Cape Peninsula
University of Technology

EXPLORATION AND APPLICATION OF MISR HIGH RESOLUTION RAHMAN-PINTY-VERSTRAETE TIME SERIES

by

ZHAO LIU

Thesis submitted in fulfilment of the requirements for the degree
Doctor of Engineering: Electrical Engineering
in the Faculty of Engineering at the
Cape Peninsula University of Technology

Supervisor: Prof. Michel M. Verstraete
Co-supervisors: Prof. Gerhard de Jager, Prof. Robert van Zyl

Bellville
June 2017

DECLARATION

I, ZHAO LIU, declare that the contents of this thesis represent my own unaided work, and that the thesis has not previously been submitted for academic examination towards any qualification. Furthermore, it represents my own opinions and not necessarily those of the Cape Peninsula University of Technology.

Signed **Date**

ACKNOWLEDGEMENTS

I wish to thank:

- Prof. Michel M. Verstraete, for his immense knowledge and patient guidance.
- Prof. Robert van Zyl and Prof. Gerhard de Jager, for their support for and promotion of this project.
- My family, for their unlimited love.
- Ms Melanie Stark, for her editing work on this thesis.

The financial assistance of French South Africa Institute of Technology (F'SATI) towards this research is acknowledged.

RESEARCH OUTPUT

Peer reviewed scientific journal paper

Title: Handling outliers in model inversion studies: A remote sensing case study using MISR-HR data in South Africa

Authors: Zhao Liu, Michel M. Verstraete and Gerhard de Jager

Status: Published by *South African Geographical Journal* in June 2017, with DOI number: 10.1080/03736245.2017.1339629.

Presented poster

Title: Handling outliers in model inversion studies: A remote sensing case study using MISR-HR data in South Africa

Authors: Zhao Liu, Michel M. Verstraete and Gerhard de Jager

Place of presentation: The 37th International Symposium on Remote Sensing of Environment held in Tshwane, South Africa 8 to 12 May 2017

ABSTRACT

Remote sensing provides a way of frequently observing broad land surfaces. The availability of various earth observation data and their potential exploitation in a wide range of socio-economic applications stimulated the rapid development of remote sensing technology. Much of the research and most of the publications dealing with remote sensing in the solar spectral domain focus on analysing and interpreting the spectral, spatial and temporal signatures of the observed areas. However, the angular signatures of the reflectance field, known as surface anisotropy, also merit attention. The current research took an exploratory approach to the land surface anisotropy described by the RPV model parameters derived from the MISR-HR processing system (denoted as MISR-HR anisotropy data or MISR-HR RPV data), over a period of 14+ years, for three typical terrestrial surfaces in the Western Cape Province of South Africa: a semi-desert area, a wheat field and a vineyard area. The objectives of this study were to explore (1) to what extent spectral and directional signatures of the MISR-HR RPV data may vary in time and space over the different targets (landscapes), and (2) whether the observed variations in anisotropy might be useful in classifying different land surfaces or as a supplementary method to the traditional land cover classification method. The objectives were achieved by exploring the statistics of the MISR-HR RPV data in each spectral band over the different land surfaces, as well as seasonality and trend in these data.

The MISR-HR RPV products were affected by outliers and missing values, both of which influenced the statistics, seasonality and trend of the examined time series. This research proposes a new outlier detection method, based on the cost function derived from the RPV model inversion process. Removed outliers and missing values leave gaps in a MISR-HR RPV time series; to avoid introducing extra biases in the statistics of the anisotropy data, this research kept the gaps and relied on gap-resilient trend and seasonality detection methods, such as the Mann-Kendal trend detection and Lomb-Scargle periodogram methods.

The exploration of the statistics of the anisotropy data showed that RPV parameter rho exhibited distinctive over the different study sites; NIR band parameter k exhibits prominent high values for the vineyard area; red band parameter Theta data are not that distinctive over different study sites; variance is important in describing all three RPV parameters. The explorations on trends also demonstrated interesting findings: the downward trend in green band parameter rho data for the semi-desert and vineyard areas; and the upward trend in blue band parameters k and Theta data for all the three study sites. The investigation on seasonality showed that all the RPV parameters had seasonal variations which differed over spectral bands and land covers; the results confirmed expectations in previous literature that parameter k varies regularly along the observation time, and also revealed seasonal variations in the parameter rho and Theta data.

The explorations on the statistics and seasonality of the MISR-HR anisotropy data show that these data are potentially useful for classifying different landscapes. Finally, the classification results demonstrated that both red band parameter rho data and NIR band parameter k data could successfully separate the three different land surfaces in this research, which fulfilled the second primary objective of this study. This research also demonstrated a classification method using multiple RPV parameters as the classification signatures to discriminate different terrestrial surfaces; significant separation results were obtained by this method.

DEDICATION

To my family, especially my two-year-old daughter Yijia Shi

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

kNN	k nearest neighbour
AVHRR	Advanced Very High Resolution Radiometer
BFAST	Breaks For Additive Seasonal and Trend
BRF	Bidirectional reflectance factor
CHRIS	Compact High Resolution Imaging Spectrometer
DAAC	Distributed Data Active Archive
EDA	Exploratory Data Analysis
EMR	Electromagnetic radiation
EO	Earth observation
FFT	Fast Fourier Transform
FPC	Foliage projective cover
GCI	Global Change Institute
IDL	Interactive Data Language
IQR	Interquartile range
JPL	Jet Propulsion Laboratory
LaRC	Langley Research Center
LCLU	Land cover and land use
MATLAB	Matrix laboratory
MISR	Multiangle Imaging SpectroRadiometer
MISR-HR	Multiangle Imaging SpectroRadiometer-High Resolution
MODIS	Moderate-resolution Imaging Spectroradiometer
NASA	National Aeronautical and Space Administration (United States)
NIR	Near-infrared
NIST	National Institute of Standards and Technology
POLDER	Polarization and Directionality of the Earth’s Reflectances
RPV	Rahman-Pinty-Verstraete
SBB	Southern Brigalow Belt
SDGs	Sustainable Development Goals
SEMATECH	Semiconductor manufacturing technology
SVM	Support vector machine

CHAPTER 1

INTRODUCTION

1.1 Research problem statement

Remote sensing[1] from space platforms has changed the way people apprehend and comprehend the dynamic evolution of their environment. This technology offers the capability of repeatedly acquiring data over very large areas, at spatial and temporal frequencies that are appropriate for a broad range of modelling efforts and applications, and at a small fraction of the cost that would be incurred with traditional methods of sending observers in the field. Earth observation (EO) started in earnest in the second half of the 20th century, though historically it evolved from earlier monitoring and surveillance approaches such as airborne photography. The availability of EO satellites stimulated numerous studies, but technology evolved rapidly, generating ever richer and more complex data (Feuerbacher & Stoewer, 2006). These, in turn, have progressively contributed to a growing range of socio-economic applications with clear benefits for societies, including, among others, environmental change detection, risk reduction and disaster management, improvements in the supply of energy and water, weather forecasting, early warning systems, sustainable agriculture, and biodiversity monitoring and conservation (Battrick, 2005).

Remote sensing technology plays a central role in promoting a global, holistic understanding of the evolution of climate and the environment (Liang, 2008). The rational sustainable development strategies proposed by policy makers and encapsulated in the United Nations' Sustainable Development Goals (SDGs, see https://www.un.org/sustainabledevelopment/sustainable-development-goals/.) ideally hinge on a general understanding of these global phenomena, which in turn, largely derives from an interpretation of EO data. The bulk of the research and the vast majority of the publications dealing with remote sensing in the solar spectral domain focus on analysing and interpreting the spectral, spatial and temporal signatures of the observed areas. However, all structured surfaces in terrestrial environments exhibit strongly varying angular reflectance characteristics, also known as reflectance anisotropy, which appears to be largely controlled by vegetation and soil structure, and may be characterised on the basis of multi-angular remote sensing data. Vegetation structure is critical in modelling the carbon cycle and global land systems (Dandois & Ellis, 2010). Since the spectral signature of terrestrial targets is not

[1] See full definition and discussion of this term in Chapter 2 below.

sensitive to structural information, their angular signature (anisotropy) offers a unique opportunity to document other aspects of the environment (Pinty et al., 2002). The Multi-angle Imaging SpectroRadiometer (MISR) is a very appropriate instrument for this purpose (Diner et al., 1999; Pinty et al., 2002). It was designed and built by the United States' National Aeronautical and Space Administration (NASA)'s Jet Propulsion Laboratory (JPL) and hosted on NASA's Terra satellite. The geophysical products describing land surfaces derived from MISR data, at spatial resolutions of 1.1 km or coarser, are freely available and accessible from the NASA Langley Research Center (LaRC)'s Distributed Data Active Archive (DAAC). These data were analysed using the MISR-High Resolution (MISR-HR) processing system to generate a set of Rahman-Pinty-Verstraete (RPV) products that represent the anisotropy of arbitrary targets, using three separate parameters (Verstraete et al., 2012).

The current study investigated the surface reflectance anisotropy in typical environments found throughout the Western Cape Province of South Africa, including a semi-desert region, a wheat field and a vineyard area. Specifically, the current research explored to what extent the description of the anisotropy of land surfaces, as provided by the RPV model parameters derived from the MISR-HR processing system, could be used to characterise those environments, as such an analysis would have interesting and important implications in a broader context, for instance improving land cover classification schemes. Previous work in the literature (Pinty et al., 2002; Armston, Scarth, Phinn & Danaher, 2007) investigated the MISR anisotropy on a particular day or for a limited set of dates. By contrast, this research project investigated the quality and performance of these MISR-HR RPV products over a period of 14+ years, from March, 2000 to May, 2014.

While this thesis focused on the exploration of the MISR-HR RPV parameters time series data for the three typical landscapes, using statistical tools to detect trend and seasonality, it became necessary to address the problem of outliers because of their impact on the estimation of time series characteristics.

1.2 Research questions

This study aims to answer the following research question:

Is the anisotropy data, described by MISR-HR RPV parameters, able to distinguish between the selected typical landscapes: a semi-desert region, a wheat field and a vineyard area?

To address this research question, the following research sub-questions should be considered:

- What are the statistics of the MISR-HR RPV time series for the three study sites?
- Is there a long-term trend in the MISR-HR RPV time series in each spectral band over the three selected targets?
- Is there a seasonal variation in the MISR-HR RPV time series for different spectral bands over different land covers?
- Are the statistics, trend or seasonality of the MISR-HR RPV time series distinctive over different landscapes?

1.3 Objectives

The primary objectives of this research were to explore:

- To what extent spectral and directional signatures of the MISR-HR RPV data may vary in space and time over different targets (landscapes).

- Whether the observed variations in anisotropy might be useful in classifying different land surfaces or as a supplementary method to the traditional land cover classification method.

Since the presence of outliers affect the statistics, trend and seasonality of the examined MISR-HR RPV data, an improper way to detect outliers may lead to biased analysis of the results. In order to complete the primary objectives, a sub-objective, namely:

- To find a method dealing with those unexpected outliers in the RPV parameters properly and rationally,

should be achieved firstly.

1.4 Significance of research

This research fills some of the gaps in the studies on the angular signatures of surface reflectance in the context of remote sensing, which are remarkably less in literature than the investments on the spectral, spatial and temporal signatures of the observed area. This study initially summarized the statistical characteristics of the MISR-HR RPV time series over different terrestrial surfaces, and revealed the trend and seasonality in the series. If successful this research could show that the anisotropy data can be used for classifying typical landscapes. It is hoped that further advances in this direction may lead to a new or better characterization of land surfaces, and improved downstream applications, such as monitoring climate changes and environmental degradation, or to support policy making and the management of natural resources.

1.5 Methodology

The objectives of this research were achieved by examining the statistics of the MISR-HR RPV time series in each spectral band over different land covers and exploring the trend and seasonality in these data. The flowchart of this research methodology is shown in Figure 1.1.

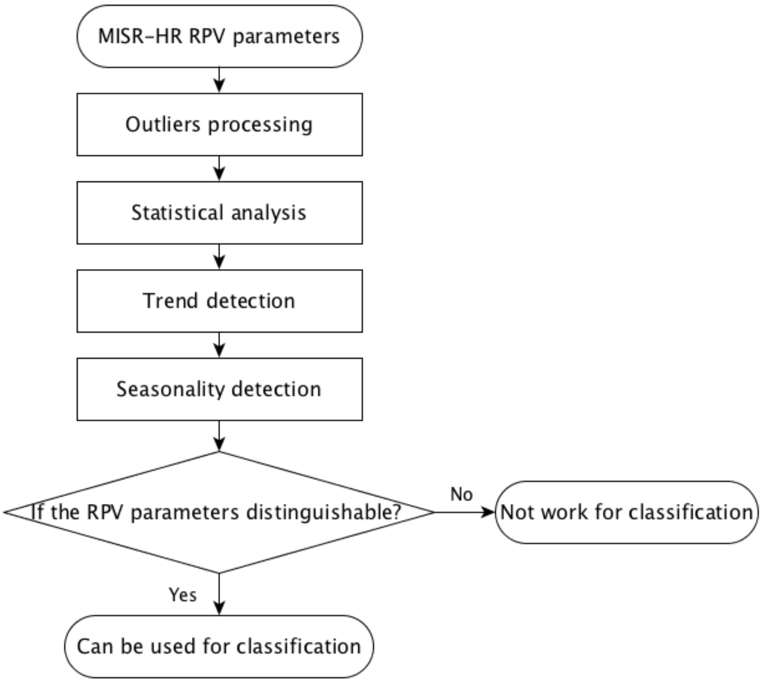

Figure 1.1: Conceptual flowchart of the research methodology.

This research used the cost function of the RPV model inversion process (Rahman, Pinty & Verstraete, 1993) to detect and manage the outliers in the MISR-HR RPV time series. This cost function indicates how well the inversion model fits the measurement data, and is thus a reasonable indicator of the degree to which variations in the data can be explained by the model. While the cost function method may not detect all the possible outliers, it does permit the elimination of data points that are spurious, whether they appear to be outliers or not. A traditional statistical outlier detection method, the box plot method, was employed to compensate the cost function method.

Missing values in remote sensing data are also quite common. To avoid introducing extra biases in the statistics of the MISR-HR time series, this research kept the missing values as gaps and relied on gap resilient trend and seasonality detection methods. The trend detection method used in this research was the Mann-Kendall test (Mann, 1945; Kendall, 1975) because this method can be applied without the assumption of “normally distributed” for the analysed time series.

Detecting seasonality in a time series may be quite challenging when there are data gaps in the series. The Lomb-Scargle periodogram was employed in the current research to detect seasonality from the unevenly spaced MISR-HR RPV time series. Once seasonality was detected, the seasonal component was retrieved using the reconstruction method proposed by Hocke and Kampfer (2009). The retrieved seasonal part of the analysed time series was used to reveal the vegetation signatures of the observed area.

Finally, the classification signatures for each environment were explored by comparing the statistics of the MISR-HR RPV time series over the different land covers, as well as their seasonal variations. In this case, the k-nearest neighbour (kNN) classification method was found to be simple and reliable. The different environments were classified by a single MISR-HR RPV product combined with the kNN classifier. Any arbitrary target can be represented by the three parameters of the MISR-HR RPV products, all of which potentially may be used as classification signatures. Classification of the different environments was also applied using the multiple MISR-HR RPV parameters data in this study.

1.6 Outline of this thesis

This thesis is structured as follows:

- Chapter 2 provides background on remote sensing, the MISR instrument and the specific sites investigated in this context, and reviews the literature.

- Chapter 3 focuses on illustrating the outlier and missing value handling methods followed in this research. A new outlier detection and processing method is proposed, based on the cost function derived from inverting the RPV model against observations.

- Chapter 4 analyses and discusses the trend and seasonal variations in the MISR-HR RPV time series. A monotonic trend detection method which is applied on the MISR-HR anisotropy data is shown. The seasonality detection method used in this research is introduced in this chapter, as well as the seasonal component reconstruction method.

- Chapter 5 describes how the temporal, spectral and anisotropy information was used in distinguishing different environments. The performance of RPV parameters to classify different land surfaces is explored and compared to more traditional approaches.

- Chapter 6 summarises the findings and suggests possible future work based on this research.

CHAPTER 2

BACKGROUND, LITERATURE REVIEW, DATA SOURCE AND TEST SITES

2.1 Definition and purpose of remote sensing

Remote sensing is broadly defined as the acquisition of information about an object without being in actual contact with it. It implies the acquisition and interpretation of signals that have interacted with the target of interest before they are absorbed in the sensor (Elachi & Van Zyl, 2006; Chuvieco & Huete, 2010). As a matter of principle, these signals could be carried out as acoustic waves, electromagnetic radiation (EMR), or variations in the gravity field (Elachi & Van Zyl, 2006), though in practice EO from space platforms relies essentially on EMR (Sabins, 1987; Kerle, Janssen & Huurneman, 2001). The expression 'remote sensing' was coined in the 1960s by geographers at the U.S. Office of Naval Research to describe the observation of the Earth from a remote platform, using the aerial photography equipment available at that time (Cracknell & Hayes, 2007; Chuvieco & Huete, 2010). As space technology developed, remote sensing began to refer specifically to the exploitation of measurements obtained from artificial satellites (Cracknell & Hayes, 2007). Remote sensing, in the general sense, can refer to the analysis of EMR signals in a broad range of spectral domains, by active or passive sensors, for the purpose of characterising any given target of interest (e.g., the atmosphere, the oceans, land surfaces, the biosphere or the cryosphere) (Tempfli, Kerle, Huurneman & Janssen, 2009). The present work, however, is concerned with the exploitation of measurements obtained exclusively in the solar spectral range, and specifically from terrestrial environments.

The primary objective of remote sensing over land surfaces is to characterise the past and current states, as well as the dynamic evolution, of terrestrial environments (Coppin et al., 2004). A central motivation for this effort arises from the expectation that a proper understanding of the processes at work may provide a basis for predicting the future evolution of the observed system, or at least some range of likely scenarios concerning the future.

2.2 Interaction of solar radiation with the atmosphere and surface targets

The atmosphere, which is composed of various gases, particulate matter (aerosols), and clouds (water droplets), plays a crucial role in Earth Observation from space (Tempfli et al., 2009).

- Clouds constitute the most significant hindrance by far to the observation of the planetary surface in the solar spectral domain, as they become quickly opaque and hide the surface (or cast shadows on it) when their optical thickness increases.

- Effectively taking aerosol effects into account presents a significant scientific challenge because of the diversity of materials (e.g., dust, smoke, pollution, sea salt, black carbon), the wide range of particle sizes and shapes, as well as the complex ways in which they interact with light (e.g., Mie scattering).

- The gaseous components exhibit their own spectral absorption bands, but these are the easiest perturbations to deal with, either because the observations that are most useful for characterising the surface are acquired outside the main absorption bands, or because these spectral bands are well-known and can be corrected for, provided the amount of gas is known, which is the case for most stable atmospheric constituents (oxygen, ozone, carbon dioxide), outside of water vapour.

Consequently, the surface can be quantitatively characterised only in the absence of clouds, and provided the radiative effects of aerosols and gases have been taken into account.

Solar radiation interacts with material objects in one of two ways: it is either absorbed or scattered (i.e., the direction of propagation is modified). In the special case where the target of interest can be represented by a planar surface, the scattering component can itself be separated in two contributions: the fraction of the radiation that is scattered back on the same side of the plane from which the light originally came (this is called the reflectance of the surface) and the fraction of the radiation that traverses the surface and continues to propagate on the other side of that plane (called the transmission). Clearly, the solar radiation that is absorbed in the environment can never be observed; only the scattered component is measurable by remote sensing.

It is common in remote sensing theoretical studies to assume that surfaces reflect solar light equally well in all directions (Kimes & Kirchner, 1982). Such surfaces are called 'Lambertian'. However, no natural surface exhibits such a property (Diner et al., 1999). In fact, it is virtually impossible even to build an artificial surface that scatters incoming light in such a way, in all spectral bands. The smoother the material surface, the more it tends to behave like a mirror (thereby reflecting incoming light very preferentially in a direction symmetrical to the incoming direction with respect to the surface normal); and the rougher the surface, the more likely it will feature a 'hot spot' in the backscattering direction (i.e., a preferential reflectance in the exact opposite direction of illumination). This latter situation is particularly noticeable with structured, three-dimensional, opaque materials (Rahman, Pinty & Verstraete, 1993). For these reasons, the reflectance of natural as well as man-made surfaces is generally called bi-directional: it strongly depends on both the illumination and observation directions.

In summary, the chemical composition of a surface target controls its spectral reflectance through the absorption bands of the materials involved, while its physical structure and properties largely determine the directional distribution of reflected light. The latter, referred to as the anisotropy of the material, is itself somewhat spectrally dependent. Hence, it is expected that the combined analysis of the spectral and directional signatures of typical terrestrial targets could lead to a better characterisation of the environment, or at the very least, a more correct retrieval of its physical and chemical properties.

2.3 Data interpretation

Data interpretation is the process of extracting useful information from (in this case, remote sensing) data. Such information is necessarily derived from the variations in the observed signals with respect to independent variables that describe how the measurements are acquired (Verstraete & Pinty, 2000). Five main sources of variability have proven useful in optical remote sensing (Gerstl, 1990): spatial, temporal, spectral, directional and polarimetric. This latter approach will not be addressed further here, as the MISR instrument used here as the main source of remote sensing data does not deliver any polarimetric data. As noted above, spectral variations provide a direct way to infer the chemical nature of the targets, while directional variations hint at their structure. The spatial and temporal variations permit the identification of the horizontal size and shape of the targets, and their dynamic evolution, respectively.

While early efforts to exploit remote sensing data focused on spatial aspects (e.g., pattern recognition, for instance for the purpose of mapping), or on spectral variations (e.g., distinguishing bare soils from vegetation) (Gerstl, 1990), intense research and significant progress have taken place over the last few decades, resulting in the design, implementation and evaluation of advanced algorithms to quantitatively characterise the physical, chemical and biological properties of terrestrial targets (Diner et al., 1999). These developments, in turn, have stimulated many new applications, as well as the design of more sophisticated instruments, with improved sampling in all four main domains of variability, with better signal-to-noise ratio, and with on-board calibration mechanisms to guarantee the long-term performance and usefulness of the data (Diner et al., 1999). As a result, the methods and approaches used early on (say before about 1990) have become totally obsolete and largely irrelevant for the analysis and exploitation of data generated by modern instruments.

To the extent that quantitative information on the radiative properties of targets can be derived from an analysis of signals acquired by space-borne instruments, these can, in turn, be exploited, together with information derived from other sources, to infer biogeophysical characteristics of the environment and serve in a multitude of applications in support of the management of natural resources, sustainable economic development, planning, etc., in addition to further research (Liu et al., 2014; Moura et al., 2012; Pisek et al., 2013).

2.4 Multi-angle remote sensing

Observing an object from different angles offers a unique opportunity to obtain information on the target that is not retrievable by standard spectral methods (Pinty et al., 2002). Diner et al. (1999) categorised multi-angle remote sensing into “simultaneous” and “sequential” observing systems. “Simultaneous” systems measure the same area on a land surface at more than one angle, and within a period of at most several minutes; “sequential” systems refer to the accumulation of data from a cross-track scanning instrument over a period of several days or weeks. The typical “simultaneous” multi-angle systems[2] include MISR, POLDER and CHRIS; “sequential” systems include AVHRR and MODIS (Liang, 2008), for instance.

2.4.1 The Multi-angle Imaging SpectroRadiometer instrument

The Multi-angle Imaging SpectroRadiometer (MISR) instrument was designed and built by NASA’s Jet Propulsion Laboratory and hosted on NASA’s Terra satellite. MISR features nine digital cameras, with one pointing at nadir (0°), while the others are arrayed forward and backward of nadir, nominally at 26.1°, 45.6°, 60.0°, and 70.5° respectively. All nine cameras observe the same area on the ground within a period of less than seven minutes. Each camera has four spectral bands, in the blue, green, red and NIR (near-infrared), respectively. This instrument therefore features 36 data channels in total. The native spatial resolution of the MISR instrument is 275 m in all data channels. However, to reduce the transmission data rate to the ground receiving station, data are typically transferred (in global mode) at the full resolution of 275 m for all four nadir spectral bands and for the eight off-nadir red channels, and averaged to 1.1 km for the other 24 data channels (non-red and off-nadir): the data compression ratio is thus 16:1 in those channels. The polar-orbiting Terra platform operates on a 16-day cycle of 233 orbits each. The MISR instrument offers a complete global coverage

[2] See List of Abbreviations for the full names of these systems and instruments.

of the planet in at most nine days, though the actual frequency of revisit is highly dependent on latitude and varies from two days near the poles to nine days at the Equator.

2.4.2 MISR data and MISR-HR products

All land surface products generated by the standard MISR processing at NASA Langley are provided at the 1.1 km spatial resolution, or coarser. However, Verstraete et al. (2012) showed that it is in fact possible to derive a whole suite of surface characteristics at the full (native) spatial resolution of the sensor (275 m). The processing system required to generate such high spatial resolution (MISR-HR) products has been installed and is operational at the Global Change Institute (GCI) of the University of the Witwatersrand in Johannesburg, South Africa. This thesis is directly and exclusively focused on the analysis and evaluation of these high-resolution products.

The MISR-HR RPV product is one of the outcomes of this processing system; it describes the anisotropy of land surfaces. This product is generated by inverting the RPV model against MISR-HR atmospherically-corrected surface reflectance data (Verstraete et al., 2012). The RPV model itself is a parametric model used to describe the bidirectional reflectance of arbitrary terrestrial surfaces, proposed by Rahman, Pinty and Verstraete (Rahman et al., 1993). This model is inverted against atmospherically-corrected bidirectional reflectance data, as explained in Verstraete et al. (2012). It requires only three independent parameters to represent the anisotropy of arbitrary natural surfaces such as bare ground or vegetation (Rahman et al., 1993). The model equations are as follows:

$$\rho_s^R(\theta_0, \theta, \phi;\ \rho_0, \Theta, k) = \ \rho_0 M(\theta_0, \theta, k) F_{HG}(g;\ \Theta_{HG}) H(\rho_0, G) \qquad (2.1)$$

$$M(\theta_0, \theta, k) = \frac{\cos^{k-1}\theta_0 \cos^{k-1}\theta}{(\cos\theta_0 + \cos\theta)^{1-k}} \qquad (2.2)$$

$$F_{HG}(\Theta_{HG}, g) = \frac{1-\Theta_{HG}^2}{[1+2\Theta_{HG}\cos g+\Theta_{HG}^2]^{3/2}} \qquad (2.3)$$

$$H(\rho_0, G) = \ 1 + \frac{1-\rho_0}{1+G} \qquad (2.4)$$

$$G = [\tan^2\theta_0 + \tan^2\theta - 2\tan\theta_0\tan\theta\cos\phi]^{1/2} \qquad (2.5)$$

$$\cos g = \cos\theta\cos\theta_0 + \sin\theta\sin\theta_0\cos\phi \qquad (2.6)$$

Here, ρ_s is the surface bidirectional reflectance factor; θ and θ_0 are the observation and illumination zenith angles; ϕ is the relative azimuth angle between the illumination and observation directions; $\rho_0, \mathrm{k}, \Theta$ are the parameters to be derived by the inversion procedure. These three parameters are denoted as rho, Theta and k, respectively, for the convenience of typing in the rest of this thesis. rho characterises the overall reflectance intensity of the surface and k indicates the shape of the surface reflectance anisotropy. Specifically, k equals to 1.0 represents a Lambertian surface, while k values lower or greater than 1.0 corresponds to a bowl-shape or bell-shape surface anisotropy pattern, respectively. The bowl-shape or bell-shape pattern anisotropy means the spectral bidirectional reflectance factor (BRF) values increase or decrease with the illumination or observation zenith angle. Parameter Theta controls the relative amount of forward and backward scattering.

Along with the generation of the three main model parameters, the cost function expresses the ability of the expected model to account for the variability in measurement data. The cost function is defined as follows:

$$J(X) = \left(\frac{1}{2}\right)[(M(X)-d)^T C_d^{-1}(M(X)-d) + \left(X - X_{prior}\right)^T C_{X_{prior}}^{-1}\left(X - X_{prior}\right)] \quad (2.7)$$

where $M(X)$ stands for the RPV model being inverted against the data, X is the vector of model parameters, d is the vector of available measurements, and C_d^{-1} and $C_{X_{prior}}^{-1}$are the inverse uncertainty covariance matrices for the data and prior values respectively, and X_{prior} represent the prior values of model parameters.

The process of inverting the RPV model against the MISR-HR reflectance values consists in minimizing this cost function, following a steepest descent algorithm, as defined by the adjoint model of J, and generates posterior values of the model parameters X_{post} (Tarantola et al., 1987; Errico, 1997; Giering & Kaminski, 1998; Tarantola, 2005). The value of the cost function indicates to what extent the model was able to explain the variability of the data: the smaller that value, the better fit between the RPV model and the observed bidirectional reflectance factors; excessively large cost function values indicate that the RPV model is incapable of 'explaining' the variability present in the measurement data, or, equivalently, that one or more measurements do not match the pattern expected by the model.

2.5 Study areas

Three different sites in the Western Cape Province of South Africa were chosen to investigate the variations of surface anisotropy data: a semi-desert area, a wheat field and a vineyard area. Figure 2.1 displays a general map of the Western Cape Province in South Africa, while

Figure 2.2 shows the specific location of the three sites within that Province. The reason for choosing the semi-desert area was to explore how MISR-HR RPV parameters vary in an uncultivated, uninhabited and presumably temporally stable environment. Compared to the semi-desert area, the wheat field is a farmed area with an obvious seasonal signature, while the vineyard area provides a more complex example of cultivated fields in the Western Cape Province.

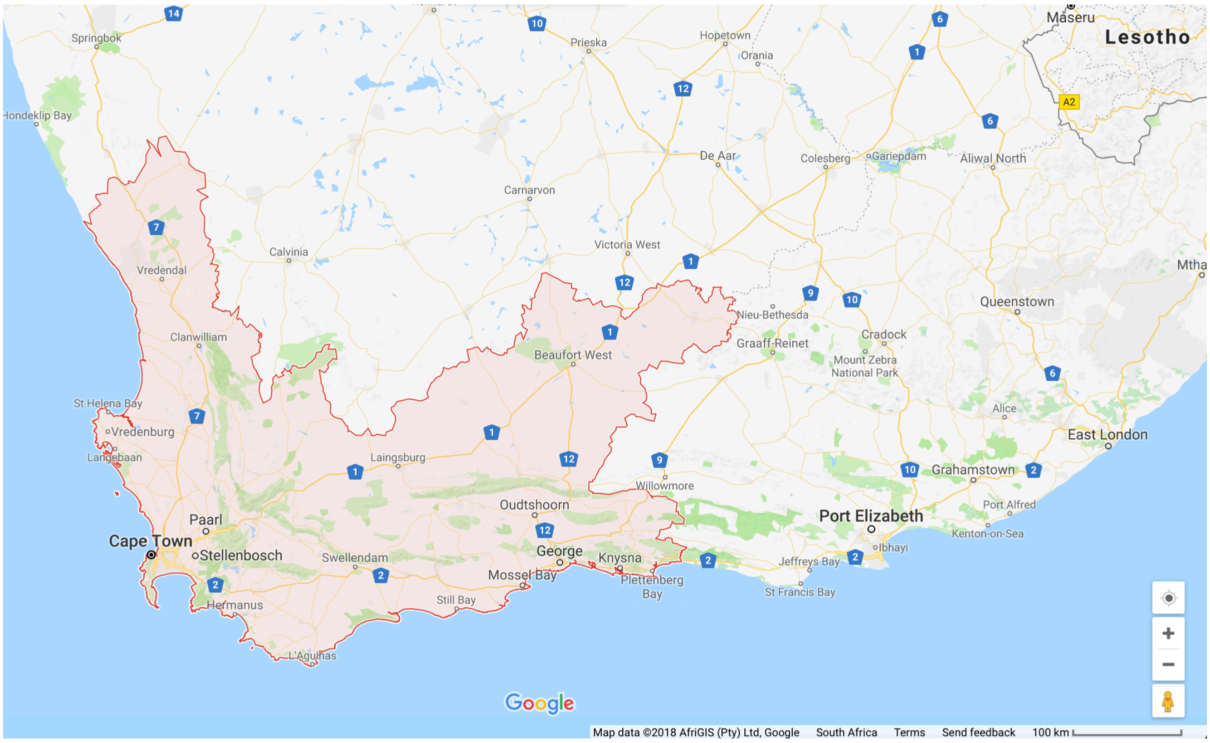

Figure 2.1: General map of the Western Cape Province of South Africa.

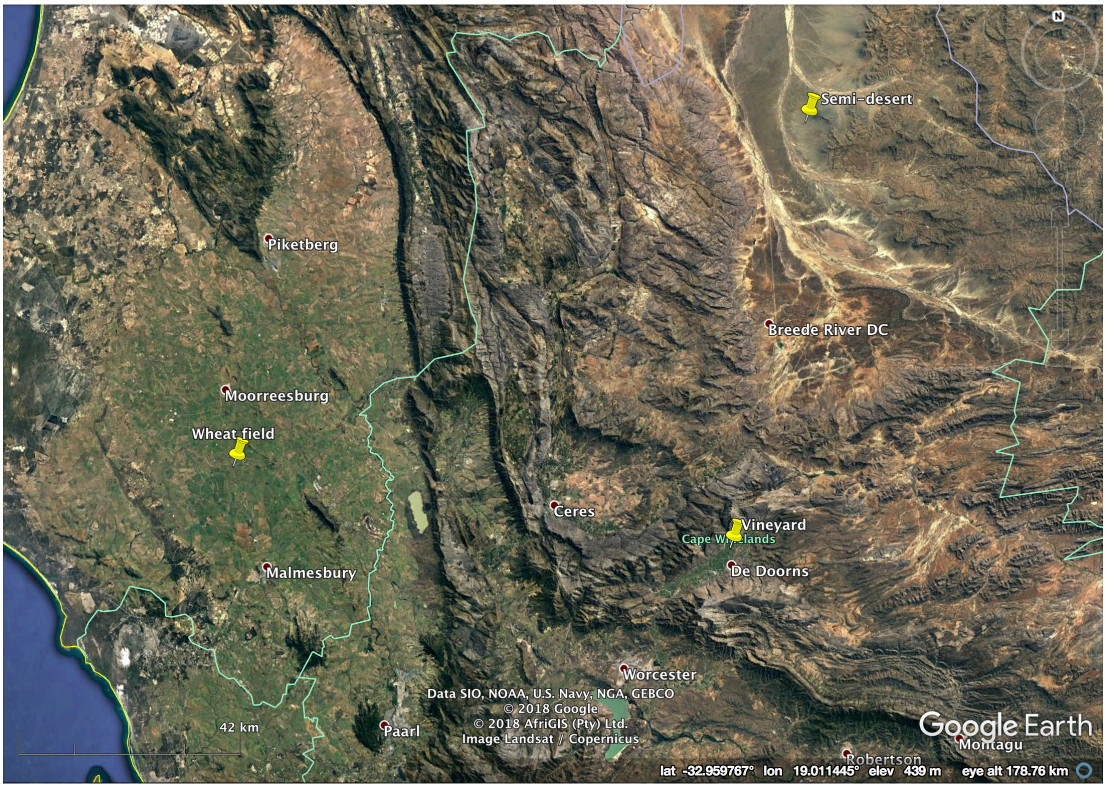

Figure 2.2: Location of the three sites used in this thesis, within the Western Cape Province.

The semi-desert study area, which lies to the south of the Tankwa Karoo National Park in South Africa, is located between 32.7528°S to 32.7312°S and 19.8305°E to 19.8631°E. The term ‘semi-desert’, also referred to as ‘steppe’ in some literature, points to ecosystems that arise between fully vegetated areas and desert regions. Vegetation in this study area is very sparse, precipitation is very limited and the sky is often clear. Figures 2.3 shows a satellite image of this area, with a central pixel and four boundary pixels marked by yellow pins. A square matrix of 11 x 11 = 121 MISR-HR pixels covers this area. The geometric coordinates of a pixel were used to distinguish each pixel in the study area: for instance, s05_+000_+000 points to the central pixel of the particular site (s05 is the index of this site). The geometric coordinates range from -005 to +005 in the across- and along- track directions (corresponding roughly to east-west and north-south, respectively). Measurements over this area were obtained from March, 2000 to May, 2014, resulting in an observation period of 14+ years. RPV model parameters were derived in the four spectral bands of the instrument, namely the blue, green, red and NIR bands.

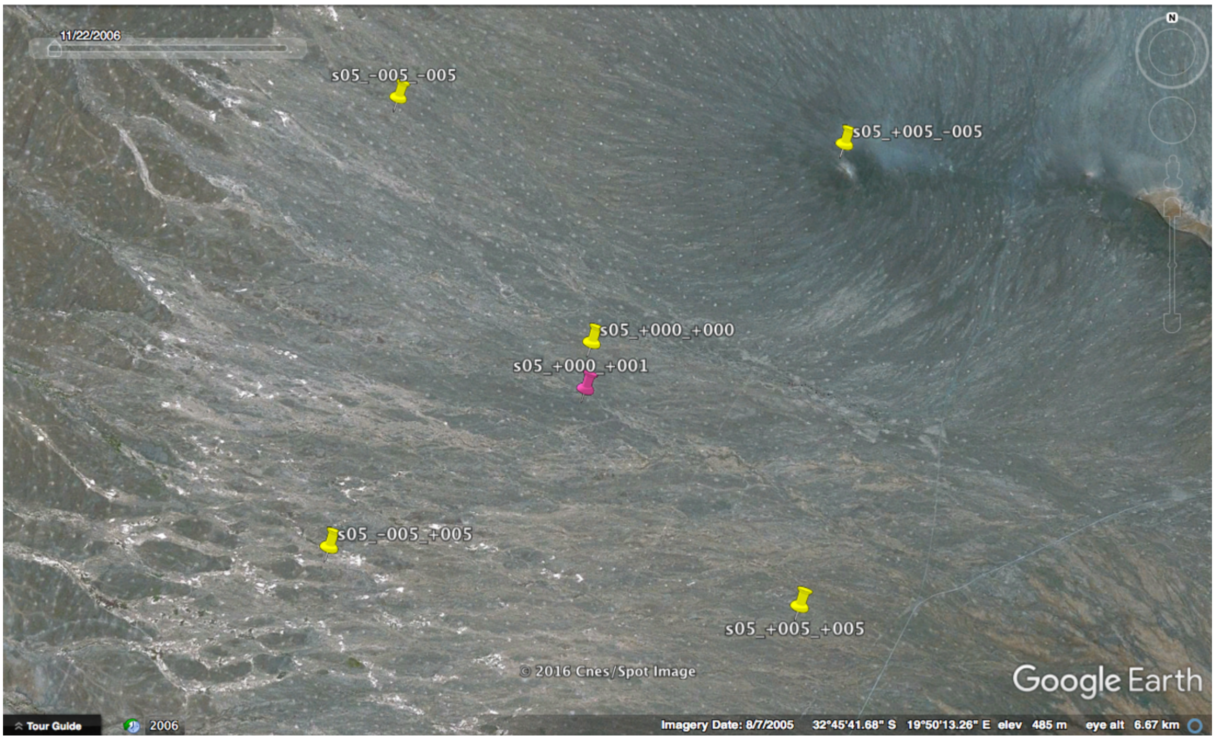

Figure 2.3: Satellite image (acquired on 7 August, 2005) of the study semi-desert area located between 32.7528°S to 32.7312°S and 19.8305°E to 19.8631°E. The distance between the two northernmost pins (or any two pins forming the sides of the square area) is 10 * 275 m = 2,750 m or 2.75 km.

Image data from Google Earth: ©2016 Cnes/Spot Image

Pixel s05_+000_+001 (marked in Figure 2.3 by a pink pin), close to the central pixel (s05_+000_+000), is located at 32.7445°S and 19.8464°E and taken to be representative of that environment. The distance between pixel s05_+000_+001 and the central pixel is 275m.

A wheat field was also studied in this investigation, since wheat is one of the major crops in South Africa and this cultivation exhibits obvious seasonal variation. Wheat is an annual winter crop, which in South Africa is usually planted between late April and early July, and harvested in November and December. The wheat field site is located to the north of the town of Malmesbury, with latitude ranging from 33.27°S to 33.29°S and longitude ranging from 18.66°E to 18.69°E. The satellite view of this area (supplied by Google Earth) is shown in Figure 2.4. Similar to the semi-desert area, there are a total of 121 pixels in the observed wheat field, and the four boundary points and central point are marked in yellow pins on the image. There are a number of fields in the observation area; as the planting time of these fields may be different, the colour of each field varies. Since the central point, s10_+000_+000, is covered with pure wheat cultivation (no trees or houses nearby), the MISR-HR anisotropy data of this pixel can be used for representing the anisotropy features of the whole area.

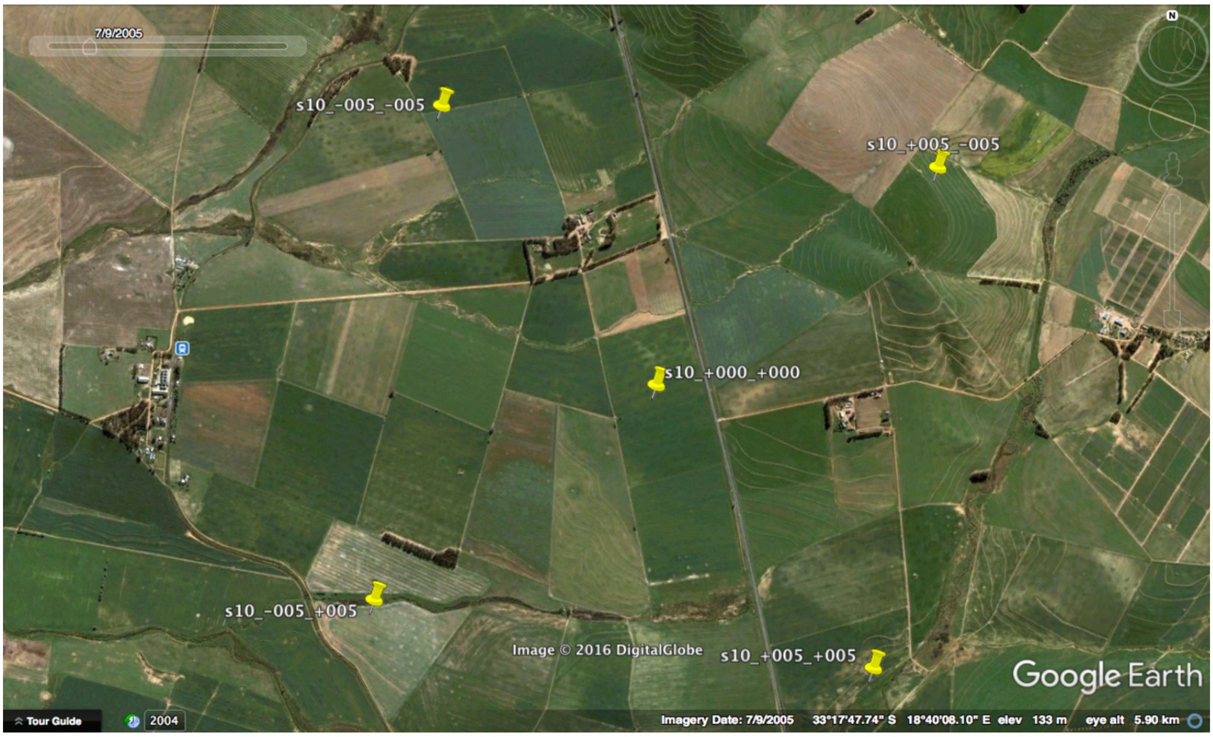
7/9/2005
s10_-005_-005
s10_+005_-005
s10_+000_+000
s10_-005_+005
Image © 2016 DigitalGlobe
s10_+005_+005
Google Earth
Tour Guide
2004
Imagery Date: 7/9/2005 33°17'47.74" S 18°40'08.10" E elev 133 m eye alt 5.90 km

Figure 2.4: Satellite image (acquired on 9 July 2005) of the study wheat field located between 33.27°S and 33.29°S, 18.66°E and 18.69°E. The distance between the two northernmost pins (or any two pins forming the sides of the square area) is 10 * 275 m = 2,750 m or 2.75 km.

Image data from Google Earth: Image © 2016 DigitalGlobe

The vineyard area is one of the most topographically and economically interesting areas in the Western Cape, South Africa. The vineyard site is located in the Hex River Valley, with latitude 33.44°S to 33.46°S and longitude 19.65°E to 19.68°E. The Hex River Valley is one of the main table grape cultivation areas in the country and has the longest harvesting period. Figure 2.5 shows satellite view of this area obtained on 2 October 2005, supplied by Google Earth. Again, the four boundary pixels and one central pixel are marked on this satellite image.

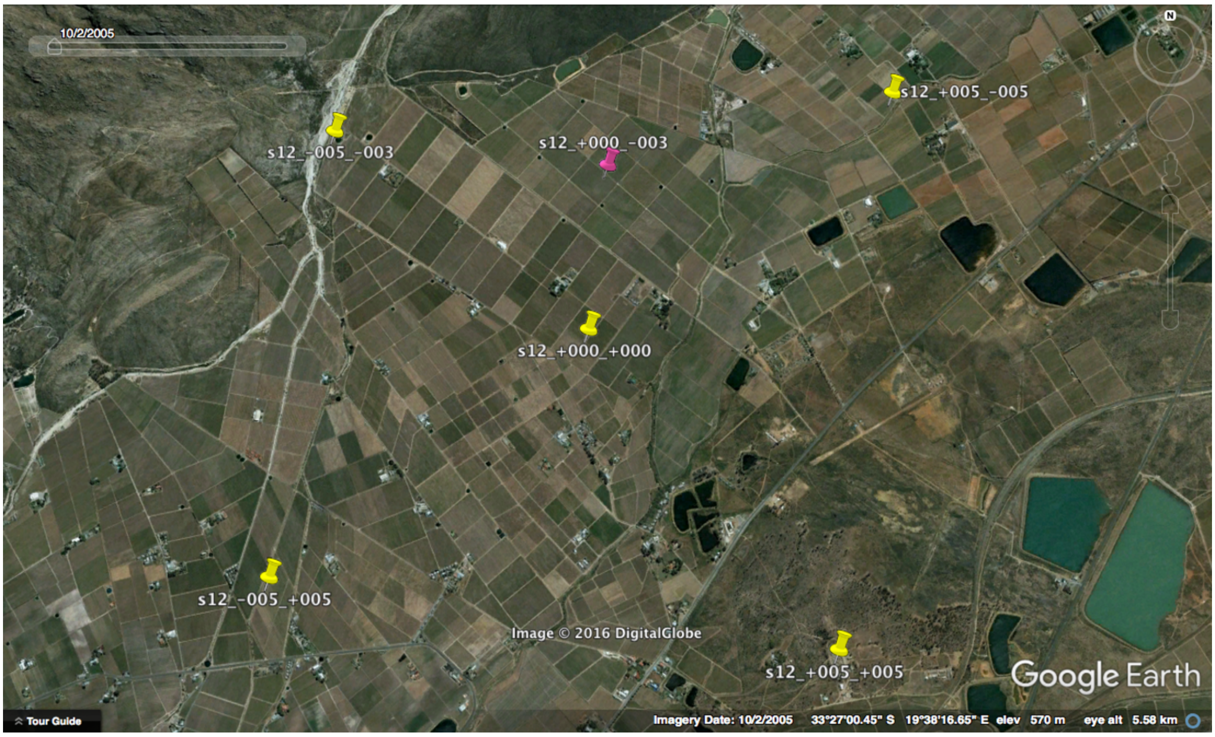

Figure 2.5: Satellite image (acquired on 2 October 2015) of the study vineyard area located between 33.44°S and 33.46°S, 19.65°E and 19.68°E. The distance between the two southernmost pins is 10 * 275 m = 2,750 m or 2.75 km.

Image data from Google Earth: ©2016 DigitalGlobe

As can be seen from this image, the vineyard area is long and narrow; therefore, it was not possible to get a square study area with 121 pixels (11 pixels in each of the along- and across-track directions) as in the semi-desert area and the wheat field. Due to partial obscuration of the surface by the local topography, only 108 pixels with full spectral bands data were available from the MISR-HR processing system for this study area. A few pixels lack all four spectral bands data, for instance, the boundary pixels in the northern part (s12_-005_-005 from the northwest to s12_+002_-005 on the northeast corner); and a few pixels lack part of the spectral bands data, for example, pixels s12_-001_-004 and s12_-002_-004 which have no data in the blue, red and NIR spectral bands. The results described below are based only on the 108 pixels available for this vineyard area.

Compared to the wheat field, the satellite images also indicate that there are more individual cultivated plots, and also more buildings beside the cultivated land, in this area. Pixel s12_+000_-003 (marked by a pink pin) was used as an example to demonstrate the variation of the anisotropy data in this area, since it represents a pure grape cultivation area without houses or planted trees nearby. This area is not exclusively a grape cultivation area, however:

there are uncultivated areas, for instance, around pixel s12_+005_+005, and probably other types of crops either mixed in with the grapevines or in adjacent fields.

2.6 Exploratory data analysis for the three selected sites

Simple Exploratory Data Analysis (EDA) tools were applied to some of the MISR–HR RPV products available to gain an appreciation for the diversity of land cover on these three sites.

2.6.1 RPV parameter rho

The RPV parameter rho describes the brightness of the reflectance in the selected spectral band. Low values are expected in the red spectral band whenever vegetation is present in the target area, because of the strong absorption capacity of chlorophyll. Rho in the red spectral band should therefore discriminate easily between vegetated and bare ground.

Figure 2.6 shows how this parameter rho varies in time, for the selected pixels of the three study sites. The outliers in the parameter rho time series were processed by the method introduced in Chapter 3. As can be seen, the time series for the semi-desert area is relatively smoother than that for the vineyard area, and both of those exhibit smaller seasonal variations than the wheat field. The average rho value for the semi-desert site is somewhat higher than for the vineyard site, as expected because a lower vegetation cover implies a larger proportion of bright bare soil.

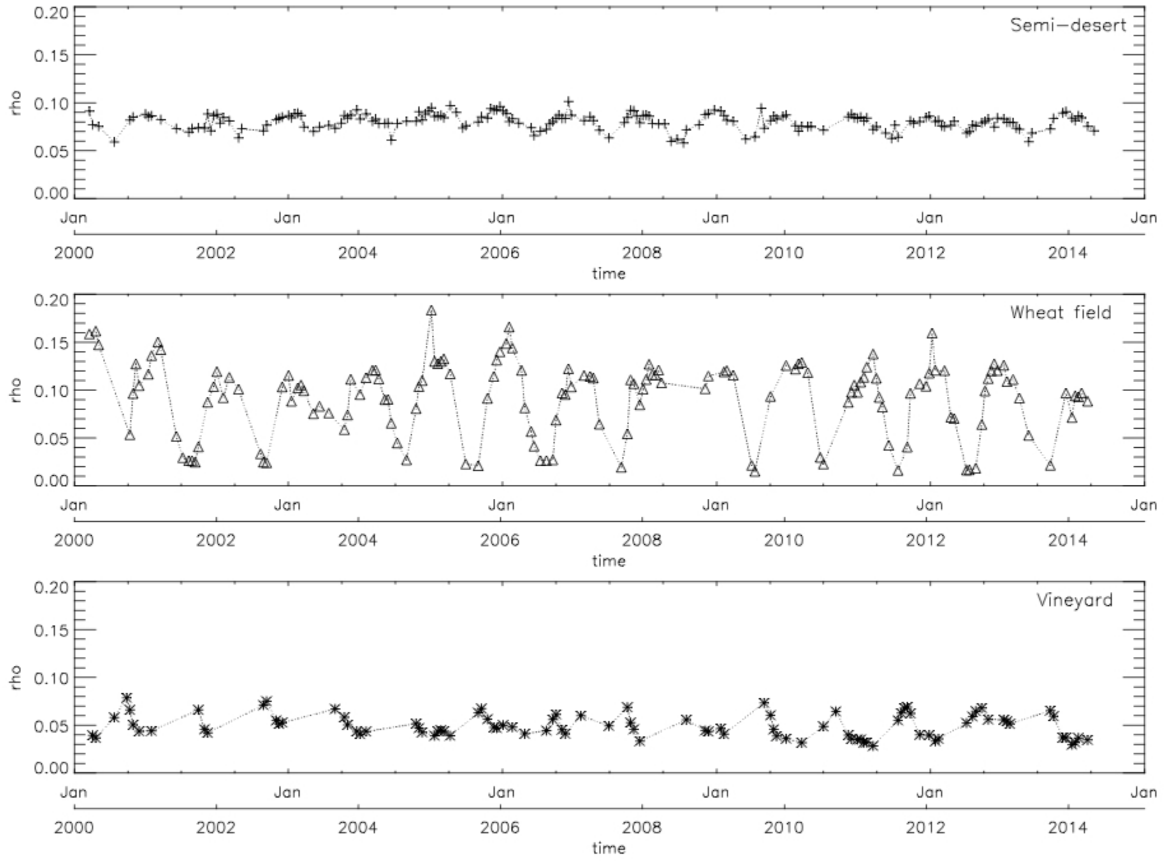

Figure 2.6: Time series plots of red band parameter rho for the semi-desert area*, the wheat field and the vineyard area***, respectively.**

* Location specified by Path 174, Block 117, Line 93 and Sample 798;
** Location specified by Path 174, Block 117, Line 358 and Sample 431;
*** Location specified by Path 174, Block 117, Line 381 and Sample 773

A box plot is an efficient and popular way of graphically representing statistics of analysed data. Figure 2.7 shows the box plots for parameter rho in the red spectral band for the three study sites. It can be seen that this parameter varies the most over the wheat field.

In this and some of the subsequent box plots, small circles aligned with the vertical line indicate values that lie outside the outer fences.

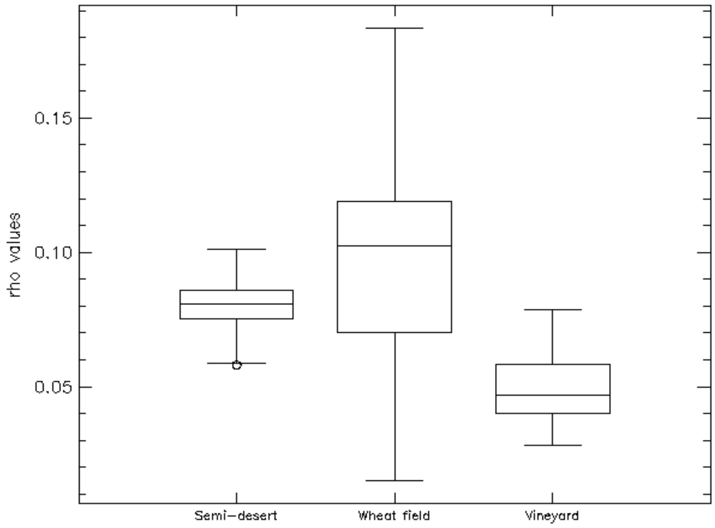

Figure 2.7: Box plot* of red band parameter rho time series, for the three study areas. Pixel locations are as specified in Figure 2.6.

* The line representing central tendency is the median; the two lines making up the box itself are the first and third quartiles; the small lines at both ends of the vertical line are the minimum and maximum values, respectively

The basic statistics of the parameter rho for the three sites are summarised in Table 2.1 below, which confirms that the vineyard area has the smallest average. For the entire observation period (March, 2000–May, 2014) there should be 324 measurements in total for a sampling frequency of 16-days, but because of the missing values and the removal of outliers (to be discussed in Chapter 3), only 94 measurements were left for the vineyard area, and 153 and 187 for the wheat field and the semi-desert area, respectively. The available number of observations for each study site is consistent with the natural environment of that area; for instance, the semi-desert area is known as a dry area with usually clear sky, so more measurements were acquired for this area. Unlike the semi-desert, the vineyard area is located in a valley with more frequent cloud obscuration, which resulted in relatively fewer measurements. The missing values and removed outliers left gaps in the time series, which may alter the true statistics of the data set; while filling the gaps in a time series may introduce extra bias to the statistics, such as the mean and variance (as discussed in Chapter 3).

Table 2.1: Basic statistics for the MISR-HR RPV rho parameter in the red spectral band, for the selected pixels in the three study areas. Pixel locations are as specified in Figure 2.6.

	Number of observations	**Mean value**	**Variance**
Semi-desert area	187	8.02×10^{-2}	6.70×10^{-5}
Wheat field	153	9.25×10^{-2}	1.49×10^{-3}
Vineyard area	94	4.94×10^{-2}	1.43×10^{-4}

The variance value for the semi-desert area is the smallest among the three study sites. This is quite reasonable since there is little or no vegetation in this semi-desert area. The variability seen in the wheat field case is due to the large variation in red spectral reflectance between the (dark) crop during the growing season and the (bright) soil after harvest. The variability observed in the vineyard case is more difficult to interpret because (1) there are fewer data points in that series (due to cloudiness) and (2) the land is apparently used for multiple cropping, with plants interspersed with vines and growing on different schedules. Figure 2.8 shows histograms of the rho parameter in the red spectral band for the three sites.

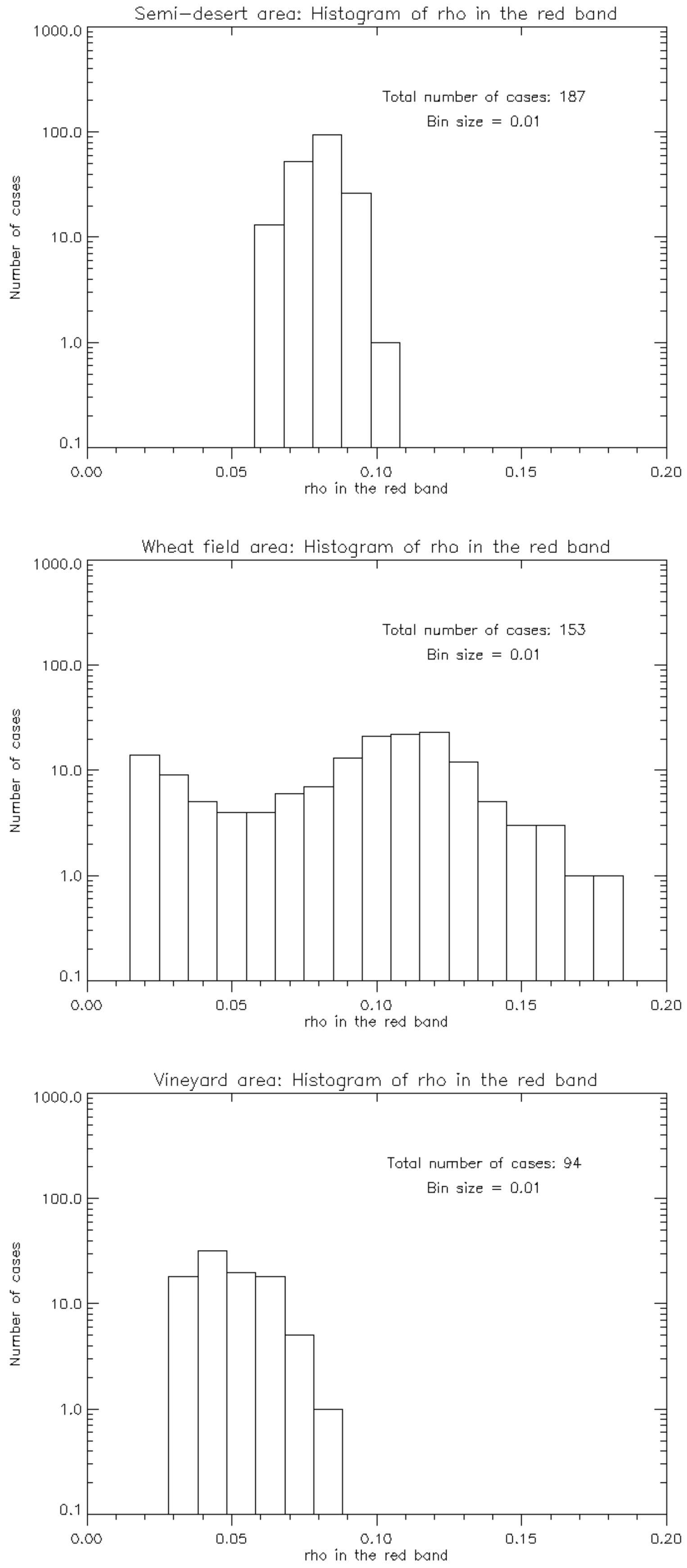

Figure 2.8: Histograms of the rho parameter in the red spectral band, for the three study areas. Pixel locations are as specified in Figure 2.6.

2.6.2 RPV parameter k

The RPV parameter k describes whether the angular reflectance distribution is bowl- or bell-shaped. Figure 2.9 exhibits the time series of that parameter in the red spectral band for the three study areas. It can be seen that the time series for the semi-desert area is relatively smooth, exhibiting a simple seasonal cycle of small amplitude. Those for the vineyard and the wheat field sites show both a higher variability and a more erratic pattern than the semi-desert site.

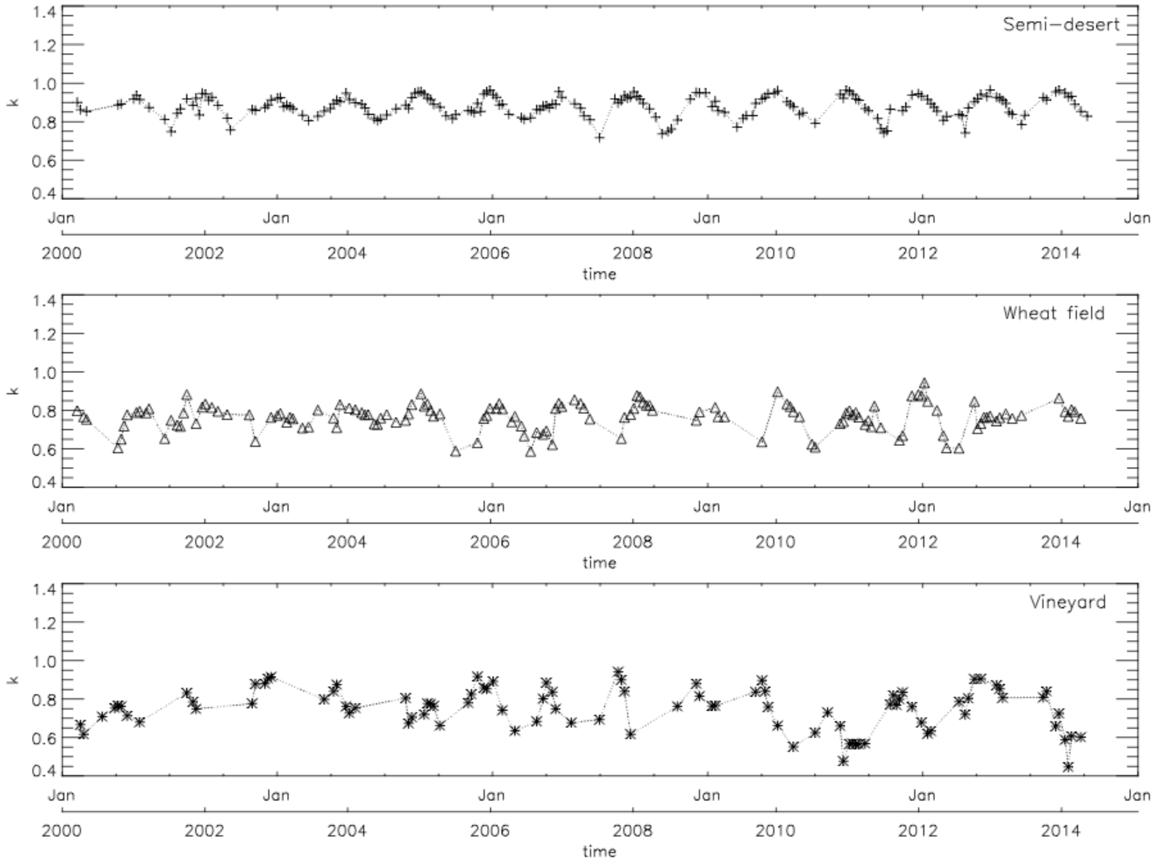

Figure 2.9: Time series of the RPV k parameter in the red spectral band for the three study areas. Pixel locations are as specified in Figure 2.6.

The corresponding box plots are shown in Figure 2.10, and the associated statistics are provided in Table 2.2. It can be seen that, in this case, it is the vineyard site that exhibits the largest k fluctuations in the red spectral band, compared to the semi-desert and the wheat field sites. The largest mean value occurs for the semi-desert area (as shown also in Table

2.2) while the mean values for the wheat and grape cultivation are lower and similar. The distribution of the k parameter in the red spectral band are shown in the histograms of Figure 2.11.

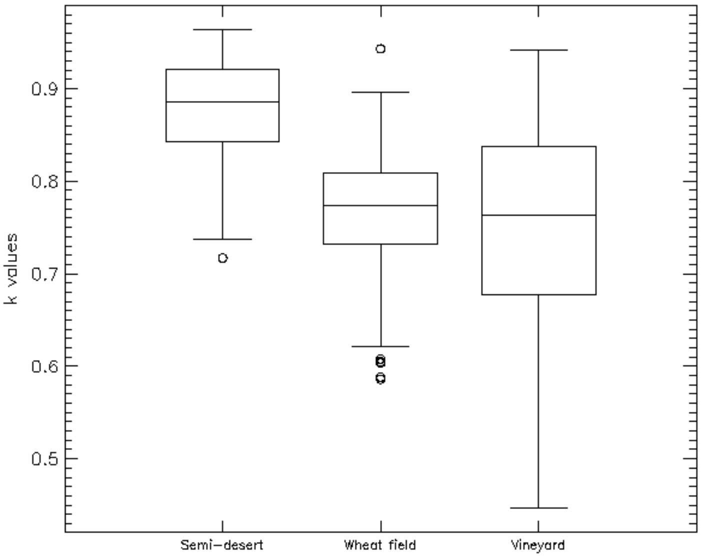

Figure 2.10: Box plot of red band parameter k time series, for the three study areas. Pixels are as specified in Figure 2.6.

Table 2.2: Basic statistics of red band parameter k time series, for the three study areas. Pixels are as specified in Figure 2.6.

	Number of observations	**Mean value**	**Variance**
Semi-desert area	190	0.879	2.96×10^{-3}
Wheat field	143	0.766	4.72×10^{-3}
Vineyard area	94	0.751	1.16×10^{-2}

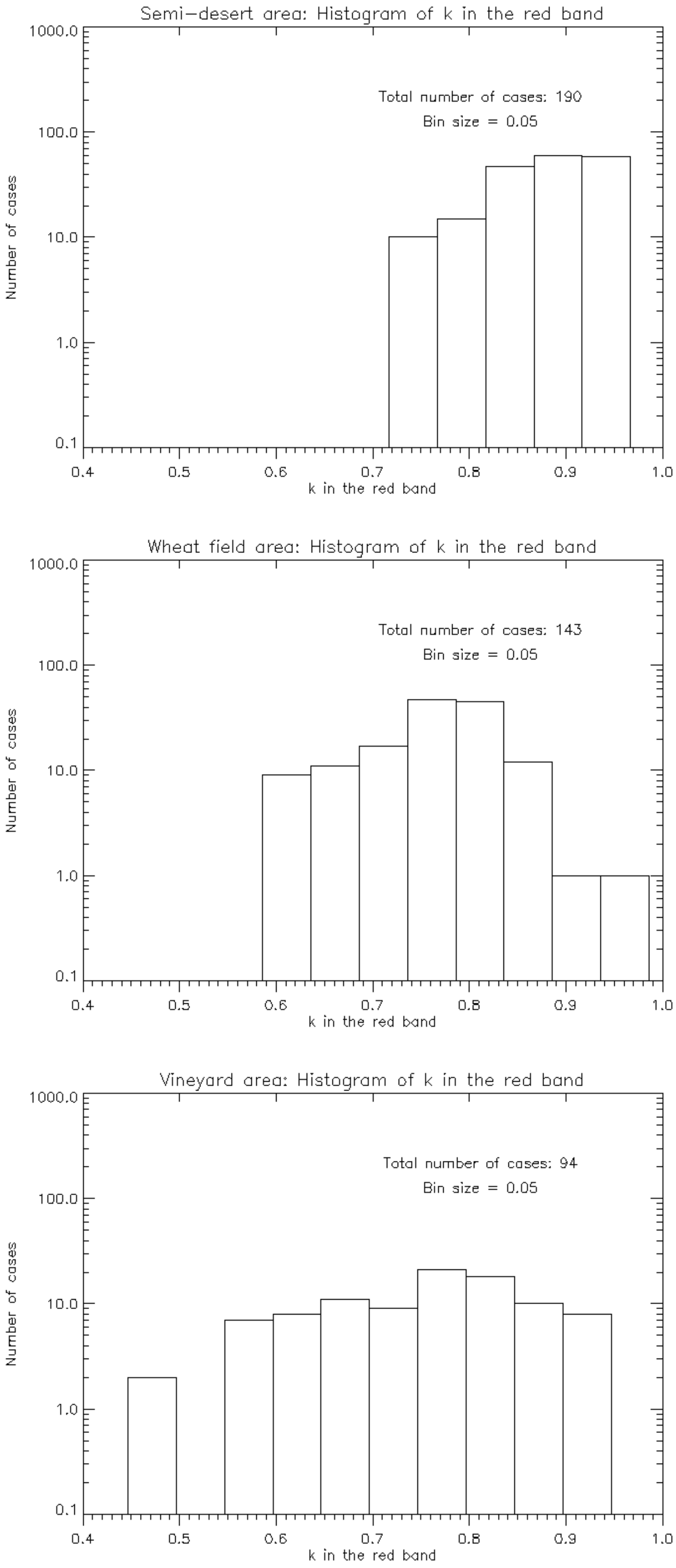

Figure 2.11: Histograms of the parameter in the red spectral band, for the three study areas. Pixel locations are as specified in Figure 2.6.

A similar exploration was conducted on the MISR-HR RPV k parameter in the NIR spectral band. A distinctive signature emerged, in that the average k values for the vineyard area were much higher (around 1.0) than the k values for the semi-desert and the wheat field, which remained under 0.8. Figure 2.12 shows the time series of this parameter in the NIR spectral band for the three study sites. Figure 2.13 shows the overall statistics for those sites as a box plot. This feature is not particular to the selected representative pixel: indeed, Figure 2.14 exhibits the statistics for the MISR-HR RPV parameter k in the red spectral band, averaged over the 108 valid pixels for the vineyard site.

A more detailed look at the temporal distribution of these k values in the NIR spectral band established that the usual bowl-shape anisotropy pattern ($k < 1.0$, near nadir reflectance lower than at larger observation angles) occurs mostly in the winter and spring (May to October), while a bell-shaped anisotropy pattern ($k > 1.0$, near nadir reflectance larger and at larger observation angles) occurs mostly in the summer and autumn (November to April). This observation might be consistent with the phenology of grapevine, but all investigations of the RPV k parameter so far have focused on the values in the red spectral band, so detailed field measurements would be required to establish or confirm any interpretation of this finding in phenological terms.

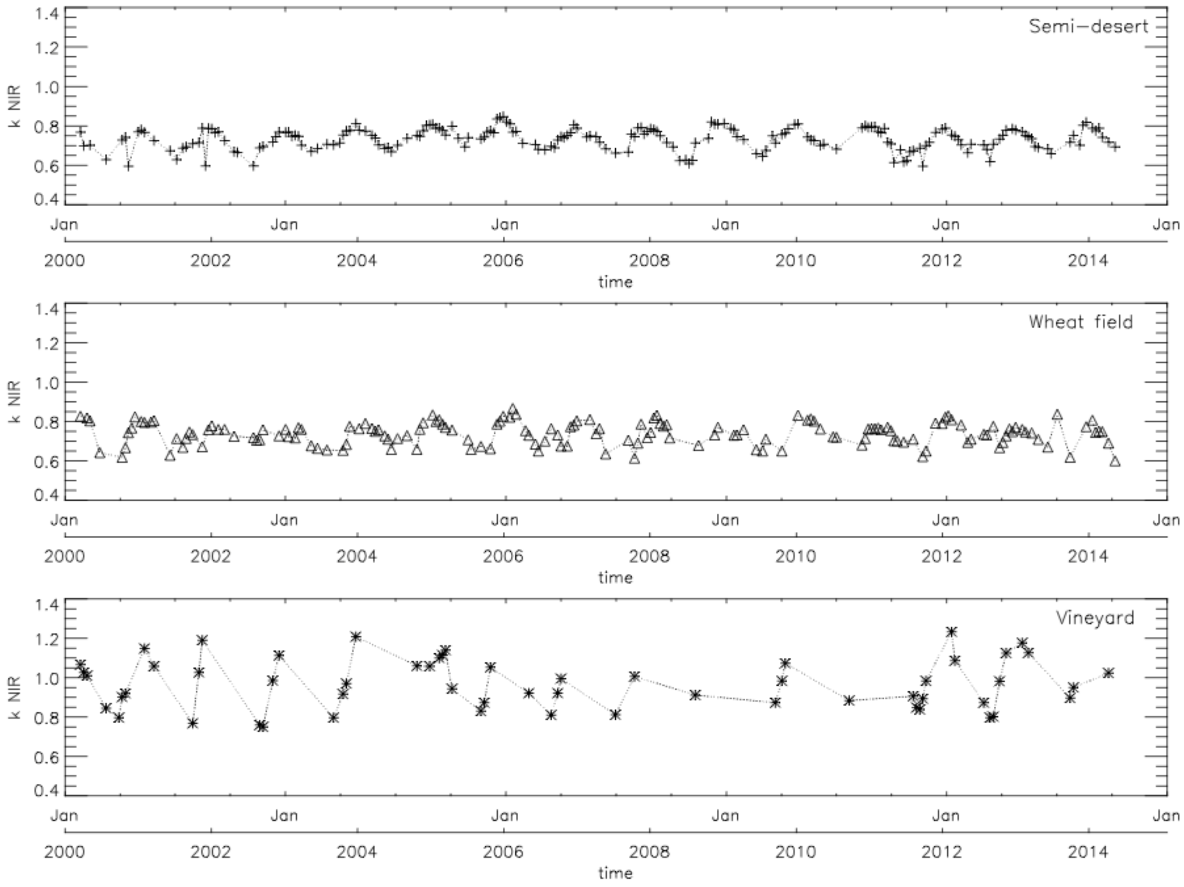

Figure 2.12: Time series plots for NIR band parameter k for the three study areas. Pixels are as specified in Figure 2.6.

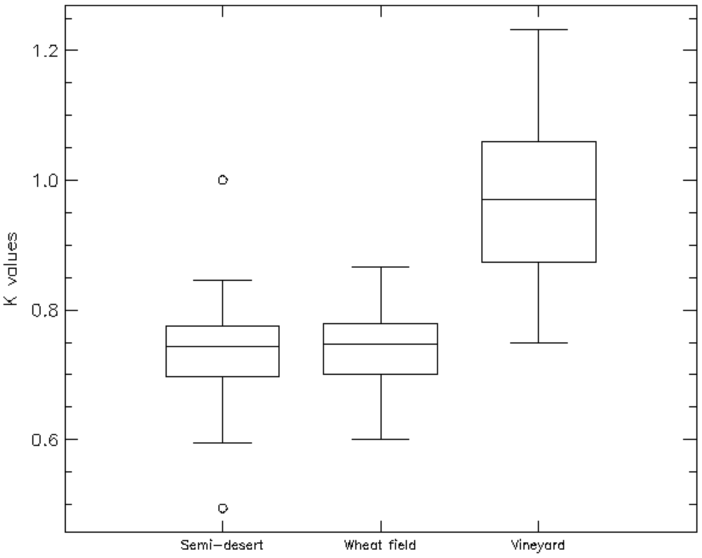

Figure 2.13: Box plot of the MISR-HR RPV k parameter in the NIR spectral band, for the three study areas. Pixel locations are as specified in Figure 2.6.

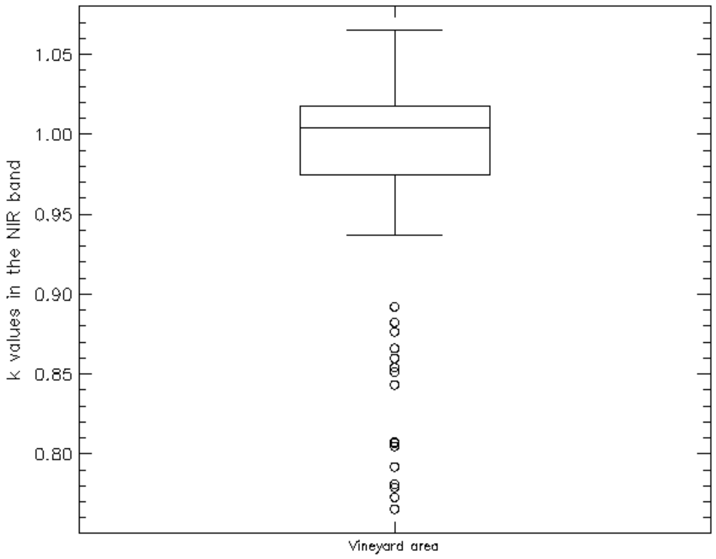

Figure 2.14: Spatial variations of the mean k time series, of all NIR band pixels in the vineyard area.

2.7 Literature review

Remote sensing investigations and applications of surface anisotropy data are relatively scant compared to those based on spectral signatures. Angular variations of the surface reflectance should be analysed because they contain information on the structure of the observed ground surface (Pinty et al., 2002). These authors proposed that parameter k can be used as an indication of surface heterogeneity at the subpixel scale. Their investigation was based on parameter k in the red spectral band, where the reflectance of vegetation and bare soil exhibit strong brightness contrasts. When tall dark vertical objects such as trees are located over a bright background surface and sufficiently separated from each other, nadir observations are influenced by the bright background and lead to relatively high reflectance observations. However, when the same structured target is observed at larger zenith angles, these dark vertical structures obscure the background, resulting in lower reflectance measurements (bell shape pattern). By contrast, a bare background or a fully covering canopy will result in more typical anisotropy shapes where the reflectance increases with the illumination or observation zenith angle (bowl shape pattern). Hence, these authors suggested that the anisotropy described by the k parameter in the red spectral band, which controls the bowl versus bell shape, could be used to define or refine land cover classification and change detection. They also foresaw that k might vary regularly with seasons.

Armston et al. (2007) assessed the relationship between the spectral directional reflectance of the land surface, which is represented by the MISR RPV data, and foliage projective cover (FPC) in Queensland, Australia. FPC is commonly used in Australian vegetation classification frameworks. That paper used multi-day MISR RPV data in 2003—2004 to display the pattern of the spectral directional reflectance variation. The results indicated that the MISR RPV data showed coincident spatial and temporal variations to known vegetation structure changes (e.g., a fire on 28 September, 2003) on the ground surface of the Southern Brigalow Belt Biogeographic Region. These encouraging results foster and justify further research on classifying different environments using the MISR RPV product.

These two studies explored possible applications of the anisotropy data represented by MISR RPV model parameters, but no studies so far have investigated the variations of the MISR-HR RPV product over a long observation period, for example the 14+ years available for this investigation. Documenting changes in the anisotropy at a finer spatial resolution and over a long-term period may help refine the description of the evolution of the environment. This thesis will take advantage of this situation and explore the potential of using the MISR-HR RPV product to characterize land surface processes in the selected sites.

2.8 Summary

This chapter introduced the definition of remote sensing, and some of the main processes involved. Sections 2.2 and 2.3 underscored why the varying angular reflectance signature should not be ignored, even though a large number of remote sensing papers focus almost exclusively on the spectral signature of the terrestrial target. Section 2.4 then showed that multi-angle remote sensing can help characterise the anisotropy of the reflectance field, and provided basic information on the MISR instrument, as well as described some of the MISR-HR data products. Section 2.5 introduced the general information of the three study sites: a semi-desert area, a wheat field and a vineyard area. Section 2.6 provided simple statistical descriptions of the MISR-HR RPV rho and k parameters in the red spectral band, while Section 2.7 surveyed some of the papers which investigated the use of the RPV model parameters in concrete applications.

CHAPTER 3

HANDLING OUTLIERS AND MISSING DATA

3.1 Background and definition

Outliers are quite common in practical measurements in various application areas. Many papers are dedicated to outlier detection: Hodge and Austin (2004) and Sreevidya et al. (2014) studied and compared a few popular outlier detection methods applied in various areas. Gupta, Gao, Aggarwal and Han (2014) surveyed various outlier detection techniques specifically applied to temporal data. These methods include statistics, classification, proximity and clustering. All of them identify and remove potential outliers because “they don’t fit” some preconceived idea or imposed criterion. The assumption is that it may be too risky to include those points in the analysis, as they do not agree with or conform to the preconceived idea, and risk generating abnormalities in the investigation. While the surveys done by Hodge and Austin (2004) and Gupta et al. (2014) revealed that there is no universal outlier detection method suitable for all types of outliers in an arbitrary data set, they also showed that the most appropriate method to be used in a particular application may depend on various factors, such as the data size, the attributes of the data and type(s) of outliers expected, etc.

In practice, outliers are usually considered data values that differ greatly from the vast majority of a data set (Triola, 2012). Although there is no universally accepted definition of an outlier, many statisticians and computer scientists adopted the oft-cited suggestion by Hawkins that an outlier is a suspicious observation, which deviates so greatly from the other observations that it raises the possibility that it was generated by a different mechanism (Hawkins, 1980). In the literature about data mining and statistics, outliers are also referred to as “anomalies”, “abnormalities” or “deviates” (Aggarwal, 2013). Outliers can occur in any data set, including those generated by space-based remote sensing instruments such as MISR.

Optical remote sensing products describing the properties of land surfaces are often affected by missing values due to a number of reasons, including obscuration by deep clouds or thick aerosol layers, or temporary instrument glitches. These events generate gaps in time series, which may interfere with the performance of the subsequent analysis, as many statistical tools and procedures to characterise time series require regular, equally spaced values. The problem is amplified by the presence of outliers, if these must be removed from the dataset.

3.2 Outlier handling process

The process of handling outliers usually involves two steps: identification and treatment (Liu, Verstraete & de Jager, 2017). Outlier detection aims to discover anomalies within a given data set. This technique has been studied extensively and applied widely in various data areas for decades (Han, Kamber & Pei, 2012). The main challenge of outlier detection is that there is no clear boundary to separate the outliers and the majority data values in the definition, which makes the detection of outliers quite subjective (Singh & Upadhyaya, 2012). Thus, outliers should be inspected with care in the process of analysing data. In any case, the detected potential outliers can be treated in the following three ways: ignoring, correcting or eliminating.

Ignoring an outlier might be applied when it has been ascertained that its presence has no significant impact on the outcome of the analysis, or when analysing tools are deemed insensitive to such outliers (Liu, Verstraete & de Jager, 2017). Correcting an outlier is obviously an ideal option, but that approach requires a careful analysis of each suspect data point. In the majority of applications, outliers can't be ignored because they influence the basic statistics of the data set, such as the mean and standard deviation (Bluman, 2012; Osborne & Overbay, 2004; Peterson, Vose, Schmoyer & Razuvaev, 1998). However, this approach is difficult or impossible to implement for very large data sets. Thus, a systematic (machine-driven) and automatic process of identification and elimination of outliers is required.

There is a vast literature about statistical data processing in general and outlier handling in particular (e.g. Grubbs (1969); Heymann, Latapy & Magnien (2012); Manoj and Senthamarai (2013); Seo (2006); Tukey (1977) amongst many others). Traditional statistical methods suffer from intrinsic limitations in identifying and eliminating outliers: firstly, they don't supply any explanation why the 'offending' points should be removed from the data set, other than 'they don't fit' some pre-defined statistical criterion; secondly, there is usually some degree of arbitrariness in setting the boundary between outliers and the bulk of the data. Too conservative a method, for instance, may remove all the potential outliers but may also lead to a significant decrease in the spatial or temporal coverage of the remaining data. On the other hand, eliminating only the extreme outliers may still result in biased outcomes or incorrect conclusions. Hence, it is highly desirable to remove suspicious outliers when there are good reasons to do so. Although this may not always be possible or sufficient, relying on objective methods to detect and handle outliers is recommended to provide explicit reasons for the elimination of data points. These concepts are now applied to the MISR-HR RPV products described in the previous Chapter.

The cost function that is generated in the process of inverting the RPV model and retrieving the model parameters indicates how well the inversion model fits the measurement data: a small value means the model fits the data very well; conversely, an excessively large value indicates that the model is incapable of explaining the variability in the data (Tarantola, 1987; Errico, 1997; Giering & Kaminski, 1998; Tarantola, 2005). This situation occurs when one or more of the data points take on values that are inconsistent with those expected by the model. In that case, the derived model parameters may be questionable because the inversion procedure forces the model to take on unreasonable values to match this (or those) unusual data point(s). Therefore, the cost function can be used to indicate the reliability of the RPV parameters retrieved by inversion, and thus to identify dubious points in the data set.

3.3 Handling outliers in MIRSR-HR RPV products

This section illustrates both the traditional statistical methods and the proposed cost function based method in handling the outliers in MISR-HR RPV parameter time series.

3.3.1 Classical statistical methods

Outliers can occur in any set of measurements or observations, for a variety of reasons, including instrumental problems, exceptional and unexpected changes in the target, or human errors. Remote sensing from space is no exception, as can be seen from the MISR-HR RPV product time series exhibited in Figure 3.1. Four points stand out of from the bulk of the data set: these are candidate outliers. It is easy to detect these values, any of the standard statistical methods will screen them out.

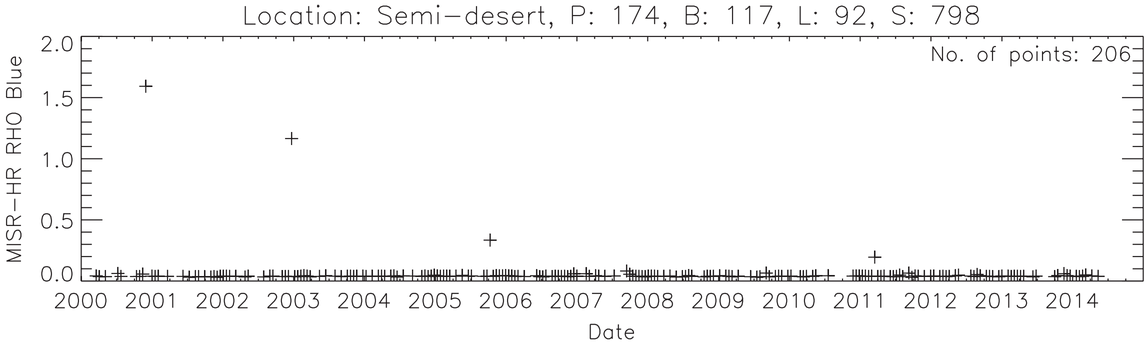

Figure 3.1: Time series of the MISR-HR RPV parameter rho in the blue spectral band for a representative pixel* located in a semi-desert area of South Africa. Four points clearly stand out of the majority data set; they are candidate outliers.

* Location specified by Path 174, Block 117, Line 92 and Sample 798

Statistical methods for outlier detection have been extensively studied in the literature and reviewed in SEMATECH (2013), in particular in Section 7.1.6. One classical example is the box plot method, initially proposed by Tukey (1977), which is a simple popular graphical tool to characterize the variability of a reasonably well-behaved data set. This method first calculates the median and the two quartiles (Q1 and Q3), corresponding to the 25 and 75 percentiles of the cumulative distribution function of the data. It then evaluates the difference between Q3 and Q1, called the interquartile range (IQR). Inner and outer fences are established as follows:

$$\text{Lower Outer Fence (LOF)} = Q1 - 3.0 \times IQR \quad (3.1)$$

$$\text{Lower Inner Fence (LIF)} = Q1 - 1.5 \times IQR \quad (3.2)$$

$$\text{Upper Inner Fence (UIF)} = Q3 + 1.5 \times IQR \quad (3.3)$$

$$\text{Upper Outer Fence (UOF)} = Q3 + 3.0 \times IQR \quad (3.4)$$

Any data values outside the inner fences but inside the outer fences are considered as mild outliers; data values outside the outer fences are considered as extreme outliers (Dawson, 2011).

Table 3.1 shows the box plot statistics for the parameter rho time series shown in Figure 3.1, as well as the other two RPV model parameters k and Theta time series retrieved for the same location. It can be seen that 14 extreme outliers in the rho time series are found by this classical method, including 4 mild high and 4 mild low outliers.

The shortcoming of these purely statistical methods is that outliers would be screened out on the basis of their deviation from some pre-defined measures of central tendency, without inspecting whether they could contain any useful information. In other words, traditional statistical methods can detect outliers but don’t provide any insight about why these values are so extreme.

Table 3.1: Box plot statistics for the RPV parameter time series* (in the blue spectral band).

Ref.	ρ		k		Θ	
	14		0		10	
UOF		.052		1.503		.084
	4		1		4	
UIF		.047		1.271		.016
	33		50		37	
Q3		.041		1.039		−.052
	52		52		52	
Q2		.039		.997		−.078
	52		52		52	
Q1		.037		.884		−.098
	47		38		49	
LIF		.031		.652		−.166
	4		11		2	
LOF		.026		.420		−.234
	0		2		0	
Total	206		206		206	

* Parameter rho time series in Figure 3.1, as well as the other two RPV model parameters k and Theta retrieved simultaneously.

3.3.2 Using the cost function to detect outliers

The main argument proposed in this Chapter is that inspecting the value of the cost function associated with the corresponding RPV parameters in the time series may provide a valid justification for eliminating outliers, since these cases reflect situations where there is a definite mismatch between the underlying model and the observations. Compared to traditional statistical approaches, which rely on the assumption that data are normally distributed (for instance), this proposed cost function based method relies on the selection of a suitable model to describe the reflectance anisotropy, where this model can be independently tested and benchmarked. In other words, this study aims to detect and reject data points in a time series based on using a proven objective, quantifiable rationale model of anisotropy, rather than assuming the unknown statistical distribution of the values in the time series to some known distribution (e.g., normal distribution). In addition, this proposed approach is able to identify and remove the dubious data points whether they appear to be outliers or not.

In the particular case exhibited in Figure 3.1, it turns out that the larger than expected rho values correspond to high cost function values in the same (blue) spectral band, which are J = 326.375 on 29 November 2000, J = 63.8284 on 21 December 2002, J = 51.8984 on 10 October 2005 and J = 37.4346 on 17 March 2011, while in the majority of cases the cost function value is well under 20.

Since the cost function indicates how well all model parameters were retrieved from the inversion process in the same spectral band, it is useful to inspect those other parameters too. Figure 3.2 exhibits the time series for all three RPV parameters for the pixel demonstrated in Figure 3.1 in the blue spectral band, as well as the corresponding cost function values. It can be seen that there are multiple cases of relatively high cost functions corresponding to normal (non-extreme) parameter rho values. Besides, this Figure also reveals that (1) all three RPV parameters show unexpected values; (2) some outliers take on extreme low (rather than extreme high) values (e.g., outliers in the k time series); (3) the three parameters often show dubious values simultaneously, but not always and (4) those apparent outliers are typically associated with high cost function values.

As mentioned in the previous Chapter, the inversion process attempts to attribute the variance in the data to the model parameters, subject to the mathematical description of the model and the constraint of minimizing the cost function. When an outlier is encountered, the unusual variance may be "explainable" by an odd combination of model parameter values, some of which may be unreasonable. It is thus possible to end up with one or more plausible model parameter values, but the cost function will nevertheless be larger than in more typical cases. Based on inspecting hundreds of cases, the net outcome of this study is that the residual cost function value obtained in an inversion procedure constitutes a natural indicator to screen out values that may be less reliable. Therefore, examining the ‘unexpected observations’ in the results of a model inversion procedure by means of the value of the residual cost function turns out to be feasible. Eliminating the dubious points by this method is reasonable since there is objective evidence of a relative mismatch between the model and the data. This approach, as importantly, will remove not only outliers that might be detected in statistical way, but also all results associated with a high cost function value, whether or not they are distinguished as outliers by traditional statistical methods.

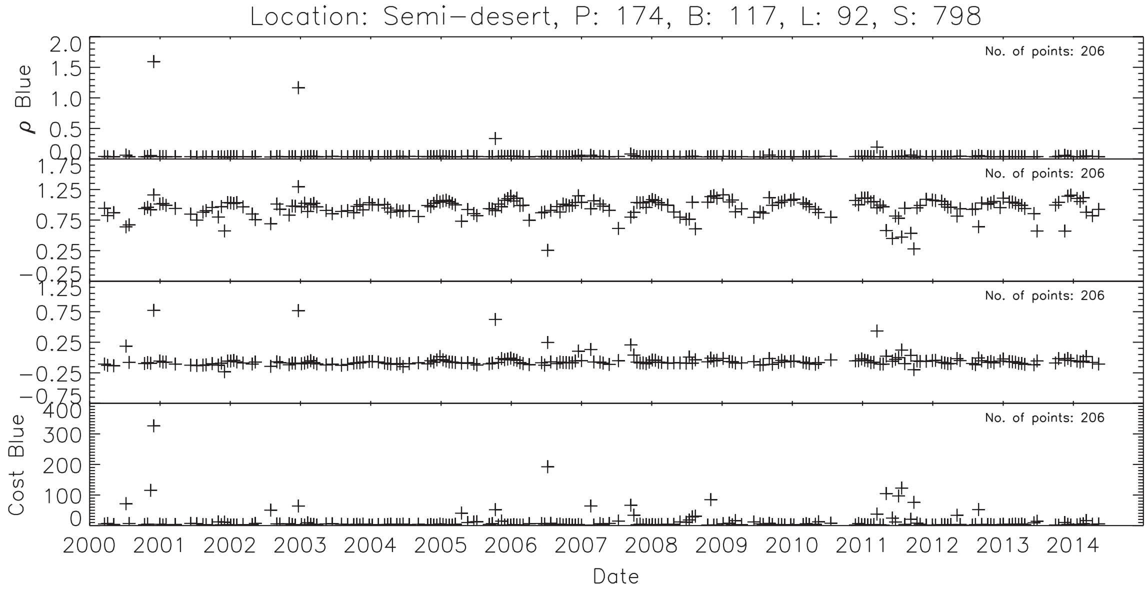

Figure 3.2: Time series for all three RPV parameters, associate with the cost function, in the blue spectral band. Pixel is as specified in Figure 3.1.

The next step for applying this cost function based method is to decide how high the cost function value needs to be to screen the dubious points out of the results of the inversion. Choosing a threshold is nontrivial, as there is no fixed universal value to separate excessive from acceptable cost function values. This threshold may also change according to location. For instance, a cost function value of 30 may be good enough to isolate the high cost function values for the semi-desert area, but it may not be suitable for the vineyard area, as this value might screen out too many measurements, resulting in a sparse time series unsuitable for further analysis. Generally, the selection of such a threshold should depend on the purpose and accuracy requirements of the downstream application. Besides, it is also important to keep in mind that there is a trade-off between reliability of the results and coverage (the number of valid data after elimination): a stricter threshold (low cost function value) may lead to an elimination of more data points which results in a poor temporal coverage.

Inspecting the histogram of residual cost function values for the time series is a useful step in this process. Figure 3.3 shows the histogram of the cost function values in the blue spectral bands for the same pixel as illustrated in Figure 3.1. This is a semi-logarithmic plot, and the majority of the cost function values are below 20. However, this value could be different for different spectral bands, since the RPV model is inverted separately in each of the four spectral bands. The histogram also indicates that the number of cases with a cost function value larger than 40 is very limited (15 in this particular case, among 206 data points, which is about 7% of the cases). Table 3.2 shows the number of the remaining data points in the time series corresponding to different cost function thresholds (again for this particular case).

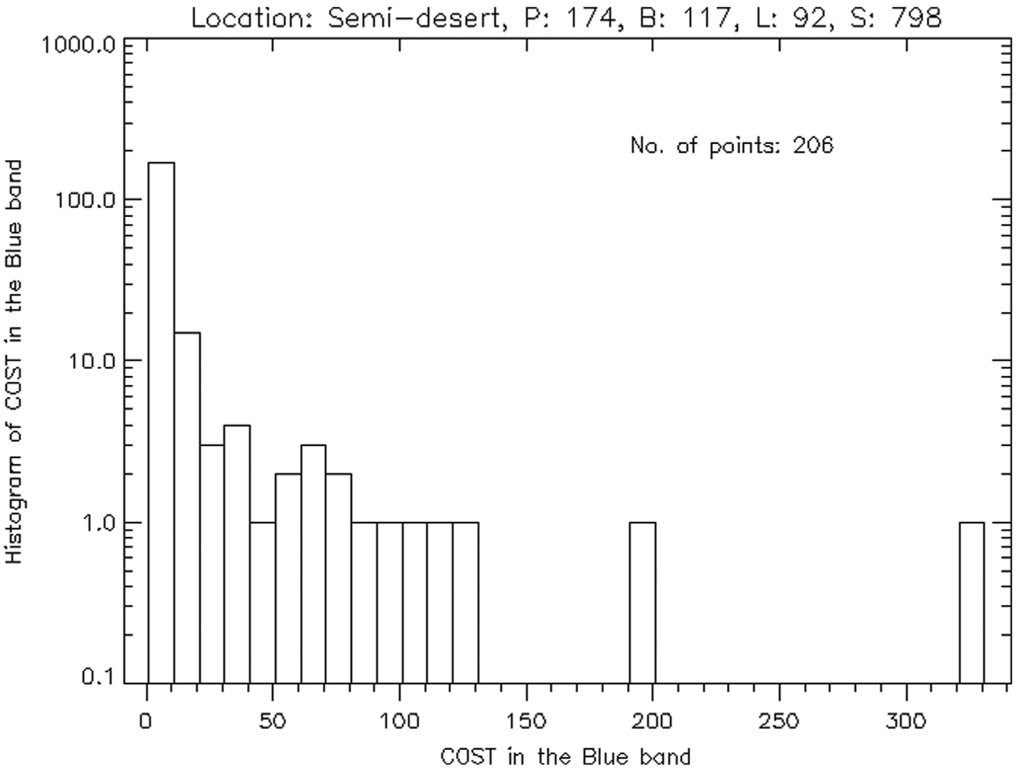

Figure 3.3: Histogram of the cost function values of inverting the RPV model against the MISR-HR surface reflectance in blue spectral bands. The pixel location is as specified in Figure 3.1

Table 3.2: Trade-off between the threshold value of the cost function, and the number of remaining data points in the time series after elimination.

Threshold	500	200	100	50	40	30	20	10
No. of points	206	205	203	191	190	186	183	167

The selection of a reasonable threshold for acceptable cost function values may vary with the specifics of each application as well as the choice of the bin size. Here are three possible rules to establish its value in practice:

(1) The lower boundary of the first empty histogram bin.
(2) The lower boundary of the first histogram bin containing less than 2 items.

(3) The upper boundary of the largest bin for which the sequence of bin populations decreases monotonically.

These suggested methods aim to keep the majority of data and to reject the points that fall in the disconnected bins from the main histogram, but the results may vary according to different bin size. Choosing a proper bin size has been discussed extensively in literature (e.g. Freedman and Diaconis, 1981 and Izenman, 1991). A generic way of calculating the bin size is by W = 2 IQR/ $N^{1/3}$, where W is the bin size, IQR is the Inter-Quartile range, and N is the number of items in the sample. Thus, a bin size of the order of 5 – 10 is suitable for the current application.

As MISR-HR products are systematically generated in all four spectral bands, it is useful to examine the data values obtained on other spectral bands as well. Figure 3.4 shows the time series of parameter rho in all four spectral bands, which are blue, green, red and near-infrared bands, respectively, for the same location as described before. It can be seen that (1) the unexpected values occur generally simultaneously in all four bands, (2) the magnitude of the deviations appears to decrease in time, and (3) that the size of the deviations also decreases with wavelength.

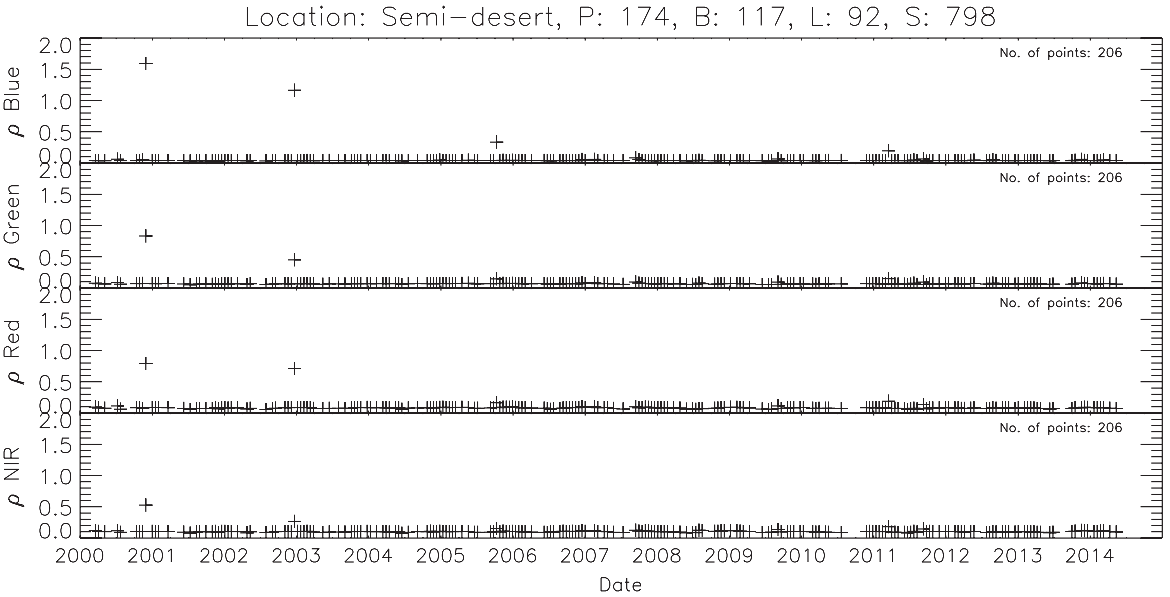

Figure 3.4: Time series of the MISR-HR RPV parameter rho in all four spectral bands. The pixel location is as specified in Figure 3.1. Suspected outliers occur simultaneously in all bands, though the deviations from the main data body decrease along with the wavelength.

Further investigations on other RPV model parameters (k and Theta) and neighbouring pixels showed (1) that unexpected values may occur simultaneously in all parameter in all four spectral bands, and (2) the extreme values were usually found in the same way in neighbouring pixels, often at the same time. Similar results were found in a number of different locations.

3.3.3 Is the cost function sufficient to identify outliers

As discussed above, the residual cost function values can be used as an indicator to identify not only the extreme values in the data set, but also those data items which are likely unreliable although they do not look like outliers. Figure 3.5 shows the time series of parameters rho, k and Theta in the blue spectral band, for the same location as before, but with outliers removed on the basis of a threshold cost function value 20. The number of data points is 183 per time series after the elimination of those unreliable points, instead of the original 206. The result of applying the standard box plot method to this revised time series is shown in Table 3.3.

It can be seen from Table 3.3 that the values of the box plot fences, as well as the three quartiles, are generally close to those shown in Table 3.1, which confirms the insensitivity of these statistics to the presence of outliers. The number of candidate outliers is now much reduced, but there are still 3 dubious larger than the upper outer fence in the rho time series. It is natural to ask whether those remaining extremes should also be eliminated from the data set, on the basis of some partly arbitrary statistical threshold. But since the underlying bidirectional reflectance factor model accounts for the implied variability of the data to an acceptable degree, and the corresponding cost function is below the maximum ‘authorized’ value, it may be appropriate to conduct a sensitivity analysis to quantify the role caused by these remaining outliers on the outcome of the investigation.

In summary, to detect the outliers in the analysed MISR-HR RPV time series, two steps were employed in this research; firstly, checking the MISR-HR RPV data’s corresponding cost function value to detect and remove the outliers; secondly, applying the box plot outlier detection method on the MISR-HR RPV data to identify remaining suspect outliers, if any, that were not found by the first step. Data values outside the inner fences of the box plot method were treated as suspect outliers in this research.

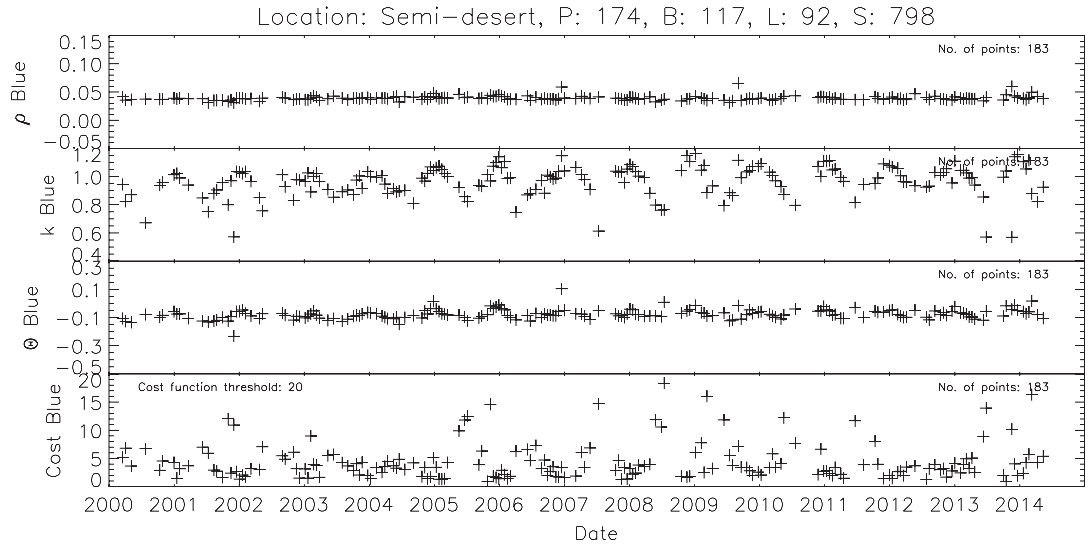

Figure 3.5: Time series of the MISR-HR RPV parameter rho and the associated cost function values in the blue spectral band, with outliers removed on the basis of a cost value threshold of 20. The pixel location is specified in Figure 3.1.

Table 3.3: Box plot statistics for the RPV parameter time series*, after eliminating out outliers based on the cost function.

Ref.	*ρ*		*k*		Θ		Cost	
	3		0		1		5	
UOF		.051		1.437		.067		14.335
	4		0		3		12	
UIF		.046		1.238		.005		9.731
	39		46		42		29	
Q3		.041		1.040		−.058		5.126
	46		46		46		46	
Q2		.038		.990		−.081		3.387
	46		46		46		46	
Q1		.037		.908		−.099		2.056
	42		40		44		45	
LIF		.032		.710		−.162		−2.548
	3		5		0		0	
LOF		.027		.512		−.224		−7.153
	0		0		1		0	

* Parameter rho time series in Figure 3.1, as well as the other two synchronized RPV model parameters k and Theta

3.4 Handling missing values in MISR-HR RPV data

Various methods have been described in the literature to replace missing values in a time series (Schneider, 2001; Troyanskaya et al., 2001; Alonso et al., 2008; Jiang, Lan & Wu, 2009; Honaker & King, 2010, Musial et al., 2011). This is not a trivial step in the whole process of data analysis, especially when there is no prior knowledge about the underlying statistical distribution of the time series. Interpolation methods are easy to implement and therefore frequently used, but they seem most reasonable when there is some degree of correlation between the observations (Jiang, Lan & Wu, 2009). Using a simple interpolation method to fill the gaps in the series may introduce an unwanted bias to the original data set (Musial et al., 2011). In the case of the MISR-HR time series data analysed in this work, the existence and strength of the correlations between the observations are unknown. However, each observation taken by MISR is separate from the adjacent measurements, both in space and time. In other words, the previous and successive measurements do not influence the current measurement. For these reasons, interpolation methods may not be ideal for pre-processing the MISR-HR time series.

Figure 3.6 displays the time series of parameter k for one particular pixel from the semi-desert area. The overall proportion of missing measurements is 40.16%. Missing values are marked out with a low value of 0.4, which is used to make it easy to see the distribution of the missing dates. It can be seen from the plot that the missing values are neither systematically clustered nor regularly spaced. Yet, there remains ample data to document clearly a seasonal cycle that repeats from year to year, even considering such a large number of gaps, and without knowing their precise distribution in time.

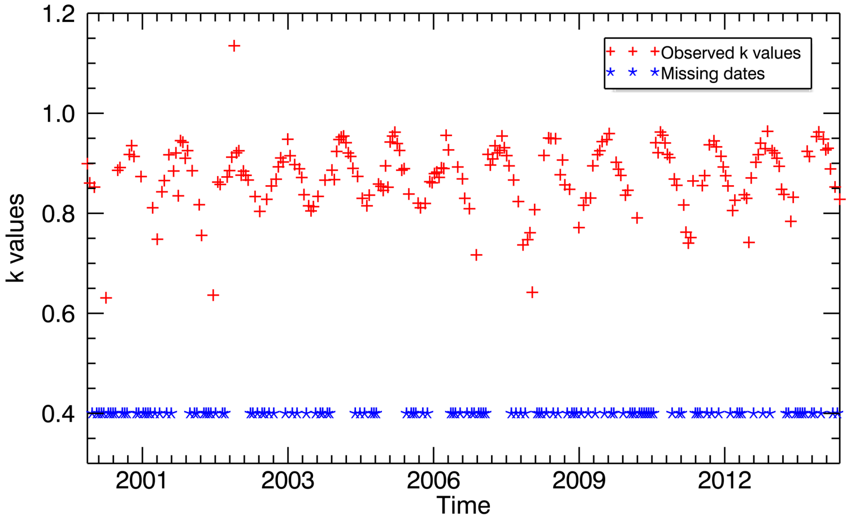

Figure 3.6: Time series of the MISR-HR RPV parameter k in the red spectral band for a pixel in the semi-desert area*, with missing values marked.

* Location specified by Path 174, Block 117, Line 93 and Sample 798

It is clear from the above discussion that attempting to replace missing values must be done with caution, especially when there is no prior knowledge about the distribution of values in the time series. Although reconstructed, continuous, evenly spaced time series would be easier to analyse with standard tools, improperly filling those gaps might introduce biases (in particular spurious frequencies) in the data. To avoid this situation, an alternative way is to employ gap-resilient or gap-insensitive methods to pursue the investigation further.

Since this thesis focuses on the identification of trends and seasonality in time series that are intrinsically not equally spaced and containing potentially large numbers of missing data points, methods that are insensitive to the temporal distribution of the data points have been selected, such as the Mann-Kendal test for trends and the Lomb-Scargle periodogram for establishing the seasonality characteristics of the time series, as will be discussed in the next Chapter. However, the application of a clustering algorithm in Chapter 5 below does require that all time series contain data for the same dates: in that case, the values reconstructed from an analysis in the frequency domain will be used to estimate missing values for that particular purpose.

3.5 Summary

Outliers are a ubiquitous feature of many datasets, especially those resulting from observations or measurements. They do occasionally hint to interesting, unexpected findings, but otherwise may skew or invalidate the analysis. This research proposed a new method to detect outliers, based on the value of the cost function at the end of the inversion of the RPV model against MISR-HR data. This cost function is an indicator of the performance of the RPV model, and therefore of the reliability of the parameters rho, k and Theta that are retrieved from this inversion process. Large values of this cost function identify questionable measurements, whether or not they appear like outliers. This is the specific merit of this cost function outlier detection method.

While this cost function outlier detection method may not detect the same potential outliers as traditional methods, a box plot method was subsequently applied to complement the cost function method. It was also shown that investigating potentially spurious measurements in their spatial, temporal or spectral context may help confirm or infirm their status as outliers.

Missing values are quite common in optical remote sensing data, and this is another critical issue. Many gap-filling methods have been reviewed in the literature, and it was argued that the optimal method to be used to handle the missing observations in a time series depends on the purpose of the research and the attributes of the data set. For the particular case of this research, and to avoid of introducing extra errors in the data set by filling the missing values, it was decided to opt for analysis methods that are insensitive to the presence of gaps in the data.

CHAPTER 4

TIME SERIES ANALYSIS OF MISR-HR RPV PRODUCTS

This chapter explores the trend and seasonality characteristics of time series of MISR-HR anisotropy data.

Time series are sequences of data (observations or measurements) that are collected at different points in time. This kind of data exists in various thematic areas, for example, astronomy, biology, economics, finance, ecology, etc. (Fan & Yao, 2003). Time series analyses have long been used in remote sensing. More recently, Salmon et al. (2011) investigated how to detect new human settlements in South Africa by analysing MODIS time series data with a Multilayer Perceptron. Verbesselt et al. (2010b) used satellite time series data to differentiate between different types of land surface changes. This latter effort focused on detecting and characterising the “Breaks For Additive Seasonal and Trend” (BFAST) by decomposing the time series data into seasonal, trend and noise components. Lhermitte, Verbesselt, Verstraeten and Coppin (2011) worked on monitoring ecosystem dynamics by measuring the time series similarity. All these applications demonstrate that satellite remote sensing data, combined with time series analysis, can be used successfully to document the evolution of terrestrial surfaces during the period of observation. Observational time series can be affected by measurement noise, missing values, and possibly outliers, however the time series analysed below have been pre-processed as explained in the previous Chapter.

Long-term trend analysis plays an important role in the whole process of time series analysis, insofar as it helps describe, model and forecast time series data (Chandler & Scott, 2011). Fitting a straight line to the analysed data set is a common way of detecting a monotonic trend. This process is also known as linear regression (Press, Teukolsky, Vetterling & Flannery, 2007). Although this linear regression method is quite popular in detecting the long-term trend, a few assumptions are required when applying this method (Hirsch, Slack & Smith, 1982). When those assumptions are unverified, or unverifiable, a non-parametric test is preferred as a general approach to detect the trend in the data sequence. The non-parametric Mann-Kendall test has long been considered effective in this regard (Hirsch, Alexander & Smith, 1991). This method does not require the data to be normally distributed, and is flexible with regard to missing values.

Seasonality detection can help in building the underlying model of a time series data (Verbesselt et al., 2010a). Of course, the intrinsic periods of a data set are not always known. Graphical techniques, for instance the run sequence plot, the seasonal sub-series plot and box plot, can help detect the seasonality when the period is known, while an autocorrelation plot is able to detect the seasonality when the period is not known (Tukey, 1977). A dedicated seasonality detection method may be required when the data set contains missing values or when the time series data are unevenly spaced. The Lomb-Scargle periodogram method has been widely used for detecting the seasonality from unevenly spaced time series (Ruf, 1999; Hocke & Kampfer, 2009; Townsend, 2010). This method, originally applied to astronomical data, was developed by Lomb and elaborated by Scargle (Press et al., 2007). In fact, the Lomb-Scargle periodogram method works well to detect frequencies in unevenly spaced data or in time series with missing values (Scargle, 1989). Once the frequencies are detected, the seasonal components can be retrieved from the data set. Based on the Lomb-Scargle periodogram, Hocke and Kampfer (2009) proposed a reconstruction method which is able to retrieve the seasonal component from a time series. This method can help rebuild the seasonal part or the entire time series, including estimating the missing values.

This Chapter explores the feasibility of applying these tools to the MISR-HR RPV time series, as well as the results obtained. Section 4.1 briefly reviews the tools to establish the presence of a trend and then applies them to the time series for the three selected sites described in Chapter 2. Section 4.2 similarly investigates the methods available to document the seasonality of a time series, and assesses the periodic nature of those same time series.

4.1 Trend analysis

4.1.1 Mann-Kendall trend detection method

The Mann-Kendall test is an effective non-parametric trend test method widely used in literature (Hirsch, Alexander & Smith, 1991). This test method is suitable for data where the measurement errors are not normally distributed, and also may be contaminated with missing values and outliers. The Mann-Kendall approach tests the null hypothesis H_0 that there is no trend in the data against the alternative hypothesis H_1 that a trend does exist, at a given significance level (Hirsch, Slack & Smith, 1982; Longobardi & Villani, 2009). H_0 is initially assumed to be true, while the Mann-Kendall test aims to provide reasonable doubt to reject H_0 and accept H_1. It can be applied under the assumption that the measurements obtained over time are independent and identically distributed, which means the observations are not serially correlated over time.

Let $X_1, X_2, \ldots, X_n$ be a sequence of measurements over time. The core of the Mann-Kendall test is to determine the sign (+/-) of all $n(n-1)/2$ possible differences $X_j - X_i$, where $j > i$ (http://vsp.pnnl.gov/help/Vsample/Design_Trend_Mann_Kendall.htm). These differences are $X_2 - X_1$, $X_3 - X_1$, $\ldots,$ $X_n - X_1$, $X_3 - X_2$, $X_4 - X_2$, $\ldots,$ $X_n - X_{n-2}$, $X_n - X_{n-1}$. Under hypothesis H_0, the Mann-Kendall statistical test is:

$$S = {}_{i=1}^{n-1}{}_{j=i+1}^{n} sgn(X_j - X_i) \qquad (4.1)$$

where

$$sgn(X_j - X_i) = \begin{cases} +1 & (X_j - X_i) > 0 \\ 0 & (X_j - X_i) = 0 \\ -1 & (X_j - X_i) < 0 \end{cases} \qquad (4.2)$$

When n 8, *S* is approximately normally distributed according to the assumptions of the Mann-Kendall test, with zero mean and variance given by:

$$\sigma^2 = \frac{n(n-1)(2n+5)}{18} \qquad (4.3)$$

Consequently, statistic *Z* follows a standardised normal distribution:

$$Z = \begin{cases} \frac{S-1}{\sigma} & if \;\; S > 0 \\ 0 & if \;\; S = 0 \\ \frac{S+1}{\sigma} & if \;\; S < 0 \end{cases} \qquad (4.4)$$

With a given significance level α, hypothesis H_0 is rejected when

$$|Z| > Z_{1-\alpha/2} \qquad (4.5)$$

Otherwise, H_0 is accepted. A significance level of 0.05 has been adopted for the current purpose. The value of $Z_{1-{}^{\alpha}/_{2}}$ for a significance level of 0.05 from the standard normal table is 1.96.

4.1.2 Sen slope estimator

The Mann-Kendall test only gives information about the monotonic trend, such as upward, downward or no significant trend. The magnitude of the trend can be measured by Sen's slope estimator (Sen, 1968).

Sen's slope is a rank-based parameter. Following the Mann-Kendall test, the slopes for all the $\mathrm{N} = n(n-1)/2$ possible differences $X_j - X_i$ are calculated by:

$$Q_k = \frac{X_j - X_i}{j-i} \quad \mathrm{k} = 1, 2, \ldots, \mathrm{N} \qquad (4.6)$$

where i, j, X_i, X_j have the same meaning as in Equation 4.1. Sen's slope is defined as the median of these N values of Q (Sen, 1968; Yadav & Mishra, 2015).

4.1.3 Illustration of the trend detection tools applied on MISR-HR anisotropy data

The Mann-Kendall test was first systematically applied to the MISR-HR RPV time series of the three selected sites to detect the presence of a monotonic trend, in which case the Sen slope estimator was also computed. A subset of the results is shown and discussed below.

Figure 4.1 illustrates the presence of a linear trend in the MISR-HR RPV k parameter in the blue spectral band, in this case for a single selected pixel in the semi-desert site (|Z| = 3.91 and Q = 0.00052). This time series, spanning the period from 18 March 2000 to 12 May 2014, contains 183 data points (out of a theoretical maximum of 324) once outliers and missing data were accounted for.

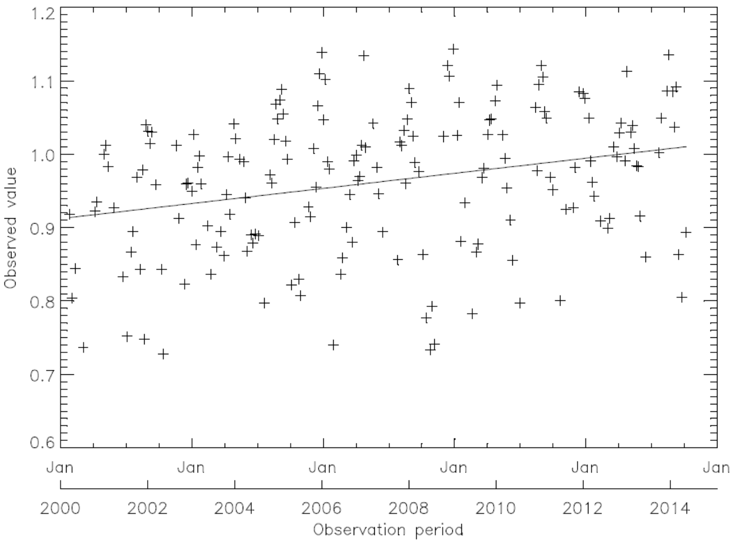

Figure 4.1: Time series of the MISR-HR RPV parameter k in the blue spectral band for a single pixel*, together with the associated linear trend line.

* Location specified by Path 174, Block 117, Line 93 and Block 798

Tables 4.1, 4.2 and 4.3 summarize these results for each of the three RPV model parameters, in each of the 4 MISR spectral bands, and for each of the three sites described in Chapter 2, when all pixels are each site are combined. Specifically, each table entry indicates the proportion of all pixels in the specified site which exhibit a significant positive trend, no significant trend, or a significant negative trend.

Table 4.1: Proportion of pixels in each of 3 significant trend categories for RPV parameter rho over the three study areas.

Study sites	**Trend categories**	**Blue** **%**	**Green** **%**	**Red** **%**	**NIR** **%**
Semi-desert area*	Upward	0.00	0.00	0.00	0.00
	No trend	100	0.82	30.58	86.78
	Downward	0.00	99.17	69.42	13.22
Wheat field*	Upward	0.00	0.00	0.00	0.00
	No trend	100	100	97.52	88.43
	Downward	0.00	0.00	2.48	11.57
Vineyard area**	Upward	0.00	0.00	0.92	3.70
	No trend	48.15	0.00	44.44	89.81
	Downward	51.85	100	54.63	6.48

* This study site covered 121 pixels
** This study site covered 108 pixels (see Section 2.5)

It can be seen that the results differ among the three sites: while no appreciable trend was detected in the rho parameter in any of the 4 spectral bands for the wheat field site, virtually all pixels in the semi-desert and vineyard sites exhibit a downward trend in the green spectral band, and more than half of those pixels experience a downward trend in the red spectral band.

Table 4.2: Proportions of pixels in each of 3 significant trend categories for RPV parameter k over the three study areas.

Study sites	Trend categories	Blue %	Green %	Red %	NIR %
Semi-desert area*	Upward	100	36.36	0.00	0.00
	No trend	0.00	63.64	100	100
	Downward	0.00	0.00	0.00	0.00
Wheat field*	Upward	66.94	6.61	14.05	0.00
	No trend	33.06	93.39	85.95	96.69
	Downward	0.00	0.00	0.00	3.31
Vineyard area**	Upward	62.96	40.74	15.74	1.85
	No trend	37.04	52.78	60.19	87.96
	Downward	0.00	6.48	24.07	10.19

* This study site covered 121 pixels
** This study site covered 108 pixels (see Section 2.5)

By contrast, in the case of the RPV parameter k, all pixels in the semi-desert site and about 2/3 of the pixels in the other two sites exhibited a significant upward trend in the blue band. Results from the other sites are more mixed, though no or very few pixels showed a downward trend at any of the sites and in any of the spectral bands, except the vineyard site in the red (24%) and NIR (10%) spectral bands.

Lastly, the results for the RPV Theta parameter show yet another pattern of trends: by and large, all three sites exhibit a strong upward trend in the blue spectral band, while the semi-desert and wheat field sites exhibit little or no significant trend in the other three spectral bands. The vineyard site differs somewhat in this case, with varying proportions of pixels showing either no trend or a positive trend, depending on the spectral band.

Table 4.3: Proportion of pixels in trend categories for parameter Theta over the three study areas.

Study sites	Trend categories	Blue %	Green %	Red %	NIR %
Semi-desert area*	Upward	100	0.00	0.00	0.00
	No trend	0.00	100	100	100
	Downward	0.00	0.00	0.00	0.00
Wheat field*	Upward	97.52	8.26	7.44	4.13
	No trend	2.48	91.74	91.74	95.87
	Downward	0.00	0.00	0.82	0.00
Vineyard area**	Upward	93.52	6.48	34.26	64.81
	No trend	6.48	93.52	55.56	35.19
	Downward	0.00	0.00	10.18	0.00

* This study site covered 121 pixels
** This study site covered 108 pixels (see Section 2.5)

The formal interpretation of these results would require access to field data to understand how these trends may be related to land cover or land use changes in those sites. However, there is another possible explanation, namely that the calibration of the MISR instrument may suffer from a slight but noticeable bias over this period. It is important, in this context, to note that the MISR instrument is calibrated on-board, through the periodic deployment of Spectralon™ panels in front of the cameras. However, that instrument was designed for a lifetime of 5 years and has now been operational for 18+ years (it is still working as of this writing). A comprehensive process of re-analysis of the entire MISR archive is underway at NASA and may be able to shed some light on this matter, so it will be interesting to re-evaluate the results described here in the context of those developments, once the reprocessing has been completed (not before 2021).

In the meantime, since the trend statistics have been computed for every pixel, in each of the 4 spectral bands and at each of the 3 sites, it is also possible to map the trend slope at each site, as measured by Sen's estimator. Figures 4.2 and 4.3 show the spatial distribution of the RPV rho parameter downward slope in the green spectral band, for the semi-desert and the vineyard sites, as examples. In these maps, the coordinates vary from 0 to 10 (11 pixels on the side) and match the locations of the original MISR-HR pixels. Hence, point (10, 10) on the slope map corresponds to pixel s05_+005_-005 (north-east corner) in the MISR-HR dataset. The yellow diamond near point (1, 2) in Figure 6.1 represents the single pixel that exhibits no significant slope in the trend, as detected by the Mann-Kendall test at a confidence level of 0.95 (and for which no Sen's slope value was computed). In Figure 4.3, the greyed-out areas near the north-west corner and along the northern border of the map correspond to the area where data are missing due to topography. The colour bars in Figures 6.1 and 6.2 both range from -1.03×10^{-4} to -1.10×10^{-5}, with darker tones indicating stronger (negative) slopes.

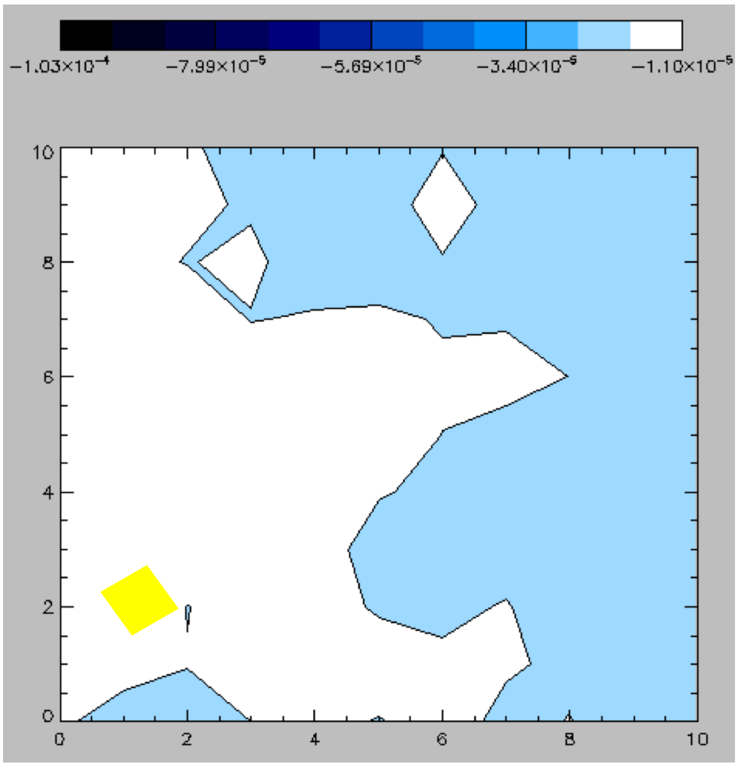

Figure 4.2: Map of the intensity of the downward slope of the trend for the MISR-HR parameter rho in the green spectral band for the semi-desert site. The yellow diamond indicates the location of the single pixel in that site that does not exhibit a significant trend.

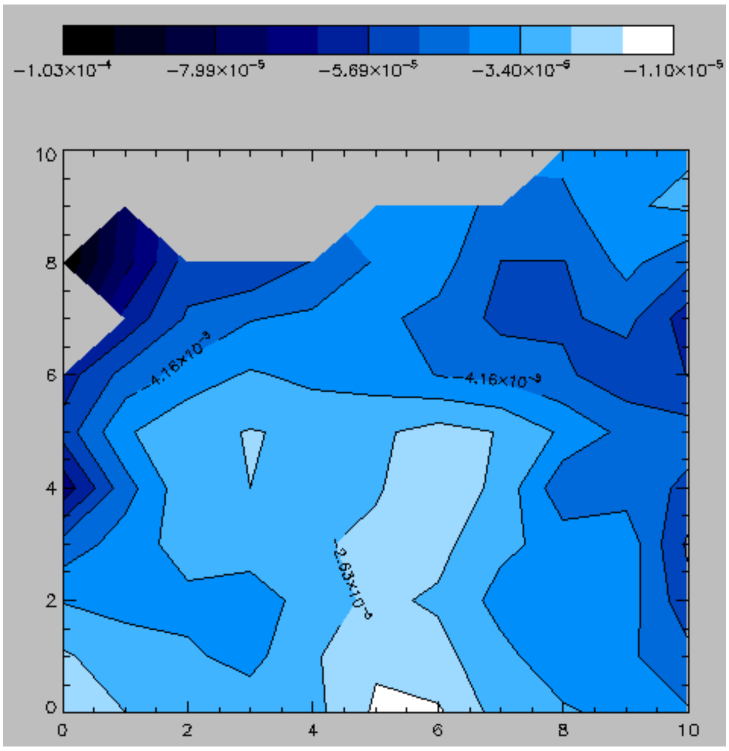

Figure 4.3: Map of the intensity of the downward slope of the trend for the MISR-HR parameter rho in the green spectral band for the vineyard site. The grey area covers the region where no data are available due to the topography.

It can be seen that the downward trend over the semi-desert site is more spatially homogeneous and weaker than for the vineyard site. These spatial variations could result from local changes in the environment (although the semi-desert site is virtually free from human interference), or from a small, progressive degradation of the instrument that is not properly characterized by the regular calibration procedure, or—more likely—by a combination of both factors. Indeed, a calibration issue should translate into a uniform or smoothly varying trend, while land cover and land use (LCLU) changes could well be responsible for local changes. As noted earlier, unpacking these potential causes would require substantial field data.

4.2 Seasonal analysis

In addition to trends, time series often display a range of fluctuations that occur at specific frequencies. This is the case also for observations acquired from space, especially over vegetated areas since they follow the natural yearly seasonal cycle. Figure 4.4 shows the time series of the MISR-HR RPV parameter k in the red spectral band for the selected pixel of the semi-desert site. The yearly seasonal cycle is clearly visible despite the fact that outliers and missing values have partly depleted the data set. This particular time series will be used as the primary example for the subsequent discussions, though the data for any other pixel of either of the test sites would clearly be as appropriate.

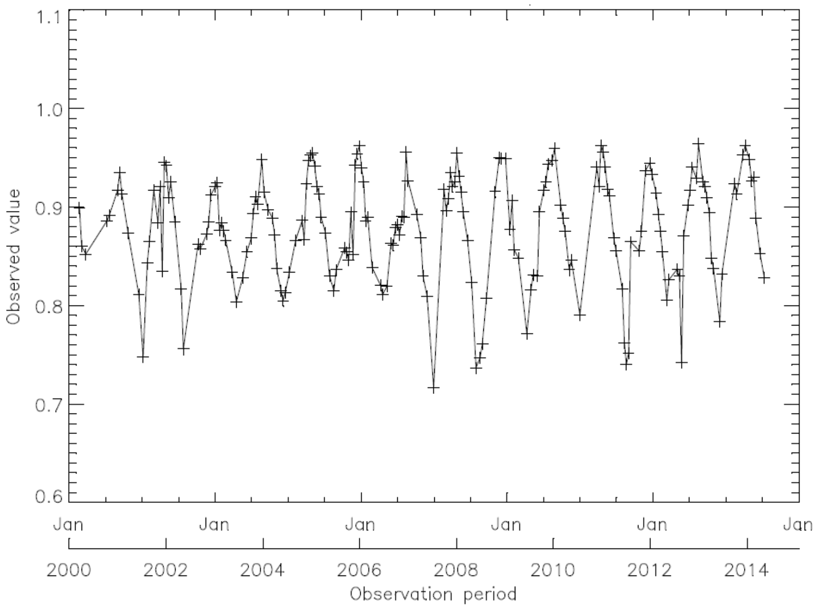

Figure 4.4: Time series of the MISR-HR RPV parameter k in the red spectral band for the selected pixel of the semi-desert site (Path 174, Block 117, Line 93 and Sample 798). Note the obvious seasonal cycle.

4.2.1 The Lomb-Scargle periodogram

Systematic periodicity is often expected in time series, as is the case with the annual seasonal cycle of vegetation, because there is a clear understanding of the causal factor at work. In other cases, it is not known a priori whether a time series may contain periodic components or not. However, classical methods to determine the properties of time series in the spectral domain typically require complete data sets (no missing values) as well as data points equally spaced in time. This is hard or impossible to achieve with observations of the natural environment, including from space with optical instruments, because of the ubiquitous nature of clouds.

Specific statistical tests such as the autocorrelation function have been developed to detect the presence of periodic components in time series (Nopiah et al., 2012). More importantly, spectral methods have been developed to analyse data sets where the observations are not equally spaced. The Lomb-Scargle periodogram method, in particular, is widely used for detecting the periodic components from unevenly-spaced time series (Ruf, 1999; Hocke & Kampfer, 2009; Townsend, 2010). This method was originally developed by Lomb and elaborated by Scargle (Press et al., 2007), who used it to analyse astronomical data that have irregular samples or regular samples with missing values (Scargle, 1989).

The periodogram is one of the statistical tools used for detecting periodic variations in a time series (Press et al., 2007). It determines the "power" over a spectrum of frequencies. Suppose there is a time series with N data points $h_i = h(t_i)$, where $i = 0, \_, N-1$. The mean and variance of the time series are given by the following equations:

$$\bar{h} = \frac{1}{N}\sum_{i=0}^{N-1} h_i \qquad (4.7)$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=0}^{N-1}(h_i - \bar{h})^2 \qquad (4.8)$$

The Lomb-Scargle periodogram, developed by Lomb (1976) and elaborated by Scargle (1982), is defined as:

$$P_N(\omega) = \frac{1}{2\sigma^2}\left\{\frac{[\sum_i (h_i-\bar{h}) \cos \omega(t_i-\tau)]^2}{\sum_i \cos^2 \omega(t_i-\tau)} + \frac{[\sum_i (h_i-\bar{h}) \sin \omega(t_i-\tau)]^2}{\sum_i \sin^2 \omega(t_i-\tau)}\right\} \qquad (4.9)$$

where the angular frequency is:

$$\omega = 2f \qquad (4.10)$$

and the frequency-dependent time offset is calculated by:

$$\tan(2\omega\tau) = \frac{\sum_i \sin 2\omega t_i}{\sum_i \cos 2\omega t_i} \quad (4.11)$$

Knowing how to calculate $P_N(\omega)$, it is interesting to quantify how significant a peak in the spectrum is. Scargle (1982) pointed out that $P_N(\omega)$ has an exponential probability distribution with unit mean at any and in the case of the null hypothesis. This means that $P_N(\omega)$ will be a value between some positive z and $z + dz$ and the probability is $\exp(-z)dz$. Suppose there are M independent frequencies, the probability that none gives values larger than z is calculated by $(1 - e^{-Z})^M$. Thus,

$$P(> z) \equiv 1 - (1 - e^{-z})^M \quad (4.12)$$

is the false-alarm probability of the null hypothesis, which means the significance level of any $P_N(\omega)$ can be measured. Since the interesting region of the significance level is usually a very small number, $\ll 1$, Equation 4.12 can be shortened to

$$P(> z) \approx Me^{-z} \quad (4.13)$$

Horne and Baliunas (1986) tested the determination of M in various cases and pointed out that M is very nearly equal to N when the time series is approximately equally spaced and when the sampled frequencies extend between 0 and the Nyquist frequency. With this assumption, the significance level of any peak in $P_N(\omega)$ can be calculated; the smaller the value of the false-alarm probability, the higher the significance that a periodic signal exists.

The Lomb-Scargle periodogram method has been used here to detect seasonality for the time series data illustrated in Figure 4.4; the periodogram is shown in Figure 4.5. It can be seen from the periodogram that there is an obvious peak with the power value almost 70. The peak corresponds to the frequency 0.044, which corresponds to a period of one calendar year. (The frequency value 0.044 was converted to a period of 22.67 data points. Because the MISR instrument collects data every 16 days, over the period of a year there are 22 or 23 observations resulting in data points, so some calendar years have 22 data points and the others have 23. This 22.67-data-points period is an average value for the observed 14+ years, thus it is reasonable to approximate the period of 22.67 data points to one calendar year.) Significance levels 0.1, 0.05 and 0.01 are popular in practice, and sometime even a significance level of 0.5 is used. The false alarm values corresponding to these significance levels were calculated according to Equation 4.13, and were marked on the periodogram plot in Figure 4.5. As the amplitude of the peak is much greater than the significance level 0.01 threshold, it can be asserted with strong confidence that seasonality exists in the time series.

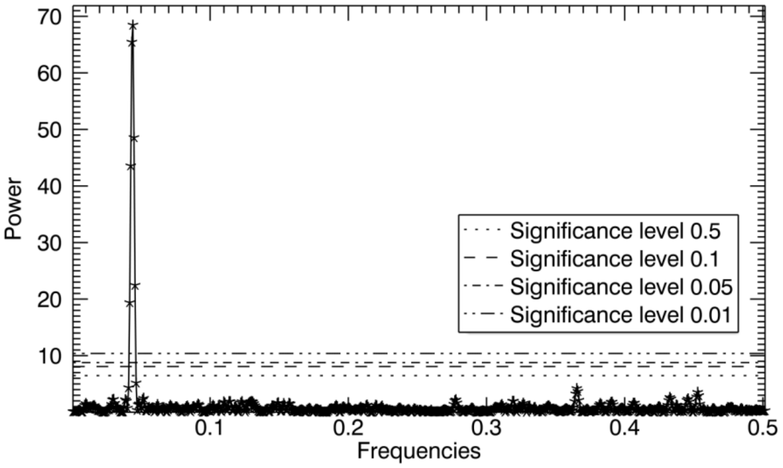

Figure 4.5: The Lomb-Scargle periodogram of the parameter k time series, as specified in Figure 4.4.

4.2.2 Seasonal component reconstruction method

Once the seasonality of the MISR-HR RPV time series was detected by the Lomb-Scargle periodogram method, it was interesting to see how the seasonal variations behaved through the observation period, because seasonal variation may help to explain phenomena on the ground or help in the time series prediction. It was also interesting to see in literature how residuals varied when the seasonal component was extracted from the time series, because analysis of the residuals may reveal stochastic change in the time series.

The seasonal part can be retrieved by the Lomb-Scargle algorithm-based reconstruction method. Scargle (1989) proposed a method of reconstructing an unevenly-spaced time series. First, he used the Lomb-Scargle method to compute the real and imaginary parts of the power spectrum which converted the time series into the frequency domain, and then reverted back to the time domain to obtain a time series of equally spaced points. Based on Scargle's idea, Hocke and Kampfer (2009) extended the method to estimate the real and imaginary part of the spectral components, since a complex Fourier spectrum can be generated on the basis of the amplitude and phase information. The inverse Fourier transform can then be applied to that spectrum to retrieve an evenly-spaced time series. To generate a simulated signal that reproduces the seasonal fluctuations only, it is sufficient to include only those frequencies that are deemed statistically significant in the inverse Fourier transform.

The mathematical algorithm of this reconstruction method is described below (Hocke & Kampfer, 2009):

The Lomb-Scargle periodogram method calculates the normalised amplitude $P_N(\omega)$. Here the reconstruction method needs to recover the spectrum amplitude A_{FT} which is calculated by:

$$A_{FT}(\omega) = \sqrt{\frac{N}{2}\sigma^2 P_N(\omega)} \qquad (4.14)$$

The phase $\varphi_{FT}$of the complex Fourier spectrum is given by:

$$\varphi_{FT} = \arctan(I, R) + \omega t_{ave} + \ \Theta \qquad (4.15)$$

where variables R, I are calculated by:

$$R(\omega) = \sum_i (h_i - \bar{h}) \cos \omega (t_i - \tau) \qquad (4.16)$$

$$I(\omega) = \sum_i (h_i - \bar{h}) \sin \omega (t_i - \tau) \qquad (4.17)$$

ωt_{ave} is the phase correction variable with the average time $t_{ave} = (t_1 + t_N)/2$; the Lomb-Scargle periodogram phase Θ is measured with the four-quadrant inverse tangent:

$$\Theta = \frac{1}{2}\arctan(\sum_i \sin(2\omega t_i), \sum_i \cos(2\omega t_i)) \qquad (4.18)$$

The real part of the Fourier spectrum is:

$$R_{FT} = A_{FT} cos\varphi_{FT} \qquad (4.19)$$

The imaginary part of the Fourier spectrum is:

$$I_{FT} = A_{FT} sin\varphi_{FT} \qquad (4.20)$$

Hocke and Kampfer (2009) used MATLAB computer language to explore this reconstruction method. The Fast Fourier Transform (FFT) algorithm of this programming language needs a complex vector *F* in the following format:

$$F = [complex(0,0), complex(R_{FT}, I_{FT}), reverse[complex(R_{FT}, -I_{FT})]\,] \qquad (4.21)$$

The first number is the zero mean of the time series, followed by the complex spectrum, and last is the reversed complex conjugated spectrum. (The MATLAB programme lspr.m supplied by Hocke and Kampfer (2009) provides full details.)

Once the complex Fourier spectrum is determined, the inverse Fourier transform can be applied. The real part of the inverse Fourier transform of *F* is the reconstructed evenly-spaced time series. This is the way of fully rebuilding the time series, which means all frequency components are taken back to the time domain. An evenly-spaced time series is reconstructed in this way, which also includes the noise component.

The seasonal part can be retrieved by modifying the Fourier spectrum before the inverse Fourier transform is applied. The modification is the process of setting the frequency power of the unwanted frequency component to zero and keeping the desired spectrum. For instance, the modification can be applied by setting the frequency power lower than significance level 0.05 to 0, and returning the remaining frequency component to the inverse Fourier transform; seasonal components with a confidence level of 0.95 are retrieved in this way.

4.2.3 Testing the reconstruction method on MISR-HR RPV products

Hocke and Kampfer (2009) supplied the complex Fourier spectrum construction computer program 'lspr.m' in MATLAB format. This program was converted into IDL (Interactive Data Language), and modified to retrieve the required seasonal component from both the evenly- and unevenly-spaced time series. The MISR-HR time series data used in the current research to test the Lomb-Scargle reconstruction method is the same data set used for illustrating the Lomb-Scargle periodogram algorithm, namely the red band parameter k time series as specified in Figure 4.4. Applying the IDL reconstruction program to this k time series, the reconstructed evenly-spaced time series is plotted in Figure 4.6. No modification was made to the Fourier spectrum in this reconstruction process, which means noise spectra were also taken into the inverse Fourier transform and were reconstructed.

The seasonal part of this k time series was retrieved by setting the Fourier spectrum threshold at the significance value equal to 0.05 in the IDL reconstruction program, since this threshold can get rid of most of the noise signals and keep the significant seasonal components. The spectral components with the amplitude larger than this threshold were reconstructed. The retrieved seasonal part as well as the original time series is plotted in Figure 4.7. It can be seen from the plot that the seasonal part consists of an annually-repeated sinusoidal wave. It is also noticeable that the middle part of the reconstructed seasonal part fits the original time series very well, but the sinusoidal waves are 'shrunk' at both ends of the time series. Hocke and Kampfer (2009) explained this is a drawback of the reconstruction method and suggested using the middle part of the reconstructed time series rather than using all of them. Musial, Verstraete and Gobron (2011) addressed this problem by applying the Kaiser-Bessel window instead of the Hamming window in Hocke's 'lspr.m' program. They pointed out that the Lomb-Scargle algorithm together with the Kaiser-Bessel window can overcome the 'shrinkage' problem and regenerate reliable reconstructed time series. Retrieving an accurate seasonal component is not the purpose of the current research. The retrieved seasonal period and the reconstructed signal pattern of each season from the MISR-HR anisotropy data are adequate for revealing the ground information of the studied areas. The Lomb-Scargle periodogram and Hocke and Kampfer's reconstruction algorithm are good enough to display the seasonal period and signal patterns of the MISR-HR RPV time series. The 'shrinkage' problem did exist in the reconstructed seasonal component but only affected the amplitudes rather than the pattern of the seasonal waves; thus, the 'shrinkage' problem was put aside in this study.

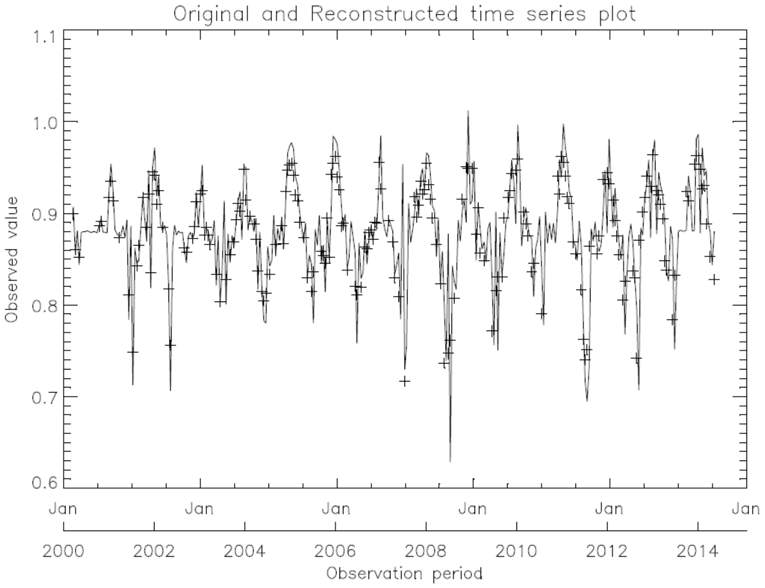

Figure 4.6: The original parameter k time series (Figure 4.4) against the reconstructed k time series derived by the Lomb-Scargle periodogram-based reconstruction method.

'+' represents the original k time series; '-' represents the reconstructed k time series

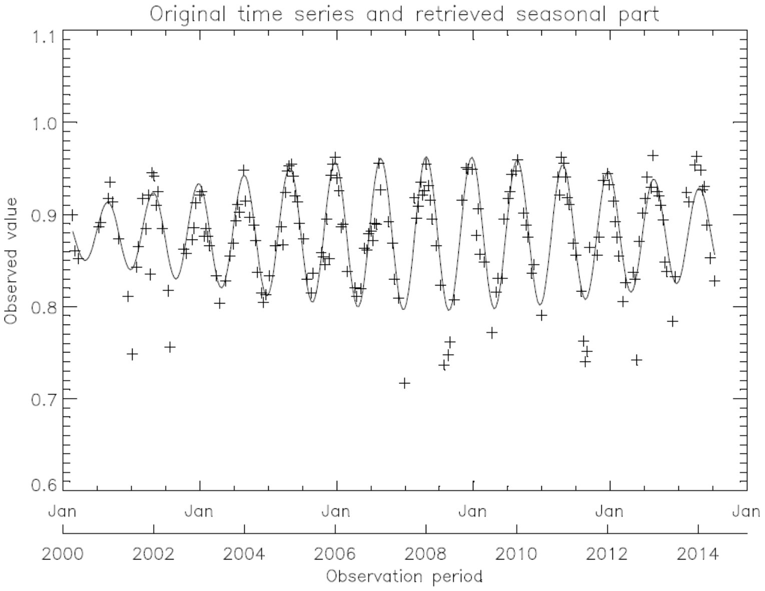

Figure 4.7: Time series of parameter k (Figure 4.4) against the reconstructed seasonal part* of the k time series.

* Six frequencies with power greater than significance value 0.05 on the periodogram (Figure 4.5) were used to reconstruct this seasonal part

4.2.4 Discussion of the reconstruction method

The previous sections showed that the Lomb-Scargle algorithm can be used to estimate the periodogram of an unevenly-distributed time series, and that this power spectrum, in turn, can be exploited to regenerate a synthetic time series very similar to the original one, which can be sampled at arbitrary intervals. This is one way to fill the gaps in an existing time series, or to resample a time series on a different temporal grid. This reconstruction method also allows rebuilding the required frequency signal by modifying the unwanted frequency spectrum to zero. The seasonal component of a time series can be retrieved in this way, by setting the power of the frequency spectrum less than a certain significance level to zero. The retrieved seasonal component can help in interpreting seasonal variations in the observed area, for instance the vegetation. The test of this reconstruction method on the illustrated parameter k time series, as well as the tests made on other RPV parameters for different spectral bands, showed that this Lomb-Scargle-based reconstruction method worked well in representing the pattern and phase of the seasonal component.

4.3 Seasonality exploration of the MISR-HR anisotropy data

As seen from the time series plots of parameters rho, k and Theta in Figures 2.6, 2.9 and 2.12, respectively, seasonal variations are obvious even though there are many missing values in each series. This section first reports on the tests for seasonality in the MISR-HR anisotropy data and then on the extracted and analysed seasonal components in the RPV time series.

4.3.1 Results of seasonality detection

To investigate regular variations such as seasonality in data, the Lomb-Scargle periodogram method was employed in the current research because it has the merit of detecting seasonality from unevenly spaced time series data, as discussed in Section 4.2.1. This Lomb-Scargle periodogram method was applied to all the pixels in the four spectral bands for all three study sites. Some degree of seasonality was detected at each of the three sites, at the significance level of 0.05; the results are summarised in Table 6.4.

As can be seen from Table 4.4, nearly all pixels for the semi-desert area show seasonal variations in the k and Theta time series, for all spectral bands; all pixels for this area exhibit seasonal variations in the parameter rho time series, but only for the red and NIR spectral bands. For the wheat field, all pixels have seasonality in the parameter rho data, for the red, green and blue bands; nearly all pixels in the wheat field show seasonal variations in the parameter k time series, for all spectral bands. For the vineyard area, the majority of pixels in both parameter rho and k data exhibit seasonal variations in the red and NIR bands.

Table 4.4: Proportion of pixels showing seasonality in the MISR-HR anisotropy data in four spectral bands, for the three study areas.

Study sites	**RPV parameters**	**Blue %**	**Green %**	**Red %**	**NIR %**
Semi-desert area*	Rho	0.00	54.54	100	100
	K	100	100	100	100
	Theta	99.17	100	100	100
Wheat field*	Rho	100	100	100	14.88
	K	100	100	99.17	100
	Theta	74.38	100	65.29	94.21
Vineyard area**	Rho	0.93	27.10	92.52	87.94
	K	59.26	71.30	98.15	99.07
	Theta	0.93	25.93	68.52	62.04

*This study site covered 121 pixels in total

** This study site covered 108 pixels in total (see Section 2.5)

4.3.2 Analysis of seasonal variations

This section analyses the seasonal component of the MISR-HR anisotropy data. This analysis can help in revealing regular variations in the anisotropy data. The method used to rebuild the seasonal variations was introduced in Section 4.2.2. The seasonal component reconstruction can give details about the pattern and the phase of the seasonal signal.

4.3.2.1 Parameter rho

Figure 4.8 displays the reconstructed seasonal component of the rho time series in the red spectral band for the selected pixels in the three study areas. The selected pixels are the same pixels identified in Chapter 2, namely s05_+000_+001 for the semi-desert area, s10_+000_+000 for the wheat field and s12_+000_-003 for the vineyard area. The detected period of parameter rho is one calendar year for all three pixels.

It can be seen from the plot that parameter rho of the wheat field peaks in summer (around January every year), while at the same time the vineyard area reaches its lowest point. This indicates the seasonal variation of the wheat field has an opposite phase to that of the vineyard area. These variations are consistent with the vegetation signatures of these two areas: wheat grows in winter with the green leaves absorbing red spectral band energy, while in summer the leaves on the mature wheat turn yellow. The growing season of grapes is just the opposite of wheat, with green leaves in summer which fall off in winter. As we know, green leaves absorb red spectral band energy so in summer in this area the reflectance of this band will be lower than in the area without green leaves. Thus, these vegetation signatures are represented by the seasonal variations of the parameter rho values.

The differences of the phase in the three areas can be calculated by retrieving the observation dates of the peak and lowest values. The first peak values for the semi-desert and wheat field were observed on 16 January, 2001 and 17 February, 2001, respectively (one month's difference); the lowest values for these two areas were observed on 24 July, 2000 and 25 August, 2000, respectively. The first peak and lowest values for the vineyard area were obtained on August 25, 2000 and February 17, 2001, which means the vineyard rho parameter variation is exactly the opposite phase as the wheat field.

These accurate phase determinations would be impossible without the use of the multiyear time series reconstruction.

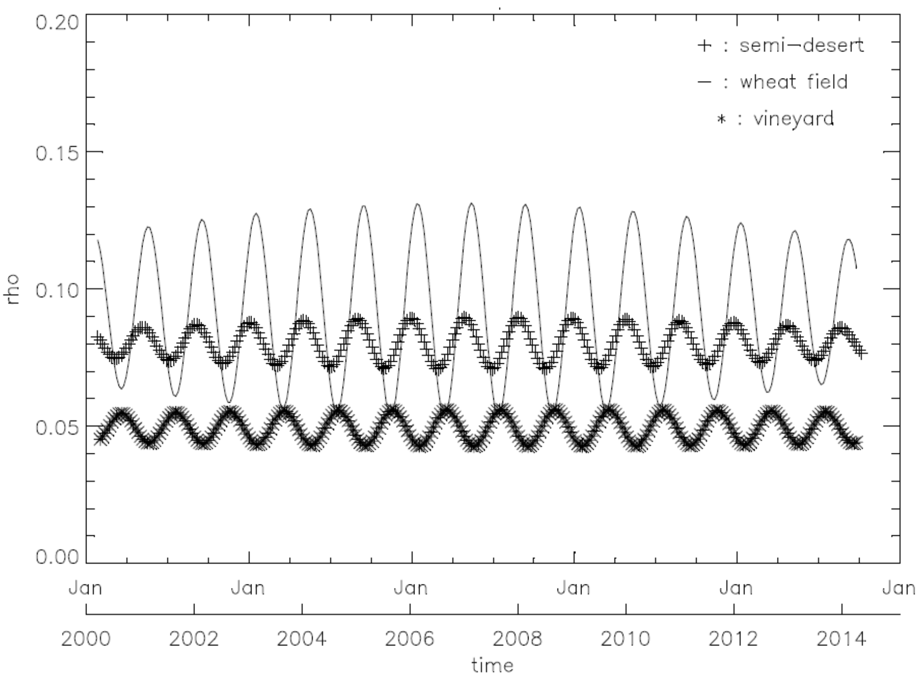

Figure 4.8: Reconstructed seasonal components of the red band parameter rho time series, for the semi-desert area*, the wheat field and the vineyard area***, respectively.**

* Location specified by Path 174, Block 117, Line 93 and Sample 798
** Location specified by Path 174, Block 117, Line 358 and Sample 431
*** Location specified by Path 174, Block 117, Line 381 and Sample 773

4.3.2.2 Parameter k

The period of parameter k is calculated as one calendar year—the same as for parameter rho—for all three study sites. The time series plots of the seasonal components of red band parameter k are shown in Figure 4.9 for all three study sites. The first lowest values were observed on 21 May, 22 June and 25 August, 2000 for the vineyard, semi-desert area and the wheat field, respectively. The first peak values were obtained approximately six months after the lowest values for these three areas. The retrieved lowest observation dates reveal a three-month shift between the vineyard and the wheat field to reach their minimum value in each period. The k values themselves are very close for these two sites, such that it is not easy to discriminate the vineyard from the wheat field by that statistics. However, the phase difference between these time series offers an opportunity to discriminate between the different land covers. Classification can be applied by using the different phases of the seasonal components as the classification signatures.

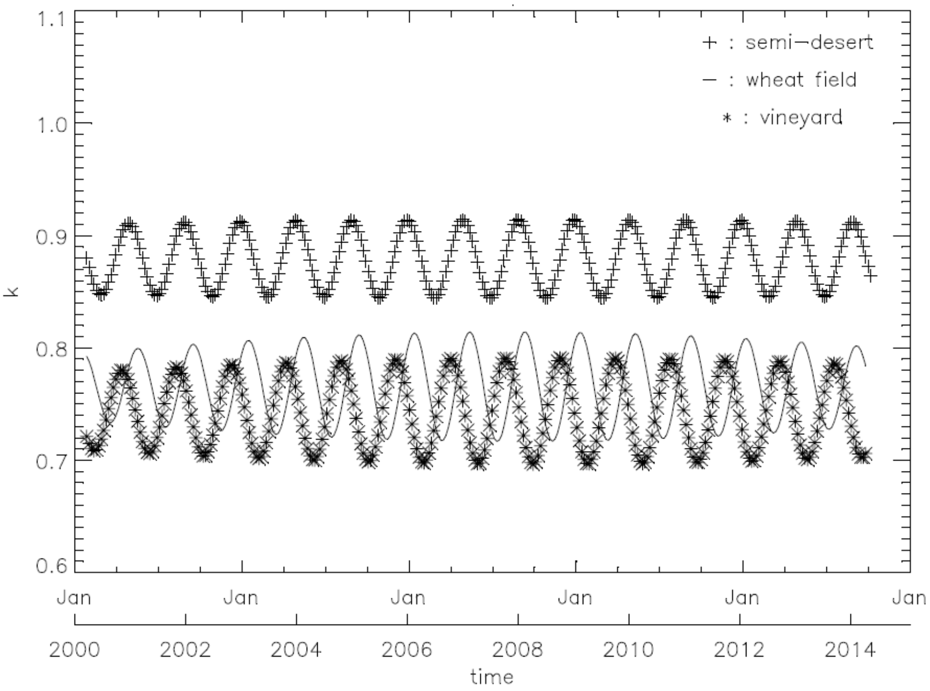

Figure 4.9: Reconstructed seasonal components of the red band parameter k time series, for the three study areas. Pixels are as specified in Figure 4.8.

Since parameter k shows higher mean values in the NIR band for the vineyard area than the other two sites, it was interesting to explore seasonal variations of the NIR band k values for all three (see Figure 4.10). It can be seen from the plot that the illustrated data sets display seasonal variations; the shape of each seasonal cycle for the three study sites is different; NIR band k data go down from the start of the observation date for all the areas; the grape cultivation data achieved the first lowest point on 5 June, 2000, and after a little bump the illustrated data reached the minimum value in the first cycle on 10 September, 2000; the vineyard area data reached the first peak value on 29 November, 2000 and the second peak value on 5 March, 2001; similar to the vineyard area, the wheat field data reached their first lowest value on 8 July, 2000, and after a tiny bump the illustrated data of this area achieved minimum value on 13 October, 2000; however, the wheat field data only have one peak value in each cycle, with the peak value being obtained on 17 February, 2001; the illustrated semi-desert area data only achieved one lowest point and one peak value in a seasonal cycle, where the lowest and peak values for this area were obtained on 8 July, 2000 and 31 December, 2000, respectively.

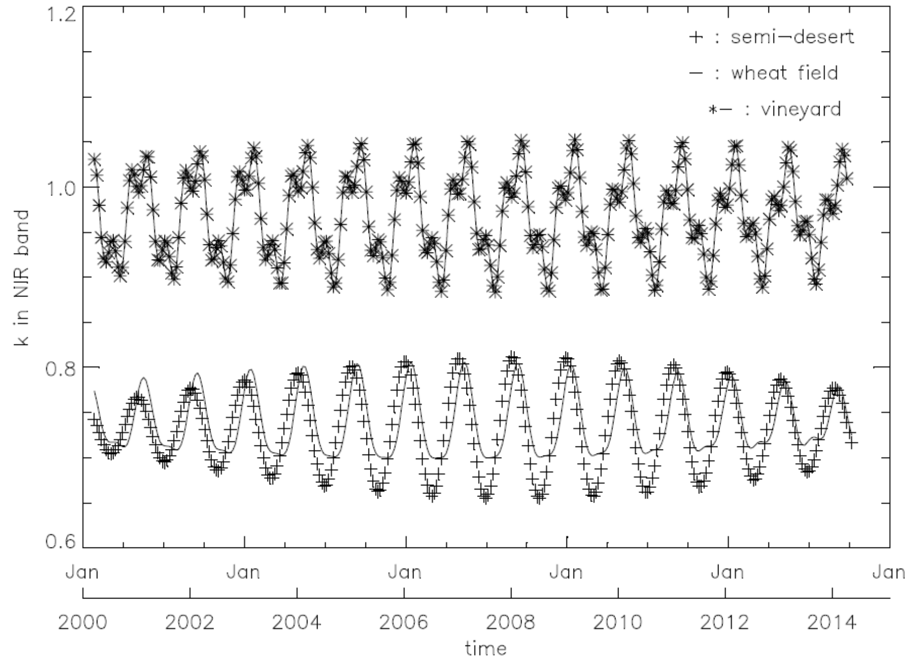

Figure 4.10: Reconstructed seasonal components for the RPV parameter k in the NIR spectral band, for the three study areas. Pixel locations are as specified in Figure 4.8.

For the illustrated vineyard area data, the time between the first lowest value (6 June, 2000) and the first peak value (29 November, 2000) was about 6 months. The period between the second lowest value (10 September, 2000) and the second peak value (5 March, 2001) was also approximately six months. The Lomb-Scargle periodogram plot of this selected time series shows that there are two peaks with significance level over 0.05, as demonstrated in Figure 4.11. However, this seasonal variation shape of the NIR band k data is not a common characteristic for this vineyard area. When the NIR band k data of other pixels in this area were examined, no similar seasonal variation shapes were found. The examined pixels were s12_-003_+005, s12_-002_-002, s12_+000_+000, s12_+000_-001 and s12_+005_-005. The selected pixel is not representative of the common seasonal variation features of the NIR band parameter k data for the entire vineyard area. There are however pixels that show similar variations in the set. This is clear from comparing two groups of pixels as shown in Figures 4.12 and 4.13. All in all, it does still show high mean values of NIR band k data.

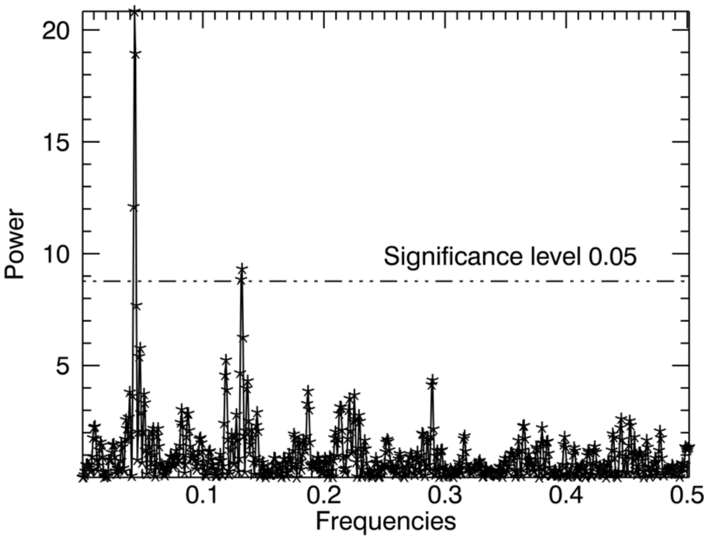

Figure 4.11: Periodogram of the illustrated RPV k parameter in the NIR spectral band, for the vineyard area. Pixel location is as specified in Figure 4.8.

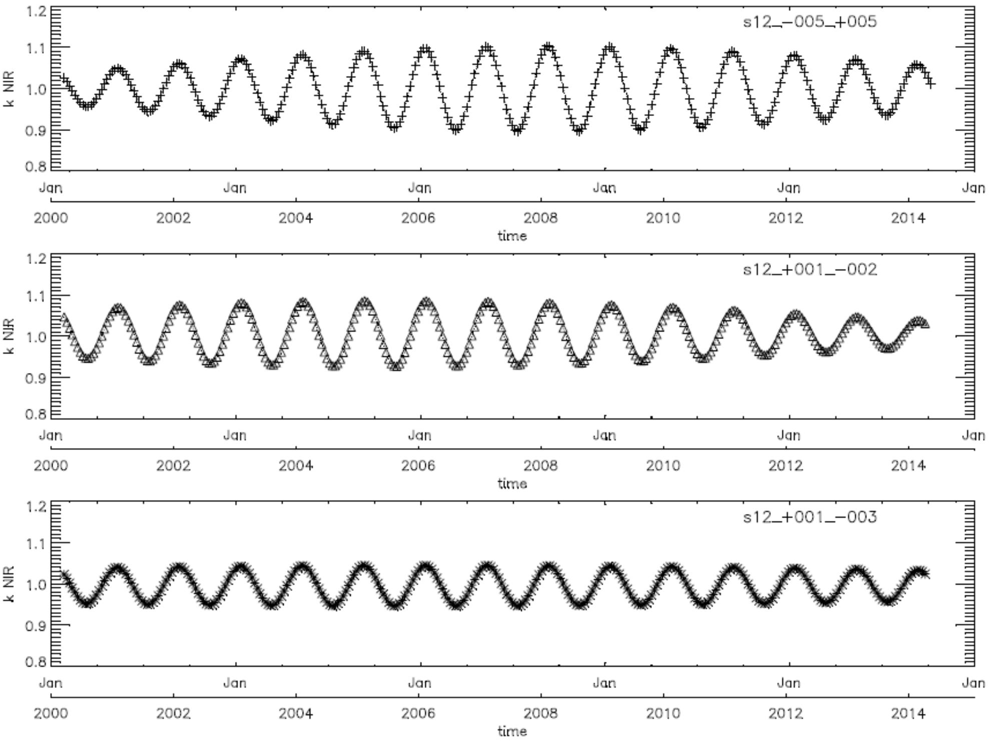

Figure 4.12: Reconstructed seasonal components for the RPV parameter k in the NIR spectral band, for pixels s12_-005_+005, s12_+001_-002 and s12_+001_-003 in the vineyard area. These pixels show different variations to pixel s12_+000_-003 demonstrated in Figure 4.10.

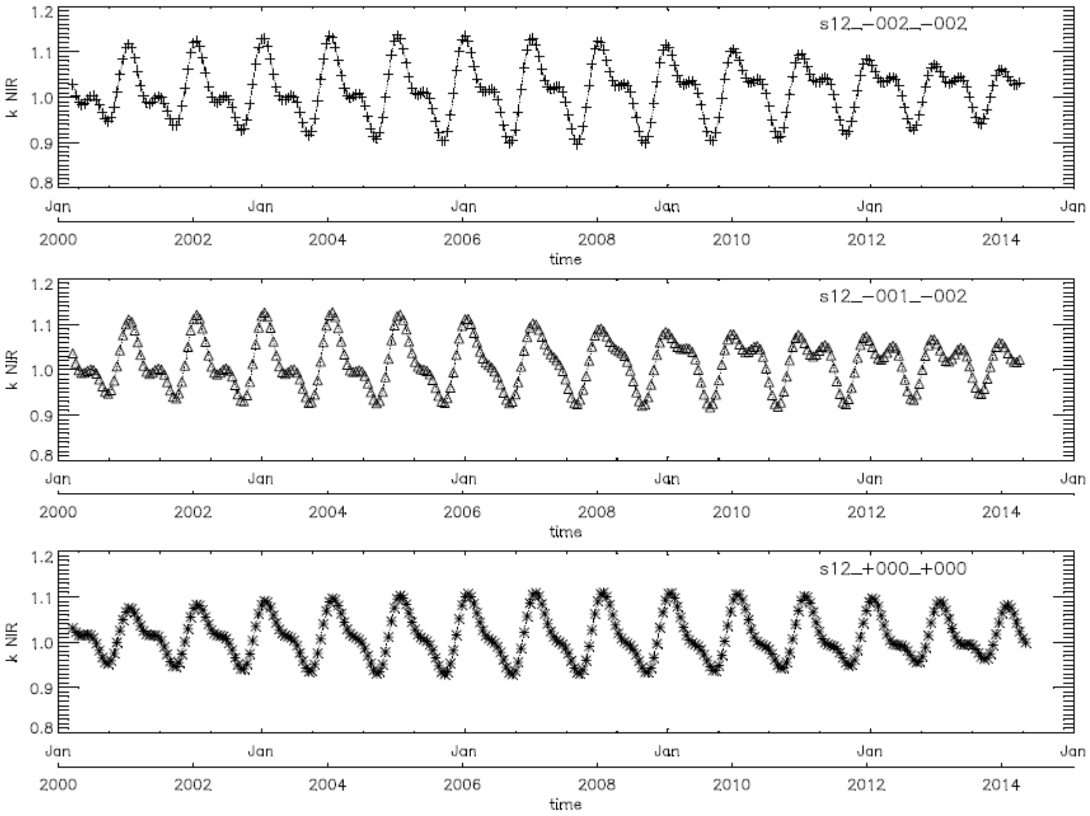

Figure 4.13: Reconstructed seasonal components for the RPV parameter k in the NIR spectral band, for pixels s12_-002_-002, s12_-001_-002 and s12_+000_+000 in the vineyard area. These pixels show similar variations as pixel s12_+000_-003 demonstrated in Figure 4.10.

4.3.2.3 Parameter Theta

The seasonal variations in parameter Theta are quite complex for the three study sites. The seasonal components of the illustrated red band Theta data are shown in Figure 4.14 for all three study areas. The periods of the Theta data for the semi-desert area and the wheat field are one calendar year and six months, respectively, retrieved by the Lomb-Scargle periodogram method with a significance level of 0.05 (Section 4.2.1). For the illustrated vineyard area, Theta data exhibit two peaks in the periodogram with significance level higher than 0.05, as seen in Figure 4.15. These two peaks show that there are two dominant seasonal variations in the Theta time series and explains why there is structure in the seasonal component plot for the vineyard area (Figure 4.14).

As seen from the seasonality summaries for the wheat field and vineyard area in Table 4.4, not all pixels for these two areas show seasonal variations in the Theta data. For pixels with seasonality, the periods are the same for the same study area in the same spectral band; for instance, some pixels in the wheat field were detected with a cycle of six months in the red spectral band, while a few pixels were detected with a cycle of one year for the same study site in the same spectral band. The diverse seasonality of the Theta data for the same study area makes it difficult to summarise a common feature for this area, as illustrated by the results from wheat field.

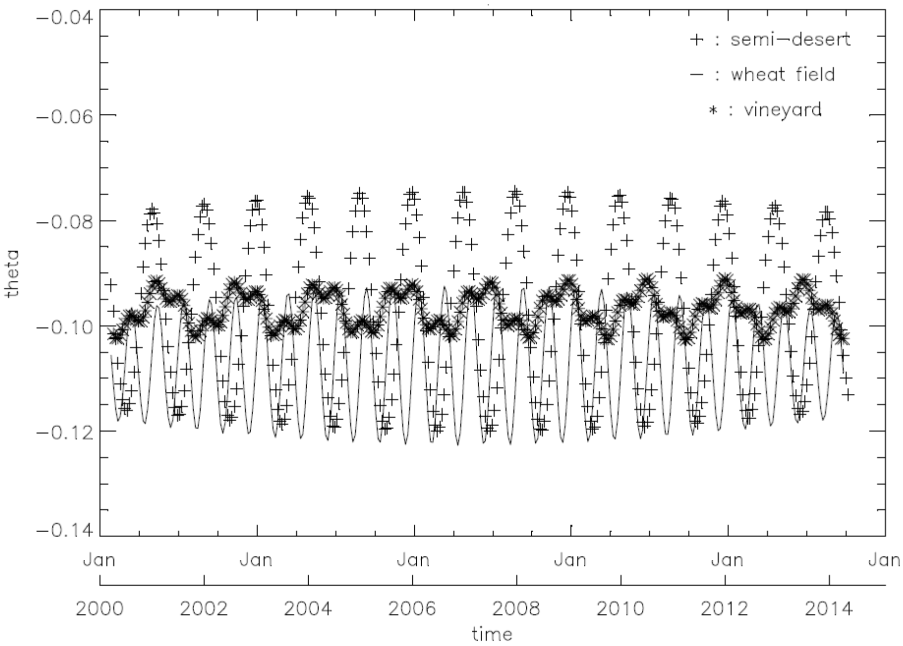

Figure 4.14: Seasonal variations of RPV parameter Theta in the red spectral band for the three study areas. Pixel locations are as specified in Figure 4.8.

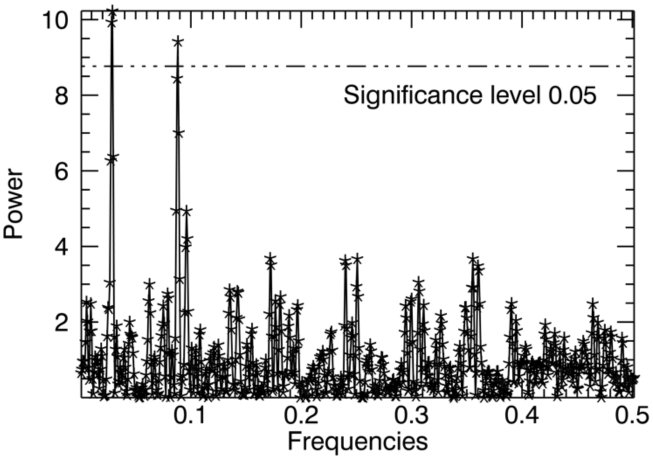

Figure 4.15: Periodogram of the illustrated red band Theta data, for the vineyard area. Pixel is as specified in Figure 4.8.

4.4 Discussion of the trend detection and seasonal variation analyses

This research undertook a preliminary exploration of the trends in the MISR-HR anisotropy data over a period of 14+ years, with the aim of revealing long-term change in the semi-desert area, the wheat field and the vineyard area, represented by time series in the RPV parameters. The downward trend in the green band parameter rho data for this area may reveal the degradation of the environment or the change in the MISR calibration system, which needs to be confirmed by further investigation. Changes in the wheat field and the vineyard area are influenced by the natural environment, MISR instrument and human activities. The trend detection results also indicated the upward trend of the blue band parameter k data for the entire semi-desert area, and more than half of the pixels covering the wheat field and the vineyard area. The interpretation of the blue band parameter k data is difficult, since no research about the application of this spectral band parameter k data was found in literature. Nor was any literature found on blue band parameter Theta data, which showed an upward trend in nearly all pixels in the three areas. Even though the current study was not able to interpret the trends in the MISR-HR anisotropy data, it did detect and represent the monotonic trends in that data.

The seasonality detection results show that all three RPV parameters have seasonal variations, which may differ between different spectral bands and study areas. Pinty et al. (2002) suggested that parameter k may change regularly with the observation time; the seasonality detection results in the current research confirmed this hypothesis and, furthermore, showed the seasonal variations in the parameters rho and Theta data also occur. The reasons that caused seasonal variation in the MISR-HR anisotropy data are complex. In theory, the RPV model parameters are not dependent on the position of the Sun under perfect conditions. But due to the limited sampling measurement (the MISR instrument only collects nine angular measurements in each spectral band), the retrieved RPV products may not able to describe thoroughly the actual anisotropy of the observed area. This may result in the RPV products still being influenced by the solar zenith angle. For instance, the author speculates that seasonality in the parameter k time series for the semi-desert area is caused by the solar zenith angle, because this area is supposed to be constant over the observation period. Seasonality of parameter rho was probably influenced by vegetation conditions on the ground, which are represented by the different seasonal phases between different vegetation types. Varying seasonal phases for different types of vegetation are potentially useful for discriminating different land surfaces. Armston et al. (2007) presented parameter rho values for a cypress pine forest, eucalypt open woodland and a grassland in their research; only the grassland showed a slight increase (1-2%) in the red and green bands of parameter rho, while the other two areas remained constant over the one-year-long observation period. These results are not in conflict with the seasonality found in the current research, since both the cypress pine and the eucalypt are evergreen vegetation with more constant reflection than seasonally-sensitive cultivation areas, such as the cultivated wheat and vineyard areas.

4.5 Summary

This research detected monotonic trends and seasonality in the MISR-HR anisotropy data over a period of 14+ years, for a semi-desert area, a wheat field and a vineyard area. The trend analysis on the MISR-HR anisotropy data could indicate long-term change of the observed ground surface, calibration drift or other factors. This is an area for further research.

Three interesting trend detection results were found. There was a downward trend of the green band rho data for the semi-desert area and the vineyard area. An upward trend of the blue band k data for the entire semi-desert area and more than half of the pixels covering the wheat field and vineyard area was detected. An upward trend of the blue band Theta data for nearly all the pixels in the three study sites was also detected.

The seasonality analysis shows that, while all the RPV parameters have seasonal variations, there is diversity between different spectral bands and land covers. The seasonal information presented by the RPV parameters corresponded to the signatures of natural phenomena (e.g., cultivation of different plan types) in the observed areas. This investigation of the seasonality confirmed that parameter k varies regularly along the observation time; and also revealed seasonal variations in the parameter rho and Theta data. Seasonal signatures represented by the RPV time series are potentially useful for classifying different land covers.

CHAPTER 5

USING MISR-HR PRODUCTS TO DISCRIMINATE LAND COVERS

Chapters 2 and 4 have highlighted to fact that the MISR-HR RPV products for the three selected sites (semi-desert, wheat field and vineyard) of South Africa's Western Cape Province exhibited significantly different signatures in the temporal, spectral and directional domains. This, in turn, suggests that these properties could be used to discriminate between such land cover types, as was speculated in the initial objectives of this thesis.

Section 5.1 reviews those outcomes and associates specific signatures to each of the three sites. Section 5.2 outlines the clustering algorithm that was selected for this study (the *k*NN classifier), while sections 5.3 and 5.4 explore the effectiveness of the previously identified spectral and directional signatures in these three cases. Section 5.5 discusses these results, and Section 5.6 investigates the combined use of multiple criteria to improve the classification performance. Lastly, Section 5.7 summarizes the findings of this Chapter.

5.1 Distinctive signatures of the MISR-HR RPV data

The MISR-HR RPV rho parameter in the red spectral band proved to be highly useful in discriminating between the three sites. This was largely expected because of the strong absorption of the chlorophyll molecules in that spectral band. In fact, most traditional classification algorithms rely on one or more spectral bands (Lu, Mausel, Batistella & Moran, 2004a), and in particular the red band, to characterize land covers, but they typically rely on the reflectance measurements themselves, which may have been corrected for atmospheric effects but rarely account for the anisotropy of the illumination field or of the surface itself. The main advantage and the relatively innovative use of the RPV rho parameter is that this product results from the inversion of a bidirectional reflectance model against multi-angular data and therefore is essentially decontaminated from such directional effects.

A simple exploratory investigation of the properties of the MISR-HR RPV k parameter in the NIR spectral band (Section 2.6.2) unveiled that this parameter exhibited very different values over the vineyard compared to the other two sites (semi-desert and the wheat field), which were otherwise quite similar. This result is particularly interesting because it contrasts with the finding of Armston et al. (2007, p. 295) who found that "the mean k parameter for the NIR band does not show any values greater than 1.0 due to the lack of spectral contrast between the canopy and the soil background" in the Southern Brigalow Belt (SBB) Biogeographic

Region of Australia (that region covers 367,404 km2, though their investigation only used 20 acquisitions, between 2002 and 2004). On the other hand, the RPV parameter Theta did not appear to be as effective in these cases than in the Australian study. This difference may be worth investigating in the future.

Lastly, Section 4.3.2 showed that the three sites also exhibited specific and different temporal signatures, with phase differences likely due to the specific phenological properties of these different environment. Hence, it can be surmised that the spectral analysis of these time series, and in particular the timing of the seasonal onset of the growing seasons, could also be exploited to differentiate between these land covers.

The second objective of this thesis will therefore focus on exploring the classification potential of the MISR-HR RPV parameters rho and k in the spectral and temporal domains.

5.2 The k-nearest neighbour classifier (kNN)

The literature on clustering and classification is replete with methods and algorithms, including the maximum likelihood classifier, support vector machine (SVM) classifier, k-means or k-nearest neighbour classifier (Lu et al., 2004a; Marcal, Borges, Gomes & Pinto da Costa, 2005; Szuster, Chen & Borger, 2011; Jia et al., 2014; Taati et al., 2014). The k-nearest neighbour classifier, in particular, is a supervised distance-based, non-parametric classifier, which is simple to understand and works very well in practice (Franco-Lopez, Ek & Bauer, 2001; McRoberts, Nelson & Wendt, 2002). It was adopted in this work, first for the purpose of exploring whether the three selected sites could be differentiated on the basis of a single MISR-HR RPV parameter.

The kNN classifier computes the distance between an arbitrary test case and a set of training instances (with known class labels). It assigns a class label to that case on the basis of its proximity to its k-nearest neighbours. Various rules can be used to estimate these distances, and selecting the class label associated with the majority of k-nearest neighbours is commonly used. Euclidian metric, one of the simplest measures, is frequently used in many applications. The reliability of the kNN classifier has been thoroughly discussed in literature (Cooper & Hart, 1967; Wang, Neskovic & Cooper, 2003; Wang, Neskovic & Cooper, 2006; Tsypin & Roder, 2007; Walpole et al., 2012). Tsypin and Roder (2007) proposed a method of measuring the confidence level for the kNN classifier that depends only on the training data set. Their formula for calculating the level of confidence assumes the existence of only two classes. When the numbers of training instances within each class are the same, the confidence for a class, for instance class 1, is calculated by:

$$P(class\ 1) = \frac{k1+1}{k1+k2+2} \qquad (5.1)$$

$$\frac{P(class\ 1)}{P(class\ 2)} = \frac{k1+1}{k2+1} \qquad (5.2)$$

$$k = k1 + k2 \qquad (5.3)$$

where $k1$ is the number of nearest neighbours belonging to class 1 and $k2$ is the number of nearest neighbours belonging to class 2. Hence, when the number k of nearest neighbours to be checked in order to classify a new item is 5, if all of those belong to the same class (say 1, for the sake of the argument), then $k1$ = 5, $k2$ = 5 - $k1$ = 0. The confidence level associated with assigning that point to class 1 is thus 6/7 or 85.71%, or, equivalently, assigning that point to class 2 would result in a confidence value of only 1/7 or 14.29%. Similarly, if one of those 5 neighbours belonged to class 2, the confidence level for assigning that test point to class 1 would drop to (4 + 1) / (4 + 1 + 2) = 5/7 or 71.43%. It is seen that this confidence level indicator only takes on a limited number of discrete values. Its maximum value $k/(k$ + 1) tends to 1.0 (or 100%) when the number k of smallest distances to training points increases and all of them happen to belong to the same class.

This investigation adopted the kNN classifier and the Euclidian distance estimator, combined with the majority voting rule to assign a class label to each point, and then used Equation 5.1 to calculate the confidence level of the classification result.

In order to use this approach, it is imperative that each item to be classified exhibits the same number of properties. In this particular case, this implies that all time series of MISR-HR RPV parameter values must have valid values on the same dates as the training pixel's time series, as the Euclidian distance estimator would return meaningless values if the length of the time series were different, or if the observation dates of the two series did not match.

Since the RPV time series discussed in Chapter 3 and 4 could contain missing values (due to cloudiness or detected outliers), such values should be replaced by reasonable estimates before applying the kNN classifier. The time series reconstruction method based on the Lomb-Scargle periodogram and described in Section 4.2.2 was used in those cases to ensure that all time series were of equal length and provided data on the same dates.

The kNN algorithm requires a set of training data to determine a priori the properties of typical classes. Too few or poorly chosen training points may lead to an unsatisfactory classification, while too many training points would significantly increase the computational cost. Ten training points from each of the three sites were selected for the purpose of this work. In the case of the semi-desert site, the training points were chosen arbitrarily because the absence of land use or human interference with the landscape implied that they could all be equally relevant. In the case of the wheat field and the vineyard, the training points were specifically picked by inspection of high spatial resolution images available on Google Earth to avoid including mixed or different targets such as houses or obviously different land covers. Table 5.1 lists the pixels selected for the training of the kNN algorithm at each site. The status of the remaining pixels (111 for the semi-desert and wheat field, and 98 for the vineyard) was the assessed with the classification algorithm.

Table 5.1 Training pixels for the three study areas.

	Vineyard	Wheat field	Semi-desert
1	s12_-005_+004	s10_-005_+000	S05_-004_-001
2	s12_-005_+005	s10_-005_+001	S05_-003_+003
3	s12_-003_-002	s10_-005_+002	S05_-002_-004
4	s12_-003_+000	s10_-004_+000	S05_-001_+005
5	s12_-003_+001	s10_-004_+001	S05_+000_+001
6	s12_-003_+005	s10_-004_+002	S05_+001_-003
7	s12_+000_-003	s10_-003_+000	S05_+002_+003
8	s12_+000_+000	s10_-003_+001	S05_+003_-002
9	s12_+001_-004	s10_-003_+002	S05_+004_-005
10	s12_+001_-002	s10_-002_+000	S05_+005_+001

5.3 Classification by spectral signature

Given the historical importance of spectral information in classifying land covers, the first application of the kNN algorithm investigated the performance of the RPV parameter rho in the red spectral band to discriminate between the three sites on the basis of 14+ years of data.

5.3.1 Distinguishing the vineyard from the wheat field

The time series of all non-training pixels for the RPV parameter rho in the red spectral band were first updated using the method described earlier to replace missing values by reasonable estimates and therefore generate sequences with the same number of data points and for the same dates. Training pixels for the vineyard were assigned to class 1, while those for the wheat field were labelled class 2. The number of the k-nearest neighbours was set to 5. Applying the kNN classifier on all (209) remaining query pixels for these two areas showed that all query pixels for the vineyard area were assigned to class 1 (grape cultivation) and all query pixels for the wheat field were attributed to class 2 (wheat field) with a confidence level of 85.71%. Non-grape cultivation pixels in the vineyard area were not separated out by red band parameter rho data in this classification test.

5.3.2 Distinguishing the vineyard area from the semi-desert area

A similar procedure was applied to explore whether the vineyard could also be discriminated from the semi-desert on the basis of the RPV parameter rho in the red spectral band. Setting again the vineyard training pixels to class 1 and the semi-desert area to class 2, and then apply the kNN classifier on all query data for these two areas, the results showed that all query pixels for the semi-desert area were assigned to class 2, with all but three pixels having a confidence level of 85.71%; pixels s12_+005_-005 and s12_+005_-003 had a confidence level of 71.43%, and pixel s12_+005_-004 a confidence level of 57.14%. For the query pixels of the vineyard area, 83 pixels were attributed to class 1 (68 pixels with confidence of 85.71% and 15 pixels with confidence levels lower than 85.71%) and 15 pixels were assigned to class 2 (9 of these 15 pixels with a confidence level lower than 85.71%). These 15 pixels (listed in Table 5.2), which the classifier allocated to the semi-desert class from the vineyard area, are located in the non-grape cultivation area, or at the boundary between grape cultivation and non-grape cultivation areas.

Table 5.2: Non-grape cultivation pixels detected when discriminating the vineyard area from the semi-desert area, with the kNN classifier combined with red band rho data.

	Pixel name		Pixel name
1	s12_-005_-003	9	s12_+004_+003
2	s12_-005_-001	10	s12_+005_-002
3	s12_-004_-004	11	s12_+005_-001
4	s12_+003_+003	12	s12_+005_+000
5	s12_+004_-001	13	s12_+005_+001
6	s12_+004_+000	14	s12_+005_+002
7	s12_+004_+001	15	s12_+005_+005
8	s12_+004_+002		

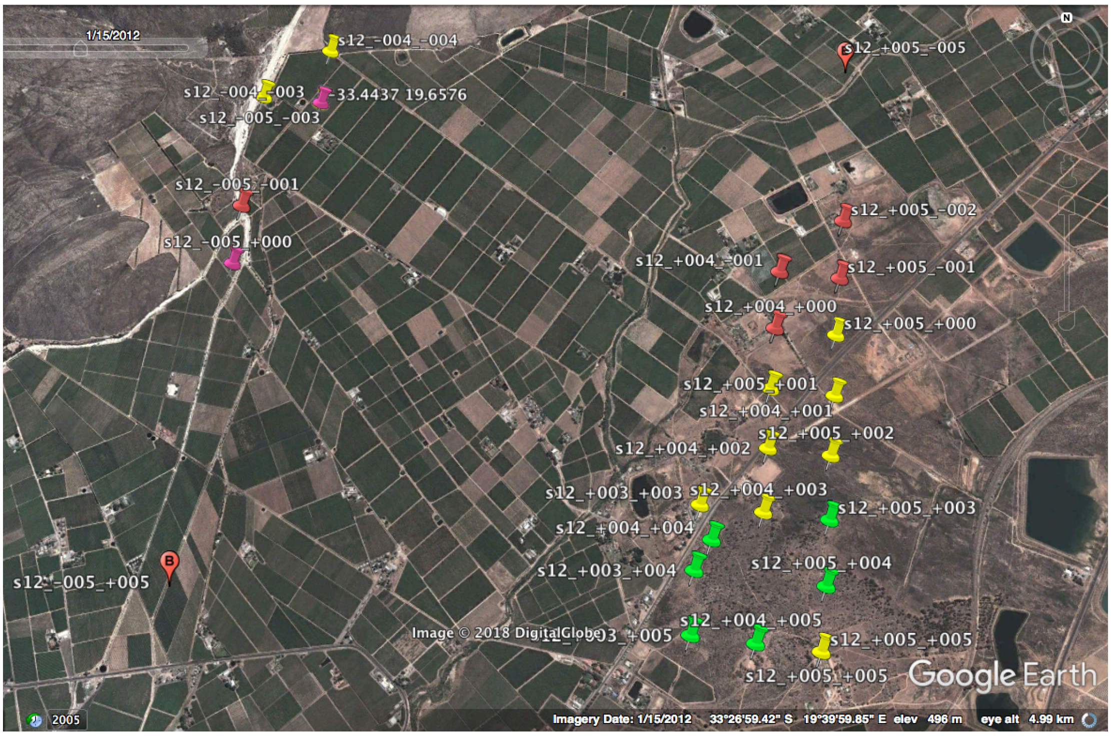

Figure 5.1: Satellite map of the vineyard area. Red balloons labelled 'B' indicate southwest and northeast boundary pixels. Yellow, red and green pins indicate pixels that were declared non-grape cultivation by both the RPV rho parameter in the red spectral band and the RPV k parameter in the NIR spectral band, by the rho red only or by the k NIR only, respectively. The pink pins identify additional non-grape cultivation pixels identified using multiple RPV parameters.

Inspection of Figure 5.1, which locates those 15 pixels on a Google Earth background map as yellow and red pins (the distinction between those will become clearer below), shows that those pixels are indeed not cultivated but rather bare ground, or perturbed by a bright surface such as a road (e.g. pixel s12_-005_-001). Hence, the classifier successfully assigned the pixels to the correct class labels and was able to separated non-grape cultivation pixels from the vineyard area.

5.3.3 Distinguishing the wheat field from the semi-desert area

The kNN classification procedure was then applied to test the feasibility of separating the wheat field (training pixels assigned to class 1) from the semi-desert (training pixels assigned to class 2). The classification results showed that all query pixels were attributed to the correct class with a confidence level of 85.71%. The kNN classifier can therefore be applied to the RPV rho parameter in the red spectral band to successfully discriminate the wheat field and the semi-desert.

5.4 Classification by directional signature

The results described in the previous section confirm that different land cover types can be distinguished on the basis of their spectral signature, in this case the RPV rho parameter in the red spectral band, as expected given earlier findings published in the refereed literature.

In this section, the same approach was applied to analyse the directional signatures of those three sites, this time characterized by the RPV k parameter in the NIR spectral band.

5.4.1 Distinguishing the vineyard from the wheat field

As for the previous case (Section 5.3), the time series of the RPV parameter k in the NIR spectral band for all pixels of the three sites were completed and made comparable by ensuring that they were of the same length and provided valid values for all applicable dates. The vineyard site was set to class 1, the wheat field to class 2, and the number of nearest neighbour to 5. The kNN classifier reported that all query pixels from the wheat field were assigned to class 2 with a confidence level of 85.71%; while 13 pixels of the vineyard area were assigned to class 2 and the remaining pixels were assigned to class 1. Two query pixels in the vineyard site (s12_+003_+003 and s12_+003_+005) were assigned a lower classification confidence level of 71.43% and 57.14%, respectively, while all other pixels of this area had a classification confidence level of 85.71%. The 13 pixels located in the vineyard area but assigned to the wheat field class by the kNN classifier are listed in Table 5.3, and marked on the satellite map in Figure 5.1 with yellow and green pins: yellow pins characterize those pixels which were correctly detected as non-vine cultivation by both the spectral and directional classification method, and green pins represent those pixels detected by the directional classification method only).

Table 5.3: Non-grape cultivation pixels detected when discriminating the vineyard area from the wheat field, with the kNN classifier combined with NIR band k data.

	Pixel name		Pixel name
1	s12_-004_-004	8	s12_+005_+000
2	s12_+003_+004	9	s12_+005_+001
3	s12_+003_+005	10	s12_+005_+002
4	s12_+004_+002	11	s12_+005_+003
5	s12_+004_+003	12	s12_+005_+004
6	s12_+004_+004	13	s12_+005_+005
7	s12_+004_+005		

As can be seen on that satellite map, pixels s12_+003_+003 and s12_+003_+005 are both located on the boundary between the vineyard fields (to the west) and uncultivated areas (to the east): this can explain their lower classification reliability. It should be remembered (1) that the geolocation accuracy of MISR data is of the same order of magnitude as the size of the high-resolution pixels, hence the actual observed area on the ground for those pixels can differ slightly from the positions indicated on the map of Figure 5.1. (see the NASA JPL web page available at https://misr.jpl.nasa.gov/Mission/geocal/migeocal.html), and (2) that the kNN algorithm was given only two classes to choose from: vineyard and wheat field. Hence anything that did not resemble a vineyard had to be classified as a wheat field, and the algorithm correctly indicated that the confidence level in this case was not very high (57.14%).

5.4.2 Distinguishing the vineyard from the semi-desert area

The kNN classification procedure was carried out once again, this time to test the separability of the vineyard (class 1) from the semi-desert (class 2), when using 5 nearest neighbours. It turned out that all the query pixels from the semi-desert area were assigned to class 2, hence with a confidence level of 85.71%; 82 pixels of the vineyard area were attributed to class 1, also with a confidence level of 85.71%, and 16 pixels of this site were allocated by the classifier to class 2. Of these 16 pixels, two (s12_-005_-003 and s12_+004_+001) were assigned a confidence level of 57.14%; another two (s12_+003_+003 and s12_+003_+005) were given a confidence level of 71.42%; and the remaining 12 pixels had a confidence level of 85.71%. These 16 pixels include all 13 non-grape cultivation pixels detected in Section 5.4.1 and three additional pixels, namely s12_-005_-003, s12_+003_+003 and s12_+004_+001. The latter are located either in the non-grape cultivation area (e.g., s12_+003_+003, s12_+004_+001) or on the boundary between the non-grape cultivation and the grape cultivation area (e.g., s12_-005_-003) (see Figure 5.1).

5.4.3 Distinguishing the wheat field from the semi-desert area

This section aimed to test whether the wheat field and the semi-desert area were separable by the kNN classifier on the basis of the RPV k parameter in the NIR spectral band. The number k of the nearest neighbours to consider was set to 5, the wheat field was assigned to class 1 and the semi-desert area to class 2. In this case, the classification results showed that all the query pixels were assigned to the correct class with a confidence level of 85.71%. This outcome suggests that the wheat field and the semi-desert area were completely separable at that level of reliability by the kNN classifier using the RPV k parameter in the NIR spectral band.

5.5 Discussion of the binary classification

As seen in Subsections 5.3.1 and 5.4.1 above, the wheat field could be separated from the vineyard area either on the basis of the RPV parameter rho in the red spectral band or using the RPV parameter k in the NIR spectral band; the main difference was that the latter approach was more sensitive than former in separating out non-grape cultivation pixels.

The semi-desert area was also separable from the vineyard area (Subsections 5.3.2 and 5.4.2) on the basis of either of these two RPV parameters. And in both binary confrontations, the kNN algorithm was able to detect the presence of pixels that could not be reliably assigned to the vineyard category. However, while the two approaches agreed in the case of 8 pixels, 4 were classified as non-grape by the RPV parameter rho in the red spectral band only, while 6 were identified as different from the vineyard by the RPV parameter k in the NIR. In other words, neither of the utilised RPV parameters was able to distinguish all the non-grape cultivation pixels in the vineyard area in this classification test situation.

Discriminating between the wheat field and the semi-desert area (Subsections 5.3.3 and 5.4.3) turned out to be relatively simple and reliable, using either RPV parameter.

Classification outcomes were also inspected when derived with a smaller number (3) of nearest neighbours. Since the confidence level is affected by the k parameter of the kNN classifier (see Section 5.2), the results generated by the 3NN classifier were associated with relatively lower confidence levels than obtained with the 5NN classifier. Nevertheless, the number of non-grape cultivation pixels detected by the 3NN classifier was slightly different from the 5NN classifier, but all other results were identical for both the 3NN and 5NN classifiers.

Classification results might change slightly if different training instances were selected as described in Section 5.2 within the same three sites. However, if that alternate training set selection involved different targets, for instance pixels containing settlements or forest patches, the results could obviously be substantially different, both in terms of the number of test pixels assigned to each class and in terms of the confidence associated with those results.

These three land cover types therefore exhibit sufficiently different time series of spectral or directional properties to be correctly identified with high confidence in binary tests involving two training sets and a single RPV model parameter, such as rho in the red or k in the NIR spectral band. On the vineyard site, the kNN algorithm isolated non-grape cultivation pixels, with slightly different results depending on which training set was available. The next Section will investigate to what extent all three land cover types could be distinguished using two RPV model parameters.

5.6 Classification by multiple MISR-HR RPV parameters

It is customary to use reflectance measurements in two spectral bands to better distinguish between multiple types of land cover. Such an approach is followed in this Section, though the clustering algorithm will exploit (1) the RPV parameter rho, a better spectral indicator than the straight reflectance because it is unaffected by the anisotropy of the surface, and (2) the RPV parameters k and Theta, to see whether land cover classification can also be derived using these characteristics. For this purpose, the time series available for each pixel were averaged, so that each location on the ground at any one of the sites was described by a single value in each of the RPV model parameters and spectral band, and the separability between the classes was evaluated visually by inspecting the distribution of pixels in 2-dimensional graphs.

5.6.1 Classification tests

The first test mimics the standard approach of inspecting the characteristics of the target in the red and NIR spectral bands. However, in this case, we used the RPV model parameter rho instead of the reflectance measurements, to take advantage of the fact that this product is decontaminated from the anisotropy of the surface (which it describes explicitly). Figure 5.2 shows how the entire set of all pixels from the three sites are distributed in the two-dimensional spectral space defined by the parameters rho in the red and the NIR spectral bands. As expected this approach works well to distinguish between these land cover types on the basis of the long-term average of their properties.

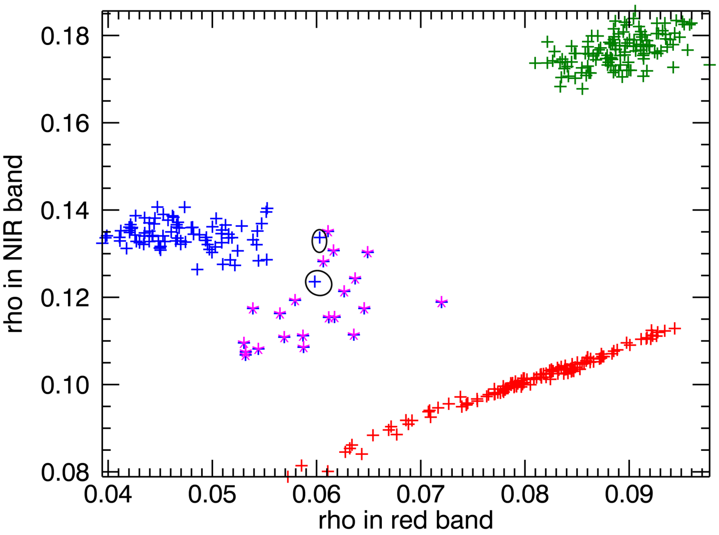

Figure 5.2: Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameter rho in the red and NIR spectral bands. The red, green and blue crosses represent the pixels of the semi-desert, wheat field and vineyard areas, respectively. Pink stars represent the non-grape cultivation pixels in the vineyard site, and the two circled blue crosses represent pixels s12_-005_+000 and s12_-004_-003, which belong to the non-grape cultivation area.

A second test was then conducted to explore further the capability of the MISR-HR RPV parameter rho in the red spectral band and the parameter k in the NIR spectral band to separate these land covers, since these indicators had been found useful earlier. Figure 5.3 shows that this particular combination of parameters also works well to distinguish these three sites. As before, (1) the red, green and blue crosses represent the temporal averages of the indicated properties of the pixels in the semi-desert, wheat field and vineyard areas, respectively, (2) the pink stars represent the non-grape cultivation pixels in the vineyard site, and (3) the circled blue crosses point to the pixels s12_-005_+000 and s12_-004_-003, respectively. It will be recalled that the former was located on the boundary of the grape cultivation and near a bright road, and that the properties of the latter may have been affected by buildings (these two pixels are marked by pink pins on the satellite map in Figure 5.1). It is noteworthy that those two pixels were not properly discriminated using the simpler method of the previous Section.

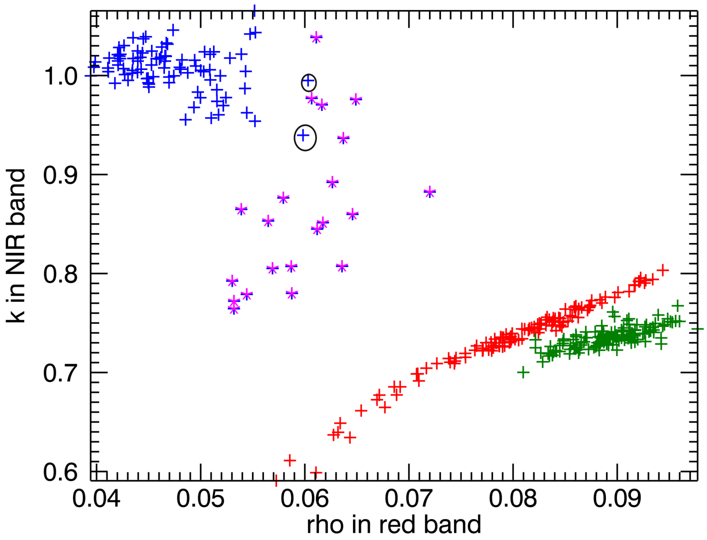

Figure 5.3: Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameters rho in the red and k in the NIR spectral bands. The colour coding is the same as for Figure 5.2.

It is also worth pointing out that the kNN algorithm may not work as expected under these conditions, because the semi-desert and the wheat field clusters are both elongated in shape but close together in distance. In this case, a different algorithm should be used, such as a SVM classifier (Press et al., 2007) for instance.

In a third test, the classification of the entire set of pixels was attempted using the MISR-HR RPV parameters rho and k, in the NIR spectral band. The result is exhibited in Figure 5.4.

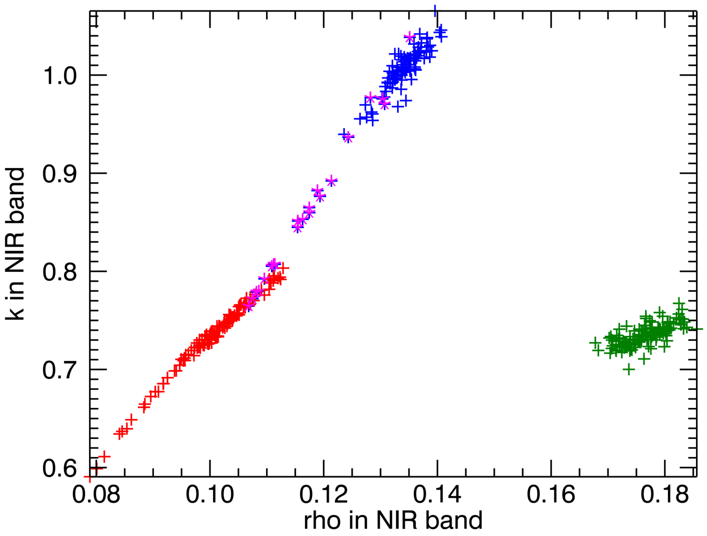

Figure 5.4: Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameters rho and k in the NIR spectral band. The colour coding is the same as for Figure 5.2.

In this case, it can be seen that all three sites are properly separated, but that the distinction between grape and non-grape pixels in the vineyard site is less satisfactory.

A fourth test was then carried out to investigate the potential of the RPV Theta parameter for classification purposes.

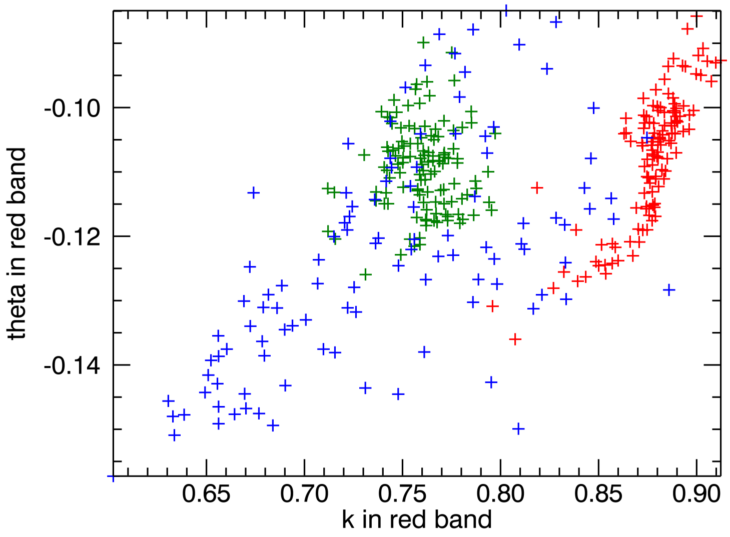

Figure 5.5: Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameters k and Theta in the red spectral band. The colour coding is the same as for Figure 5.2.

As can be seen from Figure 5.5, the vineyard and the wheat field are largely confused in this case, though the semi-desert is still differentiated from the other two sites.

On the other hand, the combination of MISR-HR RPV parameter rho in the red and Theta in the NIR spectral bands showed more promising results. Figure 5.6 exhibit the outcome of that fifth test.

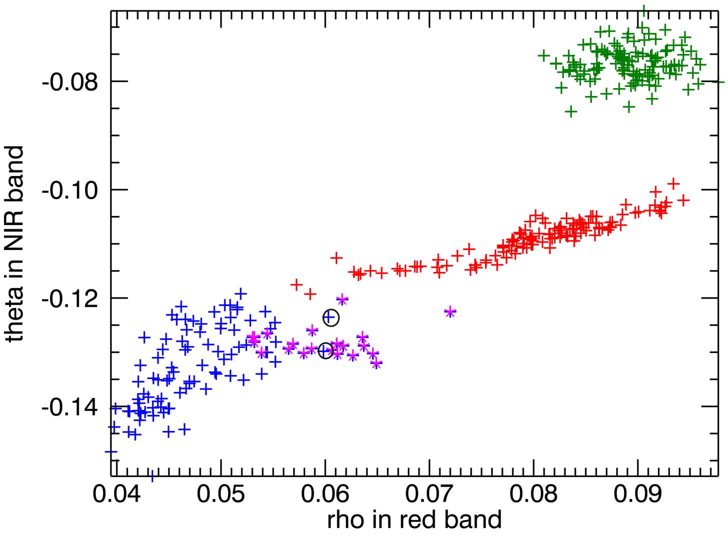

Figure 5.6. Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameters rho in the red and Theta in the NIR spectral bands. The colour coding is the same as for Figure 5.2.

Here again, it can be seen that all three land cover types are properly distinguished and that the non-grape cultivation pixels of the grapevine site exhibit anisotropic properties that are intermediary between the latter and the semi-desert sites.

Lastly, plotting the MISR-HR RPV anisotropy parameter k in the green band versus the same parameter in the blue band shows that these two model parameters are quite well correlated for all three sites. In his case, both RPV parameters carry essentially the same information, and it is not useful to try to classify the sites using both. This is shown in Figure 5.7.

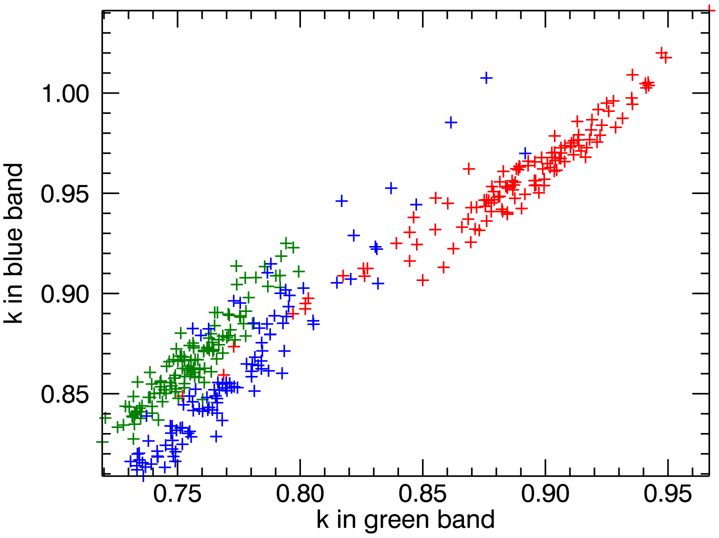

Figure 5.7 Plot of the distribution of all pixels from all three sites in the 2-dimensional space defined by the MISR-HR RPV parameter k in the green and blue spectral bands. The colour coding is the same as for Figure 5.2.

5.6.2 Discussion of the multiple RPV parameters classification method

As seen from the various classification tests in Section 5.6.1, the multiple RPV parameters classification method was able to separate the three different sites, with an accuracy that depends on the pair of RPV parameters and spectral bands used. Using two sources of information offers better opportunities to identify different clusters in the data set, as was shown with the non-grape pixels in the vineyard site. Furthermore, the special pixels identified with pink pins in Figure 5.1 and circled symbols in subsequent Figures were sometimes easier to identify and discriminate from the main clusters using this more elaborate method.

The MISR-HR RPV model parameters thus are quite suitable for the purpose of land cover classification, and offer more options or additional opportunities to discriminate between sites than using the spectral reflectance measurements alone.

It will be interesting to explore, in the future, whether some of these RPV parameters are particularly effective to identify or discriminate certain types of land cover, using a wider range of situations. In the meantime, this project has shown that quantitative information about the anisotropy of land surfaces could be used for the purpose of classifying different types of environments.

5.7 Summary

The main outcome of this Chapter is that the spectral and directional information contained in the MISR-HR RPV model parameters is adequate to perform land cover classification, and that jointly using the temporal, spectral and directional signatures of land surfaces may be very effective, especially to discriminate complex situations, such as areas that do not exhibit properties similar to the expected, clear-cut situations,

Section 5.2 described the $\boldsymbol{k}$NN algorithm, a typical example of a supervised classification algorithm. Section 5.3 explored the performance of that method in binary tests examining the separability of pixels (land areas) based on their spectral values. The same approach was then pursued in Section 5.4, but this time using directional properties of the sites instead. It was shown that both approaches worked satisfactorily, with minor differences in the results. Section 5.5 summarized and discussed those results.

Section 5.6 then proceeded to explore how land cover sites, represented by the long-term averages of their time series for the 3 RPV model parameters in each of the 4 spectral bands, were distributed in the 2-dimensional space of those parameters. It was shown that such combinations of parameters offer new opportunities to classify land cover types, and that each couple of such parameters may offer a different capacity of discriminating between those sites. Some combinations work better to separate certain land covers, while other combinations are more appropriate to discriminate other sites. This suggests that the combination of spectral and directional RPV model parameters might provide a superior capability to generate land cover maps.

This initial investigation may provide further motivation for additional exploration along those lines, especially when attempting to characterize complex environments with more than three relatively well separated land cover types. In the meantime, the goal of determining whether anisotropy may be useful to classify terrestrial environments has been conclusively achieved.

CHAPTER 6

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

Remote sensing technology plays a crucial role in understanding the evolution of climate and environment globally and holistically. It offers a unique opportunity for policy makers to formulate rational sustainable development strategies. The vast majority of studies and publications in literature on remote sensing in the solar spectral domain are focused on analysing and interpreting the spectral, spatial and temporal signatures of the observed area. However, all structured surfaces in terrestrial environments show strongly directional reflectance signatures, also known as the surface reflectance anisotropy. This is largely determined by the physical structure of the vegetation and soil in the observed area and may be documented with multi-angle remote sensing data. The structural information can be important for some applications; for instance, vegetation structure is critical in modelling the carbon cycle and global land systems. As the spectral signature of terrestrial targets is unable to represent structural information, their angular signature (anisotropy) provides a unique way of documenting aspects of the environment.

This work explored the reflectance anisotropy expressed by the MISR-HR RPV data, which is one of the MISR land surface products calculated by inverting the RPV model against MISR-HR atmospherically-corrected surface reflectance data. The author was given access to a 'data cube' of MISR-HR RPV products for 3 different sites, each containing between 108 and 121 pixels. For each of those pixels, the 3 RPV model parameters and the cost function were provided for between 100 and 200 dates, in each of the 4 spectral bands of the MISR instrument. This dataset thus collectively amounted to more than 720,000 data items. The results exhibited in the previous chapters thus represent a small selection of those that have actually been processed and investigated.

This thesis constitutes a first systematic look at the MISR-HR anisotropy data over a period of 14+ years, for three typical terrestrial environments in the Western Cape Province of South Africa, namely, a semi-desert area, a wheat field and a vineyard area. It achieved the objectives proposed at the beginning of this thesis, which were to explore (1) to what extent spectral and directional signatures of the MISR-HR RPV data varied in time and space over the three studied terrestrial targets and (2) whether the observed variations in anisotropy could be used for classifying different land surfaces or as a supplementary method to the traditional land cover classification method.

6.1 MISR-HR anisotropy data analysing tools

The MISR-HR RPV data utilised in this research were contaminated with outliers and missing values. While outliers may hint at interesting or unexpected findings, they may also skew or invalidate the results of a time series analysis.

This research started by proposing a new outlier detection method—the cost function-based outlier detection method—according to the signature of the MISR-HR RPV products (see Section 3.1). In the process of inverting the RPV model and retrieving the model parameters, the cost function, which indicates how well the RPV model fits the measurement data, is derived simultaneously. A small cost function value means the inversion model fits the data very well. Large value suggests that the model cannot adequately account for the variability present in the measurements. Thus, the cost function value can be used as an indicator of the reliability of the retrieved RPV parameters. This research used the cost function value as a threshold to detect the outliers in the MISR-HR RPV data. The merit of this cost function-based outlier detection method is that it can help in detecting unreliable RPV parameters even when the data does not 'appear' like outliers. Yet, as this cost function method may not detect all the potential outliers in the MISR-HR RPV time series, the box plot method was employed to complement this approach. Checking the contexts of the detected outliers, for instance, spatial, temporal or spectral contexts, can help in understanding the reasons causing extreme values. Overall, this proposed outlier detection method can detect the outliers effectively, which may not be detected by traditional methods and can also explain what caused extreme values in a certain level. Most significantly, this work resulted in a method capable not only to identify outliers but also to eliminate dubious results in a remote sensing data set, whether or not they look like outliers. This investigation led to a publication in the refereed literature (Liu, Verstraete & de Jager, 2018).

Another critical issue for MISR-HR anisotropy data is that of missing values. This research reviewed various methods to replace missing values by reasonable estimates (see Section 3.4). Their effectiveness depends very much on the context and purpose of the time series analysis. To avoid unintentionally introducing some bias in the MISR-HR RPV time series, much of the analysis performed avoided this approach and preferentially selected tools and methods that are gap-insensitive. The kNN clustering algorithm used in Chapter 5 is the only algorithm that required input data on identical dates, and in that case the time series reconstructed on the basis of the Lomb-Scargle method was used.

The Mann-Kendall test is a popular non-parametric trend detection method used in literature. This method can be applied without assuming that the time series elements are normally distributed. Since there was no previous knowledge about the distribution of the MISR-HR anisotropy data, this work relied the Mann-Kendall test to detect the possible presence of a monotonic trend (see Section 4.1). In addition, this trend detection method is insensitive to outliers and missing values, and therefore highly applicable to the MISR-HR RPV products investigated here. All output results generated with this method were obtained at a significance level of 0.05 (see Section 4.1).

Seasonality detection is challenging for the present data, since the RPV time series are unevenly spaced due to the variable number of missing values. The Lomb-Scargle periodogram method, which has the advantage of detecting seasonality even from unevenly spaced time series, was employed in this research (see Section 4.2.1). Hocke and Kampfer (2009) proposed a seasonal component reconstruction method based on the Lomb-Scargle algorithm (see Section 4.2.2). This reconstruction method was used on the MISR-HR anisotropy data (see Section 4.2.3). Although the reconstruction method proposed by Hocke and Kampfer (2009) fails to estimate the proper amplitudes on both ends of the seasonal component, it does retrieve the frequency and the phase of the seasonal signal in the MISR-HR RPV time series.

6.2 Statistical analysis of MISR-HR products

This work explored the statistical properties of the MISR-HR anisotropy data over a period of 14+ years for three different land surfaces. The results showed that the parameter rho exhibited distinctive values over the different study sites (see Section 2.6.1); variations of the RPV parameters are important in fully describing the RPV values, since different observed areas may have a different variation ranges for a single RPV parameter (Section 2.6.1). The exploration of the parameter k data showed that the vineyard area alone exhibited high values (around 1.0) in the NIR spectral band. This signature made the vineyard area separable from the other two study sites, as well as in a few sites studied by Armston et al. (2007) (see Section 2.6.2). It appeared that the RPV Theta parameter in the red spectral band was less discriminating and therefore perhaps less useful for the purpose of land cover mapping, although the temporal evolution of that parameter was intriguing. The spatial analysis of all three RPV parameters showed interesting correlations in the semi-desert area, which confirmed previous arguments in literature (Armston et al., 2007) on the correlation of the parameter rho and Theta data.

This work demonstrated that multiple MISR-HR RPV model parameters in various spectral bands exhibited monotonic linear trends, which would imply either changes of surface properties or a temporal drift in the calibration of the instrument over the 14+ years for which data were available. There were a number of interesting findings in the trend detection results shown in Tables 4.1, 4.2 and 4.3. For instance, all the pixels in the vineyard area and nearly all the pixels in the semi-desert area showed a downward trend in the green band parameter rho data, and upward trends in blue band parameter k and Theta for the three study sites. As the semi-desert area is supposed to be influenced by the environment alone, the trend in the green band parameter rho data might reflect either a progressive degradation of the environment (discussed in Section 4.4), or a small drift in the MISR calibration system. Unpacking these potential factors would require access to independent field data, a task that lies outside the scope of this thesis.

This project explored, for the first time, seasonality of the MISR-HR anisotropy data for a period of fourteen years, from March 2000 to May 2014, for three typical terrestrial surfaces. It turned out that all RPV parameters exhibited some seasonal variations, depending on the spectral band and land surface type (see Section 4.2). These results are congruent with earlier findings on the seasonal variation of the parameter k data, first reported by Pinty et al. (2002), and now also confirmed for the parameters rho and Theta data. The phase and/or the pattern of the seasonal component of the anisotropy data may differ over different terrestrial surfaces (see Section 4.3). Hence, these seasonal signatures of the MISR-HR anisotropy data could potentially be used for discriminating different land surfaces.

6.3 Application of anisotropy information

The diverse signatures of the MISR-HR anisotropy data over different land surfaces (see Section 5.1), motivated the classification of different terrestrial landscapes by using these RPV parameters data. The three study sites were successfully separated by the red band parameter rho data and the NIR band parameter k data combined with the kNN classifier respectively (see Sections 5.3 and 5.4). These results showed that MISR-HR anisotropy data could be used for discriminating different ground surfaces, which fulfilled the second objective of this research project.

This work explored a new method using multiple RPV parameters as the classification features to discriminate different land covers (see Section 5.6). The tests on this proposed classification method proved that the three study sites could be separated successfully (see Section 5.6.1). In addition, different terrestrial landscapes are characterised by different combinations of the RPV products. Since MISR-HR RPV products represent the angular information of the observed area, this proposed multiple RPV parameters classification method could usefully complement traditional spectral methods, especially when the land covers vary mostly in structure.

6.4 Suggestions for future research

Based on the exploration of the MISR-HR anisotropy data in this work, the following research directions could be pursued in subsequent studies:

- The current investigation on the trends of the MISR-HR anisotropy data shows that parameter rho data in the green wavelength exhibits a downward trend for the entire vineyard area and in nearly all the pixels in the semi-desert area (Table 4.1), and that there is an upward trend in the blue band parameters k and Theta data (Tables 4.2 and 4.3) for all the three study sites. Future work may focus on revealing what caused the monotonic trend in these MISR-HR RPV time series.
- The statistical analysis of parameter k (see Section 2.6.2) shows that the vineyard area has distinctively high NIR band k values compared to the NIR band k values of the semi-desert area and the wheat field in this research, and even more ground surfaces as studied by Armston et al. (2007). Subsequent study may focus on investigating what caused the very high NIR band k values in this vineyard area, to reveal what feature of the grape cultivation is represented by the NIR band k data, with a view to monitoring grape cultivation in South Africa by this RPV parameter product.
- Although the current study focused on discriminating different ground surfaces on the basis of their temporal, spectral and directional properties (see Sections 5.3, 5.4 and 5.6), more detailed and systematic studies might be able to explore these issues in greater detail, in particular using longer time series as well as more combinations of RPV parameters and spectral bands. Thus, change detection with the MISR-HR RPV time series data is expected to usefully complement the more traditional methods based largely on the spectral information only (e.g., Lu, Mausel, Brondizio & Moran, 2004b).
- This research provides an initial hint that using multiple RPV parameters can be used to effectively classify land use and land cover (see Section 5.6). It is expected that this approach will help improve the accuracy of traditional approaches, in which case the combined methods should be tested on harder problems, which cannot be solved with the current methods.

REFERENCES

Aggarwal, C.C. 2013. *Outlier analysis*. New York: Springer.

Alonso, S. et al. 2008. A consistency-based procedure to estimate missing pairwise preference values. *International Journal of Intelligent Systems*, **23**(2):155–175.

Armston, J.D., Scarth, P.F., Phinn, S.R. and Danaher, T.J. 2007. Analysis of multi-date MISR measurements for forest and woodland communities, Queensland, Australia. *Remote Sensing of Environment*, **107**:287–298.

Battrick, B. (ed). 2005. *GEOSS 10-year implementation plan reference document*. Noordwijk: ESA Publications Division.

Bluman, A.G. 2012. *Elementary statistics: A step by step approach*. 8th ed. New York: McGraw-Hill.

Chandler, R.E. and Scott, E.M. 2011. *Statistical methods for trend detection and analysis in the environmental sciences*. United Kingdom: John Wiley & Sons.

Chuvieco, E. and Huete, A. 2010. *Fundamentals of satellite remote sensing*. New York: Taylor & Francis.

Cooper, T.M. and Hart, P.E. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT-13:21–27.

Coppin, P. et al. 2004. Digital change detection methods in ecosystem monitoring: a review. *International Journal of Remote Sensing*, **25**(9):1565–1596.

Cracknell, A.P. and Hayes, L. 2007. *Introduction to remote sensing*. 2nded. New York: Taylor & Francis.

Dandois, J.P. and Ellis, E.C. 2010. Remote sensing of vegetation structure using computer vision. *Remote Sensing*, **2**:1157–1176. Doi: 10.3390/rs2041157. [Accessed 03 March 2014]

Dawson, R. 2011. How significant is a Boxplot outlier? *Journal of Statistics Education*, **19**(2). www.amstat.org/publications/jse/v19n2/dawson.pdf

Diner, D.J. et al. 1999. New directions in Earth observing: Scientific applications of multi-angle remote sensing. *Bulletin of the American Meteorological Society*, **80**:2209–2228.

Elachi, C. and Van Zyl, J.2006. *Introduction to the physics and techniques of remote sensing*. New Jersey: John Wiley & Sons.

Errico, R.M. 1997. What is an adjoint model? *Bulletin of the American Meteorological Society*, **78**:2577–2591.

Fan, J. and Yao, Q. 2003. *Nonlinear time series: Nonparametric and parametric methods*. New York: Springer.

Feuerbacher, B and Stoewer, H. (eds.). 2006. *Utilization of space: today and tomorrow*. Berlin: Springer.

Franco-Lopez, H., Ek, A.R. and Bauer, M.E. 2001. Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbor method. *Remote Sensing of Environment*, **77**:251–274.

Gerstl, S.A.W. 1990. Physics concepts of optical and radar reflectance signatures: A summary review. *International Journal of Remote Sensing*, **11**:1109–1117.

Giering, R and Kaminski, T. 1998. Recipes for adjoint code construction. *ACM Transactions on Mathematical Software*, **24**:437–474.

Grubbs, F.E. 1969. Procedures for detecting outlying observations in samples. *Technometrics*, **11**:1-21.

Gupta, M., Gao J., Aggarwal, C.C. and Han, J. 2014. Outlier detection for temporal data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, **26(9):**2250–2267. Doi: 10.1109/TKDE.2013.184. [Accessed 23 January 2014]

Han, J., Kamber, M. and Pei, J. 2012. *Data mining concepts and techniques*. 3rd ed. Waltham, MA: Morgan Kaufmann Publishers.

Hawkins, D.M. 1980. *Identification of outliers*. London: Chapman & Hall.

Helsel, D.R., Mueller, D.K. and Slack, J.R. 2006. Computer program for the Kendall family of trend tests. *Scientific Investigations Report 2005–5275*. Virginia: U.S. Geological Survey.

Heymann, S., Latapy, M., and Magnien, C. 2012. *Outskewer: Using skewness to spot outliers in samples and time series.* In Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 527–534). IEEE/ACM. Retrieved from doi:10.1109/ASONAM.2012.91 https://www-complexnetworks.lip6.fr/~magnien/ DynGraph/Docs/2012asonam.pdf

Hirsch, R.M., Alexander, R.B., and Smith, R.A. 1991. Selection of methods for the detection and estimation of trends in water quality. *Water Resources Research,* **27**:803–-814.

Hirsch, R.M., Slack, J.R. and Smith, R.A. 1982. Techniques of trend analysis for monthly water quality data. *Water Resources Research,* **18**:107–121.

Hocke, K. and Kampfer, N. 2009. Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. *Atmospheric Chemistry and Physics,* **9**:4197–4206. www.atmos-chem-phys.net/9/4197/2009/. [Access 16 May 2015]

Hodge, V.J. and Austin, J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review,* **22**(2):85–126.

Honaker, J. and King, G. 2010. What to do about missing values in time-series cross-section data. American Journal of Political Science, **54**(2):561–581.

Horne, J.H and Baliunas, S.L. 1986. A prescription for period analysis of unevenly sampled time series. *The Astrophysical Journal,* **302**:757–763. Doi: 10.1086/164037. [Accessed 08 April 2015]

Jia, K. et al. 2014. Land cover classification of Landsat data with phenological features extracted from time series MODIS NDVI data. *Remote Sensing*, **6**:11518–11532.

Jiang, Y., Lan, T. and Wu, L. 2009. A comparison study of missing value processing methods in time series data mining. *2009 International Conference on Computational Intelligence and Software Engineering*. IEEE. Doi: 10.1109/CISE.2009.5365076. [Accessed 09 August 2014]

Kendall, M.G. 1975. *Rank correlation methods*. 4th ed. London: Charles Griffin.

Kerle, N., Janssen, L.F. and Huurneman, G.C. (eds.). 2004. *Principles of remote sensing*. 3rd ed. Enschede: The International Institute for Geo-Information Science and Earth Observation (ITC).

Kimes, D.S. and Kirchner, J.A. 1982. Irradiance measurement errors due to the assumption of a Lambertian reference panel. *Remote Sensing of Environment*, **12**:141-149.

Lhermitte, S., Verbesselt, J., Verstraeten, W.W. and Coppin, P. 2011. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. *Remote Sensing of Environment*, **115**:3129–3152.

Liang, S. (ed.). 2008. *Advances in land remote sensing*. New York: Springer Science + Business Media B.V.

Liu, Q., Liang, S., Xiao, Z. and Fang, H. 2014. Retrieval of leaf area index using temporal, spectral, and angular information from multiple satellite data. *Remote sensing of environment*, **145**:25-37.

Liu, Z., Verstraete, M.M. and de Jager, G. 2017. Handling outliers in model inversion studies: a remte sensing case study using MISR-HR data in South Africa. South African geographical journal, **100(1)**:122-139. DOI: 10.1080/03736245.2017.1339629

Lomb, N.R. 1976. Least-Squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, **39**:447–462.

Longobardi, A. and Villani, P. 2009. Trend analysis of annual and seasonal rainfall time series in the Mediterranean area. *International Journal of Climatology*, **30**(10):1538–1546.

Lu, D., Mausel, P., Batistella, M. and Moran, E. 2004a. Comparison of land-cover classification methods in the Brazilian Amazon basin. *Photogrammetric Engineering and Remote Sensing*, **70**(6):723–731.

Lu, D., Mausel, P., Brondizio, E. and Moran, E. 2004b. Change detection techniques. *International Journal of Remote Sensing,* **25**(12):2365–2407.

Mann, H.B. 1945. Non-parametric tests against trend. *Econometrica*, **13**:245–259.

Marcal, A.R.S., Borges, J.S., Gomes, J.A. and Pinto da Costa, J.F. 2005. Land cover update by supervised classification of segmented ASTER images. *International Journal of Remote Sensing*, **26**(7):1347–1362.

Manoj, K., and Senthamarai, K.K. 2013. Comparison of methods for detecting outliers. *International journal of scientific and engineering research*, **4**:709-714.

McRoberts, R.E., Nelson, M.D. and Wendt, D.G. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbor technique. *Remote Sensing of Environment*, **82**:457–468.

MISR Jet Propulsion Laboratory. N.d. *MISR Instrument*. Available at http://www-misr.jpl.nasa.gov/Mission/misrInstrument/ [Accessed 26 October 2014]

Moura, Y.M., Galvao, L.S., Santos, J.R., Roberts, D.A. and Breuning, F.M. 2012. Use of MISR/Terra data to study intra- and inter-annual EVI variations in the dry season of tropical forest. *Remote sensing of environment*, **127**:260-270.

Musial, J.P., Verstraete, M.M. and Gobron, N. 2011. Technical note: comparing the effectiveness of recent algorithms to fill and smooth incomplete and noisy time series. *Atmospheric Chemistry and Physics,* **11**:7905–7923.

NIST/SEMATECH. 2013. *E-handbook of statistical methods*. Technical report, NIST. Available from http://www.itl.nist.gov/div898/handbook/. [Accessed 10 March 2013].

Nopiah, Z.M. et al. 2012. The use of autocorrelation function in the seasonality analysis for fatigue strain data. *Journal of Asian Scientific Research*, **2**(11):782–788.

Osborne, J.W. and Overbay, A. 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation,* **9**(6). Available from http://pareonline.net/getvn.asp?v=9&n=6. [Accessed 09 June 2013].

Peterson, T.C., Vose, R., Schnoyer, R. and Razuvaev, V. 1998. Global historical climatology network (GHCN) quality control of monthly temperature data. *International jounal of climatology*, **18**:1169-1179.

Pinty, B. et al. 2002. Uniqueness of multi-angular measurements – Part I: an indicator of subpixel surface heterogeneity from MISR. *IEEE Transactions on Geoscience and Remote Sensing*, **40**(7):1560–1573.

Pisek, J., Ryu, Y., Sprintsin, M., He, L., Oliphant, A. J., Korhonen, L., Kuusk, J., Kuusk, A., Bergstrom, R., Verrelst, J. and Alikas, K. 2013. Retrieving vegetation clumping index from Multi-angle Imaging SpectroRadiometer (MISR) data at 275 m resolution. *Remote sensing of environment*, **138**:126 – 133.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. 2007. *Numerical recipes.* 3rd ed. New York: Cambridge University Press.

Rahman, H., Pinty, B. and Verstraete, M.M. 1993. Coupled Surface-Atmosphere Reflectance (CSAR) Model 2. Semiempirical surface model usable with NOAA advanced very high resolution radiometer data. *Journal of Geophysical Research*, **98**(11):20791–20801.

Ruf, T. 1999. The Lomb-Scargle periodogram in biological rhythm research: analysis of incomplete and unequally spaced time-series. *Biological rhythm research*, **30**(2):178–201.

Sabins, F.F. 1987. *Remote sensing principles and interpretation*. 2nd ed. New York: W.H. Freeman and Company.

Salmon, B.P. et al. 2011. The use of a Multilayer Perceptron for detecting new human settlements from a time series of MODIS images. *International Journal of Applied Earth Observation and Geoinformation*, **13**:873–883.

Scargle, J.D. 1982. Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly sampled data. *The Astrophysical Journal*, **263**:835–853.

Scargle, J.D. 1989. Studies in astronomical time series analysis. III. Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *The Astrophysical Journal*, **343**:874–887.

Schneider, T. 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, March 2001. DOI: http://dx.doi.org/10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2. [Accessed 05 November 2014].

Sen, P.K. 1968. Estimates of the regression coefficient based on Kendall's Tau. *Journal of the American Statistical Association*, **63**(324):1379–1389.

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets (Master's thesis).* Graduate School of Public Health of the University of Pittsburgh. Available from http://d-scholarship.pitt.edu/7948/1/Seo.pdf [Accessed 17 July 2014]

Singh, K. and Upadhyaya, S. 2012. Outlier detection: applications and techniques. *International Journal of Computer Science Issues*, **9**(1):307–323.

South African National Space Agency. 2012. *Strategic Plan 2012–1017*. Available from http://www.sansa.org.za/images/resource_centre/publications/SANSA%20Strategic%20Plan.pdf. [Accessed 20 August 2014].

Sreevidya, S.S. et al. 2014. A survey on outlier detection methods. *International Journal of Computer Science and Information Technologies*, **5**(6):8153–8156.

Szuster, B.W., Chen, Q. and Borger, M. 2011. A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones. *Applied Geography*, **31**:525–532.

Taati, A. et al. 2014. Land use classification using Support Vector Machine and Maximum Likelihood algorithms by Landsat 5 TM images. *Walailak Journal*, **12**(8):681–687.

Tarantola, A. 1987. *Inverse problem theory, methods for data fitting and model parameter estimation*. New York: Elsevier Science.

Tarantola, A. 2005. *Inverse problem theory and methods for model parameter estimation*. Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Tempfli, K., Kerle, N., Huurneman, G.C. and Janssen, L.L.F. (eds.) 2009. *Principles of remote sensing*. 4th ed. Enschede: ITC. Available at https://www.itc.nl/library/papers_2009/general/principlesremotesensing.pdf. [Accessed 01 October 2015].

Townsend, R.H.D. 2010. Fast calculation of the Lomb-Scargle periodogram using graphics processing units. *The Astrophysical Journal Supplement Series*, **191**:247–253.

Triola, M.F. 2012. *Elementary statistics*, 11th ed. Boston: Pearson Education.

Troyanskaya, O. et al. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6):520–525. DOI: 10.1093/bioinformatics/17.6.520. [Accessed 16 February 2015].

Tsypin, M. and Roder, H. 2007. On the reliability of kNN classification. *Proceedings of the World Congress on Engineering and Computer Science 2007 (WCECS 2007), San Francisco*.

Tukey, J.W. 1977. *Exploratory data analysis*. Boston: Pearson Education.

United Nations' sustainable Development Goals. 2015. Available from https://www.un.org/sustainabledevelopment/sustainable-development-goals/ [Accessed 18 September 2015].

Verbesselt, J., Hyndman, R., Newnham, G. and Culvenor, D. 2010a. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, **114**:106–115.

Verbesselt, J., Hyndman, R., Zeileis, A. and Culvenor, D. 2010b. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, **114**:2970–2980.

Verstraete, M.M. and Pinty, B. 2000. Environmental Information Extraction from Satellite Remote Sensing Data. *Global Biogeochemical Cycles (AGU Monograph),* **114**:125–137.

Verstraete, M.M. et al. 2012. Generating 275 m resolution land surface products from the Multi-Angle Imaging SpectroRadiometer Data. *Geoscience and Remote Sensing, IEEE Transactions*, **50**(10):3980–3990.

Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K. 2012. *Probability and statistics for engineers and scientists.* Boston: Prentice Hall.

Wang, J., Neskovic, P. and Cooper, L.N. 2003. Partitioning a feature space using a locally defined confidence measure. *International Conference on Artificial Neural Networks and International Conference on Neural Information Processing, Istanbul, Turkey, June 2003*:200–203.

Wang, J., Neskovic, P. and Cooper, L.N. 2006. Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Pattern Recognition*, **39**:417–423.

Wardlow, B.D., Egbert, S.L. and Kastens, J.H. 2007. Analysis of time-series MODIS 250m vegetation index data for crop classification in the U.S. Central Great Plains. *Remote Sensing of Environment*, **108**(3):290–310.

Yadav, G. and Mishra, N. 2015. Air pollution trend analysis using Sen estimator method. *International Journal of Advanced Research in Computer Science and Software Engineering*, **5**(7):1073–1080