SCALE CHALLENGES IN INVENTORY OF FORESTS AIDED BY REMOTE SENSING

Dissertation to attain the degree Doctor of Philosophy (Ph. D.) of the
Faculty of Forest Sciences and Forest Ecology
Georg-August-Universität Göttingen

Submitted by:

Collins B. Kukunda

Born on 18.04.1986 in Mbarara, Uganda

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Göttingen, July 2020

1st Referee:
Prof. Dr. Christoph Kleinn
Forest Inventory and Remote Sensing
Faculty of Forest Sciences and Forest Ecology
University of Göttingen, Göttingen, Germany

2nd Referee:
Prof. Dr. Steen Magnussen
Forest Invetory and Analysis
Pacific Forestry Center
Victoria, British Columbia, Canada

3rd Referee:
Prof. Dr. Thomas Kneib
Chair of Statistics and Econometrics
University of Göttingen, Göttingen, Germany

Date of defense: May 13, 2019

In memory of dad, Eng. Azaria Byobona (*1954 – †2014), for nurturing my love for forests.

"Space and time are the framework within which the mind is constrained to construct its experience of reality."
—Immanuel Kant, *Critique of pure reason*, 1855.

Abstract

The impact of changing the scale of observation on information derived from forest inventories is the basis of scale-related research in forest inventory and analysis (FIA). Interactions between the scale of observation and observed heterogeneity in studied variables highlight a dependence on scale that affects measurements, estimates, and relationships between inventory data from terrestrial and remote sensing surveys. This doctoral research defines *"scale"* as the divisions of continuous space over which measurements are made, or hierarchies of discrete units of study/analysis in space. Therefore, the *"scale of observation"* (also known as support) refers to that integral of space over which statistics are computed and forest inventory variables regionalized.

Given the ubiquitous nature of scale issues, a case study approach was undertaken in this research (Articles I-IV) with the goal to provide fundamental understanding of responses to the scale of observation for specific FIA variables. The studied forest inventory variables are; forest stand structural heterogeneity, forest cover proportion and tree species identities. Forest cover proportion (or simply forest area) and tree species are traditional and fundamental forest inventory variables commonly assessed over large areas using both terrestrial samples and remote sensing data whereas, forest stand structural heterogeneity is a contemporary FIA variable that is increasingly demanded in multi-resource inventories to inform management and conservation efforts as it is linked to biodiversity, productivity, ecosystem functioning and productivity, and used as auxiliary data in forest inventory.

This research has two overall aims:

1. To improve the understanding of the association between the scale of observation and observed heterogeneity in inventory of forest stand structural heterogeneity, forest-cover proportions, and identification of tree species from a combination of terrestrial samples and remote sensing data.

2. To contribute knowledge to the estimation of scale-dependence in inventory of forest stand structural heterogeneity, forest-cover proportions, and identification of tree species from a combination of terrestrial samples and remote sensing data.

Different scales of observation were considered across the four case studies encompassing individual leaf, crown-part or branch, single-tree crown, forest stand, landscape and global levels of analysis. Terrestrial and remote sensing data sets from a variety of temperate forests in Germany and France were utilized across case studies. In cases where no inventory data were

available, synthetic data was simulated at different scales of observation. Heterogeneity in FIA variable estimates was monitored across scales of observation using estimators of variance and associated precision. As too much heterogeneity is hardly interpreted due to a low signal to noise ratio, object-based image analysis (OBIA) methods were used to manage heterogeneity in high-resolution remote sensing data before evaluating scale dependence or scaling across observed scales. Similarly, ensemble classification techniques were applied to address methodological heterogeneity across classifiers in a case study on classification of two physically and spectrally similar *Pinus* species. Across case studies, a dependence on the scale of observation was determined by linking estimates of heterogeneity to their respective scales of observation using linear regression and a combination of geo-statistics and Monte-Carlo approaches. In order to address scale-dependence, thresholds to scale domains were identified so as to enable efficient observation of studied FIA variables and scaling approaches proposed to bridge observations across scales. For scaling, this research evaluated the potential of different regression techniques to map forest stand structural heterogeneity and tree species wall-to-wall from remote sensing data. In addition, radiative transfer modelling was evaluated in the transfer between leaf and crown hyperspectra, and a global sampling grid framework proposed to efficiently link different stages of survey sampling.

This research shows that the scale of observation affected all studied FIA variables albeit to varying degrees, conditioned on the spatial structure and aggregation properties of the assessed FIA variable (i.e. whether the variable is extensive, intensive or scale-specific) and the method used in aggregation on support (e.g. mean, variance, quantile etc.). The scale of observation affected measurements or estimates of the studied FIA variables as well as relationships between spatially structured FIA variables. The scale of observation determined observed heterogeneity in FIA variables, affected parameter retrieval from radiative transfer models, and affected variable selection and performance of models linking terrestrial and remote sensing data. On the other hand, this research shows that it is possible to determine domains of scale dependence within which to efficiently observe the studied FIA variables and to bridge between scales of observation using various scaling methods.

The findings of this doctoral research are relevant for the general understanding of scale issues in FIA. Research in Article I, for example, informs optimization of plot sizes for efficient inventory and mapping of forest structural heterogeneity, as well as for the design of natural resource inventories. Similarly, research in Article II is applicable in large area forest (or general land) cover monitoring from sampling by both visual interpretation of high resolution remote sensing imagery and terrestrial surveys. This research is also useful to determine observation design for efficient inventory of land cover. Research in Article III contributes in many contexts of remote sensing assisted inventory of forests especially in management and conservation planning, pest and diseases control and in the estimation of biomass. Lastly, research in Article IV highlights scale-related effects in passive optical remote sensing of forests currently understudied and can ultimately contribute to sensor calibration and modelling approaches.

Zusammenfassung

Der Einfluss von unterschiedlichen Beobachtungsskalen auf Informationen aus Stichprobeinventuren ist der Ausgangspunkt für skalen-bezogene Forschung in der Waldinventur.

Wechselwirkungen zwischen den Beobachtungsskalen und der beobachteten Heterogenität der untersuchten Variablen zeigen eine Skalenabhängigkeit, welche Messungen, Schätzungen sowie den Zusammenhang von terrestrischen Inventurdaten und Fernerkundungserhebungen beeinflussen.

Die vorliegende Arbeit definiert „Skala" als die Unterteilung des kontinuierlichen Raums, oder Hierarchien von diskreten Einheiten des Studiengebiets in dem Messungen oder Analysen im Raum durchgeführt werden. Deshalb bezieht sich die „Beobachtungsskala" (auch als „support" bezeichnet) auf das Integral des Raumes, über das Statistiken berechnet und Waldinventurvariablen regionalisiert werden.

Wegen der Omnipräsenz der Skalenproblematik, wurde in der vorliegenden Arbeit ein Fallstudienansatz (Fachzeitschriftenbeitrag I-IV) verwendet, Das Ziel der Arbeit ist es das grundlegende Verständnis des Einflusses der Beobachtungsskala auf spezifische Waldinventurvariablen und letztendlich das generelle Verständnis von Skalenproblematiken in der Waldinventur zu verbessern. Die betrachten Zielgrößen sind die Heterogenität der Bestandesstruktur, die Waldfläche, sowie die Baumarten bestimmung. Waldbedeckungsgrad (Waldfläche) und Baumarten bestimmung sind grundlegende Waldinventurvariable, die üblicherweise über große Gebiete durch terrestrische Stichprobenpunkte oder Fernerkundungsdaten aufgenommen werden. Die Heterogenität der Bestandesstruktur dagegen ist eine neuere Waldinventurvariable, die vermehrt in Inventuren gefragt ist um Bewirtschaftungs- und Umweltschutzmaßnahmen zu unterstützen.

Die vorliegende Arbeit hat zwei übergreifende Ziele:

1. die Erforschung der Beziehung zwischen Beobachtungsskala und beobachteter Heterogenität der Bestandesstruktur, des Deckungsgrades und der Baumartenbestimmung durch eine Kombination aus terrestrischen Stichprobenpunkten und Fernerkundungsdaten und

2. die Einführung von Methoden zur Schätzung von Skalenabhängigkeit bei der Erfassung der Strukturheterogenität, des Deckungsgradess und der Baumartenbestimmung durch eine Kombination aus terrestrischen Stichpunkten und Fernerkundungsdaten.

In den vier Fallstudien wurden unterschiedliche Beobachtungsskalen berücksichtigt. Sie umfassen Analyseebenen von einzelnen Blättern, Kronenteilen oder Ästen, einzelnen Baumkronen, Waldbeständen, sowie die Landschafts- und Globalebene. In den Fallstudien wurden terrestrische

Inventurdaten und Fernerkundungsdaten verschiedener Waldgebiete der gemäßigten Zone in Deutschland und Frankreich verwendet. In den Fällen in denen keine Inventurdaten vorhanden waren, wurden die Daten für verschiedene Beobachtungsebenen simuliert. Die Heterogenität der Schätzungen der Waldinventurvariablen wurde auf verschiedenen Beobachtungsskalen durch Varianzschätzer und entsprechende Genauigkeitsmaße teingeschätzt. Da zu hohe Heterogenität aufgrund eines geringen Signal-zu-Rausch-Verhältnisses kaum interpretierbar ist, wurden vor der Evaluierung der Skalenabhängigkeit oder der Skalierung zwischen Beobachtungsskalen, objekt-basierte Bildanalysemethoden verwendet um die Heterogenität in hochaufgelösten Fernerkundungsdaten zu kontrollieren. Ensemble-Classification-Methoden wurden in einer Fallstudie zur Klassifikation von zwei physikalisch und spektral ähnlichen Kiefernarten angewendet, um die methodische Heterogenität zwischen Klassifizierungsverfahren zu verringern. In allen Fallstudien wurde die Skalenabhängigkeit durch den Zusammenhang der Heterogenitätsschätzwerte und ihrer jeweiligen Beobachtungsskala bestimmt. Hierfür wurden lineare Regressionen und eine Kombination von Geostatistischen und Monte-Carlo-Verfahren angewendet. Zur Erforschung der Skalenabhängigkeit wurden Grenzwerte der Skalenbereiche identifiziert, um eine effiziente Beobachtung der untersuchten Waldinventurvariablen und Skalen-Verfahren zu ermöglichen. Diese Forschungsarbeit evaluiert das Potential von verschiedenen Regressionsmethoden zur Kartierung der Heterogenitä der Bestandesstruktur und der Baumarten mithilfe von Fernerkundungsdaten. Zudem wurde die Strahlungstransfermodellierung evaluiert um zwischen der hyperspektralen Rückstrahlung des Blatts und der Krone zu skalieren und ein globales Stichproben-Bezugssystem entwickelt um effizient verschiedene Phasen der Stichprobenerhebung zu verbinden.

Diese Forschungsarbeit zeigt, dass die Beobachtungsskala alle untersuchten Waldinventurvariablen beeinflusst, wenngleich in unterschiedlichem Ausmaß. Dieses hängt von den Eigenschaften der räumlichen Struktur, der Aggregation der geschätzten Waldinventurvariable (z.B. extensive, intensive oder skalen-spezifische Variable) und der unterstützenden Aggregationsmethode (z.B. Mittelwert, Varianz, Quantil, etc.) ab. Die Beobachtungsskala beeinflusste Messungen oder Schätzungen der untersuchten Waldinventurvariablen ebenso wie das Verhältnis zwischen räumlich strukturierten Waldinventurvariablen. Die Beobachtungsskala bestimmte die beobachtete Heterogenität der Waldinventurvariablen, beeinflusste die Parameter der Strahlungstransfermodelle, sowie die Variablenauswahl und die Modellanpassung von terrestrischen Inventurdaten und Fernerkundungserhebungen. Nichtsdestotrotz zeigt diese Forschungsarbeit, dass es möglich ist Bereiche der Skalenabhängigkeit zu bestimmen in denen die erforschten Waldinventurvariablen effizient beobachtet und Beobachtungsskalen durch die Verwendung verschiedener Methoden verbunden werden können.

Die Ergebnisse dieser Doktorarbeit sind hinsichtlich des generellen Verständnisses von Skalenproblematiken in der Waldinventur relevant. Der Fachzeitschriftenbeitrag I ist für die Optimierung der Stichprobenflächengröße für eine effiziente Inventur und Kartierung der Heterogenität der Waldstruktur, sowie für das Design von Inventuren natürlicher Ressourcen relevant. Die Forschung aus Fachzeitschriftenbeitrag II ist bei großflächigem Monitoring der Waldbedeckung bzw. generell der Bodenbedeckung, durch Stichproben aus visuell interpretierten

hochaufgelösten Fernerkundungsbildern als auch terrestrischen Aufnahmen, anwendbar. Der Ansatz kann auch zur Optimierung des Beobachtungsdesigns effizienter Inventuren von Landnutzungsklassen angewendet werden. Fachzeitschriftenbeitrag III ist für viele Bereiche der fernerkundungsunterstützten Waldinventur, insbesondere für Bewirtschaftungs- und Naturschutzplanung, Schädlings- und Krankheitsbekämpfung sowie die Schätzung von Biomasse relevant. Die Ergebnisse des Fachzeitschriftenbeitrags IV heben die skalenbezogenen Effekte der passiven optischen Fernerkundung von Wäldern hervor, welche bis jetzt wissenschaftlich vernachlässigt wurden und können letztendlich zur Sensor-Kalibrierung und zu Modellierungsansätzen beitragen.

Acknowledgment

List of Articles

I. **Collins B. Kukunda**, Philip Beckschäfer, Paul Magdon, Peter Schall, Christian Wirth, Christoph Kleinn
   Scale-guided mapping of forest stand structural heterogeneity from airborne Li-DAR
   Ecological indicators 2019; 102:410–425.

II. Fehrmann Lutz, **Collins B. Kukunda**, Nils Nölke, Sebastian Schnell, Dominik Seidel, Steen Magnussen, Christoph Kleinn
   A unified framework for land cover monitoring based on a discrete global sampling grid (GSG)
   Environmental Monitoring and Assessment 2019; 191(2) 46.

III. **Collins B. Kukunda**, Joaquín Duque-Lazo, Eduardo González-Ferreiro, Hauke Thaden, Christoph Kleinn
   Ensemble classification of individual *Pinus* crowns from multispectral satellite imagery and airborne LiDAR
   International Journal of Applied Earth Observation and Geoinformation 2018; 65:12–23.

IV. **Collins B. Kukunda**, Henning Aberle, Peter P.J. Roosjen, Christoph Kleinn, Steen Magnussen, Hans Fuchs
   Scale effects in broadleaved tree reflectance at leaf and crown levels
   Manuscript submitted to Journal of Agricultural and Forest Meteorology.

The articles are referred to in the body of the thesis numerically as listed above. Research in Articles I, III and IV was based on original ideas of Collins B. Kukunda, whereas research in Article II was developed jointly with co-authors. Collins B. Kukunda was responsible for all analyses and manuscript preparation in Articles I, III and IV. The thesis author conducted the case study on observation design presented in Article II and wrote the manuscript together with co-authors. All co-authors contributed at various stages of the research, especially during review of manuscripts. Revisions from co-authors improved the quality of the final articles. Articles I-III are published and are reproduced in the annex with kind permissions from the respective publishers.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scale in forest inventory and analysis

The core of forest inventory is to enumerate a population of trees and associated variables over a specified area (Scott and Gove, 2002). A census is almost always impractical and thus forest inventory requires a *"measuring tool"*, or *"scale"* (Marceau and Hay, 1999), through which trees and associated variables are observed, enumerated, and inference made on the population characteristics. Aggregation (spatial or otherwise) is most commonly necessary to create meaningful units for mensuration and analysis (Gotway and Young, 2002). The units (or support) represent that integral of space over which statistics are computed and forest inventory variables regionalized (Malenovský et al., 2007). They are essentially the measurement/sample scale, or measurement/sample unit (e.g. intervals, areas, volumes etc.) for data (Wu and Li, 2009) collected in forest inventory and analysis (FIA). The units (or support) are what I refer to here as the *"scale of observation"*.

The general meaning of scale however, is not constrained to the scale of observation but includes other notions. There is a scale of operation/action referring to the level at which an observed process operates, is supposed to operate, or is best observed (Marceau and Hay, 1999; Malenovský et al., 2007); a cartographic/map scale referring to the ratio of the map distance to the corresponding distance on the ground; a modelling scale describing the scale of model building (i.e. the support of model inputs) vis-à-vis model derivation/application; a geographic scale representing the coverage of pattern or analysis; and a policy scale in reference to the levels of decision making or policy implementation (Wu and Li, 2009; Marceau and Hay, 1999; Malenovský et al., 2007). Various other notions of scale may exist since the term is widely used, however, of the so far listed, the scale of observation is the most apparent in FIA assisted by remote sensing.

The scale of observation plays a very important role in FIA, especially in definition of a forest (Kleinn, 2001; Magdon et al., 2014), in estimation of forest area (Magdon and Kleinn, 2013; Fehrmann et al., 2019), in estimation of forest edge length (Kleinn et al., 2011), in selection of samples (Czaplewski, 2003; Fehrmann et al., 2019), in measurement of trees, in building of models to link field to remote sensing data (Kukunda et al., 2019, for example), as

well as, in inference on tree populations (Magnussen et al., 2016; McRoberts et al., 2016; Puliti et al., 2018; Mauro et al., 2017). The integration of remote sensing into contemporary forest inventories extends the role of the scale of observation in FIA to include, spectral, temporal, directional, polarization, and radiometric dimensions to scale (Wu and Li, 2009) that together with terrestrial observation design, influence the precision of FIA variables. In fact, understanding the impact of changing the scale of observation on information derived from forest inventories, forms the basis of scale-related research in FIA.

"Scale" primarily relates to the divisions of continuous space over which measurements are made, or to the hierarchies of discrete units of study/analysis in space (Wiens, 1989; Levin, 1992; Marceau and Hay, 1999; Schneider, 2001; Malenovský et al., 2007; Gunton et al., 2014; Sandel, 2015). This basic definition of scale (Figure 1.1) suffices in all notions of scale (i.e. scale of observation, action/operation, modelling, policy, geographic, and cartographic), albeit re-projected into their respective spaces of conception or measurement. The core concept of scale is illustrated in Figure 1.1 according to Gunton et al. (2014) in the spatial context. Fundamentally, spatial scale is characterized by the area of support (commonly referred to in FIA in terms of resolution, pixel size, plot size), the extent, and the hierarchies of discrete units of study/analysis (Figure 1.1). As previously mentioned, the scale of observation in contemporary FIA assisted by remote sensing includes spatial, spectral, temporal, directional, polarization and radiometric dimensions. This thesis work focuses on the "spatial scale of observation" and will refer to it simply as "scale of observation" in the succeeding text. Table 1.1 illustrates how the core meaning of the scale of observation is similar across different dimensions.
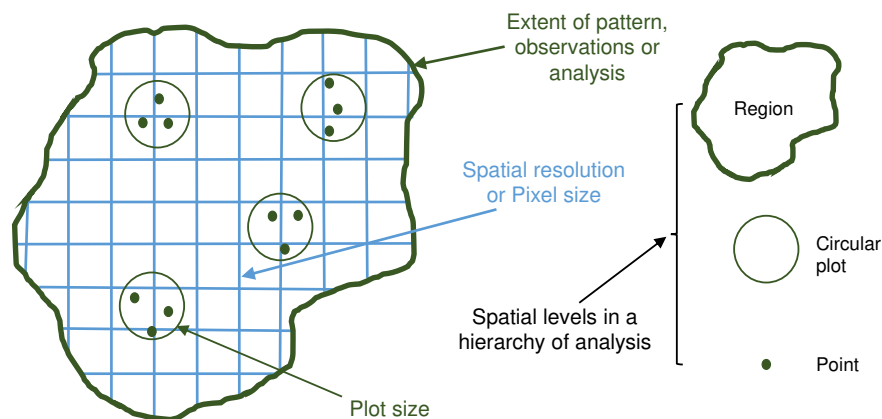
Figure 1.1: The basic concept of scale. The resolution/pixel, plot size and the extent are illustrated for a simple case of contiguous coverage. Scale as a "level" is illustrated with discrete, spatially-nested units that can also be looked at as "support". A simple illustration of sampling across space with clusters of points and fixed area circular plots is integrated. This schematic is adapted after Gunton et al. (2014).

For the reader unfamiliar with all the dimensions to scale of observation in remote sensing (Table 1.1), the spectral dimension refers to the volume of support over which spectrometric measurement is made. In a volume, the support considers more than two dimensional space

Table 1.1: Definitions of resolution and extent in the spatial, spectral, temporal, directional, and radiometric dimensions of the scale of observation as encountered within the context of remote sensing assisted inventory of forests.

Resolution	
Spatial	Smallest observable unit in space or the pixel of a remotely sensed image or plot size in a sample-based study (Figure 1.1)[a]
Spectral	Smallest unit of distinction among spectral differences in a spectral sample (Figure 1.2).
Temporal	Shortest time required to combine reflected energy into an image on the CCD array element.[b]
Directional	Smallest angle of distinction among angular spectral reflectances.[c]
Radiometric	Precision or sensor sensitivity to magnitudes of energy.
Extent	
Spatial	Total area encompassed by observations or analysis (Figure 1.1).
Spectral	Range of wavelengths included in the spectral sample (Figure 1.2).
Temporal	Time between the first and last observation at a given location.
Directional	Range of viewing directions.
Radiometric	Range of values stored in bits.

[a]By pixel re-sampling, the resolution of data representation can be different from the resolution of measurement.
[b]Temporal resolution was redefined by Malenovský et al. (2007) in order to make definitions across dimensions of the scale of observation coherent. Temporal resolution is commonly defined as the sensor revisit rate at a given geographic location - see for example Khorram et al. (2012). However, the sensor revisit rate at a given geographic location is in the strictest sense the temporal sampling interval.
[c]The directional resolution is determined by the instantaneous field of view (IFOV), the size of the CCD array, tilt, motion speed, and altitude of the sensor (Malenovský et al., 2007).

described in Figure 1.1 to include spectra (i.e. wavelength and corresponding reflectances or transmitances) in defining space over which a spectral measurement is made (Malenovský et al., 2007, see also Figure 1.2). Similarly, the temporal dimension deals with a volume of support considering a combination of space, spectra, and time over which measurement is made. Directional scale, on the other hand, deals with the strong directional behaviour of reflectances reaching the sensor element as a result of the multi-angular reflectivity of the earth's surface and scattering by atmospheric particles and gases (Roosjen et al., 2018). It relies on bi-directional reflectance distribution functions determined by sun-object-sensor geometry (Wu and Li, 2009; Malenovský et al., 2007) and optical properties of the observed object (Malenovský et al., 2007). Therefore, it combines space, spectra, time, and the solid angle of the sensor in defining the support. Similar to the directional dimension for passive optical sensors, the polarization dimension to support considers the solid angle or polarization direction of SAR data on top of space, signal and time. And lastly, the radiometric dimension is analogous to the scale of a diameter tape or caliper used in DBH measurements, and considers the translation of received energy at the sensor element into image colours and therefore deals directly with accuracy in the units of measurement. Even though the dimensions to the scale of observation in remote sensing
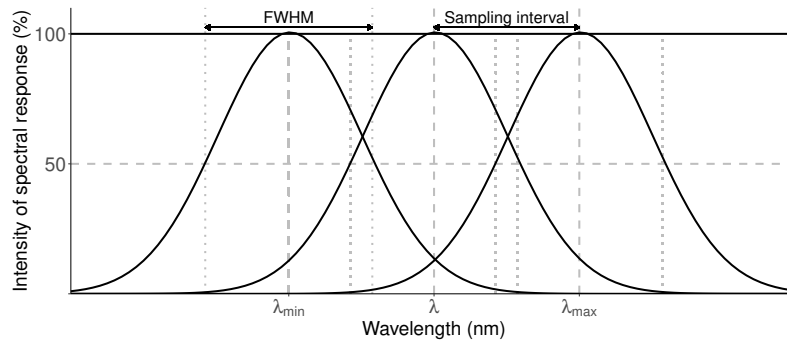
Figure 1.2: The basic concept of spectral scale. The spectral resolution is equivalent to the Full-Width-Half-Minimum (FWHM). The extent $= \lambda_{\max} - \lambda_{\min}$, the difference between the maximum and minimum wavelength. Notice that the spectral sampling interval is independent of the spectral resolution (FWHM). There is possible overlap between response functions of consecutive wavelength bands among consecutive spectral sampling intervals. The schematic is adapted after Malenovský et al. (2007).

are often described independently, it is clear that the scale dimensions are mutually inclusive as sensor systems are integrated and therefore the volume of support for any observation scale simultaneously relies on components from other dimensions.

On the other hand, the term scale is also used to mean a level in a hierarchy of analysis (Gunton et al., 2014, Figure 1.1). This scale definition holds for FIA supported by remote sensing as well as traditional terrestrial sampling schemes with respect to the observation level vis-à-vis the level of inference. While observations are made at point, line, or plot levels, inference is made for regions larger than the observed areas. Multi-phase and multi-stage sampling schemes (Köhl et al., 2006, Chapter 3), for example, and more recently hybrid (McRoberts et al., 2016) and hierarchical estimation approaches (Puliti et al., 2018), underpin the application of hierarchical scales of analysis in forest inventories.

1.2 Heterogeneity

The scale of observation and heterogeneity interact very closely that it is impossible to refer one without reference to the other. *"Heterogeneity"* primarily describes variability, complexity, or diversity in structure, composition, and functioning[1] of any system in space-time (Stein and Kreft, 2015). Heterogeneity is inherent in forest ecosystems and changes in properties relative to the scale of observation and the variable (or aggregation properties of the variable) in question. Therefore, one needs scale to define heterogeneity (Levin, 1992), and properties of the observed variable must be known a priori to achieve reasonable aggregation (mean, variance, etc.) on support.

There are generally three categories of variables with respect to aggregation properties; (1) variables with extensive properties - also termed *"extensive variables"* (e.g. number of

---

[1] *"structure"* is the spatial-temporal arrangement of components of the ecosystem, *"composition"*, the spatial-temporal identity and variety of ecosystem components, and *"function"*, the underlying stochastic or deterministic processes controlling realizations or states of populations of natural systems in space-time (Valbuena, 2015). FIA seeks to enumerate *"structure"* and *"composition"* whereas Ecology focuses on understanding *"function"*.

trees), whose quantities can be proportionally/simply summed across extents, (2) variables with intensive properties - also called *"intensive variables"* (e.g. tree densities), whose quantities are meaningless when simply summed across extents because they are averages or ratios, and (3) variables that may vary with extent but not proportional to it e.g. species richness - such variables are also termed *"scale-specific variables"*. The distinction of variables by their properties is essential in characterization of heterogeneity (Gunton et al., 2014) and its dependence on the scale of observation. In fact, it is not correct to ask, for example, whether a particular measurement is dependent on the scale of observation without also specifying how we intend to aggregate it on support (Sandel, 2015). In homogeneous systems, intensive variables remain constant across space or time (they are *"scale invariant"*), while extensive variables change linearly across space or time. For the same examples in heterogeneous systems, the opposite is observed. However, extensive variables in a heterogeneous system may exhibit domains of homogeneity related to a particular scale or scales of observation - when the spatial extent is large enough relative to the resolution of heterogeneity. In fact, finding this threshold between heterogeneity and homogeneity or simply put 'managing heterogeneity' is a key goal of data preparation in remote sensing assisted FIA; since too much variability can hardly be interpreted. For this reason, research on scale-dependence has been developed. Scale-dependence research seeks to address that trade off between detail (or heterogeneity within a group) and system predictability (Levin, 1992). Similarly, in the context of sampling studies in forest inventory, the balance between cost, between- and within-plot heterogeneity, and precision of estimation is the basis of most decisions on optimal observation designs.

1.3 Scale dependence

The recognition of interactions between the scale of observation and heterogeneity highlights a dependence on scale that, when ignored, may affect both measurements and relationships between variables in FIA. Scale dependence primarily stems from the modifiable areal unit problem MAUP (Openshaw, 1977). The integral of support, especially the way borders are drawn - i.e. how large the units are?, what shape?, where?, etc. - strongly affects the patterns observed for specific variables and the general outcome of data analysis. With the MAUP, there is variation in characteristics of observations due to changing the scale of observation - also called the *"scale effect*[2]*"* or sometimes also referred to as the *"scale problem"*, and there is variation in characteristics of observations due to alternative aggregations of areas of support at the same scale of observation - also called the *"aggregation/zoning problem"* (Marceau and Hay, 1999). Similarly, apart from the category of the variable observed and the form of aggregation on support some other factors could result in a scale dependence. For example, scale effects could result from limitations of measurement, or from differences in instrumentation applied in measurement. Scale effects could also result from errors in modelling - especially in relation to choice of an appropriate model of the underlying process,

---

[2] *"Scale effect"* refers to the relative contrast/difference in information and the respective characteristics of observations or models made on different support (Wu and Li, 2009)

or simply from the inherent heterogeneity and non-linearity in natural systems given the scale of observation. Stemming from scale effects, properties of systems become *"scale-dependent"*. However, conditioned on the variable observed and the means of aggregation, a sensitivity analysis across scales may yield patterns of scale dependence. Systems may become less scale-dependent or even scale-independent/scale-invariant in particular scale regions. *"Scale domains"* and *"scale thresholds"* can be identified and *"scaling"* within the domain made relatively easier (due to homogeneity/stability in the underlying process) - see as an example research in Article I. In this sense, the domain is the appropriate scale (or series of scales) of observation of a given process and the threshold(s) is that tipping point of the scale of observation (Wu and Li, 2009). Most research on scale-dependence ultimately aims at identifying scale domains and thresholds.

Several methods are proposed in the literature to quantitatively describe scale thresholds and domains. Among others Wu and Li (2009) describe the following; the geographic variance method (Moellering and Tobler, 1972), the wavelet transform method (Percival, 1995), the local variance method (Woodcock and Strahler, 1987), the semi-variogram based method (Garrigues et al., 2006), and the fractals method (Mandelbrot, 1967). These methods determine the threshold scale of observation based on progressive aggregation of data in space, hinged on a given optimization criterion/criteria that generally rely on local co-variation metrics. For the same purpose, in the last two-three decades, object-based image analysis (OBIA) has gained traction in remote sensing assisted FIA (Blaschke, 2010). In the same way, the approach emphasizes working within scale domains through iterative aggregation of high resolution pixel data into individual objects to identify specific entities in space. Altogether, methods describing scale thresholds and domains share a common goal i.e. to manage heterogeneity in the studied systems so as to increase predictability, interpretability, and scaling feasibility.

In forest ecosystems, the magnitude of the scale effect is expected as a function of spatial autocorrelation as site factors and growth conditions are highly correlated for trees close together than further apart. While comparing continuous surfaces of differing spatial structure, Sandel (2015) shows that indeed the magnitude of spatial auto-correlation determines the magnitude of effects from changing the scale of observation on estimated heterogeneity in sample data. Heterogeneity follows the pattern of spatial autocorrelation by increasing rapidly over a short range in low spatial autocorrelation and gradually over a longer range in high spatial autocorrelation, in aggregation of areas of support at the same resolution. For this reason, plot design in FIA relies on spatial autocorrelation.

In addition, the effect of changing the scale of observation, on relationships between spatially structured variables[3] is also of paramount interest. For example, whether a given modelled relationship between variables (e.g. $Y = f(X) + \epsilon$), is affected by changing the resolution, extent, or plot size of observation is of fundamental interest in FIA. Whether there are effects from changing the support on the coefficient(s) estimated from a modeled relationship, or effects on the variables included in the model, or effects on the overall model fit or model performance, or even effects on the functional form of the relationship, is of fundamental interest. If there

---

[3]A spatially structured variable exhibits spatial dependence due to spatial autocorrelation.

are scale effects, what causes them? Are they real effects or are they artifacts of the modeling process? For example, could there be a "perceived" scale dependence due to omission of a co-varying variable in the model, or is it a true scale dependence (Sandel, 2015)? Are the scale effects a consequence of statistical inequalities arising from the process of nonlinear averaging (Chesson, 2012)? For non-linear relationships between variables, is it also possible that changing the support (extent, plot size) changes the perceived shape of the relationship (Sandel, 2015)? Such questions (and perhaps many more) underpin analyses on scale effects on the relationships between FIA variables and their respective predictors.

1.4 Scaling

Since observations are not always made within their respective scale domains or with complete coverage of the extent of study or analysis, scaling becomes of great theoretical and applied importance. Scaling primarily relates to bridging or transferring information across scales (Marceau and Hay, 1999). It is essentially a means to compensate for scale effects, as transfer of information across scales without consideration of scale effects is susceptible to artifacts in heterogeneous systems. The transfer from "local" to "large" scales is known as *"up-scaling"* whereas the reverse, *"down-scaling"*. Figure 1.3 illustrates up- and down-scaling concepts across space.
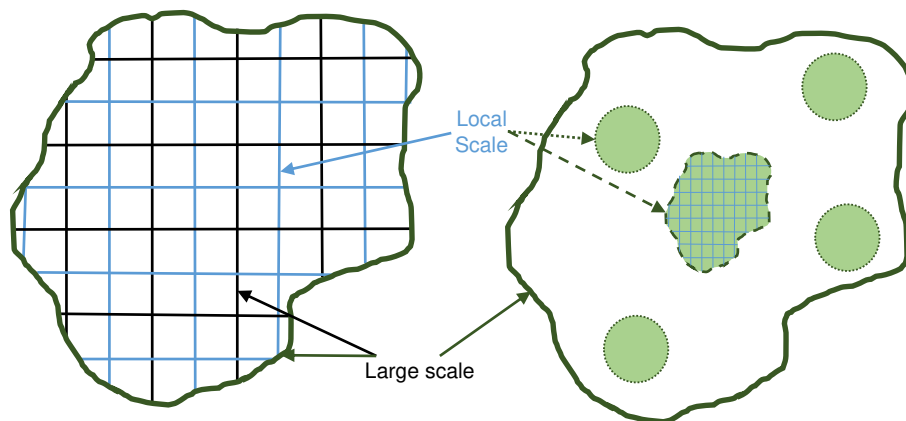
Figure 1.3: Various types of scaling. **Left**: Up/Down scaling of coarser and finer resolutions depending on the direction of transfer. **Right**: (1) Upscaling from a single small region (dashed region) to predict a quantity of a larger extent (solid enclosure), for example the case of expansion factors in forest inventory or scaling from leaf to crown levels in spectroscopy, and (2) Up scaling from a number of samples (dotted regions) that have incomplete coverage to a target region (solid enclosure) e.g. the case of remote sensing assisted inventory of forests. This schematic is adapted after Gunton et al. (2014).

Several methods have been proposed to deal with scaling of measurements, of structured relationships between variables, of retrieval models e.g. in spectroscopy, and of different products derived from remote sensing datasets e.g LAI. Wu and Li (2009) and Malenovský et al. (2007) provide a good starting point for a collection of scaling methods, with a focus on remote sensing.

On scaling of measurements in forest inventory, expansion factors are for example, the simplest and most common form of linear up-scaling or extrapolation, whereas pixel re-sampling of remote sensing imagery is the most common form of down-scaling with respect to data representation but not with respect to the resolution of measurements. Recently, Magnussen et al. (2016) proposed a scaling method for proportions and quantiles per unit area that respects the influence of spatial autocorrelation. Upscaling requires an estimate of the spatial autocorrelation of Y given X at the scale of the original spatial support. Similarly, different forms of interpolations and area weighted approaches (Liu et al., 2006), model-based and model-assisted estimation (McRoberts et al., 2013; Ståhl et al., 2016), hybrid (McRoberts et al., 2016) and hierarchical estimation (Puliti et al., 2018), provide alternatives to bridge measurements at different scales of observation to make estimates of small and large areas. On the other hand, statistical methods accounting for spatial structure of data like for example simulteneous autoregressive models (Sandel, 2015, Example 3), geo-additive models (Fahrmeir et al., 2013), structural equation models for dealing with spatial confounding (Thaden and Kneib, 2018) etc., help to account for spatial confounding effects in spatially structured data with changing of support. The spatial regression models are however often computationally intensive for high resolution or large area remote sensing datasets. On scaling of retrieval models in spectroscopy, Malenovský et al. (2007) propose three scaling approaches including, radiative transfer modelling, spectral unmixing, and data fusion. And lastly, for scaling remote sensing products, Wu and Li (2009) compiled the following, empirical regression between in-situ observations and remote sensing data, the Taylor series expansion method, the contextual parameter method, the statistical fractal and self-similar method, and process simulation methods. They conclude that no single scaling method is universal. The choice of methods is simply case specific. However, increased understanding of scale domains and thresholds would support rapid progress of research into effects from changing the scale of observation.

It is clear, therefore, that there remains a need for increased scientific understanding of scale issues in FIA. However, given the ubiquitous nature and diversity of scale issues in FIA, it is difficult to tackle the scale topic as a whole. Sandel (2015) puts it well; *"identifying and focusing on specific challenges in case studies is likely more productive and provides more fundamental understanding and improved generalization"*. Thus, this thesis presents four case studies written in article format and addressing different scale challenges related to the estimation of forest stand structural heterogeneity, general land cover monitoring with examples on the estimation of forest cover proportions by both visual interpretation of high resolution remote sensing imagery and terrestrial surveys, and the discrimination of tree species from various high resolution remote sensing datasets.

1.5 Overview of the thesis

This chapter presented an introduction to the subjects of scale and heterogeneity. It clarified the terminologies involved in the *"science of scale"* (Marceau and Hay, 1999), and presented the relevance of scale-related analyses in FIA. In Chapter 2 the aims of this thesis are specified.

Chapters 3 compiles materials and methods used across individual case studies. Chapter 4 presents a summary of results from the research, Chapter 5 discusses the findings and lastly, Chapter 6 presents overall conclusions from the research and makes recommendations for future work.

As mentioned, four case studies were conducted and are reported in articles I-IV. Briefly, research in Article I considers the mapping of forest stand structural heterogeneity from airborne LiDAR data at different spatial scales of observation. The study first tackles the problem of finding the scale domain to observe forest stand structural heterogeneity using three forest stand structural heterogeneity indices in forests of different management types and continues to assess the effects of the observed scales on the relationships between the indices of forest stand structural heterogeneity computed from in-situ data and airborne LiDAR data describing the vertical distribution of canopy heights and canopy cover. The findings of this research are relevant to determine observation design in the inventory of forest stand structural heterogeneity.

Research presented in Article II changes focus to explore the challenge of optimizing sampling and observation designs for land cover monitoring at local to global scales using both visual interpretation of high resolution remote sensing imagery and terrestrial surveys. The study addresses the development of a dynamic grid - the global sampling grid (GSG) - applicable at different spatial scales of observation from local to global levels and with specific scaling properties. Here we answer scale questions on how to lay a systematic sampling grid on a spherical approximation of the earth's surface ensuring that the total area of interest is evenly covered and represented in a sample and that a defined minimum distance between sampling locations is fixed. In a case study, the grid is tested in the estimation of proportions of forest cover from clusters of points across landscapes of differing spatial structure highlighting the effects of observation design (scale of observation) on precision of forest area estimates.

The study presented in Article III proposes ways to manage the high within-class variation/heterogeneity in pixel values of high resolution satellite imagery and airborne LIDAR data using object-based image analysis (OBIA) for the discrimination of spectrally and structurally similar tree species at individual crown level. The study presents an approach that combines OBIA and ensemble classification models in distinction of two spectrally and structurally similar *Pinus* trees.

Lastly, research in Article IV delves into scale effects in spectral reflectances of broadleaved tree species collected at leaf and crown levels. This study answers the questions on whether there are differences in spectral reflectance collected at different heirrachical scales and whether the observed scale-related differences affect clasification of the spectral reflectances by species. In addition, the study investigates factors contributing to observed variation in leaf and crown spectral reflectances using a radiative transfer modelling approach and identifies important regions of the electromagnetic spectrum for separating leaf and crown spectral reflectances by species.

1.6 Contribution of the thesis

The research presented in this thesis is of a quantitative and applied nature. The case studies considered different scales of observation and different scale topics (e.g. heterogeneity, scale-dependence, scaling), all within the general context of remote-sensing assisted FIA. In order to emphasize the contribution of this thesis to the general framework of FIA, Figure 1.4 shows the areas to which the case studies contribute knowledge and highlights the appearance of issues on the scale of observation during the data analysis phase of the FIA cycle.
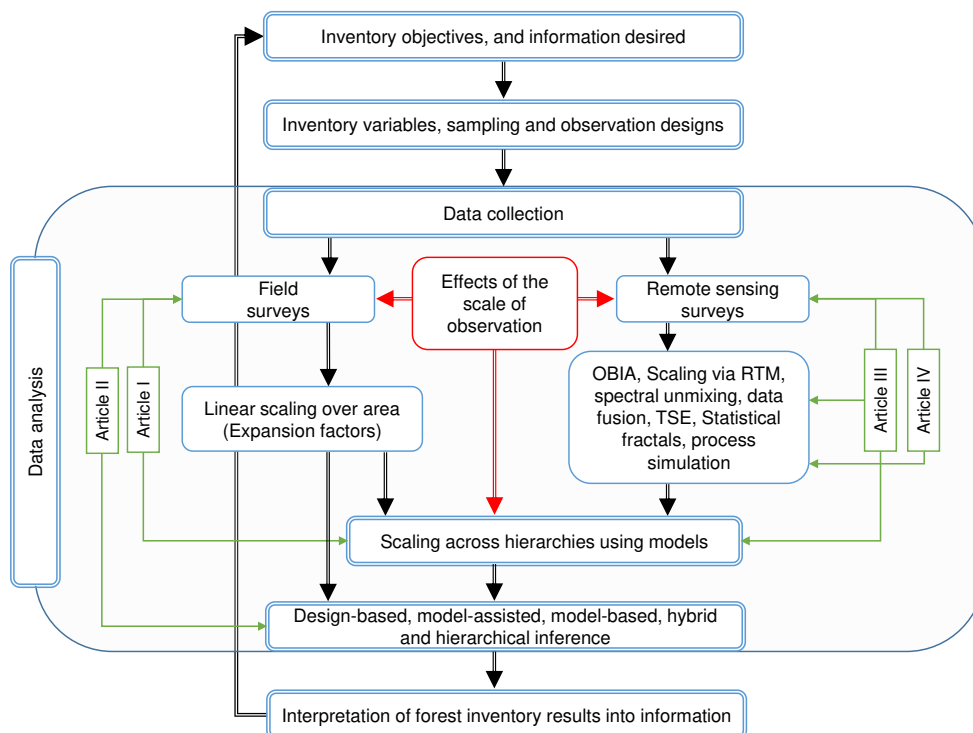
Figure 1.4: Contribution of individual case studies to FIA assisted by remote sensing. The case studies focused on the effects of the scale of observation on data collected from the field, the effects of the scale of observation on data collected by remote sensing and the effects of the scale of observation in the combination of field and remote sensing data through modelling. OBIA stands for Object Based Image Analysis, RTM; Radiative Transfer Models and TSE; Taylor Series Expansion.

Chapter 2

Aims of the thesis

The overall objectives of this cumulative thesis are; (1) to improve the understanding of the association between the scale of observation and observed heterogeneity in inventory of forest stand structural heterogeneity, forest-cover proportions, and identification of tree species from a combination of terrestrial samples and remote sensing data, and (2) to contribute knowledge to the estimation of scale-dependence in inventory of forest stand structural heterogeneity, forest-cover proportions, and identification of tree species from a combination of terrestrial samples and remote sensing data.

The specific objectives across the four case studies are:

- To identify a threshold plot size to estimate forest stand structural heterogeneity in stands under different management systems and create wall-to-wall maps at varying scales of observation from terrestrial samples and airborne LiDAR data.

- To develop and evaluate a sampling framework for land cover monitoring based on a simple and scalable global grid system that allows for simplified and unbiased inference and a straight forward observation protocol at different spatial scales of sampling.

- To evaluate the potential of ensemble models and object-based image analysis (OBIA) to manage high within-class variability encountered in the discrimination of structurally and spectrally similar tree species at individual tree scale from multispectral satellite imagery and airborne LiDAR.

- To highlight the influence of the scale of observation in analysis of spectral reflectance of tree foliage and individual tree crowns.

The technical research objectives for each case study are highlighted in the respective Articles (I-IV) appended to this thesis.

Chapter 3

Material and methods

3.1 Study areas

The case studies presented in Articles I, III and IV were conducted using field inventory and remote sensing data collected in a variety of temperate coniferous, deciduous and mixed forest landscapes, and in stands undergoing a variety of management strategies, in Germany and France. The study in Article II, on the other hand, utilized synthetic data at a global scale and simulated forest/non-forest imagery at a regional scale as shown in Figure 3.1.

The studies in Germany were conducted within the framework of the Biodiversity Exploratories (BE) project - www.biodiversity-exploratories.de, at the Hainich-Dün, Schorfheide-Chorin and Schwäbische Alb exploratories. The three study sites in Germany are distributed across the country in regions of high biodiversity and nature conservation value including two UNESCO biosphere reserves (Schorfheide-Chorin and Schwäbische Alb) and a national park (Hainich National Park in the Hainich-Dün region). Scots pine (*Pinus sylvestris*) stands with parts of beech (*Fagus sylvatica L.*), pine/beech and oak (*Quercus robur*) forests dominate the Schorfheide-Chorin area. In the Schwäbische Alb area, beech forests dominate the lower altitudes and Norway spruce (*Picea abies*) dominates the higher altitudes, and the Hainich-Dün region is dominated by beech stands. The forests in the BEs are of different management type, history and intensity as further elaborated in Article I.

The study in Article III was conducted in a coniferous forest in the South Eastern alpine region of France, the district of Barcelonette. The area is dominated by Pines (*Pinus sylvestris L. and Pinus uncinata Mill. Ex Mirb*), has a wide altitude range (1400 - 2020 m.a.s.l.), rough terrain (slope gradients between 10° and 70°), and is prone to landslides. The forests in the Barcelonette region were planted and left unmanaged for decades with the aim of stabilizing surface soils against landslides. Trees fall often on the shallow soils at steep slopes making the structure of the forest stands uneven in many parts.

Figure 3.1 shows the locations of study sites, highlighting their respective case studies in Articles I-IV. Further details on environmental and forest conditions at the study sites are elaborated or referenced to in Articles I-IV in the appendix.
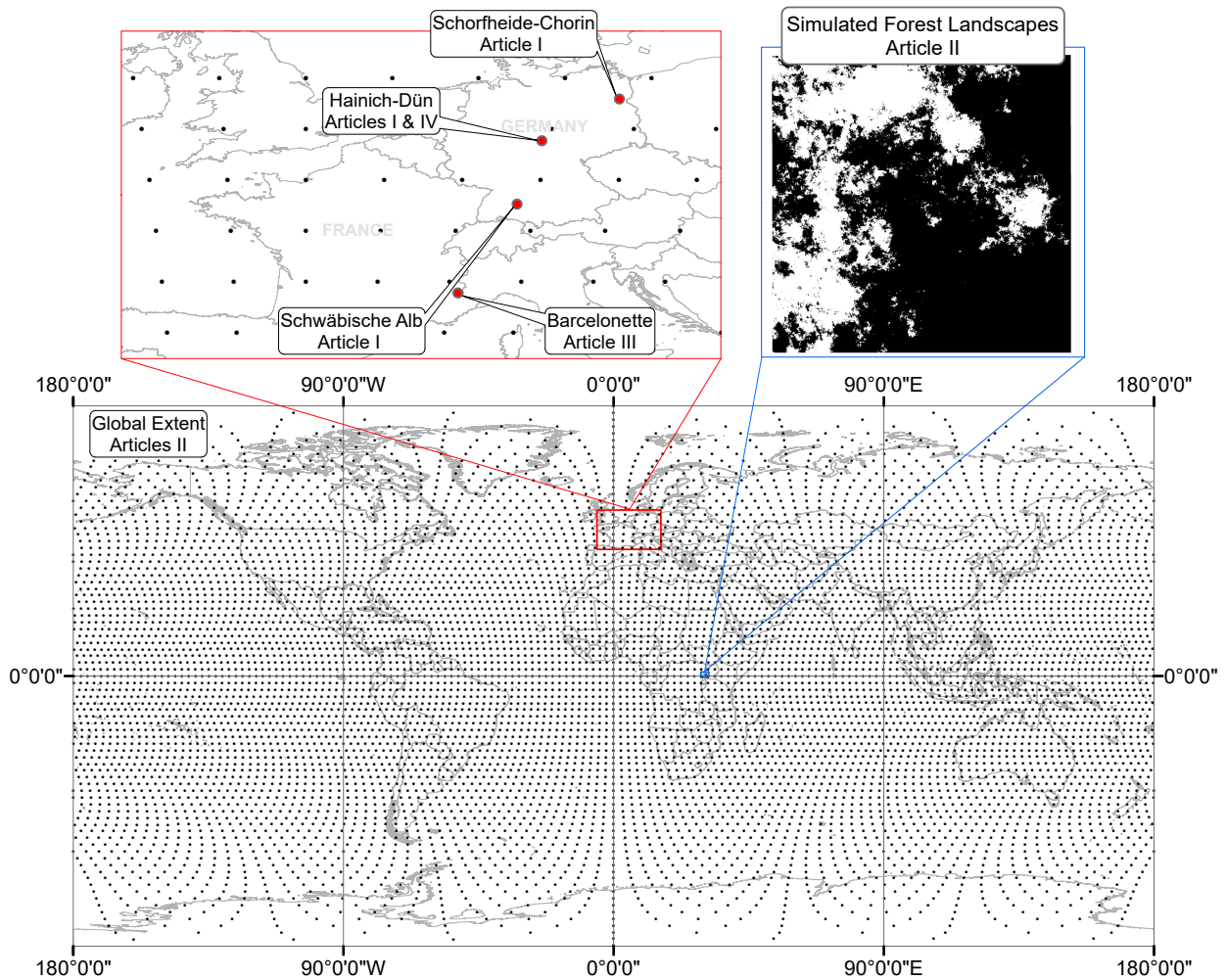
Figure 3.1: Map of the earth surface highlighting the locations of the study areas (red dots) and the respective case studies or articles. The global map shows an example Global Sampling Grid - GSG_{250} (black dots) generated in Article II and covering the entire extent of the earth's surface, with a distance between circles of latitude and points per latitude of 250 km. At the top left is a zoom-in to locations of study sites (red dots) in France and Germany, and at the top right is an example of a simulated forest landscape of 1 x 1° size at the equator.

3.2 Data

The studies utilized in-situ data from field surveys, auxiliary data from remote sensing, and synthentic/simulated data, at different scales of observation (Figures 3.1 and 3.2). Field surveys sampled typical inventory variables like species, diameter at breast height (DBH), tree height, tree locations, as well as collected non-imaging spectroscopy data at leaf and in-situ crown levels. On the other hand, remote full coverage surveys collected very high-resolution airborne LiDAR data sets, and high resolution multi- and hyper-spectral imagery. In addition, synthetic/simulated data included grids of points of varying resolutions calculated from a spherical model of the earth's surface and the so-called global sampling grids (GSGs), and artificial forest landscapes of varying degrees of spatial autocorrelation/fragmentation calculated from a Gaussian random
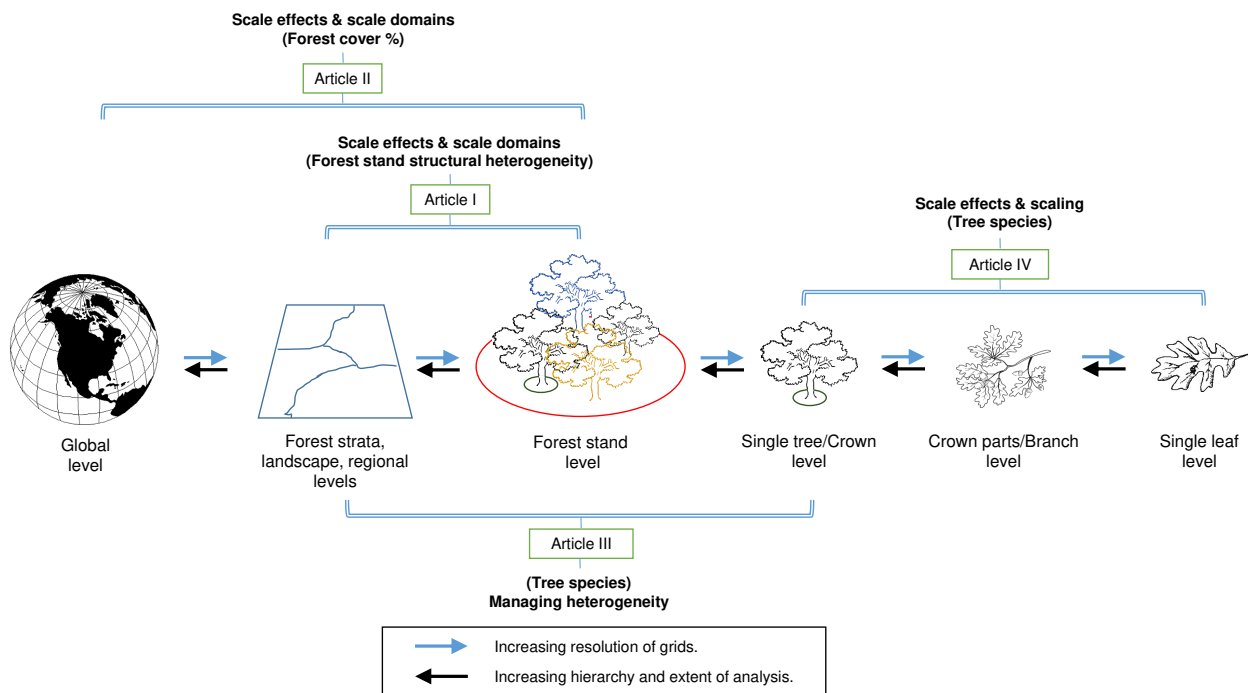
Figure 3.2: Hierarchies/Scales of observation, forest inventory variables (in brackets) and scale topics (in bold) covered by case studies. The term "Resolution of grids" refers to the pixel size of matrices of remote sensing data and does not apply for non-imaging spectroscopy.

fields model. Generally, the data sets varied in extents, spanning from single leaf to global scales and allowed for study designs on scale challenges across different levels/hierarchies (see Figure 3.2). In the following sub-sections more details are provided on each data set with emphasis on the characteristics that make the data suitable for scale-related analyses.

3.2.1 In-situ data

Different in-situ datasets were collected across case studies as listed in Table 3.1. The datasets were collected on individual trees purposely selected or encountered on selected plots during terrestrial surveys by the thesis author or by partners as specified in Articles I, III, IV. Among the in-situ survey data, individual tree positions were particularly important for scale-related analyses as these data supported the optimization of plot sizes and co-location of field and remote sensing data. In case studies reported in Articles I and IV, the Fieldmap® forest inventory system (Hédl et al., 2009, www.fieldmap.cz) was used to collect data on tree locations among the other forest inventory attributes. This state-of-the-art system guaranteed accurate location of individual stems with positional errors around ± 10 cm under full canopy cover. In Barcelonette (see Figure 3.1, Article III), a different approach was employed to collect data on individual tree positions. There, individual tree tops and crowns were initially estimated from a canopy height model (CHM) prior to field work and the respective treetops and crowns verified during fieldwork with the help of landmarks and triangulation - see section on field work in Article III

for details. A similar method (i.e. triangulation from landmarks) was used to geo-locate trees in the 2.25 ha plot collected by Aberle (2016) and used in Article IV. In comparison to the Fieldmap® approach, procedures based on triangulation from landmarks required inexpensive equipment but equally ensured accurate geo-location of trees and subsequent co-location of in-situ and remotely sensed data. They were however very time consuming, a property that restricted their application when a large number of trees were required. In addition, methods to assess precision of estimated locations with triangulation methods requires further research. Therefore, research in Article III and IV did not determine the geo-location accuracy of the estimated tree positions.

The other in-situ dataset that was suitable for scale analyses in this doctoral thesis was spectral reflectance data collected at in-situ leaf and in-situ crown levels by Aberle (2016) and illustrated in Figure 3.3. The non-imaging spectroscopy data sets were collected using the ASD FieldSpec 3© spectrometer within Hainich National Park. Leaf data was acquired with a leaf clip and in-situ reflectance measurements were collected for tree crown parts using a fore optic with an instantaneous field of view of 25°, resulting in footprints of average diameter about 0.5 m. The distance between ASD sensor and crown was about 1 m. However, in a few cases the spacing was up to about 4 m due to accessibility. Some spectral reflectance data were omitted at the in-situ crown level causing gaps (Figure 3.3, middle) due to a high noise-to-signal ratio at the water absorption bands, about 1400 nm (1351-1449) and 1850 nm (1801-1949).
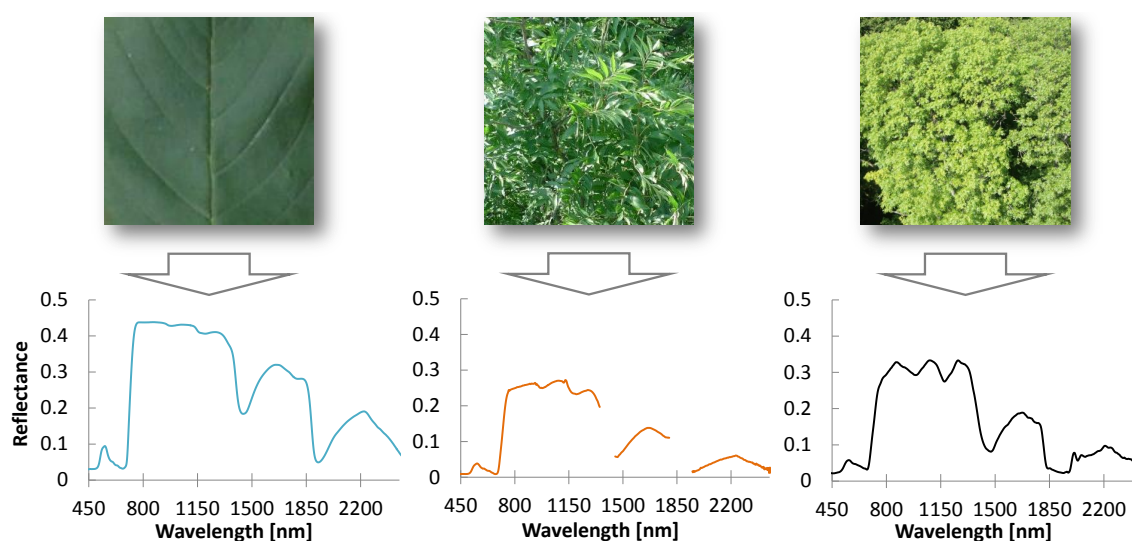
Figure 3.3: Schematic overview of the three levels of data acquisition used in this thesis: in-situ leaf (left), in-situ crown (middle) and airborne crown (right) with corresponding mean reflectance curves. At the in-situ crown level, the sensor noise induced by water vapor is masked out at about 1400 nm and 1800 nm.

3.2.2 Airborne LiDAR data

Research done in Articles I and III utilized very high density LiDAR data collected under the auspices of the University of Twente, Faculty of Geoinformation Science and Earth Observation.

Table 3.1: In-situ tree data across case studies, highlighting all in-situ data sources in table footnotes.

Attribute	Article I	Article III	Article IV
Collection date(s)	Winters 2013[a], 2014[a,b] Winter 2015[b]	Autumn 2012[d]	Winters 2014[b], 2015[b] Winter 2012[e], 24.07.2012[c]
DBH threshold(s)	1 cm[a], 7 cm[b]	7 cm[d]	7 cm[b], 15 cm[e]
Tree stem location system	Fieldmap[®][b] Amplitude Compass & Vertex[a]	SUUNTO Compass & Measuring tape[d]	Fieldmap[®][b] SUUNTO Compass & Vertex[e]
Geo-accuracy	$\pm 10^b$, ± 20 cm - ± 50 cm[a]	not calculated	$\pm 10^b$, not calculated[e]

[a]A 28.5 ha plot in Hainich National Park collected by the German Center for Integrative Biodiversity Research (iDiv), Germany.
[b]80 one ha. EP plots collected by the Biodiversity Exploratory project, Germany.
[c]In-situ hyperspectral data, collected by Aberle (2016) in Hainich National Park, Germany.
[d]Field inventory in Barcelonette, France, collected by the author.
[e]A 2.25 ha plot data collected by Aberle (2016) in Hainich National Park, Germany.

Wall-to-wall, high density, airborne LiDAR data offers opportunities for scale-related research especially because it allows for variation of pixel sizes in matrices derived from the point-cloud. When tree position data were available from field surveys, both data sets allowed for optimization of plot sizes without encountering scale-mismatches from their relation. Researches in Articles I and III illustrate such utility from airborne LiDAR data in scale-related analyses. Table 3.2 summarises the characteristics of airborne LiDAR data utilised across case studies.

Table 3.2: Characteristics of airborne LiDAR data across case studies.

Attribute	Article I	Article III
Collection date(s)	July 2015	June 2009
Sensor (Frequency)	Riegl Q780 (400 kHz)	Riegl VQ480i (300 kHz)
Altitude	940	400
Waveform	Full	Full
Potential returns per pulse	4	7
Mean return density (point/m^2)	31	160

3.2.3 Optical imagery data

High-resolution Worldview-2 multispectral satellite imagery was collected for the case study presented in Article III by the University of Twente, Faculty of Geoinformation Science and Earth Observation whereas airborne hyperspectral imagery from a combination of AISA EAGLE[©] and HAWK[©] sensors (SPECTRAL IMAGING LTD, Finland) was collected for the case study presented in Article IV by the Department of Forest Inventory and Remote Sensing, Faculty of Forest Sciences and Forest Ecology, University of Göttingen. The hyperspectral imagery are the airborne crown scale illustrated in Figure 3.3. The general characteristics of the optical data

utilized in this thesis are presented in Table 3.3.

Table 3.3: Characteristics of optical data across sensors and case studies. Worldview-2 data was utilized in Article III and AISA-HAWK data in Article IV.

Attribute	Worldview -2	AISA-HAWK
Collection date (time)	September 2010	24.07.2012 (10:30 am - 11:30 am)
Average sun elevation (azimuth) angles	48.1°(161.7°)	58.27°(168.99°)
Average platform elevation (azimuth) angles	74.8°(55.0°)	Nadir (Nadir)[a]
Processing level	L2	RAW
Spatial resolution	2 m	2 m
Spectral resolution	8 Bands	2001 Bands

[a]Flight height $=$ 1700 m above ground level.

3.2.4 Synthetic data

Synthetic data was generated for the case study presented in Article II by the thesis author. As previously mentioned, the dataset included sample locations of the global sampling grid (GSG) framework (Fehrmann et al., 2019) and forest landscapes of varying degrees of fragmentation/spatial autocorrelation produced from a Gaussian random fields model. Article II gives details on the construction of the GSG and generation of the artificial forest landscapes. An example grid - GSG_{250}, is shown in Figure 3.1. The simulated forest landscapes of varying fragmentation and spatial autocorrelation are shown in Figure 3.4. It is on the six landscapes (Figure 3.4) that cluster observation designs were optimized to estimate forest cover percentage in Article II. See Article II for further details on the Gaussian random fields model used to generate the artificial forest landscapes.

3.3 Forest inventory variables

The research presented in this doctoral thesis focused on forest stand structural heterogeneity, forest cover proportions and tree species identities. Forest area and tree species are traditional and very important variables commonly assessed over large areas using sample-based surveys and remote sensing. Quantification of change in forest area is the basis for assessments on deforestation. Information on tree species distribution supports management and conservation planning, pest and disease control, and biomass estimation. Forest stand structural heterogeneity, on the other hand, is a contemporary FIA variable that is increasingly demanded in multi-resource inventories due to its linkage to biodiversity (McElhinny et al., 2005), productivity (Juchheim et al., 2017; Dănescu et al., 2016; Bohn and Huth, 2017; Schall and Ammer, 2013), ecosystem functioning and stability (Díaz-Yáñez et al., 2017) and its importance in biomass modelling over large areas (Rödig et al., 2017). This research described and quantified the studied forest inventory variables as elaborated in the following three sub-sections.
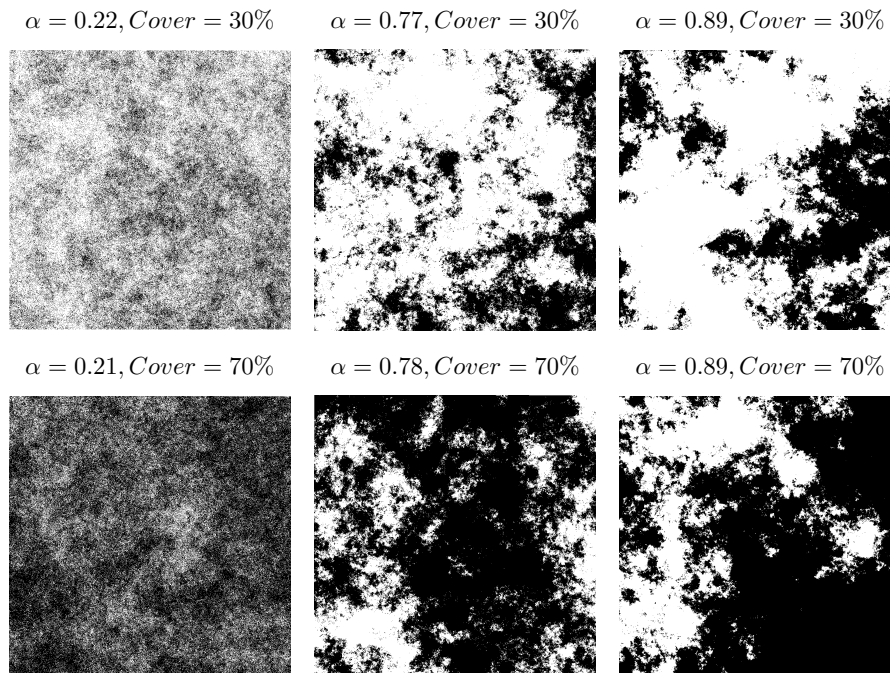
Figure 3.4: Simulated forest landscapes of varying fragmentation and spatial autocorrelation. The α parameter is a measure of patch complexity and therefore controls for fragmentation. Low α values lead to high fragmentation and vice versa.

3.3.1 Forest stand structural heterogeneity

Forest stand structure is described in this research as the spatial arrangement, size and distribution of individual trees in a volume of space lain on a horizontal ground area of one hectare. The stand area needed to be defined a priori in order to compute heterogeneity, because forest stand structural heterogeneity is scale-specific. Forest stand structural heterogeneity was quantified through indices computed from in-situ field observations and through quantitative summaries of the vertical distributions of airborne LiDAR returns. From in-situ field observations, three indices of forest stand structural heterogeneity were considered; Gini Coefficient (GC), Structural Complexity Index (SCI) and Enhanced Structural Complexity Index (ESCI). The GC is a non-spatial index whereas SCI and ESCI are spatial indices (see Figures 3.5 and 3.7). GC (Weiner and Solbrig, 1984) is in this thesis used to measure tree size inequality - and in Article I, utilized DBH as the indicator for tree sizes (Valbuena et al., 2012). GC can be calculated in two equivalent ways, of which the second was used in this research; (1) from the so-called Lorenz curves, obtained by mapping the cumulative proportion of ordered basal area to the cumulative proportion of the total number of trees encountered on the plot (Valbuena et al., 2013, see also Figure 3.5). For stands of different degrees of tree size inequality, Lorenz curves drift off the axis of absolute equality (Figure 3.5). GC is then obtained by taking the ratio of the area between the Lorenz curve and the axis of absolute equality to the total area under the axis of absolute equality. Alternatively, (2) GC was here calculated from the arithmetic mean of differences between all ordered individual tree pairs normalized by the basal area of the quadratic mean diameter (see equation 1, section 2.5 in Article I). Both ways yield the same continuous

values between 0 for theoretically perfect equality among trees and 1 for theoretically perfect
inequality as observed for example in stands with maximally bi-modal distributions (Valbuena
et al., 2013) and thus the largest possible dispersion among tree sizes. The zero to one range
of GC ordered inequality in tree sizes in a logical way able to segregate stands of different
forest structural types in the study areas when the position of quadratic mean diameter on the
Lorenz curve, also equivalent to the curve's point of inflection, was monitored together with
the value of GC (Figure 3.6 - according to Valbuena et al. (2013)).
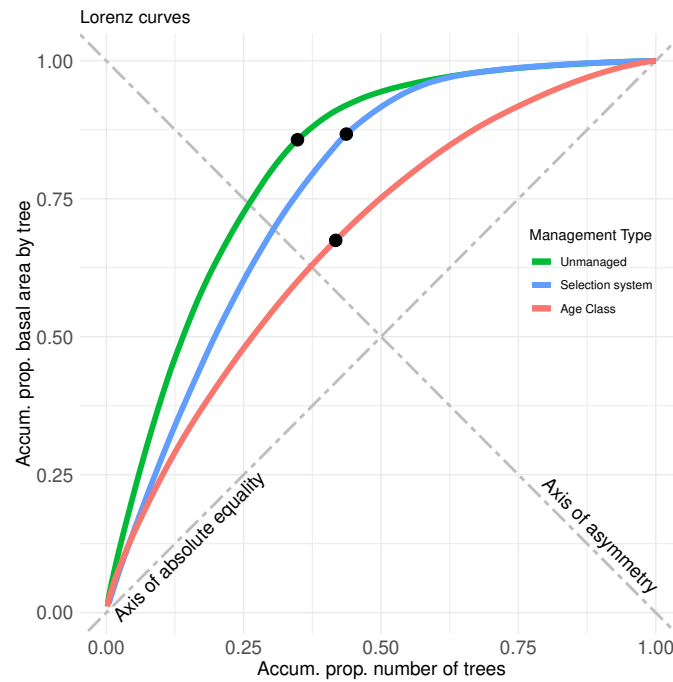
Figure 3.5: Illustration of Lorenz curves generated from three field plots of the Biodiversity Exploratory
Project each collected in a stand of different forest management type: unmanaged, selection system, age
class. "Selection system" refers to single tree selection or selective logging and age class refers to even-
aged forests representing different development stages. The graph highlights differences among forests of
varying tree size distributions. GC and the position of the quadratic mean diameter change across plots.
The straight lines show the axes of absolute equality and asymmetry. Any degree of inequality among
individual pairs drifts the Lorenz curves off the 45° line. Notice that the cumulative proportions are
computed for ordered basal areas decreasing from the largest to the smallest. This means, for example,
large trees contribute about 85% of basal area in the analyzed unmanaged forest and are about 35% of
the total number of trees when using the quadratic mean diameter as the size threshold. Such properties
are useful to discriminate the different forest management types.

The other index utilized to quantify forest stand structural heterogeneity is SCI. SCI (Zenner
and Hibbs, 2000) was employed using DBH as the indicator for tree size. Unlike GC, SCI
simultaneously measured the degree of inequality in tree sizes and variability within and between
forest patches by explicitly accounting for individual tree positions in space. SCI is a spatial index
conceptualized as the adjacency of differently sized trees and computed by linking individual
tree locations and their respective sizes in a Delaunay triangulation. The index was obtained
from the ratio of the surface area of the Delaunay triangulation to the area in projection (Figure
3.7). The computation yields continuous values from 1 when there is perfect equality among
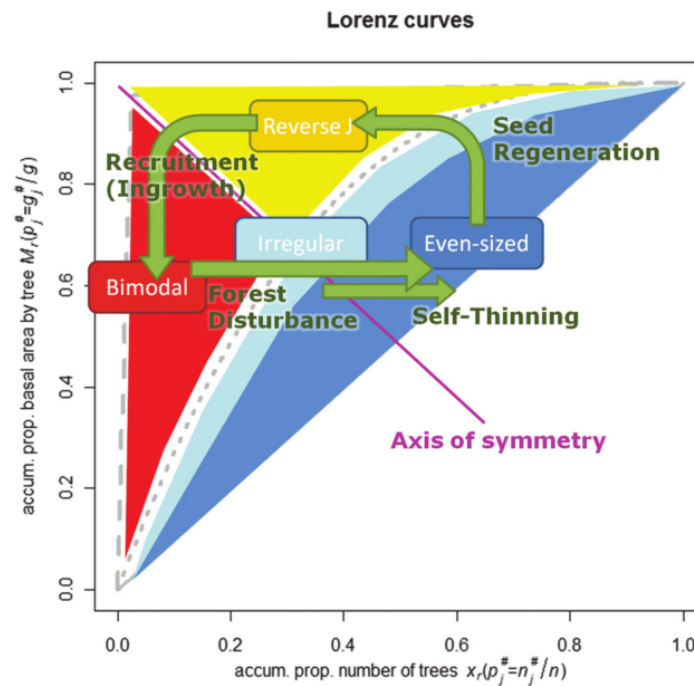
Figure 3.6: Illustration of the ability of GC to segregate forests of different tree size distributions. X-axis shows the cummulative proportion of ordered number of trees and Y-axis shows the cummulative proportion of ordered basal area. The straight lines show the axes of absolute equality and asymmetry. Any degree of inequality among individual pairs drifts the Lorenz curves off the axis of absolute equality. The different colors show the regions on Lorenz curves generated from stands of different forest structural types i.e. even-sized, irregular, reverse J, bi-modal. The regions are demarcated based on GC and the position of the quadratic mean diameter as shown in Figure 3.5. It is necessary to use both GC and the position of the quadratic mean diameter simultaneously since asymmetry of the Lorenz curves yields similar GC values for stands of different tree size distributions. In plantation forestry, different treatments as shown by the green arrows (i.e. self thinning, seed regeneration, recruitment, forest disturbance) result into transitions from even-sized to bi-modal stands as further elaborated by Valbuena et al. (2013). This figure is adapted after Valbuena et al. (2013).

trees to infinity for infinite scenarios of perfect inequality among trees exhibiting variant spatial configurations. A derivative of SCI referred to as ESCI, integrated into the computation of SCI the magnitude and orientation of individual vectors of triangles of the Delaunay triangulation to form what Beckschäfer et al. (2013) called the vector ruggedness measure, and also added a stem density term. According to Beckschäfer et al. (2013), the two additional terms in computation of SCI increase ESCI's characterization capability among structurally similar forest stands while maintaining the index's ability to linearly arrange the stands along a gradient of low to high structural heterogeneity. More detail on the computation of GC, SCI and ESCI is available in Article I and the indices' respective methodological publications as cited therein.

This research considered GC, SCI and ESCI over the numerous indices summarizing forest stand structure (Zenner and Hibbs, 2000; Pommerening, 2002; Beckschäfer et al., 2013; Valbuena et al., 2012; Ehrbrecht et al., 2017; del Río et al., 2016) mainly because the three indices concisely summarize the essential attributes describing forest stand structure (i.e. the spatial arrangement, size and distribution of individual trees in space) and because they can be

easily mapped across larger areas from remote sensing when compared to other well established indices describing the spatial arrangement, size and distribution of individual trees in space (e.g. Ripley's K (Ripley, 1977), Besag's L (Besag, 1977) and semivariograms). Ripley's K, Besag's L and semivariograms fail when large area maps are required because of their functional nature. In other words, the methods summarize point patterns into lag distances on the x-axis and index quantities on the y-axis making it challenging to relate with remote sensing data in the third dimension.

The indices utilizing field observations described the vertical (GC) as well as both the horizontal and vertical (SCI and ESCI) distributions of tree sizes in the studied forest stands. On the other hand, airborne LiDAR metrics described the vertical distribution of canopy heights and cover (Table 3.4). Article I investigates whether there is a relationships between indices computed from in-situ field inventory data and metrics computed from the airborne LiDAR point cloud at different scales of observation, with the aim to map forest stand structural heterogeneity over areas larger than what is estimated from only in-situ field surveys. To achieve this aim, different airborne LiDAR metrics (Table 3.4) describing various aspects of canopy structure like canopy cover (A_j, B_k, D_n, E_m), dominant height (Max, P_{90-99}), crown dominance (L.Skew), variation in canopy returns (ADD, CV) etc. were computed from the airborne LiDAR point cloud aggregated at different scales of observation. Article I gives further details on methods used in computation of airborne LiDAR metrics. Table 3.4 summarizes the all the computed airborne LiDAR metrics.

3.3.2 Forest cover proportion

In forest inventories, the proportion of forest cover is either estimated from sampling surveys or from maps of forest cover commonly derived from remote sensing data. Sampling surveys are of two categories including traditional terrestrial sampling and sampling from aerial photographs (airborne or satellite-borne). The case study in Article II mainly deals with monitoring of forest cover from sampling surveys (both terrestrial and remote). The inventory variable "Forest cover proportion" was in the case study defined in a similar way as done by Magdon and Kleinn (2013), as the proportion of forest covered by the vertical projection of tree crowns with overlaps not counted (Geschwanter et al., 2009; Jennings et al., 1999). From the definition of forest cover proportion, it is apparent that a clear definition of the term "forest" was required before forest cover proportion could be calculated. According to Kleinn (1991); Kleinn (2001); Magdon and Kleinn (2013), a good forest definition includes both qualitative and quantitative criteria. The quantitative criteria are definition of, minimum area, minimum crown cover percentage, minimum tree height and minimum width. Qualitative criteria include descriptions of how to deal with special features like roads, creeks, clear-cuts; how to deal with plantations or even plantations of "non forest" trees such as palms, orchards etc., or how to deal with trees outside "forests". Apart from the definition of a forest, Magdon and Kleinn (2013) mention two other important factors to consider when determining forest cover proportion from remotely sensed imagery, namely, (1) spatial resolution of the sensor, and (2) size of the reference area
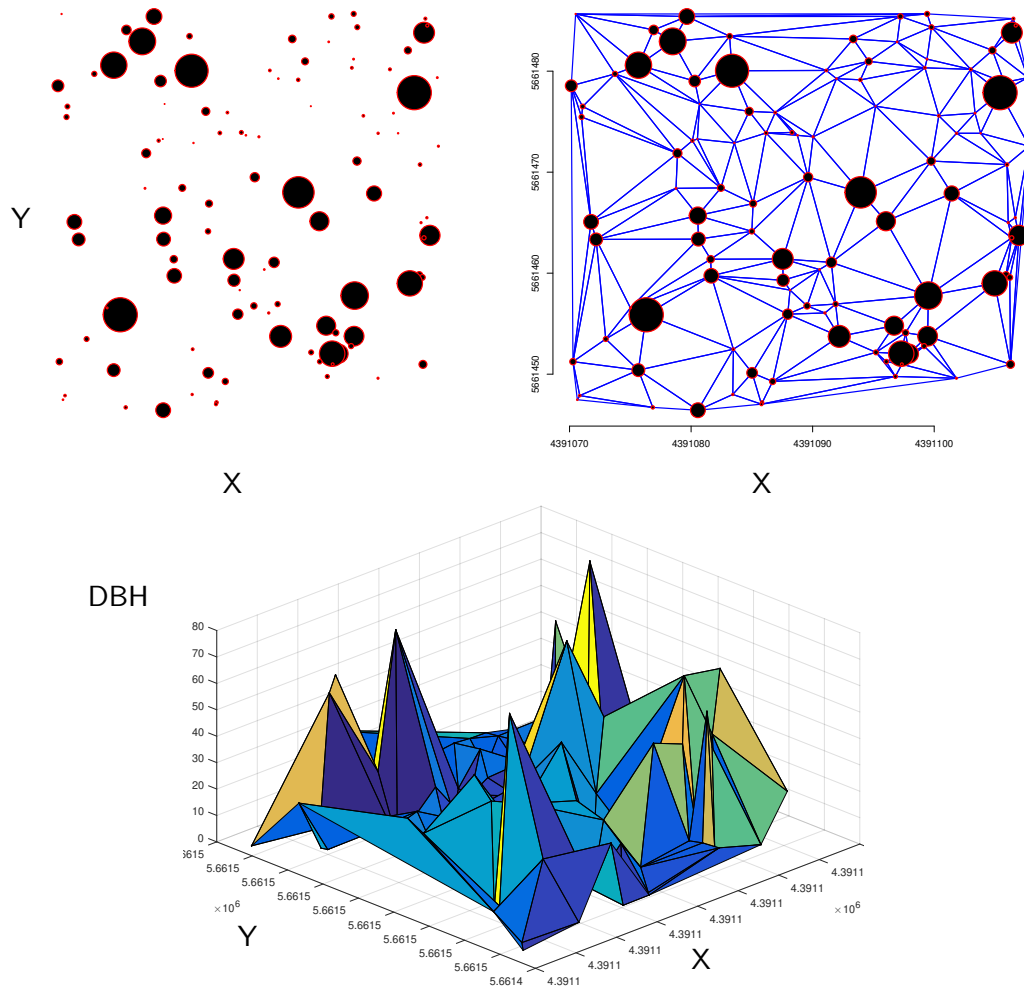
Figure 3.7: Computation of Structural Complexity Index (SCI) and Enhanced Structural Complexity Index (ESCI) from field observations using tree location and DBH in the Delaunay triangulation. Top left are the tree locations scaled by DBH. Top right is the resultant Delaunay triangulation. Bottom is the 3D representation of the surface of the Deluanay triangulation. SCI is computed as the ratio of the surface area of the Deluanay triangulation to the projection area. ESCI is computed by adding a surface ruggedness measure and a stem density term.

considered to determine forest cover proportions. The spatial resolution is important because it determines the fraction of intensity of the reflected electromagnetic radiation attributable to forest features given the size of the pixel. And the size of the reference area is important because it determines the class of a given reference area across forest/non-forest categories given the minimum crown cover criterion. Magdon and Kleinn (2013) show in a simulation setting how both the spatial resolution and the size of the reference area affect estimated values of forest cover proportion, conditioned on a minimum crown cover percentage. They go ahead to recommend defining the size of the reference area if comparable estimates of forest area are to be derived from analyses considering a given minimum crown cover percentage.

This thesis considers the above mentioned criteria when defining forest land prior to computing forest cover proportion in the case study presented in Article II. Forest land was defined

Table 3.4: Summary of LiDAR metrics used in this study (Kukunda et al., 2019).

Height related metrics	
Group	**Abbreviation**
Order statistics	Min; Max; P50
Height quantiles	P_i for $i = 1, 5, 10, 20, 30..90, 95, 99$
Central tendency	Mean; Mode
Dispersion	SD; Var; CV; IQR; AAD; $\mathrm{MAD_{mode}}$; $\mathrm{MAD_{med}}$
Skewness,Kurtosis	Skew; Kurt
L-Moments	L_i for $i = 1, 2, 3, 4$
L-ratios	L.CV; L.skew; L.kurt
Generalized means for the 2nd and 3rd power	QRT, CUB

Crown cover related metrics	
% first returns above: 7 m, mean, mode	A_j for $j = 7$ m, mean, mode
% all returns above: 7 m, mean, mode	B_k for $k = 7$ m, mean, mode
% first returns above: 0 m, mean,mode	D_n for $n = 0$ m, mean, mode
% all returns above: 0 m, mean, mode	E_m for $m = 0$ m, mean, mode
Canopy Relief Ratio	CRR
Count of returns by return number	C_i for $i = 1{:}9$, Total, Other
Count of returns above: 7 m, 0 m, mean, mode	C_{7m}, C_{0m}, C_{mean}, C_{mode}
Count of first returns above: 7 m, 0 m, mean, mode	$C1_{7m}$, $C1_{0m}$, $C1_{mean}$, $C1_{mode}$

Abbreviations: SD, Standard deviation; CV, Coefficient of variation; IQR, Inter-quartile range; AAD, Average absolute deviation; Lskew, L-moment of Skewness; Lkurt, L-Moment of Kurtosis; $\mathrm{MAD_{mode}}$, Median of the absolute deviations from the overall mode; $\mathrm{MAD_{med}}$, Median of the absolute deviations from the overall median. Further information on metrics is available in McGaughey (2014).

by a minimum area equal to 30 x 30 m^2 and minimum crown cover percentage of 100% for a reference area of 30 x 30 m^2. This definition sufficed for all synthetic forest/non-forest cover maps shown in Figure 3.4. From a remote sensing perspective, employing synthetic data helped overcome issues regarding the spatial resolution of the sensor in defining forest/non-forest areas since every 30 m pixel was either 100% forest or 100% non-forest, which was equivalent to 100% intensity of reflected electromagnetic radiation from forest features and thus a forest map with 100% overall and user accuracy. Similarly, to observe forest cover, dimensionless point observations were employed across different cluster observation designs (as further elaborated in section 3.4), masking out scale issues related with the size of the reference area in definition of a forest/non-forest class at a given location/pixel observed.

To compute forest cover proportion from samples collected over the artificial landscapes (Figure 3.4), systematic sampling designs were used across case studies presented in Article II in a similar way as done by Kleinn (1991); Bastin et al. (2017). Different point observation designs were considered including, single points and clusters of points, and therefore different estimation designs. When considering single point observations, the existence of forest land at individual sample locations was queried from each simulated forest landscape and the resulting

area estimated from the sample proportion $\hat{p} = n_f/n$, where n_f is the number of grid points over forest land and n is the sample size, i.e. the total number of grid points. When clusters of points were considered, the proportion of forest cover was estimated following Cochran (1977, Chapter 3, Pg. 66). In this case, forest cover proportion at the ith cluster was given by $p_i = f_i/m_i$, where f_i is the number of points with forest cover and m_i is the size of the cluster, i.e. the number of sub-plots. As cluster size could vary due to overlap of the sampling grid at the edge of the sampling frame, a ratio of means (rom) estimator was used for estimating forest cover proportion $\hat{p}_{rom}$ of a forest landscape from a sample of n clusters (Equation 3.1):

$$\hat{p}_{rom} = \frac{\sum\limits_{i=1}^{n} f_i}{\sum\limits_{i=1}^{n} m_i} \tag{3.1}$$

3.3.3 Tree species

The final FIA variable considered here (Articles III & IV) is tree species. Identification of tree species is important in forest inventory operations especially within forest management and conservation contexts. Depending on forest type and region, accurate identification of tree species is often possible from terrestrial surveys where morphological and anatomical characteristics are examined to visually distinguish among species identities, however, species identification remains complex from a remote sensing perspective. This is mainly because, different species may exhibit similar characteristics retrievable from remote sensing data. Additionally, the probability that a remote sensing pixel or object corresponds to the size of a single tree or stand or matches their respective extents is very low. Due to the interaction between resolution of observation, canopy closure, canopy structure etc., pixels contain background noise, shadows, mixed signal etc. As a consequence, discreet tree species classes do not directly correspond to information obtained by remote sensing data. It is therefore important to determine the optimal resolution of observation and the appropriate spatial unit of analysis on which to obtain information on tree species.

This research examines discrete classes of tree species within one genus (Article III) and among genera (Article IV) with emphasis on scale effects and scaling of species traits obtained from remote sensing data. The term "species traits" refers to aggregations of remote sensing attributes over a specific area.

3.4 Approaches used to observe at different scales

To study the influence of the scale of observation, observations across different scales are necessary. Again, Figure 3.2 shows the different scales of observation considered by this research across case studies. The case study in Article IV utilized terrestrial and remotely sensed observations at three scales of observation as shown in Figure 3.3 and relied on raw survey data. This is not always the case in most scale-research settings, especially because of cost

implications. In most cases scale-researchers rely on simulations as done in the case studies in Articles I, II and IV or green house experiments (Gara et al., 2018, or example).

In Article III, different scales of observation were monitored through OBIA. The OBIA procedure developed by González-Ferreiro et al. (2013) was used to segment a canopy height model into objects approximating individual tree crowns, after monitoring various extents around tree tops in rule-based iterations of a watershed segmentation algorithm. In the algorithm, iterations resulted in changes of extents of individual objects ("watersheds") where merging was based on statistical similarity of underlying pixels with their neighborhood. Further information on the OBIA procedure is given in section 3.6.1 and shows how the OBIA technique can be used to manage high heterogeneity in pixel data.

In order to evaluate the effect of changing observation designs in estimation of forest cover proportion (Article II), this study changed the scales of observation across one cluster observation design by employing different number of subplots (cluster size) and across different cluster observation designs maintaining the same number of subplots (see Figures 3.8 and 3.9). In addition, different cluster extents were evaluated for every combination of cluster observation design and cluster size across different forest landscapes.
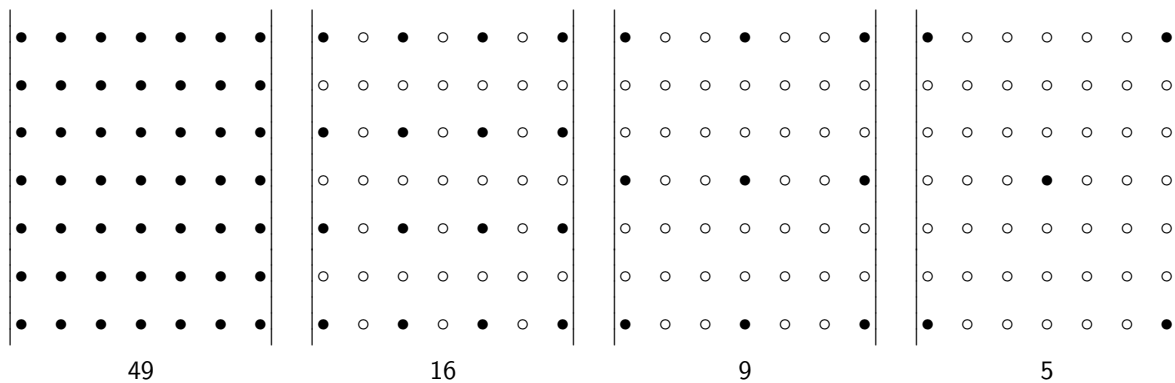
Figure 3.8: Spatial configuration and cluster size (number of subplots); ● = subplot, ○ = NULL. The total number of subplots in each cluster is shown below each design (Fehrmann et al., 2019).
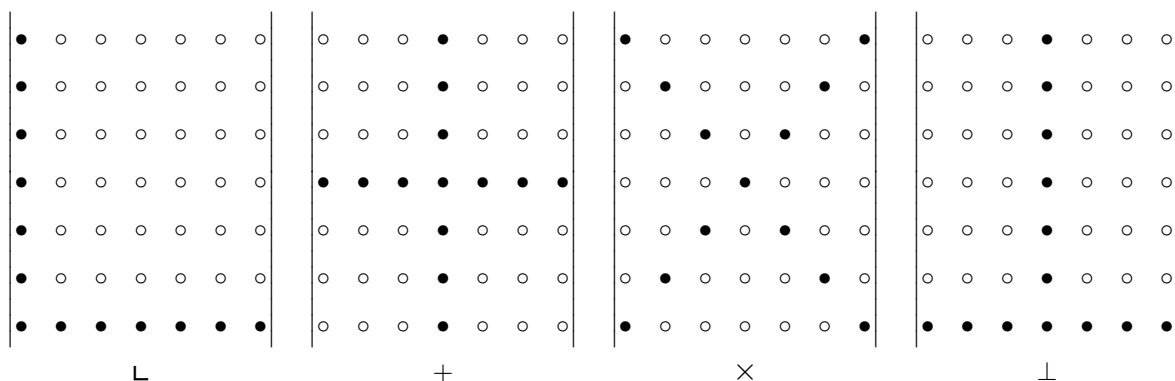
Figure 3.9: Different spatial configuration of clusters with constant number of 13 subplots. ● = subplot, ○ = NULL (Fehrmann et al., 2019).

Similarly, a simulation was used to monitor different scales of observation in the estimation

of forest stand structural heterogeneity (Article I) in which multiple plot sizes were mimicked in a moving window over plot maps. The moving window algorithm evaluated 18 plot sizes from 15 m * 15 m to 100 m * 100 m in 5 m steps as shown in Figure 3.10 and further elaborated in Article I. GC, SCI, ESCI were computed at each position of the moving window. The next section explains how heterogeneity was assessed at the different scales observed.
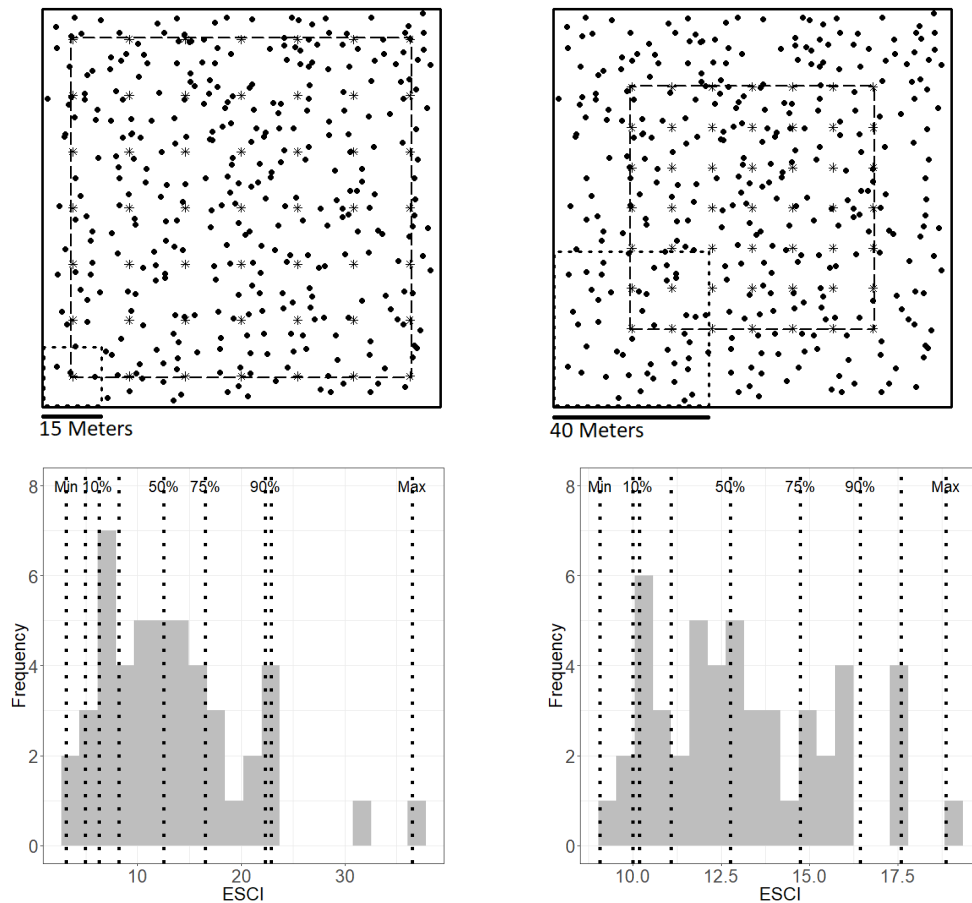
Figure 3.10: 15 m * 15 m and 40 m * 40 m plots (above) over an example EP plot (1 ha) with corresponding histograms (below) based on enhanced structural heterogeneity index (ESCI). Dotted points (above) show tree locations, asterisks (above) show plot centers for all 49 laid subplots, the dotted lines (above) show boundaries of one of the 49 plots, and the dashed lines show the sampling frame within which plot centers were laid in a way that maintained overlap between subplots and had no edge effects. The dotted lines (below) show positions of the minimum and maximum index values, and 5th, 10th, 25th, 50th, 75th, 90th, 95th percentiles of the distribution (Kukunda et al., 2019).

3.5 Quantifying heterogeneity across scales of observation

The assessment of heterogeneity in regression statistics commonly refers to description and accounting of heteroscedasticy (see for example, Zuur et al., 2009, page 20). In sampling statistics, assessment of heterogeneity refers to estimating variance of a population and the associated precision from a sample. This research focused on the assessment of heterogeneity in

respect to quantification of variance across scales of observation. Heteroscedasticity challenges could be expected across the case studies but these were not explicitly assessed by this study. For example, in Bernoulli sampling for forest cover (Article II) the variance depended on the mean (which may be scale/resolution dependent), similarly, in the assessment of spectral reflectance from foliage, "branches", and tree crowns, more variance was observed at (cut-out) wavelengths associated with water vapor. In addition, there was intrinsic heteroscedasticy in the form of unequal variances across LiDAR canopy height percentiles used in the discrimination and classification of the two similar *Pinus* species (Article III).

Sample data can be heterogeneous to varying degrees, conditioned on the scale at which they are observed vis à vis the spatial autocorrelation in the sampled landscape and the inventory variable in question. Two of the three FIA variables considered in this research have extensive properties including, forest cover proportion and tree species (i.e. in reference to pixel/spectral values per species class). The other FIA variable i.e. forest stand structural heterogeneity, is scale-specific. The extensive variables concentrate around the mean with changing extents of observation and therefore allow for "independent" comparison of heterogeneity estimates across scales of observation. However, in analysis of scale effects in estimation of forest stand structural heterogeneity (a scale-specific variable), structural heterogeneity at the different scales of observation needed to be interpreted relative to a reference scale (of one hectare).

Variance (s^2) in individual estimates of forest stand structural heterogeneity and of the pixel/spectral values per species class at each scale of observation, was computed using the simple random sampling estimator shown in Equation 3.2, for normal distributions;

$$s_{\mathrm{srs}}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\bar{y}})^2}{n-1} \tag{3.2}$$

where y_i is the individual observation, $\hat{\bar{y}}$ is the mean across observations and n is the size of the sample.

Similarly, a conservative variance estimator of $\hat{p}_{\mathrm{rom}}$ from systematic sampling with cluster plots assuming simple random sampling without replacement, was computed for cluster observations for each estimate of forest cover proportion (Article II), as there are no design-based variance estimators for systematic sampling (Magnussen and Nord-Larsen, 2019). It should be noted that the use of simple random sampling estimators for data collected according to a systematic sampling design does not provide a good example of heterogeneity across scales since the estimators do not consider the spatial configuration of the data nor take any spatial covariance process into consideration. Scale effects in sampling becomes a function of the choice of variance estimator and its ability to capture spatial covariance

$$\hat{v}(\hat{p}_{\mathrm{rom}}) = \frac{1}{n\hat{m}^2} \frac{\sum f_i^2 - 2\hat{p}\sum f_i m_i + \hat{p}^2 \sum m_i^2}{n-1}, \tag{3.3}$$

(Cochran, 1977)

Where $\hat{m} = \sum m_i/n$ in Equation 3.3 (Cochran, 1977, Chapter 3, Pg. 66) is the average number of subplots per cluster in the sample and $\hat{p}$, f_i and m_i follow from Equation 3.1 above.

To compare the performance of the different cluster configurations, the relative observed standard error was computed across $K = 1000$ iterations following Equation 3.4:

$$\mathrm{RSE_{obs}} = \frac{V_{obs}}{\bar{\hat{p}}} \times 100\%, \tag{3.4}$$

where $\bar{\hat{p}} = 1/K \sum_{k=1}^{K} \hat{p_k}$ is the mean of the generated sample distribution of $\hat{p}$ and $V_{obs} = 1/K \sum_{k=1}^{K} (\hat{p_k} - \bar{\hat{p}})^2$ is the according observed variance.

In order to estimate scale differences among species spectral reflectance (Article IV, Figure 3.3), means and confidence intervals across observations were estimated at each scale of observation with the help of Generalized Additive Models (GAMs) employing P-splines smooth functions (Eilers et al., 2015; Wood, 2011). Fitted models were of the general form;

$$y_i = f(z_i) + \epsilon = \sum_j^d \gamma_j B_j(z_i) + \epsilon_i \tag{3.5}$$

<div align="right">(Fahrmeir et al., 2013)</div>

where y_i are the reflectance values, z_i are the wavelengths, γ_j is a vector of coefficients, B_j specifies the basis functions and ϵ_i is the error term. The number of basis functions considered depended on the chosen knot configuration. Thus d is the total number of linear combinations. The P-splines employed polynomials of degree l, therefore the standard residual sum of squares estimate could not apply for the estimation of γ without considering a penalty λ that controlled for smoothness and over fitting (Equation 3.6);

$$\mathrm{PLS}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_j^d \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2. \tag{3.6}$$

<div align="right">(Fahrmeir et al., 2013)</div>

The penalized least squares estimate of the vector γ given λ, $\mathrm{PLS}(\lambda)$, was arrived at by minimizing Equation 3.6. It followed that the mean estimate $\hat{f}(z)$ from Equation 3.5 is equivalent to a weighted sum of the reflectance observations y_i pivoted at each wavelength z as shown in Equation 3.7;

$$\hat{f}(z_i) = s(z, z_i) y_i \tag{3.7}$$

<div align="right">(Fahrmeir et al., 2013)</div>

with $s(z, z_i)$ a vector of weights given the wavelengths z_i. The variance of the prediction from regression given residual error in Equation 3.5, $\mathrm{Var}(\epsilon) = \sigma^2 I$ is;

$$\mathrm{Var}(\hat{f}(z)) = \sigma^2 s(z)' s(z) \tag{3.8}$$

<div align="right">(Fahrmeir et al., 2013)</div>

And the confidence interval at $(1 - \alpha)$ level was derived as a quantity of uncertainty or heterogeneity at the different scales of observation assuming normally distributed errors as in Equation 3.9.

$$\hat{f}(z) \pm z_{1-\alpha/2} \sigma \sqrt{s(z)'s(z)} \tag{3.9}$$

(Fahrmeir et al., 2013)

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution pivoted at each wavelength. However, because of autocorrelation in pixel observations affecting the σ^2 term in Equation 3.8, uncertainty estimates at crown level were substantially downward-biased. To overcome this bias and improve comparability across scales of observation, a bootstrapping approach was used to estimate prediction intervals (James et al., 2013, pages 187-189). Bootstrapping was done by iterating 500 times over the model fitting process elaborated above while randomly selecting with replacement a sample of n = 20 to fit an average reflectance curve. Eventually, the 25th and the 95th percentiles of the distribution of mean reflectance curves were used as thresholds approximating the standard error around the mean reflectance curve including all observations. This way, non-overlapping confidence bands among leaf and crown spectral reflectance per species indicated significant differences among the spectral reflectance.

In addition, research in Article IV utilized a rather popular and relatively simple radiative transfer model to predict contributions of structure and foliar chemical contents to variability observed in leaf and crown spectral reflectance at different scales of observation. PROSAIL is a physical model that couples the leaf optical model PROSPECT-5B (Jacquemoud and Baret, 1990) and the canopy optical model 4SAIL (Jacquemoud et al., 2009). In the PROSAIL model, PROSPECT-5B describes physical processes producing directional-hemispherical reflectances and transmittances at leaf scale and feeds the information into 4SAIL together with soil optical properties, illumination and observation geometry, to model the bidirectional reflectance of the canopy (Berger et al., 2018). A spectral range of 400 - 2500 nm is covered by the model. The following parameters are considered in PROSPECT-5B including, the internal structure of the leaf mesophyll (N), chlorophyll a + b concentration (Cab), leaf water content (Cw), dry matter content (Cm), brown pigments (Cbrown), and total caretonoid content (Car). The interaction between parameters in PROSPECT-5B is modelled into specific absorption coefficients (SAC) for each spectral band giving the model the ability to predict reflectances and absorbances along the 400 - 2500 nm spectral range in 1 nm steps. By coupling PROSPECT-5B and 4SAIL, PROSAIL models bidirectional reflectance of the canopy sequentially utilizing the leaf reflectance and transmittance obtained from PROSPECT-5B along with other parameters like leaf area index (LAI), leaf inclination distribution function (LIDF), viewing geometries e.g. sun and sensor(observer) zenith angles (SZA, OZA), hotspot effect resulting from the ratio of leaf width to canopy height (Hot), the fraction of diffuse incident solar radiation (skyl), and the soil reflectance factor (α_{soil}) used to mimic moisture-induced reflectance changes of the upper soil layer (ρ_{soil}) (Berger et al., 2018). In total PROSAIL uses about 16 parameters defining leaf structure, leaf pigment and water content, canopy architecture, soil background, hot spot, solar diffusivity as well as observation geometry, to model the bidirectional reflectance of the canopy.

To predict contributions of structure and foliar chemical contents to the variation observed

in leaf and crown spectral reflectance, the PROSAIL model was inverted at both in-situ crown and airborne crown levels to obtain leaf biochemical and biophysical properties (i.e. N, Cab, Cw, Cm, Cbrown, Car) and LAI as a descriptor of crown structure. The remaining parameters in the PROSAIL model like LIDF, SZA, OZA, Hot, skyl, α_{soil}, and ρ_{soil} were set to field observed values or to default model values. Similarly, PROSPECT-5B was inverted at leaf level to estimate leaf biochemical and biophysical properties. PROSPECT-5B and PROSAIL were inverted at leaf and crown levels independently as dominant processes changed across levels. Model sensitivity analyses were used to predict contributions of structure and foliar chemical contents to observed variability in leaf and crown spectral reflectance thus highlighting the dominant processes influencing observed leaf and crown spectral reflectance and the models' abilities to accurately retrieve structure and foliar chemical contents at the different scales of observation as an indicator of their scaling potential in a vertically heterogeneous forest canopy.

3.6 Managing heterogeneity

3.6.1 Managing pixel heterogeneity through Object-Based Image Analysis (OBIA)

As too much heterogeneity is hardly interpreted because of a low signal to noise ratio, Article III utilized OBIA to reduce the high inter-pixel variation in high-resolution remote sensing data with the aim to improve the distinction of two spectrally and structurally similar *Pinus* tree species. The tree crown-approximate objects were segmented from a LiDAR canopy height model (CHM) in an iterative and rule-based procedure that involved monitoring of various extents around tree tops and amalgamation of neighborhood pixels based on statistics. The procedure considered the following five steps, CHM smoothing, segmentation, classification of canopy areas, iterative merging of segmented canopy objects into individual crown extents, and export of individual crown- and treetop-approximate objects for further analyses (González-Ferreiro et al., 2013). While following these steps, smoothing served to remove data "pits" in the CHM (Khosravipour et al., 2014) equivalent to abrupt low canopy height values compared to the local average canopy height that would introduce noise in the merging process. A chess-board segmentation algorithm was used to segment the smoothed CHM into image objects the size of a single pixel. In the next step, gaps were identified in the CHM as all objects with height <1.5m. Next, treetops were identified from crown objects using a moving window approach that employed a window equivalent to the average crown width observed in the field. And finally, in an iterative process, the treetops were used as markers for a watershed segmentation algorithm to identify crown-approximate objects. The morphology of the identified crown-approximate objects was cleaned using specific shape decision criteria that reclassified "good" and "bad" crowns at each iteration. "Good" crowns were exported for further analyses whereas the "bad" crowns were reiterated into the watershed algorithm until all objects were considered "well segmented" according to (González-Ferreiro et al., 2013). The crown-approximate objects derived from the OBIA procedure served as support to aggregate individual crown traits from various remote sensing data and ultimately reduce within-crown pixel heterogeneity.

3.6.2 Managing methodological heterogeneity through ensemble classification

The previous sub-section 3.6.1 highlights how OBIA was employed in Article III to overcome the higher intra- than inter-specific variation in crown spectral and structural features observed by remote sensing, however, a convenient multivariate form of a classification model was as well not known a priori. A consideration of different classifiers introduced a form of "methodological heterogeneity". Article III tests the utility of an ensemble approach in two ways: (1) the ability of the approach to extract as much information as possible from the remote sensing data to distinguish two spectrally and structurally similar *Pinus* species and (2) the ability of the approach to reduce methodological heterogeneity and maximize generalization capability across the applied models. Selected metrics from both airborne LiDAR and Worldview-2 data sets as shown in Tables 3.2 and 3.3 were used as predictors (Article III). Figure 3.11 shows the general workflow of the ensemble modelling approach. Essentially, individual models were trained and applied independently and then later assembled into a single prediction considered optimal to generalize.
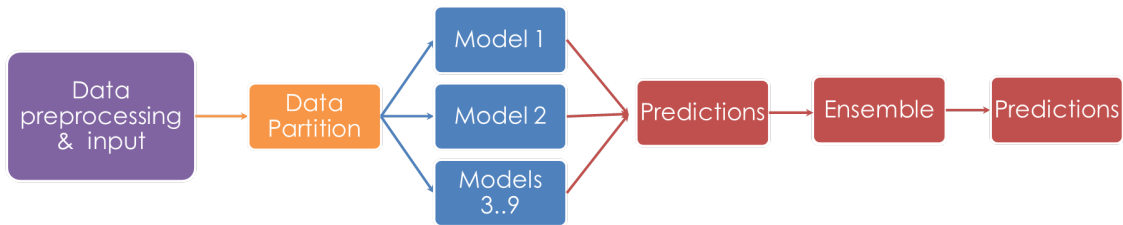
Figure 3.11: Schematic overview of the ensemble modelling approach.

The argument for an ensemble classification approach stemmed from the fact that a difference in models (parametric, semi- or non-parametric) and variables (dependent on a variable selection procedure and datasets), may affect the quality of the resultant species distribution maps because: (1) models fit data to varying degrees (see for example, Figure 3.12) and (2) a covariate effect is partially influenced by effects of other variables in the predictor set. In total nine regression techniques of logistic form were trained according to Equation 3.10;

$$\log\left(\frac{\mu(z)}{1 - \mu(z)}\right) = \hat{f}(z), \tag{3.10}$$

where $\hat{f}$ consisted: (1) Artificial Neural Networks (ANN), (2) Classification and Regression Trees (CTA), (3) Flexible Discriminant Analysis (FDA), (4) Generalized Additive Models (GAM), (5) Generalized Boosting Models (GBM), (6) Generalized Linear Models (GLM), (7) Multivariate Adaptive Regression Splines, (8) Maximum Entropy (MAXENT.Phillips) and (9) Random Forest (RF) and $\mu(z) = \Pr(y = 1|z)$. AIC and BIC (Fahrmeir et al., 2013) were used to select models by summing prediction squared errors to compromise between good fit and model complexity using step-wise linear regression approaches. AIC was defined according to Fahrmeir et al. (2013,

Page 664) for log-likelihood inference with p-dimensional parameter vector $\theta = (\theta_1, ..., \theta_p)'$ as;

$$\text{AIC} = -2l(\hat{\theta}) + 2p \tag{3.11}$$

where $l(\hat{\theta})$ is the log-likelihood and the term $2p$ penalizes for overly complex models. Models were selected on the basis of the smallest AIC value. Alternatively, a stronger penalty for model complexity given the p parameter dimensions was considered in BIC with the aim to even further minimize model complexity as follows;

$$\text{BIC} = -2l(\hat{\theta}) + \log(n)p \tag{3.12}$$

It was thought that the dependence on linear regression to select predictors to all individual models could be disadvantageous for the non-parametric approaches, and thus the consideration of another variable selection technique, the Area Under Curve Random Forest with cross validation, AUCRFcv (Calle et al., 2011). In this technique, an initial Random Forest model was built including all predictors and using the standard permutation procedure (Out of bag - OOB) of the Random Forest model, each variable ranked by the residual sum of squares. Variables were then reduced until a specified number using predictive accuracy computed as the area under the relative/reciever operating characteristic (ROC) curve also known as the Area Under Curve (AUC) statistic. Further detail is given below on the general computation of ROC and the resultant AUC.

To allow for comparison of variable importance across models, variable importance was determined from model predictions in a procedure elaborated in Article III section 2.8 on statistical models and calibration. Models were evaluated using ROC, with both cross validation and split sample techniques. ROC evaluated model accuracy. It described the nature of the probability curve that resulted from mapping model sensitivity (TPR, Equation 3.13) to the corresponding proportion or rate of false positives (FPR, Equation 3.15), when using all possible thresholds to classify the logistic response in Equation 3.10 into a confusion matrix.

$$\text{Sensitivity / TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.13}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.14}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.15}$$

Whereby TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. The resultant area under the curve quantified how good a model distinguished between distributions of the two species. Therefore, ROC = 0.5 meant no separability, and ROC = 1 indicated perfect separability between the species classes. Map quality was evaluated using the true skills statistic (TSS) as it is known to correct for the dependence of Cohen's κ statistic on class prevalence in validation data (Allouche et al., 2006). TSS compared all correct predictions (Equations 3.13 and 3.14), minus those predictions attributable to random
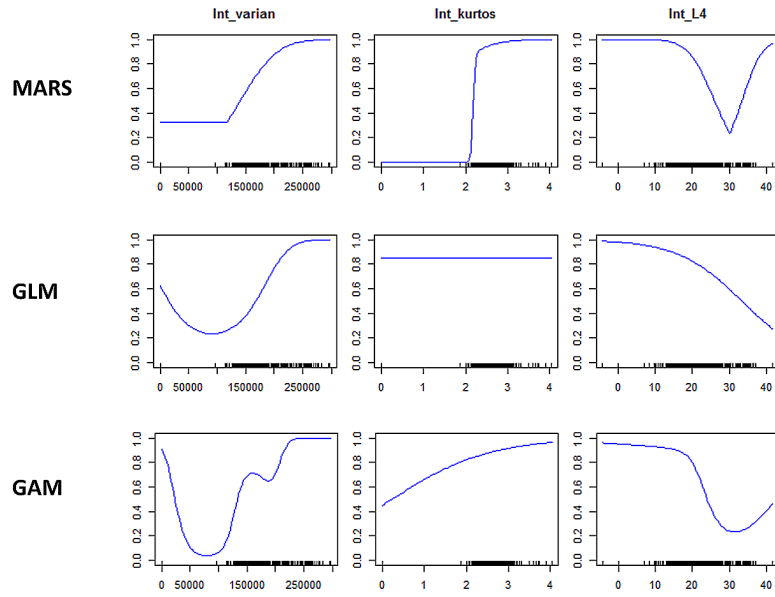
Figure 3.12: Response curves of three intensity variables across three models. The y-axis shows the probability range of the predictions. The x-axis shows the data range of the covariates. The models had a total of eight predictors and these three are chosen for illustration purposes only. The individual effects (partial effects) were computed by holding other covariates in the models at their median values. The key message here is,7 depending on a model or covariate set chosen, a covariate may or may not have a significant effect for the classification of the species. MARS stands for Multivariate Adaptive Regression Splines, GLM, Generalized Linear Models, GAM, Generalized Additive Models. Similarly, Int_varian, Int_kurtos and Int_L4 stand for the variance, kurtosis and L-moments of the vertical distribution of intensity values reflected by canopy material.

chance as shown in Equation 3.16;

$$\frac{\text{TP} * \text{TN} - \text{FP} * \text{FN}}{(\text{TP} + \text{FN})(\text{FP} + \text{TN})} = \text{Sensitivity} + \text{Specificity} - 1 \qquad (3.16)$$

Once individual models were evaluated, ensembles were computed by combining probabilities of occurrence as predicted from Equation 3.10 across all considered modelling techniques via the mean, coefficient of variation, superior and inferior confidence intervals, median, committee average and weighted mean statistics. The coefficient of variation was computed as sd/mean of probabilities over the models. In addition, the superior and inferior confidence intervals around the mean probability were computed at α level of confidence as follows;

$$\text{I}_c = \left[\bar{\text{x}} - \frac{\text{t}_\alpha \text{sd}}{\sqrt{\text{n}}} ; \bar{\text{x}} + \frac{\text{t}_\alpha \text{sd}}{\sqrt{\text{n}}} \right] \qquad (3.17)$$

where $\bar{\text{x}}$ is the mean probability of occurrence of a species across n models selected into the final ensemble given a specified ROC threshold $= 0.8$ to select only high performing models, t is the value from the t distribution and sd is the standard deviation across probabilities of occurrence. The committee average score was the average of binary predictions built on the analogy of a simple vote. In this case, each model voted for the species being either present or absent. For each site, the sum of ones was then divided by the number of models included in

the ensemble. For the weighted mean ensemble, a weighted sum of probabilities of occurrences across models was computed proportional to their ROC scores. Finally, cross validation and split-sample procedures were compared to evaluate performance of the ensemble models. Models were run 50 times in order to monitor variation in variable importance and prediction accuracies.

3.7 Determining scale dependence and scale domains

3.7.1 Determining scale dependence using linear models for fixed effects

Section 2.7 of Article I elaborates a linear mixed model used to monitor distributions of GC, SCI and ESCI to arrive at a robust threshold plot size for stable estimation of forest stand structural heterogeneity at a one hectare plot size. The method offers the advantage of observing more than the estimated mean stand structural heterogeneity across plot sizes while simultaneously monitoring differences in scale effects across stands of different management types. The technique considered grouping primarily based on parts of the index distribution (percentiles, Figure 3.10) and eventually included secondary grouping by forest management type, thus allowing for simultaneous estimation of scale effects across forests of differing management type. To avoid redundancy, the reader is referred to Article I section 2.7 for all required detail on how scale dependence was estimated using a linear mixed model approach.

3.7.2 Determining scale dependence from geo-statistics and optimization of observation design

Spatial autocorrelation directly links to independence of data (a basic tenet of statistics) and thus determines the estimated variation in probability samples. In Article II, the Moran's I index was used to determine spatial autocorrelation in simulated forest landscapes (Figure 3.4). Spatial autocorrelation was used to highlight the range of scale dependence in each landscape ultimately informing efficiency of the tested observation designs. Moran's I was calculated as a correlation coefficient between image pixels within a defined local neighborhood. The local neighbourhood was defined using moving windows of different sizes across cells of each artificial landscape to produce spatial correlograms. Window sizes ranged from 3 x 3 pixels to 100 x 100 pixels i.e. 30 x 30 m to 3 x 3 km. Moran's I was defined via a weight matrix following Anselin (1995) as shown in Equation 3.18, and yielded values between -1 for perfect clustering with dissimilar pixels in the neighborhood, 1 for perfect clustering with similar pixels in the neighborhood and 0 for perfect randomness among pixels in the neighborhood relative to the focal pixel.

$$\text{Local Moran}'\text{s I} = \frac{1}{s^2} \frac{\sum_i \sum_j (y_i - \bar{y})(y_j - \bar{y})}{\sum_i \sum_j w_{ij}} \qquad (3.18)$$

where $\bar{y} = \sum_{i=1}^{n} y_i / n$ is the mean across observations in the focal window, $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / n$ the sample variance and w_{ij} the weight matrix. The Global Moran's I for each landscape was computed as an average of Local Moran's I over all pixels in each artificial landscape. Finally, spatial correlograms were derived as Global Moran's I versus window size. The considered

window sizes were equivalent to employed cluster extents as described in Article II for cluster designs shown in Figures 3.8, 3.9. To assess scale-dependence, resolution was optimized through changing the number of subplots per cluster whereas the extent was varied through changing the distances between the subplots (Article II).

3.7.3 Determining scale dependence in relationships among spatially structured variables

To evaluate scale dependence in relationships among spatially structured variables, Spearman's correlations were used to quantify relationships between forest stand structural heterogeneity and airborne LiDAR preditors (Article I). The bivariate correlations between the response and each predictor at different scales of observation were indicators of differences in magnitudes of predictor effects. So as to study the effect of plot size on the nature of relationships between indices of forest stand structural heterogeneity and airborne predictors, Random Forest models were optimized across scales of observation as further elaborated in Section 2.9 of Article I.

3.8 Scaling

3.8.1 Scaling across hierarchies using regression models

Two case studies in this research scaled across hierarchies using regression models (Figure 3.2). Article I employed Random Forest regression models to map indices of forest stand structural heterogeneity at landscape scale from field observations at a stand scale and wall-to-wall airborne LiDAR data. Similarly, Article III utilized ensemble modelling techniques to classify tree species in a landscape from single crown field observations, multispectral satellite imagery and airborne LiDAR data. It should be noted however that regression models can be used to combine information at different scales in an optimal way for the purpose of predicting a target attribute observed at yet another scale. But strictly speaking, this may not be categorized as scaling but remains model-fitting when there are no attempts to harmonize the scales of individual predictors with the scale of the dependent variable.

3.8.2 Scaling using the global sampling grid (GSG)

The GSG presented in Article II is a sampling design with interesting scaling capabilities. First, the GSG addressed the problem of how to lay a finite population of points on the surface of a spherical approximation of the earth without having systematic over- or under-representation of parts of the earth's surface when projected on a plane, and second, addressed how to maintain a hierarchical relationship between grids at different resolutions that enables scaling across samples by combining datasets collected from different grid resolutions. At the same time, the GSG was constructed in such a way that it tackled these two scaling challenges without losing similarity to systematic sampling design and therefore maintaining the following design properties; (1) proportional representation of classes in the population, (2) homogeneity in

sampling intensity across the land area of the earth surface, (3) total coverage of any land area of interest on the earth's surface, (4) a defined minimum (spherical) distance between sampled locations in both x- and y-directions, (5) a higher precision compared to simple random sampling, and lastly (6) simplicity in construction and communication. Figure 3.13 illustrates construction of the GSG, further details on construction of the GSG can be found in Article II.
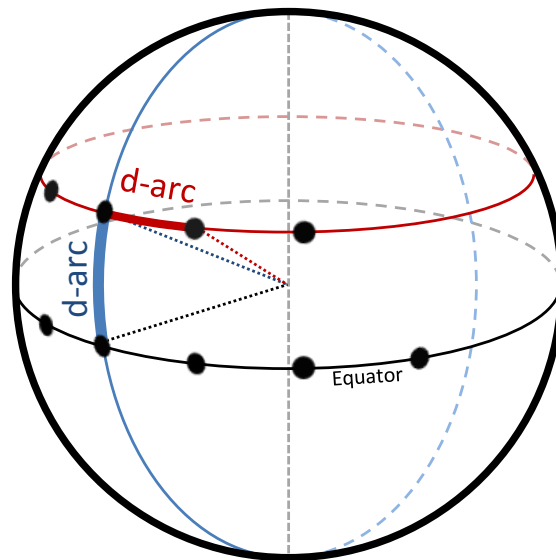
Figure 3.13: Construction of the GSG. Circles of latitudes are placed with a constant distance on the surface of a sphere in north and south direction starting from the equator. Points are placed in equal distance along latitudes in West and East direction starting from the null meridian (d-arc = spherical distance) - (Fehrmann et al., 2019).

Chapter 4

Results

4.1 Scale effects in estimates of forest stand structural heterogeneity

Figure 4.1 shows higher variability in index estimates when smaller plot sizes were used to compute indices of forest stand structural heterogeneity across studied forest stand structural types. The variance in index estimates was higher in selectively logged and unmanaged forests than in age class forests. Similarly, for all structural indices and for all parts of the index distributions, there was a bias in mean estimates of forest stand structural heterogeneity compared to the reference estimate (at one hectare), when smaller plot sizes were employed (Figures 4.1, 4.2, 4.3 and 4.4 for GC, SCI and ESCI respectively).

The influence of plot size (scale effect) varied across levels/magnitudes of stand structural heterogeneity with higher effects when plots fell in locations significantly different from the average heterogeneity on the plot (see Figure 5, Article I) and was similar across forest management systems i.e. consistently lowest in age class forest and highest in selection systems and unmanaged forests. Scale effects were similar across studied indices and were either positive or negative relative to whether the smaller plots fell in locations of lower or higher heterogeneity than the average heterogeneity on the entire plot (Figures 4.2, 4.3 and 4.4 for GC, SCI and ESCI respectively - whereby, the percentiles are the levels/magnitudes of heterogeneity in each index distribution). For easy comparison across the indices, see Figure 6 in Article I. Increasing plot size increased the magnitude of structural heterogeneity when plots were centered in locations of lower heterogeneity compared to their immediate vicinity (such as observed in the minimum percentiles of the distributions - Figures 4.2, 4.3 and 4.4 for GC, SCI and ESCI respectively), and vice versa when plots were centered in locations of higher heterogeneity than their immediate vicinity (as observed in the maximum percentiles of the index distributions - Figures 4.2, 4.3 and 4.4 for GC, SCI and ESCI respectively). The scale effects were lowest in percentiles around the mean depending on the index observed (GC = 75th, SCI = 25th, ESCI = 50th).
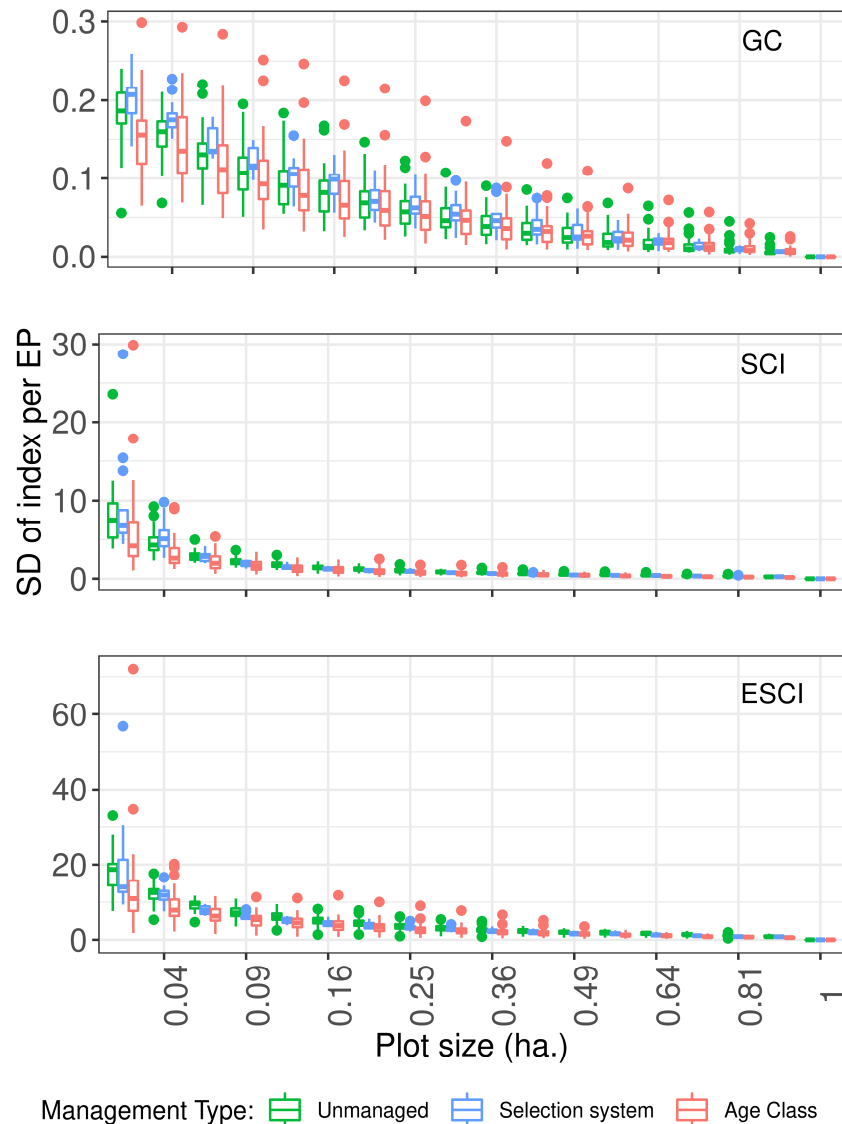
Figure 4.1: Standard deviation (SD) of index values per experimental plot (EP) across plot size. From top to bottom: GC, SCI and ESCI.

4.2 Plot size thresholds to estimate forest stand structural heterogeneity

Figures 4.5, 4.6 and 4.7 show stabilization of index values on the reference estimate for GC, SCI and ESCI across forest management types. As mentioned in the methods section, the point of index stabilization on the reference estimate was identified as a range between observed plot sizes, when the second derivative of the function describing the structure-area-curves was equivalent to zero (Figure 4 in Article I). Stabilization in index estimates depended on the region of the distribution assessed. Regions with low scale effects stabilized faster than regions with higher scale effects. Plot size thresholds ranged between 900 m^2 and 2500 m^2 depending on the region of the distribution assessed, with stable estimation of all indices across all regions
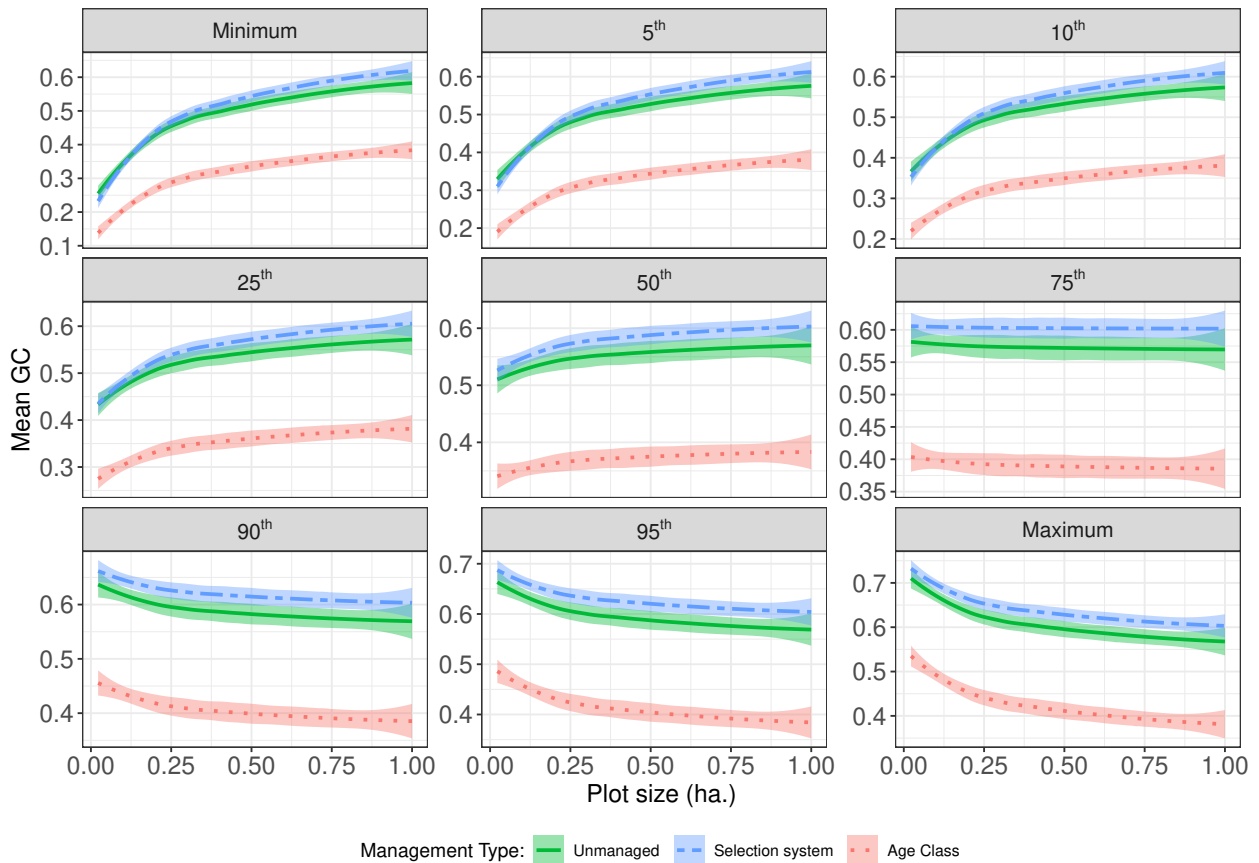
Figure 4.2: Mean per management system of the predicted values of structural heterogeneity (GC) against plot size. The different regions of the index distributions are described in percentiles. Generalized additive models were used to estimate mean index values across structure-area-curves grouped first by percentiles of the index distribution and second by the forest management type.

of the distribution possible when plots were $>=2500$ m^2. Plot size thresholds were similar across forest management types.

4.3 Scale dependence in relationships between forest stand structural heterogeneity and airborne LiDAR derivatives

Table 4.1 shows Spearman's correlations between individual LiDAR predictors and GC, SCI, ESCI. The scale of observation affected the magnitude of relationships between the indices and the predictors resulting in lower correlations at smaller plot sizes. Cover (B_{7m} and A_{7m}), dispersion (CV, SD, Var, AAD, L.CV), low height percentiles (P_1, P_{05}, P_{10}) had higher relationships with GC, SCI and ESCI compared to other LiDAR metrics. In regard to the derived random forest models (Figures 7 and 8 in Article I), similar cover metrics were consistently important predictors across indices and scales. Figures 7 and 8 in Article I show that dispersion metrics were important predictors for all the three indices but ranked comparatively higher in GC models whereas lower height percentiles were important at all observed scales. However, higher height
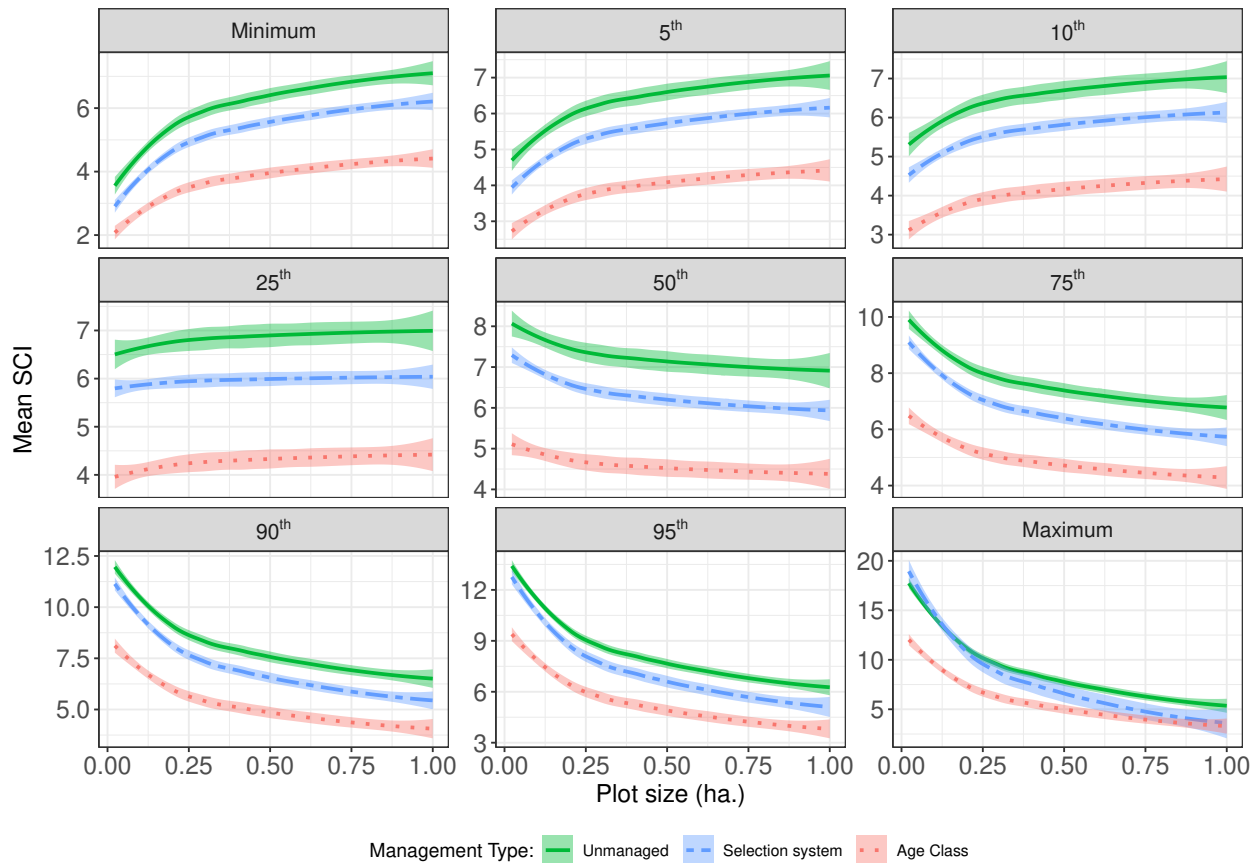
Figure 4.3: Mean per management system of the predicted values of structural heterogeneity (SCI) against plot size. The different regions of the index distributions are described in percentiles. Generalized additive models were used to estimate mean index values across structure-area-curves grouped first by percentiles of the index distribution and second by the forest management type.

percentiles (P_{99}, Max) were included into models only at the 20 m plot size. In addition, Figures 7 and 8 in Article I show that models and their respective predictive power changed with changing the scale of observation.

Table 4.1: Spearman's correlation values between individual LiDAR predictors and individual structural heterogeneity indices at varying scales of observation. Full variable names in Table 3.4.

	ESCI			SCI			GC		
LiDAR metric	100 m	50 m	20 m	100 m	50 m	20 m	100 m	50 m	20 m
A_{7m}	0.74	0.65	0.40	0.59	0.54	0.32	0.40	0.36	0.32
B_{7m}	0.70	0.63	0.41	0.64	0.58	0.37	0.44	0.43	0.36
P_{01}	0.67	0.57	0.37	0.62	0.55	0.34	0.51	0.45	0.34
P_{05}	0.65	0.51	0.33	0.60	0.50	0.30	0.41	0.33	0.26
CV	-0.57	-0.49	-0.33	-0.50	-0.44	-0.28	-0.25	-0.23	-0.22

Table 4.1: Spearman's correlation values between individual LiDAR predictors and individual structural heterogeneity indices at varying scales of observation. Full variable names in Table 3.4.

LiDAR metric	ESCI			SCI			GC		
	100 m	50 m	20 m	100 m	50 m	20 m	100 m	50 m	20 m
SD	-0.55	-0.49	-0.30	-0.35	-0.31	-0.18	-0.22	-0.22	-0.21
Var	-0.55	-0.49	-0.30	-0.35	-0.31	-0.18	-0.22	-0.22	-0.21
P_{10}	0.51	0.46	0.30	0.54	0.51	0.32	0.28	0.26	0.20
L.CV	-0.51	-0.45	-0.30	-0.46	-0.42	-0.27	-0.19	-0.18	-0.17
AAD	-0.49	-0.44	-0.29	-0.35	-0.30	-0.19	-0.15	-0.16	-0.16
C_{7m}	0.48	0.39	0.26	0.35	0.31	0.20	0.14	0.17	0.17
L_2	-0.47	-0.43	-0.28	-0.31	-0.28	-0.18	-0.12	-0.14	-0.15
Mode	0.45	0.40	0.26	0.58	0.54	0.33	0.33	0.28	0.21
Kurt	0.45	0.40	0.27	0.43	0.38	0.24	0.14	0.13	0.12
P_{20}	0.44	0.39	0.27	0.57	0.51	0.34	0.23	0.20	0.16
$C1_{7m}$	0.44	0.37	0.22	0.39	0.33	0.17	0.12	0.12	0.10
Mean	0.43	0.38	0.28	0.64	0.56	0.39	0.33	0.26	0.21
L_1	0.43	0.38	0.28	0.64	0.56	0.39	0.33	0.26	0.21
P_{25}	0.42	0.37	0.26	0.57	0.51	0.34	0.24	0.19	0.16
P_{30}	0.39	0.35	0.25	0.57	0.51	0.34	0.24	0.18	0.15
A_{mean}	0.38	0.41	0.26	0.45	0.42	0.25	0.11	0.11	0.11
P_{99}	0.37	0.26	0.17	0.65	0.53	0.35	0.51	0.36	0.22
QRT	0.37	0.33	0.25	0.61	0.55	0.38	0.32	0.24	0.19
C_{mode}	-0.37	-0.41	-0.22	-0.42	-0.45	-0.24	-0.30	-0.28	-0.20
$C1_{mean}$	0.37	0.33	0.20	0.34	0.30	0.16	0.03	0.06	0.06
$(C_{mode}/C1)*100$	-0.36	-0.38	-0.23	-0.39	-0.42	-0.24	-0.17	-0.19	-0.15
$C1_{0m}$	0.36	0.28	0.16	0.30	0.24	0.12	0.03	0.05	0.04
CUB	0.35	0.30	0.23	0.61	0.54	0.37	0.34	0.25	0.18
P_{40}	0.34	0.30	0.24	0.56	0.50	0.35	0.24	0.17	0.15
CRR	0.33	0.33	0.24	0.45	0.41	0.27	0.19	0.12	0.11
B_{mode}	-0.33	-0.36	-0.21	-0.34	-0.37	-0.22	-0.18	-0.17	-0.16
P_{95}	0.32	0.24	0.16	0.62	0.52	0.35	0.45	0.32	0.21
D_{mode}	-0.32	-0.32	-0.16	-0.31	-0.33	-0.17	-0.19	-0.16	-0.13
MAD_{mode}	-0.32	-0.27	-0.15	-0.24	-0.20	-0.09	-0.03	-0.01	-0.03

Continued on next page

Table 4.1: Spearman's correlation values between individual LiDAR predictors and individual structural heterogeneity indices at varying scales of observation. Full variable names in Table 3.4.

LiDAR metric	ESCI			SCI			GC		
	100 m	50 m	20 m	100 m	50 m	20 m	100 m	50 m	20 m
P_{50}	0.31	0.27	0.22	0.56	0.50	0.36	0.27	0.19	0.15
Skew	-0.31	-0.29	-0.22	-0.36	-0.32	-0.22	-0.03	-0.03	-0.05
L.kurt	0.31	0.29	0.21	0.34	0.29	0.20	0.03	0.03	0.07
$(C_{mean}/C1) * 100$	-0.30	-0.20	-0.06	-0.23	-0.17	-0.04	-0.06	-0.04	0.00
IQR	-0.30	-0.31	-0.22	-0.20	-0.21	-0.15	0.04	-0.00	-0.07
P_{60}	0.29	0.26	0.21	0.56	0.51	0.36	0.29	0.22	0.16
C_1	0.29	0.22	0.12	0.23	0.19	0.09	-0.04	-0.00	-0.00
P_{90}	0.28	0.23	0.16	0.58	0.51	0.34	0.39	0.28	0.18
P_{70}	0.27	0.23	0.18	0.56	0.50	0.35	0.31	0.22	0.15
P_{80}	0.26	0.21	0.17	0.55	0.49	0.34	0.33	0.23	0.16
P_{75}	0.26	0.22	0.17	0.55	0.49	0.34	0.31	0.22	0.15
L_3	0.24	0.14	0.00	-0.02	-0.07	-0.12	0.13	0.13	0.06
MAD_{med}	-0.22	-0.20	-0.15	-0.12	-0.10	-0.08	0.16	0.13	0.03
L_4	0.21	0.21	0.17	0.39	0.33	0.24	0.10	0.07	0.09
C_4	-0.21	-0.14	-0.09	-0.21	-0.15	-0.10	-0.09	-0.05	-0.04
Max	0.20	0.22	0.16	0.27	0.42	0.32	0.23	0.32	0.22
C_3	-0.16	-0.10	-0.04	-0.21	-0.15	-0.07	-0.10	-0.03	-0.01
C_5	-0.16	-0.11	-0.06	-0.15	-0.12	-0.09	-0.06	-0.05	-0.03
Min	0.16	0.20	0.22	0.11	0.24	0.21	0.03	0.21	0.23
C_2	0.14	0.09	0.07	-0.01	-0.02	0.00	-0.17	-0.07	-0.01
C_6	-0.11	-0.08	-0.05	-0.11	-0.09	-0.07	-0.01	-0.02	-0.02
C_7	-0.11	-0.02	-0.00	-0.10	-0.05	-0.03	-0.03	-0.01	-0.01
Total	0.09	0.09	0.07	-0.02	0.02	0.01	-0.20	-0.08	-0.01
L.skew	0.08	0.00	-0.09	-0.09	-0.12	-0.15	0.20	0.17	0.06
$(C_{7m}/C1) * 100$	0.07	0.12	0.10	0.07	0.10	0.10	0.26	0.25	0.20
C_{mean}	-0.04	0.06	0.09	-0.01	0.06	0.08	-0.24	-0.12	-0.04
Total	-0.03	-0.04	-0.01	-0.13	-0.11	-0.06	-0.28	-0.16	-0.09
B_{mean}	0.02	0.10	0.15	0.21	0.23	0.20	-0.06	-0.05	0.02
$C1_{mode}$	0.01	0.01	0.02	-0.04	-0.03	-0.02	-0.21	-0.13	-0.08
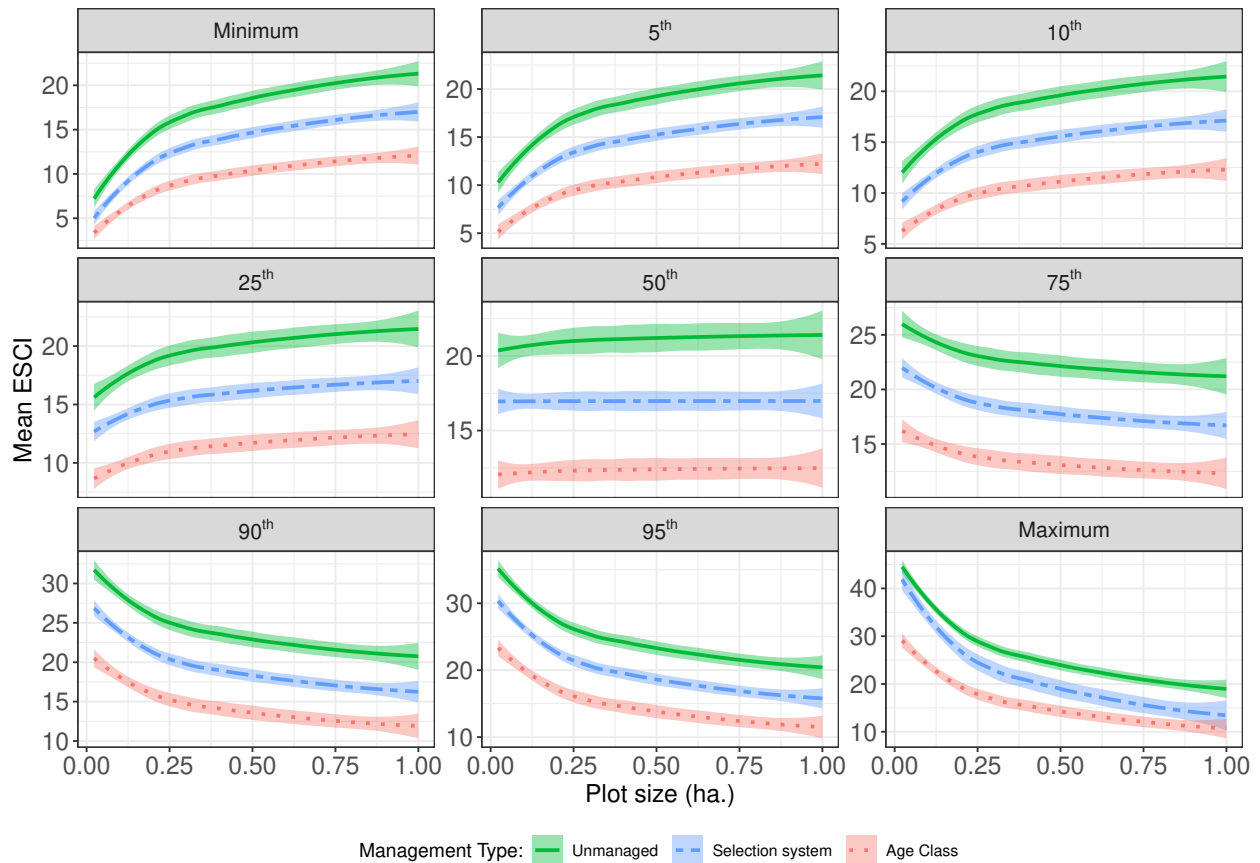
Figure 4.4: Mean per management system of the predicted values of structural heterogeneity (ESCI) against plot size. The different regions of the index distributions are described in percentiles. Generalized additive models were used to estimate mean index values across structure-area-curves grouped first by percentiles of the index distribution and second by the forest management type.

4.4 Scale effects in estimates of forest cover proportion

Results in Article II showed an asymptotic decrease in standard errors of the estimated mean forest cover proportion was observed in all studied landscapes with increasing cluster resolution/size (or number of subplots per cluster - Figure 3.8) and cluster extent (or size of the cluster) as shown in Figures 4.8 and 4.9. Notice that the term "cluster resolution" is here used in the context of the definition of scale as elaborated in section 1.1. Cluster resolution/size had a higher effect on precision of estimation with decreasing spatial autocorrelation whereas cluster extent had a higher effect on precision of estimation with increasing spatial autocorrelation. Generally, the higher the coverage and fragmentation in the landscape, the higher the precision of estimation. Cluster resolution/size and extent had lower effects on precision of estimation when coverage was higher. To see better the effect of coverage on precision of estimation across cluster sizes and designs, the reader should refer to Figures 7 and 8 in Article II where ranges of y-axes are fixed across landscapes to facilitate comparisons. When reading results on the effect of cluster design (Figure 3.9) in estimation of forest cover proportion as presented in Figure 4.9, one should keep in mind that cluster resolution/size and extent were constant

Figure 4.5: Box plots of the second derivative of GC per management system and plot size range. The different regions of the index distributions are described in percentiles. GC stabilized fastest in the 75th percentile of the distribution. A total of 2565 structure-area-curves were considered. The second derivative (y-axis) conveys the behavior of the function describing the curve of the predicted index means. The derivative is positive at a point where the curve is approaching its minimum and negative where the curve is approaching its maximum. The point of index stabilization is when the second derivative is equivalent to zero.

across the considered cluster designs and therefore only total distance between cluster subplots (Figure 3.9) changed. Increasing distance between subplots yielded similar effects on precision across cluster designs in landscapes of high fragmentation. However, as spatial autocorrelation in landscapes increased, clusters with longer total subplot distances (i.e. of design ∟ and ×) yielded more precise estimates compared to designs with shorter total subplot distances (i.e. + and ⊥ shapes).
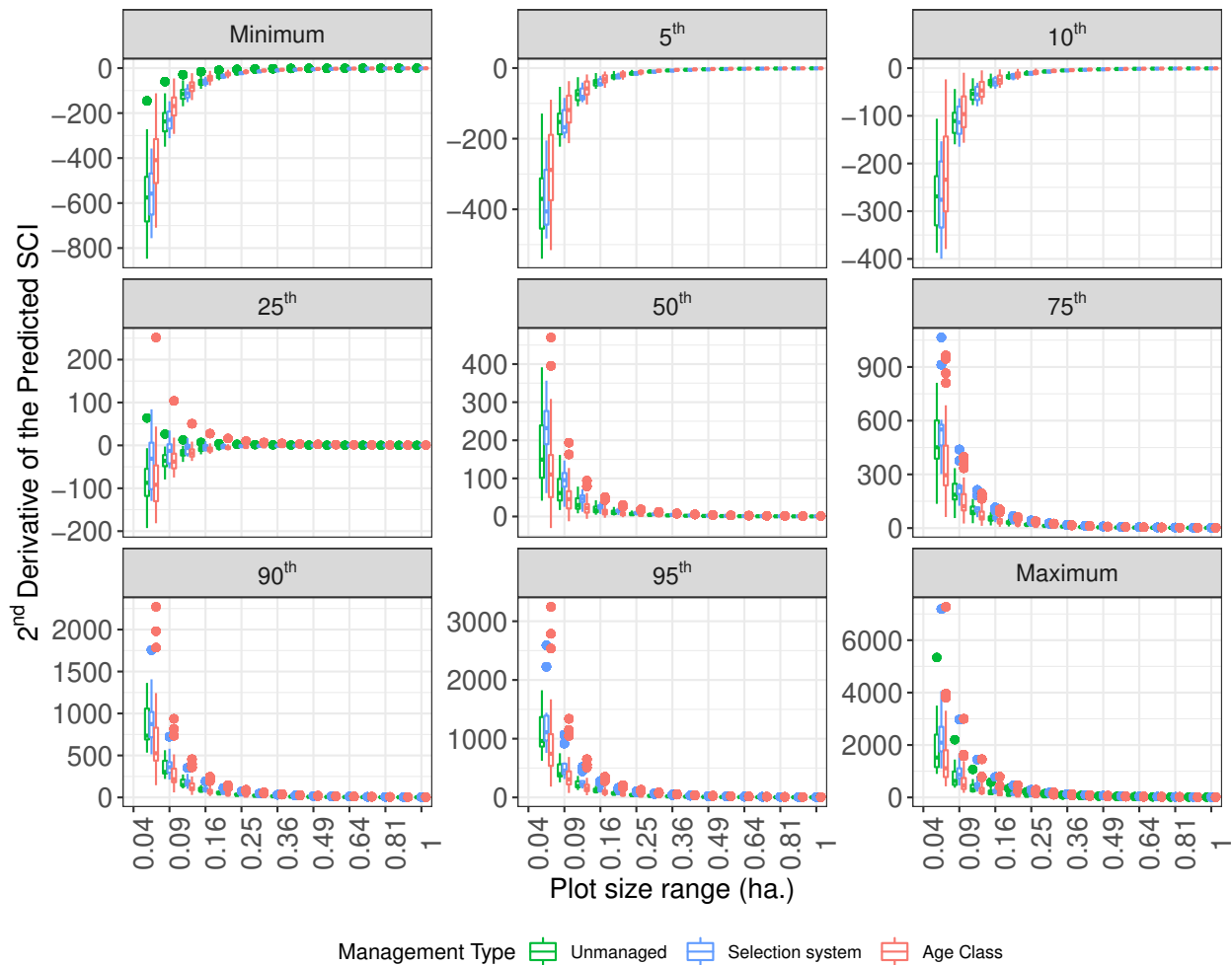
Figure 4.6: Box plots of the second derivative of SCI per management system and plot size range. The different regions of the index distributions are described in percentiles. SCI stabilized fastest in the 25th percentile of the distribution. A total of 2565 structure-area-curves were considered. The second derivative (y-axis) conveys the behavior of the function describing the curve of the predicted index means. The derivative is positive at a point where the curve is approaching its minimum and negative where the curve is approaching its maximum. The point of index stabilization is when the second derivative is equivalent to zero.

4.5 Scale effects in broadleaved tree reflectance at leaf and crown levels

Figure 4.10 shows observed and modelled leaf and crown spectral reflectance. Leaf spectral reflectance were higher than in both in-situ and airborne crown spectral reflectance and had the lowest variation among observations. The highest variation in reflectance was observed in in-situ crown data. There were differences across spectral reflectance curves at both crown scales especially in the near-infrared region of the spectrum ($\approx$ 800 - 1300 nm) where water absorption features (i.e. valleys) were pronounced at airborne crown scale. Airborne crown reflectances were higher than in-situ crown reflectance. There were significant differences in mean spectral reflectance across the entire range of the spectrum as the scale of observation changed from
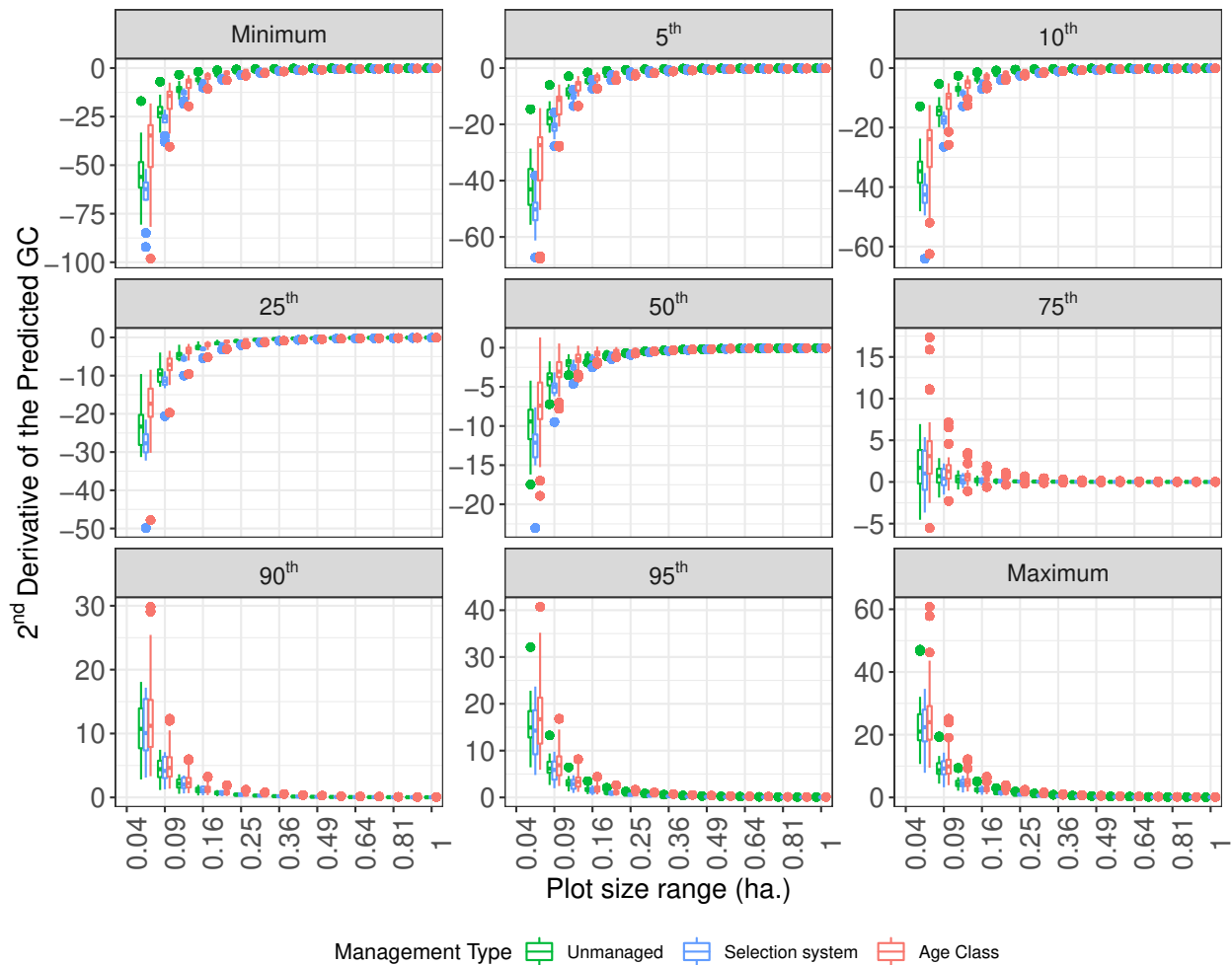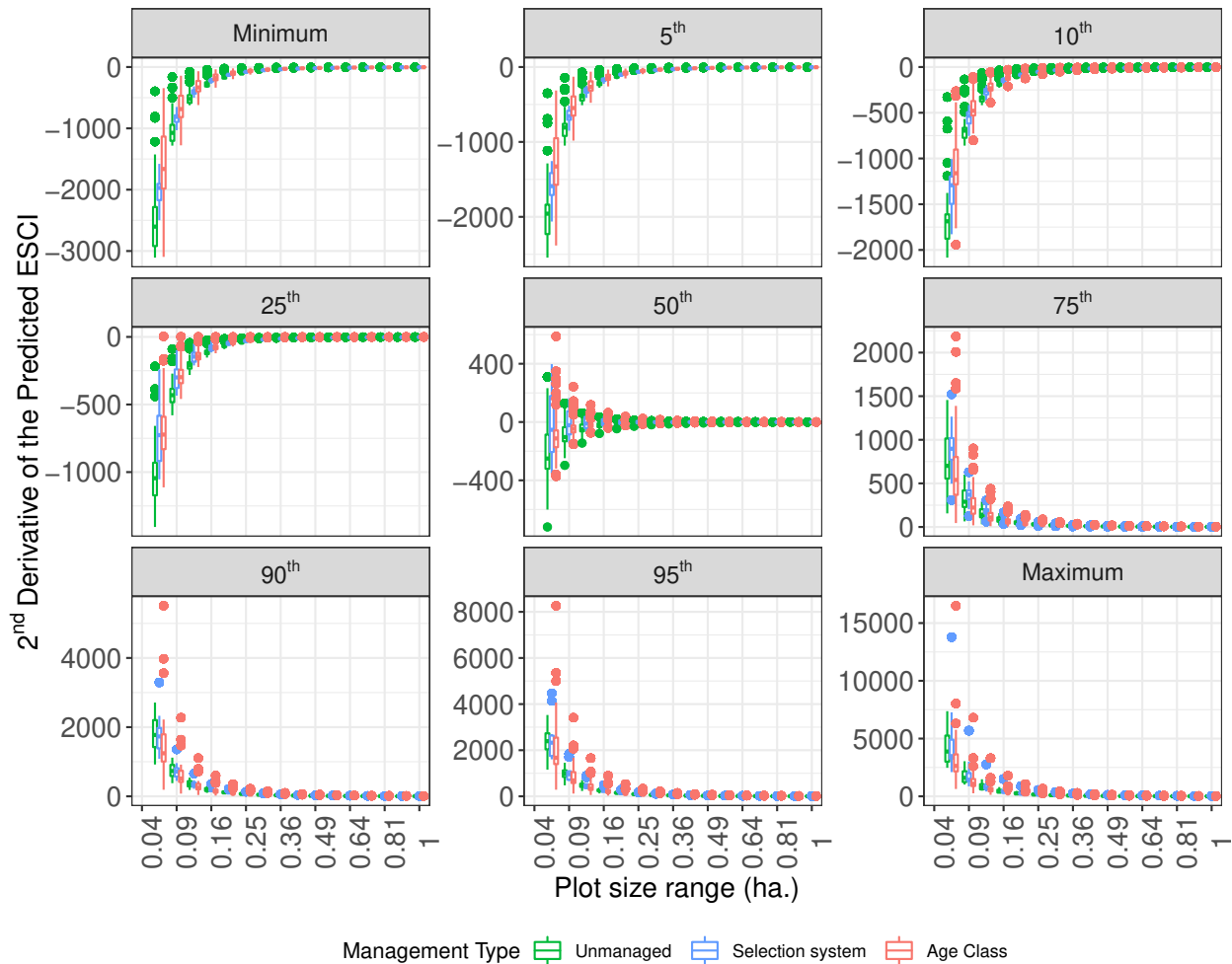
Figure 4.7: Box plots of the second derivative of ESCI per management system and plot size range. The different regions of the index distributions are described in percentiles. ESCI stabilized fastest in the 50th percentile of the distribution. A total of 2565 structure-area-curves were considered. The second derivative (y-axis) conveys the behavior of the function describing the curve of the predicted index means. The derivative is positive at a point where the curve is approaching its minimum and negative where the curve is approaching its maximum. The point of index stabilization is when the second derivative is equivalent to zero.
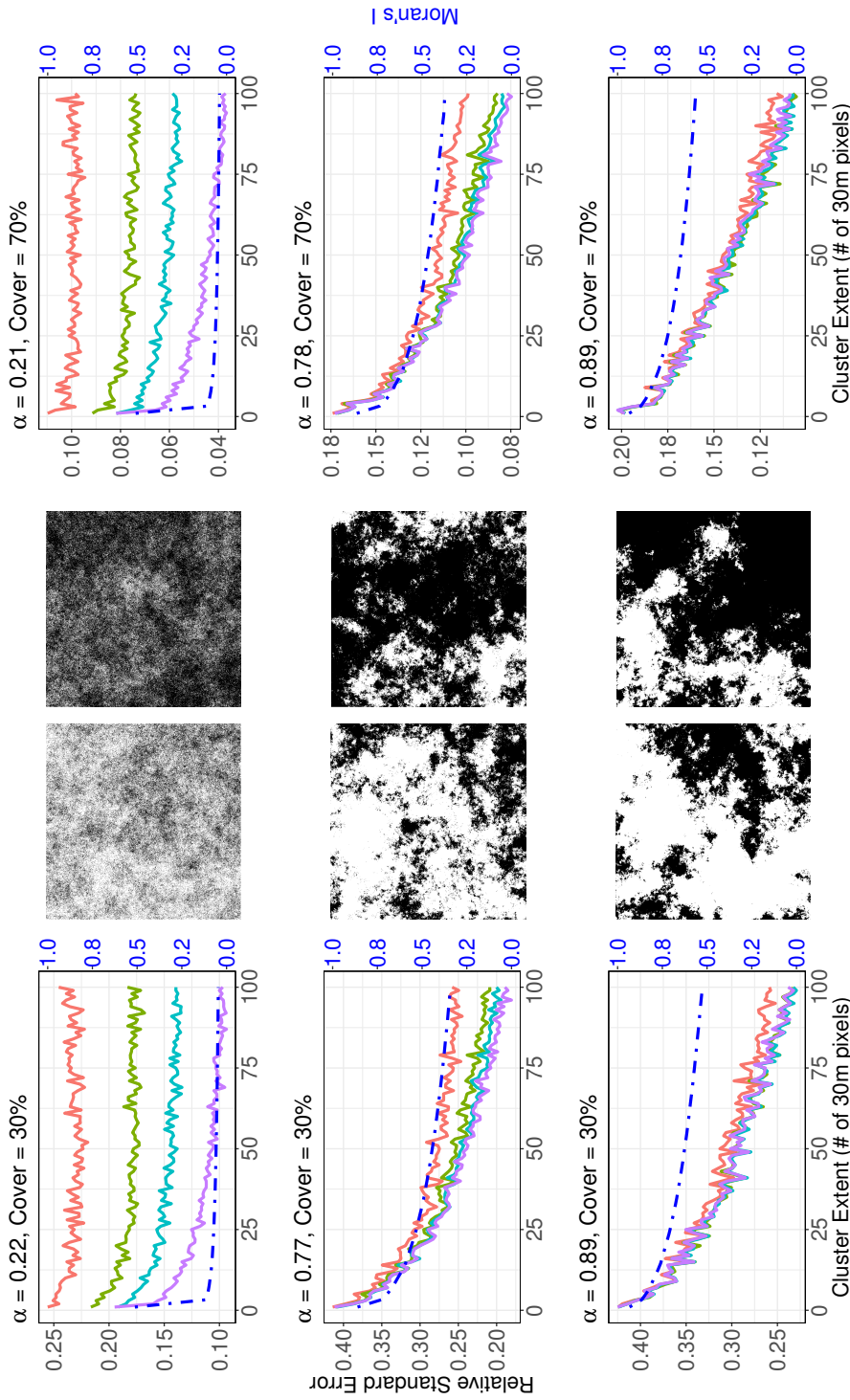
Figure 4.8: The primary y-axis (left) shows relative standard error against cluster extent for each observation design for 5, 9, 16 and 49 subplots and forest cover landscapes (see Figure 3.8 for details on the observation designs). Forest landscapes are placed adjacent to the corresponding graphs. α values in the heading of each graph represent levels of fragmentation in landscapes of 30% (left) and 70% (right) forest cover. The extent of clusters was increased by one pixel from 0 to 100 pixels (x-axis) maintaining the design configurations as shown in Figure 3.8. Therefore, relative subplot distances increased in the same magnitude for each increment in extent. Notice that clusters of 5 subplots have the longest subplot distances and clusters of 49 subplots have the shortest subplot distances. The secondary y-axis (in blue) shows Moran's I computed in moving windows of neighborhoods ranging from 1 to 100 pixels (x-axis). Scales of the y-axes across the considered six landscapes are independent to emphasize differences in considered observation designs per landscape, however, the reader interested in relative differences across landscapes, can refer to Figure 7 in Article II.
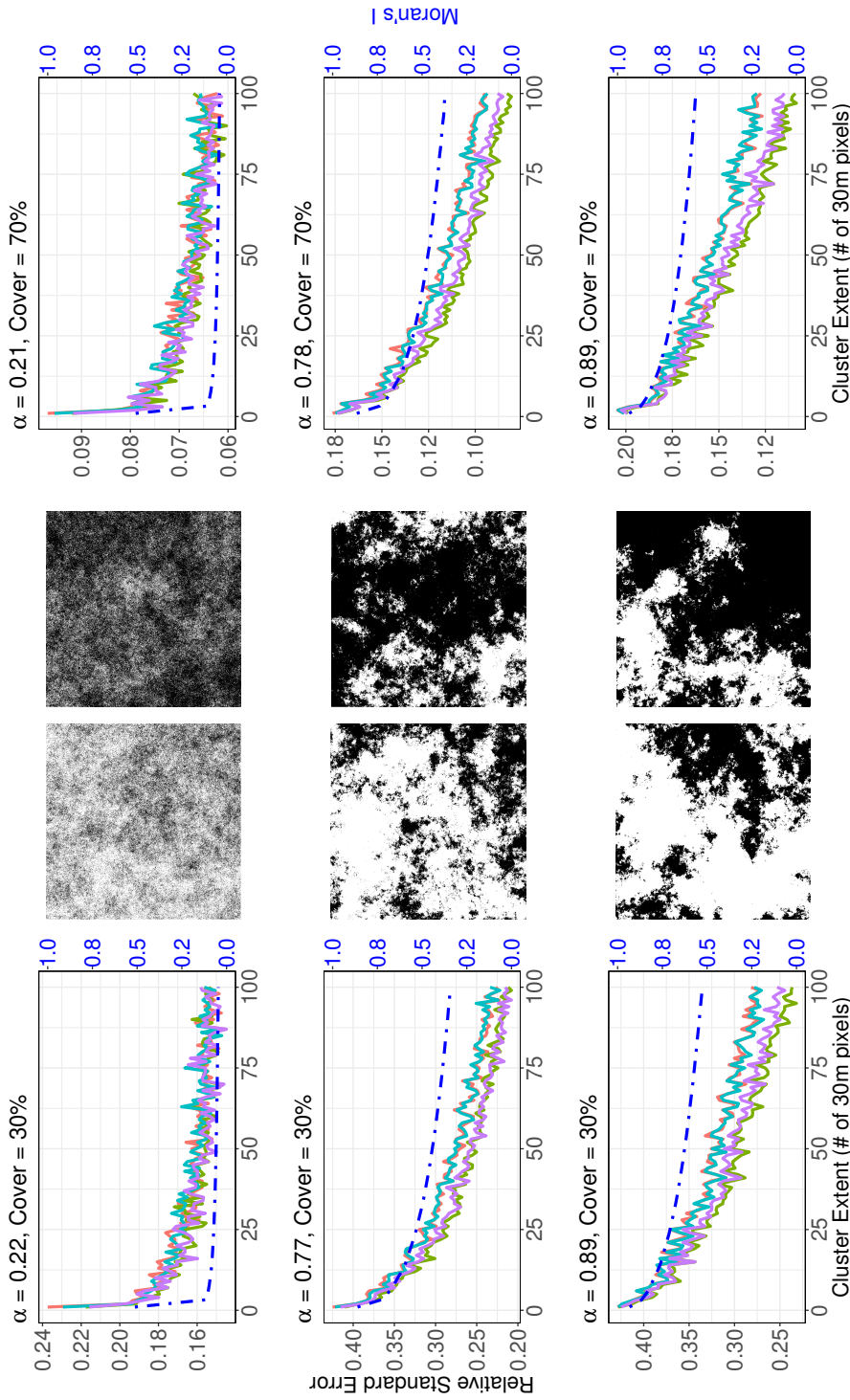
Figure 4.9: The primary y-axis (left) shows relative standard error against cluster extent for each observation design of 13 subplots arranged in L, +, ×, and ⊥ configurations and forest cover landscapes (see Figure 3.9 for details on the observation designs). Forest landscapes are placed adjacent to the corresponding graphs. α values in the heading of each graph represent levels of fragmentation in landscapes of 30% (left) and 70% (right) forest cover. The extent of clusters was increased by one pixel from 0 to 100 pixels (x-axis) maintaining the design configurations as shown in Figure 3.9. Therefore, relative subplot distances increased in the same magnitude for each increment in extent. Notice that clusters of 5 subplots have the longest subplot distances and clusters of 49 subplots have the shortest subplot distances. The secondary y-axis (in blue) shows Moran's I computed in moving windows of neighborhoods ranging from 1 to 100 pixels (x-axis). Scales of the y-axes across the considered six landscapes are independent to emphasize differences in considered observation designs per landscape, however, the reader interested in relative differences across landscapes, can refer to Figure 8 in Article II.

leaves to crowns (Fig. 4.10). The scale effects on reflectances were similar, however, the magnitude of scale effects in spectral reflectance across hierarchies of observation changed depending on the regions of the spectrum observed and were more pronounced when changing from leaf to airborne crown scales as seen from the magnitude of slopes (Fig. 4.11). The scale of observation affected separability among species spectral reflectance and contributing foliar and chemical properties (Article IV).
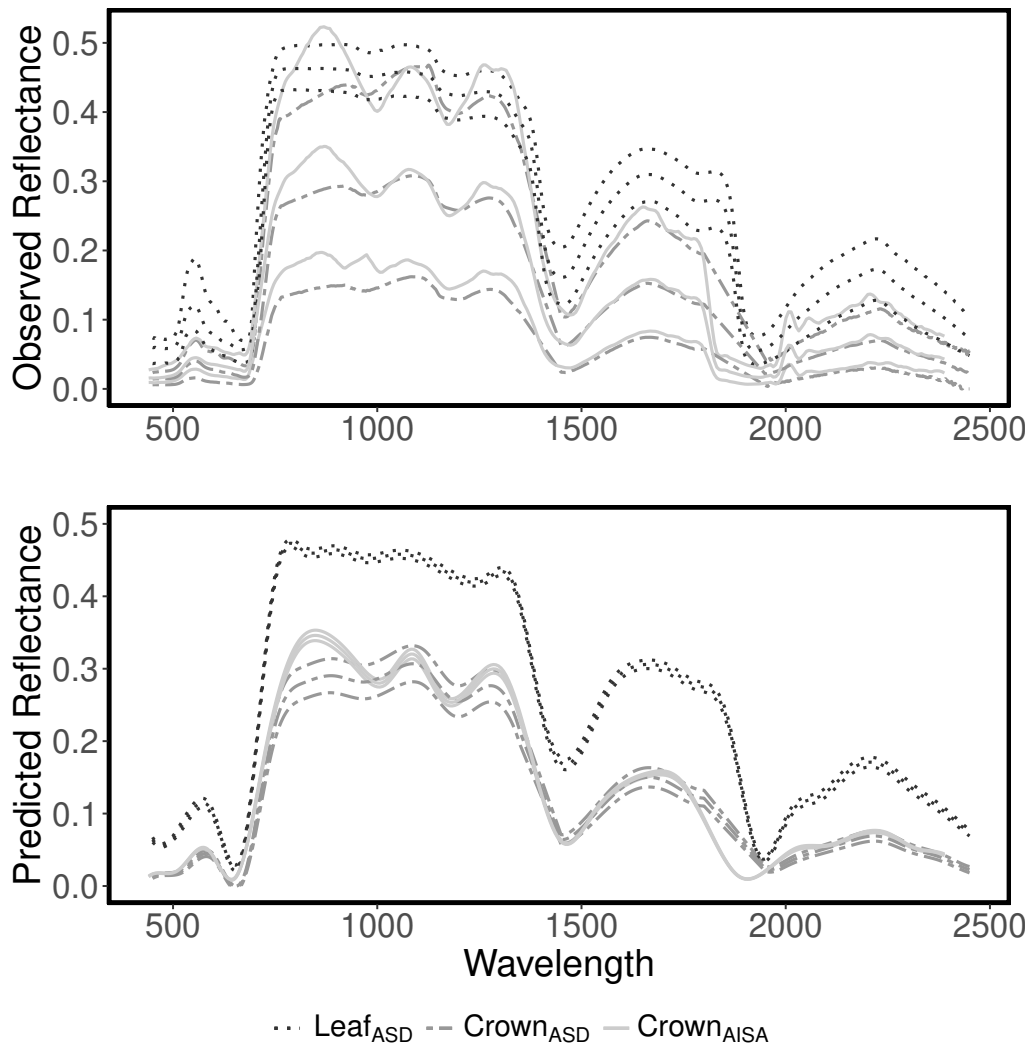
Figure 4.10: **Top:** Means, 5th and 95th percentiles of observed reflectance at $\mathrm{Leaf_{ASD}}$, $\mathrm{Crown_{ASD}}$ and $\mathrm{Crown_{AISA}}$ scales. **Bottom:** Mean modelled reflectance and the associated bootstrapping standard errors at $\mathrm{Leaf_{ASD}}$, $\mathrm{Crown_{ASD}}$ and $\mathrm{Crown_{AISA}}$ scales.

4.6 Managing both within-class pixel heterogeneity and methodological heterogeneity through OBIA and ensemble models

Crown segments generated from OBIA were later used to summarize within-crown pixel data to single crown values (Article III - Figure 2). On the other hand, Figure 3 in Article III highlights
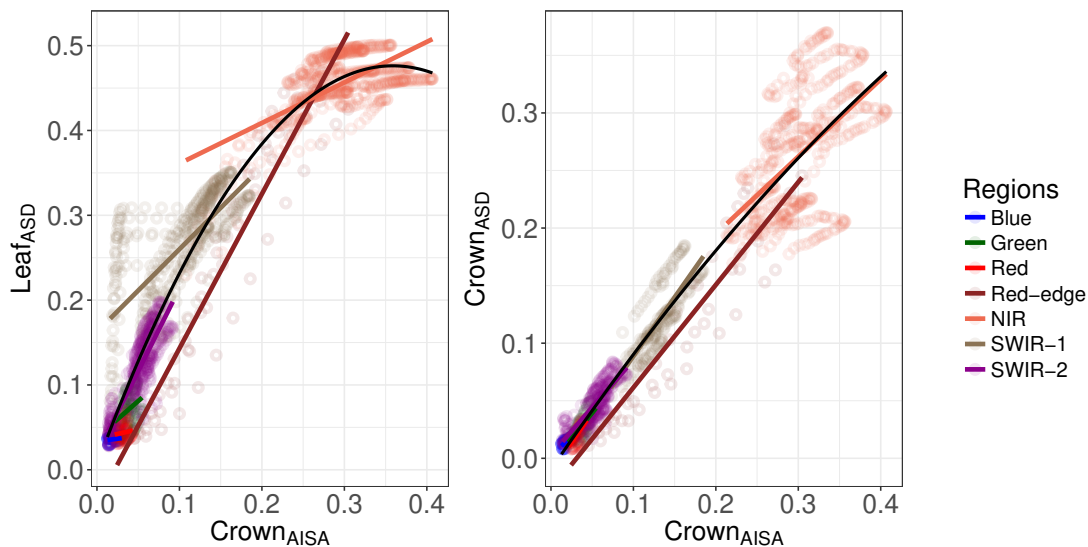
Figure 4.11: Mean observed reflectance at airborne crown scale against mean observed reflectance at in-situ crown and leaf scales. The black solid lines represent global models of second order polynomial. The coloured lines represent within region (local) linear relationships between reflectances at both scales of observation. Units of wavelengths are shown in micrometers. The band observations are slightly faded to emphasize the linear patterns while still showing the variance of observations.

precision of variable importance across models computed from 50 model runs. A comparison between Figures 4 and 5 in Article III highlights the utility of an ensemble classification approach at reducing methodological heterogeneity and maximizing generalization capability across individual models. The peculiar clustering in individual model performance showed MAXENT.Phillips, ANN, CTA and GAM to have been consistently poorly performing and unstable across multiple runs (Figure 4, Article III). Ensemble modelling, on the other hand, improved individual models by increasing both map accuracy and minimizing prediction variance (Figure 5, Article III). All individual models and datasets failed to differentiate the species when employed independently (Figure 4, Article III). Individual performance among the best five models (RF, GBM, GLM, FDA, MARS) was on average TSS $=0.55$ for LiDAR and TSS $= 0.64$ for spectral variables and about TSS $= 0.67$ when both datasets were integrated. However, assemblage of individual models increased map quality with across-model TSS averages to 0.63 for LiDAR models, to 0.68 for spectral models and to 0.73 for both spectral and LiDAR powered models (Figure 5, Article III), as evaluated by a split sample approach. Higher map accuracies were reported with a cross validation evaluation procedure. In addition, the ensemble approach reported precision of the estimated map accuracy through the across model coefficient of variation at an individual crown level as shown in Figure 6, Article III.

4.7 Scaling across hierarchies using regression models

The scale of observation affected derived models and model prediction accuracies when mapping forest stand structural heterogeneity from airborne LiDAR metrics (Figures 7 and 8 in Article I). The referred to figures show that relationships between predictors and the studied indices

as well as the prediction model accuracies (Table 4, Article I), changed with changing the scale of observation. However, within the identified scale domain (i.e. plot sizes >2500 m^2), models and model accuracies remained stable (Figure 8 and Table 4, Article I). Figure 9 in Article I shows example maps at landscape scale derived from plot observations and airborne LiDAR metrics at 2500 m^2 plot size.

In another case study, Article III used ensemble models to scale cross single crown observations collected on field plots to the landscape scale as shown in Figure 6.

4.8 Scaling using the global sampling grid (GSG)
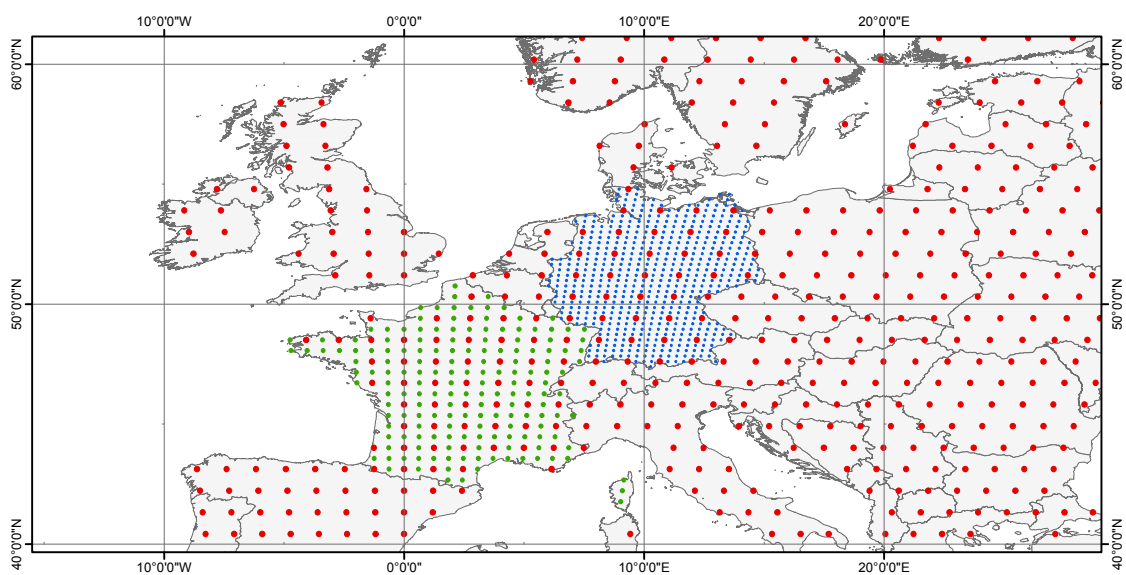
Figure 4.12: Different spatial resolutions of the affine GSG grid system over central Europe: GSG 100 (red dots), GSG 50 over France (green dots) and GSG 25 over Germany (blue dots) (Fehrmann, 2015).

Figures 3.1 and 4.12 show the scaling properties of the GSG. Figure 3.1 shows the map of the entire land surface of the earth sampled by the affine GSG_{250}. The affine GSG returned a semi systematic arrangement of sampling locations, where the distance of points was constant along the latitudes therefore maintaining a constant sampling intensity in all parts of the global land surface. In addition, because all grids generated based on the GSG system had the same origin, the sampling locations of affine grids with different resolutions coincided as shown in Figure 4.12, allowing for more efficient multi-stage sampling schemes where samples in preceding stages (resolutions) can be utilized.

Chapter 5

Discussion

5.1 Patterns of heterogeneity across scales of observation

The thesis findings show that heterogeneity changes with the scale of observation and the pattern of change can be different across FIA variables. As expected, larger areas of support reduced variance in estimates of forest cover proportion (Figures 4.8, 4.9) and within-crown pixel variation in remote sensing data sets (OBIA). Similarly, variance in estimates of forest stand structural heterogeneity reduced as the reference plot size was approached (Figure 4.1). However, increasing the scale of observation increased variance in tree spectral reflectance (Figure 4.10). This result confirms that variance among observational units will only decrease with an increase in the observational size (extent) if the attribute of interest is either a mean or a statistic purported to estimate its expected value in the population under study. The increase in the among-spectral variance in reflectance at a given wavelength is an issue closely linked to the increase in measurement errors that accompanies an increase in the field of view (resolution) and aggregation of many reflecting objects. In general, the observed patterns in heterogeneity across scales of observation confirmed findings by Sandel (2015) that a change in resolution has a higher effect on observed heterogeneity compared to a change in extent when there is low spatial autocorrelation, while extent increases in importance as spatial autocorrelation increases. Observations across scales help us to understand how to aggregate FIA variables and ultimately guide plot design. Articles I & II show how scale research can guide plot design and extensively discuss the general implications for inventory of forest stand structural heterogeneity and forest cover proportions.

5.2 Scale thresholds and domains

This research identified plot size thresholds for inventory of forest stand structural heterogeneity as shown in Figures 4.5, 4.6 and 4.7 and utilized spatial autocorrelation to understand scale domains in the inventory of forest cover proportions (Figures 4.8 and 4.9). Understanding the concept of scale domains benefits inventory of FIA variables by improving efficiency of sampling and observation designs and by simplifying scaling within domains. The case studies

(Articles I & II) show that differences across scales within the domain are subtle and dominant processes are the same or similar. Figure 8 and Table 4 in Article I show how model performance becomes stable within the identified scale domain (Figures 4.5, 4.6 and 4.7). Similarly, Figures 4.8 and 4.9 show how precision stabilizes within domains and sometimes simultaneously for different observation designs. Identifying domain thresholds informs on appropriate/efficient scales of observation. Article I proposes a relatively large plot size to enumerate forest stand structural heterogeneity, a property that is undesirable in many large-area inventories due to the costs involved, while Article II demonstrates why a high number of subplots - as for example utilized by Bastin et al. (2017, 49 subplots) - does not always translate into more efficient estimates of forest cover proportions especially in cases of high spatial autocorrelation. Precision of estimation was also a factor of coverage. Similar results are presented by Kleinn (1991). In addition, the utility of cluster designs with longer subplot distances (e.g ∟, and ×) over compact designs (e.g. + and ⊥) is illustrated in landscapes of varying spatial autocorrelation, highlighting that for the same observation effort a consideration of cluster configuration could yield better efficiency Kleinn (1991). Such findings highlight how understanding scale domains is useful to improve cost efficiency of forest inventories.

5.3 Scale of observation and relationships between terrestrial and remote sensing data

Considering that most FIA variables are spatially structured as site factors and growth conditions in forests are highly correlated for trees close together than further apart, spatial confounding effects in relationships between terrestrial and remote sensing data can be expected. Plot size affected relationships between forest stand structural heterogeneity and airborne LiDAR predictors in different ways including, changes in bivariate correlations (Table 4.1), in model performance (evaluated in Article I by R^2 and $RMSE$) and in variables included in the model. It is possible that plot size also affected model coefficients and model fit but these two properties could not be evaluated given the structure of the Random Forest model utilized and that models were allowed to change across scales of observation. Smaller plot sizes yielded weaker bivariate correlations and model performance (Table 4.1, and Figure 8, Table 4 in Article I), indicating a scale-dependence. This response can be attributed to an increase in edge effects, co-registration error, and a reduction of within-plot variance at smaller plot sizes; as larger plot sizes are more representative of forest structure and density. Ruiz et al. (2014) obtained similar effects of plot size on model fit and performance for LiDAR models of volume, biomass and basal area. Similarly, Mauya et al. (2015) obtained similar results in the estimation of biomass from airborne LiDAR data. Most FIA variables are expected to react to changes in plot size in a similar way due to short ranges of spatial autocorrelation especially when airborne LiDAR data is utilized (Mauro et al., 2017). The loss in bivariate correlation, model fit and performance at smaller plot sizes has interesting consequences for applications involving multiple regression and model selection, mainly because the coefficient estimate for a particular variable may not

change with scale, but the probability that it is included in the selected model can change (Sandel, 2015), hence resulting in a change of model utilized in estimation. Similarly, changes in variables included in the model could result from non-linear relationships between the FIA variable and predictor in question at different scales of observation. An example of such a response was presented in Article I where a non-linear relationship was observed between forest stand structural heterogeneity and crown height. Dominant height (P_{99} - Max) was important at a smaller plot size whereas gaps (P_{01} - P_{10}) became more important with larger plot sizes. Such responses imply that if a relationship is not known in advance and one wishes to use automatic methods to select the model or even the model form, plot size would matter. Yet, the influence of plot size on precision of estimates has not been extensively studied in use of airborne LiDAR models (fit and performance) in FIA (Saarela et al., 2015; Mauya et al., 2015; Mauro et al., 2017). There will also be a need to consider relative efficiency with respect to sample size, on-plot costs, travel costs and overall field inventory design.

5.4 Scaling

The simplest form of scaling is the transfer between hierarchies of analysis maintaining the same resolution of observation. Article I illustrated how transfer across hierarchies of analysis is relatively simplified when scale domains are known a priori. Identification of scale domains stabilized scaling between field plot data and the landscape. Similarly, Article III highlights the potential of OBIA and ensemble models to manage both pixel and methodological heterogeneity, to improve generalization capability and to provide a transparent assessment of mapping errors in the transfer between individual crown data and the landscape. In FIA, model-based as well as design-based model-assisted approaches for long predict across hierarchies of analysis while maintaining the same resolution of observation. However, it is only recently that researchers have began to consider implications of scale domains in assessment of precision derived from modeled populations (Mauya et al., 2015; Magnussen et al., 2016; Mauro et al., 2017). Due to spatial autocorrelation, up scaling can not be linear, but correlation functions need to be estimated from either field (Mauro et al., 2017) or model predictions (Magnussen et al., 2016) to inform the variance estimation process. A mere reliance on linear expansion factors leads to overly optimistic assessments.

Similarly, the presented scaling properties of the GSG are of importance in large area land cover assessments. One such application area of the GSG is the FAO-FRA assessment (FAO, 2009; FAO, 2010) presented in Article II where the problem of unequal sampling intensities can be tackled with the use of the GSG. Similarly, the proposed GSG can be a basis for nearly any sampling design and/or sub selection of locations from a systematic grid. Subselecting (random or systematic) from a dense base grid allows implementing any common sampling design, including stratified, two-phase, multistage sampling (Cochran, 1977) or techniques like systematic unaligned sampling or spatially balanced sampling (Stevens Jr. and Olsen, 2004).

Chapter 6

Conclusions

This chapter presents the overall conclusions from this research across studied topics. Specific conclusions from each study are presented in Articles I - IV respectively. The following conclusions were drawn from this research;

1. The scale of observation affects estimates of heterogeneity of forest stand structural heterogeneity, forest cover proportions and tree species attributes derived from terrestrial and remote sensing data sets in different ways conditioned on, the aggregation properties (i.e. whether the variable is extensive, intensive or scale-specific), the method used in aggregation on support (e.g. mean, variance, quantile etc.) and the spatial structure (spatial autocorrelation) of the studied FIA variable. However, the determination of scale domains and threshold observation units for forest stand structural heterogeneity, forest cover proportions and tree species identities enhances efficiency of their inventory and provides a means to efficiently transfer among scales of observation using regression techniques.

2. From a scaling perspective, this thesis work demonstrates that it is possible to efficiently bridge between hierarchies of survey sampling in forest (land) cover monitoring using the proposed global sampling grid - GSG (Fehrmann et al., 2019) while allowing for simplified and unbiased inference and straight-forward observation protocols such as cluster plot designs, whose efficiency for the same sampling effort is determined by the spatial autocorrelation in the landscape and the subplot distances.

3. It can be generally concluded that the scale of observation affects selection of variables and performance of models linking the studied forest variables to remote sensing data with the magnitude of scale effects determined by the spatial autocorrelation in the studied FIA variables and the dominant processes at the observed scale. To this respect, this thesis showed (1) that the importance of airborne LiDAR predictors can change with the scale of observation in the estimation of forest stand structural heterogeneity and, (2) contributions of structural and foliar properties to observed variation in spectral reflectance can change with the scale of observation.

References

Aberle, H. 2016. Hyperspectral remote sensing and field measurements for forest characteristics - A case study in the Hainich National Park, Central Germany. PhD thesis, University of Goettingen.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology* 43(6): 1223–1232.

Anselin, L. 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis* 27(2): 93–115.

Bastin, J.-f., N. Berrahmouni, A. Grainger, D. Maniatis, D. Mollicone, R. Moore, C. Patriarca, N. Picard, B. Sparrow, E. M. Abraham, K. Aloui, A. Atesoglu, F. Attore, A. Bey, M. Garzuglia, L. G. García-montero, N. Groot, G. Guerin, L. Laestadius, A. J. Lowe, and B. Mamane. 2017. The extent of forest in dryland biomes. *Science* 638(May): 1–5.

Beckschäfer, P., P. Mundhenk, C. Kleinn, Y. Ji, D. Yu, and R. Harrison. 2013. Enhanced structural complexity index: An improved index for describing forest structural complexity. *Open Journal of Forestry* 3: 23–29.

Berger, K., C. Atzberger, M. Danner, G. D'Urso, W. Mauser, F. Vuolo, and T. Hank. 2018. Evaluation of the PROSAIL Model Capabilities for Future Hyperspectral Model Environments: A Review Study. *Remote Sensing* 10(1).

Besag, J. 1977. Contribution to the discussion of Dr Ripley's paper. *Journal of the Royal Statistical Society B* 39(2): 193–195.

Blaschke, T. 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(1): 2–16.

Bohn, F. J., and A. Huth. 2017. The importance of forest structure to biodiversity-productivity relationships. *Royal Society Open Science* 4(1): 160521.

Calle, L., V. Urrea, A. Boulesteix, and N. Malats. 2011. AUC-RF: A new strategy for genomic profiling with Random Forest. *Human Heredity* 72: 121–132.

Chesson, P. 2012. Scale transition theory: Its aims, motivations and predictions. *Ecological Complexity* 10: 52–68.

Cochran, W. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Combal, B., F. Baret, M. Weiss, A. Trubuil, D. Macé, A. Pragn??re, R. Myneni, Y. Knyazikhin, and L. Wang. 2002. Retrieval of canopy biophysical variables from bidirectional reflectance using prior information to solve the ill-posed inverse problem. *Remote Sensing of Environment* 84(1): 1–15.

Czaplewski, R. L. 2003. Can a sample of Landsat sensor scenes reliably estimate the global extent of tropical deforestation? *International Journal of Remote Sensing* 24(6): 1409–1412.

Dănescu, A., A. T. Albrecht, and J. Bauhus. 2016. Structural diversity promotes productivity of mixed, uneven-aged forests in southwestern Germany. *Oecologia* 182(2): 319–333.

Díaz-Yáñez, O., B. Mola-Yudego, J. Ramón González-Olabarria, and T. Pukkala. 2017. How does forest composition and structure affect the stability against wind and snow? *Forest Ecology and Management* 401: 215 – 222.

Ehbrecht, M., P. Schall, C. Ammer, and D. Seidel. 2017. Quantifying stand structural complexity and its relationship with forest management, tree species diversity and microclimate. *Agricultural and Forest Meteorology* 242(Supplement C): 1 – 9.

Eilers, P., B. Marx, and M. Durbán. 2015. Twenty years of P-Splines. *Statistics and Operations Research Transactions* 39: 149–186.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx. 2013. Regression: Models, Methods and Applications. *Springer Berlin Heidelberg, ISBN:978-3-642-34332-2* .

FAO. 2010. Global Forest Resources Assessment 2010. Technical report.

FAO. 2009. The 2010 Global Forest Resources Assessment Remote Sensing Survey: an outline of the objectives, data, methods and approach. Technical report.

Fehrmann, L. 2015. A unified framework for environmental monitoring based on a discrete global sampling grid (GSG) system. In *Proceedings of the 5th international DAAD workshop*, ed. L. Fehrmann and C. Kleinn, 99–111. Cuvellier Verlag Göttingen.

Fehrmann, L., C. B. Kukunda, N. Nölke, S. Schnell, D. Seidel, S. Magnussen, and C. Kleinn. 2019. A unified framework for land cover monitoring based on a discrete global sampling grid (GSG). *Environmental Monitoring and Assessment* 191(2): 46.

Gara, T. W., R. Darvishzadeh, A. K. Skidmore, and T. Wang. 2018. Impact of vertical canopy position on leaf spectral properties and traits across multiple species. *Remote Sensing* 10(2): 1–17.

Garrigues, S., D. Allard, F. Baret, and M. Weiss. 2006. Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sensing of Environment* 103(1): 81–96.

Geschwanter, T., K. Schadauer, C. Vidal, A. Lanz, E. Tomppo, L. di Cosmo, N. Robert, D. Duursma, and M. Lawrence. 2009. Common tree definitions for national forest inventories in Europe. *Silva Fennica* 43(2): 303–321.

González-Ferreiro, E., U. Diéguez-Aranda, L. Barreiro-Fernández, S. Buján, M. Barbosa, J. Suárez, I. Bye, and D. Miranda. 2013. A mixed pixel- and region-based approach for using airborne laser scanning data for individual tree crown delineation in *Pinus radiata D. Don* plantations. *International Journal of Remote Sensing* 34: 7671–7690.

Gotway, C. A., and L. J. Young. 2002. Combining incompatible spatial data. *Journal of the American Statistical Association* 97(458): 632–648.

Gunton, M. R., A. R. Klenke, R. Paloniemi, Y. Gavish, C. J. Marsh, W. E. Kunin, and K. Henle. 2014. The meaning of "scale". In *Scaling in Ecology and Biodiversity Conservation*, ed. K. Henle, S. G. Potts, W. E. Kunin, Y. G. Matsinos, J. Similä, J. D. Pantis, V. Grobelnik, L. Penev, and J. Settele, chap. 19, 206 pp. 1, Sofia: Pensoft Publishers.

Hédl, R., M. Svátek, M. Dančák, A. Rodzay, A. Salleh, and A. Kamariah. 2009. A new technique for inventory of permanent plots in tropical forests: a case study from lowland dipterocarp forest in Kuala Belalong, Brunei Darussalam. *Blumea* 54: 124–130.

Jacquemoud, S., and F. Baret. 1990. PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment* 34(2): 75–91.

Jacquemoud, S., W. Verhoef, F. Baret, C. Bacour, P. J. Zarco-Tejada, G. P. Asner, C. François, and S. L. Ustin. 2009. PROSPECT + SAIL models: A review of use for vegetation characterization. *Remote Sensing of Environment* 113(SUPPL. 1): S56—-S66.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*. URL http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf

Jennings, S. B., N. D. Brown, and D. Sheil. 1999. Assessing forest canopies and understorey illumination : canopy closure , canopy cover and other measures. *Forestry* 72(1): 59–73.

Juchheim, J., C. Ammer, P. Schall, and D. Seidel. 2017. Canopy space filling rather than conventional measures of structural diversity explains productivity of beech stands. *Forest Ecology and Management* 395(Supplement C): 19 – 26.

Khorram, S., F. Koch, C. van der Wiele, and S. Nelson. 2012. *Remote sensing*. Springer. URL http://www.springer.com/978-1-4614-3102-2

Khosravipour, A., A. Skidmore, M. Isenburg, T. Wang, and Y. Hussin. 2014. Generating pit - free canopy height models from airborne lidar. *Photogrammetric Engineering and Remote Sensing* 80: 863–872.

Kleinn, C. 1991. Der Fehler von Flächenschätzungen mit Punkterastern und linienförmigen Stichprobenelementen. PhD thesis, University of Freiburg.

Kleinn, C., G. Kändler, and S. Schnell. 2011. Estimating forest edge length from forest inventory sample dataThis article is one of a selection of papers from Extending Forest Inventory and Monitoring over Space and Time. *Canadian Journal of Forest Research* 41(1): 1–10.

Kleinn, C. 2001. A cautionary note on the minimum crown cover criterion in forest definitions. *Canadian Journal of Forest Research* 31(2): 350–356.

Köhl, M., S. Magnussen, and M. Marchetti. 2006. *Sampling methods, Remote sensing and GIS Multisource Forest Inventory*. Berlin Heidelberg: Springer.

Kukunda, C. B., P. Beckschäfer, P. Magdon, P. Schall, C. Wirth, and C. Kleinn. 2019. Scale-guided mapping of forest stand structural heterogeneity from airborne LiDAR. *Ecological Indicators* 102(September 2018): 410–425.

Levin, S., A. 1992. The problem of pattern and scale in ecology. *Ecology* 73(6): 1943–1967.

Liu, Y., T. Hiyama, and Y. Yamaguchi. 2006. Scaling of land surface temperature using satellite data: A case examination on ASTER and MODIS products over a heterogeneous terrain area. *Remote Sensing of Environment* 105(2): 115–128.

Magdon, P., C. Fischer, H. Fuchs, and C. Kleinn. 2014. Translating criteria of international forest definitions into remote sensing image analysis. *Remote Sensing of Environment* 149: 252–262.

Magdon, P., and C. Kleinn. 2013. Uncertainties of forest area estimates caused by the minimum crown cover criterion: - A scale issue relevant to forest cover monitoring. *Environmental Monitoring and Assessment* 185(6): 5345–5360.

Magnussen, S., D. Mandallaz, A. Lanz, C. Ginzler, E. Næsset, and T. Gobakken. 2016. Scale effects in survey estimates of proportions and quantiles of per unit area attributes. *Forest Ecology and Management* 364: 122–129.

Magnussen, S., and T. Nord-Larsen. 2019. Design-consistent model-based variances with systematic sampling: a case study with the Danish national Forest inventory. *Communications in Statistics - Simulation and Computation* 0(0): 1–11.

Malenovský, Z. 2006. Quantitative remote sensing of Norway spruce (Picea abies (L.) Karst.): Spectroscopy from needles to crowns to canopies. PhD thesis.

Malenovský, Z., H. M. Bartholomeus, F. W. Acerbi-Junior, J. T. Schopfer, T. H. Painter, G. F. Epema, and A. K. Bregt. 2007. Scaling dimensions in spectroscopy of soil and vegetation. *International Journal of Applied Earth Observation and Geoinformation* 9(2): 137–164.

Mandelbrot, B. B. 1967. How long is the coast of Britain? Statistical self-similarity and fractal dimensions. *Science* 156: 636–638.

Marceau, D. J., and G. J. Hay. 1999. Remote Sensing Contributions to the Scale Issue. *Canadian Journal of Remote Sensing* 25(4): 357–366.

Mauro, F., V. Monleon, H. Temesgen, and L. Ruiz. 2017. Analysis of spatial correlation in predictive models of forest variables that use LiDAR auxiliary information. *Canadian Journal of Forest Research* 47(6): 788–799.

Mauya, E. W., E. H. Hansen, T. Gobakken, O. M. Bollandsås, R. E. Malimbwi, and E. a. Næsset. 2015. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. *Carbon Balance and Management* 10(1).

McElhinny, C., P. Gibbons, C. Brack, and J. Bauhus. 2005. Forest and woodland stand structural complexity: Its definition and measurement. *Forest Ecology and Management* 218(1-3): 1–24.

McGaughey, R. 2014. FUSION/LDV: software for LiDAR data analysis and visualization, v. 3.42. *USDA Forest Service, Pacific Northwest Research Station (Seattle, WA)* .

McRoberts, R. E., Q. Chen, G. M. Domke, G. Ståhl, S. Saarela, and J. A. Westfall. 2016. Hybrid estimators for mean aboveground carbon per unit area. *Forest Ecology and Management* 378: 44–56.

McRoberts, R. E., E. Næsset, and T. Gobakken. 2013. Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment* 128: 268 – 275.

Moellering, H., and W. Tobler. 1972. Geographical variances. *Geographical Analysis* 2: 34–64.

Openshaw, S. 1977. A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers* 2(4): 459–472.

Percival, D. P. 1995. On estimation of the wavelet variance. *Biometrika* 82(3): 619–631.

Pommerening, A. 2002. Approaches to quantifying forest structures. *Forestry* 75(3): 305–323.

Puliti, S., S. Saarela, T. Gobakken, G. Ståhl, and E. Næsset. 2018. Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference. *Remote Sensing of Environment* 204: 485–497.

del Río, M., H. Pretzsch, I. Alberdi, K. Bielak, F. Bravo, A. Brunner, S. Condés, M. J. Ducey, T. Fonseca, N. von Lüpke, M. Pach, S. Peric, T. Perot, Z. Souidi, P. Spathelf, H. Sterba, M. Tijardovic, M. Tomé, P. Vallet, and A. Bravo-Oviedo. 2016. Characterization of the structure, dynamics, and productivity of mixed-species stands: review and perspectives. *European Journal of Forest Research* 135(1): 23–49.

Ripley, B. D. 1977. Modelling Spatial Patterns. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(2): 172–212.

Rödig, E., M. Cuntz, J. Heinke, A. Rammig, and A. Huth. 2017. Spatial heterogeneity of biomass and forest structure of the Amazon rain forest: Linking remote sensing, forest modelling and field inventory. *Global Ecology and Biogeography* 26(11): 1292–1302.

Roosjen, P. P., B. Brede, J. M. Suomalainen, H. M. Bartholomeus, L. Kooistra, and J. G. Clevers. 2018. Improved estimation of leaf area index and leaf chlorophyll content of a potato crop using multi-angle spectral data – potential of unmanned aerial vehicle imagery. *International Journal of Applied Earth Observation and Geoinformation* 66(November 2017): 14–26.

Ruiz, L. A., T. Hermosilla, F. Mauro, and M. Godino. 2014. Analysis of the influence of plot size and LiDAR density on forest structure attribute estimates. *Forests* 5(5): 936–951.

Saarela, S., S. Schnell, A. Grafström, S. Tuominen, K. Nordkvist, J. Hyyppä, A. Kangas, and G. Ståhl. 2015. Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume. *Canadian Journal of Forest Research* 45(11): 1524–1534.

Sandel, B. 2015. Towards a taxonomy of spatial scale-dependence. *Ecography* 38(4): 358–369.

Schall, P., and C. Ammer. 2013. How to quantify forest management intenisty in Central European forests. *Eur. J. Forest Res.* 132: 379–396.

Schneider, D. C. 2001. The Rise of the Concept of Scale in Ecology. *BioScience* 51(7): 545–553.

Scott, C. T., and J. H. Gove. 2002. Forest inventory. *Encyclopedia of Environmetrics* 2: 814–820.

Ståhl, G., S. Saarela, S. Schnell, S. Holm, J. Breidenbach, S. P. Healey, P. L. Patterson, S. Magnussen, E. Næsset, R. E. McRoberts, and T. G. Gregoire. 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems* 3(1): 5.

Stein, A., and H. Kreft. 2015. Terminology and quantification of environmental heterogeneity in species-richness research. *Biological Reviews* 90(3): 815–836.

Stevens Jr., D. L., and A. R. Olsen. 2004. Spatially Balanced Sampling of Natural Resources. *Journal of the American Statistical Association* 99(465): 262–278.

Thaden, H., and T. Kneib. 2018. Structural Equation Models for Dealing With Spatial Confounding. *The American Statistician* 72(3): 239–252.

Valbuena, R. 2015. Forest structure indicators based on tree size inequality and their relationships to airborne laser scanning. PhD dissertation, University of Eastern Finland.

Valbuena, R., P. Packalén, S. Martín-Fernández, and M. Maltamo. 2012. Diversity and equitability ordering profiles applied to the study of forest structure. *Forest Ecology and Management* 276: 185–195.

Valbuena, R., P. Packalen, L. Mehtätalo, A. García-Abril, and M. Maltamo. 2013. Characterizing forest structural types and shelterwood dynamics from Lorenz-based indicators predicted by airborne laser scanning. *Canadian Journal of Forest Research* 43(11): 1063–1074.

Weiner, J., and O. Solbrig. 1984. The meaning and measurement of size hierarchies in plant-populations. *Oecologia* 61: 334–336.

Wiens, J. A. 1989. Spatial Scaling in Ecology Spatial scaling in ecology1. *Functional Ecology* 3(4): 385–397.

Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1): 3–36.

Woodcock, C. E., and A. H. Strahler. 1987. the Factor of Scale in Remote-Sensing. *Remote Sensing of Environment* 21(3): 311–332.

Wu, H., and Z. L. Li. 2009. Scale issues in remote sensing: A review on analysis, processing and modeling. *Sensors* 9(3): 1768–1793.

Zenner, E., and D. Hibbs. 2000. A new method for modelling the heterogeneity of forest structure. *Forest Ecology and Management* 129: 75–87.

Zuur, A., E. Ieno, N. Walker, A. Saveliev, and G. Smith. 2009. Mixed effects models and extensions in Ecology with R. *Springer Science+Business Media, LLC ISBN:978-0-387-87457-9* .
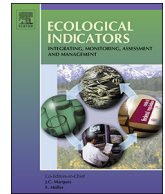
Articles I-IV in full text

Original Articles

Scale-guided mapping of forest stand structural heterogeneity from airborne LiDAR

Collins B. Kukunda[a,*], Philip Beckschäfer[a,b], Paul Magdon[a], Peter Schall[c], Christian Wirth[d,e], Christoph Kleinn[a]

[a] University of Goettingen, Forest Inventory and Remote Sensing, Faculty of Forest Sciences, Büsgenweg 5, D-37077 Göttingen, Germany
[b] Northwest German Forest Research Institute, Grätzelstrasse 2, D-37079 Göttingen, Germany
[c] University of Goettingen, Silviculture and Forest Ecology of the Temperate Zones, Faculty of Forest Sciences, Büsgenweg 1, D-37077 Göttingen, Germany
[d] University of Leipzig, Systematic Botany and Functional Biodiversity, Johannisallee 21, D-04103 Leipzig, Germany
[e] German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig, Deutscher Platz 5e, D-04103 Leipzig, Germany

ABSTRACT

Heterogeneity in forest structure, naturally occurring or induced by management, is continuous in space and time. However, measures used to quantify structure of forests are scale-variant, as they rely on bounded observations on either ecological or forest inventory observation units. The understanding of the influence of the scale of observation in mapping of forest structural heterogeneity is limited. Therefore, we researched into effects of plot size on quantifying forest structural heterogeneity, where we describe heterogeneity by three indices in stands under different management systems. In addition, we studied the performance of structural indices in separating different forest management systems across plot sizes, and created wall to wall maps of the indices using airborne LiDAR metrics describing the vertical distribution of canopy heights at different scales of observation. The studied indices are: Gini Coefficient (GC), Structural Complexity Index (SCI), and Enhanced Structural Complexity Index (ESCI). SCI and ESCI require fully mapped plots whereas GC has no information on individual tree locations. Inventory data from 95 one-hectare plots covering a range of management intensities from un-managed to age class forests were used. We quantified the three structural indices for 18 plot sizes ranging from 225 m^2 to 10,000 m^2. Linear fixed effects models were used to study the effects of plot sizes in different levels of structural heterogeneity and Random Forest (RF) models used to provide wall-to-wall maps at varying scales from airborne LiDAR data. The simulation showed that all indices were influenced by the scale of observation with larger effects for plots in forests with higher structural heterogeneity. For the data analyzed we found a threshold scale for enumerating stand structural heterogeneity between 900 m^2 and 2500 m^2. However, stable field and remote sensing predictions of stand structural heterogeneity required plots at least ⩾2500 m^2. Compared to GC, SCI and ESCI improved separation of forest structure in the three management systems and at all observed scales. A change of plot sizes affected bivariate relationships between structural indices and airborne LiDAR metrics as well as the resultant predictive models. Smaller plot sizes yielded weaker relationships and predictive models. All structure indices were predicted from airborne LiDAR with RMSE ⩽ 22% at scales equal or larger than the identified threshold plot size. These findings are relevant to optimize plot sizes for efficient inventory and mapping of forest structural heterogeneity, as well as for the design of natural resource inventories. Additionally, derived maps are useful for studies on forest structure and the link with forest growth, degradation, management intensity, productivity, and biodiversity in the regions.

1. Introduction

Structure is a relevant feature of forest ecosystems frequently linked to biodiversity (McElhinny et al., 2005), productivity (Schall and Ammer, 2013; Danescu et. al., 2016; Bohn and Huth, 2017; Juchheim et. al. 2017), ecosystem functioning and stability (Díaz-Yáñez., 2017), and is as well used as auxiliary data in forest inventory (Rödig et al., 2017). Forest management affects forest structure across scales through interventions at the scale of individual trees, larger thinnings and clearcuts. To support management decisions at a stand scale – which may

---

also affect regional and landscape scales – there is an increasing demand for efficient tools that allow answering questions such as to which extent the removal of trees results in structural changes at the stand level or how alteration of structure affects ecosystem processes such as growth and mortality. In order to support implementation of management and conservation practices, as well as to effectively monitor impacts of management regimes on forest structure across scales, both field and remote sensing techniques need to be developed and tested (Valbuena et al., 2016). This way, the indicators derived from field-plot data can be linked to remote sensing indicators of forest structure such as airborne LiDAR derivatives (Valbuena et al., 2014) to produce continuous maps of stand structural heterogeneity (Valbuena et al., 2016). These maps of stand structural heterogeneity are useful in decision making (see e.g. (Valbuena et al., 2016)) and support the general understanding of forest ecosystems functioning.

Due to the complexity of forests, approaches defining and quantifying their structure from both field and remote sensing data are manifold (Zenner and Hibbs, 2000; Pommerening, 2002; Valbuena et al., 2012; Beckschäfer et al., 2013; del Río et al., 2016; Ehbrecht et al. 2017). Forest structure is one of those forest mensuration variables that can not be ”measured” directly but needs to be captured by indices or indicators. Therefore, the quantification of forest structural heterogeneity is often approached through spatial or non-spatial indices (Pommerening, 2002) that include either individual structural components of the forest or aggregations of forest attributes (McElhinny et al., 2005; del Río et al., 2016). Various structural aspects of a forest are considered in structural indices, including, the horizontal distribution of trees, the horizontal variation in tree dimensions, stand density and size distribution, forest biomass, leaf area, canopy closure and layering, species richness, composition and diversity, stand development stages, deadwood, forest area and fragmentation (McElhinny et al., 2005; del Río et al., 2016). Of the many structural elements considered in literature, the species, spatial arrangement, and the size and distribution of individual trees fundamentally determine the structure of a forest stand as individual trees fill the volume of space on a defined area. Among the many available indices of structural heterogeneity, Ripley's K (Ripley and Patterns, 1977), Besag's L (Besag, 1977) and semivariograms summarize these essential attributes (i.e. spatial arrangement, and the size and distribution of individual trees in space), but because of their functional nature, present a challenge when maps are required at wider scales. Therefore, we here focus the study on three indices that summarize fundamental attributes into concise values easily mapped at wider scales. One of the indices is non-spatial, the Gini Coefficient (GC) (Valbuena et al., 2013), and the other two are spatial including Structural Complexity Index (SCI) (Zenner and Hibbs, 2000) and Enhanced Structural Complexity Index (ESCI) (Beckschäfer et al., 2013).

GC and conventional diversity and evenness indices (Magurran, 2004), are probably the most utilized measures to characterize forest structural heterogeneity at stand scales. Valbuena et al. (2012) compared the potential of various non-spatial indices at characterizing tree size inequality among stands. They found that GC and Lorenz asymmetry (LA) – both metrics from the Lorenz curve (Weiner and Solbrig, 1984) – were superior at characterizing differences in tree size inequality among stands especially because they are normalized and thus considered independent of the total number of trees surveyed, quadratic mean diameter or development stage of the stand. In principal, GC measures dispersion around the average pair difference in tree sizes (Valbuena et al., 2013, 2014) and LA measures the balance or skewness in the distribution of tree sizes as the proportion of basal area or stem density above the quadratic mean diameter (Valbuena et al., 2014). While studying boreal unmanaged and commercial forest plantations, Valbuena et al. (2016) show that a threshold value of $GC = 0.5$ can reliably segregate even-aged from uneven-aged stands and GC values consistently order tree size inequality in realistic magnitudes; thus GC is highly recommended over use of conventional entropy and diversity indices (Valbuena et al., 2012) in characterization of forest structure in

similar settings. Subsequent studies showed that GC can be mapped reliably (RMSE ⩽ 20%) from airborne LiDAR data (Valbuena, 2015; Valbuena et al., 2013, 2014, 2016), or from the fusion of airborne LiDAR and optical data sets (Manzanera et al., 2016). The relationship between GC and airborne LiDAR-derived variations in tree or canopy height, canopy cover, and the presence of gaps or understory (Valbuena et al., 2013, 2014, 2016; Manzanera et al., 2016; Adnan et al., 2017) is well established. Similarly, GC has been found to be related to management intensity (Valbuena et al., 2016).

The other concise and holistic characterization of structural heterogeneity in forests that should be further studied comes from Zenner and Hibbs (2000), the Structural Complexity Index. SCI, defined as the *"…juxtaposition of differently sized neighboring trees"* (Zenner, 2004), is a three-dimensional abstraction of the position of trees and their respective structural attribute (e.g. the dimensions). SCI encompasses inter-tree size variability – as does GC, and also integrates within and between-patch heterogeneity (Zenner and Hibbs, 2000); through an explicit account of tree distribution in space. The accounting of spatial distribution of trees restricts utility of SCI to mapped data as tree positions are prerequisite. Modifications to SCI were proposed by Beckschäfer et al. (2013) in their so-called Enhanced Structural Complexity Index. Beckschäfer et al. (2013) included in the computation of SCI the degree of mingling among trees – i.e. in relation to adjacency of differently sized trees in space – and called this the vector ruggedness measure. Similarly, they included a stem density term in the computation of SCI. In a case study based on simulated forest stands of different structural characteristics and inventory data from a gradient of forest types, Beckschäfer et al. (2013) showed that inclusion of a vector ruggedness measure and a density factor into SCI to form ESCI, enhanced the separability of forests of different magnitudes of structural heterogeneity. Research has also shown that both SCI and ESCI, in a similar way to GC, order structural heterogeneity in realistic magnitudes (Zenner, 2004; Beckschäfer et al., 2013). However, little is known about SCI and ESCI (in comparison to GC), especially in relation to their covariance patterns with remote sensing descriptors of forest structure such as airborne LiDAR metrics. Relating remote sensing information to forest structure metrics that characterize the structure based on the dimension and spatial distribution of individual trees is challenging particularly for structurally complex, multi-layered forests. LiDAR systems have the potential to penetrate the forest providing 3D structural information. It has been shown in many studies that LiDAR data could be used to detect individual trees in the dominant canopy layer. However, most of the studies report poor results for the sub-canopy trees (see for example (Jeronimo et al., 2018)). Therefore, we follow an area-based approach where the distribution of vegetation is characterized based on the distribution of the 3D point cloud summarized over a unified area (plots, pixels). To the best knowledge of the authors, the relationships between SCI, ESCI and any remote sensing descriptors of stand structure has not been explored yet.

Since GC measures tree size inequality, spatial indices such as SCI and ESCI that aim to incorporate more holistic definitions of forest structure (Zenner and Hibbs, 2000; Schall et al, 2018) may better characterize forest structural heterogeneity and at the same time could exhibit better relationships with airborne LiDAR descriptors of forest structure. However, in order to reliably compare across indices, the scale at which observations are made must be considered. This is because differently sized or shaped observational units might result in different and incomparable estimates of structure. Larger plot sizes may yield more reliable estimates of stand structural heterogeneity (Zenner, 2005; Zenner et al., 2015; Adnan et al., 2017) as different stand conditions aggregate on the plot and a point of index stabilization is reached (Zenner, 2005; Adnan et al., 2017). The spatial arrangement of trees equally dictates the threshold plot-size to achieve a given precision of stand structural heterogeneity (Pommerening, 2002). Observations made on multiple scales might reveal relationships of forest structure and other forest properties of interest that would otherwise be

missed by observations at a single scale (Zenner et al., 2015; Adnan et al., 2017). Similarly, as scales of observation change, the relationship between structural indices and remote sensing predictors is likely to change, affecting their predictability from remote sensing (Adnan et al., 2017). Therefore, to ensure that comparisons among indices in this study are done at stable scales of observation, we first assessed the dependency of the indices on plot size to identify a threshold-plot size for stable observation. Once the threshold-plot size was established, three scales of observation were considered for further comparative analyses i.e. 400 m², 2500 m², 10,000 m².

The study objective is to contribute to better inventory and mapping of forest stand structural heterogeneity from field and airborne LiDAR data. The technical objectives are: (1) to assess GC, SCI and ESCI dependence on plot size and identify a threshold plot size for stable observation. (2) to assess characterization of forest structural heterogeneity by three structural indices, GC, SCI and ESCI, among forests under three forest management systems including unmanaged, target diameter selection (hereafter referred to as selection system) and age class. (3) to predict GC, SCI and ESCI from airborne LiDAR metrics producing continuous maps of stand structural heterogeneity.

2. Material and methods

2.1. Study sites

The study sites are part of a large research program, the Deutsche Foschungsgemeinschaft (DFG) Infrastructure Priority Program referred to as Biodiversity Exploratories (BEs) in line with the terminology used by Fischer et al. (2010). The BEs are distributed across Germany in three regions of high biodiversity and nature conservation value (see Biodiversity Exploratories Project, www.biodiversity-exploratories.de). The first two are located in Schwäbische Alb (South-west Germany) and Schorfheide-Chorin (North-west Germany) and are both UNESCO Biosphere Reserves. The other BE is in the Hainich-Dün region (Central Germany) and encompasses the Hainich National Park (HNP) and its surroundings (Fig. 1). Information on general environmental conditions influencing site suitability of forests within the BEs such as geology and soil conditions, altitude, precipitation and temperature can be found in Fischer et al. (2010). More detailed information on soil properties is available in Solly et al. (2013). Further details on forest management in the BEs can be found in Schall et al. (2018).

2.2. Characteristics of forests at study sites

The three BEs consist of forest sites distinct in environmental conditions and forest management intensity (Schall and Ammer, 2013) and are representative of a gradient of forest types and forest management systems widely practiced in Central Europe (Table 1). Detailed information on the forest type classification used in Table 1 can be found in Hessenmöller et al., (2011); Schall and Ammer (2013). The following forest management systems are practiced at the study areas; (1) Unmanaged forests which were set aside decades ago but which still show some signs of the management formerly applied, (2) European beech (*Fagus sylvatica* L.) forests managed for decades by single tree selection resulting in uneven-aged forests and, (3) even-aged forests representing different developmental stages (age classes, see Table 1). Descriptive statistics on properties of forests under the different management systems in the study areas are shown in Table 2.

As mentioned, most of the forests in the study areas are managed to different degrees providing a gradient of even- and uneven-aged stands of varying degrees of structural heterogeneity. Forest management in all BEs has a history of thinning from below (i.e. thinning by removal of small and "weak" trees) at polewood stages promoting single layered even-aged forests, followed by thinning from above (i.e. thinning by removal of tall/larger trees that compete with target trees) promoting timber quality and vertical structure in the shortest possible time. There

are shelter wood cuttings in some parts, and in the recent decades, target diameter cuttings are common in many parts (Schall et al., 2018). Similarly, there are protected areas where cuttings have not been done since about 50 years ago. For example, the Hainich-Dün area (dominated by beech stands) is a National Park since 1997. However, the stands within the national park (in the southern part of the Hainich study area – Fig. 1) have been unmanaged since 1965 as this area was used as a military training area (Holzwarth et al., 2013). In the Schwäbische Alb area, beech forests dominate the lower altitudes and Norway spruce (*Picea abies*) the higher altitudes. There, pastoral forests with large beech individuals were protected from grazing and logging, allowing open spaces to naturally succeed into heterogeneous uneven-aged stands with various species (Schall et al., 2018). The Schorfheide-Chorin is dominated by Scots pine (*Pinus sylvestris*) stands with parts of beech, pine/beech and oak (*Quercus robur*) forests. There, the unmanaged forests were originally regular even-aged beech forests and are therefore single-layered (Schall et al., 2018). With forests of various types, under differing management strategies and management intensities, and from differing management histories, the BEs offer a wide range of different forest structures in temperate managed forest conditions, suitable for this comparative study on indices of structure and among scales of observation. In these forests, we had comprehensive data sets: one large-area plot of 28.5 hectares and 80 plots of one hectare each.

2.3. Data

2.3.1. Large-area plot

The large-area plot is in a beech-unmanaged-mature stand – located in the Weberstedter Holz in HNP (Butler-Manning, 2007). Data from this plot were provided by a joint project between the Max-Planck-Institute for Biogeochemistry, Jena, the University of Freiburg and HNP. This data set was used to study the influence of the scale of observation on estimates of structural heterogeneity at scales larger than one hectare. Additionally, 15 one-hectare plots of similar design as the experimental plots (EPs) – described in Section 2.3.2 – were cut out of the large-area plot to complement the experimental plot data. For details on the locations and design of the 15 simulated EP plots, see Fig. 1.

2.3.2. Experimental plots (EPs)

EPs are long-term research plots within the BEs. They are one hectare in size, located in stands of differing forest types and management systems (Table 1). The selection of EPs was done according to Fischer et al. (2010) after conducting preliminary large-scale soil, vegetation and land use inventories to identify the range of forest types, conditions and management intensities in the studied regions. For this study, individual tree coordinates were available for a subset of 80 out of the 150 plots originally selected by Fischer et al. (2010). Together, 95 experimental plots were constituted for this study i.e. 80 EPs plus 15 one-hectare simulated EPs (Section 2.3.1).

2.3.3. Airborne LiDAR data

Airborne LiDAR data was collected during leaf-on, cloud and fog free conditions, in July 2015. A full waveform signal and the respective intensity were recorded using the Riegl Q780 sensor at a frequency of 400 kHz. In areas with vegetation cover, a minimum of four potential returns per pulse were generated from each full waveform signal. In total, an area of approximately 100 km² was scanned wall-to-wall in each of the BEs (Fig. 1). Flight height was averagely 940 m above ground level. An average return density of 31 points/m² with an overall average inter-point spacing of 18 cm was achieved at all study plots.

2.4. Field inventory

We conducted field work on EPs in the winters of 2014/15 and 2015/16. We measured diameters at 1.3 m above the ground (DBH) of
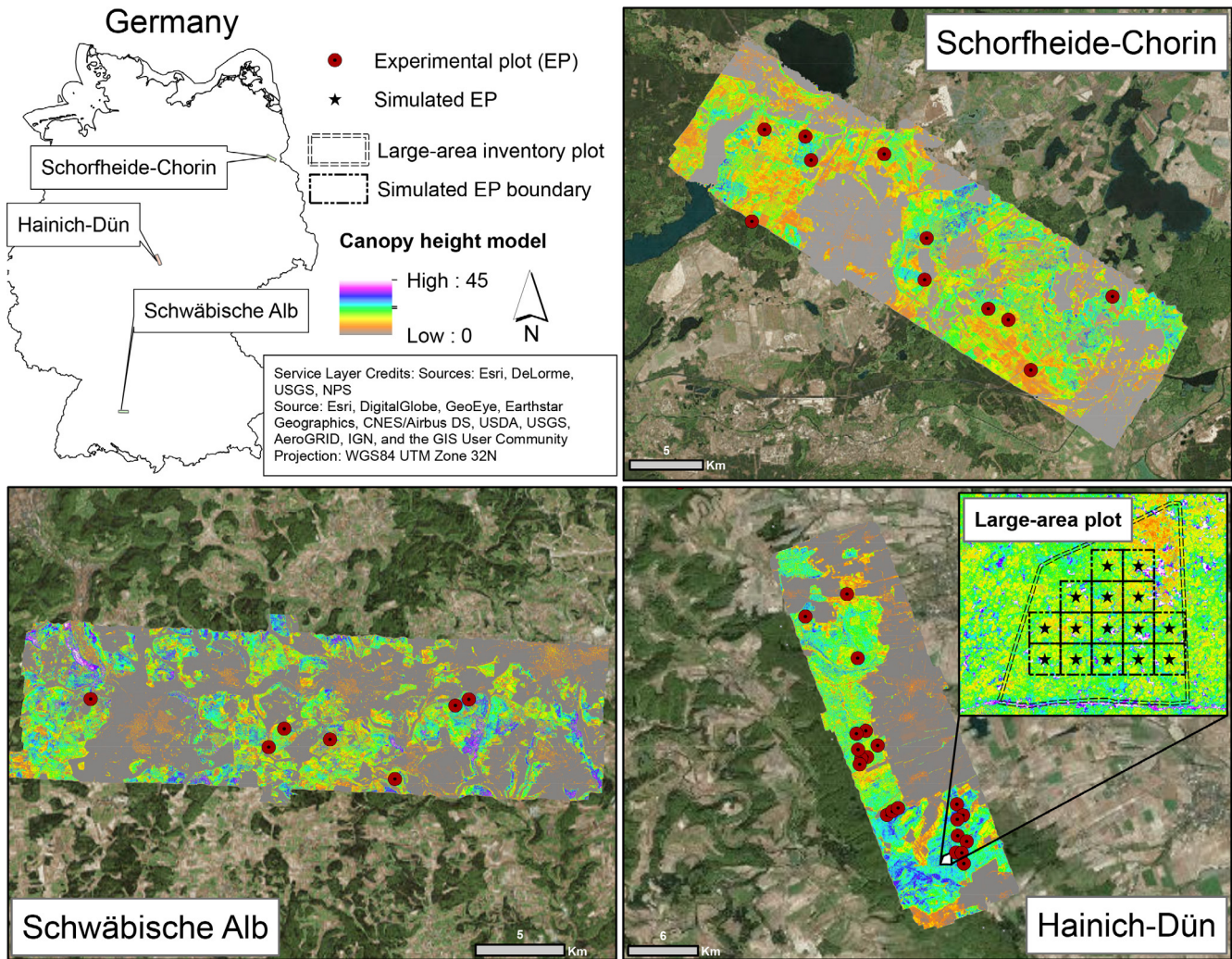
Fig. 1. Map of the study areas highlighting the utilized datasets.

Table 1

Information on forest types based on dominant species, management system and development stage of selected plots of three Biodiversity Exploratories (BE). A = Schwäbische Alb, H = Hainich-Dün, S = Schorfheide-Chorin.

Dominant species	Management system	Development stage	No. of plots (No. per BE)
Beech	Age class	Immature	10 (A = 5, H = 4, S = 1)
		Mature	11 (A = 1, H = 4, S = 6)
		Thicket	1 (A = 1)
		Thicket with shelterwood	1 (H = 1)
	Selection	Mature	13 (H = 13)
	Unmanaged	Mature	38 (A = 3, H = 28, S = 7)
Oak	Age class	Immature	1 (S = 1)
		Mature	5 (S = 5)
Pine	Age class	Immature	4 (S = 4)
		Mature	2 (S = 2)
Spruce	Age class	Immature	4 (A = 1, H = 3)
		Mature	1 (H = 1)
Pine/beech	Age class	Mature	4 (S = 4)

all trees ⩾7 cm with a diameter tape. DBH, species identities and position were recorded using the Fieldmap® inventory software v3.151 (www.fieldmap.cz). The Fieldmap system and procedure of measuring tree coordinates are described by Hédl et al. (2009). This state-of-the-art instrumentation guaranteed accurate mapping of tree positions (in reference to tree stem location). Similarly, field work on the large-area plot took place in the winter of 2013/14. The position, DBH and height for all trees with a DBH ⩾1 cm were recorded by staff of the German Center for Integrative Biodiversity Research (iDiv). For purposes of comparability with the EP dataset, we selected in the computation of the structural indices only trees with DBH ⩾7 cm. A tape was used to measure DBH. Measurement of xy-coordinates was based on a reference grid initially established in 1999 by means of a Tachymeter and reconstructed in 2013 with an amplitude compass and a VERTEX©. The position deviations incurred during grid reconstruction was determined between ± 20 cm and ± 50 cm.

2.5. Description of computed indices of structure

GC (Weiner and Solbrig, 1984) was employed as a measure of tree size inequality – using DBH (Valbuena et al., 2012) and was calculated by linking the cumulative proportion of ordered basal area to the respective cumulative proportion of the total number of trees encountered on a plot (Valbuena et al., 2013, 2014) to make a Lorenz curve (Weiner and Solbrig, 1984; Weiner, 1985). Tree sizes must be ordered prior to computing a cumulative sum of basal areas (i.e.

Table 2
Characteristics of forests by management type.

	Management system	No. of plots	Min	Mean	Max	SD
Density (stems ha^{-1})	Age class	44	117.00	338.22	575.00	110.13
	Unmanaged	38	249.00	376.41	680.00	124.19
	Selection	13	206.00	322.29	425.00	71.87
Basal area (m^2 ha^{-1})	Age class	44	16.68	30.86	42.53	6.11
	Unmanaged	38	31.46	38.57	42.76	2.73
	Selection	13	27.04	28.60	31.01	1.34
Quadratic mean diameter (cm)	Age class	44	26.92	35.13	44.94	5.19
	Unmanaged	38	25.67	37.32	45.04	5.50
	Selection	13	29.15	34.21	42.75	4.37

$g_i \leqslant g_j...g_n$). In a theoretical case of absolute equality among all tree pairs (i and j), the Lorenz curve results into a 45° line; as the rate of change of pair differences in basal area is equal to the rate of change of simultaneous counts. Conversely, the rate of change of pair differences in basal area for individuals with any degree of inequality among individual pairs drifts off the 45° line. GC was computed from the Lorenz curve as the arithmetic mean of differences between all individual tree pairs (Weiner and Solbrig, 1984; Weiner, 1985; Valbuena et al., 2014), normalized by the quadratic mean diameter – see Eq. (1) (Valbuena et al., 2013). This is equivalent to the ratio of the area between the Lorenz curve and the 45° line to the area under the 45° line.

$$GC = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |g_i - g_j|}{2n^2 \bar{g}} * \frac{n}{n-1} \qquad (1)$$

where, g_i and g_j are succeeding trees in the hierarchy of tree sizes, $\bar{g}$ is the mean basal area and n is the total number of trees encountered. GC is finite population corrected when a sample is as large as the population e.g. when all trees on a given area are surveyed. GC can take on values between 0 and 1.

SCI and ESCI were computed using DBH as the z-coordinate and individual tree positions. Both indices attribute higher weight to larger trees in the stand thereby aligning the definition of structure to that of a mature/old growth forest which relies more on horizontal than vertical tree size distribution (Zenner, 2005). Detailed descriptions of ESCI and SCI can be found in their respective methodological publications (Zenner and Hibbs, 2000; Beckschäfer et al., 2013). SCI (Eq. (2)) is calculated as the ratio of the surface area of a triangulated irregular network (TIN) resulting from a Delaunay triangulation of a set of tree xy-coordinates, and a tree attribute (in this case DBH) as the z-coordinate, to the area of the TIN as projected into the horizontal plane. SCI can take on values between 1 and infinity (Zenner, 2005).

$$SCI = \frac{A_{surface}}{A_{projection}} \qquad (2)$$

SCI describes the nearness of differently sized neighbouring trees where $A_{surface}$ stands for surface area of TIN and $A_{projection}$ is the projected area of TIN. Founded on SCI, the ESCI (see Eq. (5)) incorporates in its computation a measure of orientation of component triangles in the TIN leading to a descriptor of surface ruggedness of the TIN (Vector Ruggedness Measure, VRM – Eq. (3) and a stem density term. The changes in ESCI enhance differentiation capabilities of SCI Beckschäfer et al. (2013). VRM ranges between 1 for a smooth surface e.g. observed in a uniform stand and 2 for a stand with high intermingling of unequal tree sizes as with the unmanaged stands and is calculated following Beckschäfer et al. (2013):

$$VRM = 2 - \frac{VSTR}{n} \qquad (3)$$

where,

$$VSTR = \left| \sum_{i=1}^{n} \frac{a_i \times b_i}{|a_i \times b_i|} \right| \qquad (4)$$

The vector strength, VSTR, is the total magnitude of the vector that originates from summing up the unit vectors (a_i, b_i) centered at each triangle (i) in the TIN. ESCI is continuous in magnitude (ESCI $\in [1, \infty]$) and is computed as;

$$ESCI = SCI*VRM*(1 + No.\ of stems^{-10m^2}) \qquad (5)$$

It is notable that both SCI and ESCI are based on triangulation and therefore a minimum of three trees is required for their computation.

2.6. Testing different scales of observation

To evaluate whether and how different plot sizes lead to differences in calculated indices, we implemented a moving window algorithm for 18 plot sizes from 15 m * 15 m to 100 m * 100 m in 5 m steps. For each scale of observation between 15 m * 15 m and 95 m * 95 m, 49 index observations were made in a sliding window covering the full extent of each EP plot. In addition, a single index observation was made for each entire EP plot (100 m * 100 m) as reference. Observation windows were laid systematically within dynamic sampling frames in a way that removed edge effects (Fig. 2). In so doing, sample frames changed relative to scale of observation and ensured that observation windows precisely aligned with the EP plot boundary. Sample frames were determined as negative buffers from each EP plot extent, equivalent to half the side length of the considered scale of observation/plot size (Fig. 2). There was overlap between adjacent windows at all considered scales of observation, consistently resulting in a full census of trees on the plot. However, as the plot sizes became smaller, some plots fell within canopy gaps (hence no trees surveyed) or had ⩽ 3 trees encountered on the plot; such plots were excluded from index computations as SCI and ESCI require a minimum number of three trees.

2.7. Estimation of the influence of plot size

At each scale of observation, percentiles describing the distribution of GC, ESCI and SCI were computed i.e. the minimum and maximum index values and, 5th, 10th, 25th, 50th, 75th, 90th, 95th percentiles. Such an approach was also used by Zenner (2005) to describe the index distribution and linked percentiles to plot area via regression to form structure-area-curves. In our case, the structure-area-curves were fitted using linear models for fixed effects and for each of the 95 one-hectare plots and three indices separately. For the fixed effects models, we used nine groups of percentiles and three indices on the total 95 plots, which makes a total of 2565 structure-area-curves (i.e. 95 plots * 3 indices * 9 groups). Note that grouping crossed among scales of observation; for example, all maximum values at all considered scales in one group. Also note that all percentiles converged to the one-hectare plot estimate as a reference. The intercept and slope were allowed to vary among the groups. Natural logarithmic transformation on the independent variable (area per plot) was performed to linearize its relationship with the
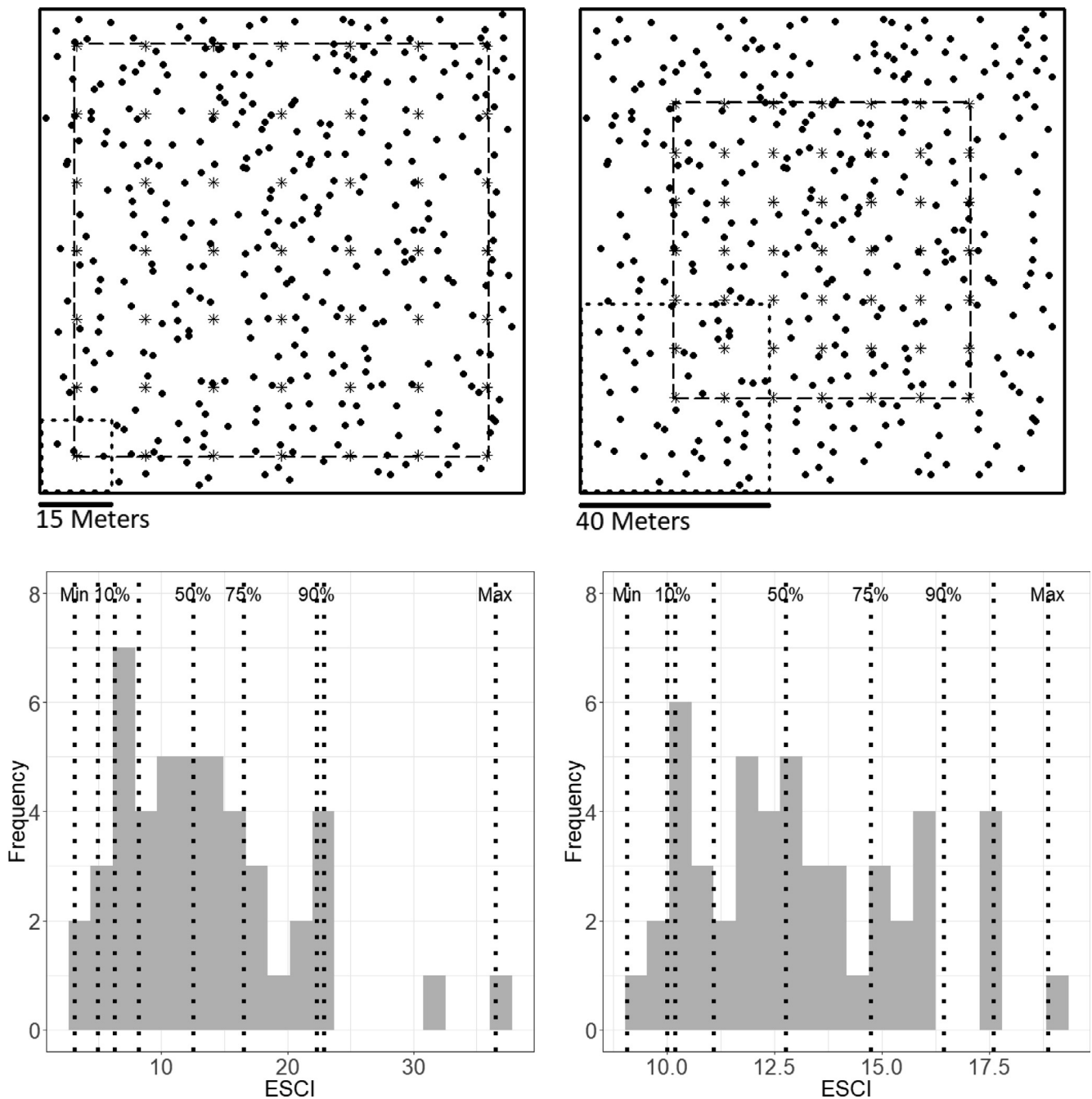
Fig. 2. 15 m * 15 m and 40 m * 40 m plots (above) over an example EP plot (1 ha) with corresponding histograms (below) of enhanced structural heterogeneity index (ESCI). Dotted points (above) show tree locations, asterisks (above) show plot centers for all 49 laid subplots, the dotted lines (above) show boundaries of one of the 49 plots, and the dashed lines show the sampling frame within which plot centers were laid in a way that maintained overlap between subplots and had no edge effects. The dotted lines (below) show positions of the minimum and maximum index values, and 5th, 10th, 25th, 50th, 75th, 90th, 95th percentiles of the distribution.

indices. Using generalized additive models, a secondary grouping variable (forest management systems) was used to estimate mean structure-area-curves and related confidence intervals for each of the forest management systems and for different regions of the index distributions as described in percentiles. In this step we treated structure-area-curves of individual plots as a set of independent observations including index values and their respective plot sizes and by means of spline bases (Eilers et al., 2015), estimated mean curves for each forest management type. Lastly, we determined threshold scales of observation for each of the 2565 curves separately by considering the point of index stabilization; equivalent to where the second derivative of the

structure-area-curve was equal to zero. As derivatives are changes between two scales (say between 15 m * 15 m and 20 m * 20 m scales), we conservatively considered the upper bound (in this example 20 m * 20 m) as the threshold plot size at which the indices can be effectively enumerated with minimal scale effects. The linear models for fixed effects were fitted using the R lm function (R Core Team, 2016) and generalized additive models were fitted using the R mgcv package (Wood, 2011). As previously mentioned, further analyses to illustrate the impact of changing scales of observation on the relationship between airborne LiDAR and the three structural indices were performed at three preselected plot sizes (i.e. 400 m^2, 2500 m^2, 10,000 m^2), chosen

in consideration of the identified threshold scale of observation.

2.8. Airborne LiDAR metrics

We used LAStools© v.170419 software (Isenburg, 2017) to pre-process the airborne LiDAR point cloud. First, we re-tiled the point cloud to tiles of half a kilometer width each, to facilitate faster paral-lelized processing. To eliminate effects from flight overlap and multiple scanning, we used *lasoverlap* and *lasoverage* tools after reclassification of the point cloud to identify ground returns by the *lasground* tool. Next, the point cloud was normalized to remove the effect of the terrain on heights of features in the landscape, and to drop returns higher than the expected tree height using the *lasheight* tool. In order to process in-dividual plot metrics, we clipped the point cloud to the boundaries of each study plot using the *lasclip* tool and for the entire study site, we calculated metrics on grids commensurate to the scale of observation of field-based indices.

Finally, we employed the FUSION© LiDAR software v.3.6 to cal-culate standard metrics on height distributions and crown cover from the point cloud as done by Valbuena et al. (2013) and here using the *CloudMetrics* (individual plots), and *GridMetrics* (entire study site) tools (McGaughey, 2014). To calculate height metrics all normalized returns were considered whereas for crown cover metrics, a height threshold of 7 m was used, corresponding to the average height of trees of DBH 7 cm estimated from a DBH-height curve of field observations. 7 cm was the minimum DBH considered during field inventory on the EPs. In addi-tion, cover metrics were computed based on pulse penetrability down to the ground (i.e 0 m in a normalized point cloud). Table 3 summarizes the computed LiDAR metrics.

Table 3
Summary of LiDAR metrics.

Height related metrics	
Group	Abbreviation
Order statistics	Min; Max; P50
Height quantiles	P_i for i = 1, 5, 10, 20, 30...90, 95, 99
Central tendency	Mean; Mode
Dispersion	SD; Var; CV; IQR; AAD; MAD_{mode};MAD_{med}
Skewness,Kurtosis	Skew; Kurt
L-Moments	L_i for i = 1, 2, 3, 4
L-ratios	L.CV; L.skew; L.kurt
Generalized means for the 2nd and 3rd power	QRT, CUB
Crown cover related metrics	
% first returns above: 7 m, mean, mode	A_j for j = 7 m, mean, mode
% all returns above: 7 m, mean, mode	B_k for k = 7 m, mean, mode
% first returns above: 0 m, mean,mode	D_n for n = 0 m, mean, mode
% all returns above: 0 m, mean, mode	E_m for m = 0 m, mean, mode
Canopy Relief Ratio	CRR
Count of returns by return number	C_i for i = 1:9, Total, Other
Count of returns above: 7 m, 0 m, mean, mode	C_{7m}, C_{0m}, C_{mean}, C_{mode}
Count of first returns above: 7 m, 0 m, mean, mode	$C1_{7m}$, $C1_{0m}$, $C1_{mean}$, $C1_{mode}$

Abbreviations: SD, Standard deviation; CV, Coefficient of variation; IQR, Inter-quartile range; AAD, Average absolute deviation; Lskew, L-moment of Skewness; Lkurt, L-Moment of Kurtosis; MAD_{mode}, Median of the absolute de-viations from the overall mode; MAD_{med}, Median of the absolute deviations from the overall median. Further information on metrics is available in McGaughey (2014).

2.9. Variable selection and Random Forest models

We explored relationships between each airborne LiDAR metric and each index of structural heterogeneity before performing variable se-lection. The bivariate relationships explained why particular variables were selected into the final predictor set. Spearman's correlations were chosen to reduce the influence of outliers and clustering (Valbuena et al., 2013) in estimating strength of bivariate relationships. Even-tually, variable selection was performed for each index and for each scale of observations independently using a step-wise regression method specifically adapted for Random Forest (RF) models (Ran-domForest R package, Liaw and Wiener (2002)) in the VSURF (Variable Selection Using Random Forest) R package (Genuer et al., 2016). More detailed information on the VSURF algorithm is available in Genuer et al. (2016). To ensure stability of the variable selection process, we iterated over the VSURF algorithm 50 times and selected only variables that were chosen into the best predictor set in each of the iterations. The RF models were trained with field data, fitted with the best pre-dictor subset, and predictions made for all the pixels in the study area.

2.10. Model validation and assessment of generated maps

Model quality was evaluated by a 6-fold cross validation procedure in which 17% of data was isolated from model development and used for validation by the root mean squared error (RMSE). The resultant RMSE conveyed the average generalization error (or expected model accuracy) when applying the models on a test data set that is within the data ranges of the training data set (Hastie et al., 2009 Pg. 242–243). Therefore, in order to evaluate the quality of the generated maps re-lative to the accuracy of the applied models, we observed whether the ranges of training data matched the ranges of model predictions (see Table 4). This was done because, with model extrapolation the expected map accuracy is unknown unless ground-truthing data is available. In contrast, the expected model accuracy is equivalent to the expected map accuracy without model extrapolation. To the advantage of the chosen model validation procedure, the nature of RF regression trees minimizes cases of model extrapolation. We complemented this pro-cedure with visual assessments using local knowledge of stands in the study areas.

Table 4
Model accuracy across scales and structural heterogeneity indices based on a 6-fold cross validation. Assessed LiDAR models are presented in Fig. 7.

	100 m		
	GC	SCI	ESCI
RMSE (%):	13.80	13.74	19.09
Training range:	0.22–0.71	2.96–9.42	7.98–29.59
Prediction range:	0.20–0.70	3.00–9.30	7.50–31.50
	50 m		
	GC	SCI	ESCI
RMSE (%):	14.84	16.87	21.70
Training range:	0.18–0.77	2.49–10.26	5.93–36.11
Prediction range:	0.23–0.64	3.00–9.30	7.50–31.50
	20 m		
	GC	SCI	ESCI
RMSE (%):	24.52	35.59	41.38
Training range:	0.05–0.81	1.18–23.47	1.63–53.68
Prediction range:	0.17–0.66	2.40–15.00	5.5–36.00

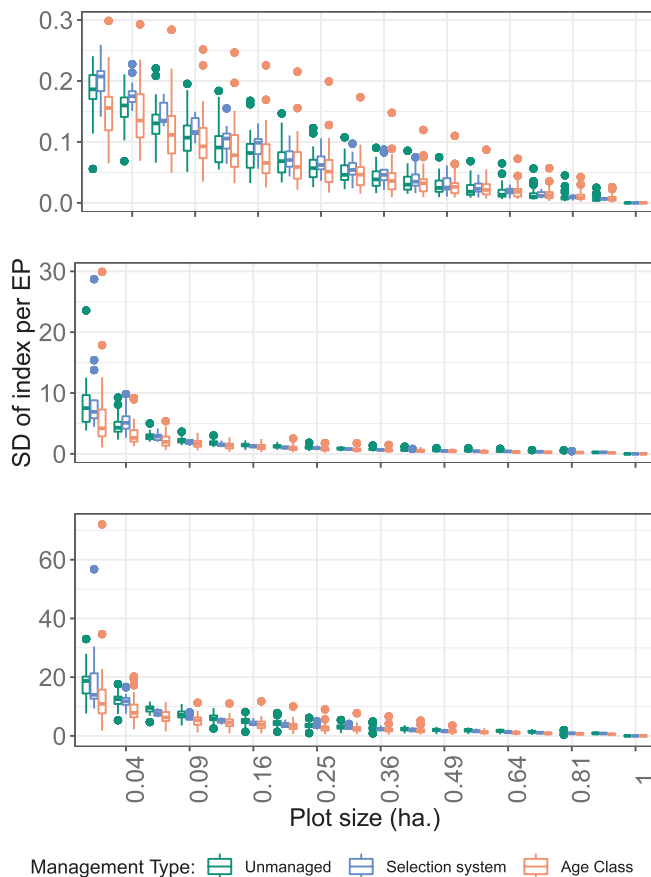**Fig. 3.** Standard deviation (SD) of index values per experimental plot (EP) across plot size. From top to bottom: GC, SCI and ESCI.

3. Results

3.1. Scale-dependence of structural indices across forest management types

Smaller plot sizes increased the variance in estimates of structural indices per plot (Fig. 3). Variance in plot estimates of structure across scales was highest in unmanaged and selection systems compared to age class forests (Fig. 3). As plot size increased, the variance in estimates of structure reduced as estimates stabilized on the reference index magnitude (i.e. at one hectare plot size). Similarly, the scale of observation/plot size affected mean estimates of percentiles of the index distribution. Generally, there was greater deviation (bias) in mean index estimates from the reference index magnitude when smaller plot sizes were employed (Fig. 6). Increasing plot size increased the magnitude of structural heterogeneity when plots were centered in locations of lower heterogeneity compared to their immediate vicinity (such as observed in the minimum percentiles of the distributions), and vice versa when plots were centered in locations of higher heterogeneity than their immediate vicinity (as observed in the maximum percentiles of the index distributions). The bias in index values reduced faster towards the reference estimate in the percentiles of the distribution around the mean (i.e. 75th for GC, 25th for SCI and 50th for ESCI). For changes in all parts of the distribution other than shown in Fig. 6, the reader is referred to Figs. A.10–A.12 in Supplementary material for GC, SCI and ESCI, respectively.

Stabilization of index values on the reference estimate with increasing plot sizes was similar among forest management systems and non-linear in nature (Fig. 4). As mentioned in methods, the point of index stabilization on the reference estimate was identified as a range between observed plot sizes, and identified when the second derivative of the function describing the structure-area-curves was equivalent to

zero. Fig. 4 shows that the point of index stabilization depended on the region of the distribution assessed. For all indices, percentiles around the mean (i.e. 25th and 75th) stabilized at smaller plot sizes compared to regions of the distribution significantly different from the mean. Stabilization across all indices ranged between $900\,m^2$ and $2500\,m^2$ depending on the region of the distribution assessed. All regions of the distribution stabilized at $2500\,m^2$ (Fig. 4). This result was the basis for considering three mapping scales; $400\,m^2$ – considered not stable in all parts of the distribution, and both $2500\,m^2$ and $10,000\,m^2$ – considered stable in all parts of the distribution. For changes in all parts of the distribution other than shown in Fig. 4, the reader is referred to Figs. A.13–A.15 in Supplementary material for GC, SCI and ESCI, respectively.

The overall magnitude of plot size effects changed given the observed part of the index distribution. Fig. 5 shows that generally, plot size effects were more pronounced when plots fell in locations significantly different from the average heterogeneity on the plot and were similar across forest management systems i.e. consistently lowest in age class forests and highest in selection systems and unmanaged forests (Fig. 5).

## 3.2. Index performance across forest management types

Across all studied scales, SCI and ESCI successfully differentiated among age class, selection and unmanaged systems whereas GC failed to differentiate between selection and unmanaged systems shown by the overlap in confidence intervals of the two classes (Figs. 6). Visually, there were greater differences for ESCI than SCI at segregating among selection and unmanaged systems in plots of very high structural heterogeneity (the maximum percentile of the distribution). Changes in all parts of the distribution other than shown in Fig. 6, are presented in Figs. A.10–A.12 in Supplementary material for GC, SCI and ESCI, respectively.

3.3. Variable importance of selected predictors in Random Forest models

A preliminary assessment of bivariate relationships between airborne LiDAR metrics and structural indices shows that, in general, selected predictors exhibited strong relationships with the structural indices (see Table A.5 in Supplementary material), assesed by the Spearman's correlations. Similary, the scale of observation affected the strengths of the bivariate relationships seen through a reduction in correlation values when smaller plot sizes were employed (Table A.5 in Supplementary material). There were differences in variables selected into the final predictor set across scales of observation. Fig. 7 shows all the variables used in prediction of indices at different scales of observation. Cover metrics (B_{7m} and A_{7m}) were consistently important across indices and scales. Similarly, dispersion metrics (Skew, L.skew, AAD, MAD_{mode}, CV) were important for all the three indices but ranked comparatively higher in GC models. Whereas lower height percentiles (P_1 and P_{10}) were important at all observed scales, we note the inclusion of higher height percentiles (P_{99}, Max) into the final predictor set at the 20 m resolution.

3.4. Accuracy of fitted models of structural heterogeneity

Fig. 8 shows scatter plots of observed versus predicted structural heterogeneity values across scales and indices. Table 4 shows the corresponding mean RMSE scores across six cross validation runs per model. Notice that all models were predicted within the data range of the training dataset. Model accuracy increased with increase in plot size. Fig. 8 also shows increasing model bias across SCI and ESCI in unmanaged forests for decreasing plot sizes. A higher bias for ESCI and SCI than for GC is seen at the 20 m resolution. GC also shows the smallest RMSE values for all plot sizes (see Table 4).
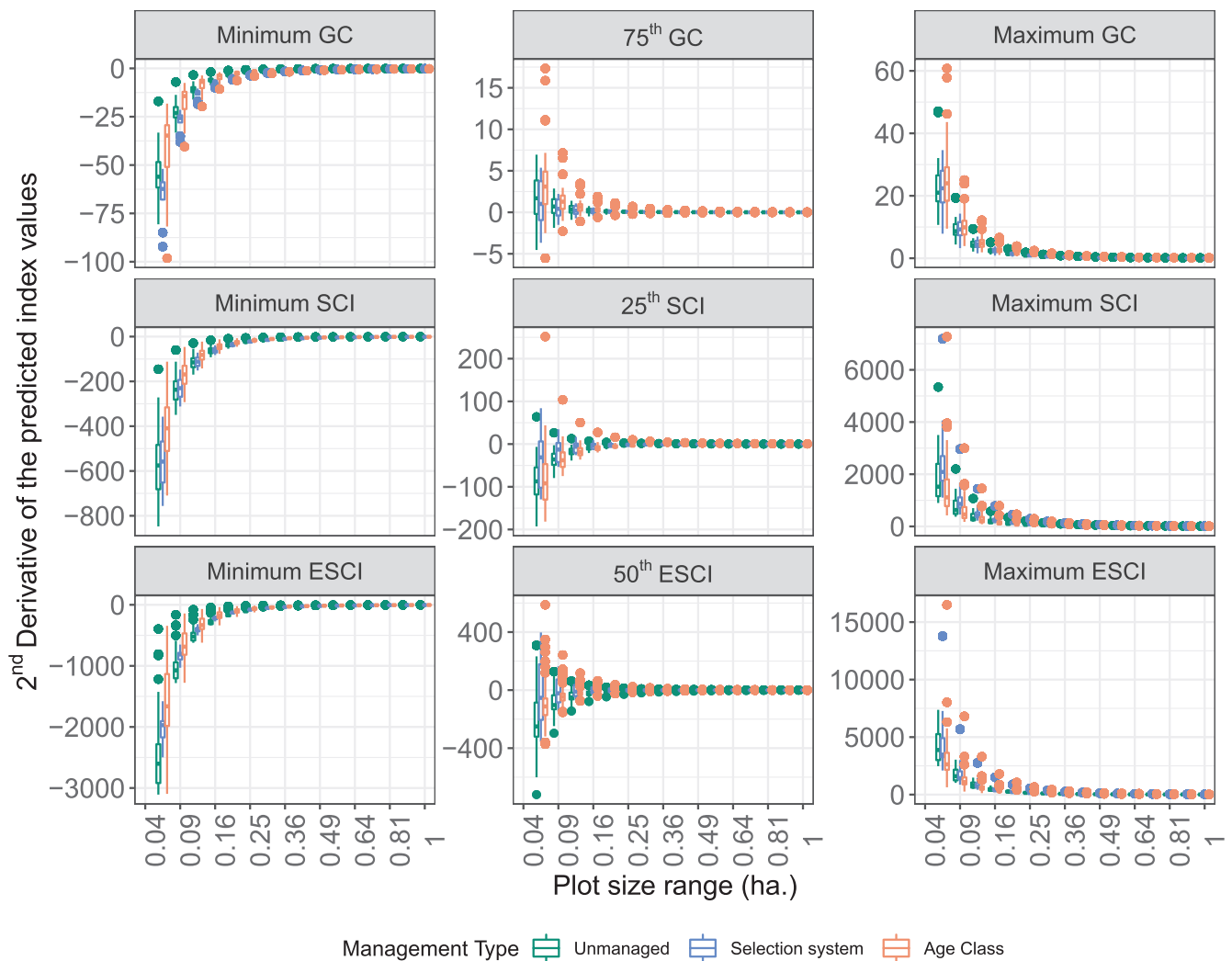
Fig. 4. Box plots of the second derivative of GC, SCI and ESCI per management system and plot size range. The different regions of the index distributions are described in percentiles. Indices stabilized fastest in different parts of the distribution. GC stabilized fastest in the 75th percentile of the distribution whereas SCI and ESCI stabilized fastest in the 25th and 50th percentiles of the distribution, respectively. A total of 2565 structure-area-curves were considered. The second derivative (y-axis) conveys the behavior of the function describing the curve of the predicted index means. The derivative is positive at a point where the curve is approaching its minimum and negative where the curve is approaching its maximum. The point of index stabilization is when the second derivative is equivalent to zero.

3.5. Structural heterogeneity at the study sites

Fig. 9 shows example maps of structural heterogeneity for the three study sites and indices at the identified threshold scale of 50 m resolution. Maps at 100 m and 20 m resolutions are appended (Figs. A.16 and A.17). The colours range from green (low heterogeneity) to red (high heterogeneity). Our maps consistently show higher differentiation capabilities for ESCI and SCI compared to GC at the three study sites. Based on a visual assessment, the 50 m and 100 m resolution maps showed high similarities in spatial structure of predicted structural heterogeneity values compared to maps at a 20 m resolution. This result is consistent with the observed trends of index stabilization in Figs. 4, A.13–A.15. Generally, the maps show that the Hainich-Dün area has more structurally complex stands than both Schorfheide-Chorin and Schwäbische Alb, however, one finds structurally complex stands in all the study areas (Fig. 9).

4. Discussion

4.1. The threshold scale of observation across structural indices and forest management types

In assessment of the threshold plot size to accurately observe the indices, we used a moving window approach and monitored the distributions of index values at each scale of observation. The shapes of the distributions across scales of observation depended on both the number of plots and the area of the plots (Magnussen et al., 2016). Therefore, we kept the number of plots per distribution constant across the evaluated plot sizes so as to remove any sample-size effects in the evaluation of the plot-size effects and in the determination of the plot size threshold. The employed moving window algorithm had the advantage of requiring no edge correction and assisted in efficient computation of airborne LiDAR metrics. In all cases, a census, rather than a sample based approach was employed. As the plot size increased, the amount of overlap between individual plots increased resulting in more correlated between-plot index estimates and thus, a reduction in variance at larger plot sizes.

Our findings show that variance in estimates of structural
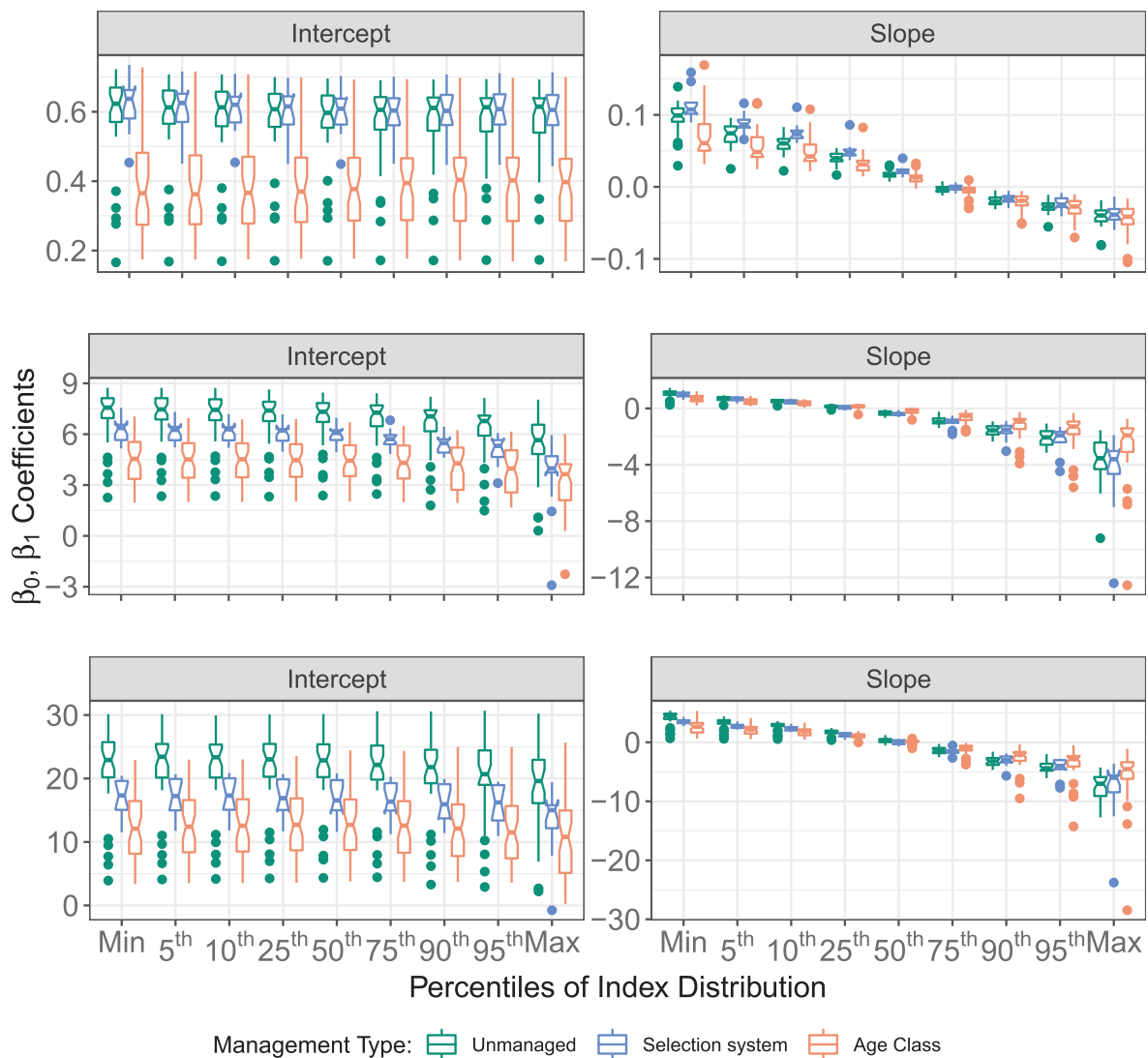
Fig. 5. Box plots of slopes (β_1) and intercepts (β_0) of GC (top), SCI (middle) and ESCI (bottom) per management system. The different regions of the index distributions are described in percentiles on the x-axis. A total of 2565 structure-area-curves were considered. Note that the slope conveys the magnitude of the linearized scale effect and the intercept conveys the degree of separation among forest management systems.

heterogeneity (Figs. 3) increases with a decreasing plot size and is higher in unmanaged and selection management systems than in age class forests. The differences in structural heterogeneity among forest types are a function of the frequency and spatio-temporal patchiness of disturbance events and competitive processes in stands (Coomes and Allen, 2007). These processes eventually determine the spatial arrangement, and the size and distribution of individual trees. Our result shows that it is possible to gain stabilization of structural index values relative to the reference estimate when more variation in structural heterogeneity is accommodated within than between plots at larger plot sizes. This is, by the way, a basic guideline in plot design in forest monitoring: trying to capture a maximum of variability per plot in order to achieve higher precision (equivalent to lower variability between plots). The threshold depends on the structural complexity of the observed stand – i.e. on the monitored region of the distribution (Fig. 4). We identified the plot size threshold of regions of the distribution around the mean and median complexities (i.e 25th, 50th and 75th percentiles) to be between 900 and 1600 m² across all studied forest management types, whereas the other regions of the distribution that are significantly different from the mean and median heterogeneity (e.g. minimum, 5th, 10th, 90th, 95th, maximum) stabilized between

1600 and 2500 m². The identified plot size thresholds compare well with thresholds in other studies such as; 250–450 m² (Adnan et al., 2017) in a boreal managed forest dominated by Scots pine, and 500 m² (Lombardi et al., 2015) in undisturbed beech stands, and both based on the mean across samples, and Zenner, 2005 250 m² based on the median and 2500 m² based on all quantiles of the distribution. Therefore, stand-level estimates of structural heterogeneity can be reliably compared at a 2500 m² scale (Zenner, 2005), for managed temperate forests. This result reinforces the approach by Zenner (2005) that emphasizes monitoring of more than the mean of structural heterogeneity estimates to arrive at a robust stand-level estimate of a threshold scale of observation while considering the differences in magnitude of scale effects in different parts of the distribution (Fig. 5 and Magnussen et al. (2016)). Only the larger common forest inventory plot sizes seem to correspond with plot sizes required for stable observation of the indices as common plot sizes in sample based forest inventories range from 500 to 1000 m². From this point it can be assumed that the presented indicators of structural heterogeneity could be integrated into forest inventory programs. However, if concentric plot designs are used, more difficulties arise since in the field trees are selected based on their sizes. Therefore, new estimators accounting for differences in selection
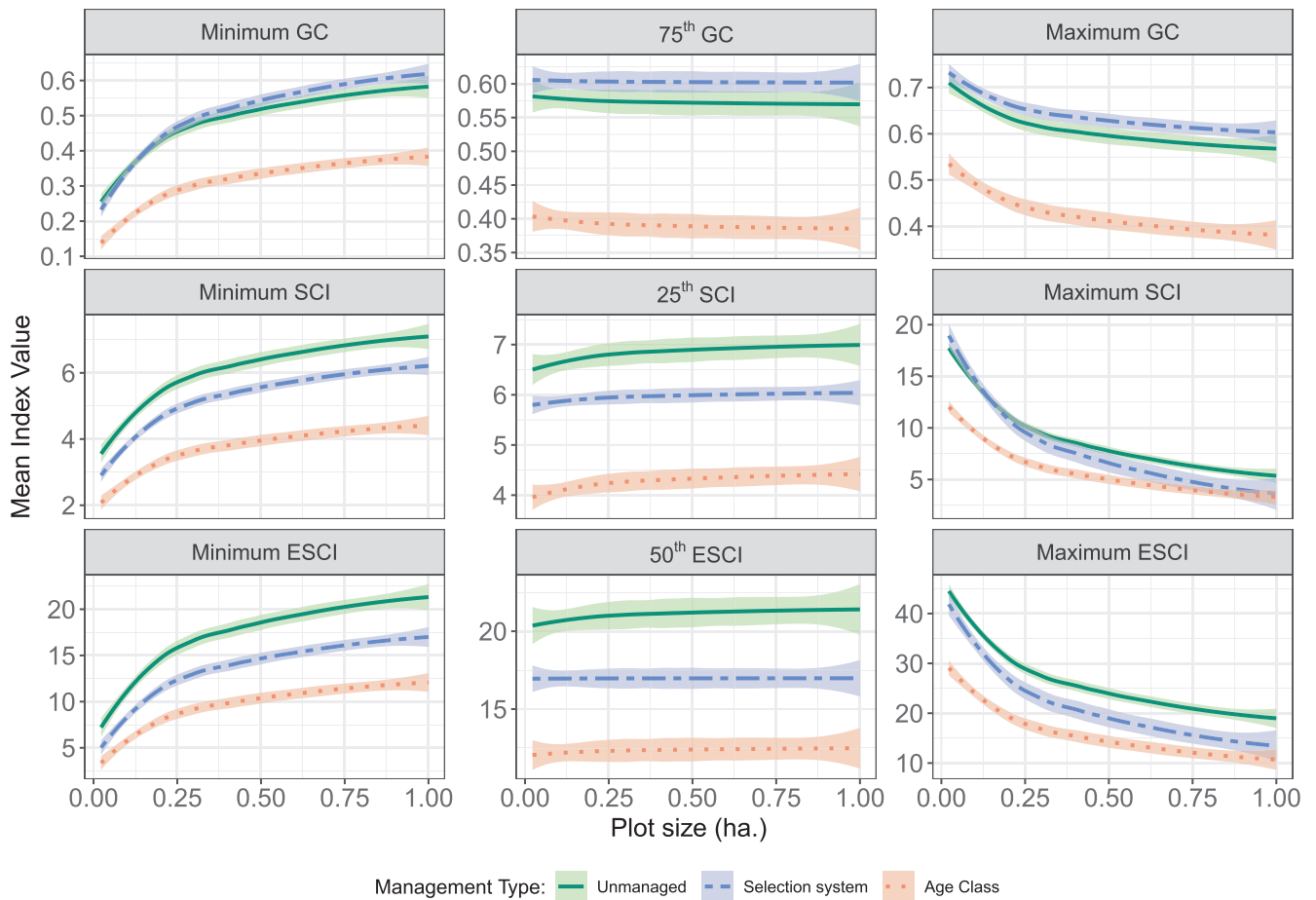
Fig. 6. Mean per management system of the predicted values of structural indices against plot size in ha. The different regions of the index distributions are described in percentiles. Generalized additive models were used to estimate mean index values across structure-area-curves grouped first by percentiles of the index distribution and second by the forest management type. Changes in all parts of the distribution other than shown in this figure are presented in Figs. A.10–A.12 in Supplementary material for GC, SCI and ESCI, respectively.
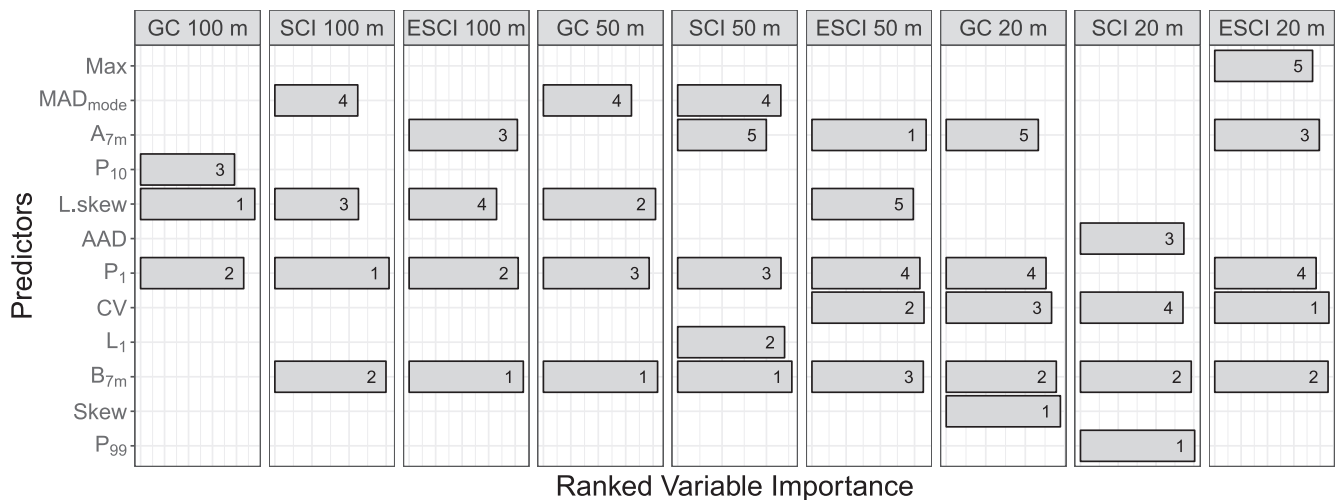
Fig. 7. Variable importance of remotely-sensed predictors in Random Forest models across scales of observation. Variable importance is ranked in an ascending order from the most to the least important variable in each model (i.e. Rank 1 = Most important variable). Computation of variable importance is based on residual sum of squares following the standard permutation procedure of the Random forest model. Since magnitudes of variable importance are not comparable across scales and indices (different models), we do not show the node purity values on the x-axis but the variable importance ranks per model. Please refer to Table 3 for a full list of acronyms.
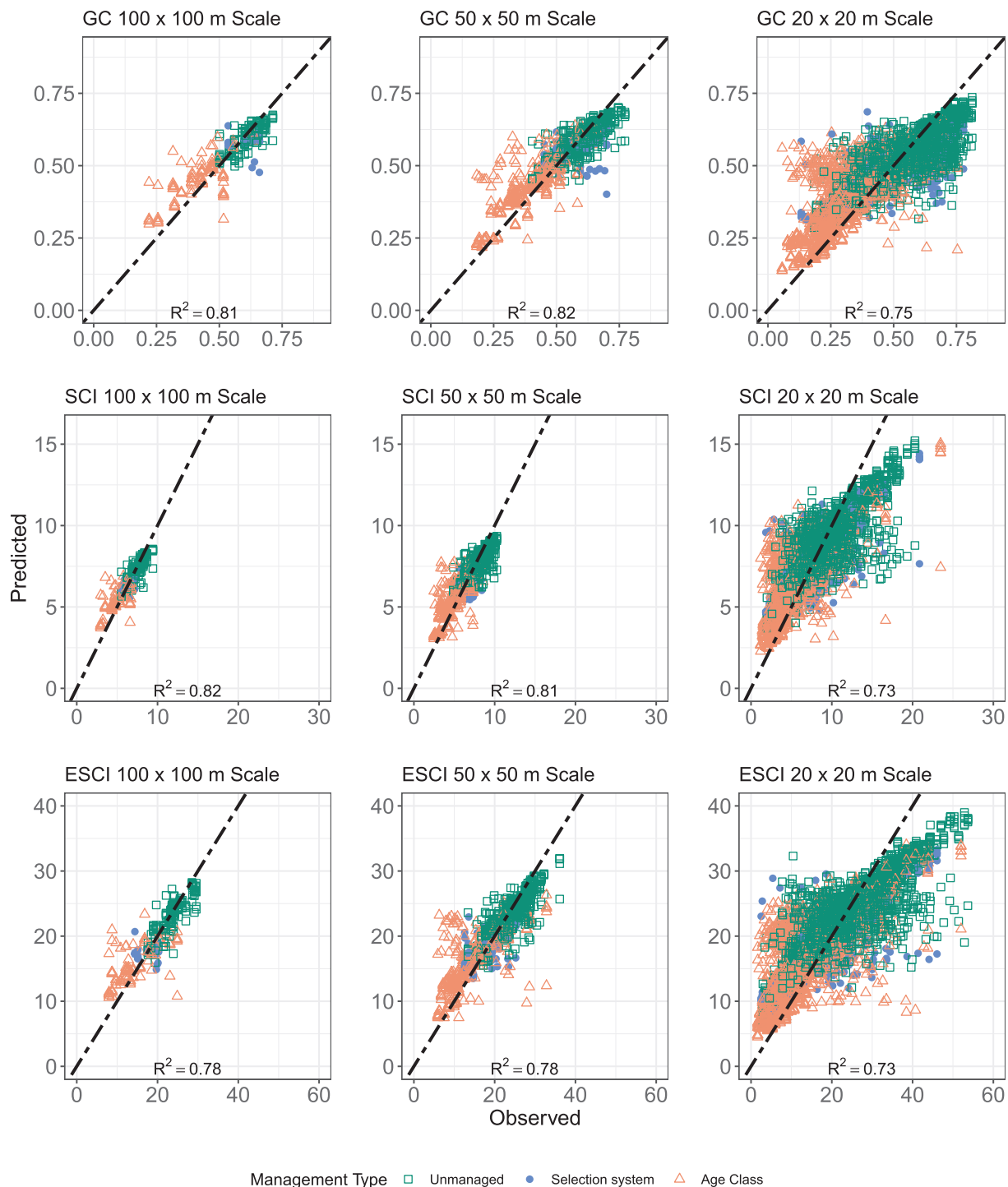
Fig. 8. Predicted vs. observed structural heterogeneity values across scales and indices for the three management systems. The dashed black line represents the 1:1 line. Assessed LiDAR models are presented in Fig. 7.

probabilities would be required or estimates of structural heterogeneity on concentric plots would be biased since smaller trees are only recorded in the smallest subplot.

The generalization of the threshold plot-size findings from this study to different forest conditions needs be done cautiously. This is because unlike most forest attributes that concentrate around the population mean with increasing plot size (Magnussen et al., 2016), an increase in plot size amplifies the probability of finding a differently-sized tree (Adnan et al., 2017) and follows an exponential distribution (Coomes and Allen, 2007); analogous to the problem of comparing estimates of

species richness or diversity collected from samples (Gotelli and Chao, 2013) at different scales of observation. For this reason, generality of threshold plot size findings to estimate heterogeneity in structure remains complex and dependent on the characteristics of the reference population in question. When using spatially explicit indices such as SCI and ESCI, stand summary statistics (e.g. tree density, as suggested by Adnan et al. (2017) may guide deductions on transferability of threshold plot sizes here identified. However, such a conclusion is hard to arrive at for non-spatial indices such as GC given that stands may have similar summary statistics (e.g. tree density) but differing spatial
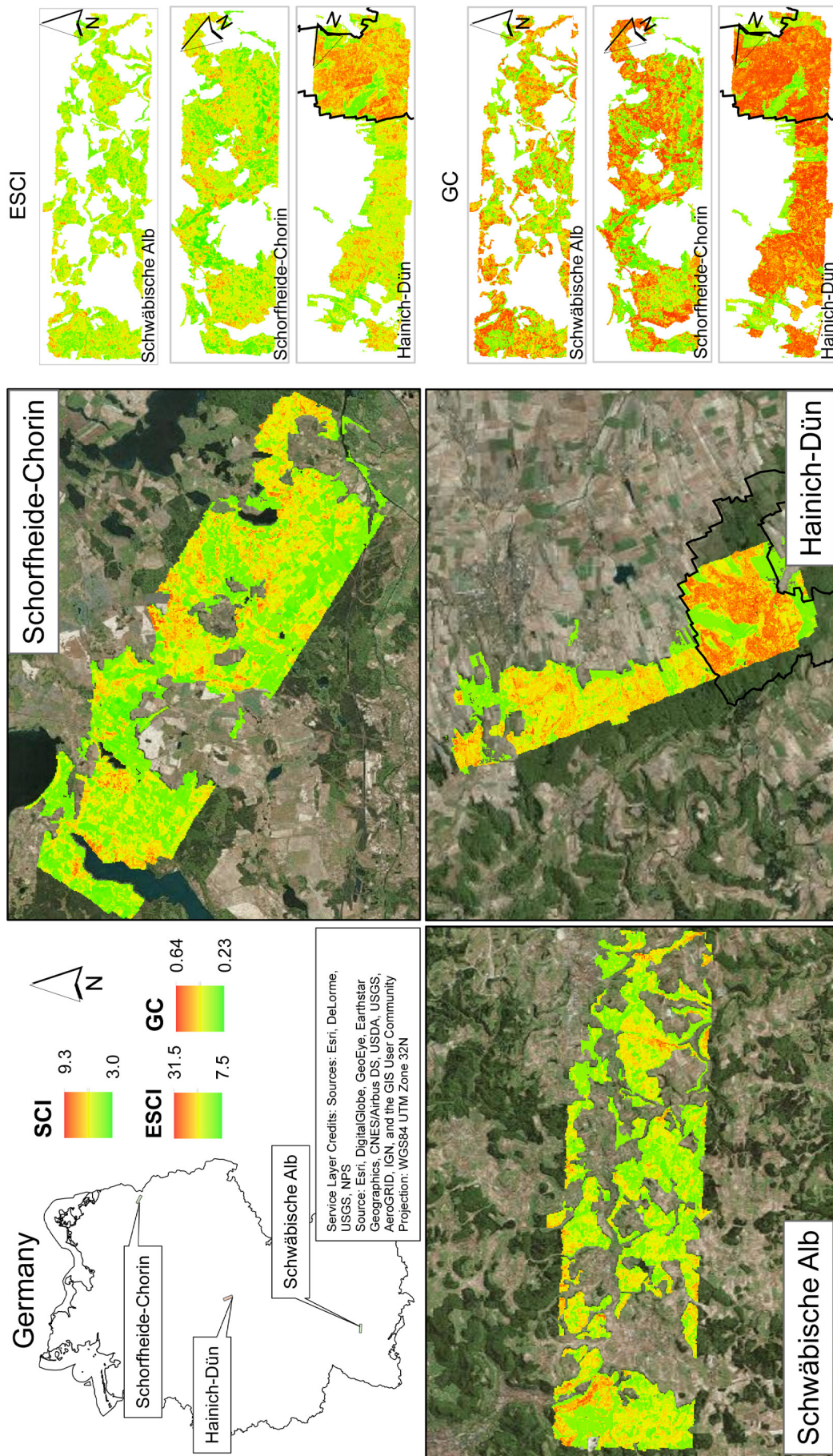
Fig. 9. Maps of SCI for the three study sites. Inset are the maps of ESCI and GC. All maps are at a 50 m resolution (the identified threshold plot size). The colours range from green (low heterogeneity) to red (high heterogeneity). The bold black line shows the boundary of Hainich National Park. Notice that the maps inset have rotated North arrows.

point patterns thus requiring different threshold plot sizes. If one aims at avoiding any plot size effects in different forest conditions as here analyzed and across all indices, plot sizes at least $2500 \, m^2$ are recommended but from a practical point of view such plot sizes are hard to implement in standard forest inventories.

In general, plot size thresholds capture only a facet of the influence of the scale of observation on forest structural heterogeneity indices: since different plot designs may for the same plot sizes and forest conditions, change the plot size thresholds here presented. Common in forest monitoring are circular plots, and in some tropical inventories, rectangular strips of say 20 m * 150 m. Whether it is a requirement to use compact plot designs – as opposed to cluster plots or strips – in order to be able to precisely observe forest structural heterogeneity, will require further study.

4.2. Performance of structural indices across forest management systems

We demonstrate that SCI and ESCI have higher differentiation capabilities compared to GC in uneven-aged stands under selective logging and without management (Fig. 6). Unmanaged stands were on average more complex than managed uneven-aged stands, and both stand types had very high heterogeneity relative to even-aged stands (Fig. 6). On the one hand, this finding supports the perception that unmanaged and managed uneven-aged stands of the study areas are highly similar, highlighting the role of target diameter cuttings in preservation of complex structures and productivity (Bohn and Huth, 2017; Schall et al., 2018). Similarly, Zenner (2004, 2005) show similar results while employing SCI to describe tree size distributions using DBH values in young, transition and old-growth stands. On the other hand, Schall et al. (2018) found contrasting results that managed uneven-aged forests in our study areas were more complex than unmanaged forests with respect to tree height distributions. The contrast in results when using DBH and tree height distributions to quantify heterogeneity is intuitive, since canopies of old-growth unmanaged-beech stands are closed and open-up with selective cuttings, thus in the long run increasing structural heterogeneity as gaps fill up.

Fig. 6 does as well highlight the importance of information on tree spatial patterns in description of stand structural heterogeneity based on tree size distributions. The spatial indices (SCI and ESCI) outperformed the non-spatial index (GC) at separating stands under different management regimes (Fig. 6). Recently, researchers looked into the importance of aggregation of several stand structural attributes in characterization of stand structure among forest types (Schall et al., 2018). Similarly, SCI and ESCI combine more than one structural attribute (horizontal distribution of trees/spatial point pattern, horizontal distinction of tree dimensions, density) and also avail a relatively simple way to quantify stand structural heterogeneity. However, and as further discussed in Section 4.3, it should be noted that combination of more structural attributes into an index does not necessarily guarantee better index performance, especially for prediction from airborne LiDAR data: as shown by worse prediction accuracies for ESCI (Table 4).

We also learn from Fig. 6 that the use of hard thresholds to separate forest management types must be done cautious of the scale of observation. For example, due to a scale-induced bias, the mean GC values were lower than the established 0.5 threshold (Valbuena et al., 2016) in the minimum percentile of selectively logged and unmanaged forests, and the mean GC values were higher than the established 0.5 threshold in the maximum percentile for age class forests when smaller plot sizes were employed. However, the 0.5 threshold sufficed in all parts of the distribution when plot sizes at least $2500 \, m^2$ were employed. This study did not delve into determining hard thresholds for distinguishing among different forest management systems using SCI and ESCI but recommends further research to this regard.

4.3. Prediction of structural indices from airborne LiDAR derivatives across scales of observation

The study results on covariance patterns between structural indices and airborne LiDAR derivatives reinforce previous findings by Valbuena et al. (2013, 2014, 2016) that extensively studied GC relationships with airborne LiDAR derivatives. This study confirms that the three indices are indeed similar in their descriptions of tree size distributions as indicators of structural heterogeneity (Fig. 7 and Table A.5). SCI, ESCI and GC are highly correlated to canopy cover metrics (B_{7m} and A_{7m}), variations in crown heights (AAD, CV), crown dominance (L.Skew), and an interplay of low (gaps or presence of understory; $P_{01}–P_{10}$) and high (dominant height; P_{99} – Max) height percentiles (Valbuena et al., 2013; Adnan et al., 2017). Our results show that the correlations generally decreased in magnitude the smaller the plot size, related to an increase in the noise/uncertainty to signal ratio (Fig. 3). Similarly, depending on the scale of observation, index relationships with some airborne LiDAR predictors may change. For example, cover predictors were consistently very important across scales of observation but presence of gaps and understory explained more variance in structural heterogeneity estimates at large scales whereas tree/crown dominance and dominant height became important descriptors of structural heterogeneity at small scales of observation (Fig. 7 and Table 4).

For prediction from airborne LiDAR, this study demonstrates that higher correlations between SCI, ESCI and airborne LiDAR derivatives compared to GC (Table A.5 in Supplementary material), did not necessarily translate into higher predictability of SCI and ESCI compared to GC (Table 4). Airborne LiDAR metrics aggregated at plot level describe the vertical distribution of crown heights which relates to employed structure metrics that describe either vertical (GC) or both vertical and horizontal (SCI and ESCI) distributions of stems in a stand. Since indices were powered by DBH and DBH is highly correlated with height (Trorey, 1932), all indices correlated well to vertical distributions of crown heights from airborne LiDAR. However, the employed height and cover airborne LiDAR metrics capture only the vertical distribution of canopy heights (integrated over the plot area) whereas SCI, ESCI explicitly capture both the vertical and horizontal distribution of stems in a stand. This may explain why predictability of SCI and ESCI was averagely lower than indicated by the bivariate correlations with airborne LiDAR metrics. Future use of airborne LiDAR metrics like rumple (Kane et al., 2010) may yield better models for SCI and ESCI than presented in this research since such formulations include both vertical and horizontal distribution of LiDAR returns. Similarly, we also note that an inclusion of a tree density term in computation of ESCI affects its estimation from airborne LiDAR (Table 4), consistent with findings by Næsset et al. (2005) on the estimation of stem density from airborne laser data.

In regard to validation of prediction models, there was an increase in prediction bias of SCI and ESCI in locations of high heterogeneity at the 20 m resolution (Fig. 8). These trends are related to a scale mismatch in computation of SCI, ESCI and airborne LiDAR derivatives. The spatial indices are confined to the convex hull of the TIN in computation of both the surface and projection areas of the TIN, whereas the airborne LiDAR metrics are computed based on the entire grid cell area equivalent to the size of the field plot. Since individual trees do not always lie at the boundary of the plot, the projection area of the convex hull connecting individual tree positions may be smaller than the field plot area used to compute airborne LiDAR metrics. For large plot sizes, the ratio between the projection areas of the convex hull and the grid cell is very small and insignificant, increasing at small scales of observation and high levels of heterogeneity and thus introducing a bias as seen at the 20 m resolution (Fig. 8). Further research will be required on ways to account for this scale mismatch.

4.4. The relevance of maps of structural heterogeneity

Maps of structural heterogeneity are important to continuously identify areas along a gradient of high and low structural heterogeneity that may eventually correspond to differences in stand management history and management intensity (Valbuena et al., 2016), or serve as indicators of disturbance and degradation (Mitchell et al., 2017), productivity and ecosystem functioning (Dănescu et al., 2016), or habitat and biodiversity (Pasher and King, 2011). For example, our maps reveal interesting patterns of structural heterogeneity in forests of the study areas that can be linked to their management history (see Section 2.2); stands within the Hainich National Park (in the southern part of the Hainich area, Fig. 9) were identified as the most structurally heterogeneous stands in the three study areas, comparatively consistent with management histories at the areas. In line with biodiversity monitoring, the maps may describe the habitat features for species' niches at scales difficult to cover through survey sampling (Estes et al., 2010), while at the same time characterizing structural heterogeneity across space in ways accounting for differences in productivity and ecosystem functioning (Dănescu et al., 2016) especially when combined with species distribution maps. In regard to forest planning and management, our maps may help in defining areas to focus silvicultural activities. Lastly and in the context of forest monitoring, multi-date maps of changes in structural heterogeneity may contribute information on forest growth and degradation typically manifesting through a change in forest structure (Mitchell et al., 2017) or may work as auxiliary data in forest inventory and biomass modelling. Rödig et al. (2017) recently show that precise estimation of spatial variation in tropical forest biomass requires to consider both small and large scale variations in forest structure.

5. Conclusions

We draw the following conclusions from this study;

1. The threshold plot size for enumerating SCI, ESCI and GC depends on the monitored region of the distribution between 900 m^2 and 2500 m^2, in forests similar to those examined in this study. Therefore, stand-level estimates of structural heterogeneity can be reliably compared and mapped from airborne LiDAR data at a 2500 m^2 scale.

2. SCI and ESCI have higher differentiation capabilities than GC among stands of target diameter selection and stands without management in the study area. In general, unmanaged stands in the study area were more structurally heterogeneous compared to selection system and age class forests in sequence. We therefore can conclude that the inclusion of spatial point patterns in description of forest structural heterogeneity enhances index ability to separate forest structure in the three management systems across plot sizes.

3. SCI, ESCI and GC can be reliably mapped from airborne LiDAR data derivatives on canopy cover (B_{7m} and A_{7m}), variations in crown heights (AAD, CV), crown dominance (L.Skew), and an interplay of low (gaps or presence of understory; P_{01}–P_{10}) and high (dominant height; P_{99} – Max) height percentiles at a threshold scale of 2500 m^2. In mapping of the indices from airborne LiDAR, inclusion of data on tree positions into SCI and ESCI did not give the two spatial indices a competitive advantage over non-spatial GC.

6. Outlook

This study's findings are important, especially as the links between forest structure and disturbance, degradation, biodiversity, productivity and ecosystem functioning, are increasingly understood. Due to the current costs involved in large area LiDAR acquisitions, efficient mapping of structural heterogeneity is still experimental and small scale. However, opportunities to obtain less expensive information on forest height, cover and biomass to map structural heterogeneity at local to global scales exist. For example; the fusion of photogrammetric point cloud data with high-resolution digital elevation model datasets, and NASA's GEDI satellite mission. Similarly, the potential of radar tomography to provide forest height and canopy cover information needs to be further researched in the context of mapping forest structural heterogeneity at wide scales.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ecolind.2019.02.056.

References

Adnan, S., Maltamo, M., Coomes, D., Valbuena, R., 2017. Effects of plot size, stand density and scan density on the relationship between airborne laser scanning metrics and the Gini coefficient of tree size inequality. Can. J. For. Res. 47, 1590–1602.

Beckschäfer, P., Mundhenk, P., Kleinn, C., Ji, Y., Yu, D., Harrison, R., 2013. Enhanced structural complexity index: an improved index for describing forest structural complexity. Open J. For. 3, 23–29.

Besag, J., 1977. Contribution to the discussion of Dr Ripley's paper. J. R. Stat. Soc. B 39 (2), 193–195.

Bohn, F.J., Huth, A., 2017. The importance of forest structure to biodiversity-productivity relationships. R. Soc. Open Sci. 4 (1), 160521 .

Butler-Manning, D., 2007. Stand structure, gap dynamics and regeneration of a semi-natural mixed beech forest on limestone in central europe: a case study (Tech. rep.).

Coomes, D.A., Allen, R.B., 2007. Mortality and tree-size distributions in natural mixed-age forests. J. Ecol. 95 (1), 27–40.

del Río, M., Pretzsch, H., Alberdi, I., Bielak, K., Bravo, F., Brunner, A., Condés, S., Ducey, M.J., Fonseca, T., von Lüpke, N., Pach, M., Peric, S., Perot, T., Souidi, Z., Spathelf, P., Sterba, H., Tijardovic, M., Tomé, M., Vallet, P., Bravo-Oviedo, A., 2016. Characterization of the structure, dynamics, and productivity of mixed-species stands: review and perspectives. Eur. J. For. Res. 135 (1), 23–49. https://doi.org/10.1007/s10342-015-0927-6.

Díaz-Yáñez, O., Mola-Yudego, B., Ramón González-Olabarria, J., Pukkala, T., 2017. How does forest composition and structure affect the stability against wind and snow? For. Ecol. Manage. 401, 215–222. URL:http://www.sciencedirect.com/science/article/pii/S0378112717302104.

Dănescu, A., Albrecht, A.T., Bauhus, J., 2016. Structural diversity promotes productivity

of mixed, uneven-aged forests in southwestern Germany. Oecologia 182 (2), 319–333.

Ehbrecht, M., Schall, P., Ammer, C., Seidel, D., 2017. Quantifying stand structural complexity and its relationship with forest management, tree species diversity and microclimate. Agric. For. Meteorol. 242 (Suppl. C), 1–9.

Eilers, P., Marx, B., Durbán, M., 2015. Twenty years of p-splines. Stat. Oper. Res. Trans. 39, 149–186.

Estes, L.D., Reillo, P.R., Mwangi, A.G., Okin, G.S., Shugart, H.H., 2010. Remote sensing of structural complexity indices for habitat and species distribution modeling. Remote Sens. Environ. 114 (4), 792–804. https://doi.org/10.1016/j.rse.2009.11.016.

Fischer, M., Bossdorf, O., Gockel, S., Hnsel, F., Hemp, A., Hessenmller, D., Korte, G., Nieschulze, J., Pfeiffer, S., Prati, D., Renner, S., Schning, I., Schumacher, U., Wells, K., Buscot, F., Kalko, E.K., Linsenmair, K.E., Schulze, E.D., Weisser, W.W., 2010. Implementing large-scale and long-term functional biodiversity research: the biodiversity exploratories. Basic Appl. Ecol. 11, 473–485.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., 2016. VSURF: variable selection using random forests. R package version 1.0.3.

Gotelli, N.J., Chao, A., 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. vol. 5.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer ISBN: 978-0-387-84857-0.

Hédl, R., Svátek, M., Danák, M., Rodzay, A., Salleh, A., Kamariah, A., 2009. A new technique for inventory of permanent plots in tropical forests: a case study from lowland dipterocarp forest in kuala belalong, brunei darussalam. Blumea 54, 124–130.

Hessenmöller, D., Nieschulze, J., von Lpke, N., Schulze, E.-D., 2011. Identification of forest management types from ground-based and remotely sensed variables and the effects of forest management on forest structure and composition. Forstarchiv 82, 171–183.

Holzwarth, F., Kahl, A., Bauhus, J., Wirth, C., 2013. Many ways to die – partitioning tree mortality dynamics in a near-natural mixed decidous forest. J. Ecol. 101, 220–230.

Isenburg, M., 2017. Lastools – efficient lidar processing software. Version 161127. Academic, rapidlasso GmbH.

Jeronimo, S., Kane, V., Churchill, D.J., Mcgaughey, R., Franklin, J.F., 2018. Applying lidar individual tree detection to management of structurally diverse forest landscapes. J. For. 116, 336–346.

Juchheim, J., Ammer, C., Schall, P., Seidel, D., 2017. Canopy space filling rather than conventional measures of structural diversity explains productivity of beech stands. For. Ecol. Manage. 395 (Suppl. C), 19–26.

Kane, V.R., McGaughey, R.J., Bakker, J.D., Gersonde, R.F., Lutz, J.A., Franklin, J.F., 2010. Comparisons between field- and lidar-based measures of stand structural complexity. Can. J. For. Res. 40 (4), 761–773. https://doi.org/10.1139/X10-024.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22. URL:http://CRAN.R-project.org/doc/Rnews/.

Lombardi, F., Marchetti, M., Corona, P., Merlini, P., Chirici, G., Tognetti, R., Burrascano, S., Alivernini, A., Puletti, N., 2015. Quantifying the effect of sampling plot size on the estimation of structural indicators in old-growth forest stands. For. Ecol. Manage. 346, 89–97. https://doi.org/10.1016/j.foreco.2015.02.011.

Magnussen, S., Mandallaz, D., Lanz, A., Ginzler, C., Næsset, E., Gobakken, T., 2016. Scale effects in survey estimates of proportions and quantiles of per unit area attributes. For. Ecol. Manage. 364, 122–129.

Magurran, A., 2004. Measuring Biological Diversity. Blackwell Science.

Manzanera, J.A., García-Abril, A., Pascual, C., Tejera, R., Martín-Fernández, S., Tokola, T., Valbuena, R., 2016. Fusion of airborne lidar and multispectral sensors reveals synergic capabilities in forest structure characterization. GISc. Remote Sens.

McElhinny, C., Gibbons, P., Brack, C., Bauhus, J., 2005. Forest and woodland stand structural complexity: its definition and measurement. For. Ecol. Manage. 218 (1–3), 1–24.

McGaughey, R., 2014. FUSION/LDV: Software for Lidar Data Analysis and Visualization, v. 3.42. USDA Forest Service, Pacific Northwest Research Station (Seattle, WA).

Mitchell, A.L., Rosenqvist, A., Mora, B., 2017. Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD +. Carbon Balance Manage. 12 (1), 9. URL:http://cbmjournal.springeropen.com/articles/10.1186/s13021-017-0078-9.

Næsset, E., Bollandsås, O.M., Gobakken, T., 2005. Comparing regression methods in estimation of biophysical properties of forest stands from two different inventories using laser scanner data. Remote Sens. Environ. 94 (4), 541–553.

Pasher, J., King, D.J., 2011. Development of a forest structural complexity index based on multispectral airborne remote sensing and topographic data. Can. J. For. Res. 41 (1), 44–58. URL:http://www.nrcresearchpress.com/doi/abs/10.1139/X10-175.

Pommerening, A., 2002. Approaches to quantifying forest structures. Forestry 75 (3), 305–323.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL:https://www.R-project.org/.

Ripley, B.D., 1977. Modelling spatial patterns. J. R. Stat. Soc. Ser. B (Methodological) 39 (2), 172–212. https://doi.org/10.2307/2984796.

Rödig, E., Cuntz, M., Heinke, J., Rammig, A., Huth, A., 2017. Spatial heterogeneity of biomass and forest structure of the Amazon rain forest: linking remote sensing, forest modelling and field inventory. Glob. Ecol. Biogeogr. 26 (11), 1292–1302.

Schall, P., Ammer, C., 2013. How to quantify forest management intenisty in Central European forests. Eur. J. Forest Res. 132, 379–396.

Schall, P., Schulze, E.D., Fischer, M., Ayasse, M., Ammer, C., 2018. Relations between forest management, stand structure and productivity across different types of Central European forests. Basic Appl. Ecol.

Solly, E., Schning, I., Muller, J., Socher, S., Trumbore, S., Schrumpf, M., 2013. Mean age of carbon in fine roots from temperate forests and grasslands with different management. Biogeisciences 10 (7).

Trorey, L., 1932. A mathematical method for the construction of diameter height curves based on site. For. Chronicle 8, 121–132.

Valbuena, R., 2015. Forest structure indicators based on tree size inequality and their relationships to airborne laser scanning (Ph.D. dissertation). University of Eastern Finland.

Valbuena, R., Packalén, P., Martín-Fernández, S., Maltamo, M., 2012. Diversity and equitability ordering profiles applied to the study of forest structure. For. Ecol. Manage. 276, 185–195.

Valbuena, R., Maltamo, M., Martín-Fernández, S., Packalén, P., Pascual, C., Nabuurs, G., 2013. Patterns of covariance between airborne laser scanning metrics and lorenz curve descriptors of tree size inequality. Can. J. Remote Sens. 39 (1), 18–31.

Valbuena, R., Vauhkonen, J., Packalén, P., Pitkänen, J., Maltamo, M., 2014. Comparison of airborne laser scanning methods for estimating forest structure indicators based on lorenz curves. ISPRS J. Photogrammetry Remote Sens. 95, 23–33.

Valbuena, R., Eerikäinen, K., Packalen, P., Maltamo, M., 2016. Gini coefficient predictions from airborne lidar remote sensing display the effect of management intensity on forest structure. Ecol. Ind. 60, 574–585.

Weiner, J., 1985. Size heirarchies in experimental populations of annual plants. Ecology 66 (3), 743–752.

Weiner, J., Solbrig, O., 1984. The meaning and measurement of size hierarchies in plant-populations. Oecologia 61, 334–336.

Wood, S., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. (B) 73 (1), 3–36.

Zenner, E., 2004. Does old-growth forest imply high live-tree structural complexity? For. Ecol. Manage. 195, 243–258.

Zenner, E., 2005. Investigating scale-dependent stand heterogeneity with structure-area-curves. For. Ecol. Manage. 209, 87–100.

Zenner, E., Hibbs, D., 2000. A new method for modelling the heterogeneity of forest structure. For. Ecol. Manage. 129, 75–87.

Zenner, E.K., Peck, J.E., Hobi, M.L., Commarmot, B., 2015. The dynamics of structure across scale in a primaeval European beech stand. Forestry 88 (2), 180–189.

CrossMark

A unified framework for land cover monitoring based on a discrete global sampling grid (GSG)

Lutz Fehrmann (ORCID) **· Collins B. Kukunda ·**
Nils Nölke · Sebastian Schnell · Dominik Seidel ·
Steen Magnussen · Christoph Kleinn

Abstract Environmental monitoring and assessment of the extent and change of land uses and their renewable natural resources over time is a key element in many international processes and one crucial basis for sustainable management. Remote sensing plays an increasingly important role in these monitoring systems, especially if the interest is in large areas. Integration of remote sensing requires comprehensive and careful preprocessing and a high level of expertise which is not always at hand in all applications. However, easy-to-implement sampling techniques based on visual interpretation are an alternative approach for utilizing remote sensing imagery, including the evolving archives of georeferenced and preprocessed data provided by virtual globes like Google Earth, Bing, and others. The goal of this paper is to propose a simple unified framework that may be used in the context of sampling studies and environmental monitoring from local to global scale. Besides the definition of a sampling design, the observation or plot design, i.e., defining how observations are to be made and recorded, has a strong influence on the precision of estimates as well as the overall efficiency of a sampling exercise. As an example, we present a simulation study focusing on the estimation of forest cover in artificial landscapes with different coverage and degree of fragmentation. The sampling units we compare are point clusters with different configuration and spatial extent.

L. Fehrmann (✉) · C. B. Kukunda · N. Nölke · C. Kleinn
Forest Inventory and Remote Sensing, Georg-August
Universität Göttingen, Göttingen, Germany
e-mail: lfehrma@gwdg.de

C. B. Kukunda
e-mail: ckukund@gwdg.de

N. Nölke
e-mail: nnoelke@gwdg.de

C. Kleinn
e-mail: ckleinn@gwdg.de

S. Schnell
Thünen Institute of Forest Ecosystems, Eberswalde, Germany
e-mail: sebastian.schnell@thuenen.de

D. Seidel
Silviculture and Forest Ecology of the Temperate Zones,
Georg-August Universität Göttingen, Göttingen, Germany
e-mail: dseidel@gwdg.de

S. Magnussen
Pacific Forestry Centre, Natural Resources Canada, Ottawa,
Canada
e-mail: steen.magnussen@canada.ca

Introduction

Motivation

Monitoring of land cover status and change is a key feature in many international and national processes

related to ecological and environmental challenges. The combat against deforestation and forest degradation (REDD+) is but one example showing that monitoring of the state and change of forest resources is an essential basis for performance-based payments that should compensate countries for their efforts in avoiding or reducing deforestation and forest degradation (Gibbs et al. 2007).

Different forest cover or carbon stock benchmark maps have been generated in order to define global deforestation levels, in particular for tropical regions (Bastin et al. 2017; Hansen et al. 2013; Saatchi et al. 2011; Achard et al. 2002, 2007). Most studies on global or pan-tropical scale are based on wall-to-wall remote sensing analysis, which is regarded as the only feasible approach to estimate deforestation rates over large areas (Achard et al. 2007; Beuchle et al. 2011).

The increasing availability of large archives of freely accessible high resolution and georeferenced imagery through virtual globes like Google Earth, Microsoft Virtual Earth (Bing), and others, appears to be under-exploited by scientific applications in the context of global environmental monitoring (Potere 2008; Yu and Gong 2012). Scientists of many disciplines are using technologies such as Google Earth, mainly as a tool for representation, dissemination, and validation of their results, but only to a limited extent also as a data source (Sheppard and Cizek 2009; Sun et al. 2012; Cracknell et al. 2013; Hu et al. 2013). Google Earth, for example, provides high resolution overview imagery with varying resolution on a global scale and the coverage is expected to further increase in future. Even if the original spectral values of the imagery are reduced to rgb (red, green, blue) only, observations of some target variables may be made from a sample-based visual interpretation of small image subsets, like, e.g., proposed by Stehman et al. (2003, 2005); Ramezani and Holm (2011); Fehrmann et al. (2014, 2015).

Google Earth and Bing imagery have proven to be a useful basis for sampling studies for different purposes. Ploton et al. (2012) demonstrated that texture and above ground biomass is about equally correlated in IKONOS and Google Earth/Bing imagery. In case that current imagery is available at suitable resolution, two-phase sampling approaches with visual interpretation of sampling locations in a first phase and field data collection in a second phase can increase the overall efficiency of sampling studies (Barrett et al. 2016).

A probabilistic sample from high resolution imagery is also a good basis for the validation of global land cover products (Stehman 1999; Olofsson et al. 2012; Pengra et al. 2015).

One prerequisite, however, is a scientifically sound sampling framework allowing for unbiased statistical estimation and a straightforward observation protocol to be applied on different spatial scales. Our aim is to propose such a framework in the form of a simple and scalable global grid system for environmental monitoring at any geographical reference area (population). While such a global grid system might be basis for the implementation of different sampling designs, we are here not aiming at comparisons of different sampling procedures, but restrict ourselves to a presentation of the grid system and a simulation evaluating different observation designs for the purpose of estimating the proportion of land cover for different target classes. In our example, we focus on the estimation of forest cover in a binary forest/nonforest map, but in principle, our findings would equally hold for other land cover types or applications of the target variable "cover" also considering more than two classes.

Discrete grid systems for environmental monitoring

Many different global grid systems for sampling on a global scale were proposed. Most attempt to tessellate the surface of a spherical approximation of the earth surface into a finite population of equally sized non-overlapping tiles arranged in a quasi systematic way. The underlying assumption was that randomization of the grid requires that it be regular and retain equal-area cells when projected on the surface of the earth. White et al. (1992); Kimerling et al. (1999); Richards et al. (2000); Sahr et al. (2003); Esseen et al. (2006); Sahr (2011); Song et al. (2002); Swinbank and Purser (2006); Youngren and Petty (2017) give a comprehensive overview of the complexity of constructing discrete grid systems satisfying the above criteria on a spherical projection of the Earth's surface. Wickman et al. (1974) outlined the construction of a system of domains for global sampling problems and listed desirable characteristics, like equal area and shape of polygonal domains that can be subdivided into a discrete grid system. They also acknowledge that a complete systematic grid on a realistic projection of the Earth's surface is not available. Their argument for the mentioned characteristics was to

avoid a potential estimation bias resulting from systematic over- or underrepresentation of parts of the Earth's surface. The seemingly trivial task of distributing points on a sphere, such that they are all the same distance apart and form the center points of a network of equally sized Thiessen polygons has not yet been solved so far (Richards et al. 2000). Even if such a tessellation would be available, it would refer to a spherical projection of the Earth's surface exclusively. Further, the construction of such grid systems is difficult to communicate and, as proposed by Sahr (2011), it is not always guaranteed that grids at different resolutions have a regular hierarchical relationship that can be exploited during analysis.

From the perspective of statistical sampling the construction of a finite population of a discrete number of completely systematically arranged tiles (sampling elements) that do not overlap and cover the target area without gaps, as suggested by other authors, is neither necessary nor efficient. In such cases, design-based sampling is rather developed under the assumption of an infinite population of dimensionless points (the continuum of the study area), from which sampling locations are selected based on a stipulated statistical sampling design (Gregoire and Valentine 2007; Mandallaz 2007). The advantage of a systematic sampling grid is that the share of different land cover classes is on average maintained in such a balanced sample (Jr. and Olsen 2004; Theobald et al. 2007; Lister and Scott 2009). Further, a systematic design with fixed distance between sampling locations ensures that the total area of interest is evenly covered and represented in the sample and that a defined minimum distance between sampling locations is given. A systematic approach is easy to communicate and of higher precision compared to simple random sampling in most cases. However, to maintain the favorable characteristics of systematic sampling on the precision of estimation, the homogeneous sampling intensity and even distribution of sampling locations is much more important than a completely constant distance between the individual sample locations (Fattorini et al. 2009).

A couple of examples exist, where global grid systems have been applied in sample-based remote sensing studies . In context of the Global Forest Resource Assessment (FRA) of the Food and Agricultural Organization of the United Nations (FAO), a global remote

sensing survey (FRA-RSS) of forest cover change based on an analysis of multitemporal remote sensing imagery has been completed FAO (2009, 2010). It was assumed that the intersections of each integer degree of longitude and latitude in a geographic coordinate system would results in a "systematic sampling design" on the Earth's surface. Accordingly, tiles of the Global Land Survey (GLS) dataset were selected as observation units at these locations (Beuchle et al. 2011). The sampling grid covers the latitude range between 75° north/south while the intensity is reduced above 60° north/south by selecting every second intersection only. Recently, FAO provided a new tool for assessing land cover with free and open-source software (Collect Earth) that is also based on a dense lat/long grid (0.04°), but can handle any other predefined grid.

However, even if there is a certain systematic procedure in the selection of sampling locations from the lat/lon grid, it is obvious that *systematic sampling* refers to different characteristics. As a result of the geographic departure, the distances of meridians along parallels of latitude are decreasing with increasing latitude and converge to zero at the poles. While the distance between resulting sampling locations along the latitudes is approximately 111 km near the equator, it is reduced to about 56 km at 60° north/south latitude. Therefore, even if the sample size per latitude is constant with this approach (below 60°), the sampling intensity in terms of sample points per area, is dramatically changing with increasing latitude (Fig. 1).

As a consequence, the estimation of any target variable and its variance for areas with a notable north-south extent, or even for the tropics or globally, becomes a methodological challenge that needs to consider unequal inclusion probabilities. The objective of this study is to overcome the problem of gradually changing sampling intensities when using the lat/lon grid and to propose a new grid system which maintains equal distances of point locations along latitudes, ensuring unbiased estimation from local to global scale. While the basic principle of the grid was briefly described in Fehrmann (2015), we here also present a simulation study on the optimization of observation units for the estimation of land cover.
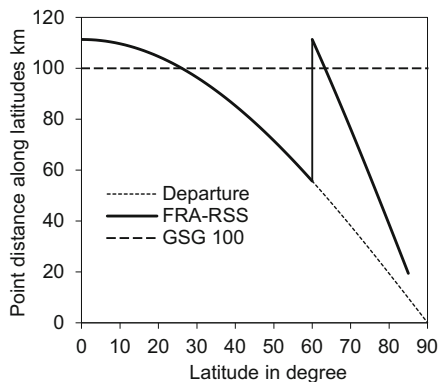
Fig. 1 Distance between sampling locations along parallels as a function of latitude resulting from the global sampling grid of FRA-RSS (solid line) and the proposed GSG, here with a 100-km point distance (horizontal dashed line). Sampling intensity of FRA-RSS grid is reduced above 60° latitude, north and south, explaining the bump in interpoint distances from 60° latitude (Fehrmann 2015)

Methods

The basic concept of the global sampling grid (GSG) we are proposing is easy to implement and to communicate. The main characteristics of the GSG are as follows: (1) the sampling intensity in terms of sampling locations per area is nearly uniform, (2) the spherical distance between sampling locations along latitudes is constant, and (3) the distance between circles of latitudes is constant. All three criteria ensure the interpretability of sampling studies as well as unbiased estimation. As the grid system is constructed on a spherical projection of the earth surface, small distortions are resulting from projecting the grid on the WGS84 or other ellipsoids.

The grid construction follows a simple approach (Fig. 2): circles of latitudes are placed with a constant distance on the surface of a sphere in north and south direction starting from the equator. The sphere has a radius of 6378.137 m (major axis of the World Geodetic System ellipsoid, WGS84). On each of these latitudes, sampling locations are placed in the same constant distance starting from the Greenwich meridian in est and east direction. This leads to a semi systematic arrangement of sampling locations, where the distance of points is constant along the latitudes.

All grids generated based on this system have the same origin, and sampling locations of affine grids with different resolutions coincide. Figures 3 and 4
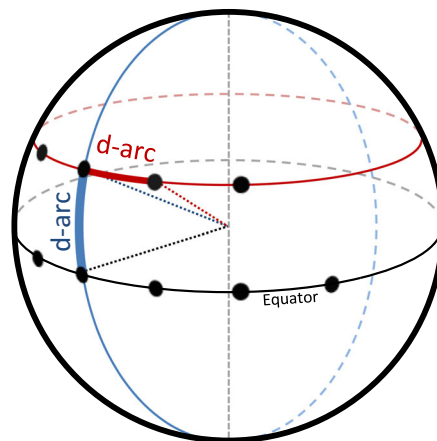
Fig. 2 Construction of the GSG. Points are placed in equal distance along latitudes in west and east direction starting from the null meridian (d-arc = spherical distance)

show the global GSG 250 grid (distance between latitudes and points along latitudes of 250 km) and country subsets.

If the GSG is draped over a spheroid, it will obviously have an irregular seam on the opposite side of the start meridian leading to an irregular sampling intensity in a narrow corridor that is affecting a small land mass in East Russia (Chukotka). However, this is not relevant for an application to the rest of the land surface (Fig. 3).

Simulating different observational designs

The GSG provides a simple and scalable systematic global sampling design that can be used for different purposes. The main idea for application in this study lies in sampling high resolution remote sensing imagery as provided by virtual globes through visual interpretation. A typical scenario, relevant in many contexts, might be the estimation of coverage for different land cover classes. For this purpose, a suitable observation design needs to be planned. If observations refer to dimensionless points, this also includes a clear definition of a reference area around point locations that is used to fell the decision about the respective land cover.

In a first example, we used the single points of the GSG as observation units to estimate the global land surface area. The existence of land at individual sample locations was queried from the *global administrative*
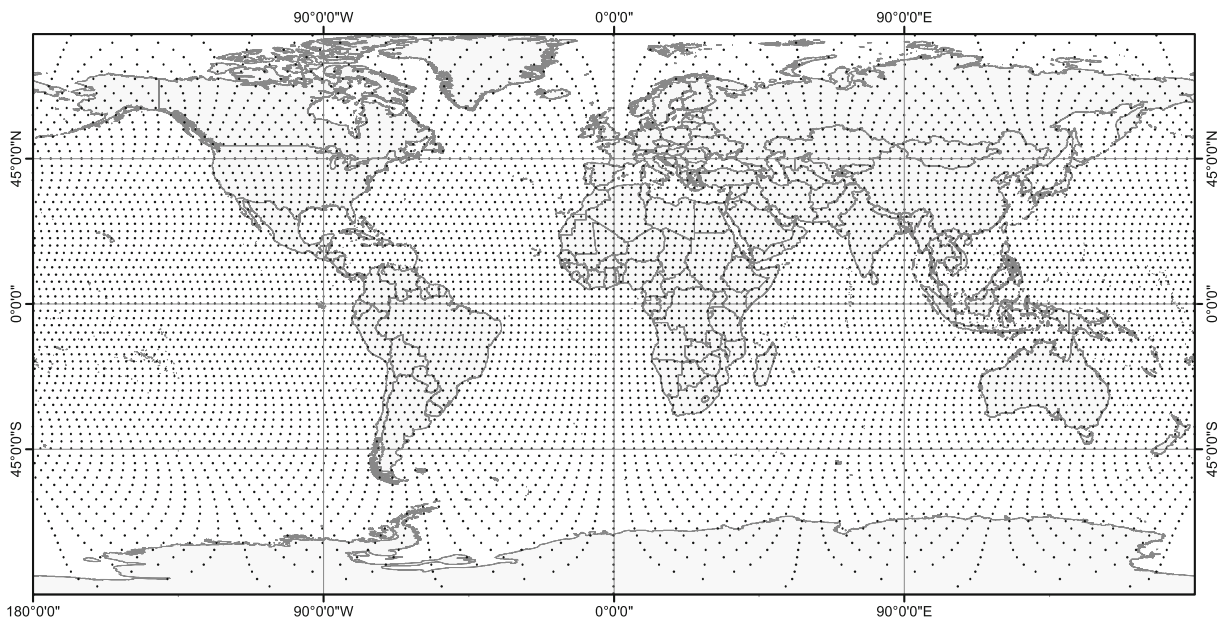
Fig. 3 GSG with a distance between circles of latitude and points per latitude of 250 km (Fehrmann 2015)

areas data set (GADM 2012) and the resulting area was estimated from the sample proportion $\hat{p} = n_L/n$, where n_L is the number of grid points over land and n is the sample size, i.e., the total number of grid points. A simple random sampling variance estimator is $\hat{v}(\hat{p}) = 1/(n-1) \times \hat{p}\hat{q}$, with $\hat{q} = 1 - \hat{p}$ (Cochran 1977). This estimator is upwards biased when applied to a systematic sample, thus resulting in a conservative estimate of variance. Systematic sampling, as for example, implemented by the GSG, is expected to be more precise (to an unknown extent) than simple random sampling (Aune-Lundberg and Strand 2014).
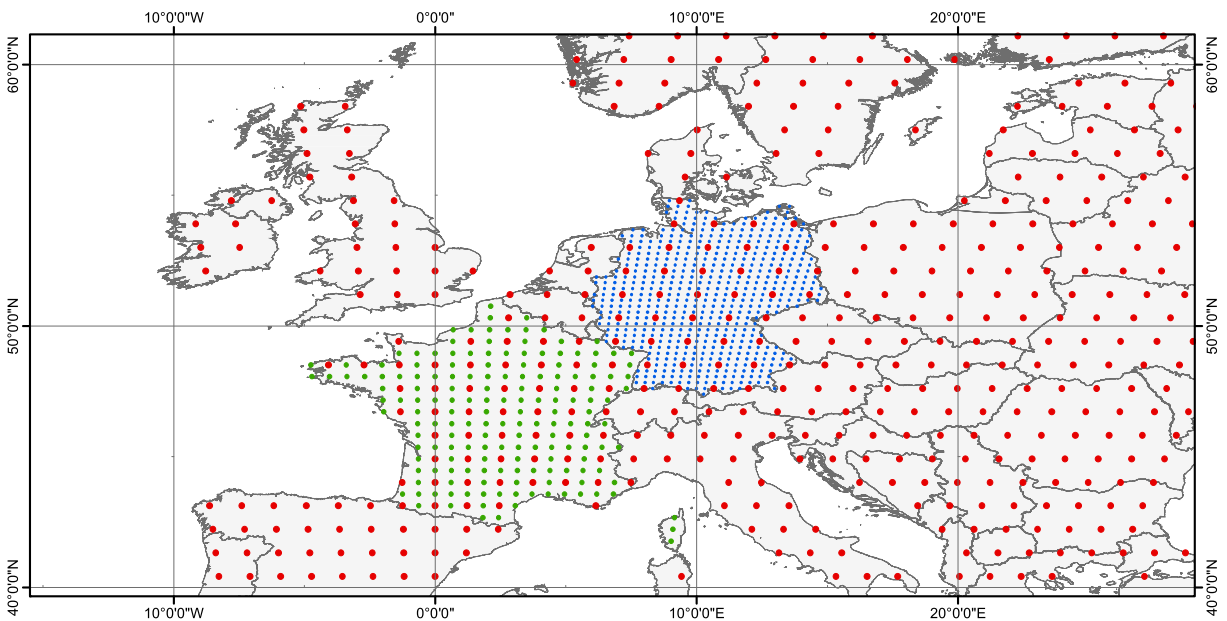
Fig. 4 Different spatial resolutions of the affine GSG grid system over central Europe: GSG 100 (red dots), GSG 50 over France (green dots), and GSG 25 over Germany (blue dots) (Fehrmann 2015)

In a second example, we used different configurations of point clusters as observation units for estimating forest cover, and evaluated their performance using Monte-Carlo simulations on artificially generated landscapes.

Binary forest cover maps simulating natural landscapes

We used artificial binary forest cover maps (forest = 1, nonforest = 0) to study the effects of observation design on the precision of proportions, estimated from GSG samples. The artificial forest maps reproduce realistic landscape patterns and were generated from Gaussian random fields, where the covariance structure was modeled using a generalized form of the Cauchy model (Schlather et al. 2015, Eq. 1).

$$C(r) = (1 + |r|^\alpha)^{-\frac{\beta = 0.3}{\alpha}} \qquad (1)$$

The model basically describes the covariance $C(r)$ of two pixel observations at distance r apart, based on two parameters, α and β. Here, $\alpha \in (0, 2]$ defines the local surface roughness (or smoothness) in fractal dimension - D, and $\beta > 0$ defines the global/long-memory dependence in the surface according to the Hurst coefficient—H (Schlather 2004). The implementation of the procedure generating continuous Gaussian random fields is simplified by the R-package *"RandomFields,"* using the function *"RMgencauchy"* (Schlather et al. 2015). The necessary theoretical and mathematical details are elaborated in Schlather (2004) and Schabenberger and Gotway (2005). For the generation of landscapes, α—a measure of patch complexity—was varied to control for fragmentation in forest/nonforest maps. Low α values led to high fragmentation and vice versa. Similar to Magdon and Kleinn (2013), highly fragmented landscapes can be achieved with $\alpha \approx 0.21$, medium fragmented with

$\alpha \approx 0.77$ and low fragmented landscapes were generated with $\alpha = 0.89$ respectively, while $\beta = 0.3$ was kept constant across all landscapes (Table 1). We used a 30-m pixel size to mimic Landsat like satellite images over a total area of 12,321 km^2 (equivalent to $1 \times 1°$ at the equator). To control the proportion of forest cover in the final imagery, the continuous Gaussian random fields were transformed to binary maps using 0.3 and 0.7 quantiles of the normal distribution to produce 30 and 70% forest cover proportions, respectively. We calculated a local Moran's I index using variable sizes of moving windows with neighborhoods ranging from one to 100 pixels, as a simple descriptor of local spatial autocorrelation in the final landscapes.

### Optimizing the cluster observation design to estimate forest cover

To study the precision of estimates from different cluster designs over artificially generated landscapes of varying spatial characteristics, we used Monte-Carlo simulations (srs) with point clusters of varying spatial configuration, cluster extent (or subplot distances), and number of subplots. Our setup conveniently mimicked application of a GSG with clusters of points in comparable landscapes. For our simulation, we used six artificial landscapes resulting from combinations of different forest cover proportions (30 and 70%) and degrees of fragmentation, equivalent to low, medium, and high spatial autocorrelation (Table 1). In order to match the scale of observation across clusters of different size (viz. number of subplots), each variant has the same spatial extent as the reference cluster with 49 subplots (7 × 7). Following this, clusters with 49, 16, 9, and 5 subplots were used (see, Fig. 5).

To further study design effects among different spatial configurations of cluster plots on sample precision,

Table 1 Characteristics of landscapes: cover %, level of fragmentation (α), and global spatial autocorrelation (Moran's I)

	Cover %	Fragmentation α	Moran's I
1.	30	0.22 (high)	0.1 (low)
2.	30	0.77 (medium–low)	0.5 (medium)
3.	30	0.89 (low)	0.7 (high)
4.	70	0.21 (high)	0.1 (low)
5.	70	0.78 (medium–low)	0.5 (medium)
6.	70	0.89 (low)	0.7 (high)

Landscapes are shown in Figs. 7 and 8
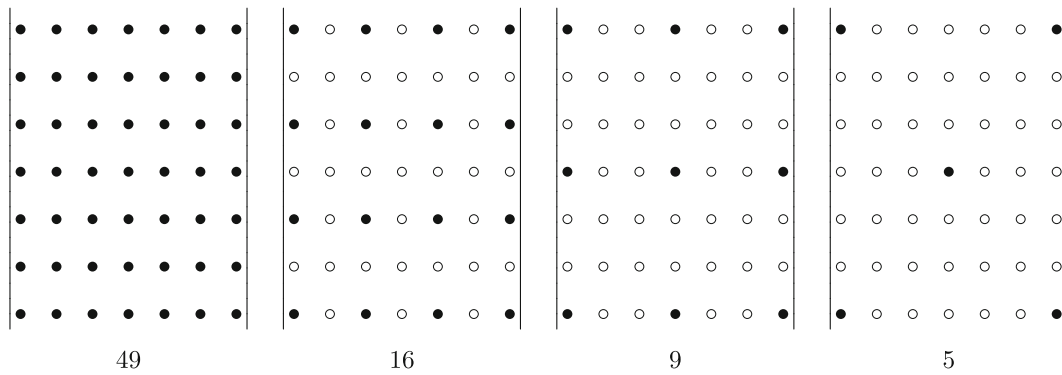
Fig. 5 Spatial configuration and cluster size (number of subplots); ● = subplot, ○ = NULL. The total number of subplots in each cluster is shown below each design
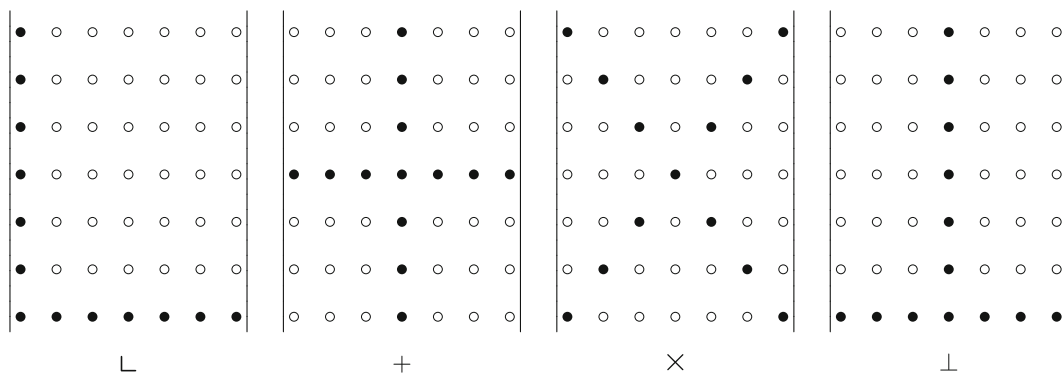
clusters with a constant number of 13 subplots were arranged according to the following shapes: ⌞, +, ×, and ⊥ (Fig. 6). For comparability to the previous designs, the scale of observation, i.e., the spatial extent of the clusters was maintained. In sampling with cluster plots, increasing the distance between subplots reduces the intracluster correlation and in turn increases precision of estimation. However, the mentioned gain in precision comes at a logistical cost (Yim et al. 2015). Therefore, an optimal distance between subplots and spatial configuration ought to be investigated in order to improve sampling efficiency. We varied distances between subplots in single-pixel steps from 1 pixel (30 m) to 100 pixels (3 km) for each spatial configuration and cluster size. During the Monte-Carlo simulation, each observation design was repeated 1000 times with a sample size of $n = 100$. At each iteration, the proportion of forest cover and its variance were estimated following (Cochran 1977, Chapter 3, Pg. 66). Forest cover proportion at

the ith cluster is given by $p_i = f_i/m_i$, where f_i is the number of points with forest cover and m_i is the size of the cluster, i.e., the number of sub-plots. As cluster size may vary due to overlap at the edge of the sampling frame, a ratio of means (rom) estimator is used for estimating forest cover proportion $\hat{p}_{\text{rom}}$ of an image from a sample of n clusters (Eq. 2):

$$\hat{p}_{\text{rom}} = \frac{\sum\limits_{i=1}^{n} f_i}{\sum\limits_{i=1}^{n} m_i} \tag{2}$$

A variance estimator for $\hat{p}_{\text{rom}}$ under the assumption of simple random sampling without replacement (srswor) is given in Eq. 3, without considering finite population correction:

$$\hat{v}(\hat{p}_{\text{rom}}) = \frac{1}{n\hat{m}^2} \frac{\sum f_i^2 - 2\hat{p}\sum f_i m_i + \hat{p}^2 \sum m_i^2}{n - 1}, \tag{3}$$

Fig. 6 Different spatial configuration of clusters with constant number of 13 subplots. ● = subplot, ○ = NULL

where $\hat{m} = \sum m_i/n$ is the average number of subplots per cluster in the sample.

To compare the performance of the different cluster configurations, the relative observed standard error was computed across $K = 1000$ iterations following Equation 4:

$$RSE_{\text{obs}} = \frac{V_{\text{obs}}}{\bar{\hat{p}}} \times 100\%, \qquad (4)$$

where $\bar{\hat{p}} = 1/K \sum_{k=1}^{K} \hat{p}_k$ is the mean of the generated sample distribution of $\hat{p}$ and $V_{\text{obs}} = 1/K \sum_{k=1}^{K} (\hat{p}_k - \bar{\hat{p}})^2$ is the according observed variance.

Results

Table 2 shows the results for global land surface estimation using different GSG grid sizes. The primary intention is to illustrate the global sample sizes for different grid resolutions.

Our simulation of different configurations of point clusters for the estimation of forest cover showed some expected but still interesting and instructive results. Figure 7 shows the estimated relative standard errors for clusters of different spatial extent and cluster size in dependence of varying forest cover and degree of fragmentation. The respective Moran's I is given as measure of local spatial autocorrelation, which is here in fact reproducing the spatial covariance structure determined by the parameter settings of the landscape generation. All single results show an asymptotically decreasing trend of standard errors over increasing cluster extent. That means, we see an increasing precision of estimation from cluster plots of larger spatial extent, where the decrease of standard errors is more pronounced for less fragmented landscapes. Another general outcome of the simulation is that the standard errors are decreasing with increasing fragmentation of the landscape. A higher fragmentation (viz. lower spatial autocorrelation) consistently leads to lower standard errors compared to more compact landscapes with low degree of fragmentation. Further, we see that differences between cluster sizes of 5, 9, 16, and 49 subplots become smaller with decreasing fragmentation. In the most extreme case of very compact landscapes with low fragmentation (lowest graphs in Fig. 7), cluster sizes of more than 5 subplots hardly show any differences in regard to the resulting precision of estimates. While the asymptotic decrease of standard errors levels off much earlier in highly fragmented landscapes, the curves still show a decreasing trend even for a large cluster extent in landscapes with low fragmentation. Thereby, clusters with less subplots tend to reach an asymptote much earlier than clusters with a larger number of subplots. Considering a constant spatial extent of the whole cluster plot, a lower number of subplots is related to larger subplot distances and vice versa.

Comparing the left and right side of the panel in Fig. 7, it is obvious that standard errors are constantly smaller for a higher coverage (70%) compared to low coverage (30%) of the target class. At the same time, the standard error is much less influenced by the respective cluster size (in terms of number of subplots) if the coverage is higher.

The comparison of different spatial cluster configurations presented in Fig. 8 shows no effect in

Table 2 Overview of the global sample size (n) for different standard GSG scales. n_L is the number of grid points on land, $\hat{p}$ the estimated proportion of the land surface and SE% is the estimated relative standard error

GSG scale	n	n_L	$\hat{p}$	SE%
500 km	2,053	589	0.287	3.48
250 km	8,173	2,344	0.287	1.74
200 km	12,785	3,649	0.285	1.40
100 km	51,133	14,675	0.287	0.70
50 km	204,489	58,810	0.288	0.35
25 km	817,979	235,344	0.288	0.17
10 km	5,112,081	1,471,349	0.288	0.07

For this example, existence of land was observed at GSG grid points using *global administrative areas* data set (GADM 2012)
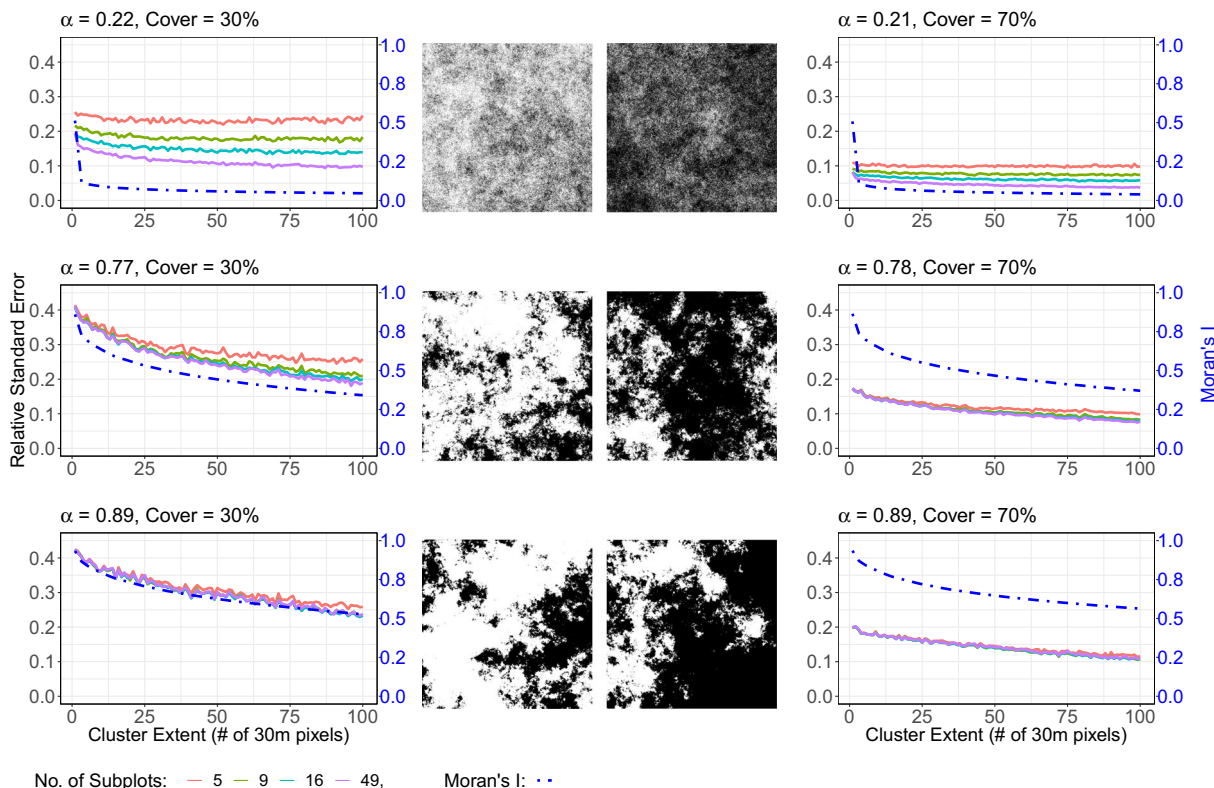
Fig. 7 The primary *y*-axis (left) shows relative standard error against cluster extent for each observation design of 5, 9, 16, and 49 subplots and forest cover landscapes (see, Fig. 5 for details on the observation designs). Forest landscapes are placed adjacent to the corresponding graphs. *α* values in the heading of each graph represent levels of fragmentation in landscapes of 30 (left) and 70% (right) forest cover. The extent of clusters was increased by one pixel from 0 to 100 pixels (*x*-axis) maintaining the design configurations as shown in Fig. 5. Therefore, relative subplot distances increased in the same magnitude for each increment in extent. Notice that clusters of 5 subplots have the longest subplot distances and clusters of 49 subplots have the shortest subplot distances. The secondary *y*-axis (in blue) shows Moran's I computed in moving windows of sizes ranging from 1 to 100 pixels (*x*-axis). Scales of the *y*-axes across the considered six landscapes are kept constant for comparability; however, the reader interested in detailed differences among considered subplot configurations per landscape can refer to Fig. A.9 in the Supplementary Material

landscapes with high fragmentation and low spatial autocorrelation. With increasing autocorrelation, the separation of two groups of cluster configurations becomes obvious, while this distinction is more pronounced in landscapes with lower coverage (30%). As the number of subplots per cluster is held constant (13 across all spatial cluster configurations), the effect is exclusively a result of varying distances between the subplots. Those configurations that include larger maximum subplot distances (⌐ and × shape) show higher precision than more compact forms (+ and ⊥ shape).

Discussion and conclusions

With the GSG, we are proposing a simple sampling grid for application from local to global scale. Compared to other grid systems (see, e.g., White et al. (1992); Kimerling et al. (1999); Richards et al. (2000); Sahr et al. (2003); Sahr (2011); Song et al. (2002); Swinbank and Purser (2006), Youngren and Petty (2017)), the construction of the GSG is straightforward and comprehensible, while maintaining the desirable characteristic of an almost uniform sampling intensity on a global scale. The fixed origin of
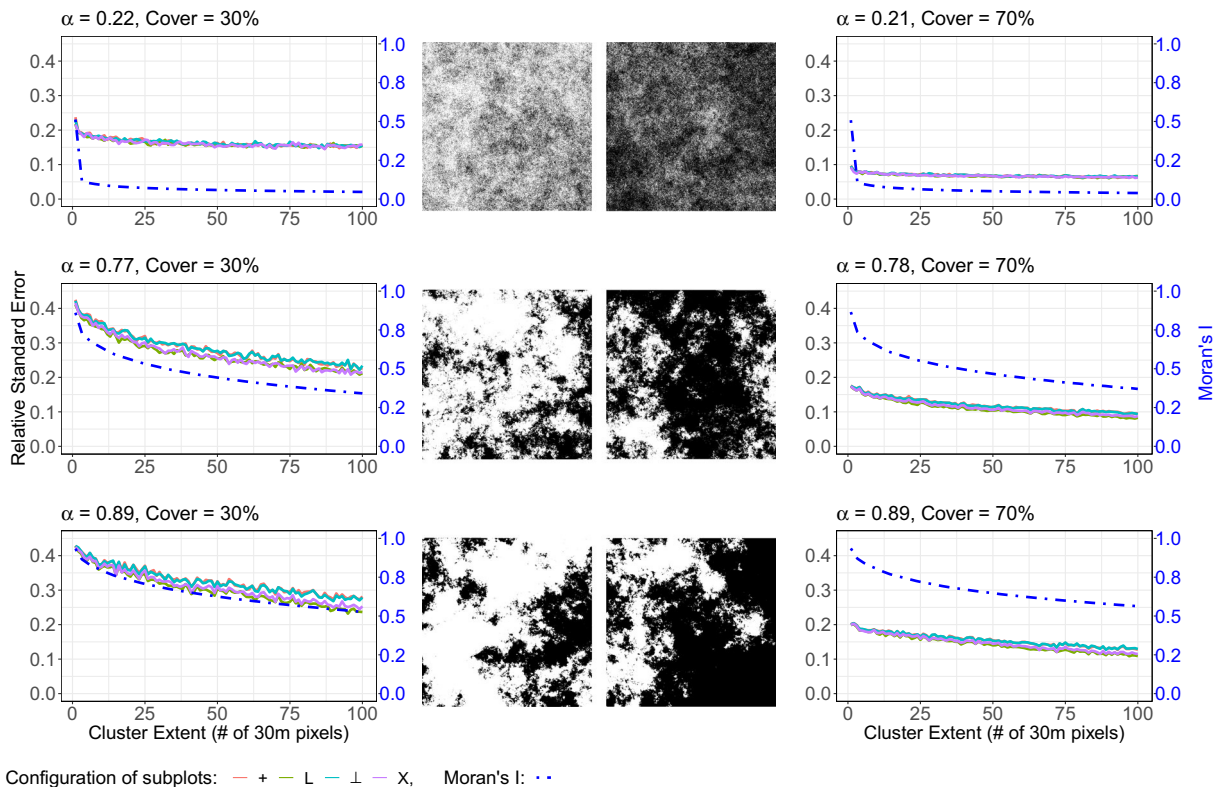
Fig. 8 The primary *y*-axis (left) shows relative standard error against cluster extent for each observation design of 13 subplots arranged in ⊔, +, ×, and ⊥ configurations and forest cover landscapes (see, Fig. 6 for details on the observation designs). Forest landscapes are placed adjacent to the corresponding graphs. α values in the heading of each graph represent levels of fragmentation in landscapes of 30 (left) and 70% (right) forest cover. The extent of clusters was increased by 1 pixel from 0 to 100 pixels (*x*-axis) maintaining the design configurations as shown in Fig. 6. Therefore, relative subplot distances

increased in the same magnitude for each increment in extent. Notice that clusters of 5 subplots have the longest subplot distances and clusters of 49 subplots have the shortest subplot distances. The secondary *y*-axis (in blue) shows Moran's I computed in moving windows of neighborhoods ranging from 1 to 100 pixels (*x*-axis). Scales of the *y*-axes across the considered six landscapes are kept constant for comparability; however, the reader interested in detailed differences among considered subplot configurations per landscape can refer to Fig. A.10 in the Supplementary Material

the grid system ensures that grids at multiple resolutions can have a regular hierarchical relationship that can be exploited during analysis (Sahr 2011). This characteristic allows flexibility in planning of individual sampling studies according to the specific user needs or target variables, while at the same time synergies between different studies through combination of observations from different fields of application can be exploited. In principle, the proposed GSG can be basis for nearly any sampling design and/or subselection of locations from the systematic grid. Subselecting (random or systematic) from a dense

base grid allows implementing any common sampling design, including stratified, two-phase, multistage sampling or techniques like systematic unaligned sampling or Spatially Balanced Sampling ().

Wrapping the grid around a sphere leads to a narrow corridor of irregular point distances at a longitude of 180°. This corridor could be shifted eastwards such that no land mass is affected by setting the start meridian of the grid to 11° east instead of using the 0-Meridian at Greenwich as start point. In favor of an easy to communicate grid construction and because just a small area is affected, we refrain from doing so.

The primary idea of the proposed GSG is to facilitate sampling studies by visual interpretation of high resolution remote sensing imagery, but it can equally be used as a basis for terrestrial sampling studies, like national or regional forest inventories or other monitoring activities. Using a uniform base grid instead of independent grids for different purposes, allows direct comparability.

While the GSG provides a comprehensive and standardized grid system, the observation design (or plot design) implemented at each sampling location needs to be planned in accordance with the respective target variables. Dependent on the goals of sampling, the choice of suitable plots can vary substantially. It is not possible to suggest a single design for all purposes. Our simulation is focused on the estimation of coverage by visual interpretation of imagery, which is a very common task usually accomplished by classification of remote sensing data.

While both, an exhaustive remote sensing based classification and a sample-based estimation from visual interpretation, are not free of errors, the latter does not require specific remote sensing or image processing skills and practically no preprocessing of data sources. It utilizes the human capacity of visual comprehension and appraisal for a more careful and detailed interpretation on small samples. A human interpreter is more flexible to cope with different image qualities, seasonal differences, varying phenology, or changing atmospheric conditions than a classification algorithm. Therefore, visual interpretation can be less demanding in regard to image quality than image classification approaches and opens the possibility to make use of global image archives provided by virtual globes, which do not allow any further processing or automatic classification of imagery. Assuming a visual interpretation that is free of measurement errors, the only uncertainty in the sample estimate of p would result from the random selection of a sample much smaller than the population. This assumption, however, is not realistic, as also a careful visual interpretation is not always unambitious.

Evaluation of cluster plot designs

In our simulation study on the optimization of point clusters for the estimation of coverage, we compared clusters of different size and spatial configuration. Single points or clusters of points have been proven to be a very efficient observation design compared to other options like delineation in fixed area plots or line sampling approaches (Fehrmann et al. 2014; Kleinn 1994). In general, the use of clusters of points instead of independently selected single points is meaningful only if logistical costs for reaching or observing the single point locations need to be taken into account. A direct comparison of performance between cluster sampling and sampling with the same number of independently selected single points will always result in higher efficiency for the latter, simply because of the much higher sample size. In context of visual interpretation of imagery, logistical costs are reduced to visualize images at each sampling location on screen, which is a fairly low effort compared to reaching sampling locations in the field. However, cluster sampling might still be meaningful as it allows obtaining information on small-scale landscape structures or spatial covariance (Kleinn 2000), or for estimating landscape metrics (Ramezani et al. 2010; Ramezani and Holm 2011). In the best case, the cluster's spatial extent should not exceed a size that can be visualized and interpreted at once for all subplots together.

Even if the results presented in Figs. 7 and 8 are not unexpected, they are very informative and useful for the optimization of cluster designs for the estimation of coverage from point clusters. The resulting precision of estimates is highly dependent on the ability of the observation design to capture the existing spatial variability of coverage of the target class in the different landscapes. Our results indicate that the "observation effort," which is here given by the number of subplots per cluster that need to be observed, is not necessarily directly linked to the resulting precision of estimates. Depending on the coverage and the spatial structure of the landscape, an assessment of a high number of subplots per cluster might not be efficient, as the related increase of precision with increasing cluster size is relatively modest. The reason is that larger clusters with more subplots tend to collect more redundant information, which reduces the overall efficiency.

In our example, we used artificial forest cover maps assuming that they mimic realistic spatial landscape patterns with different degree of fragmentation and coverage of the target class. In this case, it is quite obvious that clusters of more than nine subplots will not lead to practically relevant improvements of the final estimation if the fragmentation is relatively low.

Only if fragmentation is relatively high (and spatial autocorrelation is low), it might be advisable to observe larger clusters with more subplots. The distance between subplots (here points) inside the cluster determines to which extent spatially uncorrelated; therefore, less redundant information is assessed. An optimal subplot distance and size of the cluster is therefore dependent on the spatial covariance structure or autocorrelation in the given landscape. In our simulation, these characteristics are determined by the spatial process used to generate the artificial landscapes. Figure 7 clearly shows this effect, especially in the landscapes with high and medium fragmentation. Clusters of identical extent but fewer subplots (and therefore larger distances between subplots) tend to reach an asymptote much earlier than those with more subplots.

In regard to the comparison of equally sized clusters with 13 subplots, our results indicate that spatial configurations that include larger maximum distances between points are more efficient than those with more compact forms. In landscapes with medium or low fragmentation, the $\llcorner$- shaped clusters constantly showed the best performance, followed by $\times$- shaped plots. As the landscape were generated using a random field process, the landscape structure does not show a directional pattern. Thus, the only effect of the different cluster designs is based on the difference in the subplot distances.

One topic we have not addressed in our simulation study is the problem of rare events. If the target of estimation is in classes with relatively low coverage or small area changes, single points are not an appropriate observation design. Even if cluster plots with many subplots could increase the inclusion probability of small-scale classes or changes, fixed area plots or line elements are expected to be much more efficient in such cases (Kleinn 1994, 1996; Fehrmann et al. 2014). Adaptive cluster sampling is another alternative that can help to increase the overall efficiency for rare and clustered events, like small-scale deforestation (Magnussen et al. 2005). On the other hand, it is always an option to map classes with small area extent inside fixed area plots. The GSG as described here might also be used in context of two-phase sampling, were typically a large sample of relatively easy/cheap to obtain observations is taken from an image of auxiliary variables correlated with a target variable in a first phase and a subsample is selected for expensive observations including field work.

Publisher's note   Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Achard, F., Defries, R., Eva, H., Hansen, M., Mayaux, P., Stibig, H.J. (2007). Pan-tropical monitoring of deforestation. *Environmental Research Letters, 2*. https://doi.org/10.1088/1748-9326/2/4/045022.

Achard, F., Eva, H.D., Stibig, H.-J., Mayaux, P., Gallego, J., Richards, T., Malingreau, J.P. (2002). Determination of deforestation rates of the world's humid tropical forests. *Science*, *297*, 999–1002. https://doi.org/10.1126/science.1070656.

Aune-Lundberg, L., & Strand, G.H. (2014). Comparison of variance estimation methods for use with two-dimensional systematic sampling of land use/land cover data. *Environmental Modelling and Software*, *61*, 87–97. https://doi.org/10.1016/j.envsoft.2014.07.001.

Barrett, F., McRoberts, R.E., Tomppo, E., Cienciala, E., Waser, L.T. (2016). A questionnaire-based review of the operational use of remotely sensed data by national forest inventories. *Remote Sensing of Environment*, *174*, 279–289. http://www.sciencedirect.com/science/article/pii/S0034425715301176.

Bastin, J.-F., Berrahmouni, N., Grainger, A., Maniatis, D., Mollicone, D., Moore, R., Patriarca, C., Picard, N., Sparrow, B., Abraham, E.M., Aloui, K., Atesoglu, A., Attore, F., Bey, A., Garzuglia, M., García-montero, L.G., Groot, N., Guerin, G., Laestadius, L., Lowe, A.J., Mamane, B. (2017). The extent of forest in dryland biomes. *Science*, *638*, 1–5. http://science.sciencemag.org/content/sci/356/6338/635.full.pdf.

Beuchle, R., Eva, H.D., Stibig, H.-J., Bodart, C., Brink, A., Mayaux, P., Johansson, D., Achard, F., Belward, A. (2011). A satellite data set for tropical forest area change assessment. *International Journal of Remote Sensing*, *32*, 7009–7031. https://www.tandfonline.com/doi/full/10.1080/01431161.2011.611186.

Cochran, W.G. (1977). Sampling techniques.

Cracknell, A.P., Kanniah, K.D., Tan, K.P., Wang, L. (2013). Evaluation of MODIS gross primary productivity and land cover products for the humid tropics using oil palm trees in Peninsular Malaysia and Google Earth imagery. *International Journal of Remote Sensing*, *34*, 7400–7423. https://doi.org/10.1080/01431161.2013.820367.

http://apps.webofknowledge.com/full_record.do?product=WOS&search_mode=GeneralSearch&qid=2&SID=X1gj212wPM4ZeXzz3lz&page=1&doc=4.

Esseen, P.A., Jansson, K.U., Nilsson, M. (2006). Forest edge quantification by line intersect sampling in aerial photographs. *Forest Ecology and Management*, *230*, 32–42. https://doi.org/10.1016/j.foreco.2006.04.012.

FAO (2009). The 2010 Global Forest Resources Assessment Remote Sensing Survey: an outline of the objectives, data, methods and approach. Techical Report, FAO. http://www.fao.org/3/a-k7023e.pdf.

FAO (2010). Global forest resources assessment. Technical Report. FAO. arXiv:0404048. ISBN:978-92-5-106654-6.

Fattorini, L., Franceschi, S., Pisani, C. (2009). A two-phase sampling strategy for large-scale forest carbon budgets. *Journal of Statistical Planning and Inference*, *139*, 1045–1055. https://doi.org/10.1016/j.jspi.2008.06.014.

Fehrmann, L. (2015). A unified framework for environmental monitoring based on a discrete global sampling grid (GSG) system. In Fehrmann, L., & Kleinn, C. (Eds.) *Proceedings of the 5th international DAAD workshop* (pp. 99–111). Cuvellier Verlag Göttingen.

Fehrmann, L., Seidel, D., Krause, B., Kleinn, C. (2014). Sampling for landscape elements - a case study from Lower Saxony, Germany. *Environmental Monitoring and Assessment*, *186*, 1421–1430. https://doi.org/10.1007/s10661-013-3464-0.

GADM (2012). Gadm database of global administrative areas, version 2.0. https://www.gadm.org/.

Gibbs, H.K., Brown, S., Niles, J.O., Foley, J.A. (2007). Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters, 2*. https://doi.org/10.1088/1748-9326/2/4/045023.

Gregoire, T., & Valentine, H. (2007). Sampling strategies for natural resources and the environment. Chapman & Hall/CRC Applied Environmental Statistics. Taylor & Francis. https://books.google.de/books?id=1z71MAe3gL0C.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*, 850–853. https://doi.org/10.1126/science.1244693. http://www.sciencemag.org/content/342/6160/850.abstract. arXiv:1011.1669v3.

Hu, Q., Wu, W., Xia, T., Yu, Q., Yang, P., Li, Z., Song, Q. (2013). Exploring the use of Google Earth imagery and object-based methods in land use/cover mapping. *Remote Sensing*, *5*, 6026–6042. https://doi.org/10.3390/rs5116026. http://www.mdpi.com/2072-4292/5/11/6026.

Jr., D.L.S., & Olsen, A.R. (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, *99*, 262–278. https://doi.org/10.1198/016214504000000250.

Kimerling, J.A., Sahr, K., White, D., Song, L. (1999). Comparing geometrical properties of global grids. *Cartography and Geographic Information Science*, *26*, 271–288. http://www.tandfonline.com/doi/abs/10.1559/152304099782294186.

Kleinn, C. (1994). Comparison of the performance of line sampling to other forms of cluster sampling. *Forest Ecology and Management*, *68*, 365–373. https://doi.org/10.1016/0378-1127(94)90057-4.

Kleinn, C. (1996). Ein vergleich der effizienz von verschiedenen clusterformen in forstlichen großrauminventuren. *Forstwissenschaftliches Centralblatt vereinigt mit Tharandter forstliches Jahrbuch*, *115*, 378–390. https://doi.org/10.1007/BF02738616.

Kleinn, C. (2000). Estimating metrics of forest spatial pattern from large area forest inventory cluster samples. *Forest Science*, *46*, 548–557. https://doi.org/10.1093/1007forestscience/46.4.548.

Lister, A.J., & Scott, C.T. (2009). Use of space-filling curves to select sample locations in natural resource monitoring studies. *Environmental Monitoring and Assessment*, *149*, 71–80. https://doi.org/10.1007/s10661-008-0184-y.

Magdon, P., & Kleinn, C. (2013). Uncertainties of forest area estimates caused by the minimum crown cover criterion: - a scale issue relevant to forest cover monitoring. *Environmental Monitoring and Assessment*, *185*, 5345–5360. https://doi.org/10.1007/s10661-012-2950-0.

Magnussen, S., Kurz, W., Leckie, D.G., Paradine, D. (2005). Adaptive cluster sampling for estimation of deforestation rates. *European Journal of Forest Research*, *124*, 207–220. https://doi.org/10.1007/s10342-005-0074-6.

Mandallaz, D. (2007). *Sampling techniques for forest inventories. Chapman & Hall/CRC Applied Environmental Statistics*. Boca Raton: CRC Press. https://books.google.de/books?id=lCnIGO1rt18C.

Olofsson, P., Stehman, S.V., Woodcock, C.E., Sulla-Menashe, D., Sibley, A.M., Newell, J.D., Friedl, M.A., Herold, M. (2012). A global land-cover validation data set, part I: fundamental design principles. *International Journal of Remote Sensing*, *33*, 5768–5788. https://doi.org/10.1080/01431161.2012.674230.

Pengra, B., Long, J., Dahal, D., Stehman, S.V., Loveland, T.R. (2015). A global reference database from very high resolution commercial satellite data and methodology for application to landsat derived 30m continuous field tree cover data. *Remote Sensing of Environment*, *165*, 234–248. https://doi.org/10.1016/j.rse.2015.01.1018. http://www.sciencedirect.com/science/article/pii/S003442571500036X.

Ploton, P., Pélissier, R., Proisy, C., Flavenot, T., Barbier, N., Rai, S.N., Couteron, P. (2012). Assessing aboveground tropical forest biomass using Google Earth canopy images. *Ecological Applications : a Publication of the Ecological Society of America*, *22*, 993–1003. https://doi.org/10.1890/11-1606.1. http://www.ncbi.nlm.nih.gov/pubmed/22645827, http://www.esajournals.org/doi/abs/10.1890/11-1606.1.

Potere, D. (2008). Horizontal positional accuracy of Google Earth's high-resolution imagery archive. *Sensors*, *8*, 7973–7981. https://doi.org/10.3390/s8127973.

Ramezani, H., & Holm, S. (2011). Sample based estimation of landscape metrics; accuracy of line intersect sampling for estimating edge density and Shannon's diversity index. *Environmental and Ecological Statistics*, *18*, 109–130. https://doi.org/10.1007/s10651-009-0123-2.

Ramezani, H., Holm, S., Allard, A., Ståhl, G. (2010). Monitoring landscape metrics by point sampling: accuracy in estimating Shannon's diversity and edge density. *Environmental Monitoring and Assessment*, *164*, 403–421.

https://doi.org/10.1007/s10661-009-0902-0. http://link.spri-
nger.com/10.1007/s10661-009-0902-0.

Richards, T., Gallego, J., Achard, F. (2000). Sampling for
forest cover change assessment at the pan-tropical scale.
International Journal of Remote Sensing, *21*, 1473–1490.
https://doi.org/10.1080/014311600210272.

Saatchi, S.S., Harris, N.L., Brown, S., Lefsky, M., Mitchard,
E.T.A., Salas, W., Zutta, B.R., Buermann, W., Lewis, S.L.,
Hagen, S., Petrova, S., White, L., Silman, M., Morel, A.
(2011). Benchmark map of forest carbon stocks in tropical
regions across three continents. *Proceedings of the National
Academy of Sciences*, *108*, 9899–9904. https://doi.org/10.
1073/pnas.1019576108. http://www.pnas.org/content/early/
2011/05/24/1019576108.abstract, http://www.pnas.org/con-
tent/108/24/9899.abstract%0A, http://www.pnas.org/content/
108/24/9899.full.pdf. arXiv:1408.1149.

Sahr, K. (2011). Hexagonal discrete global grid systems for
geospatial computing. *Archives of Photogrammetry, Cartogra-
phy and Remote Sensizng*, *22*, 363–376. https://journals.
indexcopernicus.com/search/article?articleId=1435632.

Sahr, K., White, D., Kimerling, A.J. (2003). Geodesic dis-
crete global grid systems. *Cartography and Geographic
Information Science*, *30*, 121–134. https://doi.org/10.1559/
152304003100011090. http://www.tandfonline.com/doi/
abs/10.1559/152304003100011090.

Schabenberger, O., & Gotway, C.A. (2005). *Statistical methods
for spatial data analysis*. USA: Chapman & Hall/CRC.

Schlather, M. (2004). Separate fractal dimension and the hurst
effect. *SIAM REVIEW*, *46*, 269–282. https://doi.org/10.
1137/S0036144501394387.

Schlather, M., Malinowski, A., Menck, P.J., Oesting, M.,
Strokorb, K. (2015). Analysis, simulation and prediction
of multivariate random fields with package RandomFields.
Journal of Statistical Software, *63*, 1–25. http://www.
jstatsoft.org/v63/i08/.

Sheppard, S.R.J., & Cizek, P. (2009). The ethics of Google
Earth: crossing thresholds from spatial data to land-
scape visualisation. *Journal of Environmental Management*,
90, 2102–17. https://doi.org/10.1016/j.jenvman.2007.09.012.
http://apps.webofknowledge.com/full_record.do?product=
WOS&search_mode=Refine&qid=6&SID=X1gj212wPM4
ZeXzz3lz&page=1&doc=3.

Song, L., Kimerling, A.J., Sahr, K. (2002). Developing an equal
area global grid by small circle subdivision. http://www.
ncgia.ucsb.edu/globalgrids-book/song-kimmerling-sahr/.

Stehman, S., Sohl, T., Loveland, T. (2003). Statistical sam-
pling to characterize recent united states land-cover change.

Remote Sensing of Environment, *86*, 517–529. https://doi.
org/10.1016/S0034-4257(03)00129-9. http://www.science-
direct.com/science/article/pii/S0034425703001299.

Stehman, S.V. (1999). Basic probability sampling designs for
thematic map accuracy assessment. *International Journal
of Remote Sensing*, *20*, 2423–2441. https://doi.org/10.1080/
014311699212100.

Stehman, S.V., Sohl, T.L., Loveland, T.R. (2005). An eval-
uation of sampling strategies to improve precision of
estimates of gross change in land use and land cover.
International Journal of Remote Sensing, *26*, 4941–4957.
https://doi.org/10.1080/01431160500222632.

Sun, X., Shen, S., Leptoukh, G.G., Wang, P., Di, L., Lu, M.
(2012). Development of a Web-based visualization plat-
form for climate research using Google Earth. *Computers &
Geosciences*, *47*, 160–168. https://doi.org/10.1016/j.cageo.
2011.09.010. http://apps.webofknowledge.com/full_record.
do?product=WOS&search_mode=GeneralSearch&qid=3&
SID=X1gj212wPM4ZeXzz3lz&page=1&doc=5.

Swinbank, R., & Purser, R.J. (2006). Fibonacci grids: a novel
approach to global modelling. Quarterly Journal of the
Royal Meteorological Society https://doi.org/10.1256/qj.
05.227.

Theobald, D.M., Stevens, D.L., White, D., Urquhart, N.S.,
Olsen, A.R., Norman, J.B. (2007). Using GIS to gener-
ate spatially balanced random survey designs for natu-
ral resource applications. *Environmental Management*, *40*,
134–146. https://doi.org/10.1007/s00267-005-0199-x.

White, D., Kimerling, J.A., Overton, S.W. (1992). Carto-
graphic and geometric components of a global sam-
pling design for environmental monitoring. *Cartogra-
phy and Geographic Information Systems*, *19*, 5–22.
https://doi.org/10.1559/152304092783786636.

Wickman, F.E., Elvers, E., Edvarson, K. (1974). A system
of domains for global sampling problems. *Geografiska
Annaler: Series A, Physical Geography*, *56*, 201–212.
https://doi.org/10.1080/04353676.1974.11879899.

Yim, J.S., Shin, M.Y., Son, Y., Kleinn, C. (2015). Cluster
plot optimization for a large area forest resource inven-
tory in Korea. *Forest Science and Technology*, *11*, 139–146.
https://doi.org/10.1080/21580103.2014.968222.

Youngren, R.W., & Petty, M.D. (2017). A multi-resolution
HEALPix data structure for spherically mapped point data.
Heliyon https://doi.org/10.1016/j.heliyon.2017.e00332.

Yu, L., & Gong, P. (2012). Google earth as a virtual globe tool
for earth science applications at the global scale: progress
and perspectives https://doi.org/10.1080/01431161.2011.
636081.
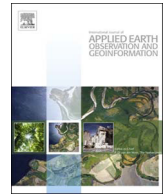
CrossMark

Ensemble classification of individual *Pinus* crowns from multispectral satellite imagery and airborne LiDAR

Collins B. Kukunda[a,*], Joaquín Duque-Lazo[b], Eduardo González-Ferreiro[a,c], Hauke Thaden[d], Christoph Kleinn[a]

[a] *University of Goettingen, Forest Inventory and Remote Sensing, Faculty of Forest Sciences, Büsgenweg 5, D-37077 Göttingen, Germany*
[b] *Department of Forestry, School of Agriculture and Forestry, University of Córdoba, Laboratory of Dendrochronology, DendrodatLab-ERSAF Edf Leonardo da Vinci, Campus de Rabanales s/n, 14071 Córdoba, Spain*
[c] *Unidade de Xestión Forestal Sostible (UXFS) – Departamento de Enxeñería Agroforestal, Escola Politécnica Superior, R/Benigno Ledo, Campus Terra, E-27002 Lugo, Spain*
[d] *University of Goettingen, Chair for Statistics, Humboldallee 3, D-37073 Göttingen, Germany*

ARTICLE INFO

ABSTRACT

Distinguishing tree species is relevant in many contexts of remote sensing assisted forest inventory. Accurate tree species maps support management and conservation planning, pest and disease control and biomass estimation. This study evaluated the performance of applying ensemble techniques with the goal of automatically distinguishing *Pinus sylvestris* L. and *Pinus uncinata* Mill. Ex Mirb within a 1.3 km^2 mountainous area in Barcelonnette (France). Three modelling schemes were examined, based on: (1) high-density LiDAR data (160 returns m^{-2}), (2) Worldview-2 multispectral imagery, and (3) Worldview-2 and LiDAR in combination. Variables related to the crown structure and height of individual trees were extracted from the normalized LiDAR point cloud at individual-tree level, after performing individual tree crown (ITC) delineation. Vegetation indices and the Haralick texture indices were derived from Worldview-2 images and served as independent spectral variables. Selection of the best predictor subset was done after a comparison of three variable selection procedures: (1) Random Forests with cross validation (AUCRFcv), (2) Akaike Information Criterion (AIC) and (3) Bayesian Information Criterion (BIC). To classify the species, 9 regression techniques were combined using ensemble models. Predictions were evaluated using cross validation and an independent dataset. Integration of datasets and models improved individual tree species classification (True Skills Statistic, TSS; from 0.67 to 0.81) over individual techniques and maintained strong predictive power (Relative Operating Characteristic, ROC = 0.91). Assemblage of regression models and integration of the datasets provided more reliable species distribution maps and associated tree-scale mapping uncertainties. Our study highlights the potential of model and data assemblage at improving species classifications needed in present-day forest planning and management.

1. Introduction

Physically and spectrally similar species such as *P. sylvestris* (*Pinus sylvestris* L.) and *P. uncinata* (*Pinus uncinata* Mill. Ex Mirb) are hard to distinguish in the field and in remote sensing imagery. In some cases microbiological analyses (Alvarez et al., 2009; Boratynska and Boratynski, 2007) are required to separate their identity. The microbiological methods provide high taxonomic precision but are inapplicable when, for example, species identification is required in forest inventories over large areas. Remotely sensed hyperspectral and/or multispectral data show potential to address this challenge. They can

accurately recognize and map continuous stochastic distributions of vegetation communities, species groups, land-use and land-cover types and individual species across different genera. However, optical remote sensing data can fail when discriminating individual tree species that have similar appearances. An early attempt by Coleman et al. (1990) using Landsat TM data failed to discriminate between *Pinus* species stands because of similarity in spectral responses. Later, Goodwin et al. (2005) using high spatial resolution airborne CASI-2 data found individual *Eucalyptus* species – "spectrally complex" – and opted for genera groups. In the time since the above mentioned studies were published, spatial and spectral resolutions of datasets have increased

---

making available less correlated sub-genus reflectance data. However, spectrally and structurally related species continue to present a challenge. Studies reported that discrimination of closely related tree species was hampered by: (1) high structural similarity undermining the utility of structural characteristics like crown shape, size and leaf area index (Goodwin et al., 2005), (2) a higher intra-species than inter-species leaf spectral variability implying a fuzzy spectral signal (Youngentob et al., 2011) and (3) mixed spectral signals resulting from the influence of canopy scale structure (e.g. vegetation to background ratio in a pixel) (Aberle, 2016), crown aspect (i.e. shaded or non-shaded crowns) (Gerard and North, 1997) and noise from tree age and phenology (Clark et al., 2005; Peerbhay et al., 2014). Such challenges explain the limited usefulness of optical structural and spectral attributes in distinguishing similar tree species.

Complementary to spectral features recorded by means of optical remote sensing are structural features of the tree crowns or forest canopies that can be derived from LiDAR (Light Detection and Ranging) point clouds (Vauhkonen et al., 2014). LiDAR data offers the opportunity of describing some differences in properties of species crowns (e.g. the amount and allocation of biomass to branches and leaves) by recording differences in point height distributions (Ørka et al., 2009; Vauhkonen et al., 2014), especially when high point densities are available. LiDAR derived height and density distributions can be isolated for individual trees by means of individual tree detection techniques (e.g. Wulder et al. (2000)) and crown segmentation methods. Ultimately, the individual crown metrics can be used as predictors of species identities in object oriented approaches (Gougeon and Leckie, 2006; Ke et al., 2010). Studies have also shown that LiDAR intensity data i.e. strength of the back scattered energy, is useful in distinguishing between tree species, particularly when used in conjunction with other structural LiDAR-derived variables (Kim et al., 2009; Suratno et al., 2009; Zhang and Lui, 2013). For example, combined intensity and structural features to distinguish Norway Spruce and Birch trees which resulted in an overall accuracy of 88%. Similarly, Zhang and Lui (2013) demonstrated the applicability of LiDAR-derived structure and intensity variables to distinguish *Nothofagus cunninghamii* (Hook.) Oerst. and *Acacia dealbata* Link using Support Vector Machines (SVM) attaining overall accuracy up to 88.6%. Suratno et al. (2009) also used both structural and intensity predictors for identifying four species of individual trees in a mixed coniferous forest and reported kappa of 56% compared to kappa of 93% based on stands. Nonetheless, some researchers argue that the lack of a spectral signal remains an important limitation of LiDAR data in identifying tree species (Deng et al., 2007; Leckie et al., 2003; Swatantran et al., 2011) and they propose integration of LiDAR and spectral datasets as a more effective method.

Irrespective of the dataset in question and the similarity of the target objects, automatic species differentiation requires a classification model to link field observations to predictors. In many cases, a convenient multivariate form of the classification model is not known. In their review article, Fassnacht et al. (2016) show that different studies have explored various classification techniques including parametric approaches like: Discriminant analysis (FDA), Maximum likelihood (ML), Spectral Angle Mapper (SAM), Bayesian regression, Generalized Linear Models (GLM), Spectral Mixture Analysis (SMA), Logistic regression, Fuzzy logic, and thresholding and non-parametric approaches like; SVM, Random Forests (RF) and Artificial Neural Networks (ANN). We argue here that for any given choice of a model (parametric, semi- or non-parametric) or any choice of a variable set (dependent on the chosen variable selection procedure and datasets), the accuracy in prediction may vary since: (1) different model families fit the data to varying degrees (Appendix 1 in supplementary material shows differences in model fit along the data range given the model type) and, (2) a predictor effect is partially influenced/confounded by effects of other covariates. Recently, some studies proposed ensemble classification approaches to address the issues related to uncertainty in prediction

across classifiers (Engler et al., 2013; Ko et al., 2014; Duque-Lazo and Navaro-Cerrillo, 2017). These techniques essentially combine decisions from several statistical classifiers with the aim of minimizing generalization error (Banfield et al., 2007; Ko et al., 2014; Engler et al., 2013). In the same way, their research demonstrates a consistent improvement in classification accuracies when employing ensemble modelling methods. However, these methods have not been tested in distinction of "spectrally and structurally similar" species and from that originates the motivation for this study.

The study contributes to improved mapping of structurally and spectrally similar tree species based on two remote sensing data sources at individual tree scale. We compared their predictive performance and later integrated high density airborne LiDAR data and high resolution Worldview-2 optical satellite data and regression classifiers. We additionally evaluated the performance of three variable selection procedures in the selection of the best predictor subset across classifiers.

2. Materials and methods

2.1. Study area

The study site is located in the South Eastern alpine part of France, district of Barcelonnette (latitude 44° 25′ 22.87″ N and longitude 6° 40′22.43″ E; Fig. 1). The area is about 1.3 km² and is covered by mainly unmanaged forests of *P. sylvestris* and *P.uncinata* (≈95% of total area). Norway spruce (*Picea abies* (L.) H.Karst.), European Larch (*Larix decidua* Mill.) and some other broadleaved species cover the remaining ≈5% of the study area. Altitude ranges between 1400 and 2020 m.a.s.l. Topography is irregular (mountainous) with slope gradients ranging between 10° and 70°. This type of terrain posses some challenges especially during the generation of elevation models and topographic normalization of spectral responses.

2.2. Characteristics of studied species

To demonstrate the potential of ensemble techniques, our study focused on the two dominant species in the study area. *P. sylvestris* and *P. uncinata* have chromatic and morphological differences that can be leveraged for their distinction (Farjon, 2010). These are: (1) the crown of *P. uncinata* is conical with narrow spreading lateral branches whereas that of *P. sylvestris* is conical-ovoid to ovoid with widely spreading to ascending lateral branches when mature but conical when young; (2) The density of the branches varies with the growth of the tree but is generally denser and grows to a lower base height for *P. uncinata* than *P. sylvestris*; (3) Visually, *P. uncinata* has a greyish-black trunk whereas *P. sylvestris*'s bark is reddish gray at the base and orange at the thin bark of the upper trunk and major branches; (4) *P. uncinata* is generally shorter (12–20 m) than *P. sylvestris* (15–35 m) at maturity and (5) the barks of both species are scaly plated or fissured with varying degrees of surface roughness but generally with the bark of *P. sylvestris* rougher than that of *P. uncinata* (Farjon, 2010). Despite these differences in crown appearance and architecture, a visual distinction of *P. sylvestris* and *P. uncinata* via needles can be difficult (Fauvart et al., 2012). This property exacerbates the challenge of their characterization from optical remote sensing as needles contribute the largest portion of reflectance in the imagery. We expected that combining LiDAR and Worldview-2 datasets as well as regression classifiers would help capture various biophysical traits and therefore increase the chance of their automatic distinction.

2.3. Data

LiDAR data, a 15 × 15 cm RGB aerial orthophoto and Worldview-2 (2 × 2 m, 8 bands and 0.5 × 0.5 panchromatic band) datasets were acquired during leaf-on and snow free conditions in June of 2009 and September of 2010, respectively. LiDAR data and the aerial orthophoto
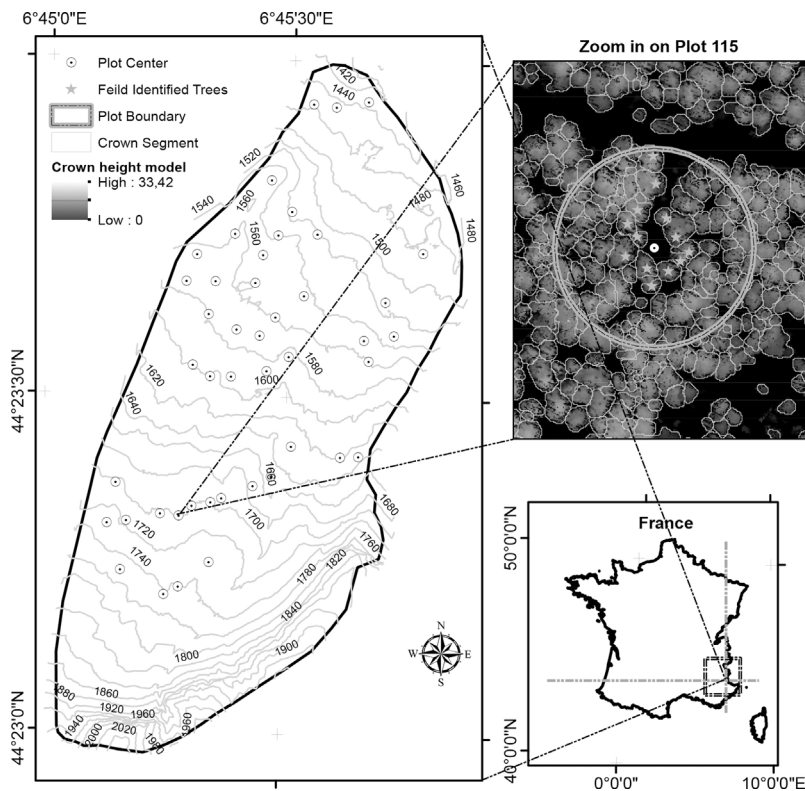
Fig. 1. Map of the study area located within the following X and Y UTM 32N coordinates (WGS84 reference system): 320,797.6; 4,916,782.5 (lower left) and 321,992.5; 4,918,787.1 (upper right).

were collected from an aircraft flying at an average 300 m above ground. A Riegl VQ480i system with a pulse repetition rate of up to 300 kHz and a rotating mirror scanning method was used. A very high density of the point cloud (160 points m^{-2}, seven discrete returns per pulse) was achieved from seven overlapping flight lines.

For the Worldview-2 imagery, level 2 post processing had been done by the vendor. The average sun elevation and azimuth angles were 48.1° and 161.7° and the average satellite elevation and azimuth angles were 74.8° and 55.0° respectively. The sun elevation and azimuth angles were used as input parameters during topographic normalization of the images to eliminate potential effects of shadows resulting from the rugged terrain.

2.4. Fieldwork

Field work was done in autumn 2012, 2- and 3-years after collection of LiDAR and Worldview-2 imagery respectively. A total of 48 circular field plots of 500 m^2 (with slope correction) were measured. Within each plot, all trees with diameter at breast height (DBH, 1.3 m above the ground) larger than 7 cm were recorded and callipered in two orthogonal directions. However, of all the trees encountered on the plots, only a subset of 544 trees were identified and their location recorded: 273 individuals of *P. sylvestris* and 271 individuals of *P. uncinata*. Note that the 544 trees are those whose crowns were visible in the dominant canopy and whose treetops could be determined on the canopy height model (CHM). We were able to verify locations of the identified dominant trees in the field using a CHM (Fig. 2), whereby the ground horizontal distance and orientation from a landmark were compared to the CHM distance and orientation. The landmarks included other isolated trees and canopy gaps.

2.5. Canopy height model generation and individual tree crown (ITC) delineation

The LiDAR Digital Elevation Model (DEM) and Digital Surface Model (DSM) were generated by griding Delaunay triangulated ground

and first returns into 15 cm resolution rasters respectively. Tree position, height, and crown width was retrieved from the CHM by using a modified version of the mixed-pixel and region-based algorithm designed by González-Ferreiro et al. (2013). This algorithm is proposed as a sequence of routines programmed in the integrated development environment (IDE) (eCognition Developer 8.7 (®Trimble GmbH, Munich, Germany)). All logical procedures in the construction of the canopy delineation algorithm were arranged into five groups: CHM smoothing, segmentation, classification of canopy areas, iterative process, and data export. For this study, one parameter of the algorithm has been tuned, to adapt to the analyzed species. Concretely, we have changed the break value for the roundness index from 0.5 to 0.25 in the shape decision criteria (see the Fig. 3 in González-Ferreiro et al. (2013)) in order to discard and control for unnatural crown shapes.

Finally, the ITC result (Fig. 2) and crown attributes were then exported as vector polygons in an ESRI™ shapefile with the associated database. Tree tops and height attributes were also exported as a point vector shapefile for subsequent analysis.

2.6. Assessment of accuracy of delineated crowns

A good ITC segmentation was prerequisite in order to accurately isolate individual tree predictors. To evaluate the accuracy of the crown delineation algorithm, we closely followed two area-based methods presented by Clinton et al. (2010). The reference polygons (assumed truth), relative to which the performance of the segmentation algorithm was judged, were manually digitized from a CHM overlain on a high resolution (0.15 × 0.15 m) aerial orthophoto collected together with the LiDAR data. Three student assistants performed this task, stratawise – i.e. the study area was divided into three parts – and each independently, on desktop screens of 1600 × 900 pixel resolution and at a common map scale of 1:50. The assistants were trained before the digitization according to a defined protocol detailing ways to identify crown boundaries based on changes in grey-color contrasts. For all study participants, 1615 individual crown polygons spread across the entire study area, were manually digitized and used in the assessment
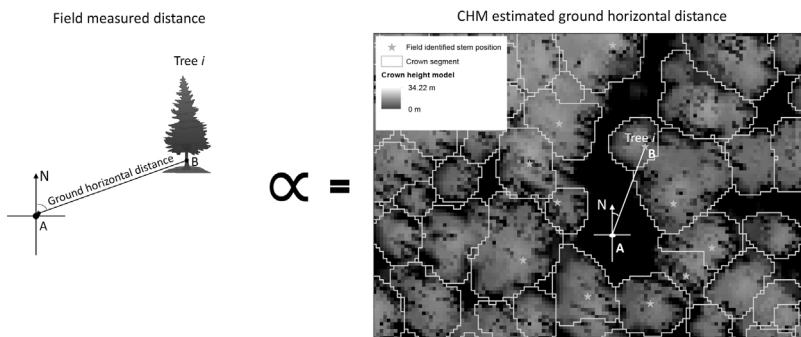
Fig. 2. Identification of individual tree stem locations in the field, whereby; A is the plot center or landmark, B is the stem position, i is the tree identification number ranging from 1 to 544 and N signifies the North direction. The identification of tree stem positions on the Canopy Height Model is shown on the right whereby, ground horizontal distance (A to B) ≈ CHM distance (A to B) × CHM image resolution. The crown segments are an output of automatic individual tree crown delineation using the Canopy Height Model as input data.

of the segmentation accuracy.

We computed the segmentation goodness metric (*Eq. (1)*) for each reference polygon based on relative overlap with the largest corresponding automatically generated segment (Clinton et al., 2010).

$$\text{segmentation goodness}_{ij} = 1 - \frac{A_{Ii}}{A_{Rj}} \quad (1)$$

A_{Ii} is the area in intersection ($A_{Ai} \cap A_{Rj}$) between an automatically generated segment and a corresponding manually delineated segment, A_{Ai} is the area of segment i and A_{Rj} is the area of the corresponding reference segment.

To compare pairwise sizes and context of automatically generated crowns relative to the reference polygons, we computed the area fit index (*AFI*) (Clinton et al., 2010) as shown in *Eq. (2)*.

$$\text{AFI}_j = \frac{A_{Rj} - A_{maxAi}}{A_{Rj}} \quad (2)$$

AFI_j is the Area Fit Index for segment j and A_{maxAi} is the area of the largest automatically generated segment intersecting with A_{Rj}.

Following Clinton et al. (2010), we classified over-segmented reference polygons (*O*) as reference objects with less than 100% overlap with automatic polygons and $AFI > 0$ and under-segmented reference polygons (*U*) as those of 100% overlap with automatic polygons and $AFI < 0$. By averaging over all over-segmented and under-segmented polygons we computed the closeness index D shown in *Eq. (3)*.

$$D = \sqrt{\frac{O_{\in k}^2 + U_{\in m}^2}{2}} \cdot \quad (3)$$

O is the mean segmentation goodness across the k over-segmented reference polygons and the term U is the mean segmentation goodness across m under-segmented reference polygons. We judged our segmentation result based on Clinton et al. (2010) that identified closeness index values of 0.3 as good segmentations.

Lastly, an overall rate of correct detection was computed as the ratio of correctly detected reference polygons and all reference polygons. Correct detection was defined by the following limits: overlap ⩾70% and $-3 \leqslant AFI \geqslant 2$. The *AFI* thresholds were determined based on visual inspection.

2.7. Preparation of predictors

The individual crown segments delineated in the previous step were used to isolate individual crown variables from both remote sensing datasets. From the height normalized point cloud, metrics describing per-crown return intensity and height distributions and variables describing sizes and shapes of individual crowns were extracted using the FUSION LiDAR Toolkit (McGaughey, 2014). As no flight trajectory file was available for this study, radiometric normalization of intensity data was done following González-Ferreiro et al. (2014) and based on a user defined standard range. Here note that Korpela et al. (2010) report a marginal gain in accuracy (2–3%) of species classification after intensity normalization while distinguishing conifers in boreal conditions.

Fassnacht et al. (2016) mention that *"this task can be quasi-impossible in mountainous terrain"*.

Standard crown size and shape variables were extracted from the points and polygon shapefiles generated in the ITC delineation. LiDAR metrics were derived from the normalized LiDAR height and intensity distributions within the limits of the delineated individual tree crowns. The minimum height threshold (MHT), which is commonly specified as the lower boundary for calculating height metrics (central tendency, dispersion, shape and percentile statistics), was established at 1 m. The height break threshold (HBT), which is the limit for separating the point cloud data into two sets to separate canopy returns from the under canopy returns, in order to estimate canopy cover metrics, was established as 5 m (based on field observation). In total 108 metrics were derived from the LiDAR data (see Table 1 for a complete description of LiDAR-derived metrics).

Before extracting per-crown optical metrics, topographic normalization based on local parameter estimation of spectral differences (Mo et al., 2015) was done using the LiDAR DEM. The topographic normalization aimed at reducing radiometric distortions brought about by shadow effects in rugged terrain especially enhanced by a low scanning and sun elevation angle at the time of image acquisition. Optical images were resampled using a nearest neighbor interpolation from 2 m to 0.5 m before computing the per crown statistics (Table 1). The following broadband greenness and leaf pigmentation products were derived: Normalized Difference Vegetation Index (NDVI), Simple Ratio Index (SRI), Enhanced Vegetation Index (EVI), Atmospherically Resistant Vegetation Index (ARVI) and Anthocyanin Reflectance Index (ARI). Additionally, per crown statistics were computed from texture indices derived from the panchromatic band including; energy, entropy, correlation, inverse distance moment, inertia, cluster shade, cluster prominence and Haralick correlation. The same statistics were calculated for each of the eight multispectral bands (Table 1). In total 230 spectral predictors were prepared.

2.8. Statistical models and calibration

The predictors described above were used to distinguish between the two tree species via multiple regression approaches. We used the default settings of the *biomod2* R-package version 3.3-7 (Thuiller et al., 2016) to ensemble results of 9 regression techniques into a final prediction. The models are described in detail in . The included modelling techniques were; Classification and Regression Trees (CTA), Generalized Linear Models (GLM), Generalized Boosting Models (GBM), Generalized Additive Models (GAM), Artificial Neural Networks (ANN), Flexible Discriminant Analysis (FDA), Multiple Adaptive Regression Splines (MARS), Random Forests (RF) and MAXENT.Phillips. The models were calibrated with species data based on the field observations.

Multi-collinearity analysis was done prior to the fitting process to assess whether two or more explanatory variables were significantly correlated. Collinearity among all explanatory variables was evaluated by the Variance Inflation Factor (VIF) (Fahrmeir et al., 2013, P. 156),

Table 1

Potential LiDAR and optical predictors. Note that descriptions of metrics are listed in the order of appearance of the acronyms and not repeated for each variable group. Metrics are grouped into: Intensity (*I*), Crown cover, shape and size (*C*), Height (*H*), Individual bands (*B*1, …, *B*8), Texture indices (*T*) and Vegetation indices (*V*). "…" refers to "in the same sequence ".

Abbreviation	Description
LiDAR metrics on cover, size and shape of crowns (C)	
C_{cr}, C_{cwr}, C_{len}, C_{cls}, C_{cbh}, C_{per}, C_{per1}, …, C_{per4}, C_{crr}, C_{r1}, …, C_{r7} .	Crown radius, Crown-width ratio, Crown perimeter length, Crown closure, Crown base height, Percentage first returns above 5 m, Percentage first returns above mean, Percentage first returns above mode, Percentage all returns above 5 m, Percentage all returns above mean, Percentage all returns above mode, Canopy relief ratio, Count of returns by return number (1-7).
LiDAR metrics on height (H) and intensity (I) distributions – dispersion statistics	
H_{IQD}, H_{SD}, H_{var}, H_{CV}, H_{skew}, H_{kur}, H_{AAD}, H_{MeAD}, H_{MoAD}, H_{L2}, …, H_{L4}, H_{LSkew}, H_{LKur}, I_{IQD}, …, I_{LKur}.	Inter-quartile distance, Standard deviation, Variance, Coefficient of variation, Skewness, Kurtosis, Average absolute deviation, Median of the absolute deviations from median, Median of the absolute deviations from mode, L-Moments (2:4), L-Moment of Skewness, L-Moment of Kurtosis.
LiDAR metrics on height (H) and intensity (I) distributions – descriptive statistics	
H_{max}, H_{min}, H_{mean}, H_{med}, H_{mode}, H_{P01}, …, H_{P99}, I_{max}, …, I_{P99}.	Maximum, Minimum, Mean, Median, Mode, Percentiles (1,5,10,20,25,30, …, 70,75,80,90,95,99).
Spectral metrics from individual bands (B) and vegetation indices	
$V_{minNDVI}$, $V_{maxNDVI}$, $V_{sumNDVI}$, $V_{cntNDVI}$, $V_{meanNDVI}$, V_{sdNDVI}, $V_{uqeNDVI}$, $V_{rangeNDVI}$, $V_{varNDVI}$, $V_{medNDVI}$, $V_{modeNDVI}$, V_{minSRI}, …, $V_{modeSRI}$, V_{minEVI}, …, $V_{modeEVI}$, $V_{minARVI}$, …, $V_{modeARVI}$, V_{minARI}, …, $V_{modeARI}$, V_{minB1}, …, V_{modeB1}, …, V_{minB8}, …, V_{modeB8}.	Minimum, Maximum, Sum, Count, Mean, Standard deviation, Unique, Range, Variance, Median, Mode of: vegetation indices (NDVI, EVI, ARVI, ARI), and individual bands (*B*1, …, *B*8).
Spectral metrics from texture indices (T)	
$T_{minInertia}$, $T_{maxInertia}$, $T_{sumInertia}$, $T_{cntInertia}$, $T_{meanInertia}$, $T_{sdInertia}$, $T_{uqeInertia}$, $T_{rangeInertia}$, $T_{varInertia}$, $T_{medInertia}$, $T_{modeInertia}$, $T_{minEner}$, …, $T_{modeEner}$, T_{minEnt}, …, $T_{modeEnt}$, T_{minCor}, …, $T_{modeCor}$, T_{minIDM}, …, $T_{modeIDM}$, T_{minCP}, …, T_{modeCP}, $T_{minHCor}$, …, $T_{modeHCor}$.	Minimum, Maximum, Sum, Count, Mean, Standard deviation, Unique, Range, Variance, Median, Mode of: Inertia, Energy (Ener), Entropy (Ent), Correlation (Cor), Inverse distance moment (IDM), Cluster prominence (CP), Haralick correlation (HCor).

considering VIF > =10 as the step-wise elimination threshold (Duque-Lazo et al., 2016).

To overcome dimensionality issues commonly faced in multiple regressions of small sample sizes to many predictor variables, and so as to select for the most robust predictor set, three variable selection procedures were considered, namely: (1) AUCRFcv (Calle et al., 2011), (2) AIC and (3) BIC (Fahrmeir et al., 2013).

A variable importance measure (VI) was used to assess the relative contributions of individual variables in a given model. To allow for comparability of variable importance across models, VI was computed from model predictions (for all model types). After obtaining a prediction (P_1), the variable of interest was shuffled while holding all the other variables in the model at their median value and a new model prediction made (P_2). A simple correlation between P_1 and P_2 was then calculated to obtain VI according to Eq. (4):

$$\mathrm{VI} = 1 - \mathrm{cor}(P_1, P_2) \tag{4}$$

A zero *VI* value meant no importance whereas the importance of a variable in a predictor set increased with VI scores (Thuiller et al., 2016).

2.9. Model and map evaluation

Model fit was evaluated using the relative operating characteristic (ROC) and map quality was assessed with the true skills statistic (TSS). Both ROC and TSS were computed from cross validation and split sample procedures. During cross validation, all observational data (544 field identified trees) was utilized where models were fit with 80% of the data and evaluated with the remaining 20% of the data. An almost 1:1 ratio, 271 training and 273 validation, was used in the split sample approach. The same training and validation set was kept constant across all runs and variable selection procedures. Models were run 50 times in order to monitor variation in variable importance and prediction accuracies. Robust models were defined by a combination of both ROC and TSS.

2.10. Ensemble prediction

A stacked assemblage of predictions across individual models was based on mean, median, inferior confidence interval, superior confidence interval, committee average and a weighted mean (Thuiller et al., 2016). Not all individual models computed per run were included in the ensembles. Different ROC quality thresholds were tested to select models into the ensemble and minimize loss in the resultant ensemble model and map accuracy brought about by inclusion of weak or less informative individual models. Eventually, ROC = 0.8 was used. The model and map evaluation described in Section 2.9 applies to ensemble models.

Finally, three modelling schemes were considered based on: (1) high-density LiDAR data, (2) spectral image layers derived from multispectral imagery (Worldview-2), and (3) both the spectral and LiDAR data. For each scheme, all variable selection and model types were evaluated.

3. Results

3.1. Individual crown detection

More than 70% of dominant crowns in the canopy were correctly delineated according to the defined criteria. Closeness index values averaged around 0.33 as shown strata-wise in (Table 2).

Table 2

Segmentation goodness across study participants.

Participant	Segmentation measures		No. of Crowns
	Correct detections (%)	Closeness index (D)	
I	73.36	0.31	517
II	60.65	0.41	310
III	77.48	0.33	826
Total	73.18	0.33	1615
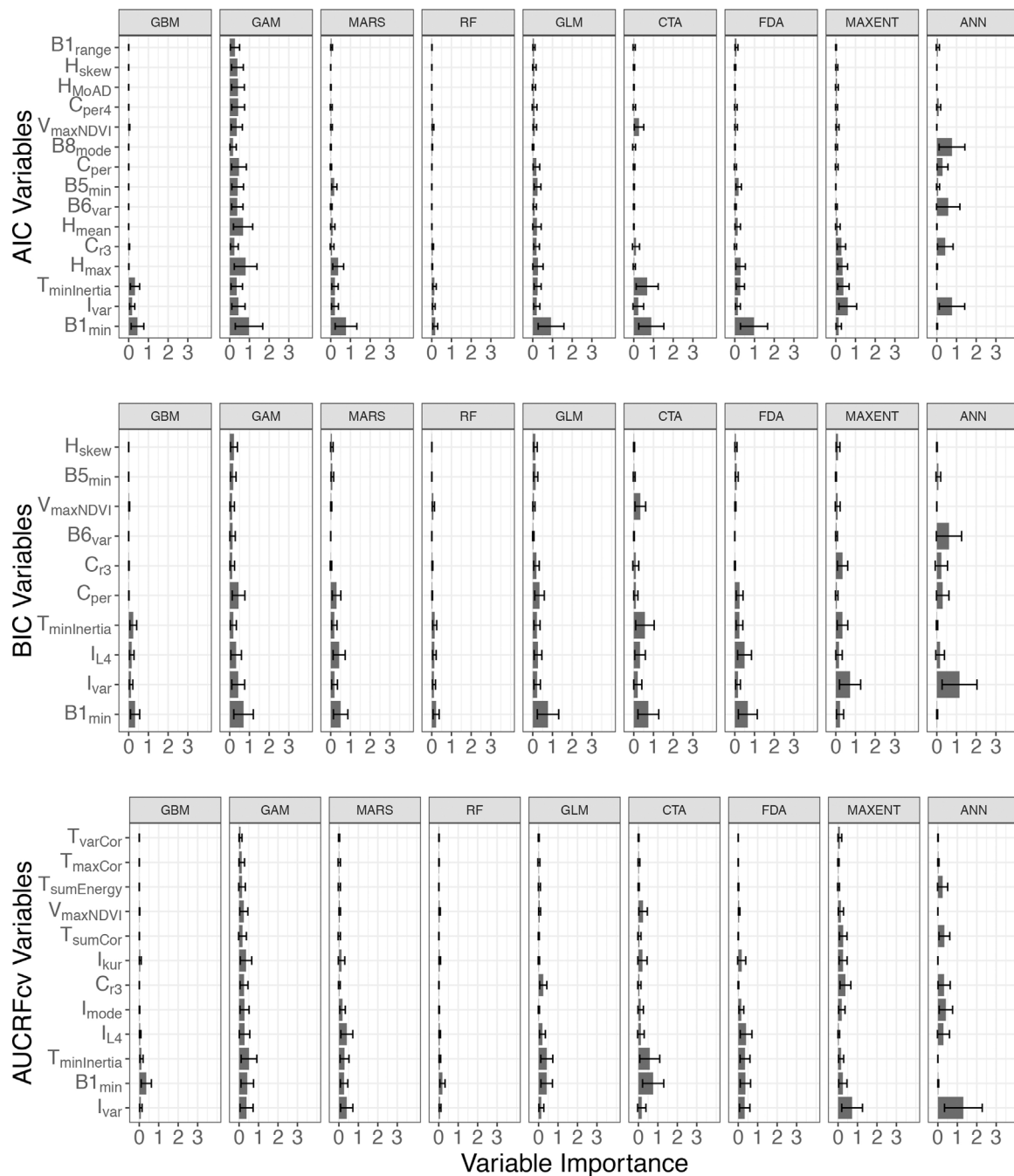
Fig. 3. Mean variable importance of spectral and LiDAR predictors across 50 model runs. Variable importance is sorted across models in an ascending order. Sorting of variable importances was global so as to identify variables of highest importance across models. Groups of metrics are: Intensity (I), Height (H), Crown cover, size and shape (C), Texture index (T), Vegetation index (V) and Individual band (B1, …, B8). For detailed predictor variable names including explanation of the subscript in each acronym, refer to Table 1. Models: Artificial Neural Networks (ANN), Classification and Regression Trees (CTA), Flexible Discriminant Analysis (FDA), Generalized Additive Models (GAM), Generalized Boosting Models (GBM), Generalized Linear Model (GLM), Multiplicative Regression Splines (MARS), Maximum Entropy (MAXENT.Phillips), Random Forests (RF).

3.2. Importances of predictor variables

Both LiDAR and optical metrics were of high importance in the computed models (Fig. 3). We suggest that identification of the most important predictors considers both the magnitude of variable importance (as shown in Fig. 3) and the consistence of selection across modelling schemes (Table 3). The reader is referred to supplementary material for importances of variables selected when the datasets were employed in isolation (individually). Amongst the LiDAR variables, the following were of high importance: I_{var}, I_{L4}, C_{per}, C_{r3}, and H_{max}, and in Worldview-2 imagery: $B1_{min}$, $B5_{min}$, $T_{minInertia}$, and $V_{maxNDVI}$.

Judged by the variable importance scores alone (i.e. without considering consistency of selection across modelling schemes), LiDAR intensity, optical imagery texture indices, and individual band reflectance characteristics received higher importance in the models compared to height, canopy cover metrics and vegetation indices (Fig. 3). Similar trends were observed when the datasets were employed in isolation (Appendices 2 and 3). There was generally agreement across individual models and variable selection procedures (Fig. 3). However, GAM and ANN assigned higher importance to variables less important in other models. Additionally, some models variable importances were more precise across runs compared to others (based on

17

Table 3

Selected predictors across modelling schemes and variable selection procedures. Metrics are grouped into: Intensity (I), Crown cover, shape and size (C), Height (H), Individual bands ($B1$, …, $B8$), Texture indices (T) and Vegetation indices (V). ☑ Refers to selected into the final predictor set given a variable selection procedure. For detailed predictor variable names including explanation of the subscript in each acronym, refer to Table 1.

	Predictor	Both			LiDAR		
		AUCRFcv	BIC	AIC	AUCRFcv	BIC	AIC
1.		I_{var}	☑	☑	☑	☑	☑	☑
2.		I_{L4}	☑	☑		☑	☑	☑
3.	Intensity	I_{kur}	☑			☑		☑
4.		I_{mode}	☑			☑		
5.		I_{L3}				☑		
6.		I_{P01}						☑
7.		C_{r3}	☑	☑	☑	☑		☑
8.		C_{r2}				☑		
9.		C_{r1}				☑		
10.	Crown cover,	C_{per}		☑	☑	☑	☑	☑
11.	size & shape	C_{per4}			☑			
12.		C_{length}				☑		
13.		C_{per3}						☑
14.		C_{cr}						☑
15.		H_{skew}		☑	☑	☑		
16.		H_{max}			☑	☑	☑	☑
17.		H_{mean}			☑	☑	☑	☑
18.	Height	H_{MoAD}			☑		☑	☑
19.		H_{L3}				☑		
20.		H_{var}				☑		
21.		H_{kur}					☑	
22.		H_{CV}						☑

	Predictor	Both			Spectral		
23.		$B1_{min}$	☑	☑	☑	☑	☑	☑
24.		$B5_{min}$		☑	☑	☑	☑	☑
25.		$B6_{var}$		☑	☑			
26.	Individual	$B1_{range}$			☑		☑	☑
27.	Band	$B8_{mode}$			☑	☑		
28.		$B4_{med}$					☑	☑
29.		$B1_{var}$						☑
30.		$B2_{mode}$						☑
31.		$B8_{var}$						☑
32.		$T_{minInertia}$	☑	☑	☑		☑	☑
33.		$T_{sumEnergy}$	☑			☑		
34.		T_{sumCor}	☑			☑		
35.		T_{maxCor}	☑			☑		
36.	Texture	$T_{varEnergy}$	☑			☑		
37.	index	T_{varCor}	☑			☑		
38.		T_{sumCP}						☑
39.		$T_{varInertia}$						☑
40.		T_{minCS}						☑
41.		T_{modeCP}						☑
42.		$V_{maxNDVI}$	☑	☑	☑			
43.	Vegetation	V_{minSRI}				☑	☑	☑
44.	index	V_{sumARI}				☑		

standard error bars in Fig. 3). There was a higher precision in *VI* from RF, GBM and GLM compared to the other model categories.

3.3. Individual model and map quality

Two clusters in performance of individual models were observed when comparing individual models (Fig. 4). In general, RF, GBM, GLM, FDA, MARS formed a cluster of the better performing models compared to MAXENT.Phillips, CTA, ANN and GAM in the cluster of lower performing models. The former cluster had considerably lower variance in both quality of model fit and map accuracies across model runs. There was preponderance of linear models when spectral predictors were employed in isolation and a majority of classification tree models when LiDAR predictors were employed (Fig. 4). Additionally, spectral models were more stable than models using LiDAR only and Spectral and LiDAR. Generally, individual model and map quality across variable selection procedures was similar (Fig. 4).

All individual models and datasets failed to differentiate the species when employed independently (Fig. 4). Individual performance among the best five models was on average TSS = 0.55 for LiDAR and TSS = 0.64 for Spectral. However, the average individual performance of the same models increased slightly to about TSS = 0.67 when both datasets were integrated. In isolation, spectral variables exhibited higher predictive power compared to LiDAR variables.

3.4. Ensemble model and map quality

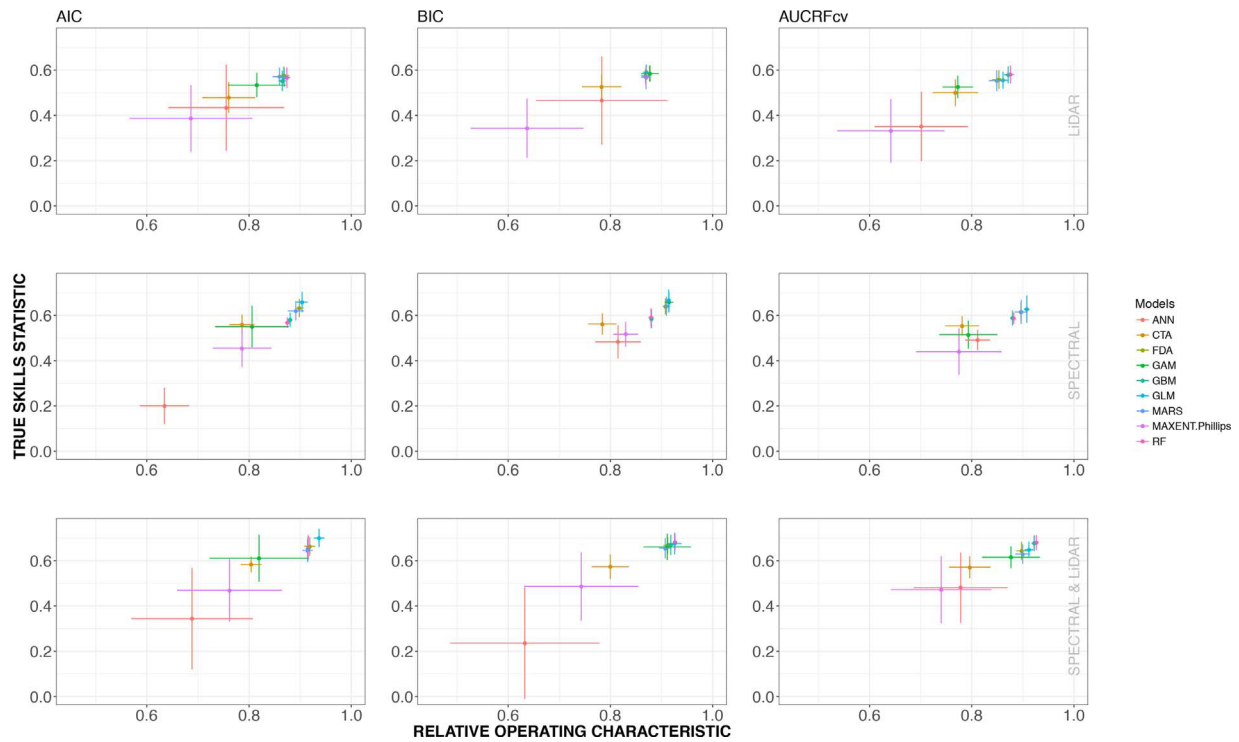In general, all ensemble models performed equally for both

Fig. 4. Performance by model type (Artificial Neural Networks – ANN, Classification and Regression Trees – CTA, Flexible Discriminant Analysis – FDA, Generalized Additive Models – GAM, Generalized Boosting Models – GBM, Generalized Linear Model – GLM, Multiplicative Regression Splines – MARS, Maximum Entropy – MAXENT.Phillips and Random Forests – RF) and variable selection procedure (Top row = LiDAR, Middle row = Spectral, Bottom row = Spectral & LiDAR). The dots represent the mean and the lines represent the associated standard errors across 50 model runs.
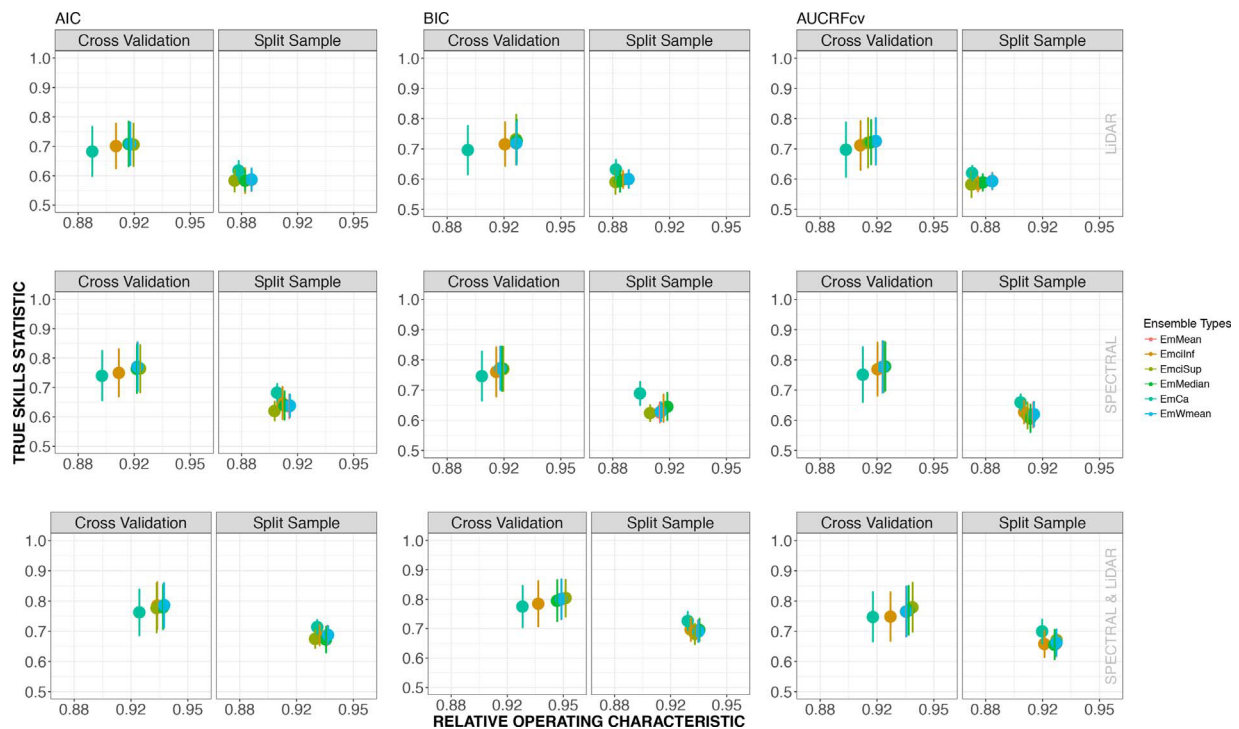
Fig. 5. Performance by ensemble model type (Mean – EmMean, Inferior confidence interval – EmciInf, Superior confidence interval – EmciSup, Median – EmMedian, Committee average – EmCa and Weighted mean EmWmean) and variable selection procedure (Top row = LiDAR, Middle row = Spectral, Bottom row = Both). The dots represent the mean and the lines, the associated standard errors across 50 model runs.

validation procedures. Assemblage of individual models increased map quality with across-model TSS averages to 0.63 for LiDAR models, to 0.68 for spectral models and to 0.73 for both spectral and LiDAR powered models (Fig. 5). Higher model and map quality values were

reported by the cross validation procedure (Fig. 5). The Committee Average ensemble (EMca) was consistently the highest performing ensemble when evaluated using a split sample approach. EMca was the worst performing ensemble model based on cross validation.
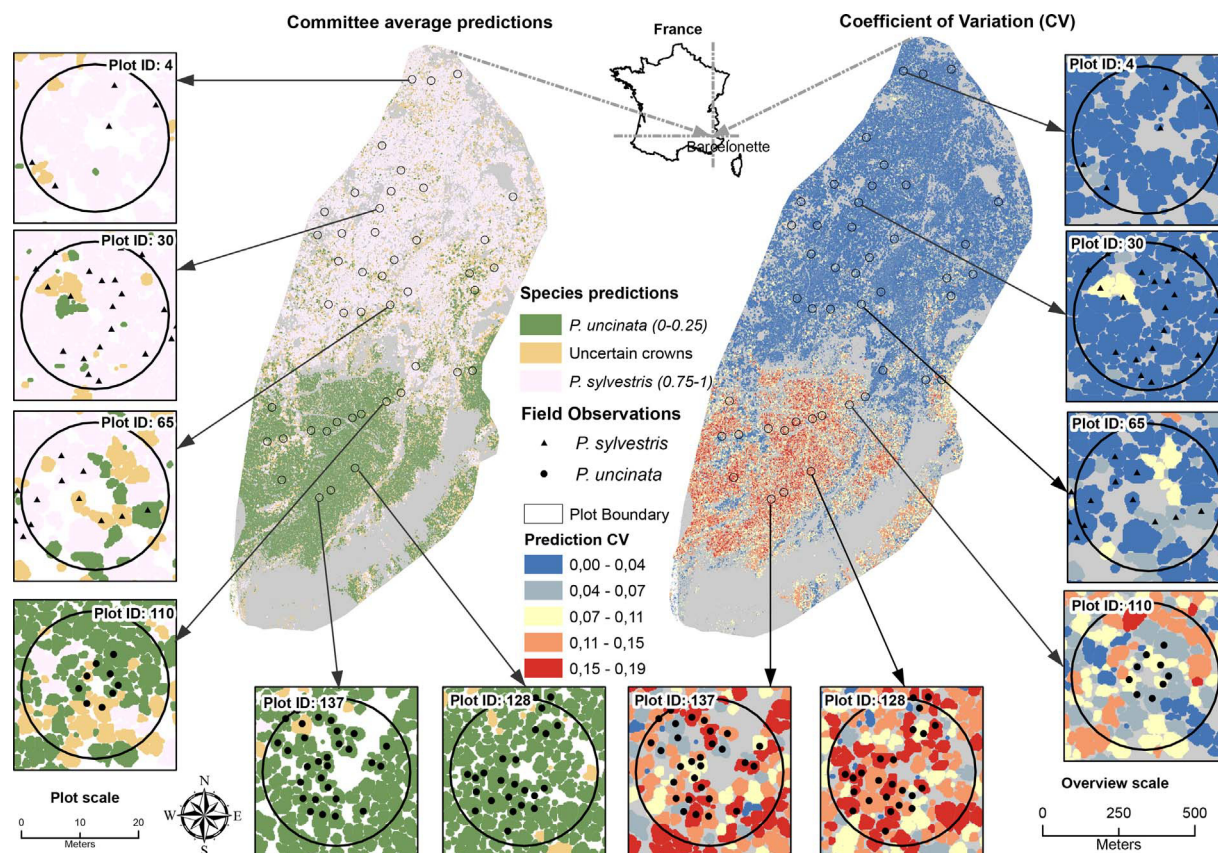
Fig. 6. An individual tree species map and associated across model uncertainty computed via the coefficient of variation.

Fig. 6 shows example maps of tree species at individual crown level. The maps show a single committee average prediction and the respective uncertainty estimate i.e. Coefficient of Variation (CV) across modelling techniques. There was higher uncertainty across models for predicting *P. uncinata* than was for predicting *P. sylvestris*. However, the uncertainty was generally low (CV ≤19%) indicating general agreement among individual models in the ensemble.

4. Discussion

Our result shows that it is possible to perform an accurate automatic tree crown segmentation (Closeness index = 0.33) of two very similar *Pinus* species in complex terrain, and that unlike the variable selection procedure, a combination of LiDAR and multispectral variables improved model prediction. More concretely, we have demonstrated that the ensemble model approach is superior to single model prediction. However, the key challenge is a high intra-species variance in classifications of individual trees.

4.1. Effect of a high intra- vs inter-species variation

A common recommendation when variance is greater within species is to employ object-oriented techniques (Gougeon and Leckie, 2006) with a higher spatial or spectral resolution, either in isolation or after some form of fusion (Leckie et al., 2003; Ke et al., 2010) in order to summarize the within class (tree) variation at a lower scale of observation. At the scale of single trees (crowns) our results have shown that the high within-tree variation can persist (Fig. 4). However, classification accuracies can improve with ensemble models (Fig. 5). The very high intra-species variation is attributed to differences in individual tree morphology and differences in local site conditions. No management planning information was available to this study, and neither was tree age data collected. We however did observe that there

were differences in stand characteristics (for example, the size of trees) suggesting differences in either site specific rates of growth or stand ages. The difference has influenced within species variation in particular related to structure of crowns. *P. sylvetris* crowns are known to change with stages of maturity (Farjon, 2010; Ross et al., 1986). Similarly, the rugged terrain introduced variations in site relative to local slope, aspect and soil conditions and thus increased tree-scale within species variation. The terrain effect is further seen in the spatial pattern of model uncertainty (Fig. 6) as dissimilarity among models increased with elevation and ruggedness (Fig. 1) in areas dominated by *P. uncinata*. In other words, errors in normalization of remote sensing datasets induced higher model uncertainty in areas of rugged terrain. It should be noted however, that we utilized very high point density LiDAR which supported generation of very high quality elevation models uncommonly possible in complex terrain. This enhanced robustness of ITC delineation and precision of within crown statistics. In common practice, a good ITC delineation is possible with densities from between 4 and 10 points m^{-2} (Hamraz et al., 2017). However, at such lower densities, it will be probable that fewer points inside of each segment (delineated crown) affect precision of computed statistics and metrics. Therefore, additional within species variation could arise from errors in ITC delineation and metrics calculation. However, the latter two sources of variation are here considered of low importance based on the high segmentation accuracy achieved and the employment of advanced topographic and intensity normalization procedures. The higher intra-species variation triggered the need to utilize individual models with several predictor variables.

4.2. Important species predictors

As expected, a combination of both LiDAR and spectral variables resulted in increased differentiation capabilities as either datasets captured varying characteristics of the species crowns. Similarly, a

comparison between prediction power of multispectral and LiDAR variables in isolation suggested higher structural than spectral similarity between the species; resulting in a higher prediction power for spectral predictors. Combined, the most important spectral predictors: $B1_{min}$, $B5_{min}$, $T_{minInertia}$, and $V_{maxNDVI}$, highlight differences in condition, pigmentation and internal structure of the needles useful for their characterization. Similar conclusions were drawn by Alvarez et al. (2009) using microscopy in analysis of epidermal characteristics of needles of both species. Additionally, we attribute the high importance of the texture index $T_{minInertia}$ to the combined effect of chromatic differences in the upper trunk and branches of the two species and their differences in branch density (refer to Section 2.2 for details). $T_{minInertia}$ measures inter-pixel contrast within each crown area. The 50 cm resolution of the Worldview-2 imagery captured inter-pixel variation in brightness of the leaves and trunks that was useful in the distinction of the species. On the other hand, LiDAR variables highlighted the importance of bark roughness (via intensity metrics, I_{var}, I_{L4}), branch density (via canopy cover metrics, C_{per}, C_{r3}) and height (H_{max}) differences in characterization of the species.

4.3. Evaluation of individual models

The peculiar clustering in individual model performance showed; MAXENT.Phillips, ANN, CTA and GAM to have been consistently poorly performing and unstable across multiple runs. This was unsurprising for GAM and ANN since they simultaneously assigned higher importance to variables less important in other models (Fig. 3). In a similar ensemble setting, akin results regarding variable usage in ANN models were reported by Marmion et al. (2009). They related this unique behavior of the ANN model to its inherent non-selective nature, given that it builds intermediate relationships between predictors and therefore risks assigning higher importance to less effective variables based on their indirect contribution to the prediction process. On the other hand, GAMs fit local splines along the data range of each available variable, estimate a single smooth curve per variable, and then additively combine the results. Both structures of the GAM and ANN models are prone to becoming exceedingly complex and may therefore face over-fitting (Marmion et al., 2009). The challenge therefore was finding the optimal threshold between individual model complexity and overall ensemble mapping accuracy. Here, more complex individual models resulted in higher map accuracies attributed to their ability to minimize model bias. However, based on the comparison between the split-sample and cross-validation results (Fig. 5), we posit that the observed marginal discrepancy between map accuracies reported by both validation procedures is testimony that over-fitting was not a significant problem since ensemble models generalized well on new data.

Further, the poor performance of the CTA model is linked to the lack of bagging and bootstrapping capabilities, disposing the model type to bias and high across run variances (Briem et al., 2007). One notes that despite their similarity in structure, the RF model consistently outperformed CTA. This can be explained by regularization, bagging and bootstrapping algorithms in the RF model affording better performance at generalization on validation datasets. On the other hand, it was unique to this study that the performance of MAXENT.Phillips was not comparable with other models, such as RF and GBM, that are known to be consistently highest performing (Elith et al., 2011). We found out that the poor performance of MAXENT.Phillips – in this specific case – is linked to its calibration with presence-absence data rather than presence only data and to a sub-optimal specification of the prevalence parameter (set here to 0.5) from which a logistic output is generated after combining presence (here *P. sylvestris*) and background (here *P. uncinata*) data (Guillera-Arroita et al., 2014). Based on several modelling techniques, Fig. 6 shows that *P. sylvestris* potentially has a higher prevalence in the study area compared to *P. uncinata*. Guillera-Arroita et al. (2014) recommend to arrive at a prevalence estimate when parameterizing MAXENT.Phillips via the sample data, however, our

individual tree selection approach did not afford the option to estimate unbiased prevalence before hand. The ideal practice would have been modelling the distribution of each species separately and employing the respective prevalence parameters independently. Nonetheless, we emphasize that the individual performance of MAXENT.Phillips had marginal to no effect on the presented ensemble results, given that an ROC threshold of 0.8 was used in selection of models into the ensemble which MAXENT.Phillips rarely passed. Remember that even within the cluster of better performing models (i.e. GLM, MARS, FDA, GBM, RF), the best individual model only achieved up to 40% better-than-random map accuracy (TSS = 0.69).

4.4. The contribution of an ensemble approach

Ensemble modelling improved individual models by increasing both map accuracy and minimizing prediction variance. The improvement was achieved through stacked fusion of predictions as well as due to the fact that the ensemble approach gives more weight to models with both good fit and effective bias-reduction. The bias minimization properties stemmed from tree-based and gradient boosting algorithms – such as RF and GBM – that sequentially arrived at a prediction by building several classifiers in a complementary tandem (Banfield et al., 2007; Briem et al., 2007). This way, our result is in agreement with Engler et al. (2013) that assembled per-pixel predictions of six species obtaining a cross validated Cohen's Kappa, $\kappa = 0.65$. For comparison with their study, the best mean cross-validation accuracy obtained in this study is TSS = 0.81 based on a combination of LiDAR and spectral datasets and BIC variable selection procedure (Fig. 5). A mean TSS of 0.81 means that the ensemble models were on average approximately 60% better-than-random at distinguishing *Pinus sylvestris* and *Pinus uncinata* at an individual crown scale. Note that TSS and κ are equivalent and that Engler et al. (2013) had to distinguish up to six species per pixel. Therefore, their expected error rate is higher. If the number of target species would have been more than two, we would have separate predictions of each species and spatially combine the result as done by Engler et al. (2013). However, we linked one species to the other and assigned class probabilities to either class in a single step. This approach had the advantage of computational efficiency but fell short for particular models (e.g. MAXENT.Phillips) as previously discussed. Similarly, it is worth explaining the discrepancy in behavior of the committee average ensemble (EmCa) given that it is the highest performing ensemble accuracy when evaluated in a split sample approach and the worst when evaluated by cross validation. This issue is related to the EmCa calculation which transforms all predictive probabilities into binaries (0 or 1) according to the maximum ROC/TSS threshold (the same as used in selection of individual models into the ensemble). Later, the new class probability is calculated as the average of the combined votes. During cross validation, a change in the training sample induces differences in model fit and therefore affects individual model votes – especially when encountering a high intra-class variation.

Further advantage of the ensemble approach rests in reporting across run variances of map accuracy scores. This is important in order to communicate the underlying precision of the estimated map accuracy. We show that with a single model, map accuracies can fluctuate across multiple runs with higher variances in cross validation than the split sample. At this point, it is important to understand that the two validation procedures convey fundamentally different assessments. While the cross validation approach estimates the expected prediction map accuracy, the split sample approach estimates both the conditional prediction map accuracy and model generalization capabilities (Hastie et al., 2009, p. 242). Expectedly, as with any model, our ensemble model generalization capabilities were lower (from TSS = 0.81 to TSS = 0.73) although better than individual models in isolation. The higher variances seen in the cross validation compared to the split sample approach are determined by the similarity of across fold training samples in the cross validation procedure (Hastie et al., 2009,

pp. 242–243). In other words, the more similar the training sample, the lower the across fold variation and therefore relates directly to the problem of intra-species variation. Certainly, the validation and test sample sizes as well affect the prediction precision when comparing between cross validation and split sample results.

Lastly, a spatial context to the precision of the estimated map accuracy is conveyed by across model coefficient of variation (Fig. 6) which communicates the level of agreement among individual models. This is a valuable resource when making inference towards the spatial variability of mapping errors. Therefore, such an output can in operational settings, for example, support selection of sites to focus map ground truthing exercises. Fig. 6 shows that there was a high agreement across models conveyed by a < 20% coefficient of variation. When such an estimate is used in tandem with the mean and variance estimates of map quality (e.g. TSS), it can enhance interpretation of the quality of the mapping product.

5. Conclusion

We have shown that assembly of regression models and integration of the datasets can provide a more reliable species distribution map with associated tree-scale mapping uncertainties. We have also shown that the approach can provide more transparent assessments of errors around modelled species distributions. Given that all the tried ensemble approaches performed equally, we do not recommend any in particular. We recommend leveraging of models and data assemblages in order to provide improved and transparent species classifications for forest planning, management and science at local to landscape scales. We speculate that the techniques and approaches used here lend themselves to other important areas such as classification of forest health conditions and forest degradation, among other classification challenges where inter-class overlap is pronounced. Similarly, the ensemble approaches presented here may perform better than conventional methods for species that are more clearly distinguishable and therefore should be studied further.

Conflict of interests

All authors declare no competing interests.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jag.2017.09.016.

References

Aberle, H., 2016. Hyperspectral remote sensing and field measurements for forest characteristics – a case study in the Hainich National Park, Central Germany. University of Goettingen (Ph.D. thesis).

Alvarez, S.G., Juaristi, C.M., Gutierrez, J.S., Garcia-Amorena, I., 2009. Taxonomic differences between *Pinus sylvestris* and *P. uncinata* revealed in the stomata and cuticle characters for use in the study of fossil material. Rev. Palaeobot. Palynol 155, 61–68.

Banfield, R., Hall, L., Bowyer, K., Kegelmeyer, W., 2007. A comparison of decision tree ensemble creation techniques. IEEE Trans. Pattern Anal. Mach. Intell. 29, 173–180.

Boratynska, K., Boratynski, A., 2007. Taxonomic differences among closely related pines *Pinus sylvestris*, *P. mugo*, *P. uncinata* and *P. uliginosa* as revealed in needle sclerenchyma cells. Flora 202, 555–569.

Briem, G., Benediktsson, J., Sveinsson, J., 2007. Multiple classifier applied to multisource remote sensing data. IEEE Trans. Geosci. Remote Sens. 29, 173–180.

Calle, L., Urrea, V., Boulesteix, A., Malats, N., 2011. AUC-RF: a new strategy for genomic profiling with Random Forest. Hum. Heredity 72, 121–132.

Clark, M., Roberts, D., Clark, D., 2005. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. Remote Sens. Environ. 96, 375–398.

Clinton, N., Holt, A., Scarborough, J., Yan, L., Gong, P., 2010. Accuracy assessment measures for object-based image segmentation goodness. Photogramm. Eng. Remote Sens. 76, 289–299.

Coleman, T., Gudapati, L., Derrington, J., 1990. Monitoring forest plantations using Landsat Thematic Mapper data. Remote Sens. Environ. 33, 211–221.

Deng, F., Li, S., Su, G., 2007. A methodology for fusion LIDAR and digital images. Proc. SPIE 6795, Second International Conference on Space Information Technology, 67952C.

Duque-Lazo, J., Navaro-Cerrillo, R., 2017. What to save, the host or the pest? The spatial distribution of xylophage insects within the Mediterranean oak woodlands SouthWestern Spain. Forest Ecol. Manag. 392, 90–104.

Duque-Lazo, J., van Gils, H., Groen, T., Navarro-Cerrillo, R., 2016. Transferability of species distribution models: the case of *Phytophthora cinnamomi* in Southwest Spain and Southwest Australia. Ecol. Model. 320, 62–70.

Elith, J., Phillips, S., Hastie, T., Dudík, M., chee, Y., Yates, C., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57.

Engler, R., Waser, L., Zimmermann, N., Schaub, M., Berdos, S., Ginzler, C., Psomas, A., 2013. Combining ensemble modeling and remote sensing for mapping individual tree species at high spatial resolution. Forest Ecol. Manag. 310, 64–73.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression: Models, Methods and Applications. Springer, Berlin, Heidelberg ISBN:978-3-642-34332-2.

Farjon, A., 2010. A Handbook of the World's Conifers. Koninklijke Brill, Leiden E-ISBN: 9789047430629.

Fassnacht, F., Latifi, H., Stereńczak, K., Modzelewska, A., Lefsky, M., Waser, L., Straub, C., Ghosh, A., 2016. Review of studies on tree species classification from remotely sensed data. Remote Sens. Environ. 186, 64–87.

Fauvart, N., Ali, A., Terral, J., Roiron, P., Blarquez, O., Carcaillet, C., 2012. Holocene upper tree-limits of *Pinus* section *sylvestris* in the western Alps as evidenced from travertine archives. Rev. Palaeobot. Palynol. 169, 96–102.

Gerard, F., North, P., 1997. Analyzing the effect of structural variability and canopy gaps on Forest BRDF using a geometric-optical model. Remote Sens. Environ. 62, 46–62.

González-Ferreiro, E., Diéguez-Aranda, U., Barreiro-Fernández, L., Buján, S., Barbosa, M., Suárez, J., Bye, I., Miranda, D., 2013. A mixed pixel- and region-based approach for using airborne laser scanning data for individual tree crown delineation in *Pinus radiata D. Don* plantations. Int. J. Remote Sens. 34, 7671–7690.

González-Ferreiro, E., Diéguez-Aranda, U., Crecente-Campo, F., Barreiro-Fernández, L., Miranda, D., Castedo-Dorado, F., 2014. Modelling canopy fuel variables for *Pinus radiata D. Don* in NW Spain with low-density LiDAR data. Int. J. Wildland Fire 23, 350–362.

Goodwin, N., Turner, R., Merton, R., 2005. Classifying eucalyptus forests with high spatial and spectral resolution imagery: an investigation of individual species and vegetation communities. Aust. J. Bot. 53, 337–345.

Gougeon, F., Leckie, D., 2006. The individual tree crown approach applied to ikonos images of a coniferous plantation area. Photogramm. Eng. Remote Sens. 72, 1287–1297.

Guillera-Arroita, G., Lahoz-Monfort, J., Elith, J., 2014. Maxent is not a presence–absence method: a comment on Thibuad et al. Methods Ecol. Evol. 5, 1192–1197.

Hamraz, H., Contreras, M., Zhang, J., 2017. Forest understory trees can be segmented accurately within sufficiently dense airborne laser scanning point clouds. Sci. Rep. 7 (1), 2045–2322.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, 2nd ed. Springer ISBN: 978-0-387-84857-0.

Ke, Y., Quackenbush, L., Im, J., 2010. Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. Remote Sens. Environ. 114, 1141–1154.

Kim, M., McGaughey, R., Andersen, H., Schreuder, G., 2009. Tree species differentiation using intensity data derived from leaf-on and leaf-off airborne laser scanner data. Remote Sens. Environ. 113, 1575–1586.

Ko, C., Sohn, G., Remmel, T., Miller, J., 2014. Hybrid ensemble classification of tree genera using airborne LIDAR data. Remote Sens ISSN 2072-4292.

Korpela, I., Ørka, H., Hyyppä, J., Heikkinen, V., Tokola, T., 2010. Range and AGC normalization in airborne discrete-return LiDAR intensity data for forest canopies. ISPRS J. Photogramm. Remote Sens. 65, 369–379.

Leckie, D., Gougeon, F., Hill, D., Quinn, R., Armstrong, L., Shreenan, R., 2003. Combined high-density lidar and multispectral imagery for individual tree crown analysis. Can. J. Remote Sens. 29, 633–649.

Marmion, M., Parviainen, M., Luoto, M., Heikkinen, R., Thuiller, W., 2009. Evaluation of consensus methods in predictive species distribution modelling. Divers. Distrib. 15, 59–69.

McGaughey, R., 2014. FUSION/LDV: software for lidar data analysis and visualization, v. 3.42. USDA Forest Service, Pacific Northwest Research Station, Seattle, WA.

Mo, D., Fuchs, H., Fehrmann, L., Yang, H., Lu, Y., Kleinn, C., 2015. Local parameter estimation of topographic normalization for forest type classification. IEEE Geosci. Remote Sens. Lett. 12, 1998–2002.

Ørka, H., Næsset, E., Bollandsås, O., 2009. Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. Remote Sens. Environ. 113, 1163–1174.

Peerbhay, K., Mutanga, O., Ismail, R., 2014. Does simultaneous variable selection and dimension reduction improve the classification of Pinus forest species? J. Appl. Remote Sens. 8, 085194-L.

Ross, J., Kellomäki, S., Oler-Blom, P., Ross, V., Vilikainen, L., 1986. Architecture of scots pine crown: phytometrical characteristics of needles and shoots. Silva Fennica 20, 91–105.

Suratno, A., Seielstad, C., Queen, L., 2009. Tree species identification in mixed coniferous forest using airborne laser scanning. ISPRS J. Photogramm. Remote Sens. 64,

683–693.

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., Blair, J., 2011. Mapping biomass and stress in the sierra nevada using lidar and hyperspectral data fusion. Remote Sens. Environ. 115, 2917–2930.

Thuiller, W., Georges, D., Engler, R., Breiner, F., 2016. biomod2: Ensemble Platform for Species Distribution Modeling. R package version 3.3-7. https://CRAN.R-project.org/package=biomod2.

Vauhkonen, J., Maltamo, M., McRoberts, R., Næsset, E., 2014. Introduction to forestry applications of airborne laser scanning. Springer, Dord ISSN: 1568-1319.

Wulder, M., Niemann, K., Goodenough, D., 2000. Local maximum filtering for the extraction of tree locations and basal area from high spatial resolution imagery. Remote Sens. Environ. 73, 103–114.

Youngentob, K., Roberts, D., Held, A., Dennison, P., Jia, X., Lindenmayer, D., 2011. Mapping two eucalyptus subgenera using multiple end member spectral mixture analysis and continuum-removed imaging spectrometry data. Remote Sens. Environ. 115, 1115–1128.

Zhang, Z., Lui, X., 2013. Support vector machines for tree species identification using LiDAR-derived structure and intensity variables. Geocarto Int. 28, 364–378.

Scale effects in broadleaved tree reflectance at leaf and crown levels

Collins B. Kukunda[a], Henning Aberle[a,d], Peter P.J. Roosjen[b], Christoph Kleinn[a], Steen Magnussen[c], Hans Fuchs[a]

a. University of Goettingen, Forest Inventory and Remote Sensing, Faculty of Forest Sciences, Büsgenweg 5, D-37077 Göttingen, Germany.

b. Laboratory of Geoinformation Science and Remote sensing, Wageningen University & Research, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands.

c. Pacific Forestry Centre, Natural Resources Canada, 506 Burnside Road West, V8Z 1M5, Victoria, British Columbia, Canada.

d. Satellite-Based Crisis and Spatial Information Service, Federal Agency for Cartography and Geodesy, Richard-Strauss-Allee 11, 60598 Frankfurt am Main, Germany

1.1 Abstract

Imaging and non-imaging spectroscopy are offering new insights and possibilities in the remote sensing of forests. The link between new hyperspectral airborne and satellite sensors and the information gathered on the ground is crucial as field data is widely used for sensor calibration and validation or for the derivation and up-scaling of biophysical parameters such as nitrogen and chlorophyll content of foliage. In that context, we are often facing several scaling issues arising from differences in spatial and spectral scales of observation. This study deals with the spectral reflectance of trees at varying spatial scales of observation. The goal is to highlight effects from changing the scale of observation in analysis of broadleaved tree spectral reflectance obtained from leaves and crowns. The study was carried out with five tree species in an old-growth broadleaved forest in the Hainich area, Germany. We used a set of simultaneously acquired in-situ and remotely sensed reflectance records at three different scales of leaves and crowns (i.e. leaf, crown-part, entire crown) incorporating field spectroradiometer as well as hyperspectral airborne imagery. At each scale of observation, generalized additive models approximated a mean spectral response curve across observations. A bootstrapping procedure was used to calculate confidence intervals around modeled spectral profiles highlighting differences brought about by changing the scale of observation. Jefferies-Matusita distances were used to show the effect of the scale of observation on separability of species spectral reflectance obtained from leaves and crowns. Finally, to predict contributions of structure and foliar chemical composition to observed spectral reflectance at the three scales of observation, the observed reflectance were inverted using PROSPECT-5B and PROSAIL and RTM sensitivity analyses conducted with reference to the LOPEX database. We show that there can be significant differences in tree spectral reflectance observed from leaves and crowns. Similarly, separability among species spectral reflectance diminished across the scale of observation from leaves to crowns. Finally, we show that a combination of leaf chemical and structural properties as well as crown structure dominated spectral characteristics observed at different scales of observation and wavelength regions. This study's findings are important, since our ability to discriminate species from hyperspectral data has been shown to depend on the scale of observation. In addition, this study's results inform sensor calibration and modeling approaches aimed at retrieval or mapping of leaf to canopy physical and chemical properties from hyperspectral data while at the same time highlight the utility of spectral libraries.

1.2 Introduction

1 Forests are dynamic and complex ecosystems that take up important social, economic and
2 ecological functions. Knowledge about the tree species composition and their condition serves
3 local and regional planning and conservation measures (Lui and Coomes, 2015; O'Connor et al.,
4 2015; Turner et al., 2015). In Germany, forests cover about one third of the land (BMEL, 2015),
5 which is comparable to the global average (UNEP et al., 2009). Since there is a development
6 towards mixed stands within Germany, forest management has become more complex with an

increased need for detailed information on tree species identities. For observing large forested areas, passive optical airborne and satellite systems are commonly used. In that context, imaging and non-imaging spectroscopy offer new insights and new facility in remote sensing of forests and tree species classification (Yang et al., 2016; Asner et al., 2015; Feilhauer et al., 2015; Féret and Asner, 2013). The new hyperspectral sensors commonly cover the electromagnetic spectrum between visible violet and shortwave infrared.

The mapping of tree species and tree physical and chemical properties are contemporary and ongoing issues in hyperspectral remote sensing of forests (Clark and Roberts, 2012; Dalponte et al., 2013; Féret and Asner, 2013). Some studies found differences in wavelength regions discriminating species (Clark and Roberts, 2012; Fassnacht et al., 2016) yet findings depend often on local circumstances. Further appropriate measures to amplify spectral features are vegetation indices (VI) derived from remotely sensed reflectance imagery (Huete et al., 1997; Schlerf et al., 2005). Hyperspectral sensors with their plethora of bands offer numerous possible VI by using ratios, normalized differences or further equations incorporating a few channels (Bannari et al., 1995; Prospere et al., 2014). Similarly, the capability of hyperspectral sensors to estimate chemical and structural variables such as leaf area and chlorophyll content is well documented, for example by Asner et al. (2015); Brantley et al. (2011); Gitelson et al. (2003) using empirical approaches, and by Ferreira et al. (2013) and Ali et al. (2016b) with inversion of physical models. However, similar to the detection of spectral features, like the red edge position at the reflectance curve inclination point (at 700 - 720 nm), the precision of VI values, leaf chemical, and structural properties are affected by choices of spectral, directional, and spatial scales (Asner, 1998; Clark et al., 2005; Malenovský et al., 2007; Roosjen et al., 2018).

Scale effects are recognized in spectroscopy (Malenovský et al., 2007; Roosjen et al., 2018). However, the effect of scale-related variation in spectral reflectance of tree species is less studied in the context of species differentiation (Cho et al., 2008). Similarly, the fundamental understanding of potential drivers of observed reflectance at different scales is understudied (Xiao et al., 2014). Overall, there are still unresolved scale issues regarding the physiological and spectral traits of trees when going from leaves to canopies and stands (Kumar et al., 2010; Malenovský et al., 2007; Schaepman et al., 2009). Ground data, acquired at different scales of observation, are still needed (Homolová et al., 2013). In-situ observations are used as reference for remotely sensed data (Groeneveld et al., 2006), for sensor calibration (Pfitzner et al., 2006; Smith and Milton, 1999), for modeling approaches (Schneider et al., 2014), and for reference in spectral libraries (Kokaly et al., 2017); it is important that we find linkages to hyperspectral imagery. We expect this study to be a relevant contribution to forest monitoring with hyperspectral remote sensing by furthering research into the scale-dependent linkages between field reference data from sample trees and hyperspectral image data.

The linkages between leaf to crown spectral reflectance are commonly modelled by radiative transfer models (RTMs) in homogeneous systems. RTMs combine chemical and physical attributes of the system with other environmental and physical characteristics such as soil reflectance, sensor viewing geometry, illumination and atmospheric conditions (Jacquemoud and Baret, 1990; Féret et al., 2008; Jacquemoud et al., 2009; Berger et al., 2018). However,

when dealing with non-imaging spectroscopy in forests, the conditions are different, and mostly more complex due to the size of trees and their crown structures (Gara et al., 2018). Some research is already done on radiative transfer within heterogeneous forest stands (Widlowski et al., 2015) leading to improved understanding of challenges presented by canopy complexity when scaling spectral reflectance between leaves and canopies. However, complex RTMs often require a large amount of in-situ data for calibration, that is often not readily available. For this reason, this study tested the utility of a leaf to canopy RTM (the coupled PROSPECT-5B and SAILH models: PROSAIL). The PROSAIL model has been successfully applied in different settings in the retrieval of both physical and chemical attributes and simulation of spectral reflectance (Berger et al., 2018). The model has shown the ability to consistently link spectral variation associated with leaf chemical properties (mainly chlorophyll, water, and dry matter contents) to the directional variation linked to the canopy structure (primarily, leaf area index, leaf angle distribution, and relative leaf size), while accounting for soil/vegetation contrast and viewing geometries. In the model, emissivity, reflectance, radiance and brightness temperature are modelled as a function of LAI, which makes model parameterization rather straight forward (Jacquemoud et al., 2009; Berger et al., 2018). However, the model's application for a comparison of spectral reflectance of tree species at different spatial scales of observation and in heterogeneous forest environments is generally limited.

The study's goal is to improve our understanding of the potential in forest monitoring of hyperspectral remote sensing data at different scales of observation. Specifically, this study aims to highlight the influence of the scale of observation in analysis of spectral reflectance of tree foliage and individual tree crowns. The focus is on five dominant species in the study area including, ash (*Fraxinus excelsior*), European beech (*Fagus sylvatica*), European hornbeam (*Carpinus betulus*), sessile oak (*Quercus petraea*) and small-leaved lime (*Tilia cordata*). The study is guided by the following technical objectives;

1. To identify differences in leaf and crown spectral reflectance of five broadleaved tree species.

2. To highlight the impact of the scale of observation on separability among species spectral reflectance.

3. To predict contributions of structure and foliar chemical contents to observed leaf and crown spectral reflectance using a radiative transfer modeling approach.

4. To identify regions of the electromagnetic spectrum suited for estimation of physical and chemical properties of trees as well as to separate tree species in optical imagery.

1.3 Material and Methods

1.3.1 Study area

The data was collected in a mixed broadleaved forest (Fig. 1.1), within the Hainich National Park (HNP), Thuringia, Germany (Latlong: 51.08° N, 10.51° E). Elevation is about 340 m a.s.l., the terrain is generally flat, precipitation averages 750 mm annually, mean temperature is 7.5°C. With a total area of 16,500 ha the Hainich is the largest continuous deciduous forest of Germany and is part of UNESCO's World Heritage "Old beech forests in Central Europe". About 7,500 ha were declared as National Park in 1997, of which about 5,000 ha has not been managed for 50 years (Knohl et al., 2003; Mund et al., 2010). The forest stands are dominated by European beech, followed by ash, maples (*Acer sp.*), oaks (*Quercus sp.*), European hornbeam and others. Figure 1.1 highlights the data sets used in this study.
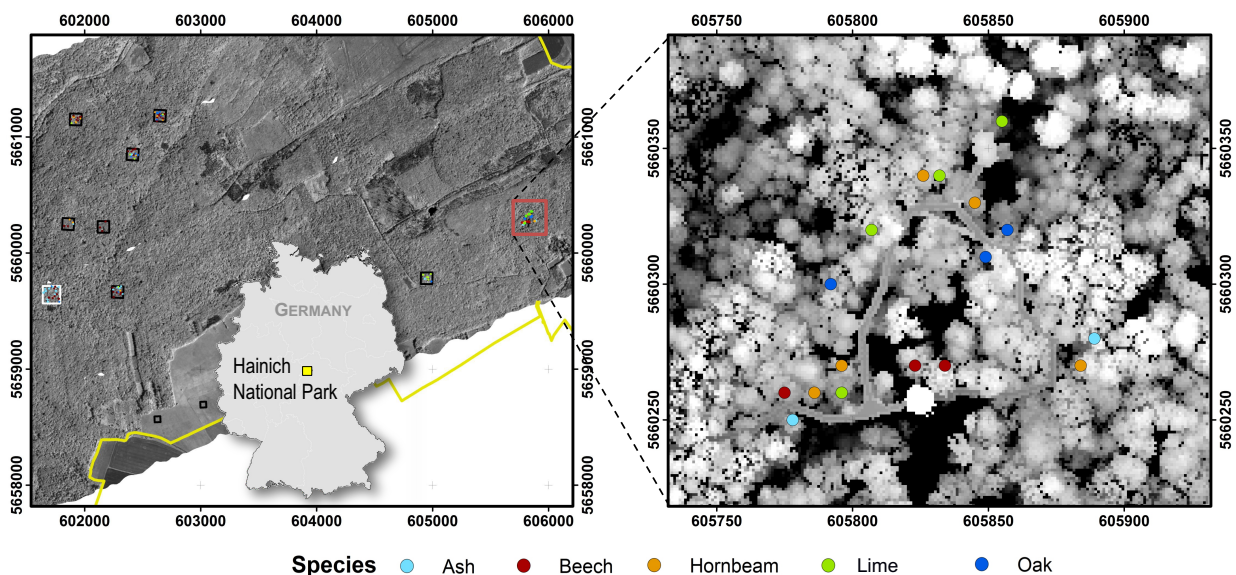
Figure 1.1: **Left**: Location of the study plots in the Hainich National Park (National Park boundary shown in a yellow line). Black boxes are combined observational and experimental plots in a huge research project on biodiversity functioning, the Biodiversity Exploratories project (www.biodiversity-exploratories.de), the white box marks the plot area from which additional data from Aberle (2016) come, and the red box highlights the data collected at the canopy walk tower. Field identified tree species are shown by color (see online version). Manually delineated individual tree crowns are shown in the respective colors of tree species. The background image shows the AISA red band at 654 nm in grayscale. **Right**: Positions of the sampled trees at the Hainich National Park canopy walk overlain on a canopy height model derived from open LiDAR data (Landesamt für Vermessung und Geoinformation, Thüringen, www.geoportal-th.de/de-de, Version 2.0, License: www.govdata.de/dl-de/by-2-0).

1.3.2 Field inventory

Inventory data on individual trees were aggregated from three data sets. The data sets include seven 1.0 ha experimental plots (EPs) from the Biodiversity Exploratories project (www.biodiversity-exploratories.de), one 2.25 ha plot from Aberle (2016, Figure 1.1), and one

collected along the canopy walk of the HNP by Aberle (2016, Figure 1.1, right). Further details on the data sets including plot selection procedures can be found in Fischer et al. (2010) and Aberle (2016). The seven EPs are a subset of 50 plots selected by Fischer et al. (2010) that covered the hyperspectral imagery utilized in this study (see section 1.3.4).

Tree measurements focused on accurate species identification and geo-location of stems to precisely link field and remotely sensed observations. Fieldwork on the EPs was done during the winters of 2014/15 and 2015/16 by the Department of Silviculture and Forest Ecology of the Temperate Zones, University of Goettingen. All trees with diameter at breast height (DBH) $\geq$ 7 cm were measured with a diameter tape. DBH, species identities and stem positions were recorded into a FieldMap$^{©}$ system (www.fieldmap.cz) based on a procedure described by Hédl et al. (2009). The approach enabled geo-location of tree positions with very high accuracy (modal deviations $\approx$ 10 cm). Trees on the 2.25 ha plot and along the canopy walk were identified and geo-located according to Aberle (2016). Trees with a DBH $\geq$ 15 cm were recorded under leaf-off conditions in the first months of 2012. Tree diameters were measured with caliper and diameter tape, where larger DBH were measured crosswise. The spatial position was acquired using compass (SUUNTO) and an ultrasonic inclinometer (HAGLÖF Inc. Vertex IV). The corners of a flux tower within the plot served as reference points. Coordinates were determined by longtime GPS logging (GARMIN GPSmap 60Cx) on the top of the tower above the canopy. On the canopy walk, measured trees were identified by GPS recordings (LEICA GS20 GPS) and by recognizing the structure of the canopy walk. The location of all trees were subsequently validated on the canopy height model (CHM) derived from aerial LiDAR data (Figure 1.1, right).

1.3.3 Acquisition and pre-processing of in-situ spectral data

The in-situ reflectance was collected from 17 individual trees representing five species, on 24.07.2012, about solar noon and in sunny conditions (Figure 1.1, right). The data was collected along the canopy walkway of the HNP using an ASD FieldSpec 3$^{©}$ high resolution field spectroradiometer (ANALYTICAL SPECTRAL DEVICES INC.). The ASD FieldSpec 3$^{©}$ consists of three sensors in the visible-near infrared (VIS-NIR, 350-1000 nm), near infrared-shortwave infrared 1 (NIR-SWIR1, 1000-1800 nm) and shortwave infrared 2 (SWIR2, 1800-2500 nm) domains of the spectrum. Leaf scale data was acquired with the ASD sensor modified with a leaf clip, plant probe and an internal halogen lamp as the light source. Each measurement contained a minimum of four readings to reduce measurement errors. In-situ reflectance measurements were collected for tree crown parts using a fore optic and this time with the sun as the light source. The instantaneous field of view of 25° resulted in footprints of average diameter about 0.5 m. The distance between ASD sensor and crown was $\approx$ 1 m, only in a few cases the spacing was up to $\approx$ 4 m. Water vapor in the air absorbs radiation at specific wavelengths causing strong noise or no signal reaching the sensor. The main water absorption bands at about 1400 nm (1351-1449) and 1850 nm (1801-1949) were cut out from the reflectance data. At the leaf scale, this noise is not apparent since there is no space between object and sensor.

135 However, in this dataset we deleted signals below 450 nm and above 2450 nm that are most
136 influenced by noise. As the ASD FieldSpec 3$^{©}$ has three sensors, the resulting reflectance curves
137 were characterized by steps or abrupt changes at the points of transition between sensors i.e.
138 at 1000 and 1800 nm. The abrupt changes need to be corrected to get a continuous signal.
139 Here, an additive jump correction for each of the corresponding wavelengths was applied as
140 implemented in the Spectral Analysis and Management System (SAMS$^{©}$) software (Rueda and
141 Wrona, 2003). Additionally, a Savitzky-Golay filter (Savitzky and Golay, 1964; Schafer, 2011)
142 was employed for smoothing the response patterns using a second order polynomial function
143 including the five previous and five subsequent bands.

1.3.4 Acquisition and pre-processing of imaging spectroscopy data

145 The hyperspectral imagery were collected about solar noon - i.e. coincident to in-situ spectral
146 data collection and using a combined airborne sensor system with AISA EAGLE$^{©}$ and HAWK$^{©}$
147 sensors (SPECTRAL IMAGING LTD, Finland). The system covers a spectral range of 400-2500
148 nm split into 368 bands resulting in spectral sampling of about 4 - 6 nm. Ground sampling
149 distance of the imagery was 2 m. Eight image stripes were acquired. Flight planning was done
150 in consultation with the Department of Computational Landscape Ecology of the Helmholtz
151 Centre for Environmental Research (UFZ), Leipzig. The following pre-processing steps were
152 performed by UFZ on delivery of the data set: (1) geometric and radiometric correction using the
153 ENVI module CaliGeo (SPECIM), (2) correction of sensor-dependent striping effects using the
154 ROME method (Reduction of Miscalibration Effects, Helmholtz Center Potsdam, GFZ German
155 Research Center for Geosciences), (3) atmospheric correction using ATCOR 4 (Atmospheric
156 and Topographic Corrections, RESE APPLICATIONS SCHLÄPFER), (4) spectral smoothing
157 by a Savitzky-Golay filter (Savitzky and Golay, 1964) incorporating six neighboring bands, and
158 (5) interpolation across water-absorption bands (about 1400 nm and 1850 nm) yielding a
159 continuous reflectance curve. After these steps, each image was manually referenced with the
160 help of ArcMap base maps (ArcGIS Desktop v.10.3, ESRI Inc.) and a step-by-step mosaicking
161 performed in ENVI (v.4.8, ITTVIS). Pixels were resampled to 0.5 m x 0.5 m using a nearest
162 neighbor approach to achieve a spatially more accurate value extraction. This allowed for
163 quasi-subpixel crown delineation while maintaining the original pixel values.

1.3.5 Crown delineation

165 We manually delineated 146 crowns across five species from the experimental plots and the
166 plot from Aberle (2016, Figure 1.1) to isolate pixel-wise spectral response patterns of individual
167 trees from the hyperspectral imagery (Table 1.1). The trees whose crowns were delineated
168 were selected from field plot data subjectively, guided by precision in geometric matching of
169 tree positions and crowns as seen from plot maps (section 1.3.2), on a LiDAR CHM (Figure
170 1.1) and in the imagery (section 1.3.4). The LiDAR CHM was used as ancillary data during
171 delineation. Crown delineation was done maintaining a single map scale across crowns. Field
172 observations (section 1.3.2) were used to assign tree species to delineated crowns. Table 1.1

173 shows the number of delineated crowns by species.

1.3.6 Scales of observation

175 All data sets described above were collected at three spatial scales of observation and two
176 spectral scales of observation. Figure 1.2 shows schematics of example reflectance collected at in-
177 situ leaf ($\mathrm{Leaf_{ASD}}$), crown-part or in-situ crown ($\mathrm{Crown_{ASD}}$) and airborne crown ($\mathrm{Crown_{AISA}}$)
178 scales of observation. Table 1.1 summarizes the number of trees and readings/total repeated
179 measurements considered in analyses at the three scales of observation. In total, 17 in-situ trees
180 and 146 tree crowns were observed. The spectral resolution of in-situ and airborne reflectance
181 differed. At $\mathrm{Leaf_{ASD}}$ and $\mathrm{Crown_{ASD}}$ scales, 1 nm steps were sampled across the range of
182 450-2450 nm yielding in total 2001 bands. However, at $\mathrm{Crown_{ASD}}$ scale, at total of 1753
183 bands remained after pre-processing (section 1.3.3). At $\mathrm{Crown_{AISA}}$ scale, spectral sampling of
184 about 4 - 6 nm yielded a total 368 bands but only 330 bands were used in our analyses after
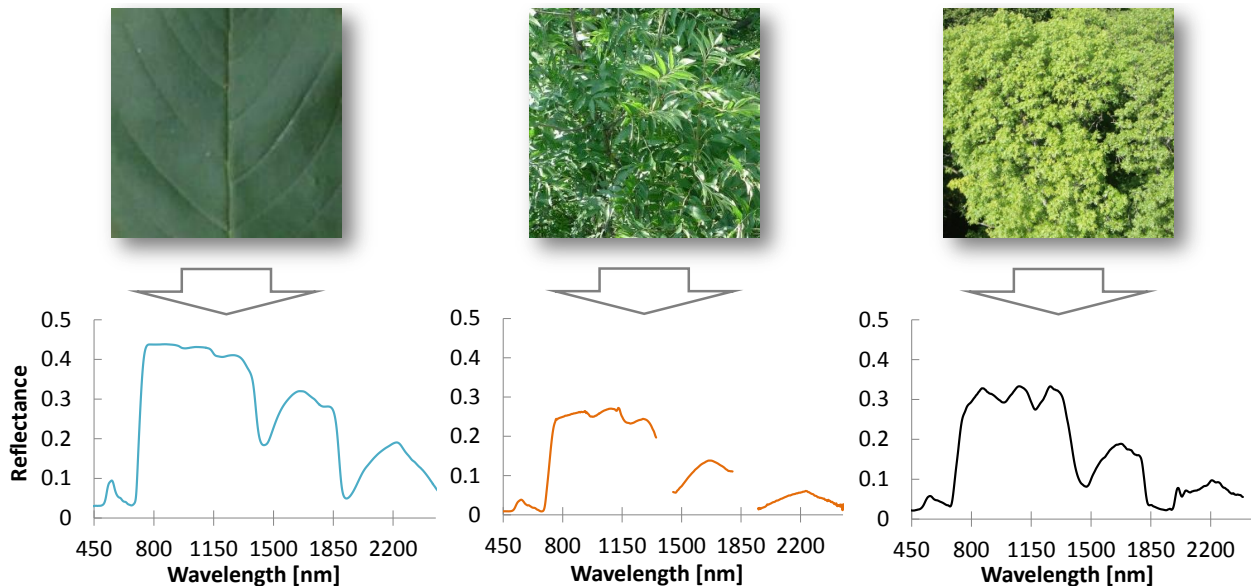185 eliminating noisy bands at edges of the spectral range.

Figure 1.2: Schematic overview of the three spatial scales of data acquisition: in-situ leaf/$\mathrm{Leaf_{ASD}}$ (left), in-situ crown/$\mathrm{Crown_{ASD}}$ (middle) and airborne crown/$\mathrm{Crown_{AISA}}$ (right). Means of observed reflectance are shown. At the in-situ crown scale, sensor noise induced by water vapor is masked out.

1.3.7 Estimating scale-related differences among leaf and crown spectral reflectance

187 We aimed to accurately model the non-linear patterns of spectral response curves and corre-
188 sponding across-scan variances at each scale of observation. To do this, we fitted Generalized
189 Additive Models (GAMs) with P-spline basis functions (Eilers et al., 2015; Wood, 2011) and
190 estimated mean reflectance curves across spectral measurements (Table 1.1) and used boot-
191 strapping to estimate confidence bands around the mean reflectance curves (James et al.,
192 2013, Pages 187-190). The GAMs had, in addition to their capability to model local non-linear

Table 1.1: Number of trees and readings (or total repeated measurements). Leaf_{ASD} refers to in-situ leaf, $\text{Crown}_{\text{ASD}}$ refers to in-situ crown and $\text{Crown}_{\text{AISA}}$ refers to airborne crown scales of observation.

	Trees		Readings	
Species	In-situ	$\text{Crown}_{\text{AISA}}$	Leaf_{ASD}	$\text{Crown}_{\text{ASD}}$
Ash	2	22	206	12
Beech	3	36	111	17
Hornbeam	5	12	66	32
Oak	3	45	51	12
Lime	4	31	48	24
Totals	17	146	482	97

patterns, the advantage of interpolating across masked water bands of in-situ crown reflectance spectra (Figure 1.2, middle) generating continuous reflectance curves, while bootstrapping helped reduce autocorrelation among reflectance observations at each wavelength. Notice that at $\text{Crown}_{\text{AISA}}$ scale, independent observations were means of all 0.5 m pixels delineated within a crown, while at Leaf_{ASD} and $\text{Crown}_{\text{ASD}}$ scales the repeated observations (i.e. readings in Table 1.1) were treated as independent observations. Bootstrapping was done by iterating 500 times over the model fitting process while randomly selecting with replacement a sample of size $n = 20$ reflectance curves to fit an average reflectance curve per scale of observation. We calculated the inter-quantile range between 95th and 5th percentiles of the distribution of mean reflectance at each wavelength and approximated standard error around the mean reflectance at each wavelength by dividing the inter-quantile range by 3.29 standard deviation units of the normal distribution. Non-overlapping confidence bands among leaf and crown spectra per species indicated significant differences among the spectra.

1.3.8 Calculating species separability among leaf and crown spectral reflectance

Pair-wise Jeffries-Matusita (J-M) distances (Schmidt and Skidmore, 2003; Vaiphasa et al., 2005; Adam and Mutanga, 2009) were computed to assess spectral differences among tree species and determine spectral regions most appropriate for species discrimination in both leaf and crown spectral reflectance. The pairwise distances were computed between mean reflectance per species and for each wavelength assuming a multivariate normal distribution (Richards, 2013; Schmidt and Skidmore, 2003), as shown in Equation 1.1;

$$\text{JM}_{\text{ij}} = 2(1 - e^{-B}) \tag{1.1}$$

where i represents the focal species, j the second species in the pair and B is the Bhattacharyya distance, computed as a square of normalized distance between the class means (Richards, 2013), as shown in Equation 1.2;

$$B = \frac{1}{8}(m_{\text{i}} - m_{\text{j}})^{\text{T}} \left(\frac{\sigma_{\text{i}} + \sigma_{\text{j}}}{2} \right)^{-1} (m_{\text{i}} - m_{\text{j}}) + \frac{1}{2} \ln \left(\frac{|\frac{(\sigma_{\text{i}}+\sigma_{\text{j}})}{2}|}{\sqrt{|\sigma_{\text{i}}| \, |\sigma_{\text{j}}|}} \right) \tag{1.2}$$

where m is the mean reflectance of species i and j, σ_i and σ_j are the variances across observations per species i and j. $|\sigma_i|$ and $|\sigma_j|$ is the determinant of the matrix of observations of species i and j. T refers to a transpose of the matrix of differences across species means. JM values will vary between zero and a maximum of 2.0 that is reached with perfect separation. Hence, a JM distance of 2.0 implies that the between-species difference is larger than the within-species difference (Schmidt and Skidmore, 2003). This procedure was performed separately for each scale of observation across pair combinations of the five studied species resulting in a total of 48 comparisons.

1.3.9 Predicting contributions of structure and foliar chemical contents to observed leaf and crown spectral reflectance

To predict contributions of structure and foliar chemical contents to observed spectral reflectance from leaves and crowns, and to identify regions of the electromagnetic spectrum potentially useful to predict physical and chemical properties of leaves, crown-parts, and entire crowns, we employed PROSAIL including PROSPECT-5B and 4SAIL RTMs (Jacquemoud et al., 2009). Figure 1.3 gives an overview of the approach.
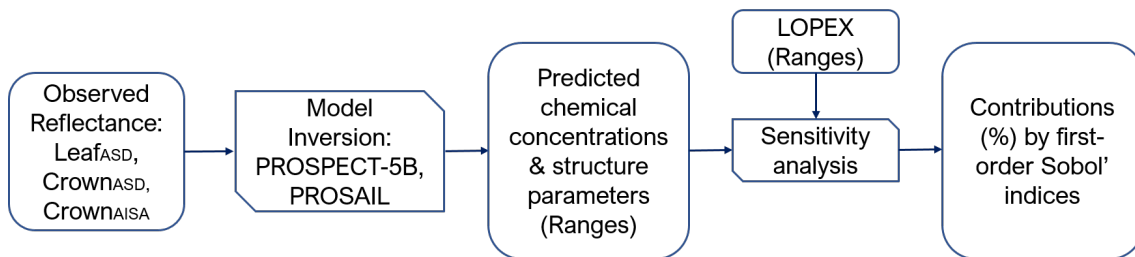

Figure 1.3: Overview of the procedure used to predict drivers of leaf and crown spectral reflectance from PROSPECT-5B and PROSAIL models.

First, we inverted $\mathrm{Leaf}_{\mathrm{ASD}}$ reflectance using PROSPECT-5B in default calibration and without transmittance data, and retrieved predictions of chemical and physical properties for each observed leaf spectrum. The default calibration of PROPECT-5B is described in Féret et al. (2008). We used a constrained Powell's search method during optimization and minimized the Root Mean Squared Error of Prediction (RMSEP) between observed and simulated spectral reflectance to find the best fit of parameters (Féret et al., 2008). The retrieved properties included, leaf structure parameter (N), Chlorophyll a+b concentration (Cab), Carotenoid concentration (Car), Leaf brown pigments (Cbrown), Equivalent water thickness (Cw) and Leaf dry matter content (Cm).

Second, we inverted $\mathrm{Crown}_{\mathrm{ASD}}$ and $\mathrm{Crown}_{\mathrm{AISA}}$ data using PROSAIL with a look-up table (LUT) approach in order to obtain predictions of leaf properties and leaf area index (LAI). The LUT varied parameters between the ranges presented in Table 1.2, left. At this step, we fixed solar zenith to 58°, observer zenith to 10°, and relative azimuth to 0°according to the time as well as sun-object-sensor geometry at acquisition of $\mathrm{Crown}_{\mathrm{AISA}}$ data. PROSAIL was inverted as done in Roosjen et al. (2018) by minimizing the RMSE between measured and modelled

reflectance in a procedure that iterated over ranges of parameters in Table 1.2 using the *Optim* function in R (R Core Team, 2019). The values of parameters at convergence were taken as the true parameters. Each collected spectrum (Table 1.1) was inverted independently.

Last, we conducted sensitivity analyses of PROSPECT-5B and PROSAIL to predict contributions of structure and foliar chemical contents to observed spectral reflectance at the different scales of observation. To do this, we monitored variation in simulated reflectance in response to variation in input physical and chemical parameters and their interactions. The sensitivity analyses were first limited to ranges of physical and chemical properties as retrieved from model inversion (Table 1.3) and later for comparison purposes, the parameters were restricted to the ranges of physical and chemical properties in the Leaf Optical Properties Experiment (LOPEX) (Hosgood et al., 1993; Xiao et al., 2014, and Table 1.2 right). The LOPEX database was selected because it included all studied genera except European hornbeam. Sensitivity analyses were performed using a Monte Carlo approach (Sobol' et al., 2007; Saltelli et al., 2010) whereby, first-order and total-order Sobol' indices were calculated for each band and at each scale of observation. The first-order indices conveyed the proportion of variance attributable to variation in individual model parameters without considering their interactions with other optimized parameters in the model, while total-order indices calculated the proportion of variance attributable to variation in optimized parameters while including their interactions. See Xiao et al. (2014) for a step-by-step account of variance partitioning using Sobol' indices. The indices were computed for each parameter, band and scale of observation as measure of contribution/importance in the observed spectral reflectance. Sensitivity analyses were done using *multisensi* and *sensitivity* packages in R (Iooss et al., 2020; Bidot et al., 2018).

Table 1.2: **Left:** Ranges of parameters optimized in PROSAIL to invert $\mathrm{Crown_{ASD}}$ and $\mathrm{Crown_{AISA}}$ reflectance using a look-up table approach. **Right:** Ranges of parameters optimized in PROSAIL to conduct sensitivity analyses with respect to the LOPEX database. The used abbreviations in full: Leaf structure parameter (N), Chlorophyll a+b concentration (Cab), Carotenoid concentration (Car), Leaf brown pigments (Cbrown), Equivalent water thickness (Cw) and Leaf dry matter content (Cm) and Leaf area index (LAI).

Parameter	Min	Max	Initial	Min	Max
N	0.8	2.5	1.65	1	2
Cab (mg. cm^{-2})	0	80	40	10	120
Car (mg. cm^{-2})	0.0001	20	10	5	30
Cbrown (mg. cm^{-2})	0	1	0.5	0	0
Cw (g. cm^{-2})	0.0001	1	0.05	0.005	0.04
Cm (g. cm^{-2})	0.0001	10	0.5	0.002	0.014
LAI	0	10	3	0	10

1.4 Results

1.4.1 Comparison of leaf and crown spectral reflectance

The observed and modelled leaf and crown spectral reflectance are presented in Figure 1.4. Reflectance at all scales of observation exhibited typical spectral reflectance patterns of vegetation but with statistical differences in magnitudes. The visible (VIS) part of the electromagnetic spectrum (450 - 700 nm) was characterized by the highest reflectance absorption features before 500 nm and between 653 - 684 nm. Absorption features of similar magnitude as observed in VIS were as well observed in the shortwave infrared region - SWIR (1450 - 2450 nm) between 1450 - 1500 nm and between 1892 - 2026 nm. In addition, there were relatively shallow reflectance absorption features at $\approx$ 1000 nm, $\approx$1250 nm in the near infrared region - NIR (701 - 1450 nm) that increased in magnitude from Leaf_{ASD} to $\text{Crown}_{\text{AISA}}$ scales. The highest reflectance across the spectrum were observed in NIR, followed by the SWIR 1 (1451 - 1850 nm).

Leaf scale reflectance were on average higher than $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ reflectance. However, individual spectral curves at $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ scales were as high as leaf spectral curves in the near infrared part of the spectrum. In addition, leaf scale reflectance had the lowest variation among observations. The highest variation among reflectance was observed in $\text{Crown}_{\text{ASD}}$ reflectance. There were significant differences in mean predicted reflectance across the entire range of the observed spectrum between Leaf_{ASD} and both $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ data (Figure 1.4). $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ reflectance were similar across the spectrum expect around 800 nm (774 - 950 nm) where the mean reflectance of $\text{Crown}_{\text{AISA}}$ data was higher than the mean reflectance of $\text{Crown}_{\text{ASD}}$ data. Spectral differences between $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ mean reflectance from 1801 nm to 1949 nm should be ignored since these bands were masked out at $\text{Crown}_{\text{ASD}}$ scale. Scale-related differences in mean reflectance were similar across studied species (Figure 1.8 in supplementary material).

1.4.2 Separability among species leaf and crown spectral reflectance

Figure 1.5 shows separability among pairs of species spectral reflectance at leaf and crown scales. Spectral reflectance at leaf scale exhibited the highest J-M values compared to the two crown scales. Separability between species pairs was highest in leaf spectral reflectance and lowest in $\text{Crown}_{\text{AISA}}$ reflectance. Different regions of the spectrum (VIS, red-edge; 700 - 720 nm, NIR, SWIR) were important to distinguish at least one species pair, however, separability generally increased within SWIR, followed by the VIS and red-edge regions at Leaf_{ASD} and $\text{Crown}_{\text{AISA}}$ scales. In contrast, VIS and red-edge regions exhibited the highest separability while NIR and SWIR regions exhibited generally similar J-M values at $\text{Crown}_{\text{ASD}}$ scale. There were local peaks in separability in VIS (at 600 nm), red-edge, and SWIR (between 1450 - 1500 nm and 1950-2100 nm) regions at all studied scales, but the peaks were more pronounced at Leaf_{ASD} and $\text{Crown}_{\text{AISA}}$ scales.

Separability among species pairs depended on the scale of observation and was generally more consistent at Leaf_{ASD} and $\text{Crown}_{\text{AISA}}$ scales than the $\text{Crown}_{\text{ASD}}$ scale. For example,
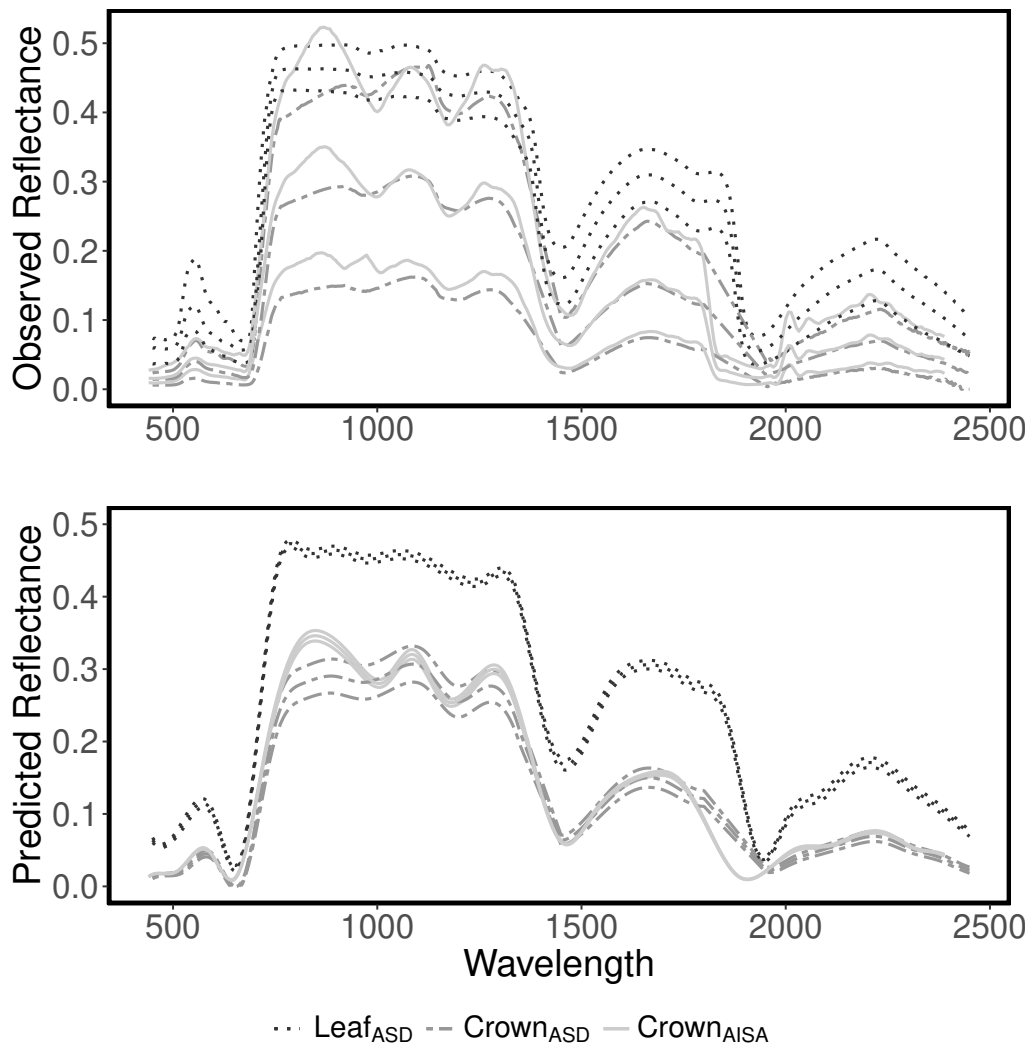
Figure 1.4: **Top:** Means, 5th and 95th percentiles of observed reflectance at Leaf$_{ASD}$, Crown$_{ASD}$ and Crown$_{AISA}$ scales. **Bottom:** Mean modelled reflectance and the associated bootstrapping standard errors at Leaf$_{ASD}$, Crown$_{ASD}$ and Crown$_{AISA}$ scales.

306 ash was separable from beech, hornbeam and lime in SWIR at both Leaf$_{ASD}$ and Crown$_{AISA}$
307 scales, while its separation from oak was best in NIR at Leaf$_{ASD}$ and Crown$_{AISA}$. The pattern
308 of spectral separation from ash changed at Crown$_{ASD}$ scale. In a similar way, beech was least
309 separable from hornbeam at Leaf$_{ASD}$ and Crown$_{AISA}$ scales but most separable from hornbeam
310 at Crown$_{ASD}$ scale (Figure 1.5, second and third rows). In addition, oak was separable from
311 beech, hornbeam and lime in SWIR, and separability between oak and ash increased in NIR
312 and VIS at both Leaf$_{ASD}$ and Crown$_{AISA}$ scales. This pattern of spectral separation from ash
313 changed at Crown$_{ASD}$ scale. (Figure 1.5, fifth row). Lastly, spectral separation from lime was
314 least consistent across the observed scales. Lime was most separable from beech and hornbeam
315 in SWIR at leaf scale but exhibited higher separability from oak and ash in VIS (at Crown$_{ASD}$
316 scale) and in VIS, NIR and SWIR at Crown$_{AISA}$ scale.
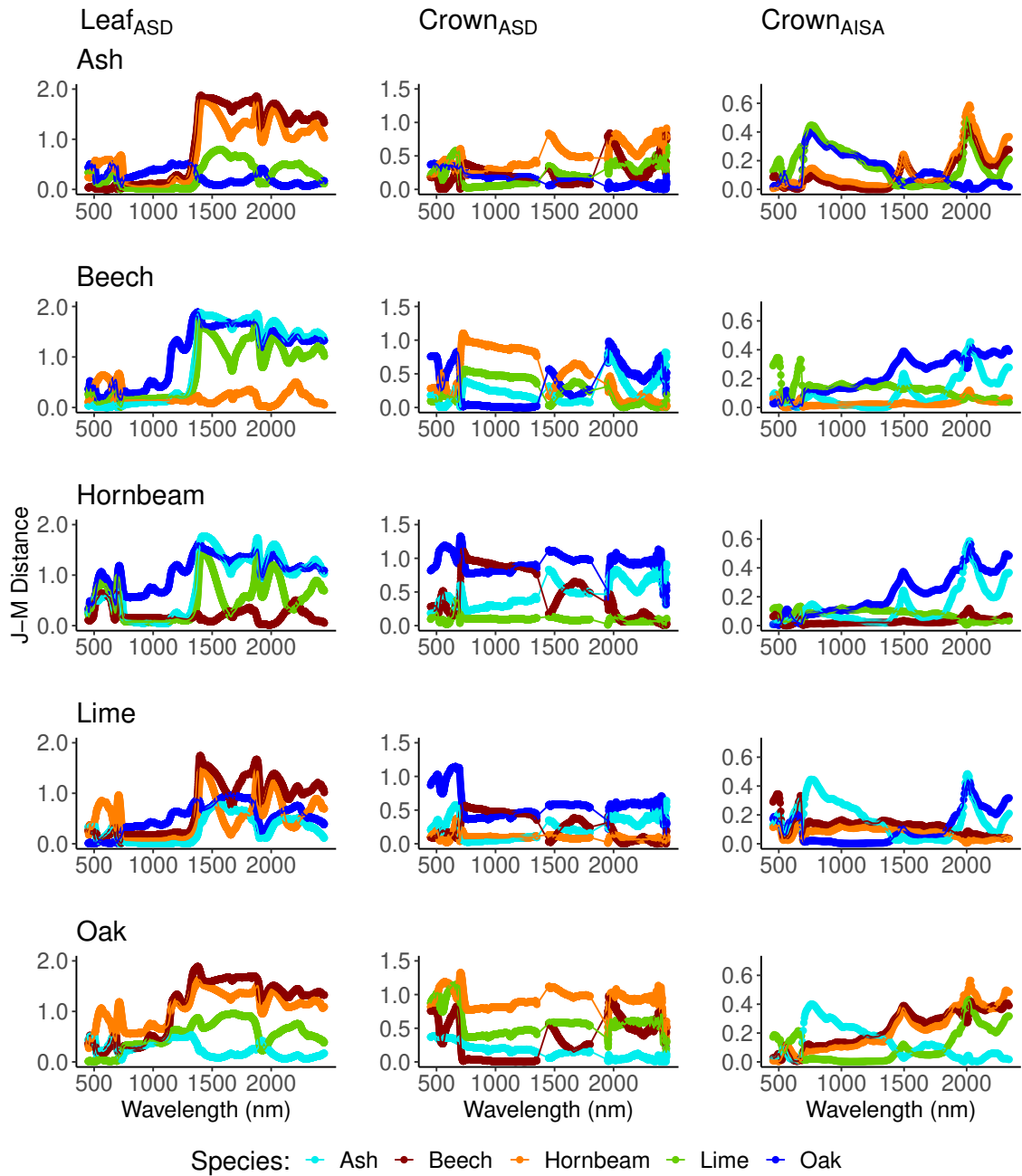
Figure 1.5: Columns from left to right show Jeffries-Matusita (J-M) distances per species pair at Leaf_{ASD}, $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ scales. The focal species is written in the row heading and the species in comparison plotted by color shown in the legend. Notice that y-axes change with scale of observation to emphasize inter-species differences at lower separability values.

1.4.3 Model performance and predicted structure and foliar chemical contents

Model performance results showed good correspondence between observed and modeled spectral reflectance justifying further analysis on contributions of predicted structure and foliar chemical contents to observed leaf and crown spectral reflectance using RTM sensitivity analyses (Section 1.4.4). A comparison of observed and predicted reflectance showed higher model performance for PROSPECT-5B at $Leaf_{ASD}$ scale compared to PROSAIL at both crown scales (Figure 1.6), but both models were generally successfull at predicting field observed spectral reflectance. However, correspondence between observed and predicted spectral reflectance highlighted a negative bias in the near infrared region at $Crown_{ASD}$ scale and higher errors especially among shortwave infrared spectra at $Crown_{AISA}$ scale.
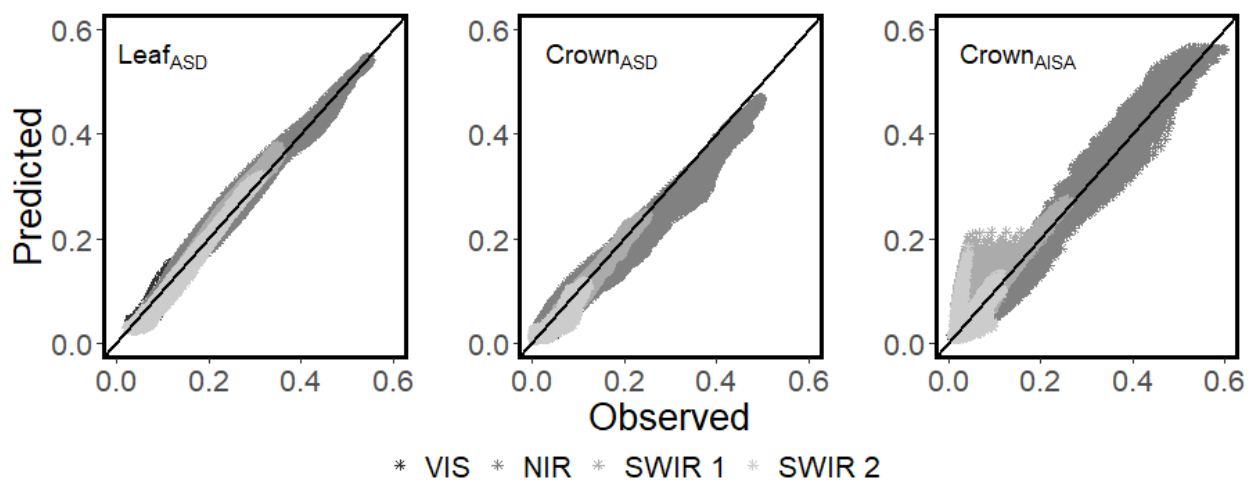
Figure 1.6: Observed versus predicted reflectance at $Leaf_{ASD}$ (left), $Crown_{ASD}$ (middle) and $Crown_{AISA}$ (right) scales. The different parts of the electromagnetic spectrum are distinguished by gray shades; VIS = Visible (450 - 700 nm), NIR = Near infrared (701 - 1450 nm), SWIR 1 = Shortwave infrared 1 (1451 - 1850 nm), SWIR 2 = Shortwave infrared 2 (1851 - 2450 nm).

Table 1.3: Estimates of structure and foliar chemical content from inversion of PROSPECT-5B and PROSAIL models at different scales of observation. LAI estimates correspond to only $Crown_{ASD}$ and $Crown_{AISA}$ scales. The used abbreviations in full: Leaf structure parameter (N), Chlorophyll a+b concentration (Cab), Carotenoid concentration (Car), Leaf brown pigments (Cbrown), Equivalent water thickness (Cw) and Leaf dry matter content (Cm) and Leaf area index (LAI).

	$Leaf_{ASD}$			$Crown_{ASD}$			$Crown_{AISA}$		
	min	mean	max	min	mean	max	min	mean	max
N	1.142	1.539	2.280	0.800	1.589	2.500	1.254	1.985	2.500
Cab	10.340	42.430	100.000	39.940	40.060	40.370	39.950	39.990	40.050
Car	2.016	7.737	23.496	9.978	10.010	10.073	9.990	10.000	10.040
Cbrown	0.000	0.078	0.273	0.000	0.230	1.000	0.000	0.085	0.497
Cw	0.005	0.012	0.021	0.008	0.023	0.117	0.012	0.022	0.049
Cm	0.002	0.007	0.016	0.000	0.018	0.118	0.000	0.005	0.020
LAI				0.406	2.261	3.999	1.283	3.105	4.285

Predictions of structure and foliar chemical contents retrieved from inversion of observed spectral reflectance using PROSPECT-5B and PROSAIL models are presented in Table 1.3.

All predictions at Leaf_{ASD} scale were similar in magnitudes and ranges to those reported in the reference dataset - the LOPEX database (Table 1.2, right). In a similar manner, ranges of N, Cw and Cm at $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ scales compared well with observations in the LOPEX database (Table 1.2, right). However, there was less variation in estimates of Cab and Car at both $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ scales compared to the Leaf_{ASD} scale and to the LOPEX database. In general, the predicted values of LAI at both $\text{Crown}_{\text{ASD}}$ and $\text{Crown}_{\text{AISA}}$ scales were low.

1.4.4 Predicted contributions of structure and foliar chemical contents to observed leaf and crown spectral reflectance

Model sensitivity at leaf scale was generally similar for parameters with ranges obtained from inversion of observed spectral reflectance with PROSPECT-5B and with ranges of the LOPEX database (Figure 1.7, Top panel). Sensitivity analyses at leaf scale showed variation in foliar chemical compounds to strongly affect reflectance in VIS with the highest contributions from chlorophyll a+b and carotenoid concentrations. The first-order indices showed the importance of chlorophyll a+b to have exceeded 75% between 547 - 709 nm while carotenoid contributions exceeded 35% between 506 - 528 nm. Within the same range of 506 - 528 nm, contributions of chlorophyll a+b content (28 - 35%) became lower than from carotenoid concentration. On the other hand, the leaf structural parameter had a generally limited effect in VIS with contributions between 3 - 21% for ranges from inversion of observed spectral reflectance and between 2 - 18% for LOPEX ranges. As leaf brown pigments did not vary in the LOPEX database (see Table 1.2), sensitivity analyses with LOPEX ranges did not include the parameter, however, leaf brown pigments had no contribution to spectral reflectance in VIS when considering ranges from inversion of observed spectral reflectance. Therefore, interactions among chlorophyll a+b , carotenoids and the leaf structure parameter contributed between 2 - 56% of observed variation when considering wider LOPEX ranges but their combined contributions reduced to between 2 - 43% with generally shorter ranges from inversion of observed spectral reflectance. In both sensitivity cases of the leaf scale, the highest contributions from interactions among chlorophyll a+b, carotenoids and the leaf structure parameter came from wavelength shorter than 510 nm.

The NIR region at leaf scale was dominated by contributions from the leaf structure parameter and dry matter contents except in the red-edge part where chlorophyll a+b contributed between 0.62 - 0.89% of the observed spectral variation and thus higher in contribution than the leaf structure parameter (at < 34%) and dry matter content (at < 1%). Contributions of the leaf structure parameter ranged between 9 - 91% in NIR compared to 9 - 34% in red-edge, while dry matter content accounted for about 0.05 - 10.6% of the observed spectral variation in NIR and < 1% of the observed spectral variation in red-edge (Figure 1.7, Top panel). Equivalent water thickness only marginally influenced observed reflectance in NIR between 950 and 1370 nm with local peaks in contribution corresponding to increased light absorption observed in Figure 1.4; at ≈ 1000 nm (950 - 1020 nm) and ≈ 1250 nm (1150 - 1250 nm). However, contributions of equivalent water thickness increased to 20 - 68% between 1369 - 1450 nm for
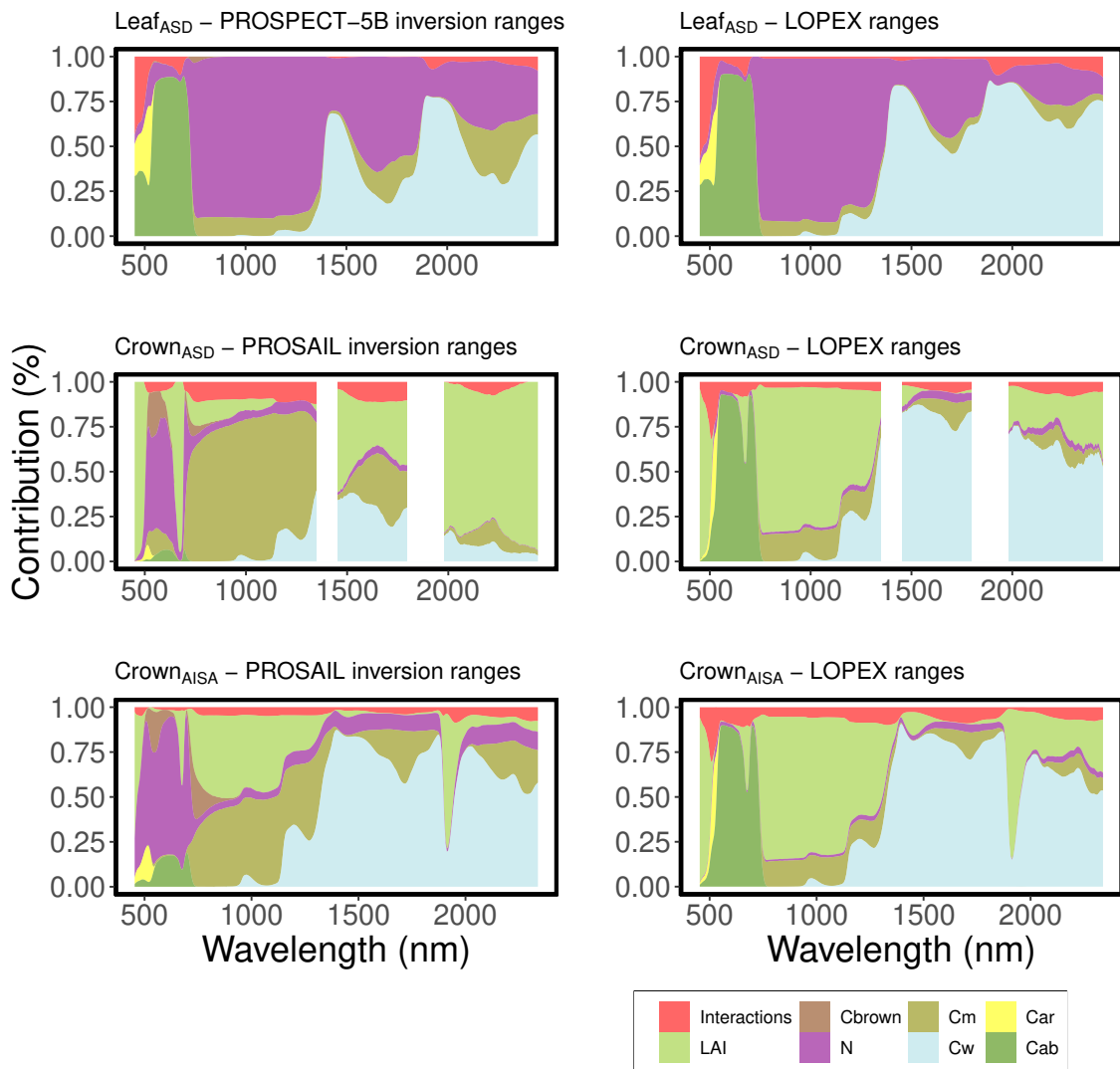
Figure 1.7: First-order sensitivity indices and interactions computed with PROSPECT-5B from ranges of predictions of structure and foliar chemical compounds retrieved from inversion of $\mathrm{Leaf_{ASD}}$ spectral reflectance (Top left) and by ranges of the LOPEX database (Top right). The middle and bottom panels show first-order sensitivity indices and interactions computed with PROSAIL from ranges of predictions of structure and foliar chemical compounds retrieved from inversion of $\mathrm{Crown_{ASD}}$ spectral reflectance (Middle left) and considering ranges of the LOPEX database (Middle right). The conditions at $\mathrm{Crown_{ASD}}$ scale are mimicked through masking of bands affected by water vapour (Middle). In the bottom panel, ranges of estimates of structure and foliar chemical compounds retrieved from inversion of $\mathrm{Crown_{AISA}}$ spectral reflectance (Bottom left) and considering ranges of the LOPEX database (Bottom right). Spectral resampling mimicked bands of $\mathrm{Crown_{AISA}}$ data. The used abbreviations in full: Leaf structure parameter (N), Chlorophyll a+b concentration (Cab), Carotenoid concentration (Car), Leaf brown pigments (Cbrown), Equivalent water thickness (Cw), Leaf dry matter content (Cm), Leaf Area Index (LAI).

narrower ranges from inversion of observed spectral reflectance and to 20 - 84% between 1321 - 1450 nm wavelength for wider LOPEX ranges. Leaf brown pigments had marginal impact on spectral reflectance in NIR with the highest contributions ranging between 3 - 3.5% from 736 - 762 nm wavelengths (Figure 1.7, Top panel left). In red-edge, contributions of leaf brown pigments content were minute (up to 0.9%). Similarly, interactions among the leaf structure parameter, dry matter content, and equivalent water thickness were of very low importance in NIR accounting for only $< 2.4\%$ of the observed variation in the region with LOPEX ranges and diminished to $< 1.7\%$ when leaf brown pigments were varied with respect to ranges from inversion of observed spectral reflectance.

Contributions in the SWIR region at leaf scale came from equivalent water thickness, the structure parameter, dry matter content and their interactions, in that sequence of importance. There were unexpected differences in contributions within SWIR at leaf scale when considering ranges from inversion of observed spectral reflectance (Figure 1.7, top left) and the LOPEX database (Figure 1.7, top right). Equivalent water thickness reduced in importance when sensitivity analyses were run with a narrower range from inversion of observed reflectance compared to the wider range of the LOPEX database. A reduction in importance of equivalent water thickness with a narrower range came with an increase in contributions of the leaf structure parameter and dry matter content. Contributions of equivalent water thickness had local maximum values at 1430 nm, 1950 nm, and > 2400 nm that coincided with higher light absorption points in Figure 1.4 and higher separability values in Figure 1.5. At these local peaks, contributions of equivalent water thickness exceeded 60%. The highest contributions of leaf structure parameter and dry matter content in the SWIR region were between 1600 - 1700 nm and between 2100 - 2350 nm and coincided with local peaks in reflectance shown in Figure 1.4. Leaf structure contributed up to 64.4% of variation for the range from inversion of observed spectral reflectance and up to 44.2% of the variation for the LOPEX range. On the other hand, leaf dry matter content contributed up to 34% of the variation for the range from inversion of observed spectral reflectance and up to 12.8% of the variation for the LOPEX range at the local peaks seen in Figure 1.4 in SWIR.

Sensitivity analyses of PROSAIL as calibrated by ranges from inversion of observed spectral reflectance and LOPEX ranges, yielded contrasting results at both $\mathrm{Crown_{ASD}}$ and $\mathrm{Crown_{AISA}}$ scales (Figure 1.7, middle and bottom panels). Chlorophyll a+b (with contributions reaching $> 75\%$) and carotenoids (with contributions up to 39.6%) dominated in importance among reflectance in VIS (between 530 - 700 nm) when sensitivity analysis was calibrated by wider LOPEX ranges (Figure 1.7 right, middle and bottom panels). However, the contribution of the two biochemical contents dropped to $< 20\%$ at both crown scales when considering narrow ranges from inversion of observed spectral reflectance (Figure 1.7 left, middle and bottom panels). The loss in importance of chlorophyll a+b and carotenoids in VIS with narrow parameter ranges was compensated by increases in contributions from the leaf structure parameter, brown pigments, dry matter content and interactions among parameters. However, predictions from sensitivity analyses considering LOPEX ranges at $\mathrm{Crown_{ASD}}$ and $\mathrm{Crown_{AISA}}$ scales (Figure 1.7 right, middle and bottom panels) were more consistent with predictions at $\mathrm{Leaf_{ASD}}$ scale

(Figure 1.7, top panel) as compared to predictions from sensitivity analyses calibrated by ranges from inversion of observed spectral reflectance (Figure 1.7 left, middle and bottom panels). Inclusion of LAI in sensitivity analyses calibrated by LOPEX ranges at both crown scales as compared to the leaf scale, reduced contributions of the leaf structure parameter, chlorophyll a+b and carotenoids and their interactions in VIS. Specifically, the contribution of LAI ranged between 48.4 - 95.2% at bands < 500 nm. LAI diminished in importance to 0% between 525 - 612 nm allowing for marginal effects from N (between 1.7 - 6.8%). It again increased up to 37% at 677 nm again coinciding with a reduction contribution from the leaf structure parameter and finally reduced to 0% from 696 - 700 nm where contributions from the leaf structure parameter were evident up to 2.4%. At bands where interaction among parameters in VIS increased in contribution (up to 31.8% in VIS), LAI had the least effect.

The red-edge and NIR regions exhibited contradictions between predictions from sensitivity analyses calibrated after inversion of observed spectral reflectance and LOPEX ranges at both crown scales. Whereas contributions from equivalent water thickness, LAI and leaf structure parameter dominated the regions when sensitivity analyses were calibrated by ranges from inversion of observed spectral reflectance (Figure 1.7 left, middle and bottom panels), chrolophyll a+b, LAI and leaf dry matter content dominated the regions when sensitivity analyses were calibrated by ranges of the LOPEX database (Figure 1.7 right, middle and bottom panels). Similar to results in VIS, red-edge and NIR predictions from LOPEX ranges at both crown scales were more consistent with results at leaf scale compared to predictions obtained after inverting observed spectral reflectance. For example, chrolophyll a+b dominated red-edge with contributions > 75% while LAI dominated NIR with contributions up to 81.7% (Figure 1.7 right, middle and bottom panels) replacing the significant contribution from N observed at leaf scale. On the other hand, contributions from the leaf structure parameter dominated red-edge between 17.6 - 57.2% at $\text{Crown}_{\text{ASD}}$ scale and between 28.2 - 62.3% at $\text{Crown}_{\text{AISA}}$ scale compensating for a loss in sensitivity of chrolophyll a+b with narrower ranges (Figure 1.7 left, middle and bottom panels). In addition, equivalent water thickness dominated NIR with contributions between 2.7 - 80.7% at $\text{Crown}_{\text{ASD}}$ scale and 0 - 48.5% at $\text{Crown}_{\text{AISA}}$ scale (Figure 1.7 left, middle and bottom panels) corresponding to large differences in retrieved ranges of leaf dry matter contents at both crown scales (Table 1.3). However, unlike other parameters contributing to observed variation in NIR, contributions of equivalent water thickness in at both crown scales and for both sensitivity analysis cases were consistent with results at the leaf scale; suggesting generally successful inversion of observed spectral reflectance with respect to equivalent water thickness (Figure 1.7). Despite the differences in magnitudes of contributions from equivalent water thickness after inversion of observed spectral reflectance at both crown scales (Figure 1.7 left, middle and bottom panels), the local peaks seen at leaf scale between 950 - 1020 nm and 1150 - 1250 nm were preserved and corresponded with increased light absorption at those scales (Figure 1.4). Equivalent water thickness at $\text{Crown}_{\text{AISA}}$ scale contributed between 70 - 90% of the observed spectral variation at 1360 - 1450 nm; consistent for both sensitivity analysis cases and with results at leaf scale.

Contributions in SWIR were generally dominated by equivalent water thickness at both

crown scales except for sensitivity analysis with ranges from inversion of observed spectral reflectance at $\mathrm{Crown}_{\mathrm{ASD}}$ scale, where LAI dominated. Again notice that a wider equivalent water thickness range was retrieved from model inversion at $\mathrm{Crown}_{\mathrm{ASD}}$ than at $\mathrm{Leaf}_{\mathrm{ASD}}$ and $\mathrm{Crown}_{\mathrm{AISA}}$ scales (Table 1.3). Contributions of equivalent water thickness at both crown scales were on average higher than observed at the leaf scale in the SWIR region (Figure 1.7 middle right and bottom panels). Similarly, local peaks in contributions of equivalent water thickness observed at leaf scale were preserved at both crown scales and coincided with increase in absorption of spectral reflectance (Figure 1.4). Inclusion of LAI into model parameters resulted into a reduction of contributions of the leaf structure parameter and dry matter contents within SWIR at both crown scales compared to the leaf scale. In addition, there was a peculiar increased in contributions of LAI (up to 83.5%) between 1850 - 1960 nm at $\mathrm{Crown}_{\mathrm{AISA}}$ scale. LAI contributions between 1850 - 1960 nm could not be assessed at $\mathrm{Crown}_{\mathrm{ASD}}$ scale as the bands were masked out.

1.5 Discussion

1.5.1 Scales of observation

This study presents results on the relative trade-off between features of tree spectral reflectance acquired by imaging and non-imaging spectroscopy and the spatial scale of observation. We only have a few studies on the linkages between non-imaging and imaging spectroscopy data (Asner et al., 1998; Clark et al., 2005). Studies with coincident spectroscopy data that removes confounding effects from differences in ages of observed vegetation and atmospheric conditions are all but lacking. There has been limited research at $\mathrm{Crown}_{\mathrm{ASD}}$ scale partly because field-measured non-imaging spectroscopy data can be challenging to collect given the sizes and canopy structures of trees. This study was greatly facilitated by the existence of a canopy walk within Hainich National Park (Figure1.1), however, recent advances in use of UAVs (with imaging sensors) in forest monitoring may facilitate studies at the $\mathrm{Crown}_{\mathrm{ASD}}$ scale.

1.5.2 Properties of broadleaved tree spectral reflectance obtained from leaves and crowns

The study shows that sampled leaves and crowns exhibited typical spectral reflectance patterns of vegetation. The reflectance absorption features in VIS are known to stem from photosynthetic pigment dominated by chlorophyll concentrations (Asner, 1998), of which discrimination of chlorophyll into a and b, and consideration of other photosynthetic pigments like carotenoids, anthocyanins (Gitelson et al., 2001), brown pigment, improve the accounting of observed vegetation reflectance properties in VIS (Féret et al., 2008, 2017). On the other hand, reflectance absorption features in SWIR are dominated by water content (Ceccato et al., 2001) while the high reflectance in NIR are determined principally by the structural properties of vegetation material (Slaton et al., 2001), and in our case excerbated by the woody nature of trees explaining why individual spectral reflectance at crown scales were as high as at the leaf scale. These

findings were generally confirmed by a comparison of predictions of contributions of structure and foliar chemical contents from RTM sensitivity analysis calibrated by ranges from inversion of observed leaf and crown spectral reflectance and the LOPEX dataset (Figure 1.7). Second, we show that means and variances of spectral reflectance shift downwards with changing the scale of observation from leaves to crowns. This finding confirms results from Clark et al. (2005) that used field-measured ASD and HYDICE spectral reflectance. We attribute the observed reduction/darkening in spectral reflectance from leaves to crowns to a reduction in leafy material and fine-scale shadows within branches. Third, our study shows increasing absorption properties in NIR (at $\approx$ 1000 nm and 1250 nm) from leaves to crowns. These absorption features can be attributed to multiple scattering of photons among leaves and other crown parts emphasizing the water absorption properties at local peaks in contributions of water content as for example seen in Figure 1.7 from leaves to crowns. A similar phenomenon was observed by Roberts et al. (2004). Lastly, we show that spectral variation across studied tree species increased from leaves to crowns and was highest at $\mathrm{Crown_{ASD}}$ scale. Asner et al. (1998) found that optical properties of woody vegetation were generally more variable than for fresh leaves explaining the increase in variability among reflectance at $\mathrm{Crown_{ASD}}$ scale where the possibility of scanning woody branches was higher. On the other hand, we attribute the reduction in variance across spectral reflectance at $\mathrm{Crown_{AISA}}$ scale to full foliar cover at the time of data collection. Nonetheless, effects from using a halogen lamp as the source of radiation at the leaf scale and the sun at both crown scales may also have affected the observed variability at leaf scale relative to both crown scales.

1.5.3 Discrimination among species spectral reflectance obtained from leaves and crowns

Variation among species spectral reflectance as well as the spectral resolution affected species separability. There was a higher spectral variation at both crown scales and a lower spectral resolution of the $\mathrm{Crown_{AISA}}$ scale (Figure 1.6). Hence, as we move to larger spatial scales and lower spectral scales, our ability to classify individual tree crowns to species decreases. A decrease is mainly due to a reduction in spectral 'purity' (Figure 1.4) when the spatial resolution of observation exceeds the size of the reflecting object (minimum mapping unit). In addition, as we perform spectral resampling and smoothing, we lose fine-scale information potentially useful to discriminate species. Therefore, a balance between spatial and spectral resolutions is necessary to reduce spectral variation and yet preserve separability among species spectral reflectance. Our results show that the $\mathrm{Crown_{ASD}}$ scale exhibited higher species separability than observed at $\mathrm{Crown_{AISA}}$ scale but the pattern of separability was not consistent with the other two scales of observation. We attribute the higher separability at $\mathrm{Crown_{ASD}}$ scale to a higher spectral resolution in comparison to $\mathrm{Crown_{AISA}}$ scale. On the other hand, species separability at the $\mathrm{Crown_{AISA}}$ scale was more consistent with the leaf scale compared to at $\mathrm{Crown_{ASD}}$ scale. This could be attributed to a reduction in leafy material as a consequence of a high spatial resolution at $\mathrm{Crown_{ASD}}$ scale. In other words, the lower spatial resolution at

$Crown_{AISA}$ scale allowed for observation of more leafy material at the respective LAI values of the sampled crowns, thus increasing spectral consistency with the leaf scale.

We showed that generally all observed regions of the electromagnetic spectrum were useful in discrimination of species but VIS, red-edge and SWIR exhibited the highest discrimination power (Figure 1.6). In addition, we showed that separability in SWIR was actually highest at bands most affected by the amount of water content while separability in VIS and red-edge was highest at bands most affected by chlorophyll concentrations (Figure 1.7). This finding highlights the importance of rarely used vegetation characteristics like water content to discriminate studied broadleaved tree species at different scales of observation. However, it can not be ruled out that species adaptation or sensitivity to drought conditions could easily bias discrimination relying only on leaf water content. Further, this result confirms the utility of chlorophyll concentrations in discrimination of tree species (Figure 1.7). Considering that we studied only broad leaved tree species, it could be expected that the NIR region is less powerful for the discrimination of tree species with similar internal foliage structures.

1.5.4 Utility of open spectral libraries for non-validated inversion of observed spectral reflectance obtained from leaves and crowns

Figures 1.6 and 1.7 indicate PROSPECT-5B to have been relatively successful at inversion of observed spectral reflectance at leaf scale. These findings confirm reports by Ali et al. (2016a); Gara et al. (2019) regarding performance of the PROSPECT model in forest settings at leaf scale. In addition, our study results indicate that PROSAIL may have the potential to robustly retrieve equivalent water thickness at $Crown_{AISA}$ scale even with difficulties to predict other model input parameters at both crown scales. This may be due to the relatively very high sensitivity of reflectance in the NIR and SWIR regions of the electromagnetic spectrum to changes in equivalent water thickness at all studied scales of observation. However, model sensitivity analysis highlighted weaknesses of PROSAIL in a heterogeneous forest environment (Figure 1.7 and Table 1.3) especially regarding the retrieval of chlorophyll a+b and carotenoids at both crown scales. A comparative assessment of model input ranges from inversion of observed spectral reflectance and ranges of the LOPEX dataset showed that wider ranges generally resulted in higher contributions to model outputs and consistency in contributions across scales of observation (Figure 1.7). For example, very narrow ranges of chlorophyll a+b and carotenoid concentrations from inversion of observed spectral reflectance by PROSAIL resulted in smaller contributions to observed spectral variation at both crown scales and were inconsistent with generalized model assumptions (as discussed in section 1.5.2) as well as results at leaf scale. With the exception of equivalent water thickness at $Crown_{AISA}$ scale, it appears that PROSAIL can not estimate input parameters from inversion of observed spectral reflectance at both crown scales. The retrieval of structure and foliar chemical contents by inversion of observed spectral reflectance using PROSAIL was ill-posed at both crown scales. This means that different combinations of input parameters may produce similar spectral reflectance curves (Figure 1.6) resulting in inaccurate retrieval of model input parameters and

in turn different attributions of model input parameters to the variation in observed spectral reflectances. Considering that PROSAIL assumes the forest canopy as a collection absorbing and scattering tissues randomly distributed on a horizontal surface, the model may have failed to account for the complexity of vertically heterogeneous forest canopies. These unrealistic assumptions in forest settings have spurned others to explore more complex models (Ali et al., 2016b; Wang et al., 2018).

To overcome model ill-position, Combal et al. (2002) recommended the use of prior information on ranges of structure and foliar chemical contents often acquired through rigorous field sampling and laboratory analysis that can be highly demanding for vertically heterogeneous forest canopies. Malenovský (2006); Ali et al. (2016a) reported significant improvement in accuracy of parameter retrieval after using prior information. However, information from non-representative samples may affect overall model specificity and sensitivity. The lack of validation data on structure and foliar chemical contents clearly limits our ability to gain further insights. However, we show that open spectral libraries combined with model sensitivity analyses can support identification of model inaccuracies stemming from theoretical uncertainty as opposed to physical uncertainty that can only be assessed with validation data (Malenovský et al., 2019). PROSAIL appears less suited for application in a heterogeneous forest environment and without validation data compared to 3D models as for example presented in Widlowski et al. (2015). In addition, our comparative approach highlights the utility of open spectral libraries like the LOPEX dataset from which typical ranges of physical and biochemical properties can be bench-marked across species. It should be explicitly noted that our approach was by no means intended to substitute field-sampled and laboratory-analyzed validation datasets in terms of accounting for model reliability and fidelity but is rather an initial check of model performance with regard to adherence to generalized assumptions such as contributions of structure and foliar chemical contents across the studied wavelengths. An additional caveat is that global sensitivity analyses like here presented will be limited to relatively simple RTMs like PROSAIL since the theoretical uncertainty of RTMs increases with increasing model complexity (Malenovský et al., 2019), making inclusion of prior information a more reliable option.

1.6 Conclusions and outlook

We draw the following conclusions from this study;

1. The leaf and crown spectral reflectance of the sampled broad leaved species are generally similar between 400 nm - 2500 nm. A change from leaf to crown scales darkens spectral reflectance and increases the across-species variability. The highest variance among spectral reflectance in this study observed at $\text{Crown}_{\text{ASD}}$ scale.

2. Separability among species spectral reflectances is a function of the spatial and spectral resolution of imaging and non-imaging spectroscopy data collected between 400 nm - 2500 nm, whereby all regions of the observed electromagnetic spectrum are important for separation with the SWIR emerging as the region highest importance.

3. Photosynthetic pigments especially chlorophyll a+b contribute most to the variation in spectral reflectance in the visible and red-edge regions of the electromagnetic spectrum while equivalent water thickness contributes most to the variation of spectral reflectances in the shortwave infrared region at all studied scales. Leaf internal structures determine variation in spectral reflectance in the near infrared region at leaf scale whereas LAI exerts a strong influence on the variation in spectral reflectance within NIR region at both crown scales.

From an applied perspective, the new hyperspectral airborne systems and the upcoming satellites such as EnMAP, PRISMA, HISUI, CHIME, - at a 30 m GSD and 420 - 2500 nm spectral interval (Guanter et al., 2015), and the general need for continuing research on (non-) imaging spectroscopy (Schaepman et al., 2009), underline the importance and necessity of more specific scale-related analyses, if their products are to be validated appropriately (Malenovský et al., 2019). We expect that issues on the scale of observation addressed in this study may become of even greater importance in the near future when up-coming imaging spectroscopy satellites become operational.

1.7 Acknowledgement

641 Exploratories"(DFG-Refno. KL894/21-1 and KL894/21-2). Field work permits were issued by

642 the responsible state environmental offices of Baden-Württemberg, Thüringen and Brandenburg

643 (according to § 72 BbgNatSchG).

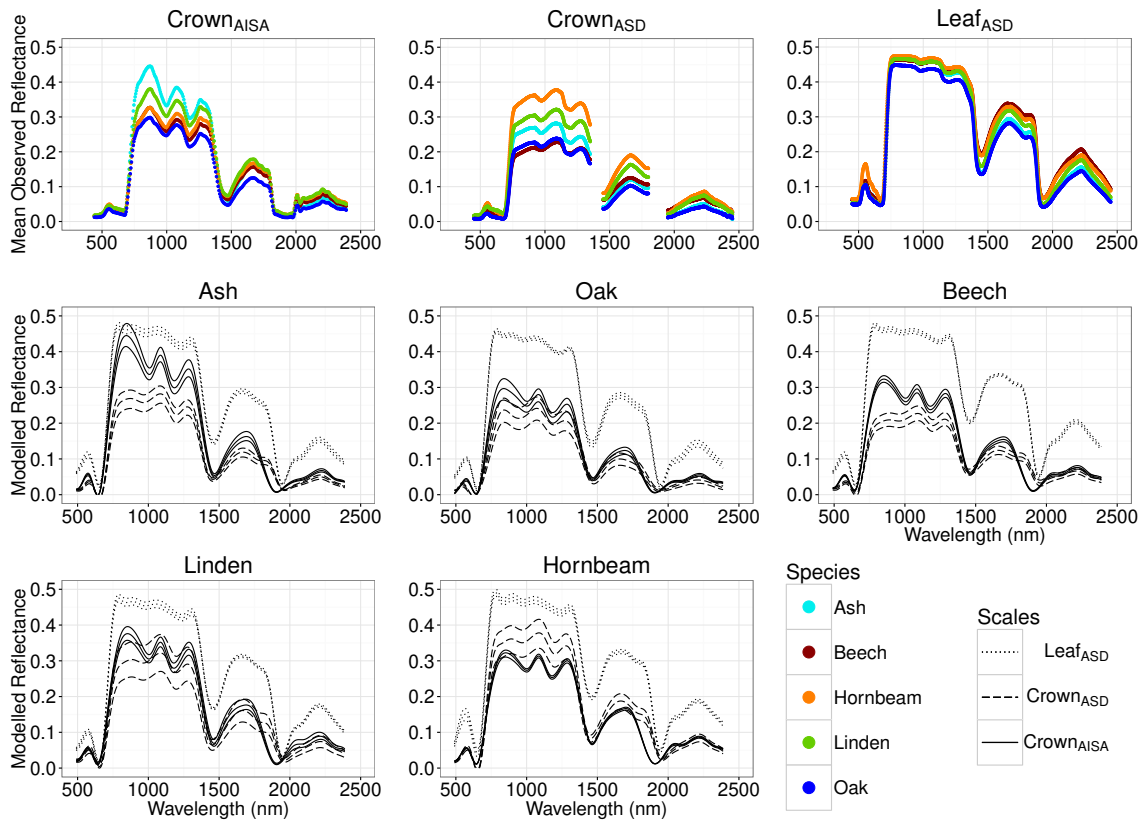## 1.8 Supplementary material

644

Figure 1.8: Top row shows mean observed reflectance at airborne crown, in-situ crown and leaf scales. The other panels show species-wise mean modelled reflectance and the associated standard errors at airborne crown, in-situ crown and leaf scales.

References

Aberle, H. 2016. Hyperspectral remote sensing and field measurements for forest characteristics - A case study in the Hainich National Park, Central Germany. PhD thesis, University of Goettingen.

Adam, E., and O. Mutanga. 2009. Spectral discrimination of papyrus vegetation (Cyperus papyrus L.) in swamp wetlands using field spectrometry. *ISPRS Journal of Photogrammetry and Remote Sensing* 64(6): 612–620.

Ali, A. M., R. Darvishzadeh, A. K. Skidmore, I. van Duren, U. Heiden, and M. Heurich. 2016a. Estimating leaf functional traits by inversion of PROSPECT: Assessing leaf dry matter content and specific leaf area in mixed mountainous forest. *International Journal of Applied Earth Observation and Geoinformation* 45: 66–76.

Ali, A. M., R. Darvishzadeh, A. K. Skidmore, and I. V. Duren. 2016b. Effects of Canopy Structural Variables on Retrieval of Leaf Dry Matter Content and Specific Leaf Area from Remotely Sensed Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9(2): 898–909.

Asner, G. P. 1998. Biophysical and Biochemical Sources of Variability in Canopy Reflectance. *Remote Sensing of Environment* 64(3): 234–253.

Asner, G. P., R. E. Martin, C. B. Anderson, and D. E. Knapp. 2015. Quantifying forest canopy traits: Imaging spectroscopy versus field survey. *Remote Sensing of Environment* 158: 15–27.

Asner, G. P., C. A. Wessman, and S. Archer. 1998. Scale Dependence of Absorption of Photosynthetically Active Radiation in Terrestrial Ecosystems. *Ecological Applications* 8(4): 1003–1021.

Bannari, A., D. Morin, F. Bonn, and A. R. Huete. 1995. A review of vegetation indices. *Remote Sensing Reviews* 13(1-2): 95–120.

Berger, K., C. Atzberger, M. Danner, G. D'Urso, W. Mauser, F. Vuolo, and T. Hank. 2018. Evaluation of the PROSAIL Model Capabilities for Future Hyperspectral Model Environments: A Review Study. *Remote Sensing* 10(1).

Bidot, C., M. Lamboni, and H. Monod. 2018. *multisensi: Multivariate Sensitivity Analysis*. R package version 2.1-1. URL `https://CRAN.R-project.org/package=multisensi`

674 BMEL. 2015. *The Forests in Germany. Selected Results of the Third National Forest Inventory.*
675 Bonn: Federal Ministry of Food and Agriculture.

676 Brantley, S. T., J. C. Zinnert, and D. R. Young. 2011. Application of hyperspectral vegetation
677 indices to detect variations in high leaf area index temperate shrub thicket canopies. *Remote*
678 *Sensing of Environment* 115(2): 514–523.

679 Ceccato, P., S. Flasse, S. Tarantola, S. Jacquemoud, and J. M. Grégoire. 2001. Detecting
680 vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of*
681 *Environment* 77(1): 22–33.

682 Cho, M. A., I. Sobhan, A. K. Skidmore, and J. D. Leeuwb. 2008. Discriminating Species Using
683 Hyperspectral Indices At Leaf and Canopy Scales. *Remote Sensing and Spatial Information*
684 *Sciences.* XXXVII(B7): 369–376.

685 Clark, M., D. Roberts, and D. Clark. 2005. Hyperspectral discrimination of tropical rain forest
686 tree species at leaf to crown scales. *Remote sensing of Environment* 96: 375–398.

687 Clark, M. L., and D. A. Roberts. 2012. Species-level differences in hyperspectral metrics among
688 tropical rainforest trees as determined by a tree-based classifier. *Remote Sensing* 4(6):
689 1820–1855.

690 Combal, B., F. Baret, M. Weiss, A. Trubuil, D. Macé, A. Pragn??re, R. Myneni, Y. Knyazikhin,
691 and L. Wang. 2002. Retrieval of canopy biophysical variables from bidirectional reflectance
692 using prior information to solve the ill-posed inverse problem. *Remote Sensing of Environment*
693 84(1): 1–15.

694 Dalponte, M., H. O. Ørka, T. Gobakken, D. Gianelle, and E. Næsset. 2013. Tree species
695 classification in boreal forests with hyperspectral data. *IEEE Transactions on Geoscience*
696 *and Remote Sensing* 51(5): 2632–2645.

697 Eilers, P., B. Marx, and M. Durbán. 2015. Twenty years of P-Splines. *Statistics and Operations*
698 *Research Transactions* 39: 149–186.

699 Fassnacht, E. F., H. Latifi, K. Stereńczak, A. Modzelewska, M. Lefsky, L. T. Waser, C. Straub,
700 and A. Ghosh. 2016. Review of studies on tree species classification from remotely sensed
701 data. *Remote Sensing of Environment* 186(August): 64–87.

702 Feilhauer, H., G. P. Asner, and R. E. Martin. 2015. Multi-method ensemble selection of spectral
703 bands related to leaf biochemistry. *Remote Sensing of Environment* 164: 57–65.

704 Féret, J.-B., and G. P. Asner. 2013. Tree Species Discrimination in Tropical Forests Using
705 Airborne Imaging Spectroscopy. *IEEE Trans. Geoscience Remote Sensing* 51(1): 73–84.

706 Féret, J. B., C. François, G. P. Asner, A. A. Gitelson, R. E. Martin, L. P. R. Bidel, S. L. Ustin,
707 G. le Maire, and S. Jacquemoud. 2008. PROSPECT-4 and 5: Advances in the leaf optical

708 properties model separating photosynthetic pigments. *Remote Sensing of Environment*
709 112(6): 3030–3043.

710 Féret, J. B., A. A. Gitelson, S. D. Noble, and S. Jacquemoud. 2017. PROSPECT-D: Towards
711 modeling leaf optical properties through a complete lifecycle. *Remote Sensing of Environment*
712 193: 204–215.

713 Ferreira, M. P., A. E. B. Grondona, S. B. A. Rolim, and Y. E. Shimabukuro. 2013. Analyzing
714 the spectral variability of tropical tree species using hyperspectral feature selection and leaf
715 optical modeling. *Journal of Applied Remote Sensing* 7(1): 73502.

716 Fischer, M., O. Bossdorf, S. Gockel, F. Hänsel, A. Hemp, D. Hessenmöller, G. Korte, J. Nieschulze,
717 S. Pfeiffer, D. Prati, S. Renner, I. Schöning, U. Schumacher, K. Wells, F. Buscot, E. K. Kalko,
718 K. E. Linsenmair, and W. W. W. E.D. Schulze. 2010. Implementing large-scale and long-term
719 functional biodiversity research: The Biodiversity Exploratories. *Basic and Applied Ecology*
720 11: 473–485.

721 Gara, T. W., R. Darvishzadeh, A. K. Skidmore, and T. Wang. 2018. Impact of vertical canopy
722 position on leaf spectral properties and traits across multiple species. *Remote Sensing* 10(2):
723 1–17.

724 Gara, T. W., R. Darvishzadeh, A. K. Skidmore, T. Wang, and M. Heurich. 2019. Evaluating
725 the performance of PROSPECT in the retrieval of leaf traits across canopy throughout the
726 growing season. *International Journal of Applied Earth Observation and Geoinformation*
727 83(June): 101919.

728 Gitelson, A. A., Y. Gritz, and M. N. Merzlyak. 2003. Relationships between leaf chlorophyll
729 content and spectral reflectance and algorithms for non-destructive chlorophyll assessment
730 in higher plant leaves. *Journal of plant physiology* 160(3): 271–282.

731 Gitelson, A. A., M. N. Merzlyak, and O. B. Chivkunova. 2001. Optical properties and nonde-
732 structive estimation of anthocyanin content in plant leaves. *Photochemistry and photobiology*
733 74(1): 38–45.

734 Groeneveld, D. P., W. M. Baugh, J. B. Silverman, and D. D. Barz. 2006. Non-imaging Airborne
735 Spectroscopy for Calibrating Satellite Imagery. Technical report.

736 Guanter, L., H. Kaufmann, K. Segl, S. Foerster, C. Rogass, S. Chabrillat, T. Kuester, A. Hollstein,
737 G. Rossner, C. Chlebek, C. Straif, S. Fischer, S. Schrader, T. Storch, U. Heiden, A. Mueller,
738 M. Bachmann, H. Mühle, R. Müller, M. Habermeyer, A. Ohndorf, J. Hill, H. Buddenbaum,
739 P. Hostert, S. van der Linden, P. Leitão, A. Rabe, R. Doerffer, H. Krasemann, H. Xi, W. Mauser,
740 T. Hank, M. Locherer, M. Rast, K. Staenz, and B. Sang. 2015. The EnMAP Spaceborne
741 Imaging Spectroscopy Mission for Earth Observation. *Remote Sensing* 7(7): 8830–8857.
742 URL http://www.mdpi.com/2072-4292/7/7/8830/

743 Hédl, R., M. Svátek, M. Dančák, A. Rodzay, A. Salleh, and A. Kamariah. 2009. A new technique
744 for inventory of permanent plots in tropical forests: a case study from lowland dipterocarp
745 forest in Kuala Belalong, Brunei Darussalam. *Blumea* 54: 124–130.

746 Homolová, L., Z. Malenovský, J. G. P. W. Clevers, G. García-Santos, and M. E. Schaepman.
747 2013. Review of optical-based remote sensing for plant trait mapping. *Ecological Complexity*
748 15: 1–16.

749 Hosgood, B., S. Jacquemound, G. Andreeoli, J. Verdebout, A. Pedrini, and G. Schmuck. 1993.
750 Leaf Optical Properties Experiment Database (LOPEX93). Data available at, `http://`
751 `ecosis.org` from the Ecological Spectral Information System (EcoSIS).

752 Huete, A. R., H. Q. Liu, K. Batchily, and W. Van Leeuwen. 1997. A comparison of vegetation
753 indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*
754 59(3): 440–451.

755 Iooss, B., A. Janon, G. Pujol, with contributions from Baptiste Broto, K. Boumhaout, S. D.
756 Veiga, T. Delage, J. Fruth, L. Gilquin, J. Guillaume, L. Le Gratiet, P. Lemaitre, A. Marrel,
757 A. Meynaoui, B. L. Nelson, F. Monari, R. Oomen, O. Rakovec, B. Ramos, O. Roustant,
758 E. Song, J. Staum, R. Sueur, T. Touati, and F. Weber. 2020. *sensitivity: Global Sensitivity*
759 *Analysis of Model Outputs*. R package version 1.17.1. URL `https://CRAN.R-project.`
760 `org/package=sensitivity`

761 Jacquemoud, S., and F. Baret. 1990. PROSPECT: A model of leaf optical properties spectra.
762 *Remote Sensing of Environment* 34(2): 75–91.

763 Jacquemoud, S., W. Verhoef, F. Baret, C. Bacour, P. J. Zarco-Tejada, G. P. Asner, C. François,
764 and S. L. Ustin. 2009. PROSPECT + SAIL models: A review of use for vegetation charac-
765 terization. *Remote Sensing of Environment* 113(SUPPL. 1): S56—-S66.

766 James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An introduction to statistical learning*.
767 URL `http://link.springer.com/content/pdf/10.1007/978-1-4614-7138-7.pdf`

768 Knohl, A., E. D. Schulze, O. Kolle, and N. Buchmann. 2003. Large carbon uptake by an
769 unmanaged 250-year-old deciduous forest in Central Germany. *Agricultural and Forest*
770 *Meteorology* 118(3-4): 151–167.

771 Kokaly, R., R. Clark, G. Swayze, K. Livo, T. Hoefen, N. Pearson, R. Wise, W. Benzel, H. Lowers,
772 R. Driscoll, and A. Klein. 2017. USGS spectral library version 7. *United States Geological*
773 *Survey Data Series 1035* 61.

774 Kumar, L., A. Skidmore, and O. Mutanga. 2010. Leaf level experiments to discriminate between
775 eucalyptus species using high spectral resolution reflectance data: use of derivatives, ratios
776 and vegetation indices. *Geocarto International* (March 2013): 37–41.

777 Lui, G., and D. Coomes. 2015. A Comparison of Novel Optical Remote Sensing-Based Tech-
778 nologies for Forest-Cover/Change Monitoring. *Remote Sensing* 7(3): 2781–2807.

Malenovský, Z. 2006. Quantitative remote sensing of Norway spruce (Picea abies (L.) Karst.): Spectroscopy from needles to crowns to canopies. PhD thesis.

Malenovský, Z., H. M. Bartholomeus, F. W. Acerbi-Junior, J. T. Schopfer, T. H. Painter, G. F. Epema, and A. K. Bregt. 2007. Scaling dimensions in spectroscopy of soil and vegetation. *International Journal of Applied Earth Observation and Geoinformation* 9(2): 137–164.

Malenovský, Z., L. Homolová, P. Lukeš, H. Buddenbaum, J. Verrelst, L. Alonso, M. E. Schaepman, N. Lauret, and J. P. Gastellu-Etchegorry. 2019. Variability and Uncertainty Challenges in Scaling Imaging Spectroscopy Retrievals and Validations from Leaves Up to Vegetation Canopies. *Surveys in Geophysics* 40(3): 631–656.

Mund, M., W. L. Kutsch, C. Wirth, T. Kahl, A. Knohl, M. V. Skomarkova, and E. D. Schulze. 2010. The influence of climate and fructification on the inter-annual variability of stem growth and net primary productivity in an old-growth, mixed beech forest. *Tree Physiology* 30(6): 689–704.

O'Connor, B., C. Secades, J. Penner, R. Sonnenschein, A. Skidmore, N. D. Burgess, and J. M. Hutton. 2015. Earth observation as a tool for tracking progress towards the Aichi Biodiversity Targets. *Remote Sensing in Ecology and Conservation* 1(1): 19–28.

Pfitzner, K., A. Bollhöfer, and G. Carr. 2006. A standard design for collecting vegetation reference spectra: Implementation and implications for data sharing.

Prospere, K., K. Mclaren, and B. Wilson. 2014. Plant Species Discrimination in a Tropical Wetland Using In Situ Hyperspectral Data. *Remote Sensing* 6: 8494–8523.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`

Richards, J. A. 2013. *Remote Sensing Digital Image Analysis: An Introduction*. 5th ed. Berlin / Heidelberg: Springer.

Roberts, D. A., S. L. Ustin, S. Ogunjemiyo, J. Greenberg, S. Z. Dobrowski, J. Chen, and T. M. Hinckley. 2004. Spectral and Structural Measures of Northwest Forest Vegetation at Leaf to Landscape Scales. *Ecosystems* 7(5): 545–562.

Roosjen, P. P., B. Brede, J. M. Suomalainen, H. M. Bartholomeus, L. Kooistra, and J. G. Clevers. 2018. Improved estimation of leaf area index and leaf chlorophyll content of a potato crop using multi-angle spectral data – potential of unmanned aerial vehicle imagery. *International Journal of Applied Earth Observation and Geoinformation* 66(November 2017): 14–26.

Rueda, C. A., and A. F. Wrona. 2003. SAMS - Spectral Analysis and Management System.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. 2010. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications* 181(2): 259–270.

Savitzky, A., and M. J. E. Golay. 1964. Smoothing and differentiation by simplied least squares procedures. *Anal. Chem.* 36(8): 1627–1639.

Schaepman, M. E., S. L. Ustin, A. J. Plaza, T. H. Painter, J. Verrelst, and S. Liang. 2009. Earth system science related imaging spectroscopy-An assessment. *Remote Sensing of Environment* 113(SUPPL. 1): S123—-S137.

Schafer, R. W. 2011. What is a savitzky-golay filter? *IEEE Signal Processing Magazine* 28(4): 111–117.

Schlerf, M., C. Atzberger, and J. Hill. 2005. Remote sensing of forest biophysical variables using HyMap imaging spectrometer data. *Remote Sensing of Environment* 95(2): 177–194.

Schmidt, K. S., and A. K. Skidmore. 2003. Spectral discrimination of vegetation types in a coastal wetland. *Remote Sensing of Environment* 85(1): 92–108.

Schneider, F. D., R. Leiterer, F. Morsdorf, J. P. Gastellu-Etchegorry, N. Lauret, N. Pfeifer, and M. E. Schaepman. 2014. Simulating imaging spectrometer data: 3D forest modeling based on LiDAR and in situ data. *Remote Sensing of Environment* 152: 235–250.

Slaton, M. R., E. Raymond Hunt Jr., and W. K. Smith. 2001. Estimating near-infrared leaf reflectance from leaf structural characteristics. *American Journal of Botany* 88(2): 278–284.

Smith, G. M., and E. J. Milton. 1999. The use of the empirical line method to calibrate remotely sensed data to reflectance. *International Journal of Remote Sensing* 20(13): 2653–2662.

Sobol', I. M., S. Tarantola, D. Gatelli, S. S. Kucherenko, and W. Mauntz. 2007. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering and System Safety* 92(7): 957–960.

Turner, W., C. Rondinini, N. Pettorelli, B. Mora, A. K. Leidner, Z. Szantoi, G. Buchanan, S. Dech, J. Dwyer, M. Herold, L. P. Koh, P. Leimgruber, H. Taubenboeck, M. Wegmann, M. Wikelski, and C. Woodcock. 2015. Free and open-access satellite data are key to biodiversity conservation. *Biological Conservation* 182: 173–176. URL `http://dx.doi.org/10.1016/j.biocon.2014.11.048`

UNEP, FAO, and UNFF. 2009. *Vital Forest Graphics*. Nairobi: United Nations Environment Programmme (UNEP), United Nations Food and Agriculture Organisation (FAO), United Nations Food and Agriculture Organisation (UNFF), United Nations Forum on Forests.

Vaiphasa, C., S. Ongsomwang, T. Vaiphasa, and A. K. Skidmore. 2005. Tropical mangrove species discrimination using hyperspectral data: A laboratory study. *Estuarine, Coastal and Shelf Science* 65(1-2): 371–379.

847 Wang, Z., A. K. Skidmore, R. Darvishzadeh, and T. Wang. 2018. Mapping forest canopy

848   nitrogen content by inversion of coupled leaf-canopy radiative transfer models from airborne

849   hyperspectral imagery. *Agricultural and Forest Meteorology* 253-254(August 2017): 247–260.

850 Widlowski, J. L., C. Mio, M. Disney, J. Adams, I. Andredakis, C. Atzberger, J. Brennan,

851   L. Busetto, M. Chelle, G. Ceccherini, R. Colombo, J. F. Côté, A. Eenmäe, R. Essery, J. P.

852   Gastellu-Etchegorry, N. Gobron, E. Grau, V. Haverd, L. Homolová, H. Huang, L. Hunt,

853   H. Kobayashi, B. Koetz, A. Kuusk, J. Kuusk, M. Lang, P. E. Lewis, J. L. Lovell, Z. Malen-

854   ovský, M. Meroni, F. Morsdorf, M. Mõttus, W. Ni-Meister, B. Pinty, M. Rautiainen, M. Schlerf,

855   B. Somers, J. Stuckens, M. M. Verstraete, W. Yang, F. Zhao, and T. Zenone. 2015. The

856   fourth phase of the radiative transfer model intercomparison (RAMI) exercise: Actual canopy

857   scenarios and conformity testing. *Remote Sensing of Environment* 169: 418–437.

858 Wood, S. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation

859   of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*

860   73(1): 3–36.

861 Xiao, Y., W. Zhao, D. Zhou, and H. Gong. 2014. Sensitivity Analysis of Vegetation Reflectance

862   to Biochemical and Biophysical Variables at Leaf, Canopy, and Regional Scales. *IEEE Trans-*

863   *actions on Geoscience and Remote Sensing* 52(7): 4014–4024.

864 Yang, X., J. Tang, J. F. Mustard, J. Wu, K. Zhao, S. Serbin, and J.-E. Lee. 2016. Seasonal

865   variability of multiple leaf traits captured by leaf spectroscopy at two temperate deciduous

866   forests. *Remote Sensing of Environment* 179: 1–12.