TUNING A STATISTICAL TRADE-OFF BETWEEN SPECTRAL AND SPATIAL DOMAINS TO PREDICT PLANT TRAITS WITH HYPERSPECTRAL REMOTE SENSING

Alby Duarte Rocha

TUNING A STATISTICAL TRADE-OFF BETWEEN SPECTRAL AND SPATIAL DOMAINS TO PREDICT PLANT TRAITS WITH HYPERSPECTRAL REMOTE SENSING

DISSERTATION

to obtain
the degree of doctor at the Universiteit Twente,
on the authority of the rector magnificus,
Prof.dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board
to be publicly defended
on Wednesday 25 September 2019 at 16.45

by

Alby Duarte Rocha

born on 10 May 1976

in São Paulo, Brazil

This thesis has been approved by
Prof. dr. A.K. Skidmore, supervisor
Dr. T.A. Groen, co-supervisor
Dr. R Darvishzadeh Varchehi, co-supervisor

ITC dissertation number 365
ITC, P.O. Box 217, 7500 AE Enschede, The Netherlands

ISBN 978-90-365-4862-5
DOI 10.3990/1.9789036548625

Cover designed by Job Duim
Printed by ITC Printing Department

Graduation committee:

Chairman/Secretary

Prof.dr.ir. A. Veldkamp	

Supervisor(s)

prof.dr. A.K. Skidmore	University of Twente

Co-supervisor(s)

dr. T.A. Groen	University of Twente
dr. R Darvishzadeh Varchehi	University of Twente

Members

prof.dr. V.G. Jetten	University of Twente
prof.dr. R.J. Boucherie	University of Twente
prof. dr. D. Tuia	Wageningen University
prof. dr. M. Herold	Wageningen University

"I cannot teach anybody anything,
I can only make them think."
Socrates

Acknowledgements

I wish to express my great appreciation for everyone who has helped me in this long and winding road that leads me to my PhD. I received valuable contributions along the way from many people inside and outside academia. Of course, the first person that comes to my mind is Sheila (my wife), to whom I devote all my gratitude and love for having encouraged me for four long winters to pursue this goal. I also want to give special thanks for my family and friends that supported me before, during and after the time spent in the Netherlands (which I will not nominate here because they are so many that for sure I would miss someone). However, I could not forget to mention my mother that had followed only part of my education journey, but she left me the feeling that I am free to decide my own path.

I would like to express my appreciation of everyone who has cooperated in carrying out this research and have contributed to this thesis. It is undeniable that the success of this project has received great contribution from my research committee composed by Andrew Skidmore, Thomas Groen, Roshanak Darvishzadeh and Louise (Wieteke) Willemen. I want to praise Thomas Groen for being ever patient and friendly, always cooperating despite all my unconventional ideas and stubbornness. I want to show my respect for Andrew Skidmore, who politely but sharply used to rise crucial points that instigated me to improve further.

I am truly glad for the opportunity of becoming part of such a plural community as ITC Faculty and meeting people from a variety of nationalities and cultures. I am also grateful to colleagues and pairs that did not hesitate to share experiences, insights, pieces of advice and (why not?) some beers. Thank you for all the staff of the Natural Resources Department, which have provided me with a great environment at work.

The current research was supported by CNPq (the Brazilian National Council for Scientific and Technological Development). My recognition for the effort of the former Brazilian government to provide opportunities for international experience, which in my case, made me realise that we also have excellent universities in our country. It would be impossible to go throughout the PhD and write this thesis without the help of them all.

Table of Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Plant traits and ecosystem dynamics

The understanding of ecological processes from patterns observed in nature is a recurrent goal in ecology and many other related fields (Legendre and Fortin, 1989). Biomass production and biogeochemical cycles are vegetation properties often linked with essential morphological, physiological and phenological plant characteristics (Van Cleemput et al., 2018). For instance, biochemical and biophysical characteristics in vegetation represented by plant traits such as leaf chlorophyll content and leaf area index (LAI) are essential to understand photosynthesis processes and net primary productivity (Kokaly et al., 2009; Schlerf et al., 2010).

Observation of plant traits enriches the understanding of the dynamics of the ecosystems (Van Cleemput et al., 2018). The monitoring of plant traits in natural environments is important for conservation (Abdullah et al., 2018; Skidmore et al., 2015). Plant trait measurements are used by agribusiness to evaluate crop yields or to fine-tune fertiliser application (Boegh et al., 2013; Hansen and Schjoerring, 2003). Approximately 40% of the total land area on Earth is covered by grassland and shrub plants. This ecosystem provides essential habitats to many species, and also regulate water quality and soil erosion (Van Cleemput et al., 2018; Wang et al., 2014). To have a better understanding of the dynamics of our planet, it is therefore essential to assess changes in plant traits in this ecosystem (Van Cleemput et al., 2018; Wang et al., 2014).

1.1.1 Measuring plant traits

The process of observing vegetation dynamics depends on the methods to measure plant trait accurately (Dutilleul, 1993; Milton et al., 2009; Pearse et al., 2016). In situ measurements of plant traits are frequently available for limited areas, as data collection is time-consuming and expensive (Milton et al., 2009). Direct measurements are often destructive, for instance, when determining chlorophyll or nitrogen concentrations by chemical analysis (Muñoz-Huerta et al., 2013). Also, biophysical plant traits such as leaf area index (LAI) require harvesting of all the leaves from sampled plants (Lee et al., 2004). The difficulty of obtaining direct measurements in more isolated or vulnerable environments restricts the availability of plant trait information for these areas (Vallejos and Osorio, 2014).

Although data on plant traits from many species at the local and global scale are available, these databases cover only about 2% of the known vascular plants (Van Cleemput et al., 2018). Also, as they are measured by different methods and instruments, their values are not usually directly comparable, and inconsistency may exit in the measurement protocols (Van Cleemput et al., 2018). The lack of comprehensive and standardised datasets adding to the

limitations on field campaigns constrains the availability of functional traits at finer temporal and spatial scales (Hoeting, 2009; Muñoz-Huerta et al., 2013; Secades et al., 2014; Wikle, 2003; Wilson et al., 2011). More efficient procedures to measure plant traits indirectly are needed to observe and monitor vegetation dynamics (Pearse et al., 2016). A common alternative for measuring indirectly plant trait are optical instruments. This method is non-destructive and can be used in situ, avoiding the necessity of physical and chemical laboratory analysis (Milton et al., 2009).

1.2 Estimating plant traits from remote sensing

Remote sensing can be used to observe vegetation over spatially continuous areas at a temporally regular pace (Manolakis et al., 2003; Van Cleemput et al., 2018). The amount of radiation emitted from a vegetation surface is captured by an optical sensor, which can be linked with structural and biochemical plant traits (Curran, 1989). Therefore, remote sensing technology creates the possibility to observe spatial and temporal changes in vegetation (Legendre and Fortin, 1989; Si et al., 2012). Many anthropogenic activities are changing biochemical processes, altering plant traits such as nitrogen or carbon concentration, without necessarily changing the land cover directly (Van Cleemput et al., 2018). Therefore, apart from land cover maps, quantitative trait maps are needed as ecosystems can be altered without any direct land use or land cover changes (Lovett et al., 2005; Secades et al., 2014).

The assessment of plant traits from a specific species at a local level to an entire ecosystem has shown to be promising with the advances of remote sensing and computer processing (Feilhauer et al., 2017; Secades et al., 2014; Van Cleemput et al., 2018). The estimation of biochemical plant traits by remote sensing relies mostly on the quantification of leaf pigments or moisture through the reflectance from certain spectral regions (Curran, 1989; Schaepman-Strub et al., 2006). It is the case for chlorophyll and water content, essential plant traits related to photosynthesis and plant stress (Buitrago et al., 2018; Clevers et al., 2010; Clevers and Kooistra, 2012). While biophysical plant traits such as leaf area index (LAI) can be estimated using optical instrument by the difference of the transmittance of visible light below and above the canopy (Pearse et al., 2016). Indirect estimations of plant traits using remote sensing have presented satisfactory accuracy for many vegetation types and environments (Boegh et al., 2013; Van Cleemput et al., 2018). These instruments can mitigate the limitations of direct measurements of plant traits and provide opportunities to collect ground references over a comprehensive range of temporal and spatial scales (Finley et al., 2014; Patenaude et al., 2008; Shen et al., 2013a; Wilson et al., 2011).

1.2.1 Hyperspectral remote sensing

Hyperspectral sensors capture a comprehensive wavelength range, divided into narrow bands (Shaw and Burke, 2003; Milton et al.,2009). Many studies have demonstrated that, in general, hyperspectral remote sensing estimate plant traits more accurately than sensors designed with broad bands around the visible spectra (Clevers and Kooistra, 2012; Lee et al., 2004). The resultant wavelengths are sequential measurements of radiation from the plant surfaces that represent interactions from physical, chemical and biological properties (Huber et al., 2008; Kokaly et al., 2009). Optical measurements are often transformed in reflectance values to estimate leaf and canopy plant traits by physical or empirical models (Curran, 1989; Manolakis et al., 2003). Hyperspectral measurements are provided by sensors with a fine spectral resolution, which capture an extended region of the electromagnetic spectrum (0.4nm to 2.5nm) dominated by solar illumination (Manolakis et al., 2003). These sensors measure the radiation reflected by the target surface at a large number of narrow wavelengths from the visible (red, green, and blue) and the invisible frequency (Manolakis et al., 2003; Vohland and Jarmer, 2008). The detection of changes in these specific regions of the spectrum allows monitoring biological processes more precisely (Lee et al., 2004; Manolakis et al., 2003; Wang et al., 2014).

Hyperspectral sensors provide a detailed spectral signature of the target vegetation (Figure 1.1), but even in a controlled laboratory environment, a distinctive and unique signature for given surface properties is unlikely(Curran, 1989; Manolakis et al., 2003). In a natural environment, reflectance depends greatly on sunlight variations observed at the moment it is captured, such as soil moisture, weather conditions or solar angle about the view of the sensor (Dutilleul, 1993). These conditions are independent of the plant characteristics, but they affect the reflectance captured by the sensor (Atkinson and Emery, 1999). In addition, space and time-dependent variations interact with the vegetation radiance, and the area imaged is often a mix of species at different stages of growing and senescence (Clevers and Kooistra, 2012; Knyazikhin et al., 2013; Martin et al., 2008).

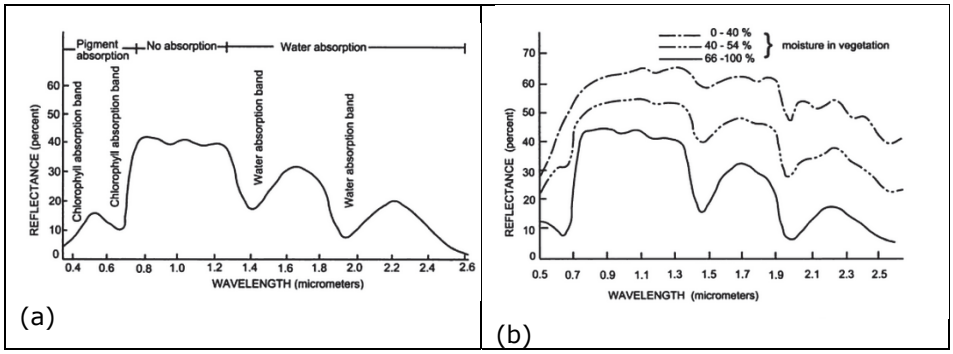


Figure 1.1– (a) Typical response curve for vegetation from a hyperspectral sensor, showing the absorption of pigmented substances (*e.g.* chlorophyll), and of a non-pigment content (*e.g.* LAI, water content), and (b) reflectance in vegetation with different levels of moisture. Extracted from McCoy (2005).

1.2.2 Space and time misalignment with remote sensing

Apart from the spectral domain, remote sensing data normally present two more dimensions, space and time. The spatial domain is determined by the resolution of the pixels and the extent of the scene captured by the sensor (Wilson et al., 2011). The temporal domain is related to the frequency at which the images are taken, and the duration of recording the radiance (cf. shutter speed with camera's). Depending on the sensor platform the area captured (instantaneous field of view) can vary from an individual pixel to a scene of thousands of pixels and many square kilometres of extent simultaneously (Manolakis et al., 2003). Regardless of the scene size, spectral measurements are not independent in space or time, and the spectral domain cannot be disassociated from the spatial and temporal domain (Webster et al., 1989). Airborne or spaceborne spectral images should be recorded as simultaneous as possible with the ground references using a similar spatial resolution to reduce misalignment and minimise variations on the reflectance unrelated to vegetation (Wilson et al., 2011). These platforms present spectral unit (pixel) more suitable to a canopy-level than to a leaf level by the great difference in spatial resolution (Huber et al., 2008). This difference in scale between reflectance and ground references of plant trait is called a change of support problems. This scale difference will include new components of variations such as soil background, canopy structure and size, shadows and mixed species in the pixel (Ullah et al., 2012).

Other components of variability related to the spatial alignment between spectra and ground references are errors in plot coordinates, upscale or downscale, distortions to departing from the nadir, among others (Manolakis et al., 2003). The discretisation of continuous domains such as spectra, space or time results in the loss of a certain amount of information (Bruce et al., 2002). The spatial resolution of a remote sensing data (pixel) or the sample

unit of a plant trait ground references (plot), rarely ever present the same size, position, aggregation method and time alignment (Atkinson and Emery, 1999). A mismatching can affect the relationship between spectra and plant trait, but some degree of mismatching is tolerable to make field campaigns feasible (Gotway and Young, 2002).

1.3 Modelling plant traits with hyperspectral data

Remote sensing can greatly boost the observation of plant traits and vegetation dynamics. However, to understand spatial-temporal patterns or to predict plant traits by remote sensing, a multidisciplinary approach is required. Optical, chemical, ecological, temporal, spatial and statistical understanding is needed to avoid incorrect inferences about the underlying process which drive the plant trait. For instance, the empirical relationship between leaf chlorophyll content and reflectance at canopy level *in situ* goes far beyond the physical explanation of radiance for a given concentration of leaf pigment. Factors directly related to the vegetation characteristics such as species composition, phenological stage or last occurrence of a fire disturbance are inherent the place and cannot be completely isolated or even measured in some cases. For indirect factors related to the environment such as soil nutrients, water availability, slope or temperature it is even more challenging to include in the modelling process (Knyazikhin et al., 2013; Martin et al., 2008). Factors completely independent of the underlying process, such as atmospheric and climatic factors that affect the radiation at the moment of capturing are the most unwanted variation in the modelling process (Schaepman-Strub et al., 2006). These factors include the intensity and position of the illumination source, sensors viewing angle, (cloud) shadows and background reflectance and radiation (Thenkabail et al., 2000).

Hyperspectral remote sensing data is very susceptible to random variation (noise) in some regions of the spectrum depending on atmospheric conditions and the capacity to control illumination and view geometry (Manolakis et al., 2003). These variations may lead to a lack of generalisation power in models, making the need for fieldwork every time a new spectra image is captured, hampers most of the gaining in scale from remote sensing applications (Verrelst et al., 2015). Suitable sampling designs for the ground references and the definition of an appropriated regression method is essential for a modelling process involving spectral, spatial and temporal variations (Wang and Gertner, 2013; Webster et al., 1989). Given the high dimensionality of the spectral part in hyperspectral data, space and time domains are neglected and commonly assumed as constant. The decision about which domains should be prioritised depends on whether the model is aimed to be explanatory or predictive (explain or predict).

1.3.1 Physical versus empirical models

The relation between plant trait and reflectance should be empirically estimated by a statistical model using ground references, or be deducted from known optical properties of the vegetation by a physical model. Plant traits can be estimated (*i.e.* retrieval) by physical models from spectral radiance (or reflectance) at leaf and canopy levels (Goodenough et al., 2006). Such models are deterministic and based on physical principals that rule the relationship between reflectance and a set of plant traits (Jacquemoud et al., 2009). Radiative transfer models (RTM), such as PROSAIL, SCOPE, DART are often used for retrieving plant traits (Verrelst et al., 2015). Despite the current knowledge of the physical relationship between spectra and plant traits, a deterministic model based only on spectral radiance remains a challenge (Combal et al., 2002). The main difficulty is to control or consistently measure all the factors required to parameterise the models, such as illumination and plant structure (Vohland and Jarmer, 2008).

In the case of spectral observations at canopy level captured under sunlight illumination at a heterogeneous area, physical models are still conceptually right but technically beyond (Combal et al., 2002; Goodenough et al., 2006). For instance, the estimation of an essential parameter to retrieve LAI in the PROSAIL model, such as leaf angle distribution, will probably be unreliable when determined for pixels with mixed canopies in a heterogeneous landscape. Therefore, remote sensing applications to predict plant traits still relying mostly on empirical relationships rather than physical relationships (Goodenough et al., 2006). The empirical relationships are often established by fitting regression models using reflectance as covariate and ground references of the plant trait as a response variable to train a model.

1.3.2 Regression methods to predict plant trait with hyperspectral data

The most commonly used modelling approaches to estimate plant traits are ordinary least square regressions. For ordinary linear regressions, it is necessary to reduce the number of hyperspectral bands drastically because of the lack of degrees of freedom and multicollinearity (Dormann et al., 2013). Often a vegetation index from a combination of two (or more) hyperspectral bands is used as a covariate (Li et al., 2011a). The index can be selected by a-priori knowledge about the capacity of explaining the variations in the target plant trait (Curran, 1989).

It is also a possibility to fit a multiple linear regression with different indices or latent variables created by grouping bands using techniques such as principal components or wavelets as covariates (Bioucas-Dias and Nascimento, 2008;

Bruce et al., 2002). The latter regression approaches are unsupervised and do not require the use of the response variable to select the covariates for the model (Kuhn and Johnson, 2013). However, these approaches require a previous step to define spectral indices or latent variables used as a covariate that can be created in an unsupervised or supervised way (James et al., 2013).

The selection of covariates for a model lacking a deep knowledge of the subject turns trick using all the hyperspectral bands without using a supervised approach. Machine learning algorithms can be easily applied using the entire spectral range, facilitating the model selection when previous knowledge is unavailable (Hastie et al., 2009). Machine learning methods such as Artificial Neural Network (ANN), Partial Least Squares (PLSR), Support Vector Machine (SVM) and Random Forest (RF) are broadly used for modelling plant traits with hyperspectral data (Abdel-Rahman et al., 2013; Mountrakis et al., 2011; Van Cleemput et al., 2018). They are often reported as being more accurate comparing to ordinary regressions (Van Cleemput et al., 2018). These regression methods are also considered supervised methods because the model is tuned using the support of the response variable (James et al., 2013). These models tend to become very complex, which decreases the capacity of interpretation and understanding of how each wavelength contribute to the model.

The spectral domain contains valuable information to estimate plant trait, but the spatial and the temporal domains can also be an important source of explanation about the plant trait variation. A spatially or temporally explicit model to estimate plant traits using the full hyperspectral range as covariates is technically hard to fit because of the high dimensionality (Hoeting, 2009; Wikle and Hooten, 2010). Therefore, one domain should be prioritised, and the others drastically reduced (hyperspectral) or considered constant (space or time). Spatial models fitted by Bayesian inference using Markov Chain Monte Carlo (MCMC) simulations are currently easily available (Banerjee and Fuentes, 2012; Bivand et al., 2015; Heaton et al., 2017). Although, for very complex spatial models, the MCMC method is still time and computationally demanding (Wang et al., 2018). The method called Integrated Nested Laplace Approximations (INLA) offers a faster and more friendly approach for fitting spatial models using spectra as covariates (Poggio et al., 2016; Rue et al., 2009).

1.4 Challenges to model plant traits with hyperspectral data

Modelling plant trains with hyperspectral data involved some challenges given the dimensionality of the spectral domain. For instance, the number of wavelengths available to use as covariates is frequently far larger than the

number of observations for model training (Zhao et al., 2013). Also, the bands are strongly correlated, being highly redundant in the same region of the spectrum when all the observations come from similar land surfaces (Dormann et al., 2013). Uncontrolled factors when capturing hyperspectral signals over sunlight provoke strong random noise in specific regions of the spectrum. All these characteristics increase the risk of spurious correlation that can be mistakenly interpreted as causality when modelling (Milton et al., 2009).

Optical sensors not necessarily capture only the reflectance of the targeted plant trait but also spatial and temporal variations (Milton et al., 2009; Pearse et al., 2016). Therefore, the observations collected *in situ* to represent the study area are not independent and identically distributed (i.i.d.) nor over space nor time (Gotway and Young, 2002; Ingebrigtsen et al., 2014). Spatial and temporal domains are not usually modelled explicitly because of the dimensionality, despite all the recognition of its importance in ecological processes. Modelling with autocorrelated observations with an unsuitable regression can result in unrealistic and non-reproducible results (Ingebrigtsen et al., 2014). Multicollinearity, model overfitting, residuals autocorrelation, lack of generality are some of the commons issues when modelling plant traits with hyperspectral data (Dormann et al., 2013; Hawkins, 2004; Zhang et al., 2005).

1.4.1 Feature selection and multicollinearity

The high dimensionality of hyperspectral data and multicollinearity provoked by the strong autocorrelated wavelengths make the selection of relevant spectral bands a complicated exercise during the modelling process (Curran, 1989). As several wavelengths can be written as linear combinations, it can falsely inflate the importance of a band in the model (Gelman and Hill, 2006; Kuhn and Johnson, 2013; Meehl, 1945). Multicollinearity is magnified when all the spectral signals were captured at similar land cover surfaces (Cho et al., 2007). This is demonstrated in Figure 1.2(b), where spectral data are captured from samples of sand collected in a specific beach location, resulting in extremely correlated bands as the main difference is resumed to the amount of moisture. In this case, it is reasonable the use of only one out of 2100 wavelengths in an empirical model. In Figure 1.2(a), using data from PROSAIL model simulating grassland, the number of bands less correlated than 0.75 were no higher than 3 out of 2100 wavelengths.

The possible solutions for selecting covariates for modelling using hyperspectral data and avoid multicollinearity include: (1) extracting spectral indices that explain causally or empirically the relationship with the target plant trait based on a-priori knowledge; (2) searching a coefficient from a combination of two or more bands that is highest correlated with the plant trait (Darvishzadeh et al., 2008); (3) combining wavelengths to create latent

variables by methods such as wavelets and principal components (Bruce et al., 2002); (4) searching an optimal combination of (non-collinear) wavebands to best explain the plant trait using a method such as stepwise regression or genetic algorithms (Ramoelo et al., 2012; Schlerf et al., 2010); (5) Tuning machine learnings or penalised regressions using the entire hyperspectral set of wavelengths. Some of these approaches are supervised methods (*i.e.* 2, 4 and 5), which select covariates to be included in the model with the support of the response variable (James et al., 2013). Supervised approaches may solve the problem of selecting the variables for the model, but increases the risk of overfitting significantly (Hawkins, 2004). Despite being a supervised method, the second option is performed in a step before modelling and usually stays apart from the assessment of prediction accuracy.

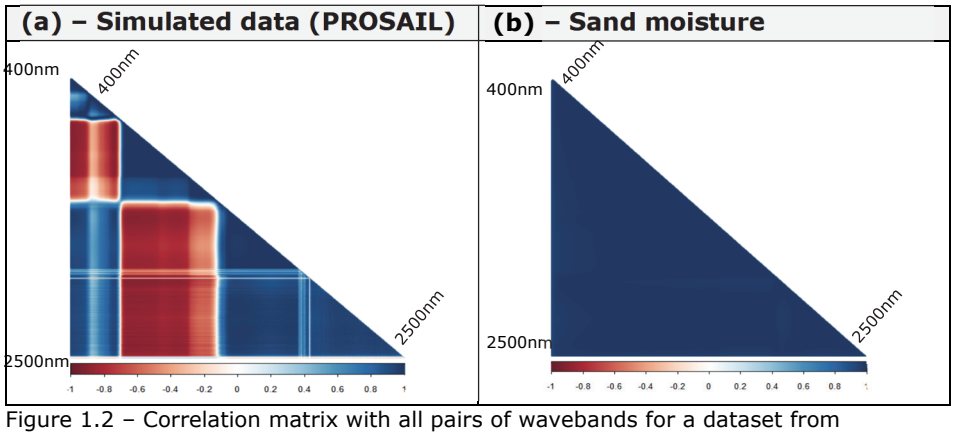


Figure 1.2 – Correlation matrix with all pairs of wavebands for a dataset from simulated grassland using PROSAIL (a–right) and other contain reflectance of beach sand with different amount of moisture (b-left) extracted from Nolet et al. (2014).

1.4.2 Model complexity and Overfitting

The number of terms included in the final selected model determines the complexity (Kuhn and Johnson, 2013). The type and number of terms per wavelength used in a fitted model vary between regression techniques. These terms can be represented by parameter coefficients, interaction, second-order terms, nodes, trees, components and many others (James et al., 2013). Model complexities are not comparable between different model techniques (Hastie et al., 2009). If a large number of wavelengths is searched with the support of the response variable, and later only the most important ones are included as covariates, the final model will still complex yet hidden (Bruce et al., 2002). This procedure may bring similar issues related to model complexity than machine learning, stepwise or other regression which model a supervised model selection. For instance, a simple linear regression using an index (one term) as a covariate, but selected from a combination (two-by-two) of 2100 hyperspectral bands (Roberts et al., 2017). Also, a multiple linear regression

fit by a stepwise procedure may select only two or three bands from the entire (hyper) spectral domain. Both cases the resultant model is mathematically very elementary, despite using (*i.e.* searching) all the bands (Kuhn and Johnson, 2013).

Although more efficient when searching for relevant wavelengths to explain plant trait variations, supervised methods increase the risk of overfitting (James et al., 2013). Overfitting occurs when spurious correlations unrelated to the underlying relationship as random and systematic errors in the data is incorporated into a model (James et al., 2013). In other words, a model may fit the training set quite perfectly, however, present significant lower accuracy when used for estimating in new samples (Gelman et al., 2001; Lee et al., 2004). Overfitting is pruned to occur when a model is overly complex, or a supervised approach was used in the previous stages to select covariates included in the final model. The risk is even higher when modelling with a large set of bands relative to the number of observations, as often the case with hyperspectral data (Hastie et al., 2009). Therefore, models complexity should be constrained or the amount of the bands available to search limited to avoid overfitting (Fassnacht et al., 2014; Kuhn and Johnson, 2013). The process to select model complexity to decrease the risk of overfitting in machine learning is called tuning (Hastie et al., 2009). This process controls the number of parameters or terms in the model, such as "cost" in support vector machine regression (Hastie et al., 2009). The model complexity is selected by fitting models increasing the level of complexity and assessing the accuracy by cross-validation (Krstajic et al., 2014; Verrelst et al., 2012).

1.4.3 Spatial dependency and autocorrelation in the model residuals

As mentioned before, autocorrelation in the wavelengths result in multicollinearity in the model and raises the (type II error) chances of masking important variables (Dormann et al., 2013). As nearby wavelengths tend to be strongly autocorrelated, pixels at close locations are also expected to be (spatial) autocorrelated (Tobler, 1970). Remote sensing imaging or field spectrometers capturing data from a continuous vegetation surface is prune to present significant spatial (and temporal) dependency (Legendre, 1993; Lobo et al., 1998). It is expected from plant traits estimated by remote sensing out of the lab be spatial dependent, disregarding environment targeted, platform, sensor, spatial resolution or extent (Hawkins, 2012; Naimi et al., 2011; Roberts et al., 2017). Spatial autocorrelation violates the assumption of independent and identically distributed observations for many modelling approaches (Dormann et al., 2013; Legendre, 1993; Wikle and Hooten, 2010).

Although spatially dependent plant traits tend to result in spatially correlated observations, this information is often neglected when modelling with hyperspectral data, assuming randomly distributed observations (Babcock et al., 2013; Wikle and Hooten, 2010). If the pattern related to spatial dependency remains in the model residuals, it may indicate biased parameters (Zhang et al., 2005). The spatial autocorrelation increases the chances of Type I error, being the null hypotheses rejected when it is true (Dormann et al., 2007; Fortin et al., 2012; Hawkins, 2012). Adding environmental and topographic covariates in the model, which partially explain the spatial dependency of the plant trait may avoid the presence of autocorrelated residual. However, the lack of these data available or enough knowledge about the underlying processes hardly ever allows it (Fortin et al., 2012). The spatiotemporal structures in remote sensing data may show patterns that are not even causally or empirically related to the target plant trait, such as changes in soil background (Cochrane, 2000). Moreover, spectral, temporal and spatial domains are all serially correlated data, because there is a logical sequence in the data, and nearby pairs of wavelengths, locations or times tend to be more similar than pairs further apart (Tobler, 1970). Model assessment on scientific publications in remote sensing has focused mainly on model fitting and overall accuracy, given little attention to the spatial distribution of model residuals (Moisen and Frescino, 2002; Zhang et al., 2005). The selection of variable under spatial autocorrelation, and its effect on the identification of the best-fitting model still also unclear (Dormann et al., 2007).

1.4.4 Explanatory versus predictive models

Regression models and exploratory statistical analysis can help to understand the variations observed on plant traits in the study area. Although, to fully understand the underlying ecological process a multidisciplinary knowledge about the environment in consideration is needed (Gelman et al., 2001). Understanding the plant processes and functions, including their impacts on ecology should be the aim, rather than modelling and predicting (Ingebrigtsen et al., 2014; Shmueli, 2010). However, it is the most difficult side, which is often ignored or done backwards by empirical associations based on the tuned model. Methods to model plant traits with hyperspectral data (as many others), present uncertainties that limit either the capacity to explain or predict the results (Van Cleemput et al., 2018). Whether or not there is a true (causal) relationship between reflectance and plant trait or a sound explanation of how it occurs, might be meaningless when the aim is a predictive model (Shmueli, 2010). Also, whether the relationship is inversely proportional, non-linear or saturates after a certain value is secondary. These relations are often masked in complex predictive models by many terms and interaction or by data transformation such as latent variables (James et al., 2013; Kuhn and Johnson, 2013).

In which extent a model design to capture small nuances of the data may interfere in the understanding, depend greatly on the complexity of the land surface to be estimated and the experience of the practitioner (Shmueli, 2010). Predictive (empirical) models have the aim to detect associations rather than causation, but even analysing primary data, it is required some knowledge about physical phenomena to avoid risks of misinterpretations of the results (Huber et al., 2008; Stroppiana et al., 2011). Imprecise measurements and complex models contribute to very specific functions that predict accurately only under the completely same conditions, if not only using the same database.

1.4.5 Prediction accuracy and model generalisation

For explanatory models, the assessment of model prediction is recommended but not required (Shmueli, 2010). However, for predictive models, it is mandatory as the selection is performed based on the minimisation of the prediction error, rather than knowledge (Kuhn and Johnson, 2013). There are different assumptions to be checked according to the model approach applied (Gelman et al., 2001). However, prediction accuracy and residual assessment should always be verified (Hastie et al., 2009). For instance, multicollinearity should be tested for ordinary least squares regression using two or more covariates, but it is not applied to machine learning or penalised regression (Dormann et al., 2013). Prediction accuracy should be assessed using an unseen data set and appropriated performance metrics to assess the quality of the fitted model (Cho et al., 2013). Metrics such as the adjusted coefficient of determination (R^2_{adj}), Bayesian Information Criterion (BIC) or Akaikes Information Criterion (AIC), which are based on the assumption of degrees of freedom, are more suitable for explanatory than to predictive models (Kuhn and Johnson, 2013). These methods penalise model complexity, but they are only valid when comparing models fitted under the same regression approach, turning meaningless for machine learnings and penalised regressions (James et al., 2013). Another common way to present model accuracy is calculating the Root Mean Squared Error (RMSE) of the observed versus predicted values of the target plant trait (Gelman et al., 2001).

Predictive models for plant traits are mostly selected by data rather than based on theory, and often elected among different regression techniques (James et al., 2013). If the model is assessed with the same data as was fitted, more complexity, directly means more accuracy, as the prediction error always reduces when the complexity increases (James et al., 2013). Consequently, it is improper to assess and report the accuracy of predictive models with the same data as used for selecting the final model. Predictive models require to split the data into training and testing (sub) sets to assess accuracy (Esbensen and Geladi, 2010). There are many alternatives, from splitting an independent

set manually to making an automatise procedure as repeated cross-validation (Roberts et al., 2017). There are also simulated methods such as bootstrapping that allows using all the original set to fit the model (Brenning, 2012). The method and the proportion of the sample to be spare for assessing the accuracy will depend on data availability, sample design and the heterogeneity of the population to be inferred (Fassnacht et al., 2014; Kuhn and Johnson, 2013).

However, when the number of observations is limited, very common situation when hyperspectral data is used for modelling, most of the data should be allocated to training the model (Hawkins, 2004). Cross-validation is a convenient method to assess model accuracy in this case as it makes multiple randomly splittings of training and testing sets, using all the data for both (James et al., 2013). However, if the estimation of the prediction accuracy from the cross-validation or testing set is significantly smaller than the generated by the training set, the model is considered overfitted, and its complexity should be reduced (Dormann et al., 2013). Although choosing a non-representative testing set or samples coming from a different population can also lead to higher prediction error, overfitting is related to the process of modelling (Hawkins, 2004). For machine learning, cross-validation is broadly used for tuning the model complexity, but there is the risk of overestimating prediction accuracy (Kuhn and Johnson, 2013).

Both testing sets or cross-validation estimations are usually originated from the same field campaign, which may limit the capacity to assess generalisation in a new sample (Kuhn and Johnson, 2013). This may occur as the previous sample might have a different data structure determined by the spatiotemporal sequence which the data were collected (Brenning, 2012; Roberts et al., 2017). Some machine learning regressions can deal well with the autocorrelation from the spectral domain (i.e. multicollinearity), but not necessarily with spatial autocorrelation from plant trait observations or remote sensing data as shown in chapter three. For this reason, it is crucial to assess the model residuals for detecting if there are any pattern but random. The "accuracy rush" is creating specific models, overfitted by an excess of parameters and complexity, lacking in generality and almost meaningless to understand the plant trait underlying process. In the literature, plant traits and species distribution are often considered spatial dependent and correlated to each other. However, this knowledge is rarely used for predicting plant traits with remote sensing. This thesis aims to address modelling issues to predict plant traits using hyperspectral while accounting spatial autocorrelation, which may replace several underlying processes often not available as covariates.

1.5 Research objectives and thesis structure

This thesis focuses on empirical predictive models of plant traits with hyperspectral data, exploring the spectral and spatial domains. The objectives of the thesis can be divided in:

1. To propose a method to deal with multicollinearity, overfitting and feature selection for the most common machine learning methods when modelling highly dimensional hyperspectral data.
2. To evaluate to what extent model prediction, using machine learning methods and linear models, are affected by spatially dependent plant traits.
3. To develop a procedure to explore the spectra-space trade-off when modelling spatially dependent plant traits using hyperspectral remote sensing to improve prediction accuracy.
4. To design a sampling strategy for predicting spatially dependent plant traits at unseen locations with remote sensing data.

The study starts exploring different hyperspectral datasets and traits to demonstrate the effects of the dimensionality and serially correlated wavelengths on the modelling process. Then, random fields of simulated grassland datasets with increasing ranges of spatial autocorrelation were used to test the prediction accuracy of machine learning methods and spatial models under different levels of spatial autocorrelation. A physically-based Radiative Transfer Model (*i.e.* PROSAIL 5B) was used to simulate hyperspectral data as collected by spectrometers in the field for this generated dataset.

This thesis is comprised of six chapters, of which four research chapters are submitted, and three are currently accepted as scientific articles to peer-reviewed ISI journals. The general outline is indicated below.

Chapter 1: the introductory chapter discusses the importance of plant traits and the role of remote sensing to monitoring and understanding the underlying process. The chapter is designed to highlight issues that need further improvement when modelling plant traits with hyperspectral data.

Chapter 2: demonstrates that empirical models using hyperspectral data to predict traits are very likely to lead to significant overfitting, even when selected by commonly used robust cross-validation. A new method named Naïve Overfitting Index Selection (NOIS) was developed to quantify overfitting while selecting model complexity (tuning). The method was tested using five hyperspectral datasets and seven machine learning regression techniques.

Chapter 3: shows that machine learning regressions using hyperspectral data are likely to lead to inaccurate predictions when significant autocorrelation is

observed. These overly complex models are inflated by redundant and noisy spectral bands which result in overestimated prediction accuracies in the presence of spatial structures in the data.

Chapter 4: demonstrates that finding a trade-off between spatial and spectral information when modelling spatially dependent plant traits with hyperspectral data, improves prediction accuracy considerably. A spatially explicitly model with spectral information (expressed by a ratio between two a-priori selected bands) as covariate exhibits higher prediction accuracy compared to machine learning algorithms and linear models when there is significant spatial autocorrelation.

Chapter 5: analyses different sampling designs to predict spatially dependent plant traits with spatial and non-spatial models using hyperspectral data. The design and size of the sampling have a strong influence on the spacing between observations, and therefore, the ability to account or avoid autocorrelation. The sampling design affects the estimation of population parameters or the prediction for unseen locations regardless of the modelling technique applied.

Chapter 6: presents a synthesis of the findings in the previous chapters, connecting the ideas and discusses opportunities for future studies. It brings an overview of the challenges and suggested alternatives for predictive modelling of plant traits using hyperspectral remote sensing data.

Chapter 2

Naïve Overfitting Index Selection (NOIS)[1] - a new method to quantify overfitting and to tune model complexity using hyperspectral data

[1] This chapter is based on: Rocha, A. D.; Groen, T. A.; Skidmore, A. K.; Darvishzadeh, R.; Willemen, L. The Naïve Overfitting Index Selection (NOIS): A new method to optimize model complexity for hyperspectral data. ISPRS J. Photogramm. Remote Sens. 2017, 133, 61–74, doi:10.1016/j.isprsjprs.2017.09.012.

Abstract

The growing number of narrow spectral bands in hyperspectral remote sensing improves the capacity to describe and predict biological processes in ecosystems. But it also poses a challenge to fit empirical models based on such high dimensional data, which often contain correlated and noisy predictors. As sample sizes, to train and validate empirical models, seem not to be increasing at the same rate, overfitting has become a serious concern. Overly complex models lead to overfitting by capturing more than the underlying relationship, and also through fitting random noise in the data. Many regression techniques claim to overcome these problems by using different strategies to constrain complexity, such as limiting the number of terms in the model, by creating latent variables or by shrinking parameter coefficients. This paper is proposing a new method, named Naïve Overfitting Index Selection (NOIS), which makes use of artificially generated spectra, to quantify the relative model overfitting and to select an optimal model complexity supported by the data. The robustness of this new method is assessed by comparing it to a traditional model selection based on cross-validation. The optimal model complexity is determined for seven different regression techniques, such as partial least squares regression, support vector machine, artificial neural network and tree-based regressions using five hyperspectral datasets. The NOIS method selects less complex models, which present accuracies similar to the cross-validation method. The NOIS method reduces the chance of overfitting, thereby avoiding models that present accurate predictions that are only valid for the data used, and too complex to make inferences about the underlying process.

2.1 Introduction

Data collection using in situ measurements is time-consuming and expensive, constraining the availability of information to limited areas and specific periods (Muñoz-Huerta et al., 2013; Plaza et al., 2009; Ramoelo et al., 2013). Remote sensing technologies can mitigate these limitations and provide opportunities to monitor biological processes over wider temporal and spatial scales (Stroppiana et al., 2011; Wilson et al., 2011). The monitoring of biological processes in ecosystems by remote sensing relies mostly on empirical models to predict a variety of biochemical and biophysical properties of vegetation, soil or water (such as nitrogen concentration, organic carbon and biomass stocks), estimated from spectral information (Huber et al., 2008; Kokaly et al., 2009; Nguyen and Lee, 2006; Thiemann and Kaufmann, 2002).

Hyperspectral images present even greater potential, as they consist of many narrow spectral bands that can detect changes in specific regions of the spectrum to which concentrations of such substances or structural characteristics of vegetation can be related (Buitrago Acevedo et al., 2017; Curran, 1989; Darvishzadeh et al., 2011; Hansen and Schjoerring, 2003;

Manolakis et al., 2003). Predictive empirical models face two important challenges when using hyperspectral data, as a result of the high dimensions involved: (1) there is a large number of predictors relative to the number of observations to fit the model (Zhao et al., 2013) and (2) there is strong multicollinearity in the predictors, resulting in highly redundant reflectance values at close spectral distances (Dormann et al. 2013). Multicollinearity is enhanced when the sample originates from a homogeneous land cover type, because similar surfaces result in more similar reflectance values across wavelengths (Cho et al., 2013). High dimensionality and multicollinearity complicate the identification of relevant spectral bands to predict the response variable and the estimation of their regression coefficients, since several explanatory variables can be written as a linear combination of the others (Gelman and Hill, 2006; James et al., 2013; Kuhn and Johnson, 2013). Also, multicollinearity can falsely increase prediction accuracy when a variable that has no correlation with the response but correlates well with another variable that does correlate with the response is used in the model (Meehl, 1945).

There are two main solutions to process high dimensional and multicollinear hyperspectral data with regression models (Stroppiana et al., 2011). Firstly, the number of predictors (bands) can be reduced before fitting an ordinary least squares (OLS) type of model. This can be achieved by selecting a spectral index based on a-priori knowledge, by grouping bands to create latent variables using techniques such as principal components and wavelets (Bioucas-Dias and Nascimento, 2008; Bruce et al., 2002), or by finding an optimal combination of bands using stepwise multiple linear regression or genetic algorithms (Darvishzadeh et al., 2008; Ramoelo et al., 2013; Schlerf et al., 2010). Secondly, models can be fitted using all explanatory variables based on non-ordinary least square techniques (non-OLS). Commonly used non-OLS regressions applied to remote sensing are: dimension reductions such as Partial Least Squares Regression (Carvalho et al., 2013; Martin et al., 2008), tree-based ensembles such as Random Forest or Boosted Regression Trees (Abdel-Rahman et al., 2013; Feilhauer et al., 2015), support vector machine regression (Feilhauer et al., 2015; Mountrakis et al., 2011), and artificial neural networks (Farifteh et al., 2007; Mirzaie et al., 2014; Skidmore, 1997).

Regardless of whether or not there is a true relationship between predictors (spectral bands) and the response variable, using a large set of predictors in relation to the number of observations with a supervised method is likely to cause model overfitting (Hastie et al., 2009). A model may fit the training set almost perfectly, but lead to lower accuracy predictions when applied to new samples or a testing set (Gelman and Hill, 2006; Lee et al., 2004).

Overfitting is the situation where overly complex models capture more than the underlying relationship, and also fit random and systematic errors (noise)

in the data (James et al., 2013). This is even more of a concern in non-OLS regression techniques that use the residuals from a model fitted in a previous step as a new response in a subsequent step (Hastie et al., 2009). Also, predictors derived from hyperspectral data may present a considerable amount of noise in some regions of the spectra, depending on the capacity to control variations in illumination and atmospheric conditions during the measurements (Manolakis et al., 2003).

Therefore, empirical models need to be constrained regarding the number of predictors or parameters included to avoid overfitting. The type and number of terms per predictor used in a fitted model varies between techniques, including parameter coefficients, interaction, second-order terms, nodes, trees, and so on (James et al., 2013). The number of terms used determines the level of model complexity (Hastie et al., 2009). The maximum model complexity to avoid overfitting depends greatly on the number of observations relative to the number of predictors used for fitting the model (Fassnacht et al., 2014; Kuhn and Johnson, 2013). The procedure to select an optimal model complexity that balances the trade-off between accuracy and overfitting is called the tuning process (James et al., 2013). This process is typically performed by adjusting or "tuning" parameters that control the number of terms in the model, such as the "number of components" in partial least squares regression or "cost" in support vector machine regression (Hastie et al., 2009).

The optimal model complexity cannot be calculated directly from the data but can be defined by fitting models with different complexities and evaluating their prediction accuracy (Krstajic et al., 2014; Verrelst et al., 2012). Some metrics to assess model accuracy, such as the adjusted coefficient of determination (R^2_{adj}), Akaikes Information Criterion (AIC), and the Bayesian Information Criterion (BIC) are inappropriate for selecting the best model complexity from different non-OLS regressions as the degrees of freedom are impossible to determine or compare between regression techniques (James et al., 2013). Often the coefficient of determination (R^2) of the simple regression between observed data and model predictions is presented as accuracy metric for non-OLS regressions. Assessing model performance with the same dataset to which it was fitted, greater complexity automatically means higher accuracy because error declines monotonically as complexity increases (James et al., 2013). Therefore, it is inappropriate to use the same dataset to select model complexity and to report the prediction accuracy, requiring a method that separates the data into training and testing (sub) sets (Esbensen and Geladi, 2010). Whether the most suitable splitting of data will be based on approaches such as cross-validation or bootstrapping or even the collection of an independent validation set, will depend on the sample design and data availability (Fassnacht et al., 2014; Kuhn and Johnson, 2013).

Independent validation can be achieved by splitting the existing data into training and testing sets, keeping the validation set apart to quantify the accuracy of each level of model complexity. In this case, the fitted model will be considered overfitted when the accuracy of an independent validation set is significantly lower than the accuracy of the training set (Dormann et al., 2013). Although non-representative samples or samples from different populations can also lead to lower accuracies, overfitting is related exclusively to the process of modelling (Hawkins, 2004). Despite being widely employed, splitting a single dataset into a training and a testing set may only have a limited ability to characterise the uncertainty in the predictions (Kuhn and Johnson, 2013). Model performance can be highly variable depending on the size of the testing set and the variability in the population that was sampled (Darvishzadeh et al., 2008; Kuhn and Johnson, 2013). In addition, when the number of observations is limited, most of them need to be allocated to calibrate the model (Hawkins, 2004). In these cases, cross-validation is an alternative approach to evaluate a model as it randomly splits off multiple combinations of training and validation sets (James et al., 2013).

Cross-validation estimation can produce a reasonable indication of overfitting, and has shown, in general, to be efficient in finding optimal model complexity, giving a satisfactory estimation of the predictive performance (Kuhn and Johnson, 2013). A widely used cross-validation method is the K-fold approach, based on the random splitting of observations into k groups of similar size (James et al., 2013). This procedure can be repeated many times, using a different selection of folds as testing set each time, to increase the robustness (Krstajic et al., 2014). Being widely accepted as tuning method, cross-validation procedures may still select overly complex model in the case of hyperspectral data. Hawkins, (2004) stated that a model overfits when it is more complex than another model that performs equally well. Also, robust cross-validation can be computationally intensive and thus time-consuming for high dimensional data such as hyperspectral datasets, depending on the number of parameters to tune (Hastie et al., 2009; Krstajic et al., 2014). Another limitation is that tuning parameters are often not comparable between different modelling methods and the available methods do not evaluate the adequacy of the model complexity selected from different non-OLS regressions (Kuhn and Johnson, 2013). In addition, cross-validation tuning methods do not quantify the amount of overfitting as the (true) maximum model contribution for a given set of predictors is normally unknown, making it difficult to fairly compare the accuracy of different regression techniques.

The novelty of this study is to present a new tuning method for modelling hyperspectral data that overcomes these limitations of existing techniques. The new method is termed Naïve Overfitting Index Selection (NOIS) and it (1) provides an efficient and structured method to tune over a range of

parameters, showing a gradual increase in model complexity, for non-OLS regressions; (2) determines the maximum level of model complexity supported by a specific data structure without overfitting; and (3) quantifies the relative amount of overfitting across regression techniques consistently, highlighting the trade-off between prediction accuracy and overfitting. The performance of models derived from this tuning method, is compared to a tuning method based on robust cross-validation, and tested using different hyperspectral datasets and regression techniques.

2.2 Methods

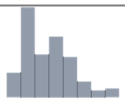
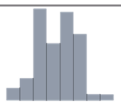
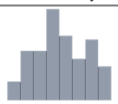
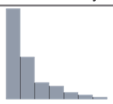
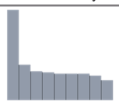
The Naïve Overfitting Index Selection (NOIS) requires three steps. Firstly, a dataset of artificial spectra is generated, having the same data structure as the original spectra, but uncorrelated with the response variable. Secondly, the amount of overfitting at different levels of model complexity is calculated using the generated spectra as predictors. Thirdly, a model complexity is selected based on an overfitting threshold that is compatible with the data structure and comparable between datasets and regression techniques. In this paper, the NOIS method is subsequently compared with a traditional cross-validation tuning method by fitting seven commonly used non-OLS regression techniques to five hyperspectral datasets.

2.2.1 Database

A selection of hyperspectral datasets (Table 2.1) composed of different surfaces and measured using diverse instruments under singular conditions is used to assess the robustness of the NOIS method. These datasets originate from various scientific contexts, representing plausible combinations of number of observations versus number of predictors. These include a dataset with a number of observations higher than the number of spectral bands (*e.g.*, the soil organic carbon dataset), as well as a dataset where the number of observations is considerably smaller than the number of spectral bands (*e.g.*, the leaf water content dataset).

The last column of Table 2.1 indicates the risk of multicollinearity in the model, as in hyperspectral data a large proportion of bands can be considered redundant when a specific surface is measured. For example, if a maximum correlation threshold of 0.75 between any pair of bands is defined as "not being sufficiently different", only a few individual bands will be considered non-redundant in all datasets, implying a strong risk of multicollinearity.

Table 2.1- Description and structure of the five selected datasets used for assessing the new tuning method NOIS.

Data structure	Vegetation traits			Soil traits	
	Leaf Area Index (LAI)	Leaf Chlorophyll Content (LCC)	Leaf Water Content (LWC)	Sand Moisture Content (SMC)	Organic Carbon Content (OCC)
Observations (n)	129	111	108	208	292
Predictors (p)	592	126	6612	2150	216
Wavelength	352-2382nm	436-2485nm	2500–16700nm	350-2500nm	350-2500nm
Instrument	GER3700	Hymap	Bruker Vertex 70	ASD Fieldspec	FieldSpec FR
Type	Field	Airborne	Laboratory	Laboratory	Laboratory
Distribution (Y)					
Redundancy (pair of bands)	ρ<0.75=3 bands ρ<0.90=8 bands	ρ<0.75=3 bands ρ<0.90=5 bands	ρ<0.75=1 band ρ<0.90=3 bands	ρ<0.75=3 bands ρ<0.90=3 bands	ρ<0.75=13 bands ρ<0.90=35 bands
Published by	Darvishzadeh et al. 2008	Darvishzadeh et al. 2011	Buitrago et al. 2016	Nolet et al. 2014	ICRAF-ISRIC Spectral Library

Note: more detailed information about each dataset can be found in the supplementary material (Appendix 2A).

2.2.2 Generating artificial spectral data

A new dataset of predictors with the same dimensions as the original dataset (Table 2.1) is generated from a multivariate normal distribution. This generated dataset preserves the number of bands and has an equivalent mean, variance and covariance to those observed in the original spectra. This procedure intends to create predictors that are completely uncorrelated with the response variable, but maintain the data structure of the original predictors (Figure 2.1). Artificial spectra were generated using the *mvrnorm* function from the MASS package in R version 3.2.5 (Venables and Ripley, 2002), R Core Team 2016). This function requires a vector of means and a positive-definite symmetric covariance matrix extracted from the original spectra. The generated data were rescaled according to the original spectra, preserving the same reflectance range of each band using the function *rescaled* from the package plotrix.

The process of generating spectral datasets gives a good indication of the amount of noise present in the predictors (all generated datasets can be found in Appendix 2B). For instance, the generated spectra for the moisture dataset present all bands as almost completely uncorrelated with the response variable (Appendix 2B), indicating low noise in the data. Because sand samples allow for well-controlled experiments to be conducted in a laboratory, precise measurements could be made for this dataset. Also, only wavelengths between 350 and 2100 nm are included in the analysis, as wavelengths over 2100 nm are considered by the data provider to have a low signal-to-noise ratio (Nolet et al., 2014, p. 201). On the other hand, the LWC dataset contains bands between 2500 and 16700 nm (thermal) and no specific pre-processing in the data has been applied to reduce the noise in the data. A high level of noise in

certain regions of the spectra for this dataset can produce generated predictors that may, by chance, still be slightly correlated with the response variable.

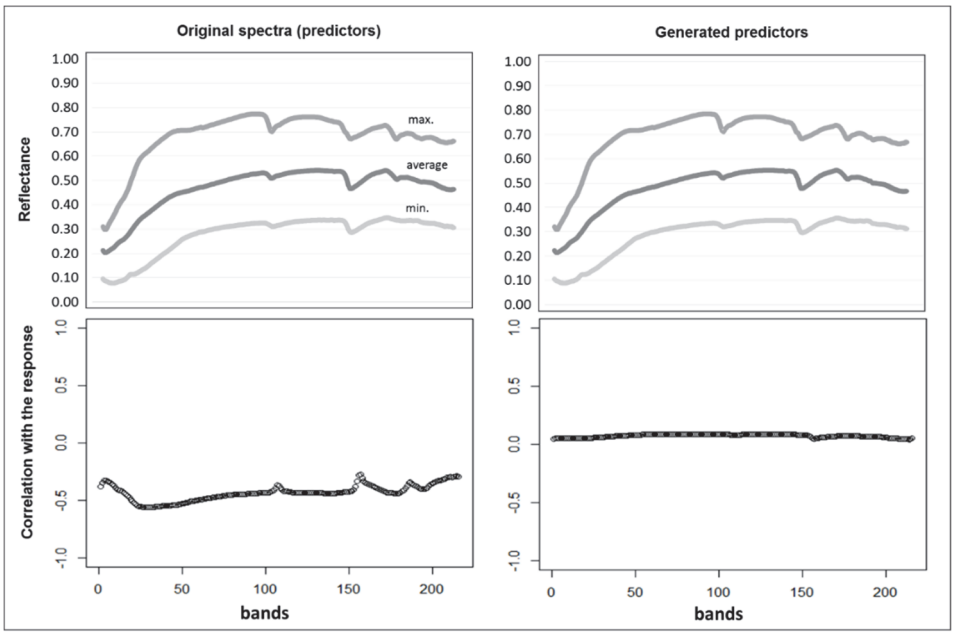


Figure 2.1 - Comparison between original and generated reflectance for the soil dataset. The average (dark grey), maximum (lighter grey) and minimum (light grey) from the original spectra (top left) and generated data (top right). And the correlation between the response variable (OCC) and predictors (bands), using original spectra (bottom left) or generated data (bottom right).

2.2.3 Quantifying overfitting

The generated predictors' dataset (X') preserves the relationship across spectral bands, but makes them uncorrelated (*i.e.*, independent) with the response variable (y). Given that y and X' are independent, the conditional distribution y|X' does not depend on the value of X', E[y|X'] = E[y], and covariance y|X' should approach zero (Cook and Weisberg, 1999). Consequently, the only information available is the mean of response variable, and any model based on generated spectra as explanatory variables will be referred as a naïve model. It implies that the mean square error of a prediction based on X' depends only on the variance of the response variable σ_y^2. Therefore, the naïve models, in theory, should not reduce predictor errors (*i.e.*, $\hat{y}_i = \bar{y}$ and $\hat{\sigma}_y^2 \cong \sigma_y^2$). Consequently, any reduction in prediction error can be attributed to an increase in the model complexity and thus to overfitting.

The amount of overfitting in a naïve model can be quantified by the difference between the prediction error and the true error (*i.e.*, variance of the response

variable), expressed by 1- ($\frac{\hat{\sigma}_y^2}{\sigma_y^2}$). When values of the predictor error ($\hat{\sigma}_y^2$) are significantly lower than the true error (σ_y^2) this will indicate model overfitting. The index will achieve a maximum of 1 when the predictor error approaches zero ($\hat{\sigma}_y^2 \to 0$). In case of no overfitting, σ_y^2 and $\hat{\sigma}_y^2$ should be equal and the overfitting index will approach 0.

2.2.4 Naïve Overfitting Index Selection (NOIS)

Since variance or mean square errors depend on the response variable range (y), the model accuracy was reported as Root Mean Square Error normalised by the range of the response variable (NRMSE). The naïve overfitting index produced by a specific level of complexity is also calculated based on NRMSE.

naïve overfitting index = 1 – ($\frac{\text{NRMSEg}}{\text{NRMSEy}}$), where:

NRMSEg is the error based on the prediction derived from the naïve model using the generated data (X'), and NRMSEy is the error based on the prediction derived from the mean of the response variable (y).

For instance, a naïve overfitting index of 0.75 indicates that the true error is falsely reduced 75% by this level of model complexity. In this case, the model complexity should be significantly constrained or the number of observations considerably increased. Negative index values indicate that the model predicts a bigger error than NRMSEy, and the model complexity is constrained excessively ("underfitted"). Because the NRMSEy is only based on the response variable (y), and no model contribution is expected from naïve models, the degree of overfitting is directly comparable between regression techniques.

2.2.5 Selecting model complexity

The optimal model complexity supported by the data is selected by increasing tuning parameter values until the naïve overfitting index drops below a pre-defined tolerance (Figure 2.2). This tolerance, expressed as a percentage of the NRMSEy, can be adjusted to avoid selecting underfitted models, where the level of complexity is excessively constrained. The tolerance is set at 0.05 in this study, based on the maximum correlation between the response variable (y) and the artificially generated spectra (see Appendix 2B).

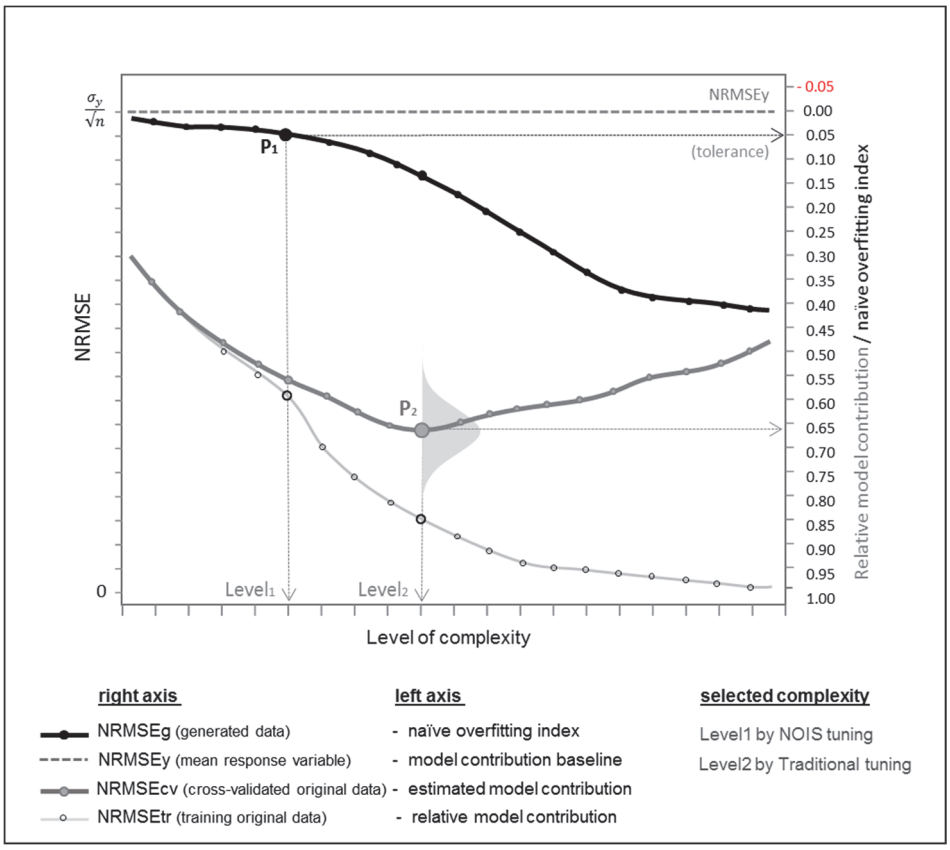


Figure 2.2 - Process to select the level of model complexity using the NOIS method and the traditional cross-validation tuning. The point P1 represents the level of complexity, where the value of the naïve overfitting index over the NRMSEg curve (generated data) approaches 0.05 (tolerance). The point P2 represents the level of complexity selected by the traditional method (based on original data), where the maximum model contribution is achieved based on minimisation of the NRMSE estimate from the cross-validation (NRMSEcv curve).

In some of the regression techniques there is more than one tuning parameter to define model complexity, requiring a repeat of the procedure for each parameter, whilst keeping other tuning parameters fixed. Because naïve models are trained and selected using generated artificial spectra, the accuracy can be assessed using the full dataset with original predictors, as opposed to the traditional cross-validation method, which requires multiple splitting of training and validation subsets. Thus, compared to traditional cross-validation, the NOIS method avoids uncertainty in the estimation of prediction errors. The naïve overfitting index is defined as the relative model contribution when using generated data (naïve model) and provides an indication of the amount of overfitting for a given level of model complexity.

2.2.6 Comparison with a traditional 'tuning' method

The NOIS method is compared with a tuning procedure using traditional cross-validation to test its consistency and reliability in the tuning process (Figure 2.3). A 10-fold cross-validation is adopted to evaluate the performance of each level of model complexity with the original spectra as predictors. This procedure is randomly repeated ten times, resulting in a combination of 100 subsets of training and validating sets of the original data. The model tuning by means of traditional cross-validation is based on minimization of the cross-validated prediction error (NRMSEcv).

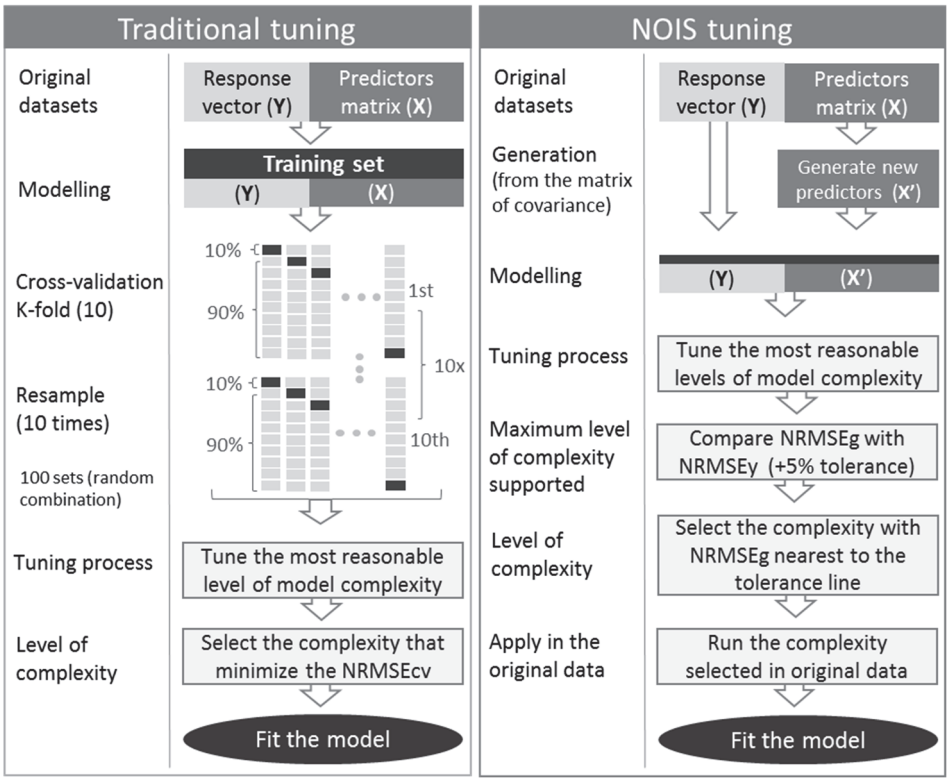


Figure 2.3 - Comparison between the proposed NOIS method and a traditional approach of cross-validation

The same approach to calculate the naïve overfitting index can be used with the traditional cross-validated tuning method to represent the relative model contribution, by replacing the NRMSEg from the naïve model (generated predictors) by the NRMSEcv estimate from the model fitted on the original data. However, as the true model contribution is unknown in this case, the model contribution may be confused with overfitting. Also, the prediction error

estimated by cross-validation is based on an average (see Figure 2.2 - P2), and the model contribution may vary significantly between sub-models.

2.2.7 Regression techniques tested

Common regression techniques for modelling hyperspectral data are used to compare the NOIS method with a traditional cross-validation method (Table 2.2). These regression techniques are often selected because they are considered to be reasonably robust regarding highly dimensional data and high multicollinearity (Kuhn and Johnson, 2013; Zhao et al., 2013).

Table 2.2 - List of regression techniques tested, R packages and functions to fit the model, and tuning parameters used for defining model complexity.

Type	**Regression Technique**	**R package (function)**	**Tuning parameters to define model complexity**
Regression Trees (ensemble)	Random Forest	randomForest (rf)	mtry (number of randomly selected predictors); maxnode (max number of terminal nodes trees)
	Boosted Trees	gbm (gbm)	n.trees (number of interactions); interaction.depth (max. of variable interactions); shrinkage (learning rate); n.minobsinnode (min. terminal node size)
Artificial Neural Network	Stuttgart Neural Network Simulator	RSNNS (mlp)	size (number of hidden units); max (no. max. of interactions); decay (weight decay, shrinkage or learning rate)
Dimension Reduction	Partial Least Squares	pls (pls)	ncomp (number of components)
Vector Machines	Support Vector Machines	e1071 (svmLinear)	cost (cost); epsilon
Penalized, Shrinkage model	The Lasso	elasticnet (lasso)	fraction (fraction of full solution - shrinkage)
	Ridge	elasticnet (ridge)	lambda (weight decay - shrinkage)

All regression methods are executed in R version 3.2.2 (The R Foundation for Statistical Computing). The package Caret (Classification and Regression Training) is used for fitting models from different regression techniques with cross-validation under the same platform (all the packages are presented in Table 2.2). The value of all selected tuning parameters for each regression technique and dataset are presented in Appendix 2C. The response and explanatory variables in each dataset are mean centred and scaled by standard deviation before fitting the models to increase comparability across techniques and datasets (Kuhn, 2008).

2.3 Results

2.3.1 Selecting model complexity

The NOIS method, in most cases, identifies lower levels of complexity as suitable than the traditional tuning process using cross-validation does (Figure 2.4). Datasets with a higher number of observations (n) in relation to the number of predictors (p) support greater model complexities (Burket, 1943; Hastie et al., 2009). This principle becomes quite clear when the model complexity is selected by the NOIS method, but is less evident when the traditional cross-validation is used. For example, the organic C dataset (grey line in Figure 2.4) is the dataset with the highest n/p ratio, namely 292 observations for 216 bands. This dataset also shows the lowest overfitting in almost all the regression techniques. Random Forest models form an exception with similar levels of overfitting occurring regardless of the differences in n/p ratios. In contrast, LWC has the lowest n/p ratio, *i.e.*,108 observations for 6612 bands, and shows overfitting at relatively low levels of model complexity.

The two tuning methods suggest similar levels of complexity for the LCC dataset for all regression methods. The Support Vector Machine tuning for the LCC dataset generated the only instance where the traditional tuning method selected a lower level of model complexity than the NOIS method did. The reason for this is that the LCC dataset has the smallest number of predictors, which are also the least correlated with the response. As well, the generated spectra present a low level of noise to fit in this dataset (see Appendix 2B). This leads to select models with low levels of complexity in both methods.

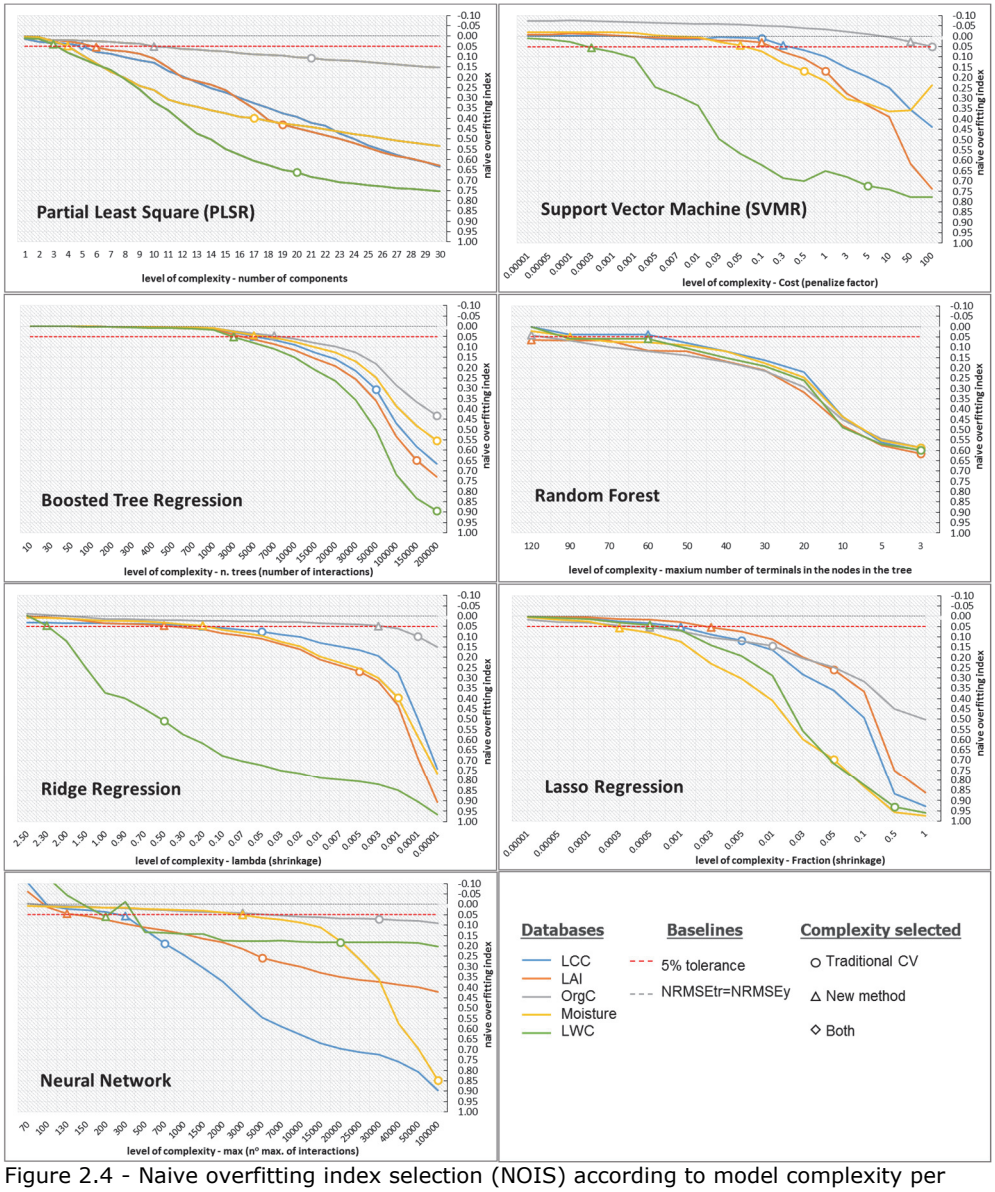


Figure 2.4 - Naive overfitting index selection (NOIS) according to model complexity per regression technique.

Note: The range of tuning parameters commonly suggested by software guides or machine learning literature seems unsuitable for the high dimensional hyperspectral data used in this study, and more constrained tuning parameters used to reduce complexity are needed to avoid overfitting. See for example James et al. (2013) and Kuhn and Johnson, (2013) or https://cran.r-project.org/ for suggested ranges of tuning parameters.

The different tuning approaches result in different levels of overfitting. For example, the PLSR model fitted to the LWC dataset is constrained to a maximum complexity of 3 components (tuning parameter of PLSR) at a naïve overfitting index of 0.04 when the NOIS method is used. On the other hand, the traditional method selects up to 20 components, at a naïve overfitting index of 0.66. Whereas the new method selected a model complexity that, when applied to the original spectra, presents a model contribution of 33%, the traditional method selected a model complexity that presents a model contribution of 48%. The model contribution suggested by cross-validation is only slightly higher than the one indicated by the new method, but has a much higher level of complexity (100,000 more parameter terms in the model). This complexity selected by cross-validation is large enough to present a model contribution of 99% for the training model (NRMSEtr=0.0037).

Some researchers suggest selecting a smaller model complexity if increasing it does not decrease the error by at least 2% (Kooistra et al. 2004, Darvishzadeh et al. 2011). In the case of PLSR, selecting a model complexity by the traditional tuning method, with this criterion, 9 rather than 20 components would be selected for optimal complexity. Although less overfitted, this still presents an NRMSEtr more than twice as small as NRMSEcv, and a level of complexity sufficient to reduce the NRMSEtr one quarter of the NRMSEy in the generated data (with a naïve overfitting index of 0.26). Also, this 2% rule is easily applied in PLSR where there is only one discrete tuning parameter, but is less applicable in many other regression techniques that present two or more non-discrete tuning parameters for selection.

2.3.2 Quantifying overfitting and model contribution

Figure 2.5 presents the cross-validated error (NRMSEcv) for models fitted to the original spectra with a level of complexity tuned by the NOIS method and the traditional method for all regression techniques. The boxplot shows the variability in NRMSEcv among the sub-models' performance by the repeated k-fold cross-validation. The results, when presented in ascending order of dimensionality (n versus p), indicate that by increasing the number of predictors relative to the observations, the distance between NRMSEtr and NRMSEcv in the traditional method increases considerably. On the other hand, the new method results in NRMSEtr values that are very similar to the NRMSEcv values. The amount of overfitting (bars on Figure 2.5) for the model complexities selected by the new method are all controlled at a tolerance around 0.05.

The model complexities selected by the traditional method present much higher levels of overfitting for a number of scenarios. In the most extreme case, a naïve overfitting index of around 0.90 was found using traditional

tuning (Random forest applied to the LWC dataset), suggesting that the error can be reduced to 10% with non-informative predictors by selecting an overly complex model. The results indicate that the new method selects models that are less likely to be overfitted, while in most cases showing similar accuracy.

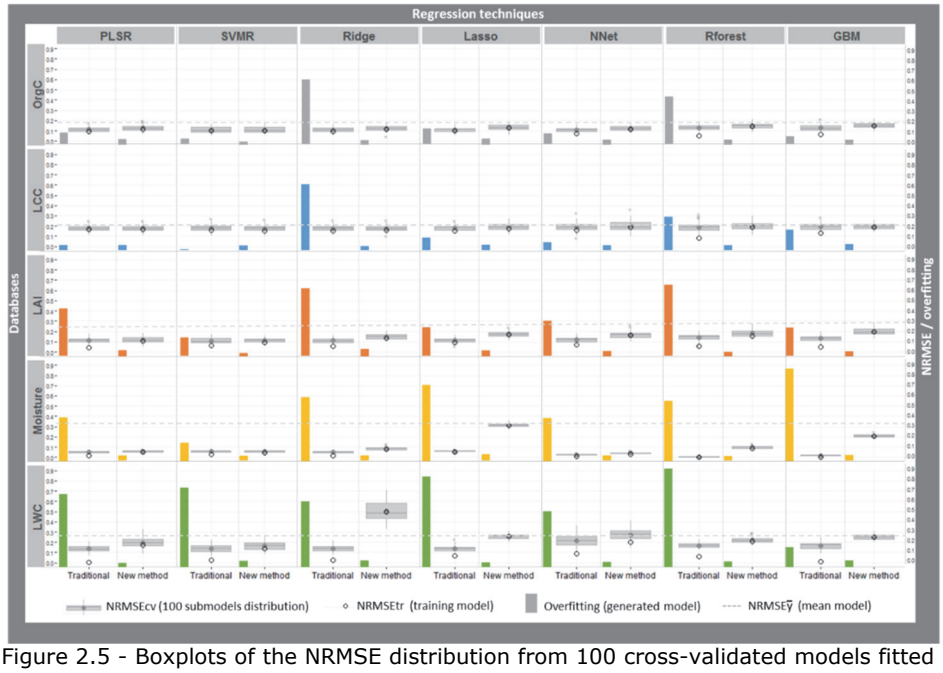


Figure 2.5 - Boxplots of the NRMSE distribution from 100 cross-validated models fitted on the original bands with a model complexity selected by the traditional and NOIS method. The bars represent the naive overfitting index of the model complexity selected. The circles indicate the NRMSEtr using all the observations.

2.3.3 Comparison between methods

PLSR and SVMR models show only small differences in performance between the levels of complexity selected by the traditional and the new method. These regression techniques also present results that are more consistent across different data structures and different capacities of explaining the response by the predictors. The distribution of NRMSEcv between models selected in both methods is mostly similar, yet the level of complexity for the NOIS method is usually significantly smaller than for the traditional method. While the model selected by the NOIS method presents a single value of prediction error, the cross-validation procedure presents an average of hundred combinations of training and validation sets. The more random noise there is present in the original spectral signal, the more uncertainty is presented in the cross-validation estimation. This is noticeable when comparing the variability of the cross-validation estimates between the Moisture and LWC datasets (Figure 2.5). This can also be derived from the higher capacity to generate artificial

predictors that are uncorrelated with the response variables in the first step of the NOIS method (see Appendix 2B).

Taking the LWC original dataset as an example, the relative model contribution of the selected models for different regressions is between 0.66 and 0.99 for the training models, while the model contribution from cross-validation is between 0.16 and 0.49. Such differences may be due to the smallest model contribution coming from an underfitted model (NRMSEcv= 0.218) and the highest from a highly overfitted model (NRMSEtr= 0.004). Based on these results, it is difficult to decide what the most reasonable estimation of accuracy is given the available predictors to explain the response variable LWC. However, concluding that choosing a model with a complexity that minimises NRMSEcv does not guarantee generalizable non-OLS model predictions. In the proposed NOIS method, the model is selected by the maximum complexity that is supported by the data structure without overfitting, and the accuracy is a single calculated value for that particular model using the original data.

Another limitation of the traditional method is the effect of intensive cross-validation, resulting in a low capacity to indicate overfitting in complex models when the number of observations is insufficient. Figure 2.6 presents the difference between cross-validation error estimates (NRMSEcv) and training model errors (NRMSEtr) for the tuning process of PLSR as an example using the traditional method.

As observed in the LCC plot (Figure 2.6e) the NRMSEcv decreases to a certain level of complexity, after which it starts to increase, while the NRMSEtr further decreases. The optimal complexity for this dataset occurs at a point where the difference between the NRNSEtr and NRMSEcv is not too great (Hastie et al. 2009, Schlerf and Atzberger, 2006). This, however, is not observed in datasets where the number of observations is much smaller than the number of predictors, such as for LWC and Moisture. After the fourth component, NRMSEtr and NRMSEcv start to bifurcate in the LWC dataset (Figure 2.6b). While NRMSEtr reduces to approximately zero when twenty or more components are included, NRMSEcv remains at an almost steady value for nine or more components. The gap between NRMSEtr and NRMSEcv demonstrates a clearly overfitted model that presents a complexity higher than supported by the data available. This complexity produces an NRMSEtr value of 0.004 for LWC, near to the nominal precision of the instruments used for measuring the spectra (photometric accuracy of 0.1% T - VERTEX 70 Spectrometer) and is probably more precise than the capacity to determine the true leaf water content values. Nevertheless, this clearly overfitted model would still be selected when using minimisation of the cross-validation error (NRMSEcv) as a tuning method.

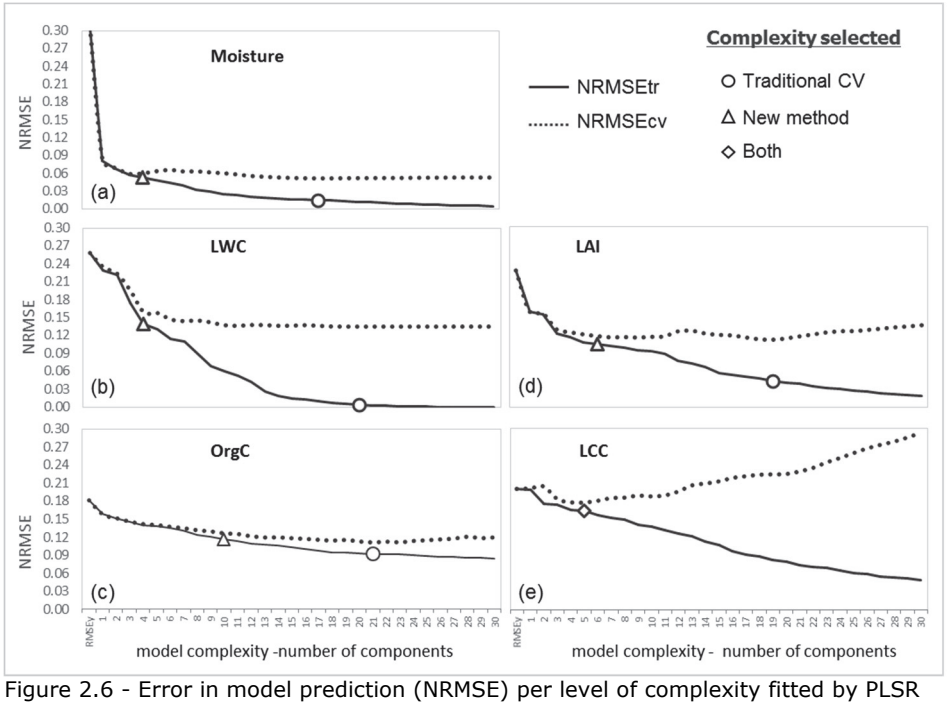


Figure 2.6 - Error in model prediction (NRMSE) per level of complexity fitted by PLSR using the traditional tuning method (original data). Solid lines represent training models (NRMSEtr) and dotted lines are cross-validation estimates (NRMSEcv).

2.4 Discussion

The naïve overfitting index selection as presented has a number of advantages. Firstly, it is based on a single error estimation (*i.e.*, NRMSEg). Secondly, it uses all the available observations to calibrate the model ensuring that no degree of freedom is lost in the tuning process. Thirdly, a comparison between different regression techniques is more reliable as the amount of overfitting can be quantified and controlled. This comparison indicates that the maximum level of complexity supported by a model before overfitting depends greatly on the data structure. Especially the number of observations (n) versus number of predictors (p), the degree of multicollinearity, and the amount of random noise in the data can increase the risk of overfitting considerably. Fourthly, model complexity is hard to standardise across regression techniques, but now the amount of overfitting can be estimated by the naïve overfitting index in a comparable way.

Traditional tuning based on cross-validation does not indicate whether the level of model complexity is appropriate for the data under consideration. The NOIS method allows more control and understanding about the effects of the model complexity in the trade-off between accuracy and overfitting. Finally, robust cross-validation can be time-consuming, requiring intensive computing time

for high dimensional data such as hyperspectral measurements. The NOIS method is considerably faster, especially for regression techniques that require tuning across a large range of parameters. Cross-validation only needs to be performed for the selected level of complexity to assess the final model accuracy.

2.4.1 Trade-off accuracy and overfitting

Some machine learning algorithms were initially designed for classification, such as the ones based on regression trees (Random Forest and GBM). Such methods normally produce training models with significantly higher accuracy than the validation models. Thus, these techniques will hardly ever present similar accuracy in the training and validation sets for models using continuous response variables, regardless of the tuning method applied. Also, Dormann et al. (2013) concluded that Random Forests consistently overfit, without there being an obvious solution to correct this. In prediction problems, it is desirable to fit models from a given sample in such a way that the most accurate predictions are produced, also when applied to other samples from the same population (Burket, 1943). However, building complex models with high dimensional data with techniques that learn from the information in the model residues can reduce the reproducibility of the prediction accuracy considerably for future samples from the same population (Kuhn and Johnson, 2013). Accuracy metrics used for model selection in non-OLS regression techniques do not take into account the lack of parsimony as common in ordinary least square regressions. The new tuning method overcomes this problem by identifying for each regression technique the maximum model complexity that is supported by the given data structure.

Seeking accurate models by minimising the prediction error has to be weighed against the risk of overfitting and producing unrealistically small errors. At times, complex models fictitiously perform better than the accuracy of the measuring system used for collecting the set of spectral signals, chemical concentrations or structural components. The random error in measurements or situations when relevant predictors are missing in the model should not be mistaken for lack of fitness (underfitting) and be a reason to increase model complexity. Predictors derived from hyperspectral data cannot be considered independent because the reflectance is measured by the same instrument, at the same time, and from nearby wavelengths (Curran, 1989). These characteristics are generally undesirable for modelling, as predictors that are not independent of each other tend to cause serious problems of multicollinearity. However, these characteristics also provide the opportunity to generate artificial spectra using the covariance matrix in such a way that the data structure is replicated, but the result is not correlated with the response variable. So, our proposed method uses these properties of

hyperspectral data to present an intuitive tuning process that permits understanding of the trade-off between accuracy and overfitting for the selected model complexity.

2.4.2 Limitations and precautions

Our proposed method is built on the assumption that the modelling algorithm conducts the same procedure for the original and for the artificially generated predictors. This is not the case for regression techniques that present an internal mechanism of feature selection for explanatory variables. Such techniques (*e.g.*, Lasso and GBM) may actually present different levels of model complexity for the same value of a tuning parameter (Hastie et al., 2009; James et al., 2013; Kuhn and Johnson, 2013). This can be seen, for example, in the lasso regression of the moisture dataset. This model, when tuned with cross-validation retains 341 out of the 2150 predictors available (others have their coefficients shrunk to zero). However, with the NOIS method it retains 571 predictors.

The process to generate artificial predictors may result in a dataset slightly correlated with the response variable when the number of predictors is extremely large and noisy. In this case, when the complexity is constrained to a level that presents no model contribution, the NOIS tuning may select an underfitted model. This is the case for the LWC dataset (see Appendix 2B), and can be seen distinctly in the ridge regression were the model coefficients were shrunken excessively resulting in an error higher than the RMSEy (Figure 2.6). In this case, the remaining correlation from the generated data can overtake the tolerance of 5%, and a higher threshold should be defined to accept more model contribution. A pre-processing filter to smooth the original spectral signal to reduce the noise before generating the artificial predictors could be applied in such cases. As this study aimed to compare the new method for different data structures, no extra pre-processing was applied on the (original) spectra and the tolerance was kept constant, despite the risk of selecting underfitted models for a particular dataset.

2.5 Conclusion

Hyperspectral data provide opportunities to monitor biological processes and structure in a natural environment over wider temporal and spatial scales. However, as demonstrated in this study, empirical models using high dimensional hyperspectral data as predictors are very likely to cause model overfitting. The traditional tuning methods fail to precisely determine the maximum level of complexity that is warranted by the used data. These methods are also unable to estimate the amount of overfitting expected given a selected model complexity. The NOIS method presented here, overcomes these problems by quantifying the relative amount of overfitting and by

selecting an optimal model complexity supported by the data. The new tuning method consistently selects a less complex model and is thus less susceptible to overfitting, while the model performance is similar to the ones selected by the traditional tuning method. The NOIS method increases the chances of fitting more generalizable models from hyperspectral data, avoiding models that perform accurately only on the data that they were trained with.

Appendix 2A

Datasets

LAI and GER3700 canopy spectra
Description extracted from Darvishzadeh et al. (2008).

Leaf Area Index – LAI
Non-destructive measurements of leaf area index were taken using a Plant Canopy Analyzer (LAI-2000), an instrument produced by LICOR Inc. (Lincoln, NE USA). The measurements were taken under clear sky conditions, with a low solar elevation, and without direct sunlight reaching the sensor. Five bellow-canopy samples and a reference above-canopy radiation were collected to represent the, average, LAI.

GER3700 canopy spectra
The canopy spectra measurements were captured in the field from June 15 to July 15 in 2005 by the spectroradiometer GER3700 (Geophysical and Environmental Research Corporation, Buffalo, New York). The wavelength range was between 350nm to 2500nm with a spectral resolution of 3nm to 16nm. Measurements were collected on clear sunny days between 11:30 and 14:00 to reduce atmospheric perturbations and BRDF effects. The sensor captured a base area about 45cm in diameter. Up to 15 measurements per plot (1m x 1m) were recorded, changing the position slightly to represent the plot area. The average of the measures was used in order to reduce noise.

LCC and Hymap Image
Description extracted from Darvishzadeh et al. (2011).

Leaf Chlorophyll Content - LCC
Leaf chlorophyll content (LCC) was measured in the field by the instrument SPAD-502 Leaf Chlorophyll Meter (Minolta, Inc.). SPAD values are unitless measurements based on the transmittance in red (650nm) and NIR (920nm) wavelength regions. Many studies have demonstrated that these values are highly correlated with the leaf chlorophyll concentrations derived from chemical processes. A total of 30 leaves of main, dominant species were measured and averaged to represent the LCC in each plot.

Hymap Image
Hyperspectral airborne HyMap sensor data were acquired over the study area on 4 July 2005. The sensor contained 126 spectral channels in a wavelength range of 436nm to 2485nm with a spectral resolution of between 13nm and 17nm and a spatial resolution of 4m.

LWC and FTIR spectrometer
Description extracted from Buitrago et al. (2016)

Leaf Water Content (LWC)
LWC was destructively measured at each stage of the experiment using leaves from the same cohort as the marked leaves, which were used for the spectral measurements. The relative gravimetric LWC was calculated using the equation: LWC = 100 ∗ (Ww − Wd)/Ww, where Ww is the weight of the fresh leaf, and Wd is the weight of the dried leaf. Leaves were dried in an oven at 65 °C. Cuticle thickness was measured from a thin transverse section of the marked leaves, using a Leitz Wetzlar microscope, with an amplification of 250×. This trait was measured at least 3 times in each leaf and the measurements averaged and expressed in µm.

Leaf spectral
All plants were measured with a Bruker Vertex 70 FTIR spectrometer, adapted with an external integrating sphere. An Infragold plate with known spectral emissivity was used to calibrate each measurement. Spectra were measured in the range 4000–600 cm−1 (2.5–16.7 µm) with a resolution of 4 cm−1. Per leaf eight samples, with 520 scans per sample, were taken. These measurements were averaged and the results were calculated per leaf. Five leaves per plant were measured in the same way for a total of 75 leaves per treatment at every stage of the experiment.

Moisture and laboratory spectroscopy
Description extracted from Nolet et al. (2014). Data are publicly accessible at doi:10.4121/uuid:866135c2-2be3-4b74-8f9c-922505285a7b.

Moisture
A representative sample of beach sand was collected from the 'Sand Motor' (GPS location: 52.0520N 4.1840E). Before the experiment, the sample was coarsely sieved (2 mm) to remove shells and constituents other than sand. The sand, composed of quartz with some feldspar, had a dry bulk density rb of 1.655 gcm. For each experiment, a sub-sample of the collected beach sand was placed in a matte black petridish (5 cm radius, 1.5 cm height), filling it up to the rim, and oven-dried for 24 hours at 105ºC. The sample was, after measuring its initial weight, slowly saturated with distilled water. The water was allowed to distribute itself uniformly throughout the sample and excess free water was drained from the surface. The sample was placed on a data-logging weighing scale with milligram precision.

Laboratory spectroscopy
A laboratory spectroscopy experiment was conducted twice to observe spectral reflectance in the optical domain (350– 2500 nm) under different moisture

conditions. The spectral reflectance was measured at 1 nm intervals using an ASD Fieldspec Pro spectrometer (Analytical Spectra Devices, Boulder, CO). A 40640 cm white Spectralon panel (LabSphere, Inc., North Sutton, NH) was used to calibrate the spectrometer. The spectrometer was fitted with a 10 FOV foreoptic which was directed at nadir at 40 cm distance from the sample. As an artificial light source, a 900 watt Quartz Tungsten Halogen (QTH) lamp was placed 70 cm from the sample at a 300 zenith angle. The spectrometer was programmed to take a measurement every 5 minutes. Each time the weight of the sample was also measured and stored.

Soil OrgC and VNIR Spectral Library

Data from: World Agroforestry Centre (ICRAF) and ISRIC - World Soil Information. 2010. ICRAF-ISRIC Soil VNIR Spectral Library. Nairobi, Kenya: World Agroforestry Centre (ICRAF). Available at http://africasoils.net/.

This spectral library consists of visible near-infrared spectra of 785 soil profiles (4,437 samples) selected from the Soil Information System (ISIS) of the International Soil Reference and Information Centre (ISRIC). The samples are all physically archived at ISRIC that soil attribute data were available in 2004.

OrgC

Soil samples were air-dried, clods crushed and the resulting sample material sieved through a 2 mm sieve prior to further analysis. Organic carbon content was determined using the Walkley-3 Black procedure. This involves wet combustion of the organic matter with a mixture of potassium dichromate and sulfuric acid at about 125°C. Soil property attributes were provided by ISRIC and had been analysed according to the ISRIC "Procedures for soil analysis" (Van Reeuwijk, 2002).

VNIR Spectral

Soil diffuse reflectance spectra were recorded for each library sample using a FieldSpec FR spectroradiometer (Analytical Spectral Devices, Boulder, CO) at wavelengths from 0.35 to 2.5 m with a spectral sampling interval of 1 nm. Samples were illuminated from below using a high-intensity source probe. About 20 g of air-dried soil, ground to pass through a 2-mm sieve, was placed into 7.4 cm diameter Duran glass Petri dishes to give a sample height of about 1 cm. To sample within-dish variation, reflectance spectra were recorded at two positions, successively rotating the sample dish through 90° between readings and an average of 25 spectra was recorded at each position to minimize instrument noise. Before reading each sample 10 white reference spectra were recorded using calibrated spectralon (Labsphere, Sutton, NH, USA) placed in a glass petri dish. Reflectance readings for each wavelength band were expressed relative to the average of the white reference readings. The 1 nm interval spectra were resampled by selecting every tenth-nanometer value from 0.35 to 2.5 μm to give a total of 216 data points for each spectrum.

Appendix 2B

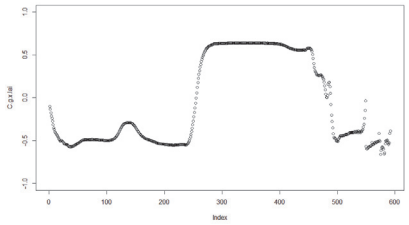
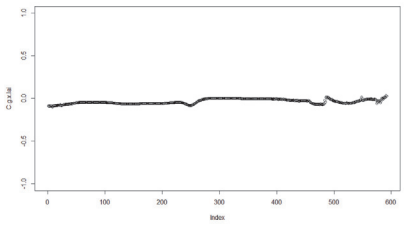
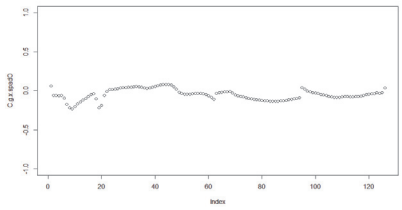
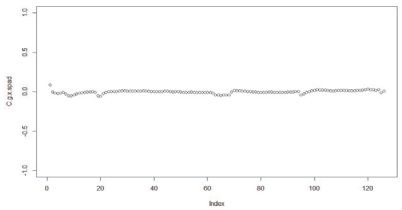
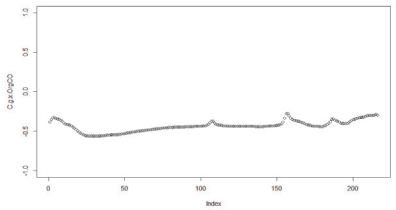
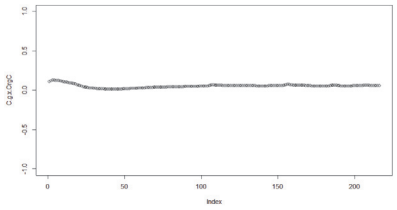
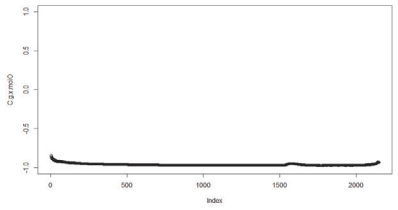
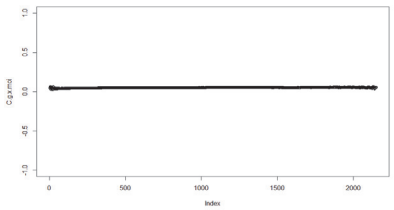
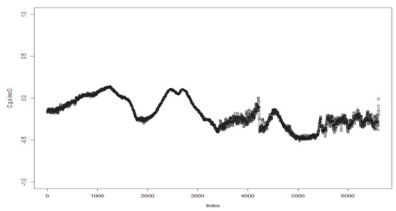
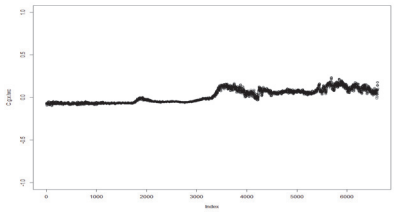
	Correlation with original spectra	Correlation with generated spectra
LAI	 Min -0.662 / Mean -0.065 / Max 0.642	 Min -0.100 / Mean -0.035 / Max 0.035
LCC	 Min -0.231 / Mean -0.048 / Max 0.083	 Min -0.056 / Mean -0.002 / Max 0.086
OrgC	 Min -0.561 / Mean -0.434/Max -0.272	 Min 0.016 / Mean 0.056 / Max 0.130
Moisture	 Min -0.975 /Mean -0.964 /Max -0.844	 Min 0.024 / Mean 0.0511 / Max 0.067
LWC	 Min -0.514 / Mean -0.181 /Max 0.145	 Min -0.101 / Mean 0.015 / Max 0.235

Figure 2.7 - Original and generated spectra for all the datasets. The average, maximum and minimum correlation between the wavelengths with the response variable.

Appendix 2C

Tuning parameters

Regression technique	Tuning parameter	OrgC		LCC		LAI		Moisture		LWC	
		NOIS	CV	NOIS	CV	NOIS	CV	NOIS	CV	NOIS	CV
PLSR	ncomp	10	21	5	5	6	19	4	17	3	20
SVMR	cost	50	100	0.3	0.1	0.1	1	0.05	0.5	0.0003	5
	epsilon	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Rforest	nodesize	120	3	60	3	120	3	90	3	60	3
	mtry	50	50	5	5	10	10	10	10	10	10
	ntree	200	200	200	200	200	200	200	200	200	200
Lasso	fraction	0.0005	0.01	0.001	0.005	0.003	0.05	0.0003	0.03	0.0005	0.5
Ridge	lambda	0.003	0.0001	0.2	0.05	0.5	0.003	0.2	0.001	2.3	0.01
GBM	n.trees	7000	200000	5000	50000	3000	150000	5000	200000	3000	200000
	i.depth	2	2	2	2	2	2	2	2	2	2
	shrinkage	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
	n.minnode	10	10	10	10	10	10	10	10	10	10
Nnet	m.interaction	3000	30000	300	700	130	5000	3000	100000	200	20000
	hidden units	4	4	4	4	6	6	3	3	4	4
	learn rate	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.005	0.005	0.0001	0.0001

Table 2.3 - Tuning parameters selected by the NOIS method and the traditional cross-validation per database and regression technique.

Chapter 3

Machine learning predicting plant traits under spatial dependency using hyperspectral data are unreliable[2]

[2] This chapter is based on: Rocha, A., Groen, T., Skidmore, A., Darvishzadeh, R., Willemen, L., 2018. Machine Learning Using Hyperspectral Data Inaccurately Predicts Plant Traits Under Spatial Dependency. Remote Sens. 10, 1263. https://doi.org/10.3390/rs10081263

Abstract

Spectral, temporal and spatial domains are difficult to model together when predicting in situ plant traits from remote sensing data. Therefore, machine learning algorithms solely based on spectral domains are often used as predictors, even when there is a strong effect of spatial or temporal autocorrelation in the data. A significant reduction in prediction accuracy is expected when algorithms are trained using a sequence in space or time that is unlikely to be observed again. The ensuing inability to generalise creates a necessity for ground references for every new area or period, provoking the propagation of "single-use" models. This study assesses the impact of spatial autocorrelation on the generalisation of plant trait models predicted with hyperspectral data. Leaf Area Index (LAI) data generated at increasing levels of spatial dependency are used to simulate hyperspectral data using Radiative Transfer Models. Machine learning regressions to predict LAI at different levels of spatial dependency are then tuned (determining the optimum model complexity) using cross-validation as well as the NOIS method. The results show that cross-validated prediction accuracy tends to be overestimated when spatial structures present in the training data are fitted (or learned) by the model.

3.1 Introduction

Remote sensing data from optical instruments are increasingly available and captured at a wide range of spectral resolutions and wavelength regions (Ortenberg, 2011). Sensors can be deployed on different platforms such as satellites, aircraft, drones or land-based vehicles (Milton et al., 2009). Optical sensors capture spectral signals from a target surface but also capture spatial and temporal variations that are not necessarily targeted, regardless of the type of platform used (Feilhauer et al., 2017). Particularly reflectance captured from a continuous area is likely to exhibit significant spatial or temporal dependency for most types of surfaces (Legendre, 1993; Lobo et al., 1998). Thus any biophysical or biochemical characteristic of vegetation estimated by remote sensing is expected to be affected by spatiotemporal autocorrelation, regardless of the type of environment, sensor, platform, spatial resolution, extent, period of collection, or sample design (Hawkins, 2012; Legendre and Fortin, 1989; Naimi et al., 2011; Roberts et al., 2017).

Many studies have demonstrated the feasibility to quantify plant traits, such as chlorophyll content, water content and leaf area index (LAI), at leaf and canopy level with satisfactory accuracy using remotely sensed data (Clevers et al., 2010; Curran, 1989; Darvishzadeh et al., 2011). Applications for plant trait estimation range from assessing agricultural productivity and fire risk to monitoring biodiversity (Boegh et al., 2013; Skidmore et al., 2015). In most cases, explanatory variables based on narrow spectral bands from a

comprehensive wavelength range generate models that more accurately predict plant traits than variables based on broad bands from visible spectra (Curran, 1989; Qi et al., 2011). Therefore, hyperspectral data from either airborne or land-based platforms are often used to predict plant traits (Milton et al., 2009). Despite the current knowledge of the physical relationship between many plant traits and reflectance, it is still a challenge in a continuous and heterogeneous landscape, to consistently measure (or estimate) all factors needed to be able to use a deterministic model based on spectral radiance (Combal et al., 2002; Goodenough et al., 2006).

For example, even though the driving "cause-effect" relation between LAI and reflectance is known, data on other essential plant traits such as leaf structure, water content and leaf orientation are needed to be able to estimate LAI from reflectance data (Jacquemoud et al., 2009). Consequently, most of the applications for estimating biochemical or biophysical characteristics of vegetation rely on empirical associations between reflectance and plant pigments or canopy structure (Goodenough et al., 2006). Such empirical models must be trained with ground references that are representative, in space and time, of the remote sensing data (Manolakis et al., 2003). Ordinary least square regression using vegetation indices from a combination of two (or more) spectral bands is commonly used to predict plant traits. However, machine learning algorithms using the entire wavelength range, such as Partial Least Squares Regression (PLSR), Support Vector Machine (SVM), Random Forest (RF), or Artificial Neural Network (ANN) are often reported as being more accurate in predicting plant traits from hyperspectral data (Buitrago et al., 2018; Carvalho et al., 2013; Feilhauer et al., 2017; Skidmore, 1997; Yuan et al., 2017).

Using these supervised methods with a large set of predictors (*i.e.*, the number of spectral bands) in relation to the number of observations is likely to cause model overfitting (Hastie et al., 2009; Rocha et al., 2017). Overfitting occurs when the model incorporates random noises and data structures unrelated to the underlying relationship (James et al., 2013). Therefore, models need to be constrained in their complexity to avoid overfitting. This is often achieved by limiting the number of terms or interactions used for learning data structures (Kuhn and Johnson, 2013). The procedure to select the optimum model complexity to reduce the risk of overfitting is called tuning (James et al., 2013). Using a non-representative sample for training or using a sample from another population may jeopardise generalisation of a model. This could occur, for instance, when a model is applied to a new place or time that does not share similar characteristics (Cochran, 1977; Roberts et al., 2017).

A common way to estimate model generalisation or prediction accuracy is to estimate the Root Mean Squared Error (RMSE) of predictions based on a testing

dataset that is kept separate from the sample set before the model is fitted. Alternatively, cross-validation techniques such as leave-one-out, k-fold subsetting or bootstrapping can be applied (Bousquet and Elisseeff, 2002; Dormann et al., 2013; James et al., 2013). Despite being widely used, both approaches are based on subsets from the same sampling effort and may present unreliable estimations of model generalisation if the observations are spatially or temporally autocorrelated (Brenning, 2012; Roberts et al., 2017). Little is known of how machine learning algorithms that are trained on hyperspectral data perform when predicting for a different but similar area or in a different timeframe without being retrained (Feilhauer et al., 2017). Spatiotemporal structures in remote sensing data may actually represent the spatial and temporal pattern and processes of the plant trait under study (Hawkins, 2012). However, these structures may also present spatial patterns that are not causally related to the target plant trait. For example, soil characteristics or moisture content are likely to provoke changes in the spatial pattern captured by a sensor, either by altering a set of plant traits (targeted or not) or by capturing changes in the (soil) background (Cochrane et al., 2000).

For instance, using a field spectrometer it can be difficult to control variations in illumination geometry, canopy height and weather conditions across time and space under natural lighting (Breunig et al., 2013). Thus, the timing and order in which locations are visited to collect data can affect plant trait measurements, spectral measurements or both (Pearse et al., 2016; Woodgate et al., 2015). This aspect will be less apparent in satellite-based spectral data. However, taking in situ plant trait measurements may be so time-consuming that the vegetation gradually changes, possibly creating an undesirable data structure in the sampling collection (Mu et al., 2015; Wang et al., 2012). Spectral, temporal and spatial domains are all serially correlated data. This means that there is a logical order in the data, where pairs of wavelengths, times or locations positioned nearby, are likely to be more similar than pairs coming from positions further apart (Tobler, 1970). While the spectral data provoke multicollinearity problems related to strong correlation among predictors in the model (bands), the other two might provoke autocorrelation within observations (Babcock et al., 2013; Wikle and Hooten, 2010).

Spatial structures are often neglected, even when it is clear that the remote sensing data or in situ plant trait measurements are not far enough apart to be considered as spatially independent observations (Boegh et al., 2013; Carvalho et al., 2013; Knyazikhin et al., 2013; Lovett et al., 2005). Autocorrelated observations violate the model assumption of independent and identically distributed observations (i.i.d) in ordinary regressions (Dormann et al., 2007; Fortin et al., 2012; Legendre, 1993). For many machine learning algorithms, explicit warnings about such assumptions are missing. However,

noisy and autocorrelated data may cause model overfitting and misleading interpretations (Hawkins, 2012; Rocha et al., 2017). Often machine learning algorithms create latent variables to explain residual variance from previously fitted models in a progressive stepwise manner. Autocorrelation may not always be detectable in the residues of the final model (Kuhn and Johnson, 2013).

Given the combination of (1) large numbers of correlated bands available in hyperspectral data relative to the number of observations, (2) plant trait measurements containing spatiotemporal structures, and (3) supervised model selection applied by machine learning algorithms: particular attention is needed when empirically estimating plant traits from hyperspectral data, to avoid fitting predictive models with a low capacity of generalisation. The objective of this study is to assess to what extent spatial autocorrelation in the landscape (and hence in the imagery) can affect the prediction accuracy of plant traits when estimated by machine learning algorithms. The assessment focusses on prediction accuracy, model generalisation and independence of residuals across increasing levels of spatial dependency. The model fitting was implemented using two tuning processes, cross-validation and the NOIS method, which both aim to reduce the effects of overfitting while optimising model complexity. Machine learning algorithms are compared to less complex linear regressions using a vegetation index to assess model generalisation under spatial dependency.

3.2 Materials and Methods

Artificial landscapes were generated with increasing levels of spatial autocorrelation. This, in order to test the effect that spatial dependency has on the accuracy of model prediction when using machine learning regressions based on hyperspectral data. The artificial landscapes generated are a hypothetical representation of vegetation with a short canopy (as in grassland). These landscapes were represented by layers of plant traits for further be used as parameters to simulate reflectance with Radiative Transfer Models (RTM). Samples were drawn from the landscapes to train empirical models and to assess prediction accuracy while varying either the level of spatial dependency (autocorrelation ranges) or the spatial configuration (a unique realisation of a landscape).

The artificial landscapes were created by (1) generating variogram models with increasing ranges of spatial autocorrelation; (2) generating values for seven plant traits at a regular grid based on these variogram models using Sequential Gaussian Simulations of random fields (*i.e.*, unconditional simulation); (3) simulating hyperspectral data using Radiative Transfer Models (RTM), as

collected by spectrometers in the field, and (4) adding random and spatial dependent noise to the response variable (Y) and the hyperspectral data (X). Of the seven plant traits generated, Leaf Area Index (LAI) was selected as the response variable to be predicted by the simulated hyperspectral data. LAI can be defined as half of the surface area of green leaves per unit of horizontal ground area (Chen and Black, 1992). This parameter was chosen since it is the primary descriptor of vegetation functioning and structure, and essential to understanding biophysical processes (Woodgate et al., 2015).

3.2.1 Simulating Plant Traits

Unconditional simulations, based on variogram models, were used to generate plant traits representing landscapes with different levels of spatial dependency at a regular grid of 100 by 100 cells. In total, 15 levels of spatial dependency were created with autocorrelation ranging from zero to 70% of the extent of the artificial landscape (Figure 3.1). In other words, the landscapes ranged from ones where all pixels were independent in space to landscapes with autocorrelation of up to seventy per cent of the grid extent. Thirty realisations of each plant trait layer were generated for each of the 15 levels of spatial dependency. Each realisation used a single random path (neighbourhood selection) through the grid locations to create a unique spatial configuration (Bivand et al., 2015). The spatial patterning of a plant trait layer from the same realisation will be more similar between the following ranges of spatial autocorrelation than between different realisations of the same range of autocorrelation. This is illustrated in Figure 3.1, where patterns are more similar along the vertical lines than along horizontal lines. Initially, 450 LAI layers were generated, corresponding to 30 different realisations for each of the 15 levels of spatial dependency (30 x 15). The levels of spatial dependency were selected to produce similar intervals between variograms curves (Figure 3.1), rather than a scale equally spaced by distance or percentage of the area extent.

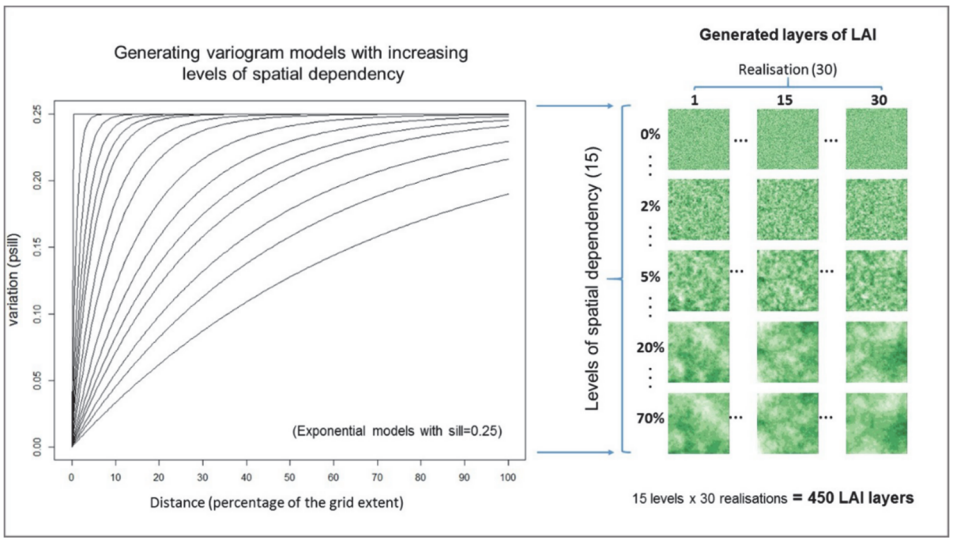


Figure 3.1 - Generation of Leaf Area Index (LAI) layers at 15 levels of spatial dependency.

Layers of plant traits were also generated based on Chlorophyll Leaf Content (Cab), Leaf Structure (N), Dry Matter Content (Cm) and Hotspot (hspot) to be used as input in a Radiative Transfer Model (RTM; PROSAIL 5B, see 2.2). The plant traits Carotenoid (Car) and Water Content (CW) were, based on their strong correlation with other plant traits, as applied by Vohland and Jarmer (2008) (Vohland and Jarmer, 2008) and Jarocińska (2014) (Jarocińska Anna M., 2014), defined as a function of Chlorophyll (Car = Cab/5) and of dry matter content (Cw = $4/Cm^{-1}$), respectively. The resultant 450 layers per plant trait were rescaled to present the same mean and standard deviation across realisations and levels of spatial dependency as presented in Table 3.1. Each plant trait layer was rescaled using the equation:

$$Trait_{rescale} = \mu_{trait} + \frac{(x_i - \bar{x}_{layer})\ \sigma_{trait}}{s_{layer}}, \quad (1)$$

where: μ_{trait} and σ_{trait} are the mean and the standard deviation of the plan trait as defined in Table 3.1;
x_i is the plant trait value of the pixel i, for i = 1, 2, …, 10,000;
$\bar{x}_{layer}$ and s_{layer} are the mean and standard deviation of the 10,000 simulated values of the layer.

This procedure standardised the data distribution while retaining the original spatial autocorrelation and spatial configuration. Random variations were then added to each plant trait layer (except LAI) to avoid linear combinations between parameters as all traits were generated from the same set of variogram models and realisation seeds. The random values added to each

pixel of the generated plant trait layers followed a normal distribution with mean zero and standard deviation according to the scale of the variable and the assumed coefficient of determination (R^2) with LAI (Table 3.1). These procedures guaranteed that the correlation between a trait and LAI was kept almost constant for all levels of spatial dependency and the different realisations considered. The R^2 with LAI was defined based on experiments found in the literature (Feret et al., 2008; International Symposium on Recent Advances in Quantitative Remote Sensing and Sobrino, 2002).

Table 3.1 - PROSAIL parameters used to simulate canopy reflectance for each 450 landscapes combination.

	Parameter	Description (unit)	Distribution	R^2 LAI
Leaf	Cab[1]	Chlorophyll a+b concentration (ug/cm^2)	~N(28,4.5)	0.36
	Car[2]	Carotenoid concentration (ug/cm^2)	~N(5,0.7)	0.35
	Cbrown[3]	Brown pigment (unitless)	0	-
	Cm[1]	Dry matter content (g/cm^{-2})	~N(0.004, 0.0005)	0.69
	Cw[2]	Equivalent water thickness (cm)	~N(0.016, 0.002)	0.66
	N[1]	Leaf structure parameter (unitless)	~N(1.5, 0.12)	0.48
Canopy	LAI[1]	Leaf Area Index (unitless)	~N(3.1, 0.6)	-
	hspot[1]	Hotspot parameter (unitless)	~N(0.05, 0.01)	0.50
	LAD[3]	Leaf angle distribution (attribute)	Erectophile (90°)	-
	psoil[3]	Dry/Wet soil factor (unitless)	0	-
Geometry	tto[4]	View zenith angle—VZA (degree)	~U(0,5)	-
	tts[4]	Solar zenith angle—SZA (degree)	~U(30, 38)	-
	psi[4]	Relative azimuth angle (degree)	~U(0,360)-U(129,252)	-

[1] simulated from plant traits by levels of spatial dependency and rescaled to present a Normal distribution ~N (mean, standard deviation). [2] a function of another parameter: Car = Cab/5 and Cw = 4/Cm^{-1}. [3] fixed values for all landscapes. [4] generated randomly from a uniform distribution with max and min ~U(min,max), varying according to hypothetical in situ measurements using a field spectrometer: where tto is the deviation from nadir (0°); tts = 90° minus the max and min sun altitude, and psi = ~U(0,360) minus the max and min solar zenith angle during the collection.

3.2.2 Simulating Spectra

A Radiative Transfer Model (RTM) was used to simulate 450 hyperspectral cubes. The PROSAIL 5B model was adopted to simulate wavelengths from 400 nm to 2500 nm with a 1 nm spectral resolution, generating in total 2100 bands. This physical model used 13 parameters divided into leaf, canopy and observation geometry properties (Berger et al., 2018; Jacquemoud et al., 2009). The model parameters were set to simulate spectra from grassland landscapes captured by a field spectrometer (Si et al., 2012). Besides the seven leaf and canopy plant traits described before, three other RTM

parameters were included (Table 3.1). These three were kept at fixed levels: Brown pigment (Cbrown) = 0, assuming that the canopies are entirely green; Leaf Angle Distribution (LAD) = Erectophile or 90°, given that is the principal orientation observed on grassland; and the soil moisture factor (psoil) = 0, assuming that moisture has no influence across space. The last three parameters in Table 3.1, related to illumination and observation geometry, were randomly generated from a uniform distribution ~U(min, max), varying solar and view angles slightly based on the hypothetical sample collection using a field spectrometer under natural light conditions.

3.2.3 Adding Variability into the Simulations

As RTM models are fully deterministic, different kinds of noises were added into the data to represent variations expected when observations are collected sequentially (rather than simultaneously) and by different instruments (spectra and ground references). It is unlikely that hyperspectral data captured by a handheld spectrometer will present the same spatial structures observed on the LAI values measured with another instrument (*e.g.*, LAI2200). For this reason, the random and spatially dependent noise was added separately to the spectra and to the LAI values that were used as the response variable in the training set (Figure 3.2). Before generating spectral cubes, spatially dependent noise N~(0, 0.25) was added to all LAI layers, using the same spatial dependency, but from a different realisation.

Random noise per waveband was also added to the spectra from a normal distribution with mean and standard deviation as estimated in a pilot experiment. This was done, because, when capturing spectra under natural light conditions, random variations in reflectance will occur as a result of the sensitivity of the sensor for specific regions of wavelengths. An experiment with a portable spectroradiometer (ASD FieldSpec® 3, Boulder, CO, USA) was conducted to estimate the magnitude of such expected random noise per band. The spectra from 40 distinct grassland surfaces were captured in similar atmospheric conditions. Each plot was measured for 30 consecutive times from spectral ranges between 400 nm and 2500 nm under natural sunlight around noon with a clear sky in summer. As some wavelengths were strongly affected by the random noise, smoothing with a Savitzky–Golay filter was applied over a length of 11 bands (Tsai and Philpot, n.d.). A spatially dependent noise of N~(0,0.1) was also added to the LAI layers before extracting data sets for model selection and validation, but with a different realisation than the ones used for simulating spectra. Random noise N~(0,0.1) was also added to create the final 450 LAI layers to be used as the response variable (Figure 3.2).

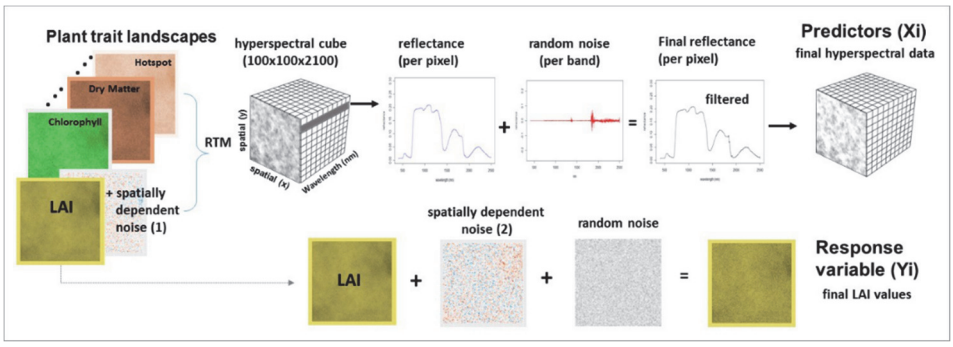
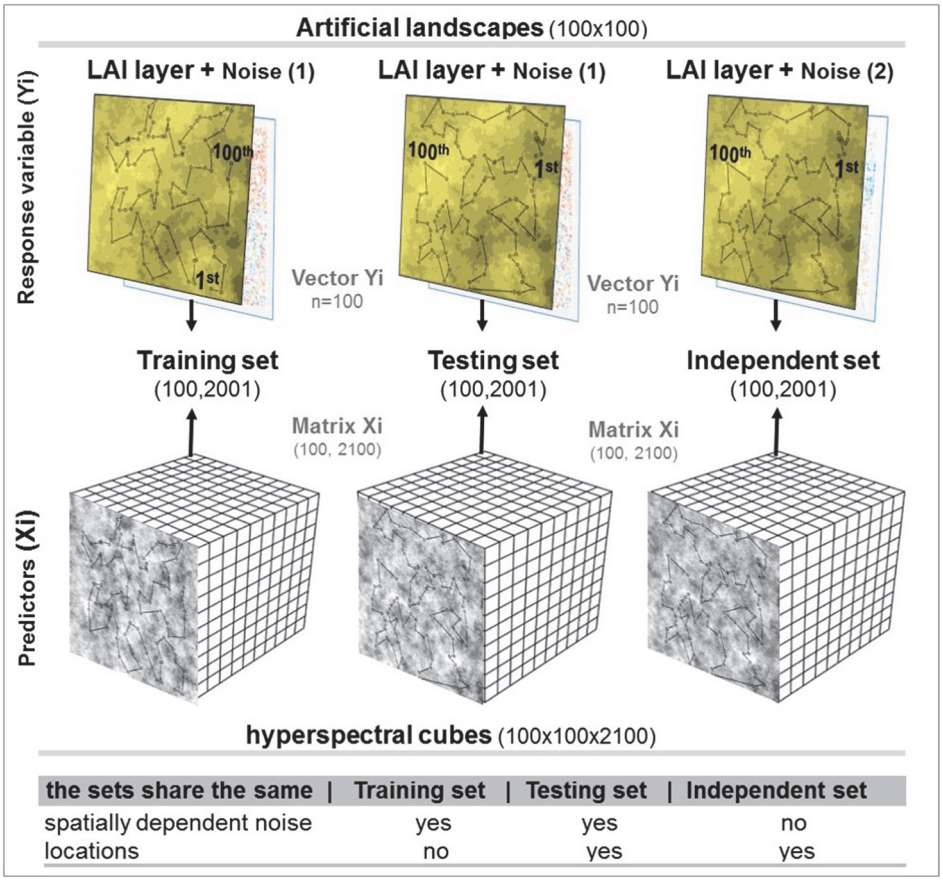


Figure 3.2 - Spectral simulation and process to generate predictors and response variable for modelling.

3.2.4 Sampling Schemes

Observations were extracted from the simulated spectral cubes and the final LAI layers to train empirical models (training set) at a hundred random x and y positions. At another hundred random locations, observations were extracted to validate the fitted models. This second set of locations was used for extracting two validation sets: a testing set and an independent set. The testing set is extracted from the same artificial landscapes (*i.e.*, same realisation and spatial dependency) as the training set, but with different values of random noise. The independent set contains both different, random and spatially dependent noise. The intention is to mimic an independent test set collected from the same landscape but in a different sampling campaign. In this case, the spatially dependent noise captured by one campaign may not match the other one used for training the model.

A path that minimises the distance travelled to collect these random points was defined for the two distinct sets of sample locations (Figure 3.3). The LAI and reflectance values were stored in this sequence of sample collection to train models, and later, to assess the presence of spatial correlation in the residues. The average distance between two consecutive points of the path was approximately 8% of the total extent of landscapes for both training and testing sets. There is an exclusive set of sample locations (training and testing) for each of the 30 realisations to reduce the risk of a particular sample distribution randomly selected causing a strong influence in the analyses.



the sets share the same	Training set	Testing set	Independent set
spatially dependent noise	yes	yes	no
locations	no	yes	yes

Figure 3.3 - Sampling spectra and Leaf Area Index (LAI) values for model training and validation sets. The testing sets share the same spatial dependent noise as the training sets, but sample at different locations and sequence. The independent sets present the same location as the testing sets but with different spatially dependent noise. Those datasets are publicly accessible at DOI: 10.4121/uuid:2016d562-cf6e-4060-ac13-5db9477b6512.

3.2.5 Modelling and Performance Assessment

Partial Least Squared Regression (PLSR) and Support Vector Machine (SVM) were selected as machine learning algorithms in this study because these are frequently used for modelling hyperspectral data, and their prediction accuracies have been reported as high, relative to other algorithms (Carvalho et al., 2013; Feilhauer et al., 2015). Also, these techniques can deal with multicollinearity and high dimensional data. Overfitting can be reduced by limiting the level of complexity of these models, such as the number of components in PLSR.

Two tuning methods were applied to select model complexity: traditional cross-validation and a novel method called Naïve Overfitting Index Selection (NOIS)

(Rocha et al., 2017). When tuning a model with cross-validation (we used 10-fold cross-validation), a model is selected with a complexity that minimises the Root Mean Squared Error (RMSE) of the predictions from the validation subsets (Hastie et al., 2009). This procedure was randomly repeated ten times, resulting in a combination of 100 subsets of training and validation sets from the original data (James et al., 2013). The NOIS method selects model complexity considering an a priori level of overfitting tolerated by the user (we used 5%; see Rocha et al., 2017 (Rocha et al., 2017) for details). The complexity selected for models tuned with cross-validation varied according to the landscape. In PLSR up to 20 components could be selected, while in SVM the tuning parameter was chosen among 11 cost values (0.00005 to 0.25).

The tuning parameters were fixed across all landscapes when tuning with the NOIS method. Partial Least Square Regression (PLSR) models were fixed with two components, while Support Vector Machine (SVM) models were parameterised with a cost of 0.0001 (Appendix 3A). Partial Least Square Regression (PLSR) models were fixed with two components, while Support Vector Machine (SVM) models were parameterised with a cost of 0.0001 (Appendix 3A). An ordinary least square (simple) regression using a two-band vegetation index was fitted to compare with the machine learning algorithms. The LAI Determining Index (LAIDI), a ratio between two wavelengths (1050 nm and 1250 nm) situated in the NIR spectral domain, was used to predict LAI using linear regression (Delalieux et al., 2008). The wavelengths were selected a priori based on literature, rather than by searching for the band combination that explained most of the variation in the response variable.

The selected model for each regression technique was assessed by the capacity to generalise with similar accuracy when predicting with a new dataset. Therefore, the RMSE calculated from the training set (RMSEtr), the testing set (RMSEtest) and the independent set (RMSEind), were compared to the estimated RMSE of cross-validation (RMSEcv). The testing and independent set were also used to assess model generalisation in a different realisation or spatial dependency (when moving across landscapes vertically or horizontally as in Figure 3.1). The Durbin Watson statistic was calculated to quantify autocorrelation in the model residues considering the observations sequentially in space, following the sampling path as depicted in Figure 3.2, reflecting the spatiotemporal autocorrelation as if the data were collected in the field. The statistic varies between 0 and 4, where values around 2 indicate no autocorrelation, values below 2 indicate positive autocorrelation and values above 2 negative autocorrelations (Hastie et al., 2009).

All the analyses are executed in R version 3.2.2 (The R Foundation for Statistical Computing). The package gstat was used for unconditional simulations, hsdar for simulations of spectra with PROSAIL 5B, and Caret for

fitting models from all regression techniques with the same cross-validation approach.

3.3 Results

3.3.1 Prediction Accuracy Estimated from the Training Set

Estimates of RMSE based on the complete dataset used for training the model (RMSEtr), as expected, were smaller than the estimation of the cross-validated partition (RMSEcv; Figure 3.4). The differences were much larger for the machine learning algorithms tuned with cross-validation (PLSRcv and SVMcv) than for those tuned with the NOIS method. A large gap between the estimated RMSEcv and RMSEtr indicated overfitting, which may be partly caused by learning (*i.e.*, fitting) spatial structures and random noises. Regardless of the level of spatial dependency, differences between RMSEcv and RMSEtr were relatively small for machine learning models that were tuned with the NOIS method and practically disappeared for the simple regression (lm). For the PLSRcv and SVMcv models, this difference slightly increased when the spatial dependency increased, because the training error (RMSEtr) decreased faster than the cross-validated error (RMSEcv). This trend is less clear for the less complex models tuned by the NOIS method or the linear models. The RMSEcv from models tuned by cross-validation was smaller, as the selection was based on the model complexity that minimises the prediction error.

The LAI values for all artificial landscapes came from the same distribution N~(4,0.5). Thus, across the fifteen levels of spatial dependency, no trend should be observed in RMSEcv. Where spatial relations were not captured (fitted) by the model, all variations should come from random differences among the thirty realisations. Model complexities should be similar for all 450 landscapes (15 x 30), as spectra were generated by the same deterministic function from the Radiative Transfer Models (RTMs). However, for instance, the PLSRcv models had levels of complexity ranging from 1 to 12 components. This is a sign that models may be fitting spatial structures and random noises because these form the only differences within the artificial landscapes. The prediction errors estimated from the training data are more affected by spatial structures and random noise in complex models such as PLSRcv and SVMcv.

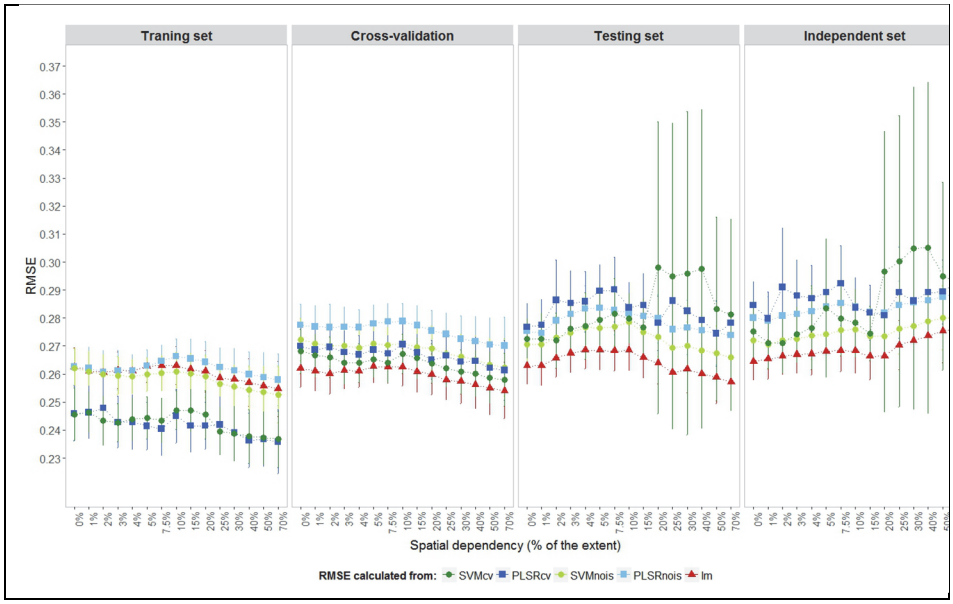


Figure 3.4 - Mean and confidence intervals for prediction error, Root Mean Squared Error (RMSE), by the level of spatial dependency estimated from the training set (RMSEtr), cross-validation subsets (RMSEcv), testing sets based on a different sample subset from the same landscape (RMSEtest), and independent testing sets (RMSEind). The different shapes represent different models: squares for PLSR, circles for SVM and triangles for the linear regression model (lm). The darker shades represent the models tuned via traditional cross-validation (PLSRcv and SVMcv), the lighter shades those tuned via the NOIS method (PLSR_NOIS and SVM_NOIS).

3.3.2 Prediction Accuracy Estimated from Validation Sets

Prediction accuracy can be estimated by observations collected in the same campaign as the training set (so from the same imagery and ground observations) but kept apart for validation instead of cross-validation. Despite being from different sample locations, the observations from the testing set (RMSEtest) contained the same underlying spatial structure as the training set. For more complex models the testing sets presented higher prediction errors than cross-validation estimation did (Figure 3.4). Simple linear models performed according to the testing set and quite similar to cross-validation, while SVM models tuned with the NOIS method presented the best machine learning performance. For cases where spatial dependency was higher than 15%, RMSEtest values for more complex models presented much higher average errors and wider confidence intervals.

The prediction error can also be estimated by an independent testing set collected in the same landscape, but in a different sampling campaign, represented here by RMSEind. In this case, the predictions presented visibly higher errors than the RMSEtest for spatial dependencies higher than 15% for

all models. This occurred because the models fitted spatial relations in the observations that were not supported by any explanatory variable present as these sets had different spatially dependent noises. Overall, prediction errors presented two general types of behaviour across levels of spatial dependency. Firstly, regardless of the type of validation data used, the error increased around levels of spatial dependency that matched the sampling distance (approximately 8% of the extent in our case). Secondly, the error decreased above 15% of spatial dependence, except for complex models validated by the testing set (RMSEtest) or any model validated with the independent set.

3.3.3 Prediction Accuracy Estimated on a New Realisation

The cause-effect relationship between the reflectance and LAI values is the same for all artificial landscapes, being defined deterministically by the Radiative Transfer Model. Therefore, any empirical model should, in theory, produce a similar accuracy when predicting another realisation. This assumption was not confirmed by this study whenever the models were complex, or the spatial dependency was strong. The gradual reduction in prediction error (RMSEtest) in landscapes with a spatial dependency higher than 15% was not observed when validated with data from another realisation. This confirms that these models are learning with the spatial distribution of the training set.

Two aspects are resulting in a lack of model generalisation. One is caused by the sampling density, resulting in higher prediction errors between 2% and 10% of spatial dependency. This behaviour indicates that sample densities similar to the spatial dependency of the plant trait may produce quite unstable models and reduce the accuracy of the predictions. The other aspect is related to models that were trained in landscapes with strong spatial dependency (more than 15%). The prediction error estimated by cross-validation, in this case, is not observed in a new landscape with the same spatial autocorrelation. In other words, these models were fitted to represent a particular spatial distribution in that dataset, rather than the underlying causal relationship.

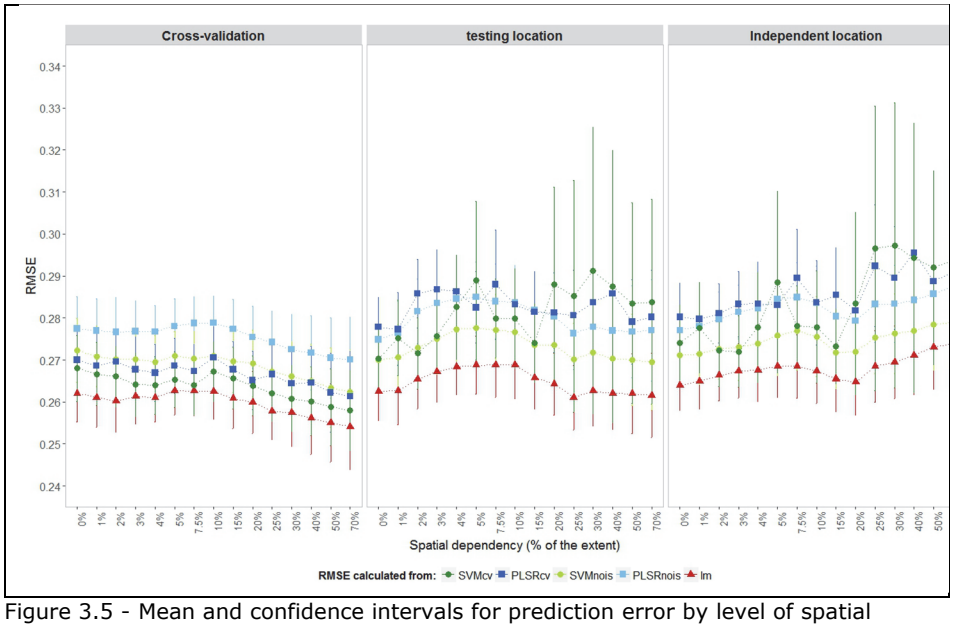


Figure 3.5 - Mean and confidence intervals for prediction error by level of spatial dependency estimated from the cross-validation subset (RMSEcv), compared to the testing and independent sets but from different realisation. The different shapes represent different models: squares for PLSR, circles for SVM and triangles for the linear regression model (lm). The darker shades represent the models tuned via traditional cross-validation (PLSRcv and SVMcv), the lighter shades those tuned via the NOIS method (PLSR_NOIS and SVM_NOIS).

3.3.4 Prediction Accuracy Estimated on a Different Spatial Dependency

Models trained in landscapes without spatial dependency, in general, produced lower prediction errors when applied to landscapes with other levels of spatial dependency as presented in Figure 3.6. The models should achieve similar accuracies when used for landscapes having different levels of spatial dependency if they capture only the true relation between reflectance and LAI. However, the higher the spatial dependency used for training a model, the more likely it is that this model will capture undesirable spatial relations from the observations. This is most visible on complex models such as SVMcv, where lower prediction accuracies are observed for models trained on the highest spatial dependency (*e.g.*, 70%, Figure 3.6).

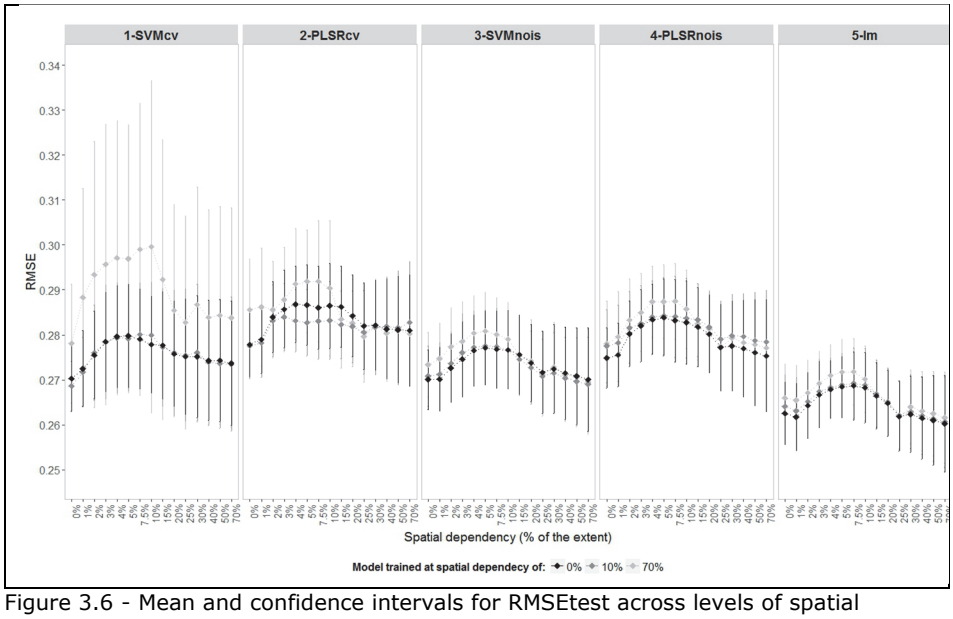


Figure 3.6 - Mean and confidence intervals for RMSEtest across levels of spatial dependency for which models are making predictions (along x-axis) when trained in landscapes with a spatial dependency of 0% (black), 10% (dark grey) and 70% (light grey) for different models and tuning methods.

Regardless of the spatial dependency in the landscape used for training the model, the RMSE increases in the landscapes with levels of spatial dependency between 2% and 10%. In this interval, the PLSRcv models present lower prediction error when trained under 10% of spatial dependency. If the sample size is reduced, the highest values of RMSE shift to landscapes with stronger spatially dependency, while the effect moves in the opposite direction when the sampling density is increased (Appendix 3B).

3.3.5 The Effect of Spatial Dependency on Model Assumptions

Model residues should be normally distributed with mean zero, but should also be randomly distributed in space and time. Non-spatial models using spatially dependent observations might not fulfil these assumptions. The Durbin Watson (DW) statistics (Figure 3.7) show the presence of significant autocorrelation in the residues for models that are trained in landscapes with the spatial dependency of 5% and above, departing clearly from the baseline represented by the value 2. The autocorrelation in the model residues is less strong for the more complex models (PLSR_CV and SVM_CV) compared to the more simple models (SVM_NOIS and LM). Autocorrelation may not be detected in the final model residues in machine learning algorithms when sufficient latent variables are created to explain all the residues. Figure 3.7 shows some outliers for machine learning trained with data that had 20% to 40% of spatial dependency

supporting this claim. If most of the spatial structures of the response variable (LAI) are explained by the spectra, the model residues might also not present significant spatial autocorrelation (Appendix 3C). This would occur if spatially dependent noise was not added in this study before modelling. However, in a real case scenario, it is unlikely that remote sensing data of a canopy will only explain the spatial dependency of the target plant trait.

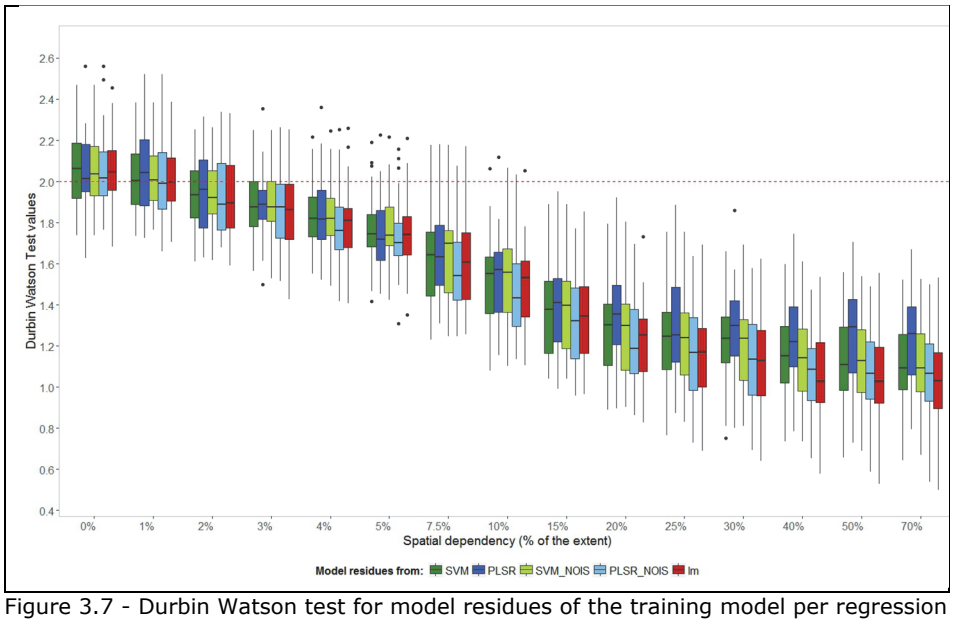


Figure 3.7 - Durbin Watson test for model residues of the training model per regression technique and tuning approach.

3.4 Discussion

3.4.1 Spatial Dependency and Prediction Accuracy

Plant traits are likely to exhibit spatial dependency in continuous landscapes regardless of the extent or the spatial resolution of the measurements (Legendre and Fortin, 1989; Roberts et al., 2017). Therefore, when the objective is to compare plant trait predictions from similar landscapes or monitoring the same landscape over time, a model needs to be carefully fitted and tested. An important consideration should be to avoid modelling spatial relations in the observations when these are not causally linked to the plant trait under investigation. Otherwise, these models should be considered to be "single-use" models that have little capacity to predict when new spectral data are available.

This study shows that training complex models when using high dimensional data, such as hyperspectral measurements, presents a considerable risk of

underestimating prediction error. This occurs because models may overfit by capturing random noise from a large number of wavelengths, often supported by insufficient observations. The risk grows when model complexity increases and in the event of spatial dependency in the plant traits. Here we showed that machine learning models tuned by cross-validation seemed to learn from spatial structures and fit spurious correlations between random noise and LAI, suggesting an increase in performance.

A linear regression using a predefined two bands ratio index reduces this risk of overfitting by decreasing the effect of multicollinearity and lack of the degree of freedom caused by a large number of predictors available. However, if a linear regression is selected through a stepwise algorithm with all spectral bands, or using a vegetation index that searches the combinations of two or more bands provides the best correlation with the plant trait (*i.e.*, supervised feature selection), the risk of overfitting is also expected to be high, regardless of how simple the final model is (James et al., 2013; Thenkabail et al., 2000). Optimistic estimation of RMSE can also be the result of using cross-validation procedures when testing data is set apart from the same sampling effort as the training data. This is especially noticeable when significant spatial structures are present in the data (compare RMSEcv with RMSEind in Figure 3.4). With a large number of predictors available, the cross-validation tuning process seemed to select models that partially fit variations in the data rather than in the underlying phenomenon.

Less complex models tuned by the NOIS method or linear models present more reliable estimations of the prediction errors, but these also tend to be underestimated when the spatial dependency is not well covered by the sampling density. In our simulations, this was the case when spatial dependency was less than the average distance between sampling locations (*i.e.*, landscapes with a spatial dependency of less than 10%). The spatial dependency of plant traits should be used to define the optimal distance between samples. However, in most cases, this information is only known after the data is collected (Kobayashi et al., 2013). If remote sensing images are available before the sampling campaign, these could be used as a reference to estimate the expected spatial dependency from wavelengths known to be correlated with the desirable plant trait. The primary goal of model fitting is to learn the empirical relationship between reflectance and plant traits for a determinate landscape. In this study, this implied that the models should approximate the function used by the Radiative Transfer Models to simulated reflectance from LAI values. When the model learns with spatial structures and random noises in the data, instead of the underlying relationships, it produces misleading inferences and results in underestimating prediction errors.

3.4.2 Beyond the Scope of These Simulations

A number of simplifications and restrictive assumptions were used to simulate spectral data that only vary spatially in relation to LAI values. For instance, all plant traits used as parameters by the RTM to simulate spectra had the same level of spatial dependency as LAI within each realisation. The spatial structure introduced by noise in both LAI and spectral data presented the same level of spatial dependency. These assumptions might be unrealistic as, despite the potential correlation among plant traits, it is more likely that spatial dependency occurs at different levels. Reflectance values may also present different spatial structures based on the wavelength regions and their sensitivity to the different plant traits. In addition, other spatial structures related to the landscape, such as soil moisture and its effect on background reflectance, will be captured in spectral datasets. Temporal structures can also be captured as optical sensors, and ground references are hardly ever collected simultaneously and therefore are not entirely free of systematic errors. All these factors and their combinations can exponentially increase the risk of overfitting. In real-life environments, which are not as controlled as in the presented study, the underestimation of the prediction error is probably even higher for complex models through learning from the spatiotemporal structure in the training data.

Although machine learning regressions are known for not requiring assumptions such as independent and identical distributed observations nor model residues, their effect on model prediction might be even stronger than with ordinary least squared regressions. The relaxation of assumptions such as the absence of multicollinearity and spatial autocorrelation, or principles such as model parsimony, does not mean that their effect on model predictions is negligible. Due to the high dimensionality of hyperspectral data, machine learning is often used for modelling plant traits based only on spectral data. Classical assessment comparing spatial and non-spatial models are based on how they manage to decrease or eliminate spatial autocorrelation in the residuals (Dormann et al., 2007). Machine learning algorithms cannot be compared in this way, as they often use the residuals to improve the model.

3.4.3 Spatiotemporal Structures in Remote Sensing

Water availability, species dominance, slopes, nutrient concentrations in the soil and many more factors can drive the spatial dependency of plant traits in nature. If measurements from continuous vegetation (landscapes) do not contain spatial structures, they are not an accurate description of nature, limiting the understanding of the target surface (Hawkins, 2004). As it is known that nature is stochastic and does not repeat a process under the same conditions, temporal structures can also affect prediction accuracy.

Spectral measurements may vary over the course of a day by capturing changes in the relation between sun altitude and viewing angles at the same location. Variations can also occur on a medium to long-term time scale, related to weather conditions or seasonal plant cycles. Organising a field campaign using a limited period of the day to control for solar azimuth may require many days to complete the data collection. This may decrease the variation in illumination geometry but will increase differences in plant phenology or maximum sun zenith. In contrast, an intensive and short campaign may have to use many hours per day, increasing variability in illumination conditions and the autocorrelation between consecutive measurements. This trade-off in sources of error will depend on the plant trait of interest, sample design and instruments used.

Both, plant traits and reflectance values, for instance, can be captured by optical sensors in the field (*e.g.*, LAI using LAI2200 and spectral data with an ASD spectrometer). In this case, the risk of undesirable spatiotemporal structures in the data is higher than when ground references is measured in the lab, and spectral data come from the same scene of a satellite image. In satellite or airborne data, the difference in geometrical distortion (changes in the field of view) within the scene or in time between two scenes (in the same swath or not) may also provoke spatiotemporal patterns. Other data structures rather than spatiotemporal, such as phylogenetic or genetic relations may also lead to dependency in the multi-species analysis (Roberts et al., 2017). In a lab experiment illumination and view angles can be well controlled, however, combinations such as repeated samples from a set of different plant species, growth stages or levels of stress may create a sequence in the data that can lead to similar effects as spatiotemporal structure if modelled together (Roberts et al., 2017).

Modelled plant traits from a landscape using hyperspectral data likely present at least three sources of spatial autocorrelation in the data: (1) the spatial pattern of the landscape determined by the underlying process that drives the plant trait, in this study represented by the different realisations of LAI values correlated in space; (2) the spatial autocorrelation in the ground measurements determined by the sampling footprint used for training, illustrated here by the path of a sequential sample; and (3) the autocorrelation related to noise from the optical sensor and ground measurements, as data is captured neither simultaneously, nor independently in space (included in this study as spatial dependent noise). The first source is of natural origin (inherent) and should be modelled with an approach that takes the spatial structure explicitly into account whenever it is not fully explained by the reflectance values (Appendix 3C). Although, to model the spatial structure of the plant trait properly, the sample design and density (second source) have to be spatially representative (Reichenau et al., 2016). The third source of

autocorrelation can mislead the second and might be considered a sort of bias or distortion that should be corrected before modelling. For instance, whether or not a spatial structure in remote sensing data caused by soil background should be treated as a systematic error or modelled as part of the underlying process, will depend on whether the plant trait under consideration is also related to process, or only affects the reflectance values.

Appropriate sampling design and a well-controlled measurement campaign can significantly reduce random and systematic noise in the measurements, but will never eliminate all noise, so model validation with new (unseen) observations that do not stem from the same sampling effort is essential to achieving model generalisation. Lastly, in this study, we showed that the choice of sample size could affect generalisation in different ways. Models trained with a small sample size may increase overfitting as a small sample size reduces the number of observations to support a large number of predictors. Larger samples reduce the distance between the points, changing the sensitivity to the spatial dependency and increasing model complexity (Appendix 3B). Remote sensing data provide an opportunity to study the spatial structure in a landscape before planning fieldwork to collect ground references. This opportunity should be grasped more often to determine sample design and point density in order to avoid or properly model spatiotemporal structures.

3.5 Conclusions

Machine learning regressions using hyperspectral data to predict plant traits are sensitive to overfitting if careful model tuning is not conducted. In the presence of strong spatial autocorrelation, the risk of overfitting increases considerably, and there is no obvious solution to correct this for machine learnings. The result is that models have lower actual prediction accuracies than those estimated by cross-validation. Spatial structures should not mistakenly be interpreted as causal relationships between spectra and the trait of interest. When spatial structures are inherent to the underlying process that drives the trait but are not completely explained by the spectra, they could be modelled with a method that accounts explicitly for the spatial structure. As illustrated in this study, the effect of the spatial dependency can be easily detected in model residues by conventional autocorrelation tests using a sequence of sampling plots. Robust model validation and tuning approaches to restricting complexity in machine learning algorithms, such as the NOIS method, can help to reduce the risk of producing "single-use" models that cannot be applied in any other area or at any other time than for which they were trained.

Appendix 3A

NOIS Method—The Naïve Overfitting Index Selection (NOIS) is implemented in three steps (see also Rocha et al. 2017 (Rocha et al., 2017)):

1st: Artificial spectra are generated from a multivariate normal distribution based on the mean and covariance matrix of original hyperspectral data. This procedure keeps the same number of observations and number of bands as the original spectra but ensures that these are uncorrelated with the response variable.

2nd: Regression models are fitted with the generated spectra as explanatory variables, but the original plant trait (LAI) as the response variable. The (naïve) models are fitted with increasing levels of complexity, for instance, from 1 to 20 components in PLRS.

3rd: An overfitting index for each level of complexity is calculated based on the contribution of the naïve model to reduce the Root Mean Square Error (RMSE) of the prediction according to the equation:

$$NOIS\ index\ = 1\ - \frac{RMSE_g}{RMSE_y}, \quad (2)$$

where: $\boldsymbol{RMSE_g}$ is the Root Mean Square Error for a given model complexity fitted with generated data (naïve model);

$\boldsymbol{RMSE_y}$ is the Root Mean Square Error when the mean of the response variable (y) is taken as a prediction.

The index has a maximum value of 1 when the predictor error approaches zero, and it should approach 0 when there is no model overfitting. For instance, a naïve overfitting index of 0.45 indicates that the prediction error is falsely reduced by 45% at the given level of complexity. Negative index values indicate that the model predicts a bigger error than RMSEy, and the model complexity is constrained excessively ("underfitted"). Because the RMSEy is solely based on the response variable (y), and no model contribution is expected from naïve models, the degree of overfitting is directly comparable between regression techniques.

The results show that after including more than two components in the PLSR models and a setting the cost variable higher than 0.00025 in the SVM models, they start to produce overfitting indices that exceed the pre-defined tolerance level of 5% (red line). The models at the highest level of complexity where the NOIS index values still stay below the tolerance line were selected (Figure A1). A unique model complexity for each regression technique was used for all realisations and spatial dependencies when the NOIS method was used, in contrast to cross-validation where for each landscape a different complexity was selected.

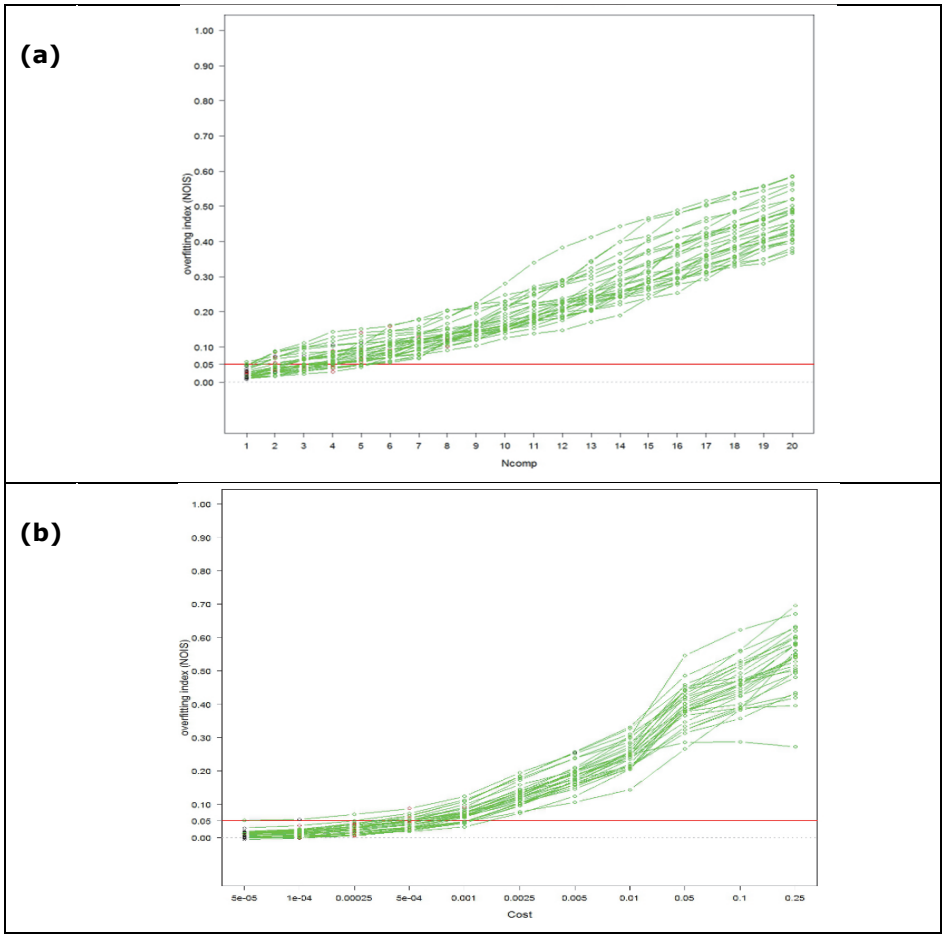


Figure 3.8 - Results of the NOIS index for PLSR (a) and SVM (b) for the landscapes without spatial dependency (0%). Green lines show the naïve overfitting index (y-axis) develops with increasing model complexity (x-axis) for each realisation. The red and black circles are the model complexity selected by the tuning process using a cross-validation approach from the original and naïve data respectively. The red line is a threshold, and the model complexity selected is set at the highest level the naïve overfitting index remains under. The script to run the NOIS method in R and the database used in this paper is publicly accessible at DOI: 10.4121/uuid:2016d562-cf6e-4060-ac13-5db9477b6512

Appendix 3B

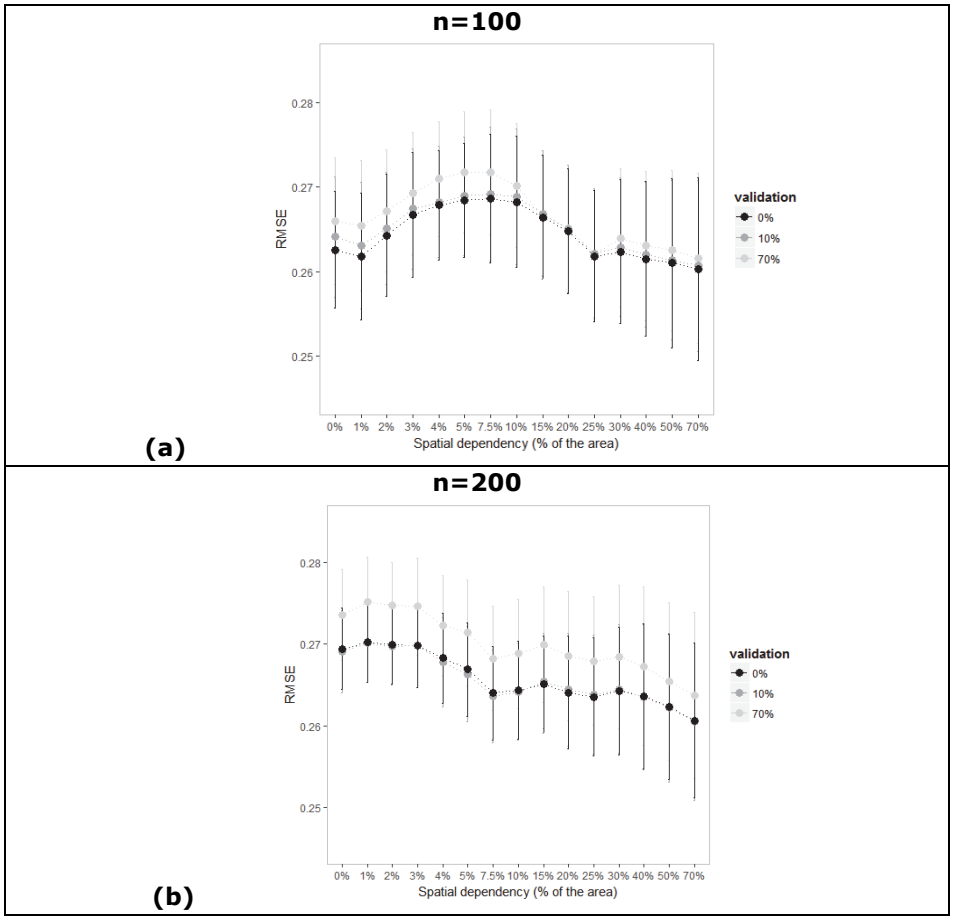


Figure 3.9 - Mean and confidence intervals for RMSEtest across levels of spatial dependency for which models are making predictions (along the x-axis) when trained in landscapes with a spatial dependency of 0% (black), 10% (dark grey) and 70% (light grey) for different sample sizes. Linear models trained with the sample size of n = 100 (a) against a sample size of n = 200 (b).

Linear models trained and tested with 100 observations present the highest RMSE (for the testing set), in landscapes with spatial dependency around 7.5% of the extent. This dependency presents a similar range of autocorrelation as the average distance calculated between consecutive points according to the sequence of the sample patch (8%). When the sample size increases, the higher values of RMSE shift to lower values of spatial dependence while the sample distance reduces to 5% of the total extent due to an increase in the density of points.

Appendix 3C

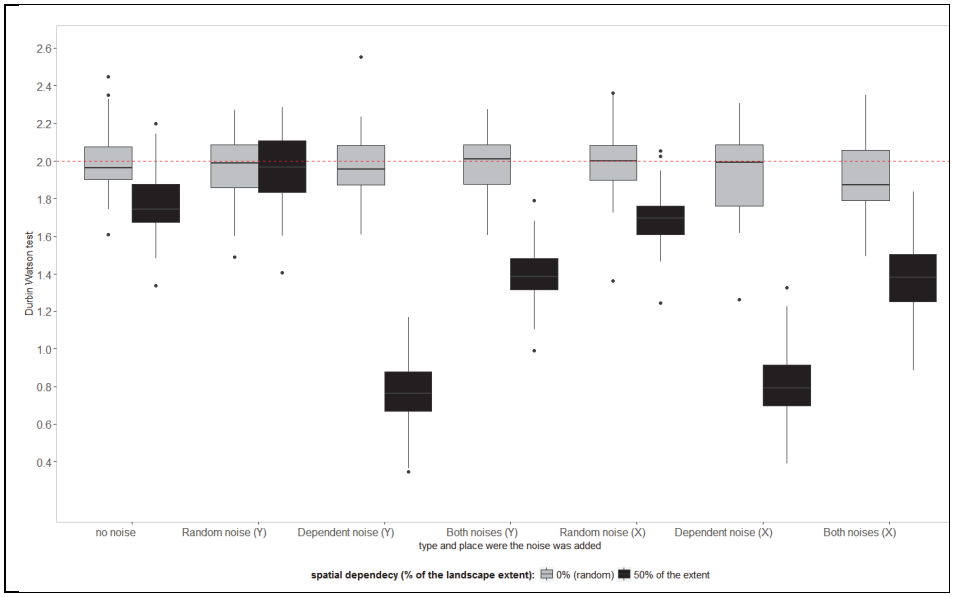


Figure 3.10 - Results of Durbin Watson test for the residues of linear models for the landscapes without spatial dependency (grey) and with 50% of the extent (black). The red line shows the expected value for ideal random residues. The model residues were tested by adding random error, spatial dependent error and both noises into the response variable (LAI) or into the explanatory variable (spectra).

The graph clearly shows that when the only difference between the spatial pattern of the plant trait LAI and the respective reflectance index is derived by a random error in the response, the model can present independent residues (free of autocorrelation). This occurs as the spatial structures in the explanatory variables and in the response variable, both describe the same pattern, and then, no residual autocorrelation is detected in the model. Otherwise, autocorrelation in the residuals is expected from non-spatial models under the presence of spatial dependency.

Chapter 4

A space-spectra tuning improves prediction of plant traits with hyperspectral data[3]

[3] This chapter is based on: Rocha, A.D.; Groen, T.A.; Skidmore, A.K., 2019. Spatially-explicit modelling with support of hyperspectral data can improve prediction of plant traits. Remote Sensing of Environment, Doi: 10.1016/j.rse.2019.05.019, (111200).

Abstract

Data from remote sensing with finer spectral and spatial resolution are increasingly available. While this allows a more accurate prediction of plant traits at different spatial scales, it raises concerns about a lack of independence between observations. Hyperspectral wavelengths are serially correlated provoking multicollinearity among the predictors. As collections of ground references for validation remains time-consuming and difficult in many environments, empirical models are trained with a limited number of observations compared to the number of wavelengths. Moreover, any set of observations collected from a continuous surface is also likely to be spatially autocorrelated. Machine learning regression facilitates the task of selecting the most informative wavelengths and then transforming them into latent variables to avoid the problem of multicollinearity. However, these regression methods do not solve the problem of spatial autocorrelation in the model residuals. In this study, we show that, when significant spatial autocorrelation is observed, models that explicitly deal with spatial information and use a spectral index as a covariate exhibit a higher prediction accuracy than machine learning regressions do. However, for these models to work, the number of (hyperspectral) bands included in the models has to be drastically reduced, and the model cannot be directly extrapolated to a new (unobserved) location in another area. We conclude that quantifying spatial autocorrelation a-priori in the data can help in deciding whether the spatial and the spectral domain should be modelled together or not.

4.1 Introduction

Plant traits such as chlorophyll content and leaf area index are essential biochemical and biophysical characteristics, related to processes such as photosynthesis and net primary productivity (Huber et al., 2008; Kokaly et al., 2009; Mirzaie et al., 2014; Schlerf et al., 2010). The monitoring of plant trait variations across landscapes is in demand from agribusiness to conservation applications such as Essential Climate & Biodiversity Variables (Abdullah et al., 2018; Manolakis et al., 2003; Skidmore et al., 2015). However, these mapping exercises require in situ observations over an extensive spatial and frequent temporal scale and are therefore laborious and expensive (Secades et al., 2014). Remote sensing technologies offer an opportunity to estimate plant traits over finer spatial and temporal resolutions (Finley et al., 2014; Patenaude et al., 2008; Shen et al., 2013b). Plant traits at canopy level can be successfully predicted by different sensor instruments and platforms in many ecosystems (Kokaly et al., 2009; Ramoelo et al., 2012; Van Cleemput et al., 2018). Despite satisfactory results, with different spectral resolutions, hyperspectral data are often used to estimate plant traits because of the specificity of some of the narrow spectral bands (Clevers and Kooistra, 2012; Huber et al., 2008; Mutanga and Skidmore, 2007). However, even with

hyperspectral sensors, a perfect spectrally based prediction for any given surface property is unlikely, even in laboratory experiments (Manolakis et al., 2003).

Estimation of plant traits by spectra depends greatly on variations in a number of elements, including leaf surface, canopy structure, the different species present in an area and the phenological stage of growth (Knyazikhin et al., 2013; Li et al., 2011a; Martin et al., 2008). Remote sensing data may also show great variability in some regions of the spectra, according to the capacity of the sensor's platform to reduce effects derived from illumination and view angle variations (Manolakis et al., 2003). Even though the physical understanding about how changes in leaf pigments (*e.g.* chlorophyll content) and canopy structure (*e.g.* LAI) affect reflectance are mainly known, it remains difficult to measure (or control) all factors required to predict accurately using a fully deterministic model (Combal et al., 2002). Therefore, remote sensing applications rely mostly on regression models based on the empirical relationship between ground references and the corresponding (leaf or canopy) reflectance (Kokaly et al. 2009).

Uncontrolled factors in experiments outside a laboratory may create a spurious correlation that is mistakenly interpreted as causality when drawing inferences about the model predictions (Milton et al., 2009). Multicollinearity, for example, is an issue in regression models using hyperspectral data, as the wavelengths are often strongly correlated (Dormann et al., 2013; Nguyen and Lee, 2006). In such cases, the selection of wavelengths that significantly contribute for predicting a plant trait may be masked by linear combinations within the large set of wavebands available (Gelman and Hill, 2006; Kuhn and Johnson, 2013). Overfitting is another common problem in regression models using hyperspectral data (Curran, 1989; Rocha et al., 2017). This occurs as models are often trained with a limited number of ground references in relation to a large number of wavelengths selected (or searched) as explanatory variables (Hansen and Schjoerring, 2003). The high dimensionality of hyperspectral data increases the risk of fitting spurious correlations due to random and systematic noise, resulting in accurate predictions only for the data which the model was trained with (Meehl, 1945; Rocha et al., 2017).

For modelling plant traits with hyperspectral data, the number of predictors (wavelengths) has to be reduced to avoid multicollinearity and overfitting (James et al., 2013). This can be achieved by an unsupervised approach (without the support of the response), or by transforming wavelengths into "latent variables" with methods such as wavelets or principal component analysis, or by using the result of previous studies which have concluded that a combination of wavebands or a vegetation index was efficient to predict this plant trait in similar conditions (Bruce et al., 2002). Supervised model

selection, such as genetic algorithms or stepwise regression, have the support of the response variable to find the most informative wavelengths (Schlerf et al., 2010; Ullah et al., 2012). Although more efficient when searching for relevant predictors to explain the response variable, supervised methods significantly raises the risk of overfitting (James et al., 2013).

Many machine learning algorithms, such as Partial Least Square Regression (PLSR), are supervised methods since they create latent variables by transforming the original data into principal components with the support of the response (Buitrago et al., 2018; Martin et al., 2008; Ramoelo et al., 2012). Broadly used for modelling plant traits with hyperspectral data, machine learning regressions overcome the problem of multicollinearity by reducing dimensionality, but not necessarily by decreasing the risk of overfitting (Rocha et al., 2017). Machine learning regression, therefore, requires a procedure to select a model complexity which counterbalances the prediction accuracy with the risk of overfitting. This procedure to restrict the number of terms included in the model, such as the number of "components" in PLSR, is called "tuning process" and it is traditionally performed by cross-validation (Kuhn and Johnson, 2013). Machine learning regression can deal with the serially correlated predictors derived from the spectral domain if properly tuned (Dormann et al., 2007). However, these algorithms may pose challenges when dealing with serially correlated observations from spatially dependent plant traits and remote sensing data (Rocha et al., 2018).

Likewise, as pairs of proximate wavelengths tend to be similar, locations close together are also expected to present more similar plant trait values than locations further apart (Tobler, 1970). Autocorrelation in the spectra provokes multicollinearity, increasing the risk of not identifying important variables (type II error). Conversely, violating the assumption of independent and identically distributed observations inflates the chance of a type I error (Babcock et al., 2013; Dormann et al., 2013, 2007; Fortin et al., 2012; Legendre, 1993; Wikle and Hooten, 2010). Spatial autocorrelation related to processes that drive plant traits is commonly ignored, assuming the observations as randomly distributed when modelling with hyperspectral data. Thus, ground references are frequently assumed independent in space even where measurements were not taken far apart to be considered free of spatial autocorrelation (Boegh et al., 2013; Lovett et al., 2005). The effect of spatial autocorrelation on variable selection has received little attention, and the implications on model predictions remain unclear (Dormann et al., 2007).

Current practice in model assessment has focused mainly on model fitting and overall accuracy, lacking attention regarding the spatial distribution of model residuals (Moisen and Frescino, 2002; Zhang et al., 2005). The growing recognition of the importance of spatial modelling in statistical fields is

undermined by the fact that countermeasures are often computationally challenging (Bakka et al., 2018). Spatial models fitted by Bayesian inference have become more popular after powerful computational methods such as Markov chain Monte Carlo (MCMC) became commonly available ((Banerjee and Fuentes, 2012; Bivand et al., 2015; Heaton et al., 2017). Despite being very flexible, for complex models or big datasets, MCMC is still computationally demanding and time-consuming, especially when it comes to spatial models (Wang et al., 2018). The development of a computationally efficient alternative to MCMC, the so-called Integrated Nested Laplace Approximations (INLA), creates an opportunity to offer a more friendly approach to fitting spatial models (Poggio et al., 2016; Rue et al., 2009). INLA has been used successfully on a large number of spatial problems (Simpson et al., 2012).

A spatially explicit model to predict plant traits using the full hyperspectral range as covariates is unreasonable given the high complexity. Therefore, under which circumstances one should prioritise either spatial or spectral domains is not a trivial decision. In this study, we assess the trade-off between these two domains while modelling plant traits by controlling the range of spatial dependency across simulated landscapes. The assessment is performed by comparing spatially explicit models using INLA (with and without a spectral index as a covariate) against machine learning algorithms using the full range of hyperspectral data. Prediction accuracy and model generalisation were assessed to evaluate which conditions would favour either approach.

4.2 Methods

The trade-off between spectral and spatial information when predicting plant traits using hyperspectral data were assessed by comparing the accuracy of spatial and non-spatial models under different levels of autocorrelation. The models were fitted using simulated landscapes of vegetation with a low canopy (*e.g.* grass and shrub) represented by layers of plant traits with increasing ranges of spatial dependency (autocorrelation). Non-spatial models were represented by machine learning regressions using the full spectra of 2100 wavelengths and by linear regression using a single vegetation index composed of two bands as a covariate (predictor). Spatial models were fitted using the same vegetation index as a covariate and Bayesian inference with the Integrated Nested Laplace Approximation (INLA) approach. The comparisons focused on prediction accuracy and model generalisation across 15 levels of spatial dependency and 30 different realisations (different spatial patterns) of each level, totalling 450 simulated landscapes. Prediction accuracies for spatial and non-spatial models were assessed taking into consideration the spatial dependency of LAI and the landscape patterns of each realisation.

4.2.1 Data simulation

The simulated data used in this study was performed in the following steps. Firstly, a set of 15 variogram models were determined by increasing the range of spatial autocorrelation along the image extent gradually (Figure 4.1). Then, according to these variograms, plant trait values were generated by unconditional simulation resulting in a regular grid (layers). Hyperspectral data-cubes were simulated by Radiative Transfer Models (RTM) as captured by field spectrometers based on the generated plant traits. Random and spatially dependent noise was introduced into the plant trait LAI (response variable). Random noise per wavelength was also introduced into the hyperspectral data-cubes. Finally, samples were randomly drawn from the LAI layers and hyperspectral data-cubes, arranging in a sequence of locations that minimises the distance to collect them in the field.

4.2.2 Plant traits

Plant trait values were generated by unconditional simulations based on 15 variogram models with increasing spatial dependency, varying from landscapes with no autocorrelation (or independent in space) to approximately 70% of the image extent autocorrelated. The result of these sequential Gaussian simulations of random fields were landscapes represented by a regular grid of 100 by 100 cells. Thirty realisations of each landscape were generated, each representing an exclusive spatial pattern. Landscapes have more similar pattern across levels of spatial autocorrelation within the same realisation (Figure 4.1 – along vertical lines) than within the same level of autocorrelation but from a different realisation (Figure 4.1 – horizontal lines).

Seven different plant traits (Table 4.1) were generated for input into a radiative transfer model to simulate hyperspectral data for each cell of the grid (data-cubes). The plant traits: Leaf Area Index (LAI), Dry Matter Content (Cm), Chlorophyll Leaf Content (Ca+b), Leaf Structure (N) and Hotspot (hspot) were simulated following the above procedure, with values that range between realistic scales for the chosen environment (*i.e.* grasslands). Water Content (CW) was linked by a function with Dry Matter while Carotenoid (Car) was linked with Chlorophyll Content (Jarocińska Anna M., 2014; Vohland and Jarmer, 2008).

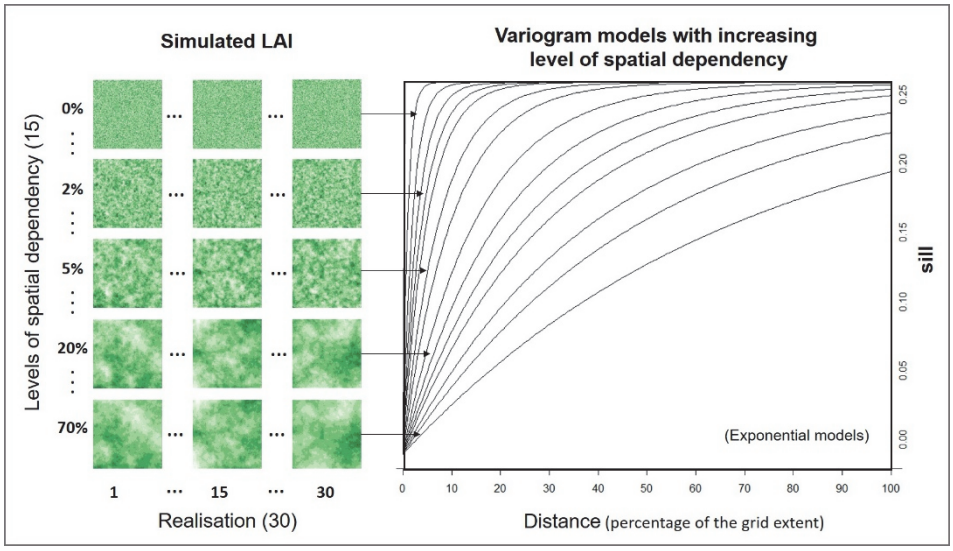


Figure 4.1 - Simulations of plant traits layers: 30 different realisations for each of the 15 variogram models with an increasing level of spatial autocorrelation.

4.2.3 Hyperspectral simulations

For representing the surface reflectance of the 450 hypothetical grassland landscapes and their spatial structures, hyperspectral data-cubes were simulated. A Radiative Transfer Model (RTM) called PROSAIL 5B was adopted to simulate hyperspectral reflectance with a spectral resolution of 1nm, resulting in 2,100 wavelengths ranging from 400nm to 2500nm (Jacquemoud et al., 2009). Leaf Area Index (LAI) was selected as the response variable to be predicted with the simulated hyperspectral data. LAI is an important proxy to describe vegetation structure and function, defined as one half of the total surface area of green leaves projected horizontally per unit of ground area (Chen and Black, 1992; Woodgate et al., 2015). A set of 13 parameters required by PROSAIL models was defined to simulate hyperspectral data as captured by a spectrometer from grassland surface under natural sun lighting (Table 4.1).

Additionally to the seven plant traits cited in the previous session, other six PROSAIL parameters were also generated as presented in Table 4.1. Brown pigment (Cbrown) and soil moisture factor (psoil) were fixed as zero, considering only the occurrence of green leaves and a homogeneous distribution of moisture across the landscape respectively. Leaf Angle Distribution (LAD) was considered as Erectophile (or 90o), assuming as the main leaf orientation for grassland. The parameters that describe illumination and observation geometry were generated from a uniform distribution ~U(min, max), varying the view and the solar angles as expected in a filed campaign with a spectrometer capturing radiance under natural sunlight.

Table 4.1 - Parameters used for PROSAIL 5B to simulate hyperspectral data from grassland landscapes.

	Parameter	Description (unit)	Distribution	R2 with LAI
Leaf	Cab[1]	Chlorophyll a+b concentration (ug/cm^2)	~N(28,4.5)	0.36
	Car[2]	Carotenoid concentration (ug/cm^2)	~N(5,0.7)	0.35
	Cbrown[3]	Brown pigment (-)	0	-
	Cm[1]	Dry matter content (g/cm^{-2})	~N(0.004, 0.0005)	0.69
	Cw[2]	Equivalent water thickness (cm)	~N(0.016, 0.002)	0.66
	N[1]	Leaf structure parameter (-)	~N(1.5, 0.12)	0.48
Canopy	LAI[1]	Leaf Area Index (-)	~N(3.1, 0.6)	-
	hspot[1]	Hotspot parameter (-)	~N(0.05, 0.01)	0.50
	LAD[3]	Leaf angle distribution (attribute)	Erectophile (90º)	-
	psoil[3]	Dry/Wet soil factor (-)	0	-
Geometry	tto[4]	View zenith angle - VZA (degree)	~U(0,5)	-
	tts[4]	Solar zenith angle - SZA (degree)	~U(30, 38)	-
	psi[4]	Relative azimuth angle (degree)	~U(0,360)-U(129,252)	-

Note: 1 assuming a normal distribution ~N(mean, standard deviation). 2 A function of another plant trait: Car=Cab/5 and Cw=4/Cm-1. 3 The same value for all 450 landscapes simulated. 4 Assuming a uniform distribution ~U(min, max). View angle deviation from the nadir (tto), variations on sun altitude (tts) relative to azimuth (psi) during the situ measurements using a field spectrometer over sunlight illumination.

As PROSAIL are fully deterministic models, random and spatially dependent noise was introduced into the LAI landscapes before simulating spectra and before sampling LAI values used as the response variable. The noise added before simulating spectra aim to reproduce variations captured by the sensor during the field campaign. A different realisation of random and spatially dependent noise was added to the LAI landscapes before sampling the response variable values, assuming that it is unlikely that the spatial structure for LAI measurements and spectral data will be completely identical. After the spectral simulation, to represent measurement variability, random noise per wavelength based on a normal distribution was introduced into the hyperspectral data. The mean and standard deviation parameters for each normal distribution were extracted from an independent experiment with a field spectrometer ASD FieldSpec® 3 (Inc., Boulder, CO, USA) to assess the reproducibility of a spectrometer per waveband when measuring grassland reflectance under natural illumination repetitively.

4.2.4 Model selection

Hundred random locations (pixels) were sampled from the layers of LAI and from the respective hyperspectral data to fit predictive models (training sets). Another set of hundred (different) locations was randomly sampled to assess the performance of the fitted models (testing sets).

These sets of random locations were organized through a "path" that optimise the travel distance to collect all the observation points. The set of observations organised in this sequence was used for training and assessing models performance. The residuals along the path can then be assessed to detect spatial or temporal autocorrelation in the field campaign observations. The average distance of two consecutive locations considering the two hundred points (training and testing set together) was around 5.5% of the total extent of the landscape, while for both the training and test sets separately, the average distance was 8.5%.

4.2.5 Non-spatial model

The machine learning algorithms such as Support Vector Machine Regression (SVMR) and Partial Least Squared Regression (PLSR) were selected to represent non-spatial model because they are commonly applied for modelling with hyperspectral data (Carvalho et al., 2013; Feilhauer et al., 2015). These algorithms were also selected by the satisfactory performances under multicollinearity and with high dimensional data. However, to reduce the risk of overfitting, the level of model complexity should be limited by tuning the parameters as "number of components" in PLSR and "cost" in SVMR algorithms. The machine learning final models were selected by model complexity that minimises the Root Mean Squared Error (RMSE) when tuned with 10-fold cross-validation (randomly repeated ten times), (Hastie et al., 2009). A univariate linear regression (*i.e.* ordinary least square) with a vegetation index as a covariate was fitted to enable comparison with the performance of machine learning. The vegetation index used was the LAI Determining Index (LAIDI), which is a ratio of two wavelengths (*i.e.* 1050 nm and 1250 nm) from the near Infrared domain (Delalieux et al., 2008). The combination of these two specific narrow bands was selected from literature (*i.e.* Delalieux et al., 2008), rather than search from the entire range of wavelengths by a supervised method that chooses bands that produce the highest correlation with the plant trait.

4.2.6 Spatial model

The sampling from the artificial landscapes no longer represents a regular grid. It is now represented by a stochastic process defined on a continuous domain, and if we draw another sample under the same conditions, we end up with a

different realisation of the same stochastic process. Assuming LAI to be normally distributed (which is known in this case), we have a continuous Gaussian Field (GF). The area corresponding to the landscape was divided into a number of non-overlapping triangles to simplify computation. This efficient representation of a spatial dependency structure is called a mesh, and it was obtained via a constrained Delaunay triangulation (Figure 4.2).

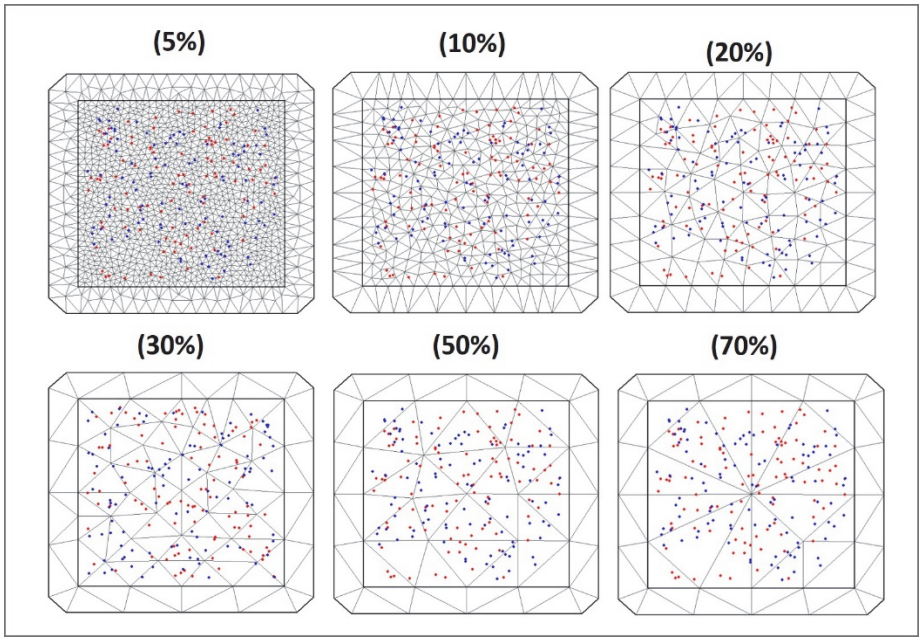


Figure 4.2 - Meshes with a maximum length of the triangle vertices from 5% (top left) to 70% (bottom right) of the extent. The area beyond the first box line is an outer area (buffer) to avoid edge effects. Blue points represent the training set, and the red points represent the testing set.

In this study, the area of a mesh was defined based on the extent of the landscape (*i.e.* the grid used to simulate the landscape), rather than the sample locations. Firstly, this guarantees that all the different random samples selected from any landscape are placed inside the mesh boundaries. Secondly, this represents the landscape more uniformly, avoiding a mesh designed to a specific sample location, which will be denser where there are more points available. The number of triangles was controlled by limiting the largest edge length allowed in the mesh. An outer area (buffer zone) in the mesh was used to reduce edge effects.

Different numbers of vertices were assessed until the density of the mesh brings no further improvement in model performance. Adding more vertices increases computation time and causes model overfitting. Based on the number of triangles (*i.e.* vertices) of the mesh, a neighbourhood area was

defined to account for spatial dependency in the model. The distribution of the weights (w) for each sample location is Gaussian, with Markov properties determined by the triangulation, leading to a sparse precision matrix (Ingebrigtsen et al., 2014). The process of selecting an optimal mesh (with associated model) is comparable to choosing a variogram model (Poggio et al., 2016). Small and regularly shaped triangles give a smaller prediction error, but the error associated with the discretisation of the continuous field (*e.g.* mesh) also depends on the range of spatial autocorrelation and smoothness of the random field (Bakka et al., 2018).

The spatial models were fitted using the Integrated Nested Laplace Approximation (INLA) approach available in the R package R-INLA. This model accounts for the spatial dependency using a mesh to represent the Matérn function and the default settings of priors on the hyper-parameters (Bakka et al., 2018). INLA is a deterministic approach that is more computationally efficient for a complex model than simulation methods using an iterative process for converging, such as Markov Chain Monte Carlo (MCMC). Spatial Latent Gaussian Models (LGMs) as fitted by INLA imply the use of Gaussian priors (Ingebrigtsen et al., 2014). In order to obtain similar results from classical and Bayesian inference for the non-spatial model, we will use a non-informative prior to avoid any influence on the result. Otherwise, a prior could be easily defined based on the LAI distribution and used for simulating the data (Table 4.1). Spatially-explicit models were fitted by INLA with the same vegetation index (LAIDI) as a covariate (or fixed effect) using a stochastic partial differential equation (SPDE) approach (see Appendix 4A for more information about INLA and SPDE).

4.2.7 Model assessment

The decision to use Bayesian inference was taken because, for spatially dependent structures, the frequentist methods are rather limited (Dormann et al., 2007). The differences in the interpretation of confidence intervals, as well as whether the model parameters are considered fixed and known, or unknown stochastic, quantities, are left out as the intention is to compare spatial and non-spatial models rather than inferential approaches. To avoid differences in metrics of model assessment, comparisons of prediction accuracy between machine learning algorithms, spatial models and linear models fitted by classical or Bayesian inference will be performed using root-mean-squares error (RMSE).

The final model selected for each regression method was assessed regarding the capacity to generalise for unseen locations within the landscape (testing set) with prediction accuracy comparable to the training model. The capacity to generalise to another realisation (landscape) was also assessed and was

presented as a "new realisation" set. Differences in accuracy between the training set and the testing set were used as an indication of overfitting (Rocha et al., 2017). When this difference increases, a model is more likely to be overfitted. Also, the residuals from all models were tested for autocorrelation by using the Durbin Watson test, considering the sequence between sample locations in the order that minimises the travel distance. The platform R version 3.2.2 (The R Foundation for Statistical Computing) was used for executing the data simulation and modelling. The unconditional simulations were performed by the package gstat, while spectra simulations using a PROSAIL 5B by hsdar. Machine learning regression was tuned by the package Caret and R-INLA package was used for fitting spatial models by the Bayesian inferential approach.

4.3 Results

4.3.1 Model accuracy and generalisation

Model accuracy starts to differ considerably between spatial and non-spatial model types when the spatial correlation is higher than 5% of the landscape extent, regardless of which data set was used for validation (Figure 4.3). The linear model shows no strong response to spatial autocorrelation and presents similar accuracy across the different datasets, suggesting a better generalisation than machine learning methods when predicting using an unseen data set. When the model is used for prediction based on datasets belonging to a different landscape (*i.e.* realisation), linear models present smaller errors (RMSE) than any other model when applied to landscapes with spatial autocorrelation levels of more than 5% of the landscape extent. This confirms the principle of parsimony (Kuhn and Johnson, 2013) that simpler models (*i.e.* with fewer parameters) can be generalised with higher accuracy than more complex models can.

Machine learning regressions, given their high complexity (*i.e.* a lot of parameters), often underestimate errors for the training set (compared with the test set). The test set has the same underlying spatial structure at similar sampling distances compared with the training set, and the two machine learning methods (PLSR and SVMR) show the greatest differences in RMSE between training and validation set. When predicting for a different realisation, the precision reduces considerably, and the error becomes unstable (wider confidence interval among the 30 realisations) for all tested models apart from the non-spatial linear model. The effect becomes stronger at higher levels of spatial autocorrelation.

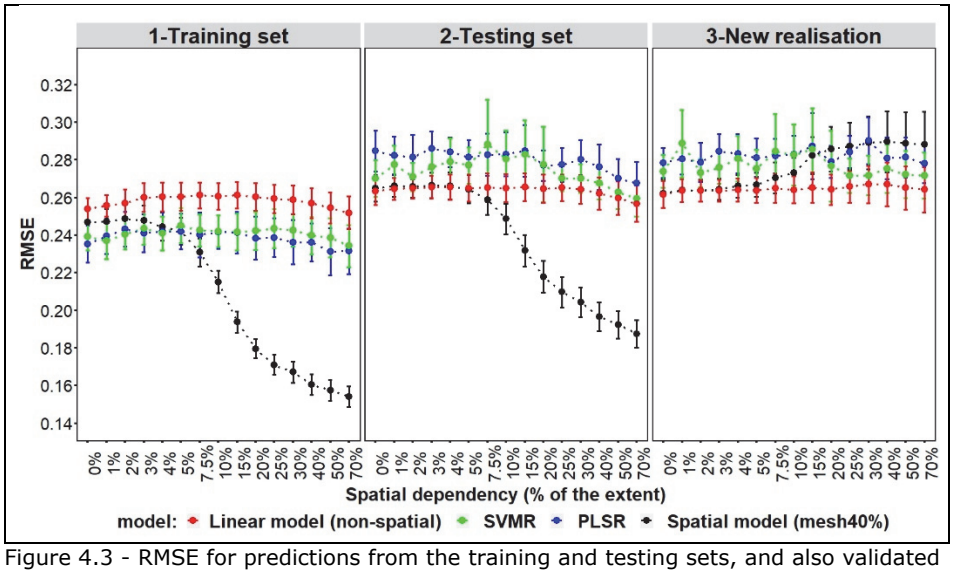


Figure 4.3 - RMSE for predictions from the training and testing sets, and also validated in a new realisation from the same spatial dependency. Four models are assessed: linear models (red circles), the machine learnings SVMR (green circles) and PLSR (blue circles), and a spatially explicit model with a mesh of 30% and a vegetation index as a covariate (black circles).

The spatially explicit model presents more accurate predictions in the cases where the spatial dependency is higher relative to the distance between the sequenced sample locations (*i.e.* spatial autocorrelation with a range larger than 7.5% of the extent and onwards; Figure 4.3), provided the test dataset originates from the same landscape. The RMSE for the spatial model predictions estimated from the test sets at the highest levels of spatial dependency is reduced, on average, by 28% compared to predictions from SVMR. However, the spatial model generalises poorly. When used to predict in another realisation with a different spatial distribution (or pattern), the RMSE of the spatially explicit model rises above all RMSE's of other modelling methods.

4.3.2 Tuning spatial parameters

Using shorter distances to define the triangulation in the mesh (*i.e.* higher-order precision matrix) results in lower prediction errors when estimating based on the training set (Figure 4.4). However, the prediction error from the testing set yields very comparable errors that are much less sensitive to the mesh density. Defining distances shorter than 40% of the extent in the mesh increases processing time considerably, but has a minor effect on prediction accuracy in the testing set (Figure 4.4). The prediction error of the training set drops to near zero when a mesh has a very dense triangulation (maximum vertices of 5% of the extent; Figure 4.4 – black marks). This occurs as the

precision matrix contains an excessive number of parameters (*i.e.* spatial weights). A mesh of 5% of the extent contains more than a thousand spatial parameters (*i.e.* triangle vertices), but the model is supported by only a hundred observations in the training set. A mesh of 70% of the extent contains no more than 50 spatial parameters, presenting a less complex model. Using the mesh of 70%, the ratio between the RMSE values from the training set and the testing set is kept roughly around 90% for all levels of spatial dependency. For the mesh of 5%, this is on average 52% but decreases to 35% when the spatial dependency is 7.5% and 10% of the extent.

The reduction in the number of spatial parameters from the densest mesh (5%) to the most sparse mesh (70%), decreases the RMSE from the testing set by only 4.8%, while the RMSE drops 43% in the training set for the stronger spatially correlated landscapes. Both meshes are probably not the best choice as the first is clearly over-parameterised and overfits the model, while the second does not capture the spatial dependency properly as it does not correct the residuals for autocorrelation (Appendix 4B). A mesh between 30% and 50% of the extent yields a similar accuracy between training and testing sets, being a choice that presents low prediction error without excessive model complexity and overfitting. The density of the mesh does not only depend on the spatial dependency of the landscapes but also strongly on the distance between the sampling locations.

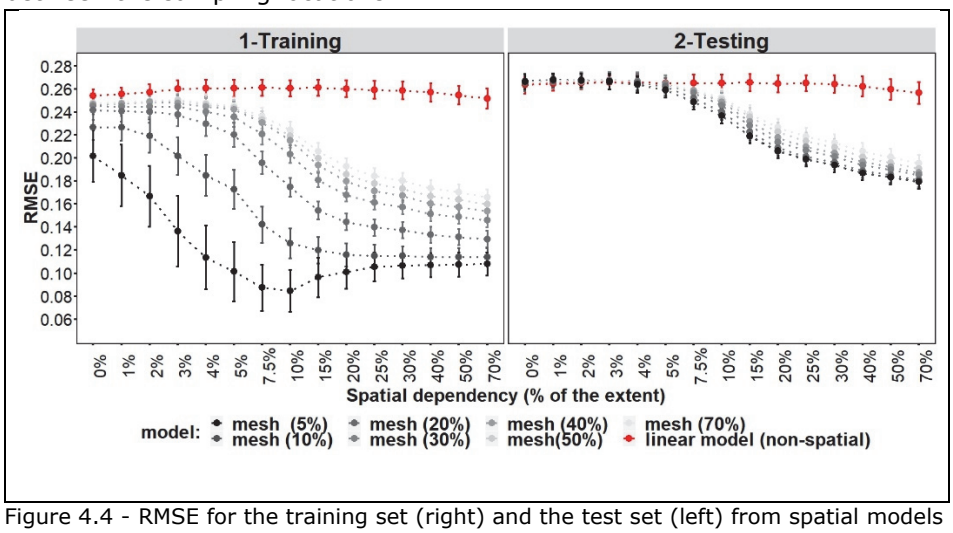


Figure 4.4 - RMSE for the training set (right) and the test set (left) from spatial models fitted on seven different mesh densities compared with the linear model with the spectral vegetation index.

4.3.3 Model residuals

The residuals of a "mean model" (using mean LAI as the predictor) show a lack of spatial independence, confirming that observations in the random samples

are autocorrelated in the scenarios with spatial dependency higher than 5% (Figure 4.5). The non-spatial models (linear models, SVMR and PLSR) present a minimal reduction in the spatial autocorrelation of the residuals compared to the mean model. Spatial models reduce the autocorrelation in the residuals significantly regardless of the spatial dependency (Figure 4.5). However, for the densest meshes, the autocorrelation changes from positive (DW values < 2) to negative (DW values > 2), mistakenly suggesting that residuals from pairs further apart are more similar than residuals from close pairs.

An indication of whether a mesh is appropriate to a specific spatial dependency can be provided by the behaviour of the residuals from a model fitted only with the spatial domain (without the vegetation index as a covariate). For instance, with a mesh of 30%, the model produces residuals that are slightly negatively autocorrelated at values of 10% of spatial dependency or larger. Among the assessed meshes, a mesh of 40% seems to account for the autocorrelation more efficiently as it seems to eliminate spatial dependency from the residual regardless of the spatial dependency in the landscape. For less dense meshes, the autocorrelation remains positive, while models produce negatively autocorrelated residuals for denser meshes.

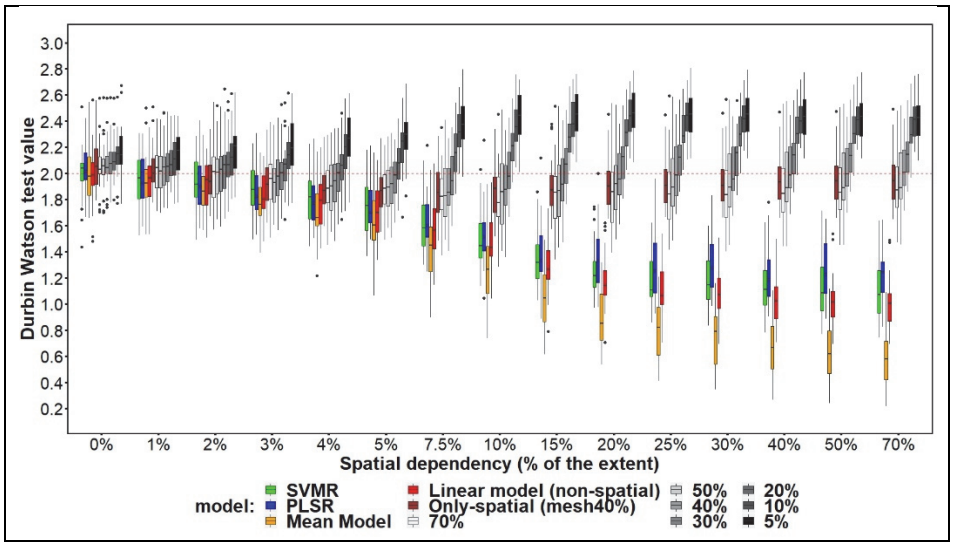


Figure 4.5 - Boxplot for the Durbin Watson test calculated from the residuals of the training model for different regression models and mesh densities in each of the 30 realisations. The red dashed line symbolises the theoretical threshold for the independence of the residuals (DW=2). Values significantly higher than 2 represent negatively autocorrelated residuals (*i.e.* the further away from the more correlated) and lower than 2 indicates positively autocorrelated residuals (*i.e.* the closer the more correlated).

4.3.4 Trade-off between spatial and spectral domains

The trade-off between using spatial or spectral domains in predicting the response becomes clear when comparing the fitted models with a model that uses only the mean of the response variable as prediction (mean model) and a model that uses only the spatial domain (Figure 4.6). Models using only the spatial domain reduce the prediction error compared to mean models for landscapes with autocorrelation stronger than 5% of the extent, but even under strong spatial dependency, they are still less accurate than the models based only on spectra. The combination of spatial and spectral domains generates models that produce the most accurate predictions when estimated from the same landscape. Therefore, a feature selected to reduce the number of bands (predictors) in the hyperspectral data, and a definition of a neighbourhood that restricts the number of spatial weights in the model, are both necessary to decrease dimensionality. The more complex a model is, the lower the generalisation of the model will be (*i.e.* applying it to another landscape/realisation), especially under significant spatial dependency. However, the spatial complexity should be sufficient to eliminate the autocorrelation in the residuals. When the residuals do not present an indication of spatial autocorrelation, the model may be based solely on spectra, and if possible, using only a few bands of the spectra that are known to be related to the plant trait of interest.

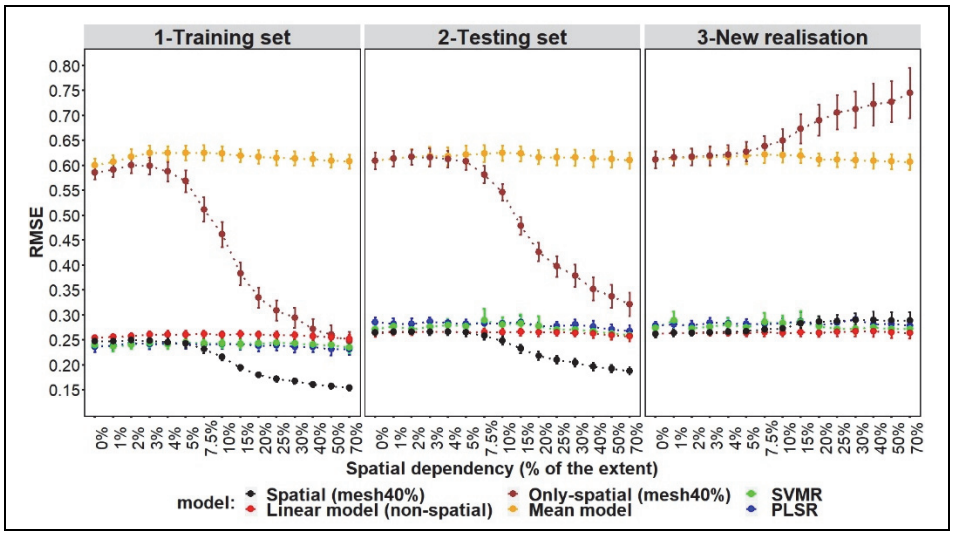


Figure 4.6 - Trade-off between spectral and spatial information to predict plant trait. RMSE for model predictions from the training, testing and new realisation sets for the mean model (orange) compared to spatial models: without covariate (only-spatial - brown) or with covariate (spatial - black), and to non-spatial models: linear model (red) or machine learnings SVMR (green) and PLSR (blue).

4.4 Discussion

4.4.1 Modelling practices and assumptions

Spatial autocorrelation is rarely taken into account when predicting plant traits from hyperspectral data, despite a common understanding that, in continuous fields under natural conditions, spatial dependency is more likely to be the rule than the exception (Legendre and Fortin, 1989). In other words, ground references measurements are often described as randomly assigned locations, though most publications do not state explicitly that the data were collected in that way. For practical reasons in situ measurements are often taken in a sequence that reduces the distance and time to collect the data, rather than samples being collected truly in random order. These practices increase the possible dependence between observations, which in such cases are an aggregation of spatial and temporal autocorrelation. Model validations in remote sensing publications also rarely mention any inspection of the residuals. An exception is presented by Wang et al. (2014), who indicated that the model used to estimate aboveground biomass of grassland using hyperspectral indices presented strong spatial autocorrelation in the residuals according to Moran's I index.

Our study demonstrates that spatial dependency affects non-spatial models in general, but complex models such as machine learning algorithms appear to be more susceptible to producing unreliable accuracy estimates. These models treat spectral bands as independent and identically distributed across wavelengths, space and time, even though these three domains are often serially correlated (Curran, 1989; Tobler, 1970). Given the high dimensionality in hyperspectral data, plant traits are commonly mistaken as randomly distributed in the study area and invariant within a specific period. The assumption of a random distribution may be valid depending on the distance between sampling locations but is probably rare in remote sensing imagery (Dalposso et al., 2013; Griffith and Chun, 2016). Even if there is no intention to predict a plant trait for the entire landscape (or remotely sensed image), autocorrelation in model residuals should be assessed and reported for any regression approach when samples are extracted from continuous fields in a specific order.

4.4.2 Spectra-space trade-off

Despite the recognition of the importance of spatial and temporal autocorrelation in ecological processes over the past decades (Gelfand, 2012), it is usually not incorporated in empirical models because of dimensionality problems (Gelman and Hill, 2006; Kuhn and Johnson, 2013). Spatial structures may raise serious concerns about the model residuals' distribution and the

reliability of predictions when only relying on spectral data, missing the opportunity to extract valuable information from the spatial dependency of the plant trait. Environmental and topographic information that explains this dependency is often not available as explanatory variables (Dormann et al., 2007). In the absence of spatial autocorrelation, non-spatial models using only spectral data may be more suitable to predict plant traits. However, as demonstrated in this study, in the case of significant autocorrelation, it is better to reduce the spectral domain and include space explicitly in the model, rather than ignore the spatial component and use the full spectrum. Models using only spatial information to predict LAI result in independent residuals, but without the spectral covariate, they present lower accuracy than spectral-based models.

Another benefit of reducing the spectral domain is that it diminishes the risk of spurious correlations due to multicollinearity and overfitting with random errors of redundant predictors. The argument that machine learning algorithms are robust under multicollinearity and that overfitting can be controlled are not supported in this study, as is shown by the differences between training and test accuracies. These issues probably affect many studies using hyperspectral remote sensing without being noticed (Figure 4.4). This was indicated by a lower RMSE for the training models, compared to the RMSE for the testing sets for these methods. Using a tuning process based on cross-validation, the number of observations to train the model is often insufficient to support large numbers of spectral bands as predictors to avoid model overfitting (Rocha et al., 2017). However, our results also show that the negative effect from overfitting is still less than the gain in accuracy that is achieved, compared to models that only include space or models that are based on just an average of the response variable.

Even for simple linear regressions with a single band as a covariate, selecting predictors through a supervised method (*e.g.* stepwise) carry the same risk of overfitting (Thenkabail et al., 2000). Therefore, searching for a combinations of two or more wavelengths to create a vegetation index which shows the highest correlation with the plant trait to use as covariate in an empirical model again has a great risk of overfitting, disregarding how simple the selected model is (James et al., 2013; Thenkabail et al., 2000). Spatial models can also suffer from dimensionality problems (Gelfand, 2012) and therefore be overfitted, increasing the prediction error in response to small alterations in the spatial pattern. This overfitting may change the trade-off between spectral and spatial model contributions, as presented in this study (Figure 4.4). For this reason, a tuning process to select an optimal mesh density (or another discretisation method) for the landscape under consideration is necessary. This can be achieved by testing different densities of the mesh or neighbourhood matrix using only the spatial component of the model. The optimal mesh

density should be selected based on the spatial complexity that can correct for most of the autocorrelation in the model residuals while minimising the prediction error in the testing set. After selecting the density of spatial weights that best represent the autocorrelation in the landscape, different covariates can be tested to improve the model. If the final model, with covariates, presents autocorrelation in the residuals, the mesh has to be adjusted again, increasing the density of the mesh in the case of positive correlation and decreasing the density in response to a negative correlation.

4.4.3 Remote sensing and environment applications

The process of predicting vegetation dynamics has changed since remote sensing data first became available, but uncertainty in modelling these processes will always remain. The advancements in remote sensing technology create even more possibilities to observe the dynamics of vegetation over a range of spatial and temporal scales. Observations with optical sensors, like any other measurement system, will always be susceptible to errors. Regardless of the platform deployed and the resolutions captured (spatial, temporal and spectral), certain levels of random and systematic error will be part of the observation values obtained. Any reflectance or emittance will capture more than vegetation radiance characteristics alone, including variation in illumination and view geometry, weather conditions, soil background and many other aspects. Such variation will hardly ever be independent in time as, for instance, sun altitude changes gradually during the day and over the year (Kumar and Skidmore, 2000). This will also hold for observations from spaceborne platforms, where geometric distortions or variations are driven by patterns in soil background and cloud cover.

Narrow and abundant, bands from hyperspectral data can yield higher accuracy when predicting plant traits comparing with broad and limited to the visible spectra commonly available for satellite image (Curran, 1989). In some cases, this higher accuracy can be an artefact created by the combination of a large set of spectra with a spatiotemporal autocorrelated plant trait. This may result in models with a low capacity of generalisation, overfitted by spurious relations with random noise or systematic patterns in the observations. Spatiotemporal structures also affect ground references, and when represented by continuous variables, these measurements should not be treated as "ground truth", but as a stochastic process. It is impossible to measure all the ground references at the same time, and it is known that nature will never repeat a process under the same conditions. Thus, when several readings are taken from the same location, different values are expected (for a continuous variable), even if the measurement system was totally free of error.

4.5 Conclusion

Accounting for the spatial domain increases prediction accuracy substantially when there is significant autocorrelation in the plant trait observations under consideration. Whether the autocorrelation can be considered strong enough to justify a spatial model depends on the relation between the spatial dependency present in the (continuous) field and the sampling density used to collect ground references for training the model. If the spatial dependency is negligible, machine learning algorithms or linear models can be fitted, providing similar accuracy to a spatial model, but with less effort. Machine learning methods, however, should be properly tuned to avoid overfitting and used only if the empirical relation between the plant trait and spectral regions is unknown. Otherwise, a less complex model such as ordinary regression is advisable to increase generalisability. Spatial models were not generalisable for landscapes with different patterns, limiting the capacity to predict at completely new locations using these models.

Appendix 4A

Plant trait values over a set of locations can be described as a spatial process under a Gaussian Field (GF) framework where values tend to depend more on the vegetation nearby than to on vegetation far away (Bivand et al., 2015). The spatial dependency in a GF can be expressed by a covariance function (*e.g.* a Matérn correlation function) that gives the strength of the dependency between two locations (Bivand et al., 2015; Ingebrigtsen et al., 2014). A challenge with this common modelling approach to capturing spatial dependency is that a GF can become impractical on a large database (Bakka et al., 2018). Given the high dimensionality, a full covariance matrix to account for spatial structures is computationally intensive when it has to consider the correlation between all pairs of locations (Banerjee and Finley, 2007).

However, instead of using a GF with a full covariance matrix, the computations could be carried out with the properties of Gaussian Markov random fields (GMRFs) using a sparse matrix (Ingebrigtsen et al., 2014; Poggio et al., 2016). GMRFs form a class of Gaussian Fields that are discretely indexed, simplifying the representation of space and facilitating the numerical calculations by considering non-neighbour elements to be zero (Lindgren et al., 2011; Wang et al., 2018). The result is a more sparse structure to account for the spatial dependency than a covariance matrix called precision matrix (Simpson et al., 2012).

The spatial sparsity structure for the precision matrix is obtained by using a mathematical approach (stochastic partial differential equation or SPDE) to link the Gaussian fields to the GMRFs (Lindgren et al., 2011). It is important to remember that in this case, the model represents real-world phenomena that exist independently of whether or not they are observed in a given location (Lindgren and Rue, 2015). Thus, the model is not solely built for discretely observed data location or a grid, but approximate the entire processes defined on continuous domains (Lindgren, 2012; Lindgren et al., 2011). The SPDE approach makes it possible to represent a continuous spatial process by a discretely indexed spatial random process, and therefore gaining computational efficiency (Bakka et al., 2018; Lindgren et al., 2011; Rue and Held, 2005). The solution to the SPDE can be approximated using a finite element method with a basic function representation defined on a triangulation of the area of interest (Wang et al., 2018).

Appendix 4B

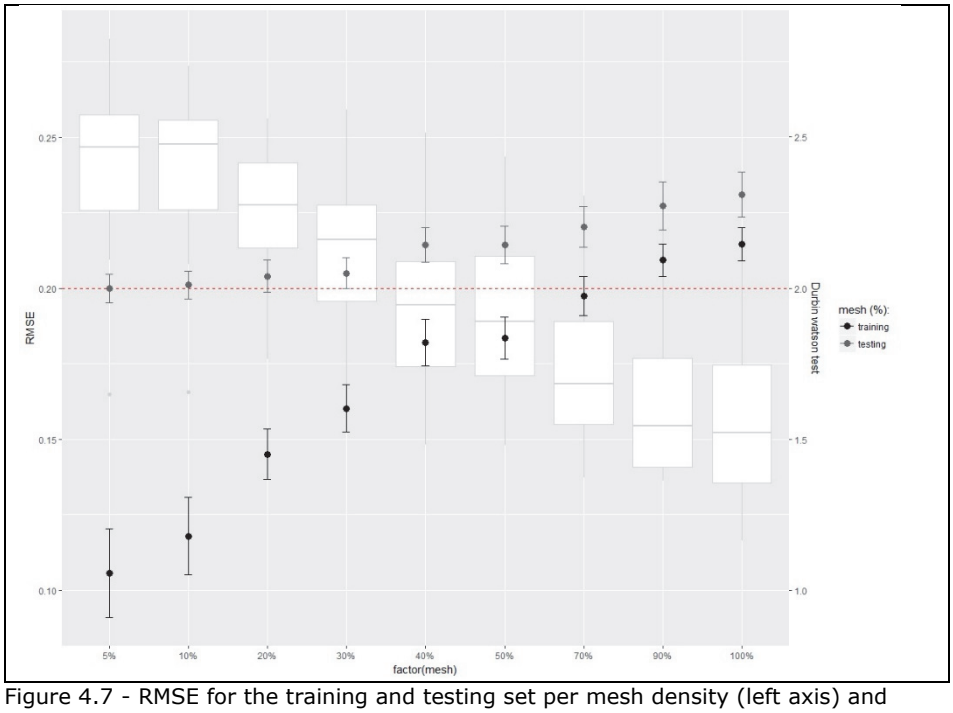


Figure 4.7 - RMSE for the training and testing set per mesh density (left axis) and Durbin Watson test values for the model residual per mesh density (right axis).

Chapter 5

Choosing a sampling design under spatial dependence when predicting plant traits with hyperspectral remote sensing[4]

[4] A modified version of this chapter was submitted in September, 2019 to the journal IEEE Transactions on Geoscience and Remote Sensing.

Abstract

Remote sensing data opens opportunities to estimate the spatial dependency of plant traits in a landscape. This knowledge can be useful to design sampling strategies for fieldwork based on whether the focus should be only for the spectral domain, or it should also consider the spatial domain. This knowledge can support the selection of the spacing between observations, either to capture most autocorrelation in the field or to completely avoid it, depending on the aim. In this study, we show the effects of different sampling designs predictions from autocorrelated plant traits using a set of simulated data with an increasing range of spatial dependency. When the sampling is designed to estimate a global parameter such as the mean and variance of the population, a random design is appropriate even where there is strong spatial autocorrelation. But in remote sensing applications, the aim usually is to predict, for unsampled locations, using only spectral information. In this case, regular or systematic sampling may offer a more efficient design. The use of close pairs of points clustered over a regular sampling design may improve the training model accuracy but generalise poorly for test samples. The hyperspectral dimension has to be drastically reduced to be modelled spatially but improves prediction accuracy significantly when used with machine learning or ordinary regression, under strong autocorrelation. Spatial models predict with similar accuracy within the training area (testing set), but the model lacks generalisation to extrapolate to landscapes with a different spatial pattern. The design and size of the sample have a strong influence on the spacing between observations. Therefore, it affects not only the ability to capture or avoid spatial autocorrelation, but also increases the variability when the sampling distances are similar to the range of the spatial dependency of the plant trait.

5.1 Introduction

Remote sensing images are measurements of electromagnetic radiation captured by optical sensors, discretising a continuous spectral signal into a specific wavelength region (Manolakis et al., 2003; Ortenberg, 2011). The spectral resolution of a sensor determines the wavelength width measured by each band, defining whether the image is a product of few broader bands or a very large number of narrow bands such as hyperspectral data (Ortenberg, 2011). One of the main application of hyperspectral remote sensing is to make indirect estimations of ecological processes such as biochemical and biophysical properties of vegetation (Kokaly et al., 2009; Lee et al., 2004; Skidmore et al., 2015). With remote sensing, ecosystems can be monitored over wide temporal and spatial scales for a variety of biochemical and biophysical properties, including primary plant traits as nitrogen concentration, leaf area index (LAI) or biomass (Stroppiana et al., 2011; Wilson et al., 2011).

Spectral data captured from landscapes of continuous vegetation are a product of interactions from physical, chemical and biological properties of the plant surface (Curran, 2001). A physical or empirical relationship needs to be established to relate the radiation measurements (mostly reflectance, but see Buitrago et al., 2018 for examples of emissivity), with the vegetation properties. The physical relation between plant traits and reflectance remains a challenge in practical applications where a heterogeneous surface is observed using the sun as the primary source of illumination (Combal et al., 2002). In this case, it is often unfeasible to measure or control all parameters required to use a deterministic model based on spectral radiance (Jacquemoud et al., 2009). Therefore, applications for estimating plant traits with remote sensing rely frequently on an empirical relationship (Combal et al., 2002; Goodenough et al., 2006).

For establishing these empirical relationships, it is common to fit regression models as both spectra and plant traits, are continuous variables (Kokaly et al., 2009). The simplest method to predict plant traits is an ordinary least squares regression with a spectral index based on a coefficient of two wavelengths as a covariate (Li et al., 2011b; Thenkabail et al., 2000). Machine learning algorithms such as Partial Least Squares Regression (PLSR), Support Vector Machine (SVM) or Artificial Neural Network (ANN) are also frequently used, especially with hyperspectral data. These algorithms consider the entire range of the wavelengths as covariates to fit a model (Feilhauer et al., 2015; Moisen and Frescino, 2002; Ramoelo et al., 2013). Wavelengths captured by hyperspectral sensors have very strong multicollinearity because bands are serially correlated (Dormann et al., 2013; Rocha et al., 2018). Also, considerable noise can be captured in specific regions of the spectrum, depending on the capacity of a sensor's platform to control variations in illumination and view geometry (Combal et al., 2002). These two characteristics of hyperspectral data may create a spurious correlation that affects model accuracy and the ability for generalisation (Manolakis et al., 2003).

Machine learning algorithms when tuned to restrict model complexity are capable of dealing with serially correlated hyperspectral wavelengths (Dormann et al., 2007; Rocha et al., 2017). However, non-spatial regression approaches face challenges when modelling with spatially autocorrelated observations derived from plant traits or remote sensing data (Rocha et al., 2018). Spatial autocorrelation is often neglected even when the in situ plant trait measurements are exceptionally close to each other, yet in analyses are considered as randomly distributed observations (Hoeting, 2009; Legendre et al., 2004). Reflectance captured from a continuous area of vegetation is likely to exhibit significant spatial dependency, regardless of sensor, platform, spatial resolution or type of ecosystem (Dormann et al., 2007; Fortin et al., 2012).

Standard statistical inference techniques, such as (non-spatial) regressions, were primarily designed assuming that the observations are drawn from independent and identically distributed (i.i.d.) random variables (Dutilleul, 1993). The problem is that this condition is violated by spatially autocorrelated data from dependent plant traits or remote sensing images (Cochran, 1977; Legendre and Fortin, 1989).

For this reason, there is a growing recognition of the importance of spatial modelling in remote sensing and ecology (Gelman et al., 2001). Spatial models can deal with observations that are not independent nor identically distributed, and use this information (*i.e.* autocorrelation) to improve prediction accuracy (Dormann et al., 2007). Despite being more computationally demanding and time-consuming, spatial models have become more commonly available with the use of methods such as Markov Chain Monte Carlo (MCMC) to fit Bayesian inferences to a model (Bakka et al., 2018; Banerjee and Fuentes, 2012; Heaton et al., 2017). For complex spatial models or big datasets, MCMC is still demanding, but a computationally more efficient alternative, the so-called Integrated Nested Laplace Approximation (INLA), creates an opportunity to fit such models (Wang et al., 2018; Poggio et al., 2016; Rue et al., 2009). Different scientific fields have been used INLA to model spatial problems with satisfactory results (Simpson et al., 2012).

An empirical model, spatial or not, has to be trained with ground references observations that represent the remote sensing data in space and time (Brus and de Gruijter, 1997). Therefore, a sampling design strategy is needed to estimate, with satisfactory precision and accuracy, population parameters, or to predict plant trait at unsampled areas using spectra data (Wang and Gertner, 2013). In the first instance, it is necessary to define an appropriate plot size and shape, which is compatible with the spatial resolution of the remote sensing pixel or area (Atkinson and Emery, 1999; Dutilleul, 1993). In the second instance, a design needs to be created that provides a sampling distance compatible with the spatial autocorrelation range, which is a combination of the sample size and the spatial distribution of plots in the ground (Legendre et al., 2004; Wilson et al., 2011). Spatial autocorrelation affects the sampling efficiency, as the variance of the estimation error might be biased depending on the choices of sample size and design (Haining, 2003; Ripley, 1981).

Intuitively, in a completely homogeneous landscape or a perfect spatially autocorrelated scenario, a sample size equal to one should be enough to infer the entire population, as there is no variability (Ding et al., 2014). Analogically, the number of sampling units can decrease if the spatial autocorrelation increases, when the goal is to estimate (globally) the population mean (Ding

et al., 2014). However, the sample size may need to be greater in the case where the aim is to estimate (locally) at unseen locations (Haining, 1988).
A possible complication frequently encountered when sampling environmental variables that need to be linked with remote sensing data is to obtain an accurate and compatible measurement unit (Atkinson and Emery, 1999; Legendre et al., 2004; Legendre and Fortin, 1989; Stevens and Olsen, 2004).

Remote Sensing images are taken over a finite but continuous area and represented by a grid of pixels (Manolakis et al., 2003). However, the pixel units are arbitrary discretisations as an infinite number of spectral point could be taken from the area (Wang and Gertner, 2013). Extended over large continuous regions, the population captured in a remote sensing image (*i.e.* grid) may include a substantial portion of non-targeted, mixed or disconnected elements, lacking a natural measurement units to represent the target plant trait (Manolakis et al., 2003; Milton et al., 2009; Stevens and Olsen, 2004). Also, areas that are intended to be sampled might be inaccessible because of physical location, safety, or lack of access permission from the landowner (Vallejos and Osorio, 2014). When a population is finite, it means that all the units (pixels) can be indexed by a set of integer numbers (Plant, 2012). In contrast, for the same area, ground references of leaf chlorophyll content can be considered infinite as it varies continuously and cannot be indexed. While, LAI of individual canopies can be considered infinitely countable, meaning that it can be indexed by a set of integers (Plant, 2012). For this reason, often the sampling is designed based on a covariate, in this case, the spectral image (*i.e.* grid), rather than on spatial locations of the response variable such as LAI (Cochran, 1977; Plant, 2012).

The alignment between the remote sensing unit and ground references unit should be as similar as possible in dimension, location and time (Atkinson and Emery, 1999). However, depending on the platform carrying the sensor and the method to measure the ground references, both units can vary greatly in their spatial and temporal resolution (Dutilleul, 1993; Legendre et al., 2004; Milton et al., 2009). The sampling unit in remote sensing is often the pixel. However this unit is invisible in the field, and rarely ever the ground references can be measured at the same spatial scale (Atkinson, 1997). Thus, aligning reflectance values from pixels with samples of plant traits from field plots in both space and time is one of the great challenges for monitoring ecological processes by remote sensing (Atkinson and Emery, 1999). A sampling design should guarantee that the ground references data is well aligned with the pixel (*i.e.* spectral measurements), while the spatial resolution of remote sensing data should adequately represent the phenomena of interest (Finley et al., 2014; Woodgate et al., 2012).

A sampling of plant traits needs to be designed to estimate the global mean value for the entire target area to infer about the population (Plant, 2012). In remote sensing applications, this sample should also be designed to allow predicting plant traits at unvisited locations that are covered by the spectral image (Wang et al., 2012). Simple random sampling design is a relatively basic approach, but it is rarely used for validating remote sensing data because of difficulties in positioning the sampling unit when carried out in the field (Fortin et al., 1990). Completely randomised designs applied to populations under the influence of autocorrelation should only be used in the extreme cases when sampling in a spatial homogeneity field at large scale, or in a spatial heterogeneity field at short scale (Dutilleul, 1993). It is generally accepted that regularly spaced designs (or lattices), when sampling for environmental populations, lead to more efficient spatial predictions than completely random designs (Matérn, 1986; Wang et al., 2012; Webster et al., 1989)

Systematic sampling is a regularly spaced design that is easier to conduct in the field and more suitable for detecting spatial autocorrelation, as the distances between successive samples are controlled (Cochran, 1977). The benefits of systematic over random sampling are that it always produces spatially balanced and evenly dispersed units across the area (Stevens and Olsen, 2004). Therefore, it results in more representative coverage, while random sampling may lead to relatively large gaps in the sampled area (Wang et al., 2012). The drawback of a systematic design is that the sampling distance may be miss-matching with the existing spatial structure, leaving the autocorrelation uncaptured (Diggle and Ribeiro, 2007; Matérn, 1986). Also, many pairs of points will be separated by the same distance, which may coincide with a frequency of a regular pattern in the landscape, causing interference (either amplification or attenuation) in the observed correlation (Legendre and Fortin, 1989).

The group of classical sampling approaches composed by simple random, stratified, systematic and cluster samplings are so-called design-based sampling designs (Cochran, 1977; Stehman, 2000). In design-based approaches, parameters are considered unknown but fixed, and therefore the main source of randomness originate from the process of drawing samples (Cochran, 1977; Diggle and Lophaven, 2006). Cluster designs are unsuitable when observations collected according to such a design are used subsequently to estimate values at unvisited locations (Corsten and Stein, 1994). Where the error is overestimated for a regular design as the variance from the sample estimation is inflated by the spatial autocorrelation, it is underestimated for a design with clustered observations (Vallejos and Osorio, 2014). Random or systematic sampling should be chosen if nothing is known about the target domain, although there is the risk of low sampling efficiency when the domain is spatially heterogeneous (Cochran, 1977; Wang et al., 2012). Many authors,

therefore, use remotely sensed images to improve sampling strategies (Atkinson et al., 1992; Wang et al., 2005; Wang and Weng, 2014).

Model-based sampling approaches are based on geostatistical theory and can be considered a combination of purposive sampling and interpolation (de Gruijter et al., 2006). In such sampling strategies, a value at any location is not fixed but random (Diggle and Ribeiro, 2007). It can assume more than one possible value and the probability of occurrence is a realisation of a random variable at the location (Wang and Weng, 2014). This approach is known to be more suitable for the prediction at unseen locations and estimation of parameters from the underlying stochastic model (Haining, 2003). When the plant trait presents strong spatial dependency, and a variogram model is available, model-based sampling should be considered when a reasonable sample size is feasible (Haining, 2003; Wang et al., 2012).

When the aim is to estimate model parameters and make spatial predictions for the unseen locations from the same sampling designed, some compromise may be needed (Chipeta et al., 2016). Some designs attempt this, such as lattice close pairs and lattice plus in-fill designs (Diggle and Ribeiro, 2007). The lattice plus close pairs design consists of locations in a regular spacing, supplemented by some close pairs of points (Diggle and Ribeiro, 2007). It was suggested by Diggle and Lophaven (2006) for improving the performance in estimating autocorrelation functions. The lattice plus in-fill design consists of locations in a regular spacing supplemented by a cluster of adjacent cells for each selected origin (Diggle and Ribeiro, 2007). Both designs present a spatially regular sample, combined with sub-sets of closely located points to capture short distance structures (Diggle and Ribeiro, 2007). Defining optimal sample size in a design-based approach requires knowledge of the population parameters of the variable of interest, while in a model-based approach is necessary to know the variogram to determine the maximum sampling distance (Wang et al., 2005). However, before drawing a sample, the population parameters and the variogram are both often unknown for the target plant trait, but frequently available for covariates coming from remote sensing data (Wang and Weng, 2014; Wang et al., 2012).

Even though statistical considerations are crucial for any inferential study in remote sensing, non-statistical considerations such as cost, time and inaccessibility of locations are often highly influential in determining the sampling design and the sample size (Legendre and Fortin, 1989). Mapping exercises require observations in situ over extensive spatial scales that are collected in a short period, which is demanding and expensive (Muñoz-Huerta et al., 2013; Secades et al., 2014). The sampling design to collect ground references in a continuous and heterogeneous surface should, therefore, rely on methods that avoid time-consuming procedures while preserving model

accuracy and generalisation (Tian et al., 2002). For establishing the empirical relationship between the plant trait observations and the (hyper)spectral measurements is needed to design a sampling that represents both. In remote sensing, the choice of the sampling design should consider the spatial dependency of the plant trait and model approach to be applied (Rocha et al., 2019). The objective of this study is to assess the effects of design-based sampling to predict plant traits with hyperspectral data when using spatial and non-spatial models under different levels of autocorrelation.

5.2 Methods

Different sampling designs and modelling approaches were tested to assess their effects on prediction accuracy while modelling spatially dependent plant traits. These plant traits were artificially simulated in a regular grid using Sequential Gaussian Simulations based on variograms with increasing ranges of spatial dependency (autocorrelation). The set of layers of plant trait was used to simulate hyperspectral data using radiative transfer models (RTM). The plant traits were given values that are representative of low canopy vegetation (*e.g.*, grasslands). Among the plant traits simulated, LAI was selected as the response variable. Essential variable for understanding vegetation functioning and structure (Woodgate et al., 2015), LAI is defined as half of the green leaves surface horizontally projected per unit of ground area (Chen and Black, 1992).

The other layers of plant traits were utilised only to simulate the hyperspectral data (see Rocha et al. 2018 for full details on the simulation). The simulated spectra were used as explanatory variables to fit spatial and non-spatial regression models to predict LAI. Machine learning algorithms base on the 2100 wavelengths available and linear regressions using as a covariate a spectral index comprised the non-spatial models used in this study. Spatial models using the same vegetation index as a covariate were fitted with Bayesian inference by Integrated Nested Laplace Approximation (INLA) approach (Rocha et al., 2019). The comparisons are focused on prediction accuracy and model generalisation of the different sample designs across the 15 levels of spatial dependency. The models were compared while varying either the spatial dependency (autocorrelation ranges) or the spatial configuration (a different realisation of a landscape).

5.2.1 Data simulation

The simulated data were created according to the following steps (1) to represent landscapes (or fields) with increasing levels of spatial autocorrelation a set variograms was created; (2) from these variograms, plant trait values based on a regular grid were generated using Sequential Gaussian Simulations of random fields; (3) using Radiative Transfer Models (RTM), hyperspectral data were simulated as measured by field spectrometers; (4) random and

spatially dependent noise was added to the plant trait (Y) and to the (hyper)spectral data (X), and (5) drawing four different sampling designs from each grid and ordering the observations in a sequence that minimizes distance to collect them.

5.2.1.1 Plant traits

Values of plant trait were generated on a 100 by 100 grid using unconditional simulations according to the set of variogram models representing 15 ranges of spatial dependency. These simulated landscapes represent spatial autocorrelations ranging from zero (or independent in space) to 70% of the image extent. Thirty realisations were generated to each level of dependency, representing a unique spatial configuration or pattern. The spatial patterning among different levels of spatial autocorrelation from the same realisation is more similar than among different realisations with the same level of autocorrelation. This relation is illustrated in Figure 5.1, where patterns are more similar along the vertical lines than along horizontal lines.

Seven different plant traits, with values ranging between realistic scales for a grasslands environment (Table 5.1), were generated as input for simulating hyperspectral data by a radiative transfer model. The plant traits: Leaf Area Index (LAI), Leaf Structure (N), Chlorophyll Leaf Content (Ca+b), Dry Matter Content (Cm) and Hotspot (hspot) were simulated following this procedure. The plant traits Carotenoid (Car) and Water Content (CW) were a function of Chlorophyll and Dry Matter respectively, as described in the caption of table 5.1 (Jarocińska, 2014; Vohland and Jarmer, 2008).

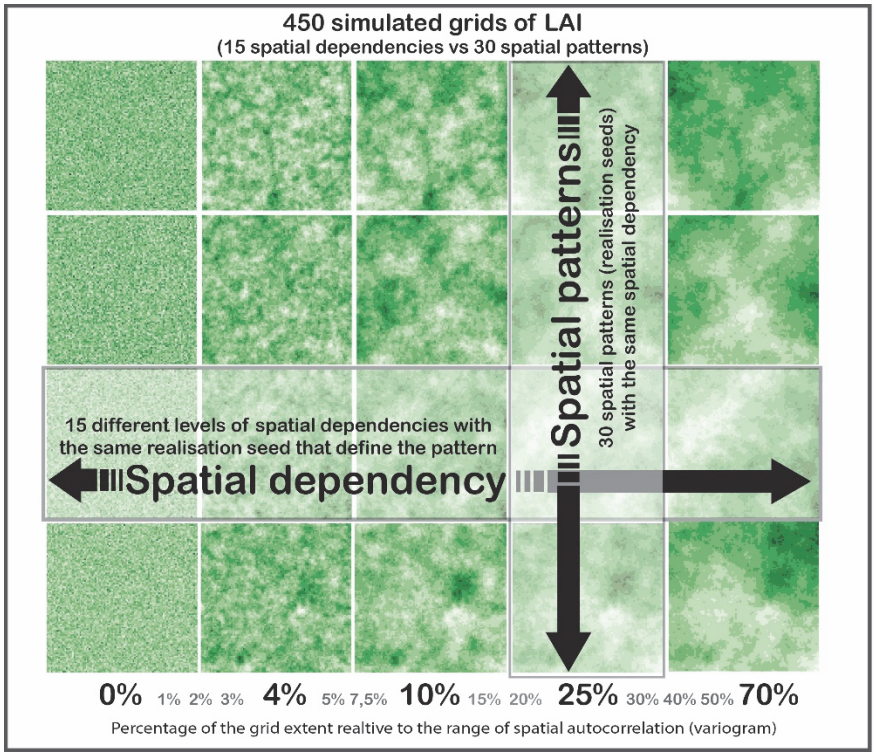


Figure 5.1 - Generation of Leaf Area Index (LAI) layers at 15 levels of spatial dependency.

5.2.1.2 Hyperspectral simulation

A Radiative Transfer Model (RTM) was used to simulate 450 hyperspectral cubes to represent the hypothetical landscapes and their spatial structures. The PROSAIL 5B model was adopted to simulate wavelengths from 400nm to 2500nm with a spectral resolution of 1nm, generating in total 2100 bands (Jacquemoud et al., 2009). Besides the above seven plant traits describing leaf and canopy properties, six other RTM parameters were defined when implementing PROSAIL 5B (Table 5.1). Of these six, three parameters were fixed: brown pigment (Cbrown) = 0, assuming entirely green canopies; leaf angle distribution (LAD) = erectophile = 90°, when considering the principal grassland orientation; and the soil moisture factor (psoil) = 0, assuming a null effect and being constant in space. The remains three RTM parameters related to illumination and geometry (Table 5.1) which were generated using a uniform distribution ~U(min, max). The solar and view angles were slightly changed according to a theoretical field campaign using a hand-based spectrometer under sun lighting.

As RTM models are fully deterministic, random and spatially dependent noise were both added into the LAI layers before simulating spectra and sampling

LAI values from these resultant layers. The noise added before simulating the spectra represent the variation expected when observations are collected by the sensor. A different realisation of this noise was added to LAI layers further sampled to be response variable, as it is unlikely to observe identical spatial structure for LAI values and spectral data. Random noise from a normal distribution was also added to each waveband to represent variability in the measurement system produced by an optical sensor. The parameters for the normal distribution, mean and standard deviation, was estimated per waveband in a pilot experiment using the spectrometer ASD FieldSpec® 3 (Inc., Boulder, CO, USA). The experiment was used for determining the reproducibility of the instrument when measuring grassland reflectance under natural illumination repetitively.

Table 5.1 - Parameters from PROSAIL used for simulating canopy reflectance for the landscape realisations.

	Parameter	Description (unit)	Distribution	R2 with LAI
Leaf	Cab[1]	Chlorophyll a+b concentration (ug/cm^2)	~N(28,4.5)	0.36
	Car[2]	Carotenoid concentration (ug/cm^2)	~N(5,0.7)	0.35
	Cbrown[3]	Brown pigment (-)	0	-
	Cm[1]	Dry matter content (g/cm^{-2})	~N(0.004, 0.0005)	0.69
	Cw[2]	Equivalent water thickness (cm)	~N(0.016, 0.002)	0.66
	N[1]	Leaf structure parameter (-)	~N(1.5, 0.12)	0.48
Canopy	LAI[1]	Leaf Area Index (-)	~N(3.1, 0.6)	-
	hspot[1]	Hotspot parameter (-)	~N(0.05, 0.01)	0.50
	LAD[3]	Leaf angle distribution (attribute)	Erectophile (90º)	-
	psoil[3]	Dry/Wet soil factor (-)	0	-
Geometry	tto[4]	View zenith angle - VZA (degree)	~U(0,5)	-
	tts[4]	Solar zenith angle - SZA (degree)	~U(30, 38)	-
	psi[4]	Relative azimuth angle (degree)	~U(0,360)-U(129,252)	-

Note: [1]plant traits simulated with spatial dependency and presented as Normal distribution ~N(mean, standard deviation). [2]plant traits correlated with another parameter, Car=Cab/5 and Cw=4/Cm^{-1}. [3]parameters with fixed values for all realisations. [4]generated from a uniform distribution with min and max ~U(min,max), based on a theoretical field campaign using a handheld spectrometer: where tto=deviation from nadir; tts=90º minus the max and min sun altitude, and psi=~U(0,360) minus the max and min solar zenith angle.

5.2.2 Sampling design

Two probabilistic design-based strategies were tested against two spatially regular sampling approaches that are supplemented by closely spaced locations, considering that no previous knowledge was available to use a model-based sampling approach. The four designs tested were represented by a simple random sampling, a systematic sampling, a lattice plus close pairs, and a lattice plus in-fill (Figure 5.2). In the random sampling designs, the locations were drawn from a list of all pixels (or cell number) of each landscape using a random number generator. This method ensures that every pixel has the same chance to be drawn (*i.e.* probabilistic sampling) and the allocation of the sample is unbiased in space. However, it may not be spatially representative, and the sampling precision (or error) is calculated assuming independent and identically distributed observations, rarely verified in a continuous spatial domain.

In the systematic sampling designs, the spacing between sample points is defined regularly, and the values were drawn based on a sequence of two dimensions (*i.e.* a grid). The origin of the sequence or the starting point in the grid was randomly selected. Thus the sampling design can be considered probabilistic. The lattice plus close pairs design is a two-stage procedure characterised by closely spaced pairs of points. In the first stage, a sample of pixels spaced as in a systematic sampling design was defined. In the second stage, for each sampled pixel, 1 out of the 16 adjacent was randomly selected from the knight, and one-cell queen moves direction (Figure 5.2c). The intention is to assess whether this design is suitable to identify the spatial covariance structure, providing adequate spatial prediction when the "true model" is unknown.

The 'plus in-fill' design is also a two-stage approach, consisting of a systematic sampling from a regular grid overlaid by in-fill squares for some selected cells or pixels (Diggle and Ribeiro, 2007). Systematic sampling was also drawn, then 6 pixels were arbitrary selected as the centre of the "squares" and filled in with all the 16 adjacent pixels from the knight, and one-cell queen moves direction (Figure 5.2d). The in-fill pixels are an attempt to estimate the spatial variation from small-scales more precisely. The selection of origin in the grid and the cells to supplement with closely spaced locations should be randomly selected to be considered probabilistic sampling and to avoid systematic bias, although in practice this is often ignored.

Observations for each sampling design were drawn from the simulated spectral cubes and LAI landscapes to train empirical models (training sets) at 50, 100 and 200 locations. Another set of observations from the same landscape were drawn at different locations to validate the fitted models (testing sets). A

sample sequence that reduces the distance made to measure the random points was stablished for each sampling realisation. The data in this sequence were used to train and to validate all the models, and then later to assess the (spatial) autocorrelation in the residuals. The spatial autocorrelation captured by the sample may vary significantly according to the sampling design and sample size, in consequence of the difference in the average distance between two consecutive points.

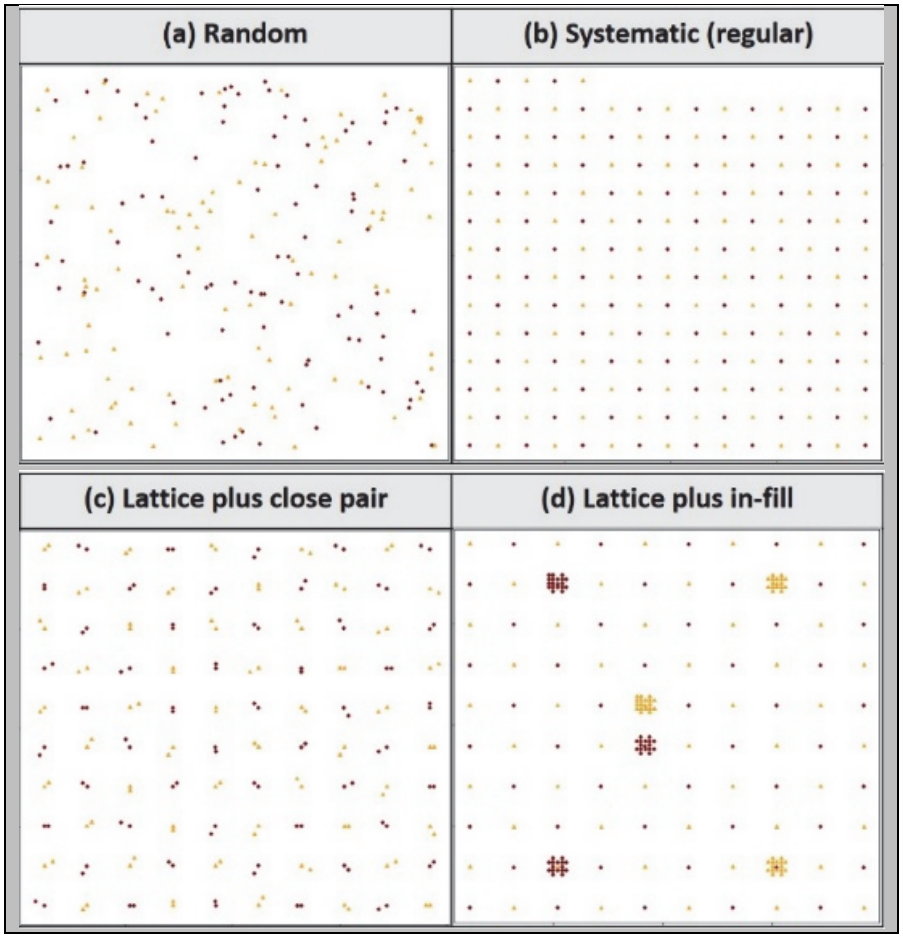


Figure 5.2 - Sampling designs: (a) random, systematic (b), lattice plus close pairs (c) and lattice plus in-fill. The dark dots are the points drawn for the training set and the lightest points for the testing set.

5.2.3 Model selection and assessment

5.2.3.1 Non-spatial models

The machine learning algorithms Partial Least Squared Regression (PLSR) and Support Vector Machine (SVMR) were chosen because of their capacity, and wide adoption, when dealing with multicollinearity and high dimensional data such as hyperspectral data (Carvalho et al., 2013; Feilhauer et al., 2015). These techniques require a tuning process to restrict their level of complexity to avoid overfitting. For instance, the "number of components" in PLSR or the "cost" value in SVMR models represent a measure of the complexity of the eventually fitted model. The machine learning models were tuned with 10-fold cross-validation, which was repeated ten times each, selecting the complexity that minimises the Root Mean Squared Error (RMSE) (Hastie et al., 2009).

In contrast to the complex machine learning algorithms, a simple linear regression model (ordinary least square) was fitted using a vegetation index as a covariate. The vegetation index used was the LAI Determining Index (LAIDI), which is the ratio between two wavelengths (1050 nm and 1250 nm) from the near-infrared (NIR) spectral domain (Delalieux et al., 2008). These wavelengths were selected a priori, rather than searching by the combination of bands that explains most of the response variable.

5.2.3.2 Spatial models

Integrated Nested Laplace Approximation (INLA) approach was applied to fit spatial models. This model uses a mesh of non-overlapping triangles (constrained Delaunay triangulation) to represent the spatial domain (Bakka et al., 2018). Based on a number of vertices, a neighbourhood area is defined to account for spatial dependency in a model. The mesh density was defined based on a spatial autocorrelation range of approximately 40% of the landscape extent. This density reduces most of the autocorrelation in the model residues, while gives a fine balance between prediction accuracy and overfitting according to previous studies (see chapter four). This mesh density was selected according to the extent of the grid area, rather than the exact sample locations. This approach represents the area more uniformly, avoiding that the mesh is constructed to a specific distribution of sample locations. The spatial models were fitted using R-INLA with the spectral index LAIDI, as a covariate used in the linear model regressions.

5.2.3.3 Model assessment

The prediction accuracy of the different model approaches was compared using the root mean square error (RMSE). The selected model for each regression technique and sampling design were assessed on their capacity to generalise

within the same landscapes where the training samples were taken from by making predictions for "unseen" locations. The capacity to generalise to another landscape (realisation) than where the training sample was taken from was also assessed and was presented as a "new realisation" set. The autocorrelation between residuals of subsequent locations where the sequence of sample locations minimises the total distance along all locations was estimated with the Durbin Watson test. The analyses were performed using R version 3.2.2 (The R Foundation for Statistical Computing). The following packages were used: gstat for unconditional simulations, hsdar for PROSAIL 5B, Caret to tune machine learning regression and R-INLA for fitting spatial models by the Bayesian inferential approach.

5.3 Results

5.3.1 The effect of sampling design on global estimation

When the intention is to estimate the global mean for the sampled area rather than predict values at unseen locations, the sampling design may have a substantial effect, depending on the sample size and spatial distribution of the points. All simulated LAI landscapes, regardless of their spatial dependency, were based on a normal distribution with a mean LAI of 3.06 and a standard deviation of 0.6. Therefore, it is expected that a representative and unbiased sample should be able to estimate these values. Drawing thirty samples with 100 observations per spatial dependency (*i.e.* a sample for each realisation), the lattice plus in-fill design shows considerable variability in the global mean of LAI (Figure 5.3a). This occurs because the clustered points over-represent areas of either low or high LAI in specific landscape realisations. Also, the in-fill design clearly underestimates the standard deviation in the presence of spatial autocorrelation (*i.e.* higher than 2% of the extent). The lattice close pairs design has more variability in the mean estimation for spatial dependency between 3% and 5% of the extent, while the systematic and random samplings tend to approach the true value as the spatial dependency increases. Except for the dependency of 1% and 2%, close pair sampling design estimates the (global) standard deviation more similar to the (true) value used for simulating the 30 realisations.

Systematic samples overestimate the standard deviation when the dependency increases, while random sampling slightly overestimated standard deviation when a spatial dependency is around the average distance of the sampling path (*i.e.* 5% of the extent). Overall, for global estimation, the lattice plus in-fill sampling should be avoided if the landscape presents significant spatial autocorrelation (roughly more than 2%). For the other sampling designs, the choice depends on the spatial dependency in the landscape, the sample size and the statistical parameter of interest (*i.e.* mean or standard deviation).

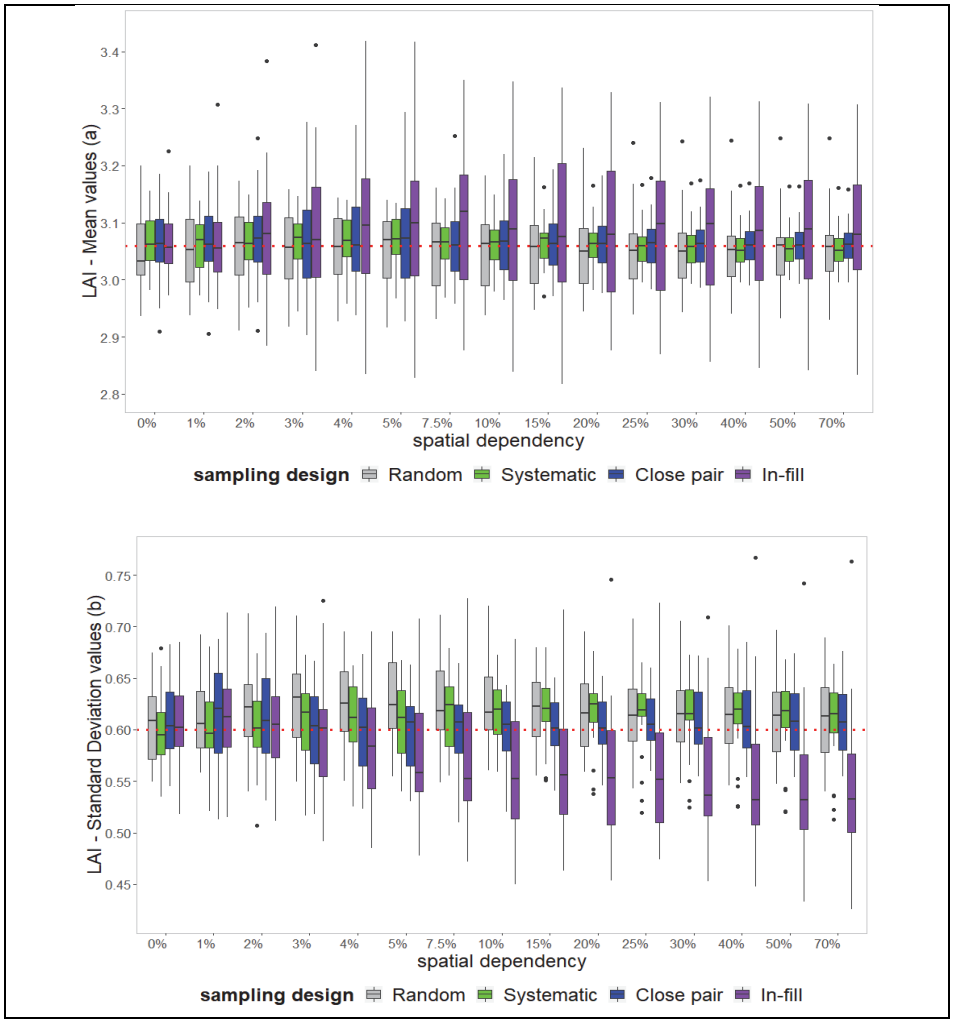


Figure 5.3 - Boxplot of the global mean (a-top) and the standard deviation (b-bottom) for the 30 realisations of LAI per sampling design. The dashed line represents the true value used for simulating the data.

5.3.2 The effect of the sampling design in the model accuracy

Assessing the accuracy for the combination of sampling design and modelling approaches, the results show that sampling designs affect the RMSE from training sets more than the RMSE from testing sets (Figure 5.4). In non-spatial models, the RMSE of the training set slightly reduces when the spatial dependency increases, while in spatial models, RMSE reduces significantly with levels of autocorrelation higher than 5% of the extent. The lattice plus in-fill design yields the lowest training errors, followed by the lattice close pairs.

However, these same sampling designs yield poor model generalisation when applied to the test data set (unseen locations from the same realisation) or for new-realisation sets.

Linear models are almost not affected by sample design or spatial dependency when generalised to unseen locations from the same realisation (test data sets) or to a new realisation (different pattern). Machine learning algorithms such as PLSR and SVMR generalise poorly compare to linear models, and the RMSE is less stable across realisations for sampling designs such as lattice close pairs and lattice plus in-fill. These models show less overfitting in the models when trained by systematic and random sampling than the other designs. The in-fill design appears to predict more accurately at lower levels of spatial autocorrelation when using the training set but is not able to generalise to unseen locations, even in the same landscape. The spatial models in combination with a systematic sampling design present the lowest testing errors at the higher levels of spatial autocorrelation.

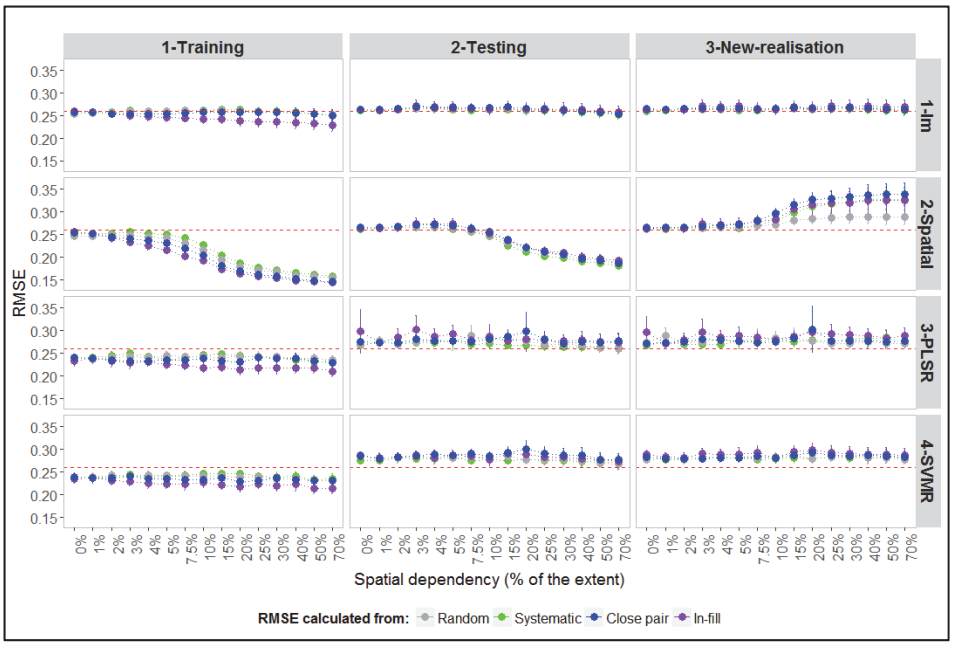


Figure 5.4 - Prediction accuracy (RMSE) per model approaches and sampling design according to the spatial dependency and the dataset used for validating de model.

When evaluating the combination of models trained and tested from all sampling designs, SVMR models are more affected by sampling designs. If lattice close pairs or in-fill are used for training, the model prediction from a landscape with spatial dependency is around 15% and is the most affected. However, when trained by random or systematic sampling, the most unstable

model occurs for landscapes with a spatial dependency of around 7.5% of the extent. Models trained by random sampling and tested by a systematic or vice-verse show no reduction on accuracy, despite the modelling approach.

The main difference amongst the spatial models trained by each sampling design is the inflexion point where the error starts to decrease and where the inclination of the slope reduces. The RMSE for the systematic design presents a very similar curve independently from which sample the test data set originates (Figure 5.5). In the spatial models, trained by the in-fill design, the error increases from the 0% to 4% of spatial dependency before starting to fall, while the random and systematic sampling remain almost flat before 5%. The error from the systematic design in the spatial model reduces more from 10% to 15% than in the other intervals. In the remaining sample designs, the reduction in the error is more gradual across the spatial dependency (Figure 5.5).

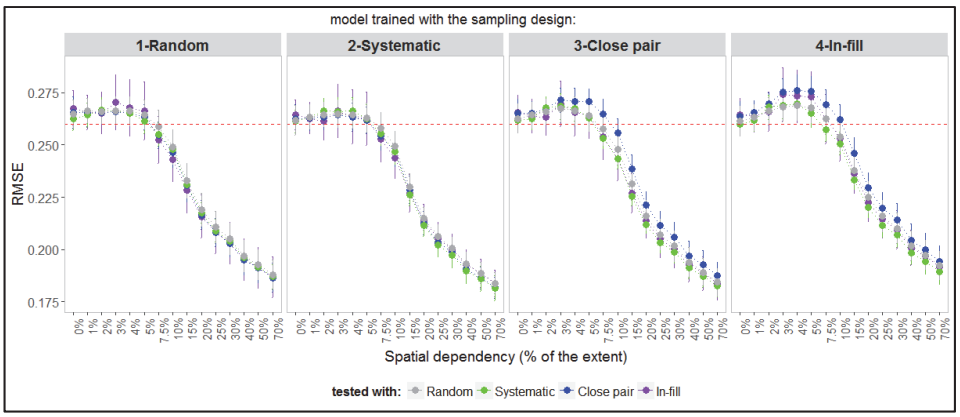


Figure 5.5 - RMSE for a spatial model trained by a sampling design (boxes one to four) and tested by all designs (colour legends) per spatial dependency.

5.3.3 The effect on the model residuals

Although the level of spatial dependency of the plant trait and the model approach determine most the remained autocorrelation in the residuals, the sampling design also shows a statistically significant effect for spatial and non-spatial models (Figure 5.6). For instance, in the spatial models trained with a sample from a systematic design, the values of the Durbin Watson test statistic slightly rise when spatial dependency increases, maintaining the average value always above the threshold line (negative autocorrelation). In contrast, with the close pair's design, the DW values reduce until 7.5% of dependency, presenting positive autocorrelation (below the threshold line) from 2% to 15%. For landscapes with strong spatial dependency, the residuals from spatial

models trained by a sample from random and systematic designs are slightly negatively autocorrelated, indicating that the mesh density should be reduced.

For the non-spatial models, the systematic sampling presents constant values of DW and approximately free of residual autocorrelation until 5% of dependency, decreasing fast up to 15% and then reducing slightly from that stage on. On the other hand, the close pair design shows autocorrelation from 1% and higher, decreasing smoothly across all range of spatial dependency. The machine learning models yield slightly less autocorrelation in the residual errors than linear models. This may occur because the algorithms use the residuals from less complex models during the tuning process as the prior information to fit the final model. The dispersal of the DW values among the 30 realisations per spatial dependency (box range) varies less in the systematic design than in the random or close pair designs for all model types.

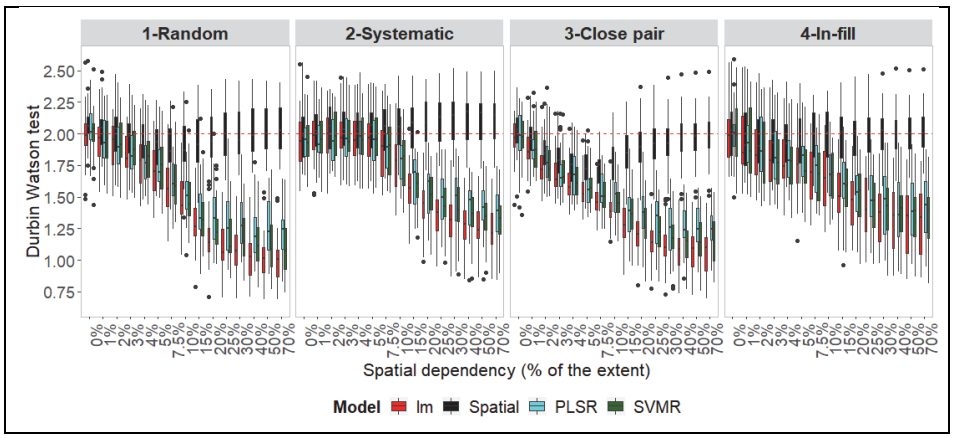


Figure 5.6 - Boxplot for the Durbin Watson statistic for the model residuals of the 30 realisations per regression type and sampling design. The horizontal red dashed line shows the threshold for residuals independence (DW=2). Values higher than two represent negative autocorrelation and lower than two indicates positive autocorrelation in the residuals.

5.3.4 The effect of sample size

As random and systematic sampling yield higher accuracy for global estimation and for predicting unseen locations, the effect of the sample size while modelling was tested for both sampling designs. A systematic sampling design yields a smaller difference between training and testing error regardless of the modelling approach, spatial dependency or sample size. This is demonstrated in Figure 5.7 by the distance between the training error (brown line) and the testing error (orange lines). The difference between training and testing increases as the sample size decreases, especially for complex models such as spatial and machine learning algorithms. This is a consequence of overfitting

with smaller sample sizes relative to the large number of parameters to train the model.

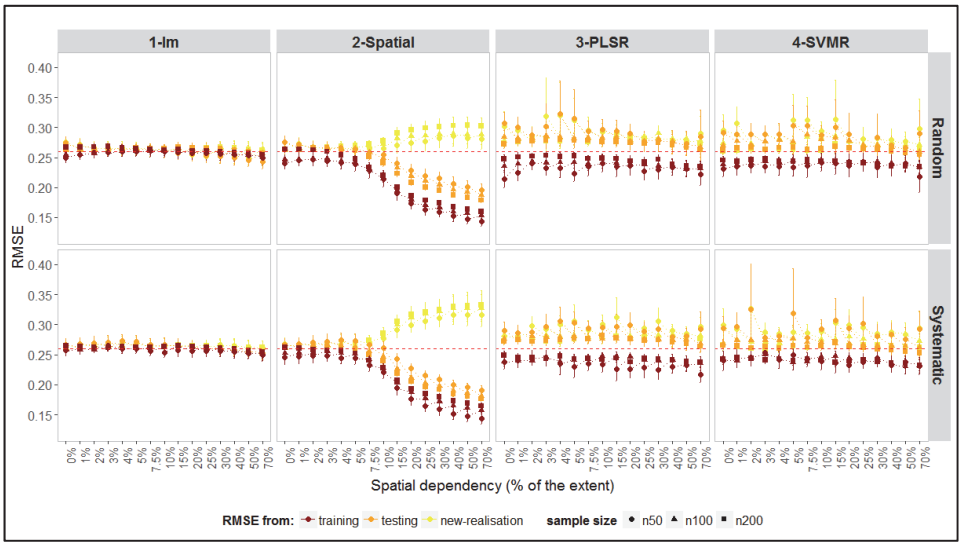


Figure 5.7 - Prediction accuracy (RMSE) per model type (vertical) according to the spatial dependency for random and systematic sampling designs (horizontal).

For the spatial model with a systematic sampling design, a sample size of two hundred observations yielded very similar test and training errors. However, the larger the sample size, the lower the capacity to generalise to this new landscape. Machine learning models present a clear division between training and testing error for all ranges of spatial dependency tested. For these models, a sample size with 50 observations was very unstable, generalising poorly for both random and systematic sampling designs. Again, linear models are more stable across sample sizes and spatial dependencies, however random samples with 50 observations showed more instability in RMSE values in both extremities of the spatial dependency axis.

The sample size alters the density of locations in the field, and thus the degree of spatial autocorrelation captured by the observed points (Figure 5.8). For random sampling, it is noticeable that for larger sample sizes, a stronger (positive) autocorrelation can be observed in the model residuals. For spatial models, the mesh can be adjusted to adapt the sample size and the spatial dependency accordingly. For instance, for random sampling of 50 observations, the residuals are negatively correlated, so the mesh density should be reduced (Figure 5.8-2), while the opposite occurs for a sample size of 200 observations. In non-spatial models, such as linear models and SVMR, a sample size of 200 observations at 2% spatial dependency yielded more autocorrelation in the residuals than a sample compared with 50 observations

at 5% spatial dependency. This demonstrates that the observed spatial autocorrelation in the landscape is related to the design and size of the sample.

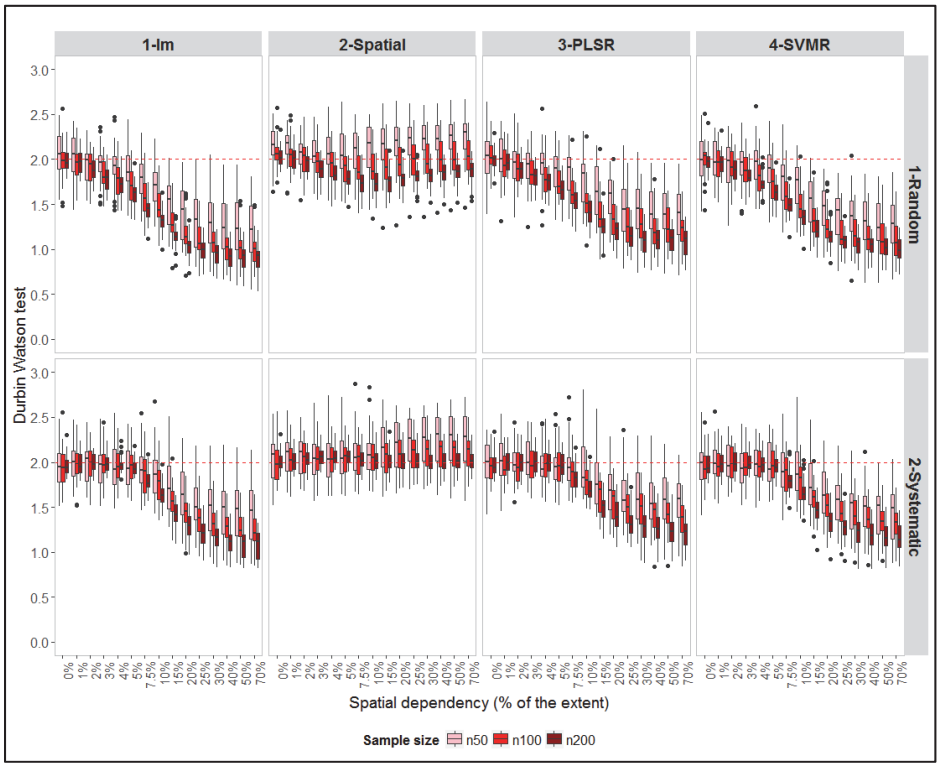


Figure 5.8 - Boxplot for the Durbin Watson statistic for the model residuals of the 30 realisations per regression type and sample size for random and systematic sampling designs. The horizontal red dashed line shows the threshold for residuals independence (DW=2). Values higher than two represent negative autocorrelation and lower than two indicates positive autocorrelation in the residuals.

5.4 Discussion

5.4.1 Sampling design and modelling approach

If a sampling design will be used to estimate global parameters such as LAI mean value, increasing the sample size beyond a certain point brings no significant improvement, but adds redundant information. In other words, when the intention is to estimate a plant trait value at an unseen location, the spatial representativeness of the sample might not be enough for a precise (local) estimation. In a landscape with significant autocorrelation, close pairs may present redundant information, while faraway pairs may lose crucial spatial information. A systematic sampling design has an advantage compared with random sampling as it guarantees that the measurements are evenly

spread over the study area. In general, this design has shown higher prediction accuracy, but where the spatial dependency coincides with the sampling distance interval, it may produce unreliable predictions. Where the spatial dependency is much smaller than the sample spacing, the spatial autocorrelation is undetectable, indicating that non-spatial models are suitable. However, if the aim is to predict across an entire image where autocorrelation does occur at a finer scale than sampled, this may produce biased values. Where there is no previous knowledge, design-based sampling may be an ideal, either to demonstrate that there is no significant spatial autocorrelation or to indicate that a regression robust to spatial autocorrelation should be used.

As important as the definition of the appropriate sampling design is the selection of a regression method that meets the assumptions for the spatial autocorrelation imposed by it. Whether the decision is between a spatial and a non-spatial linear model or machine learning algorithms, the residuals from any fitted model using remote sensing data should be assessed and reported. In the case of the spatial model, the residual error assessment is vital to evaluate whether a mesh or any other kind of neighbour matrix, accounting for the spatial domain is appropriated for the observed autocorrelation. The tuning process to find the optimal mesh is as essential as the machine learning is when selecting model complexity, because not only may it cause overfitting, but also the spatial model may transform the residuals from being positive to negative autocorrelated. When the nugget effect is unnoticeable in the variogram, the benefits of model-based approaches are diminished (Delmelle, 2009). There are two sources of nugget effect; from measurement errors or from spatial variations on a scale smaller than the shortest distance between any two points in the sampling design (Diggle and Ribeiro, 2007).

Although a systematic sampling design presents a limited combination of distances and lack samples at very close locations, this design can perform with higher accuracy compared with designs that combine a regular frame with closer pairs, or clustered locations. Lattice plus in-fill design may suffer from bias on the overall estimates, as it tends to cluster observations at locations with very high or very low LAI values where there is a significant spatial dependency. Clustered samples such as in-fill designs may commit a too large proportion of the sampling effort to access closely spaced points (Diggle and Ribeiro, 2007). For global estimates, the observations from this sampling design should be weighted or averaged to produce one value for each cluster location to prevent bias in estimation. Ripley (1981), suggested that non-stationarity and spatial anisotropy data may lead to a severe effect on the efficiency of systematic sampling. However, these assumptions are intrinsic and restrictive for most of the modelling approaches. When explanatory variables are available as a covariate, their spatial distribution will also affect

the design performance, and therefore the model accuracy (Diggle and Ribeiro, 2007).

5.4.2 Sampling design and remote sensing

When a remote sensing image is available, and there is previous knowledge of the empirical relationship with the plant trait, it is possible to design an optimal sample (*e.g.* model-based) to cover an area with higher uncertainty. Satellite images allow the use of model-based sampling designs, despite being more difficult to align the data collected in the field with the pixel captured by the sensor in orbit. The result of the sampling design used for predicting traits in the natural environment by remote sensing data is frequently constrained by the time and the cost of the fieldwork. The time factor is important not only because of the schedule but also by the effects of plant phenology, the seasonality of the natural lighting and the possible changes in weather conditions during the field campaign. In an ideal condition, spectral data and plant trait measurements should be taken simultaneously, yet it is rarely the case in practice. It is also common in remote sensing applications that samples are collected using an existing network of locations settled for another purpose, which may lead to biased inferences about the underlying spatial dependency on a continuous domain (Diggle and Ribeiro, 2007).

Spatial structure present in the spectral data may not be causally related to the plant trait being considered, such as moisture and soil characteristics captured as background by the sensor. The spatial (or temporal) autocorrelation can be created while collecting ground references according to the order and pace which those locations are measured using an optical instrument such as field spectrometer. For optical instruments such as sensors, it is challenging to maintain controlled illumination geometry and variations in weather conditions through space and time under sunlight. Long field campaigns may take so much time that plant trait values change, further increasing the spatial autocorrelation among close locations. The most common approaches to collect LAI values are by optical instruments or by measurements in the lab of the leaf surface area from canopies harvested in the field. Reflectance can be captured by sensors deployed on different platforms such as satellites, aircraft, drones, terrestrial vehicles or handheld spectrometer(Milton et al., 2009). Each combination of these direct and indirect measurements to estimate LAI imposes some constraints in sampling design, which affects the time-space alignment between both measurements. Airborne platforms, for instance, can deploy hyperspectral sensor or laser scanner, capturing larger volumes of information, but the alignment can be even more erroneous as there is no a fixed grid to be based when planning the sample.

In order to reduce the effect of support problems related to misalignment between both locations, some authors suggest averaging units in a broader "window of pixels" (Atkinson and Emery, 1999; Liang, 2005). This procedure may increase prediction accuracy, underestimating the variation (Schaepman-Strub et al., 2006). A similar procedure used to reduce the effect bidirectional reflectance factor (BRF) of the neighbour pixels in the image, may also be an artefact caused by the reduction in variability in a process known as "lumping" (Holmes et al., 2006; Lovett et al., 2005). Airborne and satellite image present some level of distortion when departing from the nadir position, which may further increase the misalignment between the pixel unit and the plot in the field (Shaw and Burke, 2003).

Vegetation indices, often used as a covariate, are the product of two or more spectral bands. There is the general notion in remote sensing that an index is a more reliable variable than the individual bands because the coefficient of the wavelengths is less unstable to oscillation from natural illumination (Liang, 2005). In the case of this study, using the two spectral bands separately, it slightly increases model accuracy (not showed). Despite not providing a significant improvement, the relation between the plant trait and reflectance may present spatial dependency for one wavelength which different to the other. As illustrated by Atkinson and Emery (1999), healthy vegetation presents, in general, a low reflectance in the blue and red wavelength regions, whereas near-infrared reflectance from a vegetation canopy is often high as the result of the internal scattering within leaves. As the red and near-infrared bands represent different physical features, it is expected that the two bands also carry different spatial information. Using the bands separately in the model as covariates can acknowledge the different spatial information and account for interactions of the two wavelengths.

5.5 Conclusion

This study showed that the sampling design affects the estimation of population parameters for unseen locations. Drawing samples from a population that are not independent nor identically distributed over space turns even more critical as the sampling design and size can determine the presence or absence of spatial autocorrelation. Additionally, it is a particular challenge to align a spatial dependent plant trait with remote sensing data. Whether a sampling design aims to capture or avoid autocorrelation (to fit either a spatial or a non-spatial model) should be defined in advance. Remote sensing provides the opportunity to test for autocorrelation before a field campaign where there is a pre-established empirical relationship between plant trait and the measured reflectance. Also, Bayesian inference opens the possibility not only to fit spatial models to predict plant trait at unseen locations as shown here but also to estimate the uncertainty of the predicted values for an entire area.

The combination of remote sensing and spatial model should guide the selection of sampling designs that improve the accuracy of the predictions at specific locations. Otherwise, when no previous information is available, a regular and probabilistic sampling as systematic design should be the primary option, with the warning about the risk of the spatial dependency accidentally agrees with the sampling spacing.

Chapter 6

Synthesis: Predicting plant traits with remote sensing

6.1 Uncertainty and Stochasticity

When predicting plant traits from a landscape with remote sensing data, there are at least three continuous domains involved: space, time and spectrum. Part of the unexplained variation in reflectance and plant trait values are inherent to the measuring system or related to patterns across space and time (Dormann et al., 2007; Legendre and Fortin, 1989). These domains are not independent, and they most probably interact considerably when continuous areas of heterogonous vegetation are imaged with sunlight as the primary source of illumination (Legendre et al., 2004; Militino et al., 2018; Roberts et al., 2017). Under these conditions, the only certainty when predicting ecological systems with remote sensing is the presence of uncertainty. The uncertainties come either due to lack of sufficient knowledge about the (deterministic) factors that affect the underlying process that drives the plant trait, or because the plant trait process and its spectral representation is partially unpredictable, but most probably both (Pan, 2018).

A lack of knowledge of the underlying mechanisms may result in uncertainties in the capacity of the sampling design to capture the data structures, or by the absence of relevant explanatory variables, or by an inappropriate model selection and parametrisation (Fortin et al., 1990; Gelman et al., 2001). To reduce uncertainties when modelling plant traits, essential variables that represent reality closely are needed. However, spatiotemporal fluctuation in environmental conditions imposes some degrees of unpredictability whether or not all essential variables are available for modelling (Militino et al., 2018). Ecological systems present inherent stochasticity or randomness, which make it somehow impossible to predict precisely their dynamics. Therefore, in most cases, it cannot be fully predicted, but its distribution can be known and the uncertainties about the prediction values assessed (Klyatskin, 2017).

It is known that weather and atmospheric systems are stochastic and do not repeat the same conditions across space and time (Franzke et al., 2015). Therefore spectral values under natural light cannot be replicated exactly, independently of the measurement system quality. Weather conditions and atmospheric systems are not entirely random. However, they are exceptionality complex and dynamic to be modelled in a fully deterministic way (Klyatskin, 2017; Schertzer and Lovejoy, 2004). As remote sensing under sunlight depends greatly on the weather and atmospheric conditions, some stochasticity in the spectral data is expected as well (Franzke et al., 2015). Plant trait values may also be driven by processes intrinsically stochastic such as temperature, precipitation or pest infestations. Modelling plant trait using remote sensing are susceptible to uncertainties not only for the influence of random (or stochastic) process but also for the limited information to predict

complex and dynamical systems using a pure spectral-based empirical model as shown in chapter four.

Spatiotemporal variations may create structures in the data according to the order in which the ground references are measured as showed in chapter three. A campaign collecting simultaneously spectral reflectance and field measurements may be unfeasible, and some degree of space and time misalignment should be expected (Finley et al., 2014; Wilson et al., 2011). Also, replication is not entirely possible as it is unlikely to revisit the same place and find the same conditions in which reflectance was captured before. Continuous variables collected *in situ* such as most of the plant traits should be considered a realisation of a stochastic process, rather than ground-truth representing fixed but unknown values (Pan, 2018). Input data with insufficient information to explain the target plant trait, or with some level of stochasticity, result in uncertainty in the model outputs. The uncertainty is neither spatially nor temporally independent regardless of the instrument used to measure spectral data and plant trait measurements (Pan, 2018). Choosing an adequate sampling design and modelling approach will not change the nature of the remote sensing data, but it does interfere in the way that it will be understood or predicted.

6.2 Spectral domain

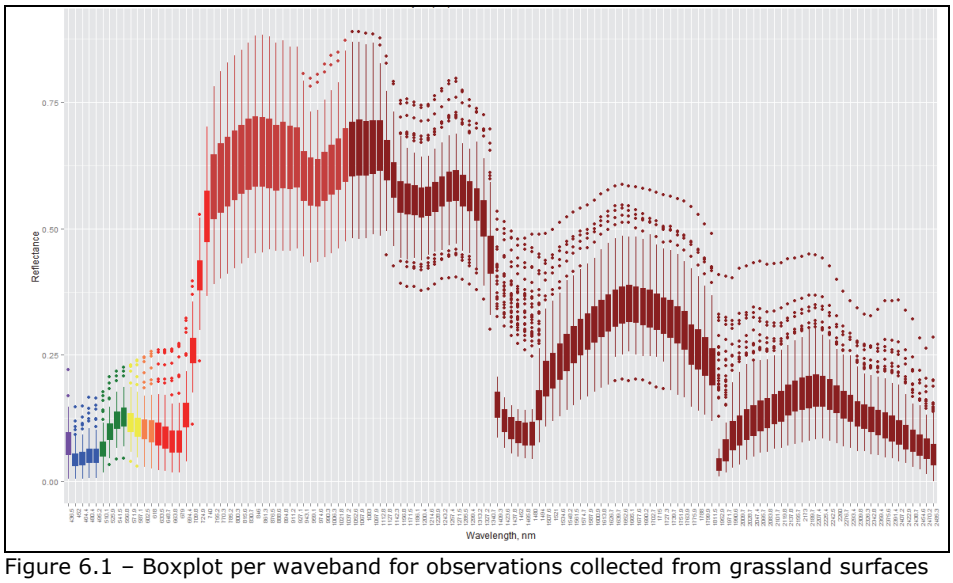


Figure 6.1 – Boxplot per waveband for observations collected from grassland surfaces using a hyperspectral airborne sensor. See more information about the data in Appendix 2A of the second Chapter.

Optical remote sensing comprises wavelengths from the visible spectrum to the thermal infrared (0.4-14 μm) (Buitrago Acevedo et al., 2017). The digital

number of each wavelength can be quantified by physical quantities such as radiance, emittance or reflectance (Liang, 2005). Commonly used for modelling, the reflectance is a (unitless) value between 0 and 1 that represents the fraction of incident light reflected by a surface in a specific wavelength (Shaw and Burke, 2003). Quantitative estimation of land surfaces, as most of the plant traits, relies greatly on a physical understanding of the remotely sensed data related to the vegetation characteristics (Curran, 1989). Reflectance captured over natural sunlight depends very much on illumination conditions and sensor geometry (Schertzer and Lovejoy, 2004). Settings such as sensor viewing angle and the sun angle relative to the zenith can considerably change the amount of light reflected into the sensor field (Liang, 2005; Pearse et al., 2016).

Atmospheric effects such as absorption and scattering also interact with the light reflected by the vegetation surface, increasing the spectral variability of the observed target (Shaw and Burke, 2003). Absorptions are mainly caused by atmospheric gases, such as water vapour, ozone, oxygen and aerosols. Clouds casting shadows or nearby objects reflecting and scattering sunlight onto a target area can also provoke substantial changes in the illumination of the land surface (Pan, 2018). Water vapour and ozone are the main concerns for multispectral sensors since the other gases absorb energy in a very narrow spectral range, while for hyperspectral sensors, gases such as oxygen can affect specific wavelengths (Liang, 2005). Remotely sensed observations might be heavily adulterated by aerosols, clouds, and their shadows according to the sensor platform (Milton et al., 2009). Given all these sources of variability, the radiance received by a sensor contains information of both, atmosphere conditions and land surface properties.

Rather than categorical variables traditionally used in remote sensing pixel classification, plant traits are quantitative variables. The reflectance values can be linked to plant traits based on the concentration of chemical substances as leaf pigment or physical structures as canopy density (Curran, 1989). The amounts of energy reflected vary according to the different regions of the spectrum, informing about the surface composition (Shaw and Burke, 2003). Therefore, topographic and environmental factors can also affect reflectance as background (noise) while capturing the spectra, such as slopes, soil type and moisture (Van Cleemput et al., 2018). Common assumptions such as random leaf angle distributions across a canopy or Lambertian leaf surfaces are not truly observed in practices but assumed for the sake of simplification (Schaepman-Strub et al., 2006). The vegetation composition in a natural environment is a function of many ecological processes, such as competition and disturbance, which can determine to a certain degree the species distribution across space and time (Legendre and Fortin, 1989). For instance, the leaf water content is very species related but can considerably vary over

space and time also when suffering from stress caused by infestation or weather conditions (Abdullah et al., 2018; Buitrago Acevedo et al., 2017). Also, open-air remote sensing, excluding crops and some temperate forests will present multiple-species canopies mixed in the same pixel (Huber et al., 2008). Despite the physical (or optical) properties of the spectral data, predicting plant trait accurately from remotely sensed data relies on the quality of the data and its statistical properties.

The reflective portion of the electromagnetic spectrum is a continuous signal which is broken in narrower continuous bands or wavelengths (Manolakis et al., 2003). These sequential segments are naturally autocorrelated, as wavelengths from the nearby region in the same spectrum signal are very redundant as shown in the introduction (Figure 1.2). As larger the number of bands is, as more serially autocorrelated are the wavelengths, resulting in strong multicollinearity while modelling as demonstrated in chapter one. The multicollinearity provoked by autocorrelation in the wavelengths increases the risk of type II error in the feature selection during the modelling process. In homogenous landscapes composed mainly by the same material, for example, measuring a grassland with a couple of thousands of narrow hyperspectral bands, the correlation between the wavelengths drastically increases. Commonly applied filters to smooth the noise in hyperspectral data such as Savitzky–Golay also corroborates to increase the autocorrelation among spectral bands.

Illumination, climatic and environmental variations introduce random and systematic noises into the spectral domain as mentioned before (Liang, 2005; Militino et al., 2018). Therefore, modelling with hyperspectral data presents a high risk to select spurious correlations rather than empirical relationships. It is questionable to what extent it is possible to correct or control most of the optical and geometric distortions in the remote sensing measurements. On the other hand, the understanding of vegetation dynamics through remote sensing will be limited if interactions between spatiotemporal and spectra domains are not considered. It is quite naïve to think that each wavelength is an independent and equally good candidate to predict a plant trait, leaving the decision of how important the band is for a model selection algorithm.

6.3 Spatial domain

Quantitative remote sensing relies primarily on the spectral signatures rather than spatial distribution to estimate plant traits in a landscape or scene. Therefore, the possibility to learn with spatial dependency to estimate more accurately plant traits is often overlooked. Environmental and topographic variables are capable of explaining the spatial dependency without the need to modelling space explicitly (Fortin et al., 2012). However, this information is

often not available, or there is not enough knowledge about the spatial scales of the underlying processes (Dormann et al., 2007). If it were possible to measure all relevant variables at multiple scales, there would be no reason to use spatial models as it will show no improvement in prediction accuracy (Hawkins, 2012). As it is impossible to measure everything at all scales, and the current knowledge is limited, this is still not the reality. Therefore, instead of measuring environmental variables at many different scales, it is easiest to use the coordinates of the collected observations, which naturally provide information about the spatial structures. In a landscape, plant traits are most likely to exhibit spatial dependency, independently on the area extent or the plot size utilised (Fortin et al., 1990). Water availability, nutrient concentrations in the soil, species dominance among other factors can drive the spatial dependency of plant traits on the environment (Van Cleemput et al., 2018). Ground references of plant trait collected over continuous vegetation that shows no spatial structures are probably not an accurate description of the reality. If spatial autocorrelation is part of nature, when trying to understand vegetation dynamics, neglecting it, or treating as some bias or distortion seems unreasonable (Hawkins, 2012).

In remote sensing data, spatial structure and patterns captured are not always related to the plant trait or even to the surface targeted, such as changes in soil background and cloud cover (Cochrane, 2000). Depending on the sensor platform geometric distortions from the nadir and the Bidirectional Reflectance Factor (BRF) effect in nearby objects can increase the spatial dependency (Schaepman-Strub et al., 2006). The spatial dependency may vary significantly if the spectra are based on (a) a collection of points taken by hand-based spectrometers, or (b) on a grid of pixel in square scenes captured by satellites at once, or (c) in stripes taken by airborne and drones. Similar to spectra, where the proximity of the wavelengths results in strongly autocorrelation, pixels located nearby are expected to be (spatial) autocorrelated as well (Tobler, 1970). In contrast to multicollinearity, spatial autocorrelation violates the assumption of independent and identically distributed (i.i.d.) observations of many modelling approaches. Breaking this assumption increases the likelihood of incorrect rejection of null hypotheses (Type I error) when it is true (Dormann et al., 2007; Fortin et al., 1990; Legendre et al., 2004)

Some simplifications and restrictive assumptions such as isotropic and stationary distribution across the image are often required for modelling space explicitly, but it may not be valid for the target plant trait or study area. In remote sensing, pixel values are not only in general spatially autocorrelated, but can also be nonstationary, non-normal, erratically spaced, and discontinuous (Liang, 2005). The spatial structure of the spectral image is not necessarily coincident with the observed in the plant traits, at least in some of

the wavelength regions (Atkinson and Emery, 1999). For instance, soil composition and moisture of a landscape may affect reflectance by these background patterns, while these same factors drive the spatial distribution of the plant trait to other configuration. Differences in pattern or range of spatial autocorrelation between plant trait and spectra may also occur. As demonstrated by Atkinson and Emery (1999), the red and near-infrared wavelengths are related to different physical features, and the two wavelengths present different spatial structure for the same area. The interaction between these two wavelengths may create a third spatial structure. Spatial structures derived from remote sensing procedures, for instance, soil background or cloud cover should be treated as an error and corrected before modelling (Militino et al., 2018). If remote sensing images are available before the sampling campaign, the spatial dependency of plant traits could be estimated and used for defining the optimal distance between samples as discussed in chapter five. The simulations in chapter three and four indicate that significant autocorrelation can be found when the average distance between sampling locations was lower than the spatial dependency.

The spatial signature of the remote sensing imagery also depends on the spatial resolution of the sensor (Shaw and Burke, 2003). Generally, for a coarser spatial resolution, there is less variation in pixel values within the image, being more susceptible to spatial autocorrelation than high-resolution ones (Liang, 2005). Estimating plant traits by remote sensing (excluding lab experiments), it is expected to detect spatial dependency. Nevertheless, which environment was targeted, platform and sensor used, or spatial resolution and extent of the image (Hawkins, 2012; Naimi et al., 2011; Roberts et al., 2017). Hyperspectral data or any remote sensing imaging captured from a continuous area (*e.g.* landscape) is likely to exhibit significant spatial or temporal dependency for most types of vegetation surfaces (Legendre, 1993; Lobo et al., 1998). The underlying process that drives the plant trait most probably impose the spatial dependency captured by a sensor. However, it is likely to be also influenced by spatial autocorrelation in the ground measurements due to sampling campaign procedures, as observations are neither captured simultaneously nor collected randomly, increasing the dependency in space and time.

6.4 Temporal domain

Although the keyword spatiotemporal is quite recurrent in satellite imagery papers, the use of spatiotemporal (stochastic) models is rare in remote sensing applications (Militino et al., 2018). In most cases, spectra are modelled while the other domains are neglected, or individual pixels are analysed through a time series, or spatial pattern from images of different periods are compared (Nguyen and Lee, 2006). These studies may investigate the autocorrelation

observed in the primary domain (modelled), but usually ignore other dependencies and their interactions (Militino et al., 2017). Temporal variations on plant trait observations and remote sensing data are not only expected when a new sample is collected or image captured, but also within of the same field campaign. It occurs, as it is unlikely to measure all the ground references and remote sensing data at the same time in continuous vegetation areas under sunlight (Atkinson and Emery, 1999). In this situation, weather and atmospheric conditions will vary in illumination intensity and scattering effects each time a new observation is made (Shaw and Burke, 2003). Optical measurements tend to vary throughout the day because of changes in the sun angle (Milton et al., 2009). Thus, when several (continuous) measurements of a plant trait are taken at the same position, different values are expected (Pan, 2018). Temporal variations can also occur on a medium to long-term related to changes in weather conditions or plant phenology (Liang, 2005). Collecting ground references during a short period of the day allows controlling the solar azimuth but may require in turn many days to finish the campaign, increasing the differences in plant seasonal variations. On the other hand, an intensive campaign may use many hours per day, increasing variability in illumination conditions and consequently the spatiotemporal dependency for consecutive locations.

Spectra captured by hand-based or drones will rarely be independent in time as sun altitude and weather conditions changes gradually during and over the days (Kumar and Skidmore, 2000). But it is not about temporal autocorrelation in a time-series of target pixels, which were captured to represent different months or seasons of the year. It is about temporal autocorrelation among observations collected in the same sampling campaign. Plant traits and reflectance can be captured both by optical sensors in the field, for instance, LAI measurements using a Plant Canopy Analyser LAI2200 (LICOR Inc., Lincoln, NE USA) and spectral data with a field spectrometer (*e.g.* ASD FieldSpec® 3, Boulder, CO, USA). In this case, temporal autocorrelation would elevate the risk of interaction between spatial and temporal structures, causing undesirable systematic noise in the data (Schaepman-Strub et al., 2006; Schertzer and Lovejoy, 2004). When plant trait observations are sampled in the field but measured in the lab, and spectra are extracted from a single satellite or airborne scene, this risk is relatively lower. In satellite or airborne data, the difference in geometrical distortion within or between scenes may cause spatiotemporal autocorrelation as well (Shaw and Burke, 2003). The recognition of the importance of spatiotemporal structures in ecology has grown recently, but it is not frequently applied because of the dimensionality for modelling explicitly the three domains (Gelfand, 2012; Gelman and Hill, 2006). Time series analysis allows monitoring and detection of changes in a large sequence of remote sensing images, but many studies neglect the spatial dependence during the time series analysis of these images (Militino et al.,

2017). Moreover, spectral, temporal and spatial domains are continuous by essence, and when discretised in many wavelengths, periods or pixels, this information becomes serially correlated. In other words, there is a logical order in the data that should be respected, where pairs of wavelengths, times or locations sampled nearby, are likely to be more similar than pairs positioned further apart (Tobler, 1970).

6.5 Sampling and measuring

6.5.1 Sampling design

The accuracy of empirical models when predicting plant traits using reflectance captured by remote sensing platforms depends significantly on the quantity and quality of the data used for training and testing, as presented in chapter two and five. A sampling design that represents the plant trait population spatially while reducing the temporal and spectral variations is essential to predict accurately (Fortin et al., 1990; Wang et al., 2005). A proper sampling design and well-controlled measurements in the field campaign can significantly reduce random and systematic noises in the data, but will never eliminate it. Less noisy data reduces the risk of spurious correlation while modelling, but a validation using unseen observations from a different sampling campaign is essential to achieve model generalisation and test the influence of spatiotemporal structures. The sampling designs to train models with remote sensing data are commonly constrained by the costs and time of field campaigns (Van Cleemput et al., 2018). A time limit is needed because changes in maximum sun azimuth during the field collection may alter reflectance values. Before defining the sampling design, it is needed to decide (or acknowledge) at which spatial resolution the sensor will capture the spectra from the land surface (Liang, 2005). With this information, the next step is to define the shape of the plot, and the number of plant trait measurements that is needed within the plot to represent the pixel or spectral area (Atkinson, 1997).

The relations between spectra (pixel) and plot area should be adequately tested at the target land surface to minimise aggregation problems and time-space misalignments (Finley et al., 2014; Wilson et al., 2011). Highly heterogeneous land surfaces make in situ measurement of coarse resolution very tough (Liang, 2005). A suitable spatial sampling scheme is vital in environmental monitoring, model calibration, or validation of remote sensing products (Tian et al., 2002). Remote sensing images can indicate the presence of spatial or temporal autocorrelation in a landscape. This information can be used for planning the field campaign through widely accepted empirical relationship between a plant trait and a spectral index (Wang and Gertner, 2013). This a-priori analysis can support the determination of sampling design

and sample size that avoid or adequately model the spatiotemporal structures. In chapter five, it was demonstrated that the sample size might affect generalisation in two different forms. Models trained with small sample sizes increase the risk of overfitting, as the number of observations is limited relatively to the number of wavelengths as showed in chapter two. While larger samples can shorten the distances between observations, changing the density of points, and consequently the sensitivity for spatial autocorrelation. The nugget effect provoked by measurement errors or spatial variations at a smaller scale than sampled can mask the correct range of spatial autocorrelation (Diggle and Ribeiro, 2007). Non-stationarity and spatial anisotropy in the data may reduce the efficiency of the systematic sampling designs, and violates the assumptions of most modelling approaches (Ripley, 1981).

Whether a sampling design aims to capture most of the autocorrelation to fit a spatial model or avoid it to use a non-spatial regression, it should be defined in advance. Sampling should be designed based on the available remote sensing data and on a model approach suitable to the characteristics of such data (Atkinson and Emery, 1999). Where there is no prior knowledge, the focus should be on a regular frame and probabilistic sampling design (Diggle and Ribeiro, 2007). Regular designs better guarantee geographically spread observations across the area. However, regular designs offer a reduced combination of distances, missing especially very close locations (Wang et al., 2012; Webster et al., 1989). In general, systematic designs present reasonable prediction accuracies where there is spatial dependency, and the range of spatial autocorrelation does not coincide with the sampling distance. Where the spatial dependency is much smaller than the spacing between samples, autocorrelation is undetectable, suggesting the use of a non-spatial model approaches (Delmelle, 2009; Wang et al., 2005). However, if the aim is to predict for the entire image where autocorrelation does occur at a finer scale than sampled, applying a non-spatial model may produce biased values in some areas of the images. As shown in chapter five, if a sample is designed to estimate a global parameter such as the mean value, increasing its size brings no significant improvement after a certain point, but redundant observations. In the presence of strong spatial autocorrelation, the sample size can be significantly reduced in the case of global parameters estimation. On the other hand, when the intention is to estimate a plant trait value at an unseen location, the size and the spatial representativeness of the sample might be expanded (Olea, 2018). If a vegetation index related to the plant trait can be extracted from an available remote sensing image, it is possible to design an optimal sample to cover locations with more uncertainty more precisely.

6.5.2 Measuring plant traits

Plant trait measurements are not easily available as field data collection is time-consuming and expensive (Milton et al., 2009). Direct measurements are mainly destructive by chemical laboratory analysis from samples of leaves, such as used to determine chlorophyll concentration. For biomass or leaf area index (LAI) physical measurements are more complicated as it is needed to harvest all the leaves from plants (Lee et al., 2004). Extensive areas are typically covered by (square) plots, in its turn are represented by many leaf or full canopies samples to be comparable with remote sensing data. The difficulty to collect direct measurements in a fragile biome or remote environments, hamper these areas to be covered by such measurements (Vallejos and Osorio, 2014). In general, plant trait datasets which are publicly available are not directly comparable as they were measured by different instruments and methods (Van Cleemput et al., 2018). The lack of comprehensive and standardised datasets diminishes the prospect of mapping plant traits at finer temporal and spatial scales (Hoeting, 2009; Muñoz-Huerta et al., 2013; Secades et al., 2014).

Indirect procedures to measure plant traits are needed to observe and to monitor vegetation dynamics more efficiently (Pearse et al., 2016). Optical instruments are the most common approach to approximate plant trait values as they are non-destructive and can be measured *in situ* (Milton et al., 2009). Although many instruments have shown efficiency in measuring plant traits, their accuracy should be assessed regarding the reproducibility and repeatability in the same environmental conditions as customary in the field campaign. Most of the measurement accuracies specified by instrument guides refer to controlled lab experiments and may present significant discrepancy in precision and bias in heterogeneous vegetation under sunlight illumination. For instance, LAI can be measured through a variety of techniques and instruments, by indirect measurements from a light-sensitive instrument such as LICOR's LAI-2200 or by analysing hemispherical photos (Liang, 2005). However, measuring LAI with optical instruments on vegetation surfaces composed of small leaves and coniferous needles is still intricate (Liang, 2005). Biochemical concentrations are usually measured in laboratories, but handheld optical instruments such as Minolta's SPAD have been widely used to measure chlorophyll, although its reliability and accuracy is often questioned (Vohland and Jarmer, 2008). Optical measurement of plant traits may suffer spatiotemporal autocorrelation according to time and location sequence, similar to what may occur with reflectance using a field spectrometer as cited before (Figure 6.2).

Figure 6.2 - Sequence of LAI values according to the order in which they were measured using the LAI2200 instrument under natural sunlight (a-right); and solar zenith angle during the LAI collection using the same sequence (b-left).

Samples of plant traits are usually described as randomly assigned locations, although most publications do not state it explicitly, data most probably were not collected in that manner. In situ campaigns usually collect the observations in a sequence to reduce time and distance, rather than genuinely collecting in random order. These practices increase the dependency between close by observations, being a combination of spatial and temporal autocorrelation. Ground references should not be so time-consuming that alteration on the vegetation properties can be observed, avoiding data structure related to the pace of the sample collection. Leaf properties such as size, mass, pigment and water content at early spring in a temperate climate, may change quite fast compared with the rhythm of the collection in timing-consuming field campaigns. For using the full potential of remote sensing to monitor plant traits, it is necessary to apply suitable sampling designs and standardised measurement procedures to make comparable observations from different instruments, institutions, places and periods (Van Cleemput et al., 2018).

6.5.3 Measuring spectra

Optical sensors are built with a particular spectral resolution, and the platform they are deployed with determines the spatial and temporal resolution (Plaza et al., 2009). Satellites, aircraft, drones or hand-based platforms vary greatly in spatial resolutions, but also in the quantity of noise provoked by factors as atmospheric condition, illumination geometry and soil background (Milton et al., 2009; Reichenau et al., 2016). The spectra captured by these platforms range from individual round areas to large scenes based on a regular (square or stripes) grid of pixels (Shaw and Burke, 2003). It may be possible to select the platform and the sensor that is most appropriate or convenient for the application, but in most of the cases, there are limited choices of different resolutions available. The sensor altitude, the scene extent and the spatial resolution will determine the amount geometric distortion as the pixel departs from the nadir (Liang, 2005). Usually, these factors are also related to the temporal resolution, defining the periodicity (regular or not) until the next

measurement in the same area. However, for instance, when using platforms such as aircrafts or drones there will be temporal variations within the same field campaign, as the full image with the mosaic of the individual stripes can take hours or days to be completed. Drones are probably the most flexible regards to spatial resolutions, while handheld spectrometers are often more flexible over temporal resolution.

Field spectrometers under natural lighting utilise a white panel reference (Lambertian surface) to convert the readings of the instrument to reflectance given a specific viewing direction (Milton et al., 2009). Field campaigns, in this case, should take measurements at similar sun altitude to avoid systematic spatiotemporal noises. Field spectrometers often have hyperspectral sensors that comprise a very comprehensive spectral range. However, some regions of the spectra present a significant amount of noise caused by the absorption of CO2 and water vapour, among other atmospheric effects, when capturing the land surface in the field (Liang, 2005; Shaw and Burke, 2003). Data pre-processing procedures such as filters to smooth spectral noise, to make atmospheric corrections and to average subplots or pixels (*e.g.* window) may reduce random and systematic error (Militino et al., 2018). On the other hand, these procedures may increase multicollinearity, spatial autocorrelation or generate aggregation problems. Using a comprehensive range of narrow bands as presented by hyperspectral data, noisy wavelength regions of the spectrum should be removed before modelling (Shaw and Burke, 2003). Being a product of two or more spectral bands, vegetation indices are often used as a covariate when modelling plant traits. The coefficient of two wavelengths is more stable to variations in natural illumination than individual bands, and thus spectral indices may be more appropriate variables for modelling (Liang, 2005).

6.5.4 Alignment of plant traits with spectra

The wider an image is, the less noise will be caused by temporal variations in illumination and atmospheric conditions. However, capturing more pixels simultaneously will probably increase the time misalignment with the ground references (Zeng et al., 2015). For instance, field spectrometers can capture individual areas (scene of one pixel at time), allowing very low time-space misalignment between spectra and ground references. Contrarily, satellites capture simultaneously large scenes based on a fixed grid of pixels, making it impossible to measure ground references at the same (time-space) pace. Sensors deployed in aircrafts or drones capture images in stripes, varying in time in two directions, within and between stripes (Atkinson, 1997). Most probably, the plant trait field collection will present a sampling path which is not coincident with airborne route. Plant trait and reflectance should be measured at the same time, but, it is rarely possible to do both simultaneously,

and time-space misalignment will always occur because of sampling design constraints.

Spectral measurements are not independent in time, nor space, and the spectral domain interacts with the spatial and temporal domain (Militino et al., 2018). Therefore, reflectance should be collected as simultaneously as possible with ground references, and at a similar spatial resolution to reduce misalignment and minimise variations unrelated to the plant trait (Atkinson and Emery, 1999). Averaging subplots and pixels to a broader size (*i.e.* window) can reduce the effect of misalignment, but may cause a smoothing effect, reducing the natural variability of the plant trait and the reflectance excessively. The geometric distortion in spectral images can enhance the mismatching between plot locations with the correspondent pixel units when it deviates from the nadir position (Liang, 2005).

The pixel size of satellite data is usually much greater than the corresponding plant trait at leaf or canopy level, necessitating the collection of points within a plot to represent the same area (Wilson et al., 2011). As spatial resolution becomes coarser, in heterogeneous land surfaces, it is needed to scale up from point to plot scale and then to pixel scale to guarantee reliable estimates of the plant trait. For instance, (effective) LAI values from remotely sensed data captured at a coarse resolution might present very distinctive (true) values if the vegetation is heterogeneous (Liang, 2005). This difference between the ideal plant trait scale and the spatial resolution of the available reflectance is called a change of support problems, which occurs independently of the quality of the spatial alignment (Ullah et al., 2012). The readings of reflectance stored in a pixel present a different aggregation method than plant trait ground references within the sample unit (*e.g.* plot), regardless the spatial alignment. Although spatial or temporal misalignment can mask the relationship between reflectance and plant trait, some degree of mismatching is accepted in exchange for an executable field campaign (Gotway and Young, 2002). However, misalignment should be small, or otherwise, the reflectance may not correspond to the vegetation surface measured as ground references (Atkinson and Emery, 1999). In the presence of strong spatial dependency, the (spatial) misalignment is attenuated as nearby locations are likely to present similar reflectance.

6.6 Modelling and assessment

Modelling plant traits using hyperspectral remote sensing data faces challenges to deal with the high dimensionality derived from the spectral and spatial domains. These challenges are amplified when the number of field observations for training the model is considerably smaller than the number of parameters needed to represent these domains (Zhao et al., 2013). Plant traits may also

be retrieved by physical models based on the spectral properties of the land surface, illumination conditions and sensor geometry (Jacquemoud et al., 2009; Van Cleemput et al., 2018). Radiative transfer models (RTM) are often used to retrieve plant traits, but it remains challenging to set realistic parameters using reflectance captured from controlled lab environments (Combal et al., 2002; Goodenough et al., 2006). Therefore, empirical models are often used to predict plant trait using remote sensing data (Goodenough et al., 2006).

6.6.1 Modelling with hyperspectral remote sensing data

Modelling with hyperspectral data is likely to present multicollinearity and overfitting issues because of the large number of narrow autocorrelated wavelengths, as was demonstrated in the second chapter. Problems with multicollinearity and model selection are even more common where all the spectra data is captured from similar land surfaces as presented in the introduction (Cho et al., 2007). The wavelengths are commonly treated as independent covariates in the model, which is a fundamental mistake, as wavebands are often redundant and strongly correlated. In other words, linear combinations of bands can falsely inflate the importance of the variable in the model (Gelman et al., 2001; Kuhn and Johnson, 2013). Extracting spectral indices related to the plant trait through a-priori knowledge is ideal, but constrained by the possibilities of the current knowledge (Liang, 2005). The uses of supervised methods (with the support of the response variable) or machine learning greatly facilitate the band selection using hyperspectral data but increase enormously the risk of overfitting as well (Lee et al., 2004).

Searching a spectral index of two or more bands which explains the response variable presents a similar risk of overfitting as a model selection procedure by stepwise method or genetic algorithms. An ordinary least squares regression using a single spectral index or a couple of bands as covariates, if performed by supervised methods, it will present a high risk of overfitting despite being a very simple model. Therefore, modelling using a large set of hyperspectral wavelengths to search for correlations with the support of the response variable causes overfitting. With higher levels of noise in hyperspectral data, it is more likely that the model will be overfitted by spurious correlation, as demonstrated in the second chapter. Overfitting occurs when the model is excessively complex or the covariates were previously selected by a supervised approach (James et al., 2013). Machine learning algorithms often applied in remote sensing were initially designed to classify pixels rather than fit regression models with continuous response variables (Liang, 2005). For classification with a limited number of categories, complex algorithms may be tolerable, but estimating continuous variables requires an understanding of the model function and its assumptions (Gelman et al., 2001). Some of the

regression techniques such as Random Forests will rarely show similar accuracy in the training and testing sets, independently of the tuning method applied as demonstrated in chapter two (Dormann et al., 2013).

Machine learning algorithms often result in overly complex models with a low capacity of understanding about the empirical relationships. Complex models suffer then from a loss of generalisation, being not transferable to another area, period, vegetation species, optical instrument or any other change. Therefore, machine learning models need to have their complexity constrained, by limiting the number of the bands selected (or searched) to avoid overfitting and thus unreliable predictions (James et al., 2013). In chapter two, a method called naïve overfitting index selection (NOIS) was developed to tune model complexity to reduce the risk of overfitting. The NOIS method generates simulated spectra using the covariance matrix of the original hyperspectral data. These data are used for tuning the maximum level of complexity supported before the model starts overfitting. This alternative tuning process allows optimisation of the model complexity according to the available data structure, balancing the trade-off between accuracy and overfitting. Tuning processes using cross-validation fail to indicate whether the model complexity is adequate to the data available, or quantify the amount of overfitting as the NOIS method.

Hyperspectral data are frequently treated as independent and identically distributed across wavelengths, but also randomly in space and in time while modelling (Babcock et al., 2013; Wikle and Hooten, 2010). As discussed earlier, it is unlikely that accurate plant trait measurements in continuous fields are free of spatial structures (Hawkins, 2012). Besides, it is commonly assumed that observations are randomly distributed in remote sensing. This assumption may be valid for widely spaced sample locations when using high-resolution remote sensing imagery (Dalposso et al., 2013; Griffith and Chun, 2016). The result of the third chapter showed that spatial autocorrelation could affect the reliability of the prediction accuracy using machine learning or non-spatial models. However, in general, non-spatial models are affected by spatial autocorrelation, and in complex models like machine learning algorithms, this effect is much stronger. The patterns provoked by the spatial dependency will remain in the model residuals as demonstrated in the third chapter. Machine learning algorithms may partially mask the autocorrelation in the final model as some of the algorithms use the residuals to improve accuracy (Hastie et al., 2009). The absence of model assumptions about spatial autocorrelation from these approaches does not imply that their effect on the predictions is marginal.

The decision about which of the three domains should be prioritised while modelling depends on how significant the spatiotemporal dependency is. In the

fourth chapter, it was demonstrated that in the presence of significant spatial dependency, it is better to reduce the number of bands in the spectral domain and include space explicitly in the model, rather than use the full spectrum. Tuning a trade-off between spectral and spatial domains improves the prediction accuracy considerably at unseen locations. Spatial models can also be overfitted by the excess of spatial parameters, where small alterations in the spatial pattern provoke strong effects in the prediction values (Gelfand, 2012). For this reason, a tuning process to select a neighbour matrix to account for the spatial information for the landscape under consideration is necessary (Bakka et al., 2018; Lindgren and Rue, 2015). The optimal neighbourhood matrix should be selected based on the spatial complexity that reduces the autocorrelation in residuals the most while minimising the prediction error of the test data set. Despite the importance in ecological processes to model the temporal domain explicitly, they are usually not incorporated in empirical models because of the complexity of spatiotemporal models (Fortin et al., 2012; Militino et al., 2018). Spectra, time and space are continuous domains, that tend to be high dimensionally and serially autocorrelated when discretised, requiring adequate modelling approaches. Metrics to assess accuracy during model selection in machine learning regressions do not take into account the lack of parsimony as usual in ordinary least square regressions. Seeking accuracy by minimising the prediction error using highly dimensional data and very complex model tends to produce unrealistically small errors. The absence of relevant explanatory variables other than spectral should not be understood as lack of accuracy related to the modelling process, justifying the inflation of model complexity with redundant wavebands.

6.6.2 Assessment and generalisation

The assessment of accuracy is strongly advised in predictive models, and in the case of high dimensional hyperspectral data, even more so. The uncertainty related to the measurement system discussed earlier, and the tendency to present spatiotemporal dependency further increases the necessity of in-depth assessment. Hyperspectral remote sensing data are inclined to introduce considerable random noise in some regions of the spectrum that provokes model overfitting by fitting spurious correlation. Therefore, tuning models based on the minimisation of prediction error using supervised methods provide unreliable accuracy. The assessment of the prediction accuracy should be made using an unseen dataset, and if possible, from a different sampling campaign to test the influence of spatiotemporal dependent errors (Gelman et al., 2001).

As discussed in the two first chapters, some metrics to assess the quality of the fitness, such as R^2adj, AIC and BIC, are not suitable for predictive models and cannot be compared over different regressions approaches. The most

straightforward way to report model accuracy is by the Root Mean Squared Error (RMSE), calculated from the differences between the prediction and the plant trait measurements. As the number of observations is often limited, it is common to allocate most of the data for training the model instead test it (Hawkins, 2004). For this reason, cross-validation is probably the most used method to assess model accuracy in remote sensing applications. As demonstrated in chapter two, for intensive tuning processes with overly complex machine learning models, this method fails in avoid overfitting. Therefore, it is essential to compare the prediction accuracy from the testing set (or cross-validated) with the training set. Large differences imply that the model is overfitted and the complexity must be reduced (Dormann et al., 2013). The method and the proportion of the observations that should be used for assessing the accuracy depend on data availability and the heterogeneity of the target population (Kuhn and Johnson, 2013). Regardless of the empirical method applied to predict plant traits, wavelengths are selected in the model mostly by the correlation found through the data (empirically), rather than based on physical theory or previous knowledge.

Although scientific publications mainly focus on model accuracy, there are plenty of assumptions to be checked depending on the regression approach applied (Gelman et al., 2001). For instance, little attention has been given to the spatial configuration of model residuals (Moisen and Frescino, 2002; Zhang et al., 2005). In chapter three and four, the effects of the spatial autocorrelation in the model residuals were presented and the risk of unreliable prediction from machine learning regressions when this occurs. The spatiotemporal autocorrelation can be assessed using the Durbin Watson or a similar test using the sequence in which the data were collected in the field (spatiotemporal order). Otherwise, Moran's I index or Geary's coefficient using the coordinates of the observations can be applied to detect spatial autocorrelation in the residuals (Diggle and Ribeiro, 2007). Complex models such as the product of machine learning algorithms seem to learn from spatial structures, suggesting an unreliable increase in performance. The assessment of these models has shown that they are unable to eliminate spatial autocorrelation in the residuals despite the lack of explicit assumptions of i.i.d. observations in these regression models.

Other assessments of model residuals such as normality and homoscedasticity of variance should also be tested to identify non-random behaviour (Gelman and Hill, 2006). Multicollinearity should be examined for ordinary least squares regression or generalised linear models using two or more covariates, but it is not necessary for machine learnings or penalised regressions (James et al., 2013; Kuhn and Johnson, 2013). These assessments are required to report a more reliable prediction accuracy and also lead to more generalisation power in the model. Both testing sets or cross-validation estimations, typically

derived from the same field campaign and the model may present limited generalisation when applied to a new sample (Roberts et al., 2017). The reason is that the sequence in which the data were collected may lead to spatially and temporally autocorrelated observations with a different data structure than the previous sample (Brenning, 2012; Roberts et al., 2017). Given all the (spatial, temporal and spectral) variations involved when using hyperspectral data, the modelling process requires not only the definition of an appropriate regression method but also a correct approach to select explanatory variables among all possible candidates, and then a careful assessment to the best-fitted model.

6.7 Applying and replicating

The process of observing, understanding and predicting vegetation dynamics has evolved fast since the start of remote sensing applications. However, the modelling uncertainties will always be present as inherent processes have a stochastic nature. The advancements in remote sensing technology create possibilities to observe the vegetation dynamics over a more comprehensive range of spatial and temporal scales. However, the dimensionality of the three domains involved when modelling plant trait makes this process a very complex task (Militino et al., 2018). The discretisation of continuous domains as the case of spectra, space or temporal certainly lose some amount of information in the aggregation's process. Over aggregated, these domains may lose the connection with the target plan trait. Contrarily, over desegregated, these domains generate redundant information.

Predicting plant trait by remote sensing at a landscape level with multiple species becomes especially useful and meaningful when it is comparable across space and (or) time. Measuring reflectance from a vegetation surface several times presumably under identical conditions will result in different values each realisation, but its (parametric) distribution can be estimated consistently based on a stochastic process (Klyatskin, 2017). Stochasticity means that the outcome of a process involves the occurrence of random events. Therefore no full control of the results or clear explanation about the patterns is possible. The unpredictability of remote sensing imagery is not comparable with the chaotic nature of the atmosphere, for instance, but is notably affected by it (Franzke et al., 2015). Variations in the spectral signal can indicate changes in biomass (leaf area) or photosynthetic trait (leaf colour). There is a huge list of uses and users for efficient methods of predicting plant traits, ranging from agribusiness to climatology. For instance, plant trait prediction can support decisions on precision agriculture and crop management associated with fertilising, pesticides or irrigation. The decisions when, where and how much to apply can be based on plant traits such as leaf area index, chlorophyll leaf content or leaf water content. Plant trait prediction for applications in conservation goes beyond the traditional detection of deforestation or changes

in land cover. It allows monitoring surrogate indicators of vegetation dynamics such as plant stress, parasites infestation or invasive species.

Remote sensing can accelerate the process of observing and monitoring plant traits, but to understand the spatiotemporal variations of vegetation, more comprehensive knowledge is required. Statistical knowledge is not enough to avoid wrong inferences about the underlying process that drives the plant trait. It needs expertise in spectral radiance, biochemical process, ecology and phenology of plants. The premise that a combination of wavelengths can (partially) explain a target plant trait is valid (Curran, 1989). However, most of the spectral indices and wavelengths from hyperspectral remote sensing used in models are selected empirically by a limited amount of ground references rather than by knowledge (Liang, 2005). Not only spectral information is fundamental to predict plant trait, but also spatial knowledge is needed. As shown in chapter four, spatial models increase the prediction accuracy significantly by tuning a trade-off between spectral and spatial domains. The definition of the sampling design may affect the decision of which regression method is suitable for the spatiotemporal autocorrelation captured in the data by the sample. For this reason, predictive models using remote sensing data should be carefully assessed and the residuals reported, indicating that the regression method is suitable for the particular data structure.

The advancement of big data and machine learning algorithms fuels the belief that everything can be modelled and predicted, even in the absence of understanding about the underlying processes (Hawkins, 2012). Models that predict accurately only for the specific dataset which they were trained are not rare. Most of these models will never be tested again or replicated in similar conditions. The lack of model generalisation creates a need for fieldwork every time a new image is captured, which invalidates the increase in speed promised by remote sensing applications. The obsession for providing the highest coefficient of regression (R^2) should be replaced for enthusiasm to replicate the same models with similar accuracy in other places and periods. It will increase the ability to produce long-term monitoring of natural vegetation over vast continuous areas. The methodologies developed and discussed in this thesis intend to contribute with good practices to predict plant trait with hyperspectral data. More than demonstrating many modelling issues and technicalities such as multicollinearity, overfitting and autocorrelation in the residuals, this study shows the potential of spectra-space tuning to increase the accuracy of predictive model significantly. The remote sensing and modelling technologies will undoubtedly improve in the future, but the nature of the data and most of the core principles discussed here will remain as currently.

Bibliography

Abdel-Rahman, E.M., Ahmed, F.B., Ismail, R., 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. Int. J. Remote Sens. 34, 712–728. https://doi.org/10.1080/01431161.2012.713142

Abdullah, H., Darvishzadeh, R., Skidmore, A.K., Groen, T.A., Heurich, M., 2018. European spruce bark beetle (Ips typographus, L.) green attack affects foliar reflectance and biochemical properties. Int. J. Appl. Earth Obs. Geoinformation 64, 199–209. https://doi.org/10.1016/j.jag.2017.09.009

Atkinson, P.M., 1997. Selecting the spatial resolution of airborne MSS imagery for small-scale agricultural mapping. Int. J. Remote Sens. 18, 1903–1917. https://doi.org/10.1080/014311697217945

Atkinson, P.M., Emery, D.R., 1999. Exploring the relation between spatial structure and wavelength: Implications for sampling reflectance in the field. Int. J. Remote Sens. 20, 2663–2678. https://doi.org/10.1080/014311699212001

Atkinson, P.M., Webster, R., Curran, P.J., 1992. Cokriging with ground-based radiometry. Remote Sens. Environ. 41, 45–60. https://doi.org/10.1016/0034-4257(92)90060-W

Babcock, C., Matney, J., Finley, A.O., Weiskittel, A., Cook, B.D., 2013. Multivariate Spatial Regression Models for Predicting Individual Tree Structure Variables Using LiDAR Data. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6, 6–14. https://doi.org/10.1109/JSTARS.2012.2215582

Bakka, H., Rue, H., Fuglstad, G.-A., Riebler, A., Bolin, D., Krainski, E., Simpson, D., Lindgren, F., 2018. Spatial modelling with R-INLA: A review. ArXiv180206350 Stat.

Banerjee, S., Finley, A., 2007. Bayesian multi-resolution modeling for spatially replicated data sets with application to forest biomass data. J. Stat. Plan. Inference 137, 3193–3205. https://doi.org/10.1016/j.jspi.2006.05.024

Banerjee, S., Fuentes, M., 2012. Bayesian modeling for large spatial datasets: Bayesian modeling for large spatial datasets. Wiley Interdiscip. Rev. Comput. Stat. 4, 59–66. https://doi.org/10.1002/wics.187

Berger, K., Atzberger, C., Danner, M., D'Urso, G., Mauser, W., Vuolo, F., Hank, T., 2018. Evaluation of the PROSAIL Model Capabilities for Future Hyperspectral Model Environments: A Review Study. Remote Sens. 10, 85. https://doi.org/10.3390/rs10010085

Bioucas-Dias, J.M., Nascimento, J.M.P., 2008. Hyperspectral Subspace Identification. IEEE Trans. Geosci. Remote Sens. 46, 2435–2445. https://doi.org/10.1109/TGRS.2008.918089

Bivand, R.S., Gómez-Rubio, V., Rue, H., 2015. Spatial Data Analysis with *R* - **INLA** with Some Extensions. J. Stat. Softw. 63. https://doi.org/10.18637/jss.v063.i20

Boegh, E., Houborg, R., Bienkowski, J., Braban, C.F., Dalgaard, T., Magliulo, V., Schelde, K., Tommasi, P.D., Vitale, L., Theobald, M.R., Cellier, P., Sutton, M.A., 2013. Remote sensing of LAI, chlorophyll and leaf nitrogen pools of crop- and grasslands in five European landscapes 29.

Bousquet, O., Elisseeff, A., 2002. Stability and Generalization. J. Mach. Learn. Res. 2, 499–526.

Brenning, A., 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. IEEE, pp. 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393

Bruce, L.M., Koger, C.H., Jiang Li, 2002. Dimensionality reduction of hyperspectral data using discrete wavelet transform feature extraction. IEEE Trans. Geosci. Remote Sens. 40, 2331–2338. https://doi.org/10.1109/TGRS.2002.804721

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–44. https://doi.org/10.1016/S0016-7061(97)00072-4

Buitrago Acevedo, M.F., Groen, T.A., Hecker, C.A., Skidmore, A.K., 2017. Identifying leaf traits that signal stress in TIR spectra. ISPRS J. Photogramm. Remote Sens. 125, 132–145. https://doi.org/10.1016/j.isprsjprs.2017.01.014

Buitrago, M.F., Groen, T.A., Hecker, C.A., Skidmore, A.K., 2018. Spectroscopic determination of leaf traits using infrared spectra. Int. J. Appl. Earth Obs. Geoinformation 69, 237–250. https://doi.org/10.1016/j.jag.2017.11.014

Burket, G.R., 1943. A study of reduced rank models for multiple prediction. Wash. UNIV SEATTLE.

Carvalho, S., Macel, M., Schlerf, M., Moghaddam, F.E., Mulder, P.P.J., Skidmore, A.K., van der Putten, W.H., 2013. Changes in plant defense chemistry (pyrrolizidine alkaloids) revealed through high-resolution spectroscopy. ISPRS J. Photogramm. Remote Sens. 80, 51–60. https://doi.org/10.1016/j.isprsjprs.2013.03.004

Chen, J.M., Black, T.A., 1992. Defining leaf area index for non-flat leaves. Plant Cell Environ. 15, 421–429. https://doi.org/10.1111/j.1365-3040.1992.tb00992.x

Chipeta, M.G., Terlouw, D.J., Phiri, K.S., Diggle, P.J., 2016. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. ArXiv160500104 Stat.

Cho, M.A., Ramoelo, A., Debba, P., Mutanga, O., Mathieu, R., van Deventer, H., Ndlovu, N., 2013. Assessing the effects of subtropical forest fragmentation on leaf nitrogen distribution using remote sensing data. Landsc. Ecol. 28, 1479–1491. https://doi.org/10.1007/s10980-013-9908-7

Cho, M.A., Skidmore, A., Corsi, F., van Wieren, S.E., Sobhan, I., 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. Int. J. Appl. Earth Obs. Geoinformation 9, 414–424. https://doi.org/10.1016/j.jag.2007.02.001

Clevers, J.G.P.W., Kooistra, L., 2012. Using Hyperspectral Remote Sensing Data for Retrieving Canopy Chlorophyll and Nitrogen Content. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 574–583. https://doi.org/10.1109/JSTARS.2011.2176468

Clevers, J.G.P.W., Kooistra, L., Schaepman, M.E., 2010. Estimating canopy water content using hyperspectral remote sensing data. Int. J. Appl. Earth Obs. Geoinformation 12, 119–125. https://doi.org/10.1016/j.jag.2010.01.007

Cochran, W.G., 1977. Sampling techniques, 3d ed. ed, Wiley series in probability and mathematical statistics. Wiley, New York.

Cochrane, M.A., 2000. Using vegetation reflectance variability for species level classification of hyperspectral data. Int. J. Remote Sens. 21, 2075–2087. https://doi.org/10.1080/01431160050021303

Combal, B., Baret, F., Weiss, M., Trubuil, A., Mace, D., Pragnere, A., Myneni, R., Knyazikhin, Y., Wang, L., 2002. Retrieval of canopy biophysical variables from bidirectional reflectance Using prior information to solve the ill-posed inverse problem. Remote Sens. Environ. 15.

Corsten, L.C.A., Stein, A., 1994. Nested sampling for estimating spatial semivariograms compared to other designs. Appl. Stoch. Models Data Anal. 10, 103–122. https://doi.org/10.1002/asm.3150100205

Curran, P.J., 2001. Remote sensing: Using the spatial domain. Remote Sens. 14.

Curran, P.J., 1989. Remote sensing of foliar chemistry. Remote Sens. Environ. 30, 271–278. https://doi.org/10.1016/0034-4257(89)90069-2

Dalposso, G.H., Uribe-Opazo, M.A., Mercante, E., Lamparelli, R.A.C., 2013. Spatial autocorrelation of ndvi and gvi indices derived from landsat/tm images for soybean crops in the western of the state of Paraná in 2004/2005 crop season. Eng. Agríc. 33, 525–537. https://doi.org/10.1590/S0100-69162013000300009

Darvishzadeh, R., Atzberger, C., Skidmore, A., Schlerf, M., 2011. Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models. ISPRS J. Photogramm. Remote Sens. 66, 894–906. https://doi.org/10.1016/j.isprsjprs.2011.09.013

Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F., Cho, M., 2008. LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. ISPRS J. Photogramm. Remote Sens. 63, 409–426. https://doi.org/10.1016/j.isprsjprs.2008.01.001

de Gruijter, J.J., Bierkens, M.F.P., Brus, D.J., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33161-1

Delalieux, S., Somers, B., Hereijgers, S., Verstraeten, W., Keulemans, W., Coppin, P., 2008. A near-infrared narrow-waveband ratio to determine Leaf Area Index in orchards. Remote Sens. Environ. 112, 3762–3772. https://doi.org/10.1016/j.rse.2008.05.003

Delmelle, E., 2009. Spatial Sampling, in: The SAGE Handbook of Spatial Analysis. SAGE Publications, Ltd, 1 Oliver's Yard, 55 City Road, London England EC1Y 1SP United Kingdom, pp. 182–206. https://doi.org/10.4135/9780857020130.n10

Diggle, P., Lophaven, S., 2006. Bayesian Geostatistical Design. Scand. J. Stat. 33, 53–64. https://doi.org/10.1111/j.1467-9469.2005.00469.x

Diggle, P., Ribeiro, P.J., 2007. Model-based geostatistics, Springer series in statistics. Springer, New York, NY.

Ding, Y., Ge, Y., Hu, M., Wang, Jinfeng, Wang, Jianghao, Zheng, X., Zhao, K., 2014. Comparison of spatial sampling strategies for ground sampling and validation of MODIS LAI products. Int. J. Remote Sens. 35, 7230–7244. https://doi.org/10.1080/01431161.2014.967889

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dormann, C.F., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628. https://doi.org/10.1111/j.2007.0906-7590.05171.x

Dutilleul, P., 1993. Spatial Heterogeneity and the Design of Ecological Field Experiments. Ecology 74, 1646–1658. https://doi.org/10.2307/1939923

Esbensen, K.H., Geladi, P., 2010. Principles of Proper Validation: use and abuse of re-sampling for validation. J. Chemom. 24, 168–187. https://doi.org/10.1002/cem.1310

Farifteh, J., Van der Meer, F., Atzberger, C., Carranza, E.J.M., 2007. Quantitative analysis of salt-affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). Remote Sens. Environ. 110, 59–78. https://doi.org/10.1016/j.rse.2007.02.005

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sens. Environ. 154, 102–114. https://doi.org/10.1016/j.rse.2014.07.028

Feilhauer, H., Asner, G.P., Martin, R.E., 2015. Multi-method ensemble selection of spectral bands related to leaf biochemistry. Remote Sens. Environ. 164, 57–65. https://doi.org/10.1016/j.rse.2015.03.033

Feilhauer, H., Somers, B., van der Linden, S., 2017. Optical trait indicators for remote sensing of plant species composition: Predictive power and seasonal variability. Ecol. Indic. 73, 825–833. https://doi.org/10.1016/j.ecolind.2016.11.003

Feret, J.-B., François, C., Asner, G.P., Gitelson, A.A., Martin, R.E., Bidel, L.P.R., Ustin, S.L., le Maire, G., Jacquemoud, S., 2008. PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments. Remote Sens. Environ. 112, 3030–3043. https://doi.org/10.1016/j.rse.2008.02.012

Finley, A.O., Banerjee, S., Cook, B.D., 2014. Bayesian hierarchical models for spatially misaligned data in R. Methods Ecol. Evol. 5, 514–523. https://doi.org/10.1111/2041-210X.12189

Fortin, M.-J., Drapeau, P., Legendre, P., 1990. Spatial autocorrelation and sampling design in plant ecology, in: Grabherr, G., Mucina, L., Dale, M.B., Ter Braak, C.J.F. (Eds.), Progress in Theoretical Vegetation Science. Springer Netherlands, Dordrecht, pp. 209–222. https://doi.org/10.1007/978-94-009-1934-1_18

Fortin, M.-J., James, P.M.A., MacKenzie, A., Melles, S.J., Rayfield, B., 2012. Spatial statistics, spatial regression, and graph theory in ecology. Spat. Stat. 1, 100–109. https://doi.org/10.1016/j.spasta.2012.02.004

Franzke, C.L.E., O'Kane, T.J., Berner, J., Williams, P.D., Lucarini, V., 2015. Stochastic climate theory and modeling: Stochastic climate theory and modeling. Wiley Interdiscip. Rev. Clim. Change 6, 63–78. https://doi.org/10.1002/wcc.318

Gelfand, A.E., 2012. Hierarchical modeling for spatial data problems. Spat. Stat. 1, 30–39. https://doi.org/10.1016/j.spasta.2012.02.005

Gelman, A., Hill, J., 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Leiden.

Gelman, A., Park, D.K., Ansolabehere, S., Price, P.N., Minnite, L.C., 2001. Models, assumptions and model checking in ecological regressions. J. R. Stat. Soc. Ser. A Stat. Soc. 164, 101–118. https://doi.org/10.1111/1467-985X.00190

Goodenough, D., Li, J., Asner, G., Schaepman, M., Ustin, S., Dyk, A., 2006. Combining Hyperspectral Remote Sensing and Physical Modeling for Applications in Land Ecosystems. IEEE, pp. 2000–2004. https://doi.org/10.1109/IGARSS.2006.518

Gotway, C.A., Young, L.J., 2002. Combining Incompatible Spatial Data. J. Am. Stat. Assoc. 97, 632–648. https://doi.org/10.1198/016214502760047140

Griffith, D., Chun, Y., 2016. Spatial Autocorrelation and Uncertainty Associated with Remotely-Sensed Data. Remote Sens. 8, 535. https://doi.org/10.3390/rs8070535

Haining, R., 1988. Estimating spatial means with an application to remotely sensed data. Commun. Stat. - Theory Methods 17, 573–597. https://doi.org/10.1080/03610928808829641

Haining, R.P., 2003. Spatial data analysis: theory and practice. Cambridge University Press, Cambridge, UK ; New York.

Hansen, P.M., Schjoerring, J.K., 2003. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. Remote Sens. Environ. 86, 542–553. https://doi.org/10.1016/S0034-4257(03)00131-7

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd ed. ed, Springer series in statistics. Springer, New York, NY.

Hawkins, B.A., 2012. Eight (and a half) deadly sins of spatial analysis: Spatial analysis. J. Biogeogr. 39, 1–9. https://doi.org/10.1111/j.1365-2699.2011.02637.x

Hawkins, D.M., 2004. The Problem of Overfitting. J. Chem. Inf. Comput. Sci. 44, 1–12. https://doi.org/10.1021/ci0342472

Heaton, M.J., Christensen, W.F., Terres, M.A., 2017. Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences. Technometrics 59, 93–101. https://doi.org/10.1080/00401706.2015.1102763

Hoeting, 2009. The Importance of Accounting for Spatial and Temporal Correlation in Analyses of Ecological Data. Ecol. Appl. 19, 574–577.

Holmes, K.W., Niel, K.V., Kendrick, G., 2006. Designs for remote sampling: review, discussion, examples of sampling methods and layout and scaling issues 72.

Huber, S., Kneubühler, M., Psomas, A., Itten, K., Zimmermann, N.E., 2008. Estimating foliar biochemistry from hyperspectral data in mixed forest canopy. For. Ecol. Manag. 256, 491–501. https://doi.org/10.1016/j.foreco.2008.05.011

Ingebrigtsen, R., Lindgren, F., Steinsland, I., 2014. Spatial models with explanatory variables in the dependence structure. Spat. Stat. 8, 20–38. https://doi.org/10.1016/j.spasta.2013.06.002

International Symposium on Recent Advances in Quantitative Remote Sensing, Sobrino, J.A. (Eds.), 2002. Proceedings of the First International Symposium on Recent Advances in Quantitative Remote Sensing: Auditori de Torrent, Spain : 16-20 september 2002. Publicacions de la Universitat de València, València.

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P.J., Asner, G.P., François, C., Ustin, S.L., 2009. PROSPECT+SAIL models: A review of use for vegetation characterization. Remote Sens. Environ. 113, S56–S66. https://doi.org/10.1016/j.rse.2008.01.026

James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), 2013. An introduction to statistical learning: with applications in R, Springer texts in statistics. Springer, New York.

Jarocińska Anna M., 2014. Radiative Transfer Model parametrization for simulating the reflectance of meadow vegetation. Misc. Geogr. 18, 5. https://doi.org/10.2478/mgrsd-2014-0001

Klyatskin, V.I., 2017. Fundamentals of Stochastic Nature Sciences, Understanding Complex Systems. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-56922-2

Knyazikhin, Y., Schull, M.A., Stenberg, P., Mottus, M., Rautiainen, M., Yang, Y., Marshak, A., Latorre Carmona, P., Kaufmann, R.K., Lewis, P., Disney, M.I., Vanderbilt, V., Davis, A.B., Baret, F., Jacquemoud, S., Lyapustin, A., Myneni, R.B., 2013. Hyperspectral remote sensing of foliar nitrogen content. Proc. Natl. Acad. Sci. 110, E185–E192. https://doi.org/10.1073/pnas.1210196109

Kobayashi, H., Ryu, Y., Baldocchi, D.D., Welles, J.M., Norman, J.M., 2013. On the correct estimation of gap fraction: How to remove scattered radiation in gap fraction measurements? Agric. For. Meteorol. 174–175, 170–183. https://doi.org/10.1016/j.agrformet.2013.02.013

Kokaly, R.F., Asner, G.P., Ollinger, S.V., Martin, M.E., Wessman, C.A., 2009. Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. Remote Sens. Environ. 113, S78–S91. https://doi.org/10.1016/j.rse.2008.10.018

Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification

models. J. Cheminformatics 6. https://doi.org/10.1186/1758-2946-6-10

Kuhn, M., 2008. Building Predictive Models in *R* Using the **caret** Package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4614-6849-3

Kumar, L., Skidmore, A.K., 2000. Radiation-Vegetation Relationships in a Eucalyptus Forest. Photogramm. Eng. 12.

Lee, K.-S., Cohen, W.B., Kennedy, R.E., Maiersperger, T.K., Gower, S.T., 2004. Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. Remote Sens. Environ. 91, 508–520. https://doi.org/10.1016/j.rse.2004.04.010

Legendre, P., 1993. Spatial Autocorrelation: Trouble or New Paradigm? Ecology 74, 1659–1673. https://doi.org/10.2307/1939924

Legendre, P., Dale, M.R.T., Fortin, M.-J., Casgrain, P., Gurevitch, J., 2004. EFFECTS OF SPATIAL STRUCTURES ON THE RESULTS OF FIELD EXPERIMENTS. Ecology 85, 3202–3214. https://doi.org/10.1890/03-0677

Legendre, P., Fortin, M.J., 1989. Spatial pattern and ecological analysis. Vegetatio 80, 107–138. https://doi.org/10.1007/BF00048036

Li, J., Li, C., Zhao, D., Gang, C., 2011a. Hyperspectral Narrowbands and Their Indices on Assessing Nitrogen Contents of Cotton Crop Applications, in: Hyperspectral Remote Sensing of Vegetation. CRC Press, pp. 579–590. https://doi.org/10.1201/b11222-34

Li, J., Li, C., Zhao, D., Gang, C., 2011b. Hyperspectral Narrowbands and Their Indices on Assessing Nitrogen Contents of Cotton Crop Applications, in: Hyperspectral Remote Sensing of Vegetation. CRC Press, pp. 579–590. https://doi.org/10.1201/b11222-34

Liang, S., 2005. Quantitative Remote Sensing of Land Surfaces. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/047172372X.ch8

Lindgren, F., 2012. Continuous Domain Spatial Models in. ISBA Bull. 19, 14–20.

Lindgren, F., Rue, H., 2015. Bayesian Spatial Modelling with *R* - **INLA**. J. Stat. Softw. 63. https://doi.org/10.18637/jss.v063.i19

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach: Link between Gaussian Fields and Gaussian Markov Random Fields. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x

Lobo, A., Moloney, K., Chic, O., Chiariello, N., 1998. Analysis of fine-scale spatial pattern of a grassland from remotely-sensed imagery and field collected data. Landsc. Ecol. 13, 111–131.

Lovett, G., Jones, C.G., Turner, M.G., Weathers, K.C., 2005. Ecosystem function in heterogeneous landscapes. Springer, New York.

Manolakis, D., Marden, D., Shaw, G.A., 2003. Hyperspectral Image Processing for Automatic Target Detection Applications 14, 38.

Martin, M.E., Plourde, L.C., Ollinger, S.V., Smith, M.-L., McNeil, B.E., 2008. A generalizable method for remote sensing of canopy nitrogen across a wide range of forest ecosystems. Remote Sens. Environ. 112, 3511–3519. https://doi.org/10.1016/j.rse.2008.04.008

Matérn, B., 1986. Spatial Variation, Lecture Notes in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4615-7892-5

McCoy, R.M., 2005. Field methods in remote sensing. Guilford Press, New York, NY.

Meehl, P.E., 1945. A Simple Algebraic Development of Horst's Suppressor Variables. Am. J. Psychol. 58, 550. https://doi.org/10.2307/1417770

Militino, A., Ugarte, M., Pérez-Goya, U., 2017. Stochastic Spatio-Temporal Models for Analysing NDVI Distribution of GIMMS NDVI3g Images. Remote Sens. 9, 76. https://doi.org/10.3390/rs9010076

Militino, A.F., Ugarte, M.D., Pérez-Goya, U., 2018. An Introduction to the Spatio-Temporal Analysis of Satellite Remote Sensing Data for Geostatisticians, in: Daya Sagar, B.S., Cheng, Q., Agterberg, F. (Eds.), Handbook of Mathematical Geosciences. Springer International Publishing, Cham, pp. 239–253. https://doi.org/10.1007/978-3-319-78999-6_13

Milton, E.J., Schaepman, M.E., Anderson, K., Kneubühler, M., Fox, N., 2009. Progress in field spectroscopy. Remote Sens. Environ. 113, S92–S109. https://doi.org/10.1016/j.rse.2007.08.001

Mirzaie, M., Darvishzadeh, R., Shakiba, A., Matkan, A.A., Atzberger, C., Skidmore, A., 2014. Comparative analysis of different uni- and multi-variate methods for estimation of vegetation water content using hyper-spectral measurements. Int. J. Appl. Earth Obs. Geoinformation 26, 1–11. https://doi.org/10.1016/j.jag.2013.04.004

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157, 209–225. https://doi.org/10.1016/S0304-3800(02)00197-7

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 66, 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Muñoz-Huerta, R., Guevara-Gonzalez, R., Contreras-Medina, L., Torres-Pacheco, I., Prado-Olivarez, J., Ocampo-Velazquez, R., 2013. A Review of Methods for Sensing the Nitrogen Status in Plants: Advantages, Disadvantages and Recent Advances. Sensors 13, 10823–10843. https://doi.org/10.3390/s130810823

Mutanga, O., Skidmore, A.K., 2007. Red edge shift and biochemical content in grass canopies. ISPRS J. Photogramm. Remote Sens. 62, 34–42. https://doi.org/10.1016/j.isprsjprs.2007.02.001

Naimi, B., Skidmore, A.K., Groen, T.A., Hamm, N.A.S., 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling: Spatial autocorrelation and positional uncertainty. J. Biogeogr. 38, 1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

Nguyen, H.T., Lee, B.-W., 2006. Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. Eur. J. Agron. 24, 349–356. https://doi.org/10.1016/j.eja.2006.01.001

Nolet, C., Poortinga, A., Roosjen, P., Bartholomeus, H., Ruessink, G., 2014. Measuring and Modeling the Effect of Surface Moisture on the Spectral Reflectance of Coastal Beach Sand. PLoS ONE 9, e112151. https://doi.org/10.1371/journal.pone.0112151

Olea, R.A., 2018. Advances in Sensitivity Analysis of Uncertainty to Changes in Sampling Density When Modeling Spatially Correlated Attributes, in: Daya Sagar, B.S., Cheng, Q., Agterberg, F. (Eds.), Handbook of Mathematical Geosciences: Fifty Years of IAMG. Springer International Publishing, Cham, pp. 375–393. https://doi.org/10.1007/978-3-319-78999-6_19

Ortenberg, F., 2011. Hyperspectral Sensor Characteristics: Airborne, Spaceborne, Hand-Held, and Truck-Mounted; Integration of Hyperspectral Data with LIDAR, in: Hyperspectral Remote Sensing of Vegetation. CRC Press, pp. 39–68. https://doi.org/10.1201/b11222-5

Pan, G., 2018. General Framework of Quantitative Target Selections, in: Daya Sagar, B.S., Cheng, Q., Agterberg, F. (Eds.), Handbook of Mathematical Geosciences. Springer International Publishing, Cham, pp. 411–435. https://doi.org/10.1007/978-3-319-78999-6_21

Patenaude, G., Milne, R., Van Oijen, M., Rowland, C.S., Hill, R.A., 2008. Integrating remote sensing datasets into ecological modelling: a Bayesian approach. Int. J. Remote Sens. 29, 1295–1315. https://doi.org/10.1080/01431160701736414

Pearse, G.D., Watt, M.S., Morgenroth, J., 2016. Comparison of optical LAI measurements under diffuse and clear skies after correcting for scattered radiation. Agric. For. Meteorol. 221, 61–70. https://doi.org/10.1016/j.agrformet.2016.02.001

Plant, R.E., 2012. Spatial data analysis in ecology and agriculture using R. CRC Press, Boca Raton.

Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., Chanussot, J., Fauvel, M., Gamba, P., Gualtieri, A., Marconcini, M., Tilton, J.C., Trianni, G., 2009. Recent advances in

techniques for hyperspectral image processing. Remote Sens. Environ. 113, S110–S122. https://doi.org/10.1016/j.rse.2007.07.028

Poggio, L., Gimona, A., Spezia, L., Brewer, M.J., 2016. Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. Geoderma 277, 69–82. https://doi.org/10.1016/j.geoderma.2016.04.026

Qi, J., Inoue, Y., Wiangwang, N., 2011. Hyperspectral Remote Sensing in Global Change Studies, in: Hyperspectral Remote Sensing of Vegetation. CRC Press, pp. 69–90. https://doi.org/10.1201/b11222-6

Ramoelo, A., Skidmore, A.K., Cho, M.A., Mathieu, R., Heitkönig, I.M.A., Dudeni-Tlhone, N., Schlerf, M., Prins, H.H.T., 2013. Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. ISPRS J. Photogramm. Remote Sens. 82, 27–40. https://doi.org/10.1016/j.isprsjprs.2013.04.012

Ramoelo, A., Skidmore, A.K., Cho, M.A., Schlerf, M., Mathieu, R., Heitkönig, I.M.A., 2012. Regional estimation of savanna grass nitrogen using the red-edge band of the spaceborne RapidEye sensor. Int. J. Appl. Earth Obs. Geoinformation 19, 151–162. https://doi.org/10.1016/j.jag.2012.05.009

Reichenau, T.G., Korres, W., Montzka, C., Fiener, P., Wilken, F., Stadler, A., Waldhoff, G., Schneider, K., 2016. Spatial Heterogeneity of Leaf Area Index (LAI) and Its Temporal Course on Arable Land: Combining Field Measurements, Remote Sensing and Simulation in a Comprehensive Data Analysis Approach (CDAA). PLOS ONE 11, e0158451. https://doi.org/10.1371/journal.pone.0158451

Ripley, B.D., 1981. Spatial Statistics: Ripley/Spatial Statistics, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. https://doi.org/10.1002/0471725218

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881

Rocha, A., Groen, T., Skidmore, A., Darvishzadeh, R., Willemen, L., 2018. Machine Learning Using Hyperspectral Data Inaccurately Predicts Plant Traits Under Spatial Dependency. Remote Sens. 10, 1263. https://doi.org/10.3390/rs10081263

Rocha, A.D., Groen, T.A., Skidmore, A.K., 2019. Space-spectra tuning improves plant trait prediction with hyperspectral data. Remote Sens. Environ.

Rocha, A.D., Groen, T.A., Skidmore, A.K., Darvishzadeh, R., Willemen, L., 2017. The Naïve Overfitting Index Selection (NOIS): A new method

to optimize model complexity for hyperspectral data. ISPRS J. Photogramm. Remote Sens. 133, 61–74. https://doi.org/10.1016/j.isprsjprs.2017.09.012

Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC, Boca Raton.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 319–392. https://doi.org/10.1111/j.1467-9868.2008.00700.x

Schaepman-Strub, G., Schaepman, M.E., Painter, T.H., Dangel, S., Martonchik, J.V., 2006. Reflectance quantities in optical remote sensing—definitions and case studies. Remote Sens. Environ. 103, 27–42. https://doi.org/10.1016/j.rse.2006.03.002

Schertzer, D., Lovejoy, S., 2004. Space–time complexity and multifractal predictability. Phys. Stat. Mech. Its Appl. 338, 173–186. https://doi.org/10.1016/j.physa.2004.04.032

Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., Schüler, G., 2010. Retrieval of chlorophyll and nitrogen in Norway spruce (Picea abies L. Karst.) using imaging spectroscopy. Int. J. Appl. Earth Obs. Geoinformation 12, 17–26. https://doi.org/10.1016/j.jag.2009.08.006

Secades, C., O'Connor, B., Brown, C., Walpole, M., UNEP World Conservation Monitoring Centre, Secretariat of the Convention on Biological Diversity, 2014. Earth observation for biodiversity monitoring: a review of current approaches and future opportunities for tracking progress towards the Aichi biodiversity targets.

Shaw, G.A., Burke, H.K., 2003. Spectral Imaging for Remote Sensing 14, 26.

Shen, G., Sakai, K., Kaji, K., 2013a. Capturing landscape changes and ecological processes in Nikko National Park (Japan) by integrated use of remote sensing images. Landsc. Ecol. Eng. 9, 89–98. https://doi.org/10.1007/s11355-011-0180-1

Shen, G., Sakai, K., Kaji, K., 2013b. Capturing landscape changes and ecological processes in Nikko National Park (Japan) by integrated use of remote sensing images. Landsc. Ecol. Eng. 9, 89–98. https://doi.org/10.1007/s11355-011-0180-1

Shmueli, G., 2010. To Explain or to Predict? Stat. Sci. 25, 289–310. https://doi.org/10.1214/10-STS330

Si, Y., Schlerf, M., Zurita-Milla, R., Skidmore, A., Wang, T., 2012. Mapping spatio-temporal variation of grassland quantity and quality using MERIS data and the PROSAIL model. Remote Sens. Environ. 121, 415–425. https://doi.org/10.1016/j.rse.2012.02.011

Simpson, D., Lindgren, F., Rue, H., 2012. Think continuous: Markovian Gaussian models in spatial statistics. Spat. Stat. 1, 16–29. https://doi.org/10.1016/j.spasta.2012.02.003

Skidmore, A.K., 1997. Performance of a Neural Network: Mapping Forests Using GIS and Remotely Sensed Data 14.

Skidmore, A.K., Pettorelli, N., Coops, N.C., Geller, G.N., Hansen, M., Lucas, R., Mücher, C.A., O'Connor, B., Paganini, M., Pereira, H.M., Schaepman, M.E., Turner, W., Wang, T., Wegmann, M., 2015. Environmental science: Agree on biodiversity metrics to track from space. Nature 523, 403–405. https://doi.org/10.1038/523403a

Stehman, S.V., 2000. Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment. Remote Sens. Environ. 72, 35–45. https://doi.org/10.1016/S0034-4257(99)00090-5

Stevens, D.L., Olsen, A.R., 2004. Spatially Balanced Sampling of Natural Resources. J. Am. Stat. Assoc. 99, 262–278. https://doi.org/10.1198/016214504000000250

Stroppiana, D., Fava, F., Boschetti, M., Brivio, P., 2011. Estimation of Nitrogen Content in Crops and Pastures Using Hyperspectral Vegetation Indices, in: Hyperspectral Remote Sensing of Vegetation. CRC Press, pp. 245–262. https://doi.org/10.1201/b11222-16

Thenkabail, P.S., Smith, R.B., De Pauw, E., 2000. Hyperspectral Vegetation Indices and Their Relationships with Agricultural Crop Characteristics. Remote Sens. Environ. 71, 158–182. https://doi.org/10.1016/S0034-4257(99)00067-X

Thiemann, S., Kaufmann, H., 2002. Lake water quality monitoring using hyperspectral airborne data—a semiempirical multisensor and multitemporal approach for the Mecklenburg Lake District, Germany. Remote Sens. Environ. 81, 228–237. https://doi.org/10.1016/S0034-4257(01)00345-5

Tian, Y., Woodcock, C., Wang, Y., Privette, J., Shabanov, N., Zhou, L., Zhang, Y., Buermann, W., Dong, J., Veikkanen, B., 2002. Multiscale analysis and validation of the MODIS LAI productII. Sampling strategy. Remote Sens. Environ. 83, 431–441. https://doi.org/10.1016/S0034-4257(02)00058-5

Tobler, W.R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Econ. Geogr. 46, 234. https://doi.org/10.2307/143141

Tsai, F., Philpot, W., n.d. Derivative Analysis of Hyperspectral Data 11.

Ullah, S., Groen, T.A., Schlerf, M., Skidmore, A.K., Nieuwenhuis, W., Vaiphasa, C., 2012. Using a Genetic Algorithm as an Optimal Band Selector in the Mid and Thermal Infrared (2.5–14 μm) to Discriminate Vegetation Species. Sensors 12, 8755–8769. https://doi.org/10.3390/s120708755

Vallejos, R., Osorio, F., 2014. Effective sample size of spatial process models. Spat. Stat. 9, 66–92. https://doi.org/10.1016/j.spasta.2014.03.003

Van Cleemput, E., Vanierschot, L., Fernández-Castilla, B., Honnay, O., Somers, B., 2018. The functional characterization of grass- and shrubland ecosystems using hyperspectral remote sensing: trends, accuracy and moderating variables. Remote Sens. Environ. 209, 747–763. https://doi.org/10.1016/j.rse.2018.02.030

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S 504.

Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J.G.P.W., Moreno, J., 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties – A review. ISPRS J. Photogramm. Remote Sens. 108, 273–290. https://doi.org/10.1016/j.isprsjprs.2015.05.005

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J.P., Camps-Valls, G., Moreno, J., 2012. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. Remote Sens. Environ. 118, 127–139. https://doi.org/10.1016/j.rse.2011.11.002

Vohland, M., Jarmer, T., 2008. Estimating structural and biochemical parameters for grassland from spectroradiometer data by radiative transfer modelling (PROSPECT+SAIL). Int. J. Remote Sens. 29, 191–209. https://doi.org/10.1080/01431160701268947

Wang, G., Gertner, G., 2013. Remote Sensing Applications for Sampling Design of Natural Resources, in: Remote Sensing of Natural Resources. CRC Press, pp. 23–44. https://doi.org/10.1201/b15159-5

Wang, G., Gertner, G., Anderson, A.B., 2005. Sampling design and uncertainty based on spatial variability of spectral variables for mapping vegetation cover. Int. J. Remote Sens. 26, 3255–3274. https://doi.org/10.1080/01431160500114748

Wang, G., Weng, Q. (Eds.), 2014. Remote sensing of natural resources, Remote sensing applications series. CRC Press, Boca Raton.

Wang, J., Brown, D.G., Bai, Y., 2014. Investigating the spectral and ecological characteristics of grassland communities across an ecological gradient of the Inner Mongolian grasslands with *in situ* hyperspectral data. Int. J. Remote Sens. 35, 7179–7198. https://doi.org/10.1080/01431161.2014.967885

Wang, J.-F., Stein, A., Gao, B.-B., Ge, Y., 2012. A review of spatial sampling. Spat. Stat. 2, 1–14. https://doi.org/10.1016/j.spasta.2012.08.001

Wang, X., Yue, Y., Faraway, J.J., 2018. Bayesian regression modeling with INLA. CRC Press, Boca Raton.

Webster, R., Curran, P.J., Munden, J.W., 1989. Spatial correlation in reflected radiation from the ground and its implications for sampling and mapping by ground-based radiometry. Remote Sens. Environ. 29, 67–78. https://doi.org/10.1016/0034-4257(89)90079-5

Wikle, 2003. Hierarchical Bayesian Models for Predicting the Spread of Ecological Processes. Ecology 84, 1382–1394.

Wikle, C.K., Hooten, M.B., 2010. A general science-based framework for dynamical spatio-temporal models. TEST 19, 417–451. https://doi.org/10.1007/s11749-010-0209-z

Wilson, A.M., Silander, J.A., Gelfand, A., Glenn, J.H., 2011. Scaling up: linking field data and remote sensing with a hierarchical model. Int. J. Geogr. Inf. Sci. 25, 509–521. https://doi.org/10.1080/13658816.2010.522779

Woodgate, W., Jones, S.D., Suarez, L., Hill, M.J., Armston, J.D., Wilkes, P., Soto-Berelov, M., Haywood, A., Mellor, A., 2015. Understanding the variability in ground-based methods for retrieving canopy openness, gap fraction, and leaf area index in diverse forest systems. Agric. For. Meteorol. 205, 83–95. https://doi.org/10.1016/j.agrformet.2015.02.012

Woodgate, W., Soto-Berelov, M., Suarez, L., Jones, S., Hill, M., Axelsson, C., Haywood, A., Mellor, A., 2012. Searching for the Optimal Sampling Design for Measuring LAI in an Upland Rainforest 11.

Yuan, H., Yang, G., Li, C., Wang, Y., Liu, J., Yu, H., Feng, H., Xu, B., Zhao, X., Yang, X., 2017. Retrieving Soybean Leaf Area Index from Unmanned Aerial Vehicle Hyperspectral Remote Sensing: Analysis of RF, ANN, and SVM Regression Models. Remote Sens. 9, 309. https://doi.org/10.3390/rs9040309

Zeng, Y., Li, J., Liu, Q., Qu, Y., Huete, A., Xu, B., Yin, G., Zhao, J., 2015. An Optimal Sampling Design for Observing and Validating Long-Term Leaf Area Index with Temporal Variations in Spatial Heterogeneities. Remote Sens. 7, 1300–1319. https://doi.org/10.3390/rs70201300

Zhang, L., Gove, J.H., Heath, L.S., 2005. Spatial residual analysis of six modeling techniques. Ecol. Model. 186, 154–177. https://doi.org/10.1016/j.ecolmodel.2005.01.007

Zhao, K., Valle, D., Popescu, S., Zhang, X., Mallick, B., 2013. Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection. Remote Sens. Environ. 132, 102–119. https://doi.org/10.1016/j.rse.2012.12.026

Summary

By monitoring biochemical and biophysical cycles in different ecosystems and relevant geolocations, vegetation dynamics can be better understood. An adequate selection of plant traits can perform a role as surrogate indicators for monitoring ecosystem dynamics. Less utopian yet important, the assessment of plant trait can guide to managing crops more efficiently and sustainable, improving food security by precision agriculture. However, it is a hard task to sample and measure plant trait, which reduces enormously the availability in the scale and time that is needed.

As direct plant trait measurements are quite a time-consuming, expensive and usually sample destructive, a need exists for other forms of indirect estimations to make an efficient monitoring system feasible. Up to now, the more acceptable alternative is measuring the vegetation surface by optical instruments. Using an optical sensor, it is possible to capture the reflection of the plant in the electromagnetic spectrum, and use it to estimate a biochemical and biophysical characterisation of the vegetation. Remote sensing has the potential to speed up the measurement (or estimation) of plant traits, allowing monitoring of vegetation dynamics over a wider range of spatial and temporal scales.

An optical instrument usually measures the reflectance of a target surface in a specific range of the spectrum, and divide it in different wavelengths. The number of wavelengths can vary from a couple of bands to thousands of them. Variations of chlorophyll leaf concentration will affect the reflectance in different regions of the spectra than water leaf content. Therefore, narrow wavebands, in general, quantify these plant traits better than wide bands. For this reason, hyperspectral remote sensing often presents more accurate estimation of plant traits than multi spectral measurements. Besides that such spectral specificity allows for higher accuracy, it also leads to problems as the data is highly dimensional with many redundant wavelengths.

Modelling a large number of serially correlated wavelengths with relatively few observations of ground references to support it, can bring serious issues. Problems such as multicollinearity and model overfitting are the most common. Overfitting leads to very specific models which lack in generalisation and are only accurate for the same dataset used to fit it. The risk of overfitting is magnified by noise from atmospheric effects and variations in certain regions of the spectra from sunlight illumination. The risk also increases using machine learning algorithms or supervised methods (with the support of the response) for model selection or tuning parameters. A new method to tuning complexity called Naïve Overfitting Index Selection (NOIS) was developed to reduce the risk of overfitting while modelling with machine learnings. The NOIS method

uses simulated data based on the covariance matrix of the original set to determine the maximum model complexity supported by the number of observations.

Remote sensing data captured from landscapes of vegetation is likely to present also spatial and temporal variations apart from the ones in the spectral domain. Often neglected while modelling because the dimensionality of the spectral domain, the spatiotemporal autocorrelation on the observations can cause serious inferential problems. Machine learning algorithms present unstable and unreliable predictions under significant autocorrelation, as was shown here by the model assessment using simulated landscapes with increscent level of spatial dependency. A spatial model is indicated in this case, but for modelling space explicit using spectral information as covariates, the number of wavelengths has to be reduced drastically. It can be achieved by spectra-space tuning process that balances the trade-off between accuracy and overfitting in both domains.

Temporal variations in sunlight illumination and atmospheric conditions have stochastic nature, and consequently, carry some degree of unpredictability. Patterns in soil fertility, slope or moisture drive the spatial dependency of plant traits, and can also be inherently stochastic. Whether modelling with hyperspectral data using the three domains explicitly or not, a certain level of uncertainties will always be present. Therefore, appropriate sampling designs and regression models are crucial but worthless without an in-depth assessment of the uncertainties coming from the three domains.

Samenvatting

Door biochemische en biofysische cycli te observeren in verschillende ecosystemen en op relevante plaatsen kan er beter begrip komen over de dynamiek van vegetatie. Een juiste keuze van planteneigenschappen kan de rol van surrogaat indicatoren vervullen voor het observeren van de dynamiek in een ecosysteem. Minder utopisch, maar even belangrijk kan het inschatten van planteneigenschappen het beheer van gewassen efficiënter en duurzamer maken in precisielandbouw, daardoor bijdragend aan voedselzekerheid. Het is echter moeilijk om planteneigenschappen te meten, en dat reduceert de mogelijkheden om waarnemingen te maken op de schaal en frequentie die gewenst zou zijn.

Omdat directe metingen van planteneigenschappen tijdrovend, duur en doorgaans desctructief zijn, zijn andere, indirecte meetmethodes nodig om een effectief observatiesysteem op te kunnen zetten. Tot nu toe is het meest acceptabele alternatief hiervoor om metingen aan vegetatie oppervlaktes te maken met optische instrumenten. Met deze instrumenten kan de reflectie van elektromagnetische straling door vegetatie gemeten worden, en gebruikt worden om een schatting te maken van biochemische en biofysische eigenschappen van die vegetatie. Deze indirecte waarnemingen (remote sensing genaamd), kunnen de meting (of schatting) van planteneigenschappen versnellen, waardoor observatie van vegetatie dynamiek over grotere gebieden en over langere periodes mogelijk wordt.

Optische instrumenten meten normaal gesproken de reflectie van een doeloppervlakte over een specifiek gedeelte van het elektromagnetisch spectrum en delen dit op in verschillende gebieden van golflengtes. Het aantal gemeten gebieden van golflengtes kan variëren van een klein aantal tot duizenden. Variaties in chlorophyl concentraties in bladeren hebben een effect op andere gedeeltes van dit spectrum dan bijvoorbeeld water concentraties. Daarom zijn smallere gebieden van het spectrum vaak beter voor het kwantificeren van planteneigenschappen dan metingen over hele brede gebieden van het spectrum. Om die reden bieden hyperspectrale metingen doorgaans betere schattingen van planteneigenschappen dan multispectrale metingen. Maar naast dat zulke specifieke gegevens accuratere schattingen mogelijk maken, kampen ze ook met bepaalde problemen omdat de gegevens uit veel metingen per observatie bestaan, waarvan veel golflengtes dezelfde informatie bevatten.

Modelleren met een groot aantal serieel gecorreleerde golflengtes, in combinatie met relatief weinig observaties van grond referenties ter ondersteuning, kan serieuze gevolgen hebben. Problemen zoals multicollineariteit en overparametrisatie zijn het meest voorkomend in deze

gevallen. Overparametrisatie leidt tot zeer specifieke modellen die niet goed veralgemeniseerd kunnen worden, en eigenlijk alleen maar accuraat zijn in het voorspellen van de dataset waarop ze gebaseerd is. Het risico op overparametrisatie neemt toe door ruis als gevolg van atmosferische condities, en variaties in bepaalde gedeelten van het spectrum door belichting van de zon. Dit risico neemt ook toe wanneer zelflerende algoritmes of gesuperviseerde methodes (waar de verklaarde variabele wordt gebruikt als basis voor model kalibratie) worden gebruikt om een selectie van modellen te maken of om modelcomplexiteit af te stemmen. Een nieuwe methode om deze complexiteit af te stemmen, de zogeheten Naïeve Overparametriserings Index Selectie (NOIS) is in dit proefschrift ontwikkeld om het risico op overparametrisering te reduceren wanneer zelflerende algoritmes worden gemaakt. De NOIS methode gebruikt gesimuleerde gegevens die zijn afgeleid van een covariantiematrix van de originele gegevens, om de maximale model complexiteit die door het aantal waarnemingen wordt ondersteund vast te stellen.

Aardobservatie gegevens van vegetatie landschappen zullen waarschijnlijk ook ruimtelijke en temporele variaties vertonen, naast de variaties in het electromagnetische spectrum. Hoewel deze dimensies vaak worden genegeerd bij het modelleren in verband met de grote hoeveelheid gegevens bij hyperspectrale studies, kan autocorrelatie in deze dimensies ook voor serieuze problemen zorgen. Zelflerende algoritmes genereren onstabiele en onbetrouwbare voorspellingen wanneer er sprake is van autocorrelatie, zoals we in deze thesis hebben laten zien door een model evaluatie gebaseerd op gesimuleerde landschappen met kunstmatig toenemende gradaties van ruimtelijke autocorrelatie. Een ruimtelijk model wordt in zulke gevallen aangeraden. Maar om ruimte expliciet te modelleren met spectrale informatie als covariabele vereist dat het aantal golflengtes dat gebruikt wordt aanzienlijk wordt gereduceerd. Dit kan bereikt worden door een afstemming van ruimte en tijd in een proces dat een afweging tussen accuraatheid en overparametrisering in beide domeinen maakt.

Variaties over tijd in belichting door de zon en atmosferische condities zijn van nature willekeurig en bijgevolg tot op zeker hoogte niet te voorspellen. Patronen in bodem vruchtbaarheid, helling of bodemvocht concentraties zijn drijvende krachten achter de ruimtelijke afhankelijkheid en kunnen ook op toevalligheden berusten. Of er nu wel of niet rekening wordt gehouden met deze drie domeinen wanneer er met hyperspectrale gegevens wordt gemodelleerd, een bepaalde mate van onzekerheid zal altijd blijven. Daarom zijn correcte bemonsterings strategien en en regressiemodellen essentieel, maar deze blijven waardeloos wanneer er geen goede inschatting wordt gemaakt van de onzekerheid die voortvloeit uit deze drie domeinen.