UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SEQUENTIAL LAND COVER CLASSIFICATION

by

Etienne Rudolph Ackermann

Submitted in partial fulfilment of the requirements for the degree

Master of Engineering (Electronic Engineering)

in the

Faculty of Engineering, the Built Environment and Information Technology

Department of Electrical, Electronic and Computer Engineering

UNIVERSITY OF PRETORIA

June 2011

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SUMMARY

SEQUENTIAL LAND COVER CLASSIFICATION

by

Etienne Ackermann

Supervisor: Prof JC Olivier

Co-supervisor: Dr AJ van Zyl (with the Department of Mathematics and Applied Mathematics)

Department: Electrical, Electronic and Computer Engineering

University: University of Pretoria

Degree: Master of Engineering (Electronic Engineering)

Keywords: Land cover classification, Sequential analysis, Sequential detection, MODIS, Remote sensing, Multispectral.

Land cover classification using remotely sensed data is a critical first step in large-scale environmental monitoring, resource management and regional planning. The classification task is made difficult by severe atmospheric scattering and absorption, seasonal variation, spatial dependence, complex surface dynamics and geometries, and large intra-class variability.

Most of the recent research effort in land cover classification has gone into the development of increasingly robust and accurate (and also increasingly complex) classifiers by constructing–often in an ad hoc manner–multispectral, multitemporal, multisource classifiers using modern machine learning techniques such as artificial neural networks, fuzzy-sets, and expert systems. However, the focus has always been (almost exclusively) on increasing the classification accuracy of newly developed classifiers. We would of course like to perform land cover classification (i) as *accurately* as possible, but also (ii) as *quickly* as possible. Unfortunately there exists a tradeoff between these two requirements, since the faster we must make a decision, the lower we expect our classification accuracy to be, and conversely, a higher classification accuracy typically requires that we observe more samples (i.e., we must wait longer for a decision).

Sequential analysis provides an attractive (indeed an optimal) solution to handling this tradeoff between the classification accuracy and the detection delay–and it is the aim of this study to apply sequential analysis to the land cover classification task. Furthermore, this study deals exclusively with the binary classification of coarse resolution MODIS time series data in the Gauteng region in South Africa, and more specifically, the task of discriminating between residential areas and vegetation is considered.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

OPSOMMING

SEKWENSIËLE KLASSIFIKASIE VAN GRONDBEDEKKING

deur

Etienne Ackermann

Studieleier: Prof JC Olivier

Mede-studieleier: Dr AJ van Zyl (Departement Wiskunde en Toegepaste Wiskunde)

Departement: Elektriese, Elektroniese en Rekenaaringenieurswese

Universiteit: Universiteit van Pretoria

Graad: Magister in Ingenieurswese (Elektroniese Ingenieurswese)

Sleutelwoorde: Klassifikasie van grondbedekking, Sekwensiële analise, Sekwensiële opsporing, MODIS, Afstandswaarneming, Multispektraal.

Klassifikasie van grondbedekking deur middel van afstandswaargeneemde data is 'n kritiese eerste stap in die grootskaalse monitering van die omgewing, hulpbronbestuur, en streeksbeplanning. Die klassifikasietaak word bemoeilik deur uiterste atmosferiese verspreiding en absorpsie, seisoenale veranderinge, ruimtelike afhanklikheid, komplekse oppervlak-dinamika en strukture, en groot intra-klas veranderlikheid.

Meeste van die onlangse navorsingswerk in grondbedekkingsklassifikasie het gefokus op die ontwikkeling van al hoe kragtiger en akkurater (maar ook meer komplekse) klassifiseerders deur–dikwels op 'n lukrake wyse–multispektrale, multitemporale multibron klassifiseerders te ontwerp met moderne masjienleertegnieke soos kunsmatige neurale netwerke, newelversamelingsleer en deskundige stelsels. Desnieteenstaande was die fokus (byna uitsluitlik) op die toenemende akkuraatheid van nuut ontwikkelde klassifiseerders. Ons sou natuurlik grondbedekkingsklassifikasie (i) so akkuraat as moontlik, maar ook (ii) so gou as moontlik wou kon doen. Ongelukkig speel die twee vereistes teen mekaar af, siende dat 'n vinniger besluit 'n laer akkuraatheid tot gevolg het; en andersom, vereis 'n hoër klassifikasie-akkuraatheid tipies dat ons meer observasies moet waarneem.

Sekwensiële analise voorsien 'n aanloklike (inderdaad 'n optimale) oplossing om die afspeeleffek tussen die akkuraatheid en die waarnemingsoponthoud te hanteer–dit is dan die doel van hierdie studie om sekwensiële analise op die grondbedekkingsklassifikasie-taak toe te pas. Verder hanteer hierdie studie uitsluitlik net die binêre klassifikasie van lae resolusie MODIS tydsreeksdata in die Gautengstreek van Suid-Afrika, en meer spesifiek word die taak om tussen residensiële areas en plantegroei te onderskei, aangepak.

LIST OF ABBREVIATIONS

ARL	Average Run Length
AVHRR	Advanced Very High Resolution Radiometer
AVIRIS	Airborne Visible/Infrared Imaging Spectrometer
BRDF	Bidirectional Reflectance Distribution Function
DN	Digital Number
EOS	Earth Observing System
ERTS-1	Earth Resources Technology Satellite
ETM+	Enhanced Thematic Mapper Plus
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
GIS	Geographic Information System
HRV	High Resolution Visible
ICA	Independent Component Analysis
ISS	International Space Station
IFOV	Instantaneous Field Of View
i.i.d.	independent and identically distributed
IR	Infrared
ISO	International Standards Organisation
LLR	Log Likelihood Ratio
MAP	maximum *a posteriori*
ML	Maximum Likelihood
MODIS	Moderate Resolution Imaging Spectroradiometer
MSS	Multispectral Scanner

NASA	National Aeronautics and Space Administration
NDVI	Normalized Difference Vegetation Index
NOAA	National Oceanographic and Atmospheric Administration
pdf	Probability Density Function
SAE	Sum of Absolute Errors
SAR	Synthetic Aperture Radar
SLAR	Side-Looking Airborne Radar
SD	Solar Diffuser
SDSM	Solar Diffuser Stability Monitor
SNR	Signal-to-Noise Ratio
SoE	Sum of Errors
SPOT	Systeme Pour d'Observation de la Terre
SPRT	Sequential Probability Ratio Test
SSE	Sum of Squared Errors
SVM	Support Vector Machine
TIROS	Television Infrared Observation Satellite
TM	Thematic Mapper
TN	True Negative
TP	True Positive
UAV	Unmanned Aerial Vehicle
UV	Ultraviolet

TABLE OF CONTENTS

Chapter 1

Introduction

> "*The mere formulation of a problem is far more often essential than its solution, which may be merely a matter of mathematical or experimental skill.*"
>
> *Albert Einstein* (1879–1955)

Land cover classification using remotely sensed data—also referred to as *thematic mapping*—is a critical first step in large-scale environmental monitoring, resource management and regional planning [80]. Land cover classification establishes a baseline map which can then be compared against subsequent classifications to detect changes in the land cover. Land use information can also be inferred from land cover data, and it is used in many situations and for various purposes, including the development of strategies to balance conservation and developmental issues.

The subtle difference between *land cover* and *land use* should probably be made explicitly clear, even though the terms are often used interchangeably in much of the literature. *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 which the land serves, such as agriculture, recreation, or wildlife habitat protection.

Approaches such as artificial neural networks and decision trees are widely used to perform multisource land cover classification (that is, from multiple sensors and/or supporting sources of information), and generally exhibit superior classification accuracy compared to single-source classification [80]. Nevertheless, the independent analysis of single-source classifiers can be extremely useful to better understand (and predict) the performance of more sophisticated classifiers, and it can also serve as an empirical lower bound for expected classification accuracy. In this present study we will restrict our attention to single-source statistical land cover classification.

Naturally, we would like to perform land cover classification (i) as *accurately* as possible, but also (ii) as *quickly* as possible. Unfortunately there exists a tradeoff between these two requirements, since the faster we must make a decision, the lower we expect our classification accuracy to be, and conversely, if we require a higher classification accuracy we must typically observe more samples. In fact, it is a well-known fact that the accuracy of a detector can be improved by increasing the signal-to-noise ratio. However, the noise power is usually fixed, so that the detector accuracy can only be improved by increasing the signal energy. In the remote sensing context, where satellite platforms typically have very limited resources on board, increasing the signal energy is particularly impractical and sometimes just plain impossible. Therefore, we must *increase the number of observations*.

Thus our main reason for only considering single-source land cover classification in this study is that we are *actually* more interested in answering the question "*how quickly can we perform land cover classification?*" than attempting to improve on the accuracy of existing classifiers which make use of many different data sources. Of course, if we can answer the question for single-source classification, we would expect that we will be able to perform classification even faster using state-of-the-art, multisource classifiers.

What then do we need to answer our question? We could try to do it purely empirically, but that would not be much fun. Instead, we will develop a simple statistical land cover model, and we will attempt to use the results from *sequential analysis* (refer to [89, 118]) to understand how quickly we can perform land cover classification.

1.1 BACKGROUND AND PROBLEM FORMULATION

Land cover classification (which can essentially be regarded as an m-class hypothesis test where m is the number of distinct land cover types) is of critical importance in many remote sensing applications including resource management, urban planning, as well as disaster warning and damage assessment, to name but a few [92]. The classification task is made difficult by (or characterised by) severe atmospheric scattering and absorption, seasonal variation, spatial dependence, complex surface dynamics and geometries, and large intra-class (within-class) variability.

1.1.1 Current approaches to land cover classification

Most of the recent research effort in land cover classification[1] has gone into the development of increasingly robust and accurate (and also increasingly complex) classifiers by constructing—often in an ad hoc manner—multispectral, multitemporal, multisource classifiers using modern machine learning techniques such as artificial neural networks, fuzzy-sets, and expert systems [80].

However, the focus has always been (almost exclusively) on increasing the classification

[1] We specifically consider *per-pixel* classification, but other areas of research include *subpixel* and *per-field* (and the related *object-oriented*) land cover classification (see [80] for a comprehensive review).

accuracy of newly developed classifiers. In fact, [22] proposed six criteria for evaluating the performance of land cover classification strategies, namely accuracy, reproducibility, robustness, ability to fully use the information content of the data, uniform applicability, and objectiveness. In [27], performance criteria including classification accuracy, computational resources, stability of the algorithm, and robustness to noise in the training data were suggested. In all this, *speed of classification* has never been considered explicitly.

1.1.2 Simple illustration of sequential classification

Consider the following simple example, shown in Figure 1.1. We are given two boxes, labelled "Class 0" and "Class 1", respectively. The labels are then removed, and a box is chosen at random (we don't know which one). Finally, we are tasked to design a classifier which must exhibit a Type II classification error of less than 0.5%. That is, the probability of deciding on Class 0 (which we denote by $\delta = 0$) when Class 1 is the actual underlying class (which is indicated by P_1) must be less than 0.005, or more concisely, $P_1(\delta = 0) < 0.005$.

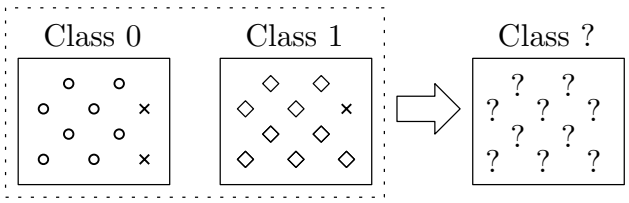


Figure 1.1: Simple example for demonstrating sequential detection or classification.

We may only draw one object from the box at a time, and each object must be replaced before another object may be drawn (i.e., consecutive observations are independent).

If we adopt a Maximum Likelihood (ML) decision rule, we quickly notice that the probability of Type I errors is exactly zero, since the ML rule will declare the box to be of Class 0 whenever anything except a $\diamond$ is observed. Furthermore, the probability of making a Type II error after observing a single observation (Z_1) is given by $P_1(\delta = 0) = P_1(Z_1 = \times) = 0.1$. Similarly, the probability of a Type II error after two observations is given by $P_1(Z_1 = \times, Z_2 = \times) = 0.01$. We notice then that in order to satisfy the Type II error requirement, $P_1(\delta = 0) < 0.005$, we must draw at least three samples from the box. Or rather, with a *fixed-size* classifier we must draw three samples from the box to guarantee $P_1(\delta = 0) < 0.005$. Not so with *sequential* methods, described next.

Unlike with fixed-size classifiers, the number of required observations is not predefined for sequential methods. Instead, sequential methods use only as many observations as are needed to guarantee the desired error probabilities. For example, whenever we observe a $\circ$, we can immediately declare the box to belong to Class 0 without taking any additional observations, and similarly, whenever we observe a $\diamond$ we can immediately declare the box to belong to Class 1. Even if our first observation happens to be a $\times$ (in which case we cannot decide with absolute certainty between the two classes yet), the second observation may be a $\circ$ or a $\diamond$, after which we will be able to decide perfectly

between Class 0 and Class 1.

From this simple example we might reasonably expect that sequential methods require significantly fewer observations than fixed sample size approaches with the same probability of error. In particular, sequential methods decide quickly in unambiguous cases and take longer in ambiguous cases, but since the ambiguous cases are usually much less likely to occur (such as in our simple example), it pays off on the average.

Sequential analysis is then simply a method of statistical inference whose characteristic feature is that the number of observations required by the procedure is not predetermined, and where the decision to terminate the experiment depends, at each stage, on the results of the previous observations [118].

An obvious advantage of the sequential method (as applied to testing statistical hypotheses) is that it helps the user reach a decision between two hypotheses after a minimal average number of experiments [89]. For this reason, it is clear that *sequential* testing is generally *less costly* than fixed sample size (i.e., *non-sequential*) testing whenever there is a nonnegative cost associated with each observation.

Sequential testing of statistical hypotheses is usually referred to as *sequential detection*, and commonly relies on the *likelihood ratio*, Λ_n, (defined below) to make a decision:

$$\Lambda_n \triangleq \prod_{k=1}^{n} \frac{q_1(Z_k)}{q_0(Z_k)} \to \begin{cases} 0, & \text{under } \mathcal{H}_0 \\ \infty, & \text{under } \mathcal{H}_1 \end{cases}, \quad \text{as } n \to \infty,$$

where it is assumed that the sequence of independent and identically distributed (i.i.d.) observations $\{Z_k; k = 1, 2, \ldots\}$ is generated by some random process with corresponding probability densities q_0 and q_1 under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively.

The test that continues sampling as long as $\Lambda_n \in (A, B)$ and then, at the first exit of Λ_n from (A, B), decides on hypothesis $\mathcal{H}_1$ if the exit is to the right of this interval, and decides $\mathcal{H}_0$ if the exit is to the left, is known as the Sequential Probability Ratio Test (SPRT) with boundaries A and B [89]. SPRTs exhibit minimal expected stopping time (i.e. minimal run length) among all sequential (and non-sequential) decision rules having given error probabilities. In particular, we have the following well-known result from Wald and Wolfowitz [119], stated here without proof:

Theorem 1 (Wald-Wolfowitz) *Suppose (τ, δ) is the sequential probability ratio test, SPRT(A, B) with $0 < A \leq 1 \leq B < \infty$, and let (τ', δ') denote any other sequential decision rule with $\max\{\mathbb{E}_0[\tau'], \mathbb{E}_1[\tau']\} < \infty$, and satisfying*

$$P_0(\delta'_{\tau'} = 1) \leq P_0(\delta_\tau = 1) \quad \textit{and} \quad P_1(\delta'_{\tau'} = 0) \leq P_1(\delta_\tau = 0),$$

with

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

Then

$$\mathbb{E}_0[\tau'] \geq \mathbb{E}_0[\tau] \quad \textit{and} \quad \mathbb{E}_1[\tau'] \geq \mathbb{E}_1[\tau].$$

Here τ and τ' are two (Markov) stopping times, δ and δ' are terminal decision rules, A and B are the optimal SPRT exit thresholds, $\mathbb{E}_i[\cdot]$ denotes the expectation under hypothesis $\mathcal{H}_i$, and $P_i(\delta_\tau = j)$ denotes the probability of deciding on hypothesis $\mathcal{H}_j$ when the true hypothesis is $\mathcal{H}_i$. That is, $P_0(\delta_\tau = 1) \equiv P(\text{choose } \mathcal{H}_1 \text{ at stopping time } \tau | \mathcal{H}_0)$.

1.1.3 Problem formulation

As mentioned previously, we would like to perform land cover classification as *accurately* as possible, but also as *quickly* as possible. So with the understanding that current approaches to land cover classification are primarily concerned with classification accuracy; and being equipped with the tools of sequential analysis (including such powerful results as the Wald-Wolfowitz theorem), we would like to be able to answer the following question:

Key question 1
How quickly can we perform land cover classification?
(With a given probability of error.)

Of course, the use of the word *quickly* above refers to the number of samples that we have to consider, and *not* to the (computational) speed of the algorithm, which we expect to be near real-time.

For most simple[2] statistical hypotheses (and those characterised by i.i.d. observations in particular), we can answer Key question 1 fairly simply by using existing results from sequential analysis. However, the complex nature of the remote sensing context makes it difficult to apply the results of sequential analysis directly to the land cover classification problem. This observation naturally leads us to ask the following question:

Key question 2
How should we adapt or extend the results from sequential analysis in order to fit the land cover classification problem?

Luckily it turns out that we do not have to adapt or extend the existing theory of sequential analysis (which is really quite general), but instead we only have to transform (or reformulate) the land cover classification problem into an equivalent, homogenised problem which lends itself to direct application of the existing sequential analysis results.

[2]The qualifier *simple* means that the hypotheses are completely specified, as opposed to partially-specified *composite* hypotheses.

The application of sequential analysis to the (homogenised) land cover classification problem will allow us to construct sequential classifiers which will exhibit minimal run lengths for any given probability of error, so that we will be able to answer Key question 1. Finally then, we can summarise the problem statement as follows:

Problem statement
A sequential land cover classification algorithm is required which must be adapted to the remote sensing context, such that the desired speed and accuracy of classification can easily be modified to suit a particular application.

1.2 OBJECTIVES OF THIS STUDY

There are two categories of objectives discussed in this section, namely *primary* objectives, which concern the key questions and core objectives directly addressed in this study, and *secondary* (or future) objectives, which include the development of an initial basis and framework for related future work.

1.2.1 Primary objectives

With reference to the problem statement given in section 1.1.3, the primary objectives of this study necessarily include the following: (i) to develop a statistical land cover model, which is required for (ii) the design of a sequential land cover classification algorithm, which in turn must be used (iii) to determine the minimum speed of classification for a given probability of error.

1.2.1.1 Objective 1: Develop statistical land cover models

We are primarily interested in classifying (as quickly as possible) the land cover type of surfaces observed by optical sensors on remote sensing platforms. These sensors commonly measure the spectral radiance of electromagnetic energy originating from the Sun, which is subsequently reflected off the surface of the Earth towards the sensors. There are, of course, many factors which influence the spectral reflectance observed by a remote sensing platform, including atmospheric scattering and absorption, seasonal variation, geographic location, and complex surface dynamics and geometries to name a few. A statistical model which takes all of these (and other) factors into account would be extremely complex and difficult to construct—luckily our objective is much simpler.

A statistical land cover model is required which must be simple enough to lend itself to easy interpretation, but which must be complex (rich) enough to allow for sufficiently accurate land cover classification. Furthermore, the statistical model must be flexible enough for use in the design and development of a land cover classification algorithm.

The development of a ‘sufficiently accurate’ model necessarily requires that we must

design, implement and evaluate a sequential classification algorithm (Objective 2), and if needed, that we must refine the statistical model in an iterative manner as illustrated in Figure 1.2, each time re-evaluating the classifier using the refined land cover model.

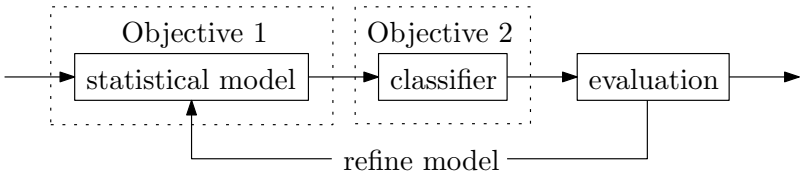


Figure 1.2: Iterative refinement of the statistical land cover model to obtain 'sufficiently accurate' classification performance.

1.2.1.2 Objective 2: Design a sequential land cover classification algorithm

The design of an accurate sequential land cover classification algorithm relies heavily on the assumption that we have sufficient (and accurate) statistical information to describe the classification process (see again Figure 1.2). Of course, if the land cover classes are not separable from the statistical information that we provide to the classifier, even the best sequential classifier can do no better than to guess.

1.2.1.2.1 Objective 2.a—To revisit the maximum likelihood classifier.

There is a common perception that statistical approaches to land cover classification (such as the maximum likelihood classifier) are not particularly useful, or are too limited for practical purposes. For example, the following excerpt is from the recent book by Tso and Mather [112, p. 61], and several other authors share this view:

> "*As the performance of the statistical maximum likelihood classifier is generally limited by frequency distribution assumptions, in recent years, more elegant classifiers such as artificial neural networks, support vector machines, fuzzy theory-based methods, and decision trees have also been introduced into the field of remote sensing imagery classification. These state-of-the-art classifiers should draw our attention, because they normally are distribution-free, and are able to show a significant level of improvements over traditional methods introduced previously...*"

In some sense this is true, since it is often difficult or impossible to construct reliable statistical models, especially when multiple sources of information have to be taken into consideration. Unfortunately, the framework of *sequential analysis* (which we will use to answer Key question 1) assumes that we do have a good statistical model of the land cover classification task (Objective 1); and so we must answer the following question:

Key question 3
Is reliable land cover classification possible using statistical methods? (With specific reference to the maximum likelihood approach.)

Fortunately, the answer to Key question 3 is 'Yes!' (see, for example, the results presented in Chapter 5), so that our objective is then to show that maximum likelihood classification can be equally (or perhaps even more) effective than competing strategies in certain settings (such as single-source classification).

Also, the observation that maximum likelihood classifiers cannot compete with state-of-the-art multisource classifiers is somewhat short-sighted, since a good maximum likelihood classifier can simply become an *additional source* in an *even-more-state-of-the-art* classifier (see [116] for a survey on combining classifiers).

Furthermore, statistical models are often more insightful than the black-box approaches of many state-of-the art classifiers. Therefore research into maximum likelihood classification remains important and easily justifiable.

1.2.1.2.2 Objective 2.b—To design a sequential classification algorithm.

With the assumption that we have a good statistical land cover model, our objective is to design, develop and evaluate a sequential land cover classification algorithm using the techniques and ideas from sequential analysis.

1.2.1.3 Objective 3: Determine the speed of classification

Naturally the task of answering Key question 1 ("*How quickly can we perform land cover classification with a given probability of error?*") is an important objective of this study, which must be answered by using the results from sequential analysis. It must be kept in mind, however, that its answer might be somewhat misleading. Firstly, we will only answer the question *for a particular model* (better models are sure to exist), and *for a particular probability of error*, which is often a rather subjective criterion. Nevertheless, part of our objective is also to show that sequential classification can yield comparable results to fixed sample size classification in a much shorter period of time.

1.2.2 Secondary (future) objectives

This study is further characterised by an important future objective, primarily concerned with moving beyond the present task of land cover classification to the more interesting (and challenging) task of *quickest detection*.

The reason for stating this objective here instead of in section 6.3 (Future research), is that this objectives had been identified even before the commencement of this study, and moreover, the task of developing quickest detection strategies for the land cover context necessitates the development of a sequential detection framework, which then became the focus of this study. In other words, this study serves as an important and natural first step towards a quickest land cover change detection strategy, and this must be kept in mind when evaluating the effectiveness or suitability of various approaches to the land cover classification task presented in the remainder of this document.

1.2.2.1 Objective 4: Apply quickest detection to the land cover change detection task

The theory of quickest detection is an attractive theoretical framework for performing optimal (in an appropriate sense) online change detection, and its application to the remote sensing context (and land cover change detection in particular) is expected to enhance our ability to effectively monitor and manage environmental resources.

In summary, the main objectives of this study are (i) to show that statistical methods can be used to perform reliable land cover classification and sometimes lead to improved understanding or insight, (ii) to apply sequential detection to the simple statistical land cover models developed in this study, and (iii) to lay the foundations for moving on to quickest land cover change detection.

1.3 PROPOSED SOLUTION AND CONTRIBUTIONS

According to Thomson Reuters' ISI Web of Knowledge, 577 papers with "land cover classification," or "change detection" and "remote sensing" in the title, abstract or keywords were published during the 5-year period from 2004 to 2008. However, only 19 of these papers fall within the subject fields of computer science, mathematics or statistics [14]. In this study we aim to revisit statistical land cover classification, and to consider the sequential detection task as well—both mathematically and practically.

In this present study we will not yet answer the more general question of "*How quickly can we perform (arbitrary) land cover classification?*", but we will lay the foundation for answering such questions by considering the related (but considerably easier) question of "*How quickly can we differentiate between two land cover types?*". That is, we will only consider the binary classification task, which can later be extended to a multi-class classifier by considering, for example, a binary tree structure as shown in Figure 1.3.

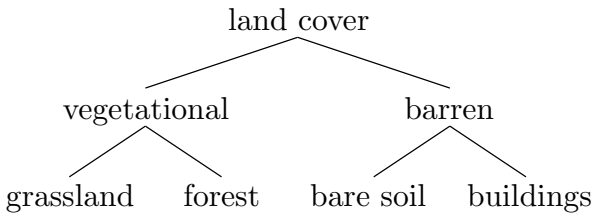


Figure 1.3: Multihypothesis ($m = 4$) land cover classification as a two-stage binary classification tree.

1.3.1 Land cover classification using MODIS surface spectral reflectance time series data

As a result of the rapid advancement of remote sensing technology, remotely sensed data are increasingly being used for land cover classification [126]. Sensors such as the Advanced Very High Resolution Radiometer (AVHRR) have commonly been used

to derive the Normalized Difference Vegetation Index (NDVI), which is useful for characterising and monitoring the phenological dynamics of terrestrial ecosystems, and to perform land cover classification. However, AVHRR was not designed for land applications, and the lack of radiometric calibration, atmospheric (cloud) screening, and poor geometric registration all contribute to excessively noisy data.

The multispectral Moderate Resolution Imaging Spectroradiometer (MODIS) sensor provides a significantly improved basis for a myriad of land applications, since it has an on-board radiometric calibration system, strategically placed narrow spectral bands, and is characterised by improved geometric registration and atmospheric screening.

The specific MODIS data that we will consider also has a relatively high temporal resolution (with a multispectral observation every eight days), but it has a low (or coarse) spatial resolution of 500×500 m, which makes it suitable for large-scale applications.

1.3.2 Development of statistical land cover models

The proposed solution comprises the development of several statistical land cover models derived from the MODIS surface reflectance time series data, followed by maximum likelihood classification and sequential analysis. Statistical multispectral models based on Probability Density Functions (pdfs) have previously been used in [31] and [32], where normally distributed data and a mixture-of-Gaussians were assumed for each land cover class. In this study we intend to refine the statistical land cover models by considering *time-varying* pdfs, such that they will take the seasonal (or temporal) information into account. In addition, we will not assume that the data must be normally distributed, but instead we will allow arbitrary (or strictly speaking, semi-parametric) pdfs, such that it will be easier to observe interesting features in the data.

Both single and multispectral models will be developed using the time-varying framework mentioned above, as well as a simplified i.i.d. framework. The classification results obtained under the i.i.d. assumption will serve as an important classification benchmark, since the assumption that observations are i.i.d. is very common in the literature.

1.3.3 Maximum likelihood classification

The newly developed statistical land cover models will be used to perform maximum likelihood classification, and the results will be compared to those obtained by linear Support Vector Machines (SVMs) in an effort to determine whether statistical methods can compete with modern machine learning alternatives.

Maximum likelihood classification is also a necessary first step in the development of a sequential classification strategy, since we need to objectively compare the classification accuracy of any sequential procedure to the classification accuracy that we could have obtained by considering more samples.

1.3.4 Sequential land cover classification

A sequential land cover classification strategy will be given, in which an i.i.d. land cover model will be used. However, the assumption of independent observations is unrealistic, and ideally we would also like to take the temporal information into account, so that we actually want to perform sequential classification on the time-varying models instead. Nevertheless, as a first attempt at sequential classification, the i.i.d. case presents many challenges of its own. A computational strategy will be given to compute the minimal cost function, $s(\pi)$, which can then be used to find optimal exit thresholds, π_{L} and π_{U}, such that the sequential test continues sampling until the posterior probability that hypothesis $\mathcal{H}_1$ is true, $\pi_n^{\pi} \notin (\pi_{\mathrm{L}}, \pi_{\mathrm{U}})$.

The structure of the optimal time-varying sequential land cover classification decision rule will be given, but there are unfortunately some remaining issues which will prevent us from implementing such a strategy in this study—the most important of which is that it is not yet entirely clear how to physically compute the minimal cost functions, or the optimal exit thresholds for the time-varying models.

1.3.5 Novel contributions of this study

There are several novel contributions of this study which are, in some sense, still in their infancy of development. To the best of my knowledge, the extension of statistical land cover models to be both semi-parametric and time-varying have not been considered previously. Ideally, the models should be extended even further by taking advantage of the information provided by the spatial and temporal dependences between observations.

A second novel contribution of this study is the application of sequential analysis to the land cover classification task, and to explicitly consider the speed-vs-accuracy tradeoff for land cover classification. However, it is expected that a larger number of land cover classes, a larger study area, and more sophisticated statistical land cover models will all contribute to more meaningful results from sequential analysis, so that the present study should only be considered to be the rudiments of sequential land cover classification.

Finally, a computational strategy for computing the minimal cost function $s(\pi)$ under the assumption of i.i.d. observations is given (without proof) in [89], but unfortunately the statement appears to be either incomplete or incorrect. A corrected computational strategy is thus proposed and subsequently proved in this study. However, a computational method for computing the cost functions of the time-varying (non-i.i.d.) case has not been completed yet, and work is ongoing to state and prove such a strategy.

1.4 PUBLICATIONS AND RELATED WORK

The following publications are not really related to this study, but have been prepared and submitted in parallel to completing this study, that is, during the completion of

my Master's degree. A paper on the topic of this study (i.e. sequential land cover classification) is also planned for the near future.

[C1] E. R. Ackermann, T. L. Grobler, J. C. Olivier, A. J. van Zyl and K. C. Steenkamp, "Minimum distance land cover separability analysis of MODIS time series data," IEEE Geoscience and Remote Sensing Symposium, Jul. 2011 (accepted).

[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, Zambia, Sept. 2011 (accepted).

[C3] E. R. Ackermann, T. L. Grobler, A. J. van Zyl and J. C. Olivier, "Belief Propagation for Nonlinear Block Codes," IEEE Africon, Zambia, Sept. 2011 (accepted).

[J1] E. R. Ackermann, J. P. de Villiers and P. J. Cilliers, "Nonlinear dynamic systems modeling using Gaussian processes: predicting ionospheric total electron content over South Africa," *Journal of Geophysical Research,* (accepted, in press).

[J2] T. L. Grobler, E. R. Ackermann, A. J. van Zyl and J. C. Olivier, "Cavalieri Integration," *Canadian Journal of Mathematics* (submitted).

1.5 ORGANISATION OF THIS DISSERTATION

The rest of this dissertation is organised as follows: Chapter 2 introduces the most important concepts related to land cover classification, including an introduction to remote sensing in the optical region (section 2.2) and current approaches to land cover classification (section 2.3). An initial example of single band land cover classification using coarse resolution MODIS surface spectral reflectance data is then presented in some detail in section 2.4.

The sequential detection of binary hypotheses is introduced next in Chapter 3 for homogeneous Markov processes, followed by two illustrative examples, namely the well-known finite-horizon secretary, selection, or marriage problem (section 3.5.2) and an infinite-horizon sequential detection problem involving Bernoulli trials (section 3.5.3).

Chapter 4 then presents the sequential land cover classification task by first describing several statistical land cover models in section 4.2, followed by the formulation of the statistical classification task (either maximum likelihood or sequential classification) for each type of model in section 4.3. Two important considerations concerning the sequential classification task, namely the extension to the time-varying case, as well as some remaining numerical issues encountered when computing the minimal cost functions, are described in section 4.4.1 and section 4.4.2, respectively.

The maximum likelihood and sequential classification results are presented in Chapter 5, along with the classification results for several linear Support Vector Machines (SVMs) in section 5.2, followed by the conclusions and directions for future research in Chapter 6.

Chapter 2

Land Cover Classification

> "*The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it, because it is beautiful. If nature were not beautiful, it would not be worth knowing, and if nature were not worth knowing, life would not be worth living.*"
>
> *Jules Henri Poincaré* (1854–1912)

REMOTE SENSING is the science—or rather art—of inferring information about the Earth's surface using data acquired from distant (remote) sensor platforms. These sensors detect and measure the electromagnetic energy which are reflected from (or emitted by) the Earth's surface, producing a *spectral signature* of the surface.

Remote sensing is not only a vibrant research field, but also finds many practical applications on a regular basis. For example, our daily weather forecasts rely heavily on information gathered by remote sensing platforms. In addition, remotely sensed data is widely used in a range of agricultural [7], atmospheric [90], oceanographic [99], terrestrial [98], and geological [117] applications to name but a few. Nevertheless, in this present study we will only be concerned with remote sensing as applied to *land cover classification* or *thematic mapping*.

2.1 INTRODUCTION

Land cover classification using remotely sensed data is a critical first step in large-scale environmental monitoring, resource management and regional planning [80]. Land cover classification establishes a baseline map which can then be compared against subsequent classifications to detect changes in the land cover, and it also serves as a basic inventory

of environmental resources for all levels of government, environmental agencies, and private industry [74]. Land use information can also be inferred from land cover data, and it is used in many situations and for various purposes, including the development of strategies to balance conservation and developmental issues.

2.1.1 Land cover and land use

As mentioned previously, *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 which the land serves, such as agriculture, recreation, or wildlife habitat protection. Land cover and land use are often very closely related, and both may benefit from accurate and timely information provided by remote sensing systems.

Whether desired land cover information is local, regional, or global in scope, remote sensing provides an efficient, cost-effective means of acquiring the necessary data in a timely manner. However, the task of and cover classification is made difficult by severe atmospheric scattering and absorption, seasonal variation, spatial dependence, complex surface dynamics and geometries, and large intra-class (within-class) variability.

Land use is even more difficult to determine, and remote sensing alone often cannot provide sufficient information to assess or classify various types of land use. For example, even though a piece of land may be agricultural in nature (i.e., its *classification)*, it is virtually impossible to tell (remotely at least) whether the piece of land is being used for commercial, or personal purposes.

Figure 2.1 lists some of the factors affecting land cover classification from remotely sensed data, as well as some factors affecting the more difficult task of land cover *change detection* (shown in grey).

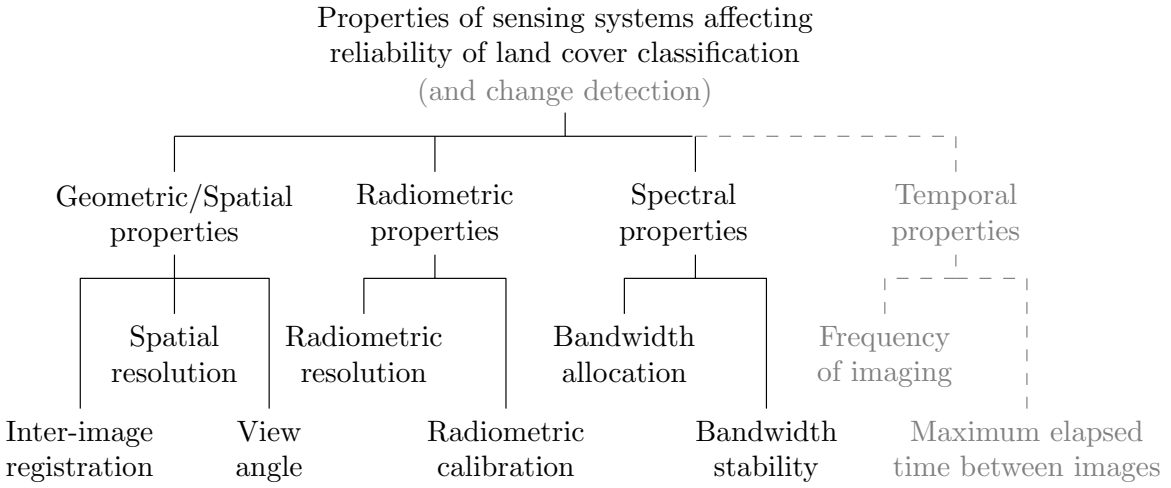


Figure 2.1: Factors affecting reliable land cover classification (and change detection) using remotely sensed data (adapted from [111]).

2.1.2 Why remote sensing?

Remote sensing is not simply the most convenient choice for land cover classification, but it is often the *only* practical choice. It would be extremely difficult, for example, to continuously send out field agents all around the world (including to inhospitable, remote or dangerous destinations such as volcanoes, war-torn countries, or Antarctica). It would not only be very time-consuming, but also incredibly expensive. Fortunately, remote sensing is both practical and very cost effective [74].

With the relatively high frequency of acquisition of many of the remote sensing platforms, it is possible to observe changes in plant phenology (growth) throughout the year, whether relating to changes in chlorophyll content (detectable with visible and infrared light) or structural changes (which is detectable with radar) [74]. Finally, remote sensing also allows us to consider near-continuous spatial coverage over very large areas, making it possible to detect regional trends and to monitor the state of our environment.

2.1.3 Data requirements

The most important characteristics of common types of remote sensing data are summarized in [6] and [73] as being spectral, radiometric, spatial, and temporal resolutions; polarization; and angularity. The scale of the study area (i.e., local, regional or global), image resolution, as well as the user's requirements are the most important factors affecting the selection of remotely sensed data [80].

At a local level, or when objects of interest are relatively small, a high resolution classification system is generally required. At a regional scale, medium spatial resolution data may suffice, while at a continental or global scale, coarse spatial resolution data such as MODIS are preferable.

The most important consideration is ultimately the purpose of the remote sensing application. For example, if global trends in climate change must be monitored, data with a relatively low temporal resolution may be sufficient. However, if a real-time fire detection system is desired, a very high temporal resolution is clearly required. Finally, even though optical remote sensing is the most common approach for land cover classification, it is severely impeded by cloud cover and other atmospheric conditions, in which case microwave remote sensing (such as radar) is expected to be more effective.

2.2 REMOTE SENSING IN THE OPTICAL REGION

There are two regions in the electromagnetic spectrum which is of particular importance in remote sensing, namely the *optical region* (including visible as well as infrared light), with wavelengths ranging from about 0.4 μm to 1 mm, and the *microwave region*, with wavelengths between 1 mm and 10 cm. We will only be concerned with remote sensing in the optical region, since this type of remote sensing is more commonly used for land cover classification than microwave remote sensing.

2.2.1 Historic overview

The French balloonist and photographer Gaspard-Félix Tournachon took the first aerial photographs of a small village near Paris from his balloon in 1858, ushering in an exciting era of Earth observation and remote sensing. Shortly thereafter, aerial photography from balloons was used to reveal key defensive positions in the American Civil War [2].

The next period of rapid development took place during World War I, where aeroplanes were used for large scale photoreconnaissance. After the war, airborne photos were used for geology, forestry, agriculture and cartography applications. During World War II, near-infrared, thermal and radar imaging were developed, again for military purposes.

The first civilian remote sensing satellite was launched in 1960, from which time satellites have become the most important platforms for remote sensing. A list of important developments in remote sensing is given in Table 2.1.

Table 2.1: A short history of remote sensing milestones (from [2]).

Year	Milestone
1800	Discovery of infrared radiation by Sir W. Herschel
1839	Beginning of photography
1847	Infrared spectrum shown by J. B. L. Foucault
1858	Photography from balloons
1873	Theory of electromagnetic radiation by J. C. Maxwell
1909	Photography from airplanes
1916	World War I: aerial reconnaissance
1935	Development of radar in Germany
1940	World War II: applications of invisible part of electromagnetic spectrum
1950	Military research and development
1959	First space photograph of the Earth (Explorer-6)
1960	First TIROS meteorological satellite launched
1970	Skylab remote sensing observations from Space
1972	Launch Landsat-1 (ERTS-1): MSS sensor
1972	Rapid advances in digital image processing
1982	Launch of Landsat-4; new generation of Landsat sensors: TM
1986	French commercial Earth observation satellite SPOT
1986	Development of hyperspectral sensors
1990	Development of high resolution spaceborne systems
First commercial developments in remote sensing	
1998	Towards cheap one-goal satellite missions
1999	Launch EOS: NASA Earth observing mission
1999	Launch of IKONOS, very high spatial resolution sensor system

2.2.2 Introduction to remote sensing

The remote sensing process is a rather complex one, in which scattered or emitted electromagnetic energy from the Earth's surface passes through (and is distorted by) the atmosphere, after which it is detected by instruments or sensors mounted on the remote sensing platform, only to be transmitted back to the Earth's surface for extensive processing and analysis in an attempt to monitor and better understand our environment here on Earth. This process is illustrated in Figure 2.2.

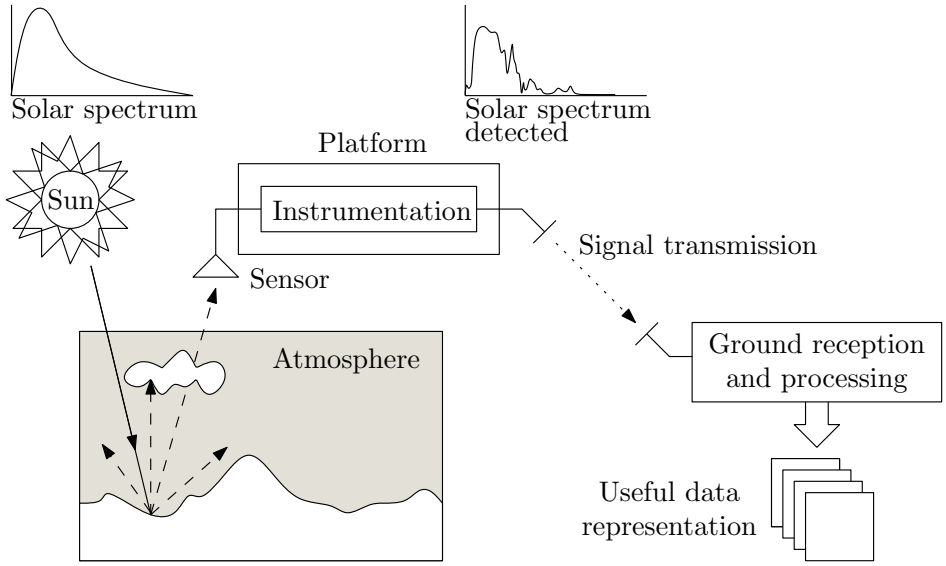


Figure 2.2: Signal and data flow in a typical remote sensing system (adapted from [93]).

2.2.3 Electromagnetic radiation

The (partial) electromagnetic spectrum is shown in Figure 2.3, where the region of particular importance for optical remote sensing is also shown. Several regions of the electromagnetic spectrum are particularly suited for specific applications. For example, the far infrared region can be used to detect thermal signatures for fire detection [61].

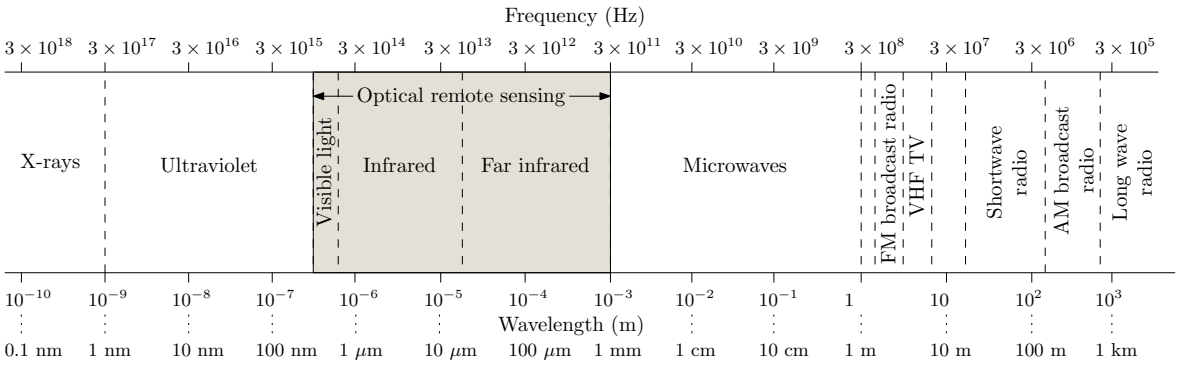


Figure 2.3: The electromagnetic spectrum, showing the region of interest for optical remote sensing.

The optical region of the electromagnetic spectrum is usually divided further into a number of smaller regions as listed in Table 2.2. These divisions are however not precise, and can vary depending on the publication. For example, according to the ISO 20473 standard, the near infrared region corresponds to 0.78–3 μm, whereas astronomers typically regard near infrared as the region 0.7–1.0 μm. Furthermore, the infrared region is sometimes divided into *reflective infrared*, typically ranging from 0.7 to 3.0 μm, and *thermal infrared*, ranging from 3 μm to about 1 mm.

Table 2.2: 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	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

2.2.3.1 Radiance, irradiance and spectral reflectance

Optical remote sensing sensors typically measure the *radiance* emitted by a target object (such as a piece of land), along with some stray sources of radiance, caused for example, by atmospheric scattering. *Reflectance* refers to the ratio of the emitted energy to the incident energy of a target object, and when the atmospheric effects and solar illumination are compensated for in digital remote sensing data, the result is the so-called *apparent reflectance* [13] (which differs from true reflectance in that shadows and directional effects are not taken into consideration).

To be more precise, radiance is a radiometric measure that describes the amount of electromagnetic energy that passes through (or is emitted from) a particular area and which falls within a given solid angle in a specified direction [13]. The SI unit of radiance is watts per steradian per square meter ($\mathrm{W{\cdot}sr^{-1}{\cdot}m^{-2}}$). Radiance characterises the *total emission or reflection*, while *spectral radiance* characterises the electromagnetic energy at a *single* wavelength or frequency. The SI units for spectral radiance are $\mathrm{W{\cdot}sr^{-1}{\cdot}m^{-3}}$ when measured per unit wavelength, and $\mathrm{W{\cdot}sr^{-1}{\cdot}m^{-2}{\cdot}Hz^{-1}}$ when measured per unit frequency interval.

Irradiance refers to the power of electromagnetic radiation incident on a surface per unit area, and it considers *all* the wavelengths or frequencies (i.e., the entire spectrum). The SI unit is $\mathrm{W/m^2}$. *Spectral irradiance* then considers a particular frequency, and has SI unit $\mathrm{W/m^3}$, or, more commonly, $\mathrm{W{\cdot}m^{-2}{\cdot}nm^{-1}}$.

Remark 1 (A misuse of terminology: "spectral reflectance") *In the remainder of this study we will consider* "spectral reflectance" *to mean the quantity obtained from the remote sensing platform, even though this would indicate that (i) we consider only a particular frequency (we will actually consider* intervals or bands of frequencies*), and (ii) that the sensor measures reflectance, which as pointed out above, it does not.*

2.2.3.2 Active vs. passive remote sensing

The electromagnetic energy that is measured by remote sensing systems (i.e., the radiance) can either be reflected sunlight, thermal energy from the Earth itself, or energy from a synthetic (man-made) energy source such as a laser or radar carried on some remote sensing platform [93].

The energy from the Sun, or more specifically the solar spectral irradiance at the upper and lower atmospheric regions, is shown in Figure 2.4, where it is clear that significant losses are incurred as the radiation passes through the atmosphere. Remote sensing sensors which rely on energy originating from the Earth or the Sun are referred to as *passive* sensors, whereas sensors whose energy sources are provided by the remote sensing platforms are said to be *active* sensors.

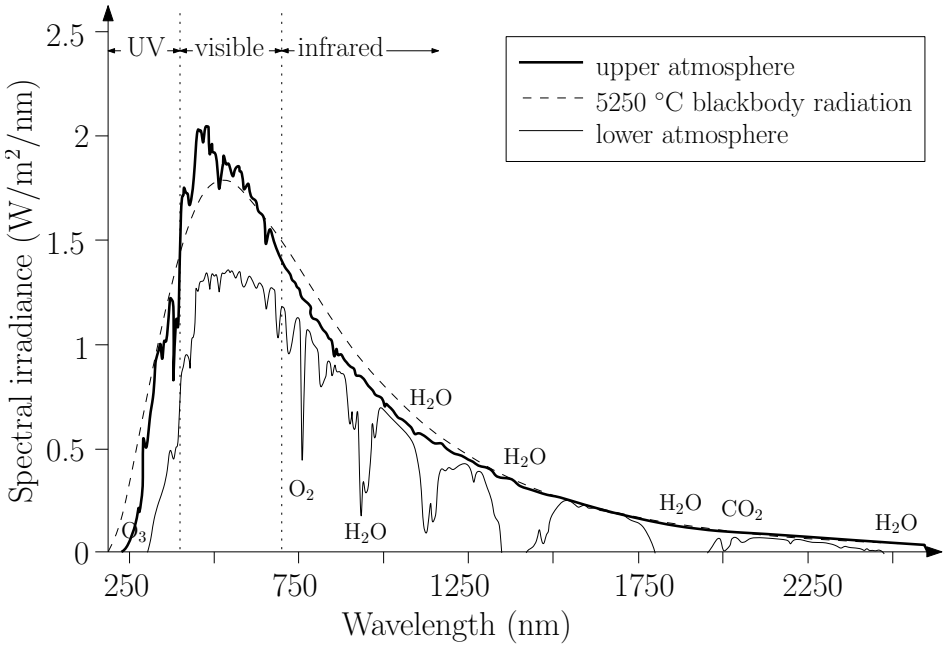


Figure 2.4: The solar radiation spectrum within the atmosphere.

Generally speaking, optical remote sensing systems employ passive sensors and are either air- or spaceborne, whereas microwave remote sensing systems make use of active sensors found on airborne platforms. This is due to the fact that the Sun emits a negligible amount of energy in the microwave region (see Figure 2.4), and even though the Earth does emit some level of microwave radiation, it is usually too small to be useful in any remote sensing applications.

2.2.4 Atmospheric interactions

Whether a sensor measures reflected sunlight or thermal energy emitted by the Earth, the electromagnetic energy must first travel through the atmosphere, which can have a profound effect on the incident energy observed at the sensor. *Absorption* and *scattering* are the two primary mechanisms by which the intensity and direction of electromagnetic radiation are altered as it passes through the atmosphere [112], and they in turn depend on the types (and concentrations) of particulates and gases present along the ray paths.

The atmosphere is a complex mixture of particulates and gases, where most particulates are usually less than 20 μm in diameter [93]. Larger particulates tend to settle on the ground fairly quickly, and are seldom in the atmosphere for significant periods of time.

2.2.4.1 Atmospheric absorption

Atmospheric absorption refers to the selective conversion of electromagnetic energy into thermal energy. As photons collide with airborne molecules in the atmosphere, some of the electromagnetic energy is absorbed by means of electron orbital transitions, induced vibrations, or atomic rotations within the molecules [13]. Oxygen, carbon dioxide, ozone and water molecules in particular attenuate the radiation significantly in certain wavebands [93]. The atmosphere is therefore characterised by so-called *atmospheric transmission windows* in which the absorption is relatively small (or equivalently, where the transmission of radiation is relatively high).

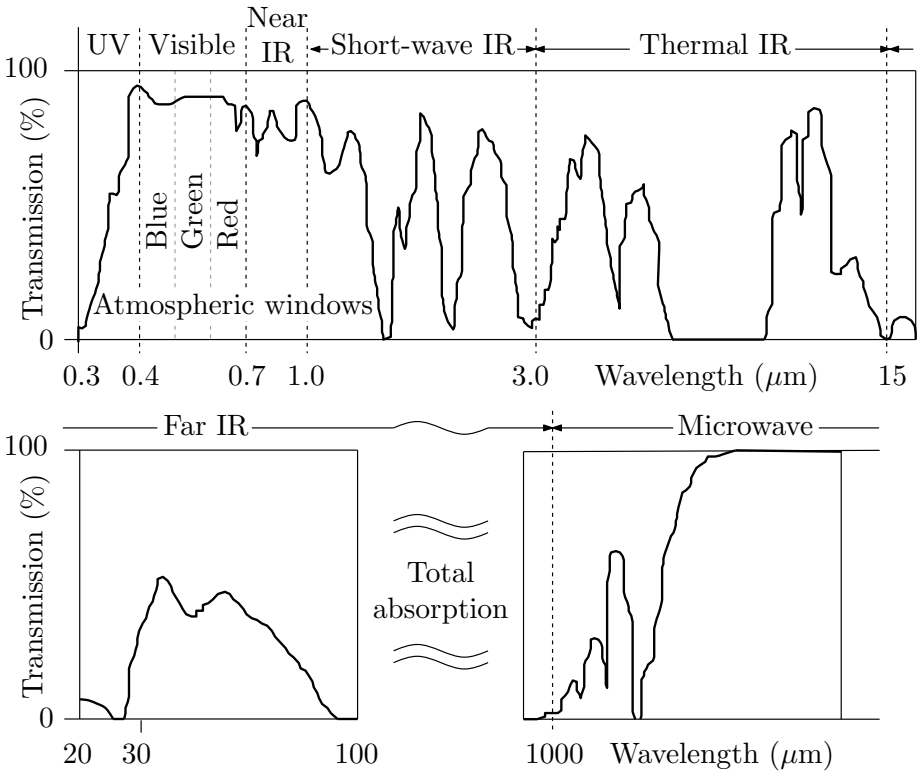


Figure 2.5: Atmospheric electromagnetic transmission windows (from [112, p. 6]).

Some typical atmospheric transmission windows are shown in Figure 2.5 as a function of wavelength, and Figure 2.4 also displays the effects of atmospheric absorption when comparing the exoatmospheric solar spectral irradiance with the lower atmospheric spectral irradiance.

Optical remote sensing sensors are therefore usually designed to operate away from the wavebands which are characterised by severe atmospheric absorption. For such sensors, the dominant mechanism that leads to radiometric distortion (apart from sensor effects and calibration errors) is then atmospheric scattering, described next.

2.2.4.2 Atmospheric scattering

Scattering occurs when electromagnetic energy collides with particulates or large gas molecules present in the atmosphere, causing the electromagnetic radiation to be redirected from its original path. The dynamics of the scattering depend on a number of factors, including the wavelength of the radiation, the concentration and surface geometry of the particulates or gases, and the distance which the radiation travels through the atmosphere. A comprehensive treatment of the complex nature of scattering and its effect on radiation propagation can be found in [19, 114].

There are three basic types of scattering which affect electromagnetic radiation, namely *Rayleigh*, *Mie* and *nonselective* scattering, as shown in Figure 2.6.

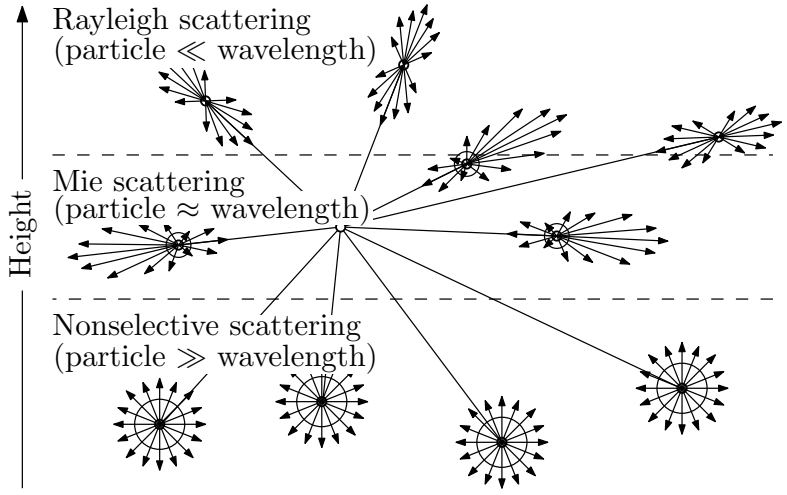


Figure 2.6: Atmospheric scattering, adapted from [13, p. 34].

2.2.4.2.1 Rayleigh scattering. Rayleigh scattering generally occurs at high altitudes, when the size of the particulates are very small relative to the wavelength of the radiation [13]. The primary components for scattering at these altitudes are atmospheric gases such as oxygen and nitrogen, or tiny specks of dust. Rayleigh scattering causes shorter wavelengths to be scattered much more than longer wavelengths, and the scattering is symmetric, with approximately equal amounts of forward and backscatter (see Figure 2.6).

The wavelength dependence of Rayleigh scattering also explains why the sky appears blue[1]: as sunlight passes through the atmosphere, the shorter visible wavelengths (blue) are scattered more than the longer visible wavelengths. At sunrise and sunset the sky often appears red, since the light has to travel much farther through the atmosphere, so that the scattering of the shorter wavelengths is more complete, leaving a greater proportion of the longer wavelengths to penetrate the atmosphere [74].

2.2.4.2.2 Mie scattering. Mie scattering occurs closer to the ground than Rayleigh scattering (up to altitudes of about 5 km) when the diameter of the particulates are about the same as the wavelength of the radiation [13]. Common particulates that are affected by Mie scattering include aerosols, dust particles, pollen, smoke and water vapour. Mie scattering primarily affects light in the visible portion of the spectrum, and is not as wavelength dependent as Rayleigh scattering. As shown in Figure 2.6, Mie scattering causes mostly forward scattering.

2.2.4.2.3 Nonselective scattering. The final scattering mechanism of importance is nonselective scattering, which occurs at low altitudes where the particles are usually much larger than the wavelength of the radiation [13]. Nonselective scattering is not really wavelength dependent, and radiation is scattered uniformly in all directions (see Figure 2.6). This type of scattering usually involves large dust particles, water droplets, ice crystals and hail; and it causes fog and clouds to appear white (since blue, green and red light are scattered in approximately equal quantities).

2.2.5 Surface material reflectance

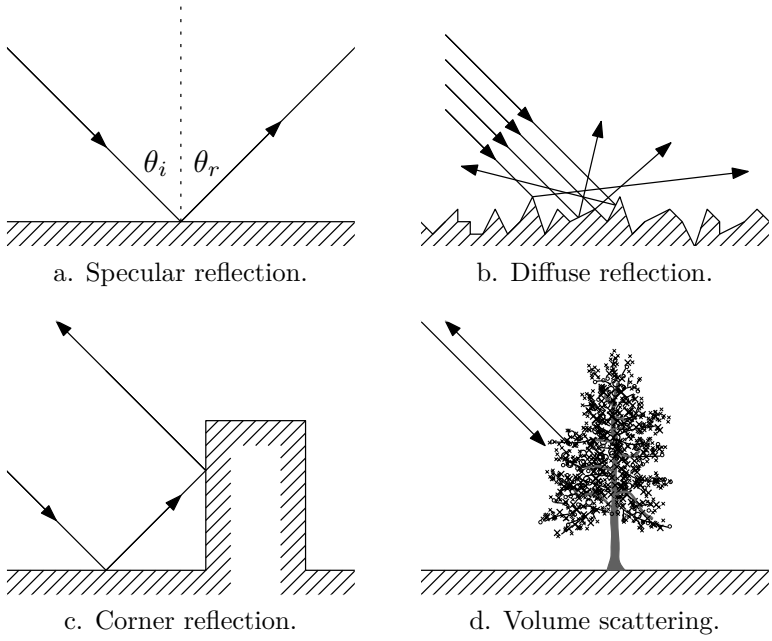
The radiance measured by optical remote sensing systems is typically reflected solar irradiance, so that the reflectance characteristics of different land cover surfaces become important. Surface roughness also affects the radiation observed by remote sensing systems, and is a wavelength-dependent phenomenon, such that surfaces appear smoother under longer wavelengths, and rougher under shorter wavelengths [112].

2.2.5.1 Surface roughness

Very smooth surfaces (such as water) act as *specular* (i.e. mirror-like) reflectors in which the reflection angle, θ_r, equals the incidence angle, θ_i, as shown in Figure 2.7.a. Consequently, such surfaces usually appear dark in optical remote sensing data, since the solar irradiance is seldom reflected directly towards the sensor. Rough surfaces act as *diffuse reflectors* (also called *Lambertian reflectors)*; which scatter the incident energy uniformly in all directions (see Figure 2.7.b). Such surfaces therefore appear lighter in remote sensing data, since some (small but positive) percentage of the incident energy is at least reflected directly towards the sensor.

[1] In fact, it was exactly for this reason—to explain the blue colour of the sky—that Lord Rayleigh first described the phenomenon of scattering in 1871 [78].

Many structures, buildings and other objects with sharp corners are characterised by so-called *corner reflection* (see Figure 2.7.c), especially under microwave surveillance [93]. Such surfaces then give a very bright response, since the energy is often reflected back directly towards the sensor (which in most microwave systems is on the same platform as the radiation source). Finally, surfaces such as vegetation canopies and sea ice are characterised by *volume scattering* [93], in which backscattered energy emerges from many hard-to-define locations within the volume, as shown in Figure 2.7.d.



a. Specular reflection. b. Diffuse reflection.

c. Corner reflection. d. Volume scattering.

Figure 2.7: Types of surface reflection, adapted from [93, p. 8].

Even though the four reflection mechanisms of Figure 2.7 can substantially modify the radiance observed at a remote sensing sensor location, the effects are the most pronounced in microwave remote sensing systems (ranging from about 30 to 300 mm). In contrast, in the visible/infrared range (from about 0.4 to 12 μm), where the sun acts as the primary energy source, the scattering is almost always diffuse, since the smaller wavelengths result in rougher surface appearances, which in turn are characterised by diffuse scattering. As a consequence, the specific type of surface reflection and surface roughness are of little importance in optical remote sensing, and the interaction between electromagnetic energy and different types of surfaces becomes more important instead.

2.2.5.2 Spectral signatures and interactions

The interaction of electromagnetic energy with different types of surfaces is wavelength dependent, and the amount of energy incident on the surface is of course a major (probably the biggest) factor affecting the observed energy at a sensor location. However, apart from the amount of incident energy, material-specific factors such as pigmentation, moisture content and cellular structure of vegetation, the mineral and moisture contents

of soil and the level of sedimentation of water are the most important factors affecting the observed electromagnetic energy reflected from a surface [93]. When these factors are considered together with the wavelength of the incident energy, we can construct *spectral signatures* of different types of surfaces, expressed in terms of *spectral reflectance*.

The spectral reflectance, $\rho(\lambda)$, of a surface object is defined as

$$\rho(\lambda) = \frac{E_r(\lambda)}{E_i(\lambda)}, \tag{2.1}$$

where $E_r(\lambda)$ is the reflected spectral energy (at a wavelength λ) from the surface object, and $E_i(\lambda)$ is the spectral irradiance (at a wavelength λ) incident on the surface object.

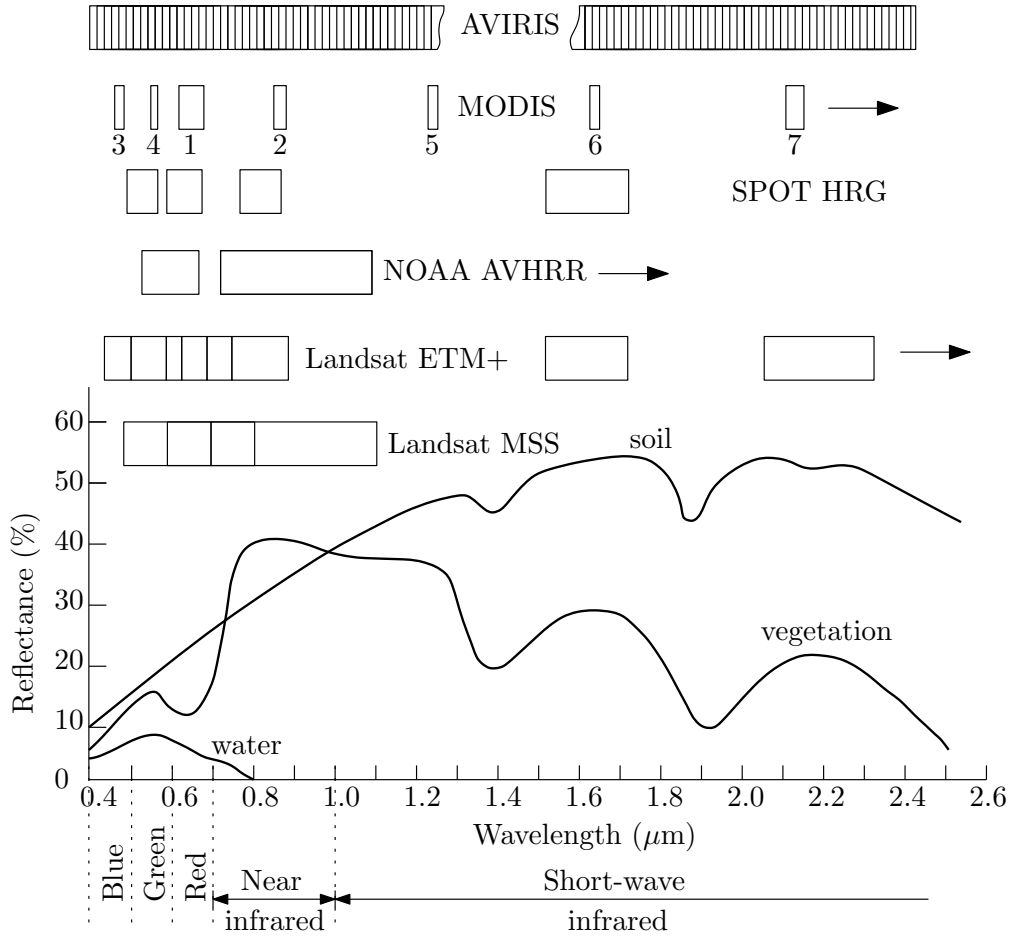


Figure 2.8: Spectral reflectance characteristics of common Earth surface materials in the visible and near-to-mid infrared range [93, p. 5]. The positions of spectral bands for common remote sensing instruments are also indicated.

2.2.5.2.1 Spectral characteristics: Vegetation. Different surface types display different spectral reflectance characteristics. For example, lush vegetation appears green, because it reflects light in the green wavelength. More specifically—with reference to Figure 2.8, which presents the spectral characteristics of three of the most important surfaces namely soil, vegetation, and water—vegetation can be seen to be characterised

by water absorption bands at 1.4 μm, 1.9 μm and 2.7 μm in the short-wave infrared range, high reflectance throughout the near-infrared range (and extending up to about 1.3 μm) which is caused by plant cell structures acting as diffuse reflectors at these wavelengths, and strong absorption of blue and red wavelengths by chlorophyll—leaving chlorophyll-pigmented plants to appear (reflect) green [93].

Vegetation index: NDVI. Spectral ratioing—or image division—is one of the most common image transforms applied to remotely sensed data in an attempt to improve its interpretability. By ratioing the data from two different spectral bands, subtle variations in the spectral responses of various surface covers can be enhanced which might otherwise have been masked by the pixel brightness variations in each of the bands [74].

Variations in scene illumination due to topographic effects can be reduced significantly by spectral ratioing, since the topographic effects are expected to affect both spectral bands equally. For example, although we might expect the absolute reflectances for forest-covered slopes to be dependent on their orientation relative to the sun, the ratio of their reflectances (between the two bands) are expected to be very similar.

One of the most widely used image transforms (which makes use of spectral ratioing) is the NDVI, which is defined as

$$\text{NDVI} = \frac{(\text{Near-infrared Band}) - (\text{Red Band})}{(\text{Near-infrared Band}) + (\text{Red Band})}. \tag{2.2}$$

The NDVI index is based on the observation that vegetation typically has a high reflectance in the near-infrared region, and a lower reflectance in the (visible) red waveband [112], while other surfaces (such as soil and water) show near equal reflectances in both the near-infrared and red regions (refer to Figure 2.8).

In this way NDVI can be derived for a multitude of multispectral sensors (assuming of course that these instruments have wavebands in the near-infrared and red regions). For example, it can be derived from bands 7 and 5 of the Landsat Thematic Mapper, bands 2 and 3 of SPOT HRV, and bands 2 and 1 of the MODIS sensor (see Figure 2.8). In this study we will only consider NDVI as derived from the MODIS bands:

$$\text{NDVI}_{\text{MODIS}} = \frac{(\text{Band 2}) - (\text{Band 1})}{(\text{Band 2}) + (\text{Band 1})}. \tag{2.3}$$

NDVI is considered to be relatively insensitive to changes in atmospheric conditions, and has therefore been widely applied for vegetation monitoring [48, 112]. NDVI has also been used for land cover classification and image segmentation (see for example, [75]). Finally, an overview of the usefulness of different vegetation indices (including NDVI) derived specifically from MODIS is given in [51].

2.2.5.2.2 Spectral characteristics: Water. Most of the radiation incident upon water is either absorbed or transmitted, while for soil the majority of the incident energy is either absorbed or reflected. Water also absorbs energy with longer wavelengths more than those with shorter wavelengths, such that it appears blue-green in the visible spectrum, and dark in the red to infrared range. Suspended sediments or shallow water bodies may cause increased reflection to occur, including small amounts of energy in the near-infrared range [93].

2.2.5.2.3 Spectral characteristics: Soil. Different types of soil are also characterised by the same water absorption bands as vegetation, centred at about 1.4 μm, 1.9 μm and 2.7 μm (see Figure 2.8), which are almost unnoticeable in extremely dry soil. Furthermore, soils containing hydroxyl (such as clay soil) are characterised by small absorption bands at 1.4 μm and 2.2 μm [93]. Other important factors affecting the spectral characteristics of soils include organic matter content, texture, structure, and iron oxide content [112]. Some materials such as rocks and minerals even fluoresce or emit visible light when illuminated by UV radiation [74].

An excellent review and discussion—including physical and biological factors—of the spectral reflectance characteristics of various common Earth surfaces (including vegetation, soils, water, snow and clouds) can be found in [46], as well as in the *Manual of Remote Sensing* [92]. Finally, a short but useful discussion on the *thermal* signatures of different surfaces (such as surface temperatures or forest fires) is given in [93].

2.2.6 Remote sensing platforms

Remote sensing can be carried out from a number of sensor locations using a variety of platforms. In many ways the sensors often have similar characteristics, but their different altitudes and stability characteristics can lead to very different observations [93].

2.2.6.1 Ground-based, airborne, and spaceborne platforms

There are three main types of remote sensing platforms, namely *ground-based*, *airborne* and *spaceborne* platforms, of which airborne and spaceborne platforms are arguably the most important.

Ground-based platforms may be used to record detailed information about a particular surface for comparison with information collected from airborne or spaceborne sensors [74]. However, the relatively small surface areas which can be observed by ground-based sensors make their widespread use both expensive and impractical.

Airborne platforms include balloons, aircraft, and more recently, Unmanned Aerial Vehicles (UAVs), whereas spaceborne platforms are predominantly satellites, although some sensors can also be found on spacecraft and the International Space Station (ISS).

Optical (passive) and microwave (active) remote sensing systems can be found on both

airborne and spaceborne platforms. For example, the most common type of (active) radar imaging is known as Side-Looking Airborne Radar (SLAR), and is found on many airborne platforms. However, a modified version - which uses the movement of its spaceborne platform to create an artificial antenna - is known as Synthetic Aperture Radar (SAR). Optical remote sensing systems include, for example, the MODIS sensor found on the (spaceborne) Aqua and Terra satellite platforms, or the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) hyperspectral sensor on airborne platforms.

2.2.6.2 Remote sensing instruments (sensors)

There is a wide variety of sensors currently in use on many remote sensing platforms, and each one is typically designed for a specific purpose or application.

For example, thermal sensors may be suitable to monitor the surface temperature on the Earth, or they can perhaps be used to detect forest fires or other thermal events provided that the temporal resolution is high enough. Optical sensors are commonly used to perform land cover classification, or to determine the relative health of vegetation. However, cloud cover and other aerosols in the atmosphere sometimes make it impossible to use optical sensors. In such cases, radar imaging can be used very effectively, since clouds, fog and smoke do not really affect radiation in the microwave range.

An important characteristic of a remote sensing system is its resolution. The resolution of a remote sensing instrument can be expressed in terms of its *spectral*, *spatial*, *temporal* and *radiometric resolution*, described next.

2.2.6.2.1 Spectral resolution. The *spectral resolution* of a sensor refers to the width or range of each spectral band being recorded. As an example, a panchromatic sensor (which records a broad range of visible wavelengths) will not be as sensitive to vegetation stress as a narrow band in the red wavelengths, where chlorophyll strongly absorbs electromagnetic energy.

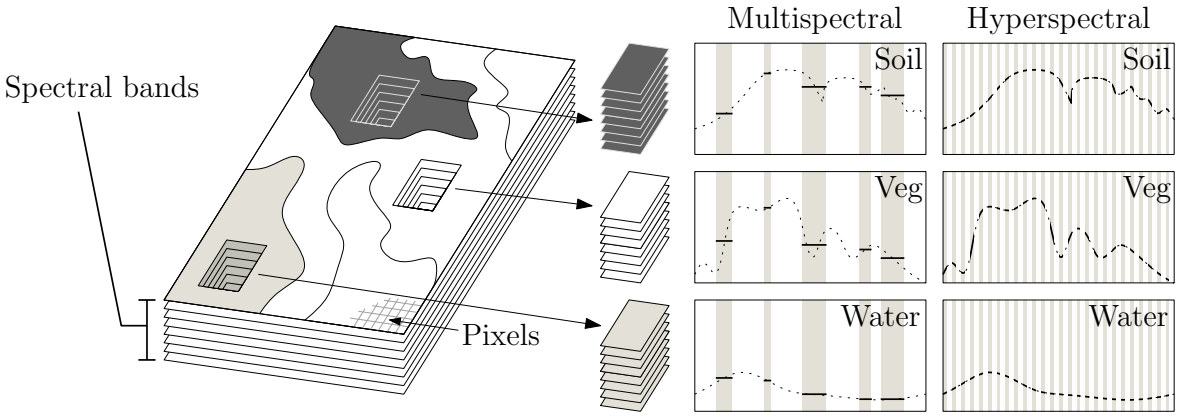


Figure 2.9: Interaction of reflected light with surface materials, showing multispectral and hyperspectral signatures (adapted from [13, p. 47]).

Figure 2.9 illustrates the difference between hyperspectral sensors (which are characterised by high spectral resolution across many contiguous bands, typically 1–15 nm wide), and multispectral sensors, where the spectral resolution is typically lower than for hyperspectral sensors (with typical bandwidths of 50–150 nm per band), and where gaps between the different spectral bands are common.

Although a high spectral resolution may be necessary to distinguish between closely related surface features, it comes at the cost of increased system complexity.

2.2.6.2.2 Spatial resolution. Higher *spatial resolutions* enable us to discern smaller details in ground objects. The spatial resolution (of passive sensors at least) depends primarily on the Instantaneous Field Of View (IFOV) of the sensor, 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.

One might reasonably expect that the higher the spatial resolution the better, but this is not always the case [112]. Increasing the spatial resolution by narrowing the instrument's IFOV means that less energy is received by the sensor (since the area from which energy is collected is smaller), and so the Signal-to-Noise Ratio (SNR) is effectively decreased. Even though such a decrease in the SNR can be mitigated by increasing the scanning bandwidth, this will cause a reduction in the spectral resolution.

2.2.6.2.3 Temporal resolution. The *temporal resolution* of a remote sensing system refers to the time interval between images, although this is usually a characteristic of the remote sensing platform, and not the sensor itself (the sampling frequency or temporal resolution of the sensors themselves are usually much higher).

Several applications require a high temporal resolution, such as the detection of oil spills and forest fires, or the monitoring of sea ice motion [112]. Other applications require very low temporal resolutions, such as seasonal crop identification, the annual quantification of forest insect infestations, or the once-off mapping of geological structures.

The *effective temporal resolution* of an optical sensor is affected by atmospheric conditions such as cloud cover, which may obscure targets from view. In some areas of the world, particularly the tropics, this is virtually a permanent condition, and microwave remote sensing must be considered instead.

2.2.6.2.4 Radiometric resolution. The *radiometric resolution* of a remote sensing system refers the sensitivity of the sensor to small differences in electromagnetic energy within each spectral band, and is quantified by the number of bits. If a sensor has a radiometric resolution of 12 bits, for example (as is the case for the MODIS sensor), it can detect and store $2^{12} = 4096$ unique levels of radiation.

Remote sensing data is then also commonly expressed as a Digital Number (DN) ranging from 0 to $2^b - 1$, where b is the radiometric resolution of the sensor, in bits.

2.3 CLASSIFICATION OF REMOTELY SENSED DATA

There are essentially two classes of analytical techniques commonly used to create thematic or land cover maps, namely *photointerpretation* and *machine analysis* (from here onwards referred to simply as *classification*) [93, p. 361]. Photointerpretation typically relies on the use of image enhancement procedures for improving the visual interpretability of multispectral images which are then subsequently interpreted by a human analyst, whereas classification is usually based on statistical or other forms of numerical algorithms for labelling regions within multispectral datasets.

Photointerpretation and classification serve different purposes, but they are often complementary. In fact, photointerpretation is greatly facilitated by computer-based image enhancement, and many classification approaches are developed using spectral information or knowledge first derived from photointerpretation.

Table 2.3: A comparison between photointerpretation (by a human analyst) and machine analysis, from [93, p. 68].

Photointerpretation (by a human analyst)	**Classification or machine analysis (by computer)**
On a large scale relative to pixel size	At individual pixel level (or larger)
Inaccurate area estimates	Accurate area estimates possible
Only limited multispectral analysis	Can perform true multispectral analysis
Can discern only a limited number of distinct brightness levels	Can make use of all available brightness levels (e.g. 4096 in MODIS)
Shape determination is easy	Shape determination involves complex software decisions
Spatial information is easy to use in a qualitative sense	Limited techniques available for making use of spatial information

A comparison between the tasks of photointerpretation and classification is given in Table 2.3, from where it can be concluded that photointerpretation is best suited for spatial assessment, but poor in quantitative accuracy, whereas classification has poor spatial ability, but high quantitative accuracy [93]. The status and research priorities of land cover classification for large geographic areas were discussed in [21], while land cover classification approaches for remotely sensed data with medium spatial resolution were considered in [38]. More recently, [80] performed a comprehensive review of modern approaches to land cover classification which make use of machine learning techniques such as artificial neural networks, fuzzy-sets, and expert systems. In fact, most of the recent research effort in land cover classification has focused on the development of increasingly robust and accurate (and also increasingly complex) classifiers by constructing—often in an ad hoc manner—multispectral, multitemporal, multisource classifiers using these modern machine learning techniques.

In this present study, we will concern ourselves only with the classification (or machine analysis) task, and in particular, we will consider a supervised, per-pixel, semi-parametric[2], hard-decision, spectral classifier. These considerations are described next.

2.3.1 Classification approaches

In general, land cover classification approaches can be grouped as being supervised or unsupervised, per-pixel, subpixel, per-field or object-oriented, parametric or non-parametric, hard or soft (fuzzy), and spectral, contextual or spectral-contextual [80]. The grouping of land cover classification approaches is shown in Figure 2.10, and each of these categories is briefly described below.

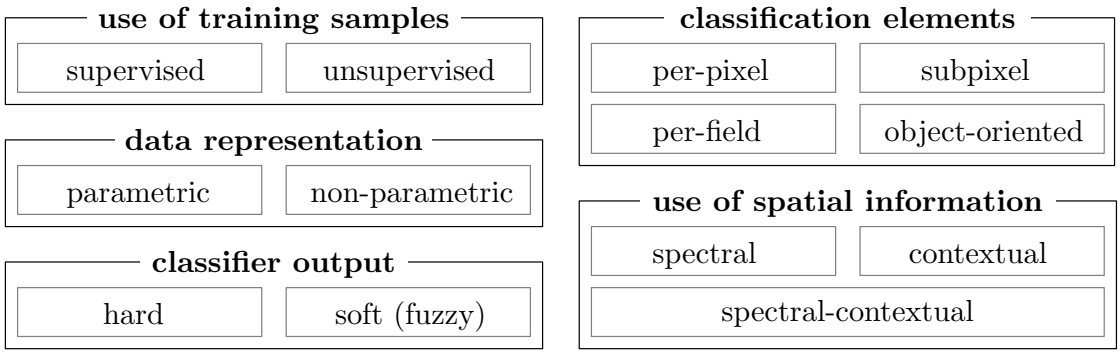


Figure 2.10: A taxonomy of fixed-sample-size land cover classification methods.

It should also be mentioned that there is no single best classification strategy, but instead the choice of classification approach depends on a number of factors, such as data availability, spatial and radiometric resolution, operating environment and purpose of classification, to name a few.

2.3.1.1 Use of training samples

Probably the most fundamental difference between various land cover approaches stems from the usage (or absence) of training data. Classification approaches in which training data is used to construct classifiers are referred to as *supervised* classification approaches, whereas those which do not make use of training data are referred to as *unsupervised* classification approaches. There is also a third, hybrid approach, called *reinforcement learning*, but it is not commonly used in the land cover classification context.

2.3.1.2 Classification elements

The most common approach to land cover classification is *per-pixel* classification, in which every pixel is classified as belonging to a specific land cover class. This approach assumes that every pixel is homogeneous, but obviously, as the spatial resolution is

[2]That is, we will consider a *statistical* model, but without the usual restriction that the data must be normally distributed.

decreased (i.e., in medium and coarse resolution images), this assumption is no longer realistic. For this reason, *subpixel* approaches have been developed, which attempt to take the mixed classes within each pixel into account, so that subpixelic regions are classified, rather than the entire pixel.

Unfortunately, high resolution data also comes with its own set of problems. In particular, when objects of interest are much larger than the pixel size, the set of corresponding pixels are rarely homogeneous, but instead exhibits so-called "salt and pepper" noise. By considering each object of interest instead of every individual pixel, much better classification accuracies are typically obtained. When the objects are described by vector data (such as in most GIS applications), the approach is called *per-field* classification, whereas it is known as *object-oriented* classification if the objects are described by raster data.

2.3.1.3 Data representation

Classifiers are considered to be either *parametric* or *non-parametric*. This distinction is unfortunately somewhat misleading, since *all* classifiers are necessarily parametric in some sense. For example, neural networks (which are fully characterised by a set of weights and an architecture) are considered non-parametric, even though the weights and architecture constitute the parameters describing the classifier. Historically though, parametric classifiers assume that a normally distributed dataset exists, and that the statistical parameters (i.e., mean and covariance) are representative of the dataset [80]. In this view, a classifier based on the assumption of an underlying Poisson distribution, for example, would be considered non-parametric, which is absurd. The distinction therefore really reduces to whether the data is "normally distributed" or not.

A better categorization might be to label classification approaches as being either *statistical* or *empirical* in nature. Statistical approaches would then include all approaches in which any assumptions have to be made about the statistical distribution of the underlying datasets, whereas empirical approaches would not need to make any such assumptions. Nevertheless, to facilitate this literature review we will keep with the historic interpretation of parametric and non-parametric approaches.

2.3.1.4 Classifier output

Classifiers can either provide *hard-decision* outputs, or *soft* (fuzzy) outputs. A hard decision only provides the class membership, whereas a fuzzy output provides a degree of confidence, along with class membership. The additional information provided by fuzzy classification approaches are often very important for more advanced procedures, such a when taking spatial (contextual) information into account.

Fuzzy approaches are also particularly well suited to subpixel classification, where the percentages of subpixel constituents might not be known exactly, and where a single pixel may need to be classified as a fuzzy mixture (linear combination) of several classes.

2.3.1.5 Use of spatial information

Spatial information, although difficult to use, can be incredibly beneficial to classification performance. Classification approaches that do not make explicit use of spatial information, but instead only use spectral information, are referred to simply as *spectral* classifiers. In the other extreme, classification approaches which make *exclusive* use of spatial information are called *contextual* approaches, but these are not used very often. Much more common though, are so-called *spectral-contextual* approaches, in which both spectral and spatial information are exploited.

2.3.2 Literature review: Land cover classification

A truly comprehensive survey of the land cover classification literature would be prohibitively large, especially since many of the published results present nothing new in terms of algorithmic improvements, but instead only give results for different datasets or study regions. Nevertheless, the 2007 review article by [80] (which does not consider common classification approaches such as maximum likelihood classification, ISODATA, or k-means clustering) already contains a staggering 376 references, most of which have used divergent classification strategies, as well as different datasets and study regions.

In addition, since land cover classification is essentially a special case of image classification, much of the literature concerning image processing and classification are also of relevance here. Nevertheless, we will restrict our attention to studies *directly* related to remote sensing (and land cover classification in particular), and from this reduced subset we will only consider the most important, innovative, or seminal contributions.

Of general importance though, is a summary of spectral reflectance characteristics of common Earth surfaces by [46], as well as early discussions on land cover classification [47, 68]. More recent treatments of land cover classification include [69] and as previously mentioned, the excellent review by [80].

Following roughly the approach presented in [80], we will organise the most important literature into the following categories: per-pixel, subpixel, per-field, contextual-based, dimensionality reduction, and multisource classification. Each of these categories may contain either supervised or unsupervised approaches (or both), with hard-decision or fuzzy outputs.

2.3.2.1 Per-pixel algorithms

Per-pixel classification is probably the most common type of land cover classification, and definitely the oldest. Most supervised per-pixel classifiers develop a *spectral signature* of each land cover class by combining the spectra of all the corresponding training-set pixels [80]. Of course, such a training strategy necessarily assumes that the every pixel in the training set is homogeneous.

Per-pixel classifiers can be either parametric or non-parametric, and one of the most widely applied (parametric) classification approaches is the *maximum likelihood* classifier, in which the data is assumed to be normally distributed. The maximum likelihood classifier is discussed in almost every textbook on remote sensing, and the interested reader is referred to [93, Chapter 8]. Obvious limitations to the maximum likelihood classifier include the difficulty of incorporating ancillary data from other (often non-statistical) sources, as well as the requirement that data must be normally distributed.

An interesting extension to the maximum likelihood classifier involves multispectral classification using probability density functions, where the underlying data distributions are considered to be *mixtures of Gaussians* [31, 32]. Both per-pixel and per-field approaches are considered.

The assumption of normally distributed spectral classes is of course not always realistic, and so many alternative (non-parametric) methods for per-pixel classification have since been considered. Among the most important of these are *artificial neural networks* (see for example [87]), *decision trees* (first described in [108]), *support vector machines* [37, 49], and expert systems. Since no statistical parameters are used to describe the data, parametric classifiers are especially suitable for the incorporation of non-spectral data into a classification procedure [80].

A comprehensive review of the use of multilayer perceptron neural networks in remote sensing (with specific reference to the back propagation learning rule) is given in [87], where it is also shown that such non-parametric approaches typically provide better classification accuracy than maximum likelihood classifiers. Bagging, boosting, or a hybrid of both techniques may also be used to further improve the classification performance of non-parametric classifiers [5], and it has been used successfully for this purpose in both decision trees [70] as well as support vector machines [63].

Additional advanced algorithms which have found recent application in land cover classification (that is, within the last decade) include the *spectral angle classifier* [105], which has both supervised and unsupervised variants, unsupervised classification based on *independent component analysis* (ICA) as presented in [72] and [102], a *model-based approach* to unsupervised classification [65, 66], and several *nearest-neighbour* approaches [24, 43], and [42].

2.3.2.2 Subpixel algorithms

Subpixel approaches are particularly suited to medium or coarse resolution remote sensing data, where heterogeneous pixels are very common. In fact, the heterogeneity of such pixels has been recognised as a major source of error in coarse-resolution per-pixel classifications [25, 34]. Two of the most popular approaches to subpixel classification include *fuzzy-set* techniques (see for example [35] or [125]), and *spectral mixture analysis* [1, 79, 95], where it is usually assumed that the spectrum of each pixel can be described as a linear combination (or mixture) of the spectra of its constituent land cover types [1].

Other noteworthy approaches to subpixel classification include artificial neural networks (a supervised neuro-fuzzy approach is compared to a hard (non-fuzzy) approach in [82]), Dempster-Shafer theory, certainty factors [11], a maximum likelihood approach [101] and IMAGINE's subpixel classifier [53]. More recently, *a priori*-known structural information from high resolution imagery was also taken into account in a spectral mixture analysis context in [96] and [97], where both supervised and unsupervised methods were presented.

Land cover *area estimation* is another important task which is negatively affected by low spatial resolution, and subpixel methods typically provide a more appropriate representation and accurate area estimate for different land covers than per-pixel approaches [36, 94, 123].

The objective and consistent evaluation of subpixel classification approaches remains a challenge however, since the confusion matrix approach is clearly not adequate to describe fuzzy classification results. Various other performance measures have been proposed, such as the Kappa coefficient, conditional entropy and mutual information [80]. Nevertheless, it is often very difficult to collect sufficient reference data to evaluate fuzzy classification approaches [80].

2.3.2.3 Per-field algorithms

Environmental heterogeneity leads to large intra-class variability within land cover classes, which at high spatial resolutions, may result in very noisy image data (at the pixel level). At lower spatial resolutions the pixels are not as noisy, but may be composed of several land cover types, as described in section 2.3.2.2.

One approach to deal with the environmental heterogeneity at higher spatial resolutions is to consider so-called *per-field classification* (see for example, [3, 77]), in which the goal is no longer to classify individual (noisy) pixels, but rather environmental "fields", which should preferably be large relative to the pixel size. In this way, the noise is averaged out over larger areas, such that fields within the same land cover class are much more homogeneous than their constituent pixels would suggest.

In per-field classification, images are first subdivided into fields using vector data from Geographic Information Systems (GIS), after which classification is performed [54]. Therefore, factors such as the size and shape of the fields, the definition of field boundaries, and the land cover classes chosen affect the performance of per-field classifiers. It can also be difficult to integrate vector and raster (pixel-based) data, so that the related technique of *object-oriented classification* may sometimes be preferred.

In object-oriented classification, "objects" consisting of groups of pixels are identified during the image segmentation phase, similar to fields in per-field classification. The primary difference between object-oriented and per-field classification is then that object-oriented methods only use raster data, whereas per-field methods employ both

raster and vector data. Object-oriented classification has been shown to exhibit superior classification accuracy as compared to per pixel methods, especially for high spatial resolution data (see [10], [41] and [120], and the referenced contained therein). One of the most commonly used methods for object-oriented classification is known as eCognition, as described in, for example [121].

2.3.2.4 Contextual-based approaches

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 [81]. A selection of early (and ad hoc) contextual-based classification methods are compared in [83], and a number of statistical methods for contextual-based classification has been developed by [64, 109], and [62]. However, more recently, Markov and Gibbs random fields (proposed in [39]) have been shown to be effective methods to take the spatial information from images into account.

One of the earliest treatments of the Markov random field approach in the remote sensing context is given by [56], while [106] have successfully used the approach to incorporate both spatial and temporal information. The Markov random field approach was also used in [58], where it has been shown that the incorporation of spatial information resulted in improved classification accuracy.

There are also spectral-contextual classifiers, which exploit both spectral and contextual information ([107]), contextual-based classifiers for high and low resolution data ([60]), hierarchical maximum *a posteriori* classification approaches ([50]), and many more. For a brief review of some of the many approaches to contextual-based classification, the reader is again referred to [80].

2.3.2.5 Dimensionality reduction

The efficient estimation of statistical parameters requires a representative sample with enough data points; consequently, as the number of parameters is increased (with the number of data points fixed), the estimation efficiency is decreased. The effectiveness of a classifier will therefore begin to decrease once a certain number of dimensions is reached [52], since it quickly becomes impractical to collect and use correspondingly larger datasets. The relationship between dimensionality and the training sample size must therefore be kept in mind when designing classifiers with many parameters [112].

A powerful technique to reduce the requirement of an excessively large training dataset is *dimensionality reduction*, in which unimportant dimensions are discarded, or new coordinate axes are sought which can capture the important data more compactly (i.e., with fewer dimensions). Dimensionality reduction is very closely related to *feature selection* or *extraction* (which itself might be a form of dimensionality reduction), but it serves a slightly more general purpose, including for example dimensionality reduction for data visualisation and connectivity analysis. Spectral connectivity analysis is a

technique for transforming data into a coordinate system that efficiently reveals the geometric structure (and the "connectivity", in particular) of the data. Such an analysis can be extremely beneficial to better understand the complex interrelationships between various spectral bands and surface types.

Several approaches to dimensionality reduction exist, including principal component analysis, minimum noise fraction transform, discriminant analysis [30, 84, 86], decision boundary feature extraction [9], Gaussian mixture model feature extraction [65], wavelet transform [113], and spectral mixture analysis [80, 91]. A comparison between multispectral and hyperspectral observations of vegetation, soil and dry carbon cover in arid regions is given in [4], from which it is also evident that dimensionality reduction is especially important when hyperspectral (e.g. AVIRIS) or multisource data is used.

Finally, after dimensionality reduction has been performed, classification proceeds as per usual, in which one of the approaches presented earlier, or a combination of some of those can be used in a multisource classifier context, described next.

2.3.2.6 Recent trends and state of the art: Multisource, multitemporal, object-oriented classification

Based on a review of 15 years of peer-reviewed publications on land cover classification, [122] claimed that no appreciable improvement in land cover classification accuracy could be observed, even though many new classification methods were proposed and implemented during this period. The reason for this apparent failure, according to [55], is that most researchers still attempt to improve classification accuracy using only spectral information, which on its own is insufficient for reliable land cover classification, irrespective of which classifier is used.

State-of-the-art classifiers therefore incorporate ancillary data from multiple sources, and exploit as much of the information contained in the data as possible. This would include, for instance, the spatial, temporal, spectral and structural information in remote sensing data, and whichever useful information can be extracted from the ancillary sources of information.

The observed trend in the land cover classification literature is then clearly to move away from traditional pixel-based techniques [76], and to focus instead on contextual, object-oriented methods which take the spatial and structural information into consideration.

An unsupervised multisource classification approach using Dempster-Shafer evidence theory is described in [71], and [15] provides a comparative study of statistical methods (modified to handle ancillary information through prior probabilities) and neural networks for multisource classification [80].

Even though contemporary classifiers try to take as many factors into consideration as possible, it seems as though the temporal resolution of many (in fact most) modern

classifiers is still underutilised.

Some of the advantages of considering a time-varying approach to land cover classification were recently pointed out in [44], where a so-called "multi-seasonal" discriminant analysis strategy was followed for large-scale land cover classification. More specifically, four Landsat7 ETM+ images corresponding to March, May, June and September were used for classification: initially the classifier only used the image corresponding to March, then March and May, and so on, and the improvement in classification accuracy was noted each time an additional image was added to the classification task. Other multitemporal approaches have also been considered, such as the neural-statistical, multitemporal, multisource approach presented in [16]. However, remote sensing data is often *hypertemporal* (i.e., characterised by a fine temporal resolution), such that true time series analysis is expected to become very important in the near future.

2.4 COARSE RESOLUTION LAND COVER CLASSIFICATION: A FIRST EXAMPLE

In this first example, three minimum distance land cover classifiers will be designed and compared on coarse resolution MODIS surface reflectance data for a two-class classification problem. It will be shown that good class separability can be achieved using only the seasonal component of NDVI, or the mean component of several other MODIS land bands. It will also be shown why particular classifiers might be expected to fail in certain spectral bands. Finally we will give an initial analysis of the usefulness of the various MODIS bands, in which we will show that after NDVI, band 2 is the most separable, and band 5 the least separable of all the MODIS land bands.

2.4.1 Motivation

Our motivation for including this first set of land cover classification results here and not together with the rest of the results (contained in Chapter 5), is that this section serves as (i) an introduction to the study area and dataset, and (ii) the method presented here does not rely on a statistical land cover model as assumed in the rest of this dissertation. In fact, the preliminary data analysis and visualisation presented here provided the necessary motivation for the development of the statistical models presented in section 4.2.

2.4.2 Problem statement

Supervised learning is the most widely used technique for land cover classification of remote sensing images. Approaches such as artificial neural networks and decision trees are widely used to perform multispectral (and often multi-source) classification, and generally exhibit superior performance to single spectrum, single-source classification. Nevertheless, the analysis of single band, single-source classifiers can be very useful to better understand (and predict) the performance of more complicated classifiers, and it can also be used to infer an empirical lower bound for classification accuracy.

Fourier—or spectral—analysis (on NDVI data in particular) has been used extensively for land cover classification (see for example [57] and [75]), and it has been noted that reliable class separation can be achieved even when considering only the mean and seasonal spectral components [75]. When concerned with the classification of either vegetation or residential areas (which are characterised by both buildings and vegetation), we will show that reliable class separation can in fact be achieved using *only* the seasonal component of NDVI, or the mean component of several other spectral bands from multitemporal coarse resolution MODIS data.

Our task is therefore to design and compare three minimum distance classifiers using single spectral bands from a coarse resolution MODIS surface reflectance time series. Furthermore, it should be established whether reliable class separation is in fact possible when using only the seasonal component of NDVI, and the relative usefulness of the various MODIS land bands should be determined for our particular classification task.

2.4.3 Moderate resolution imaging spectroradiometer

Throughout this study we will use coarse resolution MODIS time series data to perform land cover classification. However, to explain why we specifically decided to use MODIS, a short overview of the sensor characteristics in comparison to other commonly used sensors such as AVHRR is first described in this section.

Multispectral data from the AVHRR on board the National Oceanographic and Atmospheric Administration (NOAA) satellite has been used for global land cover classification using monthly composites describing seasonal variations in the photosynthetic activity of vegetation [26]. However, AVHRR was not specifically designed for land applications, and is characterised by a lack of radiometric calibration, poor geometric correction and registration, and inadequate cloud screening. Consequently, it was found that data obtained from AVHRR is insufficient to distinguish subtle differences in vegetation types with similar annual phenologies [12]. Nevertheless, AVHRR was, until recently, one of the only sources from which global NDVI data could be derived.

Other remote sensing systems which *had been* designed for land applications such as the Landsat Thematic Mapper (TM), and which recorded data at a higher spectral resolution than AVHRR, were not particularly useful for global (or even regional) land cover classification either, since they were characterised by incomplete spatial coverage, low temporal resolution, and inevitable cloud contamination [26].

The MODIS sensors on board the Aqua and Terra satellites of NASA's Earth Observing System (EOS) series of missions were designed to provide increased spectral and radiometric resolution (including accurate on-board radiometric calibration), global geographical coverage (at a spatial resolution of between 250 m and 1 km), and improved geometric and atmospheric corrections, while preserving a temporal resolution comparable to that of AVHRR [18, 111]. Therefore, data obtained from the MODIS sensor provides a substantially improved basis for studying ecosystem processes [126].

Table 2.4: MODIS sensor characteristics (see http://modis.gsfc.nasa.gov).

Band	**Wavelength (μm)**	**IFOV (m) [at nadir]**	**Primary use**	**Spectral region**
Band 1	0.62–0.67	250 × 250	Land/Cloud/Aerosols Boundaries	Visible (Red)
Band 2	0.841–0.876	250 × 250	Land/Cloud/Aerosols Boundaries	Near IR
Band 3	0.459–0.479	500 × 500	Land/Cloud/Aerosols Properties	Visible (Blue)
Band 4	0.545–0.565	500 × 500	Land/Cloud/Aerosols Properties	Visible (Green)
Band 5	1.230–1.250	500 × 500	Land/Cloud/Aerosols Properties	Short Wave IR
Band 6	1.628–1.652	500 × 500	Land/Cloud/Aerosols Properties	Short Wave IR
Band 7	2.105–2.155	500 × 500	Land/Cloud/Aerosols Properties	Short Wave IR
Band 8	0.405–0.420	1000 × 1000	Ocean Colour/Phytoplankton/ Biogeochemistry	Visible (Blue)
Band 9	0.438–0.448	1000 × 1000		Visible (Blue)
Band 10	0.483–0.493	1000 × 1000		Visible (Blue)
Band 11	0.526–0.536	1000 × 1000		Visible (Green)
Band 12	0.546–0.556	1000 × 1000		Visible (Green)
Band 13	0.662–0.672	1000 × 1000		Visible (Red)
Band 14	0.673–0.683	1000 × 1000		Visible (Red)
Band 15	0.743–0.753	1000 × 1000		Near IR
Band 16	0.862–0.877	1000 × 1000		Near IR
Band 17	0.890–0.920	1000 × 1000	Atmospheric Water Vapour	Near IR
Band 18	0.931–0.941	1000 × 1000	Atmospheric Water Vapour	Near IR
Band 19	0.915–0.965	1000 × 1000	Atmospheric Water Vapour	Near IR
Band 20	3.660–3.840	1000 × 1000	Surface/Cloud Temperature	Mid Wave IR
Band 21	3.929–3.989	1000 × 1000	Surface/Cloud Temperature	Mid Wave IR
Band 22	3.929–3.989	1000 × 1000	Surface/Cloud Temperature	Mid Wave IR
Band 23	4.020–4.080	1000 × 1000	Surface/Cloud Temperature	Mid Wave IR
Band 24	4.433–4.498	1000 × 1000	Atmospheric Temperature	Mid Wave IR
Band 25	4.482–4.549	1000 × 1000	Atmospheric Temperature	Mid Wave IR
Band 26	1.360–1.390	1000 × 1000	Cirrus Clouds Water Vapour	Near IR
Band 27	6.535–6.895	1000 × 1000	Cirrus Clouds Water Vapour	Mid Wave IR
Band 28	7.175–7.475	1000 × 1000	Cirrus Clouds Water Vapour	Long Wave IR
Band 29	8.400–8.700	1000 × 1000	Cloud Properties	Long Wave IR
Band 30	9.580–9.880	1000 × 1000	Ozone	Long Wave IR
Band 31	10.780–11.280	1000 × 1000	Surface/Cloud Temperature	Long Wave IR
Band 32	11.770–12.270	1000 × 1000	Surface/Cloud Temperature	Long Wave IR
Band 33	13.185–13.485	1000 × 1000	Cloud Top	Long Wave IR
Band 34	13.485–13.785	1000 × 1000	Cloud Top	Long Wave IR
Band 35	13.785–14.085	1000 × 1000	Cloud Top	Long Wave IR
Band 36	14.085–14.385	1000 × 1000	Cloud Top	Long Wave IR

Terra MODIS and Aqua MODIS take between one and two days to cover the entire Earth's surface, with a complete 16-day repeat cycle. Each of the MODIS sensors spans a broad range of the electromagnetic spectrum by 36 narrow spectral bands (ranging from 0.459 to 14.385 μm). It has been shown that such narrow spectral bands can lead to improved classification of vegetation characteristics [110]. Their spatial resolution is 250 m for bands 1 and 2, 500 m for bands 3 to 7, and 1 km for bands 8 to 36. Refer to Table 2.4 for a summary of the 36 spectral bands, including their spatial and spectral resolutions, and primary design purposes. Each of these spectral bands is also calibrated by a Solar Diffuser (SD) and a Solar Diffuser Stability Monitor (SDSM) system [124].

A comparison between the AVHRR and MODIS sensor characteristics is given in Table 2.5, in which MODIS-N refers to a nadir-mounted sensor, and MODIS-T to a tilting sensor (i.e., one which can take off-nadir measurements). Notice that the bandwidths of AVHRR range between 100 nm and roughly 400 nm, whereas for MODIS the bandwidths range between 20 nm and 50 nm.

Table 2.5: Comparison of the principal sensor characteristics of MODIS and AVHRR for land cover classification (from [111]).

	AVHRR	MODIS-N		MODIS-T
		Center	Bandwidth	
Spectral bands for land cover applications	580–680 nm 725–1100 nm 1580–1750 nm	470 nm 555 nm 659 nm 865 nm 1240 nm 1640 nm 2130 nm	20 nm 20 nm 20 nm 40 nm 20 nm 20 nm 50 nm	32 bands (400–880 nm)
	3 thermal bands	9 thermal bands		
IFOV (nadir)	1.1 km	500 m 250 m (659 and 865 nm)		1.1 km
Swath width	2700 km	2330 km		1500 km
Calibration	absent	lunar		lunar, solar
Radiometric quantization	10 bit	12 bit		12 bit
Global frequency	1–2 days	1–2 days		2 days
View angle	55.4°	55°		45°
Tilt capability	none	none		±50°

We can also see from Table 2.5 that AVHRR has a radiometric resolution of 10 bits or 1024 levels, whereas MODIS has a 12 bit radiometric resolution, corresponding to 4096 unique levels of radiation.

Notice further that NDVI derived from AVHRR contains a small amount of radiation in the visible green wavelengths, in which the reflectance for vegetation is typically higher

than in the visible red wavelengths. However, since 80 % of the "red" band corresponds to the visible red range, the contribution of the green wavelengths is expected to be relatively small. Nevertheless, with MODIS, we have narrow spectral bands centred in the red and near-infrared regions, as required for the derivation of NDVI.

Table 2.6: MODIS land products.

Product code	Platform	Description
Radiation balance product suite		
MOD09/MYD09	Aqua/Terra	Surface reflectance
MOD11/MYD11	Aqua/Terra	Surface temperature and emissivity
MOD43/MYD43/MCD43	Aqua/Terra/Combined	BRDF/Albedo
Vegetation product suite		
MOD13/MYD13	Aqua/Terra	Vegetation indices
MOD15/MYD15/MCD15	Aqua/Terra/Combined	Leaf Area Index – FPAR
MOD17/MYD17	Aqua/Terra	Gross primary productivity
Land cover product suite		
MOD12/MCD12	Aqua/Combined	Land cover type
MOD14/MYD14	Aqua/Terra	Thermal anomalies and fire
MOD44	Aqua	Vegetation continuous fields

The data acquired by the MODIS sensor is used to generate multiple (land) products at different preprocessing stages [18], as listed in Table 2.6, and we will specifically use the MCD43A4 product (see [100] for details), which is a level 4, eight-daily composite of 500 m, Bidirectional Reflectance Distribution Function (BRDF)-corrected surface reflectance data, available from the MODIS data product website.

2.4.4 Study area & data description

Two classes of land cover type, namely *residential* and *natural vegetation* is considered in this study. Every pixel within each class has eight associated time series, with observations every eight days. The first seven time series correspond to the seven MODIS land bands, while the 8th time series corresponds to the derived NDVI.

More specifically, the time series dataset was obtained from the MODIS MCD43A4 BRDF-corrected 500 m land surface reflectance product corresponding to a total area of approximately 230 km^2 (sampled from a region of roughly 17,000 km^2) in Gauteng, South Africa (26.12°S, 28.08°E). This area is shown in Figure 2.11.

The dataset consists of 925 MODIS pixels – identified by means of visual interpretation of high resolution Landsat and SPOT images between 2000 and 2008 – each containing eight time series (seven MODIS bands, and NDVI), with $N = 368$ observations, which

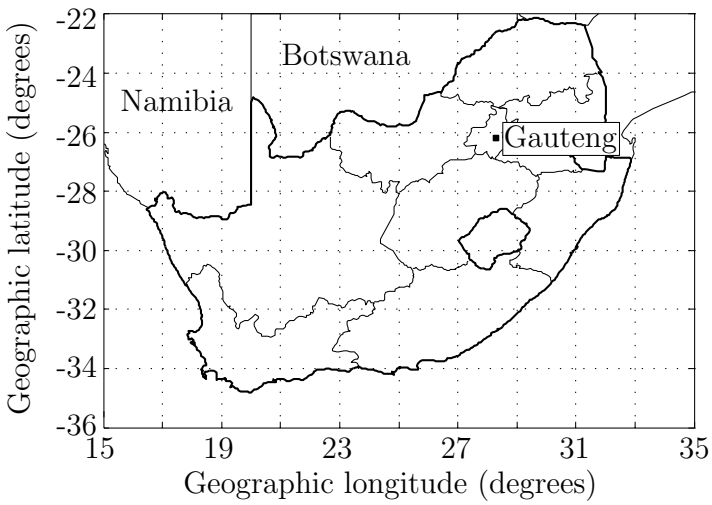


Figure 2.11: Experimental study area: Gauteng, South Africa.

correspond to just more than 8 years of 8-daily observations. The dataset is divided into two classes, namely residential (consisting of 333 pixels) and natural vegetation (consisting of 592 pixels).

Generally speaking, the residential class contains pixels consisting of about 50% buildings, and 50% vegetation, whereas the vegetation class contains pixels with more than 90% vegetation. However, the two classes are characterized by very large intra-class variability, since no distinction was made between different types of buildings or vegetation.

The DN corresponding to the surface spectral reflectance of the MODIS data has a radiometric resolution of 12 bits, which gives 4096 unique levels ranging from 0 to 4095. When presenting the data in this study, we will either give the DN directly, or we will express the data in terms of % reflectance, computed as

$$\%\ \text{reflectance} = \frac{\text{DN}}{4095}. \tag{2.4}$$

2.4.5 Minimum distance classification

We will now consider the task of deciding between two simple statistical hypotheses, $\mathcal{H}_0$ and $\mathcal{H}_1$, using information from only one single spectral band at a time. Suppose we have a sequence $\{x[n;\boldsymbol{\theta}]; n = 1, 2, \ldots\}$ of real-valued observations generated by one of two statistical hypotheses:

$$\mathcal{H}_0 \quad : x\Big[n; \boldsymbol{\theta} = (\text{band, residential})\Big], \quad n = 1, 2, \ldots$$

versus

$$\mathcal{H}_1 \quad : x\Big[n; \boldsymbol{\theta} = (\text{band, vegetation})\Big], \quad n = 1, 2, \ldots$$

where the two classes (residential and vegetation) are as described in section 2.4.4, and $\boldsymbol{\theta}$ = (spectral band, class) is a parameter vector describing the time series $x[n;\boldsymbol{\theta}] = x[n]$.

We will further assume that the observations $x[n]$ consist of an underlying (noise-free) signal $y[n]$, corrupted with some zero mean noise process $w[n]$ as follows:

$$x[n] = \begin{cases} y_0[n] + w[n] & \text{under } \mathcal{H}_0 \text{ (residential)}, \\ y_1[n] + w[n] & \text{under } \mathcal{H}_1 \text{ (vegetation)}. \end{cases}, \tag{2.5}$$

where the dependence on $\boldsymbol{\theta}$ has been made implicit for notational simplicity. Here the noise process is also expected to account for intra-class variability, so that consecutive noise samples are assumed to be statistically dependent.

The *ensemble average* or signal model, $s[n]$, is determined for each spectral band (including NDVI) at every observation period during the year:

$$s[n;\boldsymbol{\theta}] = \mathbb{E}\Big\{x[n;\boldsymbol{\theta}]\Big\}, \quad n = 1, \ldots, 45, \tag{2.6}$$

where the dependence on $\boldsymbol{\theta}$ has been made explicit. Of course, with $k = 45$ observation periods in a year, we would have $s[n+k] = s[n]\ \forall n$, since it is assumed that the underlying signal is periodic with a period equal to one year.

The model error, $\epsilon[n]$, is defined as the difference between the observed signal $x[n]$, and the signal model $s[n]$ as shown in Figure 2.12. The minimum distance classifier then simply chooses the class which minimizes this error, or more generally, a function of this error. That is, $\boldsymbol{\theta}$ is chosen such that $f(\epsilon[n;\boldsymbol{\theta}]) = f(x[n;\boldsymbol{\theta}] - s[n;\boldsymbol{\theta}])$ is a minimum.

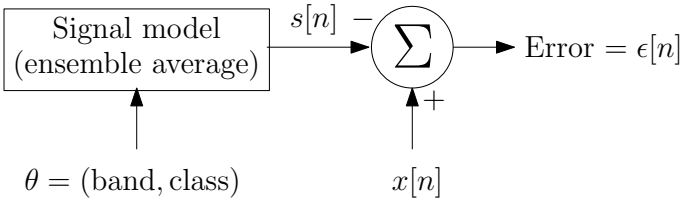


Figure 2.12: Error estimation within a single spectral band, for a particular class.

As an initial analysis of the separability of the time series data we consider three simple distances related to the error, namely the sum of errors (SoE), the sum of absolute errors (SAE), and the sum of squared errors (SSE):

$$\text{SoE}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \Big(x[n] - s[n]\Big) = \sum_{n=1}^{N} \epsilon[n], \tag{2.7}$$

$$\text{SAE}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \Big|x[n] - s[n]\Big| = \sum_{n=1}^{N} \Big|\epsilon[n]\Big|, \tag{2.8}$$

$$\text{SSE}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \Big(x[n] - s[n]\Big)^2 = \sum_{n=1}^{N} \epsilon^2[n]. \tag{2.9}$$

Each corresponding classifier then simply selects the class which minimises the appropriate distance. Note that even though the SoE given in (2.7) might seem a rather natural choice to characterize the error, it is not formally a metric (it can be negative, for example). On the other hand, the SAE in (2.8) is simply the L_1-norm, and the SSE in (2.9) is the square of the L_2-norm on the model error, ϵ.

2.4.6 Experimental results

The observed time series, $x[n]$, of two randomly selected sample pixels is shown in Figure 2.13 for a duration of one year, from which we might already expect band 5 to exhibit poor separability between the two classes under consideration.

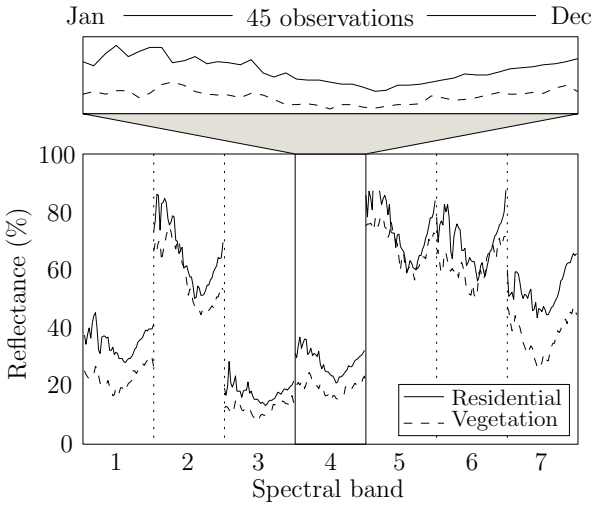
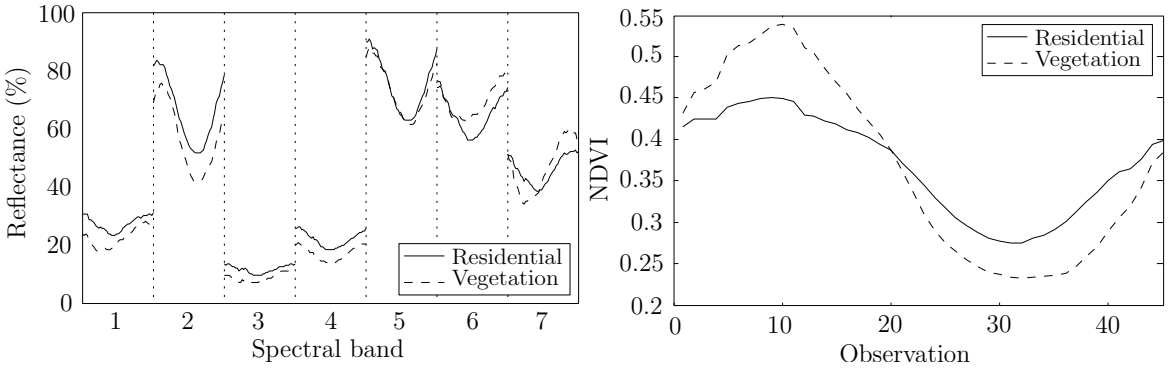


Figure 2.13: Two sample pixels: $x[n;\boldsymbol{\theta}], n = 1, \ldots, 45$; across all seven spectral bands.

The ensemble averages, $s[n]$, were approximated empirically, and are shown in Figure 2.14.a (for the seven land bands), and in Figure 2.14.b for NDVI.



a. Annual ensemble averages for the land bands.

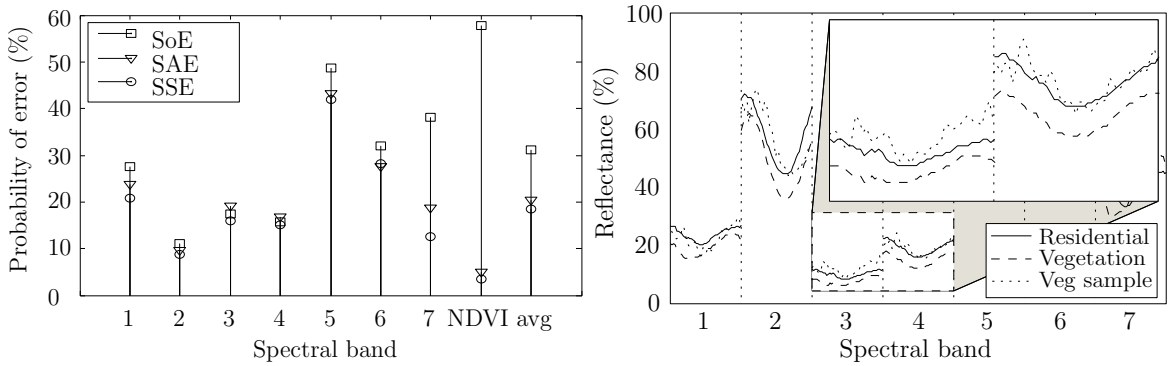
b. Two class ensemble average: NDVI.

Figure 2.14: Annual ensemble averages, $s[n;\boldsymbol{\theta}]$, for all the MODIS spectral bands.

With reference to Figure 2.14.a, we can see that bands 1, 2, 3, 4 and 6 should be easily separable by considering only the mean component of each particular band, whereas it seems very difficult to distinguish between the two classes using band 5. However, band 7, as well as NDVI shown in Figure 2.14.b, will not be separable when considering only the mean component, but should be separable when using only the seasonal (amplitude) component.

We can therefore already expect a classifier based on the SoE given in (2.7) to fail in band 7 and NDVI, since it cannot differentiate well between different amplitudes

(the positive and negative error contributions cancel each other out). This behavior is confirmed in Figure 2.15.a, in which all the classifiers have similar performance, except in band 7 and NDVI.



a. Classifier performance, $N = 368$. b. Classification errors in bands 3 and 4.

Figure 2.15: Minimum distance classifier performance comparison and examples of failure modes.

We also see that the SSE classifier performs the best overall, and especially when using NDVI, where an error probability of only 7% was obtained.

Of course classification errors are unavoidable as shown in Figure 2.15.b, where no classifier can be expected to correctly classify the vegetation sample in bands 3 and 4. This is partly due to the large intra-class variability, and situations like these motivate the development of more sophisticated classifiers.

The confusion matrices for all three classifiers, across all the bands (including NDVI) is given in Table 2.7 (rounded to the nearest integer). We can see that the SSE classifier consistently does better than the other classifiers, and after NDVI, band 2 has the highest classification accuracy.

Table 2.7: Confusion matrices (in %) for minimum distance classifiers. Top left: true positive (vegetation), top right: false positive, bottom left: false negative, bottom right: true negative (residential). V = vegetation, R = residential.

		Band 1		**Band 2**		**Band 3**		**Band 4**		**Band 5**		**Band 6**		**Band 7**		**NDVI**	
		V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R
SoE	V	80	35	90	13	89	24	90	21	54	51	65	29	53	29	47	62
	R	20	65	10	87	11	76	10	79	46	49	35	71	47	71	53	38
SAE	V	87	35	91	10	91	29	91	24	56	43	70	25	85	23	99	09
	R	13	65	09	90	09	71	09	76	44	57	30	75	15	77	01	91
SSE	V	87	29	91	09	91	23	90	20	58	42	69	26	90	15	97	04
	R	13	71	09	91	09	77	10	80	42	58	31	74	10	85	03	96

Previous research has found that adequate classification accuracy could be achieved by considering several Fast Fourier Transform (FFT) components – especially the mean and seasonal components, since they tend to carry the majority of the signal energy [57, 75]. In this section we have shown that when using NDVI, the seasonal component alone allows for good class separability, whereas in most other MODIS land bands the mean component alone provides useful information.

It was also shown that the SSE classifier exhibits the highest classification accuracy, particularly when using NDVI, and that after NDVI, band 2 seems to be the best band on which to perform classification. Band 5 on the other hand seems to be the most difficult band on which to perform classification.

A key underlying assumption of this first assumption was of course that the MODIS bands could be viewed in isolation, which is ultimately not entirely true. As a consequence, it would be interesting to investigate the *joint* behavior of the various spectral bands, which will enable the construction of much better classifiers than presented here. Such an approach is considered in Chapter 4.

2.5 SUMMARY

In this chapter we have presented the most pertinent principles of remote sensing in the optical region, with specific emphasis on the interaction between electromagnetic radiation, particulates in the atmosphere, and surfaces on the ground. Although this material is of general importance in many applications of remote sensing, we have limited our attention to the important task of land cover classification—the different approaches and common techniques for land cover classification were also discussed.

Our motivation for considering the MODIS MCD43A4 spectral reflectance product was also discussed briefly, where the main advantages included the on-board radiometric calibration system, the improved atmospheric, BRDF and geometric correction, as well as the high radiometric and temporal resolution of the MODIS sensor.

Finally, an initial example of land cover classification using the MODIS spectral reflectance time series data was given. In particular it was shown that reliable discrimination between residential and vegetation pixels is possible by considering only the mean or the annual components of the FFT, depending on which spectral band is used.

Chapter 3

Sequential Detection

> "*Use theory to provide insight; use common sense and intuition where it is suitable. A balance cannot be achieved unless one has both common sense and a facility with theory.*"
>
> *David Gries*, Associate Dean for Undergraduate Programs, College of Engineering, Cornell University

SEQUENTIAL DETECTION is an alternative to the classical, fixed-sample-size methods of statistical hypothesis testing, whose characteristic feature is that the number of observations required by the procedure is not predetermined. Instead, the decision to terminate the experiment depends, at each stage, on the results of the previous observations [118].

It is a well-known fact that the performance of a detector (in terms of accuracy) can be improved by increasing the signal-to-noise ratio. However, the noise power is usually fixed, so that the detector performance can only be improved by increasing the signal energy. In a remote sensing context, where satellite platforms typically have very limited resources on board, increasing the signal energy is particularly impractical and sometimes just plain impossible. Therefore, we must *increase the number of observations*. In most practical applications, it is desirable to minimise the number of observations that are needed to make a reliable (or accurate) decision. The theory of *sequential analysis* is used to formally treat—and solve—this challenging problem.

The so-called *sequential tests* (first introduced by Abraham Wald [118]) usually require significantly fewer observations than fixed-sample-size approaches with the same probability of error. In particular, sequential tests help the user make a decision between two hypotheses after a minimal average number of observations by deciding quickly in unambiguous cases, and taking longer in ambiguous cases [89].

3.1 INTRODUCTION

Sequential analysis (which includes the study of sequential detection problems) is a rich and vibrant area of research, which, according to [40], has its origins in the works of Huyghens, Bernoulli, DeMoivre and Laplace on the gambler's ruin problem [67]. Perhaps most importantly, however (at least for our purpose), was Wald's development of the Sequential Probability Ratio Test (SPRT) during the 1940s, which was subsequently proven to be optimal by Wald and Wolfowitz in 1948 [119].

Sequential detection problems between two simple statistical hypotheses (in which case the observation time is not fixed), give rise to so-called *optimal stopping problems*, and more specifically in the problems considered here, to *Markov optimal stopping problems*. A tradeoff exists between the error probabilities (which can be made arbitrarily small by taking sufficiently many observations), and the observation time or *detection delay* [59].

In contrast with the classical Neyman-Pearson fixed sample size test, in which the optimisation criterion is to maximise the power $P_1(\delta = 1)$ for a given sample size N and Type I error bound α, the Wald-Wolfowitz criterion is to minimise the expected detection delay under both hypotheses, subject to Type I and Type II error constraints.

3.1.1 Problem statement

In this chapter we consider the task of deciding between two simple statistical hypotheses. That is, hypotheses that are completely specified. In particular, we would like to decide *as quickly as possible*, subject to some constraint(s) on the quality of our final decision. In other words, we seek to solve the sequential detection problem for simple statistical hypotheses.

3.1.2 Chapter overview

We first present the Bayesian formulation of the sequential detection problem in section 3.2, followed by a quick overview of the fixed probability of error formulation (also referred to simply as Wald's formulation) in section 3.3.

Additional considerations are discussed in section 3.4, where the relationship between the Bayesian and Wald's formulation is given, and where approximate expressions for the Average Run Length (ARL) as well as the expected probability of error is given. A short discussion on alternative solutions to the sequential detection problem such as found in the literature of dynamic programming, generalised parking, and free-boundary problems is also given.

Finally section 3.5 gives several illustrative examples of optimal stopping problems including the well-known finite horizon secretary problem (section 3.5.2), and the sequential detection of simple hypotheses (section 3.5.3) where a coin is determined to be either fair or biased.

3.2 HYPOTHESIS TESTING: BAYESIAN FORMULATION

In the first part of this section, we will roughly follow the approach presented in [89]. Let $\mathcal{Z}_n := \{Z_k; k = 1, 2, \ldots, n\}$ be a sequence of i.i.d. real observations that obey one of two statistical hypotheses:

$$\begin{aligned} \mathcal{H}_0 &: Z_k \sim Q_0, \quad k = 1, 2, \ldots \\ \text{versus} & \\ \mathcal{H}_1 &: Z_k \sim Q_1, \quad k = 1, 2, \ldots \end{aligned}$$

where Q_0 and Q_1 are two probability distributions with associated probability densities q_0 and q_1, respectively. Further assume that hypothesis $\mathcal{H}_1$ occurs with prior probability π, and $\mathcal{H}_0$ with prior probability $1 - \pi$.

We would like to decide between these hypotheses in a way that minimises an appropriate measure (to be introduced later) of the error probability and the sampling cost.

If we observe $\{Z_k; k = 1, 2, \ldots, n\}$, we can decide between the two hypotheses by considering the limiting value of the likelihood ratio:

$$\Lambda_n(\mathcal{Z}_n) \triangleq \prod_{k=1}^{n} \frac{q_1(Z_k)}{q_0(Z_k)} \to \begin{cases} 0, & \text{under } \mathcal{H}_0 \\ \infty, & \text{under } \mathcal{H}_1 \end{cases}, \quad \text{as } n \to \infty, \tag{3.1}$$

which naturally leads to the following decision rule (known as the *maximum a posteriori* or MAP rule) when taking the prior probabilities of the hypotheses into account:

$$\delta_n = \begin{cases} 0 \ \text{(i.e. } \mathcal{H}_0 \text{ is true)}, & \text{if } \pi_n^\pi \leq 0.5 \\ 1 \ \text{(i.e. } \mathcal{H}_1 \text{ is true)}, & \text{if } \pi_n^\pi > 0.5 \end{cases} \tag{3.2}$$

where π_n^π is simply the posterior probability that $\mathcal{H}_1$ is true:

$$\pi_n^\pi = \frac{\pi \prod_{k=1}^{n} q_1(Z_k)}{\pi \prod_{k=1}^{n} q_1(Z_k) + (1 - \pi) \prod_{k=1}^{n} q_0(Z_k)}. \tag{3.3}$$

Following [89], we rewrite (3.3) as the following recursion:

$$\pi_n^\pi = \frac{\pi_{n-1}^\pi q_1(Z_n)}{\pi_{n-1}^\pi q_1(Z_n) + (1 - \pi_{n-1}^\pi) q_0(Z_n)}. \tag{3.4}$$

Suppose, now, that we observe $\{Z_k; k = 1, 2, \ldots\}$ sequentially, generating the filtration[1] $\{\mathcal{F}_k; k = 1, 2, \ldots\}$, with

$$\mathcal{F}_k = \sigma(\mathcal{Z}_k) = \sigma(Z_1, Z_2, \ldots, Z_k), \quad k = 1, 2, \ldots, \quad \mathcal{F}_0 = (\Omega, \emptyset), \tag{3.5}$$

[1] A sequence of σ-fields $\{\mathcal{F}_1, \mathcal{F}_2, \ldots\}$ such that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ is called a *filtration*.

and that there is a cost $c \geq 0$ per sample taken. Then clearly the *quality* of a decision between $\mathcal{H}_0$ and $\mathcal{H}_1$ will improve with an increasing number of samples, but the cost of making (or rather, obtaining) the decision will increase too.

Suppose further that there are costs $c_0 > 0$ and $c_1 > 0$ to the events of falsely rejecting hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Then, for any *sequential decision rule* (τ, δ) with $\tau \in \mathcal{T}$ a stopping time (with respect to the filtration $\{\mathcal{F}_k\}$) and $\delta_k \in \mathcal{D}$ an $\mathcal{F}_k$-measurable terminal decision rule, we define the *average cost of errors*:

$$c_e(\tau, \delta) = (1 - \pi)c_0 P_0(\delta_\tau = 1) + \pi c_1 P_1(\delta_\tau = 0), \tag{3.6}$$

where P_0 and P_1 are probability measures on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ such that, under P_j, $Z_1, Z_2, \ldots$ are i.i.d. with marginal distribution Q_j, for $j = 0, 1$ [89]. If we further consider the probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^\infty, \mathcal{B}^\infty, P_\pi)$ where

$$P_\pi = (1 - \pi)P_0 + \pi P_1, \tag{3.7}$$

we can define the *expected cost of sampling* as

$$c\mathbb{E}_\pi\{\tau\}, \tag{3.8}$$

where $\mathbb{E}_\pi\{\cdot\}$ denotes expectation under the measure P_π. Note that, as pointed out in [89], the observations are no longer independent under the measure P_π. This can be understood by realising that past observations affect our belief of the true underlying distribution, so that we expect future observations to support (follow) this belief.

The *total cost* incurred by (or *Bayes risk* of) any sequential decision rule (τ, δ) is then defined as the sum of the average cost of errors and the expected cost of sampling:

$$c_e(\tau, \delta) + c\mathbb{E}_\pi\{\tau\}. \tag{3.9}$$

So, naturally, we would like to choose a sequential decision rule to solve the following optimisation problem:

$$s(\pi) = \inf_{\tau \in \mathcal{T}, \delta \in \mathcal{D}} \Big[c_e(\tau, \delta) + c\mathbb{E}_\pi\{\tau\}\Big], \quad \pi \in [0, 1], \tag{3.10}$$

where $s(\pi)$ is known as the *minimal expected cost*, or simply the *minimal cost* function.

Following the approach presented in [89, Chapter 4], we use the following proposition to convert (3.10) into a Markov optimal stopping problem.

Proposition 1 *For any $\tau \in \mathcal{T}$, we have*

$$\inf_{\delta \in \mathcal{D}} c_e(\tau, \delta) = \mathbb{E}_\pi\Big\{ \min\Big[c_1 \pi_\tau^\pi, c_0(1 - \pi_\tau^\pi)\Big]\Big\}, \tag{3.11}$$

where the sequence $\{\pi_n^\pi\}$ is defined as in (3.4). Moreover, the infimum in (3.11) is achieved by the terminal decision rule

$$\delta_n = \begin{cases} 0, & \text{if } \pi_n^\pi < c_0/(c_0 + c_1) \\ 1, & \text{if } \pi_n^\pi \geq c_0/(c_0 + c_1) \end{cases} \tag{3.12}$$

Proof. The proof can be found in [89, Proposition 4.1, p. 67]. ■

Proposition 1 reduces the problem (3.10) to the alternative problem

$$\inf_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{ \min\Big[c_1\pi_\tau^\pi, c_0(1-\pi_\tau^\pi)\Big] + c\tau\Big\}, \tag{3.13}$$

in which we no longer have to search over all possible terminal decision rules, $\delta \in \mathcal{D}$.

Because of the recursivity of $\{\pi_n^\pi\}$ this new problem can be embedded in a Markov optimal stopping problem[2]. In particular, we have the following result.

Theorem 2 (Optimal i.i.d. sequential decision rule) *Consider the optimisation problem of (3.13), or equivalently, the problem of (3.10). The optimal solution is given by the sequential decision rule* (τ, δ) *with*

$$\tau_{opt} = \inf\Big\{n \geq 0 | \pi_n^\pi \notin (\pi_L, \pi_U)\Big\} \tag{3.14}$$

and

$$\delta_n = \begin{cases} 0, & \text{if } \pi_n^\pi \leq c_0/(c_0+c_1) \\ 1, & \text{if } \pi_n^\pi > c_0/(c_0+c_1) \end{cases} \tag{3.15}$$

where the exit thresholds π_L *and* π_U *are given by*

$$\pi_L = \sup\{0 \leq \pi \leq 1 | s(\pi) = c_1\pi\} \tag{3.16}$$

and

$$\pi_U = \inf\{0 \leq \pi \leq 1 | s(\pi) = c_0(1-\pi)\} \tag{3.17}$$

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

Proof. The proof follows from [89, Proposition 4.3, pp. 69–70], combined with the proof of Proposition 1. ■

We will now use the following theorem, adapted from [103], to obtain a *computational method* for the optimal cost $s(\pi)$ and for the thresholds π_L and π_U given in Theorem 2.

Theorem 3 *Consider a homogeneous Markov process* $x = \{X_k; k = 0, 1, \ldots\}$*, and let the functions* $g(x)$ *and* $c(x)$ *satisfy*

$$|g(x)| \leq G < \infty, \quad 0 \leq c(x) \equiv c < \infty, \forall x. \tag{3.18}$$

If we have the following optimisation task:

$$v(x) = \sup_{\tau\in\mathcal{T}} \mathbb{E}_x\Big\{ g(X_\tau) - \sum_{v=0}^{\tau-1} c\Big\}, \tag{3.19}$$

[2]Note that the sequence $\{\pi_n^\pi\}$ is a homogeneous Markov process [59].

then

$$v(x) = \lim_{N\to\infty} \mathcal{Q}^N g(x) \tag{3.20}$$

where $\mathcal{Q}^N$ *is the* N*th power of the operator*

$$\mathcal{Q}f(x) = \max\Big\{f(x), \mathcal{R}f(x) - c\Big\} \tag{3.21}$$

and where the operator $\mathcal{R}$ *is defined as*

$$\mathcal{R}f(x) = \mathbb{E}_x\Big\{f(X_1)\Big\}. \tag{3.22}$$

In addition, the stopping time

$$\tau = \inf\Big\{n \geq 0 | g(X_n) = v(X_n)\Big\} \tag{3.23}$$

is optimal.

Proof. The proof can be found in [103, Theorem 23, p. 94] by setting $\alpha = 1$, $c(x) \equiv c \geq 0$, and by assuming that $P_x(\tau < \infty) = 1$. In our case this is indeed a reasonable assumption; see Remark 5 on page 59. ■

We now give two additional intermediate but important results, which we will use, together with Theorem 3, to derive a computational method for the minimal cost, $s(\pi)$, and for the optimal exit thresholds, π_L and π_U.

Proposition 2 ($\mathcal{R}$ is a positive linear operator) *Let* $\{X_k; k = 0, 1, \ldots\}$ *be a homogeneous Markov process with values in the state space* $(E, \mathcal{B})$*, and let* f *and* g *be* $\mathcal{B}$*-measurable functions. The operator* $\mathcal{R}$ *defined as*

$$\mathcal{R}f(x) := \mathbb{E}_x\Big\{f(X_1)\Big\} \tag{3.24}$$

is a positive linear operator.

Proof. For $\mathcal{R}$ to be a positive linear operator, the following conditions must be satisfied:

1. $f \geq 0 \implies \mathcal{R}f \geq 0$ [for positivity], and
2. $\mathcal{R}\Big[f - h\Big](x) = (\mathcal{R}f)(x) - (\mathcal{R}h)(x)$ [for linearity].

We first show that $\mathcal{R}$ is a *positive* operator, i.e. that $\mathcal{R} \geq 0$ (see Remark 2 on the next page). Let f be a (positive) function such that

$$f(x) \geq 0, \forall x \in \mathcal{D}(f), \tag{3.25}$$

where $\mathcal{D}(f)$ denotes the *domain* of f. Then, clearly, we also have

$$f[X_1(\omega)] \geq 0, \forall \omega \in \Omega. \tag{3.26}$$

Finally we note that

$$\begin{aligned}(\mathcal{R}g)(x) &= \mathbb{E}_x[f(X_1)] \\ &= \int_\Omega f(X_1|X_0 = x)dP \geq 0,\end{aligned} \tag{3.27}$$

whence it follows that $\mathcal{R}$ is a positive operator.

To show that $\mathcal{R}$ is a linear operator, consider the following:

$$\begin{aligned}\mathcal{R}[f-h](x) &= \mathbb{E}_x[f-h](X_1) \\ &= \mathbb{E}_x[f(X_1)] - \mathbb{E}_x[h(X_1)] \\ &= (\mathcal{R}f)(x) - (\mathcal{R}h)(x),\end{aligned} \tag{3.28}$$

where the second equality simply follows from the linearity of expectation. ■

Remark 2 *Note that $\mathcal{R} \geq 0$ means that $\mathcal{R}f \geq 0$ for all functions f such that $f \geq 0$.*

Lemma 1 *If T is a positive linear operator, then*

$$T\Big[\max\{g, Tg\}\Big](x) \geq Tg(x). \tag{3.29}$$

Proof. Using the positivity and linearity of T we have:

$$\begin{aligned}\max\Big\{g(x), Tg(x)\Big\} \geq g(x) \implies & \max\Big\{g(x), Tg(x)\Big\} - g(x) \geq 0 \\ \implies & T\Big[\max\{g(x), Tg(x)\} - g(x)\Big] \geq 0 \\ \implies & T\Big[\max\{g(x), Tg(x)\}\Big] - Tg(x) \geq 0 \\ \implies & T\Big[\max\{g, Tg\}\Big](x) \geq Tg(x),\end{aligned} \tag{3.30}$$

as required. ■

Finally, by using Theorem 3, Proposition 2, and Lemma 1, we obtain the following computational method for finding the minimal cost function, $s(\pi)$, and the optimal exit thresholds, π_L and π_U, as given in Theorem 2.

Proposition 3 (Computing the minimal cost function for an i.i.d. sequence) *The minimal cost $s(\pi) = \inf_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{h(\pi_\tau^\pi) + c\tau\Big\}$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$, is the monotone pointwise limit from above of the sequence of functions*

$$s_n(\pi) = \min\Big\{h(\pi), \mathcal{R}s_{n-1}(\pi) + c\Big\}, \quad n = 1, 2, \ldots \tag{3.31}$$

with $s_0(\pi) = h(\pi)$, and where the operator $\mathcal{R}$ is defined by

$$\begin{aligned}\mathcal{R}f(\pi) &= \mathbb{E}_\pi[f(\pi_1^\pi)] \\ &= \int_{-\infty}^{\infty} f\left(\frac{\pi q_1(Z_1)}{\pi q_1(Z_1) + (1-\pi)q_0(Z_1)}\right) \cdot \Big[\pi q_1(Z_1) + (1-\pi)q_0(Z_1)\Big]\, dZ_1,\end{aligned}$$

such that

$$\mathcal{R}s_{n-1}(\pi) = \mathbb{E}_\pi\Big\{s_{n-1}(\pi_1^\pi)\Big\}. \tag{3.32}$$

Proof. We first write the minimal cost $s(\pi)$ in such a way that we can apply Theorem 3:

$$\begin{aligned} s(\pi) &= \inf_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{h(\pi_\tau^\pi) + c\tau\Big\} \\ &= -\sup_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{-h(\pi_\tau^\pi) - c\tau\Big\} \\ &= -\sup_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{g(\pi_\tau^\pi) - c\tau\Big\} \\ &= -v(\pi), \end{aligned} \tag{3.33}$$

where $g(x) := -h(x)$, and $v(x) = \sup_{\tau\in\mathcal{T}} \mathbb{E}_x\{g(X_\tau) - c\tau\}$ as in Theorem 3.

We notice that $s(\pi)$ in (3.33) is now such that we can easily apply Theorem 3 (to approximate $v(\pi) = -s(\pi)$). In addition, $\{\pi_n^\pi\}$ is a homogeneous Markov process (see for example [59]), and $g(x)$ and $c(x) \equiv c$ clearly satisfy the conditions given in (3.18). Therefore, by Theorem 3 it follows that

$$\begin{aligned} v(\pi) &= \sup_{\tau\in\mathcal{T}} \mathbb{E}_\pi\Big\{g(\pi_\tau^\pi) - c\tau\Big\} \\ &= \lim_{N\to\infty} \mathcal{Q}^N g(\pi), \end{aligned} \tag{3.34}$$

with

$$\begin{aligned} v_1(\pi) = \mathcal{Q}g(\pi) &= \max\Big\{g(\pi), \mathcal{R}g(\pi) - c\Big\} \\ &= -\min\Big\{h(\pi), \mathcal{R}h(\pi) + c\Big\} \\ &= -\min\Big\{h(\pi), \mathcal{R}s_0(\pi) + c\Big\} \\ &= -s_1(\pi), \end{aligned} \tag{3.35}$$

where $\mathcal{R}s_0(\pi) = \mathbb{E}_\pi\Big\{s_0(\pi_1^\pi)\Big\}$. Also, since by Lemma 1 we have that

$$T[\max\{g, Tg\}] \geq Tg \tag{3.36}$$

for any linear operator T, and since $\mathcal{R}$ is a linear operator (Proposition 2), we have

$$\begin{aligned} v_2(\pi) = \mathcal{Q}^2 g(\pi) &= \max\Big\{\mathcal{Q}g(\pi), \mathcal{R}\mathcal{Q}g(\pi) - c\Big\} \\ &= \max\Big\{\max[g(\pi), \mathcal{R}g(\pi) - c], \mathcal{R}\Big[\max\Big(g(\pi), \mathcal{R}g(\pi) - c\Big)\Big] - c\Big\} \\ &= \max\Big\{g(\pi), \mathcal{R}\Big[\max\Big(g(\pi), \mathcal{R}g(\pi) - c\Big)\Big] - c\Big\} \\ &= \max\Big\{g(\pi), \mathcal{R}\mathcal{Q}g(\pi) - c\Big\} \\ &= \max\Big\{-h(\pi), -\mathcal{R}s_1(\pi) - c\Big\} \\ &= -\min\Big\{h(\pi), \mathcal{R}s_1(\pi) + c\Big\} \\ &= -s_2(\pi), \end{aligned} \tag{3.37}$$

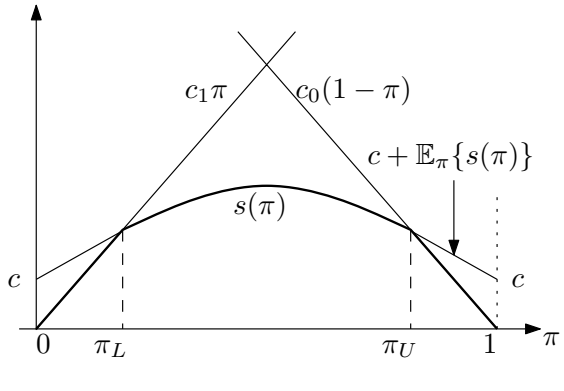
where $\mathcal{R}s_1(\pi) = \mathbb{E}_\pi\{s_1(\pi_1^\pi)\}$. Continuing in this manner we see that $-\mathcal{Q}^n g(\pi), n = 1, 2, \ldots$ corresponds to the sequence of functions $s_n(\pi)$ given in (3.31), as required. ■

Proposition 3 constitutes a means of computing the minimal cost function, $s(\pi)$, after which it is relatively straightforward to determine π_{L} and π_{U} using (3.16) and (3.17).

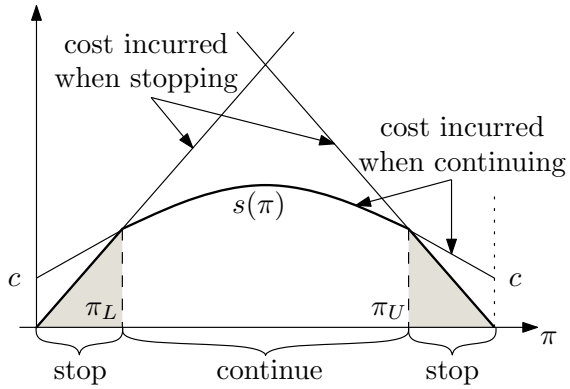
3.2.1 On the structure of the minimal cost function

It can be shown (see for example [103] or the references contained therein) that the minimal cost function $s(\pi)$ is concave, and is bounded as $0 \leq s(\pi) \leq h(\pi)$, where $h(\pi) = \min\{c_1\pi, c_0(1-\pi)\}$ as defined in Proposition 3. In addition, it always holds true that $s(0) = s(1) = 0$. Notice that the prior probability that $\mathcal{H}_1$ is true (i.e., π) does not affect the computation of the minimal cost function. That is, $s(\pi)$ holds for any $\pi \in [0, 1]$.

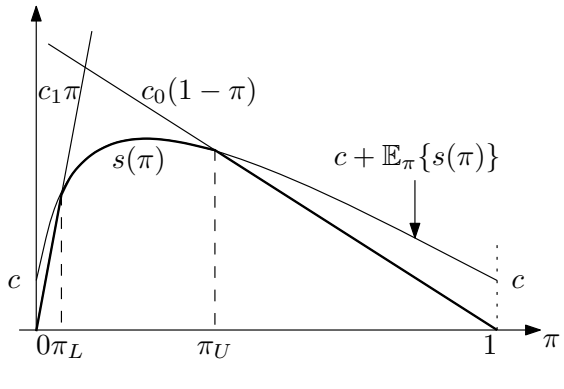
A typical minimal cost function is shown in Figure 3.1.a, where equal costs of error ($c_0 = c_1$) causes $s(\pi)$ to be symmetric about the line $\pi = 1/2$.



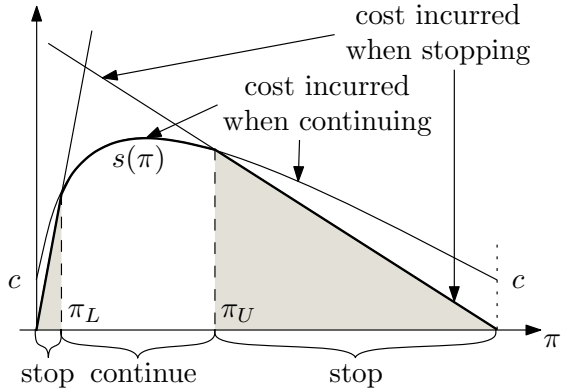
a. Structure of the cost function, $s(\pi)$, $c_0 = c_1$.



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



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



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

Figure 3.1: Typical structure and behaviour of the minimal cost function, $s(\pi)$.

With reference to Figure 3.1.b, we notice that $s(\pi)$ can be divided into regions of action

(i.e., continue sampling, or stop sampling). More specifically, we can see that $s(\pi)$ simply corresponds, at each $\pi \in [0,1]$, to the minimum between the cost incurred when continuing to sample (corresponding to $c + \mathbb{E}_\pi\{s(\pi)\}$ in Figure 3.1.a) and the cost incurred when stopping the experiment (which is simply $h(\pi)$). The same applies to Figure 3.1.c and Figure 3.1.d, where the only difference is that the costs of errors (c_0 and c_1) are no longer equal.

3.3 HYPOTHESIS TESTING: WALD'S FORMULATION

Wald's formulation of the sequential detection problem can be presented very intuitively. Consider again a sequence $\mathcal{Z}_n := \{Z_k; k = 1, 2, \ldots, n\}$ of i.i.d. real observations, defined as in section 3.2. That is,

$$\begin{aligned} \mathcal{H}_0 &: Z_k \sim Q_0, \quad k = 1, 2, \ldots \\ \text{versus} & \\ \mathcal{H}_1 &: Z_k \sim Q_1, \quad k = 1, 2, \ldots \end{aligned}$$

where Q_0 and Q_1 are two probability distributions with associated probability densities q_0 and q_1, respectively; and let hypothesis $\mathcal{H}_1$ occur with prior probability π, and $\mathcal{H}_0$ with prior probability $1 - \pi$.

We would, again, like to decide between these hypotheses in a way that minimises an appropriate measure of the error probability and sampling cost. However, this time we will go right ahead and state (informally) the exact problem that we would like to solve.

Wald's formulation solves the following problem:
We would like to stop sampling *as quickly as possible* (this seems a very natural choice, doesn't it?), given some (usually small) permissible probability of error.

Remark 3 *It may be somewhat surprising at first, but the above problem statement is* exactly *the same as given (3.10) in the Bayesian framework, with the exception of the explicit incorporation of the prior probabilities, of course, but these can be added to Wald's formulation very easily.*

We will now derive Wald's much simpler (or rather, more intuitive) solution to our original problem (3.10). We again define the likelihood ratio as

$$\Lambda_n(\mathcal{Z}_n) \triangleq \prod_{k=1}^{n} \frac{q_1(Z_k)}{q_0(Z_k)} \to \begin{cases} 0, & \text{under } \mathcal{H}_0 \\ \infty, & \text{under } \mathcal{H}_1 \end{cases}, \quad \text{as } n \to \infty, \tag{3.38}$$

We now normalise the joint likelihoods (such that $p_0^n(\mathcal{Z}_n) + p_1^n(\mathcal{Z}_n) = 1$), after which we'll talk about probabilities, even though this is perhaps a serious misuse of terminology.

In this way, we define

$$p_1^n(\mathcal{Z}_n) = \frac{\prod_{k=1}^n q_1(Z_k)}{\prod_{k=1}^n q_0(Z_k) + \prod_{k=1}^n q_1(Z_k)}, \tag{3.39}$$

as the probability of observing the sequence $\mathcal{Z}_n = \{Z_1, Z_2, \ldots, Z_n\}$ under $\mathcal{H}_1$, and

$$p_0^n(\mathcal{Z}_n) = \frac{\prod_{k=1}^n q_0(Z_k)}{\prod_{k=1}^n q_0(Z_k) + \prod_{k=1}^n q_1(Z_k)}, \tag{3.40}$$

as the probability of observing the sequence $\mathcal{Z}_n$ under hypothesis $\mathcal{H}_0$.

Remark 4 (Probability ratios and likelihood ratios) *When we consider a probability ratio between (3.39) and (3.40), we note that their denominators are the same, so that it is equivalent to considering the likelihood ratio given in (3.38). That is why we will refer to a Wald-type sequential tests as a SPRT, even though the test only requires us to consider likelihoods and likelihood ratios.*

Wald's test, similar to the Bayesian formulation, continues sampling until certain thresholds (often called the *exit thresholds*) are crossed. In the Bayesian formulation we monitored the posterior sequence $\{\pi_n^\pi\}$, and continued sampling as long as $\pi_n^\pi \in (\pi_\mathrm{L}, \pi_\mathrm{U})$. In Wald's test, we continue sampling as long as $\Lambda_n \in (A, B)$, such that

$$\tau = \inf\{n \geq 0 | \Lambda_n \notin (A, B)\}. \tag{3.41}$$

We now introduce the exit thresholds (A and B) for the SPRT(A, B) as follows:

$$A \leq \prod_{k=1}^n \frac{q_1(Z_k)}{q_0(Z_k)} \leq B, \tag{3.42}$$

so that we continue sampling until $\Lambda_n \leq A$ (in which case we stop and choose $\mathcal{H}_0$), or $\Lambda_n \geq B$ (in which case we stop and choose $\mathcal{H}_1$).

The question naturally arises, "*how can we determine the thresholds A and B?*"—the answer is both complex and simple. To find the *exact* values of A and B turns out to be rather tricky in general (a computational method is proposed in [118]), but luckily it is extremely easy to find really good *approximations* to A and B, which typically work very well in practice [118]. We will now consider these approximations in more detail.

When $\Lambda_n \geq B$, we stop sampling and decide on hypothesis $\mathcal{H}_1$. We then clearly have

$$\prod_{k=1}^n q_1(Z_k) \geq B \cdot \prod_{k=1}^n q_0(Z_k) \implies p_1^n(\mathcal{Z}_n) \geq B \cdot p_0^n(\mathcal{Z}_n) \tag{3.43}$$

$$\implies P_1(\delta = 1) \geq B \cdot P_0(\delta = 1) \tag{3.44}$$

which we can interpret as saying that "*the probability of having observed the sequence $\mathcal{Z}_n$ under hypothesis $\mathcal{H}_1$ is at least B times as great as under hypothesis $\mathcal{H}_0$.*" Furthermore,

since we have decided on hypothesis $\mathcal{H}_1$, the probability of having chosen the wrong hypothesis is equal to $P_0(\delta = 1) = \alpha$. That is, α is the probability of incorrectly choosing hypothesis $\mathcal{H}_1$ instead of $\mathcal{H}_0$, which is called a Type I error. If we now define the Type II error probability as $P_1(\delta = 0) = \beta$, we have from (3.44)

$$B \leq \frac{1-\beta}{\alpha}, \tag{3.45}$$

so that $(1-\beta)/\alpha$ constitutes an *upper limit* for B. In a similar fashion we can derive a *lower limit* for A as

$$A \geq \frac{\beta}{1-\alpha}. \tag{3.46}$$

Note that a more rigorous proof of the above can be found in [89, Proposition 4.10]. Finally, when we replace the inequalities in (3.46) and (3.45) by equalities, we arrive at the celebrated *Wald's approximations* to A and B, respectively.

3.4 ADDITIONAL CONSIDERATIONS

In this section we will present a number of interesting additional considerations regarding the theory of sequential detection, such as the relationship between the Bayesian formulation and Wald's formulation, as well as approximate expressions for the probability of error and the ARL of a particular SPRT.

3.4.0.1 Bayesian vs Wald's sequential detection

The Bayes optimal stopping time defined in (3.14) can alternatively be expressed as

$$\tau_{\text{opt}} = \inf\Big\{n \geq 0 | \Lambda_n \notin (A, B)\Big\}, \tag{3.47}$$

where the thresholds A and B are given by

$$A = \frac{1-\pi}{\pi}\frac{\pi_{\text{L}}}{1-\pi_{\text{L}}} \Longleftrightarrow \pi_{\text{L}} = \frac{\pi A}{1-\pi(1-A)}, \tag{3.48}$$

and

$$B = \frac{1-\pi}{\pi}\frac{\pi_{\text{U}}}{1-\pi_{\text{U}}} \Longleftrightarrow \pi_{\text{U}} = \frac{\pi B}{1-\pi(1-B)}, \tag{3.49}$$

so that we notice that the Bayes optimal sequential decision rule with thresholds π_{L} and π_{U} corresponds to Wald's SPRT with thresholds A and B as defined above.

It follows therefore, that the Wald-Wolfowitz theorem (Theorem 1 on page 4) also applies to Bayes optimal sequential decision rules (since they are equivalent to SPRTs), and that no other test (sequential or otherwise) will have smaller expected run lengths with the same probability of error.

Remark 5 *A consequence of the Wald-Wolfowitz theorem is that the stopping times of SPRTs (and hence also of Bayes optimal decision rules) have finite expectations under both $\mathcal{H}_0$ and $\mathcal{H}_1$ (see for example [104, Proposition 8.21]). In other words, we have that $P_\pi(\tau < \infty) = 1$, which is a necessary condition for the ϵ-optimal stopping time described in Theorem 23 of [103] to be an optimal stopping time as presented in Theorem 3.*

3.4.0.1.1 On the difficulty of choosing realistic values for the cost of sampling. It is considerably more difficult to formulate a realistic problem in the Bayesian framework than in Wald's formulation, since it remains somewhat unclear as to how we should choose an appropriate cost of sampling, c.

In most practical applications we are usually concerned with the more intuitive concept of the *probability of error*, and so it is much more convenient to specify such problems in Wald's framework. However, it is not clear how we can infer the cost of sampling from the desired probability of error without having to search for the cost experimentally.

We will therefore proceed to give expressions for the probability of error and ARL only in terms of A, B, π_{L}, π_{U}, α and β, and not in terms of the costs c_0, c_1 and c.

3.4.0.2 Estimating the probability of error

By using Wald's approximations to (3.45) and (3.46) we can solve (approximately) for α and β to obtain:

$$\alpha \approx \frac{1-A}{B-A} = \frac{\pi_{\mathrm{L}} - \pi}{\pi - 1} \cdot \frac{\pi_{\mathrm{U}} - 1}{\pi_{\mathrm{L}} - \pi_{\mathrm{U}}}, \tag{3.50}$$

and

$$\beta \approx A\frac{B-1}{B-A} = \frac{\pi_{\mathrm{L}}}{\pi} \cdot \frac{\pi - \pi_{\mathrm{U}}}{\pi_{\mathrm{L}} - \pi_{\mathrm{U}}}. \tag{3.51}$$

The probability of error is then finally given simply as

$$\begin{aligned} P_{\mathrm{e}} &= (1-\pi)P_0(\delta_\tau = 1) + \pi P_1(\delta_\tau = 0) \\ &= (1-\pi)\alpha + \pi\beta, \end{aligned} \tag{3.52}$$

which can be determined approximately by substituting (3.50) and (3.51) into (3.52).

3.4.0.3 Estimating the average run length (ARL)

We have the following useful proposition for computing the approximate ARL of an SPRT with Type I and Type II error probabilities α and β:

Proposition 4 *Suppose the random variable $\log \Lambda_1$ has finite means d_0 and d_1 under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Then*

$$\mathbb{E}_0[\tau] \geq d_0^{-1}\left[\alpha \log\left(\frac{1-\beta}{\alpha}\right) + (1-\alpha)\log\left(\frac{\beta}{1-\alpha}\right)\right], \tag{3.53}$$

and

$$\mathbb{E}_1[\tau] \geq d_1^{-1}\left[(1-\beta)\log\left(\frac{1-\beta}{\alpha}\right)+\beta\log\left(\frac{\beta}{1-\alpha}\right)\right]. \tag{3.54}$$

Proof. The proof can be found in [89, Proposition 4.11]. ■

Note that when the desired probabilities of error (α and β) are sufficiently small, then the bounds in Proposition 4 are relatively tight. Furthermore, when the random process that we observe is continuous, Proposition 4 holds with equality (see [89] for details).

3.4.0.4 Alternative methods for sequential detection

It is worthwhile to note that there are several alternative ways to compute the optimal thresholds for the Bayes sequential decision rule. Among these, are the direct computational approach presented in Proposition 3, a method by which the problem is first transformed into a *free-boundary problem* (see for example, [88]), the method of *generalised parking* in which the sequential detection problem is formulated as an Itō stochastic differential equation [8], and the contemporary techniques of dynamic programming, Bellman equations and Stefan problems.

3.5 ILLUSTRATIVE EXAMPLES

Several illustrative examples of optimal stopping problems (and their solutions) are given in this section, including a problem whose solution can be obtained by *backward induction* (section 3.5.1), the well-known finite horizon secretary (or marriage) problem discussed in section 3.5.2, and finally the infinite-horizon sequential detection or classification task presented in section 3.5.3, which is of course of primary concern in this study.

3.5.1 Backward induction

Amongst the easiest optimal stopping problems to solve (although only for relatively small problems) are problems for which we have complete information about the probabilities and the exact values of all potential future observations, in which case the technique of *backward induction* is of particular importance.

3.5.1.1 Problem formulation

Consider the following simple backward induction problem, stated and solved in [45]. We are given the opportunity to win some money by rolling a standard, fair, six-sided die at most five times. We may stop whenever we want, at which time we receive as a reward the number of Krugerrands corresponding to the number of dots shown on the die. Our objective is simply to find a stopping rule that will maximise the number of Krugerrands that we can expect to win.

3.5.1.2 Solution

We will use backward induction to solve this problem, but first, consider the following. If we decide to stop with the first roll, the expected reward is simply the expected value of a random variable that takes the values 1, 2, 3, 4, 5, and 6 with probability 1/6 each, which is equal to 3.5 Krugerrands. It is also clearly not optimal to stop if the first roll is a 1, and similarly, it is *always* optimal to stop with a 6.

The idea behind backward induction is simple: it is clearly optimal to stop whenever the value of the current roll is greater than the expected value of the roll if we continue. In this way, the optimal strategy for a five-roll game would be to stop whenever we observe a value greater than the expected value of the four-roll game. The four-roll problem in turn has an optimal strategy which involves stopping whenever the observed value is greater than the expected value of the three-roll game, and so on. Finally, by realising that we already know the optimal strategy for a one-roll game, we can work backwards (from 1 to 5) to determine the optimal strategy for the five-roll game.

In the one-roll game there is only one strategy, namely to stop, for which the expected reward is 3.5. This information now yields the optimal strategy for the two-roll problem—stop on the first roll only if the observed value is more than you expect to win if you continue, i.e. 3.5, so that the strategy for the two-roll game is "stop at the first roll only if it is a 4,5, or a 6". We can now calculate the expected reward for a two-roll game as follows: $4(1/6) + 5(1/6) + 6(1/6) + (1/2)(3.5) = 4.25$, which in turn can be used to determine the optimal strategy for a three-roll game.

The optimal strategy for a three-roll game would be to "stop if the first roll is a 5 or 6 (that is, more than 4.25), otherwise continue and stop only if the second roll is a 4, 5, or a 6, as for the two-roll game." Knowing the optimal strategy for a three-roll game allows us to compute our expected reward (4.67), which in turn can be used to determine the optimal strategy for a four-roll game. Continuing (backwards) in this manner we can finally determine the optimal strategy for a five-roll game.

The optimal strategies, along with the expected winnings are shown in Table 3.1.

Table 3.1: Optimal stopping strategies obtained by backward induction.

Total number of rolls	**Stop if initial roll is:**	**Average optimal expected reward**
1	$\{1, 2, 3, 4, 5, 6\}$	3.5
2	$\{4, 5, 6\}$	4.25
3	$\{5, 6\}$	4.67
4	$\{5, 6\}$	4.94
5	$\{5, 6\}$	5.13

Although backward induction is very versatile (and it works equally well if the process values are not independent as they were assumed to be in this example), we do not always

have sufficient information about a particular problem to apply backward induction. An example where we cannot apply backward induction for a lack of information, but where an optimal strategy can still be found, is described next.

3.5.2 Finite horizon secretary problem

One of the most famous optimal stopping problems is the so-called secretary, marriage, dowry, or best-choice problem. The problem and its many variants have a long and rich history, which is superbly presented in the delightful article by Ferguson [33]. The problem is easy to state, and has a striking—and remarkably simple—solution.

We will specifically consider the version of the problem as stated in [20], which is the most common (and arguably the simplest) of the various secretary problems. However, the objective in [20] is to maximise the expected rank of the chosen applicant, whereas we will consider the task of maximising the probability of selecting the best applicant.

Furthermore, we will state—but not derive—the solution to this problem, since our aim with this example is simply to show one of the many types of problems commonly encountered in optimal stopping theory, and besides, the solution is somewhat technical but can easily be found in several sources by the interested reader (see e.g. [89]).

3.5.2.1 Problem formulation

The following problem statement follows closely that of [20]. Suppose $N > 2$ people apply for a certain (secretarial) position, and that the applicants are rank ordered according to the ranks $1, 2, \ldots, N$ with 1 denoting the highest ranking. The applicants present themselves one by one, in random order, and when the ith applicant appears we can only observe her rank, X_i, relative to her $i-1$ predecessors. We may either select the ith applicant, in which case the process ends, or we may reject her and go on to the $(i+1)$th applicant; in which case the ith applicant cannot be recalled. We are very particular, and will be satisfied with nothing but the very best, so that our reward is 1 if we choose the best of the N applicants, and 0 otherwise. Our task is then to maximise the probability of selecting the very best applicant.

3.5.2.2 Solution

The solution (optimal stopping time) of this problem is given as (see for example [89]):

$$\tau_{\text{opt}} = \begin{cases} \inf\{k \geq k^* | X_k = 1\} & \text{if } \min\{X_{k^*}, \ldots, X_N\} = 1 \\ N & \text{otherwise} \end{cases} \tag{3.55}$$

where

$$k^* = \min\left\{ k \in \{2, 3, \ldots, N\} \middle| \sum_{l=k+1}^{n} \frac{1}{l-1} \leq 1 \right\}. \tag{3.56}$$

That is, we reject the first $k^* - 1$ applicants (but we do need to observe and rank them), and then we select the next applicant who ranks first among her predecessors. It can be shown that $k^* \approx N/\mathrm{e}$, and that the probability of selecting the highest-ranking applicant is asymptotically $1/\mathrm{e}$. A MATLAB script for this problem is given in Listing 3.1 below.

Listing 3.1: MATLAB code for a simulation of the secretary (or marriage) problem.

```
N = 1000;                                % maximum number of applicants to interview
app_rank = randperm(N);                  % randomly assign unique ranks to applicants

% determine number of applicants to reject, k_star:
v = ones(1, N)./(1:N);
v = cumsum(v(end-1:-1:1));
v = v(end:-1:1);
[val, idx] = find(v<=1);
k_star = min(idx);
rejected = app_rank(1:1:k_star-1);       % reject first k_star-1 applicants
min_rejected = min(rejected);            % best ranking, rejected applicant
cset = app_rank(k_star:1:end);           % set of applicants to consider
[val,idx] = find(cset<=min_rejected,1); % first remaining applicant whose rank is
                                         % higher than the best, rejected applicant
if isempty(idx)
   choose_idx = N;                % choose the last applicant, irrespective of her rank
else
   choose_idx = idx+k_star-1;            % choose applicant number (idx+k_star-1)
end
fprintf('You chose applicant %g. with rank_T = %g\n', choose_idx,app_rank(choose_idx));
```

3.5.3 Infinite horizon simple hypothesis testing

We will now consider the Bayesian sequential detection task of deciding whether a certain coin is fair, or biased in a particular way.[3] This simple problem can be used to investigate a number of properties of sequential tests more easily and transparently than is possible with more complex data such as the land cover model, presented later.

3.5.3.1 Problem formulation

Consider again a sequence $\{Z_k; k = 1, 2, \ldots, n\}$ of i.i.d. observations generated by either

$$\mathcal{H}_0 \quad : Z_k \sim Q_0, \quad k = 1, 2, \ldots \quad \text{(fair coin)}$$

versus

$$\mathcal{H}_1 \quad : Z_k \sim Q_1, \quad k = 1, 2, \ldots \quad \text{(biased coin)}$$

where Q_0 and Q_1 are two probability distributions with associated probability mass functions q_0 and q_1, respectively; and where hypothesis $\mathcal{H}_1$ occurs with prior probability π, and $\mathcal{H}_0$ with prior probability $1 - \pi$. We also define the random variable Z_k as

$$Z_k(\omega) = \begin{cases} 0 & \text{if } \omega = \text{tails} \\ 1 & \text{if } \omega = \text{heads} \end{cases}, \quad k = 1, 2, \ldots. \tag{3.57}$$

[3] We say "biased in a particular way" because we only consider simple hypotheses here. If, instead, we want to test the hypotheses that the coin is either fair (a simple hypothesis), or not (a composite hypothesis), we must consider other methods that are beyond the scope of this study.

Given some costs c_0 and c_1 corresponding to the cost of falsely rejecting $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively; and a cost of sampling, c, we want to minimise the Bayes risk:

$$\inf_{\tau\in\mathcal{T},\delta\in\mathcal{D}} \Big[c_e(\tau,\delta) + c\mathbb{E}_\pi\{\tau\}\Big], \tag{3.58}$$

where the average cost of errors and the expected cost of sampling are as defined in (3.6) and (3.8), respectively.

Table 3.2: Experimental parameters and results of simple Bayesian sequential detection.

	Case I.a	**Case I.b**	**Case II**	**Case III**
Experimental parameters:				
hypothesis $\mathcal{H}_0$ (fair): $[p(0\|\mathcal{H}_0), p(1\|\mathcal{H}_0)]$	$[0.5, 0.5]$	$[0.5, 0.5]$	$[0.5, 0.5]$	$[0.5, 0.5]$
hypothesis $\mathcal{H}_1$ (biased): $[p(0\|\mathcal{H}_1), p(1\|\mathcal{H}_1)]$	$[0.4, 0.6]$	$[0.4, 0.6]$	$[0.45, 0.55]$	$[0.4, 0.6]$
prior probability that $\mathcal{H}_1$ is true, π	0.5	0.7	0.5	0.5
cost of sampling, c	2×10^{-3}	2×10^{-3}	2×10^{-3}	2×10^{-3}
cost of Type I error, c_0	1	1	1	1
cost of Type II error, c_1	1	1	1	3
Optimal thresholds:				
upper threshold, π_{U}	0.860	0.860	0.640	0.830
lower threshold, π_{L}	0.140	0.140	0.360	0.040
Performance results:				
Experimental ARL	70	52	37	96
Experimental-theoretical ARL	69	52	36	95
Approximate-theoretical ARL	64	48	32	90
Experimental probability of error	0.130	0.131	0.352	0.109
Theoretical probability of error	0.140	0.140	0.360	0.116
Experimental Type I probability of error	0.131	0.337	0.355	0.188
Theoretical Type I probability of error, α	0.140	0.363	0.360	0.198
Experimental Type II probability of error	0.129	0.043	0.350	0.031
Theoretical Type II probability of error, β	0.140	0.044	0.360	0.033

3.5.3.2 Experimental parameters

We will consider the solution to three distinct cases, briefly defined below.

3.5.3.2.1 Case I The first case that we will consider (which is actually Case I.a) is characterised by equal costs of error, $c_0 = c_1 = 1$, a cost of sampling of $c = 2 \times 10^{-3}$, equiprobable hypotheses ($\pi = 1 - \pi = 0.5$), and the following probability mass functions:

$$q_0(x) = \begin{cases} 0.5 & \text{if } x = 0 \ (\omega = \text{tails}) \\ 0.5 & \text{if } x = 1 \ (\omega = \text{heads}) \end{cases} \quad \text{and} \quad q_1(x) = \begin{cases} 0.4 & \text{if } x = 0 \ (\text{tails}) \\ 0.6 & \text{if } x = 1 \ (\text{heads}) \end{cases}. \tag{3.59}$$

Note that Case I.b is similarly defined, except that the hypotheses are no longer equiprobable: $\pi = 0.7$.

3.5.3.2.2 Case II The second case is also characterised by equal costs of error, $c_0 = c_1 = 1$, a cost of sampling of $c = 2 \times 10^{-3}$, and equiprobable hypotheses ($\pi = 1 - \pi = 0.5$), but the probability mass functions are more difficult (that is, they are more alike than Case I):

$$q_0(x) = \begin{cases} 0.5 & \text{if } x = 0 \ (\omega = \text{tails}) \\ 0.5 & \text{if } x = 1 \ (\omega = \text{heads}) \end{cases} \quad \text{and} \quad q_1(x) = \begin{cases} 0.45 & \text{if } x = 0 \ (\text{tails}) \\ 0.55 & \text{if } x = 1 \ (\text{heads}) \end{cases}. \tag{3.60}$$

3.5.3.2.3 Case III The final case that we will consider is characterised by unequal costs of error, $c_0 = 1$, $c_1 = 3$, which places a disproportionate cost on making Type II errors, a cost of sampling of $c = 2 \times 10^{-3}$, equiprobable hypotheses ($\pi = 1 - \pi = 0.5$), and the probability mass functions are again as defined in (3.59).

Table 3.2 on the previous page summarises the experimental parameters, as well as the simulation results, of all three cases presented above.

3.5.3.3 Solution

The solution to (3.58) is given by Theorem 2, and Proposition 3 provides a computational strategy to determine the cost function $s(\pi)$, from which the optimal thresholds, π_{L} and π_{L}, can be determined.

In addition, we can compute the approximate ARL by applying Proposition 4 as follows. Assuming that $\log \Lambda_1$ has finite means d_0 and d_1 under hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, we can compute

$$\begin{aligned} d_0 &= \mathbb{E}_0\Big[\log \Lambda_1\Big] \\ &= \mathbb{E}_0\left[\log\left(\frac{q_1(Z_1)}{q_0(Z_1)}\right)\right] \\ &= \mathbb{E}_0\Big[\log q_1(Z_1) - \log q_0(Z_1)\Big] \\ &= \mathbb{E}_0\Big[\log q_1(Z_1)\Big] - \mathbb{E}_0\Big[\log q_0(Z_1)\Big]. \end{aligned} \tag{3.61}$$

The expectations in (3.61) can further be expanded to yield

$$\begin{aligned} \mathbb{E}_0\Big[\log q_0(Z_1)\Big] &= \log\Big(q_0(0)\Big) \times P(Z_1 = 0|\mathcal{H}_0) + \log\Big(q_0(1)\Big) \times P(Z_1 = 1|\mathcal{H}_0) \\ &= \log\Big(q_0(0)\Big) \times q_0(0) + \log\Big(q_0(1)\Big) \times q_0(1), \end{aligned} \tag{3.62}$$

and

$$\begin{aligned}\mathbb{E}_0\Big[\log q_1(Z_1)\Big] &= \log\Big(q_1(0)\Big) \times P(Z_1 = 0|\mathcal{H}_0) + \log\Big(q_1(1)\Big) \times P(Z_1 = 1|\mathcal{H}_0) \\ &= \log\Big(q_1(0)\Big) \times q_0(0) + \log\Big(q_1(1)\Big) \times q_0(1). \end{aligned} \tag{3.63}$$

So we finally have

$$d_0 = q_0(0) \times \log\left(\frac{q_1(0)}{q_0(0)}\right) + q_0(1) \times \log\left(\frac{q_1(1)}{q_0(1)}\right). \tag{3.64}$$

In a similar fashion we may obtain d_1 as

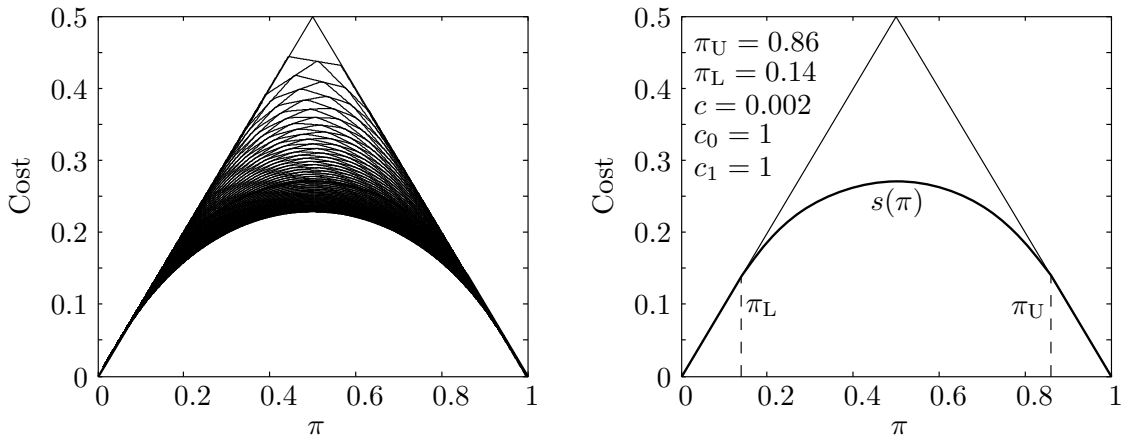
$$d_1 = q_1(0) \times \log\left(\frac{q_1(0)}{q_0(0)}\right) + q_1(1) \times \log\left(\frac{q_1(1)}{q_0(1)}\right). \tag{3.65}$$

Substituting (3.64) and (3.65) into Proposition 4, and by using of Wald's approximations to α and β (by considering (3.50) and (3.51) as equalities), we can compute the ARL.

3.5.3.4 Simulation results

Table 3.2 summarizes the most important experimental results for all three cases.

3.5.3.4.1 Case I: Detailed simulation results The numerical approximations to the minimal cost function corresponding to Case I are given in Figure 3.2, where Figure 3.2.a shows the first $N = 100$ iterations, $s_1(\pi), s_2(\pi), \ldots, s_{100}(\pi)$, obtained after applying Proposition 3 to the sequential detection problem considered in Case I. The converged minimal cost function, $s(\pi)$, as well as the optimal thresholds, π_{L} and π_{U}, are shown in Figure 3.2.b.

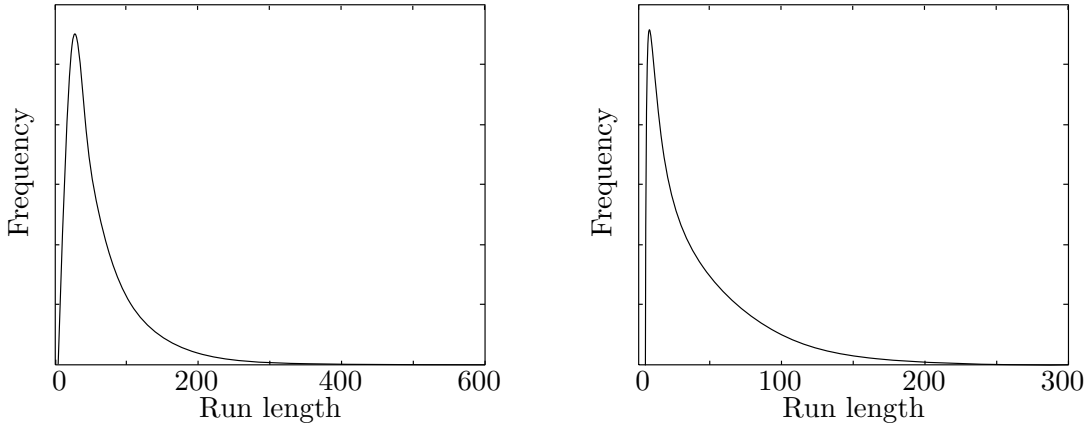


a. Case I: The first $N = 100$ iterations after applying Proposition 3 to the coin toss example.

b. Case I: Minimal expected cost function, $s(\pi)$, with optimal thresholds $\pi_{\mathrm{L}} = 0.14$ and $\pi_{\mathrm{U}} = 0.86$.

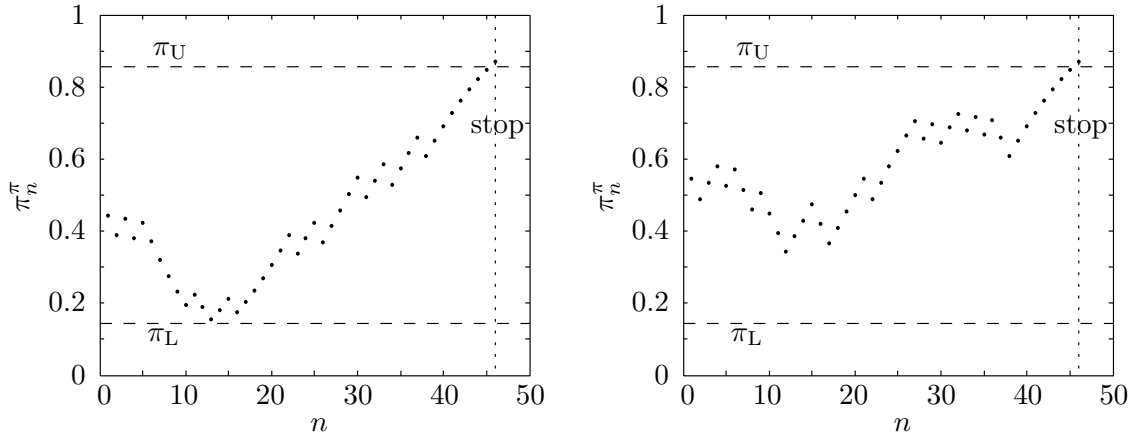
Figure 3.2: Case I: Minimal expected cost, $s(\pi)$, corresponding to both Case I.a and b.

Even though the minimal cost functions of Case I.a and Case I.b are identical, their ARLs are not. The distribution of the ARL as a function of the sampling cost is shown in Figure 3.3.a and Figure 3.3.b for Case I.a and Case I.b, respectively. The reason why we might expect Case I.b to have a lower ARL than Case I.a, is that we have additional *a priori* information; and so the posterior sequence reaches the exit thresholds faster.



a. Case I.a: Experimental ARL distribution. b. Case I.b: Experimental ARL distribution.

Figure 3.3: Case I: Experimental distributions of the ARLs.



a. Case I.a: Successful sequential test, $\tau = 46$. b. Case I.b: Successful sequential test, $\tau = 46$.

Figure 3.4: Case I: Experimental realisations of the sequential detection test.

Two realisations of the sequential detection test with $\pi_U = 0.86$ and $\pi_L = 0.14$ are given in Figure 3.4.a and Figure 3.4.b, both of which happened to terminate at $\tau = 46$ (compare with the experimental ARL of 70 and 52 for Case I.a and Case I.b, respectively). The correct decision was also made in both cases. It is interesting to note, however, how close the sequence in Figure 3.4.a came to being incorrectly classified (at $n = 13$).

The experimental probability of error as a function of the sampling cost is shown in Figure 3.5, along with the theoretical probability of error (using Wald's approximations).

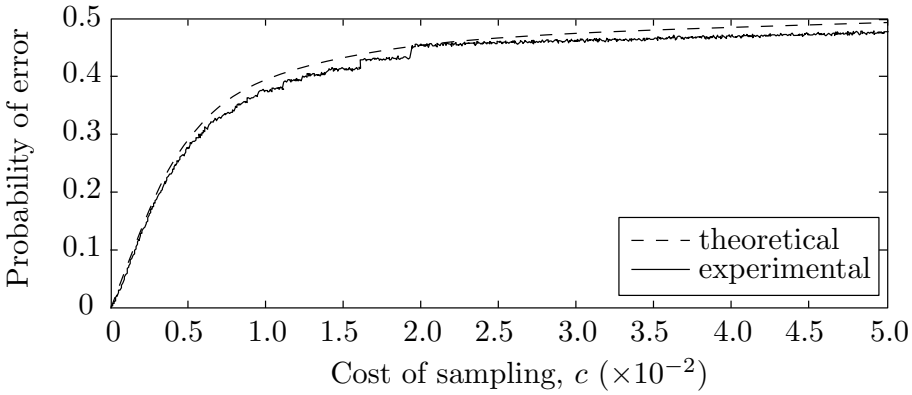


Figure 3.5: Case I.a: Probability of error as a function of the cost of sampling, c.

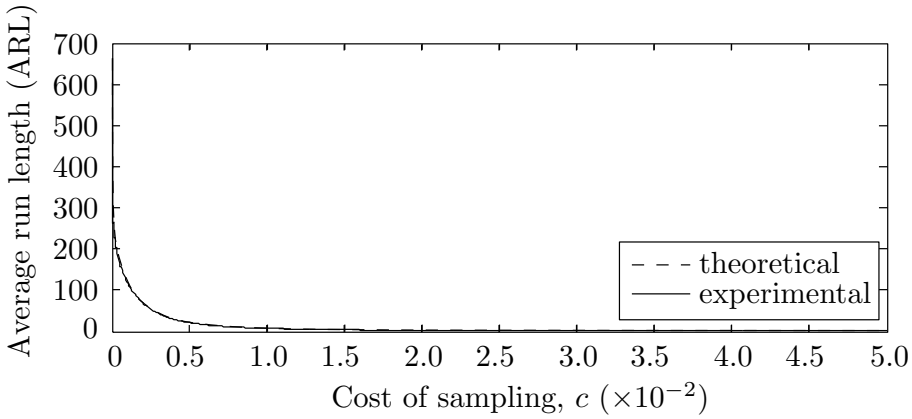


Figure 3.6: Case I.a: Experimental ARL as a function of the cost of sampling, c.

We notice that the experimental and theoretical results correspond fairly well over the entire range of costs considered. We can now use Figure 3.5 to determine an appropriate sampling cost for a desired probability of error. Recall that the results listed in Table 3.2 correspond to a cost of sampling $c = 2 \times 10^{-3} = 0.2 \times 10^{-2}$.

The ARL as a function of the sampling cost for Case I.a is presented in Figure 3.6, where it should be noted that the "theoretical" ARL has been computed using the experimental probabilities of error, and not Wald's approximations.

As we might have expected, the relationship between the ARL (of Figure 3.6) and the probability of error (given in Figure 3.5) is nonlinear. In this way, a small relaxation of the permissible probability of error can sometimes lead to a significant decrease in the ARL. For example, the ARL is reduced from roughly 650 down to just 100 with a only a 10 % increase in the probability of error. However, to further decrease the ARL from 100 down to 20 would require an additional 20 % increase to the probability of error.

Finally, the optimal thresholds, π_{U} and π_{L}, are given against the cost of sampling in Figure 3.7 on the next page. Once again we note that the thresholds are symmetric about the line $\pi = 1/2$, since equal costs of error were assumed in Case I.

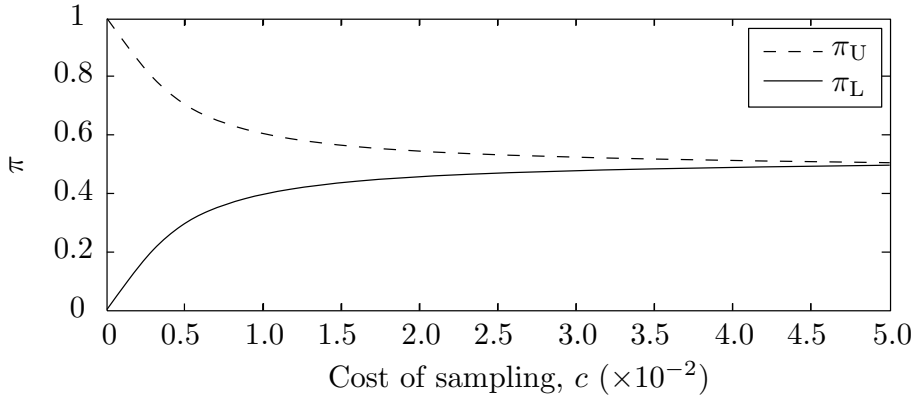


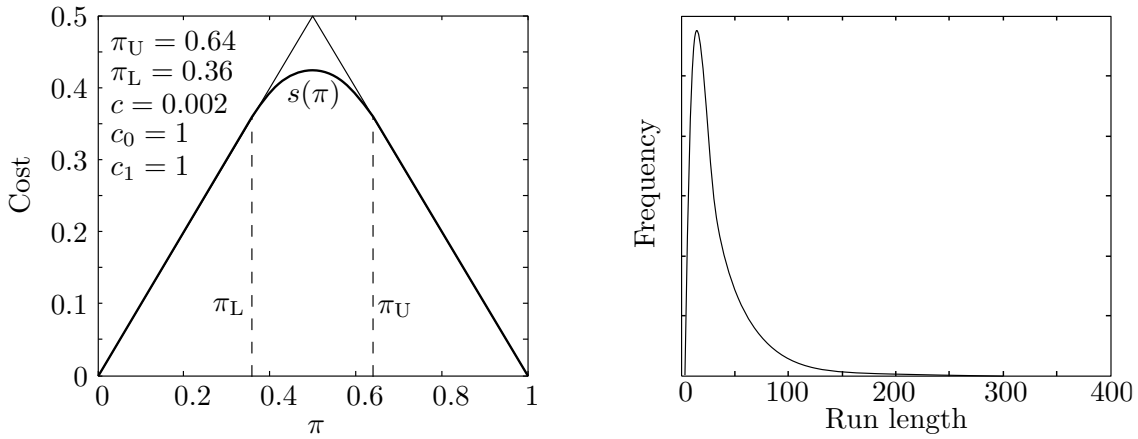
Figure 3.7: Case I.a: Optimal exit thresholds, π_{U} and π_{L}, as a function of the cost of sampling, c.

Remark 6 *With reference to Figure 3.7 we note that* $0 \leq \pi_L \leq 0.5$ *and* $0.5 \leq \pi_U \leq 1$. *According to [89, (4.38) and (4.39), p. 71] this result holds in general. However, it would appear as though the thresholds should instead be bounded according to*

$$0 \leq \pi_L \leq \frac{c_0}{c_0 + c_1} \quad \text{and} \quad \frac{c_0}{c_0 + c_1} \leq \pi_U \leq 1. \tag{3.66}$$

Furthermore, whenever $c_0 = 0$ *we set* $\pi_U = 0$, *and when* $c_1 = 0$ *we set* $\pi_L = 1$. *This scenario is however of very little (if any) practical importance.*

3.5.3.4.2 Case II: Detailed simulation results The minimal cost function for Case II is presented in Figure 3.8.a, along with the distribution of the ARL in Figure 3.8.b.



a. Case II: Minimal expected cost function, $s(\pi)$, with optimal thresholds $\pi_{\mathrm{L}} = 0.36$ and $\pi_{\mathrm{U}} = 0.64$.

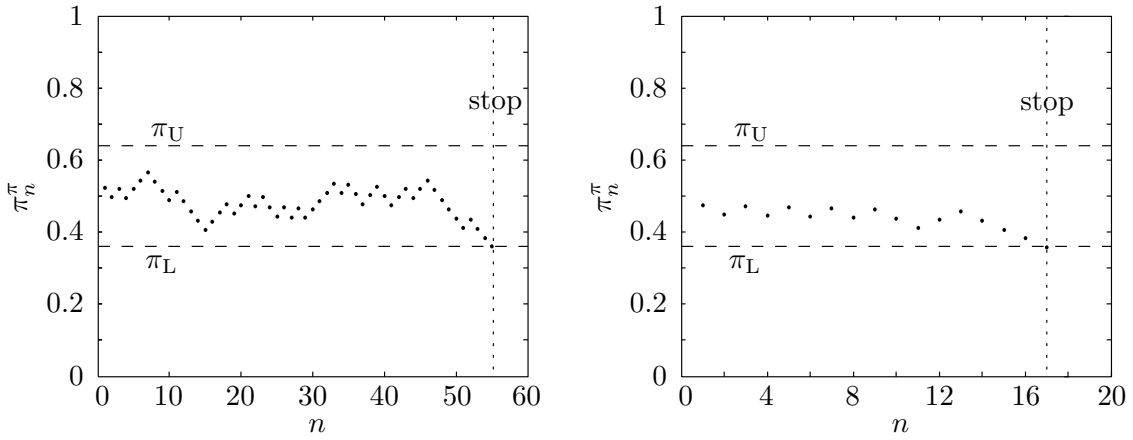
b. Case II: Experimental ARL distribution.

Figure 3.8: Case II: Minimal expected cost, $s(\pi)$, and the distribution of the ARL.

We immediately notice that the optimal thresholds, $\pi_{\mathrm{L}} = 0.36$ and $\pi_{\mathrm{U}} = 0.64$, are closer together than for Case I. Essentially this can be interpreted as meaning that, at

the time of stopping, we cannot be as sure about our decision as we were in Case I. Therefore we can expect either (i) shorter ARLs, or (ii) higher probabilities of error, or both. From Table 3.2 we see that Case II is indeed characterised by both shorter ARLs and larger probabilities of error.

Two realisations of the sequential test for Case II with optimal thresholds $\pi_L = 0.36$ and $\pi_U = 0.64$ are given in Figure 3.9, with stopping times $\tau = 56$ (slower than average) and $\tau = 17$ (faster than average). Both realisations resulted in the correct decision.



a. Case II: Successful sequential test, $\tau = 56$. b. Case II: Successful sequential test, $\tau = 17$.

Figure 3.9: Case II: Experimental realisations of the sequential detection test.

Figure 3.10 presents the probability of error as a function of the sampling cost for Case II. We once again notice a remarkable agreement between the theoretical and experimental results.

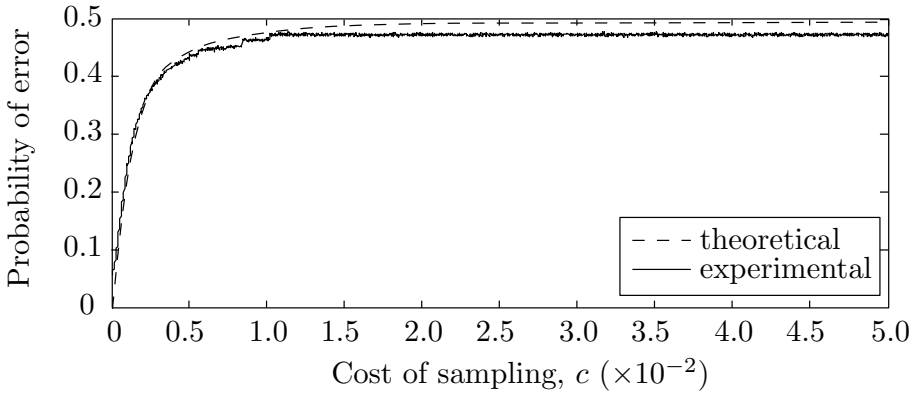


Figure 3.10: Case II: Probability of error as a function of the cost of sampling, c.

In comparison with the results presented for Case I (See Figure 3.5), we further notice that the probability of error rises much faster as the cost of sampling is increased. This might have been expected, since it is more difficult to distinguish between the two hypotheses under the probability mass functions considered in Case II than under those considered in Case I.

The ARL as a function of the sampling cost for Case II is presented in Figure 3.11. With reference to Figure 3.11, we may conclude that a "reasonable" range of sampling costs to consider for Case II might be anything less than about 0.005, which may even already exhibit too large a probability of error (> 0.4). Nevertheless, we have chosen to keep the range of sampling costs fixed for all the simulations presented here in order to facilitate an easy comparison between the different cases.

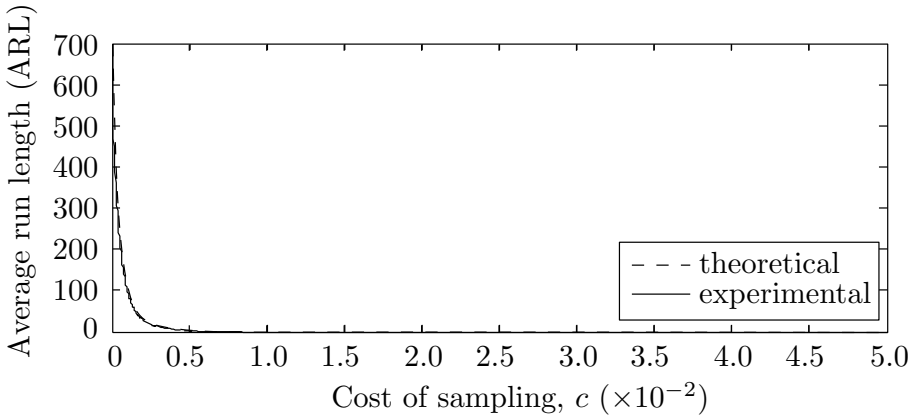


Figure 3.11: Case II: Experimental ARL as a function of the cost of sampling, c.

The optimal thresholds, π_{L} and π_{U}, given in Figure 3.12, once again satisfies $0 \leq \pi_{\mathrm{L}} \leq 0.5$ and $0.5 \leq \pi_{\mathrm{U}} \leq 1$, since $c_0 = c_1$, which also causes the thresholds to be symmetric about the line $\pi = 0.5$, as indicated previously.

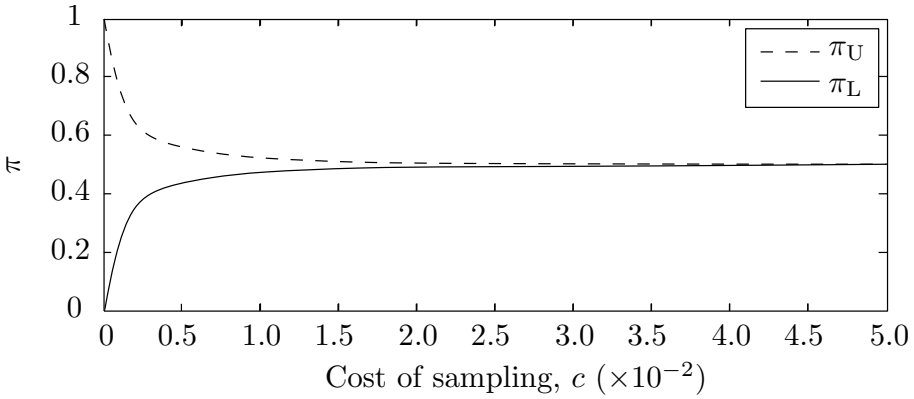
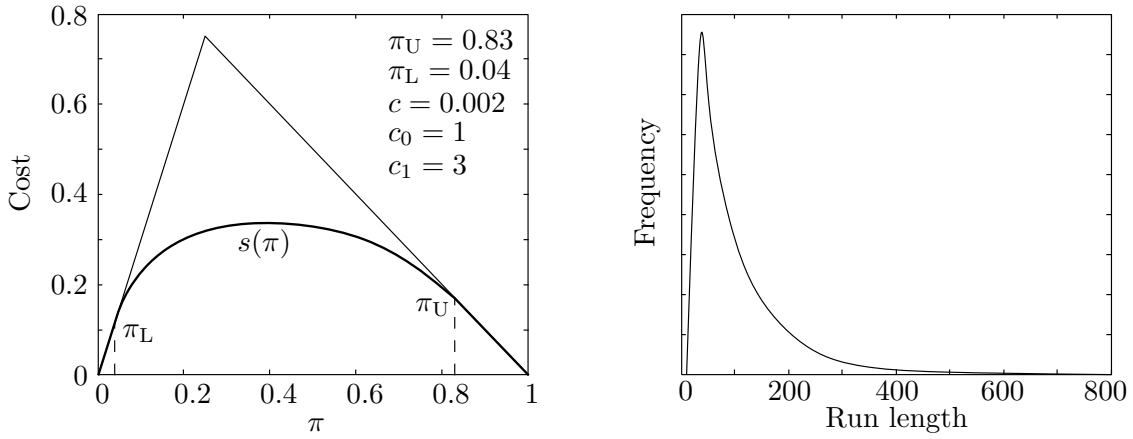


Figure 3.12: Case II: Optimal exit thresholds, π_{U} and π_{L}, as a function of the cost of sampling, c.

Remark 7 *It is very interesting to note that in both Figure 3.7 (of Case I) and Figure 3.12 (of Case II), the lower exit threshold, π_L, is exactly equal to the probability of error, shown in Figure 3.5 and Figure 3.10, respectively. This result definitely does not hold in general, but it might hold for equal costs of error—it remains to be established.*

3.5.3.4.3 Case III: Detailed simulation results Case III is characterised by unequal costs of error, namely $c_0 = 1$, and $c_1 = 3$. We therefore expect to see a much lower Type II probability of error than a Type I probability of error. Recall that a Type II error is "falsely rejecting $\mathcal{H}_1$", or equivalently, "falsely accepting $\mathcal{H}_0$". From this we may expect π_{L} to be relatively small (so that we won't accept $\mathcal{H}_0$ too hastily) and similarly we may expect π_{U} to be relatively small, so that we won't easily reject $\mathcal{H}_1$. This behaviour is reflected in the cost function for Case III, shown in Figure 3.13.a.

Since the overall cost of error in Case III is larger than for Case I ($1 + 3 > 1 + 1$),

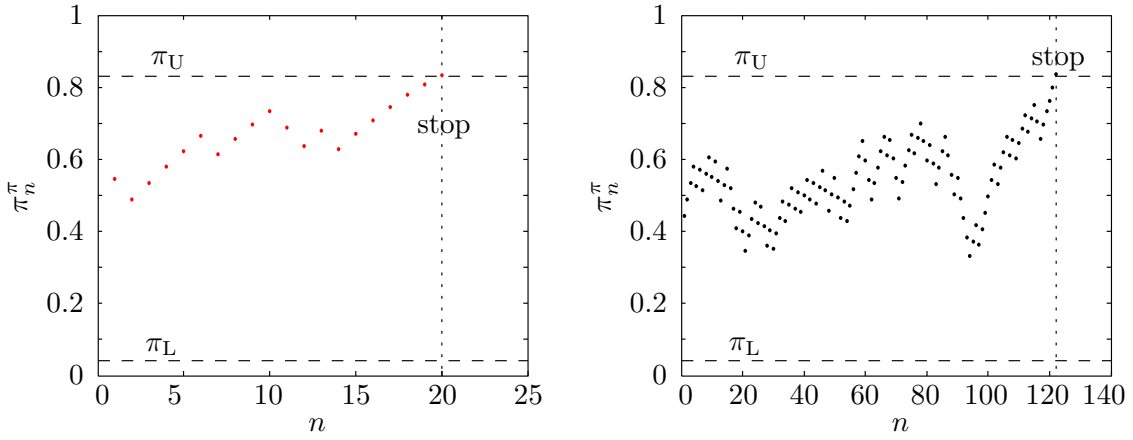


a. Case III: Minimal expected cost function, $s(\pi)$, with optimal thresholds $\pi_{\text{L}} = 0.04$ and $\pi_{\text{U}} = 0.83$.

b. Case III: Experimental ARL distribution.

Figure 3.13: Case III: Minimal expected cost, $s(\pi)$, and the distribution of the ARL.

the sequential detection problem of Case III effectively demands a lower probability of error, and we may consequently expect a larger ARL than for Case I. With reference to the results presented in Table 3.2 we see that Case III is indeed characterised by a lower probability of error (and a correspondingly larger ARL) than Case I, and the experimental ARL distribution for Case III is given in Figure 3.13.b.



a. Case III: Failed sequential test, $\tau = 20$.

b. Case III: Successful sequential test, $\tau = 122$.

Figure 3.14: Case III: Experimental realisations of the sequential detection test.

Two realisations of the sequential test for Case III, with optimal thresholds $\pi_{\mathrm{L}} = 0.04$ and $\pi_{\mathrm{U}} = 0.83$, is given in Figure 3.14.a and Figure 3.14.b. The test in Figure 3.14.a stopped at $\tau = 20$, where it decided (incorrectly!) in favour of $\mathcal{H}_1$—a Type I error. However, since the cost of making a Type I error is much lower than the cost of making a Type II error, this may actually not be *so* undesirable. The test in Figure 3.14.b on the other hand, stopped at $\tau = 121$, at which time it (correctly) declared $\mathcal{H}_1$.

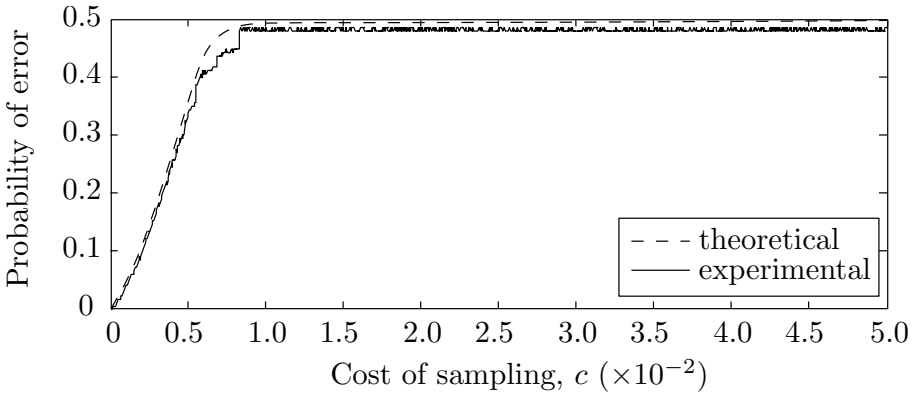


Figure 3.15: Case III: Probability of error as a function of the cost of sampling, c.

It might at first seem somewhat surprising that the probability of error (shown in Figure 3.15) grows more rapidly in Case III than in Case I (see Figure 3.5). However, as previously mentioned, the cost of errors in Case III weighs twice as much as in Case I. Therefore, a sampling cost of, say, $c = 0.01$ in Case III, would induce the same probability of error as a cost, $c = 0.02$ in Case I.

The experimental ARL for Case III is given in Figure 3.16, which once again corresponds very well with the theoretical results.

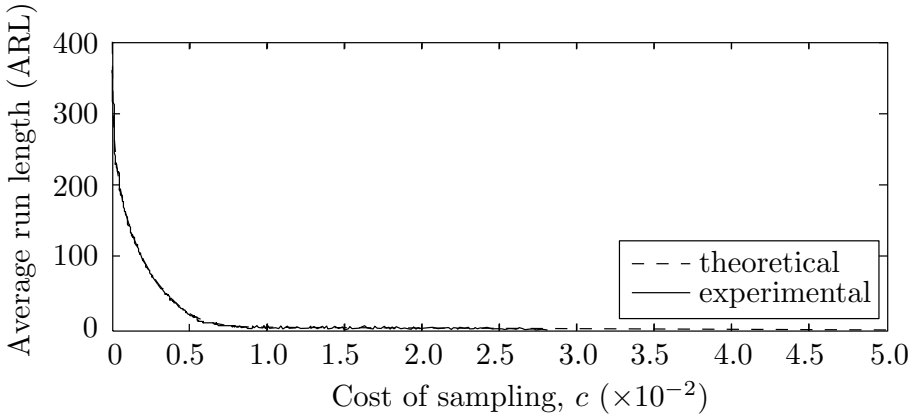


Figure 3.16: Case III: Experimental ARL as a function of the cost of sampling, c.

Finally, the optimal thresholds π_{L} and π_{U} are given in Figure 3.17, from which it is clear that the thresholds are indeed bounded as in (3.66). That is, $0 \leq \pi_{\mathrm{L}} \leq 0.25 \leq \pi_U \leq 1$.

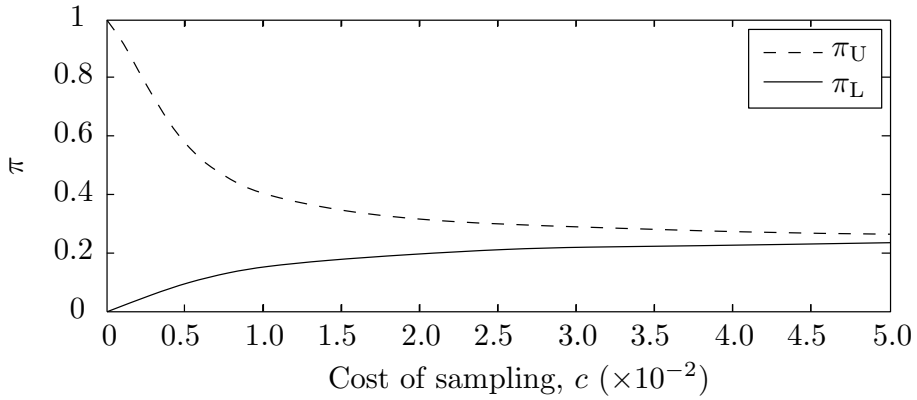


Figure 3.17: Case III: Optimal exit thresholds, π_U and π_L, as a function of the cost of sampling, c.

3.6 SUMMARY

In this chapter we have presented the sequential detection task of deciding between two simple statistical hypotheses. The optimal solution was given in Theorem 2; and Proposition 3 presented a computational method to determine the minimal cost function, $s(\pi)$, from which the optimal thresholds, π_L and π_U, could be derived.

The relationship between the Bayesian formulation and Wald's formulation was also discussed briefly, followed by a number of illustrative examples of sequential detection for a simple coin-tossing problem.

Chapter 4

Sequential Land Cover Classification

> "*Engineering is the art of modelling materials we do not wholly understand, into shapes we cannot precisely analyse so as to withstand forces we cannot properly assess, in such a way that the public has no reason to expect the extent of our ignorance.*"
>
> *A. R. Dykes*
> British Institution of Structural Engineers, 1976.

We have seen that current research efforts in land cover classification are primarily concerned with improving the classification accuracy, and that a need exists to be able classify *as quickly as possible*. We have also seen that sequential detection provides an attractive (indeed, an optimal) solution to the problem of minimising the expected detection (or classification) delay for a given probability of error.

Our task in this chapter is then to combine these two observations, to arrive at a *sequential land cover classification* strategy, in which the classification delay and probability of error can easily be adjusted to suit a particular application.

4.1 INTRODUCTION

The first step towards the development of a sequential land cover classification strategy necessarily requires some sort of statistical model for each of the land cover classes to be considered. To keep the development simple, we will restrict our attention to the two-class problem as presented in Chapter 3. More specifically, we will consider the two classes described in section 2.4.4, namely *residential* areas and *natural vegetation*.

4.1.1 Problem statement

A sequential land cover classification strategy is required. More specifically, we are interested in the development of two-class sequential land cover classifiers in which the relative importance between the classification delay and the probability of error must be adjustable.

The development of such classifiers is made difficult by the heterogeneity of land cover types, which are characterised by large intra-class variations, as well as the complex remote sensing observation process. It is expected that the multispectral observations considered in this study will be spectrally, temporally, as well as spatially dependent on one another, making the development of a representative statistical model very difficult.

4.1.2 Chapter overview

The development of statistical land cover models will be considered in section 4.2, where an i.i.d. model (section 4.2.1), as well as a time-varying model (section 4.2.2) will be presented. Thereafter, the formulation of the land cover classification task will be presented in section 4.3, where both maximum likelihood classification, as well as sequential classification will be considered in section 4.3.1 and section 4.3.3, respectively. Finally, the foundation for a multispectral, time-varying sequential classification strategy will be developed in section 4.4.1.

4.2 STATISTICAL LAND COVER MODELS

Two types of statistical land cover models will be presented in this section, namely (i) models for which it will be assumed that we obtain i.i.d. observations after applying appropriate preprocessing, and (ii) time-varying models in which the probability density functions will not be assumed identical for different times of the year.

Statistical land cover models will be developed for *residential* as well as *vegetation* classes derived from MODIS spectral reflectance data, as presented in section 2.4.4, where each MODIS pixel forms a multispectral time series as shown in Figure 4.1.

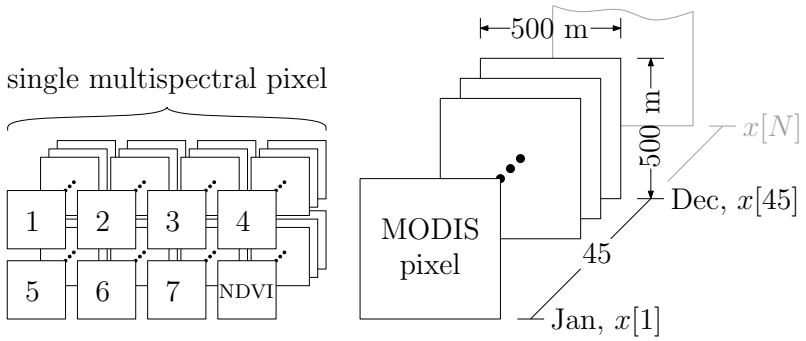


Figure 4.1: Multispectral time series data representation for a single MODIS pixel.

4.2.1 Semi-parametric stationary land cover model

Our primary aim for the first statistical land cover model is to describe (and model) a sequence of i.i.d. real-valued observations $\{Z_k; k = 1, 2, \ldots\}$ for each of the eight spectral bands. To arrive at such a sequence of observations, we will have to preprocess the observations from the MODIS sensor first.[1] More specifically, for each spectral band we will subtract the ensemble average (for a particular class) from the sequence of observations to transform the time series into a pseudo-stationary, zero mean process. Note that this preprocessing step was motivated by the preliminary data analysis of the MODIS data, presented in section 2.4. Such a transformation obviously does not cause arbitrary data to become independent or identically distributed, such that the underlying assumption of independent observations remains important.

Note further that in order to develop a land cover classifier in which these preprocessing steps are first applied, we will have to consider *two* sequences of observations for each spectral band i; each preprocessed according to the assumed underlying land cover class. That is, we must consider $\{Z_k^\theta; k = 1, 2, \ldots\}$ where $\theta \in \{0, 1\}$ is the land cover class.

Recall that each MODIS pixel (which we will denote by $\mathcal{X}$) has eight associated time series, such that $\mathcal{X}_j = \{X_k; k = 1, 2, \ldots\}$, where $X_k \in \mathbb{R}^8$ is the multispectral observation at time k, for a particular pixel $\mathcal{X}_j$. To be able to refer to single or collections of spectral bands, we first need to define the projection operator pr on the sequence $x = (x_1, x_2, \ldots, x_N)$ as follows:

$$\underset{i\in\mathcal{I}}{\text{pr}}\, x := (x_i)_{i\in\mathcal{I}}, \tag{4.1}$$

such that $\text{pr}_i\, x$ is simply the ith component of the sequence x. In this way we can express a single band i or an ordered collection of bands $\mathcal{I} = (i_1, i_2, \ldots, i_n)$ within a particular pixel $\mathcal{X}$ as

$$\mathcal{X}(\mathcal{I}) := \left\{ \underset{i\in\mathcal{I}}{\text{pr}}\, X_k; k = 1, 2, \ldots \right\}. \tag{4.2}$$

Furthermore, the kth multispectral observation X_k corresponding to a particular pixel $\mathcal{X}$ is then simply given by $\text{pr}_k\, \mathcal{X}$, such that $\text{pr}_k\, \mathcal{X}_j(i) \equiv X_k(i) \subseteq \mathcal{X}_j$ corresponds to the kth observation in the ith spectral band of the jth pixel.

The expectation $\mathbb{E}_{(k,\theta)}\{X\}$ is then used to denote the ensemble average of the multispectral observation X for a particular class θ, and similarly, $\mathbb{E}_\theta\{X\}$ is its mean. We therefore preprocess a sequence of multispectral observations $\mathcal{X} = \{X_k; k = 1, 2, \ldots\}$ in the following manner

$$Z_k^\theta = X_k - \mathbb{E}_{(k,\theta)}\{X\}, \quad \theta \in \{0, 1\}, \quad k = 1, 2, \ldots, \tag{4.3}$$

to obtain the multispectral i.i.d. sequences $\mathcal{Z}_\theta = \{Z_k^\theta; k = 1, 2, \ldots\}$, for $\theta \in \{0, 1\}$.

[1] Note that by the "observations from the MODIS sensor" we refer to the already-preprocessed-in-some-manner eight-daily composite spectral reflectance data from the MCD43A4 product, and not the raw data from the sensor itself.

We can now (finally!) proceed to develop a statistical land cover model to describe the sequences $\mathcal{Z}_\theta := \{Z_k^\theta; k = 1, 2, \ldots\}$ for each of the two land cover classes.

Let $\mathcal{D}_\theta$ denote the set of pixels, that is, the set of all sequences of multispectral observations $\{X_k; k = 1, 2, \ldots, N\}$ belonging to the land cover class θ. Furthermore, we define the set of pixels $\mathcal{D}_\theta(\mathcal{I})$ corresponding to a subset $\mathcal{I}$ of spectral bands as

$$\mathcal{D}_\theta(\mathcal{I}) := \Big\{\mathcal{X}(\mathcal{I})\Big|\mathcal{X} \in \mathcal{D}_\theta\Big\}, \quad \theta \in \{0, 1\}. \tag{4.4}$$

Next we introduce the sets (corresponding to $\theta = 0$ and $\theta = 1$) of all unordered, preprocessed observations corresponding to a collection $\mathcal{I}$ of spectral bands as

$$\mathcal{G}_\theta(\mathcal{I}) = \Big\{Z \in \mathcal{Z}_\theta(\mathcal{I})\Big|\mathcal{X}(\mathcal{I}) \in \mathcal{D}_\theta(\mathcal{I})\Big\}, \quad \theta \in \{0, 1\}, \tag{4.5}$$

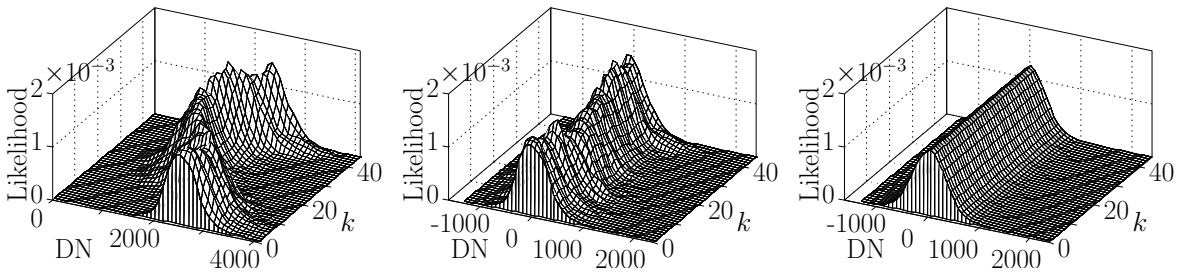
where $\mathcal{Z}_\theta$ and $\mathcal{X}$ are related by (4.3), and $\mathcal{Z}_\theta(\mathcal{I})$ is defined in the obvious manner:

$$\mathcal{Z}_\theta(\mathcal{I}) = \left\{ \operatorname*{pr}_{i\in\mathcal{I}} Z_k^\theta; k = 1, 2, \ldots \right\}, \quad \theta \in \{0, 1\}. \tag{4.6}$$

Note that the elements of $\mathcal{D}_\theta$ are multispectral *sequences*, whereas the elements of $\mathcal{G}_\theta$ are multispectral observations, or *vectors*, and finally, that the elements of $\mathcal{G}_\theta(i), 1 \leq i \leq 8$ are real-valued *scalars*.

Finally then, we only have to determine the joint pdf of each set $\mathcal{G}_\theta(\mathcal{I})$ to describe (statistically) the data within each collection $\mathcal{I}$ of spectral bands, and for each land cover class θ.

The preprocessing procedure (for the vegetation class, band 2) is shown in Figure 4.2.



a. Time-varying model. b. Model after preprocessing. c. Stationary (i.i.d.) model.

Figure 4.2: Preprocessing steps and i.i.d. model for vegetation class, band 2.

With the dataset described in section 2.4.4, $\mathcal{D}_0$ (corresponding to the residential class) contains 333 pixels, each consisting of 368 multispectral observations, such that after preprocessing, $\mathcal{G}_0$ consists of $333 \times 368 = 122,544$ multispectral observations (which is equal to $122,544 \times 8 = 980,352$ sample points). Similarly, $\mathcal{G}_1$, which corresponds to the vegetation class, contains $592 \times 368 = 197,136$ multispectral observations (or $1,577,088$ sample points).

The single-band pdfs obtained in this manner are shown in Figure 4.3 for all eight spectral bands, both land cover classes, and using all $N = 368$ observations. It should be mentioned that the model is termed "semi-parametric", since the data was not assumed to be normally distributed. Instead, we employed *kernel density estimation*, so that interesting features in the data could be captured more accurately.

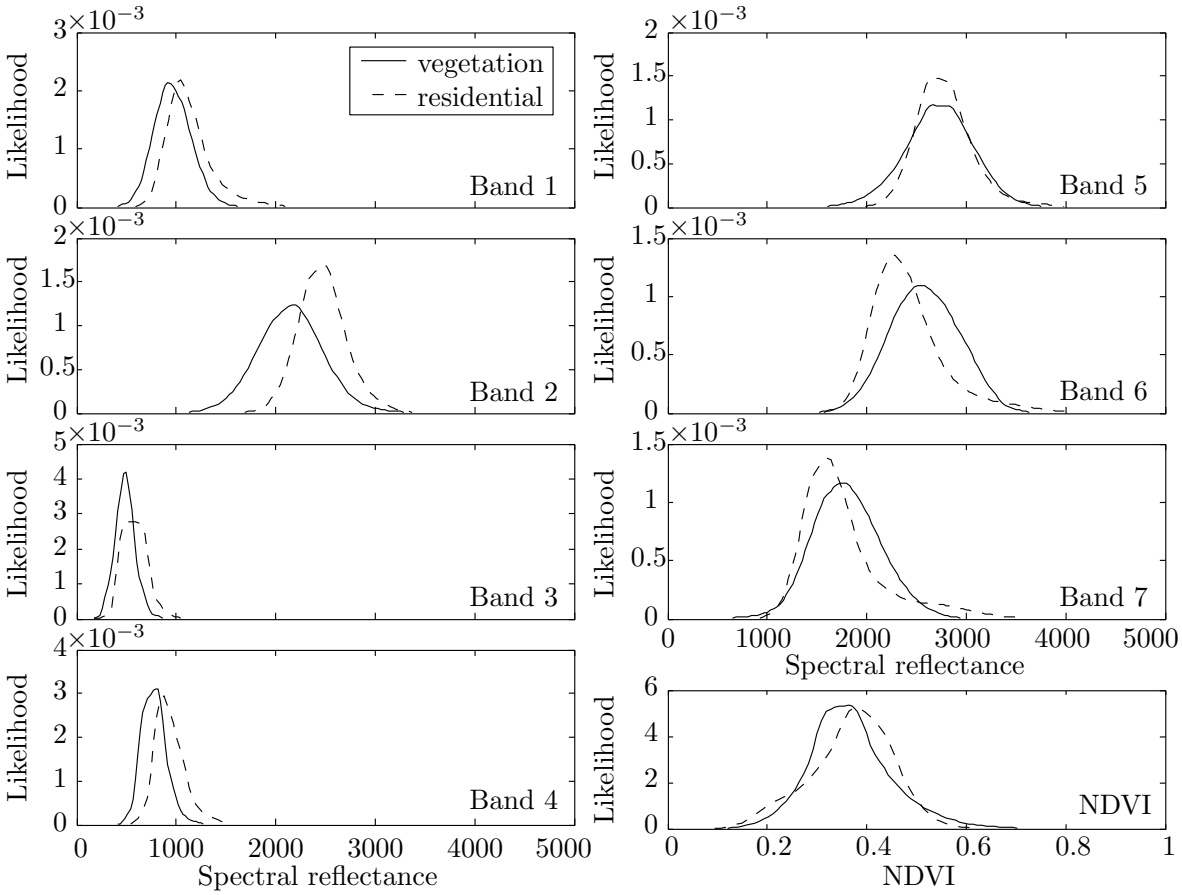


Figure 4.3: Marginal probability density functions for the i.i.d. land cover model.

With reference to Figure 4.3 we can already make a few observations: firstly, it seems reasonable to expect that band 2 will perform better than other spectral bands during classification, since the dichotomy between the two land cover classes seems to be the greatest in this band. Similarly we might expect band 5 to be among the worst bands to use for classification, since the classes appear to be the least separable in this band.

Secondly, it is interesting (albeit expected) to see that NDVI also provides poor separability between the two land cover classes considered. Unfortunately this was caused by the preprocessing of (4.3)—as shown in Figure 2.14.b on page 44, the mean components of the two land cover classes are approximately equal under NDVI, so that we lose almost all of the valuable information when we subtract the ensemble average.

4.2.2 Semi-parametric time-varying land cover model

The second model that we will consider will be modified from section 4.2.1 to allow for different pdfs at different times of the year. In other words, we will consider a *time-varying* land cover model. More specifically we will assume that there is no inter-annual variation within the two land cover classes considered, but that the sequence of multispectral observations $\{X_k; k = 1, 2, \ldots\}$ are independently generated, at each time step k, by a pdf q_θ^k, where θ is the true underlying land cover class. Furthermore, since we assume no inter-annual variation, we have that $q_\theta^k = q_\theta^{k+j}, \forall k = 1, 2, \ldots,$ and where j is the (constant) number of observation periods during the year. In this study, we have $j = 45$ observation periods during the year (see again Figure 4.1 on page 76).

As defined previously, let $\mathcal{D}_\theta$ denote the set of all sequences of multispectral observations $\mathcal{X} = \{X_k; k = 1, 2, \ldots, N\}$ belonging to the land cover class θ.

Then, for each observation period k, $1 \leq k \leq j$ (with j the number of periods in a year), and for each land cover class $\theta \in \{0, 1\}$, we define the set $\mathcal{G}_{k,\theta}$ as follows:

$$\mathcal{G}_{k,\theta} = \left\{ X \in \underset{k+nj}{\text{pr}}\, \mathcal{X} \middle| \mathcal{X} \in \mathcal{D}_\theta \right\}, \quad n = 0, 1, \ldots, j \in \mathbb{N}, 1 \leq k \leq j, \tag{4.7}$$

such that $\mathcal{G}_{k,\theta}$ corresponds to the set of all multispectral observations for a particular time k during the year, and a particular land cover class, θ. Note that $x \in \mathcal{G}_{k,\theta} \implies x \in \mathbb{R}^8$.

Next, we consider the set of observations $\mathcal{G}_{k,\theta}(\mathcal{I})$ corresponding to a particular time, k, during the year, a particular class, θ, and a selection of p spectral bands, $\mathcal{I}$:

$$\mathcal{G}_{k,\theta}(\mathcal{I}) = \left\{ \underset{i \in \mathcal{I}}{\text{pr}}\, X = X(\mathcal{I}) \middle| X \in \mathcal{G}_{k,\theta} \right\}, \tag{4.8}$$

for which it clearly it holds that $\dim y = p, \forall y \in \mathcal{G}_{k,\theta}^{\mathcal{I}}$.

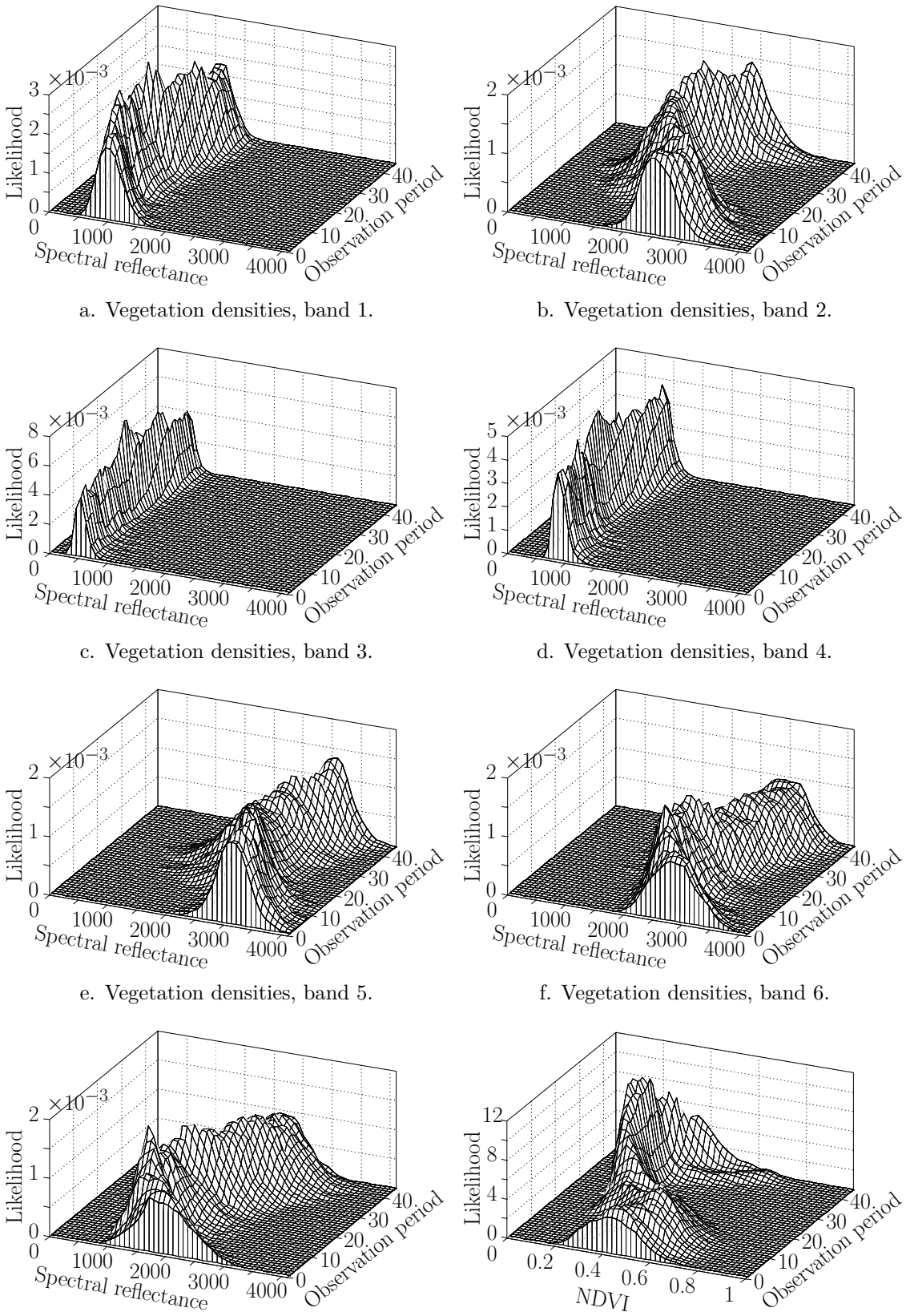
The multispectral time-varying land cover model is then obtained by finding the joint pdf of $\mathcal{G}_{k,\theta}(\mathcal{I})$ for each k, $1 \leq k \leq j = 45$, and for each land cover class $\theta \in \{0, 1\}$. In particular, we have for any single spectral band i: $\mathcal{G}_{k,\theta}(i) = \{X(i) | X \in \mathcal{G}_{k,\theta}\}$, which is used to derive a marginal pdf for the particular observation period and land cover class.

Time-varying land cover models of the type described in this section have been designed for every possible combination of spectral bands in order to experimentally determine the best single bands, as well as the best combinations of bands for classification.[2] The marginal pdfs for the vegetation class are presented in Figure 4.4 on the following page, and for the residential class in Figure 4.5 on page 82.

4.3 STATISTICAL LAND COVER CLASSIFICATION

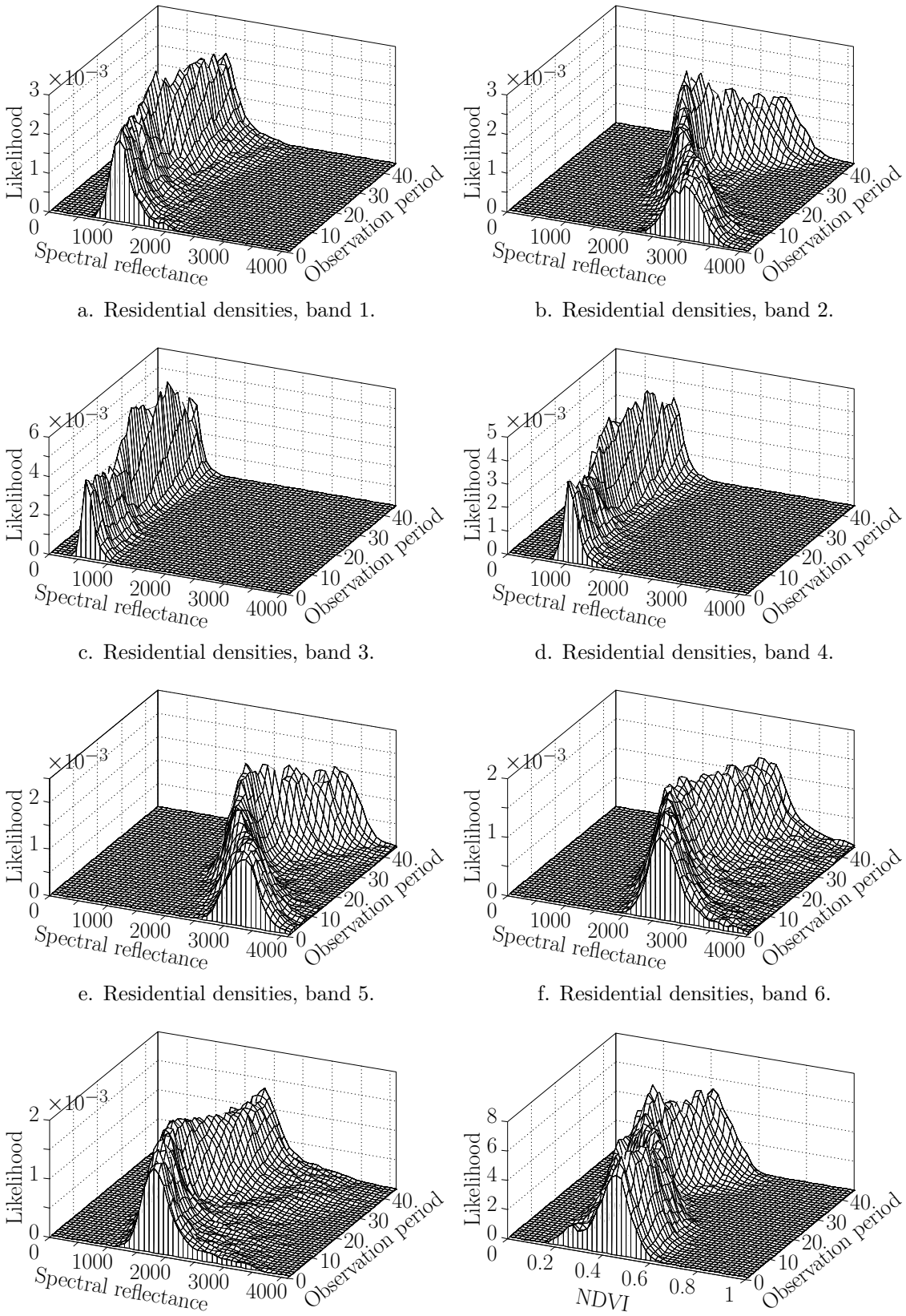
In this section we will formulate the land cover classification task using the statistical models developed in section 4.2. In particular, we will perform maximum likelihood

[2] I.e, a total of $2 \times (8 + 28 + 56 + 70 + 56 + 28 + 8 + 1) = 510$ time-varying models has been designed.



a. Vegetation densities, band 1.
b. Vegetation densities, band 2.
c. Vegetation densities, band 3.
d. Vegetation densities, band 4.
e. Vegetation densities, band 5.
f. Vegetation densities, band 6.
g. Vegetation densities, band 7.
h. Vegetation densities, NDVI.

Figure 4.4: Marginal probability density functions for vegetation class.



a. Residential densities, band 1.

b. Residential densities, band 2.

c. Residential densities, band 3.

d. Residential densities, band 4.

e. Residential densities, band 5.

f. Residential densities, band 6.

g. Residential densities, band 7.

h. Residential densities, NDVI.

Figure 4.5: Marginal probability density functions for residential class.

classification on (i) all single band stationary i.i.d. models, (ii) various multispectral stationary i.i.d. models, (iii) all single band time-varying models, (iv) all combinations of time-varying multispectral models, and we will (v) perform sequential detection on all the single band stationary i.i.d. models.

Due to several factors (the most important of which are the current time constraints, as well as some remaining numerical difficulties discussed in section 4.4.2), we will not perform multispectral sequential classification in this study, even though the most important theoretical results and framework have already been given. The implementation and evaluation of multispectral sequential methods are instead planned for future work.

4.3.1 Single band maximum likelihood classification

Maximum likelihood classification (which coincides with MAP classification under the assumption of equiprobable classes) has been performed on both the stationary and time-varying models. In each case, it is assumed that we observe a (truncated) sequence of real-valued observations $\{X_k(i); k = 1, 2, \ldots, N\}$, where i is the particular spectral band, which will be dropped from our notation in the rest of this section. That is, for some spectral band i, we want to classify a sequence of observations $\{X_k; k = 1, 2, \ldots, N\}$ as belonging either to class $\theta = 0$ (residential), or to class $\theta = 1$ (vegetation).

4.3.1.1 Stationary (i.i.d.) land cover model

Consider the sequence of real-valued observations $\{X_k; k = 1, 2, \ldots, N\}$ corresponding to some spectral band i. After applying the preprocessing step given in (4.3), we have two sequences $\{Z_k^\theta; k = 1, 2, \ldots, N\}$, $\theta = 0, 1$ of preprocessed, i.i.d., real-valued observations that obey one of two statistical hypotheses:

$$\mathcal{H}_0 \quad : Z_k^0 \sim Q_0, \quad k = 1, 2, \ldots$$

versus

$$\mathcal{H}_1 \quad : Z_k^1 \sim Q_1, \quad k = 1, 2, \ldots$$

where Q_0 and Q_1 are two probability distributions with associated probability densities q_0 and q_1, respectively. Further assume that hypothesis $\mathcal{H}_1$ occurs with prior probability π, and $\mathcal{H}_0$ with prior probability $1 - \pi$. To account for the preprocessing step of (4.3), we must modify the hypothesis test as follows. We redefine the posterior sequence as

$$\pi_n^\pi = \frac{\pi_{n-1}^\pi q_1(Z_k^1)}{\pi_{n-1}^\pi q_1(Z_k^1) + (1 - \pi_{n-1}^\pi) q_0(Z_k^0)}, \quad n = 1, 2, \ldots \tag{4.9}$$

where $\pi_0^\pi = \pi$, and where both preprocessed sequences are now used to compute π_n^π.

The maximum likelihood solution to the classification task is then given by

$$\delta_n = \begin{cases} 0 \ \text{(i.e. residential)}, & \text{if } \pi_n^\pi \leq 0.5 \\ 1 \ \text{(i.e. vegetation)}, & \text{if } \pi_n^\pi > 0.5. \end{cases} \tag{4.10}$$

We might reasonably expect the classification accuracy of this model to be slightly better than that of the minimum distance classifiers presented in section 2.4 (except for bands such as NDVI, in which the preprocessing step removed most of the valuable information), but the expected dependence between spectral bands, as well as the dependence between consecutive observations have not been adequately dealt with. Nevertheless, the classification results obtained when using this model are given in section 5.1

4.3.1.2 Time-varying land cover model

With the time-varying model we once again obtain a sequence of real-valued observations $\{X_k; k = 1, 2, \ldots, N\}$ corresponding to some spectral band i, that obeys one of two statistical hypotheses:

$$\mathcal{H}_0 \quad : X_k \sim Q_0^k, \quad k = 1, 2, \ldots$$

versus

$$\mathcal{H}_1 \quad : X_k \sim Q_1^k, \quad k = 1, 2, \ldots$$

where for each observation period k, Q_0^k and Q_1^k are two probability distributions with associated probability densities q_0^k and q_1^k, respectively. Further assume that hypothesis $\mathcal{H}_1$ occurs with prior probability π, and $\mathcal{H}_0$ with prior probability $1 - \pi$.

Since we use the sequence of observations $\{X_k; k = 1, 2, \ldots\}$ directly, and since each time period has a corresponding set of densities (depending on the underlying land cover class), we need to once again redefine the posterior sequence in the following manner:

$$\pi_n^\pi = \frac{\pi_{n-1}^\pi q_1^n(X_n)}{\pi_{n-1}^\pi q_1^n(X_n) + (1 - \pi_{n-1}^\pi) q_0^n(X_n)}, \quad n = 1, 2, \ldots, \tag{4.11}$$

and where, as per usual, $\pi_0^\pi = \pi$.

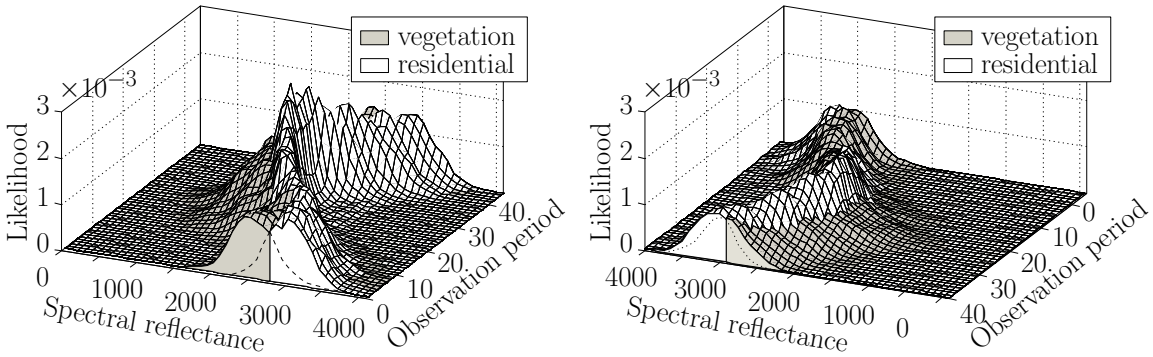
The maximum likelihood solution of the time-varying classification task is then, similar to (4.10), given by

$$\delta_n = \begin{cases} 0 \ \text{(i.e. residential)}, & \text{if } \pi_n^\pi \leq 0.5 \\ 1 \ \text{(i.e. vegetation)}, & \text{if } \pi_n^\pi > 0.5, \end{cases} \tag{4.12}$$

with the only difference being the computation of the posterior sequence π_n^π.

One immediately apparent advantage of the time-varying models considered here (as opposed to the stationary models of the previous section), is that the seasonal biases in classification ability is preserved to some extent. For example, Figure 4.6 shows the time-varying behaviour of the residential and vegetation classes for band 2, from which it is clear that some periods of the year (such as spring and early summer) are more separable than others.

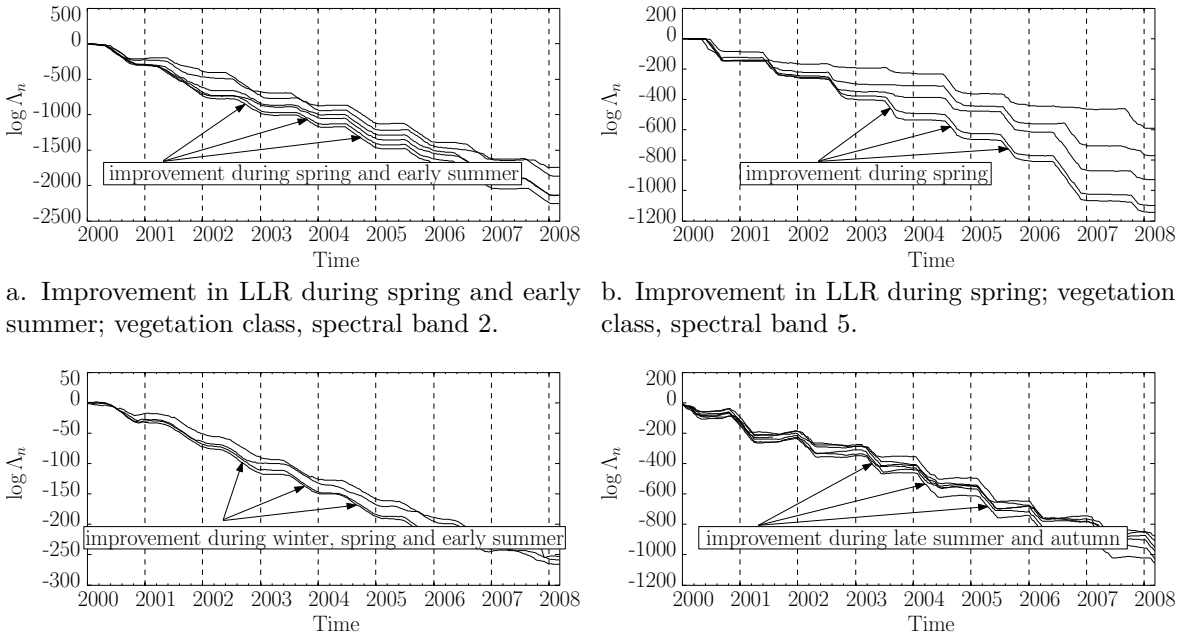
In fact, different bands have different periods during which they are easily separable. Figure 4.7 shows the Log Likelihood Ratios (LLRs) as a function of time for several



a. View A. b. View B.

Figure 4.6: Time-varying probability density function ambiguity for MODIS band 2.

vegetation pixels in four spectral bands, in which a sharp decline in the LLR indicates that the classes are easily separable, and a plateau indicates that the classes are difficult to separate.



a. Improvement in LLR during spring and early summer; vegetation class, spectral band 2.

b. Improvement in LLR during spring; vegetation class, spectral band 5.

c. Improvement in LLR during winter, spring and early summer; vegetation class, spectral band 4.

d. Improvement in LLR during late summer and autumn; vegetation class, NDVI.

Figure 4.7: Improvement in the classification ability of the vegetation class during different times of the year, for several spectral bands.

We can also evaluate the class separability based purely on the pdfs. The Bhattacharya coefficient is commonly used as an indicator of class separability, and is given as:

Definition 1 (Bhattacharya coefficient) *Let p and q denote two probability density functions (not necessarily distinct). Then the Bhattacharya coefficient (BC) between p*

and q is defined as:

$$BC(p,q) = \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, dx. \tag{4.13}$$

However, the Bhattacharya coefficient does not satisfy the triangle inequality, so we choose to consider the related Hellinger distance instead (which *does* satisfy the triangle inequality). The Hellinger distance ($0 \leq HD \leq 1$) between two (continuous or discrete) probability density functions can be expressed in terms of the Bhattacharya coefficient:

Definition 2 (Hellinger distance) *Let p and q denote two probability density functions. Then the Hellinger distance between p and q is defined as:*

$$\begin{aligned} HD(p,q) &= \sqrt{1 - BC(p,q)} \\ &= \left[1 - \int_{-\infty}^{\infty} \sqrt{p(x)q(x)}\, dx\right]^{1/2}. \end{aligned} \tag{4.14}$$

A Hellinger distance of $HD \approx 0$ indicates that the classes are not separable, whereas a distance $HD \approx 1$ indicates that the classes are trivially separable. Figure 4.8 gives the Hellinger distances between the residential and vegetation classes for the entire year. Overall the results correspond well to those presented in Figure 4.7. However, the plateaus of Figure 4.7 can be attributed to the temporal dependencies between observations, which were not captured in the pdfs used to derive the Hellinger distances.

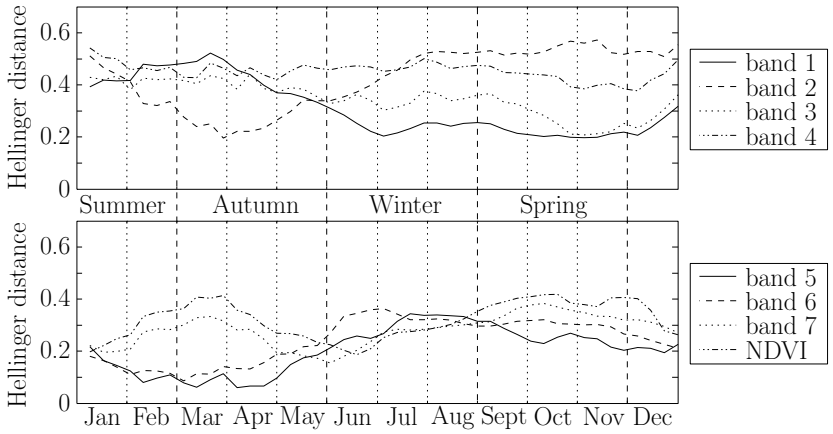


Figure 4.8: Hellinger distances indicating class separability over time.

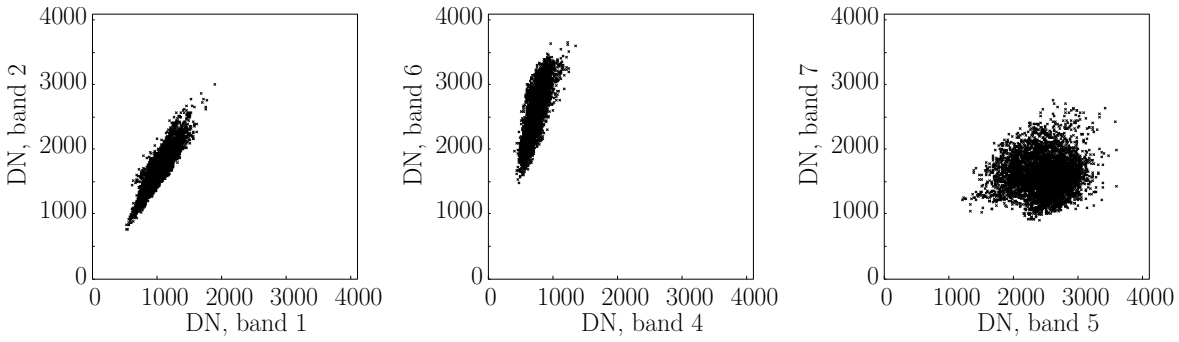
The classification results obtained when using this model are also given in section 5.1.

4.3.2 Multispectral maximum likelihood classification

The extension to the multispectral classification task is really quite simple (but the task of estimating high dimensional probability density functions is definitely not). Nevertheless,

assuming that we have representative multispectral models, the classification task is exactly the same as in the single band case, except that the sequence of observations $\{X_k; k = 1, 2, \ldots\}$ is now understood to contain multispectral observations: $X_k \in \mathbb{R}^p$, for some ordered collection $\mathcal{I} = (i_1, i_2, \ldots, i_p)$ of p spectral bands.

It is also perhaps worthwhile to give some empirical support or motivation for considering multispectral models here. Figure 4.9 presents three randomly selected scatter diagrams for dual-band multispectral observations at different times of the year.



a. Scatter diagram ($N = 592$) between bands 1 and 2, $k = 35$.
b. Scatter diagram ($N = 592$) between bands 4 and 6, $k = 32$.
c. Scatter diagram ($N = 592$) between bands 5 and 7, $k = 22$.

Figure 4.9: Scatter diagrams of several spectral bands, at different times of the year.

From Figure 4.9.a we can clearly see that there is a strong correlation between spectral bands 1 and 2 at time $k = 35$, and that there is little correlation between bands 5 and 7 at time $k = 22$. Furthermore, bands 4 and 6 exhibit some correlation at time $k = 32$ (see Figure 4.9.b), but not as much as shown in Figure 4.9.a. Of course, these are just some of the many thousands of possibilities to consider—but it is almost impossible to give a comprehensive account of the dependencies between the various spectral bands here. Nevertheless, our aim was not to completely describe these inter-band dependencies, but rather to point out that dependencies do indeed exist, so that we really should consider multispectral models.

4.3.2.1 Stationary (i.i.d.) land cover model

The maximum likelihood classification using the stationary i.i.d. multispectral models proceeds exactly as described in section 4.3.1.1, where it is now understood that X_k is a multispectral observation (and hence $Z_k^\theta, \theta = 0, 1$ are also multispectral observations), and that q_0 and q_1 are the corresponding joint probability density functions.

The classification results (i.e., confusion matrices) of all the dual-band stationary i.i.d. classifiers are given in Table 5.1 on page 94.

4.3.2.2 Time-varying land cover model

The maximum likelihood classification using the time-varying multispectral models also proceeds exactly as in section 4.3.1.2, where once again it is now understood that X_k is a multispectral observation, and that q_0 and q_1 are corresponding joint pdfs.

All of the possible combinations of spectral bands have been considered for the time-varying multispectral classification task, and the most important results are summarised in Table 5.2 on page 95 and Figure 5.1 on page 95.

We would of course expect the time-varying multispectral models to perform better than the single band models presented previously (as well as the stationary multispectral models), since they model—at least partially—the dependence between various spectral bands. As expected, Figure 5.1 on page 95 confirms that the time-varying multispectral models perform better than the other models. However, the multispectral models still fail to take the dependence (both temporal and spatial) between different multispectral observations into account.

4.3.3 Sequential land cover classification

We will perform sequential land cover classification only for the stationary (i.i.d.) single band models, since it is still not clear how to compute the minimal cost functions for the time-varying formulation, or how to evaluate the multidimensional integrals accurately enough—these issues are discussed in section 4.4.1 and section 4.4.2, respectively.

4.3.3.1 Stationary (i.i.d.) land cover model

The sequential land cover classification task requires us to first specify a set of costs: c_0, c_1, and c. As discussed in 3.4.0.1.1, there are unfortunately no guidelines for choosing the cost of sampling, c, in order to obtain a specific probability of error. Nevertheless, assuming that we have a set of costs, the optimal threshold π_{L} and π_{U} can be obtained for each spectral band by applying Proposition 3, and using the probability density functions corresponding to the residential and vegetation classes, respectively. The sequential classification task for a sequence of observations $\{X_k(i); k = 1, 2, \ldots\}$ for some spectral band i then proceeds as follows.

At each time step $k = 1, 2, \ldots$, an observation $X_k(i)$ is obtained, after which the preprocessing step given in (4.3) is applied. The posterior sequence, π_k^{π} is then computed according to (4.9), and the following decision rule is applied:

$$\delta_k = \begin{cases} 0, & \text{if } \pi_k^{\pi} \leq \pi_{\mathrm{L}} \\ 1, & \text{if } \pi_k^{\pi} \geq \pi_{\mathrm{U}} \\ \text{continue sampling} & \text{otherwise.} \end{cases} \tag{4.15}$$

The results obtained are given in section 5.3, but refer to section 4.4.2 for a short discussion on the current computational limitations of sequential land cover classification.

4.4 ADDITIONAL CONSIDERATIONS

There are two important considerations concerning the sequential land cover classification task that will be discussed in this section. The first concerns the difficulty of extending the sequential classification algorithm to the time-varying land cover models, and the second concerns numerical issues encountered when computing the optimal thresholds π_{L} and π_{U}, even in the stationary (i.i.d.) case.

4.4.1 Modified time-varying sequential detection

The sequential classification task for the time-varying land cover models is considerably more difficult than for the stationary i.i.d. land cover models. In fact, it has not quite been solved yet. Nevertheless, we *can* obtain the *structure* of the optimal sequential decision rule from Theorem 3. First, however, recall that the posterior sequence (π_n^π) for a time-varying sequence of observations $\mathcal{Z}_k = \{Z_1, Z_2, \ldots\}$ was defined in (4.11) as

$$\pi_n^\pi = \frac{\pi_{n-1}^\pi q_1^n(Z_n)}{\pi_{n-1}^\pi q_1^n(Z_n) + (1-\pi_{n-1}^\pi) q_0^n(Z_n)}, \quad n = 1, 2, \ldots, \tag{4.16}$$

which is clearly an inhomogeneous Markov process. To be able to apply Theorem 3, we first need to homogenise the posterior sequence, for which we will roughly be following the approach presented in [89]. Consider a homogeneous Markov process $\{X_k; k = 0, 1, \ldots\}$, with state space $E = [0,1] \times \mathbb{Z}$. Let $X_0 = \binom{\pi}{m}$, and consider a family of measures $\{P_{(\pi,m)}; \binom{\pi}{m} \in E\}$ such that

$$P_{(\pi,m)}\left(X_0 = \binom{\pi}{m}\right) = 1, \tag{4.17}$$

and let $\mathbb{E}_{(\pi,m)}\{\cdot\}$ denote expectation under $P_{(\pi,m)}$. Let A_1 be a Borel subset of $[0,1]$. For any A Borel subset of E, we have that $A = A_1 \times \mathbb{Z}_1$, where $\mathbb{Z}_1$ is a subset of the integers. We have that

$$P_{(\pi,m)}\left(X_k = \binom{\pi_k^x}{m+k} \in A \;\middle|\; X_0 = \binom{\pi}{m}\right) = P_{(\pi,m)}\left(\pi_k^x \in A_1 \;\middle|\; X_0 = \binom{\pi}{m}\right), \tag{4.18}$$

where the posterior probability, given the previous state $x = (\pi, m)$, is defined by

$$\pi_n^{x=(\pi,m)} = \frac{\pi_{n-1}^x q_1^{m+n}(Z_{m+n})}{\pi_{n-1}^x q_1^{m+n}(Z_{m+n}) + (1-\pi_{n-1}^x) q_0^{m+n}(Z_{m+n})}, \quad n = 1, 2, \ldots. \tag{4.19}$$

Now let the optimal expected reward $v(x)$ for the sequential time-varying land cover classification task be defined as

$$v(x) = \sup_{\tau \in \mathcal{T}} \mathbb{E}_x\Big\{g(X_\tau) - c\tau\Big\} \tag{4.20}$$

for some function $g : \mathbb{R} \times \mathbb{Z} \to \mathbb{R} \times \mathbb{Z}$, which is equivalent to writing

$$v(\pi, m) = \sup_{\tau \in \mathcal{T}} \mathbb{E}_{(\pi,m)} \left\{ g \begin{pmatrix} \pi_\tau^x \\ m + \tau \end{pmatrix} - c\tau \right\}. \tag{4.21}$$

Then, by Theorem 3 we have the following optimal stopping time for any $m \in \mathbb{Z}$:

$$\tau_m^* = \inf \left\{ k \geq 0 \middle| g \begin{pmatrix} \pi_k^x \\ m + k \end{pmatrix} = v \begin{pmatrix} \pi_k^x \\ m + k \end{pmatrix} \right\}. \tag{4.22}$$

By setting $m = 0$ (which can be interpreted as requiring that we start observing from time $m + 1 = 1$), we obtain the time-varying optimal stopping time:

$$\begin{aligned} \tau_{\text{opt}} &= \inf \left\{ k \geq 0 \middle| g \begin{pmatrix} \pi_k^\pi \\ k \end{pmatrix} = v \begin{pmatrix} \pi_k^\pi \\ k \end{pmatrix} \right\} \\ &= \inf \left\{ k \geq 0 \middle| g_k(\pi_k^\pi) = v_k(\pi_k^\pi) \right\}, \end{aligned} \tag{4.23}$$

which is reminiscent of the optimal stopping time (3.14) for the i.i.d. case. The only difference here is that for the time-varying case $g_k(x)$ and $v_k(x)$ are functions of both the state variable x and the time step k, while in the i.i.d. case $g(x)$ and $v(x)$ are functions of only the state variable x. Consequently we must now find a sequence of optimal thresholds to determine the optimal stopping time for the time-varying land cover classification task:

$$\tau_{\text{opt}} = \left\{ k \geq 0 \middle| \pi_k^\pi \notin \left(\pi_{\text{L}}(k), \pi_{\text{U}}(k) \right) \right\}. \tag{4.24}$$

The structure of the optimal stopping rule (4.24) can be confirmed in [23], in which a very different approach was followed than what we have used here. However, a computational method to determine the optimal thresholds in (4.24) has not yet been completed. In fact, [23] used the time-varying (what he referred to as generally distributed) framework to develop the optimal SPRT with dependent observations, but as pointed out in [85], the problem of computing its corresponding time-varying thresholds remains unsolved.

It is also worth mentioning that the optimal stopping of an inhomogeneous Markov process is considered in some detail in [88, pp. 15–19], but it seems as though the problem of computing the optimal thresholds was again not given adequate attention.

4.4.2 Numerical sensitivity of sequential classification

The computational strategy (Proposition 3) to compute the minimal cost function $s(\pi)$ for a sequence of i.i.d. observations first requires that we find the limit of a sequence of functions: $\lim_{n\to\infty} s_n(\pi) \to s(\pi)$, after which we can obtain the optimal thresholds π_{L} and π_{U} by evaluating the *infimum* and *supremum* expressions as given in (3.16) and

(3.17), respectively. However, both steps necessatrily involve numerical approximations; first to approximate $s(\pi)$, and then to approximate π_{L} and π_{U}.

The cost of sampling, c, together with the costs c_0 and c_1 further place a lower bound on π_{L}, and an upper bound on π_{U} as follows:

$$\pi_{\mathrm{L}} \geq \{\pi | c_1\pi = c\}, \quad \text{and} \quad \pi_{\mathrm{U}} \leq \{\pi | c_0(1-\pi) = c\}, \tag{4.25}$$

since $s(\pi) \geq \min\Big\{\min\{c_1\pi, c_0(1-\pi)\}, c\Big\}$. The smallest possible cost function is therefore given by $s(\pi) = \min\Big\{\min\{c_1\pi, c_0(1-\pi)\}, c\Big\}$, which is shown in Figure 4.10.

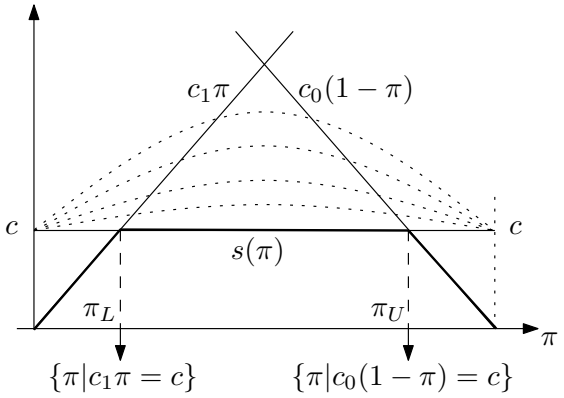


Figure 4.10: The structure of the smallest possible minimal cost function $s(\pi)$, for which $\pi_{\mathrm{L}} = \{\pi | c_1\pi = c\}$ and $\pi_{\mathrm{U}} = \{\pi | c_0(1-\pi) = c\}$ are uniquely defined.

The above observation can be used as a guideline to determine an appropriate cost of sampling for a particular problem, but it also highlights some remaining difficulties with the sequential land cover classification task, described next.

As mentioned previously, there are several 'difficult' numerical steps involved with the computation of $s(\pi)$, including the approximation of $s(\pi)$ by evaluating only a finite number of intermediate functions $s_n(\pi)$, $n = 1, 2, \ldots, N$, the numerical evaluation of integrals when computing the necessary expectations, the approximation of the intermediate functions $s_n(\pi)$ by cubic spline interpolants (we must be able to evaluate $s_n(\pi)$ for any $\pi \in [0, 1]$), as well as the approximation of infima and suprema by minima and maxima, respectively. All of these approximations limit the numerical accuracy that our current MATLAB implementation can attain, which for the land cover classification task at least, is apparently not yet sufficient (see the results presented in section 5.3).

We can easily show (experimentally) that in the sequential classification formulation task presented in section 4.3.2.1, a log likelihood ratio of about $+15$ or -15 is required for decent (but not great) classification accuracy. Such an accuracy requires that $\log_{10}\Big(\pi_{\mathrm{L}}/(1-\pi_{\mathrm{L}})\Big) \leq -15 \implies \pi_{\mathrm{L}} \leq 1 \times 10^{-15} \approx 0$. Similarly it can be shown that $\pi_{\mathrm{U}} \geq 1 - 1 \times 10^{-15} \approx 1$. By (4.25) this further implies that the cost of sampling is bounded by $c \leq 1 \times 10^{-15}$. However, the current MATLAB implementation often returns thresholds outside of the bounds specified by (4.25), which although leading to consistent classification results, are ultimately not particularly useful or insightful.

4.5 SUMMARY

In this chapter we have presented several statistical land cover models, including stationary (i.i.d.), time-varying, single band and multispectral models. The pdfs of several models were shown, from where we could form an initial idea of which bands would perform well during classification, and which would not. The seasonal variation of the datasets (and hence the motivation for developing time-varying models) were also shown, along with the times during which the land cover classes are easily separable and times in which they are not.

The maximum likelihood classification task for each type of model was then formulated, followed by the formulation of the single band, stationary (i.i.d.) sequential land cover classification task.

We further introduced the time-varying sequential classification task along with the structure of the optimal sequential decision rule, and finally we pointed out a number of remaining issues concerning the numerical approximation of the minimal cost functions, and their corresponding optimal thresholds.

Chapter 5

Experimental Results

> "*If we know that our individual errors and fluctuations follow the magic bell-shaped curve* exactly*, then the resulting estimates are known to have almost all the nice properties that people have been able to think of.*"
>
> *John W. Tukey*, 1965

The experimental results for several of the classifiers that were described in Chapter 4, as well as the experimental results obtained when using a linear Support Vector Machine (SVM), are presented and discussed in this chapter. There are several interesting observations that can be made, the most important of which are (i) that statistical methods can indeed be used to perform reliable land cover classification, (ii) that the time-varying models perform better than the stationary models, (iii) that there is little gain in using more than a year's worth of data, and lastly (iv) that multispectral models drastically increase the classification accuracy as compared to single band models.

Throughout this chapter we will present the classification results for each classifier in terms of its corresponding *confusion matrix*:

$$\mathbf{C} = \begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{FN} & \mathrm{TN} \end{bmatrix},$$

where it will be understood that the True Positives (TPs) refer to the percentage of vegetation pixels correctly classified, and that the True Negatives (TNs) refer to the percentage residential pixels correctly classified. The False Positives (FPs) and False Negatives (FNs) are then defined similarly. Clearly it must hold that $\mathrm{TP} + \mathrm{FN} = 100$, and that $\mathrm{FP} + \mathrm{TN} = 100$. Furthermore, we will associate with each confusion matrix a *classification metric*: $(\mathrm{FP} + \mathrm{FN})/2$, which can be interpreted as the probability of error.

5.1 MAXIMUM LIKELIHOOD CLASSIFICATION

5.1.1 Stationary (i.i.d.) classification

The results for the maximum likelihood classification of the stationary (i.i.d.) classifiers as described in section 4.3.1.1 (for the single band classifiers) and section 4.3.2.1 (for the multispectral classifiers) are summarised in Table 5.3 on page 96. The results are only shown for single and dual band classifiers, since it would take an impractical amount of space to report on all the multispectral classifiers. Nevertheless, the confusion matrix (in %) for the i.i.d. multispectral classifier using all the spectral bands, including NDVI, is given as

$$\begin{bmatrix} 97.3 & 3.6 \\ 2.7 & 96.4 \end{bmatrix}, \tag{5.1}$$

with an associated classification metric of (FP+FN)/2 = 3.15 %. This is a good indication of the best possible performance of a multispectral maximum likelihood classifier using the stationary (i.i.d.) models.

With reference to the classification results presented in Table 5.3, we can clearly see that band 2 has the best classification accuracy, while band 7 and NDVI have the worst. The poor performance of band 7 and NDVI is mainly as a result of the preprocessing step of (4.3), since both classes are characterised by noisy sinusoidal signals with the same mean, and differing amplitudes. Furthermore, we can see that in general, dual band classifiers perform much better than the single band classifiers, and that dual band classifiers using bands 2 & 5, and bands 2 & 7 exhibit the best classification accuracy.

The classification metrics for the single and dual band stationary (i.i.d.) classifiers are given in Table 5.1, from which it is perhaps easier to quickly gauge classifier performance.

Table 5.1: Classification metrics for the stationary, i.i.d. land cover model, defined as (FP + FN)/2, for maximum likelihood classification with $N = 368$ observations.

	Band 1	**Band 2**	**Band 3**	**Band 4**	**Band 5**	**Band 6**	**Band 7**	**NDVI**
Band 1	24.3	6.3	13.6	7.1	18.1	2.5	1.9	6.3
Band 2	6.3	8.2	6.6	7.8	5.3	7.7	5.0	6.2
Band 3	13.6	6.6	21.5	10.8	10.8	2.6	4.4	5.3
Band 4	7.1	7.8	10.8	13.7	8.7	1.6	4.4	5.6
Band 5	18.1	5.3	10.8	8.7	27.9	10.4	11.2	14.4
Band 6	2.5	7.7	2.6	1.6	10.4	28.1	16.9	19.0
Band 7	1.9	5.0	4.4	4.4	11.2	16.9	46.9	16.9
NDVI	6.3	6.2	5.3	5.6	14.4	19.0	16.9	51.3
Average	10.0	6.6	9.5	7.5	13.4	11.1	13.5	15.6

5.1.2 Time-varying maximum likelihood classification

The classification results for the time-varying classifiers described in section 4.3.1.2 and section 4.3.2.2 are summarised in Table 5.2 and Table 5.4 on page 97. Once again we see that band 2 is a pretty good band to consider, but unlike the results presented previously in section 5.1.1, band 7 and NDVI are now among the best bands to consider. In fact, the single band classifier using NDVI has the best performance of all the single band classifiers. The best dual band classifiers for the time-varying case are those corresponding to bands 4 & 6, followed closely by bands 1 & 7.

Table 5.2: Classification metrics, defined as (FP + FN)/2, for time-varying maximum likelihood classification with $N = 368$ observations.

	Band 1	**Band 2**	**Band 3**	**Band 4**	**Band 5**	**Band 6**	**Band 7**	**NDVI**
Band 1	16.0	2.6	9.6	5.5	8.1	2.3	1.8	2.5
Band 2	2.6	6.9	4.1	5.1	2.4	4.2	3.5	2.4
Band 3	9.6	4.1	15.0	9.3	7.9	2.3	3.0	3.2
Band 4	5.5	5.1	9.3	12.2	6.7	1.1	2.9	3.3
Band 5	8.1	2.4	7.9	6.7	18.8	8.2	6.9	4.3
Band 6	2.3	4.2	2.3	1.1	8.2	26.4	6.9	5.9
Band 7	1.8	3.5	3.0	2.9	6.9	6.9	14.0	6.9
NDVI	2.5	2.4	3.2	3.3	4.3	5.9	6.9	6.5
Average	6.1	3.9	6.8	5.8	8.0	7.2	5.7	4.4

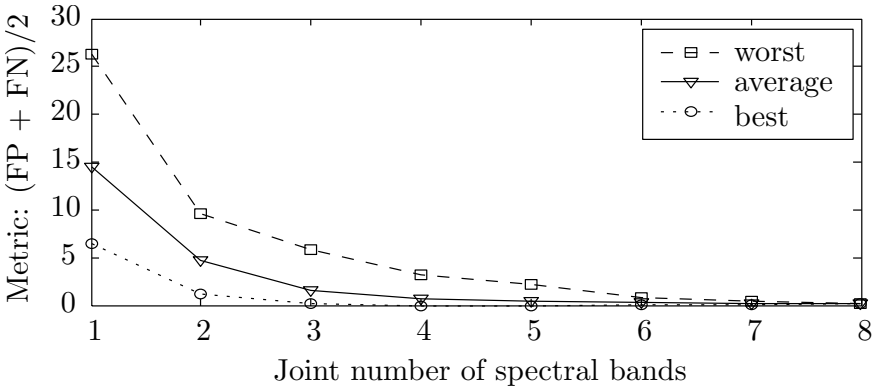


Figure 5.1: Multispectral maximum likelihood classification performance with an increasing number of spectral bands.

To see the effect of considering more spectral bands, all the possible combinations of multispectral classifiers were designed and evaluated, and the results are summarised in Figure 5.1, from which we can see that the worst combination of three bands performs as well as the best single band classifier.

Table 5.3: Confusion matrices (in %) for marginal and 2-dimensional joint maximum likelihood classification with $N = 368$ observations for the stationary, i.i.d. land cover model. Top left: true positive (vegetation), top right: false positive, bottom left: false negative, bottom right: true negative (residential). Note also that 'V' denotes vegetation, and 'R' denotes residential.

		Band 1		**Band 2**		**Band 3**		**Band 4**		**Band 5**		**Band 6**		**Band 7**		**NDVI**		**Average**	
		V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R
Band 1	V	77	25	94	07	88	16	92	06	83	20	97	02	98	02	94	06	90	10
	R	23	75	06	93	12	84	08	94	17	80	03	98	02	98	06	94	10	90
Band 2	V	94	07	93	09	93	06	91	07	100	10	100	15	99	09	94	07	95	09
	R	06	93	07	91	07	94	09	93	00	90	00	85	01	91	06	93	05	91
Band 3	V	88	16	93	06	89	32	89	11	92	13	97	02	97	06	93	04	92	11
	R	12	84	07	94	11	68	11	89	08	87	03	98	03	94	07	96	08	89
Band 4	V	92	06	91	07	89	11	88	16	92	09	98	02	97	06	94	05	93	08
	R	08	94	09	93	11	89	12	84	08	91	02	98	03	94	06	95	07	92
Band 5	V	83	20	100	10	92	13	92	09	63	18	98	19	96	19	96	25	90	17
	R	17	80	00	90	08	87	08	91	38	82	02	81	04	81	04	75	10	83
Band 6	V	97	02	100	15	97	02	98	02	98	19	72	29	80	14	81	19	90	13
	R	03	98	00	85	03	98	02	98	02	81	28	71	20	86	19	81	10	87
Band 7	V	98	02	99	09	97	06	97	06	96	19	80	14	92	85	94	28	92	14
	R	02	98	01	91	03	94	03	94	04	81	20	86	08	15	06	72	08	86
NDVI	V	94	06	94	07	93	04	94	05	96	25	81	19	94	28	46	48	92	16
	R	06	94	06	93	07	96	06	95	04	75	19	81	06	72	54	52	08	84

Table 5.4: Confusion matrices (in %) for marginal and 2-dimensional joint maximum likelihood classification with $N = 368$ observations for the time-varying land cover model. Top left: true positive (vegetation), top right: false positive, bottom left: false negative, bottom right: true negative (residential). Note also that 'V' denotes vegetation, and 'R' denotes residential.

		Band 1		**Band 2**		**Band 3**		**Band 4**		**Band 5**		**Band 6**		**Band 7**		**NDVI**		**Average**	
		V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R
Band 1	V	86	18	97	02	91	10	93	04	92	08	98	02	97	01	97	02	94	06
	R	14	82	03	98	09	90	07	96	08	92	02	98	03	99	03	98	06	94
Band 2	V	97	02	95	09	95	04	93	04	100	05	100	08	98	05	97	02	97	05
	R	03	98	05	91	05	96	07	96	00	95	00	92	02	95	03	98	03	95
Band 3	V	91	10	95	04	89	19	90	09	94	09	97	02	97	03	95	02	94	07
	R	09	90	05	96	11	81	10	91	06	91	03	98	03	97	05	98	06	93
Band 4	V	93	04	93	04	90	09	89	14	93	06	98	01	97	03	95	02	94	05
	R	07	96	07	96	10	91	11	86	07	94	02	99	03	97	05	98	06	95
Band 5	V	92	08	100	05	94	09	93	06	77	15	99	16	97	11	99	07	94	10
	R	08	92	00	95	06	91	07	94	23	85	01	84	03	89	01	93	06	90
Band 6	V	98	02	100	08	97	02	98	01	99	16	74	27	93	07	98	10	95	09
	R	02	98	00	92	03	98	02	99	01	84	26	73	07	93	02	90	05	91
Band 7	V	97	01	98	05	97	03	97	03	97	11	93	07	95	23	99	13	97	08
	R	03	99	02	95	03	97	03	97	03	89	07	93	05	77	01	87	03	92
NDVI	V	97	02	97	02	95	02	95	02	99	07	98	10	99	13	99	12	97	06
	R	03	98	03	98	05	98	05	98	01	93	02	90	01	87	01	88	03	94

5.2 SUPPORT VECTOR MACHINE CLASSIFICATION

In an attempt to verify—albeit somewhat crudely—the results obtained by the statistical land cover classification methods presented previously, several linear Support Vector Machines (SVMs) were trained, and their results are presented in this section.

5.2.1 Introduction to linear support vector machines

Statistical learning theory and SVMs in particular were largely developed by Vapnik starting in the late nineteen seventies [115], and SVMs have since been applied successfully to a myriad of classification and regression problems. For classification, SVMs typically map non-separable data into a higher dimensional *feature space*, in which the data hopefully becomes linearly separable, and then proceed to classify based on a 'learned' separating hyperplane directly in the feature space. To map the data into some higher dimensional feature space, *polynomial kernels* and *radial basis functions* are commonly used. However, we will only consider linear SVMs here.

Assume that we are given a linearly separable training dataset $\mathcal{D}$ consisting of N input-output pairs $(\mathbf{x}_i, \theta_i), i = 1, 2, \ldots, N$:

$$\mathcal{D} = \Big\{(\mathbf{x}_i, \theta_i) | \mathbf{x}_i \in \mathbb{R}^p, \theta_i \in \{-1, 1\}\Big\}, \tag{5.2}$$

where θ_i is a label denoting class membership, and $\mathbf{x}_i$ is a p-dimensional real-valued feature vector. Suppose now that we have some hyperplane which separates the positive and negative examples perfectly. The points $\mathbf{x}$ which lie on the hyperplane satisfy $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is normal to the hyperplane, $b/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathbf{w}\|$ is the Euclidean norm of $\mathbf{w}$. If we denote by d^+ and d^- the shortest distances from the separating hyperplane to the closest positive and negative data points, respectively, then the classifier *margin* is defined as $d^+ + d^-$.

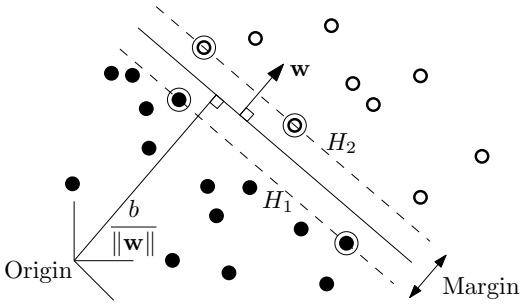


Figure 5.2: Linear separating hyperplanes for a separable dataset; the support vectors are circled. Adapted from [17].

The linear SVM (also sometimes called the *optimal margin classifier*) then simply finds the separating hyperplane with the largest margin. In other words, $\mathbf{w}$ and b are chosen such that the distance between $\mathbf{w} \cdot \mathbf{x} + b = 1$ (which corresponds to hyperplane H_2) and $\mathbf{w} \cdot \mathbf{x} + b = -1$ (which corresponds to hyperplane H_1) is maximised. Note further that the data points on H_1 and H_2 are called the *support vectors* (circled in Figure 5.2).

The problem of maximising the distance between hyperplanes H_1 and H_2 (i.e. maximising the geometric margin) can be reduced to the following optimisation problem:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{5.3}$$

$$\text{such that} \quad \theta_i(\mathbf{w}\cdot\mathbf{x}+b) \geq 1, \quad i = 1, 2, \ldots, N \tag{5.4}$$

which is commonly referred to as the *primal problem*. The solution is actually obtained by solving the so-called *dual problem*, which corresponds to the Lagrangian reformulation of the primal problem. There are two reasons for rather considering the dual problem, namely the constraints (5.4) are replaced by constraints on the Lagrange multipliers themselves, which is much easier to handle, and in the dual problem the data only appear in the form of inner products in both the training and testing algorithms, which allow for the efficient generalisation to nonlinear SVMs. Nevertheless, the dual problem will not be discussed here, and interested readers are instead referred to the excellent and comprehensive tutorial by Burges [17].

We must however mention that the above formulation holds only for *linearly separable* datasets, and we will have to consider linear SVMs for non-separable data as well. To do this, we can reformulate the primal optimisation problem (5.3) by introducing slack variables, $\xi_i \geq 0$, which measure the degree of misclassification of the data point $\mathbf{x}_i$:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{5.5}$$

$$\text{such that} \quad \theta_i(\mathbf{w}\cdot\mathbf{x}+b) \geq 1-\xi_i, \quad i = 1, 2, \ldots, N \tag{5.6}$$

where C is a user-defined penalty of misclassification. Note that if $\mathbf{x}_i$ is misclassified, then $\xi_i > 1$, so that $\sum_i \xi_i$ is an upper bound on the number of training errors [17].

5.2.2 Feature selection

Motivated by the preliminary data analysis of the MODIS data presented in section 2.4, the mean and amplitude of the noisy sinusoidal signals within each spectral band were identified as possible features on which to perform classification. These features can easily be estimated by the Fast Fourier Transform (FFT). In this way, a single band linear SVM will have a 2-dimensional feature space, and a dual band SVM will have a four dimensional feature space, consisting of the individual means and amplitudes of the two spectral bands. A major advantage of the SVM is that it can effectively handle very large dimensional feature spaces—which other classifiers typically can not.

The estimation accuracy of the FFT improves with an increasing number of observations, so that we would expect that the classification accuracy of a linear SVM which uses these features must also improve with an increasing number of observations. Consequently, since our aim with the development of linear SVMs is primarily to validate the maximum likelihood classification results of the previous sections, we will only consider the full-length (i.e. $N = 368$) SVM results here.

A total of 100 randomly sampled training sets were extracted from the dataset, and each one consisted of roughly 50 % of the data. In each case the remaining data were of course used for validation or testing purposes. The classification results presented here (in Table 5.5 and Table 5.6) are then the average performance of the 100 trials.

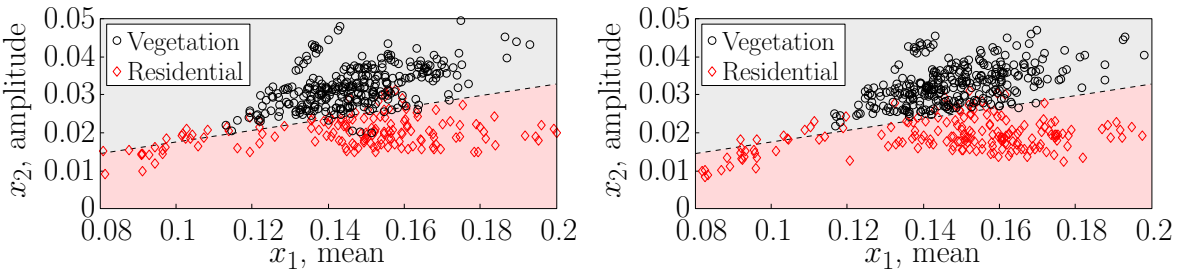
5.2.3 Support vector machine classification results

With reference to Table 5.5, we can see that band 7 and NDVI are the best to consider for single band classification using linear SVMs, which might be due to the fact that the most intra-class variability occurs along the mean components, and not along the amplitudes of the various signals (this is also clear, at least for NDVI, from Figure 5.3).

Table 5.5: Classification metrics, defined as (FP + FN)/2, for the single and dual band linear support vector machine classifiers with $N = 368$ observations.

	Band 1	**Band 2**	**Band 3**	**Band 4**	**Band 5**	**Band 6**	**Band 7**	**NDVI**
Band 1	23.5	8.5	12.1	10.7	20.9	1.2	1.7	15.9
Band 2	8.5	14.3	9.3	9.8	2.5	4.8	2.7	9.5
Band 3	12.1	9.3	19.9	14.6	15.6	2.3	1.1	9.6
Band 4	10.7	9.8	14.6	16.3	13.5	2.7	0.9	10.3
Band 5	20.9	2.5	15.6	13.5	48.0	13.4	1.9	46.3
Band 6	1.2	4.8	2.3	2.7	13.4	34.0	1.5	32.3
Band 7	1.7	2.7	1.1	0.9	1.9	1.5	8.5	8.8
NDVI	15.9	9.5	9.6	10.3	46.3	32.3	8.8	5.6
Average	11.8	7.7	10.6	9.8	20.3	11.5	3.4	17.3

Figure 5.3 shows one instance of a training (Figure 5.3.a) and validation set (Figure 5.3.b), where $N = 368$ observations were used to estimate the mean and amplitude of each of the sequences of observations for NDVI.



a. 2-Dimensional feature space: training set.

b. 2-Dimensional feature space: validation set.

Figure 5.3: Single band linear support vector machine—training and validation sets in the 2-dimensional feature space for NDVI.

That is, each point in the feature space was derived from a total of 368 observations. We can see that the two classes are indeed separable (although not perfectly) in this feature space, but that a nonlinear kernel might have made the data even more separable in some other space.

5.2.4 Support vector machines vs. sequential classification

SVMs and sequential methods have quite distinct strengths and weaknesses: SVMs can easily support high dimensional data, such that it is almost effortless to extend the current project to hyperspectral data (as opposed to the multispectral data considered here). In addition, SVMs are relatively well suited to the multiclass classification task. This is in direct contrast with the sequential methods, for which it is often extremely difficult (and sometimes impossible) to reliably infer very high dimensional probability distributions, and for which it is not clear how to find optimal exit thresholds for the multiclass problem, and it is actually still unknown if such optimal thresholds even exist for $m \geq 3$ classes.

Nevertheless, the SVMs are not particularly well suited to answer the question of how quickly we can perform reliable classification, since it is inherently a fixed window size approach. As mentioned earlier, the sequential methods excel at this task (indeed, they provide the optimal solution to the delay-accuracy tradeoff), and we further know that sequential methods (and SPRTs in particular) exhibit the smallest expected runlength of all methods (both sequential and otherwise, including SVMs) for a given probability of error, assuming of course that we are concerned with the binary classification task with accurate statistical information.

5.3 SEQUENTIAL CLASSIFICATION

The results for the sequential classification task presented in section 4.3.3 are given and discussed in this section. However, before presenting these results, it is perhaps worthwhile to consider the time-varying maximum likelihood classification task as a function of the number of observations, which is shown in Figure 5.4. Note that this classification task is different from the sequential classification task, since here we stop after a fixed number of observations, whereas with sequential classification we only stop once the likelihood ratio exceeds some predetermined threshold.

With reference to Figure 5.4 we can see that there is very little gain (in terms of classification accuracy) after the first year has elapsed. This result seems plausible, since the spectral responses of both residential and vegetation surfaces are expected to be roughly the same from one year to the next. Even though this result was not strictly obtained in the sequential classification framework, it supports the idea that sequential classification is a good idea, and furthermore, it seems to suggest that we can expect a good sequential classification strategy to classify within the first year of observation.

Table 5.6: Confusion matrices (in %) for single and dual band linear support vector machine classification with $N = 368$ observations. Top left: true positive (vegetation), top right: false positive, bottom left: false negative, bottom right: true negative (residential). Note also that 'V' denotes vegetation, and 'R' denotes residential.

		Band 1		**Band 2**		**Band 3**		**Band 4**		**Band 5**		**Band 6**		**Band 7**		**NDVI**		**Average**	
		V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R	V	R
Band 1	V	91	39	93	10	90	18	91	13	90	32	99	02	98	03	95	26	93	18
	R	09	61	07	90	10	82	09	87	10	68	01	98	02	97	05	74	07	82
Band 2	V	93	10	92	16	92	11	90	10	98	04	97	08	98	03	93	16	94	10
	R	07	90	08	84	08	89	10	90	02	96	03	92	02	97	07	84	06	90
Band 3	V	90	18	92	11	91	31	91	17	93	22	99	03	99	02	94	10	93	14
	R	10	82	08	89	09	69	09	83	07	78	01	97	01	98	06	90	07	86
Band 4	V	91	13	90	10	91	17	89	20	93	15	98	02	99	02	90	08	93	11
	R	09	87	10	90	09	83	11	80	07	85	02	98	01	98	10	92	07	89
Band 5	V	90	32	98	04	93	22	93	15	89	85	94	21	96	03	88	80	93	33
	R	10	68	02	96	07	78	07	85	11	15	06	79	04	97	12	20	07	67
Band 6	V	99	02	97	08	99	03	98	02	94	21	90	53	98	02	89	50	95	18
	R	01	98	03	92	01	97	02	98	06	79	10	47	02	98	11	50	05	82
Band 7	V	98	03	98	03	99	02	99	02	96	03	98	02	93	12	94	12	97	05
	R	02	97	02	97	01	98	01	98	04	97	02	98	07	88	06	88	03	95
NDVI	V	95	26	93	16	94	10	90	08	88	80	89	50	94	12	99	15	93	27
	R	05	74	07	84	06	90	10	92	12	20	11	50	06	88	01	85	07	73

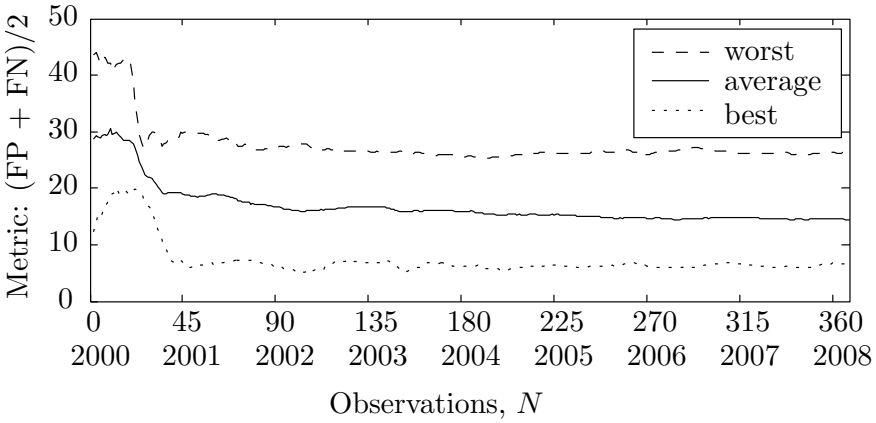


Figure 5.4: Maximum likelihood classification metric, (FP + FN)/2, over all the single band classifiers (including NDVI) as a function of the number of observations, N.

Four sets of sequential classification results (in increasing order of sampling cost) are presented in Table 5.7 on page 106. With reference to the results for bands 2 and 4, we see that in Cases I, II and III the sequential classification tests terminate after about one year, and their classification accuracies are comparable to that of the maximum likelihood classifiers, also presented in Table 5.7. However, the sampling cost of Case IV is clearly too large for these particular bands, since the classification accuracy drops significantly for both bands (which might have been expected with the tests only taking about three observations on average).

Note also the interesting behaviour of the sequential tests, in that the average number of observations for band 7 and NDVI are drastically lowered with an increase in sampling cost, without really affecting the classification accuracy. For example, in Case III, band 7 has an ARL of 11.9, and NDVI has an ARL of only 8.3, compared to roughly 100 observations each for bands 1 and 3, and about 50 each for bands 2 and 4.

Even though the results presented in Table 5.7 generally seem to be consistent with what we might expect (such as that an increase in the cost of sampling must lead to a reduction in the ARL as well as a decrease in the classification accuracy), they are nevertheless not quite satisfactory.

Firstly, the optimal thresholds seem to be larger (or smaller) than suggested by the bounds of (4.25). For example, with a cost of sampling $c = 10^{-20}$ (as in Case I), and assuming unit costs $c_0 = c_1 = 1$, we have that $-19.7 \leq \log_{10}\big(\pi/(1-\pi)\big) \leq 19.7$, but clearly most of the optimal thresholds returned by our current implementation (i.e. -36 and ∞) exceed these boundaries. In particular we also notice the apparent dependence of the optimal boundaries on the tolerance, $\text{tol} = 10^{-36}$, which is the smallest discernible unit of measure that is used when approximating the infima and suprema. This would suggest that with a smaller threshold, we would see correspondingly smaller thresholds, i.e. the thresholds tend to 0 and 1, respectively, which cannot be correct.

Secondly, the observation that such extreme thresholds are required to obtain adequate

classification accuracy is somewhat disconcerting. One possible explanation for the apparent overconfidence of the sequential classification tests is that the temporal dependence between consecutive observations are not taken into consideration. In other words, the independence assumption might lead to overoptimistic likelihood ratios, and hence might lead to such extreme thresholds as given in Table 5.7. Nevertheless, it is extremely difficult to diagnose or determine the exact origin of the above mentioned issues, but future work will focus on the systematic analysis of each part of the algorithm.

In spite of the remaining issues with the sequential classification implementation, we nevertheless determined the optimal thresholds for a range of sampling costs, and the results (for π_U at least) are presented in Figure 5.5.

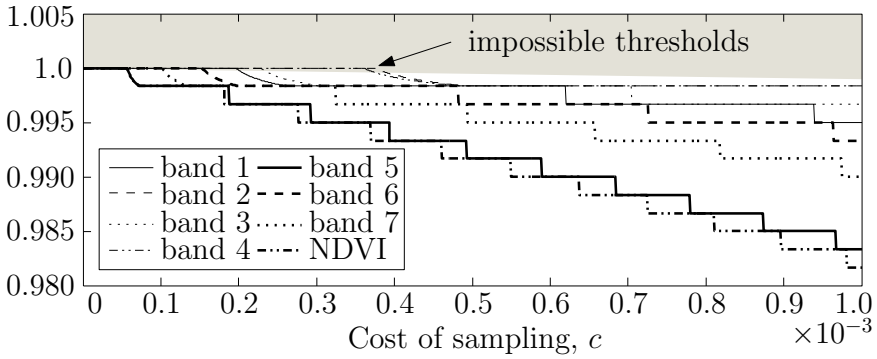


Figure 5.5: Optimal upper exit thresholds, π_U, for all the single band classifiers (including NDVI) as a function of the cost of sampling, c.

From Figure 5.5 we can see that the thresholds decrease monotonically with an increase in the cost of sampling—which is an expected result. However, what is *not expected* is the apparent step-wise structure of the thresholds, along with the smooth behaviour for $\pi \in [0.998, 1.0]$. Also note that some of the thresholds (marked as 'impossible') violate the bounds given in (4.25). The lower thresholds are characterised by similar behaviour.

The classification metrics for the various bands are shown in Figure 5.6, where we once again see the step-wise behaviour. Such step-wise behaviour can also be seen in several of the bands presented in Table 5.7—for example, band 4 has the following cost-metric pairs: $(c, m) = (2 \times 10^{-20}, 11.8)$, $(2 \times 10^{-15}, 11.8)$, $(5 \times 10^{-4}, 11.8)$, and $(5 \times 10^{-3}, 22.3)$. It is also interesting to note that the large jumps in the classification metrics are closely correlated with the first instance when the thresholds of Figure 5.5 drop below 1.0.

The ARLs as a function of the cost of sampling are shown in Figure 5.7. Notice that in general it seems as though the higher the classification metric, the sooner the ARL decreases to some steady state value. In other words, for most choices of the cost of sampling, few observations are wasted on bands that are not promising—their tests terminate early with large errors, but the expectation is that more observations would not really have improved the accuracy of the tests anyway. This behaviour can also clearly be seen from the results of band 7 and NDVI presented in Table 5.7, for which their ARLs are all less than 12 observations for Cases III and IV.

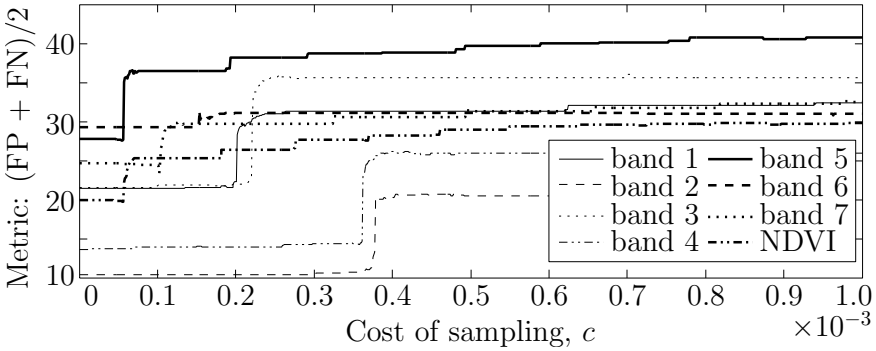


Figure 5.6: Sequential classification metric, (FP + FN)/2, over all the single band classifiers (including NDVI) as a function of the cost of sampling, c.

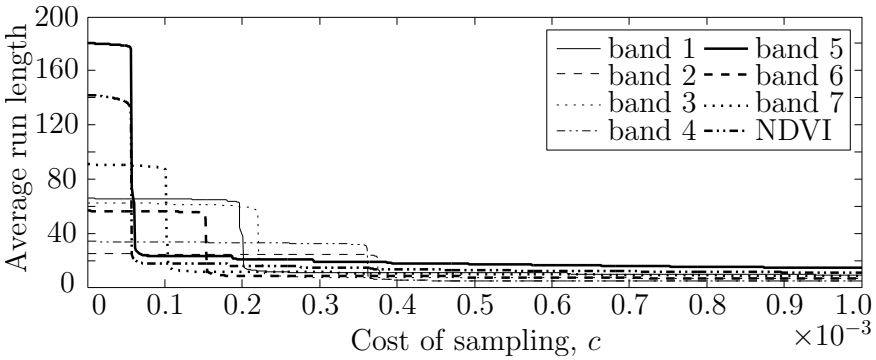


Figure 5.7: ARLs of all the single band classifiers as a function of the cost of sampling, c.

Without conducting more experiments on different datasets or perhaps verifying the MATLAB implementation in a systematic manner, it is difficult to assert which observations of this chapter hold true in general, and which are simply artifacts of the specific dataset or current implementation. Nevertheless, many of the observations are at least plausible, and sequential classification was used successfully to lower the number of observations substantially, without decreasing the classification accuracy significantly.

Table 5.7: Sequential classification results for the single band stationary (i.i.d.) land cover models.

	Band 1	**Band 2**	**Band 3**	**Band 4**	**Band 5**	**Band 6**	**Band 7**	**NDVI**	**Average**
Maximum likelihood classification									
Average run length	368	368	368	368	368	368	368	368	368
Classification metric	24.4	8.2	21.5	13.7	27.9	28.1	46.9	51.3	27.8
Sequential classification: Case I	($c = 2 \times 10^{-20}$, $c_0 = 1$, $c_1 = 1$, tol $= 1 \times 10^{-36}$, $N = 350$, numint $= 120$)								
Optimal threshold: $\log\left(\pi_{\mathrm{L}}/(1-\pi_{\mathrm{L}})\right)$	-36.00	-36.00	-36.00	-36.00	-35.70	-36.00	-36.00	-35.70	–
Optimal threshold: $\log\left(\pi_{\mathrm{U}}/(1-\pi_{\mathrm{U}})\right)$	15.95	∞	15.95	∞	∞	∞	∞	∞	–
Average run length	204.1	49.8	198.4	49.9	316.8	242.4	197.5	182.2	180.1
Classification metric	24.1	13.3	20.4	11.8	28.3	28.4	46.1	45.3	27.2
Sequential classification: Case II	($c = 2 \times 10^{-15}$, $c_0 = 1$, $c_1 = 1$, tol $= 1 \times 10^{-36}$, $N = 350$, numint $= 120$)								
Optimal threshold: $\log\left(\pi_{\mathrm{L}}/(1-\pi_{\mathrm{L}})\right)$	-36.00	-36.00	-36.00	-36.00	-36.00	-36.00	-36.00	-36.00	–
Optimal threshold: $\log\left(\pi_{\mathrm{U}}/(1-\pi_{\mathrm{U}})\right)$	15.95	15.95	15.95	15.95	15.95	15.95	15.95	15.95	–
Average run length	98.6	49.8	102.0	49.9	239.9	77.9	136.7	158.9	114.2
Classification metric	25.3	13.3	21.0	11.8	31.3	29.2	35.2	45.5	26.6
Sequential classification: Case III	($c = 5 \times 10^{-4}$, $c_0 = 1$, $c_1 = 1$, tol $= 1 \times 10^{-36}$, $N = 350$, numint $= 120$)								
Optimal threshold: $\log\left(\pi_{\mathrm{L}}/(1-\pi_{\mathrm{L}})\right)$	-4.01	-35.67	-34.94	-35.65	-2.08	-2.38	-2.38	-2.08	–
Optimal threshold: $\log\left(\pi_{\mathrm{U}}/(1-\pi_{\mathrm{U}})\right)$	∞	15.95	15.95	15.95	2.08	2.38	2.38	2.08	–
Average run length	109.5	49.3	100.1	49.7	17.1	9.0	11.9	8.3	44.4
Classification metric	28.0	13.6	20.9	11.8	54.6	37.5	26.8	44.4	29.7
Sequential classification: Case IV	($c = 5 \times 10^{-3}$, $c_0 = 1$, $c_1 = 1$, tol $= 1 \times 10^{-36}$, $N = 350$, numint $= 120$)								
Optimal threshold: $\log\left(\pi_{\mathrm{L}}/(1-\pi_{\mathrm{L}})\right)$	-1.59	-1.90	-1.59	-1.90	-0.92	-1.41	-1.24	-0.92	–
Optimal threshold: $\log\left(\pi_{\mathrm{U}}/(1-\pi_{\mathrm{U}})\right)$	1.59	1.90	1.59	1.90	1.00	1.41	1.24	0.93	–
Average run length	9.7	3.1	3.9	3.3	6.6	5.3	6.4	4.4	5.3
Classification metric	26.4	25.6	35.8	22.3	56.1	42.6	35.6	44.0	36.0

Chapter 6

Conclusions and Future Research

> When a traveller reaches a fork in the road,
> the ℓ_1-norm tells him to take either one way or the other,
> but the ℓ_2-norm instructs him to head off into the bushes.
>
> *John F. Claerbout and Francis Muir*, 1973

THREE PRIMARY OBJECTIVES were identified at the outset of this study, namely (i) to develop statistical land cover models, (ii) to design a sequential land cover classification algorithm, and (iii) to determine how quickly we can perform reliable land cover classification using coarse resolution MODIS surface spectral reflectance time series data. This chapter details the extent to which these objectives have been satisfied, highlights what has been achieved, and discusses what still remains to be done.

6.1 DISCUSSION OF WORK

This section details the extent to which our three primary objectives have been satisfied.

6.1.1 Development of statistical land cover models

Several statistical land cover models for residential as well as vegetation classes were developed in section 4.2, including single band, multispectral, stationary (i.i.d.), as well as time-varying models. These models adequately capture the inter-band dependence, but do not take the spatial or temporal dependence between observations into account, which lead to difficulties during the sequential classification task. For example, the naïve assumption of i.i.d. observations makes it difficult to interpret the 'optimal' thresholds, or to estimate a realistic probability of error.

These models are further described by several (possibly joint) pdfs, so that they are well suited to the maximum likelihood classification task. With reference to the results presented in section 5.1, we can now confidently answer Key question 3 posed in section 1.1.3. That is, reliable land cover classification *is* possible by using statistical methods such as the maximum likelihood approach. However, it should be kept in mind that statistical methods are not particularly well suited to take ancillary data into consideration, nor are they able to handle very high-dimensional data such as hyperspectral (or even just high-dimensional multispectral) data.

Nevertheless, statistical models can prove very useful to get an idea of the behaviour of the data (and consequently of the behaviour of the underlying process), which is sometimes more difficult to do with other types of models such as e.g. neural networks.

6.1.2 Design of a sequential classification algorithm

A sequential classification algorithm was developed for the stationary (i.i.d.) land cover models, in which it is possible to adjust the tradeoff between the classification accuracy and the detection delay by specifying a set of costs, $\mathbf{c} = \{c_0, c_1, c\}$. In this way, the problem statement posed in section 1.1.3 has been satisfied.

However, as discussed in section 4.4 and section 5.3, several issues still exist which makes the reliable application of sequential classification difficult. among these issues are the numerical sensitivity of the algorithm (which is made worse by the incorrect assumption of i.i.d. observations), the difficulties of evaluating multidimensional integrals for multispectral sequential classification, as well as the uncertainty of how to physically compute the cost functions (and optimal thresholds) for the time-varying land cover models.

6.1.3 The speed of land cover classification

To answer Key question 1 ("*How quickly can we perform land cover classification?*"), we have considered an experimental, fixed size approach, whose results are presented in Figure 5.4 from which we can see that—for our particular land cover models—one year (or 45 observations) seems to be sufficient for good classification accuracy, and that the classification accuracy improves very slowly after the first year.

We have also attempted to answer Key question 1 by using sequential classification, for which some of the results are summarised in Table 5.7, which also seems to suggest that an average of one year leads to comparable classification accuracy as that obtained when using all eight years.

It should be kept in mind, however, that these results were obtained by using the single band, stationary (i.i.d.) land cover models, and that Figure 5.1 would suggest that we will be able to decide much quicker with multispectral models.

6.2 CONCLUSIONS

In this study it was shown that coarse resolution MODIS surface spectral reflectance time series data can be used to effectively distinguish between residential and vegetation land cover classes, and that discrimination is even possible by considering only the mean FFT component (for bands 1 through 6) and the seasonal component for band 7 and NDVI. Furthermore, it was shown that land cover classification is easier (more reliable) during some periods of the year than others, and that these periods are different for each of the various spectral bands.

A computational strategy was also proposed and proved to compute the minimal cost function (and the associated optimal thresholds) for the i.i.d. sequential classification task, but a more robust approach is perhaps required for the multispectral land cover classification task, since numerical difficulties were encountered even for the single band classification task.

Finally, we conclude that reliable statistical land cover classification is indeed possible—albeit rather tricky—and that sequential classification can significantly reduce the required number of observations without affecting the classification accuracy too much.

6.3 FUTURE RESEARCH

There are several directions in which the current study can be extended, apart from the ultimate goal of moving over to change detection.

- The temporal and spatial dependence of the data should ideally be incorporated into the statistical models, and it is expected that such models would give more realistic optimal thresholds, as well as lead to improved classification accuracy.
- It is worthwhile to implement and investigate the multispectral sequential classification task by using Monte Carlo methods or something similar, since it is expected that the sequential classification of multispectral models will be able to terminate much sooner than the results presented here.
- The extension of sequential classification to the time-varying case should also be considered, since we have already shown the superiority of this model as compared to the i.i.d. model in terms of classification accuracy.
- It might also be of great practical significance to extend and investigate the sequential land cover classification task for more than two hypotheses, in which case the work by Dragalin et al. [28] might be useful, in which a complete generalisation of multiple-hypothesis SPRTs (MSPRTs) are given, which are asymptotically optimal w.r.t. the sample size, as well as any positive moment of the stopping time distribution. They also show how their approach can be utilised with nuisance parameters (with composite hypotheses), and in their companion paper [29] they give accurate asymptotic expansions for the expected sample size.

REFERENCES

[1] J B Adams, D E Sabol, V Kapos, R A Filho, D A Roberts, M O Smith, and A R Gillespie. Classification of multispectral images based on fractions of endmembers: application to land cover change in the Brazilian Amazon. *Remote Sensing of Environment*, 52:137–154, 1995. (Cited on page 33.)

[2] Shefali Aggarwal. Principles of remote sensing. In M V K Sivakumar, P S Roy, K Harmsen, and S K Saha, editors, *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, pages 23–38, Dehradun, India, 7–11 July 2003. World Meteorological Organisation. (Cited on page 16.)

[3] P Aplin and P M Atkinson. Sub-pixel land cover mapping for per-field classification. *International Journal of Remote Sensing*, 22:2853–2858, 2001. (Cited on page 34.)

[4] G P Asner and K B Heidebrecht. Spectral unmixing of vegetation, soil and dry carbon cover in arid regions: comparing multispectral and hyperspectral observations. *International Journal of Remote Sensing*, 23:3939–3958, 2002. (Cited on page 36.)

[5] J S Bailly, M Arnaud, and C Puech. Boosting: A classification method for remote sensing. *International Journal of Remote Sensing*, 28(7):1687–1710, 2007. (Cited on page 33.)

[6] M J Barnsley. Digital remote sensing data and their characteristics. In P Longley, M Goodchild, D J Maguire, and D W Rhind, editors, *Geographical Information Systems: Principles, techniques, applications, and management*, pages 451–466. John Wiley & Sons, New York, 2nd edition, 1999. (Cited on page 15.)

[7] Wim G M Bastiaanssen, David J Molden, and Ian W Makin. Remote sensing for irrigated agriculture: examples from research and possible applications. *Agricultural Water Management*, 46(2):137 – 155, December 2000. (Cited on page 13.)

[8] M Beibel and H R Lerche. A new look at optimal stopping problems related to mathematical finance. *Statistica Sinica*, 7:93–108, 1997. (Cited on page 60.)

[9] J A Benediktsson, J R Sveinsson, and K Arnason. Classification and feature extraction of AVIRIS data. *IEEE Transactions on Geoscience and Remote Sensing*, 33:1194–1205, 1995. (Cited on page 36.)

[10] U C 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 & Remote Sensing*, 58:239–258, 2004. (Cited on page 35.)

[11] I Bloch. Information combination operators for data fusion: a comparative review with classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 26:52–67, 1996. (Cited on page 34.)

[12] J S Borak and A H Strahler. Feature selection and land cover classification of a MODIS-like data set for a semiarid environment. *International Journal of Remote Sensing*, 20:919–938, 1999. (Cited on page 38.)

[13] Marcus Borengasser, William S Hungate, and Russel Watkins. *Hyperspectral Remote Sensing: Principles and Applications*. Taylor & Francis Series in Remote Sensing Applications. CRC Press, 2008. (Cited on pages 18, 20, 21, 22 and 27.)

[14] Alexander Brenning. Land Cover Classification by Multisource Remote Sensing: Comparing Classifiers for Spatial Data. In H Locarnek-Junge and C Weihs, editors, *Classification as a Tool for Research*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag Berlin Heidelberg, 2010. (Cited on page 9.)

[15] L Bruzzone, C Conese, F Maselli, and F Roli. Multisource classification of complex rural areas by statistical and neural-network approaches. *Photogrammetric Engineering and Remote Sensing*, 63:523–533, 1997. (Cited on page 36.)

[16] L Bruzzone, D F Prieto, and S B Serpico. A neural-statistical approach to multitemporal and multisource remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensin*, 37:1350–1359, 1999. (Cited on page 37.)

[17] Christopher J C Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998. (Cited on pages 98 and 99.)

[18] Hugo Carrão, Paulo Gonçalves, and Mário Caetano. Contribution of multispectral and multitemporal information from MODIS images to land cover classification. *Remote Sensing of Environment*, 112:986–997, 2008. (Cited on pages 38 and 41.)

[19] S Chandrasekhar. *Radiative Transfer*. Dover, Mineola, NY, 1960. (Cited on page 21.)

[20] Y S Chow, S Moriguti, Herbert Robbins, and S M Samuels. Optimal selection based on relative rank (the "Secretary Problem"). *Israel Journal of Mathematics*, 2(2):81–90, June 1964. (Cited on page 62.)

[21] J Cihlar. Land cover mapping of large areas from satellites: status and research priorities. *International Journal of Remote Sensing*, 21:1093–1114, 2000. (Cited on page 29.)

[22] J Cihlar, Q Xiao, J Chen, J Beaubien, K Fung, and R Latifovic. Classification by progressive generalization: a new automated methodology for remote sensing multispectral data. *International Journal of Remote Sensing*, 19:2685–2704, 1998. (Cited on page 3.)

[23] Jiří Cochlar and Ivan Vrana. On the optimum sequential test of two hypotheses for statistically dependent observations. *Kybernetika*, 14(1):57–69, 1978. (Cited on page 90.)

[24] M J Collins, C Dymond, and E A Johnson. Mapping subalpine forest types using networks of nearest neighbor classifiers. *International Journal of Remote Sensing*, 25:1701–1721, 2004. (Cited on page 33.)

[25] A P Cracknell. Synergy in remote sensing – what's in a pixel? *International Journal of Remote Sensing*, 19:2025–2047, 1998. (Cited on page 33.)

[26] R DeFries and A S Belward. Global and regional land cover characterization from satellite data: An introduction to the Special Issue. *International Journal of Remote Sensing*, 21:1083–1092, 2000. (Cited on page 38.)

[27] R S DeFries and J C Chan. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. *Remote Sensing of Environment*, 74:503–515, 2000. (Cited on page 3.)

[28] Vladimir P Dragalin, Alexander G Tartakovsky, and Venugopal V Veeravalli. Multihypothesis sequential probability ratio tests, Part I: Asymptotic optimality. *IEEE Transactions on Information Theory*, 45:2448–2461, November 1999. (Cited on page 109.)

[29] Vladimir P Dragalin, Alexander G Tartakovsky, and Venugopal V Veeravalli. Multihypothesis sequential probability ratio tests, Part II: Accurate asymptotic expansions for the expected sample size. *IEEE Transactions on Information Theory*, 46:1366–1383, July 2000. (Cited on page 109.)

[30] Q Du and C Chang. A linear constrained distance-based discriminant analysis for hyperspectral image classification. *Pattern Recognition*, 34:361–373, 2001. (Cited on page 36.)

[31] H Erol and F Akdeniz. A multi-spectral classification algorithm for classifying parcels in an agricultural region. *International Journal of Remote Sensing*, 17:3357–3371, 1996. (Cited on pages 10 and 33.)

[32] H Erol and F Akdeniz. A new supervised classification method for quantitative analysis of remotely-sensed multi-spectral data. *International Journal of Remote Sensing*, 19(4):775–782, 1998. (Cited on pages 10 and 33.)

[33] Thomas S Ferguson. Who Solved the Secretary Problem? *Statistical Science*, 4(3):282–289, August 1989. (Cited on page 62.)

[34] P Fisher. The pixel: a snare and a delusion. *International Journal of Remote Sensing*, 18:679–685, 1997. (Cited on page 33.)

[35] G M Foody. Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data. *International Journal of Remote Sensing*, 17:1317–1340, 1996. (Cited on page 33.)

[36] G M Foody and D P Cox. Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. *International Journal of Remote Sensing*, 15:619–631, 1994. (Cited on page 34.)

[37] G M Foody and A Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42:1336–1343, 2004. (Cited on page 33.)

[38] S E Franklin and M A Wulder. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. *Progress in Physical Geography*, 26:173–205, 2002. (Cited on page 29.)

[39] S Geman and D Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–740, 1984. (Cited on page 35.)

[40] B K Ghosh. A brief history of sequential analysis. In B K Ghosh and P K Sen, editors, *Handbook of Sequential Analysis*, pages 1–19. Dekker, New York, 1991. (Cited on page 48.)

[41] I Z Gitas, G H Mitri, and G Ventura. Object-based image classification for burned area mapping of Creus Cape Spain, using NOAA-AVHRR imagery. *Remote Sensing of Environment*, 92:409–413, 2004. (Cited on page 35.)

[42] R Haapanen, A R Ek, M E Bauer, and A O Finley. Delineation of forest/nonforest land use classes using nearest neighbor methods. *Remote Sensing of Environment*, 89:265–271, 2004. (Cited on page 33.)

[43] P J Hardin. Parametric and nearest-neighbor methods for hybrid classification: a comparison of pixel assignment accuracy. *Photogrammetric Engineering and Remote Sensing*, 60:1439–1448, 1994. (Cited on page 33.)

[44] Michael Heinl and Ulrike Tappeiner. The benefits of considering land cover seasonality in multi-spectral image classification. *Journal of Land Use Science*, (1):1–19, 2011. (Cited on page 37.)

[45] Theodore P Hill. Knowing when to stop: How to gamble if you must—the mathematics of optimal stopping. *American Scientist*, 97(2):126–133, March–April 2009. (Cited on page 60.)

[46] R M Hoffer. Biological and physical considerations in applying computer-aided analysis techniques to remote sensor data. In P.H. Swain and S.M. Davis, editors, *Remote Sensing: The Quantitative Approach*. McGraw-Hill, New York, 1978. (Cited on pages 26 and 32.)

[47] R M Hoffer. Computer Aided Analysis Techniques for Mapping Earth Surface Features. Technical Report 020179, Purdue University, Laboratory for Applications of Remote Sensing, West Lafayette, Indiana, 1979. (Cited on page 32.)

[48] B N Holben and D Kimes. Directional relectance response in AVHRR red and near-infrared bands for three cover types and varying atmospheric conditions. *Remote Sensing of Environment*, 19:213–226, 1986. (Cited on page 25.)

[49] C Huang, John R G Townshend, and L S Davis. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4):725–749, February 2002. (Cited on page 33.)

[50] L Hubert-Moy, A Cotonnec, Chardin A Du, L Le, and Perez. A comparison of parametric classification procedures of remotely sensed data applied on different landscape units. *Remote Sensing of Environment*, 75:174–187, 2001. (Cited on page 35.)

[51] A Huete et al. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, 83(1–2):195–213, November 2002. (Cited on page 25.)

[52] Gordon F Hughes. On the mean accuracy of statistical pattern classifiers. *IEEE Transactions on Information Theory*, 14(1):55–63, January 1968. (Cited on page 35.)

[53] R L Huguenin, M A Karaska, D V Blaricom, and J R Jensen. Subpixel classification of Bald Cypress and Tupelo Gum trees in Thematic Mapper imagery. *Photogrammetric Engineering and Remote Sensing*, 63:717–725, 1997. (Cited on page 34.)

[54] L F Janssen and M Molenaar. Terrain objects, their dynamics and their monitoring by integration of GIS and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 33:749–758, 1995. (Cited on page 34.)

[55] J R Jensen. Thematic information extraction: Pattern recognition. In Keith C Clarke, editor, *A Remote Sensing Perspective*, Geographic Information Science, pages 337–406. Prentice Hall, 3rd edition, 2005. (Cited on page 36.)

[56] B Jeon and D A Landgrebe. Classification with spatio-temporal interpixel class dependency contexts. *IEEE Transactions on Geoscience and Remote Sensing*, 30:663–672, 1992. (Cited on page 35.)

[57] R Juarez and W Liu. FFT analysis on NDVI annual cycle and climatic regionality in northeast Brasil. *International Journal of Climatology*, 21(14):1803–1820, Nov. 2001. (Cited on pages 38 and 46.)

[58] Y Jung and P H Swain. Bayesian Contextual Classification based on Modified M-estimates and Markov Random Fields. *IEEE Transactions on Geoscience and Remote Sensing*, 34:67–75, 1996. (Cited on page 35.)

[59] Thomas Kailath and H Vincent Poor. Detection of Stochastic Processes. *IEEE Transactions on Information Theory*, 44(6):2230–2259, October 1998. (Cited on pages 48, 51 and 54.)

[60] B Kartikeyan, B Gopalakrishna, M H Kalubarme, and K L Majumder. Contextual techniques for classification of high and low resolution remote sensing data. *International Journal of Remote Sensing*, 15:1037–1051, 1994. (Cited on page 35.)

[61] Y J Kaufmann, C J Tucker, and I Fung. Remote sensing of biomass burning in the Tropics. *Journal of Geophysical Research*, 95:9927–9939, 1990. (Cited on page 17.)

[62] N Khazenie and M M Crawford. Spatial-Temporal Autocorrelation Model for Contextual Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 28:529–539, 1990. (Cited on page 35.)

[63] H Kim, S Pang, H Je, D Kim, and S Y Bang. Constructing support vector machine ensemble. *Pattern Recognition*, 36:2757–2767, 2003. (Cited on page 33.)

[64] J Kittler and D Pairman. Contextual Pattern Recognition Applied to Cloud Detection and Identification. *IEEE Transactions on Geoscience and Remote Sensing*, 23:855–863, 1985. (Cited on page 35.)

[65] 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*, 20:123–167, 2001. (Cited on pages 33 and 36.)

[66] A Koltunov and E Ben-Dor. Mixture density separation as a tool for high-quality interpretation of multi-source remote sensing data and related issues. *International Journal of Remote Sensing*, 25:3275–3299, 2004. (Cited on page 33.)

[67] Tze Leung Lai. Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, 11:303–408, 2001. (Cited on page 48.)

[68] D A Landgrebe. Analysis Technology for Land Remote Sensing. *Proceedings of the IEEE*, 69:628–642, 1981. (Cited on page 32.)

[69] D A Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing*. Wiley, New Jersey, 2003. (Cited on page 32.)

[70] R Lawrence, A Bunn, S Powell, and M Zmabon. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote Sensing of Environment*, 90:331–336, 2004. (Cited on page 33.)

[71] Sylvie Le Hégarat-Mascle, I Bloch, and D Vidal-Madjar. Application of Dempster-Shafer evidence theory to unsupervised classification in multisource remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 35(4):1018–1031, July 1997. (Cited on page 36.)

[72] Te-Won Lee, Michael S Lewicki, and Terrence J Sejnowski. ICA Mixture Models for Unsupervised Classification of Non-Gaussian Classes and Automatic Context Switching in Blind Signal Separation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1078–1089, October 2000. (Cited on page 33.)

[73] M A Lefsky and W B Cohen. Selection of remotely sensed data. In M A Wulder and S E Franklin, editors, *Remote Sensing of Forest Environments: Concepts and case studies*, pages 13–46. Kluwer Academic Publishers, Boston, 2003. (Cited on page 15.)

[74] Noam Levin. Fundamentals of remote sensing. Technical report, Canada Centre for Remote Sensing. (Cited on pages 14, 15, 22, 25 and 26.)

[75] S Lhermitte et al. Hierarchical image segmentation based on similarity of NDVI time series. *Remote Sensing of Environment*, 112(2):506–521, Feb. 2008. (Cited on pages 25, 38 and 46.)

[76] Yan Li, Li Yan, and Jin Liu. Remote sensing image classification development in the past decade. In *Proceedings of SPIE - The International Society for Optical Engineering*, volume 7494, 2009. (Cited on page 36.)

[77] C D Lloyd, S Berberoglu, P J Curran, and P M Atkinson. A comparison of texture measures for the per-field classification of Mediterranean land cover. *International Journal of Remote Sensing*, 25:3943–3965, 2004. (Cited on page 34.)

[78] Lord Rayleigh (J. W. Strutt). *Philisophical Magazine*, 41:107–120, 274–279, 1871. (Cited on page 22.)

[79] D Lu, E Moran, and M Batistella. Linear mixture model applied to Amazonian vegetation classification. *Remote Sensing of Environment*, 87:456–469, 2003. (Cited on page 33.)

[80] D Lu and Q Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, January 2007. (Cited on pages 1, 2, 13, 15, 29, 30, 31, 32, 33, 34, 35 and 36.)

[81] S Magnussen, P Boudewyn, and M Wulder. Contextual classification of Landsat TM images to forest inventory cover types. *International Journal of Remote Sensing*, 25:2421–2440, 2004. (Cited on page 35.)

[82] B Mannan and A K Ray. Crisp and fuzzy competitive learning networks for supervised classification of multispectral IRS scenes. *International Journal of Remote Sensing*, 24:3491–3502, 2003. (Cited on page 34.)

[83] E Mohn, N L Hjort, and G O Storvik. A Simulation Study of Some Contextual Classification Methods for Remotely Sensed Data. *IEEE Transactions on Geoscience and Remote Sensing*, 25:796–804, 1987. (Cited on page 35.)

[84] S W Myint. A robust texture analysis and classification approach for urban land-use and land-cover feature discrimination. *Geocarto International*, 16:27–38, 2001. (Cited on page 36.)

[85] Ruixin Niu and Pramod K Varshney. Sampling schemes for sequential detection with dependent observations. *IEEE Transactions on Signal Processing*, 58(3):1469–1481, March 2010. (Cited on page 90.)

[86] G S Okin, D A Roberts, B Murray, and W J Okin. Practical limits on hyperspectral vegetation discrimination in arid and semiarid environments. *Remote Sensing of Environment*, 77:212–225, 2001. (Cited on page 36.)

[87] J D Paola and R A Schowengerdt. A review and analysis of back propagation neural networks for classification of remotely sensed multispectral imagery. *International Journal of Remote Sensing*, 16:3033–3058, 1995. (Cited on page 33.)

[88] Goran Peskir and Albert Shiryaev. *Optimal Stopping and Free-Boundary Problems*. Birkhäuser Verlag, 2006. (Cited on pages 60 and 90.)

[89] H Vincent Poor and Olympia Hadjiliadis. *Quickest Detection*. Cambridge University Press, United Kingdom, 2009. (Cited on pages 2, 4, 11, 47, 49, 50, 51, 58, 60, 62, 69 and 89.)

[90] P Rairoux, H Schillinger, S Niedermeier, M Rodriguez, F Ronneberger, R Sauerbrey, B Stein, D Waite, C Wedekind, H Wille, L Wöste, and C Ziener. Remote sensing of the atmosphere using ultrashort laser pulses. *Applied Physics B: Lasers and Optics*, 71:573–580, 2000. 10.1007/s003400000375. (Cited on page 13.)

[91] T Rashed, J R Weeks, M S Gadalla, and A G 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*, 16:5–15, 2001. (Cited on page 36.)

[92] A N Renez and R A Ryerson. *Manual of Remote Sensing*. Remote Sensing for the Earth Sciences, 3rd edition, 1999. (Cited on pages 2 and 26.)

[93] John A Richards and Xiuping Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer, 4th edition, 2006. (Cited on pages 17, 19, 20, 23, 24, 25, 26, 29 and 33.)

[94] C Ricotta and G C Avena. The influence of fuzzy set theory on the areal extent of thematic map classes. *International Journal of Remote Sensing*, 20:201–205, 1999. (Cited on page 34.)

[95] D A Roberts, M Gardner, R Church, S Ustin, G Scheer, and R O Green. Mapping chaparral in the santa monica mountains using multiple endmember spectral mixture models. *Remote Sensing of Environment*, 65:267–279, 1998. (Cited on page 33.)

[96] Amandine Robin, Sylvie Le Hegarat-Mascle, and Lionel Moisan. A multiscale multitemporal land cover classification method using a bayesian approach. In Lorenzo Bruzzone, editor, *Image and Signal Processing for Remote Sensing XI*, volume 5982, pages 38–49. SPIE, 2005. (Cited on page 34.)

[97] Amandine Robin, Sylvie Le Hégarat-Mascle, and Lionel Moisan. Unsupervised subpixelic classification using coarse-resolution time series and structural information. *IEEE Transactions on Geoscience and Remote Sensing*, 46(5):1359–1374, 2008. (Cited on page 34.)

[98] S W Running, C O Justice, V Salomonson, D Hall, J Barker, Y J Kaufmann, A H Strahler, A R Huete, J P Muller, V Vanderbilt, Z M Wan, P Teillet, and D Carneggie. Terrestrial remote sensing science and algorithms planned for EOS/MODIS. *International Journal of Remote Sensing*, 15(17):3587–3620, November 1994. (Cited on page 13.)

[99] Shubha Sathyendranath and Trevor Platt. The Spectral Irradiance Field at the Surface and in the Interior of the Ocean: A Model for Applications in Oceanography and Remote Sensing. *Journal of Geophysical Research*, 93(C8):9270–9280, 1988. (Cited on page 13.)

[100] C Schaaf et al. First operational BRDF, albedo nadir reflectance products from MODIS. *Remote Sensing of Environment*, 83(1-2):135–148, Nov. 2002. (Cited on page 41.)

[101] R A Schowengerdt. On the estimation of spatial-spectral mixing with classifier likelihood functions. *Pattern Recognition Letters*, 17(13):1379–1387, November 1996. (Cited on page 34.)

[102] C A Shah, M K Arora, and P K Varshney. Unsupervised classification of hyperspectral data: an ICA mixture model based approach. *International Journal of Remote Sensing*, 25:481–487, 2004. (Cited on page 33.)

[103] Albert N Shiryaev. *Optimal Stopping Rules*. Springer-Verlag, New York, 2nd edition, 1978. (Cited on pages 51, 52, 55 and 59.)

[104] D Siegmund. *Sequential Analysis*. Springer-Verlag, New York, 1985. (Cited on page 59.)

[105] Y Sohn and N S Rebello. Supervised and unsupervised spectral angle classifiers. *Photogrammetric Engineering and Remote Sensing*, 68:1271–1281, 2002. (Cited on page 33.)

[106] A H S Solberg, T Taxt, and A K Jain. A Markov Random Field Model for Classification of Multisource Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 34:100–113, 1996. (Cited on page 35.)

[107] J Stuckens, P R Coppin, and M E Bauer. Integrating contextual information with per-pixel classification for improved land cover classification. *Remote Sensing of Environment*, 71:282–296, 2000. (Cited on page 35.)

[108] P H Swain and H Hauska. The Decision Tree Classifier: Design and Potential. *IEEE Transactions on Geoscience Electronics*, 15:142–147, 1977. (Cited on page 33.)

[109] P H Swain, S B Varderman, and J C Tilton. Contextual Classification of Multispectral Image Data. *Pattern Recognition*, 13:429–441, 1981. (Cited on page 35.)

[110] P S Thenkabail, R B Smith, and E De Pauw. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote Sensing of Environment*, 71:158–182, 2000. (Cited on page 40.)

[111] John R G Townshend, Christopher Justice, Wei Li, Charlotte Gurney, and Jim McManus. Global Land Cover Classification by Remote Sensing: Present Capabilities and Future Possibilities. *Remote Sensing of Environment*, 35(2-3):243–255, February 1991. (Cited on pages 14, 38 and 40.)

[112] Brandt Tso and Paul M Mather. *Classification Methods for Remotely Sensed Data*. CRC Press, 2nd edition, 2009. (Cited on pages 7, 20, 22, 25, 26, 28 and 35.)

[113] M O Ulfarsson, J A Benediktsson, and J R Sveinsson. Data fusion and feature extraction in the wavelet domain. *International Journal of Remote Sensing*, 24:3933–3945, 2003. (Cited on page 36.)

[114] H C van de Hulst. Evaluating the light from the sun. *Optical Spectra*, 6(3):32–35, 1981. (Cited on page 21.)

[115] V Vapnik. *Estimation of Dependences Based on Empirical Data [in Russian]*. Nauka, Moscow, 1979. (English translation: Springer Verlag, New York, 1982). (Cited on page 98.)

[116] C A O Vieira. *Accuracy of remote sensing classification of agricultural crops: A comparative study*. PhD thesis, School of Geography, The University of Nottingham, Nottingham, United Kingdom, 2000. (Cited on page 8.)

[117] Robert K Vincent. *Fundamentals of geological and environmental remote sensing*. Prentice Hall, Upper Saddle River, New Jersey, 1997. (Cited on page 13.)

[118] Abraham Wald. *Sequential Analysis*. Wiley, New York, 1947. (Cited on pages 2, 4, 47 and 57.)

[119] Abraham Wald and J Wolfowitz. Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, 19(3):326–339, 1948. (Cited on pages 4 and 48.)

[120] V Walter. Object-based classification of remote sensing data for change detection. *ISPRS Journal of Photogrammetry & Remote Sensing*, 58:225–238, 2004. (Cited on page 35.)

[121] L Wang, W P Sousa, P Gong, and G S Biging. Comparison of IKONOS and QuickBird images for mapping mangrove species on the Caribbean coast of panama. *Remote Sensing of Environment*, 91:432–440, 2004. (Cited on page 35.)

[122] G G Wilkinson. Results of implications of a study of fifteen years of satellite classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, 43(3):433–440, 2005. (Cited on page 36.)

[123] C E Woodcock and S Gopal. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. *International Journal of Geographic Information Science*, 14:153–172, 2000. (Cited on page 34.)

[124] X Xiong, J Sun, J Esposito, B Guenther, and W L Barnes. MODIS reflective solar bands calibration algorithm and on-orbit performance. In *Proceedings of SPIE – Optical Remote Sensing of the Atmosphere and Clouds III*, volume 4891, pages 95–104, 2003. (Cited on page 40.)

[125] J Zhang and G M Foody. Fully-fuzzy supervised classification of sub-urban land cover from remotely sensed imagery: statistical neural network approaches. *International Journal of Remote Sensing*, 22:615–628, 2001. (Cited on page 33.)

[126] X Zhang, R Sun, B Zhang, and Q Tong. Land cover classification of the North China Plain using MODIS_EVI time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(4):476–484, July 2008. (Cited on pages 9 and 38.)