}$$

$$\tilde{N}(y) = 1 + \sum_{k=1}^{m} A_k \cdot (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(y_k-y-(\theta-\frac{1}{2}))^2} \cdot \tilde{N}(y_k), \tag{3.29}$$

where $\Phi(y) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt$, and, A_k and y_k are, respectively, the weights and roots of the Gaussian quadrature for the interval $[-a, h]$. The θ subscript is dropped to avoid clutter. Equation 3.28 can be replaced by the following system of linear equations

$$A \cdot \tilde{P} = \tilde{B},$$

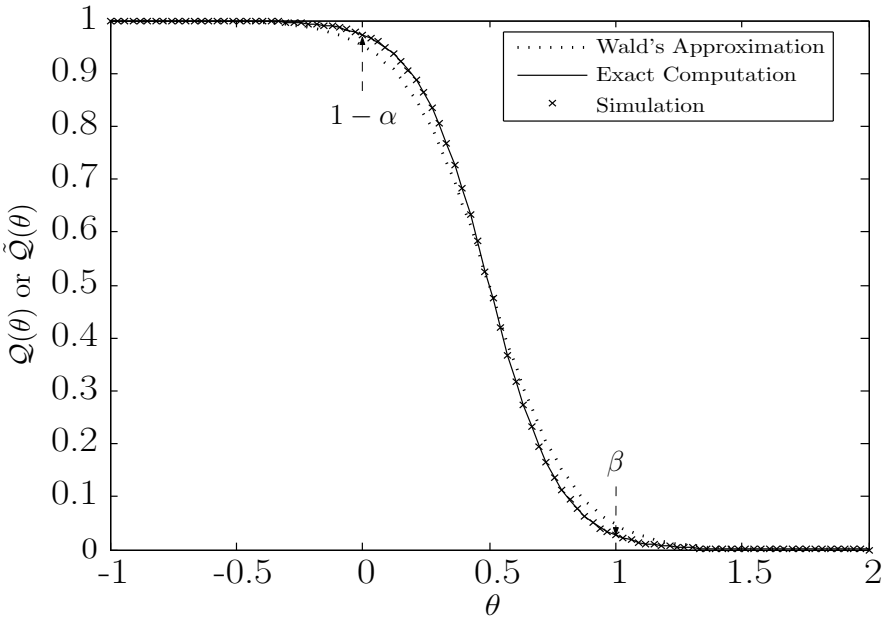


Figure 3.3: The exact OC function and Wald's approximated OC function for the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to 3 and -3.

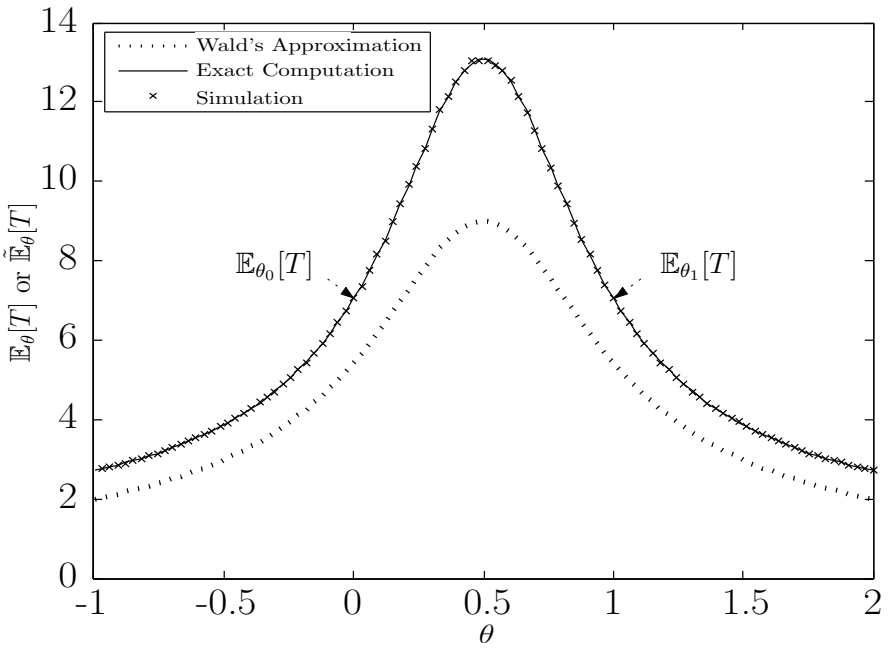


Figure 3.4: The exact ASN function and Wald's approximated ASN function for the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to 3 and -3.

where the matrix $A(m \times m)$ and column vectors $\tilde{P}(m \times 1)$ and $\tilde{B}(m \times 1)$ are defined by

$$\begin{aligned} A &= (a_{ij}),\ i,j = 1,\cdots,m; \\ \tilde{P}^T &= [\tilde{P}(y_1),\cdots,\tilde{P}(y_m)], \\ \tilde{B}^T &= \left[\Phi\left(-a-y_1-\left(\theta-\frac{1}{2}\right)\right),\cdots,\Phi\left(-a-y_m-\left(\theta-\frac{1}{2}\right)\right)\right], \end{aligned}$$

with

$$\begin{aligned} a_{ij} &= -A_j\psi(y_j,y_i) \text{ for } i \neq j, \\ a_{ii} &= 1 - A_i\psi(y_j,y_i), \\ \psi(y_j,y_i) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(y_j-y_i-(\theta-\frac{1}{2}))^2}{2}}. \end{aligned}$$

The column vector $\tilde{P}$ can now be solved easily with $\tilde{P} = A^{-1} \cdot \tilde{B}$. The OC function $\mathscr{Q}(\theta)$ is obtained by substituting the column vector $\tilde{P}$ into Equation 3.28 and setting y to naught. Similarly, $\tilde{N}$ can be ascertained by replacing Equation 3.29 with the linear system

$$A \cdot \tilde{N} = I,$$

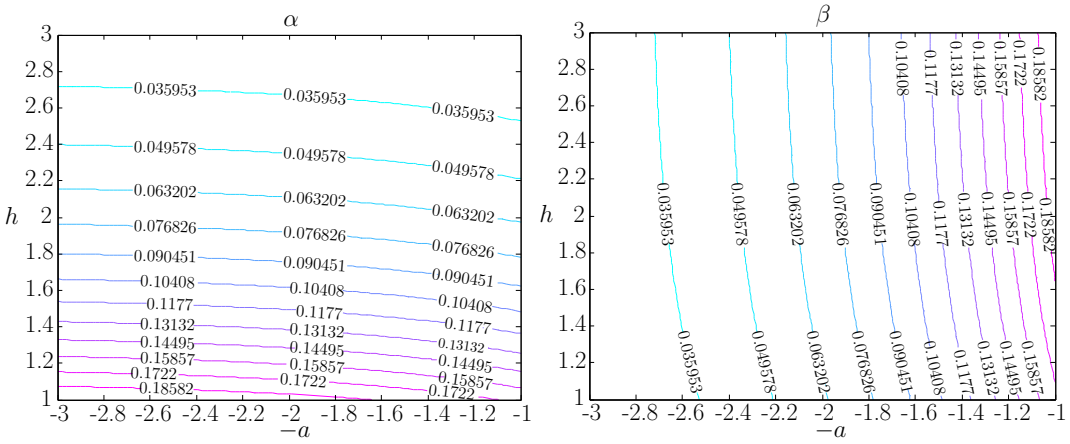
where A is as before, I is an $m \times 1$ unit vector and $\tilde{N}$ is the column vector $\tilde{N}^T = [\tilde{N}(y_1),\cdots,\tilde{N}(y_m)]$. As before, the column vector $\tilde{N}$ can be solved with $\tilde{N} = A^{-1} \cdot I$. The ASN function $\mathbb{E}_\theta[T]$ is obtained by substituting the column vector $\tilde{N}$ into Equation 3.29 and setting y to naught. The exact OC and ASN functions for the Gaussian example are presented in Figure 3.3 and Figure 3.4. Note that the curve obtained through simulation fits precisely on the exact theoretical curve.

The exact OC and ASN functions can be used to do a sweep of the exit boundaries. The type I and type II error, as well as the ASN of the SPRT algorithm, are presented in Figure 3.5 for the unit variance Gaussian example with exit boundaries in the range of $[1,3]$.

3.3.6 Example: Bernoulli random variable

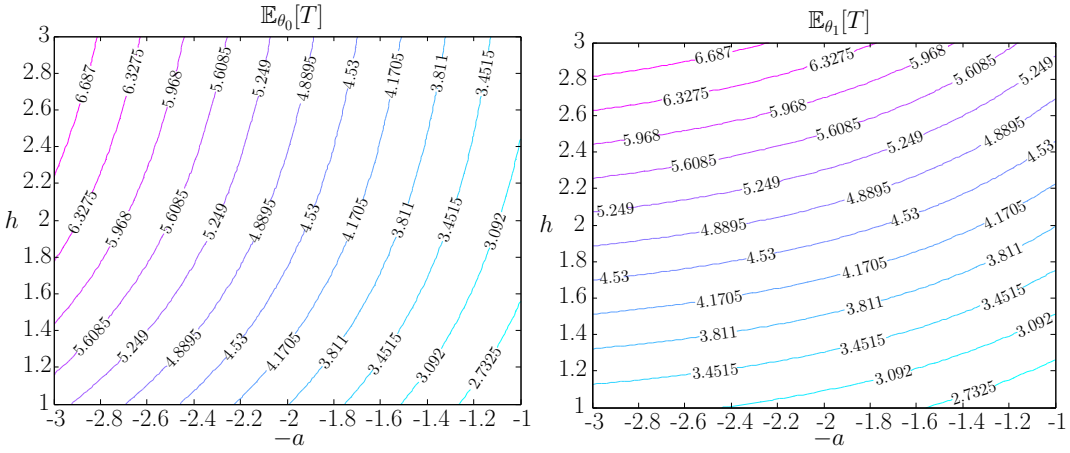
Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) generated by a Bernoulli random variable with probability mass function

$$q_p(z_1) = \begin{cases} p & \text{if } z_1 = 1 \\ 1-p & \text{if } z_1 = 0. \end{cases}$$



(a) Type I error.

(b) Type II error.



(c) $\mathbb{E}_{\theta_0}[T]$.

(d) $\mathbb{E}_{\theta_1}[T]$.

Figure 3.5: The general performance of the unit variance Gaussian example with mean $\theta = 0$ and $\theta = 1$ under $\mathscr{H}_0$ and $\mathscr{H}_1$, respectively, and exit boundaries in the range of $[1,3]$.

The problem is to choose one of two fixed hypotheses so that the data sequence fits that hypothesis best. The following two hypotheses are under consideration:

$$\mathscr{H}_0 : p = p_0 = y = 0.4$$

versus

$$\mathscr{H}_1 : p = p_1 = 1 - y = 0.6.$$

For this example the increment sequence **s** becomes

$$s_k = \begin{cases} \ln \frac{1-y}{y} & \text{if } z_k = 1 \\ \ln \frac{y}{1-y} & \text{if } z_k = 0, \end{cases}$$

with probability mass function equal to

$$f_p(s_1) = \begin{cases} p & \text{if } s_1 = \ln\frac{1-y}{y} \\ 1-p & \text{if } s_1 = \ln\frac{y}{1-y}. \end{cases}$$

and is thus also an i.i.d. Bernoulli sequence. To make the meaning of the OC and ASN functions clearer, they will be calculated in the following sections by using Wald's approximation, which for the above problem also produces the exact solution. The series S_k can only increase or decrease by $\ln\frac{1-y}{y}$ or $-\ln\frac{1-y}{y}$ and as such if the exit thresholds are chosen as integer multiples of $\ln\frac{1-y}{y}$ there will be no overshoot. When there is no overshoot, Wald's approximations are exact as they are derived by ignoring the overshoot.

3.3.6.1 Wald's approximation

As stated before, the first step in calculating the OC and ASN functions is to determine $\omega_0(p)$ by using Equation 3.12. By applying Equation 3.12 the following is attained:

$$\begin{aligned} \mathbb{E}[e^{-\omega_0(p)s_1}] &= e^{-\omega_0(p)\cdot\ln\frac{1-y}{y}}\cdot p + e^{-\omega_0(p)\cdot\ln\frac{y}{1-y}}\cdot(1-p) \\ &= e^{-\mathscr{Y}}\cdot p + e^{\mathscr{Y}}\cdot(1-p) \\ &= 1. \end{aligned} \tag{3.30}$$

Equation 3.30 can be solved by using a simple substitution, namely $e^{\mathscr{Y}} = \mathscr{X}$, as is done below:

$$\begin{aligned} e^{2\mathscr{Y}}\cdot(1-p) - e^{\mathscr{Y}} + p &= 0 \\ \mathscr{X}^2\cdot(1-p) - \mathscr{X} + p &= 0. \end{aligned}$$

The usable value of $\mathscr{X}(p)$ is equal to

$$\mathscr{X}(p) = \begin{cases} \frac{1+\sqrt{1-4(1-p)\cdot p}}{2(1-p)} & \text{if } 0 < p \leq 0.5 \\ \frac{1-\sqrt{1-4(1-p)\cdot p}}{2(1-p)} & \text{if } 0.5 < p < 1 \end{cases} \tag{3.31}$$

and is a direct result of the quadratic formula. From Equation 3.31, $\omega_0(p)$ is obtained trivially, as $e^{\omega_0(p)\cdot\ln\frac{1-y}{y}}$ is equal to $\mathscr{X}(p)$ so that

$$\omega_0(p) = \frac{\ln\mathscr{X}(p)}{\ln\frac{1-y}{y}} \text{ if } 0 < p < 1$$

$$\omega_0(p) \approx 2.47\ln\mathscr{X}(p). \tag{3.32}$$

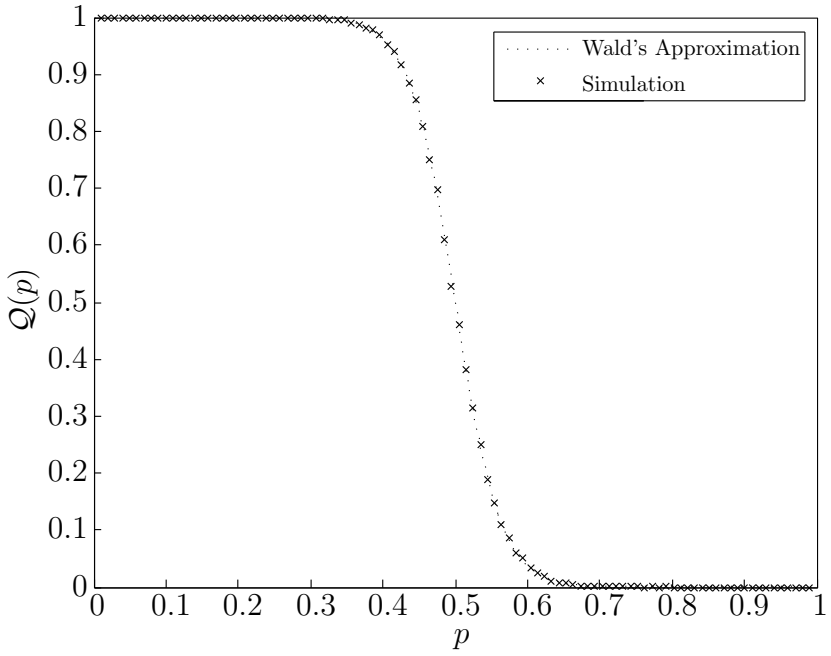


Figure 3.6: The exact OC function derived using Wald's approximation for the Bernoulli example with $p_0 = 0.4$ and $p_1 = 0.6$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to $8\ln\frac{0.6}{0.4}$ and $-8\ln\frac{0.6}{0.4}$.

Note that when $p \to 0$, $\omega_0(p) \to -\infty$ and when $p \to 1$, $\omega_0(p) \to \infty$.

The OC function is calculated by substituting Equation 3.32 into Equation 3.11 to obtain

$$\tilde{\mathscr{Q}}(p) = \begin{cases} \frac{e^{-2.47\ln\mathscr{X}h}-1}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}} & \text{when } p \neq \frac{1}{2} \\ \frac{h}{h+a} & \text{when } p = \frac{1}{2}, \end{cases} \tag{3.33}$$

and is presented in Figure 3.6. Note that the p of $\mathscr{X}(p)$ is implied.

The values $\mathbb{E}_p[s_1]$ and $\mathbb{E}_p[s_1^2]$ need to be calculated before computing the ASN function. For this example, $\mathbb{E}_p[s_1] = (2p+1)\ln\frac{1-y}{y} \approx 0.41(2p+1)$ and $\mathbb{E}_p[s_1^2] = (\ln\frac{1-y}{y})^2 \approx 0.16$. The ASN function is calculated by substituting $\mathbb{E}_p[s_1]$, $\mathbb{E}_p[s_1^2]$ and Equation 3.33 into Equation 3.18 which gives

$$\tilde{\mathbb{E}}_p[T] = \begin{cases} \frac{1}{0.41(2p+1)}\left[\frac{1-e^{2.47\ln\mathscr{X}a}}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}}h - \frac{e^{-2.47\ln\mathscr{X}h}-1}{e^{-2.47\ln\mathscr{X}h}-e^{2.47\ln\mathscr{X}a}}a\right] & \text{when } p \neq \frac{1}{2} \\ \frac{ah}{0.16} & \text{when } p = \frac{1}{2}. \end{cases}$$

The exact ASN function for the Bernoulli example can be found in Figure 3.7.

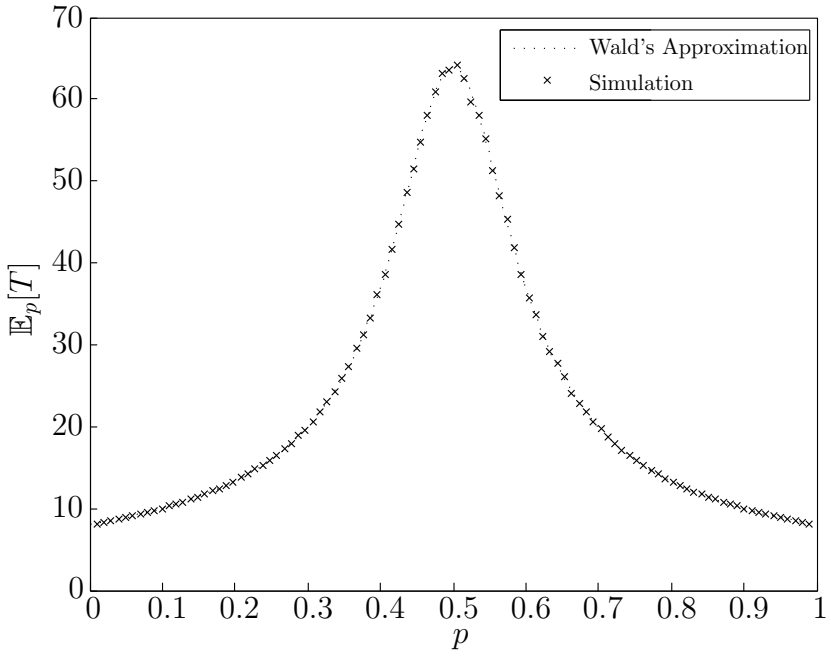


Figure 3.7: The exact ASN function function derived using Wald's approximation for the Bernoulli example with $p_0 = 0.4$ and $p_1 = 0.6$ under $\mathscr{H}_0$ and $\mathscr{H}_1$ respectively. The exit thresholds are equal to $8\ln\frac{0.6}{0.4}$ and $-8\ln\frac{0.6}{0.4}$.

3.4 HYPOTHESIS TESTING: BAYESIAN FORMULATION

The following section closely follows the notation of [48]. Once again, consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations (adapted to the filtration $\mathscr{F}_k$) following one of two hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots;$$

where Q_0 and Q_1 are two probability distributions with associated densities q_0 and q_1, respectively. Further assume that hypothesis $\mathscr{H}_1$ occurs with prior probability π and hypothesis $\mathscr{H}_0$ occurs with prior probability $1-\pi$. Instead of asking what the exit boundaries should be to obtain a *certain probability of error*, as is done in the case of Wald, the problem could be restated in terms of different *costs*, that should be minimised concurrently. Naturally three different costs are important, namely the cost incurred by observing an observation $c \geq 0$, the cost of making a type I error $c_0 > 0$ and the cost of making a type II error $c_1 > 0$. The power of this approach is flexibility, as the objective is not

to have the lowest probability of error but rather to minimise a cost function, which takes into account c, c_0 and c_1. This alternative problem formulation is known as the *Bayesian sequential detection problem*.

As already stated, any sequential test consists of a sequential decision rule (T, δ), where T is a stopping time and δ is a decision function that can be evaluated after each observation. From the sequential decision rule it follows that the *average cost of error* can be expressed as

$$\begin{aligned} c_e(T,\delta) &= (1-\pi)c_0 P_0(\delta_T = 1) + \pi c_1 P_1(\delta_T = 0) \\ &= (1-\pi)c_0\alpha + \pi c_1\beta. \end{aligned} \tag{3.34}$$

Complementary to the average cost of error is the *cost of sampling*, which can be expressed as

$$c\mathbb{E}_\pi[T] = c\cdot[(1-\pi)\mathbb{E}_0[T] + \pi\mathbb{E}_1[T]], \tag{3.35}$$

where $\mathbb{E}_\pi[\cdot]$ denotes expectation under the probability measure $P_\pi = (1-\pi)P_0 + \pi P_1$. The average cost of error reduces to the *average probability of error* P_e when $c_0 = c_1 = 1$. When $c = 1$ the average cost of sampling reduces to the Average Run Length (ARL). Normally the ARL is associated with either hypothesis $\mathscr{H}_0$ or $\mathscr{H}_1$ [6], but here it refers to the general expected run length of the experiment and could therefore be seen as a misuse of terminology. To avoid ambiguity the term ASN (which is closely related to the ARL) is only used when working in Wald's framework.

The *total cost* incurred by (or *Bayes risk* of) any sequential decision rule is thus equal to the sum of the average cost of error and the cost of sampling and is expressed mathematically as

$$c_e(T,\delta) + c\mathbb{E}_\pi[T].$$

Naturally, the best sequential decision rule would be the rule that minimises the total cost, which can be stated as

$$g(\pi) = \inf_{T\in\mathscr{T},\delta\in\mathscr{D}} \left[c_e(T,\delta) + c\mathbb{E}_\pi[T]\right], \tag{3.36}$$

where $g(\pi)$ is known as the *minimal expected cost function*, and $\mathscr{T}$ and $\mathscr{D}$ are the set of all valid stopping times and decision rules, respectively. Through simple mathematical manipulation Equation 3.36 can be reformulated as

$$g(\pi) = \inf_{T\in\mathscr{T}} \mathbb{E}_\pi\left[\min\{c_1\pi_T^\pi, c_0(1-\pi_T^\pi)\} + cT\right], \tag{3.37}$$

where π_k^π is the posterior probability that $\mathscr{H}_1$ is true, given all the information up to observation k, and is expressed as

$$\begin{aligned}\pi_k^\pi &= \frac{\pi\prod_{i=1}^k q_1(z_i)}{\pi\prod_{i=1}^k q_1(z_i)+(1-\pi)\prod_{i=1}^k q_0(z_i)}\\ &= \frac{\pi_{k-1}^\pi q_1(z_k)}{\pi_{k-1}^\pi q_1(z_k)+(1-\pi_{k-1}^\pi)q_0(z_k)},\end{aligned}$$

with $\pi_0^\pi = \pi$. The optimal sequential decision rule satisfying Equation 3.36 or Equation 3.37 is given by the following theorem [23, 48]:

Theorem 3 (Optimal i.i.d. sequential decision rule) *Consider the optimisation problem of Equation 3.36 or Equation 3.37. The optimal solution is given by the sequential decision rule (T,δ) with*

$$T = \inf\{k \geq 0 | \pi_k^\pi \notin (\pi_L, \pi_U)\} \tag{3.38}$$

and

$$\delta_k = \begin{cases} 0 \text{ if } \pi_k^\pi \leq c_0/(c_0+c_1) \\ 1 \text{ if } \pi_k^\pi > c_0/(c_0+c_1), \end{cases}$$

where the exit thresholds π_L and π_U are given by

$$\pi_L = \sup\{0 \leq \pi \leq 1 | g(\pi) = c_1\pi\} \tag{3.39}$$

and

$$\pi_U = \inf\{0 \leq \pi \leq 1 | g(\pi) = c_0(1-\pi)\} \tag{3.40}$$

respectively. That is, the optimal sequential decision rule continues sampling until $\pi_k^\pi \notin (\pi_L, \pi_U)$, at which time it chooses hypothesis $\mathscr{H}_1$ if $\pi_k^\pi \geq \pi_U$ and $\mathscr{H}_0$ otherwise.

The minimal cost function $g(\pi) = \inf_{T\in\mathscr{T}} \mathbb{E}_\pi\{h(\pi_T^\pi) + cT\}$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$, can be calculated easily, since $g(\pi)$ is the monotone point-wise limit from above of the sequence of functions

$$g_k(\pi) = \min\{h(\pi), \mathscr{R}g_{k-1}(\pi) + c\},\ k = 1, 2, \cdots \tag{3.41}$$

with $g_0(\pi) = h(\pi)$, and where the operator $\mathscr{R}$ is defined by

$$\begin{aligned}\mathscr{R}r(\pi) &= \mathbb{E}_\pi[r(\pi_1^\pi)]\\ &= \int_{-\infty}^{\infty} r\left(\frac{\pi q_1(z_1)}{\pi q_1(z_1)+(1-\pi)q_0(z_1)}\right)\cdot\left[\pi q_1(z_1)+(1-\pi)q_0(z_1)\right]dz_1,\end{aligned}$$

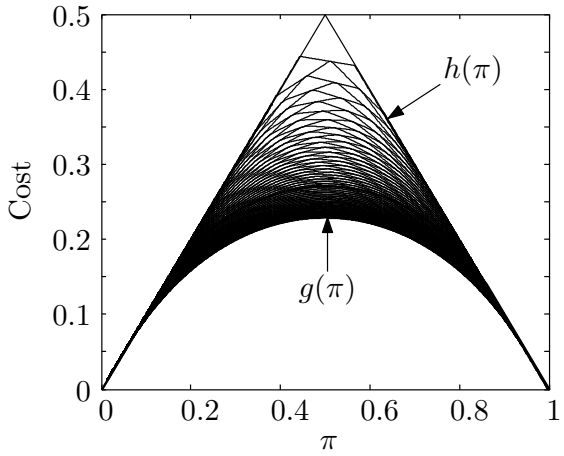


Figure 3.8: The limiting procedure used to calculate $g(\pi)$, in the case where the observations generated by $\mathscr{H}_0$ and $\mathscr{H}_1$ are different Bernoulli random variables.

such that

$$\mathscr{R}g_{k-1}(\pi) = \mathbb{E}_\pi[g_{k-1}(\pi_1^\pi)].$$

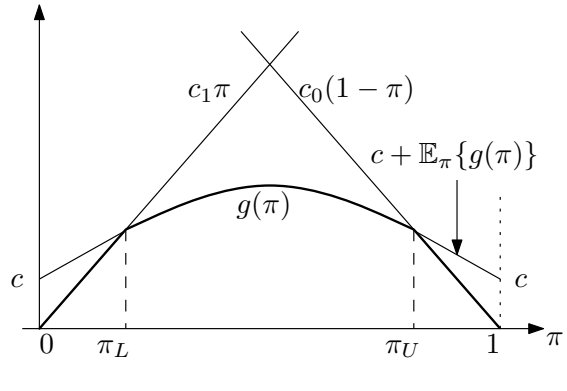
After $g(\pi)$ has been computed, the exit boundaries π_L and π_U are respectively calculated with Equation 3.39 and Equation 3.40. The limiting procedure used to calculate $g(\pi)$ is illustrated in Figure 3.8 [23, 66].

3.4.1 On the structure of the minimal cost function

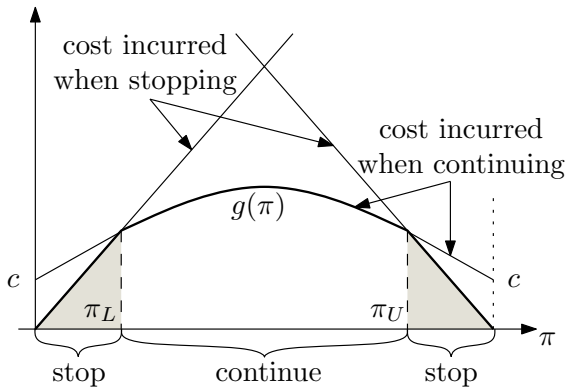
As shown in [48], the minimal cost function $g(\pi)$ is concave, and is bounded by $0 \leq g(\pi) \leq h(\pi)$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$, as mentioned in Section 3.4. Furthermore, $g(0) = g(1) = 0$. Interestingly enough, the prior probability that $\mathscr{H}_1$ is true (i.e., π) is not used to determine the minimal cost function. That is, the same $g(\pi)$ is used for any $\pi \in [0,1]$.

A classic minimal cost function is shown in Figure 3.9a, which is symmetric about the line $\pi = 1/2$, since $c_0 = c_1$.

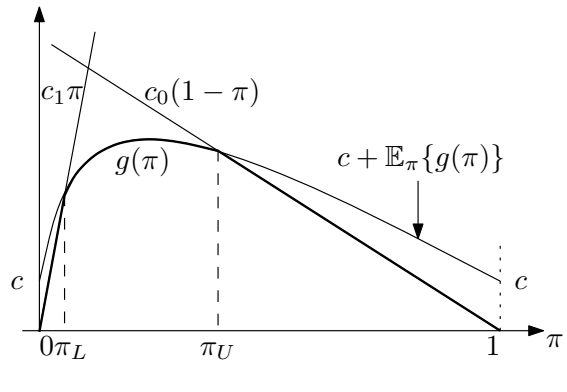
Figure 3.9b indicates that $g(\pi)$ can be divided into a continue sampling region and a stop sampling region. More specifically, $g(\pi)$ represents the minimum between the cost incurred when continuing to sample (corresponding to $c + \mathbb{E}_\pi\{g(\pi)\}$ in Figure 3.9a) and the cost incurred when terminating the experiment. The same applies to Figure 3.9c and Figure 3.9d, where the only difference is that the costs of errors (c_0 and c_1) are no longer equal.



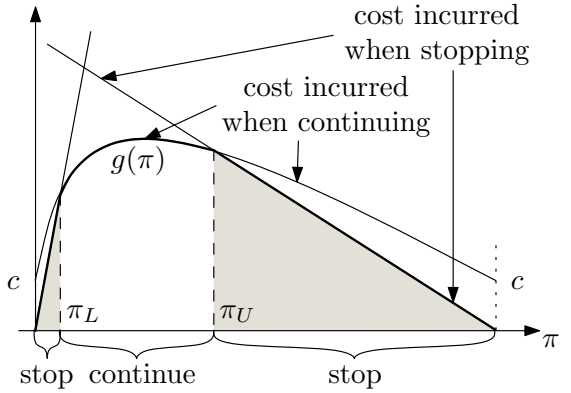
(a) Structure of the cost function, $g(\pi)$, $c_0 = c_1$.



(b) Decision regions of the cost function, $c_0 = c_1$.



(c) Structure of the cost function, $g(\pi)$, $c_0 \neq c_1$.



(d) Decision regions of the cost function, $c_0 \neq c_1$.

Figure 3.9: Typical structure and behaviour of the minimal cost function, $g(\pi)$ (from [23]).

3.4.2 Bayesian versus Wald's formulation

The boundaries π_{L} and π_U can be converted to Wald's exit boundaries with

$$A = \frac{1-\pi}{\pi}\frac{\pi_{\mathrm{L}}}{1-\pi_{\mathrm{L}}} \Longleftrightarrow \pi_{\mathrm{L}} = \frac{\pi A}{1-\pi(1-A)}, \tag{3.42}$$

and

$$B = \frac{1-\pi}{\pi}\frac{\pi_U}{1-\pi_U} \Longleftrightarrow \pi_U = \frac{\pi B}{1-\pi(1-B)}, \tag{3.43}$$

implying that Wald's SPRT stopping time (Equation 3.4) is nothing more than the Bayesian optimal stopping time (Equation 3.38) [23]. The relationship that exists between the Bayesian formulation and Wald's approach makes it possible to express the approximate type I and type II errors as a function of π_L and π_U. The approximate type I error is equal to

$$\tilde{\alpha} = \frac{1-A}{B-A} = \frac{\pi_L - \pi}{\pi - 1}\cdot\frac{\pi_{\mathrm{U-1}}}{\pi_L - \pi_U}, \tag{3.44}$$

while the approximate type II error becomes

$$\tilde{\beta} = A\frac{B-1}{B-A} = \frac{\pi_L}{\pi}\cdot\frac{\pi-\pi_U}{\pi_L-\pi_U}. \tag{3.45}$$

3.4.3 Example

Let $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ be an i.i.d. sequence of real observation (adapted to the filtration $\mathscr{F}_k$) following one of two equiprobable ($\pi = 0.5$) hypotheses:

$$\mathscr{H}_0 : z_k \sim Q_0,\ k = 1,2,\cdots,n$$

versus

$$\mathscr{H}_1 : z_k \sim Q_1,\ k = 1,2,\cdots,n$$

where Q_0 and Q_1 are two probability distributions with associated probability mass functions q_0 and q_1, respectively. The probability mass functions q_0 and q_1 are equal to

$$q_0(z_1) = \begin{cases} 0.4 \text{ if } z_1 = 1 \\ 0.6 \text{ if } z_1 = 0, \end{cases}$$

and

$$q_1(z_1) = \begin{cases} 0.6 \text{ if } z_1 = 1 \\ 0.4 \text{ if } z_1 = 0. \end{cases}$$

For this example $c_0 = 1$ and $c_1 = 1$ and $c \in [0, 0.05]$. When working in the Bayesian framework, the instinctive question arises, what should the values of c_0, c_1 and c be to obtain a certain type I and type II error? The solution to this problem turns out to be quite difficult, as there is no direct link between the costs and the probability of error. Without this link the choice of c_0, c_1 and c is quite arbitrary and of no real practical value. To find this link for the current problem, c will be traversed to determine the effect of c on α and β, while keeping c_1 and c_2 constant. See [23] for a greater variety of examples with different initial conditions. In particular, [23] investigates the case when the hypotheses are not equiprobable, as well as the case when $c_0 \neq c_1$. The focus here is however to provide an extensive example that would enable the reader to link the costs to the probability of error, for an arbitrary choice of π, c_0, c_1 and c. The exit boundaries π_U and π_L are displayed in Figure 3.10 as a function of c for the above-mentioned example. The probability of error P_e (Equation 3.34) and the ARL (Equation 3.35) is displayed in Figure 3.11. The step-like nature of the P_e and the ARL is due to the fact that for the example the exit boundaries can only be discrete functions (limited

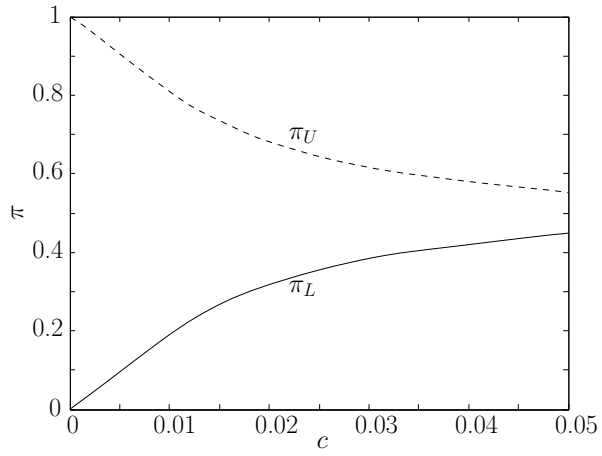
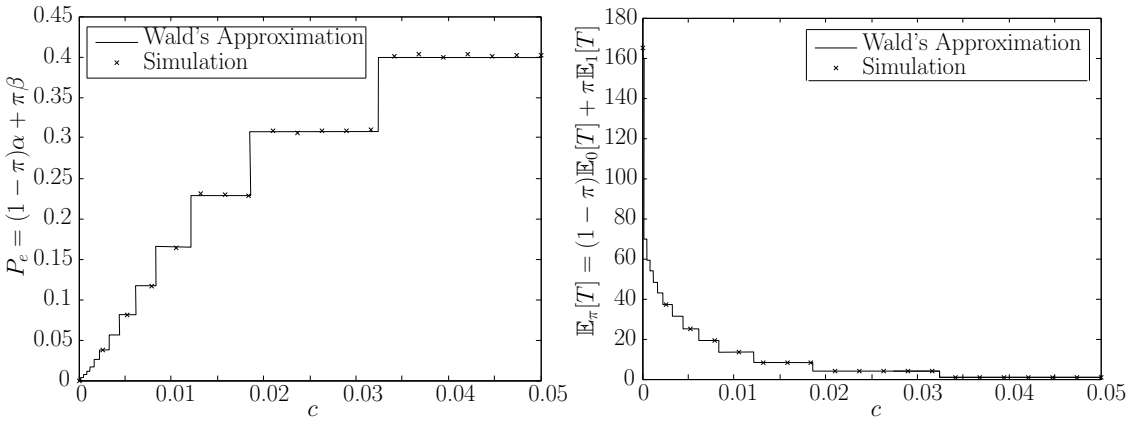


Figure 3.10: The value for π_L and π_U as a function of c with $c_0 = c_1 = 1$ for the Bernoulli example.

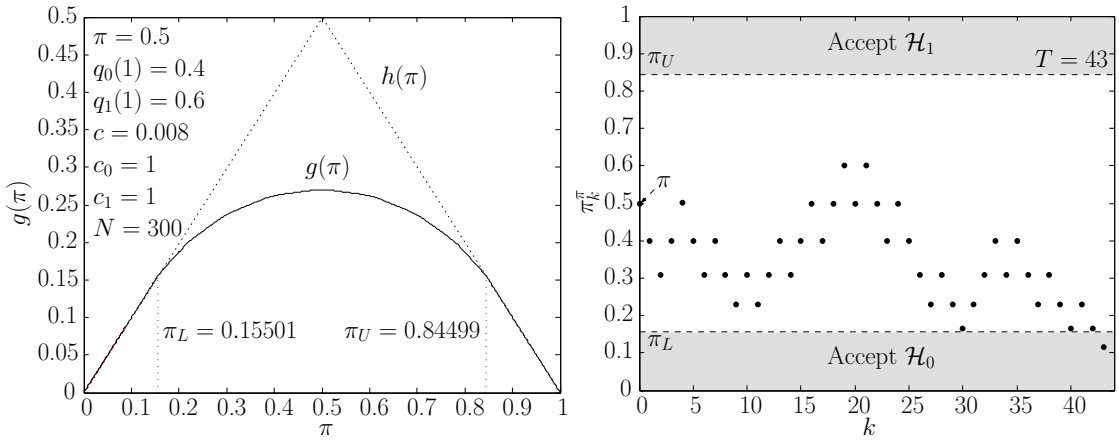


(a) Average probability of error, P_e.

(b) ARL, $\mathbb{E}_\pi[T]$.

Figure 3.11: The P_e and $\mathbb{E}_\pi[T]$ as a function of c for the Bernoulli example.

number of values), which implies that there is no overshoot, causing Wald's approximation to be exact (see Section 3.3.6 for more details). The fact that there is no overshoot should actually be taken into account when computing π_L and π_U, but is not done here for the sake of simplicity and to be compatible with [23]. The value of c is now restricted to 0.008 in order to show the reader how to obtain the curves in Figure 3.10 and Figure 3.11 (where c was traversed). When the value of c is fixed, the values for π_L and π_U are calculated by first determining $g(\pi)$ with Equation 3.41 by letting $k \to \infty$ (in practice 300 iterations were used) and then applying Equation 3.39 and Equation 3.40. The calculated function $g(\pi)$ and thresholds $\pi_L = 0.15501$ and $\pi_U = 0.84499$ for $c_0 = 1, c_1 = 1$ and $c = 0.008$ can be found in Figure 3.12a. The values of π_L and π_U can be converted to A and B with Equation 3.42 and Equation 3.43, which gives 0.1834 and 5.4512, respectively. The exit boundaries A and B need to be converted to $\bar{A}$ and $\bar{B}$, as there is no overshoot for this problem. The first step in



(a) The minimal cost function, $g(\pi)$.

(b) The sequence $\boldsymbol{\pi}^{\pi}$.

Figure 3.12: The minimal cost function and an example posterior sequence for the Bernoulli example.

calculating $\bar{A}$ and $\bar{B}$ is to calculate the constant integer k via

$$k = \left\lceil \frac{\ln B}{\ln \frac{q_1(1)}{q_0(1)}} \right\rceil .$$

Only B is used, as using A in a similar fashion would produce the same integer k. Now

$$\begin{aligned} \bar{A} &= e^{-k*\ln \frac{q_1(1)}{q_0(1)}}, \\ &= \left(\frac{q_1(1)}{q_0(1)}\right)^{-k}, \end{aligned}$$

and

$$\begin{aligned} \bar{B} &= e^{k*\ln \frac{q_1(1)}{q_0(1)}}, \\ &= \left(\frac{q_1(1)}{q_0(1)}\right)^{k}. \end{aligned}$$

Substituting $\bar{A}$ and $\bar{B}$ into Equation 3.44 and Equation 3.45 yields the correct value for $\alpha = 0.1164$ and $\beta = 0.1164$ as there is no overshoot. The average probability of error can now finally be calculated with $P_e = (1-\pi)\alpha + (\pi)\beta$, which is equal to 0.1164. Next compute the averages $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ with Equation 3.22 and Equation 3.23, where $\tilde{\alpha} = \alpha$ and $\tilde{\beta} = \alpha$ is already known and

$$\mathbb{E}_0[s_1] = q_0(0) \times \ln\left(\frac{q_1(0)}{q_0(0)}\right) + q_0(1) \times \ln\left(\frac{q_1(1)}{q_0(1)}\right),$$

$$\mathbb{E}_1[s_1] = q_1(0) \times \ln\left(\frac{q_1(0)}{q_0(0)}\right) + q_1(1) \times \ln\left(\frac{q_1(1)}{q_0(1)}\right).$$

With $\mathbb{E}_0[T]$ and $\mathbb{E}_1[T]$ the ARL is easily calculated and is equal to 14.4290. An example sequence $\boldsymbol{\pi}^{\pi}$ for the current problem under $\mathscr{H}_0$ is displayed in Figure 3.12b.

3.5 BAYESIAN QUICKEST DETECTION

The following section closely follows the notation of [48]. In Section 3.3 and Section 3.4 the focus was on sequential detection or rather classification of an observed sequence with no fixed sample size. A more general problem will be studied in the remainder of the chapter. In the more general case the observed sequence is allowed to switch from one hypothesis to another, and the aim is to detect this change as quickly as possible, while simultaneously minimising the probability of a false alarm. This section is called *Bayesian quickest detection*, since the distribution of the change point is known beforehand. In Section 3.6 the case where the change point distribution is unknown is investigated. The Bayesian quickest detection problem is also known as *Shiryaev's disruption* problem, since Shiryaev solved it [67].

Shiryaev's disruption problem is now introduced formally. Consider the sequence $\mathbf{z} = \{z_k\}_{\{k=1,2,\cdots\}}$ of i.i.d. real observations with a random change point τ. Further assume that conditioned on τ, $\mathbf{z}$ is an independent sequence where $\mathbf{z}^{-\tau} = \{z_1, z_2, \cdots, z_{\tau-1}\}$, is i.i.d. with marginal distribution Q_0, and $\mathbf{z}^{+\tau} = \{z_\tau, z_{\tau+1}, \cdots\}$ is also i.i.d. with marginal distribution Q_1. The associated densities of Q_0 and Q_1 are q_0 and q_1, respectively. A probability distribution P_π is considered that describes both the (prior) distribution of τ and the distribution of $\mathbf{z}$ induced by this prior and above conditional behaviour. Moreover, the observations $\{z_k\}_{\{k=1,2,\cdots\}}$ generate the filtration $\mathscr{F}_k$, with

$$\mathscr{F}_k = \sigma(\{z_k\}_{\{k=1,2,\cdots\}}, \{\tau = 0\}),\ k = 1,2,\cdots$$

and $\mathscr{F}_0$ contains not only Ω (the sample space) but also the set $\{\tau = 0\}$. The case where τ is geometrically distributed will be considered, and consequently,

$$P_\pi\{\tau = k\} = \begin{cases} \pi & \text{if } k = 0 \\ (1-\pi)(1-\rho)^{k-1}\rho & \text{if } k = 1,2,\cdots. \end{cases}$$

Let $T \in \mathscr{T}$ be a stopping time and let $\mathscr{T}$ be the set consisting of all valid stopping times, then T is actually the time at which the alarm is sounded to signal that a change in distribution has occurred. The optimal choice of T is the T that minimises jointly the *probability of a false alarm*

$$P_\pi\{T < \tau\} \tag{3.46}$$

and the *expected delay*

$$\mathbb{E}_\pi[(T-\tau)^+] = \mathbb{E}_\pi[\max\{T-\tau, 0\}], \tag{3.47}$$

where $\mathbb{E}_\pi$ denotes expectation under the probability measure P_π.

A convenient way of implementing a joint minimisation between Equation 3.46 and Equation 3.47 is to seek $T \in \mathscr{T}$ to solve the optimisation problem

$$g(\pi) = \inf_{T \in \mathscr{T}} \left[P_\pi\{T < \tau\} + c \cdot \mathbb{E}_\pi[(T-\tau)^+] \right], \tag{3.48}$$

where $c > 0$ is a constant controlling the relative importance of the two performance indices and $g(\pi)$ is known as the *minimal expected cost*, or simply the *minimal cost function*. Through simple mathematical manipulation Equation 3.48 can be reformulated as

$$g(\pi) = \mathbb{E}_\pi \left[1 - \pi_T^\pi + c \cdot \sum_{k=0}^{T-1} \pi_k^\pi \right], \tag{3.49}$$

where π_k^π is the posterior probability that a change did occur before or at k given all the observations up to k and is expressed as

$$\pi_k^\pi = \frac{[\pi_{k-1}^\pi + (1-\pi_{k-1}^\pi)\rho] q_1(z_k)}{[\pi_{k-1}^\pi + (1-\pi_{k-1}^\pi)\rho] q_1(z_k) + [(1-\pi_{k-1}^\pi)(1-\rho)] q_0(z_k)}, \tag{3.50}$$

with $\pi_0^\pi = \pi$.

The optimal stopping time satisfying Equation 3.48 or Equation 3.49 is given by the following theorem [48]:

Theorem 4 (Bayes optimal stopping time) *Consider the optimisation problem of Equation 3.48 or Equation 3.49. The optimal solution is given by*

$$T = \inf\{k \geq 0 | \pi_k^\pi \geq \pi^*\}$$

where the exit boundary π^* is given by

$$\pi^* = \inf\{0 \leq \pi \leq 1 | g(\pi) = 1 - \pi\}. \tag{3.51}$$

That is continue sampling until $\pi_k^\pi \geq \pi^*$, at which time a change is declared.

The minimal cost function $g(\pi) = \inf_{T \in \mathscr{T}} \mathbb{E}_\pi\{h(\pi_T^\pi) + c \cdot \sum_{k=0}^{T-1} \pi_k^\pi\}$, where $h(\pi) = 1 - \pi$, can be calculated easily, since $g(\pi)$ is the monotone point-wise limit from above of the sequence of functions

$$g_k(\pi) = \min\{h(\pi), \mathscr{R} g_{k-1}(\pi) + c\pi\}, \; k = 1, 2, \cdots$$

with $g_0(\pi) = h(\pi)$, and where the operator $\mathscr{R}$ is defined by

$$\begin{aligned}\mathscr{R}r(\pi) &= \mathbb{E}_\pi[r(\pi_1^\pi)] \\ &= \int_{-\infty}^{\infty} r(\pi_1^\pi) \cdot \big[[\pi + (1-\pi)\rho]q_1(z_1) + (1-\pi)(1-\rho)q_0(z_1)\big]dz_1,\end{aligned}$$

with

$$\pi_1^\pi = \frac{[\pi + (1-\pi)\rho]q_1(z_1)}{[\pi + (1-\pi)\rho]q_1(z_k) + [(1-\pi)(1-\rho)]q_0(z_1)},$$

such that

$$\mathscr{R}g_{k-1}(\pi) = \mathbb{E}_\pi[g_{k-1}(\pi_1^\pi)].$$

Once $g(\pi)$ is known, the exit boundary π^* is calculated with Equation 3.51 [66].

3.6 NON-BAYESIAN QUICKEST DETECTION

This section closely follows the notation from [48, 59]. In this section the quickest detection algorithms have no prior change point distribution. Two measures of *detection delay* are investigated, namely *Lorden's performance measure* [52] and *Pollak's performance measure* [56].

3.6.1 Lorden's performance measure

Consider a measurable space $(\Omega, \mathscr{F})$, consisting of a sample space Ω and a σ-field $\mathscr{F}$ of events [48]. Further consider a family $\{P_\tau | \tau \in [1, 2, \cdots, \infty]\}$ of probability measures on $(\Omega, \mathscr{F})$ and a random sequence $\mathbf{z} = \{z_k; k = 1, 2, \cdots \infty\}$, such that, under P_τ, $\mathbf{z}^{-\tau} = \{z_1, z_2, \cdots, z_{\tau-1}\}$ are independent and identically distributed (i.i.d) with a fixed marginal distribution Q_0 and $\mathbf{z}^{+\tau} = \{z_\tau, z_{\tau+1}, \cdots, \infty\}$ are i.i.d with marginal distribution Q_1 and are independent of $\mathbf{z}^{-\tau}$. The probability densities associated with Q_0 and Q_1 are q_0 and q_1 respectively. A procedure is desired that can detect a change in the underlying distribution of $\mathbf{z}$ (when $\mathbf{z}$ is sampled from Q_1 instead of Q_0), if it occurs (i.e. if $\tau < \infty$), as quickly as possible after it occurs. As a set of detection strategies, it is natural to consider the set $\mathscr{T}$ of all (extended) stopping times with respect to the filtration $\{\mathscr{F}_k\}$ where $\mathscr{F}_k$ denotes the smallest σ-field with respect to which $z_0, z_1, \cdots, z_k$ are measurable. Thus, when the stopping time T takes on the value k, the interpretation is that T has detected the existence of a change point τ at or prior to time k. It is of interest to penalise *expected delay via its worst case* value (also known as *Lorden's performance measure*)

$$d_l(T) = \sup_{\tau \geq 1} \operatorname{ess\,sup} \mathbb{E}_\tau\{(T - \tau + 1)^+ | \mathscr{F}_{\tau-1}\}, \tag{3.52}$$

where $\mathbb{E}_\tau\{\cdot\}$ denotes expectation under the distribution P_τ and $(T-\tau+1)^+ = \max\{T-\tau+1, 0\}$. Note that ess sup $\mathbb{E}_\tau\{(T-\tau+1)^+|\mathscr{F}_{\tau-1}\}$ is the worst case average delay under P_τ, where the worst case is taken over all realization of $\mathbf{z}^{-\tau}$. In other words, it is the same as measuring the average detection delay when the first sample already belongs to the changed distribution. The desire to make $d_l(T)$ small must be balanced with a constraint on the false alarm rate. The fact that false alarms will occur is accepted, however the rate at which they occur is fixed. The false alarm rate is quantified by the *mean time between false alarms*

$$f(T) = \mathbb{E}_\infty\{T\}. \tag{3.53}$$

A useful design criterion is then given by

$$\inf_{T\in\mathscr{T}} d_l(T) \text{ subject to } f(T) \geq \lambda, \tag{3.54}$$

where λ is a positive, finite constant. A stopping time is desired that minimises the worst case expected delay within a lower-bound constraint on the mean time between false alarms. A possible stopping time that meets the requirements of Equation 3.54 is Page's CUSUM stopping time [6]. In particular, for $h \geq 0$ the CUSUM stopping time is defined as

$$T_h^{\text{CUSUM}} = \inf\{k \geq 0 | g_k \geq h\},$$

where

$$g_k = \begin{cases} (g_{k-1} + s_k)^+ & \text{if } k > 0 \\ y \in \mathbb{R}^+ & \text{if } k = 0, \end{cases} \tag{3.55}$$

and

$$s_k = \ln \frac{q_1(z_k)}{q_0(z_k)}. \tag{3.56}$$

Under normal CUSUM operating conditions y is set to 0. As it turns out T_h^{CUSUM} is the optimal choice, as indicated by the theorem below [48, 53, 54]:

Theorem 5 (Optimality of CUSUM) *Choose $h \geq 0$. Then, the stopping time T_h^{CUSUM} solves Equation 3.54 with $\lambda = f(T_h^{CUSUM})$. That is,*

$$f(T) \geq f(T_h^{CUSUM}) \implies d_l(T) \geq d_l(T_h^{CUSUM}).$$

3.6.1.1 The ARL function of CUSUM

As explained in Section 3.3.1, the parameter θ determines the distribution of $\mathbf{z}$ and when $\theta = \theta_0$ density q_0 is obeyed (no change) and when $\theta = \theta_1$ density q_1 is obeyed (change occurred). The

average run length function $\mathscr{L}(\theta)$ is the expected number of samples required for an algorithm (for example CUSUM) to terminate as a function of θ when the exit threshold(s) (for example h) is/are fixed. It turns out that in the case of CUSUM, when $\theta = \theta_0$ then $\mathscr{L}(\theta_0) = f(T_h^{\text{CUSUM}})$ and when $\theta = \theta_1$ then $\mathscr{L}(\theta_1) = d_l(T_h^{\text{CUSUM}})$. The function $\mathscr{L}(\theta)$ can be calculated with [59]

$$\mathscr{L}(\theta) = \frac{N_\theta(0)}{1 - P_\theta(0)}, \tag{3.57}$$

where $N_\theta(0) = \mathbb{E}_\theta[T|0]$ and $P_\theta(0) = P_\theta(-a|0)$ were defined in Section 3.3.3. Equation 3.57 is only valid when $-a = 0$. The exact value of $\mathscr{L}(\theta)$ can thus be calculated by solving Equation 3.24 and Equation 3.25 with $-a = 0$. *Wald's approximation* of $\tilde{\mathscr{L}}(\theta)$ can be derived by evaluating the following limit

$$\tilde{\mathscr{L}}(\theta) = \lim_{a \to 0} \frac{\tilde{E}_\theta[T|0]}{1 - \tilde{P}_\theta(-a|0)}, \tag{3.58}$$

where $\tilde{P}_\theta$ is Wald's approximated OC function and $\tilde{E}_\theta$ is Wald's approximated ASN function with SPRT exit boundaries $-a$ and h. After evaluating the limit, Equation 3.58 becomes [59]

$$\tilde{\mathscr{L}}(\theta) = \begin{cases} \frac{1}{\mathbb{E}_\theta[s_k]}\left(h + \frac{e^{-\omega_0(\theta)h}}{\omega_0(\theta)} - \frac{1}{\omega_0(\theta)}\right) & \text{if } \mathbb{E}_\theta[s_k] \neq 0 \\ \frac{h^2}{\mathbb{E}_\theta[s_k^2]} & \text{if } \mathbb{E}_\theta[s_k] = 0. \end{cases} \tag{3.59}$$

However *Siegmund's approximation* is much better than Wald's approximation, as Siegmund incorporates an approximation of the overshoot. Siegmund's approximation is equal to [59]

$$\hat{\mathscr{L}}(\theta) = \begin{cases} \frac{1}{\mathbb{E}_\theta[s_k]}\left(h + \delta^+ + \delta^- + \frac{e^{-\omega_0(\theta)(h+\delta^+ + \delta^-)}}{\omega_0(\theta)} - \frac{1}{\omega_0(\theta)}\right) & \text{if } \mathbb{E}_\theta[s_k] \neq 0 \\ \frac{(h+\delta^+ + \delta^-)^2}{\mathbb{E}_\theta[s_k^2]} & \text{if } \mathbb{E}_\theta[s_k] = 0, \end{cases} \tag{3.60}$$

where

$$\delta^+ \approx \mathbb{E}_\theta[S_T - h | S_T - h \geq 0],$$

$$\delta^- \approx \mathbb{E}_\theta[S_T | S_T \leq 0].$$

3.6.1.2 Example: Gaussian random variable

Suppose there is an observed sequence $\mathbf{z}$, such that z_k is drawn from density $q_0 \sim \mathcal{N}(0,1)$ before change point τ. From time point τ, z_k is drawn from density $q_1 \sim \mathcal{N}(1,1)$. Assuming $\mathbf{z}$, it follows that $\mathbf{s}$ is also i.i.d and is characterised by density $f_0 \sim \mathcal{N}\left(-\frac{1}{2}, 1\right)$ before the change and $f_1 \sim \mathcal{N}\left(\frac{1}{2}, 1\right)$ after the change occurred. In general $\mathbf{s}$ is characterised by $f_\theta \sim \mathcal{N}(\theta - \frac{1}{2}, 1)$ (see Section 3.3.5). The CUSUM sequence $\mathbf{g}$ is derived from $\mathbf{s}$. As soon as $\mathbf{g}$ crosses h a change can be declared. An example

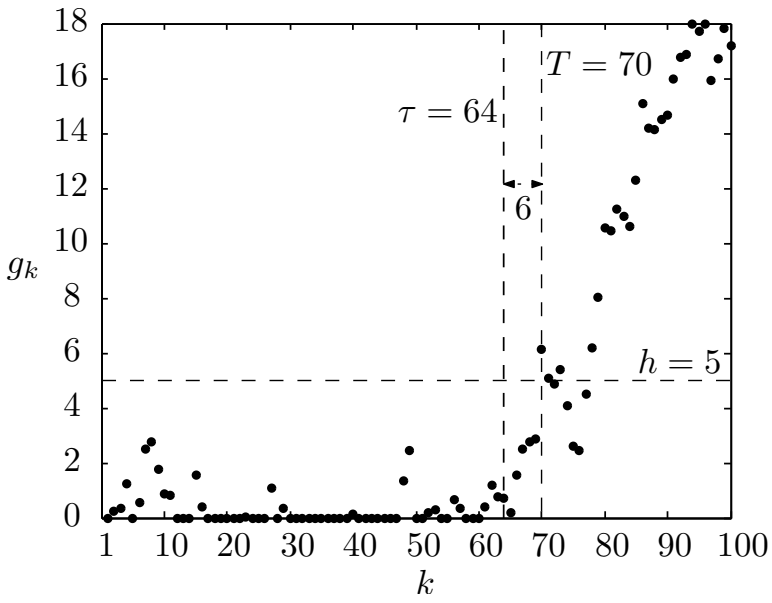


Figure 3.13: An example CUSUM sequence $\mathbf{g}$ with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$, $h = 5$ and change point $\tau = 64$.

of $\mathbf{g}$ can be found in Figure 3.13. The measures defined in Equation 3.52 and Equation 3.53 can be calculated for every h by respectively setting θ equal to either 0 or 1 in Equation 3.57. The exact values of $\mathscr{L}(0)$ and $\mathscr{L}(1)$ are calculated by using the same approach as discussed in Section 3.3.5.2. The exact values of Equation 3.52 and Equation 3.53 are displayed in Figure 3.14 for $h \in [1,5]$.

Furthermore, $\mathscr{L}(\theta)$ can be calculated by traversing θ in Equation 3.57 and fixing h. By substituting $\mathbb{E}_\theta[s_1], \mathbb{E}_\theta[s_1^2]$ and $\omega_0(\theta)$ (calculated in Section 3.3.5.1) into Equation 3.59 Wald's approximation of $\mathscr{L}(\theta)$ is obtained, which is equal to

$$\tilde{\mathscr{L}}(\theta) = \begin{cases} \frac{e^{-2(\theta-\frac{1}{2})h} - 1 + 2(\theta-\frac{1}{2})h}{2(\theta-\frac{1}{2})^2} & \text{if } \theta \neq \frac{1}{2} \\ h^2 & \text{if } \theta = \frac{1}{2}. \end{cases}$$

Siegmund's approximation is obtained by substituting $\mathbb{E}_\theta[s_1], \mathbb{E}_\theta[s_1^2]$ and $\omega_0(\theta)$ into Equation 3.60, which results in

$$\hat{\mathscr{L}}(\theta) = \begin{cases} \frac{e^{-2[(\theta-\frac{1}{2})h + 1.166(\theta-\frac{1}{2})]} - 1 + 2[(\theta-\frac{1}{2})h + 1.166(\theta-\frac{1}{2})]}{2(\theta-\frac{1}{2})^2} & \text{if } \theta \neq \frac{1}{2} \\ (h+1.166)^2 & \text{if } \theta = \frac{1}{2}. \end{cases} \tag{3.61}$$

Equation 3.61 could be calculated since in the Gaussian case $\delta^+ + \delta^- = 2\varsigma$, where [59]

$$\varsigma = -\pi^{-1} \int_0^\infty x^{-2} \ln\left[\frac{2}{x^2}(1 - e^{-\frac{1}{2}x^2})\right] dx \approx 0.583. \tag{3.62}$$

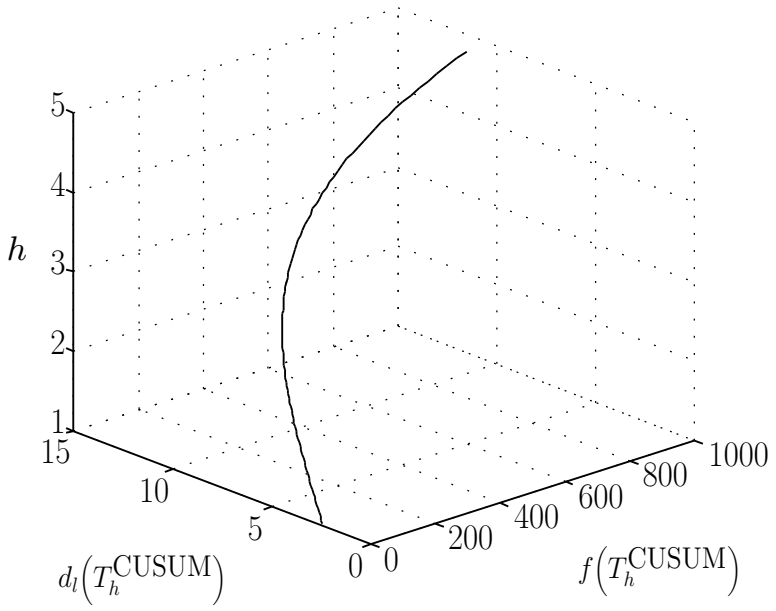


Figure 3.14: Exact values for Equation 3.52 and Equation 3.53 with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$ and $h \in [1,5]$.

The different ARL functions for $h = 5$ can be found in Figure 3.15.

3.6.2 Pollak's performance measure

The delay measure $d_l(T)$ introduced in Section 3.6.1 can also be replaced by *Pollak's performance measure*, which is equal to [56]

$$d_p(T) = \sup_{1 \leq \tau < \infty} \mathbb{E}_\tau[T - \tau | T \geq \tau], \tag{3.63}$$

which transforms the optimisation problem in Equation 3.54 into

$$\inf_{T \in \mathscr{T}} d_p(T) \text{ subject to } f(T) \geq \lambda. \tag{3.64}$$

A few stopping times have been proposed to solve Equation 3.54. The first stopping time of interest is the *Shiryaev-Roberts* stopping time, which is defined as [51, 55]

$$T_v^{\text{SR}} = \inf\{k \geq 0 | R_k \geq v\},$$

where

$$R_k = \begin{cases} (1 + R_{k-1}) \cdot s_k & \text{if } k > 0 \\ 0 & \text{if } k = 0. \end{cases}$$

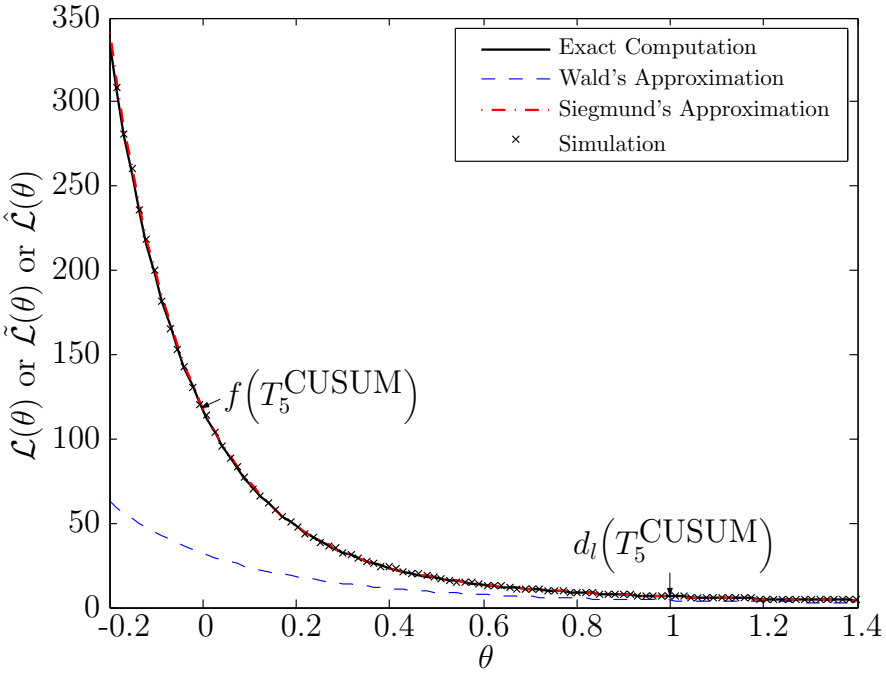


Figure 3.15: The different ARL functions with $q_0 \sim \mathcal{N}(0,1)$, $q_1 \sim \mathcal{N}(1,1)$, $h = 5$ and $\theta = [-0.2, 1.4]$.

The value of R_k can also be calculated non-iteratively via

$$R_k = \sum_{i=1}^{k} \prod_{j=i}^{k} s_j.$$

The *Shiryaev-Roberts-Pollak* stopping time is closely related to Equation 3.65. The only difference is that R_0 is not initialised with 0 but is assumed to be random with distribution equal to the quasi-stationary distribution of R_k. Another stopping time is the *deterministic Shiryaev-Roberts* stopping time. The deterministic Shiryaev-Roberts method once again considers the SR statistic R_0 to be deterministic, but not necessarily equal to zero [58]. It was shown that the Shiryaev-Roberts stopping time and the Shiryaev-Roberts-Pollak stopping time are suboptimal solutions to Equation 3.64 [57]. The Shiryaev-Roberts-Pollak stopping time is however an asymptotically optimal ($\lambda \to \infty$) solution of Equation 3.64 in an $\mathcal{O}(1)$ sense [56].

3.7 CONCLUSION

Many different algorithms were investigated in this chapter, of which only CUSUM (Section 3.6.1) and a Bayesian sequential detection (Section 3.4) variation (called *time-varying maximum likelihood classification* [23]) will be used in the remaining chapters. It is important to realize that even though

only a few of the algorithms are used directly in the remaining chapters, most of the theory in this chapter is important and is provided to derive (or understand) the algorithms that are used in the later chapters. The Neyman-Pearson (Section 3.1) result is critical to include in this chapter as it is the fundamental building block on which sequential analysis rests. The SPRT (Wald's formulation– Section 3.3) must be included as it provides the mathematical background needed to understand CUSUM, which is merely a repeated SPRT [59]. The SPRT algorithm also helps shed light on the Bayesian sequential detection problem. Section 3.5 (Bayesian quickest detection) and Section 3.6.2 (the Shiryaev-Roberts stopping time and its variants) are the only two sections that can be seen as non-critical and are included for the sake of completeness. The Bayesian quickest detection algorithm was not implemented on the datasets in Section 2.8, as the change point of the datasets was not geometrically distributed. As mentioned in Section 6.3 the Shiryaev-Roberts stopping time could still turn out to be useful (in the remote sensing field).

Listing 3.1: The pseudo-code for determining the OC and ASN functions via simulation.

```
N = 100000; %amount of sequences to generate for each theta
%parameter determining the density of the observable sequence
set theta equal to an experimental range;
S = 0; %the sum of the log-likelihood ratios
accept_H_0 = 0; %the amount of times H_0 was accepted
teller = 0; %samples required before a decision is made
%density_H_0(z) and ..._H_1(z) are the density functions
%of H_0 and H_1
delay = zeros(1,N); %vector of delays for each theta
Q = zeros(1,length(theta)); %the OC function
E_T = zeros(1,length(theta)); %the ASN function
fix h; fix a; %upper and lower boundaries of SPRT
for k = 1:length(theta) %iterate through theta
    accept_H_0 = 0; delay = zeros(1,N);
    for n = 1:N %perform N experiments
        exit = false; teller = 0; S = 0;
        while !exit %continue until exit boundaries are crossed
            teller = teller + 1;
            draw a z from density with parameter theta(k);
            s = log(density_H_1(z)/density_H_0(z)); S = S + s;
            if S >= h %crossed upper boundary
                delay(n) = teller; exit = true;
            end%if
            if S <= -a %crossed lower boundary
                accept_H_0 = accept_H_0 + 1;
                delay(n) = teller; exit = true;
            end%if
        end%while
    end%for
    Q(k) = accept_H_0/N; E_T(k) = mean(delay);
end%for
```

CHAPTER 4

HYPERTEMPORAL TECHNIQUES

The chapter provides the technical details of all the sequential and non-sequential hypertemporal classification and change detection algorithms that were investigated in this thesis. The chapter is divided into three main sections, namely *simulation* (Section 4.1), *classification* (Section 4.2) and *change detection* (Section 4.3).

Simulation is the creation of synthetic data in such a way that the synthetic data accurately represent real world data or phenomena. Simulated datasets are used for algorithm development, testing and validation, as well as for optimising instrument specifications. Simulated data are a valuable tool and is often used by the remote sensing community [26, 68]. Most remote sensing simulators are constructed by using a deductive approach, which means that they rely on the biophysical laws that govern the reflection of light [26, 27]. In Section 4.1.2, an inductive multispectral hypertemporal reflectance simulator is proposed. In contrast to deductive simulators, an inductive simulator uses a mathematical model that is built from the statistical properties of an existing dataset. The fact that an inductive model is built up from the statistical properties of an existing dataset enables an inductive model to augment datasets. The inductive simulator from Section 4.1.2 will be used to generate data for the data-intensive CUSUM algorithm presented in Section 4.3.3. The inductive model that will be used consists of two components, namely an SHO [3] (Section 4.1.1) to model the deterministic underlying noise-free signal and the Ornstein-Uhlenbeck process [4] (Section 4.1.2.1) to model the residual after the SHO has been subtracted. The two-component model will be referred to, in this chapter, as the CSHO [2], which is discussed in detail in Section 4.1.2.2. The possibility of using the parameters of the CSHO model as features for classification is discussed in Section 4.2.4.2.

Classification is the act of arranging or organising according to class or category. Land cover classi-

fication using remotely sensed data is a critical first step in large-scale environmental monitoring, resource management and regional planning [14]. A good review of different classification approaches is given in [14]. As mentioned in Chapter 1 the thesis focuses on hypertemporal classifiers. The minimum distance classifier [16], the time-varying maximum likelihood classifier [23], and the three feature groups $\boldsymbol{\iota}, \boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are discussed in Section 4.2.2, Section 4.2.3 and Section 4.2.4.2 respectively. The classification results obtained after applying these approaches to the datasets in Section 2.8 can be found in Chapter 5.

Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times. Essentially, it involves the ability to quantify temporal effects using multitemporal data [69]. There have been quite a number of reviews on change detection in the remote sensing field, namely [13, 69–73]. As mentioned in Chapter 1 the thesis also focuses on hypertemporal change detection techniques. The band differencing algorithm [7] and the windowless CUSUM algorithm [6] are discussed in Section 4.3.2 and Section 4.3.3 respectively. The change detection results obtained after applying the aforementioned techniques to the datasets in Section 2.8 are presented in Chapter 5.

4.1 SIMULATION

As stated in the previous section, most remote sensing simulators are constructed by using a deductive approach, which means that they employ the biophysical laws that govern the reflection of light [26, 27]. The simulator proposed in this chapter, however uses an inductive approach. An inductive approach tries to fit a mathematical model on the observed time-series directly, which is then used to simulate realistic reflectance values. The CSHO simulator proposed in this chapter is based on a stochastic inductive model. A stochastic inductive model tries to model the observed stochastic process, not just the noise-free underlying signal. The proposed simulator is not the first such approach used in the remote sensing literature [31]. Usually it is applied to a single time-series to enable forecasting, as is the case in [31]. The proposed simulator supplements [31], by making the concurrent simulation of multiple dependent time-series (multispectral) possible. On the other side of the inductive spectrum lies the complementary noise-free inductive models [74], which are used for noise reduction. The noise-free signals are then used to extract phenological markers. These two inductive approaches do not compete against each other, since they have different aims, noise reduction versus time-series generation.

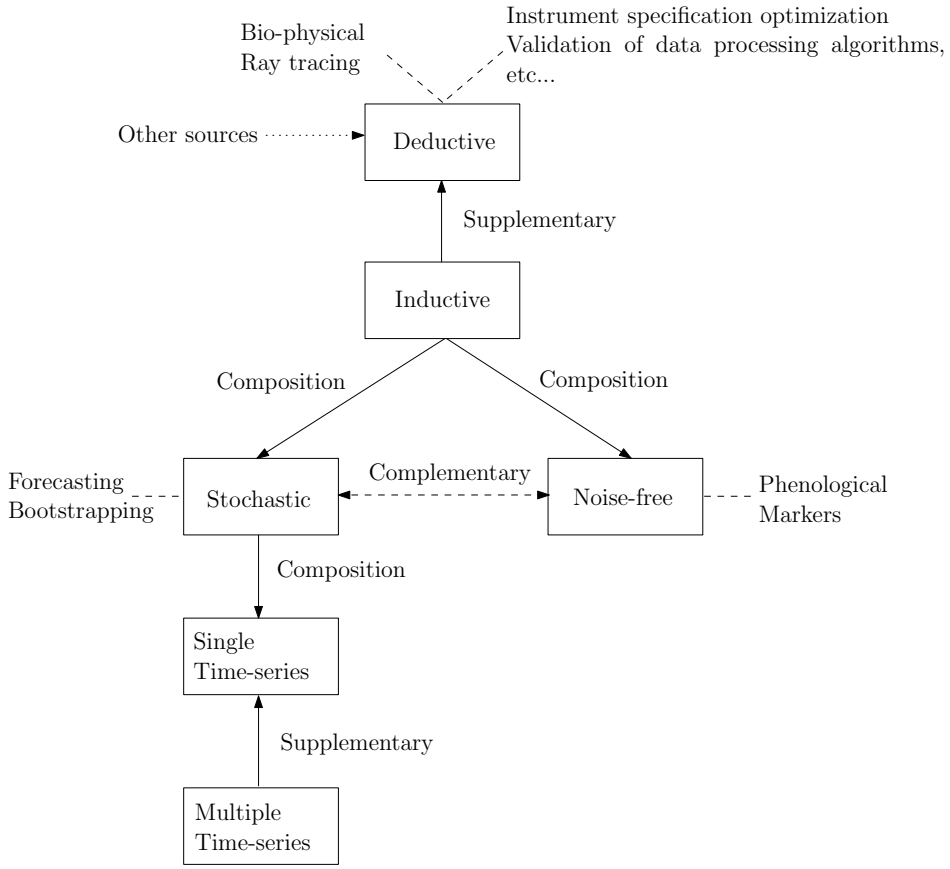


Figure 4.1: Functional (scientific) positioning of the proposed simulator.

The deductive simulators that were mentioned earlier use bio-physical parameters such as chlorophyll content, which can be derived from ancillary sources (for example direct measurement). An inductive simulator is an example of one such an ancillary source (in specific applications). For example, an inductive simulator could be used to forecast Leaf Area Index (LAI), which is a parameter required by the PROSPECT + Scattering by Arbitrary Inclined Leaves (PROSAIL) deductive simulator [27, 75]. Since an inductive simulator is actually a possible deductive simulator input source, they are not directly comparable. It is important to point out that inductive simulators are supplementary tools when used with deductive simulators, as they are not required by deductive simulators, which can function independently from inductive simulators. Figure 4.1 illustrates the scientific positioning of the simulator proposed in this chapter relative to existing simulators and models.

4.1.1 Noise-free inductive models

In this section, a short overview will be given of some of the different noise-free inductive models that are currently in use. The proposed stochastic inductive simulator uses an underlying inductive noise-free model (deterministic part) as its base. To make the simulator stochastic, a stochastic model is added to the deterministic base, which is the primary differentiating factor between noise-free modelling and stochastic modelling. In stochastic modelling the statistical properties of an observed class are replicated, while deterministic modelling wants to determine the shape of the underlying noise-free signal, and as such provides complementary functionality. The stochastic model does not necessarily require the best possible underlying noise-free model, as long as the model used for the residual preserves the statistical properties of the original signal.

The SHO model is an example of a noise-free inductive model [3] and is given by

$$A\sin(2\pi f_s t+\phi)+C, \tag{4.1}$$

where

$$\{A,C\}$$

are the harmonic features proposed by [5, 10] and $T_s=\frac{1}{f_s}$ is the period of the model. Many other models have been proposed as an improvement on the SHO model [31, 74, 76–78].

In particular, Carrão *et al.* [74] modelled MODIS time-series with a harmonic non-linear solution of a chaotic attractor

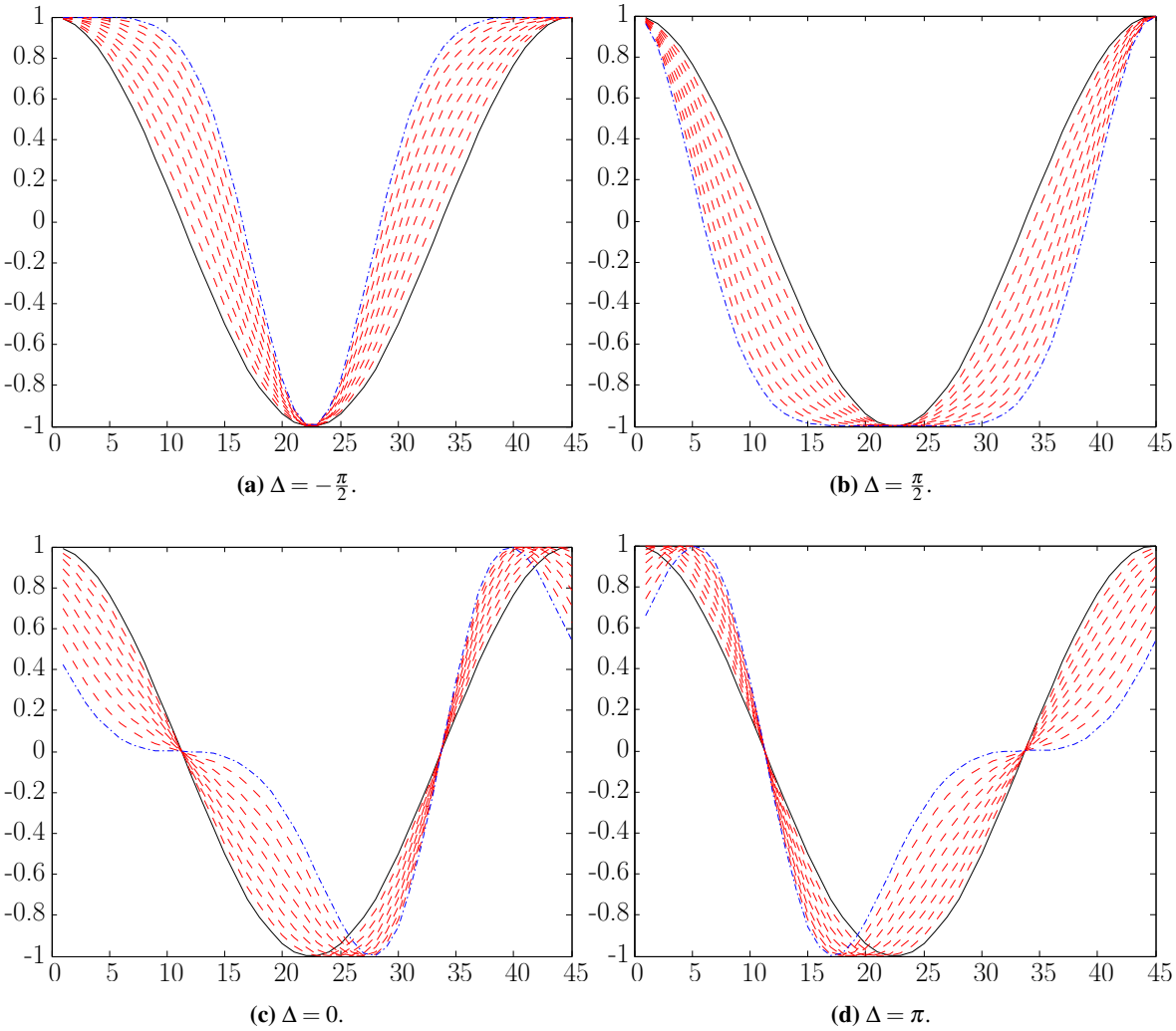
$$C+A\cos(2\pi f_s t+\phi+\alpha\cos(2\pi f_s t+\zeta)).$$

The function of each parameter used by Carrão's model is discussed below:

- C is a linear parameter that represents the annual mean of the model.
- A is the amplitude for the sine wave that fixes the peak deviation from the annual mean of the model.
- ϕ is the annual phase (produces a specific season of a given land cover class).
- α controls the non-linear strength of the model. When $\alpha=0$, the model reduces to a simple harmonic oscillator, whereas $\alpha>0$ introduces non-symmetry (bi-annual behaviour) in the model.

- ζ is the annual nonlinear phase. This phase allows time to "slow down" and to "accelerate" in order to reproduce asymmetries in variations (increases versus decreases).

Figure 4.2 illustrates the effect of the model parameters, as well as some of the different wave shapes that Carrão's model can represent.



(a) $\Delta = -\frac{\pi}{2}$. **(b)** $\Delta = \frac{\pi}{2}$. **(c)** $\Delta = 0$. **(d)** $\Delta = \pi$.

Figure 4.2: The equation for Carrão's model is, $C + A\cos(2\pi f_s t + \phi + \alpha\cos(2\pi f_s t + \zeta))$ and $\Delta = \phi - \zeta$. The blue line is when $\alpha = 1$, while the black line is when $\alpha = 0$ and the red lines are for $0 < \alpha < 1$. When the parameter $\alpha > 1$ Carrão's model will exhibit bi-annual variation. To get the specific graphs $\phi = 0$, which only functions as a translation parameter.

Kleynhans *et al.* [78] modelled NDVI time-series with a triply modulated cosine function

$$A(t)\sin(2\pi f_s t + \phi(t)) + C(t).$$

Jönsson *et al.* [76] modelled vegetation index time-series using Asymmetric Gaussian (AG) functions, while Zhang *et al.* [77] used piecewise-defined local Double Logistic (DL) functions. The AG and DL global model functions both have the following form

$$y(t) = \sum_{h=1}^{H} c_{1,h} + c_{2,h} g(t; A_h),$$

where H denotes the number of local model functions to use. The linear parameters $c_{1,h}$ and $c_{2,h}$ respectively govern the mean and the amplitude of the local function $g(t, A_h)$, while the meta-parameter $A_h = \{a_1, \cdots, a_r\}$ determines the shape of the local function $g(t, A_h)$. The AG local function is equal to

$$g(t, a_1, \cdots, a_5) = \begin{cases} \exp\left[-1\left(\frac{t-a_1}{a_2}\right)^{a_3}\right], \text{ if } t \geq a_1 \\ \exp\left[-1\left(\frac{a_1-t}{a_4}\right)^{a_5}\right], \text{ if } t < a_1. \end{cases} \tag{4.2}$$

In Equation 4.2, a_1 determines the position of the maxima (or minima) of g, while a_2 and a_3 determine the width and flatness of the right half of the function g. Similarly a_4 and a_5 determine the width and flatness of the left half of the function g. The DL local function is represented by

$$g(t; a_1, \cdots, a_4) = \frac{1}{1 + \exp\left(\frac{a_1 - t}{a_2}\right)} - \frac{1}{1 + \exp\left(\frac{a_3 - t}{a_4}\right)},$$

where a_1 and a_3 determine, respectively, the position of the left and right inflection points and a_2 and a_4 fix the rates of change at those points. The global AG and DL functions therefore respectively require $H \times 7$ and $H \times 6$ parameters.

Verbesselt *et al.* [19, 20] proposed a seasonal-trend model

$$C_1 + C_2 t + \sum_{j=1}^{k} A_j \sin(2\pi j f_s t + \phi_j),$$

where C_1 is the mean of the model, C_2 is the slope of the linear trend and A_j and ϕ_j are responsible for reproducing the seasonal behaviour.

From all the inductive models discussed up to this point the SHO was selected as the deterministic base of the proposed CSHO simulator. It is a well-known fact that the SHO model is a good first order noise-free model of a remotely sensed time-series [3].

A short discussion follows below to justify the SHO as the underlying noise-free model for the proposed simulator. In the discussion, reasons are provided for not selecting the other noise-free inductive models (in this section). Kleynhans's model is not a possible candidate for the deterministic base of the proposed simulator, since it is not a parsimonious model. The stochastic model that will be used

in the end should be parsimonious so that it can also be used to extract classification features. The seasonal-trend model is also not suitable, as the trend term of the model implies that the model should be used with a window. The model used in the end should be parsimonious and be able to simulate a multi-year time-series. It has been shown that Carrão's model is better than the remaining models, i.e. better than AG and DL, as using Carrão's model leads to lower fitting errors [74]. Carrão's model can be used as a simulator by adding white noise to it. All of the models mentioned (in particular Carrão's model) are definitely more accurate than the SHO model over a one-year window, but are also computationally more intensive than the Fourier transform used by the SHO.

In particular Carrão's method uses phase unwrapping, Levenberg-Marquardt $\times 2$ and Ordinary Least Squares (OLS) as functional blocks for estimating the parameters of the model [79, 80]. When the time-series becomes multi-year and there is inter-annual variation in the data, the long-term fitting-error made by the SHO is on average far less when compared to most of the other shapes that can be produced by Carrão's model (the SHO is one of the shapes Carrão's model can produce and is obtained when $\alpha = 0$). A summary of some of the shapes Carrão's model can generate can be found in Figure 4.2.

In other words, when the time-series becomes multi-year the SHO is actually a very good model candidate, while the extra versatility offered by Carrão's model becomes redundant (especially if the first harmonic component dominates the remaining harmonic components). In the case of multi-year time-series the increased accuracy (if any when compared to an SHO owing to the possibility of local minima, which is relevant for Levenberg-Marquardt), benefit obtained by using Carrão's model no longer outweighs the computational cost of the parameter estimation technique used by Carrão's method (compared to the SHO). It is also important to realize that when Carrão's model is used on each year individually, its parsimoniousness is compromised.

4.1.2 Proposed simulator

The proposed simulator uses the CSHO model. The CSHO consists of two components, a deterministic component and a stochastic component. The SHO is used for the deterministic component, while the Ornstein-Uhlenbeck process is used for the stochastic component. The Ornstein-Uhlenbeck process is used to model the remaining residual after the SHO is subtracted from the observed time-series. As the SHO is very general, there will be a high degree of dependence between the observations of the residual. The Ornstein-Uhlenbeck process can model a time-series with dependent observations

(first order), since this process is the continuous-time analogue of the discrete-time AR(1) process. The dependence implies colouredness, which is where the name of the simulator comes from. The Ornstein-Uhlenbeck process can be used to generate coloured noise as well as white noise [81]. The harmonic parameters of the SHO are estimated with the Fourier transform, while the parameters of the Ornstein-Uhlenbeck process are estimated with maximum likelihood parameter estimation. The objective of the CSHO simulator is to simulate multispectral time-series with an inherent correlation structure. In this thesis the simulator is used to augment datasets for data-intensive classification and change detection algorithms (Section 4.3.3). In selective cases, statistical inductive models similar to the CSHO have been used to forecast a single time-series [31]. The complex issue of incorporating multispectral correlation into a simulator was however not addressed in [31]. The CSHO simulator incorporates the average class noise correlation between the different spectral bands and reproduces class-specific spectral behaviour (spectral dependence) by enforcing the statistical restrictions imposed by different parameters (like mean and seasonal amplitude) of each spectral band in a class on one another.

In Section 4.1.2.1 the Ornstein-Uhlenbeck process is discussed, which is followed by Section 4.1.2.2 that discusses the CSHO in detail. Section 4.1.2.3 describes the algorithm used to estimate the parameters of the CSHO. The algorithm for simulating a MODIS pixel with the CSHO is presented in Section 4.1.2.7. The details of how the CSHO simulator enforces spectral dependence and correlation are presented in Section 4.1.2.4, Section 4.1.2.5 and Section 4.1.2.6.

4.1.2.1 Ornstein-Uhlenbeck

The Ornstein-Uhlenbeck process is widely used in mathematical finance for the modelling of the dynamics of interest rates and volatilities of asset prices. The Ornstein-Uhlenbeck process is the continuous-time analogue of the discrete time AR(1) process and, when initialised with the equilibrium distribution, is also stationary, Gaussian, Markovian and mean reverting. A stochastic process $\eta(t)$ is

- stationary if, for all $t_1 < t_2 < \cdots < t_n$ and $h > 0$, the random n-vectors $(\eta(t_1), \eta(t_2), \cdots, \eta(t_n))$ and $(\eta(t_1+h), \eta(t_2+h), \cdots, \eta(t_n+h))$ are identically distributed;
- Gaussian if, for all $t_1 < t_2 < \cdots < t_n$, the n-vector $(\eta(t_1), \eta(t_2), \cdots, \eta(t_n))$ is multi-variate normally distributed;

- Markovian if, $\forall B \in \mathbb{R}$ and for all $t_1 < t_2 < \cdots < t_n$, $P(\eta(t_n) \leq B|\eta(t_1), \eta(t_2), \cdots, \eta(t_{n-1})) = P(\eta(t_n) \leq B|\eta(t_{n-1}))$ (in lay man terms it means that the future is determined only by the present and not the past).

Moreover, the Ornstein-Uhlenbeck stochastic process satisfies the following stochastic differential equation:

$$d\eta(t) = \lambda(\mu - \eta(t))dt + \sigma dW(t), \tag{4.3}$$

where $\lambda > 0$ is the rate of mean reversion, μ is the long-term mean of the stochastic process, $\sigma > 0$ is the volatility or average magnitude, per square-root time, of the random fluctuations and $W(t)$ is a standard Brownian motion on $t \in [0, \infty]$, implying that $dW(t) \sim \mathcal{N}(0, \sqrt{dt})$. The solution to Equation 4.3 is given by

$$\eta(t) = \eta(0)e^{-\lambda t} + \mu(1 - e^{-\lambda t}) + \int_0^t \sigma e^{\lambda(s-t)} dW(s),$$

where the integral on the right-hand side is an Itô integral. The equilibrium density of the Ornstein-Uhlenbeck process is equal to $\mathcal{N}(\mu, \frac{\sigma^2}{2\lambda})$. If the random fluctuations in the process are ignored, it becomes clear that $\eta(t)$ has an overall drift towards the process mean μ. The process $\eta(t)$ reverts to the mean exponentially, at a rate λ, with a magnitude in direct proportion to the distance between the current value of $\eta(t)$ and μ [82].

4.1.2.2 Coloured Simple Harmonic Oscillator

Let $\mathbf{x}_c(t) = \{x_c^b(t)\}_{b \in \{1,\cdots,7\}}$ denote a MODIS pixel at time t with assigned class label $c \in \mathscr{C}$, where $x_c^b(t)$ denotes the b^{th} spectral band's reflectance at time t. The c is omitted if the class of the MODIS pixel is unknown.

Each observed signal belonging to the same class is a sample path of a stochastic process $X_c^b(t)$. Each MODIS class c is therefore modelled as a set of correlated (spectrally) stochastic processes $\mathbf{X}_c(t) = \{X_c^b(t)\}_{b \in \{1,\cdots,7\}}$. Since $X_c^b(t)$ is a stochastic process, an analytic expression can be assigned (if such an expression exists) to each sample path (MODIS pixel) $x_c^b(t; \boldsymbol{\theta}_c^b)$ of $X_c^b(t)$, where $\boldsymbol{\theta}_c^b$ is a set of random values with a joint probability density function. It is important to realise that real world MODIS pixels are also spatially correlated, while the proposed model assumes spatial independence.

The proposed analytic expression for each MODIS pixel in each band (sample path) is given by

$$x_c^b(t;\boldsymbol{\theta}_c^b) = s_c^b(t;\{A_c^b,\phi_c^b,C_c^b\}) + \eta_c^b(t;\{\mu_c^b,\lambda_c^b,\sigma_c^b\}), \tag{4.4}$$

where $s_c^b(t;\{A_c^b,\phi_c^b,C_c^b\})$ is the SHO model given in Equation 4.1 with period $T_s = \frac{1}{f_s} = 45$. The noise process $\eta_c^b(t;\{\mu_c^b,\lambda_c^b,\sigma_c^b\})$ is an Ornstein-Uhlenbeck process that satisfies the stochastic differential equation given in Equation 4.3.

For each class and band, it is expected that μ_c^b will be insignificant relative to C_c^b, as $\mu_c^b = 0$ if the parameter C_c^b is estimated without error. For convenience $\boldsymbol{\theta}_c^b$ will sometimes be omitted from $x_c^b(t;\boldsymbol{\theta}_c^b)$.

The distribution of $\boldsymbol{\theta}_c^b$ is determined by the parameter set $\{A_c^b,\phi_c^b,C_c^b,\lambda_c^b,\sigma_c^b\}$ and it follows that $\boldsymbol{\theta}_c = \{\boldsymbol{\theta}_c^b\}_{b\in\{1,\cdots,7\}} = \{A_c^b,\phi_c^b,C_c^b,\lambda_c^b,\sigma_c^b\}_{b\in\{1,\cdots,7\}} = \{\theta_1,\cdots,\theta_{35}\}$. The probability density function associated with $\boldsymbol{\theta}_c$ is denoted with $f_c(\boldsymbol{\theta}_c)$. When NDVI is included in the parameter set the notation $\tilde{\boldsymbol{\theta}}_c$ will be used. The same convention applies for $\tilde{\mathbf{X}}_c(t)$ and $\tilde{\mathbf{x}}_c(t)$. NDVI is excluded when constructing the probability density function $f_c(\boldsymbol{\theta}_c)$, since NDVI must be constructed from bands 1 and 2. The notation for a MODIS pixel (plus NDVI) is represented graphically in Figure 4.3 (where $\tilde{\mathbf{x}}[i]$ is the discrete analogue of $\tilde{\mathbf{x}}(t)$).

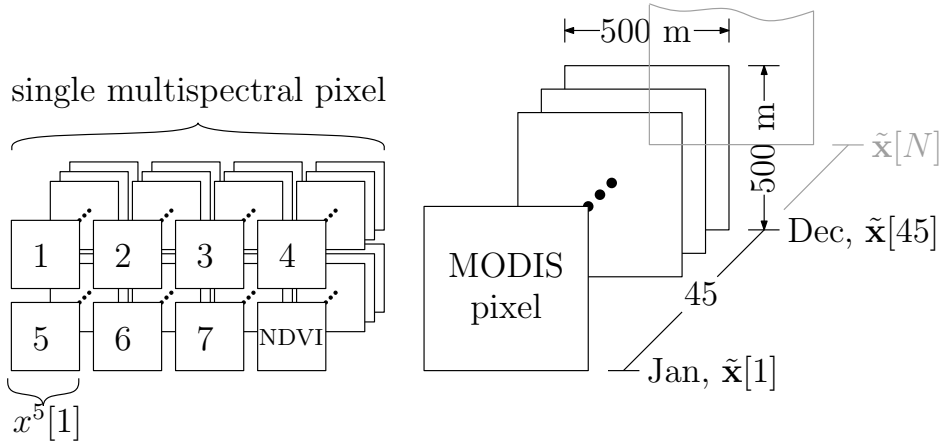


Figure 4.3: Time-series data representation for a single pixel, plus NDVI [2] © IEEE 2012.

The ensemble mean for $\tilde{\mathbf{X}}_c(t)$ is defined as

$$\tilde{\mathbf{y}}_c(t) = \{\mathbb{E}\left[X_c^b(t)\right]\}_{b\in\{1,\cdots,7,\text{NDVI}\}}, \tag{4.5}$$

and is assumed to be periodic, i.e. $\tilde{\mathbf{y}}_c(t) = \tilde{\mathbf{y}}_c(t+45j),\ \forall j \in \mathbb{N}$.

The autocorrelation of $\tilde{\mathbf{x}}_c(t)$ is defined as $\tilde{\mathbf{R}}_c(\tau) = \{\mathscr{R}_c^b(\tau)\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$, where

$$\mathscr{R}_c^b(\tau) = \frac{(x_c^b(t) - \mathbb{E}[x_c^b(t)])(x_c^b(t+\tau) - \mathbb{E}[x_c^b(t)])}{\text{var}(x_c^b(t))}. \tag{4.6}$$

Although the spatial correlation is not fully incorporated into the CSHO model it can still be quantified with the same notation. The spatial correlation of class c in band $b \in \{1,\cdots,7,\text{NDVI}\}$ can be represented with a correlation matrix $\boldsymbol{\rho}_b^c$, with elements

$$\rho_{b_{m,n}}^c = \frac{\mathbb{E}[(x_{m,c}^b(t) - \mathbb{E}[x_{m,c}^b(t)])(x_{n,c}^b(t) - \mathbb{E}[x_{n,c}^b(t)])]}{\text{std}(x_{m,c}^b(t))\text{std}(x_{n,c}^{b_n}(t))},$$

where $x_{m,c}^b(t)$ is the m-th pixel in a set of P MODIS pixels belonging to class c. The average spatial correlation is then equal to

$$\tilde{\boldsymbol{\rho}}^c = \mathbb{E}\{\{\boldsymbol{\rho}_b^c\}_{b\in\{1,\cdots,7,\text{NDVI}\}}\}. \tag{4.7}$$

The CSHO does enforce a limited amount of spatial correlation through $f_c(\boldsymbol{\theta}_c)$ (for instance the sample paths of the CSHO pixels are reasonably in phase, have slight differences in long-term mean and seasonal amplitude). As such CSHO pixels are less correlated (spatially) than the actual MODIS pixels.

4.1.2.3 Parameter estimation

To estimate the harmonic parameters of Equation 4.4 the Fourier transform is used, while the noise parameters will be estimated via maximum-likelihood parameter estimation. The Fourier transform $\mathscr{F}$ of an observed MODIS pixel $\tilde{\mathbf{x}}(t)$ is defined as

$$\tilde{\mathbb{X}}(f) = \{\mathscr{F}[x^b(t)]\}_{b\in\{1,\cdots,7,\text{NDVI}\}}.$$

The subscript c is omitted here, since the class to which the MODIS pixel belongs is unknown.

For each band b the harmonic parameters $\{\hat{A}^b, \hat{\phi}^b, \hat{C}^b\}$ are estimated as follows:

$$\hat{A}^b = 2|\mathscr{F}[x^b(t)](f_s)|$$

$$\hat{\phi}^b = \arg(\mathscr{F}[x^b(t)](f_s))$$

$$\hat{C}^b = |\mathscr{F}[x^b(t)](0)|.$$

In practice $\hat{\phi}^b$ is calculated by using a minimum squared error sinusoidal fit. The mean and amplitude of the sinusoid used to calculate $\hat{\phi}^b$ are set to $\hat{A}^b$ and $\hat{C}^b$ respectively.

The parameters $\hat{\mu}^b, \hat{\lambda}^b$ and $\hat{\sigma}^b$ for $x^b(t)$ are estimated by using maximum likelihood parameter estimation. The first step is to calculate the residual by using

$$\hat{\eta}^b(t) = x^b(t) - \hat{A}^b \sin(2\pi f t + \hat{\phi}^b) + \hat{C}^b.$$

Now let $\eta^b[i]$ be the discrete time analogue of $\eta^b(t)$, with Δt being the time step of $\eta^b[i]$, i.e. $t = i\Delta t$, and I the total number of discrete time samples that are available of $\eta^b(t)$. The log-likelihood function of $\eta^b[i]$ is given by [83]

$$L(\mu^b, \lambda^b, \bar{\sigma}^b) = -\frac{I}{2}\ln(2\pi) - I\ln(\bar{\sigma}^b) - \cdots$$
$$\cdots \frac{1}{2(\bar{\sigma}^b)^2}\sum_{i=1}^{I}\left[\eta^b[i] - \eta^b[i-1]\alpha^b - \mu^b(1-\alpha^b)\right]^2, \tag{4.8}$$

where

$$(\bar{\sigma}^b)^2 = (\sigma^b)^2\frac{1-e^{2\alpha^b}}{2\lambda^b} \tag{4.9}$$

and

$$\alpha^b = e^{-\lambda^b \Delta t}. \tag{4.10}$$

By respectively setting the partial derivative of Equation 4.8 with respect to $\mu^b, \lambda^b, \bar{\sigma}^b$ equal to 0 and respectively solving for $\mu^b, \lambda^b, \bar{\sigma}^b$, such that μ^b is independent of λ^b and $\bar{\sigma}^b$, the following maximum likelihood estimators are obtained

$$\hat{\mu}^b = \frac{\eta_l\eta_{kk} - \eta_k\eta_{kl}}{I(\eta_{kk} - \eta_{kl}) - (\eta_k^2 - \eta_k\eta_l)},$$

$$\hat{\lambda}^b = -\frac{1}{\Delta t}\ln\frac{\eta_{kl} - \hat{\mu}^b\eta_k - \hat{\mu}^b\eta_l + I(\hat{\mu}^b)^2}{\eta_{kk} - 2\hat{\mu}^b\eta_k + I(\hat{\mu}^b)^2},$$

$$\hat{\bar{\sigma}}^b = \frac{1}{I}[\eta_{ll} - 2\hat{\alpha}^b\eta_{kl} + (\hat{\alpha}^b)^2\eta_{kk}\cdots$$
$$-2\hat{\mu}^b(1-\hat{\alpha}^b)(\eta_l - \hat{\alpha}^b\eta_k) + I(\hat{\mu}^b)^2(1-\hat{\alpha}^b)^2],$$

with

$$\eta_k = \sum_{i=1}^{I}\hat{\eta}^b[i-1],\ \eta_l = \sum_{i=1}^{I}\hat{\eta}^b[i],$$

$$\eta_{kk} = \sum_{i=1}^{I}\hat{\eta}^b[i-1]^2,\ \eta_{kl} = \sum_{i=1}^{I}\hat{\eta}^b[i-1]\hat{\eta}^b[i],\ \eta_{ll} = \sum_{i=1}^{I}\hat{\eta}^b[i]^2,$$

where the relation between $\hat{\sigma}^b$ and $\hat{\bar{\sigma}}^b$ is defined in the same way as in Equation 4.9 and $\hat{\alpha}^b$ is defined in the same manner as in Equation 4.10. The estimated parameters can now be used as classification features.

4.1.2.4 Parameter probability density function

All the estimated parameters (of all pixels in a specific class) are represented by the vector $\boldsymbol{\Theta}_c = \{\Theta_1, \Theta_2, \cdots, \Theta_{35}\}$, where Θ_i is a random variable and θ_i (or rather $\hat{\theta}_i$) is a realisation of it. Note that NDVI is excluded from the parameter probability density function, as it is created from MODIS land bands 1 and 2. The joint density of $\boldsymbol{\Theta}_c$ is assumed to be Gaussian distributed and expressed with

$$f_c(\boldsymbol{\theta}_c) = \frac{1}{\sqrt{(2\pi)^{|\boldsymbol{\theta}_c|}|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_c - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_c - \boldsymbol{\mu})\right]. \tag{4.11}$$

In Equation 4.11, $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\Theta}_c]$ and $\boldsymbol{\Sigma}$ is the covariance matrix with elements $\Sigma_{n,m} = \mathbb{E}[(\Theta_n - \mu_{\Theta_n})(\Theta_m - \mu_{\Theta_m})]$, $\forall m,n \in \{1, \cdots, |\boldsymbol{\theta}_c|\}$.

4.1.2.5 Parameter and noise correlation

The parameter correlation matrix $\boldsymbol{P}_p^c$ has elements $P_{n,m} = \frac{\mathbb{E}[(\Theta_n - \mu_{\Theta_n})(\Theta_m - \mu_{\Theta_m})]}{\sigma_{\Theta_n}\sigma_{\Theta_m}}$, $\forall m,n \in \{1, \cdots, |\boldsymbol{\theta}_c|\}$. The parameter correlation matrix $\boldsymbol{P}_p^c$ is used to get an indication of the dependence between the model parameters of each class and is used to model class-specific spectral behaviour.

In addition to $\boldsymbol{P}_p^c$, the noise correlation $\boldsymbol{P}_\eta^c$ is measured between the different MODIS bands. To determine the noise correlation, $dW^b(t)$ from Equation 4.3 needs to be estimated, since $dW^b(t)$ induces the random behaviour in the noise. To estimate $dW^b(t)$, $\eta^b(t)$ is discretised with timesteps of length Δt. An exact formula that holds for $\Delta t = 1$ is [83]

$$\eta^b[i] = e^{-\lambda^b}\eta^b[i-1] + (1 - e^{-\lambda^b})\mu^b + \sigma^b\sqrt{\frac{(1 - e^{-2\lambda^b})}{2\lambda^b}}\Delta W^b[i], \tag{4.12}$$

where $\Delta W^b[i] \sim \mathcal{N}(0,1)$ and is equal to $\Delta W^b[i] = W^b[i] - W^b[i-1]$.

By making $\Delta W^b[i]$ the subject of Equation 4.12, it can be used to estimate (or approximate) the *independent*, normally distributed innovation terms for each timestep of each MODIS band. This, in turn, allows the computation of the correlation matrix $\boldsymbol{P}_\eta^c$ of the innovation terms across the spectral bands with $P_{n,m} = \frac{\mathbb{E}[(\Omega_n - \mu_{\Omega_n})(\Omega_m - \mu_{\Omega_m})]}{\sigma_{\Omega_n}\sigma_{\Omega_m}}$, $\forall m,n \in \{1, \cdots, 7\}$, where Ω_n is the random variable with realisations ΔW^n and n refers to the MODIS band.

4.1.2.6 Generating correlated innovations

Independent, correlated innovations are generated by following the approach presented in [84]. Consider d independent standard (i.e. unit variance) white noise processes $\overline{\Delta W}^1, \ldots, \overline{\Delta W}^d$ each of length I, where I is the number of observations that needs to be simulated. Let furthermore a (deterministic and constant) matrix

$$\boldsymbol{\delta} = \begin{bmatrix} \delta_{11} & \delta_{12} & \cdots & \delta_{1d} \\ \delta_{21} & \delta_{22} & \cdots & \delta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{71} & \delta_{72} & \cdots & \delta_{7d} \end{bmatrix}$$

be given, and consider the seven-dimensional process $\boldsymbol{\Delta W}_c$, defined by

$$\boldsymbol{\Delta W}_c = \boldsymbol{\delta}\overline{\boldsymbol{\Delta W}}, \tag{4.13}$$

where

$$\boldsymbol{\Delta W}_c = [\Delta W_c^1, \cdots, \Delta W_c^7]^T.$$

Moreover, assume that the rows of $\boldsymbol{\delta}$ have unit length, i.e.

$$\|\delta_{i\#}\|_2 = 1, \quad i = 1, \ldots, 7. \tag{4.14}$$

Then each of the components $\Delta W_c^1, \cdots, \Delta W_c^7$ separately is also a standard (i.e. unit variance) white noise process, with instantaneous correlation given by

$$\boldsymbol{P}_\eta^c = \boldsymbol{\delta}\boldsymbol{\delta}^*. \tag{4.15}$$

Given a positive definite correlation matrix $\boldsymbol{P}_\eta^c$, $\boldsymbol{\delta}$ can be obtained by using Cholesky factorisation (Section A.3) [40], such that Equation 4.14 is automatically satisfied.

4.1.2.7 Simulating a MODIS pixel

Let $\boldsymbol{\sigma}_c = \{\sigma_c^b\}_{b\in\{1,\cdots,7\}}$, $\boldsymbol{\lambda}_c = \{\lambda_c^b\}_{b\in\{1,\cdots,7\}}$, $\boldsymbol{\mu}_c = \{\mu_c^b\}_{b\in\{1,\cdots,7\}}$, $\boldsymbol{C}_c = \{C_c^b\}_{b\in\{1,\cdots,7\}}$, $\boldsymbol{A}_c = \{A_c^b\}_{b\in\{1,\cdots,7\}}$, $\boldsymbol{\phi}_c = \{\phi_c^b\}_{b\in\{1,\cdots,7\}}$, $\mathbf{s}_c(t) = \{s_c^b(t)\}_{b\in\{1,\cdots,7\}}$ and $\boldsymbol{\eta}_c(t) = \{\eta(t)_c^b\}_{b\in\{1,\cdots,7\}}$. If the CSHO model is used to simulate a MODIS pixel which belongs to class c the following steps are required:

1. Draw $\boldsymbol{\theta}_c$ randomly from $f_c(\boldsymbol{\theta}_c)$ (assuming that $f_c(\boldsymbol{\theta}_c)$ has already been constructed by using the procedure discussed in Section 4.1.2.4).

2. Generate correlated seven-dimensional innovations $\boldsymbol{\Delta W}_c$ that are characterised by the correlation matrix $\boldsymbol{P}_\eta^c$ (assuming that $\boldsymbol{P}_\eta^c$ has already been estimated via the procedure discussed in Section 4.1.2.5) by using the procedure discussed in Section 4.1.2.6.

3. Use $\boldsymbol{\sigma}_c$ and $\boldsymbol{\lambda}_c$ (from $\boldsymbol{\theta}_c$) together with $\boldsymbol{\Delta W}_c$ and Equation 4.12 to generate $\boldsymbol{\eta}_c(t)$ under the assumption that $\boldsymbol{\mu}_c = 0$ and $\boldsymbol{\eta}_c(0) = \mathbf{0}$. The first 45 observations must be ignored, to allow $\boldsymbol{\eta}_c(t)$ to reach a state of equilibrium.

4. Use $\boldsymbol{C}_c$, $\boldsymbol{A}_c$ and $\boldsymbol{\phi}_c$ (from $\boldsymbol{\theta}_c$) and Equation 4.1 to generate $\mathbf{s}_c(t)$.

5. Generate $\mathbf{x}_c(t)$ using $\mathbf{s}_c(t)$,$\boldsymbol{\eta}_c(t)$ and Equation 4.4.

6. Generate NDVI from $x_c^1(t)$ and $x_c^2(t)$.

4.2 CLASSIFICATION

As mentioned in the chapter introduction, *classification* is the act of arranging or organising according to class or category. Land cover classification using remotely sensed data is a critical first step in large-scale environmental monitoring, resource management and regional planning [14]. At this point it is prudent to point out the subtle difference between *land cover* and *land use*. Land cover refers to the (physical) surface cover, such as vegetation, urban infrastructure, water, bare soil etc., whereas land use refers to the (functional) purpose that the land serves, such as agriculture, recreation, or wildlife habitat [23].

The main focus of this section will be on land cover classification. In Section 4.2.1 a short literature review is given of land cover classification techniques, followed by the presentation of three hypertemporal classifiers in Section 4.2.2, Section 4.2.3 and Section 4.2.4. The CSHO feature set is discussed in Section 4.2.4.2.

4.2.1 Literature review

According to [85] there are two types of analytic approaches for creating land cover maps, namely *photointerpretation* and *machine analysis*. Photointerpretation relies on a human analyst to interpret an enhanced image. Machine analysis on the other hand, uses statistical or numerical algorithms to perform the labelling of multispectral datasets. A good review of different machine analysis techniques (henceforth described only as classification techniques) is available in [14]. According to [14], remote sensing classification approaches can be grouped using a taxonomy. The proposed taxonomy in [14] can be found in Figure 4.4. A short description of each category found in Figure 4.4 is given in [14, 23]. In [14], *the classification elements* are used as the primary attribute for grouping classification techniques together. In this section the classification of elements will also be used as the primary attribute for grouping classification techniques together. The different classification elements

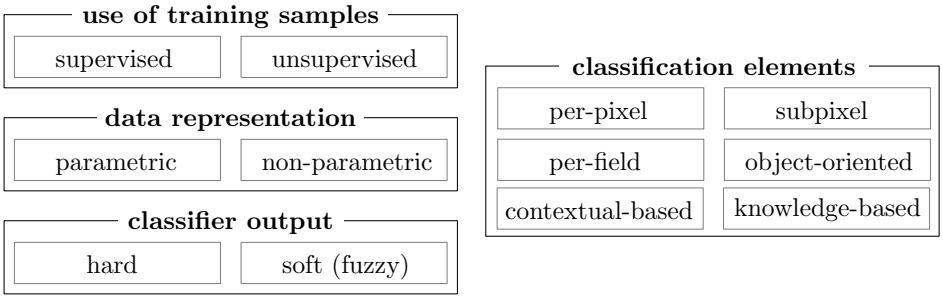


Figure 4.4: A taxonomy of fixed sample size land cover classification techniques (from [23]).

and some of the most popular algorithms used by each type of element are discussed in the following list:

1. In *per-pixel* classification, each pixel is classified as belonging to a specific land cover class. Per-pixel classification clearly assumes homogeneous pixels, which becomes unrealistic when the spatial resolution is decreased [23]. Per-pixel classifiers can be parametric or non-parametric. *Maximum likelihood classification* is probably the most commonly used *per-pixel* parametric classification approach and is presented clearly in [85]. The most frequently used non-parametric per-pixel classifiers are *Artificial Neural Networks (ANNs)* [86], *decision trees* [87, 88] and *SVMs* [89–91]. There are some remaining advanced per-pixel-based classification algorithms worth mentioning, which include: the *spectral angle classifier* [92], *Independent Component Analysis (ICA)* [93, 94], a *model-based approach* [95, 96] and several *nearest neighbour approaches* [97–99].

2. *Subpixel* techniques are especially used with medium or coarse resolution remote sensing data, since heterogeneous pixels are quite common at those spatial resolutions. Typically subpixel classification is done with either *fuzzy sets* [100, 101] or *Spectral Mixture Analysis (SMA)* [102–104]. Other prominent approaches to subpixel classification include ANNs [105], *Dempster-Shafer theory*, *certainty factors* [106] and a maximum likelihood approach [107].

3. One way to to handle pixel heterogeneity is to employ *per-field* classification. In per-field classification pixels are no longer evaluated individually, but in "fields" consisting of the same land cover type, such that the noise can be averaged out over larger areas, implying that the fields are more homogeneous than the pixels that make up the fields (see for example [108, 109]). *Object-oriented* classification is similar to per-field-based classification. The main difference is that object-oriented methods use only raster data, whereas per-field approaches use vector and raster data. The reference list [110–112], provides additional information on object-oriented classification. A frequently used object-oriented approach is eCognition, which is described in (among others) [113].

4. *Contextual-based* approaches to land cover classification take the spatial distribution of pixels into account in an attempt to minimise the effects of intra-class variations [114]. In [115] a selection of early ad hoc contextually based classifiers are compared. More recently it has been shown that the Markov and Gibbs random fields are effective approaches that can use spatial information [116, 117]. Markov and Gibbs random fields were introduced to image processing by the seminal paper [118]. There are also spectral-contextual classifiers of which [119] is a good example.

5. *Knowledge-based* methods use ancillary data sources (such as a digital elevation map, a soil map, housing, etc.) on top of the contextual information that is available for a region to perform classification (see [120] for an example).

4.2.1.1 Dimensionality reduction

The large amount of training data that hyperspectral data provide needs to be reduced, as classifiers that use large training datasets become impractical very quickly [23]. An effective way of reducing training datasets is to use *dimensionality reduction*, which is closely related to *feature extraction*. Dimensionality reduction algorithms have to be able to select the most prevailing elements from a

dataset, while skipping the unimportant elements. Several approaches to dimensionality reduction exist, including *Principal Component Analysis (PCA)*, *minimum noise fraction transform*, *discriminant analysis* [121–123], *decision boundary feature extraction* [124], *Gaussian mixture model feature extraction* [95], *wavelet transform* [125] and SMA [126].

4.2.1.2 Hypertemporal classification

Most of the classification techniques discussed in the literature review up to now have been single-date classifiers. It has been shown that multitemporal and hypertemporal classification is more reliable than single-date classification [15, 127, 128], since single-date reflectance values between different classes may be unseparable due to the fact that land-cover classes could have similar spectral characteristics during certain times of the year [15]. A second reason that motivates hypertemporal classification is that most of the earth (landmass) is covered by vegetation. Vegetation species have unique phenologies, which make remote classification possible [39]. The most prominent hypertemporal classification techniques in literature are PCA [17,129,130], phenological metrics [18,131,132], Fourier analysis [5, 133–136], wavelet analysis [137], minimum distance classification [16, 23] and time-varying maximum likelihood classification [23].

The chapter focuses on hypertemporal classification techniques. In particular it revolves around the parameters of the CSHO. The parameters of the CSHO will be used as features that will be fed into an SVM classifier. The proposed technique extends the approach in [5], which is based on Fourier features. In [5], it is shown that efficient separability can in fact be achieved when using only the mean and seasonal harmonic components. The Ornstein-Uhlenbeck process, which is a component of the CSHO, summarises the less important Fourier features that by themselves do not contribute significantly to classification up with two average model parameters that could possibly contribute significantly to classification accuracy. The SVM classifier with CSHO features is compared to the minimum distance classifier [16], the time-varying model classifier [23] and SVMs fed with temporal and harmonic features in Section 5.3.

4.2.2 Minimum distance classifier

The minimum distance classifier classifies the observed signal $\tilde{\mathbf{x}}(t)$ as class c by choosing the class with the lowest model error [16, 23]. Where the model error for each class c is defined as the accumulated euclidean distance between the observed signal $\tilde{\mathbf{x}}(t)$ and the signal model (yearly ensemble

mean) $\tilde{\mathbf{y}}_c(t)$, mathematically it can be written as, find a c such that the following optimisation problem is minimised:

$$\inf_{c\in\mathscr{C}} \int_0^I \|\tilde{\mathbf{x}}(t) - \tilde{\mathbf{y}}_c(t)\|_2 \, dt.$$

Any subset of $\tilde{\mathbf{x}}(t)$ and $\tilde{\mathbf{y}}_c(t)$ can be used for classification, as long as both subsets are constructed from the same spectral bands. The euclidean differences are normalised with the difference between the maximum and minimum observed value in each band.

4.2.3 Time-varying maximum likelihood classifier

The time-varying maximum likelihood classifier uses the time-varying model [23]. The background theory used in this section was discussed in detail in Section 3.4. The time-varying model is a discrete model and $\tilde{\mathbf{X}}_c(t)$ therefore needs to be discretised. Let the discretised form of $\tilde{\mathbf{X}}_c(t)$ be denoted by $\{\tilde{\mathbf{X}}_c[k]\}_{k=\{1,2,\cdots\}}$. The time-varying model is equivalent to the first order statistical description of $\{\tilde{\mathbf{X}}_c[k]\}_{k=\{1,2,\cdots\}}$. Usually the phrase "first order statistical description" is only associated with a single stochastic process, but here the first order statistical description is connected with a set of stochastic processes. The first order statistical description of $\{\tilde{\mathbf{X}}_c[k]\}_{k=\{1,2,\cdots\}}$ is equal to the set of probability density functions at each time step k, $\{q_k^c\}_{k=\{1,2\cdots\}}$. If it is assumed that the MODIS data contain no inter-annual variation, in other words it is assumed that the MODIS time-series is periodic (45 observations in a year), then it is true that $q_k^c = q_{k+45n}^c$, $n = \{1,2,\cdots\}$. Note that q_k^c is an eighth-dimensional density and that the density at k can also be constructed for a smaller number of bands. When only a subset of the bands is used the notation $q_k^{c,\mathbf{b}}$ is used, where $\mathbf{b}$ can be any subset of $\{1,\cdots,7,\text{NDVI}\}$. The same rule applies for $\tilde{\mathbf{X}}_c^{\mathbf{b}}(t)$ and $\tilde{\mathbf{x}}_c^{\mathbf{b}}(t)$. Assume now that the class label c is equal to either a $v \equiv 0$ or a $s \equiv 1$ if the observed MODIS pixel belongs to either the vegetation or settlement class. Any unlabelled MODIS pixel $\{\tilde{\mathbf{x}}^{\mathbf{b}}[k]\}_{k\in\mathbb{N}}$ obeys one of two statistical hypotheses:

$$\mathscr{H}_0 : \tilde{\mathbf{x}}^{\mathbf{b}}[k] \sim Q_k^{0,\mathbf{b}},\ k = 1,2,\cdots$$

versus

$$\mathscr{H}_1 : \tilde{\mathbf{x}}^{\mathbf{b}}[k] \sim Q_k^{1,\mathbf{b}},\ k = 1,2,\cdots;$$

where for each time step k, $Q_k^{0,\mathbf{b}}$ and $Q_k^{1,\mathbf{b}}$ are two $|\mathbf{b}|$-dimensional probability distributions with associated densities $q_k^{0,\mathbf{b}}$ and $q_k^{1,\mathbf{b}}$, respectively. Further assume that hypothesis $\mathscr{H}_1$ occurs with prior probability π and $\mathscr{H}_0$ with prior probability $1-\pi$.

Now define the posterior sequence to be

$$\pi_k^{\pi} = \frac{\pi_{k-1}^{\pi} q_k^{1,\mathbf{b}}(\tilde{\mathbf{x}}^{\mathbf{b}}[k])}{\pi_{k-1}^{\pi} q_k^{1,\mathbf{b}}(\tilde{\mathbf{x}}^{\mathbf{b}}[k]) + (1-\pi_{k-1}^{\pi}) q_k^{0,\mathbf{b}}(\tilde{\mathbf{x}}^{\mathbf{b}}[k])}, \; k = 1,2,\cdots,$$

where $\pi_0^{\pi} = \pi$. The maximum likelihood classification of the time-varying classification task is then given by

$$\delta_k = \begin{cases} 0, \text{ if } \pi_k^{\pi} \leq 0.5 \\ 1, \text{ if } \pi_k^{\pi} > 0.5. \end{cases}$$

If thresholds are introduced to the time-varying maximum likelihood classifier then the time-varying maximum likelihood classifier becomes sequential in nature. Let $\{\pi_{\mathrm{U}}, \pi_{\mathrm{L}}\}$ be those thresholds. If π_k^{π} crosses $\{\pi_{\mathrm{U}}, \pi_{\mathrm{L}}\}$ a decision can be made. The decision rule now becomes

$$\delta_k = \begin{cases} 0, \text{ if } \pi_k^{\pi} \leq \pi_{\mathrm{L}} \\ 1, \text{ if } \pi_k^{\pi} > \pi_{\mathrm{U}}. \end{cases}$$

It can easily be shown (see Section 3.4.2 for more details) that the sequential time-varying maximum likelihood classifier is equivalent to the time-varying SPRT (in terms of classification accuracy and delay), where the time-varying SPRT is obtained by casting the classification problem presented in this section into the likelihood domain. Currently the classification problem is solved using a posterior sequence.

4.2.4 Support Vector Machine

An SVM works by creating a hyperplane or set of hyperplanes in a high or infinite dimensional space, and as such can be used for classification, regression, or to perform other similar functions [1, 138, 139]. An SVM works on the principle of finding a hyperplane, such that the hyperplane has the furthest distance from the training data of any class (which is known as the functional margin). The training data for the classifier are a set of n points of the form

$$\mathscr{D} = \{(\mathbf{r}^{(i)}, \psi^{(i)} | \mathbf{r}^{(i)} \in \mathbb{R}^p, \psi^{(i)} \in \{-1,1\}\} \tag{4.16}$$

where $\psi^{(i)}$ is a label denoting class membership, and $\mathbf{r}^{(i)}$ is a p-dimensional real feature vector. If the data are linearly separable, a maximum-margin hyperplane is calculated to divide the data into points belonging to the class with label -1 or 1 perfectly. The maximum-margin hyperplane is represented by the following

$$\mathbf{w}^T \cdot \mathbf{r} + b = 0, \tag{4.17}$$

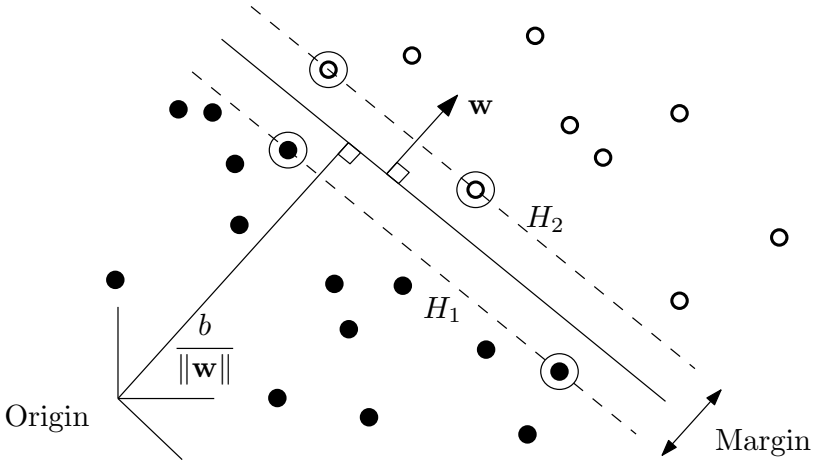


Figure 4.5: Example of a maximum-margin hyperplane of a linear SVM (from [23]).

where vector $\mathbf{w}$ is perpendicular to hyperplane, Equation 4.17, and $\frac{b}{||\mathbf{w}||}$ is the offset of hyperplane, Equation 4.17, from the origin in the direction of $\mathbf{w}$. The maximum-margin hyperplane is calculated by choosing $\mathbf{w}$ and b to maximise the distance between the hyperplanes $\mathbf{w}^T \cdot \mathbf{x} + b = -1$ (which corresponds to hyperplane H_1 in Figure 4.5) and $\mathbf{w}^T \cdot \mathbf{x} + b = 1$ (which corresponds to hyperplane H_2 in Figure 4.5). These two hyperplanes are as far a part as possible although they still correctly classify each training data point. The problem of maximising the distance between the hyperplanes $\mathbf{w}^T \cdot \mathbf{x} + b = -1$ and $\mathbf{w}^T \cdot \mathbf{x} + b = 1$ reduces to the following optimisation problem:

$$\begin{aligned} &\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \\ &\text{s.t. } \psi_i(\mathbf{w}^T \cdot \mathbf{r} + b) \geq 1,\ i = 1, \cdots, n. \end{aligned} \tag{4.18}$$

Equation 4.18 is known as the *primal problem*. The optimisation problem is actually solved by using the so-called *dual problem*, which is the Lagrangian reformulation of the primal problem. There are mainly two reasons for rather solving the dual problem, namely the constraints of Equation 4.18 are supplanted by constraints of the Lagrange multipliers themselves, which are much easier to deal with, and in the dual problem inner products are used in both the training and testing algorithms, which makes it possible to effortlessly generalise to non-linear SVMs [23]. Readers interested in the dual problem are referred to [139], which is a comprehensive tutorial on SVMs. To extend the

approach to non-separable datasets the optimisation problem is reformulated to obtain:

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t. } \psi_i(\mathbf{w}^T \cdot \mathbf{r} + b) \geq 1 - \xi_i,\ i = 1,\cdots,n$$
$$\xi_i \geq 0,\ i = 1,\cdots,n.$$

The method entails, introducing slack variables, ξ_i, which measure the degree of misclassification of $\mathbf{r}_i$. The parameter C controls the relative weighting between the slack variables and the goal $\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$.

An SVM was chosen as classification technique since SVMs, unlike neural networks, are robust to the over-fitting problem (increased spectral view increases feature set sizes). The first documented use of SVMs in remote sensing was in [140]. A good review of the application of SVMs in the remote sensing field can be found in [91], of which [89, 141–143] are worth singling out. It is also worth mentioning [15, 144–148], as these works applied SVMs to MODIS data.

4.2.4.1 Example

Consider the following linearly separable binary classification problem:

$$\boldsymbol{\rho} = [w_1, w_2, b]^T, \mathbf{w} = [w_1, w_2]^T.$$
$$\mathbf{r} = [r_1, r_2]^T.$$

The aim is to find a hyperplane $\mathbf{w}^T\mathbf{r} + b = 0$ that separates the binary classes perfectly, while the margin between $\mathbf{w}^T\mathbf{r} + b = -1$ and $\mathbf{w}^T\mathbf{r} + b = 1$ is also maximised. The following six training examples are given,

$$\begin{aligned}
\mathbf{r}^{(1)} &= [1,1]^T = (1,1),\\
\mathbf{r}^{(2)} &= [1,2]^T = (1,2),\\
\mathbf{r}^{(3)} &= [2,1]^T = (2,1),\\
\mathbf{r}^{(4)} &= [3,3]^T = (3,3),\\
\mathbf{r}^{(5)} &= [2,4]^T = (2,4),\\
\mathbf{r}^{(6)} &= [4,5]^T = (4,5).
\end{aligned}$$

With classification given by the label $\psi^{(i)} \in \{-1,+1\}$ for each training example,

$$\boldsymbol{\psi} = [-1,-1,-1,1,1,1].$$

The given binary classification problem is represented graphically in Figure 4.6. The hyperplane

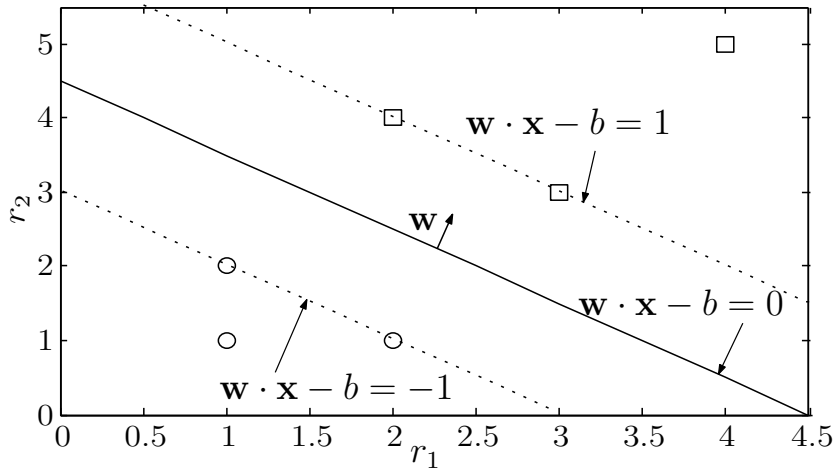


Figure 4.6: An example of an SVM classification problem.

$\mathbf{w}^T\mathbf{r} + b = 0$ for this example can be found by solving the following primal minimisation problem.

$$\begin{aligned}
\text{Minimize} \quad & f(\boldsymbol{\rho}) = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^T\mathbf{w}, \\
\text{such that} \quad & g_1(\boldsymbol{\rho}) = w_1 + w_2 + b + 1 \leq 0, \\
& g_2(\boldsymbol{\rho}) = w_1 + 2w_2 + b + 1 \leq 0, \\
& g_3(\boldsymbol{\rho}) = 2w_1 + w_2 + b + 1 \leq 0, \\
& g_4(\boldsymbol{\rho}) = -2w_1 - 4w_2 - b + 1 \leq 0, \\
& g_5(\boldsymbol{\rho}) = -3w_1 - 3w_2 - b + 1 \leq 0, \\
& g_6(\boldsymbol{\rho}) = -4w_1 - 5w_2 - b + 1 \leq 0,
\end{aligned} \tag{4.19}$$

which can be transformed to have only *equality constraints* by introducing the auxiliary variables, κ_j. After introducing the auxiliary variables Equation 4.19 reduces to:

$$\begin{aligned}
\text{minimise} \quad & f(\boldsymbol{\rho}) = \frac{1}{2}\|\mathbf{w}\|^2 \\
\text{such that} \quad & h_1(\boldsymbol{\rho}) = w_1 + w_2 + b + 1 + \kappa_1^2 = 0, \\
& h_2(\boldsymbol{\rho}) = w_1 + 2w_2 + b + 1 + \kappa_2^2 = 0, \\
& h_3(\boldsymbol{\rho}) = 2w_1 + w_2 + b + 1 + \kappa_3^2 = 0, \\
& h_4(\boldsymbol{\rho}) = -2w_1 - 4w_2 - b + 1 + \kappa_4^2 = 0, \\
& h_5(\boldsymbol{\rho}) = -3w_1 - 3w_2 - b + 1 + \kappa_5^2 = 0, \\
& h_6(\boldsymbol{\rho}) = -4w_1 - 5w_2 - b + 1 + \kappa_6^2 = 0.
\end{aligned} \tag{4.20}$$

The constraints in Equation 4.19 follow from the requirement that

$$\psi^{(i)}\left(\mathbf{w}^T\mathbf{r}^{(i)} + b\right) \geq 1, \ \forall i = 1, \ldots, 6. \tag{4.21}$$

Now the Lagrange function (Section A.4) [40] of Equation 4.20 can easily be constructed, which is equal to

$$\mathscr{L}(\boldsymbol{\rho}, \boldsymbol{\kappa}, \boldsymbol{\lambda}) = f(\boldsymbol{\rho}) + \sum_{j=1}^{6} \lambda_j (g_j(\boldsymbol{\rho}) + \kappa_j^2)$$

with $\boldsymbol{\lambda}$ being the Lagrange multiplier. The Lagrange function can be used for solving Equation 4.19 by setting each of its partial derivatives to zero and solving the set of equations formed. Taking the partial derivative of the Lagrange function with respect to each of its variables yields the following

set of equations:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial w_1} &= w_1 + \lambda_1 + \lambda_2 + \lambda_3 - 2\lambda_4 - 3\lambda_5 - 4\lambda_6, \\
\frac{\partial \mathscr{L}}{\partial w_2} &= w_2 + \lambda_1 + 2\lambda_2 + \lambda_3 - 4\lambda_4 - 3\lambda_5 - 5\lambda_6, \\
\frac{\partial \mathscr{L}}{\partial b} &= \lambda_1 + \lambda_2 + \lambda_3 - \lambda_4 - \lambda_5 - \lambda_6, \\
\frac{\partial \mathscr{L}}{\partial \kappa_j} &= 2\lambda_j \kappa_j \ \forall i \ 1 \ldots 6, \\
\frac{\partial \mathscr{L}}{\partial \lambda_1} &= w_1 + w_2 + b + 1 + \kappa_1^2, \\
\frac{\partial \mathscr{L}}{\partial \lambda_2} &= w_1 + 2w_2 + b + 1 + \kappa_2^2, \\
\frac{\partial \mathscr{L}}{\partial \lambda_3} &= 2w_1 + w_2 + b + 1 + \kappa_3^2, \\
\frac{\partial \mathscr{L}}{\partial \lambda_4} &= -2w_1 - 4w_2 - b + 1 + \kappa_4^2, \\
\frac{\partial \mathscr{L}}{\partial \lambda_5} &= -3w_1 - 3w_2 - b + 1 + \kappa_5^2, \\
\frac{\partial \mathscr{L}}{\partial \lambda_6} &= -4w_1 - 5w_2 - b + 1 + \kappa_6^2.
\end{aligned}$$

Setting the above equations to zero and solving produces the following result:

$$\begin{bmatrix} w_1^* \\ w_2^* \\ b^* \\ \kappa_1^* \\ \kappa_2^* \\ \kappa_3^* \\ \kappa_4^* \\ \kappa_5^* \\ \kappa_6^* \\ \lambda_1^* \\ \lambda_2^* \\ \lambda_3^* \\ \lambda_4^* \\ \lambda_5^* \\ \lambda_6^* \end{bmatrix} = \begin{bmatrix} 2/3 \\ 2/3 \\ -3 \\ 0.8165 \approx 49/60 \\ 0 \\ 0 \\ 0 \\ 0 \\ \sqrt{2} \\ 0 \\ 0 \\ 4/9 \\ -2/9 \\ 2/3 \\ 0 \end{bmatrix} \tag{4.22}$$

The first three solutions of Equation 4.22 are also the solution of Equation 4.19, $\boldsymbol{\rho}^* = [2/3, 2/3, -3]$, which completely describes the hyperplane $\mathbf{w}^T\mathbf{r} + b = 0$. The hyperplane can be rewritten in terms of r_1 and r_2.

$$\begin{aligned} \mathbf{w}^T\mathbf{r} + b &= w_1 r_1 + w_2 r_2 + b \\ &= 2/3 r_1 + 2/3 r_2 - 3 \\ &= 0 \end{aligned}$$

leading to the normal straight line $r_2 = -r_1 + 4.5$.

4.2.4.2 Proposed features

Three main sets of SVM features will be presented. The first feature set consists of the harmonic components of Equation 4.4 and is denoted by

$$\tilde{\boldsymbol{\imath}} = \{C^b, A^b\}_{b \in \{1, \cdots, \text{NDVI}\}}. \tag{4.23}$$

Any spectral subset of $\tilde{\boldsymbol{\imath}}$ can also be selected, and is denoted by $\boldsymbol{\imath}^{\mathbf{b}}$, where $\mathbf{b}$ is any subset of $\{1, \cdots, \text{NDVI}\}$. Fourier (or spectral) analysis, on NDVI time-series in particular, has been used extensively for land cover classification (see for example [5, 78, 134, 149]), and it has been shown that reliable class separation can be achieved even when considering only the mean and seasonal spectral components [5, 78], i.e. Equation 4.23.

The second feature set consists of noise-harmonic features, i.e. consists of all the parameters in Equation 4.4 and is represented mathematically with

$$\tilde{\boldsymbol{\theta}} = \{C^b, A^b, \phi^b, \lambda^b, \sigma^b\}_{b \in \{1, \cdots, \text{NDVI}\}}.$$

As in the case of $\tilde{\boldsymbol{\imath}}$, a spectral subset of $\tilde{\boldsymbol{\theta}}$ is denoted by $\boldsymbol{\theta}^{\mathbf{b}}$. The benefit of $\tilde{\boldsymbol{\theta}}$ when compared to $\tilde{\boldsymbol{\imath}}$ is that $\tilde{\boldsymbol{\theta}}$ also includes the parameters of the Ornstein-Uhlenbeck process, which are $\tilde{\boldsymbol{\lambda}}$ and $\tilde{\boldsymbol{\sigma}}$. The Ornstein-Uhlenbeck process summarises the less important Fourier features that by themselves do not contribute significantly to classification up with two model parameters that could contribute significantly to classification accuracy.

The third feature set is composed of temporal features. Selecting temporal features for classification purposes is a well-known approach [15]. If the most relevant reflectance values of a MODIS pixel $\tilde{\mathbf{x}}(t)$ are to be chosen, then those reflectance values from $\tilde{\mathbf{x}}(t)$ where the annual ensemble mean of two

different classes are at a maximum distance from each other need to be selected. Mathematically it can be expressed as follows: a τ should be selected such that the following optimisation problem is maximised.

$$\sup_{\tau\in\{1,\cdots,45\}} \|\tilde{\mathbf{y}}_{c_1}(\tau) - \tilde{\mathbf{y}}_{c_2}(\tau)\|_2,$$

where $\tilde{\mathbf{y}}$ represents the annual ensemble mean. The solution τ can be extended to a sequence $\boldsymbol{\tau}$; since the annual ensemble mean is periodic, a maximum can be attained more than once during the observation period I. Now the actual reflectance values from the observed MODIS pixel are selected,

$$\tilde{\boldsymbol{\zeta}} = \tilde{\mathbf{x}}(\boldsymbol{\tau}) = \{x_b(\boldsymbol{\tau})\}_{b\in\{1,\cdots,7,\text{NDVI}\}}.$$

A smaller $\boldsymbol{\zeta}^{\mathbf{b}}$ can be constructed by using subsets of $\tilde{\mathbf{y}}_{c_1}(t)$, $\tilde{\mathbf{y}}_{c_2}(t)$ and $\tilde{\mathbf{x}}(t)$, as long as the subsets are constructed by using the same spectral bands. Now, $\tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\theta}}$ or $\tilde{\boldsymbol{\zeta}}$ can be substituted into $\boldsymbol{r}^{(i)}$ (see Equation 4.16). As mentioned in Chapter 1, the shorthand notation $\boldsymbol{\iota}, \boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ will be used to respectively refer to each feature set group, namely harmonic features, noise-harmonic features and temporal features.

4.3 CHANGE DETECTION

As stated in the chapter introduction, *change detection* is the process of identifying differences in the state of an object or phenomenon by observing it at different times. Essentially, it involves the ability to quantify temporal effects using multitemporal data [69]. Usually land cover changes are categorised into *land cover conversion* and *land cover modification* [71]. Land cover conversion is described in [71] as "complete replacement of one cover type by another", whereas land cover modification is described as "more subtle changes that affect the character of land cover without changing its overall classification".

In Section 4.3.1 a short literature review is given of land cover change detection techniques, followed by the presentation of two hypertemporal change detection techniques in Sections 4.3.2 and 4.3.3.

4.3.1 Literature review

There have been quite a number of reviews on change detection in the remote sensing field, namely [13, 69–73, 150]. The review written by Singh [69] in 1989 laid much of the foundation needed

to categorize remote sensing change detection approaches effectively [151]. The following main categories were proposed by Singh, namely *univariate image differencing*, *image regression*, *image ratioing*, *vegetation index differencing*, PCA, *post-classification comparison*, *direct multi-date comparison* and *Change Vector Analysis (CVA)*. Later Lu *et al.* [13] improved on the categories proposed by Singh (more organised) [151]. The categories proposed by Lu *et al.* are briefly discussed below:

- *Algebra* – this category includes the methods that depend on a change metric, which is subsequently compared to a threshold value in order to declare a change or not. The change metric can be computed in a variety of ways, namely image differencing [152–155], image regression [156], image rationing [157], vegetation index differencing [158] and CVA [159].

- *Transformation* – this category comprises methods that reduce data dimensionality. Some of the possible approaches to reducing data redundancy are PCA [160], *Kauth-Thomas* [161], *Gramm-Schmidt* [161] and the *Chi-square transformation* [162].

- *Classification – post-classification comparison* [163], *Expectation Maximization* [164,165] and ANN [166] are some of the constituent techniques that make up the classification category. The methods in this category use classified images and require a large amount of training data.

- *Advanced models* – this category includes, among others, the *Li-Strahler reflectance model* [167], SMA [102], and the *biophysical parameter estimation model* [168, 169]. The fundamental idea behind the methods in this category is that the reflectance values are converted to biophysical parameters, which are more interpretable than the original raw reflectance values.

- *Geographic Information System (GIS)* – the integrated GIS and remote sensing method [170] and the standard GIS approach [171] are some of the algorithms that fall into this category.

- *Visual interpretation* – visual interpretation requires manual interpretation of remote sensing images at different times followed by on-screen digitation of change polygons [172].

- *Other methods* – many categories have now been suggested to group the different change detection techniques together. There are however some techniques that do not fall into any of the above categories, namely *measures of spatial dependence* [173], *knowledge-based vision systems* [174], *change curves* [175], *generalised linear models* [176], the *curve theorem-based*

approach [177], *structure based approach* [178] and *spatial statistics-based approach* [179].

Most of the categories proposed by Lu *et al.* consist of methods that are multitemporal, which normally require only two images as input. There are however some reviews that explicitly discuss an additional category called *hypertemporal change detection techniques* or *temporal trajectory analysis* [71, 73]. The other categories proposed by [71, 73], will not be adopted, as the multitemporal techniques are grouped sufficiently using the categories proposed by [13]. All the methods that are applied to hypertemporal time-series fall into this additional category.

4.3.1.1 Hypertemporal techniques

When considering multi-date change detection, a serious consideration is the selection of optimal image dates. This problem can be circumvented by considering a hypertemporal time-series [10]. The last decade has seen a dramatic increase in the number of papers published in the field of hypertemporal change detection (remote sensing) [7, 12, 19–21, 28, 151, 180–193], some of which are discussed briefly below.

Temporal change metrics are used in [180] to detect land cover changes. The temporal change metrics are computed by computing the annual difference (year2-year1) of the annual maximum, annual minimum and annual range. In addition to the above metrics, the magnitude of the multitemporal change vector is also calculated. These metrics are then compared with a threshold to determine whether a change has occurred or not. In [7], a change is detected by identifying abnormal pixel behaviour. These pixels are identified by selecting pixels that show a significant deviation in the annual difference of the yearly total NDVI relative to other pixels from the same class and study area. In [184, 185] a disturbance index is computed to detect large-scale ecosystem disturbances. The disturbance index is calculated on an annual basis by dividing two ratios. The top ratio is calculated by dividing the annual maximum land surface temperature LST_{max} with the annual maximum Enhanced Vegetation Index (EVI) EVI_{max}, while the bottom ratio is calculated by dividing the multi-year mean of LST_{max} with the multi-year mean of EVI_{max}. The disturbance index is then compared with a predetermined threshold to determine whether a change flag is required. The departure from a model algorithm, the recursive merging algorithm and the yearly delta algorithm are some of the multitemporal techniques proposed in [28, 187]. A generic change detection approach is proposed in [19, 20] for NDVI time-series by detecting and characterising Breaks for Additive Seasonal and Trend (BFAST). Lastly, it is worth mentioning the sliding window approach documented in [21], the autocorrelation approach

proposed in [189, 190] and the Kalman filter approach presented in [191].

Most of the remote sensing hypertemporal change detection algorithms in the literature use some form of windowing, in other words only recent data are used to detect change. In contrast the hypertemporal change detection algorithm proposed in Section 4.3.3 (Page's original CUSUM algorithm [6]) is windowless, consequently no step is required to determine the window length.

There are several metrics by which a change detection algorithm should be evaluated. An obvious one is *detection delay*, which is the time taken for the change detection algorithm to declare that a change has occurred, given that a change in the data actually occurred. Then there is the question of how likely it is for the algorithm to declare that a change has occurred, given that the change in the data did in fact occur, a metric that is referred to as either *probability of detection* or the True Positive Rate (TP). There are more metrics that need to be considered. For example, there is the possibility that the algorithm will declare change, even though no change has occurred in the data, which can be referred to as either the *probability of false detection (alarm)* or the False Positive Rate (FP). As this chapter is presented in a statistical framework, the detection theory terms TP and FP will not be adopted. Then there is the question of how to eliminate the need for a windowing mechanism, in the sense that the proposed algorithm is *on-line* or sequential, i.e. it uses all the past data. This is possible when the algorithm has the property that it only starts behaving differently when an actual change has occurred. However, it is not common for the above four change detection criteria to be considered simultaneously in a remote sensing change detection context, and in that respect the proposed algorithm is novel, since it can sequentially detect change (vegetation pixels that are changed into settlement pixels) as accurately and quickly as possible, while staying below a certain probability of false alarm.

The proposed change detection algorithm (Section 4.3.3) uses Page's original CUSUM algorithm in order to process samples sequentially [6]. Windowed versions of the CUSUM algorithm have been used with MODIS in the past, typically in a bootstrapping [194] or in an in-control process mean context [28, 29]. The problem with using only recent data, which was extracted using a window, is that the average pixel behaviour might not be captured if the window is not long enough. The next example highlights the main drawback of using a window.

In Figure 4.7, the time-series of a vegetation pixel in Gauteng (that did not change from land cover type) over 8 years and its filtered output (which could be considered as the in-control process mean)

is presented in Figure 4.7. Clearly to select a proper history period is quite difficult in the case of

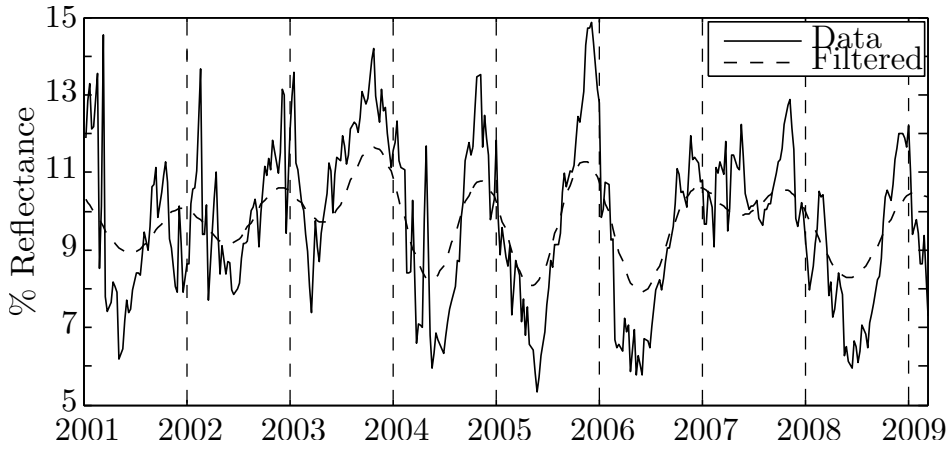


Figure 4.7: Temporal behaviour of a vegetation pixel (Gauteng) in MODIS band 1, and its filtered output (first 10 FFT components).

Figure 4.7. An improper estimated in-control process mean (due to a bad history period) will lead to wrongly estimated residuals (larger than they should be). Larger residuals cause an unnecessary amount of false alarms. Which is why CUSUM performs so badly for the approach presented in [28, 187].

CUSUM can be implemented without using a window, because in Page's original form the CUSUM statistic is derived from log-likelihood ratios, which can be obtained from densities estimated at every time-step of the year. The densities at each time-step thus circumvents intra-annual variation. The densities are constructed by using the CSHO, which can replicate average pixel behaviour which implies that the effect of inter-annual variation is also minimised (see Section 5.2.4.4) [2, 30, 32]. The CUSUM change detection algorithm is compared with the popular band differencing approach (Section 4.3.2) in Section 5.4 [7, 10, 28].

4.3.2 Lunetta et al.'s scheme

Let $\tilde{\mathbf{x}}_p[k] = \{x_p^b[k]\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$ be the p-th discrete MODIS pixel in a set of P unlabelled pixels. The c subscript is omitted as the class of the pixel is unknown. The change detection scheme proposed by Lunetta *et al.* can be implemented with the following steps [28]:

1. The signal $\tilde{\mathbf{x}}_p[k]$ is first filtered, by keeping only the first ν components of an I point Fast Fourier Transform (FFT), where I is equal to the temporal dimension of $\tilde{\mathbf{x}}_p[k]$.

2. For each pixel p compute the annual sum for each year of data (of which there are Y years). Let $\{\mathbf{a}_{p1},\cdots,\mathbf{a}_{pY}\}$ correspond to this list of annual sums, where $\mathbf{a}_{p1}=\{\sum_{k=1}^{45}x_p^b[k]\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$, etc.

3. For each pixel p compute the difference between consecutive annual sums, i.e., $\{\mathbf{a}_{p2}-\mathbf{a}_{p1},\mathbf{a}_{p3}-\mathbf{a}_{p2},\mathbf{a}_{pY}-\mathbf{a}_{pY-1}\}$, where $\mathbf{a}_{p2}-\mathbf{a}_{p1}=\{a_{p2}^b-a_{p1}^b\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$, etc. Let $\mathbf{d}_{pj}=\mathbf{a}_{pj+1}-\mathbf{a}_{pj}$.

4. For each pixel p compute the $z-score$ $\left\{\mathbf{z}_{pj}=\frac{\mathbf{d}_{pj}-\boldsymbol{\mu}_j}{\boldsymbol{\sigma}_j}\right\}$ for each of the $Y-1$ values in $\{\mathbf{d}_{p1},\mathbf{d}_{p2},\cdots,\mathbf{d}_{pY-1}\}$. This is done for each $\mathbf{d}_{pj}$ by subtracting the mean ($\boldsymbol{\mu}_j=\mathbb{E}\{\mathbf{d}_{1j},\mathbf{d}_{2j},\cdots,\mathbf{d}_{Pj}\}=\{\mathbb{E}\{d_{1j}^b,d_{2j}^b,\cdots,d_{Pj}^b\}\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$) and dividing by the standard deviation ($\boldsymbol{\sigma}_j=\sqrt{\mathbb{E}\{(\mathbf{d}_{1j})^2,(\mathbf{d}_{2j})^2,\cdots,(\mathbf{d}_{Pj})^2\}-(\boldsymbol{\mu}_j)^2}$, where $(\mathbf{d}_{1j})^2=\{(d_{1j}^b)^2\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$, etc.). Note that the mean and the standard deviation are computed across space. Let $\{\mathbf{z}_{p1},\mathbf{z}_{p2},\cdots,\mathbf{z}_{pY-1}\}$ correspond to this list of z-scores.

5. For each pixel p compute the change score $\mathbf{c}_p=\{\max\{|z_{p1}^b|,|z_{p2}^b|,\cdots,|z_{pY-1}^b|\}\}_{b\in\{1,\cdots,7,\text{NDVI}\}}$. A change or no change decision can now be reached in every band by comparing $\mathbf{c}_p$ with the eight-dimensional threshold $\mathbf{h}_l$. If $\mathbf{c}_p>\mathbf{h}_l$ a change is declared.

4.3.3 Cumulative Sum

The CUSUM algorithm is discussed in detail in Section 3.6. To apply the CUSUM algorithm to MODIS time-series it needs to be modified slightly. The modified CUSUM algorithm that is presented in this section is also applied to the first order statistical description of $\{\tilde{\mathbf{X}}_c[k]\}_{k=\{1,2,\cdots\}}$, which was introduced in Section 4.2.3. There is however a slight difference; the CUSUM algorithm will only be applied to individual bands, so that a fair comparison with Lunetta et al.'s scheme is possible (no multispectral densities are considered). The modified CUSUM stopping time is given by

$$\mathbf{T}_{\mathbf{h}}^{\text{CUSUM}}=\inf\{\mathbf{k}\geq\mathbf{0}|\mathbf{g}_{\mathbf{k}}\geq\mathbf{h}_c\},$$

where

$$\mathbf{g}_{\mathbf{k}}=\begin{cases}\{(g_{k-1}^b+s_k^b)^+\}_{b\in\{1,\cdots,7,\text{NDVI}\}} & k\neq 0\\ \mathbf{y}\in\mathbb{R}^{+8} & k=0\end{cases}$$

and

$$s_k^b=\ln\frac{q_k^{1,b}(x^b(k))}{q_k^{0,b}(x^b(k))}. \tag{4.24}$$

Under normal CUSUM operating conditions $\mathbf{y} = \mathbf{0} = \{0,0,0,0,0,0,0,0\}$. As soon as $g_k^b \geq h_c^b$ a change can be declared in band b. It is important to realise that the optimality of CUSUM (and the optimality of the sequential time-varying classifier) can no longer be guaranteed because of the following list of shortcomings:

1. The identically distributed assumption is violated by Equation 4.24.

2. The MODIS time-series does not consist of independent observations.

3. The densities $q_k^{0,b}$ and $q_k^{1,b}$ are estimated and not known beforehand.

4. In reality the MODIS pixels are spatially correlated.

The densities at each time-step can be estimated from ground truth data or via a trained CSHO simulator. Furthermore, it should also be obvious to the reader that the CUSUM algorithm presented here is nothing more than a repeated time-varying SPRT (see Section 4.3.3) [59].

4.4 CONCLUSION

The chapter presented the details of all the sequential and non-sequential hypertemporal classification and change detection algorithms that were investigated in this thesis. The chapter was divided into three main sections, namely simulation (Section 4.1), classification (Section 4.2) and change detection (Section 4.3). The chapter primarily dealt with a new stochastic inductive model, the CSHO (Section 4.1.2.2) and the model's possible application in simulation (Section 4.1.2.2), classification (Section 4.2.4.2) and change detection (Section 4.3.3). The experimental results of all the algorithms presented in this chapter will be given in Chapter 5. Note that the time-varying maximum likelihood classifier (with thresholds) in Section 4.2.3 and the CUSUM algorithm in Section 4.3.3 are sequential approaches.

CHAPTER 5

RESULTS

The chapter starts with preliminary data analysis results obtained from the datasets introduced in Section 2.8. These results can be used to predict the performance of the different classification and change detection approaches. In Section 5.2 the inductive simulator discussed in Section 4.1.2 is validated. The final sections of the chapter give the classification and change detection accuracies and rankings of the different sequential and non-sequential hypertemporal classification and change detection algorithms discussed in Chapter 4.

5.1 PRELIMINARY DATA ANALYSIS: GAUTENG AND LIMPOPO

In this section a preliminary investigation of the datasets introduced in Chapter 2 is performed. The knowledge gained from the preliminary analysis is used to explain the classification and change detection results of Section 5.3 and Section 5.4. The data analysis is conducted under the following headings: yearly ensemble mean (Section 5.1.1), temporal Hellinger distance (Section 5.1.2), CSHO parameters (Section 5.1.3), noise correlation (Section 5.1.4) and spatial correlation (Section 5.1.5). An important data manipulation technique used in the remainder of the section is temporal grouping of multispectral observations and is expressed in mathematical notation below.

Recall from Chapter 4 that each MODIS pixel (which is denoted here explicitly by $\bar{\mathbf{x}}_c$) has eight associated time-series, such that $\bar{\mathbf{x}}_c = \{\tilde{\mathbf{x}}_c[k]\}_{k=\{1,2,\cdots\}}$, where $\tilde{\mathbf{x}}_c[k] \in \mathbb{R}^8$ is the multispectral observation at time k. To group certain time steps together, the projection operator is required and is defined as

$$\underset{i\in\mathscr{I}}{\text{pr}}\ \boldsymbol{\psi} = \{\psi_i\}_{i\in\mathscr{I}}, \tag{5.1}$$

such that $\text{pr}_i\boldsymbol{\psi}$ is the i-th component of the sequence $\boldsymbol{\psi}$. In Equation 5.1, $\mathscr{I}$ denotes the index set.

Let $\mathscr{D}_c$ denote the set of all $\bar{\mathbf{x}}_c$ belonging to land cover class c. Then, for each observation period i, $1 \leq i \leq j = 45$ (with j the number of observations in a year), and for each land cover class c, the set $\mathscr{G}_{i,c}$ is defined as [23]

$$\mathscr{G}_{i,c} = \left\{ \mathscr{X} \in \underset{i+jn}{\text{pr}} \bar{\mathbf{x}}_c | \bar{\mathbf{x}}_c \in \mathscr{D}_c \right\},\ n = 1,2,\cdots,N,\ 1 \leq i \leq j,$$

such that $\mathscr{G}_{i,c}$ denotes the set of all multispectral observations for a specific time i during the year, and a particular land cover class c (with N being the number of years). Furthermore, let $\mathscr{G}_{i,c}^{\mathbf{b}}$ denote the set of observations corresponding to a particular time of the year i, a particular class c, and a selection of $|\mathbf{b}|$ spectral bands, $\mathbf{b} \subseteq \{1,\cdots,7,\text{NDVI}\}$.

5.1.1 Yearly ensemble mean

The yearly ensemble mean $\tilde{\mathbf{y}}_c(t)$ is defined by Equation 4.5. The yearly ensemble mean for each class is estimated by taking the average at each observation time step over all pixels and then over all years. In other words, inter-annual variability is not ignored but averaged to obtain the yearly ensemble mean for each class.

5.1.1.1 Yearly ensemble mean: Gauteng

The estimated yearly ensemble means $\tilde{\mathbf{y}}_v(t)$ and $\tilde{\mathbf{y}}_s(t)$ for the Gauteng data set are presented in Figures 5.1 and 5.3. The v and s subscripts respectively refer to the vegetation and settlement class. The result of fitting sinusoids on the estimated yearly ensemble means is displayed in Figures 5.2 and 5.4.

5.1.1.2 Yearly ensemble mean: Limpopo

The estimated yearly ensemble means $\tilde{\mathbf{y}}_v(t)$ and $\tilde{\mathbf{y}}_s(t)$ for the Limpopo dataset are presented in Figures 5.5 and 5.6 respectively.

5.1.1.3 Discussion of yearly ensemble mean

The average absolute distance (for each band b) between the yearly ensemble means of the vegetation and settlement classes is defined as

$$\bar{y}^b = \frac{1}{45} \sum_{k=1}^{45} |y_v^b[k] - y_s^b[k]|.$$

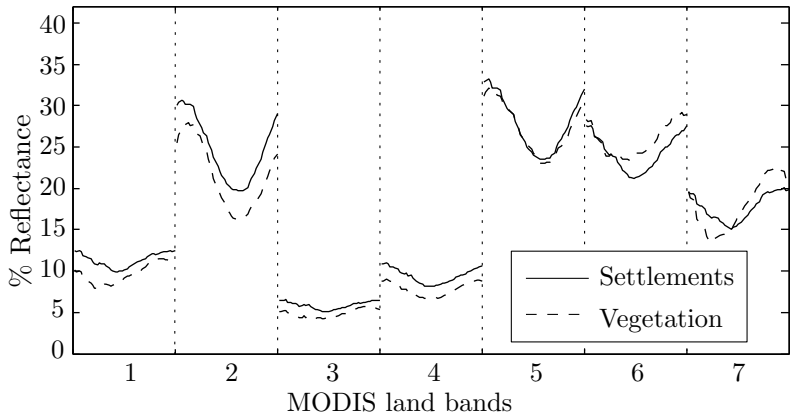


Figure 5.1: The yearly ensemble mean of the MODIS land bands for the vegetation and settlement classes (Gauteng) [2] © IEEE 2012.

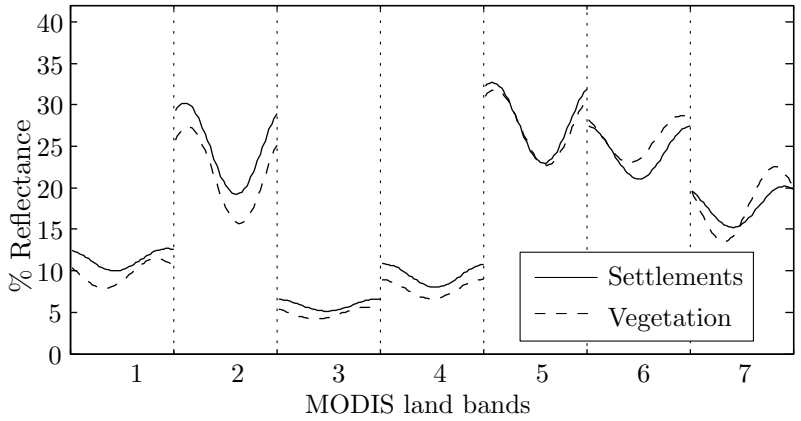


Figure 5.2: Sinusoidal fits on the yearly ensemble mean of the MODIS land bands for the vegetation and settlement class (Gauteng).

The average absolute distance between the yearly ensemble means of the vegetation and settlement classes (for both data sets) can be found in Table 5.1. The values in Table 5.1 were computed with the DN reflectance values and not the scaled values that are used in Figure 5.1 to Figure 5.6.

The average standard deviation (for each band) about the yearly ensemble mean is given in Table 5.2 (for each class and dataset). The average standard deviation is calculated by first grouping all observation time steps (over multiple years and pixels) together and then computing the standard deviation at each time step in a year. To obtain the average standard deviation, the average is then taken over all the time steps in a year of the computed standard deviations. Mathematically speaking it can be

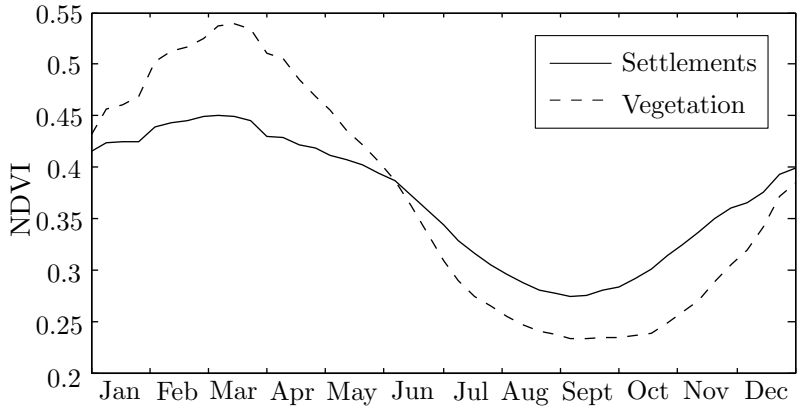


Figure 5.3: The yearly ensemble mean of NDVI for the vegetation and settlement classes (Gauteng) [2] © IEEE 2012.

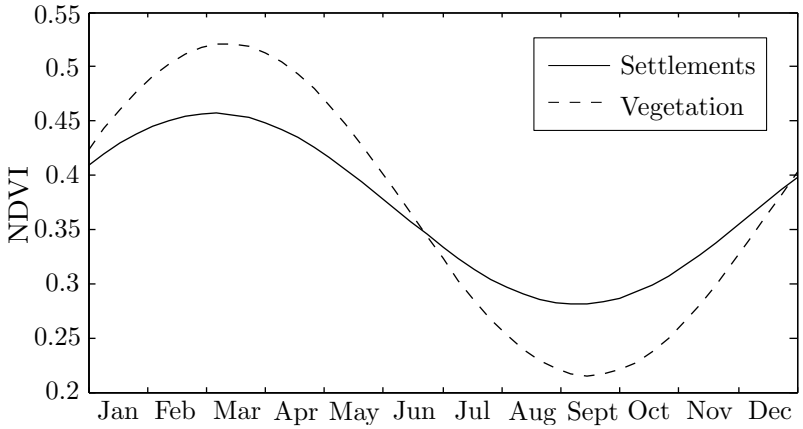


Figure 5.4: Sinusoidal fits on the yearly ensemble mean of NDVI for the vegetation and settlement class (Gauteng).

expressed as

$$\{\mathbb{E}[\{\text{std}(\mathscr{G}_{i,c}^{b})\}_{i\in\{1,\cdots,45\}}]\}_{b\in\{1,2,\cdots,7,\text{NDVI}\}}.$$

The following observations and conclusions can be made from the average yearly ensemble mean results:

1. In the case of the Gauteng dataset the average absolute distance between the yearly ensemble means of the vegetation and settlements class across all bands is equal to 144.07. For the Limpopo dataset the average absolute distance corresponds to 289.10. The absolute distance between the vegetation and settlement yearly ensemble means is therefore larger for the Limpopo dataset. This can be visually verified by using Figure 5.1 to Figure 5.6.

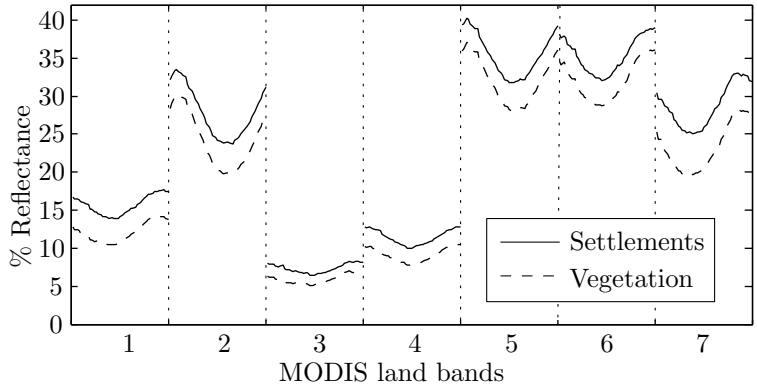


Figure 5.5: The yearly ensemble mean of the MODIS land bands for the vegetation and settlement classes (Limpopo) [2] © IEEE 2012.

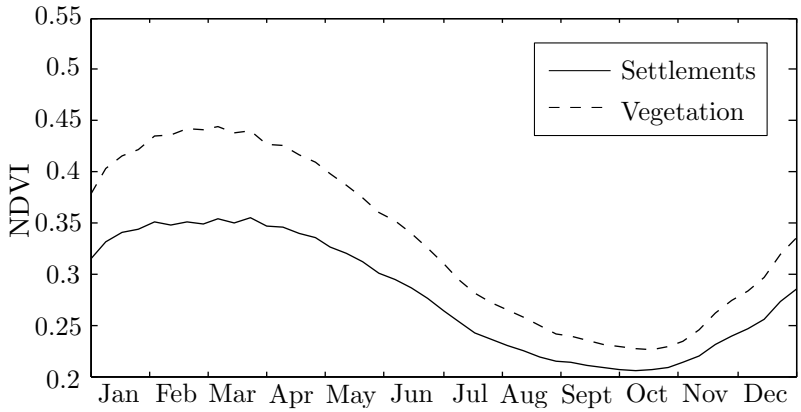


Figure 5.6: The yearly ensemble mean of NDVI for the vegetation and settlement classes (Limpopo) [2] © IEEE 2012.

2. The average standard deviation across all bands for the the vegetation class is respectively equivalent to 216.67 and 280.48 for the Gauteng and Limpopo datasets. In the case of the settlements class 223.67 and 287.35 are respectively obtained. On average the Limpopo data set has a higher standard deviation about the yearly ensemble mean when compared with the Gauteng dataset. Furthermore, in general the settlement classes also have a higher standard deviation around the yearly ensemble mean when compared with the vegetation classes.

3. The higher average absolute distance observable between the yearly ensemble means of the vegetation and settlements class for the Limpopo dataset implies that the Limpopo dataset is more separable than the Gauteng dataset. On the other hand, the higher average standard deviation about the yearly ensemble mean found in the Limpopo dataset, would suggest that the

Table 5.1: Average absolute distance between the yearly ensemble means of the vegetation and settlements class (for both datasets).

	Band							
Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NDVI**
Gauteng	166.12	317.82	97.09	162.18	65.07	168.71	175.55	0.0474
Limpopo	357.06	363.48	142.63	232.65	341.54	340.75	534.64	0.0525

Table 5.2: Average standard deviation about the yearly ensemble mean.

	Band							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**NDVI**
Vegetation								
Gauteng	181.47	329.06	93.08	117.83	336.67	339.28	335.88	0.0794
Limpopo	249.79	343.11	110.85	147.04	423.00	488.04	481.93	0.0734
Settlement								
Gauteng	225.74	239.93	106.97	140.07	280.65	370.95	424.97	0.0799
Limpopo	248.40	311.77	123.35	165.93	405.72	515.95	527.60	0.0617

Limpopo dataset is less separable than the Gauteng dataset. Looking at the average distance of the yearly ensemble means between classes or at the average standard deviations about the yearly ensemble means separately is not enough to predict high or low separability between classes. In contrast, the temporal Hellinger distance defined in Section 5.1.2 takes into account the average distance and the average standard deviation to determine to what extent two classes are separable.

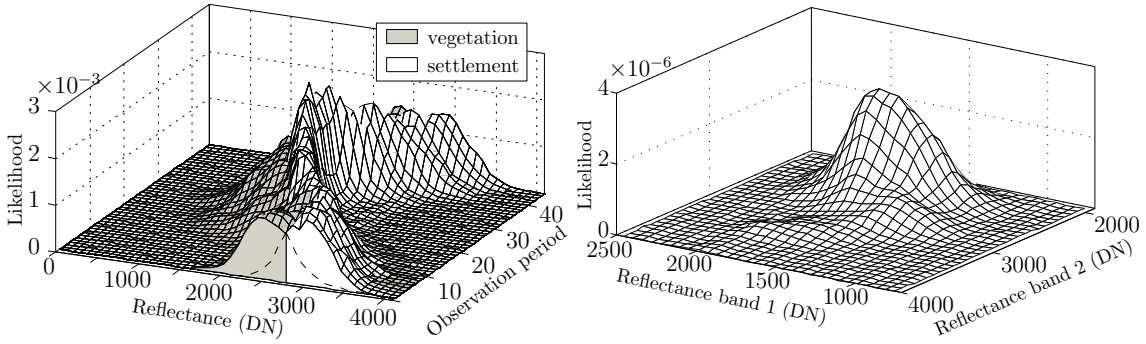
5.1.2 Temporal Hellinger distance

The Hellinger distance between probability density functions p and q is a value between 0 and 1 and is defined as

$$HD(p,q) = \sqrt{1 - \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, dx}.$$

A Hellinger distance of $HD(p,q) \approx 0$ indicates that the densities are not separable, whereas a distance of $HD(p,q) \approx 1$ indicates that the densities are trivially separable.

The temporal Hellinger distance is the Hellinger distance between the single-band time-varying models of the settlement and vegetation classes and is a measure of the degree of separability between the two classes. In other words, the closer the temporal Hellinger distance is to 1 at a specific time step in a year the better a temporal feature classifier will be able to distinguish a settlement observation from a vegetation observation at that specific time step in the year (in theory). The time-varying model was briefly discussed in Section 4.2.3. The time-varying model is constructed by estimating the density of $\mathscr{G}_{i,c}^{\mathbf{b}}$ at each time step in a year. The densities were estimated by using Kernel Density Estimation (KDE), employing a Gaussian kernel and Silverman's rule of thumb as the bandwidth selection rule (Section A.5) [195, 196]. The densities were constructed via the *KDE toolbox for Matlab* [197]. Recall from Section 4.2.3 that the time-varying model densities are denoted with $\{q_i^{c,\mathbf{b}}\}_{i\in\{1,\cdots,45\}}$. The estimated Gauteng dataset time-varying model for both the vegetation and settlement class in land band 2 is depicted in Figure 5.7a, i.e. $\{q_i^{v,2}\}_{i\in\{1,\cdots,45\}}$ and $\{q_i^{s,2}\}_{i\in\{1,\cdots,45\}}$. The estimated multispectral density at time step $i=1$ between land bands 1 and 2 for the settlements class is displayed in Figure 5.7b.



(a) $\{q_i^{v,2}\}_{i\in\{1\cdots45\}}$ and $\{q_i^{s,2}\}_{i\in\{1\cdots45\}}$ [198] © IEEE 2012. **(b)** $q_1^{s,\{1,2\}}$.

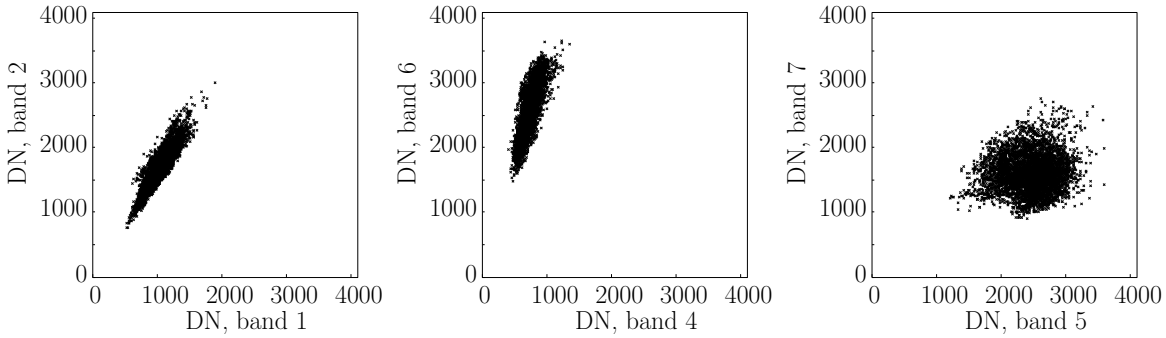
Figure 5.7: Single-band and multiband time-varying models for the Gauteng data set.

The temporal Hellinger distance in band b is defined as

$$H^b[i] = \mathrm{HD}(q_i^{v,b}, q_i^{s,b}). \tag{5.2}$$

At this point it is perhaps worthwhile to give some empirical incitement for using multispectral models. In Figure 5.8 there are three dual-band scatter diagrams at three different times of the year (for the Gauteng vegetation dataset).

Figure 5.8a indicates that there is a strong correlation at time $i=35$ between bands 1 and 2, while Figure 5.8c shows a weak correlation between bands 5 and 7 at time step $i=22$. Figure 5.8b testifies



(a) Scatter diagram ($N = 592$) between bands 1 and 2, $i = 35$.

(b) Scatter diagram ($N = 592$) between bands 4 and 6, $i = 32$.

(c) Scatter diagram ($N = 592$) between bands 5 and 7, $i = 22$.

Figure 5.8: Scatter diagrams of several spectral bands, at different times of the year.

that there is a moderate degree of correlation present between bands 4 and 6 at time $i = 32$. The degree of correlation in Figure 5.8b is less than the degree of correlation in Figure 5.8a, but more than that attested by Figure 5.8c. The reason for only giving three scatter diagrams (out of thousands of possibilities), is that the aim here is not to give a comprehensive description of the dependencies between the spectral bands, but to motivate the use of multispectral models.

5.1.2.1 Temporal Hellinger distance: Gauteng

The temporal Hellinger distance between the single-band time-varying models of the vegetation and settlement classes for the Gauteng dataset is depicted in Figure 5.9.

5.1.2.2 Temporal Hellinger distance: Limpopo

The temporal Hellinger distance between the single-band time-varying models of the vegetation and settlement classes for the Limpopo dataset is displayed in Figure 5.10.

5.1.2.3 Discussion of temporal Hellinger distance

The average and maximum temporal Hellinger distance (across all time steps in a year) between the time-varying models of the vegetation and settlement classes is given in Table 5.3.

The following comments pertain to the temporal Hellinger distance metric:

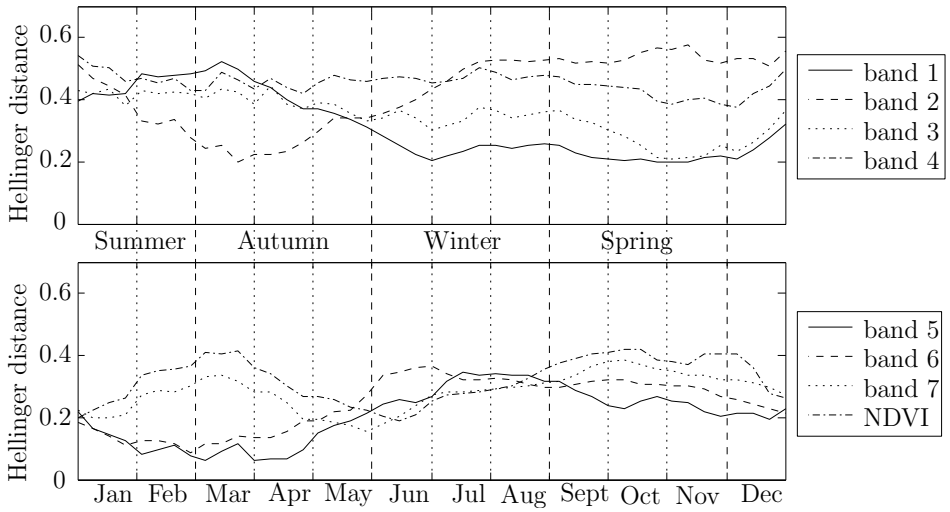


Figure 5.9: The temporal Hellinger distance between the single-band time-varying models of the vegetation and settlement classes for the Gauteng dataset [198] © IEEE 2012.

1. In the case of the Gauteng dataset the average Hellinger distance (across all bands) is equal to 0.32, while the average maximum Hellinger distance (across all bands) corresponds to 0.45.

2. In the case of the Limpopo dataset the average Hellinger distance and the average maximum Hellinger distance respectively correspond to 0.37 and 0.43.

3. During certain times of the year the time-varying models of the Gauteng dataset are more separable than the time-varying models of the Limpopo dataset (based on the average maximum temporal Hellinger distance between the time-varying models of the vegetation and settlement classes), while the time-varying models of the Limpopo dataset are on average more separable than the time-varying models of the Gauteng dataset (based on the average temporal Hellinger distance between the time-varying models of the vegetation and settlement classes). Table 5.3 predicts that the accuracy performance of a temporal classifier applied to the Gauteng or Limpopo datasets would be similar.

4. From Table 5.3 it is clear that in the case of the Gauteng dataset bands $\{1,2,3,4\}$ provide a higher degree of separability (between the time-varying models of the vegetation and settlement classes) than bands $\{5,6,7,\text{NDVI}\}$, which implies that a temporal classifier using bands $\{1,2,3,4\}$ should perform better than a temporal classifier using bands $\{5,6,7,\text{NDVI}\}$ (theoretically speaking). The vegetation and settlement classes are most separable in band 2 and least separable in band 5.

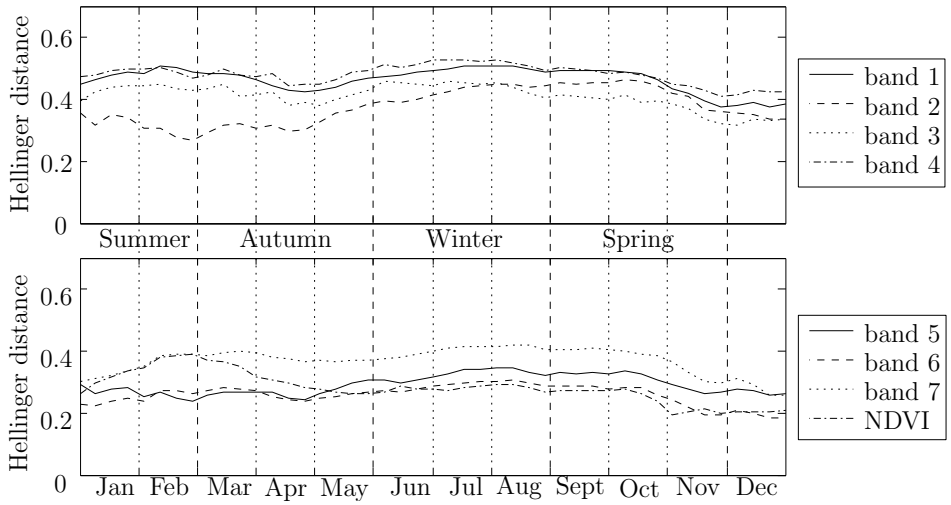


Figure 5.10: The temporal Hellinger distance between the single-band time-varying models of the vegetation and settlement classes for the Limpopo dataset.

5. It is also clear from Table 5.3 that a temporal classifier applied to the Limpopo dataset should perform better if it used data from bands $\{1,2,3,4\}$ than data from bands $\{5,6,7,\text{NDVI}\}$. The temporal classifier should produce its most accurate results if it uses data from band 1 and its most inaccurate results when it uses data from band 6.

5.1.3 CSHO model parameters

The CSHO is introduced in Section 4.1.2.2. A lot can be ascertained about the Gauteng and Limpopo datasets by looking at the parameters of the CSHO. The parameters are investigated in three different ways. Firstly, the parameter Hellinger distance between the probability density functions of the different CSHO parameters is investigated. Then a closer look is taken at the densities of $\tilde{\boldsymbol{\lambda}}$. Finally the correlation between the different parameters is investigated.

The parameter Hellinger distance between the densities of the vegetation and settlement classes for the parameter θ_i is expressed mathematically as

$$\text{HD}(f_v(\theta_i), f_s(\theta_i)),$$

where $f_c(\theta_i)$ is the marginal probability density function of $f(\tilde{\boldsymbol{\theta}})$ defined in Equation 4.11. This metric predicts which parameters would be good features to use as input to an SVM or an ANN. The parameter Hellinger distance measures the separability between the probability density functions of the parameters of the CSHO. In other words, a parameter Hellinger distance close to 1 implies that

Table 5.3: The average and maximum temporal Hellinger distance between the single-band time-varying models of the vegetation and settlement classes for the Gauteng and Limpopo datasets.

	Band							
Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NDVI**
$\mathbb{E}[\{H^b[i]\}_{i\in\{1,\cdots,45\}}]$								
Gauteng	0.31	0.42	0.35	0.45	0.21	0.24	0.28	0.32
Limpopo	0.46	0.37	0.41	0.48	0.29	0.26	0.37	0.28
$\sup\{\{H^b[i]\}_{i\in\{1,\cdots,45\}}\}$								
Gauteng	0.52	0.57	0.44	0.54	0.34	0.36	0.38	0.42
Limpopo	0.51	0.46	0.46	0.53	0.35	0.31	0.42	0.39

the parameter is a good feature to use to differentiate between the vegetation and settlement classes, while a a parameter Hellinger distance close to 0 implies the parameter is not a good feature to use to differentiate between the vegetation and settlement classes.

To a certain extent the $\tilde{\boldsymbol{\lambda}}$ parameters measure the degree of dependence between the observations of the MODIS time-series (temporal dependence). This is true since the λ parameter of the Ornstein-Uhlenbeck process regulates the coefficient of $\eta^b[i-1]$ in Equation 4.12. The influence of the previous observation on the current observation increases as $\lambda \to 0$ and decreases as $\lambda \to \infty$.

The parameter correlation matrix $\tilde{\boldsymbol{P}}_p^c$ defined in Section 4.1.2.5 measures the correlation between the parameters of the CSHO and the parameter correlation matrix $\tilde{\boldsymbol{P}}_p^c$ is thus also a measure of spectral dependence (under the Gaussian assumption). Note that $\tilde{\boldsymbol{P}}_p^c$ implies the inclusion of NDVI, whereas $\boldsymbol{P}_p^c$ does not include NDVI. Up to now the mathematical definitions of different correlation matrices were given, $\tilde{\boldsymbol{P}}_p^c, \tilde{\boldsymbol{P}}_\eta^c$ and $\tilde{\boldsymbol{\rho}}^c$. What is lacking at this point however is a computational approach for computing a correlation matrix $\boldsymbol{R}$ from a set $\{\mathscr{R}_x\}_{x=1,2,\cdots,k}$, with $\mathscr{R}_x[i]$ being observations of $\mathscr{R}_x$ and $i = 1,2,\cdots,n$. The correlation matrix $\boldsymbol{R}$ has entries

$$r_{xy} = \frac{\sum_{i=1}^{n}(\mathscr{R}_x[i]-\bar{\mathscr{R}}_x)(\mathscr{R}_y[i]-\bar{\mathscr{R}}_y)}{\sqrt{\sum_{i=1}^{n}(\mathscr{R}_x[i]-\bar{\mathscr{R}}_x)^2\sum_{i=1}^{n}(\mathscr{R}_y[i]-\bar{\mathscr{R}}_y)^2}},\ x,y = 1,2,\cdots,k. \tag{5.3}$$

5.1.3.1 CSHO model parameters: Gauteng

The parameter Hellinger distance between the probability density functions of the parameters of the CSHO for the vegetation and settlement classes of the Gauteng dataset is displayed in Figure 5.11. The probability density functions of $\tilde{\boldsymbol{\lambda}}_v$ and $\tilde{\boldsymbol{\lambda}}_s$ are given in Figure 5.12. The matrices $\tilde{\boldsymbol{P}}_p^v$ and $\tilde{\boldsymbol{P}}_p^s$ are presented in Figure 5.13.

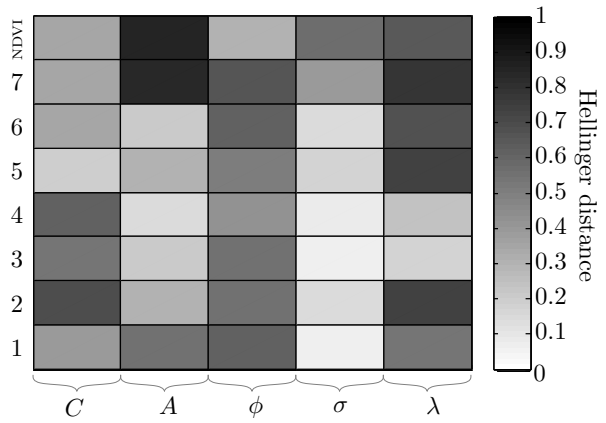
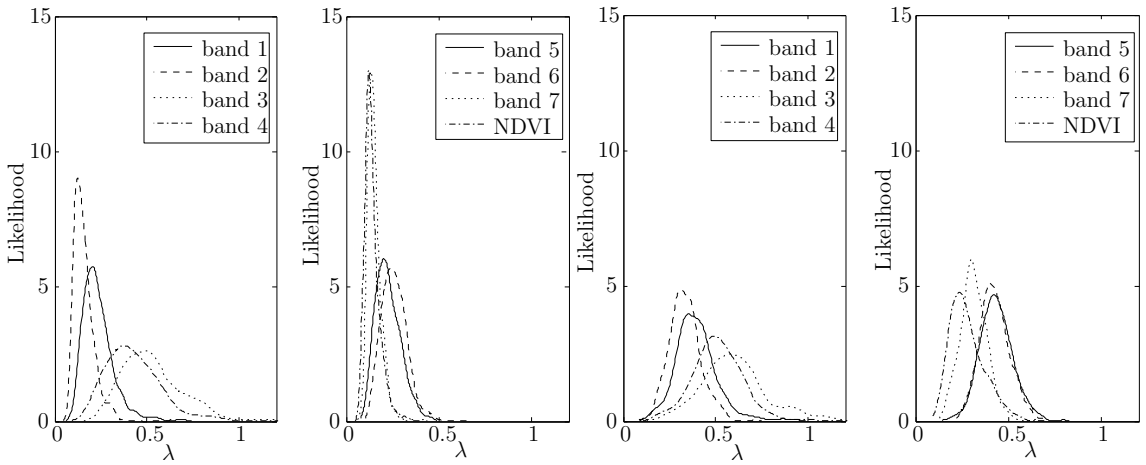


Figure 5.11: HD$(f_v(\theta_i), f_s(\theta_i))$ (Gauteng) [2] © IEEE 2012.

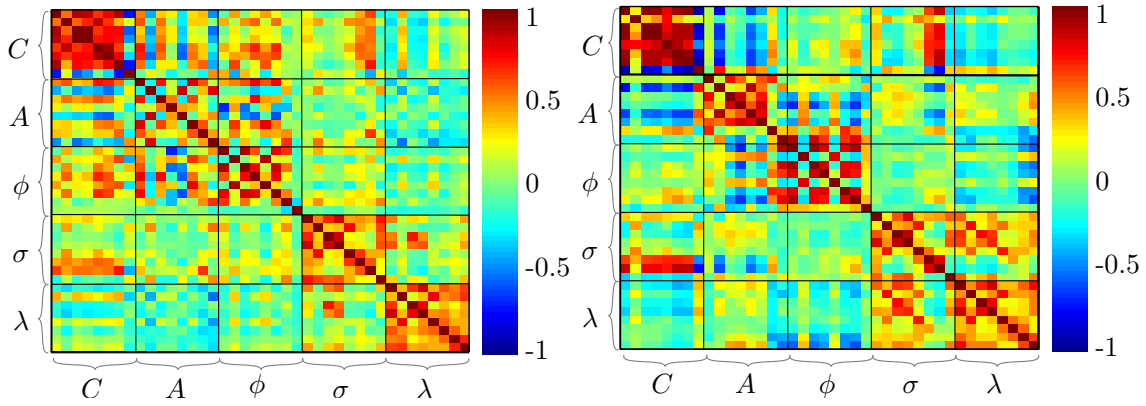


(a) $\tilde{\boldsymbol{\lambda}}_v$. **(b)** $\tilde{\boldsymbol{\lambda}}_s$.

Figure 5.12: Probability density functions of $\tilde{\boldsymbol{\lambda}}_v$ and $\tilde{\boldsymbol{\lambda}}_s$ (Gauteng).

5.1.3.2 CSHO model parameters: Limpopo

The parameter Hellinger distance between the probability density functions of the parameters of the CSHO for the vegetation and settlement classes of the Limpopo dataset is displayed in Figure 5.14.



(a) $\tilde{\boldsymbol{P}}_p^v$ [32] © IEEE 2012. **(b)** $\tilde{\boldsymbol{P}}_p^s$.

Figure 5.13: $\tilde{\boldsymbol{P}}_p^v$ and $\tilde{\boldsymbol{P}}_p^s$ (Gauteng)

The probability density functions of $\tilde{\boldsymbol{\lambda}}_v$ and $\tilde{\boldsymbol{\lambda}}_s$ for the Limpopo dataset are given in Figure 5.15. The matrices $\tilde{\boldsymbol{P}}_p^v$ and $\tilde{\boldsymbol{P}}_p^s$ are presented in Figure 5.16 for the Limpopo dataset.

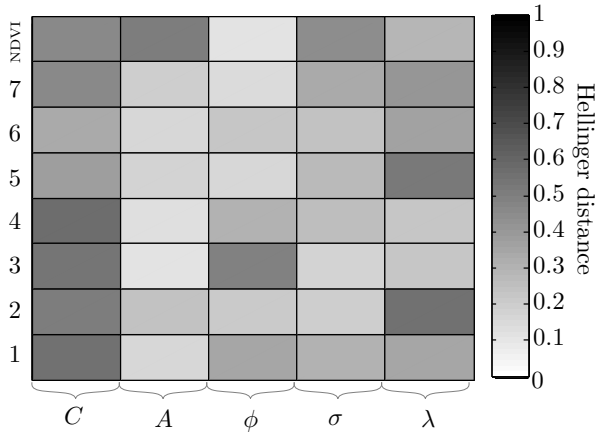
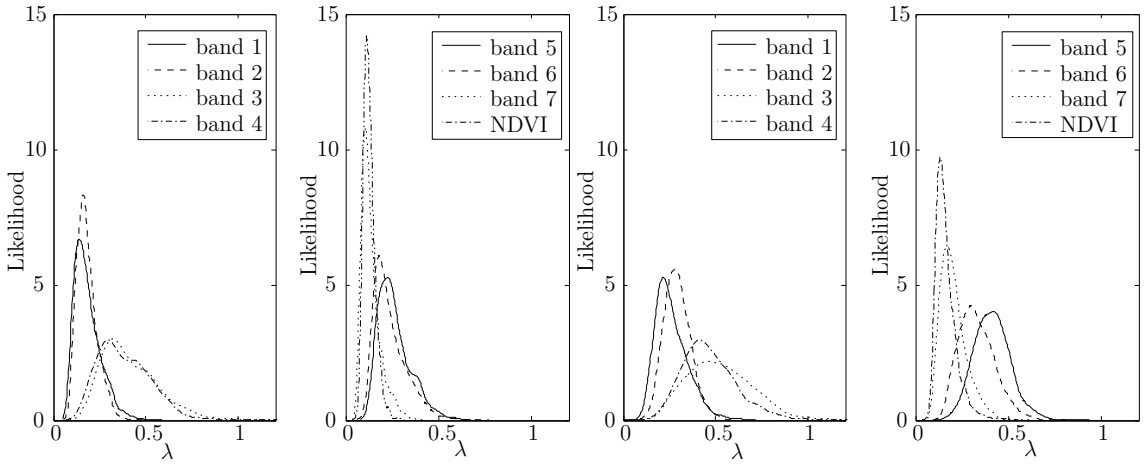


Figure 5.14: HD$(f_v(\theta_i), f_s(\theta_i))$ (Limpopo) [2] © IEEE 2012.

5.1.3.3 Discussion of CSHO model parameters

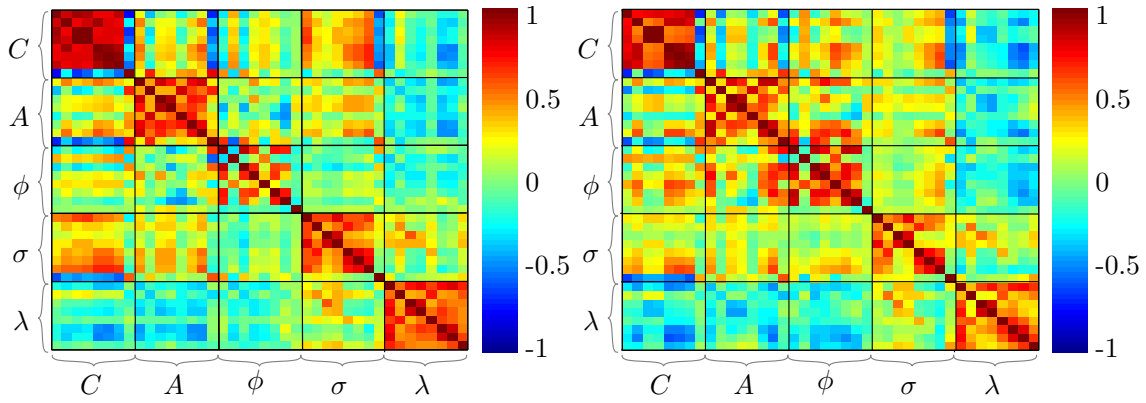
The average parameter Hellinger distances across bands and parameters are given in Table 5.4 and Table 5.5 respectively for the Gauteng and Limpopo datasets. The average values of λ for each band and class for the Gauteng and Limpopo datasets are given in Table 5.6.

The following observations and conclusion can be made from the figures and tables pertaining to the parameters of the CSHO:



(a) $\tilde{\boldsymbol{\lambda}}_v$. **(b)** $\tilde{\boldsymbol{\lambda}}_s$.

Figure 5.15: Probability density functions of $\tilde{\boldsymbol{\lambda}}_v$ and $\tilde{\boldsymbol{\lambda}}_s$ (Limpopo).



(a) $\tilde{\boldsymbol{P}}_p^v$. **(b)** $\tilde{\boldsymbol{P}}_p^v$.

Figure 5.16: $\tilde{\boldsymbol{P}}_p^v$ and $\boldsymbol{P}_p^s$ (Limpopo).

1. The average parameter Hellinger distance (across bands) can be calculated from Table 5.4 and is equal to 0.34 and 0.29 in the case of the Limpopo and Gauteng datasets. The higher average parameter Hellinger distance of the Gauteng dataset implies that the potential differentiability between the vegetation and settlement classes is higher for the Gauteng dataset than the Limpopo dataset when the parameters of the CSHO are used as classification features.

2. Based on the average parameter Hellinger distances in Table 5.4, a higher degree of differentiability between the vegetation and settlement class is possible (plausible) for the Gauteng dataset if bands $\{1,2,7,\text{NDVI}\}$ (band 7 provides the highest degree of differentiability) are used (when the parameters of the CSHO are used as classification features). The remaining

Table 5.4: Average Hellinger distance (across bands) between the densities of the parameters of the CSHO for the Gauteng and Limpopo datasets.

	Band							
Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**NDVI**
Gauteng	0.37	0.41	0.27	0.26	0.33	0.34	0.51	0.46
Limpopo	0.29	0.31	0.26	0.25	0.25	0.22	0.26	0.33

Table 5.5: Average Hellinger distance (across parameters) between the densities of the parameters of the CSHO for the Gauteng and Limpopo datasets.

	Parameter				
Dataset	C	A	ϕ	σ	λ
Gauteng	0.39	0.38	0.48	0.20	0.51
Limpopo	0.43	0.19	0.25	0.25	0.33

bands (band 4 provides the lowest degree of differentiability) provide less differentiability than $\{1,2,7,\text{NDVI}\}$). The Limpopo dataset has exactly the same division of its bands (as the Gauteng dataset). For the Limpopo dataset NDVI provides the highest amount of differentiability, while band 6 provides the lowest amount.

3. According to Table 5.5 the CSHO parameters can be sorted as follows: $\{\lambda,\phi,C,A,\sigma\}$ (by using the Hellinger distance values across parameters). The parameters are now in descending order (in terms of their potential usefulness as classification features). Similarly, for the Limpopo class the parameter list $\{C,\lambda,\phi,\sigma,A\}$ is constructed.

4. For the Gauteng dataset the following is observable when inspecting Figure 5.11: Firstly, the potential discerning capability (classification capability) of **C** (see Section 4.1.2.7) in bands 2 and 4 is high, since the parameter Hellinger distance is high in bands 2 and 4. Similarly, the potential classification capability that bands 7 and NDVI can provide in the case of **A** is also quite good. The phase parameters $\boldsymbol{\phi}$ are also theoretically capable of good class differentiability in all the MODIS land bands, while the estimated noise parameters provide good potential class discernment in bands 2, 5, 7 and NDVI. It is noteworthy to mention that generally band 5 cannot

Table 5.6: The average values of λ in each band and class for the Gauteng and Limpopo datasets.

	Band							
Average value of λ_v^b								
Gauteng	0.24	0.16	0.53	0.43	0.23	0.26	0.15	0.14
Limpopo	0.18	0.18	0.42	0.38	0.26	0.24	0.13	0.13
Average value of λ_s^b								
Gauteng	0.40	0.33	0.61	0.51	0.44	0.43	0.31	0.28
Limpopo	0.26	0.29	0.53	0.47	0.41	0.33	0.21	0.17

be used to provide good class differentiability except in the case of the mean reversion rate of the noise. Finally, the volatility of the noise provides poor potential class differentiability, because the volatility has an HD ≈ 0 in almost all of the MODIS bands.

5. According to Figure 5.14 the mean component is the best parameter to use as a classification feature, while the seasonal component would be the worst parameter to use, except in the case of NDVI (for the Limpopo dataset). However the most important result from Figure 5.14 is that the noise parameters can also be used to differentiate between classes.

6. According to Table 5.6 the average value of λ is equal to 0.34 (in the case of the Gauteng dataset), while the average value of λ is equivalent to 0.29 for the Limpopo dataset. As stated before, the higher the value of λ the less the dependence is between the observations in the dataset. Since $0.34 > 0.29$ it can be inferred that the dependence between the observations in the Gauteng dataset is less than for the Limpopo dataset.

7. For the Gauteng dataset, the lowest amount of dependence between the observations of the vegetation class is observable in bands $\{1,3,4,6\}$ (the lowest amount of dependence is observable in band 3), while the highest amount of dependence between the observations is observable in bands $\{2,5,7,\text{NDVI}\}$ (the highest amount is seen in the case of NDVI). For the settlement class, temporal dependence is more prominent in bands $\{3,4,5,6\}$ (temporal dependence is most prominent in band 3) than in bands $\{1,2,7,\text{NDVI}\}$ (The lowest amount of dependence is observable in the case of NDVI).

8. In the case of the Limpopo dataset, the lowest amount of dependence between the observations

of the vegetation class is observable in bands $\{3,4,5,6\}$ (the lowest amount of dependence is observable in band 3), while the highest amount of dependence between the observations is observable in bands $\{1,2,7,\text{NDVI}\}$ (the highest amount is seen in the case of NDVI). For the settlement class, temporal dependence is more prominent in bands $\{3,4,5,6\}$ (temporal dependence is most prominent in band 3) than in bands $\{1,2,7,\text{NDVI}\}$ (The lowest amount of dependence is observable in the case of NDVI).

9. From Figure 5.13 it is clear that for the Gauteng dataset the mean parameters $\mathbf{C}$ of the CSHO are the most correlated. Note that under a Gaussian assumption correlation implies dependence; however the terminology is not adopted here for the sake of generality. For the settlement class band 2 shows a significantly lower correlation for $\mathbf{C}$ than in the vegetation class. The amplitude parameters $\mathbf{A}$ are more correlated in the settlement class than in the vegetation class. There is also a high degree of correlation between $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ respectively; for the settlement class $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ are also highly correlated. The correlation profiles (patterns) of $\mathbf{C},\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ are especially similar for the Gauteng dataset.

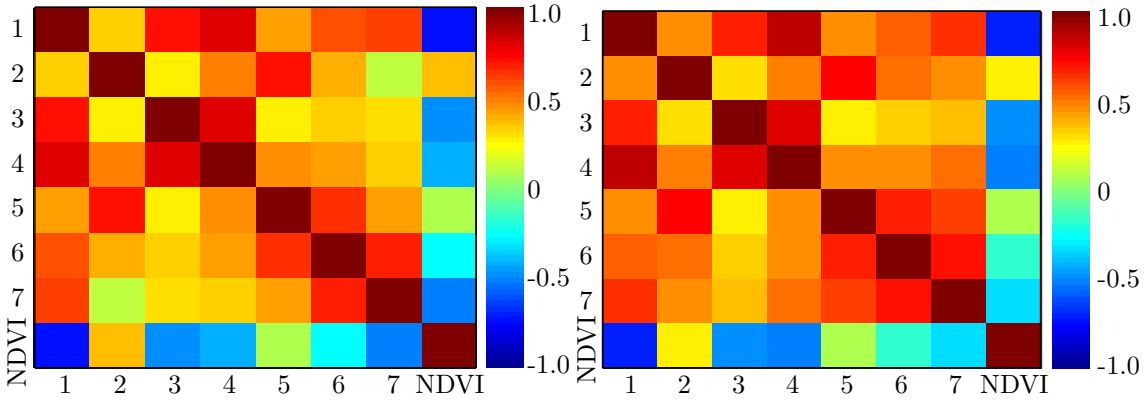
10. As can be seen in Figure 5.16, similar to the Gauteng dataset, the mean parameters of the CSHO are the most correlated. In contrast to the Gauteng dataset, the amplitude parameters $\mathbf{A}$ are also highly correlated. The correlation between $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma}$ is less in the Limpopo dataset than in the Gauteng dataset. The correlation profiles of $\mathbf{C},\mathbf{A},\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$ are especially similar for the Limpopo dataset.

5.1.4 Noise correlation

The noise correlation matrix $\tilde{\boldsymbol{P}}_{\eta}^{c}$ is discussed in Section 4.1.2.5 and measures the degree of correlation that exists (in the noise) between the different spectral bands.

5.1.4.1 Noise correlation: Gauteng

The noise correlation matrices $\tilde{\boldsymbol{P}}_{\eta}^{v}$ and $\tilde{\boldsymbol{P}}_{\eta}^{s}$ are displayed graphically in Figure 5.17 for the Gauteng dataset.

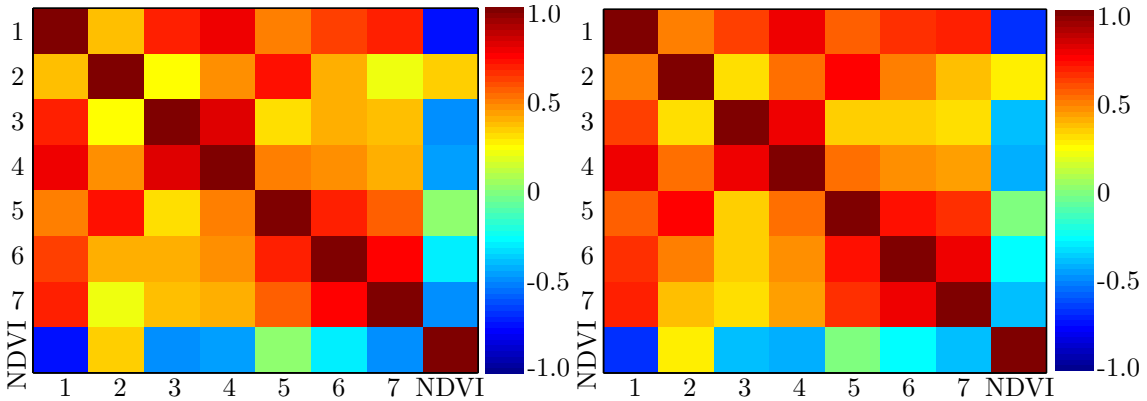


(a) $\tilde{\boldsymbol{P}}_{\eta}^{v}$ [32] © IEEE 2012. **(b)** $\tilde{\boldsymbol{P}}_{\eta}^{s}$.

Figure 5.17: $\tilde{\boldsymbol{P}}_{\eta}^{v}$ and $\tilde{\boldsymbol{P}}_{\eta}^{s}$ (Gauteng).

5.1.4.2 Noise correlation: Limpopo

The noise correlation matrices $\tilde{\boldsymbol{P}}_{\eta}^{v}$ and $\tilde{\boldsymbol{P}}_{\eta}^{s}$ are displayed graphically in Figure 5.18 for the Limpopo dataset.



(a) $\tilde{\boldsymbol{P}}_{\eta}^{v}$. **(b)** $\tilde{\boldsymbol{P}}_{\eta}^{s}$.

Figure 5.18: $\tilde{\boldsymbol{P}}_{\eta}^{v}$ and $\tilde{\boldsymbol{P}}_{\eta}^{s}$ (Limpopo).

5.1.4.3 Discussion of noise correlation

The following is conspicuous when inspecting Figures 5.17 and 5.18:

1. The parameter correlation matrices and the noise correlation matrices have similar profiles (patterns), especially in the case of the noise parameters of the CSHO.

2. The noise correlation matrices of the different classes are very similar. On average the most correlated two-band pairs are $\{1,3\},\{1,4\},\{1,6\},\{1,7\}$, $\{2,5\},\{3,4\},\{5,6\},\{6,7\}$.

3. On average the least correlated two-band pairs are $\{1,2\},\{2,3\},\{2,6\},\{2,7\}$, $\{3,5\},\{3,6\},\{3,7\},\{4,6\}$.

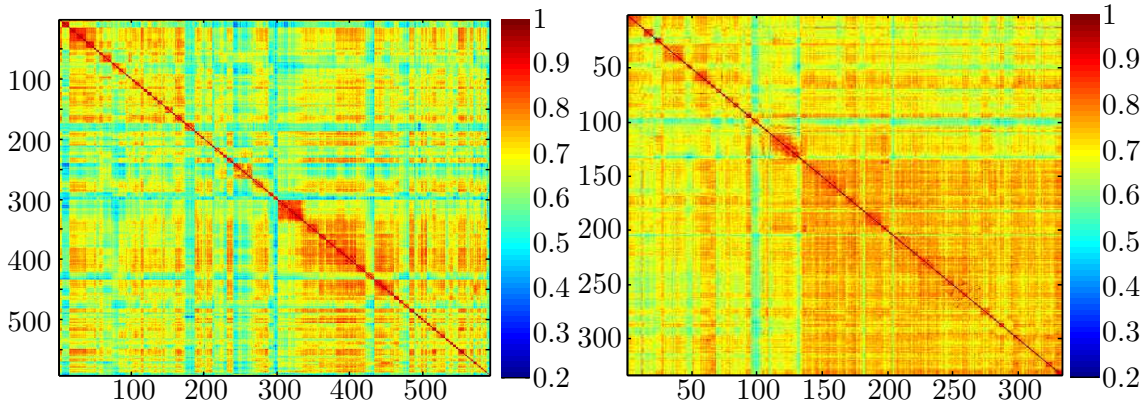
4. The following band pairs vary most between classes and datasets: $\{1,2\},\{2,6\},\{3,7\},\{5,7\}$.

5.1.5 Spatial correlation

The average spatial correlation matrix $\tilde{\boldsymbol{\rho}}^c$ is defined in Equation 4.7 and measures the average spatial correlation that exists between the pixels of a specific class. When inspecting Equation 4.7 notice that the spatial correlation is calculated by computing the average correlation between pixels and not via Euclidean distances. According to Section 4.3.3, one of the reasons that CUSUM's optimality cannot be guaranteed is because of spatial correlation (see Section 5.4.4 for more details).

5.1.5.1 Spatial correlation: Gauteng

The matrices $\tilde{\boldsymbol{\rho}}^v$ and $\tilde{\boldsymbol{\rho}}^s$ are displayed in Figure 5.19 for the Gauteng dataset.

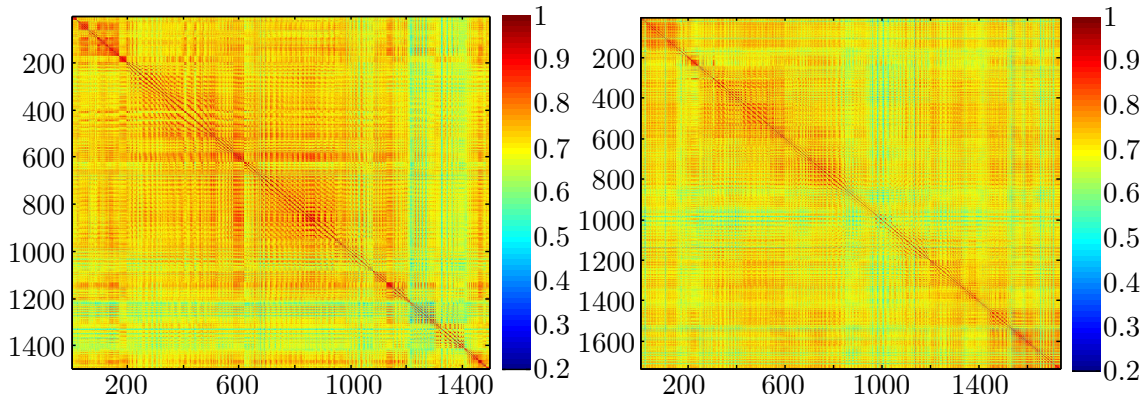


(a) $\tilde{\boldsymbol{\rho}}^v$. **(b)** $\tilde{\boldsymbol{\rho}}^s$.

Figure 5.19: $\tilde{\boldsymbol{\rho}}^v$ and $\tilde{\boldsymbol{\rho}}^s$ (Gauteng)

5.1.5.2 Spatial correlation: Limpopo

The matrices $\tilde{\boldsymbol{\rho}}^v$ and $\tilde{\boldsymbol{\rho}}^s$ are presented graphically in Figure 5.20 for the Limpopo dataset.



(a) $\tilde{\boldsymbol{\rho}}^v$. **(b)** $\tilde{\boldsymbol{\rho}}^s$.

Figure 5.20: $\tilde{\boldsymbol{\rho}}^v$ and $\tilde{\boldsymbol{\rho}}^s$ (Limpopo).

5.1.5.3 Discussion of spatial correlation

The following observations can be made about Figures 5.19 and 5.20:

1. In the case of the Gauteng dataset, the average spatial correlation (computed by taking the average of $\tilde{\boldsymbol{\rho}}^c$) is equivalent to 0.66 and 0.71 for the vegetation and settlement class respectively. In the case of the Limpopo dataset the average spatial correlation is equal to 0.71 and 0.70 for the vegetation and settlement class respectively.

2. By using Figures 5.19 and 5.20 as a visual aid, it can be confirmed that on average the spatial correlation is higher in the Limpopo dataset than in the Gauteng dataset.

5.2 SIMULATOR RESULTS: GAUTENG AND LIMPOPO

The algorithm for simulating a MODIS pixel is discussed in detail in Section 4.1.2.7. As mentioned in Chapter 1, the main purpose of the inductive simulator is to augment datasets for the data-intensive sequential algorithms, especially CUSUM. The inductive simulator is used together with the CUSUM algorithm in Section 5.4.4. Figure 5.21 contains a true vegetation pixel from the Gauteng dataset and its recreated (simulated) counterpart.

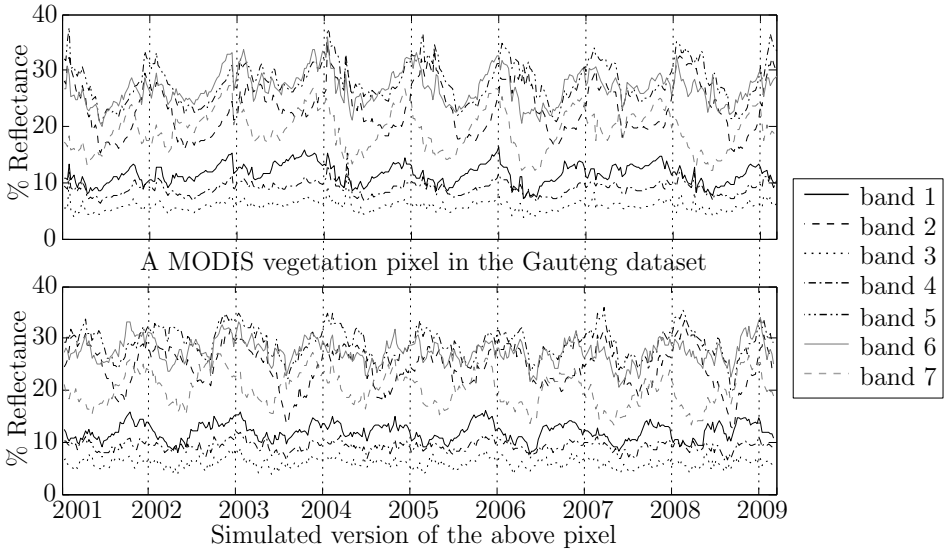


Figure 5.21: A true vegetation pixel belonging to the Gauteng dataset as well as its recreated (simulated) counterpart.

The following observations can be made from Figure 5.21:

1. From the real vegetation pixel graph it is clear that there is correlation between spectral bands (if the reflectance goes up in band 1 it usually also goes up in band 4) and spectral dependence (consider, for instance the mean of each spectral band).

2. The simulated pixel does not replicate the temporal behaviour exactly.

3. The long-term mean and seasonal components of the real and simulated pixels are however similar.

4. The real vegetation pixel has increment difference outliers, while the simulated pixel does not.

As mentioned in Chapter 4, the spectral signature for each class is encapsulated by $f(\boldsymbol{\theta}_c)$ (see Section 4.1.2.4) and $\boldsymbol{P}_{\eta}^{c}$ (see Section 4.1.2.5). The metrics introduced in this section are used to determine if the class signature is replicated adequately by the simulator.

The experimental procedure used to validate the simulator is discussed in Section 5.2.1. The different metrics and the reason for selecting each metric are discussed in Section 5.2.3. There are two main metric types. The details of each type are given in Section 5.2.4 and Section 5.2.5 respectively. The results obtained via the experimental procedure discussed in Section 5.2.1 are presented in

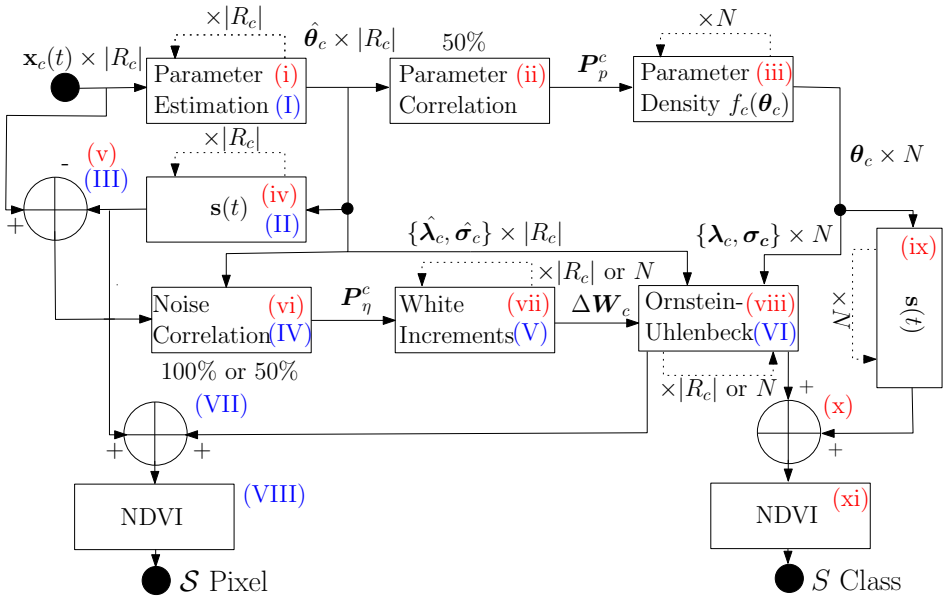


Figure 5.22: Flow diagram illustrating how S and $\mathscr{S}$ are generated. When there are two possibilities at a block, the first option relates to the generation of the pixel set $\mathscr{S}$, while the second option is used to create the class set S. The capital roman numerals are the steps needed to create the pixel set $\mathscr{S}$, while the lowercase roman numerals are the steps required to create the class set S [32] © IEEE 2012.

Section 5.2.6.

5.2.1 Simulator validation

The simulator is validated by using *class* and *pixel* metrics. The class metrics are used to determine whether the simulated dataset has the same statistical attributes as the original dataset and are important, since class attributes are used by classifiers to distinguish between classes [2, 23]. The pixel metrics, in contrast, are used to verify that the simulator can also reproduce any given pixel accurately by comparing every real world pixel to its simulated counterpart. The construction procedures of the simulated datasets to which the two types of metrics are applied differ and are illustrated in Figure 5.22. The construction procedures are an extension of the algorithm presented in Section 4.1.2.7.

The steps required to generate the dataset S ($|S| = N$) to which the class metrics are applied are summarised below:

i Estimate the parameters of R_c (all the pixels in R belonging to class c).

ii Select a random 50% of the estimated parameters to construct $\boldsymbol{P}_p^c$ (discussed in Section 4.1.2.5). The pixels associated with the selected parameters form the training set. The remaining pixels in R_c belong to the validation set.

iii Create $f_c(\boldsymbol{\theta}_c)$ from $\boldsymbol{P}_p^c$ (actually the parameters are used directly) using Equation 4.11 and draw $N \times \boldsymbol{\theta}_c$ from it.

iv Calculate $\mathbf{s}(t)$ with Equation 4.1, by using the harmonic parameters of step i.

v Determine the residual by subtracting $\mathbf{s}(t)$ from $\mathbf{x}(t)$.

vi Compute $\boldsymbol{P}_\eta^c$ from the residual, by using the same training set as in step ii and Equation 4.12.

vii Calculate N time-series of correlated increments $\boldsymbol{\Delta W}_c$ using $\boldsymbol{P}_\eta^c$, Equation 4.13 and Equation 4.15.

viii Generate correlated noise by using the noise parameters of $\boldsymbol{\theta}_c$ (drawn in step iii), $\boldsymbol{\Delta W}_c[i]$ and Equation 4.12.

ix Create the simulated harmonic component by using the harmonic parameters of $\boldsymbol{\theta}_c$ and Equation 4.1.

x Add the correlated noise to the harmonic component.

xi Generate NDVI from the simulated data by using band 1 and 2.

The pixel metrics simulated dataset $\mathscr{S}$ is constructed by using a different approach. The steps required to generate the dataset $\mathscr{S}$ ($|\mathscr{S}| = |R_c|$) to which the pixel metrics are applied are summarised below:

I-III. Follow steps i, iv and v of the class generation algorithm.

IV. Execute step vi of the class generation algorithm, but use all of the pixels in R_c.

V. Perform step vii of the class generation algorithm, but generate $|R_c|$ time-series instead of N.

VI. Generate correlated noise by using the estimated noise parameters derived in step i instead of the noise parameters of $\boldsymbol{\theta}_c$.

VII. Add the correlated noise to the harmonic signal generated in step ii.

VIII. Generate NDVI from the simulated data by using bands 1 and 2.

5.2.2 Preliminary validation results

At this point it would be useful to provide initial visual evidence of the validity and usefulness of the CSHO simulator. In Figure 5.23 the temporal Hellinger distance between the single-band time-varying models of S_v and S_s are displayed for the Gauteng dataset. The datasets S_v and S_s have 1000 elements and are single random instances generated with a modified version of the class metric simulated dataset algorithm presented in Section 5.2.1. The modification entails that all the data of R_c are used (not only 50%) as training data. The subscripts v and s are only added to S if their absence causes ambiguity.

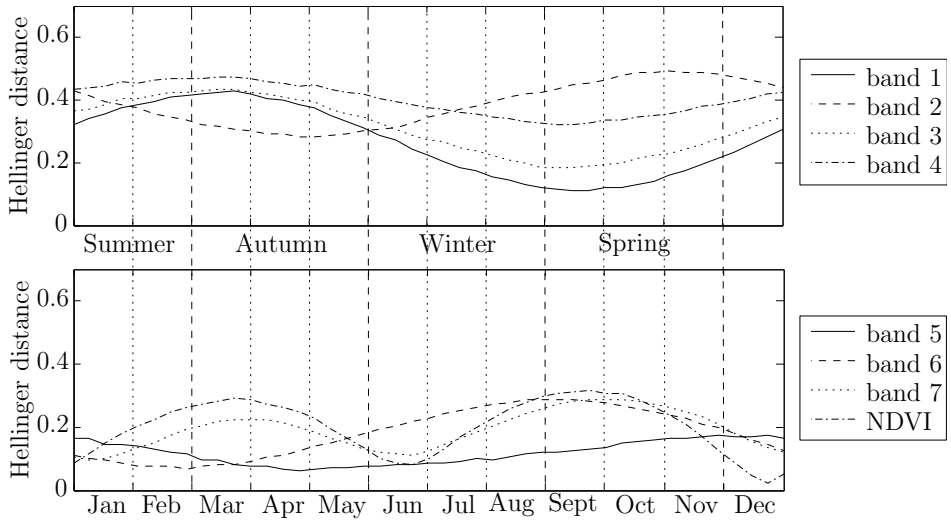


Figure 5.23: The temporal Hellinger distance between the single-band time-varying models of S_v and S_s (Gauteng).

Another important visual aid is Figure 5.24, which displays the temporal Hellinger distance between the single-band time-varying models of R_v and S for the Gauteng dataset. To avoid repetition the other combinations of R_c and S are not displayed. Figure 5.24 is very important, as it shows how accurately the simulated dataset replicated the original dataset in terms of the yearly average temporal behaviour.

Similar to Figure 5.23, Figure 5.25 displays the temporal Hellinger distance between the single-band

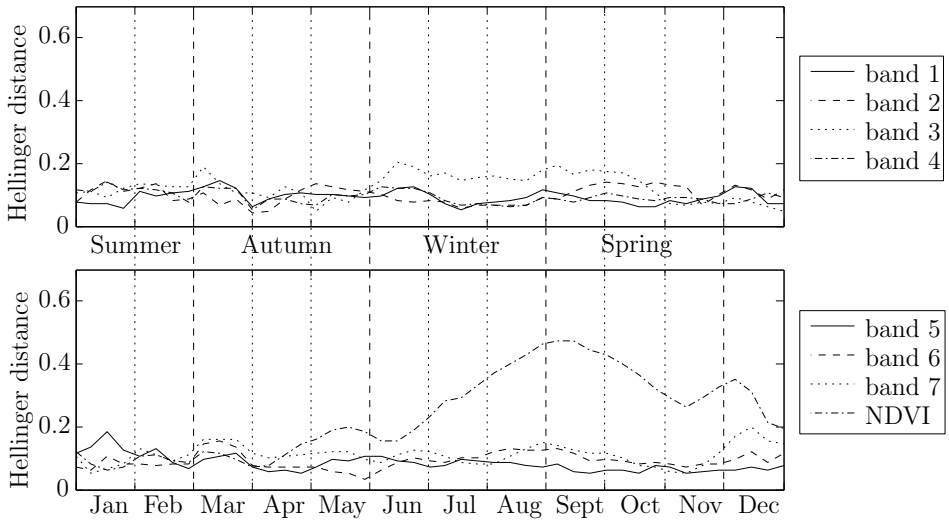


Figure 5.24: The temporal Hellinger distance between the single-band time-varying models of R_v and S (Gauteng).

time-varying models of S_v and S_s for the Limpopo dataset.

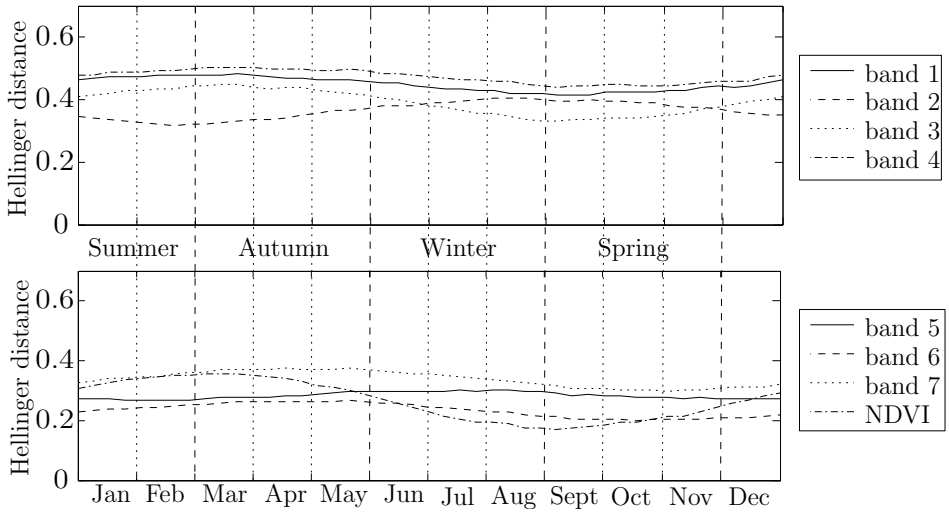


Figure 5.25: The temporal Hellinger distance between the single-band time-varying models of S_v and S_s (Limpopo).

The following observations can be made when inspecting Figure 5.23, Figure 5.24 and Figure 5.25:

1. The curves in Figure 5.23 and Figure 5.9 have similar profiles. The same can be said for Figure 5.25 and Figure 5.10.

2. The Hellinger distance for most bands in Figure 5.24 stay below 0.2. The exception is NDVI, which stays below 0.5. For most of the year however the Hellinger distance of NDVI is close to zero. The small Hellinger distance implies that the yearly temporal behaviour is replicated sufficiently by the CSHO simulator. The other possible combinations of R_c and S (although not presented) exhibit similar behaviour than that of Figure 5.24. Simular results are obtained for the Limpopo dataset.

5.2.3 Discussion of metric selection

The metrics in this section are based on the metrics proposed in [199]. Two underlying metrics are used, namely the Sum of Squared Error (SSE) and Hellinger distance (except for the power spectral density metric that measures power). In both cases a value *close to zero is desirable*. When "Hellinger" is not part of the metric name, it indicates that the SSE was used as the base metric. Each metric was chosen to verify that the simulator reproduces three important characteristics, namely temporal dynamics, spectral behaviour and accurate noise.

5.2.3.1 Temporal dynamics

There are two types of temporal dynamics to account for, namely intra-annual and inter-annual variation. The main reason for intra-annual variation is due to seasonality, which is caused by a wide range of factors including plant phenology. Inter-annual variation can be caused by many factors, including a drought or a flood. The *total model error* metric is a first-order statistic and is used to verify whether the average seasonal behaviour is replicated correctly. The *average temporal Hellinger distance* is probably the most important metric from the perspective of Section 4.2.3 and Section 4.3.3, as it measures the difference between the first-order statistical description of the CSHO and the true dataset. The average temporal Hellinger distance therefore measures whether the CSHO sufficiently replicates intra-annual variation and how effective the CSHO is in compensating for inter-annual variation. The *autocorrelation* metric is a second order statistic which measures whether the CSHO also models the temporal behaviour of any given pixel properly.

5.2.3.2 Spectral behaviour

As discussed in Section 4.1.2, the main aim of the simulator is to replicate spectral behaviour. It is well know that each class has a unique spectral signature within a certain allowable margin of varia-

tion [39]. The proposed simulator encapsulates and models the spectral signature for each class by using Equation 4.11. Equation 4.11 enforces the class-specific statistical restrictions imposed by the different CSHO model parameters of each spectral band on one another. The *parameter correlation metric* measures how effective the simulator is in reproducing spectral dependence (under the assumption of Gaussianity), while the *average parameter Hellinger distance* measures how trustworthy the joint Gaussian assumption of $f_c(\tilde{\boldsymbol{\theta}}_c)$ is.

Furthermore, the model also enforces noise correlation by using the approach presented in Sections 4.1.2.5 and 4.1.2.6. The *noise correlation metric* measures how well the noise correlation is modelled.

5.2.3.3 Accurate noise

A widely used assumption for remotely sensed time-series noise is that it is white [74, 78] if all information-carrying frequency components have been extracted [31]. The different *power spectral density* metric values reveal whether a white or coloured assumption is more appropriate when using an SHO as the underlying deterministic model. The *average noise increment Hellinger distance* determines whether the noise increments of each pixel are similar to the increments of the Ornstein-Uhlenbeck process.

5.2.4 Class metrics

The total model error, the average parameter Hellinger distance, the parameter and noise correlation and the average temporal Hellinger distance are respectively discussed in Section 5.2.4.1, Section 5.2.4.2, Section 5.2.4.3 and Section 5.2.4.4. A few figures are presented in this section to aid the reader in understanding the different class metrics. These figures were generated by comparing mostly R_v and S. The dataset S used by the figures is constructed by using the same procedure detailed in Section 5.2.1. Again, to avoid repetition, only the graphs for the Gauteng vegetation class are presented.

5.2.4.1 Total model error

The equation for the total model error is given by

$$\int_0^I \|\tilde{\mathbf{y}}_c^{R_c}(t) - \tilde{\mathbf{y}}_c^{S}(t)\|_2^2 dt, \tag{5.4}$$

where $\tilde{\mathbf{y}}_c(t)$ is the yearly ensemble mean of c and is defined in Equation 4.5. To determine the SSE of each time step in the year Equation 5.4 needs to be divided by 45. To give some insight into the total model error metric, Figure 5.26 presents the yearly ensemble means of R^v and S for the Gauteng dataset.

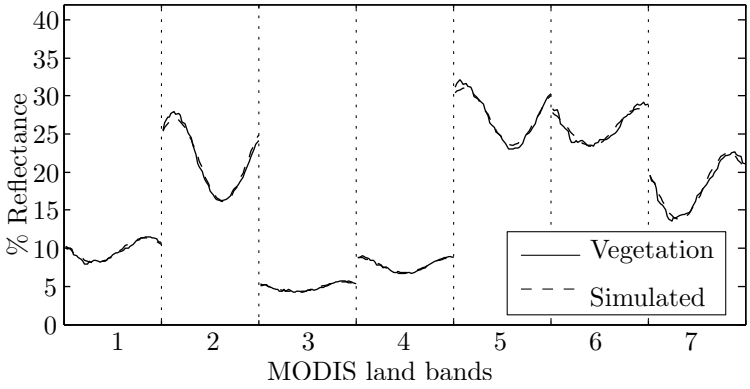


Figure 5.26: The yearly ensemble mean of R^v and S (Gauteng).

It is clear from Figure 5.26 that the total model error between R^v and S (for Gauteng) is close to zero.

5.2.4.2 Average parameter Hellinger distance metric

The equation for the average parameter Hellinger distance is equal to

$$\frac{1}{|\tilde{\boldsymbol{\theta}}_c|} \sum_{k=1}^{|\tilde{\boldsymbol{\theta}}_c|} \mathrm{HD}(f_c^{R_c}(\theta_k), f_c^{S}(\theta_k)),$$

where $f_c(\theta_k)$ is the marginal probability density function of $f_c(\tilde{\boldsymbol{\theta}}_c)$ and $\mathrm{HD}(f_c^{R_c}(\theta_k), f_c^{S}(\theta_k))$ represents the Hellinger distance between $f_c^{R_c}(\theta_k)$ and $f_c^{S}(\theta_i)$. Figure 5.27 ought to make the definition of the average parameter Hellinger distance clearer. According to Figure 5.27 the parameter Hellinger distance is greatest for **A**.

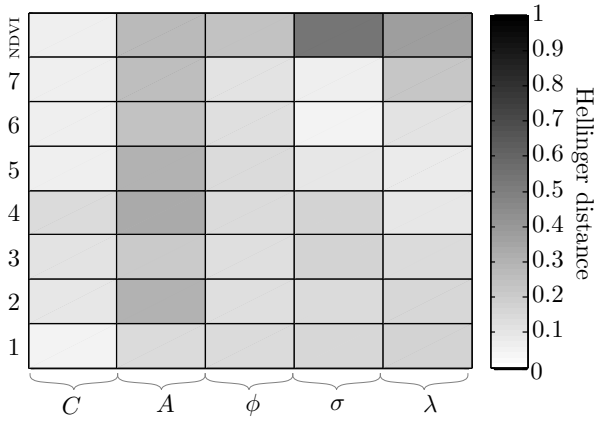


Figure 5.27: $\mathrm{HD}(f_v^{R_v}(\theta_i), f_v^{S}(\theta_i))$ (Gauteng).

5.2.4.3 Parameter and noise correlation metrics

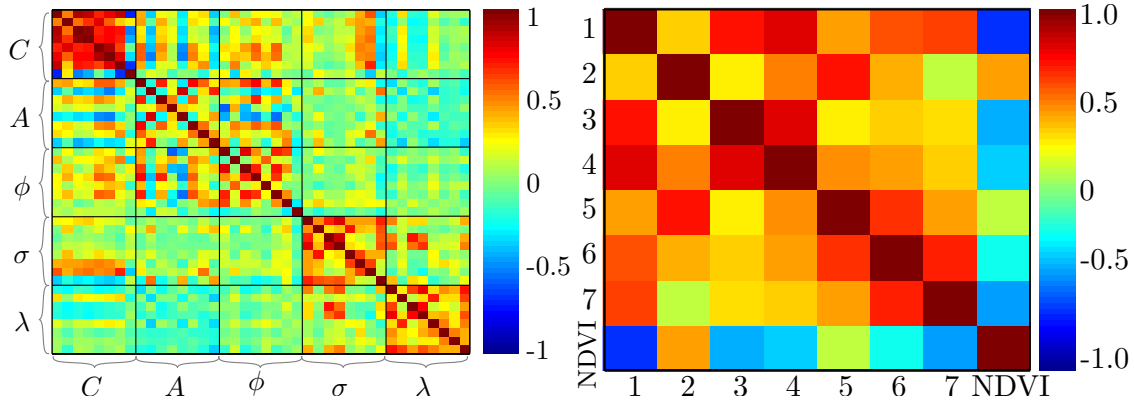
The equations for the noise and parameter correlation metrics are given by

$$\|\tilde{\boldsymbol{P}}^c_{p_{R_c}} - \tilde{\boldsymbol{P}}^c_{p_S}\|_2^2$$

and

$$\|\tilde{\boldsymbol{P}}^c_{\eta_{R_c}} - \tilde{\boldsymbol{P}}^c_{\eta_S}\|_2^2.$$

The noise correlation metric needs to be divided by 8×8 (NDVI was added for completeness), while the parameter correlation metric needs to be divided by 40×40 to determine the average SSE. For convenience the matrices $\tilde{\boldsymbol{P}}^v_{p_S}$ and $\tilde{\boldsymbol{P}}^v_{\eta_S}$ are displayed in Figure 5.28. There is not much difference between Figure 5.28a and Figure 5.13a. The same goes for Figure 5.28b and Figure 5.17a.



(a) $\tilde{\boldsymbol{P}}^v_{p_S}$. **(b)** $\tilde{\boldsymbol{P}}^v_{\eta_S}$.

Figure 5.28: $\tilde{\boldsymbol{P}}^v_{p_S}$ and $\tilde{\boldsymbol{P}}^v_{\eta_S}$ (Gauteng).

5.2.4.4 Average temporal Hellinger distance metric

The equation for the average temporal Hellinger distance is expressed as

$$\frac{1}{8}\sum_{b=1}^{\text{NDVI}}\frac{1}{I}\int_0^I \text{HD}(f^{R_c}_{x_c^b(t)}, f^{S}_{x_c^b(t)})\,dt,$$

where $f_{x_c^b(t)}$ is the probability density function in band b at time step t. Figure 5.29 and Figure 5.30 are visual aids to help explain the average temporal Hellinger distance metric. The reflectance probability density functions at time step 3 (of 45) for all eight years in MODIS band 1 for R_v and S (Gauteng) are displayed in Figure 5.29. The probability functions of S seem to be symmetrical about the mean (and almost identical) of the mean values of the densities of R_v. The temporal Hellinger distance for time steps $3+45n,\ n=\{0,\cdots,8\}$, in band 1 is calculated by determining the Hellinger distance between the probability density functions in Figure 5.29. The Hellinger distance between the probability density functions of the observation time steps of MODIS band 4 for R_v and S (Gauteng) is displayed in Figure 5.30. The mean and variance for the curve in Figure 5.30 are respectively 0.1926 and 0.0051, which is closer to zero than one.

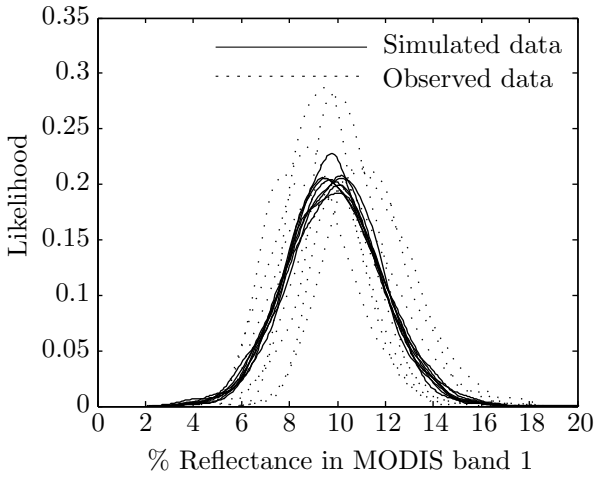


Figure 5.29: The reflectance probability density functions at time step period 3 (of 45) for all eight years in MODIS band 1 for R_v and S (Gauteng).

5.2.5 Pixel metrics

The autocorrelation metric, the average noise increment Hellinger distance and the power spectral density metric are respectively discussed in Section 5.2.5.1, Section 5.2.5.2 and Section 5.2.5.3. A few figures are displayed in this section to help the reader comprehend the different pixel metrics.

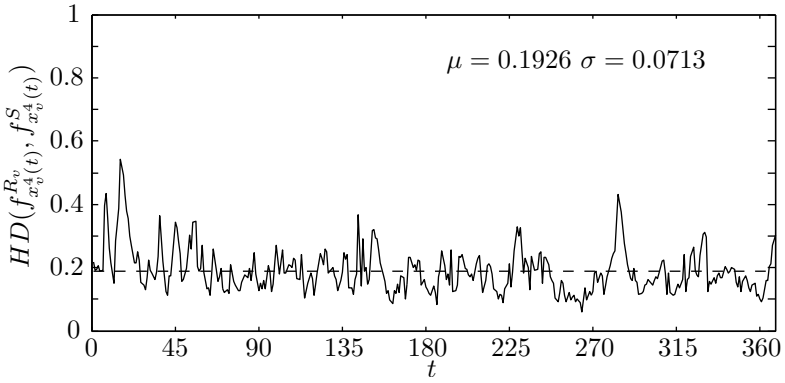


Figure 5.30: The Hellinger distance between the probability density functions of the observation time steps of MODIS band 4 for R_v and S (Gauteng).

These figures were created by comparing mostly random pixels from R_v and $\mathscr{S}$. The dataset $\mathscr{S}$ used to create the figures is constructed by using the same procedure described in Section 5.2.1. Again, to avoid repetition, only the graphs for the Gauteng vegetation class are presented.

5.2.5.1 Autocorrelation metric

The autocorrelation metric is expressed mathematically as

$$\frac{1}{|R_c|}\sum_{p=1}^{|R_c|}\int_0^I \|\tilde{\mathbf{R}}_c^{R_c(p)}(\tau) - \tilde{\mathbf{R}}_c^{\mathscr{S}(p)}(\tau)\|_2^2 \, d\tau, \tag{5.5}$$

where $\tilde{\mathbf{R}}_c^{R_c(p)}$ is the autocorrelation (defined in Equation 4.6) of the p-th pixel in R_c, while $\tilde{\mathbf{R}}_c^{S(p)}$ is defined similarly. To determine the average SSE per lag value Equation 5.5 needs to be divided by 368 (number of observations). Figure 5.31 displays the curves of a random vegetation pixel in Gauteng, as well as its replicated counterpart (belonging to $\mathscr{S}$). Clearly the two curves in Figure 5.31 are very similar.

5.2.5.2 Average noise increment Hellinger distance metric

The equation for the average noise increment Hellinger distance is defined as

$$\frac{1}{8}\sum_{b=1}^{\text{NDVI}}\frac{1}{|R_c|}\sum_{p=1}^{|R_c|}\text{HD}(f_{\Delta\eta^b}^{R_c(p)}, f_{\Delta\eta^b}^{\mathscr{S}(p)}),$$

where $f_{\Delta\eta^b}^{R_c(p)}$ is the density function of the noise increments $\eta^b[t+1]-\eta^b[t]$ for pixel p in dataset R_c, while $f_{\Delta\eta^b}^{\mathscr{S}(p)}$ is defined similarly.

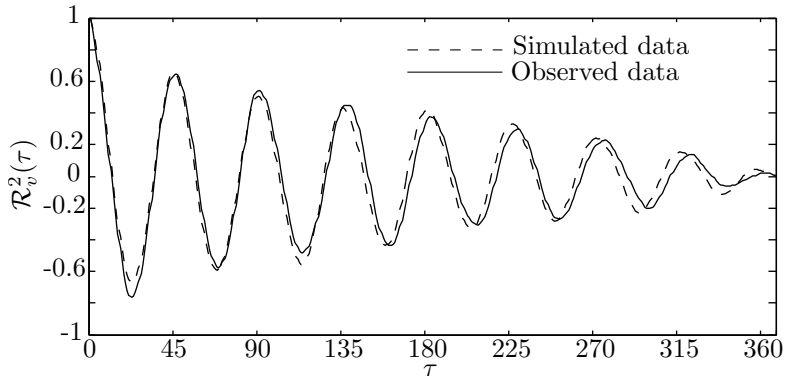


Figure 5.31: The autocorrelation function of a real and simulated vegetation pixel in band 2 (Gauteng).

5.2.5.3 Power spectral density metric

The equation for the power spectral density metric is equal to

$$\frac{1}{8}\sum_{b=1}^{\text{NDVI}}\frac{1}{|R_c|}\sum_{p=1}^{|R_c|}\int_0^{0.1} D_b^{R_c^\eta(p)}(f)\,df,$$

where $D_b^{R_c^\eta(p)}(f)$ is the power spectral density of the estimated noise of pixel p in dataset R_c in band b. The same metric can be applied to $\mathscr{S}^\eta$ and W^η, where $\mathscr{S}^\eta$ is the coloured noise (Ornstein-Uhlenbeck) representation of R_c^η (R_c after subtracting the SHO) and W^η is the white noise representation of R_c^η. Figure 5.32 displays $D_2^{R_v^\eta(p)}(f)$, $D_2^{\mathscr{S}^\eta(p)}(f)$ and $D_2^{W^\eta(p)}(f)$ for a random vegetation pixel p in the Gauteng dataset. From Figure 5.32 it is clear that $\int_0^{0.1} D_2^{R_v^\eta(p)}(f)\,df \approx \int_0^{0.1} D_2^{\mathscr{S}^\eta(p)}(f)\,df$, while $\int_0^{0.1} D_2^{R_v^\eta(p)}(f)\,df \neq \int_0^{0.1} D_2^{W^\eta(p)}(f)\,df$.

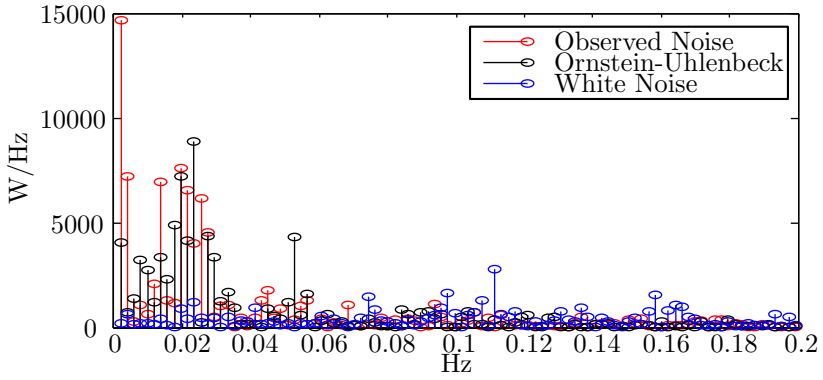


Figure 5.32: The power spectral density of the estimated noise of a vegetation pixel in MODIS band 2 along with its coloured and white representations.

5.2.6 Discussion of simulator results

Random split cross-validation was performed to create 50 different class metric simulated data sets, with the algorithm presented in Section 5.2.1 and $N = 1000$. No cross-validation was used for the pixel metrics, since all the pixels in R_c were used. For the pixel metrics 50 independent experiments were also conducted. Each data set for each experiment was independently created via the algorithm in Section 5.2.1. The class and pixel metrics were then applied to each experiment to produce the results in Table 5.7 and Table 5.8.

Table 5.7: Difference metrics between R_c, S and $\mathscr{S}$ (50 experiments) for the Gauteng dataset. The v index indicates vegetation, while the s index indicates settlement.

Metric Name	R_v	R_s
Model Error	58.3917 ± 8.1991	52.8838 ± 22.9432
Parameter Hellinger Distance	0.1817 ± 0.0070	0.2563 ± 0.0081
Parameter Correlation	19.6224 ± 3.6121	32.5042 ± 5.7380
Noise Correlation	0.0371 ± 0.0085	0.0629 ± 0.0101
Temporal Hellinger Distance	0.2349 ± 0.0042	0.2266 ± 0.0094
Autocorrelation	31.0137 ± 0.2431	34.3992 ± 0.4749
Noise Hellinger Distance	0.1674 ± 0.0003	0.1755 ± 0.0003
Power in R_η^c	92.7649	38.3202
Power in S_η	94.6440 ± 0.6564	46.0178 ± 0.2901
Power in W_η	22.9897 ± 0.0642	11.7971 ± 0.0542

The following observations can be made from Table 5.7 and Table 5.8:

1. Relative to the definitions of the SSE and Helinger distance metrics, the results in Table 5.7 and Table 5.8 are close to zero, implying that the simulator accurately replicates the temporal dynamics and spectral characteristics of the MODIS datasets. The small variances in Table 5.7 and Table 5.8 imply that the metric results are stable and reliable.

2. Relative to the other classes the Gauteng settlement class has a higher standard deviation on its metric results, which can be explained by the fact that the Gauteng settlement dataset is much smaller than the other datasets.

Table 5.8: Difference metrics between R_c, S and $\mathscr{S}$ (50 experiments) for the Limpopo dataset. The v index indicates vegetation, while the s index indicates settlement.

Metric Name	R_v	R_s
Model Error	68.5056 ± 11.7454	56.8919 ± 11.3777
Parameter Hellinger Distance	0.2045 ± 0.0054	0.2097 ± 0.0041
Parameter Correlation	15.4586 ± 2.6663	12.9429 ± 1.8421
Noise Correlation	0.0422 ± 0.0063	0.0421 ± 0.0064
Temporal Hellinger Distance	0.2147 ± 0.0047	0.1939 ± 0.0033
Autocorrelation	31.6738 ± 0.1975	30.4443 ± 0.1531
Noise Hellinger Distance	0.1630 ± 0.0002	0.1582 ± 0.0002
Power in R_η^c	110.6438	80.5593
Power in S_η	111.0334 ± 0.45182	85.4707 ± 0.3112
Power in W_η	26.8485 ± 0.0513	21.7033 ± 0.0368

3. Relative to the metric definitions the autocorrelation metric has the largest value, which is understandable since the non-stationarity that is present in the MODIS data shows up in the autocorrelation metric. The error incurred due to parameter estimation also affects the autocorrelation metric.

4. For the power metric in general $R_\eta^c \approx S_\eta$, while $R_\eta^c \neq W_\eta$, which implies that a coloured noise model is more appropriate for the current datasets.

5. There is no standard deviation for the power in R_η^c as the power of R_η^c obviously only needs to be calculated once.

5.3 CLASSIFICATION RESULTS: GAUTENG AND LIMPOPO

The focus of this section is on the classification performance of the noise-harmonic feature group $\boldsymbol{\theta}$ derived from Equation 4.4 and discussed in Section 4.2.4.2. The feature group $\boldsymbol{\theta}$ under investigation extends the feature group $\boldsymbol{\iota}$ proposed in [5]. The feature group $\boldsymbol{\theta}$ is also compared with the temporal feature group $\boldsymbol{\zeta}$, which is discussed in Section 4.2.4.2 [15]. The different feature groups are used as inputs to SVM classifiers (discussed in Section 4.2.4). The SVM classifiers are also compared

with two benchmarking techniques, namely the minimum distance classifier and the time-varying maximum likelihood classifier, which are discussed in Section 4.2.2 and Section 4.2.3 respectively. The time-varying maximum likelihood classifier is especially important, as it is based on sequential analysis. This section starts by introducing the different classification accuracy metrics employed, followed by the presentation of classification results. The classification results of the Gauteng dataset are presented in Section 5.3.3, Section 5.3.4 and Section 5.3.5. The classification results of the Limpopo dataset are presented in Section 5.3.6.

5.3.1 Classification accuracy metrics

Two classification accuracy metrics are used in this section, namely *Overall Accuracy (OA)* and the κ-*coefficient*. The κ-coefficient is especially useful, since it can determine whether the values contained in an error matrix represent a result significantly better than random [200, 201]. The two metrics in question are computed by using an error matrix, which is a matrix containing the number of pixels classified correctly and incorrectly for each class under consideration. An example error matrix is presented in Table 5.9.

Table 5.9: Error matrix used to explain the definition of OA and the κ-coefficient.

	Class 0	**Class** 1	
Class 0	x_{11}	x_{12}	x_{1*}
Class 1	x_{21}	x_{22}	x_{2*}
	x_{*1}	x_{*2}	N

OA is a percentage (obviously the closer the metric is to 100 the better the classifier) and is defined as

$$100 \times \sum_{i=1}^{r} x_{ii}/N, \tag{5.6}$$

while the κ-coefficient is computed with

$$\kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i*} \times x_{*i}}{N^2 - \sum_{i=1}^{r} x_{i*} \times x_{*i}}. \tag{5.7}$$

In Equation 5.6 and Equation 5.7, N is the total number of pixels in the error matrix, r is the number of rows in the matrix, x_{ii} is the number in row i and column i, x_{i*} is the total for row i, and x_{*i} is the total for column i (see Table 5.9 for more details). Note that x_{11} refers to the number of pixels belonging to class 0 classified correctly and x_{22} refers to the number of pixels belonging to class

1 which were classified correctly, while x_{21} refers to the number of pixels belonging to class 0 which were not classified correctly and x_{12} refers to the number of pixels belonging to class 1 which where not classified correctly. The value of κ can be smaller than or equal to 1. The magnitude guidelines for κ published in [202] are adopted here. According to [202], a larger κ indicates that the classifier can easily discern the different classes, while a smaller (or even negative) κ indicates that the classifier cannot easily discern the different classes.

5.3.2 Structure used for accuracy metrics

In this section a three-dimensional structure is presented that is used to organise the cross-validation classification results of Section 5.3.3 to Section 5.3.7.

The κ-coefficients and the OA percentages generated by the cross-validation experiments in Section 5.3.3 to Section 5.3.6 can be organised into a three-dimensional irregular structure $\Xi_{\upsilon,\mu}$, where

$$\upsilon \in \{\text{`}\kappa\text{'}, \text{`OA'}\}$$

and

$$\mu \in \{\text{Min Dist,TVML}, \boldsymbol{\theta}, \boldsymbol{\iota}, \boldsymbol{\zeta}\} \tag{5.8}$$

with elements

$$\xi^{\mu}_{x,y(x),z} \in \{\kappa, \text{OA}\},$$

where

$$\begin{aligned} x &\in \{1,2,\cdots,8\}; \\ y(x) &\in \left\{1,2,\cdots,\binom{x}{8}\right\}; \\ z &\in \{1,2,\cdots,e\}. \end{aligned} \tag{5.9}$$

In Equation 5.8, Min Dist,TVML, $\boldsymbol{\theta}, \boldsymbol{\iota}$ and $\boldsymbol{\zeta}$ are respectively associated with the minimum distance classifier, the time-varying maximum likelihood classifier, the noise-harmonic feature group, the harmonic feature group and the temporal feature group defined in Section 4.2.4.2. When μ is omitted it implies that the classification procedure that generated Ξ_{υ} is unknown.

In Equation 5.9, x represents the band restriction value (the amount of bands that may be used for classification), while each y is associated with a unique band combination given the restriction of x (which explains the combination notation used to define y). The z index points to a specific cross-validation experiment and $e \in \{1,2,\cdots\}$ denotes the amount of experiments performed.

Recall that each MODIS pixel consists of a time-series. The notation $\Xi_\upsilon[n]$ should be interpreted as representing a structure similar to Ξ_υ, with the only difference being that it consists of classification accuracy metric elements obtained by using truncated ($n \in \{1,2,\cdots,368\}$) MODIS pixels. It should be clear that if $[n]$ is omitted from $\Xi_\upsilon[n]$ then it implies that no truncation was performed before classification commenced.

The notation $\Phi_\upsilon = \mathbb{E}_\omega[\Xi_\upsilon]$ should be interpreted as the sample mean of Ξ_υ along the dimension $\omega \in \{x,y,z\}$ (which is similar to the Matlab *mean* command). The resulting structure Φ_υ has only two dimensions, since $\mathbb{E}_\omega[]$ eliminated a dimension of Ξ_υ. The notation $\Theta_\upsilon = \mathbb{E}_{\eta\neq\omega}[\mathbb{E}_\omega[\Xi_\upsilon]]$ should be interpreted as the sample mean of Ξ_υ along the dimension ω, followed by the sample mean of Φ_υ along the dimension $\eta \in \{x,y,z\}$. Note that the resulting structure Θ_υ has only one dimension. The structures $\sigma_\omega\{\Xi_\upsilon\}$ and $\sigma_{\eta\neq\omega}\{\sigma_\omega\{\Xi_\upsilon\}\}$ should be interpreted in a similar way, except for the fact that σ indicates that standard deviation (which is similar to the Matlab *std* command) should be used instead of the sample mean when deriving $\sigma_\omega\{\Xi_\upsilon\}$ and $\sigma_{\eta\neq\omega}\{\sigma_\omega\{\Xi_\upsilon\}\}$. The notation

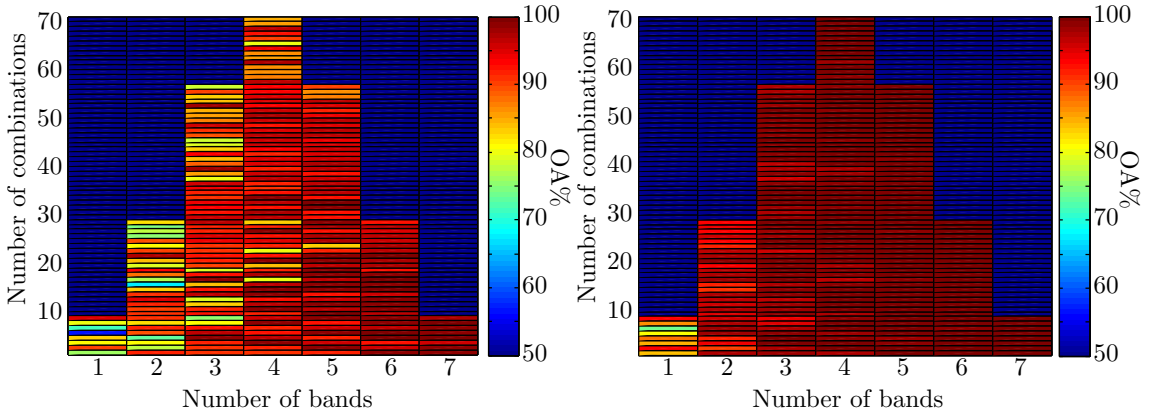
$$\Sigma_\upsilon = \Xi_\upsilon^{\{1,2,\cdots,7\},*,1}$$

should be interpreted as a substructure of Ξ_υ which only contains the elements in Ξ_υ for which $x = \{1,2,\cdots,7\}$ and $z = 1$. The $*$ is a wild card which indicates all valid values y can ascertain. If one of the dimensions of a structure Σ_υ is equal to one, then $(\Sigma_\upsilon)^@$ should be interpreted as the structure obtained after the redundant dimension is removed. When two structures Ξ_{υ,μ_1} and Ξ_{υ,μ_2} are compared, then classification approach μ_1 performs better than μ_2 if $\xi^{\mu_1}_{x,y(x),z} > \xi^{\mu_2}_{x,y(x),z}$ for more than 50% of all the possible index values (see Equation 5.9).

5.3.3 Preliminary benchmark classification results: Gauteng

In this section the results of the benchmark classification approaches are presented. The benchmark classifiers that were used are the minimum distance classifier and the time-varying maximum likelihood classifier presented in Section 4.2.2 and Section 4.2.3 respectively [23]. These two approaches were selected, as they are frontier hypertemporal approaches. As the algorithms are only used for

benchmarking, the full vegetation and settlement datasets were used for training and validation (implying that $e = 1$). The estimation procedure for the yearly ensemble means used by the minimum distance classifier is discussed in Section 5.1.1, while the estimation of the time-varying model used by the time-varying maximum likelihood classifier is discussed in Section 5.1.2. Moreover, no sequential thresholds were used for the time-varying maximum likelihood classifier and $\pi = 0.5$. The OA classification results of the benchmarking approaches for each possible band combination are found in Figure 5.33.



(a) $\left(\Xi^{\{1,2,\cdots,7\},*,1}_{\text{'OA'},\text{Min Dist}}\right)^{@}$. **(b)** $\left(\Xi^{\{1,2,\cdots,7\},*,1}_{\text{'OA'},\text{TVML}}\right)^{@}$.

Figure 5.33: The classification results (OA) of the benchmarking approaches of the Gauteng dataset.

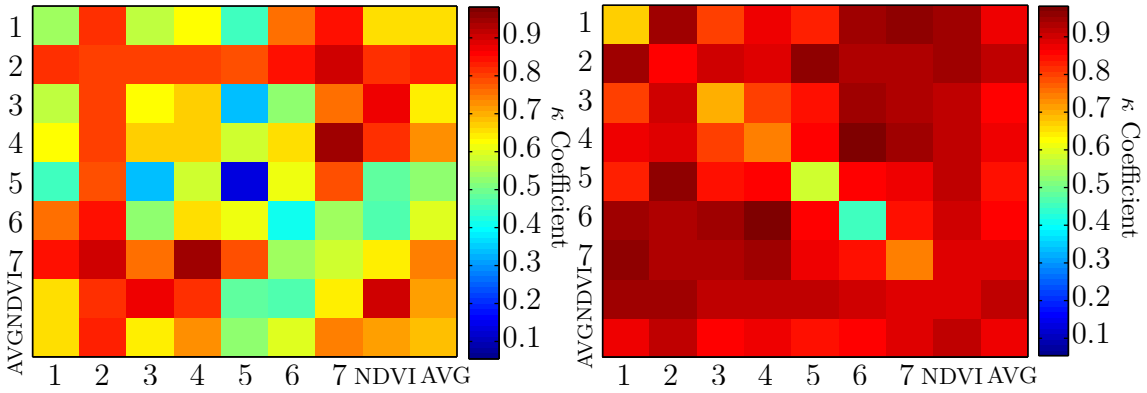
Each possible two-band combination κ-coefficient for the benchmarking approaches is presented in Figure 5.34.

The following observations and conclusions can be made from Figure 5.34:

1. With respect to the minimum distance classifier, the bands that in combination perform best are $\{2,4,7,\text{NDVI}\}$, while bands $\{1,3,5,6\}$ perform worst (in terms of classification capability). The band that in combination performs best is band 2, while band 5 performs worst. The band combination that separates the two classes best is $\{4,7\}$, while the lowest κ value is produced by $\{1,5\}$.

2. The results are exactly the same for the time-varying maximum likelihood classifier, except that $\{4,6\}$ is the band combination that performs best, while band combination $\{1,3\}$ performs worst.

3. For both approaches NDVI performs best (single-band). The worst performing bands are res-



(a) $\left(\Xi^{\{1,2\},*,1}_{\text{`}\kappa\text{'},\text{Min Dist}}\right)^{@}$ packed into a symmetrical matrix (plus a mean row and column). **(b)** $\left(\Xi^{\{1,2\},*,1}_{\text{`}\kappa\text{'},\text{TVML}}\right)^{@}$ packed into a symmetrical matrix (plus a mean row and column).

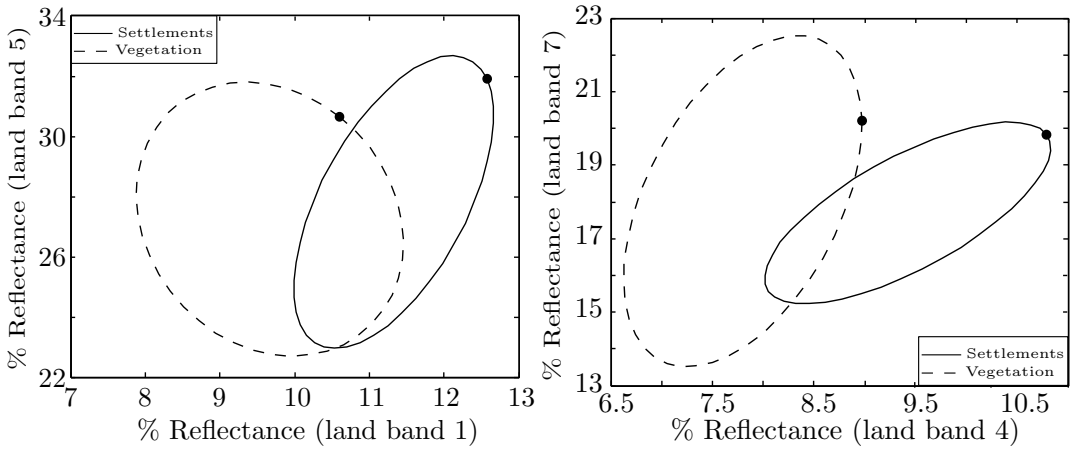
Figure 5.34: The two-band classification results (κ) of the benchmarking approaches of the Gauteng dataset.

pectively band 5 and 6 in the case of the minimum distance classifier and the time-varying maximum likelihood classifier.

4. Recall from Section 5.1.2.3 that the temporal Hellinger metric predicted that a temporal classifier using bands $\{1,2,3,4\}$ would perform better than a temporal classifier using bands $\{5,6,7,\text{NDVI}\}$. As seen in Figure 5.34, this prediction is reasonably close to the actual observed behaviour of the two temporal benchmarking classifiers investigated in this section. A possible reason for the discrepancy between the observed behaviour and the predicted behaviour is discussed in Section 5.3.4.

5. Generally the time-varying maximum likelihood classifier outperforms the minimum distance classifier, as it uses a superior metric, namely the posterior sequence, which incorporates the yearly ensemble mean as well as the inter-class variance.

As an interesting side note, a closer look is taken at bands $\{4,7\}$ and $\{1,5\}$ from Figure 5.34a. The yearly ensemble means (after fitting appropriate sinusoids) of $\{4,7\}$ and $\{1,5\}$ are displayed in Figure 5.35a and Figure 5.35b, respectively. It is clear from Figure 5.35 that, in the case of bands 4 and 7, the yearly ensemble means of the settlement and vegetation classes have a greater distance between them than in bands 1 and 5. The increased distance observable when inspecting the yearly ensemble means of $\{4,7\}$ and $\{1,5\}$ helps to explain why the minimum distance classifier using bands $\{4,7\}$ outperforms the minimum distance classifier using bands $\{1,5\}$.



(a) Yearly ensemble mean models of bands 1 and 5. **(b)** Yearly ensemble mean models of bands 4 and 7.

Figure 5.35: Two-dimensional yearly ensemble mean models of Gauteng.

In the last part of this section the focus will shift to the time-varying maximum likelihood classifier. The time-varying maximum likelihood classifier is unique in the sense that it is based on sequential analysis (discussed in Section 3.4) and was first introduced (to the remote sensing field) in [23]. As the time-varying maximum likelihood classifier is based on sequential analysis it can be truncated, which allows the classifier to make a decision after each observation is received by using all the observations received up to that point. The focus of [23] was on the trade-off between classification accuracy and classification delay. The classification study in Section 5.3 however focused solely on increasing classification accuracy. One of the aims of this thesis, which was stated in Chapter 1, is to verify the sequential results presented in [23]. The most important result of [23] was therefore reproduced and can be found in Figure 5.36. Figure 5.36 presents the mean single-band classification κ-coefficient or average single-band classification performance of the time-varying maximum likelihood classifier as a function of time or sample size. If the notation from Section 5.3.2 is used then the mean single-band classification κ-coefficient (as a function of sample size) can be expressed as

$$\mathbb{E}_{y(1)}\left[\left(\Xi^{1,*,1}_{\text{`}\kappa\text{'},\text{Min Dist}}[n]\right)^{@}\right].$$

The delay and accuracy measures discussed in Section 4.2.3 are not calculated, as the datasets under consideration have finite sizes and Figure 5.36 thus displays an alternative fixed-sample-size performance measure.

Generally the average κ-coefficient in Figure 5.36 increases with time. A steep increase is observable in the first year and a smaller increase during the second year. After the second year however the increase of κ is small. The temporal dynamics of the average κ-coefficient in Figure 5.36 suggests that

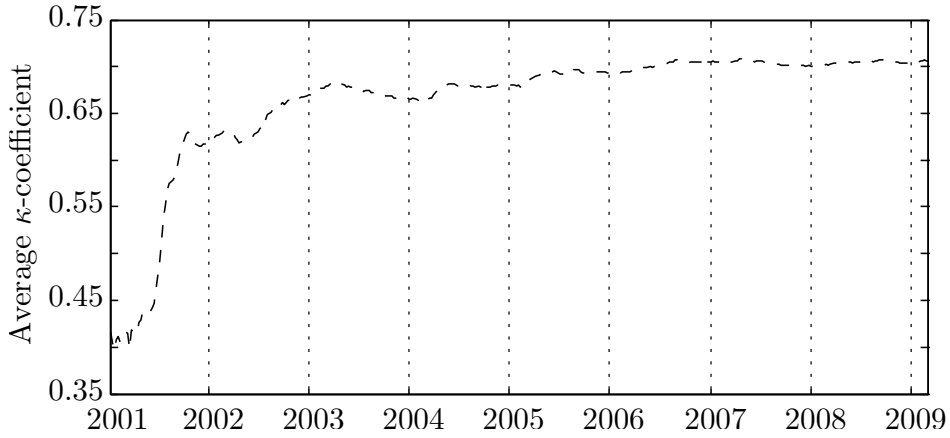


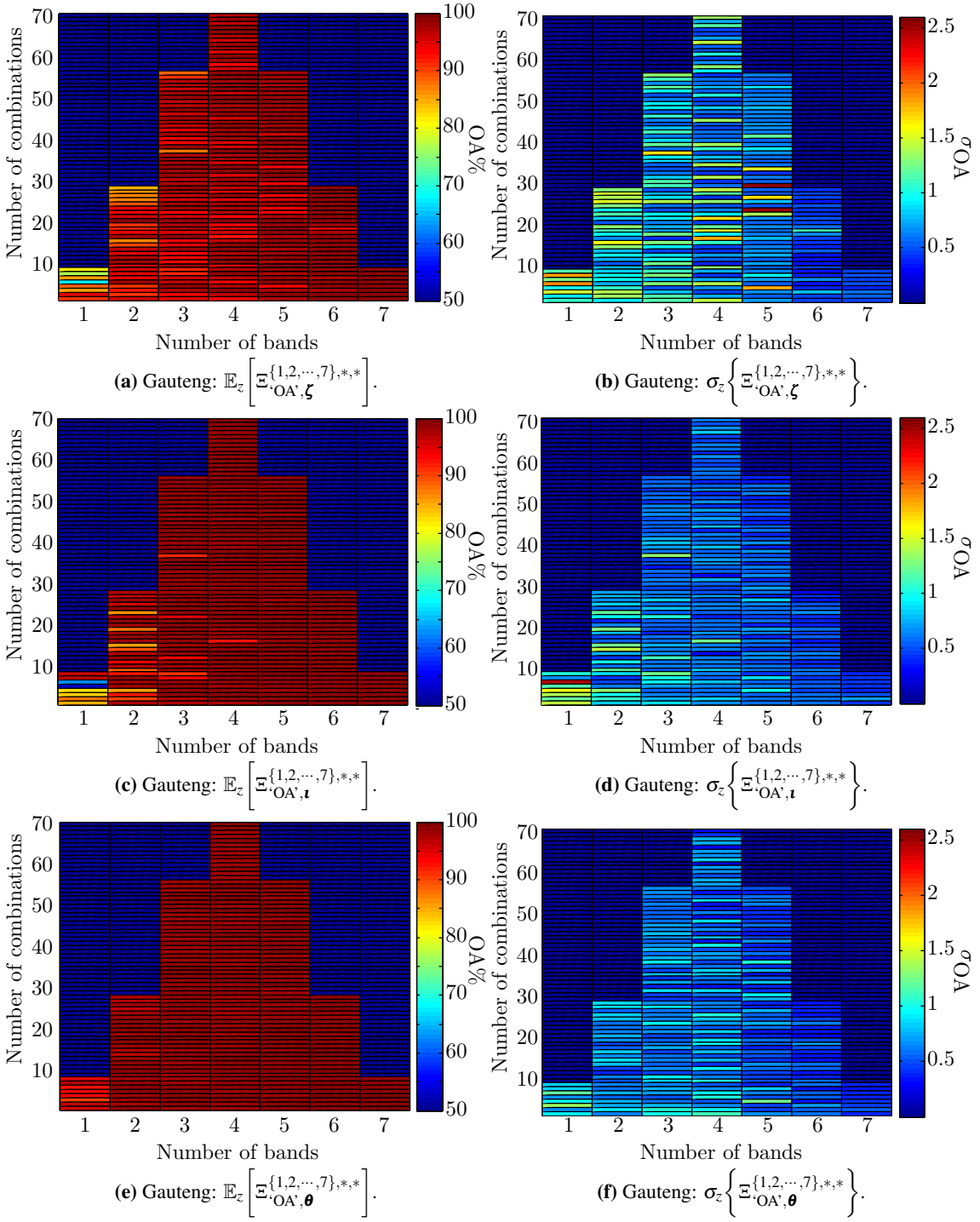
Figure 5.36: The average single-band classification performance of the time-varying maximum likelihood classifier for the Gauteng dataset as a function of time.

for the Gauteng dataset specifically the thresholds of the time-varying maximum likelihood classifier should be chosen in such a way that the classifier on average experiences (excluding outliers) at least a one-year delay (preferably two) before it can classify an observed sequence. The same result was obtained in another independent study, namely [78].

5.3.4 Preliminary SVM classification results: Gauteng

The SVM classifier and the proposed feature groups are discussed in Section 4.2.4 and Section 4.2.4.2 respectively. Recall from Section 4.2.4.2 that three feature groups are proposed, namely temporal features $\boldsymbol{\zeta}$, harmonic features $\boldsymbol{\iota}$ and noise-harmonic features $\boldsymbol{\theta}$. A linear SVM is used, since it produced sufficient classification results. The SVM is realised with the *SVM and Kernel Methods Matlab Toolbox* [203]. The SVM and Kernel Methods Matlab Toolbox requires two input parameters, which are determined via a standard grid search algorithm. The two input parameters that need to be set are (c, λ). The parameter c sets the bound on the Lagrangian multipliers, while λ is a conditioning parameter for the quadratic programming method used to determine the SVM hyperplane. Random split cross-validation (50% for training and 50% for validation) was employed by the grid search algorithm (50 independent experiments). The cross-validation OA results of the Gauteng dataset for all the possible band combinations of $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$ are presented in Figure 5.37. The standard deviation in Figure 5.37 for each feature group is small, indicating that the classification results are reliable and stable.

The Gauteng two-band classification results for $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$ are presented in Figure 5.38. The κ-

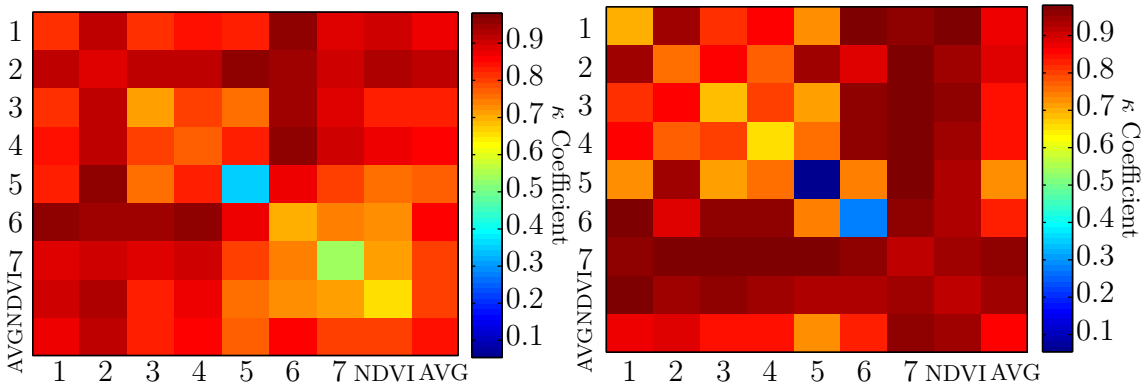


(a) Gauteng: $\mathbb{E}_z\left[\Xi_{\text{'OA'},\boldsymbol{\zeta}}^{\{1,2,\cdots,7\},*,*}\right]$. **(b)** Gauteng: $\sigma_z\left\{\Xi_{\text{'OA'},\boldsymbol{\zeta}}^{\{1,2,\cdots,7\},*,*}\right\}$.

(c) Gauteng: $\mathbb{E}_z\left[\Xi_{\text{'OA'},\boldsymbol{\iota}}^{\{1,2,\cdots,7\},*,*}\right]$. **(d)** Gauteng: $\sigma_z\left\{\Xi_{\text{'OA'},\boldsymbol{\iota}}^{\{1,2,\cdots,7\},*,*}\right\}$.

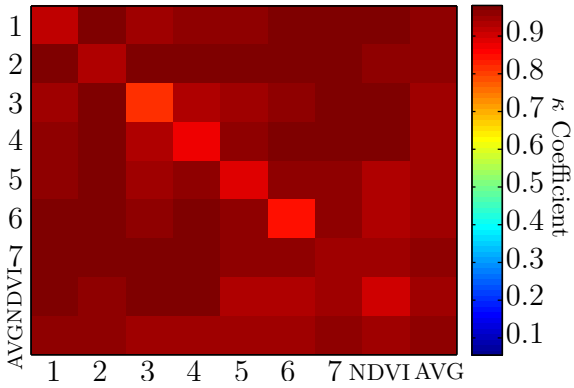
(e) Gauteng: $\mathbb{E}_z\left[\Xi_{\text{'OA'},\boldsymbol{\theta}}^{\{1,2,\cdots,7\},*,*}\right]$. **(f)** Gauteng: $\sigma_z\left\{\Xi_{\text{'OA'},\boldsymbol{\theta}}^{\{1,2,\cdots,7\},*,*}\right\}$.

Figure 5.37: The cross-validation results [with $(\infty, 0.05)$ and $e = 50$] of the Gauteng dataset for all the possible band combinations of $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$.

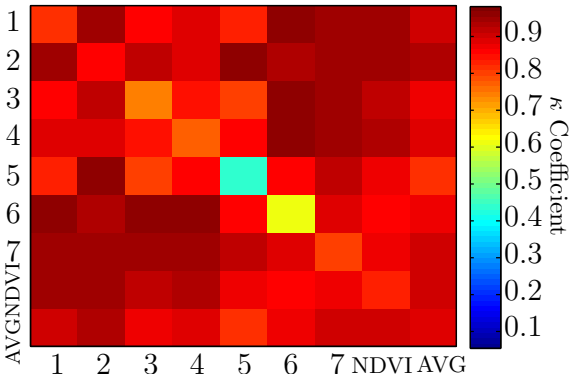


(a) Gauteng: $\mathbb{E}_z\left[\Xi^{\{1,2\},*,*}_{\kappa,\boldsymbol{\zeta}}\right]$ packed into a symmetrical matrix.

(b) Gauteng: $\mathbb{E}_z\left[\Xi^{\{1,2\},*,*}_{\kappa,\boldsymbol{\iota}}\right]$ packed into a symmetrical matrix.



(c) Gauteng: $\mathbb{E}_z\left[\Xi^{\{1,2\},*,*}_{\kappa,\boldsymbol{\theta}}\right]$ packed into a symmetrical matrix.



(d) Average between $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$ results for Gauteng.

Figure 5.38: The two-band classification results (average κ) of $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$ for the Gauteng dataset [with $(\infty, 0.05)$ and $e = 50$].

coefficients reported in Figure 5.38 are the average of 50 random split cross-validation experiments. The following observations and conclusions can be made from Figure 5.38:

1. When the SVM classifier using $\boldsymbol{\zeta}$ uses data from bands $\{1,2,4,6\}$ it classifies better than when the SVM classifier using $\boldsymbol{\zeta}$ uses data from bands $\{3,5,7,\text{NDVI}\}$. The SVM classifier using $\boldsymbol{\zeta}$ obtains its best classification results when data from band 2 are used in combination, while the SVM classifier using $\boldsymbol{\zeta}$ performs at its worst when data from band 5 are used in combination. The two-band combination $\{2,5\}$ performs best, while the two-band combination $\{7,\text{NDVI}\}$ performs worst of all the two-band combinations. The best and worst single-bands that the SVM classifier using $\boldsymbol{\zeta}$ can use are bands 2 and 5 respectively.

2. When the SVM using the harmonic features $\boldsymbol{\iota}$ uses data from bands $\{1,2,7,\text{NDVI}\}$ it achieves

better classification results than when it uses data from bands $\{3,4,5,6\}$. When the SVM using the harmonic features $\boldsymbol{\iota}$ use data from band 7 in combination it produces the best classification results, while using band 5 in combination leads to the worst classification results. Band combination $\{3,5\}$ performs best, while band combination $\{2,7\}$ performs worst of all the possible two-band combination. The best and worst single-bands that the SVM classifier using $\boldsymbol{\iota}$ can use are NDVI and band 5 respectively.

3. The SVM using the noise-harmonic features $\boldsymbol{\theta}$ perform better when it uses data from bands $\{1,2,4,7\}$ than when it uses data from bands $\{3,5,6,\text{NDVI}\}$. The highest average κ values are achieved when band 2 is used in combination, while the lowest average κ values are achieved when using band 3 in combination. The two-band combination that perform best is $\{1,2\}$, while $\{3,4\}$ perform the worst among all the possible two-band combinations. The best and worst single-bands that the SVM classifier using $\boldsymbol{\theta}$ can use are bands 2 and 3 respectively.

4. Generally for the Gauteng dataset, the SVM using $\boldsymbol{\theta}$ outperforms the SVMs using $\boldsymbol{\zeta}$ and $\boldsymbol{\iota}$. The SVM using $\boldsymbol{\iota}$ outperforms the SVM using $\boldsymbol{\zeta}$. The fact that the feature group $\boldsymbol{\theta}$ outperforms $\boldsymbol{\iota}$ is as expected, since the parameter Hellinger distance indicated in Section 5.1.3.3 that the noise and phase parameters of the CSHO are extra discerning features, which implies that they can be used to extend the classification potential of $\boldsymbol{\iota}$.

5. The average κ-coefficients of the SVMs using $\boldsymbol{\zeta}, \boldsymbol{\iota}$ and $\boldsymbol{\theta}$ are presented in Figure 5.38d, from which it is clear that on average combining $\{1,2,3,4\}$ with $\{6,7,\text{NDVI}\}$ produces high κ values. Furthermore, generally the following band combinations, $\{1,2\},\{5,2\}$ and $\{3,2\}$, also perform well.

6. Recall from Section 5.1.2.3 that for the Gauteng dataset the temporal Hellinger distance metric predicted that a classifier using data from bands $\{1,2,3,4\}$ would provide better class differentiability than a classifier using data from bands $\{5,6,7,\text{NDVI}\}$ when the classifier in question relies on temporal features. Similarly, recall from Section 5.1.3.3 that the parameter Hellinger distance metric predicted that a classifier using data from bands $\{1,2,7,\text{NDVI}\}$ would provide better class differentiability than a classifier using bands $\{3,4,5,6\}$ when the classifier in question employs the parameters of the CSHO as classification features. As seen in Figure 5.38, these two predictions are close to the actual observed behaviour of the three SVM classifiers investigated in this section. The correlation between the predictions and reality implies that the

separability metrics introduced in Section 5.1 can be used to select classification features. The small discrepancy between the observed behaviour and the predicted behaviour can be ascribed to the fact that the relation between actual performance and predicted separability based on a single metric is not a perfect one-to-one relation, as the single metric does not incorporate all the factors (for instance spectral dependence is not measured by the Hellinger metric) that influence the performance of a specific classifier.

5.3.5 Classification results: Gauteng

Finally, the bar graph representing the $\mathbb{E}_{y(x)}\left[\mathbb{E}_z\left[\Xi_{\text{`}\kappa\text{'},\mu}\right]\right]$ values for each classification approach mentioned (for all μ) is presented in Figure 5.39. Note that there is no cross-validation experiments for the benchmarking approaches. In the case of the benchmarking approaches only one sample mean is taken over all the band combinations. The values $\mathbb{E}_{y(x)}\left[\sigma_z\left\{\Xi_{\text{`}\kappa\text{'},\mu}\right\}\right]$ for the SVM approaches are displayed in Figure 5.39b. Recall that the index $x \in \{1,2,\cdots,8\}$ is used to indicate the band restriction value. So if $x = 5$, the band restriction value is equal to 5. The following observations can be made from Figure 5.39a and Figure 5.39b:

1. The SVM classifier using the noise-harmonic features outperform all the other classifiers when $x \leq 3$. When $x > 3$ however the time-varying maximum likelihood classifier performs better than the SVM classifier using the noise-harmonic features. Furthermore, when $x > 3$ the SVM classifier using the harmonic features produces similar classification results as the SVM classifier using the noise-harmonic features.

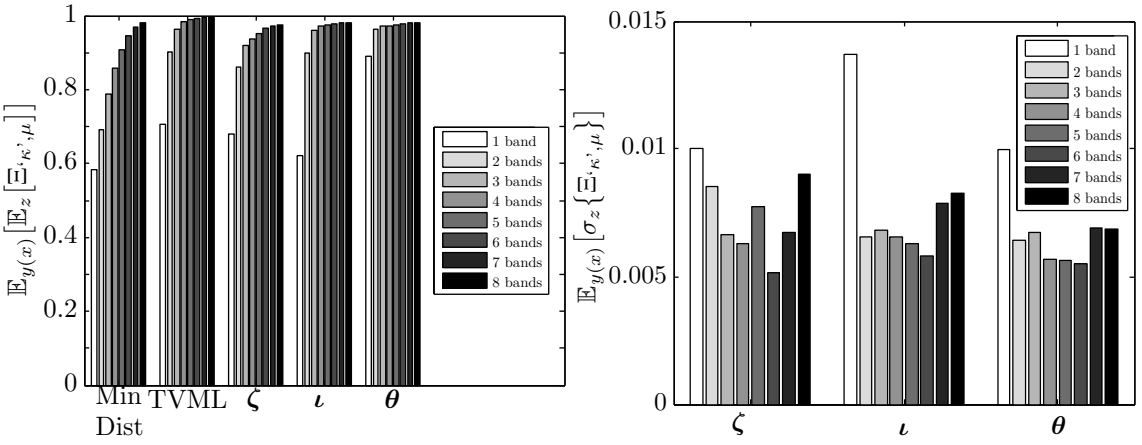
2. For all values of x, the time-varying maximum likelihood classifier on average achieves higher classification accuracies compared to the remaining classifiers (when the SVM using $\boldsymbol{\theta}$ is excluded).

3. The SVM classifier using the harmonic features classifies better than the SVM using $\boldsymbol{\zeta}$ and the minimum distance classifier except when $x = 1$, then $\boldsymbol{\zeta}$ produces better classification results than $\boldsymbol{\iota}$ or the minimum distance classifier.

4. The SVM classifier using the temporal features outperforms the minimum distance classifier when $x < 8$. When $x = 8$ the minimum distance classifier outperforms the SVM using $\boldsymbol{\zeta}$.

5. For all classifiers, increasing the spectral dimension increases classification accuracy.

6. The standard deviation for each SVM classifier (for all x) is small, implying that the classification results are stable and reliable.



(a) $\mathbb{E}_{y(x)}\left[\mathbb{E}_z\left[\Xi_{\text{`}\kappa\text{'},\mu}\right]\right]$ for the Gauteng dataset.

(b) $\mathbb{E}_{y(x)}\left[\sigma_z\left\{\Xi_{\text{`}\kappa\text{'},\mu}\right\}\right]$ for the Gauteng dataset.

Figure 5.39: The bar graph representing the $\mathbb{E}_{y(x)}\left[\mathbb{E}_z\left[\Xi_{\text{`}\kappa\text{'},\mu}\right]\right]$ values for the benchmarking and SVM classifiers [with $(\infty, 0.05)$ and $e = 50$] (Gauteng).

5.3.6 Classification results: Limpopo

To avoid repetition, only the most important classification results for the Limpopo data set will be mentioned. As mentioned in Section 5.3.3, the time-varying maximum likelihood classifier is an important facet of the thesis and the equivalent of Figure 5.36 is therefore presented in Figure 5.40 for the Limpopo dataset.

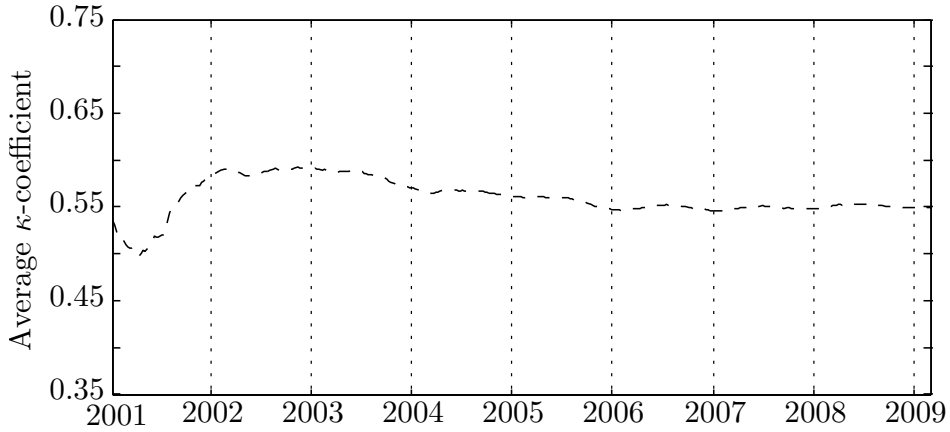
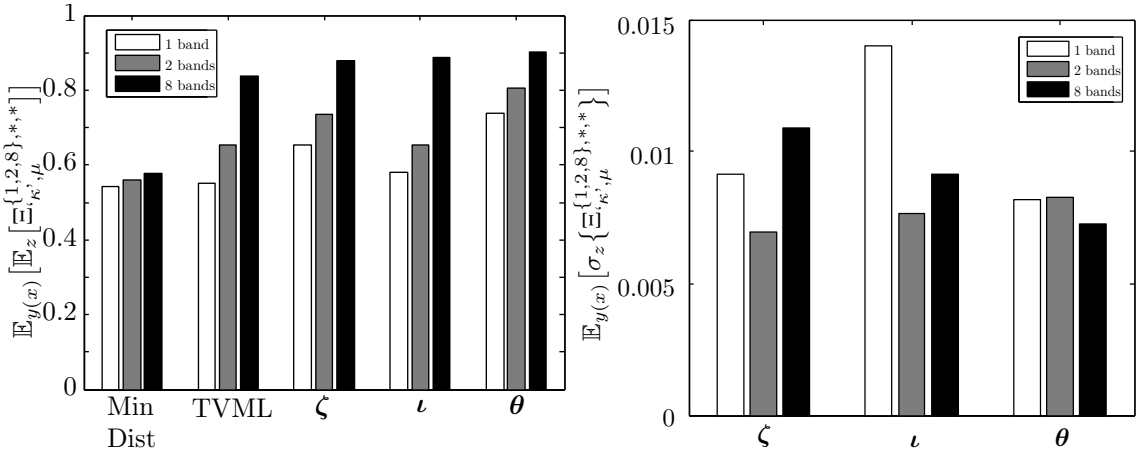


Figure 5.40: The average single-band classification performance of the time-varying maximum likelihood classifier for the Limpopo dataset as a function of time.

Generally the average κ-coefficient in Figure 5.40 increases during the first two years. A steep increase is observable in the first year and a smaller increase during the second year. After the second year however the average κ-coefficient starts to decrease steadily. The temporal dynamics over the first two years of the average κ-coefficient in Figure 5.40 strengthens the one-year (preferably two-year) sequential detection delay rule formed in Section 5.3.3, while inspecting Figure 5.36. The behaviour in Figure 5.36 and Figure 5.40 however differs from 2002 onwards. The decrease in κ observable in Figure 5.40 implies that the vegetation and settlement classes in the case of the Limpopo dataset become less separable over time. The opposite behaviour is seen in Figure 5.36, implying that the separability does not decrease in the case of the Gauteng dataset. The loss in separability in the case of the Limpopo dataset is not directly observable when inspecting Table 5.8, which implies that the separability loss happens gradually.

The equivalent of Figure 5.39 is presented in Figure 5.41 for the Limpopo dataset, as it sums up the performance of each classifier on the Limpopo dataset. As mentioned in Section 5.3.4, a grid search is required to obtain the SVM software parameters (c, λ). Random split cross-validation (50% for training and 50% for validation) was also employed by the grid search algorithm for the Limpopo dataset (10 independent experiments). The content details of Figure 5.41 are discussed in Section 5.3.5. Also recall from Section 5.3.5 that $x \in \{1, 2, 8\}$ indicates the band restriction value.



(a) $\mathbb{E}_{y(x)}\left[\mathbb{E}_z\left[\Xi_{\kappa,\mu}^{\{1,2,8\},*,*}\right]\right]$ for the Limpopo dataset.

(b) $\mathbb{E}_{y(x)}\left[\sigma_z\left\{\Xi_{\kappa,\mu}^{\{1,2,8\},*,*}\right\}\right]$ for the Limpopo dataset.

Figure 5.41: The bar graph representing the $\mathbb{E}_{y(x)}\left[\mathbb{E}_z\left[\Xi_{\kappa,\mu}\right]\right]$ values for the benchmarking and SVM classifiers [with $(\infty, 1)$ and $e = 10$] (Limpopo).

The following observations can be made from Figure 5.41:

1. The SVM classifier using the noise-harmonic features on average achieves higher classification accuracies than any of the other classifiers for all x.

2. The SVM classifier using the temporal features on average produces better classification results if compared to all the classification results produced by the remaining classifiers (if the SVM classifier using $\boldsymbol{\theta}$ is excluded) except when $x = 8$. When $x = 8$ the SVM classifier using $\boldsymbol{\iota}$ outperforms the SVM using $\boldsymbol{\zeta}$.

3. The time-varying maximum likelihood classifier and the SVM classifier using the harmonic features on average achieve better classification accuracies if compared with the classification results produced by the minimum distance classifier.

4. When $x = 1$ or $x = 8$ the SVM using the harmonic features outperforms the time-varying maximum likelihood classifier. When $x = 2$ exactly the opposite behaviour is observed.

5. For all classifiers, increasing the spectral dimension improves the accuracy of the classifier.

6. The standard deviation for each SVM classifier (for all x) is small, implying that the classification results are reliable and stable.

7. According to the classification results the Limpopo dataset is less separable than the Gauteng dataset.

5.3.7 Important classification conclusions

The following important conclusions can be made from the results presented in Section 5.3:

1. As the conclusions in Section 5.3.3 and Section 5.3.4 show, the metrics proposed in Section 5.1 can be used to choose efficient spectral bands and classification features.

2. As mentioned in Section 5.3.5 and Section 5.3.6, increasing the spectral dimension of a classifier improves its classification accuracy.

3. The feature group $\boldsymbol{\theta}$ is an extension of $\boldsymbol{\iota}$, not only in size but also in terms of its classification capability. This is clearly seen in Figure 5.11, Figure 5.14 and Figure 5.38. Moreover, Figure 5.39 and Figure 5.41 show that the SVM classifier that uses $\boldsymbol{\theta}$ outperforms all the other

classifiers investigated (except when $x > 2$, in the case of the Gauteng datset, which is when the time-varying maximum likelihood classifier starts achieving a slightly better classification accuracy). It is also significant to mention that the SVM classifier using $\boldsymbol{\theta}$ performs much better than the other classifiers when the spectral dimension of the classifier is low. This unique characteristic of $\boldsymbol{\theta}$ is significant, since it is obviously more advantageous to classify more accurately without having to increase the spectral view.

4. The Limpopo dataset is less separable than the Gauteng dataset, according to the classification results in Section 5.3.5 and Section 5.3.6. This is not surprising for the temporal or the CSHO feature classifiers. Even though the temporal Hellinger distance metrics in Table 5.3, are similar for the Gauteng and Limpopo datasets it is clear from Figure 5.40 that the separability of the Limpopo dataset deteriorates over time, explaining the weaker performance (in the case of the Limpopo dataset) of the minimum distance classifier, the time-varying maximum likelihood classifier and the SVM using $\boldsymbol{\zeta}$ in spite of good separability in the time-varying models of the Limpopo dataset. For $\boldsymbol{\iota}$ and $\boldsymbol{\theta}$ it is clear from Figure 5.11 and Figure 5.14 that the parameters of the CSHO provide better classification capability in the case of the Gauteng dataset than for the Limpopo dataset.

5. The most important sequential result found in [23] was reproduced, which is stated next. The temporal dynamics of the average κ-coefficient in Figure 5.36 and Figure 5.40 suggest that in general the thresholds of the time-varying maximum likelihood classifier should be chosen in such a way that the classifier on average experiences (excluding outliers) at least a one-year delay (preferably two) before it can classify an observed sequence.

5.4 CHANGE DETECTION RESULTS: GAUTENG AND LIMPOPO

This section focuses on the performance of the sequential change detection algorithm presented in Section 4.3.3, namely CUSUM. The CUSUM algorithm is benchmarked against the popular band differencing algorithm discussed in Section 4.3.2. This section starts by introducing the different change detection metrics employed, followed by the presentation of the change detection results of band differencing and CUSUM.

5.4.1 Change detection accuracy metrics

As change detection metrics P_{D}, P_{FA}, $\mathbb{E}\{(T-\tau)^{+368}\}$ and A_e are employed as metrics instead of $d_l(T)$ and $f(T)$, which were introduced in Section 3.6, since metrics that can be fairly compared to non-sequential change detection algorithms are required. Here P_{D} is the probability of correctly detecting a change within the eight-year observation period, P_{FA} is the probability of detecting a change when there was no change in the eight-year period, $\mathbb{E}\{(T-\tau)^{+368}\} = \mathbb{E}\{\min\{\max\{T-\tau,0\},368-\tau\}\}$ is the positive expected delay truncated to 368 observations and $A_e = \frac{1}{2}[(1-P_{\text{D}})+P_{\text{FA}}]$.

5.4.2 Results of Lunetta et al.'s scheme: Gauteng and Limpopo

The scheme developed by Lunetta et al. is discussed in detail in Section 4.3.2 and is also known as the band differencing algorithm [7]. The band differencing algorithm is a popular time-series change detection benchmarking method, which requires two parameters as input, the amount of frequency components ν to preserve and the decision threshold h_l^b [10,28]. A grid search was performed to find suitable values for ν and h_l^b. In this section the algorithm is applied to the Gauteng and Limpopo datasets. The values for ν, h_l^b and the change detection accuracies can be found in Table 5.10 and Table 5.11 for the Gauteng and Limpopo datasets, respectively. As the band differencing approach is used for benchmarking, the entire Gauteng and Limpopo datasets are used for training and validation. In a previous study [30], the band differencing algorithm was applied to the Gauteng and Limpopo datasets. In [30], the focus was on preserving the structure of the original signal and for that reason ν was set equal to 10. Figure 5.42 displays the effect of ν, which should make the reason for setting ν equal to 10 in the previous study clear. In this section, however the structure of the original signal is not one of the design criteria, and ν is thus found via a grid search algorithm. The number of frequency components ν to keep (as found via the grid search algorithm) is respectively set to two and three for the Gauteng and Limpopo datasets.

The following observations and conclusions can be made from Table 5.10 and Table 5.11:

1. In Table 5.10 and Table 5.11 the lower value of ν increases the accuracy of the band differencing change detector. The remaining observations assume that the lower value of ν is used.

2. For the Gauteng dataset it is clear from Table 5.10 that when the band differencing algorithm employs bands $\{1,3,4,\text{NDVI}\}$ the accuracy of the band differencing change detector is higher

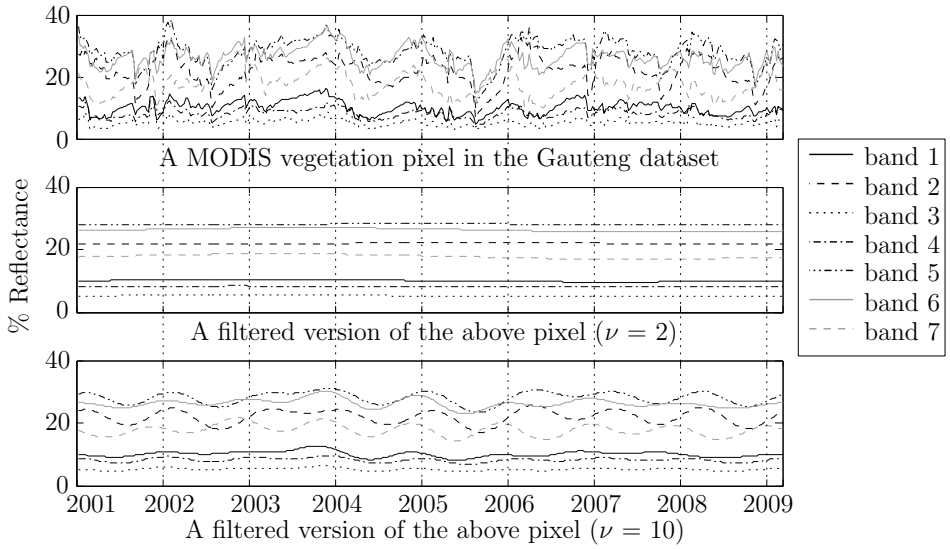


Figure 5.42: A true vegetation pixel belonging to the Gauteng dataset and two different filtered versions of the vegetation pixel.

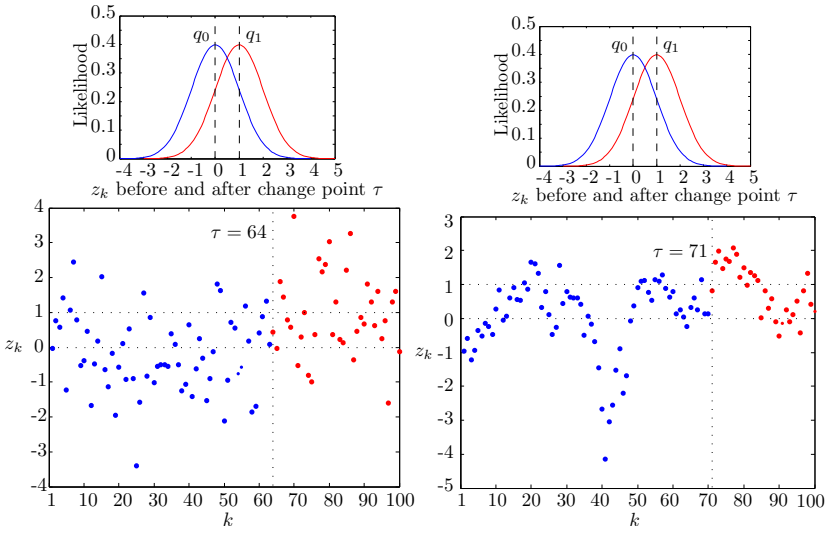
than if it uses bands $\{2,5,6,7\}$. Using band 1 leads to the lowest average error A_e, while using band 5 leads to the highest average error.

3. Exactly the same performance is observable for the Limpopo dataset as for the Gauteng dataset, except that for the Limpopo dataset, using band 4 leads to the lowest average error, while using band 2 leads to the highest average error.

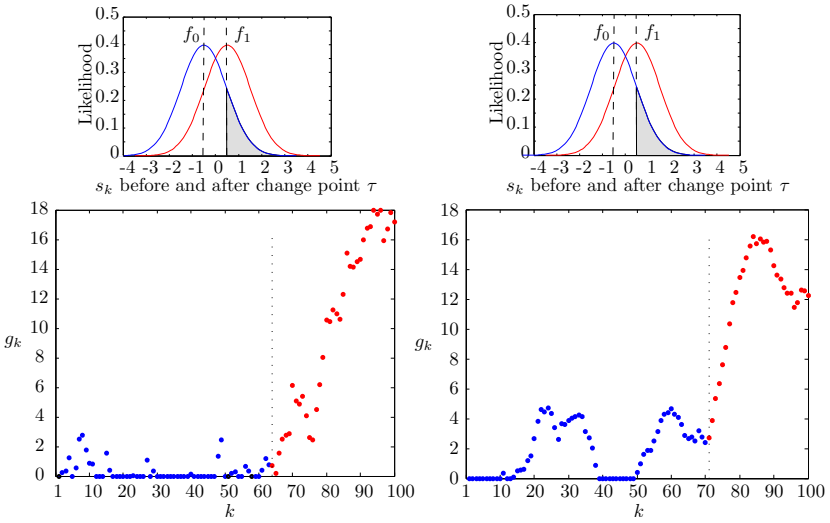
4. Generally the Gauteng dataset produces lower average errors when it is compared with the Limpopo dataset.

5.4.3 Temporal dependence and the CUSUM threshold

The influence of temporal dependence on the CUSUM threshold needs to be investigated before CUSUM can be applied to MODIS data. The influence of temporal dependence on the CUSUM threshold is explained in this section with the aid of an example. Assume therefore that there is a sequence z_k which is an i.i.d. sequence. The sequence is drawn independently from q_0 before change point τ and from τ onwards drawn from q_1. The density $q_0 \sim \mathcal{N}(0,1)$ and the density $q_1 \sim \mathcal{N}(1,1)$. An example of such a sequence can be found in Figure 5.43a with $\tau = 64$. The CUSUM statistic g_k generated from z_k with Equation 3.55 when $y = 0$ in Equation 3.55 is displayed in Figure 5.43c. Note that Equation 3.55 employs the log likelihood ratio s_k defined in Equation 3.56. While g_k stays



(a) z_k is independent. **(b)** z_k is not independent.



(c) g_k for independent case. **(d)** g_k for dependent case.

Figure 5.43: g_k for an independent and dependent example.

Table 5.10: Band differencing applied to the Gauteng dataset.

MODIS band	h_l^b	P_{D}	P_{FA}	A_e
$\frac{\nu=2}{\nu=10}$				
1	$\frac{1.2598}{1.8874}$	$\frac{0.7901}{0.6133}$	$\frac{0.0693}{0.2247}$	$\frac{0.1396}{0.3057}$
2	$\frac{1.3699}{0.7477}$	$\frac{0.6077}{0.9613}$	$\frac{0.1774}{0.8986}$	$\frac{0.2848}{0.4687}$
3	$\frac{1.1827}{1.5350}$	$\frac{0.7735}{0.7790}$	$\frac{0.1233}{0.3919}$	$\frac{0.1749}{0.3064}$
4	$\frac{1.2598}{1.5681}$	$\frac{0.7735}{0.7182}$	$\frac{0.0591}{0.3311}$	$\frac{0.1428}{0.3064}$
5	$\frac{1.4635}{0.7918}$	$\frac{0.5359}{0.9392}$	$\frac{0.2010}{0.8733}$	$\frac{0.3326}{0.4670}$
6	$\frac{1.5460}{2.7407}$	$\frac{0.5746}{0.2210}$	$\frac{0.1233}{0.0591}$	$\frac{0.2744}{0.4191}$
7	$\frac{1.2763}{1.7222}$	$\frac{0.6961}{0.6796}$	$\frac{0.1470}{0.2787}$	$\frac{0.2254}{0.2996}$
NDVI	$\frac{1.2047}{1.5185}$	$\frac{0.8011}{0.8122}$	$\frac{0.1622}{0.4645}$	$\frac{0.1805}{0.3262}$

below a threshold h no change is declared. After crossing the threshold h a change is declared. If z_k is associated with q_0 and q_1 then s_k will be associated with $f_0 \sim \mathcal{N}(-\frac{1}{2},1)$ and $f_1 \sim \mathcal{N}(\frac{1}{2},1)$ because

$$\begin{aligned} s_k &= \ln \frac{q_1(z_k)}{q_0(z_k)} \\ &= ln \frac{e^{-\frac{1}{2}(z_k-1)^2}}{e^{-\frac{1}{2}z_k^2}} \\ &= z_k - \frac{1}{2}. \end{aligned}$$

The densities f_0 and f_1 are displayed in Figure 5.43c.

The focus now shifts to the discretised Ornstein-Uhlenbeck process (see Section 4.1.2.1), which is a dependent sequence with generating equation

$$z_k = e^{-\lambda} z_{k-1} + (1-e^{-\lambda})\mu + \sigma \sqrt{\frac{1-e^{-2\lambda}}{2\lambda}} \eta_k,$$

where $\lambda > 0$ determines the degree of dependence (as well as the mean reversion rate), μ is the long-term mean, $\sigma > 0$ is the volatility of the random fluctuations and η_k is i.i.d. and has density $\mathcal{N}(0,1)$. Recall from Section 4.1.2.1 that the dependent sequence z_k is distributed according to density $\mathcal{N}\left(\mu, \frac{\sigma^2}{2\lambda}\right)$ (the equilibrium density) as long as z_0 is also distributed according to the equilibrium density. The *closer* λ *gets to zero* the *higher the temporal dependence* in the sequence z_k. To mi-

Table 5.11: Band differencing applied to the Limpopo dataset.

MODIS band	h_l^b	P_{D}	P_{FA}	A_e
$\frac{\nu=3}{\nu=10}$				
1	$\frac{2.2287}{2.4379}$	$\frac{0.6496}{0.4957}$	$\frac{0.0822}{0.0788}$	$\frac{0.2163}{0.2915}$
2	$\frac{2.4600}{1.7553}$	$\frac{0.3077}{0.4615}$	$\frac{0.0741}{0.3634}$	$\frac{0.3832}{0.4509}$
3	$\frac{1.8323}{2.2673}$	$\frac{0.7436}{0.5641}$	$\frac{0.1757}{0.0969}$	$\frac{0.2160}{0.2664}$
4	$\frac{2.2452}{2.0085}$	$\frac{0.6325}{0.6410}$	$\frac{0.0601}{0.1844}$	$\frac{0.2138}{0.2717}$
5	$\frac{1.9314}{2.1682}$	$\frac{0.5812}{0.4188}$	$\frac{0.2224}{0.1777}$	$\frac{0.3206}{0.3794}$
6	$\frac{1.9204}{2.4049}$	$\frac{0.6752}{0.3761}$	$\frac{0.2151}{0.0888}$	$\frac{0.2699}{0.3564}$
7	$\frac{1.7497}{1.6011}$	$\frac{0.7436}{0.8120}$	$\frac{0.2445}{0.4502}$	$\frac{0.2504}{0.3191}$
NDVI	$\frac{2.0305}{1.7828}$	$\frac{0.6923}{0.7778}$	$\frac{0.1383}{0.3180}$	$\frac{0.2230}{0.2701}$

mic the independent case, the variance of the equilibrium density should be equal to 1. By choosing $\sigma = \frac{1}{\sqrt{2}}$ and $\lambda = \frac{1}{4}$ a variance of 1 is obtained. Furthermore, assume that before change point τ, $\mu = 0$ and from τ onwards $\mu = 1$. The dependent sequence z_k is shown in Figure 5.43b ($\tau = 71$) and z_k's CUSUM sequence g_k is shown in Figure 5.43d. It is clear from Figure 5.43 that the change detection threshold h will usually be higher for the dependent case than the independent case (even though both cases are equally separable), since the *higher temporal dependence* found in the dependent case causes a *larger noise floor*. A larger noise floor is observable for the dependent case, as the probability of s_k to be positive increases due to the dependence (see the gray area in Figure 5.43d).

5.4.4 Results of the CUSUM test: Gauteng and Limpopo

In this section CUSUM is applied to the Gauteng and Limpopo datasets in order to detect when vegetation pixels in the study areas change into settlement pixels. The CUSUM algorithm is discussed in detail in Section 3.6 and Section 4.3.3. The CUSUM result section is divided into two parts or phases. In the first part of the section an off-line optimisation algorithm is used to determine the best threshold h by performing a sweep of h from 1 to 100 on simulated data to establish an intuitive base of the performance of CUSUM on MODIS data in the study areas. The simulated data that are used during this first phase are generated by the CSHO simulator discussed in Section 4.1.2.7 and Section 5.2.1. In the second part of the section the performance of the off-line determined h (using random split cross-validation) is evaluated on real world MODIS change data and is compared to the

band differencing method (on the same data). The following algorithm, with input vector (j,k,l,m,n), is proposed to determine the best off-line threshold h:

1. Use j pixels of the no-change vegetation data (real world no-change data) to learn the parameters needed by the simulator (training set).

2. Use k no-change settlement pixels (real world no-change data) to estimate the parameters needed by the simulator (training set).

3. Now using the trained simulator, simulate l pixels of each class, and use them to create the 45 probability density functions that span a year.

4. Simulate m pixels of each class, and use those to create simulated change data, where the change point τ has density U$[1,300]$. The change is simulated by using linear blending over a six-month period [10].

5. Simulate n pixels of no-change vegetation pixels.

6. For each threshold h perform the CUSUM algorithm on each band and determine P_{D}, P_{FA}, $\mathbb{E}\{(T-\tau)^{+_{368}}\}$ and A_e using the simulated change and no-change data.

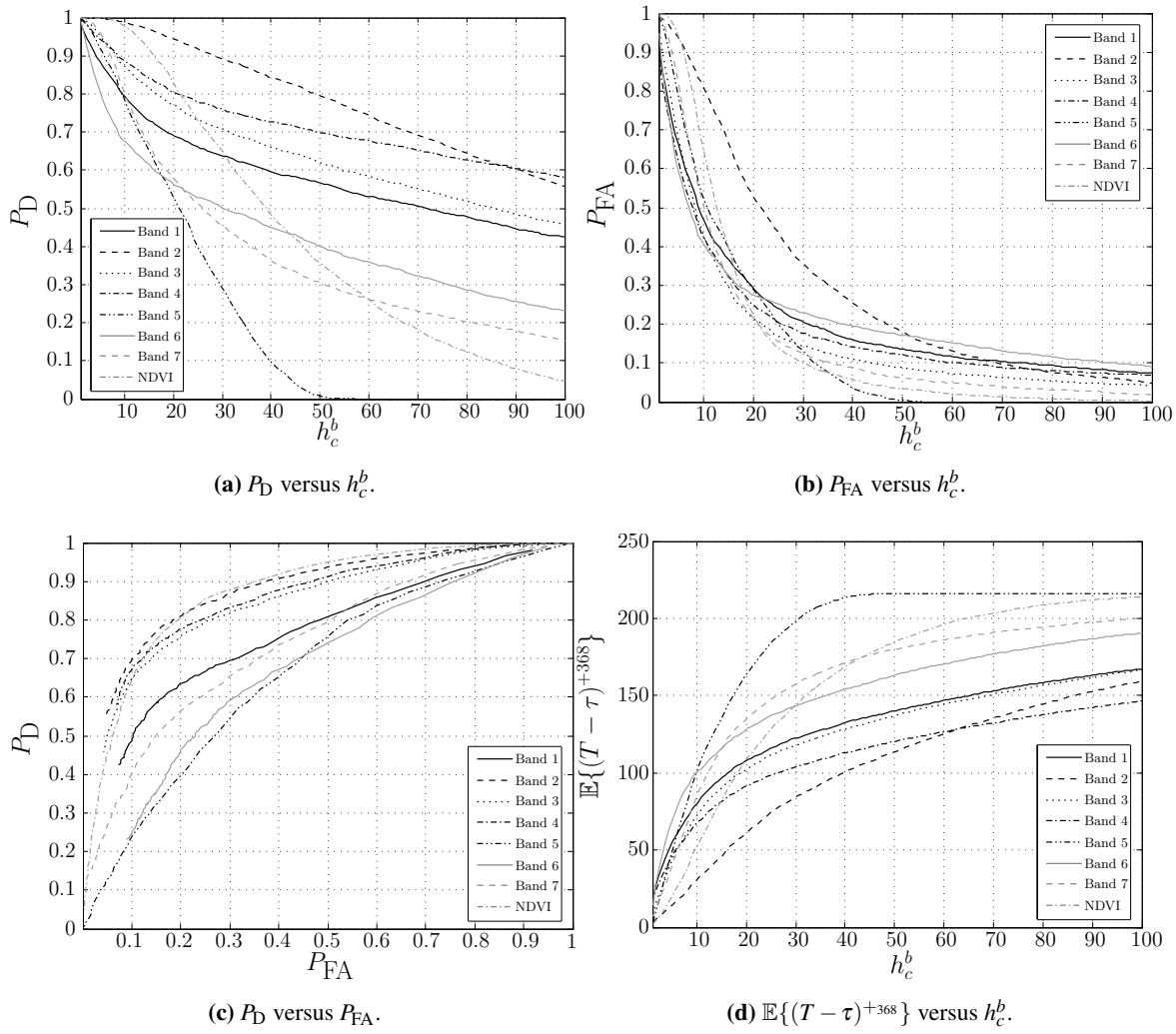
7. To determine the best h for each band, calculate the κ-coefficient Equation 5.7 (based on the number of correctly detected changes and the number of incorrectly detected changes) at each h in the sweeping interval and then select the h value that produces the largest κ-coefficient.

As an example of the output generated by the off-line optimisation algorithm, the resulting P_{D}, P_{FA} and $\mathbb{E}\{(T-\tau)^{+_{368}}\}$ metrics, determined for the Gauteng data set with input vector $(592,333,3000,3000,3000)$, are shown in Figure 5.44.

The resulting P_{D}, P_{FA} and $\mathbb{E}\{(T-\tau)^{+_{368}}\}$ generated by the off-line optimisation algorithm for the Limpopo dataset with input vector $(1497,1735,3000,3000,3000)$ is displayed in Figure 5.45.

The following observations can be made from Figure 5.44 and Figure 5.45:

1. The effect of h_c^b on the metrics P_{D}, P_{FA} and $\mathbb{E}\{(T-\tau)^{+_{368}}\}$ is different for each value of b and each dataset.



(a) P_D versus h_c^b.

(b) P_{FA} versus h_c^b.

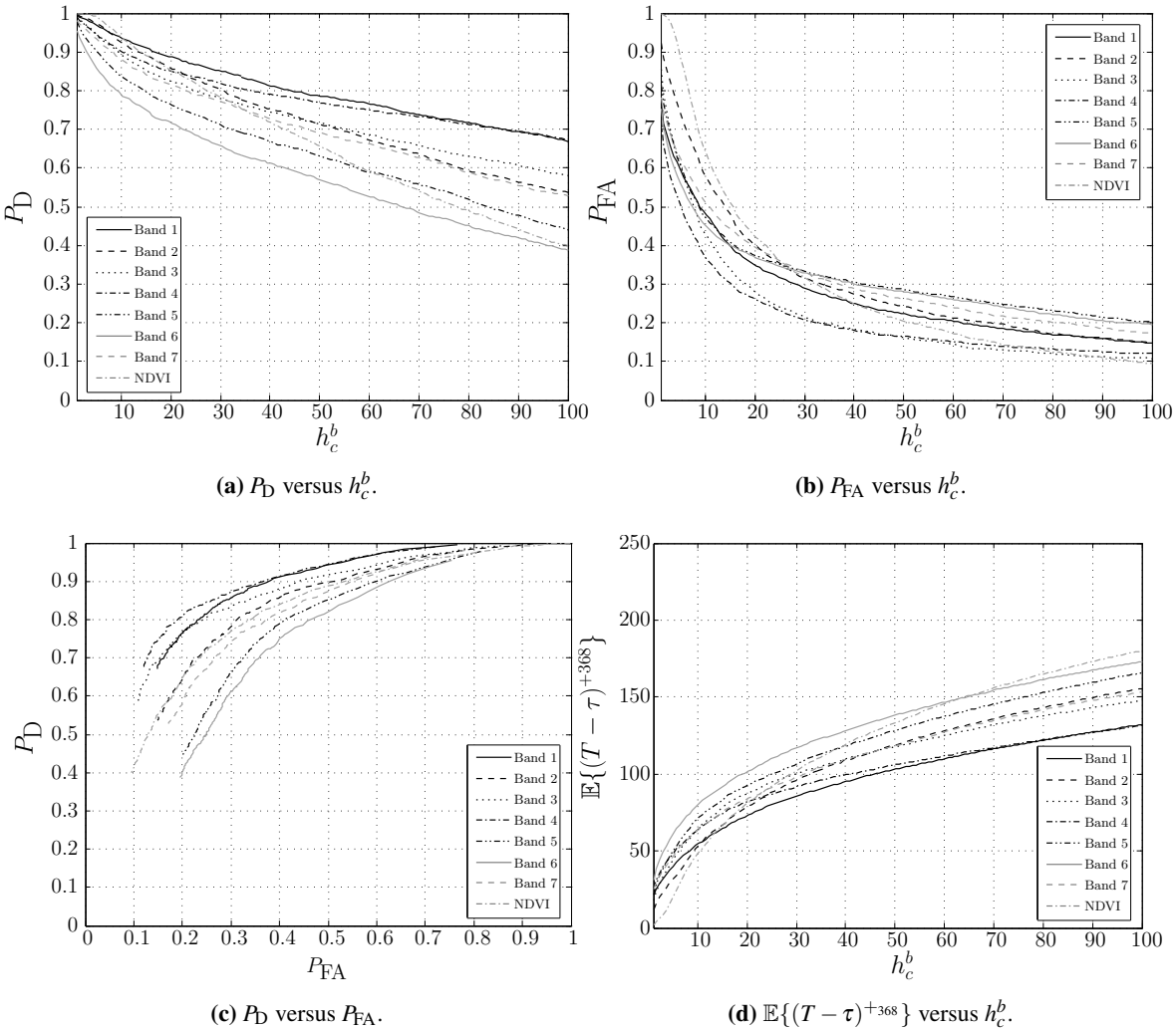
(c) P_D versus P_{FA}.

(d) $\mathbb{E}\{(T-\tau)^{+368}\}$ versus h_c^b.

Figure 5.44: Measured P_D, P_{FA} and $\mathbb{E}\{(T-\tau)^{+368}\}$ values for the simulated data in Gauteng [30] © IEEE 2012.

2. The most important graphs in Figure 5.44 and Figure 5.45 are the Receiver Operating Curves (ROCs) in Figure 5.44c and Figure 5.45c, since they display the probability of correctly detecting a change in the eight-year observation period against declaring a change during the observation period if none occurred. It is clear from Figure 5.44c that the CUSUM change detector produces higher change detection accuracies when using simulated data from bands $\{2,3,4,\text{NDVI}\}$ than if it uses simulated data from bands $\{1,5,6,7\}$. In Figure 5.45c it is clear that the CUSUM change detector achieves better change detection results if it uses simulated data from bands $\{1,2,3,4\}$ than if it uses simulated data from bands $\{5,6,7,\text{NDVI}\}$.

3. It is important to realise that the delay metric $\mathbb{E}\{(T-\tau)^{+368}\}$ presented in Figure 5.44d and



(a) P_D versus h_c^b.

(b) P_{FA} versus h_c^b.

(c) P_D versus P_{FA}.

(d) $\mathbb{E}\{(T-\tau)^{+368}\}$ versus h_c^b.

Figure 5.45: Measured P_D, P_{FA} and $\mathbb{E}\{(T-\tau)^{+368}\}$ values for the simulated data in Limpopo.

Figure 5.45d is an inefficient delay metric, as it is a truncated metric (and is easily affected by outliers). The metric is only reported here to be compatible with [30].

4. Better conclusions about the value of the threshold h_c^b, to optimise the delay of the CUSUM algorithm, can be made from the delay metric discussion in Section 5.3.6. The CUSUM algorithm and the time-varying maximum likelihood algorithm are based on the same underlying principles (see Chapter 3, Section 4.2.3 and Section 4.3.3). Since both approaches use the same underlying principles (CUSUM is merely a repeated time-varying SPRT–Section 4.3.3) both approaches are limited by the same expected decision delay imposed on them by the separability of the dataset to which they are applied. The one approach cannot suddenly be ten times faster than the other approach, as the underlying engines are based on the same design. As

seen in Figure 5.36 and Figure 5.40, the sequential threshold of the time-varying maximum likelihood classifier should be chosen in such a way that the classifier experiences a delay of at least a year (preferably two) before making a decision. A good design criterion for the CUSUM threshold h_c^b would therefore be to choose h_c^b in such a way that the expected decision delay (excluding outliers) of the CUSUM algorithm is at least one year (preferably two). Of course this delay is based on the fact that the change was instantaneous. The threshold should obviously be adapted to compensate for gradual change if required (meaning the expected delay of the detector should be increased).

To evaluate the performance of CUSUM on real world data, the metrics P_{D}, P_{FA} and A_e are used. No delay could be measured, as the true change points of the Gauteng and Limpopo datasets are unknown. The following methodology is proposed to determine the effectiveness of CUSUM on the Gauteng and Limpopo datasets:

1. Use the off-line optimisation algorithm with input vectors $(296, 333, 1000, 1000, 1000)$ and $(749, 1735, 1000, 1000,\ 1000)$ for the Gauteng and Limpopo datasets respectively, to determine the threshold h for each study region. Note that only 50% (random 50%) of the no change vegetation data was used to learn the parameters needed by the simulator and 50% of the real data was left for validation.

2. Apply the best h value to the no-change real vegetation data (validation dataset) and the real change data to determine P_{D} and P_{FA} (for each study region).

Random split cross-validation is performed by repeating the above experiment 50 times. The results of the random split cross validation experiments are displayed in Table 5.12. A training dataset and a validation dataset (equal in size), which were least correlated with each other (from all possible training and validation data sets) are also investigated. Since spatial independence is assumed by CUSUM, the worst case experiment is required to investigate whether spatial independence could be assumed (without detrimental effects) for the datasets under consideration. The results of the worst case experiment are found in Table 5.13.

The following observations and conclusion can be made from Table 5.12 and Table 5.13:

1. In the case of the real change Gauteng dataset the CUSUM algorithm achieves better change detection accuracies when applied to bands $\{1, 2, 3, 4\}$ than when the CUSUM algorithm is

Table 5.12: Random split cross-validation of CUSUM (50 experiments, 50% for training and 50% for validation) applied to the Gauteng and Limpopo datasets.

MODIS band	h_c^b	$\sigma_{h_c^b}$	P_{D}	$\sigma_{P_{\mathrm{D}}}$	P_{FA}	$\sigma_{P_{\mathrm{FA}}}$	A_e
Gauteng **Limpopo**							
1	37.33 49.77	6.34 8.03	0.9835 0.7695	0.0047 0.0167	0.1932 0.2483	0.0381 0.0281	0.1048 0.2394
2	55.88 39.19	5.87 6.46	0.8718 0.6736	0.0030 0.0144	0.1225 0.3070	0.0196 0.0281	0.1254 0.3167
3	28.67 36.90	5.55 8.42	0.9846 0.6638	0.0023 0.0376	0.1737 0.1920	0.0348 0.0307	0.0946 0.2641
4	38.02 38.53	7.29 10.77	0.9835 0.6866	0.0008 0.0308	0.1593 0.1930	0.0271 0.0311	0.0879 0.2532
5	16.30 22.78	2.78 5.90	0.7884 0.6860	0.0652 0.0360	0.5363 0.4517	0.0552 0.0352	0.3740 0.3829
6	18.05 19.06	5.21 5.30	0.2136 0.6986	0.0809 0.0416	0.3471 0.4700	0.0558 0.0456	0.5668 0.3857
7	19.27 40.72	2.88 7.45	0.7114 0.7236	0.0730 0.0281	0.2495 0.3180	0.0853 0.0408	0.2690 0.2972
NDVI	21.21 35.53	1.65 5.17	0.8170 0.7499	0.0502 0.0488	0.1484 0.4316	0.0561 0.0519	0.1657 0.3409

applied to bands $\{5,6,7,\mathrm{NDVI}\}$. Exactly the same behaviour is observed in the case of the real change Limpopo dataset. The CUSUM change detection algorithm performs much better on the Gauteng dataset than on the Limpopo dataset. The top performing band sets obtained for the off-line optimization algorithm (discussed earlier in this section) are practically the same as the top performing bands sets obtained when CUSUM is applied to the real change datasets. The discrepancy can be explained by the fact that the simulated change dataset is certainly not a carbon copy of the real change dataset and as such a guarantee can therefore not be given that CUSUM will produce exactly same results when applied to the simulated change dataset and the real change dataset.

2. Recall from Section 5.1.2.3 that for the Gauteng dataset the temporal Hellinger distance metric predicted that a change detector using data from bands $\{1,2,3,4\}$ would provide better change detection accuracies than a change detector that uses data from bands $\{5,6,7,\mathrm{NDVI}\}$ when the change detector in question relies on temporal features. From Table 5.12 it is clear that this prediction is accurate. This is no surprise, as a similar result is found in Section 5.3.3 and Section 5.3.4. A similar conclusion can be made when inspecting the real change Limpopo dataset.

3. As mentioned in Section 5.4.3, the CUSUM threshold is affected by the amount of temporal

Table 5.13: CUSUM applied to the worst possible correlated training set (for the Gauteng and Limpopo datasets).

MODIS band	h_c^b	P_{D}	P_{FA}	A_e
Gauteng / Limpopo				
1	23.59 / 60.80	0.9890 / 0.8034	0.2500 / 0.3178	0.1305 / 0.2572
2	65.45 / 44.19	0.8729 / 0.6752	0.1554 / 0.3284	0.1412 / 0.3266
3	29.57 / 61.46	0.9834 / 0.6410	0.1250 / 0.2510	0.0708 / 0.3050
4	39.54 / 32.23	0.9834 / 0.7778	0.1216 / 0.3418	0.0691 / 0.2820
5	11.63 / 26.91	0.9558 / 0.5983	0.7770 / 0.3578	0.4106 / 0.3798
6	12.96 / 34.22	0.2099 / 0.5299	0.3243 / 0.3845	0.5572 / 0.4273
7	14.29 / 43.52	0.8287 / 0.7009	0.5845 / 0.3204	0.3779 / 0.3098
NDVI	22.26 / 37.54	0.8122 / 0.8120	0.3311 / 0.5901	0.2595 / 0.3891

dependence in the data. The higher the dependence, the higher the threshold should be. This threshold phenomenon is observable in Table 5.12 when inspecting the top performing bands $\{1,2,3,4\}$ of the real change Gauteng dataset. The bands in $\{1,2,3,4\}$ that have a higher amount of dependence between their observations (which can be estimated from Table 5.6) have higher thresholds than the bands that have a lower amount of dependence between their observations. A similar conclusion can be drawn from the real change Limpopo dataset. It is however important to realise that the λ value is only an estimate of the amount of dependence between the observations of a band and can therefore not always be trusted to predict the CUSUM threshold as is seen in the case of the NDVI threshold for the real change Gauteng dataset.

4. As the results in Table 5.12 and Table 5.13 are similar, it shows that a spatial independent assumption is an allowable assumption for the current datasets.

5.4.5 Important change detection conclusions

The following important conclusions can be made from the results presented in Section 5.4:

1. From Table 5.10 and Table 5.11 it is clear that the band differencing algorithm performs better

if structure preservation is ignored when choosing the value of ν.

2. From Table 5.12 and Table 5.13 it is clear that the sequential change detection algorithm CUSUM can be effectively applied to MODIS. According to Figure 5.36 and Figure 5.40, a good design criterion for the CUSUM threshold h_c^b would be to choose h_c^b in such a way that the expected decision delay (excluding outliers) of the CUSUM algorithm is at least one year (preferably two).

3. According to Table 5.10 and Table 5.12 the CUSUM algorithm outperforms band differencing in the case of the Gauteng dataset. The exact opposite happens in the case of the Limpopo dataset. This is not surprising, since Figure 5.40 shows that the separability of the Limpopo dataset deteriorates over time. Since the CUSUM approach is a sequential algorithm, it is severely affected by this deterioration (more than band differencing that relies on detecting pixel outliers).

5.5 CONCLUSION

The chapter presented the classification and change detection accuracies and rankings of the different sequential and non-sequential hypertemporal classification and change detection algorithms investigated in this thesis. The most important conclusions from this chapter are summarised in Section 6.1.

CHAPTER 6

CONCLUSION

The results and conclusion of Chapter 5 are summarized in this chapter.

6.1 MAIN CONCLUSIONS

The following table highlights the most important sections contained in the thesis:

Table 6.1: The most important sections of the thesis.

Algorithm	Description	Results	Main conclusions
Simulation	Section 4.1.2	Section 5.2	Section 5.2.6
Minimum distance classifier	Section 4.2.2	Section 5.3	Section 5.3.7
Time-varying maximum likelihood classifier	Section 4.2.3		
$\boldsymbol{\theta}, \boldsymbol{\iota}$ and $\boldsymbol{\zeta}$	Section 4.2.4.2		
Band differencing	Section 4.3.2	Section 5.4.2	Section 5.4.5
CUSUM	Section 4.3.3	Section 5.4.4	

The sections listed in the right most column of Table 6.1 contain the most important conclusions made in the thesis.

6.2 SUMMARY OF WORK

In this thesis new hypertemporal techniques were proposed for the settlement detection problem in South Africa. The hypertemporal techniques were applied to study areas in the Gauteng and Limpopo

provinces of South Africa. To be more precise, new sequential (windowless) and non-sequential hypertemporal techniques were investigated. The time-series employed by the new hypertemporal techniques were obtained from the MODIS sensor, which is on board the earth observation satellites Aqua and Terra. One MODIS dataset was constructed for each province.

An SVM that uses a novel noise-harmonic feature set was implemented to detect existing human settlements. The noise-harmonic feature set is a non-sequential hypertemporal feature set and was constructed by using the CSHO. The CSHO consists of an SHO which is superimposed on the Ornstein-Uhlenbeck process. The noise-harmonic feature set is an extension of the classic harmonic feature set. The classic harmonic feature set consists of a mean and a seasonal component. For the case studies in this thesis, it is observed that the noise-harmonic feature set not only extends the harmonic feature set, but also improves on its classification capability.

The noise-harmonic feature SVM was also compared with the minimum distance classifier, the time-varying classifier (which is based on sequential analysis) and a temporal feature SVM. In general the noise-harmonic feature SVM outperformed the minimum distance classifier, the time-varying classifier and the temporal feature SVM. It is also worth mentioning that the noise-harmonic feature SVM performs much better than the other classifiers when the spectral dimension of the classifiers are low.

The CUSUM algorithm was developed by E.S. Page in 1954. In its original form it is a sequential (windowless) hypertemporal change detection technique. Windowed versions of the algorithm have been applied in a remote sensing context. In this thesis CUSUM was used in its original form to detect settlement expansion in South Africa and is benchmarked against the classic band differencing change detection approach of R.S. Lunetta et al. In the case of the Gauteng study area, the CUSUM algorithm outperformed the band differencing technique.

Sequential hypertemporal techniques are data-intensive and an inductive MODIS simulator was consequently also developed (to augment datasets). The proposed simulator is also based on the CSHO. Two case studies showed that the proposed inductive simulator accurately replicates the temporal dynamics and spectral dependencies found in MODIS data.

The main result of this thesis is that the noise-harmonic feature set and the CUSUM algorithm are promising hypertemporal techniques that may achieve good results (competitive) when applied in a remote sensing context. The main academic contribution of this thesis however was the successful

application of sequential hypertemporal techniques in the remote sensing field.

6.3 FUTURE WORK

The following is a list of possible work that can be done in the future to extend the results in this thesis:

1. The Shiryaev-Roberts stopping time, which is a sequential change detection algorithm, is defined in Equation 3.65. It would be interesting to compare the performance of the Shiryaev-Roberts stopping time when it is applied to the datasets in Section 2.8 with the performance results of the CUSUM (Section 4.3.3) algorithm presented in Section 5.4.4.

2. Figure 5.43 seems to indicate that if there is dependence between observations then the CUSUM statistic derived from the dependent observations exhibit sinusoidal behaviour. It would be worth investigating whether applying different filtering techniques to the CUSUM statistic could improve the performance of the CUSUM algorithm whenever it is applied to a dataset with dependent observations.

3. The AR(p) model is defined as

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t,$$

where $\{\varphi_1, \ldots, \varphi_p\}$ are the parameters of the model, c is a constant, and ε_t is white noise. Since the Ornstein-Uhlenbeck proces is actually the continues-time analogue of the AR(1) process, it would make sense to also try and model the residual $\eta_c^b(t)$ with a higher order AR process. It would be interesting to determine whether the set $\{\varphi_1, \ldots, \varphi_p\}$ can provide good class discernibility.

REFERENCES

[1] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[2] T. Grobler, E. R. Ackermann, J. C. Olivier, A. J. van Zyl, and W. Kleynhans, "Land-cover separability analysis of MODIS time-series data using a combined Simple Harmonic Oscillator and a mean reverting stochastic process," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 3, pp. 857–866, 2012.

[3] J. Hermance, R. Jacob, B. Bradley, and J. Mustard, "Extracting phenological signals from multiyear AVHRR NDVI time series: Framework for applying high-order annual splines with roughness damping," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 10, pp. 3264–3276, 2007.

[4] G. Uhlenbeck and L. Ornstein, "On the theory of the Brownian motion," *Physical Review*, vol. 36, no. 5, p. 823, 1930.

[5] S. Lhermitte, J. Verbesselt, I. Jonckheere, K. Nackaerts, J. van Aardt, W. Verstraeten, and P. Coppin, "Hierarchical image segmentation based on similarity of NDVI time series," *Remote Sensing of Environment*, vol. 112, no. 2, pp. 506–521, 2008.

[6] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 241–257, 1954.

[7] R. Lunetta, J. Knight, J. Ediriwickrema, J. Lyon, and L. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sensing of Environment*, vol. 105, no. 2, pp. 142–154, 2006.

[8] J. Foley, R. DeFries, G. Asner, C. Barford, G. Bonan, S. Carpenter, F. Chapin, M. Coe, G. Daily, H. Gibbs *et al.*, "Global consequences of land use," *Science*, vol. 309, no. 5734, pp. 570–574, 2005.

[9] G. Benítez, A. Pérez-Vázquez, M. Nava-Tablada, M. Equihua, and J. Álvarez-Palacios, "Urban expansion and the environmental effects of informal settlements on the outskirts of Xalapa city, Veracruz, Mexico," *Environment and Urbanization*, vol. 24, no. 1, pp. 149–166, 2012.

[10] W. Kleynhans, "Detecting land-cover change using MODIS time-series data," Ph.D. dissertation, University of Pretoria, 2011.

[11] "United Nations division for sustainable development – national information – South Africa," 2010. [Online]. Available: `http://www.un.org/esa/agenda21/natlinfo/countr/safrica/human_settlements`

[12] B. Salmon, "Improved hyper-temporal feature extraction methods for land cover change detection in satellite time series," Ph.D. dissertation, University of Pretoria, 2012.

[13] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.

[14] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance," *International Journal of Remote Sensing*, vol. 28, no. 5, pp. 823–870, 2007.

[15] H. Carrão, P. Gonçalves, and M. Caetano, "Contribution of multispectral and multitemporal information from MODIS images to land cover classification," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 986–997, 2008.

[16] E. Ackermann, T. Grobler, A. van Zyl, K. Steenkamp, and J. Olivier, "Minimum error land cover separability analysis and classification of MODIS time series data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*. Vancouver, Canada: IEEE, 2011, pp. 2999–3002.

[17] R. Lasaponara, "Estimating interannual variations in vegetated areas of Sardinia island using

SPOT/VEGETATION NDVI temporal series," *Geoscience and Remote Sensing Letters, IEEE*, vol. 3, no. 4, pp. 481–483, 2006.

[18] D. Alcaraz, J. Paruelo, and J. Cabello, "Identification of current ecosystem functional types in the Iberian Peninsula," *Global Ecology and Biogeography*, vol. 15, no. 2, pp. 200–212, 2006.

[19] J. Verbesselt, R. Hyndman, G. Newnham, and D. Culvenor, "Detecting trend and seasonal changes in satellite image time series," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 106–115, 2010.

[20] J. Verbesselt, R. Hyndman, A. Zeileis, and D. Culvenor, "Phenological change detection while accounting for abrupt and gradual trends in satellite image time series," *Remote Sensing of Environment*, vol. 114, no. 12, pp. 2970–2980, 2010.

[21] B. Salmon, J. Olivier, W. Kleynhans, K. Wessels, F. van den Bergh, and K. Steenkamp, "The use of a multilayer perceptron for detecting new human settlements from a time series of MODIS images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 6, pp. 873–883, 2011.

[22] C. Schaaf, F. Gao, A. Strahler, W. Lucht, X. Li, T. Tsang, N. Strugnell, X. Zhang, Y. Jin, J. Muller *et al.*, "First operational BRDF, albedo nadir reflectance products from MODIS," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 135–148, 2002.

[23] E. Ackermann, "Sequential land cover classification," Master's thesis, University of Pretoria, 2011.

[24] B. Ghosh and P. Sen, Eds., *Handbook of Sequential Analysis*. New York: Dekker, 1991, vol. 118.

[25] T. Lai, "Sequential analysis: Some classical problems and new challenges," *Statistica Sinica*, vol. 11, pp. 303–408, 2001.

[26] L. Guanter, K. Segl, and H. Kaufmann, "Simulation of optical remote-sensing scenes with application to the ENMAP hyperspectral mission," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 47, no. 7, pp. 2340–2351, 2009.

[27] W. Verhoef and H. Bach, "Coupled soil-leaf-canopy and atmosphere radiative transfer modeling to simulate hyperspectral multi-angular surface reflectance and TOA radiance data," *Remote Sensing of Environment*, vol. 109, no. 2, pp. 166–182, 2007.

[28] S. Boriah, "Time series change detection: Algorithms for land cover change," Ph.D. dissertation, The University of Minnesota, 2010.

[29] J. Verbesselt, A. Zeileis, and M. Herold, "Near real-time disturbance detection in terrestrial ecosystems using satellite image time series: drought detection in Somalia," Faculty of Economics and Statistics, University of Innsbruck, Working Papers, 2011. [Online]. Available: `http://EconPapers.repec.org/RePEc:inn:wpaper:2011-18`

[30] T. Grobler, E. Ackermann, A. van Zyl, J. Olivier, W. Kleynhans, and B. Salmon, "Using Page's Cumulative Sum test on MODIS time series to detect land-cover changes," *Geoscience and Remote Sensing Letters, IEEE*, vol. 10, no. 2, pp. 332–336, March 2013, DOI: 10.1109/LGRS.2012.2205556.

[31] B. Jiang, S. Liang, J. Wang, and Z. Xiao, "Modeling MODIS LAI time series using three statistical methods," *Remote Sensing of Environment*, vol. 114, no. 7, pp. 1432–1444, 2010.

[32] T. Grobler, E. Ackermann, A. van Zyl, J. Olivier, W. Kleynhans, and B. Salmon, "An inductive approach to simulating multispectral MODIS surface reflectance time series," *Geoscience and Remote Sensing Letters, IEEE*, vol. 10, no. 3, pp. 446–450, May 2013, DOI: 10.1109/LGRS.2012.2208446.

[33] T. Lillesand and R. Kiefier, *Remote Sensing and Image Interpretation*, 3rd ed. John Wiley & Sons, Inc., 1994.

[34] J. Cambell, *Introduction to Remote Sensing*, 4th ed. The Guilford Press, 2008.

[35] P. Gibson, *Introductory Remote Sensing, Principles and Concepts*, 1st ed. Routledge, 2000.

[36] E. Liebenberg and C. Vlok, "The Interpretation of maps, aerial photographs and satellite images," Study guide2 for GGH203–V, University of South Africa.

[37] B. Ramachandran, C. Justice, and A. Abrams, Eds., *Land Remote Sensing and Global Envi-*

ronmental Change, NASA's Earth Observing System and the Science of ASTER and MODIS. Springer, 2011.

[38] J. Qu, W. Gao, M. Kafatos, E. Robert, and V. V. Salomonson, Eds., *Earth Science Satellite Remote Sensing, Science and Instruments, Volume 1*. Springer, 2006.

[39] S. Davis, D. Landgrebe, T. Phillips, P. Swain, R. Hoffer, J. Lindenlaub, and L. Silva, *Remote Sensing, the Quantitative Aproach*, 1st ed. McGraw-Hill, 1978.

[40] M. Hazewinkel, Ed., *Encyclopedia of Mathematics*. Springer, 2001.

[41] T. García-Mora, J. Mas, and E. Hinkley, "Land cover mapping applications with MODIS: a literature review," *International Journal of Digital Earth*, vol. 5, no. 1, pp. 63–87, 2012.

[42] Q. Weng, Ed., *Advances in Environmental Remote Sensing, Sensors, Algorithms, and Applications*. CRC Press, Taylor & Francis Group, 2011.

[43] R. Colditz, "Time series generation and classification of MODIS data for land cover mapping," Ph.D. dissertation, Julius-Maximilians-Universität Würzburg, September 2007.

[44] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933.

[45] A. Wald, "Sequential tests of hypotheses," *Annals of Mathematical Statistics*, vol. 16, pp. 117–186, 1945.

[46] ——, *Sequential Analysis*. John Wiley & Sons, Inc., 1947.

[47] A. Wald and J. Wolfowitz, "Optimum character of the sequential probability ratio test," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 326–339, 1948.

[48] H. Poor and O. Hadjiliadis, *Quickest Detection*. Cambridge University Press, 2009.

[49] W. Shewhart, *Economic Control of Quality of Manufactured Product*. Princeton, NJ: D. Van Nostrand Reinhold, 1931.

[50] M. Girshick and H. Rubin, "A Bayes approach to a quality control model," *The Annals of Mathematical Statistics*, vol. 23, no. 1, pp. 114–125, 1952.

[51] A. Shiryaev, "The problem of the most rapid detection of a disturbance in a stationary process," in *Soviet Mathematics Doklady*, vol. 2, no. 795-799, 1961.

[52] G. Lorden, "Procedures for reacting to a change in distribution," *Annals of Mathematical Statistics*, vol. 42, pp. 1897–1908, 1971.

[53] G. Moustakides, "Optimal stopping times for detecting changes in distributions," *The Annals of Statistics*, vol. 14, no. 4, pp. 1379–1387, 1986.

[54] Y. Ritov, "Decision theoretic optimality of the CUSUM procedure," *The Annals of Statistics*, pp. 1464–1469, 1990.

[55] S. Roberts, "A comparison of some control chart procedures," *Technometrics*, pp. 411–430, 1966.

[56] M. Pollak, "Optimal detection of a change in distribution," *The Annals of Statistics*, pp. 206–227, 1985.

[57] A. Polunchenko, "Quickest change detection with applications to distributed multi-sensor systems," Ph.D. dissertation, University of Southern California, 2009.

[58] G. Moustakides, A. Polunchenko, and A. Tartakovsky, "A numerical approach to performance analysis of quickest change-point detection procedures," *Statistica Sinica*, vol. 21, pp. 571–596, 2011.

[59] M. Basseville, I. Nikiforov *et al.*, *Detection of Abrupt Changes: Theory and Application*. Prentice Hall Englewood Cliffs, 1993, vol. 15.

[60] F. James, *Statistical Methods in Experimental Physics*, 2nd ed. World Scientific Pub Co Inc, 2006.

[61] S. Kay, *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, 1998.

[62] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd ed. Cambridge University Press, 1992.

[63] A. Goel and S. Wu, "Determination of the ARL and a contour nomogram for CUSUM charts to control normal mean," *Technometrics*, vol. 13, no. 2, pp. 221–230, 1971.

[64] D. Brook and D. Evans, "An approach to the probability distribution of CUSUM run length," *Biometrika*, vol. 59, no. 3, pp. 539–549, 1972.

[65] L. Kantorovich and V. Krylov, *Approximate Methods of Higher Analysis*. Interscience Groningen, 1964.

[66] A. Shiryaev, *Optimal Stopping Rules*. New-York: Springer, 1978.

[67] ——, "On optimum methods in quickest detection problems," *Theory of Probability and its Applications*, vol. 8, no. 1, pp. 22–46, 1963.

[68] J. Schott, S. Brown, R. Raqueno, H. Gross, and G. Robinson, "An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development," *Canadian Journal of Remote Sensing*, vol. 25, no. 2, pp. 99–111, 1999.

[69] A. Singh, "Digital change detection techniques using remotely-sensed data," *International Journal of Remote Sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[70] D. Mouat, G. Mahin, and J. Lancaster, "Remote sensing techniques in the analysis of change detection," *Geocarto International*, vol. 8, no. 2, pp. 39–50, 1993.

[71] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Digital change detection methods in ecosystem monitoring: A review," *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1565–1596, 2004.

[72] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: A systematic survey," *Image Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 294–307, 2005.

[73] G. Jianya, S. Haigang, M. Guorui, and Z. Qiming, "A review of multi-temporal remote sensing data change detection algorithms," *The International Archives of the Photogrammetry, Remote*

Sensing and Spatial Information Sciences, vol. 37, pp. 757–762, 2008.

[74] H. Carrão, P. Gonalves, and M. Caetano, "A nonlinear harmonic model for fitting satellite image time series: Analysis and prediction of land cover dynamics," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 4, pp. 1919–1930, 2010.

[75] S. Jacquemoud, W. Verhoef, F. Baret, C. Bacour, P. Zarco-Tejada, G. Asner, C. François, and S. Ustin, "PROSPECT+ SAIL models: A review of use for vegetation characterization," *Remote Sensing of Environment*, vol. 113, pp. S56–S66, 2009.

[76] P. Jönsson and L. Eklundh, "Seasonality extraction by function fitting to time-series of satellite sensor data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 40, no. 8, 2002.

[77] X. Zhang, M. Friedl, C. Schaaf, A. Strahler, J. Hodges, F. Gao, B. Reed, and A. Huete, "Monitoring vegetation phenology using MODIS," *Remote Sensing of Environment*, vol. 84, no. 3, pp. 471–475, 2003.

[78] W. Kleynhans, J. Olivier, K. Wessels, F. Van den Bergh, B. Salmon, and K. Steenkamp, "Improving land cover class separation using an extended Kalman filter on MODIS NDVI time-series data," *Geoscience and Remote Sensing Letters, IEEE*, vol. 7, no. 2, pp. 381–385, 2010.

[79] K. Levenberg, "A method for the solution of certain problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[80] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[81] E. Bibbona, G. Panfilo, and P. Tavella, "The Ornstein–Uhlenbeck process as a model of a low pass filtered white noise," *Metrologia*, vol. 45, p. S117, 2008.

[82] Z. Brzeźniak and T. Zastawniak, *Basic Stochastic Processes: A Course Through Exercises*. Springer Verlag, 1999.

[83] T. van den Berg, "Calibrating the Ornstein-Uhlenbeck model," May 2011, `http://www.sitmo.com/article/calibrating-the-ornstein-uhlenbeck-model/`.

[84] T. Björk, *Arbitrage Theory in Continuous Time*. Oxford University Press, USA, 2009.

[85] J. Richards and X. Jia, *Remote Sensing Digital Image Analysis: An Introduction*. Berlin, Germany: Springer Verlag, 2006.

[86] J. Paola and R. Schowengerdt, "A review and analysis of backpropagation neural networks for classification of remotely-sensed multi-spectral imagery," *International Journal of Remote Sensing*, vol. 16, no. 16, pp. 3033–3058, 1995.

[87] P. Swain and H. Hauska, "The decision tree classifier design and potential," *IEEE Transactions on Geoscience Electronics*, vol. 15, no. 3, pp. 142–147, 1977.

[88] R. De Fries, M. Hansen, J. Townshend, and R. Sohlberg, "Global land cover classifications at 8 km spatial resolution: The use of training data derived from Landsat imagery in decision tree classifiers," *International Journal of Remote Sensing*, vol. 19, no. 16, pp. 3141–3168, 1998.

[89] C. Huang, L. Davis, and J. Townshend, "An assessment of Support Vector Machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[90] G. Foody and A. Mathur, "A relative evaluation of multiclass image classification by Support Vector Machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 6, pp. 1335–1343, 2004.

[91] G. Mountrakis, J. Im, and C. Ogole, "Support Vector Machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[92] Y. Sohn and N. Rebello, "Supervised and unsupervised spectral angle classifiers," *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 12, pp. 1271–1282, 2002.

[93] T. Lee, M. Lewicki, and T. Sejnowski, "ICA mixture models for unsupervised classification of non-Gaussian classes and automatic context switching in blind signal separation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1078–1089, 2000.

[94] C. Shah, M. Arora, and P. Corresponding, "Unsupervised classification of hyperspectral data:

An ICA mixture model based approach," *International Journal of Remote Sensing*, vol. 25, no. 2, pp. 481–487, 2004.

[95] A. Koltunov and E. Ben-Dor, "A new approach for spectral feature extraction and for unsupervised classification of hyperspectral data based on the Gaussian mixture model," *Remote Sensing Reviews*, vol. 20, no. 2, pp. 123–167, 2001.

[96] ——, "Mixture density separation as a tool for high-quality interpretation of multi-source remote sensing data and related issues," *International Journal of Remote Sensing*, vol. 25, pp. 3275–3299, 2004.

[97] M. Collins, C. Dymond, and E. Johnson, "Mapping subalpine forest types using networks of nearest neighbour classifiers," *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1701–1721, 2004.

[98] P. Hardin, "Parametric and nearest-neighbor methods for hybrid classification: A comparison of pixel assignment accuracy," *Photogrammetric Engineering and Remote Sensing*, vol. 60, no. 12, pp. 1439–1448, 1994.

[99] R. Haapanen, A. Ek, M. Bauer, and A. Finley, "Delineation of forest/nonforest land use classes using nearest neighbor methods," *Remote Sensing of Environment*, vol. 89, no. 3, pp. 265–271, 2004.

[100] G. Foody, "Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data," *International Journal of Remote Sensing*, vol. 17, no. 7, pp. 1317–1340, 1996.

[101] J. Zhang and G. Foody, "Fully-fuzzy supervised classification of sub-urban land cover from remotely sensed imagery: Statistical and Artificial Neural Network approaches," *International Journal of Remote Sensing*, vol. 22, no. 4, pp. 615–628, 2001.

[102] J. Adams, D. Sabol, V. Kapos, R. Almeida Filho, D. Roberts, M. Smith, and A. Gillespie, "Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon," *Remote Sensing of Environment*, vol. 52, no. 2, pp. 137–154, 1995.

[103] D. Lu, E. Moran, and M. Batistella, "Linear mixture model applied to Amazonian vegetation classification," *Remote Sensing of Environment*, vol. 87, no. 4, pp. 456–469, 2003.

[104] D. Roberts, M. Gardner, R. Church, S. Ustin, G. Scheer, and R. Green, "Mapping chaparral in the Santa Monica Mountains using multiple endmember spectral mixture models," *Remote Sensing of Environment*, vol. 65, no. 3, pp. 267–279, 1998.

[105] B. Mannan and A. Ray, "Crisp and fuzzy competitive learning networks for supervised classification of multispectral IRS scenes," *International Journal of Remote Sensing*, vol. 24, no. 17, pp. 3491–3502, 2003.

[106] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 26, no. 1, pp. 52–67, 1996.

[107] R. Schowengerdt, "On the estimation of spatial-spectral mixing with classifier likelihood functions," *Pattern Recognition Letters*, vol. 17, no. 13, pp. 1379–1387, 1996.

[108] P. Aplin and P. Atkinson, "Sub-pixel land cover mapping for per-field classification," *International Journal of Remote Sensing*, vol. 22, no. 14, pp. 2853–2858, 2001.

[109] C. Lloyd, S. Berberoglu, P. Curran, and P. Atkinson, "A comparison of texture measures for the per-field classification of Mediterranean land cover," *International Journal of Remote Sensing*, vol. 25, no. 19, pp. 3943–3965, 2004.

[110] U. Benz, P. Hofmann, G. Willhauck, I. Lingenfelder, and M. Heynen, "Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3, pp. 239–258, 2004.

[111] I. Gitas, G. Mitri, and G. Ventura, "Object-based image classification for burned area mapping of Creus Cape, Spain, using NOAA-AVHRR imagery," *Remote Sensing of Environment*, vol. 92, no. 3, pp. 409–413, 2004.

[112] V. Walter, "Object-based classification of remote sensing data for change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 58, no. 3, pp. 225–238, 2004.

[113] L. Wang, W. Sousa, P. Gong, and G. Biging, "Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of Panama," *Remote Sensing of Environment*, vol. 91, no. 3, pp. 432–440, 2004.

[114] S. Magnussen, P. Boudewyn, and M. Wulder, "Contextual classification of Landsat TM images to forest inventory cover types," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2421–2440, 2004.

[115] E. Mohn, N. Hjort, and G. Storvik, "A simulation study of some contextual classification methods for remotely sensed data," *Geoscience and Remote Sensing, IEEE Transactions on*, no. 6, pp. 796–804, 1987.

[116] B. Jeon and D. Landgrebe, "Classification with spatio-temporal interpixel class dependency contexts," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 30, no. 4, pp. 663–672, 1992.

[117] Y. Jhung and P. Swain, "Bayesian contextual classification based on modified M-estimates and Markov random fields," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 34, no. 1, pp. 67–75, 1996.

[118] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 6, pp. 721–741, 1984.

[119] J. Stuckens, P. Coppin, and M. Bauer, "Integrating contextual information with per-pixel classification for improved land cover classification," *Remote Sensing of Environment*, vol. 71, no. 3, pp. 282–296, 2000.

[120] M. Hodgson, J. Jensen, J. Tullis, K. Riordan, and C. Archer, "Synergistic use of LIDAR and color aerial photography for mapping urban parcel imperviousness," *Photogrammetric Engineering and Remote Sensing*, vol. 69, no. 9, pp. 973–980, 2003.

[121] Q. Du and C. Chang, "A linear constrained distance-based discriminant analysis for hyperspectral image classification," *Pattern Recognition*, vol. 34, no. 2, pp. 361–373, 2001.

[122] S. Myint, "A robust texture analysis and classification approach for urban land-use and land-

cover feature discrimination," *Geocarto International*, vol. 16, no. 4, pp. 29–40, 2001.

[123] G. Okin, D. Roberts, B. Murray, and W. Okin, "Practical limits on hyperspectral vegetation discrimination in arid and semiarid environments," *Remote Sensing of Environment*, vol. 77, no. 2, pp. 212–225, 2001.

[124] J. Benediktsson, J. Sveinsson, and K. Amason, "Classification and feature extraction of AVIRIS data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 33, no. 5, pp. 1194–1205, 1995.

[125] M. Ulfarsson, J. Benediktsson, and J. Sveinsson, "Data fusion and feature extraction in the wavelet domain," *International Journal of Remote Sensing*, vol. 24, no. 20, pp. 3933–3945, 2003.

[126] T. Rashed, J. Weeks, M. Gadalla, and A. Hill, "Revealing the anatomy of cities through spectral mixture analysis of multispectral satellite imagery: A case study of the Greater Cairo region, Egypt." *Geocarto International*, vol. 16, no. 4, pp. 7–18, 2001.

[127] R. Lunetta and M. Balogh, "Application of multi-temporal Landsat 5 TM imagery for wetland identification," *Photogrammetric Engineering and Remote Sensing*, vol. 65, no. 11, pp. 1303–1310, 1999.

[128] M. Heinl and U. Tappeiner, "The benefits of considering land cover seasonality in multispectral image classification," *Journal of Land Use Science*, vol. 1, pp. 1–19, 2011.

[129] J. Townshend, T. Goff, and C. Tucker, "Multitemporal dimensionality of images of Normalized Difference Vegetation Index at continental scales," *Geoscience and Remote Sensing, IEEE Transactions on*, no. 6, pp. 888–895, 1985.

[130] M. Hall-Beyer, "Comparison of single-year and multiyear NDVI time series principal components in cold temperate biomes," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 41, no. 11, pp. 2568–2574, 2003.

[131] R. Colditz, M. Schmidt, C. Conrad, M. Hansen, and S. Dech, "Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions," *Remote Sensing of Environment*, vol. 115, pp. 3264–3275, 2011.

[132] V. Simonneaux, B. Duchemin, D. Helson, S. Er-Raki, A. Olioso, and A. Chehbouni, "The use of high-resolution image time series for crop classification and evapotranspiration estimate over an irrigated area in central Morocco," *International Journal of Remote Sensing*, vol. 29, no. 1, pp. 95–116, 2008.

[133] N. Viovy, "Automatic classification of time series (ACTS): A new clustering method for remote sensing time series," *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1537–1560, 2000.

[134] R. Juárez and W. Liu, "FFT analysis on NDVI annual cycle and climatic regionality in Northeast Brazil," *International Journal of Climatology*, vol. 21, no. 14, pp. 1803–1820, 2001.

[135] M. Jakubauskas, D. Legates, and J. Kastens, "Crop identification using harmonic analysis of time-series AVHRR NDVI data," *Computers and Electronics in Agriculture*, vol. 37, no. 1, pp. 127–139, 2002.

[136] D. de Castro Victoria, A. da Paz, A. Coutinho, J. Kastens, and J. Brown, "Cropland area estimates using MODIS NDVI time series in the state of Mato Grosso, Brazil," *Pesquisa Agropecuária Brasileira, Brasília*, vol. 47, no. 9, pp. 1270–1278, 2012.

[137] G. Galford, J. Mustard, J. Melillo, A. Gendrin, C. Cerri, and C. Cerri, "Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil," *Remote Sensing of Environment*, vol. 112, no. 2, pp. 576–587, 2008.

[138] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag New York Inc, 2000.

[139] C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[140] J. Gualtieri and R. Cromp, "Support Vector Machines for hyperspectal remote sensing classification," in *Proceedings of the SPIE, 27th AIPR Workshop: Advances in Computer Assisted Recognition*, Washington, DC, October 1998, pp. 221–232.

[141] G. Zhu and D. Blumberg, "Classification using ASTER data and SVM algorithms;: The case study of Beer Sheva, Israel," *Remote Sensing of Environment*, vol. 80, no. 2, pp. 233–240, 2002.

[142] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with Support Vector Machines," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 8, pp. 1778–1790, 2004.

[143] M. Pal and P. Mather, "Support Vector Machines for classification in remote sensing," *International Journal of Remote Sensing*, vol. 26, no. 5, pp. 1007–1011, 2005.

[144] Y. Liu, Y. Xu, R. Shi, and Z. Niu, "Evaluation of various classifiers on regional land cover classification using MODIS data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2005 IEEE International*, vol. 2, July 2005, pp. 1281–1283.

[145] J. Guo, J. Zhang, Y. Zhang, and Y. Cao, "Study on the comparison of land cover classification for multitemporal MODIS images," in *International Workshop on Earth Observation and Remote Sensing Applications*, Beijing, June 2008, pp. 1–6.

[146] C. Ye, Y. Liu, J. Peng, P. Song, and D. Zhao, "Improving MODIS land cover classification using NDVI time-series and Support Vector Machines in the Poyang Lake Basin, China," in *Wireless Communications Networking and Mobile Computing*, Chengdu, September 2010, pp. 1–4.

[147] H. Cai and S. Zhang, "Regional land cover classification from MODIS time-series and geographical data using Support Vector Machines," in *IEEE Youth Conference on Information Computing and Telecommunications*, Beijing, November 2010, pp. 102–105.

[148] S. Li, Y. Zhu, J. Feng, P. Ai, and X. Chen, "Comparative study of three feature selection algorithms for regional land cover classification using MODIS data," in *Congress on Image and Signal Processing*, 2008, pp. 565–569.

[149] B. Salmon, J. Olivier, K. Wessels, W. Kleynhans, F. Van den Bergh, and K. Steenkamp, "Unsupervised land cover change detection: Meaningful sequential time series analysis," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, no. 99, pp. 1–9, 2011.

[150] R. Lasaponara and A. Lanorte, "Satellite time-series analysis," *International Journal of Remote Sensing*, vol. 33, no. 15, pp. 4649–4652, 2012.

[151] A. Srivastava, "A novel approach of land cover change mapping using hypertemporal images," Master's thesis, University of Twente, 2011.

[152] T. Sohl, "Change analysis in the United Arab Emirates: an investigation of techniques," *Photogrammetric Engineering and Remote Sensing*, vol. 65, no. 4, pp. 475–484, 1999.

[153] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: Application to Tsunami-damage assessment," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 45, no. 6, pp. 1658–1670, 2007.

[154] T. Celik, "A Bayesian approach to unsupervised multiscale change detection in synthetic aperture radar images," *Signal Processing*, vol. 90, no. 5, pp. 1471–1485, 2010.

[155] T. Celik and K. Ma, "Unsupervised change detection for satellite images using dual-tree complex wavelet transform," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 48, no. 3, pp. 1199–1210, 2010.

[156] C. Jha and N. Unni, "Digital change detection of forest conversion of a dry tropical Indian forest region," *International Journal of Remote Sensing*, vol. 15, no. 13, pp. 2543–2552, 1994.

[157] A. Prakash and R. Gupta, "Land-use mapping and change detection in a coal mining area-a case study in the Jharia coalfield, India," *International Journal of Remote Sensing*, vol. 19, no. 3, pp. 391–410, 1998.

[158] J. Townshend and C. Justice, "Spatial variability of images and the monitoring of changes in the Normalized Difference Vegetation Index," *International Journal of Remote Sensing*, vol. 16, no. 12, pp. 2187–2195, 1995.

[159] E. Lambin and A. Strahler, "Indicators of land-cover change for Change Vector Analysis in multitemporal space at coarse spatial scales," *International Journal of Remote Sensing*, vol. 15, no. 10, pp. 2099–2119, 1994.

[160] J. Deng, K. Wang, Y. Deng, and G. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *International Journal of Remote Sensing*, vol. 29, no. 16, pp. 4823–4838, 2008.

[161] J. Collins and C. Woodcock, "An assessment of several linear change detection techniques for mapping forest mortality using multitemporal Landsat TM data," *Remote Sensing of Environment*, vol. 56, no. 1, pp. 66–77, 1996.

[162] M. Ridd and J. Liu, "A comparison of four algorithms for change detection in an urban environment," *Remote Sensing of Environment*, vol. 63, no. 2, pp. 95–100, 1998.

[163] C. Munyati and T. Kabanda, "Using multitemporal Landsat TM imagery to establish land use pressure induced trends in forest and woodland cover in sections of the Soutpansberg Mountains of Venda region, Limpopo Province, South Africa," *Regional Environmental Change*, vol. 9, no. 1, pp. 41–56, 2009.

[164] L. Bruzzone and S. Serpico, "An iterative technique for the detection of land-cover transitions in multitemporal remote-sensing images," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 35, no. 4, pp. 858–867, 1997.

[165] G. Moser, S. Serpico, and G. Vernazza, "Unsupervised change detection from multichannel SAR images," *Geoscience and Remote Sensing Letters, IEEE*, vol. 4, no. 2, pp. 278–282, 2007.

[166] S. Gopal and C. Woodcock, "Remote sensing of forest change using Artificial Neural Networks," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 34, no. 2, pp. 398–404, 1996.

[167] S. Macomber and C. Woodcock, "Mapping and monitoring conifer mortality using remote sensing in the Lake Tahoe Basin," *Remote Sensing of Environment*, vol. 50, no. 3, pp. 255–266, 1994.

[168] O. Samain, J. Roujean, and B. Geiger, "Use of a Kalman filter for the retrieval of surface BRDF coefficients with a time-evolving model based on the ECOCLIMAP land cover classification," *Remote Sensing of Environment*, vol. 112, no. 4, pp. 1337–1346, 2008.

[169] M. Chen, S. Liu, L. Tieszen, and D. Hollinger, "An improved state-parameter analysis of ecosystem models using data assimilation," *Ecological Modelling*, vol. 219, no. 3-4, pp. 317–326, 2008.

[170] J. Slater and R. Brown, "Changing landscapes: Monitoring environmentally sensitive areas using satellite imagery," *International Journal of Remote Sensing*, vol. 21, no. 13-14, pp. 2753–2767, 2000.

[171] W. Michalak, "GIS in land use change analysis: Integration of remotely sensed data into GIS," *Applied Geography*, vol. 13, no. 1, pp. 28–44, 1993.

[172] T. Stone and P. Lefebvre, "Using multi-temporal satellite data to evaluate selective logging in Para, Brazil," *International Journal of Remote Sensing*, vol. 19, no. 13, pp. 2517–2526, 1998.

[173] G. Henebry, "Detecting change in grasslands using measures of spatial dependence with Landsat TM data," *Remote Sensing of Environment*, vol. 46, no. 2, pp. 223–234, 1993.

[174] F. Wang, "A knowledge-based vision system for detecting land changes at urban fringes," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 31, no. 1, pp. 136–145, 1993.

[175] R. Lawrence and W. Ripple, "Calculating change curves for multitemporal satellite imagery: Mount St. Helens 1980–1995," *Remote Sensing of Environment*, vol. 67, no. 3, pp. 309–319, 1999.

[176] J. Morisette, S. Khorram, and T. Mace, "Land-cover change detection enhanced with generalized linear models," *International Journal of Remote Sensing*, vol. 20, no. 14, pp. 2703–2721, 1999.

[177] T. Yue, S. Chen, B. Xu, Q. Liu, H. Li, G. Liu, and Q. Ye, "A curve-theorem based approach for change detection and its application to Yellow River Delta," *International Journal of Remote Sensing*, vol. 23, no. 11, pp. 2283–2292, 2002.

[178] Q. Zhang, J. Wang, X. Peng, P. Gong, and P. Shi, "Urban built-up land change detection with road density and spectral information from multi-temporal Landsat TM data," *International Journal of Remote Sensing*, vol. 23, no. 15, pp. 3057–3078, 2002.

[179] J. Read and N. S.-. N. Lam, "Spatial methods for characterising land cover and detecting land-cover changes for the tropics," *International Journal of Remote Sensing*, vol. 23, no. 12, pp. 2457–2474, 2002.

[180] J. Borak, E. Lambin, and A. Strahler, "The use of temporal metrics for land cover change detection at coarse spatial scales," *International Journal of Remote Sensing*, vol. 21, no. 6-7, pp. 1415–1432, 2000.

[181] R. Kennedy, W. Cohen, and T. Schroeder, "Trajectory-based change detection for automated characterization of forest disturbance dynamics," *Remote Sensing of Environment*, vol. 110, no. 3, pp. 370–386, 2007.

[182] A. Ramoela, "An innovative method to map land cover changes at a country level utilising hyper-temporal satellite images: A case study of Portugal," Master's thesis, International Institute for Geo-Information Science and Earth Observation, 2007.

[183] C. De Bie, M. Khan, A. Toxopeus, V. Venus, and A. Skidmore, "Hypertemporal image analysis for crop mapping and change detection," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 37, pp. 803–814, 2008.

[184] D. Mildrexler, M. Zhao, and S. Running, "Testing a MODIS global Disturbance Index across North America," *Remote Sensing of Environment*, vol. 113, no. 10, pp. 2103–2117, 2009.

[185] N. Coops, M. Wulder, and D. Iwanicka, "Large area monitoring with a MODIS-based Disturbance Index (DI) sensitive to annual and seasonal variations," *Remote Sensing of Environment*, vol. 113, no. 6, pp. 1250–1261, 2009.

[186] J. Beltran-Abaunza, "Method development to process hyper-temporal remote sensing images for change mapping," Master's thesis, International Institute for Geo-Information Science and Earth Observation, 2009.

[187] S. Boriah, V. Mithal, A. Garg, V. Kumar, M. Steinbach, C. Potter, and S. Klooster, "A comparative study of algorithms for land cover change," in *Proceedings of the Conference on Intelligent Data Understanding*, 2010.

[188] F. Van Den Bergh, K. Wessels, S. Miteff, T. Van Zyl, A. Gazendam, and A. Bachoo, "HiTempo: A platform for time-series analysis of remote-sensing satellite data in a high-performance computing environment," *International Journal of Remote Sensing*, vol. 33, no. 15, pp. 4720–4740, 2012.

[189] W. Kleynhans, B. Salmon, J. Olivier, K. Wessels, and F. Van den Bergh, "An autocorrelation analysis approach to detecting land cover change using hyper-temporal time-series data," in *Geoscience and Remote Sensing Symposium (IGARSS), 2011 IEEE International*, 2011, pp. 94–97.

[190] W. Kleynhans, B. Salmon, J. Olivier, F. van den Bergh, K. Wessels, T. Grobler, and K. Steenkamp, "Land cover change detection using autocorrelation analysis on MODIS time-series data: Detection of new human settlements in the Gauteng province of South Africa," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 5, no. 3, pp. 777–783, 2012.

[191] W. Kleynhans, J. Olivier, K. Wessels, B. Salmon, F. Van den Bergh, and K. Steenkamp, "Detecting land cover change using an extended Kalman filter on MODIS NDVI time-series data," *Geoscience and Remote Sensing Letters, IEEE*, no. 99, pp. 506–510, 2011.

[192] V. Chandola and R. Vatsavai, "A Gaussian process based online change detection algorithm for monitoring periodic time series," in *Proceedings of SIAM Data Mining Conference*, 2011, pp. 95–116.

[193] Y. Kang, "Real-time change detection in time series based on growing feature quantization," in *The 2012 International Joint Conference on Neural Networks*. IEEE, 2012, pp. 1–6.

[194] J. Kučera, P. Barbosa, and P. Strobl, "Cumulative Sum charts - A novel technique for processing daily time series of MODIS data for burnt area mapping in Portugal," in *International Workshop on the Analysis of Multi-temporal Remote Sensing Images*, Leuven, August 2007.

[195] B. W. Silverman, *Density Estimation*. Chapman and Hall, 1986.

[196] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[197] A. Ihler and M. Mandel, "Kernel Density Estimation Toolbox for MATLAB," 2003. [Online]. Available: `http://www.ics.uci.edu/ ihler/code/kde.html`

[198] T. Grobler *et al.*, "Sequential classification of MODIS time series," in *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*. Munchen, Germany: IEEE, 2012,

pp. 6236–6239.

[199] S. Lhermitte, J. Verbesselt, W. Verstraeten, and P. Coppin, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," *Remote Sensing of Environment*, 2011.

[200] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[201] J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*. Upper Saddle River, New Jersey, USA: Prentice Hall, Inc., 1996.

[202] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, pp. 159–174, 1977.

[203] S. Canu, Y. Grandvalet, V. Guigue, and A. Rakotomamonjy, "SVM and Kernel Methods Matlab Toolbox," Perception Systèmes et Information, INSA de Rouen, Rouen, France, 2005. [Online]. Available: `http://asi.insa-rouen.fr/enseignants/ arakotom/toolbox/index.html`

APPENDIX A

MATHEMATICAL BACKGROUND

This appendix contains the mathematical background that is needed to understand Chapter 3, Chapter 4 and Chapter 5. In Section A.1 an overview is given of stochastic calculus. In the last five sections of the appendix the following mathematical concepts or ideas are summarised: Gaussian quadrature, Cholesky factorisation, the method of Langragian multipliers and KDE.

A.1 STOCHASTIC CALCULUS

The basic definitions used in Chapter 3 are defined in this section and can also be found in [48, 59, 82].

Definition 1 (A σ-field $\mathscr{F}$) *Let Ω be a non-empty set. A σ-field $\mathscr{F}$ on Ω is a family of subsets of Ω such that*

1. *the empty set $\emptyset$ belongs to $\mathscr{F}$;*
2. *if A belongs to $\mathscr{F}$, then so does the complement $\Omega \backslash A$;*
3. *if $A_1, A_2, \cdots$ is a sequence of sets in $\mathscr{F}$, then their union $A_1 \cup A_2 \cup \cdots$ also belongs to $\mathscr{F}$.*

Definition 2 (Family of Borel sets) *The family of Borel sets $\mathbb{B} = \mathscr{B}(\mathbb{R})$ is a σ-field on $\mathbb{R}$. The Borel σ-field in $\mathbb{R}$ is the smallest σ-field containing all intervals in $\mathbb{R}$.*

Definition 3 (Probability measure) *Let $\mathscr{F}$ be a σ-field on Ω. A probability measure P is a function*

$$P : \mathscr{F} \to [0,1]$$

such that

1. $P(\Omega) = 1$;

2. *if* $A_1, A_2, \cdots$ *are pairwise disjoint sets (that is,* $A_i \cap A_j = \emptyset$ *for* $i \neq j$*) belonging to* $\mathscr{F}$*, then*

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots .$$

The triple $(\Omega, \mathscr{F}, P)$ is known as a *probability space*. Each set A contained in $\mathscr{F}$ is referred to as an *event*. Whenever the $P(A) = 1$ then event A is said to occur *almost surely*.

Definition 4 (Random variable) *If* $\mathscr{F}$ *is a* σ*-field on* Ω*, then a function* $\xi : \Omega \to \mathbb{R}$ *is said to be* $\mathscr{F}$*-measurable if*

$$\{\xi \in B\} \in \mathscr{F}$$

for every Borel set $B \in \mathscr{B}(\mathbb{R})$*. If* $(\Omega, \mathscr{F}, P)$ *is a probability space, then such a function* ξ *is called a random variable.*

The short-hand notation $\{\xi \in B\}$ in the above definition represents the inverse image $\xi^{-1}(B)$. In expanded form $\{\xi \in B\}$ is written as

$$\{\omega \in \Omega : \xi(\omega) \in B\}.$$

Definition 5 (The σ field generated by a random variable) *The* σ*-field* $\sigma(\xi)$ *generated by a random variable* $\xi : \Omega \to \mathbb{R}$ *consists of all sets of the form* $\{\xi \in B\}$*, where* B *is a Borel set in* $\mathbb{R}$*.*

Definition 6 (The σ field generated by a family of random variables) *The* σ*-field* $\sigma\ \{\xi_i : i \in I\}$ *generated by a family* $\{\xi_i : i \in I\}$ *of random variables is defined to be the smallest* σ*-field containing all the events of the form* $\{\xi_i \in B\}$*, where* B *is a Borel set in* $\mathbb{R}$ *and* $i \in I$ *(*I *is an index set).*

Definition 7 (The distribution of a random variable and the cumulative distribution function) *Every random variable* $\xi : \Omega \to \mathbb{R}$ *gives rise to a probability measure*

$$P_\xi(B) = P\{\xi \in B\}$$

on $\mathbb{R}$ *defined on the* σ*-field of Borel sets* $B \in \mathscr{B}(\mathbb{R})$*. The probability measure* P_ξ *is called the distribution of* ξ*. The function* $F_\xi : \mathbb{R} \to [0,1]$ *defined by*

$$F_\xi(x) = P\{\xi \leq x\} = P_\xi((-\infty, x]),\ x \in \mathbb{R}$$

is called the cumulative distribution function of ξ.

Definition 7 implies that the measurability of ξ ensures that ξ generates a probability measure P_ξ on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, such that $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P_\xi)$ is also a probability space. If P_ξ is known, then the structure of $(\Omega, \mathscr{F}, P)$ is no longer needed to describe the behavior of ξ.

Definition 8 (Probability density function) *If there is a Borel function* $f_\xi : \mathbb{R} \to \mathbb{R}$ *such that for any Borel set* $B \subset \mathbb{R}$

$$P\{\xi \in B\} = \int_B f_\xi(x)dx,$$

then ξ *is said to be a random variable with absolutely continuous distribution and* f_ξ *is called the density of* ξ. *If there is a (finite or infinite) sequence of pairwise distinct real number* $x_1, x_2, \cdots$ *such that for any Borel set* $B \subset \mathbb{R}$

$$P\{\xi \in B\} = \sum_{x_i \in B} P\{\xi = x_i\},$$

then ξ *is said to have discrete distribution with values* $x_1, x_2, \cdots$ *and mass* $P\{\xi = x_i\}$ *at* x_i.

Definition 9 (Essential supremum) *Let* $(\xi_i)_{i \in I}$ *(where* I *is an index set) be a family of real-valued random variables on* $(\Omega, \mathscr{F}, P)$, *bounded by another variable. The essential supremum of* $(\xi_i)_{i \in I}$ *is* Ξ *(denoted by* $\Xi = \ ess\ sup_I\ \xi_i$*) if*

$$(\forall i \in I)\ \xi_i \leq \mathscr{X}\ P - almost\ surely \Leftrightarrow \Xi \leq \mathscr{X}\ P - almost\ surely.$$

Definition 10 (Expectation) *A random variable* $\xi : \Omega \to \mathbb{R}$ *is said to be integrable if*

$$\int_\Omega |\xi| dP < \infty.$$

Then

$$\mathbb{E}[\xi] = \int_\Omega \xi dP$$

exists and is called the expectation of ξ.

The expectation $\mathbb{E}[\xi]$ can also be expressed as a Riemann integral as follows [48]:

$$\begin{aligned}
\mathbb{E}[\xi] &= \int_\Omega \xi dP \\
&= \int_{\mathbb{R}} x dP_\xi \\
&= \int_{-\infty}^{\infty} x dF_\xi(x) \\
&= \int_{-\infty}^{\infty} x f_\xi(x) dx.
\end{aligned}$$

It is also important to note that $\frac{d}{dx}F_\xi(x) = f_\xi(x)$.

Definition 11 (Conditioning on an event) *For an integrable random variable ξ and any event $B \in \mathscr{F}$ such that $P(B) \neq 0$ the conditional expectation of ξ given B is defined by*

$$\mathbb{E}[\xi|B] = \frac{1}{P(B)}\int_B \xi dP.$$

Definition 12 (Conditioning on a discrete random variable) *Let ξ be an integrable random variable and let η be a discrete random variable. Then the conditional expectation of ξ given η is defined to be a random variable $\mathbb{E}[\xi|\eta]$ such that*

$$\mathbb{E}[\xi|\eta](\omega) = \mathbb{E}[\xi|\{\eta = y_n\}] \text{ if } \eta(\omega) = y_n$$

for any $n = 1, 2, \cdots$.

By using Definition 12 the following proposition can be derived (stated here without proof) [82]:

Proposition 1 *If ξ is an integrable random variable and η is a discrete random variable, then*

1. *$\mathbb{E}[\xi|\eta]$ is $\sigma(\eta)$-measurable;*

2. *For any $A \in \sigma(\eta)$*

$$\int_A \mathbb{E}[\xi|\eta]dP = \int_A \xi dP. \tag{A.1}$$

Definition 13 (Conditioning on an arbitrary random variable) *Let ξ be an integrable random variable and let η be an arbitrary random variable. Then the conditional expectation of ξ given η is defined to be a random variable $\mathbb{E}[\xi|\eta]$ such that*

1. *$\mathbb{E}[\xi|\eta]$ is $\sigma(\eta)$-measurable;*

2. *For any $A \in \sigma(\eta)$*

$$\int_A \mathbb{E}[\xi|\eta]dP = \int_A \xi dP. \tag{A.2}$$

Definition 14 (Conditioning on a σ-field) *Let ξ be an integrable random variable on a probability space $(\Omega, \mathscr{F}, P)$, and let $\mathscr{G}$ be a σ-field contained in $\mathscr{F}$. Then the conditional expectation of ξ given $\mathscr{G}$ is defined to be a random variable $\mathbb{E}[\xi|\mathscr{G}]$ such that*

1. $\mathbb{E}[\xi|\mathscr{G}]$ *is* $\mathscr{G}$*-measurable*

2. *For any* $A \in \mathscr{G}$

$$\int_A \mathbb{E}[\xi|\mathscr{G}]dP = \int_A \xi dP.$$

Definition 15 (Sample path) *The sequence of number* $\xi_1(\omega), \xi_2(\omega), \cdots$ *for any fixed* $\omega \in \Omega$ *is called a sample path.*

Definition 16 (Filtration) *A sequence of* σ*-fields* $\mathscr{F}_1, \mathscr{F}_2, \cdots$ *on* Ω *such that*

$$\mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}$$

is called a filtration.

Definition 17 (Adapted to a filtration) *A sequence of random variables* $\xi_1, \xi_2, \cdots$ *is adapted to a filtration* $\mathscr{F}_1, \mathscr{F}_2 \cdots$ *if* ξ_n *is* $\mathscr{F}_n$*-measurable for each* $n = 1, 2, \cdots$.

Definition 18 (Stopping time) *A random variable* T *with values in the set* $\{1, 2, \cdots\} \cup \{\infty\}$ *is called a stopping time (with respect to a filtration* $\mathscr{F}_n$*) if for each* $n = 1, 2, \cdots$

$$\{T = n\} \in \mathscr{F}_n.$$

At this point Wald's identities can be presented (and is stated here without proof) [48]. Wald's identities were used in Chapter 3 to derive the OC and ASN functions.

Theorem 6 (Wald's identities) *Suppose* $\{s_k; k = 1, 2, \cdots\}$ *is an i.i.d. sequence adapted to the filtration* $\{\mathscr{F}_k\}$*, and let* S_k *denote the sequence of cumulative sums,* $S_k = \sum_{i=1}^{k} s_i$*. Then the following statements are true:*

1. *Suppose* $\mathbb{E}[s_1]$ *is finite, then for every stopping time* T *satisfying* $\mathbb{E}[T] < \infty$, $\mathbb{E}[S_T] = \mathbb{E}[s_1]\mathbb{E}[T] \implies \mathbb{E}[T] = \frac{\mathbb{E}[S_T]}{\mathbb{E}[s_1]}$ *when* $\mathbb{E}[s_1] \neq 0$.

2. *Suppose* $\mathbb{E}[s_1^2]$ *is finite, then for ever stopping time* T *satisfying* $\mathbb{E}[T] < \infty$, $\mathbb{E}[S_T - T\mathbb{E}[s_1]]^2 = \mathbb{E}[T]\mathbb{E}[s_1 - \mathbb{E}[s_1]]^2 \implies \mathbb{E}[T] = \frac{\mathbb{E}[S_T^2]}{\mathbb{E}[s_1^2]}$ *when* $\mathbb{E}[s_1] = 0$.

3. *For scalars* $a, h > 0$ *define the stopping time* $T_{-a}^{h} = \inf\{k | S_k \notin (-b, a)\}$*. Suppose* $\omega \neq 0$ *is such that* $\mathbb{E}[e^{-\omega s_1}] < \infty$*, then* $\mathbb{E}[e^{\omega S_T}(\mathbb{E}[e^{-\omega s_1}])^{-T}] = 1$*, holds for any stopping time* T *such that* $P(T \leq T_{-a}^{h}) = 1$.

Definition 19 (Stochastic process) *A stochastic process is a family of random variables $\xi(t)$ parameterised by $t \in \mathscr{T}$, where $\mathscr{T} \subset \mathbb{R}$. When $\mathscr{T} = \{1,2,\cdots\}$ then $\xi(t)$ is a stochastic process in discrete time (a sequence of random variables). When $\mathscr{T}$ is an interval in $\mathbb{R}$ (typically $\mathscr{T} = [0,\infty]$) then $\xi(t)$ is a stochastic process in continuous time. Moreover, for every $\omega \in \Omega$ the function*

$$\mathscr{T} \ni t \to \xi(t,\omega)$$

is called a sample path of $\xi(t)$.

Definition 20 (Brownian motion) *The Wiener process (or Brownian motion) is a stochastic process $W(t)$ with values in $\mathbb{R}$ defined for $t \in [0,\infty)$ such that*

1. *$W(0) = 0$ almost surely;*

2. *the sample paths $t \to W(t)$ are almost surely continuous;*

3. *for any finite sequence of times $0 < t_1 < \cdots < t_n$ and Borel sets $A_1, \cdots, A_n \subset \mathbb{R}$*

$$P\{W(t_1) \in A_1, \cdots, W(t_n) \in A_n\} = \int_{A_1} \cdots \int_{A_n} p(t_1,0,x_1)p(t_2-t_1,x_1,x_2)\cdots p(t_n-t_{n-1},x_{n-1},x_n)dx_1\cdots dx_2,$$

where

$$p(t,w,y) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-y)^2}{2t}}$$

defined for any $x,y \in \mathbb{R}$ and $t > 0$ is called the transition density.

From Definition 20 the following theorem can be derived (stated here without proof) [82]:

Theorem 7 *A stochastic process $W(t)$, $t \geq 0$, is a Wiener process if and only if the following conditions hold:*

1. *$W(0) = 0$ almost surely;*

2. *the sample paths $t \to W(t)$ are almost surely continuous;*

3. *$W(t)$ has stationary (the distribution of $X(s+t) - X(s)$ does not depend on s for all $s,t > 0$) independent ($\mathbb{E}[(W(u) - W(t))(W(s) - W(r))] = 0$ for any $0 \leq r \leq s \leq t \leq u$) increments;*

4. *the increment $W(t) - W(s)$ has the normal distribution with mean 0 and variance $t - s$ for any $0 \leq s \leq t$.*

Definition 21 (Random step process) *A process $f(t), t \geq 0$ is a random step process if there is a finite sequence of numbers $0 = t_0 < t_1 < \cdots < t_n$ and square integrable random variables $\eta_0, \eta_1, \cdots, \eta_{n-1}$ such that*

$$f(t) = \sum_{j=0}^{n-1} \eta_j \mathbf{1}_{[t_j, t_{j+1}]}(t),$$

where η_j is $\mathscr{F}_{t_j}$-measurable for $j = 0, 1, \cdots, n-1$. The set of random step processes will be denoted by M^2_{step}. In Equation A.3, $\mathbf{1}(t)$ represents the indicator function.

Definition 22 (Stochastic integral of a random step process) *The stochastic integral of a random step process $f \in M^2_{step}$ is defined by*

$$I(f) = \sum_{j=0}^{n-1} \eta_j (W(t_{j+1}) - W(t_j)).$$

Definition 23 *M^2 is the class of stochastic processes $f(t)$, $t \geq 0$ that satisfy*

$$\mathbb{E}\left[\int_0^\infty |f(t)|^2\right] < \infty$$

and

$$\lim_{n\to\infty} \mathbb{E}\left[\int_0^\infty |f(t) - f_n(t)|^2\right] = 0. \tag{A.3}$$

In this case the sequence of step processes $f_1, f_2, \cdots$ approximates f in M^2.

Definition 24 (Itô stochastic integral) *$I(f) \in L^2$ (L^2 is the space of square integrable random variables) is called the Itô stochastic integral [from 0 to ∞] of $f \in M^2$ if*

$$\lim_{n\to\infty} \mathbb{E}[|I(f) - I(f_n)|^2] = 0$$

for any sequence $f_1, f_2, \cdots \in M^2_{step}$ of random step processes that approximates f in M^2 (i.e. such that Equation A.3 is satisfied). $I(f)$ and $\int_0^\infty f(t)dW(t)$ are interchangeable.

A.2 GAUSSIAN QUADRATURE

The Gauss-Legendre quadrature rule is formally expressed as [40]

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2} \sum_{i=1}^{n} w_i f\left(\frac{b-a}{2} z_i + \frac{a+b}{2}\right),$$

where z_i is the i-th root of the Legendre polynomial $P_n(z) = \frac{1}{2^n n!} \frac{d^n}{dz^n}\left[(z^2-1)^n\right]$ and

$$w_i = \frac{2}{\left(1 - z_i^2\right)\left[P_n'(z_i)\right]^2}.$$

A.3 CHOLESKY FACTORISATION

If $\mathbf{A}$ has real entries and is symmetric ($\mathbf{A} = \mathbf{A}^T$) and positive definite ($\mathbf{z}^T\mathbf{A}\mathbf{z}$ is positive, for any column vector $\mathbf{z}$), then $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{L}\mathbf{L}^*,$$

where $\mathbf{L}$ has positive diagonal entries and is a lower triangular matrix, and $\mathbf{L}^*$ is equal to the conjugate transpose of $\mathbf{L}$. Writing $\mathbf{A}$ as the product $\mathbf{L}\mathbf{L}^*$ is known as Cholesky decomposition [40].

The Cholesky algorithm, used to calculate the decomposition matrix $\mathbf{L}$ is described next. The recursive Cholesky algorithm starts by setting $i = 1$ and $\mathbf{A}(1) = \mathbf{A}$. At step i, the matrix $\mathbf{A}(i)$ then has the following form:

$$\mathbf{A}^{(i)} = \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & a_{i,i} & \mathbf{b}_i^* \\ 0 & \mathbf{b}_i & \mathbf{B}^{(i)} \end{pmatrix},$$

where $\mathbf{I}_{i-1}$ is the identity matrix of dimension $i-1$.

If matrix $\mathbf{L}_i$ is defined as

$$\mathbf{L}_i = \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & \sqrt{a_{i,i}} & 0 \\ 0 & \frac{1}{\sqrt{a_{i,i}}}\mathbf{b}_i & \mathbf{I}_{n-i} \end{pmatrix},$$

then $\mathbf{A}(i)$ can be written as

$$\mathbf{A}^{(i)} = \mathbf{L}_i\mathbf{A}^{(i+1)}\mathbf{L}_i^*$$

where

$$\mathbf{A}^{(i+1)} = \begin{pmatrix} \mathbf{I}_{i-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \mathbf{B}^{(i)} - \frac{1}{a_{i,i}}\mathbf{b}_i\mathbf{b}_i^* \end{pmatrix}.$$

Note that $\mathbf{b}_i\mathbf{b}_i^*$ is an outer product. If the above is repeated enough times then at step n (which is also the dimension of the matrix $\mathbf{A}$), $\mathbf{A}(n+1) = \mathbf{I}$. Hence, the lower triangular matrix $\mathbf{L}$ is equal to

$$\mathbf{L} = \mathbf{L}_1\mathbf{L}_2\ldots\mathbf{L}_n.$$

A.4 LAGRANGE MULTIPLIERS

The method of Lagrange multipliers is employed to solve the following optimization problem:

$$\begin{aligned} &\max_{x,y} f(x,y) \\ &\text{s.t. } g(x,y) = c, \end{aligned} \tag{A.4}$$

where f and g are functions that have continuous first order partial derivatives and c is a constant. The Lagrange function is derived from $f(x,y), g(x,y), c$ and a new variable λ (the Lagrange multiplier) and is defined as

$$\mathscr{L}(x,y,\lambda) = f(x,y) + \lambda \cdot \Big(g(x,y) - c\Big).$$

If x_0 and y_0 are solutions of Equation A.4, then there exists an λ_0 such that (x_0, y_0, λ_0) is a stationary point of $\mathscr{L}(x,y,\lambda)$ [40].

A.5 KERNEL DENSITY ESTIMATION

If $(x_1, x_2, \cdots, x_n)$ are i.i.d. samples drawn from a distribution with an unknown density f, then the kernel density estimator of f is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \quad = \frac{1}{nh}\sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big),$$

where $K(x)$ is a symmetric but not necessarily positive function that integrates to one (and is known as the kernel), $h > 0$ is called the bandwidth (functions as a smoothing parameter) and $K_h(x) = 1/hK(x/h)$ (and is known as the scaled kernel). If Gaussian kernels are used to approximate univariate data, and the underlying density being estimated is itself a Gaussian then Silverman's rule of thumb is the optimal choice of h. Silverman's rule of thumb is:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-1/5},$$

where $\hat{\sigma}$ is the standard deviation of the samples $(x_1, x_2, \cdots, x_n)$. The multivariate case is approached in a similar way [195].