UNIVERSITÁ DEGLI STUDI DEL MOLISE

Dipartimento di Bioscienze e Territorio

Dottorato di ricerca in

Scienze agro-forestali, delle tecnologie agro-industriali e del territorio rurale. I sistemi forestali.

(XXVIII ciclo)

S.S.D. AGR/05

TESI

ESTIMATING AND MAPPING FOREST STRUCTURE DIVERSITY USING AIRBORNE LASER SCANNING DATA

Dottorando:
Matteo Mura
(Matr. 149342)

Supervisor:
Prof. Gherardo Chirici

Coordinatore:
Prof. Marco Marchetti

Anno accademico 2014/2015

ABSTRACT

The topic of this doctoral thesis is the investigation of the most effective approaches and techniques that can be used to predict and map indicators of forest structural diversity, in a perspective of a more comprehensive assessment, management and monitoring of biodiversity in forest environments.

The thesis is subdivided in two main sections, made up of five different but interdependent and organically connected studies, represented by as many published peer-reviewed original research articles, hereafter reported in Roman numerals as Studies I-V.

The first section comprises the studies I-II-III. The contents of this section set the basis of methods and know-how that are subsequently used to estimate and map forest structure diversity in Studies IV and V.

Several international cooperation projects has been stipulated in order to cope with the issue of the constantly loss of biodiversity at global scale, and because of the relevant influence that forest structure has on biodiversity, forest structure diversity needs to be to assessed and monitored on large areas.

In Study I is demonstrated how this achievement can be efficiently tackled coupling ground data, such as those measured during forest inventory surveys, and remotely sensed data, in particular the ones derived from airborne laser scanning (ALS), which has proved to be a reliable source to characterize forest structure. The specific case of Study I presents how ALS data support the estimates of a common forest parameter, in such case forest above ground biomass (AGB), using field data gathered in a novel two-phase tessellation stratified sampling (TSS) design.

In order to be used as a valid source of information for planning conservation strategies, along with the estimation, a detailed map showing the spatial patterns of structural diversity is of great usefulness. Study II presents an extensive meta-analysis carried out during the doctoral time frame where is demonstrated that the non-parametric k-NN is, among the others, the most used and effective technique to spatial predict and map forest attributes, alone or combined together to form synthetic indices. This technique can be further improved implementing an optimization step aimed to set the k-NN parameters in order to achieve the best prediction performance possible. Study III demonstrates that, if an optimization phase is carried out before running the k-NN procedure, the performance in the predictions improved sensibly.

In the second and last section, the methods experimented in the first section are applied in two different research studies. Study IV describes the use of ALS data and ground data for the areal estimate of mean values of two forest structural diversity indices in a model-assisted framework. Along with the areal estimates, the study proposes the calculation of the confidence intervals of such estimates and the mapping of the investigated indices. Study V is framed as a methodological paper that takes a step further than Study IV, showing how, using the capability of an optimized k-NN techniques in predict simultaneously different parameters, is possible to map a more comprehensive structural diversity index (SDI) combining different forest structural diversity indices.

Keywords: biodiversity, forest structure diversity, remote sensing, airborne laser scanning, forest inventory, model-assisted inference, k-Nearest Neighbors, estimates, mapping, k-NN optimization, design-based inference, tessellation stratified sampling.

LIST OF ORIGINAL ARTICLES

In its overall structure, the thesis is composed as a sequence of original research articles that are logically interconnected each other, and where the papers in the first section (Studies I-II-III) are preparatory for the papers in the second section (Studies IV-V).

The papers are referred as Studies following Roman numerals.

I. Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sensing of Environment*, *174*, 1–9. doi:10.1016/j.rse.2015.11.010

II. Chirici G., Mura M., McInerney D., Py N., Tomppo E. O., Waser L. T., Travaglini D., McRoberts R. E. (2016). A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment*.

III. McRoberts R. E., Næsset E., Gobakken T., Domke G. M., Chirici G., Mura M., Chen Q. (in review). The benefits of optimizing nearest neighbor configurations for lidar-assisted estimation of forest volume and biomass. *Remote Sensing of Environment*.

IV. Mura M., McRoberts R. E., Chirici G., Marchetti M. (2015). Estimating and mapping forest structural diversity using airborne laser scanning data. *Remote Sensing of Environment*, *170*, 133–142. doi:10.1016/j.rse.2015.09.016

V. Mura M., McRoberts R. E., Chirici G., Marchetti M. (in review). Statistical inference for multiple-variable, forest structural diversity indices using airborne laser scanning data and the k-Nearest Neighbors technique. *Remote Sensing*.

INDEX

1. Background motivation

According to the United Nations Convention, biological diversity can be defined as "*the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems.*" (CBD, 1992).

Such a definition leads to extend the concept of biodiversity across different levels and thoughts.

Wilson (1987), Leveque (1994) and Gaston and Spiecer (2004) recognize three levels of assessment ranging from genetic, specific and ecosystem biodiversity. Whittaker (1972) differentiates biodiversity from ecosystem to landscape level, distinguishing in alpha (α), beta (β) and gamma (γ) biodiversity. Noss (1990) gives a more structured view of biodiversity including compositional, structural and functional aspects of ecosystem. Others authors focused in both functional and phylogenetic aspects (Tilman et al., 1997).

As emerged from an analysis of several zones and habitats classifications (Holdridge, 1947, 1967, Dinerstein et al. 1995), one half of the wide range of habitats classified are trees-dominated lands, thus forest and wooded lands are the richest ecosystems from a biological and genetic point of view.

The concept of biodiversity in forestry goes beyond to "just the trees", and it includes all others living organisms as a fundamental pieces of ecosystem functionality (Hunter, 1990), thus the conservation of forest habitats become crucial for many reasons (Hunter, 1999).

Forest biodiversity studies are often based on compositional and/or on 3-D structural parameters, and the debate on which one influence the most the diversity of communities rose up since the early studies (MacArthur and MacArthur, 1961).

For some communities, plant species composition seems to be a better predictor than structural variables for habitat assessment. For instance, it seems to work better for arthropod communities (Schaffers et al. 2008; Ter Braak and Schaffers, 2004), although other studies reported contrasting results and favoring structural parameters (Halaj et al., 2000). In others communities (e.g. birds) the structure plays a relevant role (Müller et al., 2010), but it may be influenced by plant species composition (Rosenzweig, 1995) so that the relative importance of structure and composition appears sensibly influenced by the habitat type of the study area (Fleishman and MacNally 2006).

Even if composition is still an important factor, studies on others taxa such as primates, reptiles, amphibians and arthropods (Halaj et al., 2000; McGraw, 1994; Salter et al., 1985; Shine et al., 2002; Welsh and Lind, 1996) confirm that the habitat structure is directly or indirectly related to the presence/absence of the investigated species. In their review, McElhinny et al. (2005), reported that the contribution of forest structure for biodiversity assessment is more relevant than the composition factor. They come out that a diversified stand structure is likely to have more niches, thus allowing to host more species for a better efficiency use of the resources available (McElhinny et al., 2005).

Also in the Quantitative Pan-European indicators describing the Criterion 4 `Biological Diversity` (MCPFE, 2011), the structural factor plays a major role when applied in a sustainable forest management contest (Puumalainen et al., 2003). Same findings are confirmed by Kuuluvainen (2009), Lähde et al. (1999)

and in other studies relying on structural parameters other than the ones of living and standing trees (e.g deadwood in Rondeaux and Sanchez, 2010).

In the last two decades LiDAR (Light Detection and Ranging) technology has shown to be a reliable and valuable remote sensing tool in gathering and assessing forest structure parameters (Lefsky et al., 2002; Lim et al., 2003; Zimble et al., 2003; Wulder et al., 2008), thus providing a potential great support in forest biodiversity studies given the important role played by forest structure in this field.

Bergen et al. (2009) exposed the importance of some remotely sensed variables and physical requirement of spaceborne active sensors data, as well as LiDAR, in detecting vegetation 3-D structure.

Diversity and abundance of birds are the most investigated issues in forest biodiversity studies (Tews et al., 2004). Vogeler et al. (2014), Clawges et al. (2008), Lesak et al. (2011), Goetz et al. (2007), Flashpoler et al. (2010), Vierling et al. (2013) and Müller et al. (2009, 2010) all report a good relationship between LiDAR-derived structural parameters and bird communities. Although the horizontal distribution of trees in the forest is in any case important, the vertical arrangement of canopy layers has shown to be stronger related to animal habitat, thus to their distribution, than the horizontal stem location (Vierling et al. 2008), and the capability of LiDAR to detect with high accuracy the vertical canopy profile pose such instrument as a key tool in supporting forest biodiversity studies.

LiDAR data has been used also to characterize habitat for other taxa such as terrestrial birds (Graf et al., 2009), beetles (Müller and Brandl, 2009), spiders (Vierling et al. 2011), and mammals like bats (Jung et al. 2012), deers (Ewald et al., 2014) and squirrels (Nelson at al., 2005).

Further, the potential availability of wall-to-wall LiDAR coverage allows also to use the geographic layers of LiDAR metrics to map habitats across wide areas. Several studies show the efficacy of habitat mapping using LiDAR-derived layers alone (Hyde et al., 2005; Martinuzzi et al., 2009) or in combination with other remotely sensed data (Swatantran et al. 2012; Hyde et al., 2006), among which LiDAR-derived layers are the best single predictors.

The structural information derived from LiDAR data has been proven to be effective not only in the prediction of animal habitat but also for the assessment of plant richness, composition and diversity (Lucas et al., 2010; Leutner et al., 2012; Simonson et al. 2012).

In this field the potentiality of optical imagery has been explored deeper than the use of LiDAR data. Hernandez-Stefanoni & Ponce-Hernandez (2004) mapped plant diversity indices (Uniformity Index ranging from 0.62-0.89) according the relationships among α and β diversity indices and the vegetation classes obtained by multi-spectral satellite image classification. Dogan & Dogan (2006) modeled and mapped Shannon–Wiener, Simpson, and number of species (NS) indices in Turkish forest ecosystems, and they found that Wiener and NS models could be successful to reveal the richness aspect of species diversity, while Simpson model might be acceptable to delineate the evenness aspect indicating single dominant land cover types. Rocchini (2007) used species richness (α-diversity) as a proxy of diversity relying on the correlation with image heterogeneity (Quickbird: r=0.69; Aster: r=0.43; Landsat ETM+: r=0.67; resampled 60 m Landsat ETM: r=0.69), which has been proven to be dependent on the geometrical resolution of the image. Other studies underlined the great support of hyperspectral imagery in mapping α-diversity indices (Shannon Index)

(Oldelan et al. 2010; Vaglio Laurin et al., 2014) and the scale-dependent relationship between spectral diversity and species diversity (Rocchini et al., 2014).

Just a few studies used LiDAR data, alone or in combination with optical imagery, to study biodiversity indices.

Leutner et al. (2012) investigated the spatial patterns of vascular plant community composition and α-diversity in a temperate montane forest in Germany and compared the predictive power of LiDAR and hyperspectral datasets, alone and combined together, and used the final models for spatial predictions (species richness accuracy of R^2 = 0.26 to 0.55, depending on the forest type). Others authors demonstrated the utility of LiDAR-derived forest structure maps for studies on floristic diversity indices (Alberti et al., 2013).

As pointed out from the literature review, there is a lack of studies that link indices of forest structural diversity to LiDAR metrics. Such a gap in the research could be driven by two main reasons: (i) when the purpose is to predict biodiversity indices for plant communities, although the above exposed importance of structural aspects on biodiversity, the role played by the compositional factor seems to be still predominant, and given the fact that it is better predict through multispectral imagery, LiDAR data in this contest has been so far underused; (ii) on the other hand, when LiDAR data were available, its capability to depict forest structure has been linked mainly to animal habitat rather than forest structural diversity itself, referring in part to the previous point.

If the current literature, as seen, is plenty of studies which related LiDAR products to habitat mapping and assessment, only Leutner et al. (2012) used such a products to spatial predict biodiversity indices, and in their attempt they focused just on plant community composition and α-diversity, leaving other interesting indices out of the game (Neumann & Starlinger, 2001).

Maps of diversity indices would provide support to managers who need spatially-explicit information concerning patterns of biodiversity in order to plan adequate conservation strategies. Mapping forest structure, habitat and diversity indices over large areas allows assessing biological diversity in remote and impervious areas that cannot be reached by the field crews.

An interesting but non exhaustive set of practical application could include e.g. (i) the description of the status and trends of the components of biodiversity, making possible to detect changes in habitats, ecosystems and biomes, thus in animal and plant species composition, and act as consequence, (ii) identify trends in invasive alien species and apply protection strategies to the threatened native animal and plant species or even (iii) locate spots of high biodiversity values that need to be preserved by *ad hoc* management strategies.

A problem with the map products derived from remotely sensed data is that often their performances are assessed by error matrices and measures such as overall accuracy, users' and producers' accuracies, Kappa index correlation and coefficients of correlation (e.g. R or R^2), lacking then in a scientific assessment of inference (McRoberts, 2011). In the aforementioned studies of mapping biodiversity indices by LiDAR-derived products, there are not attempts to assess the scientific inference of the maps produced.

From the exposed literature review can be drawn several consideration. Firstly, forest diversity should be seen not just as an aggregate of woody species based only on plant species composition, but its structure must be included as well. Secondly, giving strength to the first statement, forest structure has shown to have a relevant influence for animal habitat and plant communities' assessment. Thirdly, the proven capability of LiDAR to depict the structural components of a forest, pose this remote sensing technique as a key tool in the research of forest biological diversity. Fourthly and lastly, in order to base management and planning strategies on reliable data, a scientific-based uncertainty for the maps of diversity indices must be developed.

These considerations, driven by the exposed literature review, uncover two gaps in the current research on the topic, that are the aims to which I progressively addressed in this PhD work: i) spatially predict a set of biodiversity indices relying on the relationship between LiDAR-derived metrics of forest structure and diversity indices; ii) assess the accuracy of the derived maps accounting for their uncertainty from a scientific inference point of view.

1.1. References

Alberti G., Boscutti F., Pirotti F., Bertacco C., De Simon G., Sigura M., Cazorzi F., Bonfanti P., (2013). A LiDAR-based approach for a multi-purpose characterization of Alpine forests: an Italian case study. iForest 6: 156-168 [online 2013-04-08] URL: http://www.sisef.it/iforest/contents/?id=ifor0876-006.

Bergen, K. M., Goetz S. J., Dubayah R. O., Henebry G. M., Hunsaker C. T., Imhoff M. L., Nelson R. F., Parker G. G., Radeloff V. C. (2009). Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions, J. Geophys. Res., 114, G00E06.

Clawges R, Vierling K, Vierling L, Rowell E (2008). The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. Remote Sensing of Environment 112:2064–2073.

Dinerstein, E., Olson, D. M., Graham, D. J., Webster, A. V., Primm, S. A., Bookbinder, M. P., & Ledec, G., (1995). Una evaluación del estado de conservación de las ecoregiones terrestres de América Latina y el Caribe. Publicado en colaboración con el Fondo Mundial para la Naturaleza. Washington, DC: Banco Mundial.

Dogan H. M., Dogan M., (2006). A new approach to diversity indices – modeling and mapping plant biodiversity of Nallihan (A3-Ankara/Turkey) forest ecosystem in frame of geographic information systems. Biodiversity and Conservation, 15:855–878.

Ewald, M.; Dupke, C.; Heurich, M.; Müller, J.; Reineking, B., (2014). LiDAR Remote Sensing of Forest Structure and GPS Telemetry Data Provide Insights on Winter Habitat Selection of European Roe Deer. Forests 2014, 5, 1374-1390.

Fleishman E., Mac Nally R., (2006). Patterns of spatial autocorrelation of assemblages of birds, floristics, physiognomy, and primary productivity in the central Great Basin, USA. Divers Distrib 12:236–243.

Flashpoler DJ, Giardina CP, Asner GP, Hart P, Price JT, Lyons CK, Castaneda X (2010). Long-term effects of fragmentation and fragment properties on bird species richness in Hawaiian forests. Biological Conservation 143:280–288.

Gaston, K.J., & Spicer, J.I. (2004). Biodiversity: An introduction. Blackwell Publishing. 2nd ed.

Goetz S, Steinberg D, Dubayah R, Blair B (2007). Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. Remote Sensing of Environment 108:254–263.

Graf R. F., Mathys L., Bollmann K., (2009). Habitat assessment for forest dwelling species using LiDAR remote sensing: Capercaillie in the Alps. Forest Ecology and management, Volume 257, Issue 1, 20 January 2009, Pages 160-167.

Halaj, J., D. W. Ross, and A. R. Moldenke (2000), Importance of habitat structure to the arthropod food-web in Douglas-fir canopies, Oikos, 90, 139–152.

Hernandez-Stefanoni J. L., Ponce-Hernandez R., (2004). Mapping the spatial distribution of plant diversity indices in a tropical forest using multi-spectral satellite image classification and field measurements. Biodiversity & Conservation, December 2004, Volume 13, Issue 14, pp 2599-2621.

Holdridge, L. R., (1947). Determination of world plant formation from simple climatic data. Science, 105, 367–368.

Holdridge, L. R., (1967). Life zone ecology. San Jose: Tropical Science Center.

Hunter, M. L., Jr. (1990). Wildlife, forests, and forestry: Principles of managing forests for biological diversity (p. 370). Englewood Cliffs: Prentice-Hall.

Hunter, M. L., Jr. (Ed.). (1999). Maintaining biodiversity in forest ecosystem (p. 698). Cambridge: Cambridge University Press.

Hyde P., Dubayah R., Peterson B., Blair J.B., Hofton M., Hunsaker C., Knox R., Walker W.(2005). Mapping forest structure for wildlife habitat analysis using waveform lidar: validation of montane ecosystems. Remote Sensing of Environment, 96: 427–37.

Hyde P., Dubayah R., Walker W., Blair B. J., Hofton M., Hunsaker C., (2006). Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM plus, Quickbird) synergy. Remote Sensing of Environment 102: 63–73.

Jung K, Kaiser S, Böhm S, Nieschulze J, Kalko EKV (2012) Moving in three dimensions: effects of structural complexity on occurrence and activity of insectivorous bats in managed forest stands. J Appl Ecol 49:523–531.

Lassau SA, Hochuli DF (2008) Testing predictions of beetle community patterns derived empiri- cally using remote sensing. Divers Distrib 14:138–147

Lassau SA, Cassis G, Flemons PKJ, Wilkie L, Hochuli DF (2005a) Using high-resolution multi- spectral imagery to estimate habitat complexity in open-canopy forests: can we predict ant community patterns? Ecography 28:495–504

Lassau SA, Hochuli DF, Cassis G, Reid CAM (2005b) Effects of habitat complexity on forest beetle diversity: do functional groups respond consistently? Divers Distrib 11:73–82.

Lähde E., Laiho O, Norokorpi Y., Saksa T.,(1999). Stand structure as the basis of diversity index, Forest Ecology and Management, Volume 115, Issues 2–3, 22 March 1999, Pages 213-220.

Lefsky M. A., Cohen W. B., Parker G. G., Harding D. J., (2002). Lidar Remote Sensing for Ecosystems Studies. BioScience 52(1):19-30. 2002.

Lesak AA, Radeloff VC, Hawbaker TJ, Pidgeon AM, Gobakken T, Contrucci K (2011) Modeling forest song bird species richness using LiDAR-derived measures of forest structure. Remote Sensing of Environment 115:2823–2835.

Leutner B.F., Reineking B., Müller J., Bachmann M., Beierkuhnlein C., Dech S., Wegmann M. (2012). Modelling forest α-diversity and floristic composition – on the added value of LiDAR plushyperspectral remote sensing. Remote Sensing 4:2818–2845.

Leveque, C. (1994). Environement et diversité du vivant. Pocket Sciences, Collection Explora. pp. 127.

Lim K., Treitz P., Wulder M., St-Onge B., Flood M. (2003). LiDAR remote sensing of forest structure. Progress in Physical Geography, 27:1, pp. 88-106.

Lucas K.L., Raber G.T., Carter G.A. (2010). Estimating vascular plant species richness of Horn Island, Mississippi using small-footprint airborne LiDAR. J Appl Remote Sensing 4:033545

MacArthur RH, MacArthur J (1961) On bird species diversity. Ecology 42:594–598.

Martinuzzi S., Vierling L. A., Gould W. A., Falkowski M. J., Evans J. S., Hudak A. T., Vierling K. T., (2009). Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability, Remote Sensing of Environment, Volume 113, Issue 12, 15 December 2009, Pages 2533-2546.

McGraw, S. (1994), Census, habitat preference, and polyspecific associa-tions of six monkeys in the Lomako Forest, Zaire, Am. J. Primatol., 34, 295–307.

MCPFE (2011). State of Europe's Forest: Status & Trends in Sustainable Forest Management in Europe.

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

McRoberts R.E., (2011). Satellite image-based maps: Scientific inference or pretty pictures?, Remote Sensing of Environment, Volume 115, Issue 2, 15 February 2011, Pages 715-724.

Müller J, Brandl R (2009) Assessing biodiversity by remote sensing and ground survey in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. J Appl Ecol 46:897–905.

Müller J, Moning C, Bässler C, Heurich M, Brandl R (2009) Using airborne laser scanning to model potential abundance and assemblages of forest passerines. Basic Appl Ecol 10:671–681.

Müller J, Stadler J, Brandl R (2010). Composition versus physiognomy of vegetation as predictors of bird assemblages: the role of lidar. Remote Sens Environ 114:490–495.

Nelson R, Keller C, Ratnaswamy M. (2005). Locating and estimating the extent of Delmarva fox squirrel habitat using an airborne LiDAR profiler. Remote Sensing of Environment 96: 292–301.

Neumann M, Sterlinger F., (2001). The significance of different indices for stand structure and diversity in forests. Forest Ecology and Management 145 (2001) 91-106.

Noss, R. F. (1990). Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology, 4(4), 355–364.

Oldeland J., Wesuls D., Rocchini D., Schmidt M., Jürgens N., (2010). Does using species abundance data improve estimates of species diversity from remotely sensed spectral heterogeneity?, Ecological Indicators, Volume 10, Issue 2, March 2010, Pages 390-396.

Puumalainen J, Kennedy P, Folving S., (2003). Monitoring forest biodiversity: a European perspective with reference to temperate and boreal forest zone. J Environ Manage. 2003 Jan;67(1):5-14.

Rocchini D. (2007). Effects of spatial and spectral resolution in estimating ecosystem α-diversity by satellite imagery, Remote Sensing of Environment, Volume 111, Issue 4, 28 December 2007, Pages 423-434.

Rocchini D., Dadalt L., Delucchi L., Neteler1 M., Palmer M.W., (2014). Disentangling the role of remotely sensed spectral heterogeneity as a proxy for North American plant species richness. Community Ecology 15(1): 37-43, 2014.

Rondeux J., Sanchez C., (2010). Review of indicators and field methods for monitoring biodiversity within national forest inventories. Core variable: deadwood. Environ Monitoring Assessment 2010 May;164(1-4):617-30.

Rosenzweig ML (1995) Species diversity in space and time. Cambridge University Press, Cambridge/New York.

Salter, R. E., N. A. Mackenzie, N. Nightingale, K. M. Aken, and P. K. P. Chai (1985), Habitat use ranging behavior and food habits of the Proboscis monkey Nasalis-larvatus in Sarawak, Primates, 26, 436 – 451

Schaffers AP, Raemakers IP, Sýkora KV, ter Braak CJF (2008) Arthropod assemblages are best predicted by plant species composition. Ecology 89:782–794.

Simonson W.D., Allen H.D., Coomes D.A. (2012). Use of an airborne Lidar system to model plant species composition and diversity of Mediterranean oak forests. Conservation Biology 26:840–850.

Shine, R., E. G. Barrott, and M. J. Elphick (2002), Some like it hot: Effects of forest clearing on nest temperatures of montane reptiles, Ecology, 83, 2808 – 2815.

Swatantran A., Dubayah R., Goetz S., Hofton M., Betts M.G., Sun M., Simard M., Holmes R., (2012). Mapping migratory bird prevalence using remote sensing data fusion. PLoS ONE Volume 7, Issue 1, 3 January 2012.

Ter Braak CJF, Schaffers AP (2004) Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology 85(3):834–846.

Tilman D., Knops J., Weldin D., Reich P., Ritchie M., Sieman E. (1997). The influence of functional diversity and composition on ecosystem processes. Science 277:1300–1302.

United Nations Convention on Biological Diversity (CBD), 1992, Rio de Janeiro.

Uuttera, J., Tokola, T., Maltamo, M., 2000. Differences in the structure of primary and managed forests in East Kalimantan Indonesia. Forest Ecol. Manage. 129, 63–74.

Vaglio Laurin G., Chan J. , Chen Q., Lindsell J.A., Coomes D.A., Guerriero L., Del Frate F., Miglietta F., Valentini R., (2014). Biodiversity Mapping in a Tropical West African Forest with Airborne Hyperspectral Data. PLoS ONE 9(6): e97910. doi:10.1371/journal.pone.0097910.

Vierling KT, Vierling LA, Gould WA, Martinuzzi S, Clawges RM (2008) Lidar: shedding new light on habitat characterization and modeling. Front Ecol Environ 6:90–98.

Vierling KT, Bässler C, Brandl B, Vierling LA, Weiß I, Müller I (2011) Spinning a laser web: predicting spider distributions using lidar. Ecol Appl 21:577–588.

Vierling LA, Vierling KT, Adam P, Hudak AT (2013) Using Satellite and Airborne LiDAR to Model Woodpecker Habitat Occupancy at the Landscape Scale. PLoS ONE 8(12).

Vogeler J. C., Hudak A. T., Vierling L. A., Evans J., Green P., Vierling K. T., (2014). Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. Remote Sensing of Environment, Volume 147, 5 May 2014, Pages 13-22.

Welsh, H. H., and A. J. Lind (1996), Habitat correlates of the southern torrent salamander, Rhyacotriton variegatus (Caudata: Rhyacotritonidae), in northwestern California, J. Herpetol., 30, 385–398.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. Taxon, 21, 213–251.

Wilson, E. O. (1987). An urgent need to map biodiversity. Scientist 1:11.

Wulder, M.A.; Bater, C.W.; Coops, N.C.; Hilker, T.; White, J.C. (2008). The role of LiDAR in sustainable forest management. The Forestry Chronicle 84(6): 807-826.

Zimble D. A., Evans D. L., Carlson G. C., Parker R. C., Grado S. C., Gerard P. D., (2003). Characterizing vertical forest structure using small-footprint airborne LiDAR. Remote Sensing of Environment, Volume 87, Issues 2–3, 15 October 2003, Pages 171-182.

2. Section I (Studies I-II-III)

This first section comprises three studies, namely Studies I-II-III, published as original research articles (Study I and Study III) and as a review article (Study II), to which the PhD candidate has actively participated :

I. Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sensing of Environment*, *174*, 1–9. doi:10.1016/j.rse.2015.11.010

II. Chirici G., Mura M., McInerney D., Py N., Tomppo E. O., Waser L. T., Travaglini D., McRoberts R. E. (2016). A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment*.

III. McRoberts R. E., Næsset E., Gobakken T., Domke G. M., Chirici G., Mura M., Chen Q. (in review). The benefits of optimizing nearest neighbor configurations for lidar-assisted estimation of forest volume and biomass. *Remote Sensing of Environment*.

The contents of this section reflects the issues addressed by the three organically organized studies.

The aim of the studies in this section, thus its contents, is to set the basis of methods and know-how that are subsequently used to estimate and map forest structure diversity in Studies IV and V of Section II.

Study I investigates how to efficiently coupling ground data, such as those measured during forest inventory surveys, and remotely sensed data, in particular the ones derived from airborne laser scanning (ALS), for forest parameter estimates and mapping. The specific case of Study I presents how ALS data support the estimates of a common forest parameter, in such case forest above ground biomass (AGB), using field data gathered in a novel two-phase tessellation stratified sampling (TSS) design.

Study II presents an extensive meta-analysis carried out during the doctoral time frame where is demonstrated that the non-parametric k-NN is, among the others, the most used and effective technique to spatial predict and map forest attributes, alone or combined together to form synthetic indices. This technique can be further improved implementing an optimization step aimed to set the k-NN parameters in order to achieve the best prediction performance possible.

Study III demonstrates that, if an optimization phase is carried out before running the k-NN procedure, the performance in the predictions improved sensibly.

2.1. Study I

From:

[1]Chirici, G., [2]McRoberts, R. E., [3]Fattorini, L., [4]Mura, M., & [4]Marchetti, M. (2016). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. *Remote Sensing of Environment*, *174*, 1–9. doi:10.1016/j.rse.2015.11.010

[1] Department of Agricultural, Food and Forestry Systems, Università degli Studi di Firenze, Italy

[2] Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota USA

[3] Department of Economics and Statistics, Università di Siena, Italy

[4] Dipartimento di Bioscienze e Territorio, Università degli Studi del Molise, Pesche (IS), Italy

Abstract

Among the forestry-related applications for which airborne laser scanning (ALS) data have been shown to be beneficial, forest inventory has been investigated as much if not more than other applications. Metrics extracted from ALS data for spatial units such as plots and grid cells are typically of two forms: echo-based metrics derived directly from the three-dimensional distribution of the point cloud data and metrics derived from a canopy height model (CHM). For both cases, a large number of metrics can be calculated and used to construct parametric and non-parametric models to predict forest variables.

We compared model-assisted estimates of total forest aboveground biomass (AGB) obtained using echo-based and CHM-based height metrics with two prediction methods: (i) a parametric linear model, and (ii) the non-parametric k-Nearest Neighbors (k-NN) technique. Model-Assisted (MA) estimators were used with sample data obtained using a two-phase, tessellation stratified sampling (TSS) framework to estimate population parameters. The study was conducted in Regione Molise in central Italy.

For the four combination of metrics and prediction technique, estimates of total biomass were similar, in the range 1.96-2.1 million t, with standard error estimates that were also similar, in the range 0.20-0.21 t. Thus, the CHM-based metrics produced AGB estimates that were similar to and as accurate as those for the echo-based metrics, regardless of whether the parametric or the non-parametric prediction method was used. Additionally, the proposed MA estimator was more accurate than the estimator that did not use auxiliary data.

Keywords: airborne laser scanning metrics, forest biomass, model-assisted estimator, k-Nearest Neighbors, tessellation stratified sampling

2.1.1. Introduction

The utility of airborne laser scanner (ALS) data for contributing to and enhancing forestry-related applications is nowadays both indisputable and well-documented (Corona et al., 2012; Lim et al., 2003a; Maltamo et al., 2014; Montaghi et al., 2013; Wulder et al., 2008). Further, when estimation of aboveground biomass (AGB) is the main goal, the predictive power of ALS data as auxiliary information has been shown to be more effective with respect to increasing the accuracy of estimates than many other sources of remotely sensed data (Zolkos et al., 2013). This capability can be attributed to strong correlations between forest AGB and forest height variables derived from ALS metrics (Lefsky at al., 2014). These correlations, in turn, are due to the nature of the ALS data itself, i.e., a cloud of geo-referenced 3-dimensional points characterized as returns or echoes. Multiple modeling techniques are used to predict AGB using ALS metrics of which the most common are the parametric regression (Næsset & Gobbaken, 2008; McRoberts et al., 2013) and the non-parametric k-Nearest Neighbors (k-NN) technique (McRoberts et al., 2015).

Independently of the prediction criterion, two main approaches are used to extract metrics from ALS data for use as covariates when constructing AGB prediction models: (i) echo-based metrics, and (ii) canopy height model (CHM)- based metrics. Echo-based metrics are descriptive statistics directly extracted from the ALS point cloud. This approach requires the availability of raw ALS data, often provided in the form of geo-referenced vector points in 3-dimensional space. If the ALS data are not acquired specifically for forestry applications, raw ALS data may not be available in which case echo-based metrics cannot be calculated (Montaghi et al., 2013). In this case, interpolations of the raw ALS echo heights are often available in the form of two gridded raster layers: ground height for each pixel characterized as a digital elevation model (DEM) and absolute height of objects above ground characterized as a digital surface model (DSM). For forestry applications, a CHM consisting of top canopy height for each pixel is constructed from the difference between the DSM and DEM (Kraus & Pfeifer, 1998).

To construct parametric or non-parametric prediction models, AGB observations are required and are typically obtained from field plots. Corresponding plot-level ALS metrics extracted from either the echo heights or from the CHM are used as predictors for estimating AGB. Only a limited number of metrics such as minimum, maximum, average and standard deviation values of height are typically extracted from a CHM grid (Barbati et al., 2009). Many more metrics can be extracted from distributions of echo heights including the minimum, mean, maximum, standard deviation, skewness, kurtosis, and coefficient of variation; the 0, 10, …, 90 distribution percentiles; canopy density metrics calculated as the proportions of first echo heights above the 0, 10, …, 90 quantiles of the first echo height distributions (Næsset, 2002; Lim et al, 2003b; Næsset, 2004a; Næsset, 2004b; Næsset & Gobbaken, 2005; Næsset & Gobbaken, 2008; Hawbaker et al, 2010; Gobbaken at al., 2012); and canopy relief ratio (Evans et al., 2009).

Although large numbers of both echo-based and CHM-based metrics can be derived from ALS data, modeling applications often require only a few of them. Næsset (2002) found that only 2-4 metrics from a set of 46 metrics were required to obtain for volume models for both young and mature forests on both good and poor sites. The selected metrics were mean and maximum echo heights, several canopy height percentiles and metrics related to canopy density; similar metrics were selected by Næsset (2004a, 2004b). Næsset and Gobbaken (2008) found that canopy height percentiles produced the greatest accuracies for above- and below-

ground biomass models in boreal forests. Two ALS metrics, the 90th percentile of laser canopy height and canopy density, plus variables representing different areas, age classes, and tree species composition, produced . Gobbaken et al. (2012) reported that mean height, percentiles of heights from first and/or last pulses and canopy density metrics, coupled with altitude information, produced for volume models for young/mature and productive/non-productive forests.

Corona et al. (2008) obtained R2=0.78 when predicting plot-level forest volume using the sum of the CHM heights raised to a power for a temperate broadleaved forest; Barbati et al. (2009) used the same methodology for a coastal Mediterranean pine forest and obtained R2=0.88. Fusco et al. (2008) obtained R2=0.76 for circular plots in a broadleaved forest. Finally, for Norway spruce in an alpine environment, Floris et al. (2010) constructed regression models for predicting standing volume in sample plots using CHM metrics and found that CHM mean height, excluding pixels with height less than 2 m, produced R2=0.94.

In recent years, an interesting debate has emerged regarding the advantages and disadvantages of the two kinds of ALS-based metrics. Although both kinds of metrics are commonly used, no peer review reports of direct comparisons for predicting volume in the same study area are known, other than possibly Gaulton and Malthus (2010) who compared them for detecting canopy gaps.

Model-assisted (MA) estimation exploits auxiliary information to augment ground data for purposes of enhancing estimation (Särndal et al., 1992). Although many sources of auxiliary information can be used (e.g., Corona et al., 2009), remotely-sensed data have been found to be a particularly useful for forestry applications.

A broad range of sampling designs have been used with MA estimators for exploiting ALS auxiliary information. Corona and Fattorini (2008) proposed MA estimation of forest standing volume using CHM height as auxiliary information when field plots are randomly and independently located. Ene et al. (2012) and McRoberts et al. (2013) applied a MA estimator using a systematic sampling design. Gregoire et al. (2011) and Gobakken et al. (2012) developed a MA regression estimator of AGB for a two-stage sampling design, and Næsset et al. (2013) tested a MA estimator for both two-phase and two-stage sampling designs. Finally, Saarela et al. (2015) described the use of MA estimators with a systematic cluster sampling design.

In designing schemes for sampling large areas, such as for forest inventory, limited financial resources for ground sampling suggest that the smallest but most representative sample should be selected. Uniform random sampling (URS), the random and independent selection of points on a continuous surface, is the simplest scheme to locate ground plots (Fattorini, 2015). However, like its finite population analogue, simple random sampling with replacement (SRSWR), URS may lead to uneven coverage of the study area which, in turn, makes the sample less representative of the study area. To avoid these drawbacks, spatially stratified schemes can be adopted (Fattorini, 2015). With tessellation stratified sampling (TSS), the study area A is covered by a region $R \supset A$ of size R constisting of N non-overlapping regular polygons R1, …, RN of equal size such that $Ri \cap A \neq \emptyset$ for all i = 1, …, N. Then, for each polygon Ri, a point is randomly selected from within the polygon (Fattorini, 2015). In a two-phase application of a TSS for forest inventory, the objective of the first phase is to uniformly spread points throughout the study area and classify them with respect to forest/non-forest. In the second-phase, a sub-sample of points within the forest class is selected using a finite-population

sampling scheme and then visited by the field crews. With this sampling design, estimators that both do not and do use auxiliary information extracted from ALS data are posssible, although the latter have not been investigated. For this study, a MA estimator using ALS-based metrics in a two-phase TSS was investigated.

The objectives of the study were threefold: (i) to compare large area estimates of total AGB obtained using echo-based and CHM-based metrics as predictor variables, (ii) to compare estimates obtained using parametric linear regression and the non-parametric k-Nearest Neighbors (k-NN) technique, and (iii) to evaluate the utility of MA estimators in a two-phase TSS framework.

We adopted the k-NN approach because it has emerged as very popular for spatial estimation of NFI variables using remotely sensed auxiliary data (Chirici et al., 2016). Despite the large number of other reported methods, including machine learning approaches, k-NN is the only approach that has been adopted by NFI programs for large area operational applications: in Finland (Tomppo, 1990), in Sweden (SLU Forest Map, 2013), in Canada (Beaudoin et al. 2014) and in the USA (Wilson et al., 2013).

2.1.2. Materials and Methods

Study area

The study area is located in the southwestern part of Molise Region in central Italy and includes 36,360 ha (Figure 1). Based on a local forest map (Chirici et al., 2011), forests cover more than 20,518 ha and comprise approximately 56% of the study area, where the COST (European Cooperation in Science and Technology) Action E43 definition of forest was used (Vidal et al., 2008). The forested area is dominated mainly by deciduous oaks (*Quercus cerris, Quercus pubescens*) covering approximately 60% of the forest area, hop hornbeam (*Ostrya carpinifolia*) covering approximately 18% and beech (*Fagus sylvatica*) covering approximately 9%. The oak and hop hornbeam forests are mainly privately-owned and are managed in a coppice with standards system. Rotation ages are usually between 18 and 25 years, with most clear-cuts 1–2 ha wide and 100-200 standards/ha. Conversely, most of the beech forests are managed with the shelterwood system or are unmanaged.

Field data

TSS was carried out in the study area. The area was tessellated into 437 hexagons, each with an area of 1 km^2. In the first phase, a point was randomly selected in each hexagon and classified as “forest” or “non-forest” based on interpretation of high-resolution aerial ortho-photography. Of the 437 points, 197 were classified as “forest” (Figure 1). In the second phase, 62 points were selected from the 197 “forest” points (sampling rate ≈ 30%) by means of simple random sampling without replacement (SRSWOR) and surveyed in the field during 2009–2011.

Plots were configured as two concentric circular plots with radii of 4 and 13 m. In the 4-m radius plot, all trees with diameter at breast-height (DBH, 1.3 m) of at least 2.5 cm were measured, and in the 13-m radius plot, all trees with DBH of at least 9.5 cm were measured. Heights (H) were measured for a sub-sample of plot trees and estimated for the remaining trees using a model of the DBH-H relationship for the

trees with measured H. In total, 4,745 trees were measured on the 62 forest plots in the second-phase sample.

For each tree, volume was estimated using DBH and H with the double-entry tables constructed for the first Italian National Forest Inventory (Castellani et al., 1984) and then scaled to plot-level and per unit area (ha) values. AGB (t/ha) was calculated as,

$$AGB = GS \times BEF \times WBD, \quad (1)$$

where *GS* is the growing stock (m^3/ha), *BEF* is the biomass expansion factor, and *WBD* is the wood basic density (t/m^3). Values of *BEF* and *WBD* were extracted from Federici et al. (2008). Although the tree- and plot-level volumes are estimates, their uncertainties were considered negligible and ignored for this study.

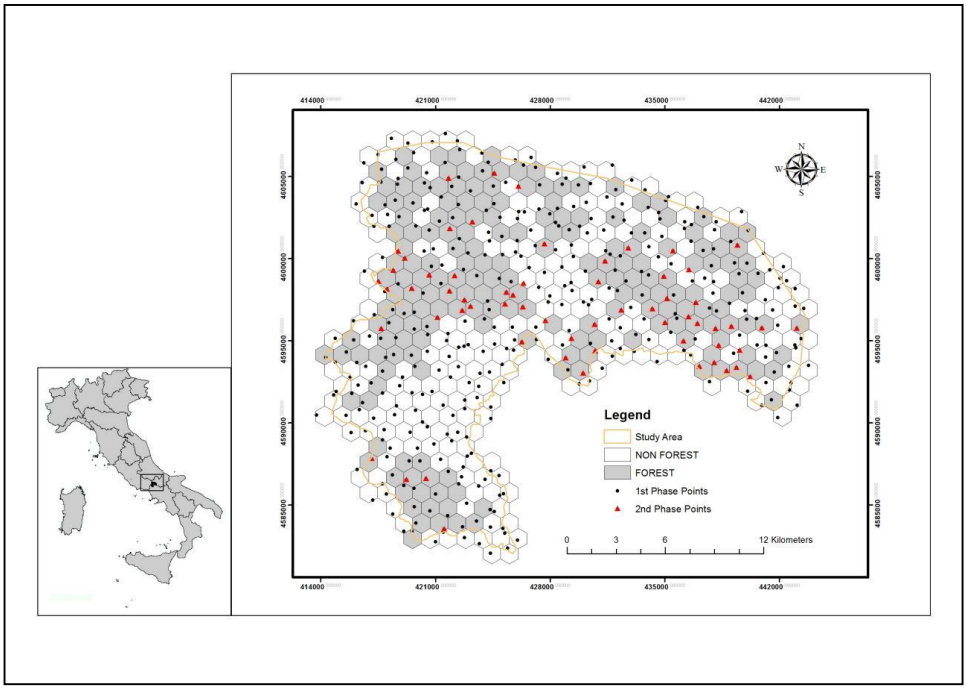


Figure 1 Sampling design and plots locations in the study area.

Airborne Laser Scanning (ALS) data

ALS data were acquired under leaf-on canopy conditions in June 2010. A fixed-wing aircraft PartenaviaP68 was used. The LiDAR instrument was an Optech Gemini LiDAR, a two-return range detection system that records a maximum of two echoes per laser pulse. The sensor was set with a maximum scan angle of 15° and a pulse frequency of 70 KHz, resulting in an average density of 1.5 pulses/m^2. The combination of the 1.5 pulses/m^2 density and the 169-m^2 plots and grid cells yields approximately 250 pulses/plot which is adequate for deriving percentiles of the echo distributions (Vauhkonen et al., 2014, Section 1.3.2.1).

Common procedures for pre-processing of ALS data included removal of outliers, ground/non-ground classification, and computation of normalized height. Firstly, air points that were clearly higher than the median elevation of surrounding points and isolated points with few neighbors resulting from sensor errors

or backscatter by flying objects were removed. Subsequently, a ground surface model was constructed by classifying ground points on the basis of the adaptive TIN model algorithm (Axelsson, 2000) and used to calculate the relative height above ground for each backscattered echo. Subsequently, a 1-m × 1-m resolution CHM in the form of a raster layer was constructed.

For each sample plot measured in the field, a set of 22 echo-based ALS height and density metrics, and seven metrics derived from the CHM were extracted and used as covariates for constructing prediction models. The echo-based metrics were canopy cover (*cov*) calculated as the proportion of first echoes above 1.30 m on all first returns. Canopy density metrics calculated as the proportions of all echoes at heights greater than 1.30 m (*dns*), the proportion (d_{00}) and the count (c_{00}) of echoes between 1.30 m and 10 m. The reference height of 1.30 m was chosen because, according to the Italian national forest inventory (INFC 2005), plants taller than this threshold are no longer considered regeneration. Canopy height metrics were the percentiles of the canopy height distribution (p_{10}, p_{20},..., p_{90}) and height summary statistics such as minimum (H_{min}), maximum (H_{max}), average (H_{avg}), standard deviation (H_{std}), coefficient of variability (H_{cv}), skewness (H_{ske}) and kurtosis (H_{kur}). Further, canopy relief ratio (*CRR*), a quantitative measure of the relative shape of the canopy, describing the proportion of all returns above the mean value of echoes heights was calculated as,

$$CRR = \frac{H_{avg} - H_{min}}{H_{max} - H_{min}} \quad (2)$$

CRR ranges between 0 and 1 and reflects the degree to which outer canopy surfaces are in the upper or lower half of the height range (Parker & Russ, 2004).

The seven CHM-based metrics extracted for each sample plot were the minimum (H_{min}), maximum (H_{max}), average (H_{avg}), standard deviation (H_{std}), coefficient of variability (H_{cv}), range (H_{max}-H_{min}) and *CRR* of the 1-m × 1-m pixel values that were inside or intersected by the boundary of the 13-m radius plot. The same echo-based and CHM-based metrics were calculated for the 23-m × 23-m forest pixels that tessellated the study area. This pixel size was chosen to mimic the field sample plot size as suggested by Magnussen and Boudewyn (1998) and Næsset (2002).

Model predictions

Two approaches for representing the relationship between the response and predictor variables were used: a linear regression model, and the k-NN technique. A linear model of the relationship was formulated as,

$$AGB_i = \beta_0 + \beta_1 \cdot x_{ij} + \cdots + \beta_p \cdot x_{pi} + \varepsilon_i \quad (3),$$

where *i* indexes plots, *j* indexes predictor variables, x_{ji} is the observation of the j^{th} predictor variable, the β_js are parameters to be estimated, and ε_i is a residual term. Once the parameter estimates $\tilde{\beta}_j$ s are obtained from the reference set, the prediction for the i^{th} plot is calculated as

$$\tilde{AGB}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{ij} + \cdots + \tilde{\beta}_p x_{pi} \quad (4)$$

Using k-NN terminology, the predictor variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the set of plots selected in the second phase, for which observations of both response and feature variables are available is designated the *reference set*; and the set of plots for which predictions are desired is designated the *target set*. For the k-NN technique, the prediction for the i^{th} plot is calculated as,

$$\tilde{AGB}_i = \sum_{l=1}^{k} w_{il} AGB_l^i \quad (5)$$

where $\left\{AGB_l^i, l = 1, 2, ..., k\right\}$ is the set of observations for the *k* reference plots that are most similar or nearest to the i^{th} target plot in feature space with respect to a distance metric, and w_{il} is the weight assigned to the l^{th} nearest neighbor with $\sum_{l=1}^{k} w_{il} = 1$. The most common approach to weighting neighbors is to use $w_{il} \prec d_{il}^t$ where d_{il} is the distance between the l^{th} reference plot and the i^{th} target plot, and $t \in [0, -2]$. For this study, Euclidean distance was used, and values of *k* and *t* were selected that minimized the residual sum of squares in the reference set using the leave-one-out method (Chirici et al., 2008).

The four combinations of the two sets of predictor variables (echo-based metrics and CHM-based metrics) and the two prediction approaches (parametric linear regression and non-parametric k-NN) led to four different prediction models for estimating forest AGB. A stepwise selection method was used to select predictors for inclusion in the linear model or feature variables for inclusion in the distance metric for the k-NN procedure. Stepwise selection of predictor variables is appropriate when there is a large number of potential predictor variables and no underlying theory on which to base selection (Efroymson, 1960). At each stage, the procedure entails iteratively cycling through all remaining predictor variables to select the one additional variable that optimizes a selected criterion. After selection of a new variable, all variables already selected are checked to determine if any can be deleted without adversely affecting the criterion. The procedure terminates when no predictor variable can be added that further optimizes the criterion or can be deleted without adversely affecting the criterion. For this study, the criterion was the F-test statistic formulated as,

$$F(df_1, df_2)_{1-\alpha} = \frac{(SS_{p-1} - SS_p)/df_1}{SS_p/df_2},$$ (6)

where SS_{p-1} and S_p are the sum of squared residuals when linear or k-NN predictions adopt *p-1* or *p* feature variables, respectively; $df_1=1$, and $df_2=n-p$ where *n* is the sample size (number of reference plots). Significance levels of p=0.05, 0.10, and 0.15 were investigated, but all three produced the same results. Because stepwise selection is conducted using the reference plots which, in turn, are sampling outcomes, the procedure entails a further source of uncertainty beyond that induced by sampling (e.g., Burnham and Anderson, 1998). The stepwise selection uncertainty was considered negligible and ignored for this study.

Estimation

Notation

Let *T* denote total forest AGB in the study area, and let *N* denote the number of hexagons covering the study area where the number of hexagons equals the number of first-phase points. Let *R* denote the total size of the *N* hexagons, and let U denote the population of the *N* first-phase points. Let $U_f \subset U$ denote the subpopulation of the $N_f \leq N$ first-phase points classified as forest. Both U and U_f are random being the results of the random selection of points within hexagons. Let *a* denote the size of the plots and let AGB_i denote the total AGB for plot *i*. Thus, $\hat{T}_i = \frac{R}{a} AGB_i$ is the Horvitz-Thompson (HT)-like estimator of T based on plot *i* (Gregoire and Valentine, 2008, Section 7.4). Note that $\hat{T}_i$ would be an unbiased estimator of *T* if the point *i* was randomly selected from the entire grid covered by the *N* hexagons (URS) instead of randomly selected within the hexagon *i* (TSS).

On the basis of TSS theory, if all the *N* first-phase points were visited and if all the $\mathrm{AGB_i s}$ were recorded, the first-phase unbiased estimator of *T* would be the arithmetic mean of the $\hat{T}_i$ s, while the variance of the $\hat{T}_i$ s divided by *N* would be a conservative estimator of the sampling variance. Usually, non-forest points are neglected, tacitly supposing $\hat{T}_i = 0$ for each of them. However, because it is usually prohibitive to visit all the forest points, these estimators are only virtual and are considered as the theoretical basis for developing second-phase estimators.

Second-phase, design-based estimation

Let S denote the sample of the *n* second-phase points selected from the U_f by means of SRSWOR. If the second-phase sample was obtained using SRSWOR, the HT estimator of *T* is given by

$$\hat{\bar{T}}_{(2)} = \frac{N_f}{N}\frac{1}{n}\sum_{i\in S}\hat{T}_i \quad . \tag{7}$$

Conditional on the population U (and hence U_f) of points selected in the first-phase, from SRSWOR the variance of $\hat{\bar{T}}_{(2)}$ is,

$$\mathrm{Var}_2(\hat{\bar{T}}_{(2)} \mid S) = \frac{N_f(N_f-n)}{N^2}\frac{S_f^2}{n} \quad , \tag{8}$$

where S_f^2 is the variance of the $\hat{T}_i$ s in U_f. Accordingly, the variance of $\hat{\bar{T}}_{(2)}$ is

$$\mathrm{Var}(\hat{\bar{T}}_{(2)}) = \mathrm{E}_1\left\{\mathrm{Var}_2(\hat{\bar{T}}_{(2)} \mid U)\right\} + \mathrm{Var}_1(\hat{\bar{T}}_{(1)}) \tag{9}$$

Henceforth, subscript 1 denotes expectation and variance with respect the random placement of points within polygons (first phase), subscript 2 denotes expectation and variance with respect to the random selection of the *n* points (second phase) conditional to the set of points selected in the first phase, and no subscript denotes expectation and variance with respect to both phases.

Under SRSWOR, an unbiased estimator of the first term in Eq. (10) is,

$$\frac{N_f(N_f-n)}{N^2}\frac{1}{n(n-1)}\sum_{i\in S}(\hat{T}_i - \hat{\bar{T}}_2)^2 \quad , \tag{10}$$

while, from Eq. (A4) in the Appendix, the second term is estimated conservatively as,

$$\frac{N_f}{N^2 n}\sum_{i\in S}\hat{T}_i^2 - \frac{2}{N^2(N-1)}\frac{N_f(N_f-1)}{n(n-1)}\sum_{h>i\in S}\hat{T}_i\hat{T}_h \quad . \tag{11}$$

Accordingly, the second-phase conservative estimator of the variance of $\hat{\bar{T}}_{(2)}$ is given by,

$$\hat{V}_{(2)} = \frac{N_f(N_f-n)}{N^2}\frac{1}{n(n-1)}\sum_{i\in S}(\hat{T}_i - \hat{\bar{T}}_2)^2 + \frac{N_f}{N^2 n}\sum_{i\in S}\hat{T}_i^2 - \frac{2}{N^2(N-1)}\frac{N_f(N_f-1)}{n(n-1)}\sum_{h>i\in S}\hat{T}_i\hat{T}_h \quad , \tag{12}$$

from which $SE = \sqrt{\hat{V}_{(2)}}$ is the standard error estimate, $RSE = SE/\hat{\bar{T}}_{(2)}$ is the relative standard error estimate and $\hat{\bar{T}}_{(2)} \pm 1.96 \times SE$ is the confidence interval at a nominal level of 95%.

Second-phase model-assisted estimation

For this study, we used parametric linear regression and the non-parametric k-NN technique to attempt to improve the estimation in the study area. If for each $i \in$ U$_f$, a vector of auxiliary variables has been recorded in such a way that the model predictions of the $\mathrm{AGB_i s}$, denoted $\tilde{AGB}_i$ s, are available for each forest plot. To use the difference estimator, these predictions must also be calculated for the second-phase points, $i \in$ S, even if they are actually known. Thus, the empirical difference estimator adopted in Baffetta et al (2009) can be adopted also in this case. Once the $\tilde{AGB}_i$ s are obtained for each $i \in$ U$_f$, the $\hat{T}_i$ predictions are readily obtained by means of $\tilde{T}_i = \frac{R}{a}\tilde{AGB}_i$ for each $i \in$ U$_f$.

Under SRSWOR, the second-phase model-assisted estimator of *T* is,

$$\tilde{\tilde{T}}_{(2)} = \frac{N_f}{N}\left\{\frac{1}{N_f}\sum_{i\in \mathrm{U}_f}\tilde{T}_i + \frac{1}{n}\sum_{i\in \mathrm{S}} e_i\right\}, \qquad (13)$$

where $e_i = \hat{T}_i - \tilde{T}_i$ for each $i \in$ S. Conditional on the population U (and hence U$_f$) of points selected in the first-phase, from SRSWOR the variance of $\tilde{\tilde{T}}_{(2)}$ is approximated by,

$$\mathrm{Var}_2(\tilde{\tilde{T}}_{(2)} \mid \mathrm{U}) = \frac{N_f(N_f - n)}{N^2}\frac{S_e^2}{n}, \qquad (14)$$

where S_e^2 is the variance of the e_j^0 s in U$_f$, and $e_j^0 = \hat{T}_j - \mathrm{E}(\tilde{T}_j)$ (Baffetta et al., 2009). Accordingly, the variance of $\tilde{\tilde{T}}_{(2)}$ is.

$$\mathrm{Var}_2(\tilde{\tilde{T}}_{(2)}) = \mathrm{E}_1\left\{\mathrm{Var}_2(\tilde{\tilde{T}}_{(2)} \mid \mathrm{U})\right\} + \mathrm{Var}_1(\hat{\tilde{T}}_{(1)}). \qquad (15)$$

Under SRSWOR the first term in Eq. (16) is estimated by,

$$\frac{N_f(N_f - n)}{N^2}\frac{1}{n(n-1)}\sum_{i\in \mathrm{S}}(e_i - \bar{e})^2, \qquad (16)$$

where $\bar{e}$ is the arithmetic mean of the e_j s (Baffetta et al, 2009, Appendix A4) while estimation of the second term remains unchanged as in Eq. (12). Accordingly, from Eqs. (12) and (17), the second-phase conservative estimator of the variance of $\tilde{\tilde{T}}_{(2)}$ is,

$$\tilde{\tilde{V}}_{(2)} = \frac{N_f(N_f - n)}{N^2}\frac{1}{n(n-1)}\sum_{i\in \mathrm{S}}(e_i - \bar{e})^2 + \frac{N_f}{N^2 n}\sum_{i\in \mathrm{S}}\hat{T}_i^2 - \frac{2}{N^2(N-1)}\frac{N_f(N_f - 1)}{n(n-1)}\sum_{h>i\in \mathrm{S}}\hat{T}_i\hat{T}_h, \qquad (17)$$

from which $SE = \sqrt{\tilde{\tilde{V}}_{(2)}}$ is the standard error estimate, $RSE = SE / \tilde{\tilde{T}}_{(2)}$ is the relative standard error estimate and $\tilde{\tilde{T}}_{(2)} \pm 1.96 \times SE$ is the confidence interval at a nominal level of 95%.

2.1.3. Results

For each combination of the two kinds of metrics and the two prediction methods, the stepwise procedure selected only a single predictor variable. The variables selected were similar for both the linear models and the k-NN technique, regardless of whether echo-based or CHM-based metrics were used (Table 1). The linear regression model and the p_{60} echo-based predictor variable produced the most accurate overall results with R^2 = 0.58 (Figure 2). For the k-NN technique, p_{70} was selected as the single echo-based predictor and produced R^2 = 0.54. For both the linear regression model and the k-NN technique, the H_{avg} CHM-based metric was selected as the predictor variable and produced R^2 = 0.56 and R^2 = 0.48, respectively.

Graphs of observations versus predictions (Figures 2-5) suggest no general lack of fit of the models (predictions) to the observations, although as expected substantial heteroskedasticity was observed.

Table 1. Model-assisted estimates.

Predictor variables	Prediction technique	Variables selected	Total estimate	SE(RSE) estimate	95% confidence interval
Echoes	Linear	$p60$	1,961,886	205,904 (10%)	1,558,314-2,365,458
	k-NN	$p70$	2,029,560	209,493 (10%)	1,618,954-2,440,166
CHM	Linear	H_{avg}	2,017,132	207,072 (10%)	1,611,271-2,422,993
	k-NN	H_{avg}	2,119,152	208,941 (10%)	1,709,628-2,528,676

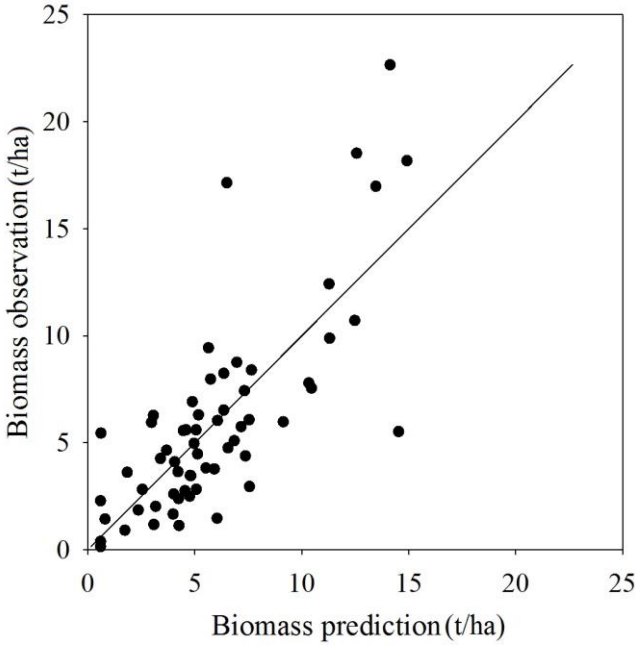


Figure 2. Observations versus predictions for the linear model with the p_{60} echo-based predictor variable.

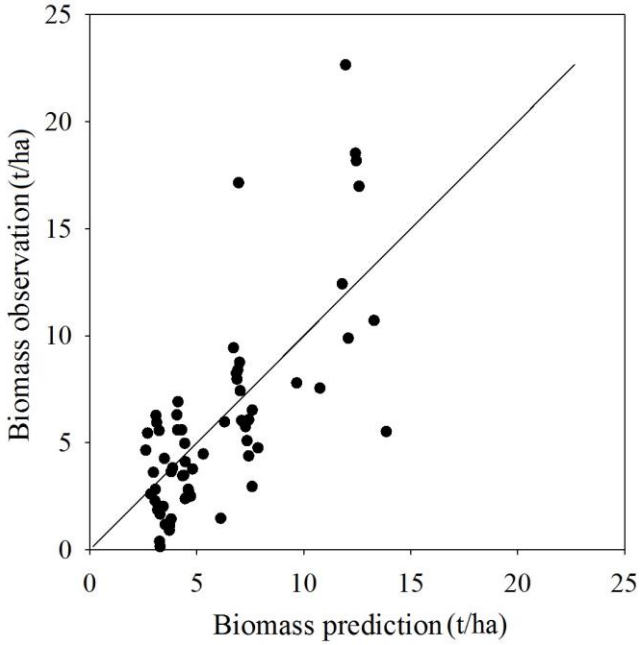


Figure 3. Observations versus predictions for k-NN with the p_{70} echo-based predictor variable.

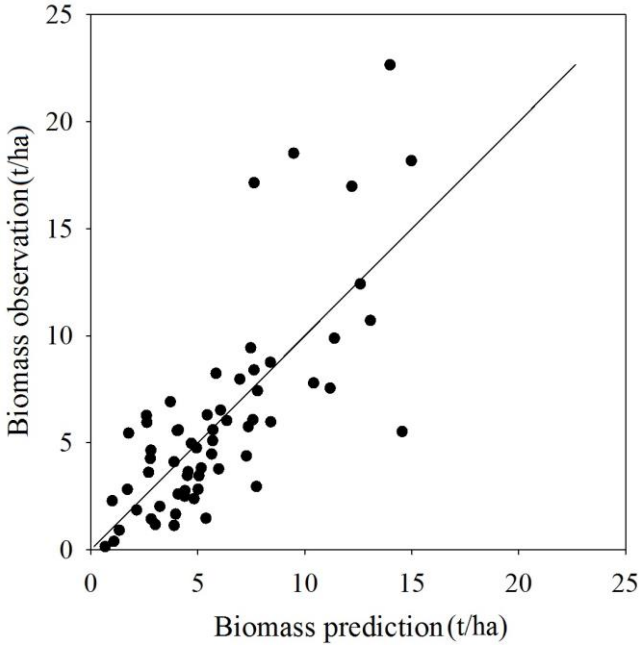


Figure 4. Observations versus predictions for linear model with the H_{avg} CHM-based predictor variable.

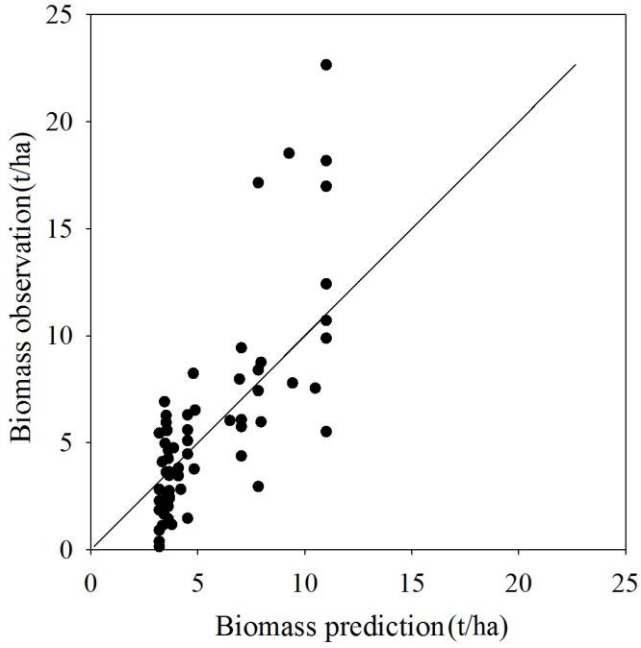


Figure 5. Observations versus predictions for k-NN with the H_{avg} CHM-based predictor variable.

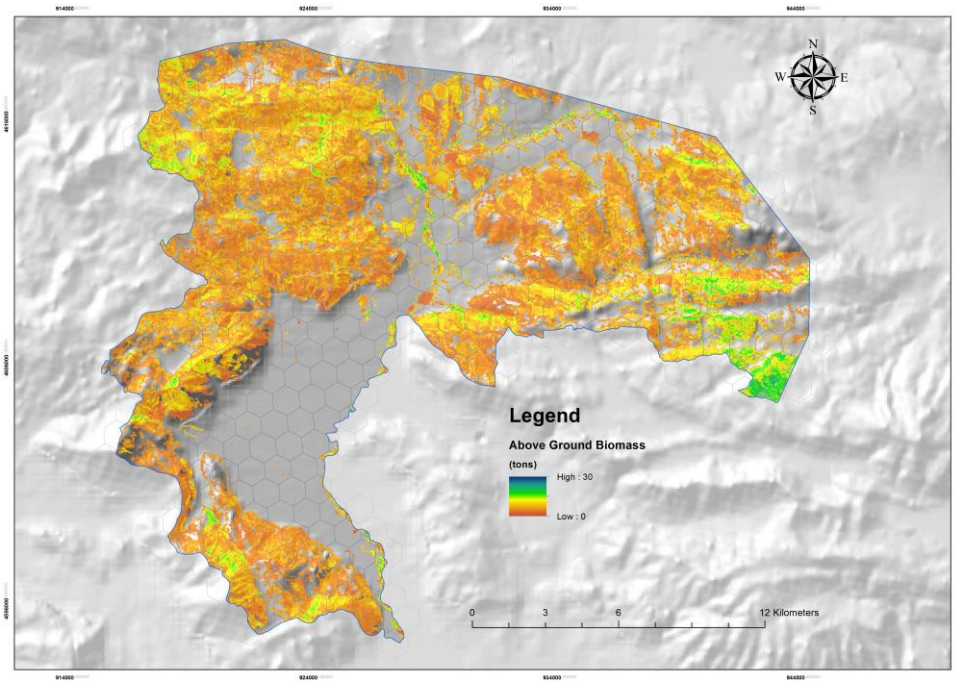


Figure 6. Map of forest AGB predictions obtained for each hexagon using the linear regression model and the *p60* echo-based predictor variable.

The design-based, second-phase HT estimate of the total AGB was $\hat{\bar{T}}_{(2)} = 2{,}277{,}061$ *t* with standard error estimate $\mathrm{SE} = 255{,}134$ *t* and corresponding relative standard error estimate RSE=11%. The 95% confidence interval was [1,766,793 t, 2,787,329 t]. The model-assisted, second-phase estimates for the four combinations are reported in Table 1. All the sampling strategies were consistent among themselves, with all the estimates within the confidence intervals of the others. Model-assisted estimates were invariably smaller than the HT estimate, and based on the SE estimates, the model assisted estimators were consistently more accurate than the HT estimator.

2.1.4.Discussions and Conclusions

The study had multiple objectives: to compare model-assisted estimates of total forest AGB obtained using four approaches: (i) a linear model with echo-based metrics as predictor variables, (ii) a linear model with CHM-based metrics as predictor variables, (iii) the k-NN technique with echo-based metrics as feature variables, and (iv) the k-NN technique with CHM-based metrics as feature variables. Additionally, estimation using the MA estimator in a two-phase TSS framework was to be evaluated.

Among the large number of both echo-based and CHM-based metrics available, only a single variable was selected for each of the four approaches. Further, they were very similar for both the linear regression

and k-NN models, namely the 60th percentile (p_{60}), and the 70th percentile (p_{70}) echo-based metrics, respectively, and the H_{avg} CHM-based metric.

For both prediction techniques, the average height, H_{avg}, of the CHM pixels in the plot area was chosen as the single predictor. These results are consistent with the current literature where only a few, albeit different, echo-based and CHM-based metrics were selected for prediction models. As previously noted, Næsset (2002) and Næsset and Gobbaken (2008) developed predictive models of forest volume using only two echo-based metrics; Corona et al. (2008) and Barbati et al. (2009) used only a single CHM-based metric; and Floris et al. (2010) used the same H_{avg} CHM-based metric to predict forest volume in an alpine environment.

The similarity in estimates obtained using the echo-based and CHM-based metrics can be attributed to the strong relationships between the H_{avg} CHM-based metric and the H_{avg}, p_{60}, p_{70}, and p_{50} echo-based metrics (Figure 7). This finding is especially relevant for users who do not have the skill or software/hardware resources for manipulating raw LiDAR pulses, or who have access only to a raster CHM rather than raw pulse data.

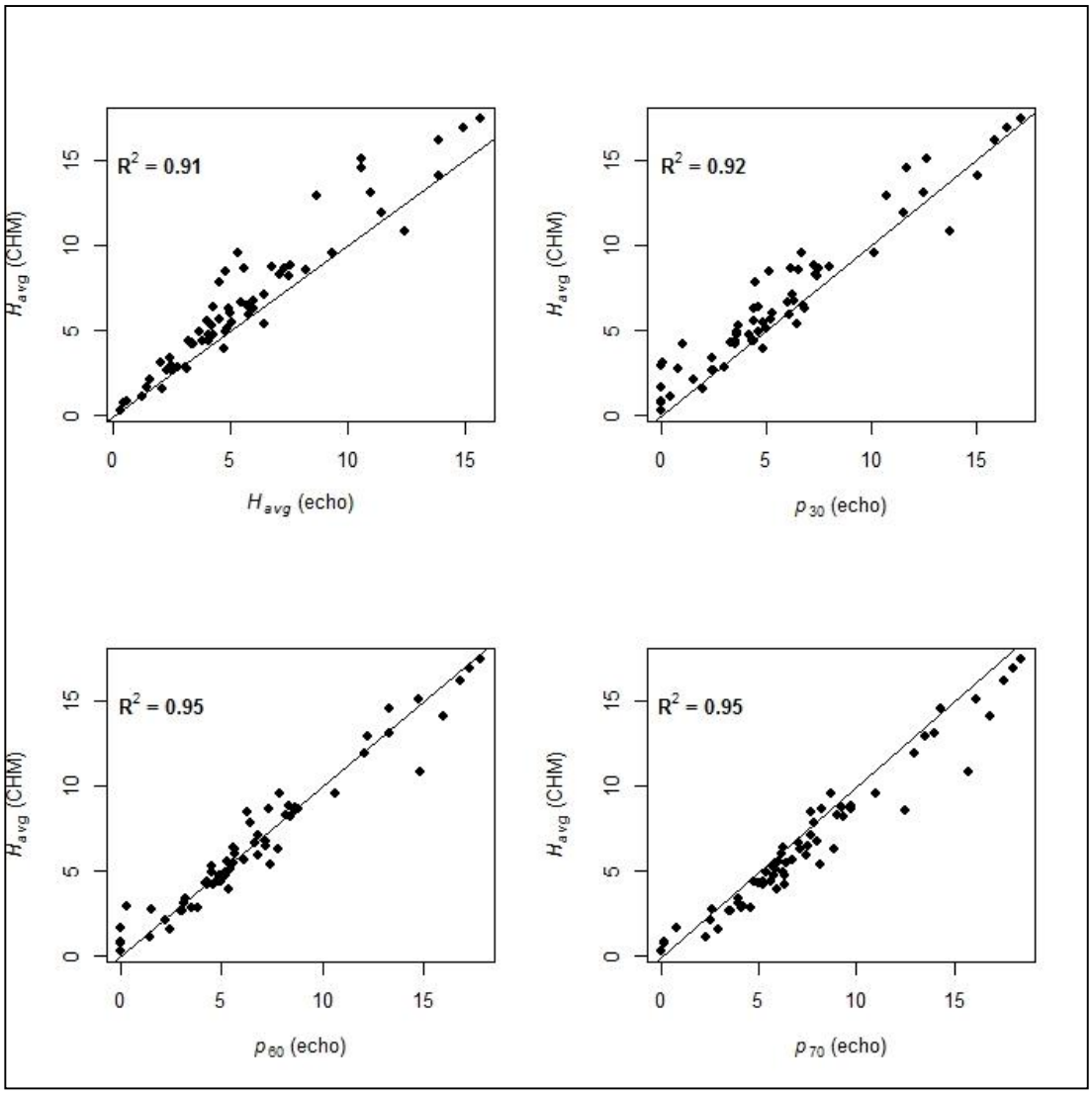


Figure 7. The H_{avg} CHM-based metric versus, H_{avg}, p_{60}, p_{70}, and p_{50} echo-based metrics.

Use of the MA estimator in a TSS design-based framework produced estimates of total forest AGB for the four combinations of metrics and prediction methods that were consistent with each other and with the design-based HT estimate. The small reduction in the SE estimates achieved using the MA estimators with respect to the HT design-based estimator (from 11% to 10%) was probably due to the relative weakness of correlations between AGB and prediction variables (R^2 always smaller than 0.6). Greater reductions should be expected with stronger correlations.

Multiple conclusions were drawn from the study. First, the echo-based and CHM-based metrics produced no substantial differences in estimates of total forest AGB. For the mainly broadleaved forest types in the study area, echo-based and CHM-based metrics can both be used to reliably estimate total forest AGB. If both metrics are available, the user can choose either depending on experience, technical expertise and software skills. This result is important because the raster CHM can be derived from sources other than ALS

data such as satellite radar imagery (Sexton et al., 2009), satellite LiDAR (Iqbal et al., 2013) or multi-angular aerial photography (Koukal & Atzberger, 2012).

Second, the parametric linear and non-parametric k-NN prediction techniques produced similar SE estimates for all the MA estimators which were consistently smaller than the SE estimate produced by the design-based HT estimator. In this case, the choice of which prediction to use for assisting estimation depends on the modeling skills of the researcher.

Comparisons of metrics and modeling approaches for this study were with respect to estimates of total forest AGB, a variable that is relatively easy to estimate using ALS data because of its strong dependence on forest height. We do not exclude the possibility of different results when estimating other variables; on the contrary we encourage further testing for additional forest response variables.

Third, even when using a sampling design such as TSS that is well-suited for forest inventory purposes, use of the MA estimator with auxiliary information, here in the form of remotely sensed data, is recommended to improve the precision of the estimates.

2.1.5. References

Axelsson, P. (2000). DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing, 33(B4).

Baffetta, F, Fattorini, L, Franceschi, S, & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment, 113(3), 463-475.

Barbati, A., Chirici, G., Corona, P., Montaghi, A., & Travaglini, D. (2009) - Area-based assessment of forest standing volume by field measurements and airborne laser scanner data. International Journal of Remote Sensing, 30(19), 5177–5194.

Beaudoin, A.; Bernier, P.Y.; Guindon, L.; Villemaire, P.; Guo, X.J.; Stinson, G.; Bergeron, T.; Magnussen, S.; Hall, R.J. 2014. Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. Can. J. For. Res. 44:521-532.

Burnham, K.P. & Anderson, D.R. (1998) Model Selection and Inference. Springer, New York, 353 p.

Castellani, C, Scrinzi, G, Tabacchi, G, & Tosi, V, (1984). Inventario Forestale Nazionale Italiano (I.F.N.I.) - Tavole di cubatura a doppia entrata. Istituto Sperimentale per l'Assestamento Forestale e per l'Alpicoltura, Trento, 83.

Chirici G., Mura, M., McInerney, D., Py N., Tomppo, E., Waser, Lars, T., Travaglini, D., McRoberts, R.E., submitted. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sensing of Environment.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F., & Bertini, R. (2008). Non-parametric and parametric methods using satellite images for estimating growing stock volume in alpine and Mediterranean forest ecosystems. Remote Sensing of Environment, 112(5), 2686–2700.

Chirici, G., Di Martino, P., Ottaviano, M., Santopuoli, G., Chiavetta, U., Tonti, D., Garfì, V., & Marchetti, M. (2011). La carta forestale su basi tipologiche. In V. Garfì, & M. Marchetti (Eds.), Tipi forestali e preforestali della regione Molise (pp. 145-152): Edizioni dell'Orso.

Corona P., & Fattorini L. (2008). Area-based LiDAR-assisted estimation of forest standing volume. Canadian Journal of Forest Research, 38(11), 2911-2916.

Corona, P., Lamonaca, A., Chirici, G., Travaglini, D., Marchetti, M., Minari, E., & Montaghi, A. (2008) Estimation of growing stock of broadleaved forests by airborne laser scanning, 39-44. In: Gianelle, D., Travaglini, D., Mason, F., Minari, E., Chirici, G., & Chemini, C. (eds.). Canopy analysis and dynamics of a floodplain forest. Rapporti Scientifici, 3. Centro Nazionale per lo studio e la conservazione della Biodiversità Forestale - Bosco della Fontana. Cierre Grafica Editore, Verona, 96.

Corona, P., Fattorini, L., & Franceschi, S. (2009). Estimating the volume of forest growing stock using auxiliary information derived from relascope or ocular assessments. Forest Ecology and Management, 257(10), 2108–2114.

Corona, P., Cartisano, R., Salvati, R., Chirici, G., Floris, A., di Martino, P., Marchetti, M., Scrinzi, G., Clementel, F., Travaglini, D., & Torresan, C. (2012). Airborne laser scanning to support forest resource management under

alpine, temperate and mediterranean environments in Italy. Italian Journal of Remote Sensing / Rivista Italiana Di Telerilevamento, 45(1), 27–37.

Efroymson, M.A. (1960). Multiple regression analysis, mathematical methods for digital computers. Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

Ene, L.T., Næsset, E., Gobakken, T., Gregoire, T. G., Ståhl, G., & Nelson, R. (2012). Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sensing of Environment, 123, 579–592.

Evans, J.S., Hudak, A.T., Faux, R., & Smith, A.M.S. (2009). Discrete Return Lidar in Natural Resources: Recommendations for Project Planning, Data Processing, and Deliverables. Remote Sensing, 1(4), 776–794.

Fattorini, L, (2015). Design-based methodological advances to support national forest inventories: a review of recent proposals. iForest, 8, 6-11.

Federici, S., Vitullo, M., Tulipano, S., De Lauretis, R., & Seufert, G. (2008). An approach to estimate carbon stocks change in forest carbon pools under the UNFCCC: the Italian case. iForest,1, 86-95.

Floris, A., Clementel, F., Farruggia, C., & Scrinzi, G. (2010) Stima su base LiDAR delle provvigioni legnose forestali: uno studio per la Foresta di Paneveggio. Rivista Italiana di Telerilevamento, 42(3),15-32.

Fusco, S., Pflugmacher, D., Kirschbaum, A., Cohen, W., Chiatante, D., & Montagnoli A. (2008) Uso di dati LiDAR per stima della biomassa forestale in un bosco misto di latifoglie: un caso studio in Valsassina (LC). Atti 12a Conferenza Nazionale ASITA, L'Aquila, 21-21 Ottobre 2008 (with English summary).

Gaulton, R., & Malthus, T. J. (2010). LiDAR mapping of canopy gaps in continuous cover forests: A comparison of canopy height model and point cloud based techniques. International Journal of Remote Sensing, 31(5), 1193–1211.

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Holmd, S., Ørka, H.O., & Astrup, R. (2012). Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment, 123, 443–456.

Gregoire, T.G., Valentine, H.T. (2008) Sampling Strategies for Natural Resources and the Environment. Chapman & Hall, Boca Raton (FL) , 474 p

Gregoire, T.G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., & Holm, S. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County. Canadian Journal of Forest Research, 41(1), 83–95.

Hawbaker, T.J., Gobakken, T., Lesak, A., Trømborg, E., Contrucci, K., & Radeloff, V.C. (2010) Light detection and ranging-based measures of mixed hardwood forest structure. Forest Science, 56(3), 313–326.

Koukal, T., & Atzberger, C. (2012). Potential of Multi-Angular Data Derived From a Digital Aerial Frame Camera for Forest Classification. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal, 5, 30-43

Kraus, K., & Pfeifer, N. (1998) Determination of terrain models in wooded areas with airborne laser scanner data. ISPRS Journal of Photogrammetry and Remote Sensing, 53(4), 193-203.

Iqbal, I.A., Dash, J., Ullah, S., & Ahmad, G. (2013). A novel approach to estimate canopy height using ICESat/GLAS data: A case study in the New Forest National Park, UK. International Journal of Applied Earth Observation and Geoinformation, 23, 109-118

Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A., & Gower, S.T., (2014). Lidar remote sensing of above-ground biomass in three biomes. Global Ecology and Biogeography, 11(5), 393–399.

Lim, K., Treitz, P., Wulder, M., St-Ongé, B., & Flood, M. (2003a). LiDAR remote sensing of forest structure. Progress in Physical Geography, 27(1), 88–106.

Lim, K., Treitz, P., Baldwin, K., Morrison, I., & Green, J. (2003b). Lidar remote sensing of biophysical properties of tolerant northern hardwood forests. Canadian Journal of Remote Sensing, 29(5), 658–678.

Magnussen, S., & Boudewyn, P. (1998). Derivations of stand heights from airborne laser scanner data with canopy-based quantile estimators. Canadian Journal of Forest Research, 28(7),1016–1031.

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

McRoberts, R.E., Næsset, E., & Gobakken, T. (2015). Optimizing the k-Nearest Neighbors technique for estimation of forest aboveground biomass using airborne laser scanning data. Remote Sensing of Environment. http://dx.doi.org/10.1016/j.rse.2015.02.026.

Maltamo, M., Næsset, E., & Vauhkonen, J. (Eds.). (2014). Forestry applications of airborne laser scanning. Springer, Dordrecht. 464 p.

Montaghi, A., Corona, P., Dalponte, M., Gianelle, D., Chirici, G., & Olsson, H. (2013). Airborne laser scanning of forest resources: An overview of research in Italy as a commentary case study. International Journal of Applied Earth Observation and Geoinformation, 23, 288–300.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. Remote Sensing of Environment, 80(1), 88–99.

Næsset, E. (2004a). Effects of different flying altitudes on biophysical stand properties estimated from canopy height and density measured with a small-footprint airborne scanning laser. Remote Sensing of Environment, 91, 243–255.

Næsset, E. (2004b). Practical large-scale forest stand inventory using a small-footprint airborne scanning laser. Scandinavian Journal of Forest Research, 19,164–179.

Næsset, E., & Gobakken, T. (2005). Estimating forest growth using canopy metrics derived from airborne laser scanner data. Remote Sensing of Environment, 96, 453–465.

Næsset, E., & Gobakken, T. (2008). Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sensing of Environment, 112(6), 3079–3090.

Næsset, E., Gobakken, T., Bollandsås, O. M., Gregoire, T. G., Nelson, R., & Ståhl, G. (2013). Comparison of precision of biomass estimates in regional field sample surveys and airborne LiDAR-assisted surveys in Hedmark County, Norway. Remote Sensing of Environment, 130, 108–120.

Parker, G.G., & Russ, M.E. (2004). The canopy surface and stand development: Assessing forest canopy structure and complexity with near-surface altimetry. Forest. Ecology and. Management, 189, 307–315.

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K., & Hyyppä, J. (2015) Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. Remote Sensing of Environment, 158, 431–440.

Särndal, C.E., Swensson, B., & Wretman, J. (1992). Model Assisted Survey Sampling. New York, New York, USA. Springer Verlag.

Sexton, J.O., Bax, T., Siqueira, P., Swenson, J.J., & Hensley, S. (2009). A comparison of lidar, radar, and field measurements of canopy height in pine and hardwood forests of southeastern North America. Forest Ecology and Management, 257, 1136-1147.

SLU Forest Map (2013). Available at: http://www.slu.se/en/collaborative-centres-and-projects/swedish-national-forest-inventory/services-and-products/interactive-services/knn-sweden/. (Last accessed: March 015).

Tomppo, E. (1990). Designing a satellite image-aided national forest survey in Finland. In Proceedings: The usability of remote sensing for forest inventory and planning. Swedish University of Agricultural Sciences, Remote Sensing Laboratory. Report 4, 43-47.

Vauhkonen, J., Maltamo, M., McRoberts, R.E., & Næsset, E. (2014). Introduction to forestry applications or airborne laser scanning. In: Maltamo, M., Næsset, E., & Vauhkonen, J. (Eds.). Forestry applications of airborne laser scanning. Springer, Dordrecht.

Vidal, C., Lanz, A., Tomppo, E., Schadauer, K., Gschwantner, T., Di cosmo, L., & Robert, N. (2008). Establishing forest inventory reference definitions for forest and growing stock: A study towards common reporting. Silva Fennica, 42, 247–266.

Wilson, B.T., Lister, A.J., & Riemann, R.I. (2013). Forest carbon stocks of the contiguous United States (2000-2009). Newtown Square, PA: USDA Forest Service, Northern Research Station. Available at: www.fs.usda.gov/rds/archive/Product/RDS-2013-0004. (Last accessed: March 2015).

Wulder, M.A., Bater, C.W., Coops, N.C., Hilker, T., & White, J.C. (2008). The role of LiDAR in sustainable forest management. Forestry Chronicle, 84(6), 807–826.

Zolkos, S.G., Goetz, S.J., & Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment, 128, 289–298.

2.2. Study II

From:

[1]Chirici G., [2]Mura M., [3]McInerney D., [4]Py N., [5]Tomppo E. O., [6]Waser L. T., [1]Travaglini D., [7]McRoberts R. E. (2016). A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. *Remote Sensing of Environment*.

[1] Università degli Studi di Firenze, Department of Agricultural, Food and Forestry Systems, Via San Bonaventura, 13 – 50145. Firenze, Italy

[2] Università degli Studi del Molise, Dipartimento di Bioscienze e Territorio, Pesche (IS), Italy

[3] Coillte Teoranta, Ireland.

[4] Institut National de l'Information Géographique et Forestière, France

[5] The Natural Resources Institute Finland

[6] Swiss Federal Research Institute WSL, Switzerland.

[7] United States Department of Agriculture, Forest Service, Northern Research Station

Abstract

The k-Nearest Neighbor (k-NN) technique is a popular method for producing spatially contiguous predictions of forest attributes by combining field and remotely sensed data. In the framework of Working Group 2 of COST Action FP1001, we reviewed the scientific literature for forestry applications of k-NN. Information available in scientific publications on this topic was used to populate a database that was then used as the basis for a meta-analysis. We extracted qualitative and quantitative information from 260 experimental tests described in 148 scientific papers. The papers represented a geographic range of 26 countries and a temporal range from 1981 to 2013. Firstly, we describe the literature search and the information extracted and analyzed. Secondly, we report the results of the meta-analysis, especially with respect to estimation accuracies reported for k-NN applications for different configurations, different forest environments, and different input information. We also provide a summary of results that may reasonably be expected for those planning a k-NN application using remotely sensed data from different sensors and for different forest attributes. Finally, we identify some methodological publications that have advanced the state of the science with respect to k-NN.

Keywords: k-Nearest Neighbors, forestry applications, review, meta-analysis

2.2.1. Introduction

Nearest neighbors techniques can be considered a class of multivariate, non-parametric approaches to continuous or categorical prediction. The multivariate property of these techniques has made them particularly popular for use with remotely sensed and national forest inventory (NFI) data. With these techniques, predictions are calculated as linear combinations of observations for population units in a sample that are similar or nearest in a space of auxiliary variables to population units requiring predictions. Nearest neighbors techniques are appealing because they can be used for both univariate and multivariate prediction; they are non-parametric in the sense that no assumptions regarding the distributions of response or auxiliary variables are necessary; they are synthetic in the sense that they can readily use information external to the geographic area of interest; and they can be used with a wide variety of data sets. When used with remotely sensed and spatially referenced NFI field data, nearest neighbors techniques can produce spatially continuous predictions (maps) of forest variables rather than just large area aggregations of plot data. These finer resolution map products add a new and useful dimension to NFIs by facilitating small area estimation, increased precision for large area estimation, and support for forest management, planning and monitoring.

Nearest neighbors techniques were first introduced in an unpublished U.S. Air Force report by Fix and Hodges (1951) as a non-parametric discriminant technique for classification into populations whose distributions are unknown. Much of the early foundational work on nearest neighbors techniques for classification purposes appears in the pattern recognition and machine learning literature. Within the natural resources area, these techniques were developed for the Finnish NFI in seminal papers by Tomppo (1990, 1991, 2008) based on earlier proposals by Kilkki and Päivinen (1987) and the ideas used with aerial photos by Poso (1972). McRoberts (2012) documented the broad international extent of the technique's use for a wide range of forestry applications including imputation of missing values for forest inventory and monitoring databases, mapping, small area estimation, and support for statistical inference. Commonly estimated forest response variables include growing stock volume, forest/non-forest, forest type, and commonly used remotely sensed feature variables include Landsat spectral bands and increasingly airborne laser scanning metrics. Recent forestry investigations have begun to emphasize foundational work on diagnostics (McRoberts, 2009), efficiency (e.g., Finley & McRoberts, 2008), optimization (e.g., Tomppo & Halme, 2004), and inference (e.g., McRoberts et al., 2007; Baffetta et al., 2009).

Variations of nearest neighbors techniques have been used operationally in both Europe and North America. In Finland, the first operational implementation of k-Nearest Neighbors (k-NN) was based on NFI, satellite and digital map data in 1990 (Tomppo, 1990, 1991). The primary initial purpose was forest resource estimation for small administrative units. The basic technique has since been enhanced using digital map data for stratification and genetic algorithms to weight feature variables as a means of increasing prediction accuracy (Tomppo & Halme 2004). The resulting municipality-level estimates are included in the official NFI statistics in Finland (Metinfo, 2007; Metla, 2013). In Sweden, the k-NN technique has been used to map forest variables such as wood volume, age, and height using NFI, satellite and digital map data (Reese et al., 2005). Basic end products include raster datasets for age, height, total wood volume, and volume by common species (SLU Forest Map, 2013). Additional products include dominant tree species, stand delineation, and base information for property taxation.

The k-NN technique has also been used operationally in North America. In Canada, Beaudoin et al. (2014) used the k-NN technique to produce continuous maps of 127 forest attributes to support regional policy and management issues. Reference data consisted of standardized observations from NFI photo plots, and feature variables were obtained from geospatial data layers that included MODIS spectral data, climatic and topographic variables. The map products provide unique baseline information for strategic analyses of Canadian forests (https://nfi.nfis.org). For the United States of America (USA), Wilson et al. (2012, 2013b) used nearest neighbor techniques with NFI plot data and vegetation phenology derived from multi-temporal MODIS imagery and other auxiliary variables to map live tree basal area for individual species across the eastern United States and to map individual carbon stocks for all of the contiguous states of the USA (Wilson et al., 2013 a,c). In the Pacific Northwest region of the USA, Ohmann et al. (2002, 2014) used nearest neighbors techniques to map and assess biodiversity, wildland fuels, and species composition and to monitor change in older forests, biomass and carbon. The maps have been widely used for research, land management, forest monitoring, and conservation planning applications (http://lemma.forestry.oregonstate.edu/). Thus, the widespread popularity of nearest neighbors techniques for both research and operational purposes justifies a review of the literature on the topic and identification of important methodological advances along with issues regarding practical and scientific application.

COST (European Cooperation on Science and Technology) is a European framework for promoting and facilitating scientific cooperation among scientists and researchers (COST, 2014). COST Action FP1001 focuses on European approaches for using multi-source NFIs to improve information on the potential supply of wood resources. Within COST Action FP1001, Working Group 1 focused on NFI sampling designs and estimation techniques with an emphasis on harmonization; Working Group 2 focused on methods for combining remotely sensed and NFI field data to improve estimates of wood resources; and Working Group 3 focused on the exchange of inventory volume and consumption information with emphasis on wood markets (COST FP1001, 2014).

The popularity of k-NN for use with forest inventory and remotely sensed data motivated Working Group 2 of COST Action FP1001 to conduct a comprehensive literature review of forestry applications. The review was implemented as a meta-analysis of the most relevant studies published in peer-review journals, book chapters and conference proceedings.

A meta-analysis is a quantitative analysis based on sound and reliable approaches aimed at providing an objective summary of results that may be helpful for other researchers in support of future applications. The usefulness of this kind of investigation, if compared to narrative or qualitative reviews, has been demonstrated for both ecological studies (Arnqvist and Wooster, 1995) and more recently for remote sensing applications in forestry (Garbulsky et al., 2011, Zolkos et al., 2013).

The study objectives were fourfold: (1) to document development and application of nearest neighbors techniques with respect to multiple factors including response and feature variables, distance metrics, algorithm characteristics, geographical regions of applications, accuracy and uncertainty measures, and results achieved in terms of prediction accuracy; (2) to provide a range of benchmark accuracy results that may reasonably be expected for combinations of factors such as response variable and forest type; (3) to provide guidelines for prospective users; and (4) to identify and briefly summarize methodological papers that

have advanced the state of the science. Thus, the paper provides more support for practical nearest neighbors implementations and future research than Eskelson et al. (2009), McRoberts et al. (2010) or the literature review section of McRoberts (2012).

The k-Nearest Neighbors technique

For notational purposes, Y is commonly used to denote a possibly multivariate vector of response variables with observations for a sample of size n from a finite population of size N, and X is used to denote a vector of auxiliary variables with observations for all population units. In the terminology of nearest neighbors techniques, the auxiliary variables are designated *feature variables* and the space defined by the feature variables is designated the *feature space*; the set of sample population units for which observations of both response and feature variables are available is designated the *reference set*; and the set of population units for which predictions of response variables are desired is designated the *target set*. All population units for both the reference and target set are assumed to have a complete set of observations for all feature variables.

For continuous response variables such as biomass or growing stock volume, the nearest neighbors prediction, $\hat{y}_i$, for the i^{th} target set unit is calculated as,

$$\hat{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i \quad (1)$$

where $\left\{ y_j^i, j = 1, 2, \ldots, k \right\}$ is the set of response variable observations for the k reference set units that are nearest or most similar to the i^{th} target set unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the j^{th} nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1$. Weights for neighbors are often of the form $w_{ij} \propto d_{ij}^{-t}$ with $0 \le t \le 2$ where d_{ij} is the distance in feature space between the i^{th} target unit and the j^{th} nearest neighbor. For categorical variables such as forest/non-forest or forest type, the predicted class of the i^{th} target set unit is the most heavily weighted class among the k nearest neighbors, a weighted median or mode in case of ordinal scale variables, or a mode in the case of nominal variables.

Implementation of nearest neighbors techniques requires three selections: (i) the distance metric, d, to assess similarity, (ii) the number, k, of nearest neighbors to be used when calculating predictions, and (iii) a scheme to weight individual neighbors when calculating predictions. Multiple distance metrics have been proposed ranging from simple unweighted Euclidean distance to more complex metrics that attempt to optimize the selection and/or weighting of the feature variables. Many familiar metrics can be expressed in matrix form as,

$$d^2_{ij} = (\mathbf{X}_i - \mathbf{X}_j)' \, \mathbf{M} \, (\mathbf{X}_i - \mathbf{X}_j), \quad (2)$$

where i denotes a target set unit for which a prediction is sought, j denotes a reference set unit, $\mathbf{X}_i$ and $\mathbf{X}_j$ are vectors of observations of feature variables for the i^{th} and j^{th} units, respectively, and **M** is a square, positive definite matrix. When **M** is the identity matrix, Euclidean distance results; when **M** is a non-identity diagonal matrix, weighted Euclidean distance results; and when **M** is the inverse of the covariance matrix of the feature variables, Mahalanobis distance results. Metrics based on canonical correlation and canonical correspondence analyses can also be expressed in matrix form.

The value of k is often selected as an arbitrarily small number in the range 1-10, although some approaches attempt to optimize the selection with respect to criteria such as classification accuracy or root mean square error. Less attention has been paid to neighbor weighting schemes. The term *k-Nearest Neighbors* (k-NN) is generic and refers to any nearest neighbor technique regardless of the distance metric, value of k, or neighbor weighting scheme.

2.2.2. Materials and Methods

Bibliographic resources for this review were obtained from systematic searches of the most important scientific databases and search engines: Scopus (http://www.scopus.com/home.url), Thomson Reuters Web of Science, Science Direct (http://www.sciencedirect.com/), IEEE Xplore (http://ieeexplore.ieee.org/Xplore/home.jsp), and Google Scholar (http://scholar.google.com). The search was conducted in 2013 using English keywords that refer to the integrated use of forest inventory and remotely sensed data for prediction using the k-NN technique. The main keywords used were: k-nearest neighbor (or neighbor), connected using the logical "or" operator with the following keywords estimation, imputation, forest management, forest inventory, and remote sensing. The search returned 148 peer-reviewed contributions from scientific peer-reviewed journals, conference proceedings, and book chapters that reported 260 experimental applications and evaluations.

We then constructed a database with 24 fields and populated it with the qualitative and quantitative information extracted from the literature review that described the 260 experiments. The resulting matrix, after the deletion of duplicated studies, had 24 fields/columns and 260 records/rows and served as the information source for our meta-analysis.

2.2.3. Results

General characteristics of studies

The main source of information was articles published in scientific journals (81.1%) with only 16.2% from conference proceedings. Nearly one-third of papers (31.1%) were published in Remote Sensing of Environment with the majority of the remainder published in Forest Ecology and Management, the Canadian Journal of Forest Research, the Scandinavian Journal of Forest Research, and the International Journal of Remote Sensing. Most papers (82.4%) reported applications of well-documented existing methods, but 7.4% reported more methodological contributions which were generally supported with practical applications or simulations.

The first paper identified from the literature search dates back to 1981 (Short & Fukunaga, 1981). Papers published in the 1980s often focused on the theoretical advantages of k-NN, while papers from the 90s and into the 21st century increasingly reported applications using forest inventory and remotely sensed data. The number of publications per year increased continuously until 2009 (Figure 1).

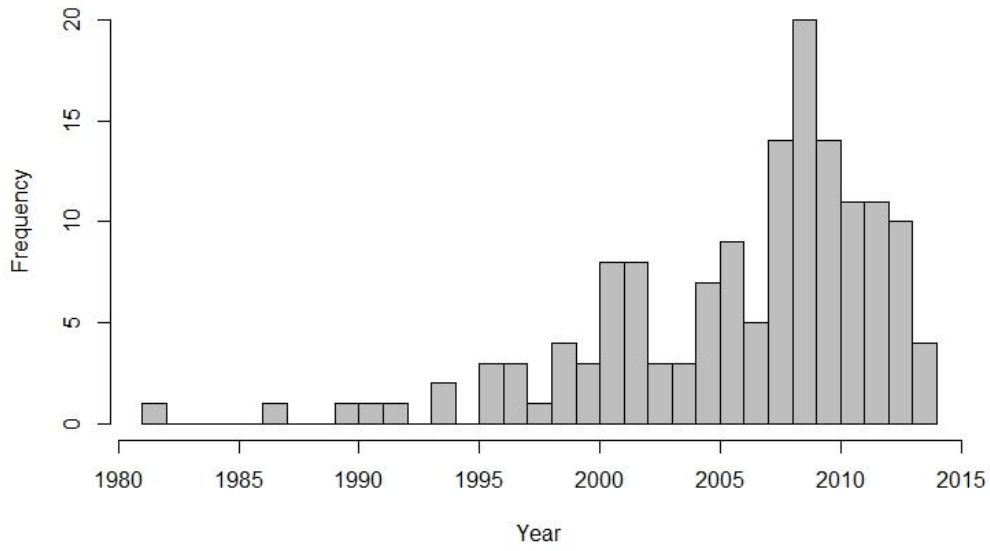


Figure 1 Number of publications per year.

The geographical coverage of published papers was mainly Europe and North America (Figure 2). The total number of countries for which applications have been reported is 26 and includes papers from six continents (Austria, Brazil, Canada, Chile, China, Costa Rica, Ecuador, Estonia, Finland, Germany, Ghana, Ireland, Italy, Japan, Korea, Lithuania, Mexico, Namibia, New Zealand, Norway, Portugal, Russia, Scotland, South Korea, Sweden, USA).

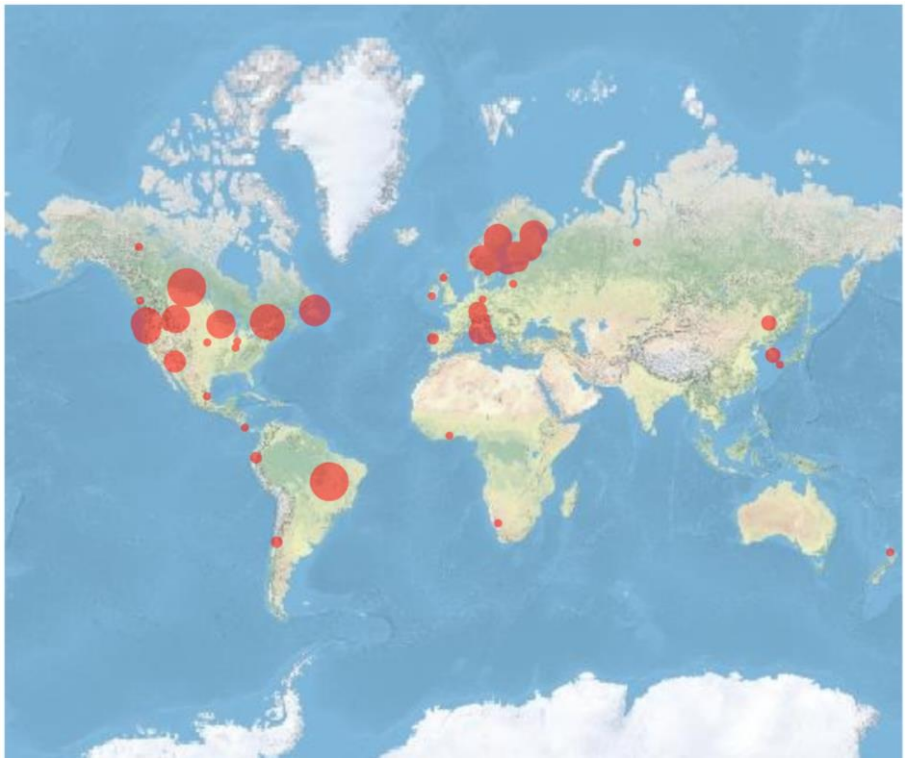


Figure 2. Locations of the studies according to the sub-country geographic area reported in the paper (the size of the dots are proportional to the area covered by the study). Base map from Esri ArcGIS online.

The forest environments most frequently investigated were boreal coniferous forests (36.5%) mainly located in Europe, and temperate continental (18.8%) and temperate mountain (11.5%) forests, mainly in the USA. Some European studies focused on temperate oceanic forests (6.9%) and Mediterranean forests (5%), with additional contributions for boreal mountain forests in the Nordic region (5.4%) (Figure 3). A few studies were classified as "continental" because they covered very large areas such as the entirety of Canada (Beaudoin et al., 2014), the entire USA (Wilson et al., 2013a), and the eastern part of the USA (Wilson at al., 2012).

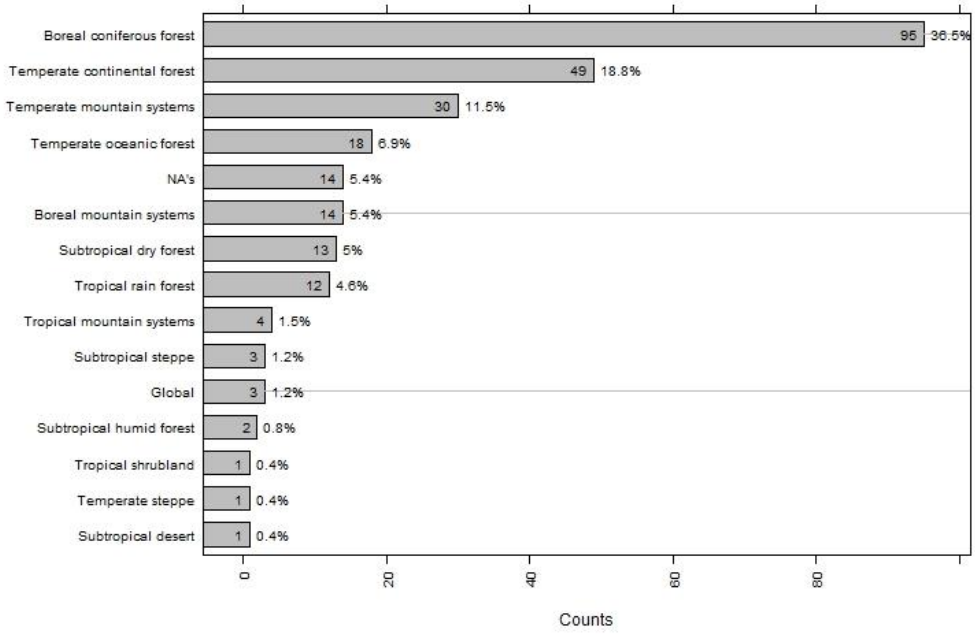


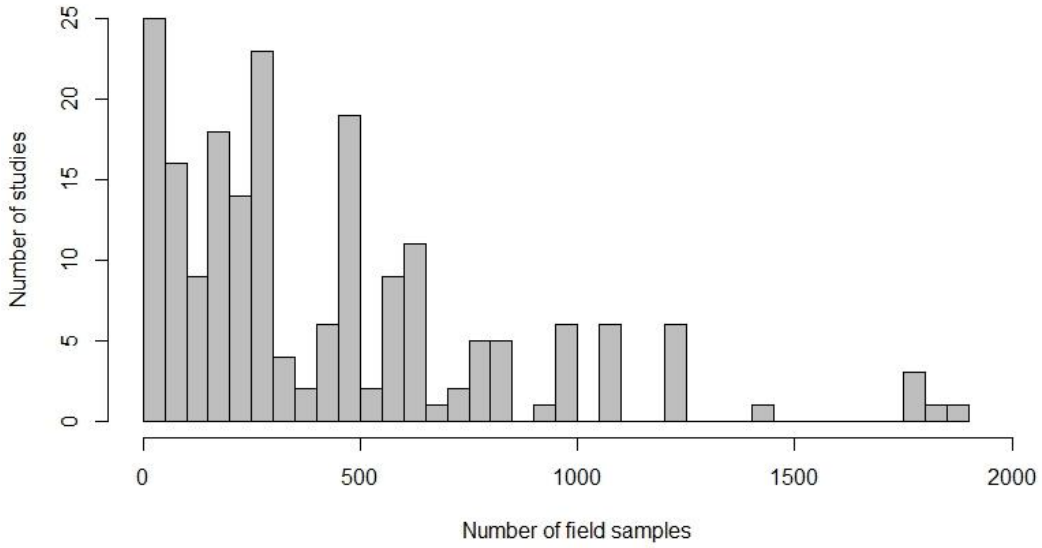
Figure 3: number of publications per Global Ecological Zones (FAO 2012). NA for not available.

The most frequently investigated response variables, representing 55.4% of studies, were the closely related growing stock volume, biomass and carbon stock variables, with other continuous response variables such as basal area (9.2%) and tree density (3.8%) also reported. Categorical response variables such as land use/land cover and forest type were investigated in 13.1% of studies.

Among feature variables, Landsat imagery was undoubtedly the most frequent source of remotely sensed data, being used in 48.8% of studies. Airborne laser scanning (ALS) metrics were used in almost 7% of the studies, digital aerial imagery in 6.5%, and SPOT imagery in 5.4%. More than 16% of the studies integrated remotely sensed data from multiple sources, while 25% of studies used data from non-remote sensing-based digital maps (e.g., digital elevation models) as feature variables or as auxiliary information for stratification purposes.

All studies used field data, mostly from local inventories or NFIs. For the 235 studies that reported the number of plots, sample sizes ranged from a minimum of 9 to a maximum of 190,888, with an average of 2,575. For 85% of these studies, reference set (sample) sizes for the field data were less than 2,000, and for 50% of the studies, the sizes were less than 400 (Figure 4).

a



b

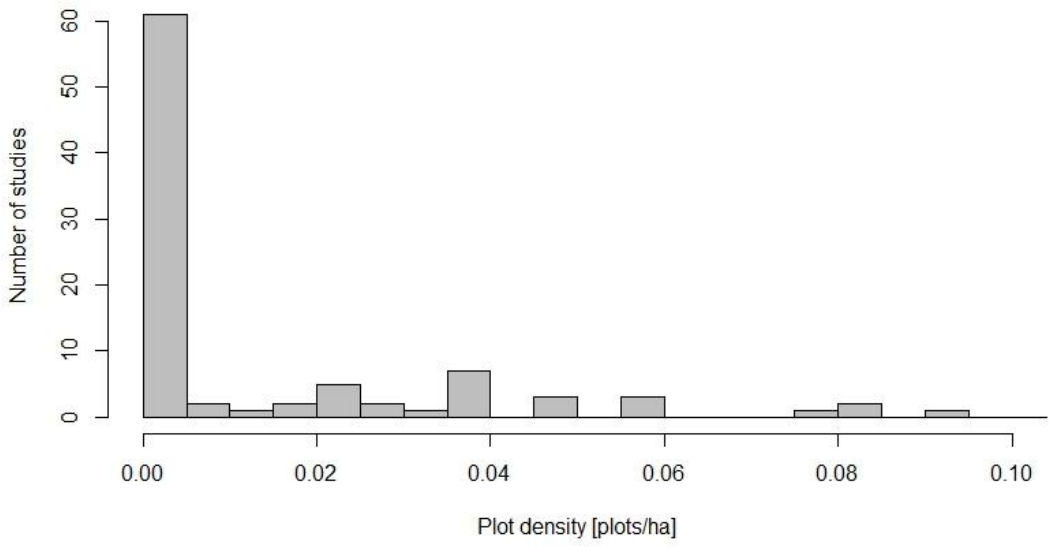


Figure 4: a) Number of field samples used in studies. The maximum value was limited to 2000 for display reasons. b) Plot density (number of plots per hectares) of the studies. The maximum value was limited to a density of 0.10 for display reasons.

k-NN configurations used

Among the k-NN distance metrics used, the Euclidean metric was most often reported (70%), followed by the Mahalanobis metric (3.5%), the canonical correlation analysis metric (1.9%), and Random Forest (1.5%). Values of k generally ranged between 1 and 10 with k = 5 being used most often and k>10 used only rarely (Figure 5). Selection of the k-value appeared to be independent of forest environment and the choice of a distance metric, although k=1 is often selected for use with the canonical correlation analysis and canonical correspondence metrics.

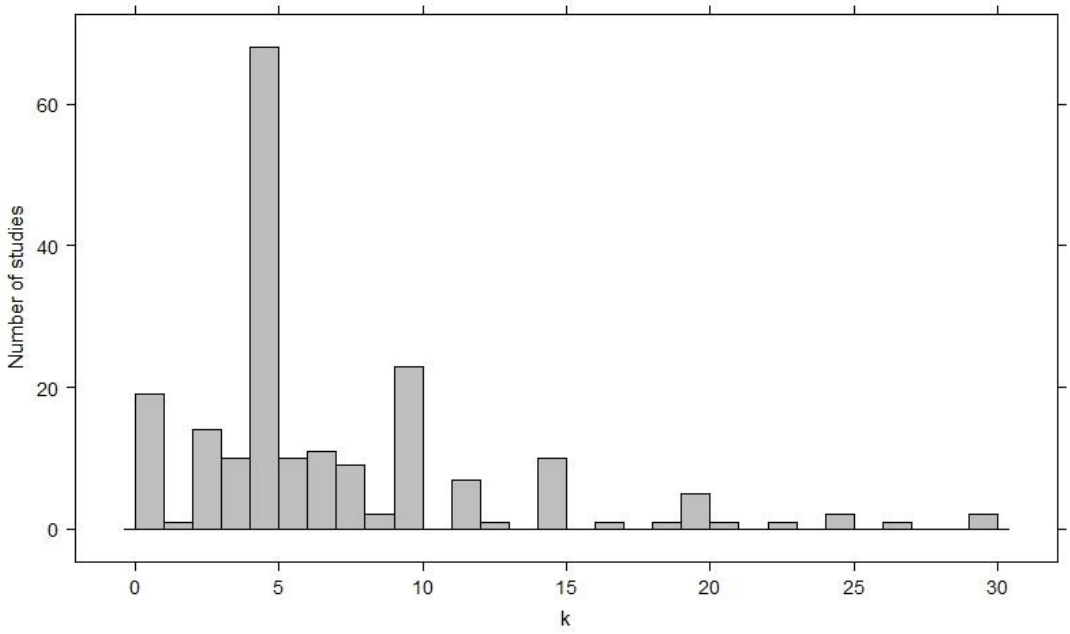


Figure 5: distribution of numbers of k's in the investigated studies.

Accuracy assessment

For continuous response variables, the most commonly used measures for assessing the accuracy of k-NN predictions were root mean square error (RMSE) and the standard error (SE) of estimates, in both absolute and relative terms (72%). Other measures included the correlation coefficient (r), coefficient of determination (R2), and both absolute and relative mean deviation between observations and predictions (3.3%). For categorical response variables, overall accuracy or the Kappa index of agreement were most often used (9.2%). From a spatial scale perspective, accuracy assessments were conducted for aggregations of pixel-level estimates (e.g., plots, compartments) for almost 65% of studies, whereas more than the 26% of accuracy assessments were at the pixel-level. A few papers (3.4%) reported confidence intervals for estimates of population parameters.

A subset of 110 of the 260 studies was used to assess the accuracy of k-NN predictions using RMSE% which expresses RMSE as a percentage of the estimated mean, or occasionally the observed mean. For these studies, RMSE% varied between 0.3% and 170.6%, with an average of approximately 33% (standard deviation

of ± 29.3%). Because average RMSE% at pixel-level scale and aggregated for larger scales areas were generally comparable (39% and 31%, respectively), we do not distinguish between them in the following analyses.

Independently of other factors such as response variable, value of k, and distance metric, the most accurate results were obtained when the source of feature variables was high resolution aerial imagery (15 studies with average RMSE% of 26% ± 18.7%) or fused multi-sensor data (27 studies with average RMSE% of 27% ± 19.6%), followed by synthetic aperture radar (SAR) and lidar-based airborne laser scanning (ALS). Independently of other factors, the average RMSE% for estimating basal area (9 studies) was 23% (± 22.4%), for tree density (2 studies) was almost 25% (± 12.7%), while for growing stock volume/biomass/carbon (81 studies) the average RMSE% was 37% (± 31.6%).

The choice of distance metric did not substantially influence estimation performance, even though four studies using the canonical correlation analysis metric and nine studies using the Mahalanobis metric reported smaller RMSE% results (18.78% ± 11.4% and 21.49% ± 22.9% respectively) than the average for studies using the Euclidean metric (RMSE% of almost 35% ± 31.4%). Of importance, however, studies using the Euclidean metric are more numerous (77) and represent a much broader range of forest types, response variables and feature variables.

To better understand the relationship between k-NN prediction accuracies, feature variables and reference set size, we selected a homogeneous and comparable subset of 53 studies that used growing stock volume/biomass/carbon as the response variable, a single type of feature variable, and RMSE% as the measure of prediction accuracy. For studies that used only feature variables based only on aerial photography, the median reference set size of 332 produced median RMSE%=48.5; for studies that used only satellite-based spectral feature variables, reference set sizes less than 1000 (median 280) produced median RMSE%=40.0, whereas reference set sizes of 1000 or greater (median 3117) produced median RMSE%=15.81; for studies that used only lidar-based feature variables, the median reference set size of 124 produced median RMSE%=31.26; and finally for studies that used only radar-based feature variables, median reference set size of 300 produced median RMSE%=44.8. This subset of studies confirms that large reference sets are necessary to obtain small RMSE% when using satellite spectral data. However, these results also depend to some degree on the complexity and diversity of the forests; in particular, most of the studies using satellite spectral feature variables were for somewhat less complex boreal and northern temperate forests. These studies also confirm the growing popularity of lidar-based feature variables for obtaining small RMSE%, even with small reference sets.

Over all 260 studies, transforming raw remotely sensed input feature data using dimension reduction techniques such as principal component analysis (PCA) or band ratio indices such as the normalized difference vegetation index (NDVI) had no effect on accuracy: the average RMSE% for the 80 studies using transformations was 41.5% (± 32.7%) contrasted with average RMSE% of 36.8% (± 32.8%) for the 33 studies that used no transformations.

Although for some studies, accuracy tended to decrease with greater values of k, generally we found any type of relationship between the final accuracy and the value of k selected as optimal.

Summary of meta-analysis

The k-NN technique is well-affirmed for integrating forest inventory field data and remotely sensed data. In this bibliographic meta-analysis we analyzed 260 experimental tests published in 148 papers since 1981, all dealing with application of k-NN for a variety of forest environments across the world. In the 1990s and the 21st century most studies were from Finland and the northern parts of the USA; since then, the technique has been successfully tested and applied in 26 countries on six continents.

The year 2009 had the largest number of published k-NN papers. Subsequently, the number of publications decreased, perhaps because the technique had become well-known and well-documented. Nowadays, the technique is widely adopted, but it is less frequently a direct research objective. However, other papers not analyzed for this study because they did not provide details on methods or the k-NN configuration, reported using the k-NN technique as a standard procedure for constructing spatially contiguous forest information to be used separately or integrated with other information for further geographical analysis.

More than half the k-NN studies had growing stock volume, wood biomass or its carbon content as the response variable. However, k-NN was also applied for estimating a very large number of other continuous and categorical response variables related to functional attributes such as leaf area index or structural attributes such as basal area or trees per unit area, or to levels of ecosystem disturbance such as defoliation.

Expected results

Because the meta-analysis did not reveal a particular k-NN configuration that could be considered optimal for all the cases, a reasonable conclusion is that an optimization to calibrate any configuration phase using the available data should be considered. However, some guidelines may be proposed. Firstly, the selection of feature variables is related to the geographical resolution of the analysis (pixel size) and the availability of remotely sensed data. Almost half the studies used Landsat imagery alone, but the majority of the studies augmented Landsat data with data from other similar multispectral satellites such as SPOT, IRS and ASTER. Studies with finer geometric resolution were frequently based on aerial photography while those at coarser resolution were frequently based on MODIS images. The advent of ALS technology since 2008 has changed this trend with the number of studies using ALS metrics alone or integrated with optical imagery increasing rapidly.

Secondly, the simplest distance metric, Euclidean distance, was most commonly used (70%), presumably because of its simpler coding implementation and because more complex metrics did not produce better results. For values of k, the majority of the studies arbitrarily adopted a value of approximately 5 as a compromise between larger values that may produce greater accuracy and smaller values that tend to retain better the reference set variability and are computationally less intensive.

Thirdly, in terms of accuracy, the k-NN technique may be expected to produce RMSE% in the range of 20-40% for common response variables (69% of cases RMSE% < 40%). However, despite the great variety of k-NN configurations and local environmental conditions, only 22% of studies reported successful use of k-NN

with RMSE% less than 10%, while a small number of studies reported unsuccessful uses with RMSE% as great as 100% or more (Figure 6).

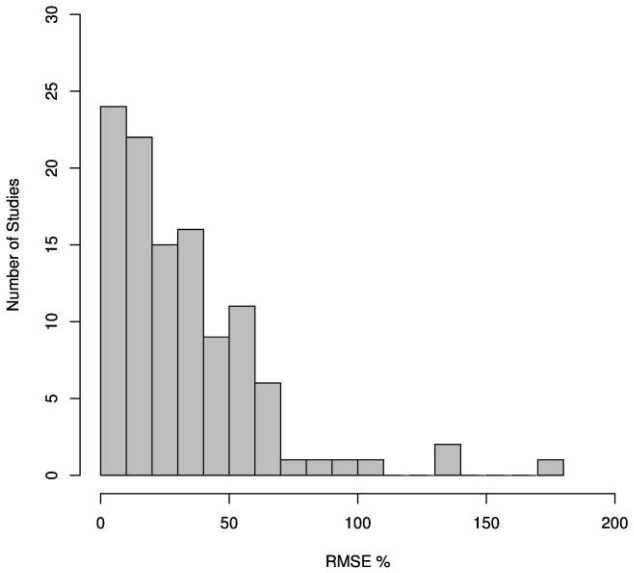


Figure 6: distribution of the accuracies of 110 studies where the RMSE% of the k-NN estimates was reported. The average is 33%.

Of interest, relative to the results of experimental tests conducted in the 1990s, the scientific community in the period 2000-2014 was able to achieve a better tuning of the k-NN method that then produced increasingly more accurate results than in prior years (Figure 7). This result can probably be attributed to greater visibility of k-NN studies via articles published in the scientific literature and presentations at international conferences (e.g., the ForestSAT series conferences and the k-NN workshops). In addition, greater prediction accuracy can also be at least partially attributed to the recent availability of remotely sensed data of greater quality from both passive and active sensors and to improved global navigation satellite systems for the geolocation of field inventory plots.

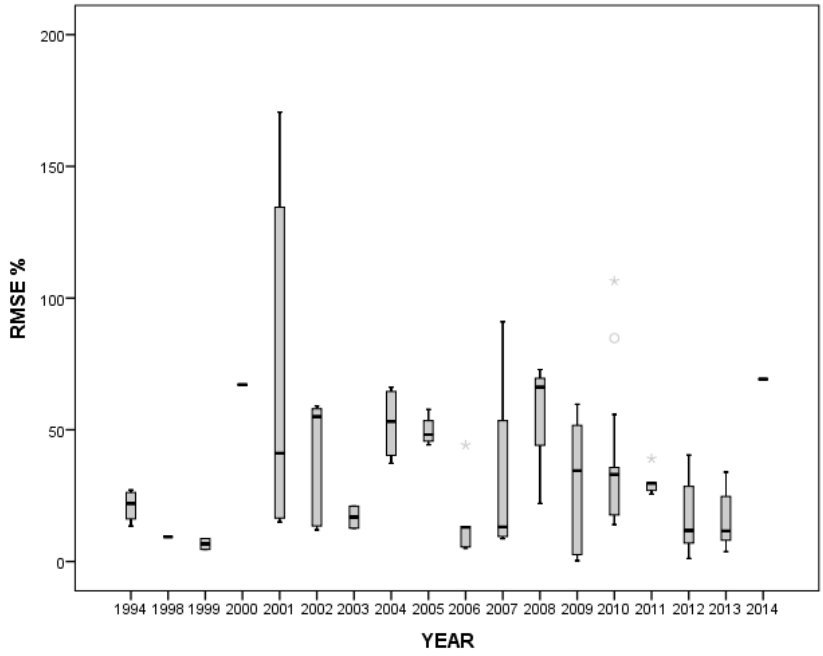


Figure 7: RMSE% distribution of the accuracies for k-NN studies for the different years.

Fourthly, the vegetation characteristics of the study areas where k-NN was used did not substantially affect estimation performance. For boreal study areas which had the greatest number of studies (59), the average RMSE% was 35%, while the average RMSE% for other vegetation zones ranged from 6% for temperate mountain systems (only 5 studies) to 42% for temperate oceanic forests (15 studies) (Figure 8).

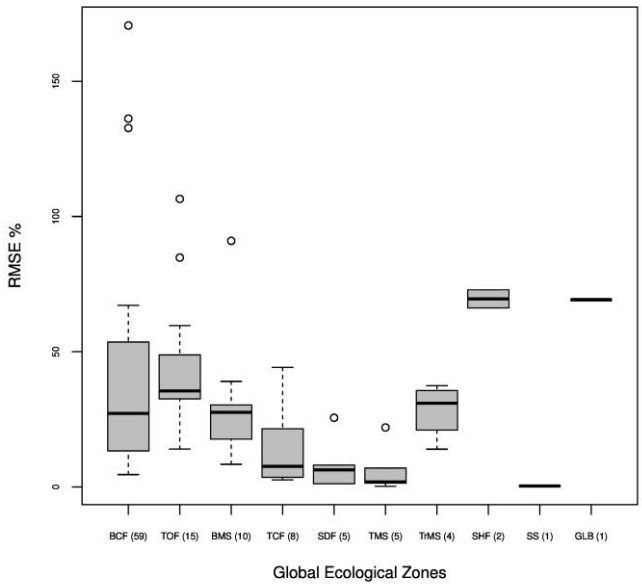


Figure 8: RMSE% distribution of the accuracies for k-NN studies in different Global Ecological Zones (FAO, 2012). SS: Subtropical steppe, TMS: Temperate mountain systems, SDF: Subtropical dry forest, TCF: Temperate continental forests, TrMS: Tropical mountain systems, BMS: boreal mountain systems, BCF: boreal coniferous forests, TOF: Temperate oceanic forests, GLB: global, SHF: Subtropical humid forests.

Fifthly, the accuracy of k-NN predictions was not greatly influenced by the selection of the feature variables. Accuracies obtained using feature variables from sensors other than Landsat were all within the range of accuracies using Landsat-based feature variables (Figure 9). However, Landsat-based studies produce a wide range of results, presumably because they were the earlies and consequentially the largest in number. The recent advent of ALS feature variables may substantially alter this finding; in particular, ALS metrics, alone or in conjunction with optical imagery, seem to be among the most promising feature variables for producing k-NN preditions.

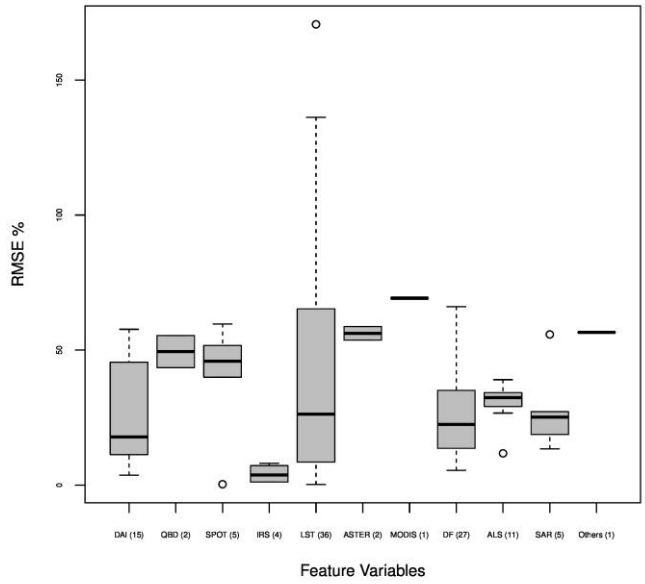


Figure 9: RMSE% distribution of the accuracies for k-NN studies based on feature variables from different remotely sensed data. IRS: Indian for Remote Sensing, DAI: digital aerial imagery, DF: data fusion from different sensors, SAR: synthetic aperture radar, SPOT: *Satellite Pour l'Observation de la Terre,* LST: Landsat, QBD: Quick Bird, ASTER: ASTER, MODIS: MODIS Terra and Aqua.

Sixthly, on the basis of the meta-analysis over all the considered 260 studies, we found that field reference set sizes did not affect the accuracy of k-NN predictions. Thus, the k-NN technique can be successfully used from local scales through to continental-level investigations.

Methodological advances

Citations and meta-analyses are useful for documenting the scope, range, geographic distribution, and history of k-NN forest applications. Further, they are useful for determining the most frequently cited papers as a measure of information sources that others have found useful and relevant. Table 1 reports the most frequently cited papers published during or before 2013 based on the number of citations standardized to reflect elapsed time since publication.

As noted in previous sections, the vast majority of published papers focused on specific applications whereas only a much smaller number focused on methodological issues. For the purposes of assessing methodological advances, a limited e-mail survey was conducted to augment the literature review and meta-analysis. The survey solicited opinions from k-NN users and researchers regarding methodological publications that have advanced the state of the science for forest applications. The following summary represents an admittedly subjective consensus of the results of the survey.

Table 1. Nearest neighbors articles ranked by number of citations normalized by years*.

Rank	Average number of citations per year	Publication
1	16	Ohmann et al. (2002)
2	15	Franco-Lopez et al. (2001)
3	14	Crookston et al. (2008)
4	11	McRoberts & Tomppo (2007)
5	11	McRoberts et al. (2007)
6	9	Tomppo & Halme (2004)
7	8	Katila & Tomppo (2001)
7	8	McRoberts et al. (2002)
8	7	Tomppo et al. (2002)
9	6	Tokola et al. (1996)
10	6	Trotter et al. (1997)
11	5	Pretzsch (1997)
12	5	Fazakas et al .(1999)
13	4	Maltamo & Kangas (1998)

*Ranked by Scopus citations.

Optimization

Optimization of the k-NN technique entails selecting a distance metric and values for k and t. Nearly all optimization efforts have focused on selecting or formulating a distance metric with much less effort focused on optimizing k and t. Although many metrics have been proposed, the unweighted Euclidean distance and the metric based on canonical correlation analysis have been used most widely. The unweighted Euclidean distance metric, which can be expressed using an identity matrix in Eq. (2), is the simplest, most intuitive, and easiest to implement. The only degree of optimization for this metric pertains to the particular feature variables to be used. Selection of feature variables can be accomplished by comparing all combinations of all numbers of feature variables (McRoberts, 2012), stepwise selection (Chirici et al., in review), or use of genetic algorithms (Tomppo & Halme, 2004; Tomppo et al., 2009).

The canonical correlation analysis metric was proposed by Moeur and Stage (1995) and has been used fairly widely (e.g., Maltamo et al., 1998, 2003; LeMay & Temesgen, 2005). The metric is based on an optimal relationship between a linear combination of response variables and a linear combination of feature variables. The configuration consisting of this metric and k=1 has been characterized as the Most Similar Neighbor approach (Moeur & Stage, 1995). The metric based on canonical correspondence analysis was proposed by Ohmann and Gregory (2002). The configuration consisting of this metric and k=1 has been characterized as

the Gradient Nearest Neighbor approach. However, both metrics have also been used with k>1 (e.g., Maltamo et al., 2003; Ohmann et al., 2014).

The weighted Euclidean distance metric, which can be expressed using a diagonal matrix in Eq. (2), is similar to the unweighted metric, albeit with possibly unequal values on the diagonal of the matrix (Tomppo & Halme, 2004). Because optimization of this metric is computationally intensive, its use has not been reported sufficiently frequently to appear in lists of most cited publications; nevertheless, recent technological advances suggest it has considerable optimization potential. Optimization entails selection of the diagonal values, one for each feature variable. For a large number of feature variables, optimization can be excessively computationally intensive. A two-step alternative is to first select a small subset of feature variables and then to optimally select the corresponding diagonal values of the matrix corresponding to the selected feature variables. The first step can be accomplished using the same methods noted for the unweighted Euclidean metric, while the second step would typically be accomplished using a genetic algorithm (Tomppo & Halme, 2004).

Little attention has been devoted to optimizing the k-NN technique with respect to selection of the k and t values. Small values of k, typically in the range $1 \leq k \leq 10$, have been common along with arbitrary selections of t=0, t=1, and occasionally t=2. Optimization of the distance metric followed by arbitrary selections of k and t may be self-defeating; specifically, the beneficial effects of optimizing the distance metric may be mitigated by the adverse effects of arbitrary selections of k and t. McRoberts (2012) and McRoberts et al. (2015) are the only known attempts to optimize the distance metric, k, and t simultaneously.

Inference

Multiple variations of nearest neighbors techniques have been shown to be useful and effective for both prediction and mapping. However, the ultimate judgment is whether these techniques can contribute to the construction of an inference in the form of a confidence interval for a population parameter. Two modes of inference are common, design-based inference and model-based inference. McRoberts et al. (2002) demonstrated the utility of nearest neighbors techniques as a key component of design-based, stratified estimation, and Baffetta et al. (2009, 2011) demonstrated its utility for design-based, model-assisted estimation. For model-based inference, McRoberts et al. (2007) derived parametric estimators for means and variances and demonstrated their inferential utility. To circumvent the complexity and computational intensity associated with parametric estimators, McRoberts et al. (2011) proposed and demonstrated a bootstrapping approach to model-based variance estimation. Magnussen et al. (2009) noted and McRoberts (2015) confirmed that parametric estimates may be unreliable for $k \leq 7$. In the latter case, the bootstrapping approach should be used.

Enhancements

Little attention has been devoted to developing or using comprehensive sets of k-NN diagnostics. Common diagnostics include comparisons of RMSEs for different combinations of feature variables and graphs

of observations versus predictions for assessing quality of fit. McRoberts (2009, 2012) and McRoberts et al. (2015) proposed additional diagnostics for identifying influential outliers in the reference set and for assessing the degree to which gains achieved by optimization in the reference set are realized in the target set.

Several approaches have been proposed to improve the accuracy of k-NN predictions. Katila and Tomppo (2002) introduced constraints on the selection of neighbors as a means of increasing accuracy. The constraints include a maximum distance in geographic space between target and reference units and restriction of neighbors to the same land use stratum as the stratum of the target unit. Tomppo and Halme (2004) augmented feature space with additional variables whose values were constructed as interpolations of field plot observations over the target space.

Beyond simply reducing the dimension of feature space by deleting some auxiliary variables, selecting optimal diagonal values for the weighted Euclidean distance matrix can be an extremely computationally intensive task, even for small numbers of feature variables. Tomppo and Halme (2004) proposed genetic algorithms for this purpose. Genetic algorithms are heuristic search procedures that theoretically converge to optimal or near optimal solutions. The search mechanism mimics natural selection using techniques inspired by natural evolution such as inheritance, mutation, selection, and crossover.

A unique feature of k-NN techniques is that no prediction can be smaller than the smallest reference set observation nor larger than the largest reference set observation. Therefore, unlike regression models, k-NN techniques cannot extrapolate predictions beyond the range of the response variable in the reference set, even if ranges of feature variables in the target set may be greater than ranges in the target set. Magnussen et al. (2010) developed an approach that uses a local linear model to extrapolate predictions beyond the range of reference set feature variables.

2.2.4. Discussions and Conclusions

The study was motivated by the popularity of the k-NN technique for use with forest inventory and remotely sensed data. The analyses were conducted within the framework of Working Group 2 of COST Action FP1001 and focused on a review of the scientific literature for forestry applications. Information available in the scientific publications was used to populate a database that served as the basis for a meta-analysis.

Multiple conclusions drawn from previous experiences are useful as background for future forest implementations and to stimulate future research. Nowadays, k-NN can be considered a useful, well-affirmed technique for a broad scope of international forest inventory applications using remotely sensed data. Previous experiences have demonstrated that k-NN can be used in all vegetation zones and at spatial scales ranging from local applications based on a limited number of field observations to large national and continental applications. The k-NN technique is frequently used to estimate growing stock volume, tree biomass, and carbon stock, but it can be used also with large variety of response variables.

Feature variables can be derived from the outputs of traditional multi-spectral optical sensors as well as active radar and ALS systems. Performance of the k-NN technique is dependent on an optimization phase

aimed at selecting the set of feature variables and their weights, the value of k, neighbor weighting, and the multidimensional distance metric.

The meta-analysis revealed that the simplest k-NN configuration frequently produced excellent results. For this reason we suggest a first tentative implementation of k-NN using the Euclidean distance metric, values of k ranging between 3 and 10, and t=0 t=1, and t=2. If the pixel-level or area-level k-NN predictions produce RMSE% less than 30%, the result should be considered in line with results reported in the literature. Starting from this simple k-NN configuration, multiple diagnostic tests and investigations of variations of the parameters and feature variables can be conducted to obtain more accurate results. Examples include transformations of feature variables, more sophisticated distance metrics, and optimization of k and t. There is no consensus regarding the order of the optimization steps that produce the greatest gains in accuracy. A reasonable order would be first to select the feature variables that reduce the effects of the curse of dimensionality by eliminating redundant feature variables. Then, depending on the nature of the feature variable set, more sophisticated distance metrics such as the weighted Euclidean or canonical correspondence analysis metrics can be tested. Finally the k value should be optimized on the basis of the leave-one-out approach.

Integration of remotely sensed data from different sensors, particularly both passive and active sensors, and non-remote sensing variables (such as elevation or climatic maps) can be considered to facilitate more accurate prediction of both structural and functional forest attributes.

Finally, we strongly encourage scientists using the k-NN technique to report results obtained for all multiple configurations considered in the optimization phase. Do so will facilitate future meta-studies and both researchers and practitioners involved in the operational implementation of the k-NN technique.

2.2.5. References

Arnqvist, G., & Wooster, D. (1995). Meta-analysis: Synthesizing research findings in ecology and evolution. Trends in Ecology and Evolution, 10 (6), pp. 236-240.

Baffetta, F., Corona, P., & Fattorini, L. (2011). Design-based diagnostics for k-NN estimators of forest resources. Canadian Journal of Forest Research, 40(1), 59-72.

Baffetta, F., Fattorini, L., Franeschi, S., & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment ,113(3), 463-475.

Chirici, G., McRoberts, R.E., Fattorini, L., Mura,M., & Marchetti, M. (in review). Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework, Remote Sensing of Environment (in review).

COST, (2014). European Cooperation in Science and Technology. Available at: http://www.cost.eu/; last accessed 5 October 2015.

COST FP1001 (2014). USEWOOD. Improving data and information on potential supply of wood resources, a European approach from multisource national forest inventories. Available at: https://sites.google.com/site/costactionfp1001/. (Last accessed October 2015).

Crookston, N.L., & Finley, A.O. (2008). yaImpute" An R Package for kNN Imputation. Journal of Statistical Software, 23(10), 1-16.

Eskelson, B.N.I., Temesgen, H., Lemay, V., Barrett, T.M., Crookston, N.L., Hudak, A.T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. Scandinavian Journal of Forest Research, 24,235-246.

Fazakas, Z., Nilsson, M., & Olsson, H. (1999). Regional forest biomass and wood volume estimation using satellite data and ancillary data. Agricultural and Forest Meteorology, 98-99, 417-425.

Finley, A.O., & McRoberts, R.E. (2008). Efficient k-nearest neighbor searches for multi-source forest attribute mapping. Remote Sensing of Environment, 112, 2203-2211.

Fix, E.,& Hodges, J.L. (1951). Discriminatory analysis. Nonparametric discrimination: consistency properties. Report Number 4, Project 21-49-004. USAF School of Aviation Medicine. Randolph Field, Texas. Reprinted as: Fix, E., and Hodges, J.L. (1989). Discriminatory analysis – nonparametric discrimination: consistency properties. International Statistical Review, 57, 238-247.

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

Garbulsky, M.F., Peñuelas, J., Gamon, J., Inoue, Y., & Filella, I. (2011). The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies. A review and meta-analysis. Remote Sensing of Environment, 115(2), 281-297.

Beaudoin, A.; Bernier, P.Y.; Guindon, L.; Villemaire, P.; Guo, X.J.; Stinson, G.; Bergeron, T.; Magnussen, S.; Hall, R.J. 2014. Mapping attributes of Canada's forests at moderate resolution through kNN and MODIS imagery. Canadian Journal of Forest Research, 44, 521-532.

Katila, M., & Tomppo, E. (2002). Stratification by ancillary data in multisource forest inventories employing k-nearest neighbor estimation. Canadian Journal of Forest Research, 32(9), 1548-1561.

Kilkki, P., & Päivinen, R. (1987) Reference sample plots to combine field measurements and satellite data in forest inventory. Department of Forest Mensuration and Management, University of Helsinki. Research Notes 19, 210-215.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science, 51(2), 109-119.

Maltamo, M., & Kangas, A. (1998). Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Canadian Journal of Forest Research, 28(8), 1107-1115.

Maltamo, M., Malinen, J., Kangas, A., Härkönen, S. & Pasanen, A.-M. (2003). Most similar neighbour based stand variable estimation for use in inventory by compartments in Finland. Forestry, 76, 449-464.

Magnussen, S., McRoberts, R.E., & Tomppo, E.O. (2009). Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. Remote Sensing of Environment 113, 476-488.

Magnussen, S., Tomppo, E., & McRoberts, R.E. (2010). A model-assisted kNN approach to remove extrapolation bias. Scandinavian Journal of Forest Research, 25, 174-184.

Metinfo. (2007). Metinfo-forest information services, Finnish Forest Research Institute. Available at: http://www.metla.fi/metinfo/index-en.htm. (Last accessed, October 2015).

Metla, 2013. File service for publicly available data. Available at: http://kartta.metla.fi/index-en.html. (Last accessed: October 2015).

Moeur, M., & Stage, A. R. (1995). Most similar neighbor: an improved sampling inference procedure for natural resource planning. Forest Science, 41, 337-359.

McRoberts, R.E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment, 113, 489-499.

McRoberts, R.E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management, 272, 3-12.

McRoberts, R.E., Cohen, W.B., Næsset, E., Stehman, S.V., & Tomppo, E.O. (2010). Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scandinavian Journal of Forest Research, 25, 340-367

McRoberts, R.E., Næsset, E., & Gobakken, E. (2015). Optimizing the k-Nearest Neighbors technique for estimation of forest aboveground biomass using airborne laser scanning data. Remote Sensing of Environment, 168, 252-264.

McRoberts, R.E., Magnussen, S., Tomppo, E.O., & Chirici, G. (2011). Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sensing of Environment, 115, 3165-3174.

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

McRoberts, R.E., Tomppo, E.O., Finley, A.O., & Heikkinen, J. (2007). Estimating areal means and variances using the k-nearest neighbors technique and satellite imagery. Remote Sensing of Environment, 111, 466-480.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor - an improved sampling inference procedure for natural resource planning. Forest Science, 41(2), 337–359.

Ohmann, J.L., & Gregory, M.J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon. U.S.A. Canadian Journal of Forest Research 32, 725-741.

Ohmann, J.L., Gregory, M.J, & Roberts, H.M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. Remote Sensing of Environment, 151, 3-15.

Pretsch, H. (1997). Analysis and modeling of spatial stand structures. Methodological considerations based on mixed beech-larch stands in Lower Saxony. Forest Ecology and Management, 97(3), 237-253.

Poso, S. (1972). A method of combining photo and field samples in forest inventory. Communicationes Instituti Forestalis Fenniae, 76, 1-133.

Reese, H., Granqvist-Pahlén, T., Egberth, M., Nilsson, M., & Olsson, H. (2005). Automated estimation of forest parameters for Sweden using Landsat data and the kNN algorithm. Proceedings of the 31st International Symposium on Remote Sensing. June 20-24, 2005. St. Petersburg, Russia. Available at: http://www.isprs.org/proceedings/2005/ISRSE/html/papers/652.pdf. (Last accessed: October 2015).

Short, R., & Fukunaga, K. (1981). The optimal distance measure for nearest neighbor classification. IEEE Transactions on Information Theory, 27(5), 622-627.

SLU Forest Map (2013). Available at: http://www.slu.se/en/collaborative-centres-and-projects/swedish-national-forest-inventory/services-and-products/interactive-services/knn-sweden/. (Last accessed: October 2015).

Tokola, T., Pitkanen, J., Partinen, S., & Muinonen, E. (1996). Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. International Journal of Remote Sensing,17, 2333-2351.

Tomppo, E. (1990). Designing a satellite image-aided national forest survey in Finland. In Proceedings: The usability of remote sensing for forest inventory and planning. Swedish University of Agricultural Sciences, Remote Sensing Laboratory. Report 4, 43-47.

Tomppo, E.. (1991). Satellite image-based national forest inventory of Finland. In: Proceedings of the symposium on global and environmental monitoring, techniques and impacts. 17-21 September 1990.

Victoria, British Columbia, Canada. International Archives of Photogrammetry and Remote Sensing, 28, 419-424.

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of k-NN estimation: a genetic algorithm approach. Remote Sensing of Environment, 92, 1-20.

Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. Remote Sensing of Environment, 82(1), 156-171.

Tomppo, E., Gagliano, C., De Natale, F. Katila, M., & McRoberts, R.E. (2009). Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. Remote Sensing of Environment, 113, 500-517.

Tomppo, E., Haakana, M., Katila, M., & Peräsaari, J. (2008). Multi-source national forest inventory – Methods and applications. Managing Forest Ecosystems 18. Springer. 374 p.

Trotter, C. M., Drymond, J. R., & Goulding, C. J. (1997). Estimation of timber volume in a coniferous plantation forest using Landsat TM. International Journal of Remote Sensing, 18, 2209-2223.

Wilson, B.T., Lister, A.J., & Riemann, R.I. (2012). A nearest neighbor imputation approach to mapping tree species over large areas using forest inventory plots and moderate resolution raster data. Forest Ecology and Management, 271, 182-198.

Wilson, B.T., Lister, A.J., & Riemann, R.I. (2013a). Live tree species basal area of the contiguous United States (2000-2009). Newtown Square, PA: USDA Forest Service, Northern Research Station. Available at: http://www.fs.usda.gov.rds/archive/Product/RDS-2013-0013. (Last accessed: October 2015).

Wilson, B.T., Woodall, C.W., & Griffith, D.M. (2013b). Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. Carbon Balance and Management 2013, 8, 1.

Wilson, B.T., Lister, A.J., & Riemann, R.I. (2013c). Forest carbon stocks of the contiguous United States (2000-2009). Newtown Square, PA: USDA Forest Service, Northern Research Station. Available at: www.fs.usda.gov/rds/archive/Product/RDS-2013-0004. (Last accessed: October 2015).

Zolkos, S.G., Goetz, S.J.,& Dubayah, R. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment, 128, 289-298.

2.3. Study III

From:

[1]McRoberts R. E., [2]Næsset E., [2]Gobakken T., [1]Domke G. M., [3]Chirici G., [3]Mura M., [4]Chen Q. (in review). The benefits of optimizing nearest neighbor configurations for lidar-assisted estimation of forest volume and biomass. *Remote Sensing of Environment*.

[1] Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota USA

[2] Department of Ecology and Natural Resource Management, Norwegian University of Life Sciences, Ås, Norway

[3] Department of Agricultural, Food and Forestry Systems, Università degli Studi di Firenze Firenze, Italy

[4] Department of Geography, University of Hawai'I, Honolulu, Hawai'i, USA

Abstract

The k-Nearest Neighbors (k-NN) technique is a non-parametric approach that calculates predictions as linear combinations of observations for sample units that are nearest in a space of auxiliary variables to the population unit for which a prediction is desired. When implementing a nearest neighbors algorithm, four choices are necessary: a distance metric, the specific auxiliary variables to be used, the number of nearest neighbors, and a scheme for weighting the nearest neighbors. Often the algorithm is implemented with arbitrary choices such as the Euclidean distance metric, all auxiliary variables, k=1 or k=5, and equal weighting of neighbors. However, as the k-NN technique has matured, methods for optimizing the four choices have begun to emerge, although few reports of rigorous comparisons of optimization methods or the benefits of optimization have been reported. The objective was to compare optimization methods with respect to the accuracy of airborne laser scanning-assisted predictions of forest volume or biomass and with respect to inferences for population mean of volume or biomass per unit area. Four study areas were used, two in Norway, one in Italy, and one in the United States of America. The primary results were twofold: first, with appropriate optimization, multiple methods produced similar predictions; and second, optimization produced considerably greater precision for estimates of population means than common arbitrary choices. Therefore, as the k-NN technique continues to mature, users are under greater obligation to justify decisions not to optimize and to assess the consequences of those decisions.

Keywords: k-Nearest neighbors, optimization, airborne laser scanning, accuracy assessment

2.3.1. Introduction

Nearest neighbors techniques are non-parametric, multivariate approaches to estimation. Population unit predictions are calculated as linear combinations of sample observations for units designated neighbors that are nearest or most similar in a space of auxiliary variables to units for which predictions are desired. Nearest neighbors techniques have received considerable attention for mapping and areal estimation of forest attributes, particularly when used with forest inventory and remotely sensed data. Application of a nearest neighbors algorithm requires choices for the distance or similarity metric, the particular auxiliary variables, the number of neighbors, and a scheme for weighting the neighbors.

Most efforts to optimize nearest neighbour algorithms focus on the distance metric, although only a few comparisons of distance metrics have been reported. For predicting forest stand attributes using variables obtained from aerial photography, LeMay and Temesgen (2005) reported that a metric based on canonical correlation analysis was superior to Euclidean distance and Manhattan distance. For predicting forest attributes from Landsat-based variables, Chirici et al (2008) reported that distance metrics giving greater weights to reference units whose response variable observations are closer to the mean of the observations was superior to Euclidean, Mahalanobis, and two other metrics that weight feature variables with respect to relationships with the response variables. Latifi et al. (2010 compared Euclidean, Mahalanobis, canonical correlation, and Random Forests distance metrics for predicting forest volume and biomass using lidar, Landsat, and aerial image data. The results were mixed with different metrics producing more optimal results for different response variables. The only general conclusion that can be drawn from these studies is that metrics that are optimized using observations of the response variable tend to produce the most accurate predictions.

The number of nearest neighbors, k, is often arbitrarily selected as k=1 or k=5 but may also be selected to optimize a criterion such as root mean square error. Neighbors are often equally weighted although they are also often weighted inversely to the distance or distance squared between units requiring predictions and sample units.

No reports are known of comprehensive efforts to optimize a nearest neighbors algorithm by comparing and simultaneously selecting distance metrics, auxiliary variables, number of neighbors, and neighbor weighting scheme. The objective of the study was to compare combinations of levels for these four factors with respect to the accuracy of airborne laser scanning (ALS)-based predictions of response variables forest volume or biomass and with respect to inferences for the population mean per unit area for the response variables. Data were used for four study areas, two in Norway, one in Italy, and one in the United States of America (USA).

2.3.2. Materials and Methods

Study areas

Hedmark, Norway

The 1259-km^2 study area was mostly in the municipalities of Åmot and Stor-Elvdal in Hedmark County, Norway (Figure 1). Dominant tree species are Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.). Field measurements were acquired for 250-m^2 Norwegian NFI field plots located at the intersections of a 3-km x 3-km grid (Tomter et al., 2010). Data for only the 145 plots measured within one year of the ALS acquisition dates were used for this study. Thus, the study area was defined as the geographic area represented by the portion of the Latin Square sampling design used by the Norwegian NFI inventoried between 2005 and 2007 (Figure 1). On each plot, all trees with diameters at-breast-height (dbh, 1.3 m) of at least 5 cm were callipered. Tree heights were measured on an average of 10 sample trees per plot selected with probability proportional to stem basal area, and heights for the remaining trees were predicted using height-dbh models (Fitje & Vestjordet 1977; Vestjordet 1968). The volume of each sample tree was estimated using species-specific volume models with dbh and either measured height or predicted height as independent variables (Braastad 1966; Brantseg 1967; Vestjordet 1967). The ratio of the mean volume estimate for trees with predicted heights and the mean volume estimate for trees with measured heights was used to adjust the former volume estimates. Volume estimates for individual trees were added to produce plot-level totals which were then scaled to a per unit area basis (m^3/ha) and considered to be observations without error (McRoberts & Westfall, 2014).

Wall-to-wall airborne lidar data were acquired between 15 July 2006 and 12 September 2006 with average point density of 0.7 pulses m^{-2}. Data for only single echoes or the first of multiple echoes were used. For each plot and population unit, height distributions were estimated for first echoes from tree canopies, i.e. heights greater than 2 m. Echoes with heights less than 2 m were considered to have been reflected from non-tree objects such as shrubs, grass, or the ground. For each plot and population unit, heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions were calculated and denoted h_1, h_2, …, h_{10}, respectively. Canopy densities were calculated as the proportions of echoes with heights greater than 0%, 10%, …, 90% of the range between 2 m above ground and the 95th height percentile and were denoted d_0, d_1, …, d_9, respectively (Gobakken & Næsset, 2008).

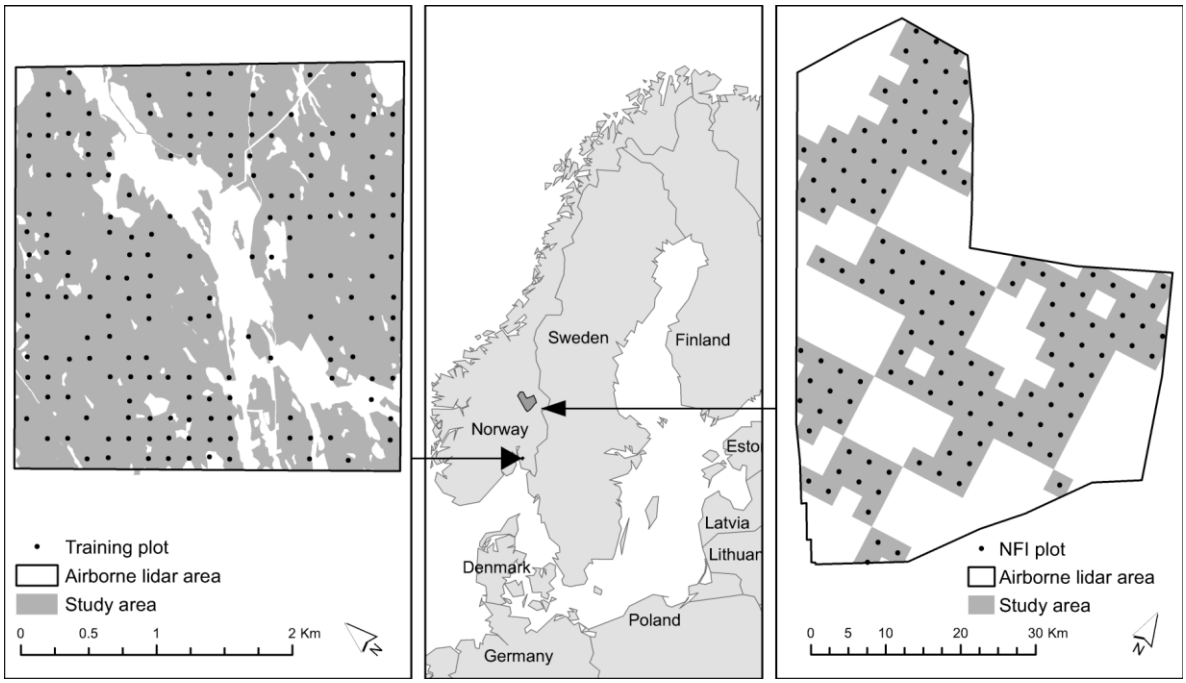


Figure 1. Våler (left) and Hedmark (right) study areas in Norway.

Våler, Norway

The 8.53-km^2 study area was located in a boreal forest region in Våler Municipality in southeastern Norway (Figure 1). Norway spruce (*Picea abies* (L.) Karst.) and Scots pine (*Pinus sylvestris* L.) are the dominant species, with younger stands having large proportions of deciduous species. Forests in the study area are actively managed with clear-cutting and commercial thinning on productive sites and selective logging on poor sites.

Measurements were obtained for 176 systematically-distributed, circular, 200-m^2 forest inventory plots. Tree-level AGB was estimated for 1999 using statistical models based on field observations of species and measurements of diameter at-breast-height (1.3m) and height (Marklund, 1988). Plot-level AGB was estimated as the sum of individual tree AGB predictions, scaled to a per unit area basis (Mg/ha), and considered to be observations without error (McRoberts & Westfall, 2014).

Wall-to-wall ALS data were acquired for the study area with pulse density of approximately 1.2 pulses per m^2. Distributions of first echo heights were constructed for the 200-m^2 plots and 200-m^2 square cells that tessellated the study area. A threshold of 1.3 m above the ground surface was used to remove the effects of echoes from ground vegetation whose biomass is not included in tree-level AGB. For each plot and cell, heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions were calculated as were canopy densities calculated as the proportions of echoes with heights greater than 0%, 10%, …, 90% of the range between 1.3 m above ground and the 95th height percentile (Gobakken & Næsset, 2008). Næsset et al. (2013) provide more details for the study area and the dataset.

Molise, Italy

The 363.6-km^2 study area is in the southwestern part of Molise Region in central Italy (Figure 2). Approximately 56% of the area, or 20,518 ha is covered by forests of which approximately 60% is dominated by deciduous oaks (*Quercus cerris, Quercus pubescens*), approximately 18% is dominated by hop hornbeam (*Ostrya carpinifolia*), and approximately 9% is dominated by unmanaged beech (*Fagus sylvatica*) forests with structures approaching natural, old-growth forest status.

The study area was tessellated into 437 hexagons, each with area of 1 km^2. A point was randomly selected in each hexagon and classified as "forest" or "non-forest" based on interpretation of high-resolution aerial ortho-photography. From the 197 points classified as forest, 62 were randomly selected and served as centers for 13-m radius field plots. Preliminary analyses led to deletion of two observations as outliers, assumed to be the result of plot disturbance such as harvest between the dates of measurement and ALS acquisition. For each plot, diameter at breast-height (dbh, 1.5 m) was measured for all trees with dbh of at least 9.5 cm. Heights (ht) were measured for a sub-sample of plot trees and estimated for the remaining trees using a model of the ft-dbh relationship constructed using data for the measured trees. National models developed by Tabacchi et al. (2011) were used to predict AGB for individual trees. The predictions were added to produce plot-level totals, scaled to a per unit area basis (Mg/ha),), and considered to be observations without error (McRoberts & Westfall, 2014).

Wall-to-wall airborne laser scanning (ALS) data were acquired under leaf-on canopy conditions in June 2010 as part of a project focusing on the use of lidar for Italian forest applications (Scrinzi et al., 2013). Mean pulse density was 1.5 echoes/m^2. The ALS metrics included heights corresponding to the 10th, 20th, …, 90th, 99th percentiles of the height canopy distribution and the maximum, average, standard deviation, coefficient of variability, skewness and kurtosis of the distribution of echo heights. All metrics were calculated for 23-m x 23-m cells that mimicked the plot area of approximately 531 m^2 and that served as population units.

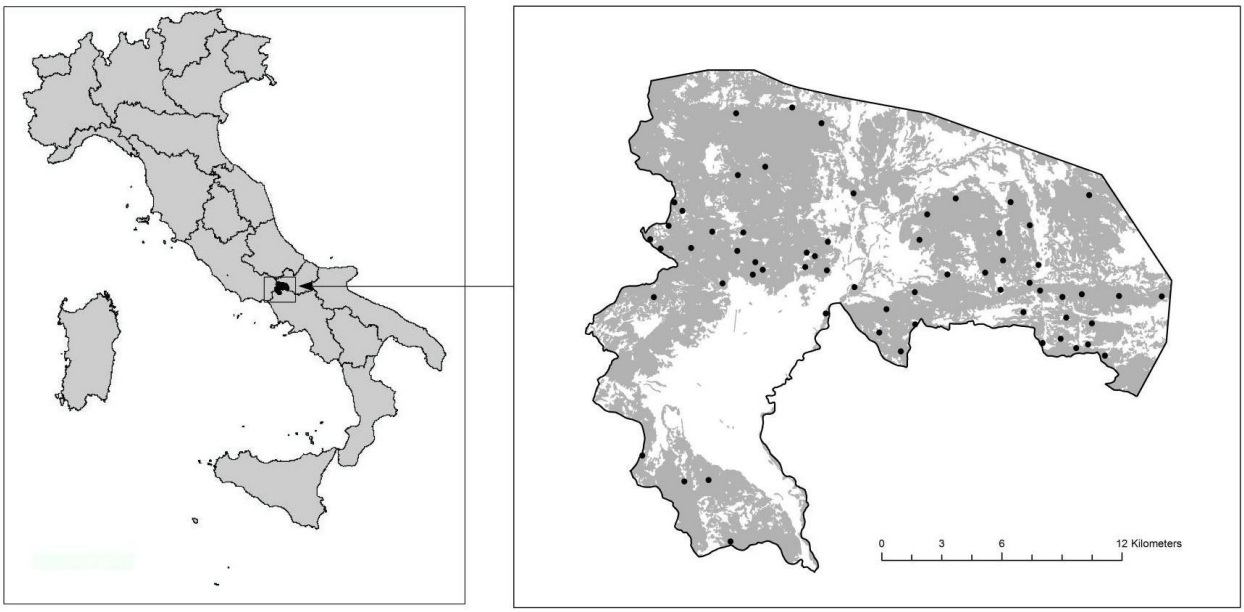


Figure 2. Molise study area in Italy.

Itasca County, Minnesota, USA

The 7,583-km^2 study area was located in north central Minnesota in the USA (Figure 3) and was characterized as approximately 80% forest land. Land cover includes water, wetlands and forest consisting of upland deciduous mixtures, pines (*Pinus* spp.) spruce (*Picea* spp.) and balsam fir (*Abies balsamea*) and with lowlands with spruce (*Picea* spp.), tamarack (*Larix laricin*), cedar (*Cedrus* spp.) and black ash (*Fraxinus nigra*). Data were obtained for plots established by the Forest Inventory and Analysis (FIA) program of the U.S. Forest Service which conducts the NFI of the USA (McRoberts et al., 2005). Each FIA plot consists of four 7.32-m (24-ft) radius circular subplots that are configured as a central subplot and three peripheral subplots with centers located at 36.58 m (120 ft) and azimuths of 0^o, 120^o, and 240^o from the center of the central subplot. Centers of forested, partially forested, or previously forested plots were determined using global positioning system receivers with sub-meter accuracy, whereas centers of non-forested plots were verified using aerial imagery and digitization methods. Field crews observe species and measure diameter at breast height (dbh, 1.37 m, 4.5 ft) and height for all trees with dbh≥12.7 cm (5 in). These data and statistical models were used to estimate individual tree volumes which were aggregated to obtain subplot-level volume estimates (m^3/ha) and were considered to be observations without error (McRoberts & Westfall, 2014). Data were used for plots measured in 2014 because this was the only year for which GPS receivers with sub-meter accuracy were available. Further, data for only the central subplot of each plot were used to avoid issues of spatial correlation among subplot observations.

Wall-to-wall lidar data were acquired in April 2012 with a nominal pulse spacing of 1.5 m using Leica ALS and Optech Gemini sensors. The average flying height above ground was 6000-7000 m, and the field of view was 40 degrees. The data were collected to meet a vertical accuracy of 5.0 cm RMSE. The data provider classified ground returns from the lidar point cloud and constructed a digital terrain model via interpolation using the Tiffs (Toolbox for Lidar Data Filtering and Forest Studies) software (Chen, 2007). Distributions of all first echo heights were constructed for the 168.3-m^2 plots and 169-m^2 square cells that tessellated the study area. For each plot and cell, the mean, standard deviation, skewness, and kurtosis of the distributins were calculated as was quadratic mean height (Lefsky et al.,1999; Chen et al., 2012). In addition, heights corresponding to the 10th, 20th, …, 100th percentiles of the distributions were calculated, and canopy densities were calculated as the proportions of echoes with heights greater than 0%, 10%, …, 90% of the range between 1.3 m above ground and the 95th height percentile (Gobakken & Næsset, 2008).

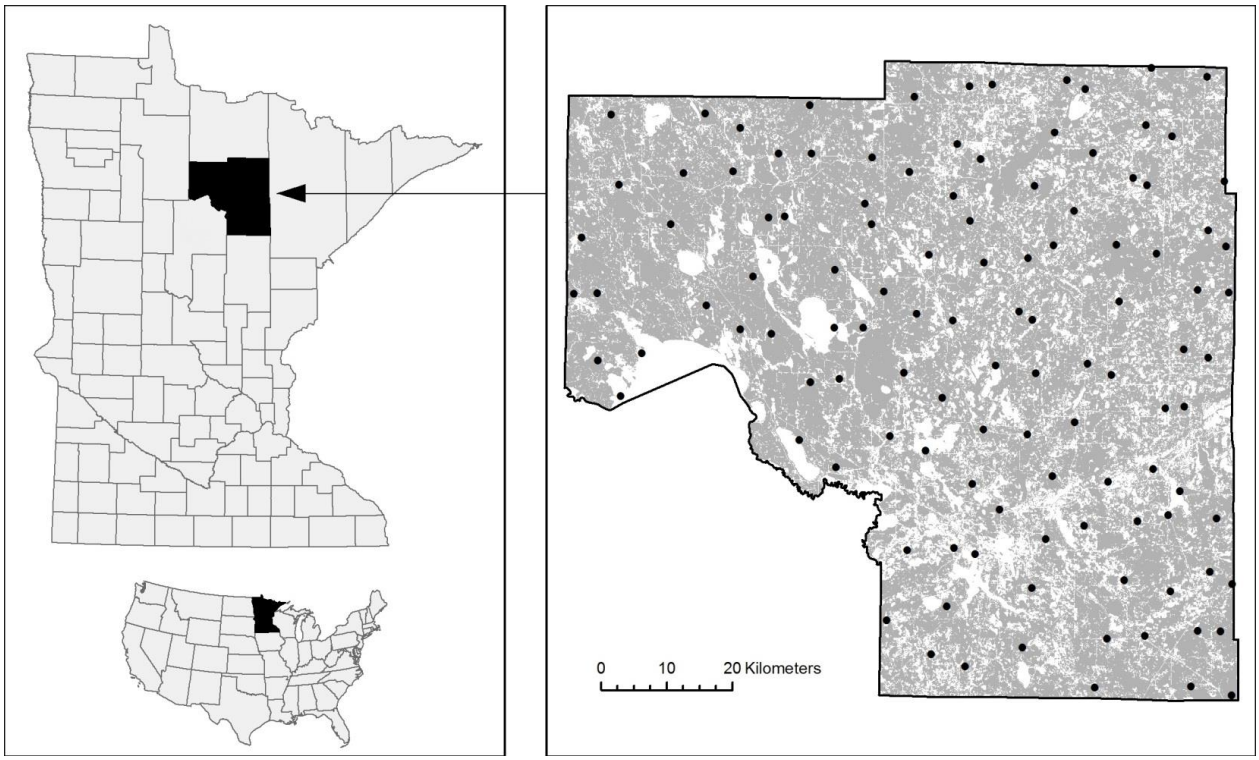


Figure 3. Itasca study area in Minnesota, USA.

Nearest neighbors techniques

Terminology and notation

For notational purposes, Y commonly denotes a possibly multivariate vector of response variables observed for a sample, and **X** denotes a vector of auxiliary variables with observations for the entire population. In the terminology of nearest neighbors techniques, the auxiliary variables are designated *feature variables*; the space defined by the feature variables is designated the *feature space*; the sample of population units for which observations of both response and feature variables are available is designated the *reference set* with size denoted n; and the set of population units for which predictions of response variables are desired is designated the *target set* with size denoted N. All population units for both the reference and target set are assumed to have a complete set of observations for all feature variables.

For continuous response variables, the nearest neighbors prediction, $\hat{y}_i$, for the i^{th} target unit is calculated as,

$$\hat{y}_i = \sum_{j=1}^{k} w_{ij} y_j^i \qquad (2)$$

where $\left\{ y_j^i, j = 1, 2, \ldots, k \right\}$ is the set of response variable observations for the k reference set units that are most similar or nearest to the i^{th} target unit in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the j^{th} nearest neighbor with $\sum_{j=1}^{k} w_{ij} = 1.$ For categorical variables such as forest/non-forest and

forest type, the predicted class of the i^{th} target unit is the most heavily weighted class among the k nearest neighbors, a weighted median or mode in the case of ordinal scale response variables, and a mode in the case of nominal response variables.

Distance metrics

Many familiar nearest neighbors distance metrics can be expressed in matrix form as,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \; M \left(X_i - X_j\right)}, \tag{3}$$

where i denotes a target unit for which a prediction is desired, j denotes a reference unit, $\mathbf{X}_i$ and $\mathbf{X}_j$ are vectors of observations of feature variables, and **M** is a square, positive definite matrix.

When considering nearest neighbors distance metrics, two feature space properties are particularly relevant. The first property, characterized by Bellman (1961) as the *curse of dimensionality*, is that the multi-dimensional size of feature space increases exponentially as the number of feature variables increases linearly. Three important detrimental consequences follow: (i) nearest neighbors are at greater distances from target units (Schaal et al., 1998); (ii) the distance to the nearest neighbor approaches the distance to the farthest neighbor (Beyer et al., 1998); and (iii) extrapolations beyond the ranges of the feature variables in the reference set are more probable (McRoberts, 2009). McRoberts (2015) showed that when using ALS data to predict AGB, and presumably also related response variables such as forest volume, the effects of the first two consequences are not severe. The second property is that inclusion of feature variables that are unrelated to the response variables has detrimental effects. Langley and Iba (1993) and Blum and Langley (1997) characterize such feature variables as *irrelevant*. Irrelevant feature variables introduce randomness into distance calculations and contribute to selection of spurious neighbors and less accurate predictions. Thus, optimization of nearest neighbors distance metrics should focus on simultaneously reducing the number of feature variables and weighting feature variables in proportion to their relevance in the distance metric.

A recent study (Chirici et al., 2016) conducted under the auspices of Action FP1001 of the European program Cooperation in Science in Technology (COST, 2015) identified the Euclidean metric, the Mahalanobis metric, and the canonical correlation analysis metric as the most frequently used metrics. These three metrics, plus the weighted Euclidean distance metric which is increasingly reported, were investigated for this study.

Euclidean metric

With the Euclidean (EUCL) distance metric, the **M** matrix from Eq. (3) is the identity matrix, **I**, and expresses distance as,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \; I \left(X_i - X_j\right)}. \tag{4}$$

The EUCL metric is the simplest, most intuitive, and probably the most frequently used metric.

Weighted Euclidean metric

The Weighted Euclidean (WEUCL) distance metric is similar to the EUCL metric, except **M** from Eq. (3) is a diagonal matrix, **D**, whose diagonal elements are permitted to differ,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \; D \left(X_i - X_j\right)}. \qquad (5)$$

Optimization of the metric entails selection of optimal values for the matrix diagonal elements and can be computationally intensive, even for relatively small numbers of feature variables.

Genetic algorithms (GA) have emerged as an increasingly common technique for optimizing selection of the diagonal elements of the **D** matrix. GAs are search heuristics that mimic natural selection to solve optimization problems (Holland, 1975). The process is iterative and starts from a population of randomly generated individuals, each consisting of multiple genomes, with the population in each iteration called a generation. For k-NN applications, a genome is a value for one element of the diagonal matrix that characterizes the WEUCL distance metric. A full set of genomes or diagonal values constitutes an individual. In each generation, each individual in the population is evaluated with respect to its fitness which for k-NN applications is typically a criterion related to the sum of squared errors for continuous response variables or classification accuracy for categorical response variables. Individuals with more optimal values of the criterion are characterized as more fit. Each subsequent generation consists of the more fit individuals from the previous generation and modifications of them plus a small number of new randomly generated individuals representing migration into the population. Modifications take multiple forms: (i) *averaging* consisting of combinations of two individuals by averaging their genomes, (ii) *cross-over* consisting of a combination of two individuals obtained by randomly selecting one genome from each pair to construct a new individual, and (iii) *random mutation* consisting of small random perturbations of randomly selected genomes for randomly selected individuals. The new generation of individuals is then used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced or a satisfactory fitness level has been reached for at least one individual in the population.

GAs could be used to optimize the full matrix, **M**, from Eq. (3) rather than just the diagonal matrix, **I**. However, the requirement that matrices associated with distance must be positive definite (Weinberger & Saul, 2009) induces considerable additional computational intensity. For example, as long as the diagonal values for a diagonal matrix are positive, the matrix will be positive definite; such is not the case for full matrices, even if they are symmetric. Thus, each new individual must be checked for positive definiteness with the result that large numbers are rejected. Therefore, for this study, GAs were used to optimize only the WEUCL metric. Tomppo and Halme (2004), McRoberts (2012, 2015), Tomppo et al. (2009), and Holopainen et al. (2010) all used GAs to select diagonal elements as a means of optimizing the WEUCL distance metric, and McRoberts (2008) and Latifi et al. (2010) used GAs to select feature variables.

Mahalanobis metric

With the Mahalanobis (MAHA) distance metric, **M** from Eq. (3) is the inverse of the feature variable covariance matrix, **V**,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \; V^{-1} \left(X_i - X_j\right)} \qquad (6)$$

(Mahalanobis, 1936). The MAHA metric is often used for comparison purposes, but seldom is selected as the optimal metric for forestry applications (Maltamo et al., 2003; Latifi et al., 2010; Ver Hoef & Temesgen, 2013).

Canonical correlation analysis metric

With the canonical correlation analysis (CCA) metric, a system of linear models is solved to obtain estimates of coefficient vectors, □ and □, that maximize the correlation between $U = \alpha_1 \cdot Y_1 + \cdots + \alpha_p \cdot Y_p$ and $V = \beta_1 \cdot X_1 + \cdots + \beta_q \cdot X_q$ where Y_j denotes the j^{th} response variable, X_j denotes the j^{th} feature variable, and p and q are the numbers of response and feature variables, respectively. The solutions are obtained using canonical decompositions for which the eigenvectors, also designated canonical correlation coefficients, are denoted □□□and the corresponding eigenvalues, also designated canonical correlations, are denoted □□ Feature space distances with this metric are calculated as,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' \; \Gamma \Lambda^2 \Gamma' \left(X_i - X_j\right)}$$

, (7) (6)

where the diagonal elements of the diagonal matrix, □□ are the squares, □², of the canonical correlations.

The CCA metric assumes a linear relationship between each of the response variables and the feature variables. The user has no control over the □ and □ vectors, and although the □ vector is of little consequence, the □ vector may not combine the response variables in a relevant manner. The metric was first proposed by Moeur and Stage (1995) who used only a single neighbor and characterized the combination of the CCA metric and k=1 as the *Most Similar Neighbor* technique (LeMay & Temesgen, 2005). However, the metric has also been used with multiple neighbors (Maltamo et al. 2003, 2009; Packalén & Maltamo, 2007).

Neighbor weighting

t-weighting

The most common approach to weighting neighbors when calculating k-NN predictions is to weight neighbors inversely proportionally to a power of the distance, d_{ij}, between the j^{th} reference unit and the i^{th} target unit,

$$w_{ij} = \frac{d_{ij}^{-t}}{W}, \tag{8}$$

where $W = \sum_{j=1}^{k} d_{ij}^{-t}$ and $t \geq 0$. Commonly, t=0, t=1, or t=2 is arbitrarily selected. Of importance, t=0 corresponds to weighting all neighbors equally. For this study, the special case of t=0 or is characterized as *c-weighting*, whereas weighting schemes corresponding to t>0 are characterized as *t-weighting*. Other than McRoberts (2012, 2015), no reports of attempts to optimize the selection of t are known. For small numbers of feature variables and/or large reference sets, $d_{ij}=0$ may occur in which case Eq. (8) leads to computational errors. For this study, if $d_{ij}=0$ for j=1,…,k, then all distances are arbitrarily reset to 1, i.e., $d_{ij}=1$ for all neighbors. If $d_{ij}=0$ for j=1,.., k' where k'<k, then all 0-distances are arbitrarily reset to half the smallest non-zero distance, i.e.,

$$d_{ij} = \frac{d_{ik'+1}}{2}, \tag{9}$$

for j=1,.., k'.

d-weighting

Dudani (1976) proposed a weighting scheme that is modified for this study to,

$$w_{ij} = \frac{\dfrac{d_{ik+1} - d_{ij}}{d_{ik+1} - d_{i1}}}{W}, \tag{10}$$

where $W = \sum_{j=1}^{k} \dfrac{d_{ik+1} - d_{ij}}{d_{ik+1} - d_{i1}}$. For notational purposes, this metric is characterized as *d-weighting*. Note that with the formulation of Eq. (10), calculation of weights for k neighbors requires distances for k+1 neighbors. For small numbers of feature variables and/or large reference sets, $d_{ik'}=d_{ik+1}$ for k'<k+1 may occur in which case, $w_{ij}=0$ for j=k',..., k. For this study, if $d_{i1}=d_{ik+1}$, then all distances are reset to 1, i.e., $d_{ij}=1$ for all neighbors. If $d_{ij}=d_{ik+1}$ for j=k',..., k where 1< k'<k+1, then all such distances are reset to the mean of the k+1st distance and the greatest distance that differs from the k+1st distance,

$$d_{ij} = \frac{d_{ik'-1} + d_{ik+1}}{2}, \tag{11}$$

for j= k',..., k. For example, suppose k=5 and the six smallest distances for the i^{th} target unit are $d_{i1}=1$, $d_{i2}=2$, $d_{i3}=3$ $d_{i4}=5$,$d_{i5}=5$,$d_{i6}=5$. Using only Eq. (1), the weights for the fourth and fifth neighbors would be $w_{i4}=w_{i5}=0$ which would exclude the observations from the fourth and fifth neighbors when calculating the k-NN prediction. However, by applying Eq. (11), the fourth nd fifth distances are rest to $d_{i4}=d_{i5}=4.0$ with the result that W=2.75 and $w_{i1}=0.364$, $w_{12}=0.272$, $w_{i3}=0.182$, $w_{i4}=0.091$, and $w_{i5}=0.091$.

Number of nearest neighbors, k

The value of k may be selected to optimize multiple criteria either individually or in combination. For k-NN variations that permit k>1, smaller values of k are generally preferred as a means of reducing complexity and computational intensity. However, caution must be exercised when selecting small values of k because such values may yield root mean square errors that are greater than the standard deviations of the response variable observations, meaning that the overall mean as a prediction for every target unit would better maximize accuracy than the k-NN predictions. Typically, a graph of a criterion of interest versus values of k is characterized by quite sub-optimal criterion values for very small values of k, a gradual approach to the optimal criterion value as k increases, and then quite sub-optimal criterion values again for larger values of k (McRoberts et al., 2002, Fig. 2; McRoberts, 2012, Fig. 2). Although the optimal value of k may be large, often a much smaller value can be selected that produces only a small proportional deviation from the optimal value of the criterion. For many reported applications, the value of k has been arbitrarily selected as k=1 or k=5 based on values reported elsewhere in the literature. The rationale for such decisions is uncertain, if not confusing, at least without assessing the consequences; for example, if a regression model were used, parameter estimates reported elsewhere in the literature would certainly not be arbitrarily selected for a new application that used the same model form and predictor variables.

Analyses

Optimization

For each distance metric and neighbor weighting scheme, the value of k that minimized the sum of squared residuals, SS_{res}, was determined for each combination of each number of feature variables, m, beginning with m=1. For each m, the combination of feature variables with the smallest SS_{res} was selected. For each value of m, beginning with m=3, three F-tests of significance were conducted: (i) SS_{res}^{m} was compared to SS_{res}^{m-1}, : (ii) SS_{res}^{m-1} was compared to SS_{res}^{m-2},and (iii) SS_{res}^{m} was compared to SS_{res}^{m-2}. The second and third tests were conducted because for several combinations of datasets, metrics and weighting schemes, the first test corresponding to inclusion of a single new feature variable indicated no statistically significant decrease in SS_{res}, but inclusion of a second new feature variable did produce a statistically significant decrease in SS_{res}. For each test, the statistic was calculated as,

$$F = \frac{\frac{SS_{res}^{m_1} - SS_{res}^{m_2}}{m_2 - m_1}}{\frac{SS_{res}^{m_2}}{(n - m_2)}}, \tag{12}$$

where m_1 and m_2 are the smaller and larger number of feature variables, respectively. The statistic was compared to the critical value for the $F(m_2 - m_1, n - m_2)$ distribution at the α=0.01 significance level. The relatively small α=0.01 significance level was selected rather than larger values for three reasons: (i) to partially compensate for the stringent 3-test selection criterion. (ii) to partially compensate for multiple applications of the test, and (iii) to err on the side of selecting fewer feature variables because, as shown by McRoberts (2015), with larger numbers of feature variables less of the optimization achieved in the reference set is actually realized in the target set. If all three tests indicated non-significance, the combination with m-2 feature variables was selected, and no combinations with greater numbers of feature variables were considered.

The significance level of this F-test is only approximate. First, for regression applications, the test assumes that the model with m-1 predictor variables is nested within the model with m predictor variables in the sense that the set of m-1 predictor variables is completely included in the set of m predictor variables. For k-NN applications, this criterion is not always satisfied. For example, for the Molise dataset, none of the feature variables selected for m=3 were included among those selected for m=4. Second, the leave-one-out technique does not necessarily produce the same distributions of SS_{res} as would a regression model. Third, the multiple applications of the test may require adjustment of the significance level as is done for statistical multiple comparisons analyses (Miller, 1981, Section 3.1). Nevertheless, the approach is an automated and objective technique for selecting feature variables.

A stepwise variable selection procedure is an alternative to considering all possible combinations of each number of feature variables. However, stepwise algorithms are known to perform poorly when the feature variables are strongly correlated (Harrell, 2001, pp. 64-65). For this study, preliminary investigations produced decidedly sub-optimal results when using stepwise procedures, presumably because of the strong correlations among the lidar height and density metrics used as feature variables.

Although feature variable selection was based on SS_{res}, results are reported using a pseudo-R^2 denoted and calculated as,

$$R^{2*} = \frac{SS_{mean} - SS_{res}}{SS_{mean}}. \quad (13)$$

Arbitrary selections

Many implementations of nearest neighbors techniques feature arbitrary selections of k and t, often using all feature variables. For example, the popular Most Similar Neighbor variation of k-NN uses the CCA metric and k=1 with no attempt to optimize other than the optimization that is inherent in the metric (Moeur & Stage, 1995). A reasonable question pertains to the degree of sub-optimality that results from such arbitrary selections. As a second example, metrics that incorporate weighting such as the WEUCL and CCA metrics ostensibly circumvent the necessity of selecting feature variables because they have the potential to mitigate the effects of irrelevant variables by giving them negligible weights. Thus, a second reasonable question pertains to the degree to which these metrics realize this potential.

To address these and related issues, consideration was given to inferences for the mean per unit area of the response variable for each dataset expressed as,

$$\hat{\mu} \pm t_{1-\alpha/2} \cdot SE(\hat{\mu}), \quad (14)$$

where $\hat{\mu}$ is the estimate of the mean per unit area, $SE(\hat{\mu}) = \sqrt{\hat{Var}(\hat{\mu})}$ is the standard error of $\hat{\mu}$, $t_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of Student's t-distribution, and □ is the significance level. The particular focus of the analyses was estimation of the mean and the SE of the mean using simple random sampling and model-assisted regression estimators.

The simple random sampling (SRS) estimator of the mean is,

$$\hat{\mu}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \quad (15)$$

where n is the reference set size, i indexes the reference units (plots), and y_i is the reference unit observation. The estimator of the variance of $\hat{\mu}_{SRS}$ is,

$$\hat{Var}(\hat{\mu}_{SRS}) = \frac{1}{n \cdot (n-1)}\sum_{i=1}^{n}(y_i - \hat{\mu}_{SRS})^2 \ . \quad (16)$$

For systematic samples, as used for this study, variances may be slightly overestimated relative to estimates based on a simple random sample (Särndal et al., 1992, p. 83). The primary advantages of the SRS estimators are that they are intuitive and unbiased, but the disadvantage is that variances may be large, particularly for small sample sizes and/or highly variable populations.

Model-assisted regression estimators use models based on auxiliary data to enhance inferences but rely on the probability sample for validity (Särndal et al., 1992). An initial estimator of the mean is designated the *synthetic* estimator and is formulated as,

$$\hat{\mu}_{Syn} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \ , \quad (17)$$

where N is the target set (population) size and $\hat{y}_i$ is the k-NN prediction. Systematic prediction errors induce bias into this estimator which can be estimated as,

$$\hat{\text{Bias}}(\hat{\mu}_{\text{Syn}}) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i , \tag{18}$$

where $\varepsilon_i = \hat{y}_i - y_i$. The model-assisted, generalized regression (GREG) estimator is then defined as,

$$\hat{\mu}_{\text{GREG}} = \hat{\mu}_{\text{Syn}} - \hat{\text{Bias}}(\hat{\mu}_{\text{Syn}})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) , \tag{19}$$

with variance estimator,

$$\hat{\text{Var}}(\hat{\mu}_{\text{GREG}}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2, \tag{20}$$

where $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$ (Särndal et al., 1992; Särndal, 2011). Despite the label characterizing the estimator, prediction techniques other than regression can be used. The degree to which the auxiliary information increases precision and thereby shortens the confidence interval is often calculated using relative efficiency,

$$\text{RE} = \frac{\hat{\text{Var}}(\hat{\mu}_{\text{SRS}})}{\hat{\text{Var}}(\hat{\mu}_{\text{GREG}})} . \tag{21}$$

The primary advantage of the GREG estimators is that they capitalize on the relationship between the reference set observations and their corresponding predictions to reduce the variance of the estimate of the population mean.

For each dataset, estimates of the mean per unit area for the response variable, estimates of SEs, and REs were compared for five approaches: (i) the SRS estimators; (ii) the GREG estimators based on k-NN predictions obtained using the WEUCL metric, the optimal value of k, and d-weighting; (iii) the GREG estimators based on k-NN predictions obtained using the CCA metric, the optimal value of k, and d-weighting; (iv) the GREG estimators based on k-NN predictions obtained using the WEUCL metric with all feature variables and arbitrary selections of k and t; and (v) the GREG estimators based on k-NN predictions obtained using the CCA metric, all feature variables, and arbitrary selections of k and t. For approaches (iv) and (v), arbitrary selections of k were k=1 and k=5 and arbitrary selections of t for use with k=5 were t=0, t=1, and t=2. The EUCL and MAHA metrics were not used for these analyses because the WEUCL and CCA metrics were deemed preferable.

2.3.3. Results

Optimization

For each dataset, optimization included four components: selection of a distance metric, selection of feature variables, selection of a value of k, and selection of a neighbor weighting scheme. Overall, the primary results of optimization were twofold. First, optimization for the WEUCL, MAHA, and CCA metrics produced

slightly greater accuracies as assessed by R^{2*} than the EUCL metric (Table 1). This result is as expected because the former metrics all feature greater potential for optimization. Second, optimization for the three former metrics produced accuracies that were generally similar to each other.

Table 1. Prediction accuracies.

Dataset	Distance metric†	c-weighting			t-weighting				d-weighting		
		Number of feature variables	k	R^{2*}	Number of feature variables	k	t	R^{2*}	Number of feature variables	k	R^{2*}
Hedmark	EUCL	3	6	0.74	3	6	0.0	0.74	3	10	0.73
	WEUCL	3	3	0.84	3	3	0.6	0.84	3	4	0.84
	MAHA	4	2	0.82	6	3	0.7	0.83	4	8	0.83
	CCA	7	3	0.85	7	4	0.7	0.85	7	3	0.85
Våler	EUCL	5	4	0.76	3	5	0.7	0.75	4	7	0.76
	WEUCL	5	11	0.83	3	4	0.0	0.84	4	11	0.82
	MAHA	6	5	0.80	6	5	1.1	0.81	6	8	0.81
	CCA	5	5	0.83	4	13	0.8	0.84	5	6	0.83
Molise	EUCL	3	2	0.78	3	4	1.6	0.82	2	3	0.78
	WEUCL	3	2	0.80	3	4	1.6	0.82	2	4	.0.78
	MAHA	3	3	0.79	3	3	1.1	0.80	3	4	0.79
	CCA	3	2	0.79	4	6	1.2	0.80	4	2	0.80
Itasca	EUCL	5	2	0.84	4	2	0.6	0.84	4	2	0.83
	WEUCL	5	2	0.85	4	2	0.4	0.85	4	2	0.84
	MAHA	4	1	0.86	6	9	3.6	0.88	5	2	0.87
	CCA	8	2	0.87	8	2	0.4	0.88	8	3	0.87

†EUCL: Euclidean; WEUCL: Weighted Euclidean; MAHA: Mahalanobis; CCA: Canonical correlation analysis.

Optimization for the WEUCL, MAHA, and CCA metrics produced R^{2*} values in the range 0.82-0.85 for the Hedmark dataset, in the range 0.80-0.84 for the Valer dataset, in the range 0.79-0.82 for the Molise dataset, and in the range 0.84-0.88 for the Itasca dataset. Relative to R^{2*} values for the EUCL metrics, R^{2*} values for the three former metrics represented increases in the range 0.08-0.12 for the Hedmark dataset, in the range 0.04-0.09 for the Våler dataset, in the range 0.00-0.02 for the Molise dataset, and 0.01-0.04 for the Itasca dataset. These small ranges indicate that optimization of all three metrics via selection of feature variables, k, and weighting scheme produced comparable results.

Although optimization produced similar R^{2*} values, each of the WEUCL, MAHA, and CCA metrics has disadvantages. For the WEUCL metric, optimization can be computationally intensive, particularly with large reference sets and large numbers of feature variables. In particular, GA optimization is unrealistic, if not impossible, for large numbers of feature variables (Tomppo et al., in review). Within the small range of R^{2*}values for the three metrics, the MAHA metric produced the smallest R^{2*} values for three of the four datasets. In addition, as noted in the literature review, supervised metrics which are optimized using observations of the response variable are generally more accurate than unsupervised metrics. In this context, the WEUCL and CCA metrics are characterized as supervised, whereas the MAHA metric is characterized as unsupervised. In addition, the MAHA metric makes no provision for minimizing the influence of irrelevant feature variables. Both the MAHA and CCA metrics are susceptible to computational problems resulting from the strongly correlated ALS metrics that serve as feature variables. Finally, for the CCA metric, optimization produced larger numbers of feature variables, a result that can likely be attributed to near negligible weights associated with mostly irrelevant feature variables that otherwise would not have been selected. The disadvantage is that

larger numbers of feature variables means that optimization requires consideration of more combinations of feature variables, thereby increasing computational intensity.

Optimal values of k tended to be small, never greater than k=13 and less than k=10 for all except four of the 48 combinations of dataset, metric, and neighbor weighting scheme. Values of k much larger than approximately k=10 would probably not actually be used. Rather, as graphs of optimization criteria versus k typically reveal, considerably smaller values of k often produce only slight changes in the optimization criterion (e.g., McRoberts et al., 2002, Figure 2; McRoberts 2012, Figure 2).

The t-and d-weighting schemes produced little increase in R^2* relative to c-weighting. This result can likely be partially attributed to the relatively small values of k. Among the three neighbor weighting schemes, c-weighting and d-weighting require no optimization and are simple to implement, whereas optimization of t-weighting can be computationally intensive.

Overall, based on the potential for optimization, the WEUCL and CCA metrics are preferable to the EUCL and MAHA metrics. In addition, based on both potential for optimization and ease of implementation, d-weighting is slightly preferable to c- and t-weighting. Therefore, the WEUCL and CCA metrics in combination with d-weighting were used for comparison purposes for assessing the effects on inferences of using all feature variables in combination with arbitrary selections of k and neighbor weighting.

Inference

Estimates of population means per unit area obtained using the SRS estimators and the GREG estimators with the optimized WEUCL and CCA metrics were generally similar; in particular, all GREG estimates were within two SRS SEs of the SRS estimates (Table 2). The similarity of the SRS and GREG estimates indicates that lack of correspondence between the sample observations and their k-NN predictions was not a serious issue and that the distributions of the feature variables in the reference and target sets were similar. Differences in the means for different datasets are attributed to use of volume as the response variables for the Hedmark and Itasca datasets and use of biomass as the response variable for the Våler and Molise. In addition, whereas the study areas for the Våler and Molise datasets were limited to forest land, the study areas for the Hedmark and Itasca datasets included both forest and non-forest lands. Finally, despite similarity in estimates of means, the GREG SEs were less than half the SRS SEs which indicates the utility of the ALS auxiliary information for increasing the precision of estimates of the population means.

Table 2. Simple random sampling and model-assisted regression estimates of population means per unit area.

Dataset	Simple random sampling (SRS) estimators		Model-assisted regression (GREG) estimators			
			Weighted Euclidean metric†		Canonical correspondence analysis metric†	
	$\hat{\mu}_{SRS}$	$SE(\hat{\mu}_{SRS})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$
Hedmark	74.26	7.45	75.23	3.01	83.23	2.89
Våler	112.39	5.00	106.03	2.13	105.46	2.04
Molise	108.23	10.94	105.98	5.08	107.58	4.88
Itasca	50.63	5.96	51.07	2.41	54.04	2.15

†Metric was in combination with optimal subset of feature variables, optimal value of k, and d-weighting.

Overall, the effects of optimization were twofold. First, for each dataset, estimates of means per unit area for the response variables were generally similar, although the estimates obtained using the optimal combinations were generally slightly smaller than the estimates obtained using the arbitrary combinations (Tables 3, 4). Second, optimization produced substantially smaller SEs and substantially larger REs. The latter result suggests that arbitrary selections and/or failure to optimize will be increasingly difficult to justify. This conclusion, while perhaps placing an additional burden on researchers, represents continuing maturation of the k-NN technique.

Table 3. Estimates of population means obtained using the Euclidean (EUCL) metric with all feature variables.

Dataset	k	t	R^{2*}	$\hat{\mu}_{syn}$	$\hat{Bias}(\hat{\mu}_{syn})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$	RE
Hedmark	1	0	0.45	87.90	4.24	83.65	5.53	1.82
	5	0	0.61	86.93	3.34	83.58	4.63	2.59
	5	1	0.60	87.20	3.63	83.57	4.71	2.51
	5	2	0.59	87.29	4.02	83.27	4.78	2.43
	4	Optimal†	0.84	78.01	2.82	75.23	3.01	6.12
Våler	1	0	0.58	109.15	2.76	106.40	3.24	2.38
	5	0	0.68	109.89	-0.10	109.99	2.83	3.13
	5	1	0.68	109.89	0.17	109.72	2.82	3.15
	5	2	0.68	109.86	0.58	109.28	2.83	3.12
	11	Optimal†	0.82	106.74	0.72	106.03	2.13	5.53
Molise	1	0	0.63	110.93	2.09	108.84	6.62	2.73
	5	0	-0.13	122.23	-1.81	124.04	11.61	0.89
	5	1	-0.01	113.10	-2.41	115.51	11.00	0.99
	5	2	0.02	106.02	-3.30	109.31	10.85	1.02
	3	Optimal†	0.78	107.10	1.12	105.98	5.08	4.63
Itasca	1	0	0.65	51.47	2.68	48.80	3.52	2.86
	5	0	0.78	51.39	-1.92	53.31	2.80	4.52
	5	1	0.78	51.62	-0.95	52.57	2.77	4.62
	5	2	0.78	51.76	0.02	51.74	2.80	4.54
	2	Optimal†	0.84	51.14	0.07	51.07	2.41	6.14

†WEUCL metric, optimal selection of feature variables, optimal value of k, and d-weighting as per Table 1.

Table 4. Estimates of population means obtained using the canonical correlation analysis (CCA) metric with all feature variables.

Dataset	k	t	R^{2*}	$\hat{\mu}_{syn}$	$\hat{Bias}(\hat{\mu}_{syn})$	$\hat{\mu}_{GREG}$	$SE(\hat{\mu}_{GREG})$	RE
Hedmark	1	0	0.61	87.00	0.94	86.07	4.64	2.58
	5	0	0.75	79.47	-1.83	81.30	3.74	3.96
	5	1	0.69	78.24	-4.18	82.43	4.10	3.30
	5	2	0.63	77.95	-4.39	82.34	4.53	2.71
	3	Optimal†	0.85	81.51	-1.72	83.23	2.89	7.45
Våler	1	0	0.66	100.59	-4.47	105.06	2.90	2.97
	5	0	0.72	102.36	-6.55	108.91	2.58	3.76
	5	1	0.74	103.37	-6.94	110.31	2.51	3.96
	5	2	0.71	103.67	-6.46	110.13	2.64	3.58
	6	Optimal†	0.83	105.53	0.07	105.46	2.04	5.98
Molise	1	0	0.63	110.93	2.09	108.84	6.62	2.73
	5	0	0.68	102.96	-11.35	114.30	6.04	3.28

	5	1	0.64	103.64	-8.19	111.83	6.45	2.88
	5	2	0.59	103.74	-5.50	109.63	6.94	2.49
	5	Optimal†	0.80	105.18	-2.39	107.58	4.88	5.04
Itasca	1	0	0.72	52.89	-0.41	53.30	3.14	3.60
	5	0	0.70	48.57	-7.71	56.28	3.16	3.56
	5	1	0.71	50.21	-6.16	56.37	3.16	3.55
	5	2	0.67	50.88	-5.50	56.38	3.38	3.10
	8	Optimal†	0.87	51.95	-2.09	54.04	2.15	7.68

†Optimal selection of feature variables, optimal value of k, and d-weighting as per Table 1.

2.3.4. Discussions and Conclusions

Four conclusions may be drawn from the study. First, use of the auxiliary airborne laser scanning data, the k-NN technique, and the model-assisted regression estimator substantially increased the precision of estimates of the means per unit area for the response variables. Although similar conclusions have been reported on numerous previous occasions, failure to have observed this phenomenon would have invalidated the entire study.

Second, optimization for the weighted Euclidean, Mahalanobis, and canonical correlation analysis metrics via selection of feature variables, k, and neighbor weighting produced generally comparable prediction accuracies. Despite this comparability, the weighted Euclidean and canonical correlation analysis metrics have greater potential for optimization and are therefore preferable to the Euclidean and Mahalanobis metrics. In addition, the canonical correlation analysis metric produced greater precision for estimates of the population means per unit area than the weighted Euclidean metric. Among the neighbor weighting schemes, c-weighting has little potential for optimization relative to t- and d-weighting, but optimization of t-weighting is computationally intensive, whereas no optimization is necessary for d-weighting. Further, given the considerable variety among the forest conditions represented by the four datasets, a reasonable degree of generalization is warranted for these findings. Thus, the second conclusion is that the combination of the canonical correlation metric and d-weighting, when optimized by selecting optimal subsets of feature variables and optimal values of k, merits serious consideration when estimating volume and biomass using airborne laser scanning data.

Third, despite the potential of the weighted Euclidean and canonical correlation analysis metrics to circumvent selection of feature variables by assigning negligible weights to irrelevant feature variables, selection of smaller numbers of feature variables produced greater precision for estimates of population means per unit area than use of all feature variables.

Fourth, and most importantly, optimization of k-NN configurations via selection of feature variables, k, and neighbor weighting, regardless of the distance metric, produced considerably more precise estimates of population means per unit area than use of all feature variables and arbitrary selections of k and t. Although arbitrary selections may be reasonable under unique situations, authors should justify any decisions not to optimize and report assessments of the degree to which such arbitrary selections produce sub-optimal prediction accuracies and inferences.

2.3.5.References

Bellman, R . (1961). Adaptive control processes: a guided tour. Princeton University Press. Princeton, New Jersey.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1998). When is "nearest neighbor" meaningful? In: Beeri, C., & Buneman, P. (Eds.). Proceedings of the 7th International Conference on Database Theory (ICDT), January 10–12, 339 1999, Jerusalem, Israel; pp. 217-235.

Blum, A.L.,& Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, 245-271.

Braastad, H. (1966). Volume tables for birch. Meddelser norske SkogforsVes., 21, 265-365. (In Norwegian with English summary).

Brantseg, A. (1967). Volume functions and tables for Scots pine. South Norway. Meddelser norske SkogforsVes, 22, 689-739. (In Norwegian with English summary).

Chen, Q. (2007). Airborne lidar data processing and information extraction. Photogrammetric Engineering and Remote Sensing, 73(12), 1355-1365.

Chen,Q., Vaglio Laurin, G., Battles, J.J., & Saah, D. (2012). Integration of airborne lidar and vegetation types derived from aerial photography for mapping aboveground live biomass. Remote Sensing of Environment, 121, 108-117.

Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F., & Bertini, R. (2008). Non-parametric and parametric methods using satellite imagery for estimating growing stock volume in alpine and Mediterranean forest ecosystems. Remote Sensing of Environment, 112, 2686–2700.

Chirici, G., Mura, M., McInerney, D., Py,Nicolas, Tomppo, E.O., Waser, L.T., Travaglini, D., & McRoberts, R.E. (in review). A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sensing of Environment.

COST. (2015). COST Action FP1001 – USEWOOD: Improving data and information on the potential supply of wood resources. Available at: https://sites.google.com/site/costactionfp1001/. Last accessed: February 2015.

Dudani, S.A. (1976). The distance-weighted k-Nearest-Neighbor rule. IEEE Transactions on SMC-6(4), 325-327.

Elisseeff, A., & Pontil, M. (2002). Leave-one-out error and stability of learning algorithms with applications. In: Advances in Learning Theory: Methods, Models, and Applications. NATO Advanced Study Institute on Learning Theory and Practice. pp. 111-1130.

Gobakken, T., & Næsset, E. (2008). Assessing effects of laser point density, ground sampling intensity, and field plot sample size on biophysical stand properties derived from airborne laser scanner data. Canadian Journal of Forest Research, 38, 1095-1109.

Harrell, F.E. (2001). Regression modeling strategies with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag.

Holland, J.H. (1975). Adaptation in natural and artificial system. Ann Arbor, Michigan. University of Michigan Press.

Holopainen, M., Haapanen, R., Karjalainen, M., Vastaranta, M., Hyyppä, J., Yu, X., Tuominen, S., & Hyyppä, H. (2010). Comparing accuracy of airborne laser scanning and TerraSAR-X radar Images in the estimation of plot-level forest variables Remote Sensing, 2, 432-445.

Langley, P., & Iba, W. (1993) Average-case analysis of a nearest neighbor algorithm. In: Proceedings International Joint Conference on Artificial Intelligence-93. Chambery, France. pp. 889-894.

Latifi, H., Nothdurft, A., & Koch, B. (2010)_. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. Forestry, 83(4), 395-407.

Lefsky, M. A., Harding, D., Cohen, W. B., Parker, G., & Shugart, H. H. (1999). Surface lidar remote sensing of basal area and biomass in deciduous forests of eastern Maryland, USA. Remote Sensing of Environment, 67, 83-98.

LeMay, V., & Temesgen, H. (2005). Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science, 51(2), 109-119.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. Proceeding of the National Institute of Sciences of India, 2, 49-55.

Maltamo, J., Malinen, J., Kangas, A., Härkönen, and Pasanen, A.-M. (2003). Most similar neighbour-based stand variable estimation for use in inventory by compartments in Finland. Forestry, 76, 449-463.

Maltamo, M., Peuhkurinen, J., Malinen, J., Vauhkonen, J., Packalén, P. & Tokola, T. (2009). Predicting tree attributes and quality characteristics of Scots pine using airborne laser scanning data. Silva Fennica, 43(3): 507–521.

Marklund, L.G. (1988). Biomass functions for pine, spruce, and birch in Sweden. Umeå: Swedish University of Agricultural Sciences, Department of Forest Survey (in Swedish).

McRoberts, R.E. (2008). Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. Remote Sensing of Environment, 112, 2212-2221.

McRoberts, R.E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment, 113, 489-499.

McRoberts, R.E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecology and Management, 272, 3-12.

McRoberts, R.E., & Westfall, J.A. (2014). The effects of uncertainty in model predictions of individual tree volume on large area volume estimates. Forest Science, 60, 34-43.

McRoberts, R. E., Nelson, M. D., &Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sensing of Environment, 82, 457-468.

McRoberts, R.E., Næsset, E., & Gobakken, T. (2015). Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. Remote Sensing of Environment, 163, 13-22.

Miller, R. (1981). Simultaneous statistical inference, 2nd Ed. Springer-Verlag. New York. 299 p.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor — an improved sampling inference procedure for natural resource planning. Forest Science, 41(2), 337–359.

Næsset, E., Bollandsås, O.M., Gobakken, T., Gregoire, T.G., & Ståhl, G. (2013). Model-assisted estimation of change in forest biomass over an 11 year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data." Remote Sensing of Environment, 128, 299-314.

Packalén, P., & Maltamo, M. (2007). The k-MSN method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of Environment, 109, 328-341.

Schaal, S., Vijayakumar, S., & Atkeson, C.G. (1998). Local dimension reduction. In: Jordan, M.I., Kearns, J.J., & Solla, S.A. (Eds.). Advances in Neural Information Processing Systems 10. Cambridge, MA. MIT Press. pp. 1-7.

Scrinzi, G., Clementel, F., Colle, G., Corona, P., Floris, A., Maistrelli, F., Chirici, G., Mura, M., Oradini, A., Bertani, R., Barbati, A., Quatrini, A., & Marchetti, M. (2013). Impiego di dati LiDAR di pubblica disponibilità per il monitoraggio forestale a grande e piccola scala: il progetto ITALID. In: Proceedings of the 9th SISEF National Congress "Multifunzionalità degli Ecosistemi Forestali: Sfide e Opportunità per la Ricerca e lo Sviluppo", Bolzano (Italy) 16-19 Sep 2013.

Tabacchi G., Di Cosmo L., Gasparini P., Morelli S. (2011). Stima del volume e della fitomassa delle principali specie forestali italiane. Equazioni di previsione, tavole del volume e tavole della fitomassa arborea epigea. Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di Ricerca per il Monitoraggio e la Pianificazione Forestale. Trento. 412 pp.

Temesgen, H., LeMay, V.M., Froese, K.L., & Marshall, P.L. (2003). Imputing tree-lists from aerial attributes for complex stands of south-eastern British Columbia. Forest Ecology and Management, 177, 277-285.

Tomppo, E., & Halme, M. (2004). Using coarse scale forest variables as ancillary information and weighting of k-NN estimation: a genetic algorithm approach. Remote Sensing of Environment, 92, 1-20.

Tomppo, E., Kuusinen, N., Mäkisara, K., & Katila, M. (in review). The effect of plot field type on the uncertainties of the ALS-based forest resource estimates. Remote Sensing of Environment.

Tomppo, E.O., Gagliano, C., De Natale, F., Katila, M., & McRoberts, R.E. (2009). Predicting categorical forest variables using an improved k-Nearest Neighbour estimator and Landsat imagery. Remote Sensing of Environment, 113, 500-517.

Ver Hoef, J.M., & Temesgen, H. (2013). A comparison of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. PLOS ONE, 8(3), e59129.

Vestjordet, E. (1967). Functions and tables for volume of standing trees. Norway spruce. Meddelser norske SkogforsVes., 22, 539-574. (In Norwegian with English summary).

Weinberger, K., & Saul, L.K. (2009). Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research, 10, 207-244.

3. Section II (Studies IV-V)

This second section comprises two studies, namely Study IV and Study V published as original research articles to which the PhD candidate participated as lead author:

IV. Mura M., McRoberts R. E., Chirici G., Marchetti M. (2015). Estimating and mapping forest structural diversity using airborne laser scanning data. *Remote Sensing of Environment*, *170*, 133–142. doi:10.1016/j.rse.2015.09.016

V. Mura M., McRoberts R. E., Chirici G., Marchetti M. (in review). Statistical inference for multiple-variable, forest structural diversity indices using airborne laser scanning data and the k-Nearest Neighbors technique. *Remote Sensing*.

In this second and last section, the methods experimented in the first section are applied in two different research studies.

Study IV describes the use of ALS data and ground data for the areal estimate of mean values of two forest structural diversity indices in a model-assisted framework. Along with the areal estimates, the study proposes the calculation of the confidence intervals of such estimates and the mapping of the investigated indices.

Study V is framed as a methodological paper that takes a step further than Study IV, showing how, using the capability of an optimized k-NN techniques in predict simultaneously different parameters, is possible to map a more comprehensive structural diversity index (SDI) combining different forest structural diversity indices.

3.1. Study IV

From:

[1]Mura M., [2]McRoberts R. E., [3]Chirici G., [1]Marchetti M. (2015). Estimating and mapping forest structural diversity using airborne laser scanning data. *Remote Sensing of Environment*, *170*, 133–142. doi:10.1016/j.rse.2015.09.016

[1] Dipartimento di Bioscienze e Territorio, University of Molise, Contrada Fonte Lappone, Pesche (IS), 86090 Italy

[2] Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota 55108 USA

[3] Department of Agricultural, Food and Forestry Systems, University of Florence, Florence 50145 Italy

Abstract

Among the wide array of terrestrial habitats, forest and wooded lands are the richest from both biological and genetic points of view because of their inherent structural and compositional complexity and diversity. Although species composition is an important biodiversity feature, forest structure may be even more relevant for biodiversity assessments because a diversified structure is likely to have more niches, which in turn, host more species and contribute to a more efficient use of available resources. Structure plays a major role as a diversity indicator for management purposes where maps of forest structural diversity are of great utility when planning conservation strategies. Airborne laser scanning (ALS) data have been demonstrated to be a reliable and valid source of information for describing the three-dimensional structure of forests. Using ALS metrics as predictor variables, we developed regression models for predicting indices of forest structural diversity for a study area in Molise, Italy. The study had two primary objectives: (i) to estimate indices of structural diversity for the entire study area, and (ii) to construct maps depicting the spatial pattern of the structural diversity indices. Our results demonstrate the utility of simple linear models using ALS data for improving areal estimates of mean structural diversity, and the resulting maps capture the patterns of structural diversity in the study area.

Keywords: Airborne laser scanning, forest structural diversity, GREG estimator, model-assisted estimator

3.1.1. Introduction

Definitions of biodiversity extend across both levels and concepts. For example, Wilson (1987), Leveque (1994) and Gaston and Spiecer (2004) all recognize three levels of assessment: genetic, specific and ecosystem biodiversity. Whittaker (1972) distinguishes among alpha (α) diversity, which refers to ecosystem diversity; beta (β) diversity, which refers to change in diversity between ecosystems; and gamma (γ) diversity, which relates to the overall diversity for different ecosystems within a region. Noss (1990) provides a more structured view of ecosystem biodiversity that includes compositional, structural and functional aspects, while other authors focus on both functional and phylogenetic aspects (Tilman et al., 1997).

An analysis of zonal and habitat classifications (Holdridge, 1947, 1967; Dinerstein et al. 1995) reveals that half of the wide range of habitat classes are located in tree-dominated lands, thus confirming forest and wooded lands as the richest terrestrial ecosystems from both biological and genetic points of view. Because the concept of biodiversity in forestry goes beyond "just the trees" to include all other living organisms as fundamental components of ecosystem functionality (Hunter, 1990), the conservation of forest habitats becomes crucial for many reasons (Hunter, 1999). The key role of forest biodiversity in a global context is also evinced in numerous international agreements such as the Convention on Biological Diversity of the 1992 United Nations Conference on Environment and Development, the Ministerial Conference on the Protection of Forests of Europe (MCPFE 2002) and the Montréal Process (2006).

Forests host a wide variety of animal species, and studies of biodiversity in forest environments often focus on habitat characterization, which is well known to be associated with animal presence/absence (Brokaw& Lent, 1999). These studies further focus on both compositional and/or three-dimensional (3-D) structural parameters and the continuous debate since early studies regarding factors that most influence the diversity of communities (MacArthur & MacArthur, 1961). For some animal communities, plant species composition seems to be a better predictor of habitat quality than structural variables (Ter Braak & Schaffers, 2004; Schaffers et al. 2008); conversely for other communities structure plays a major role (McGraw, 1994; Salter et al., 1985; Welsh & Lind, 1996; Halaj et al., 2000; Shine et al., 2002; McElhinny et al., 2005; Müller et al., 2010), although it may be influenced by plant species composition as well (Rosenzweig, 1995). Thus, the relative importance of structure and composition may reasonably be assumed to be influenced by the habitat type of the study area (Fleishman & MacNally 2006).

Even if composition is an important factor, McElhinny et al. (2005) reported in their review that forest structure is more relevant for biodiversity assessments than composition. They assert that a diverse stand structure is likely to have more niches, which would host more species and contribute to a more efficient use of available resources (McElhinny et al., 2005). Structure plays a major role in the Quantitative Pan-European indicators describing Criterion 4 `Biological Diversity` (MCPFE, 2011) when applied in a sustainable forest management context (Puumalainen et al., 2003). Kuuluvainen (2009) and Lähde et al. (1999) confirmed these findings in other studies focusing on non-tree structural parameters such as deadwood (Rondeaux & Sanchez, 2010).

From a management perspective, maps depicting spatially explicit patterns of structural diversity would be of great use when planning conservation strategies. Further, mapping forest structure would

facilitate habitat and diversity assessments for large, remote and steep areas that cannot be safely reached by field crews. For this purpose, remotely sensed data become an indispensable tool for constructing synoptic views of large areas (O'Neil et al., 1997). In the last two decades, active remote sensing technologies such as LiDAR (Light Detection and Ranging) have been shown to be reliable and valuable sources of information for estimating and assessing forest structural parameters (Lefsky et al., 2002; Lim et al., 2003; Zimble et al., 2003; Wulder et al., 2008), thus contributing great potential support for forest biodiversity studies.

Bergen et al. (2009) reviewed the use of LiDAR and RaDAR for characterizing 3-D vegetation structure for biodiversity and habitat science analyses. Their review had several focal points: relationships between vegetation structure, diversity and habitat and remotely sensed data; technical capabilities of new sensors and their potential for application to biodiversity and habitat studies; and useful and feasible variables that can be derived from spaceborne LiDAR and RaDAR observations along with their accuracy and precision requirements.

Habitat studies focusing on the diversity and abundance of birds are common (Tews et al., 2004). Vogeler et al. (2014), Clawges et al. (2008), Lesak et al. (2011), Goetz et al. (2007), Graf et al., (2009), Flashpoler et al. (2010), Vierling et al. (2013) and Müller et al. (2009, 2010) all report strong relationships between bird communities and LiDAR-based estimates of structural parameters. LiDAR data have also been used to characterize habitat for other forest taxa such as beetles (Müller & Brandl, 2009), spiders (Vierling et al. 2011), and mammals such as bats (Jung et al. 2012), deer (Ewald et al., 2014) and squirrels (Nelson at al., 2005).

Structural information that can be extracted from LiDAR data has been shown to be effective not only for the prediction of animal habitat but also for the assessment of plant richness, composition and diversity (Lucas et al., 2010; Leutner et al., 2012; Simonson et al. 2012; Alberti et al. 2013).

The common aspect of the aforementioned studies is that they characterize the diversity of specific habitats rather than directly linking indices of forest structural diversity to LiDAR metrics, leaving the latter aspect mostly unexplored. Wall-to-wall LiDAR coverage permits construction of LiDAR metrics for use as a base to develop relationship models that can then be used to map habitats across wide areas. Multiple studies have shown the efficacy of habitat mapping using LiDAR-based layers alone (Hyde et al., 2005; Martinuzzi et al., 2009) or in combination with other remotely sensed data (Swatantran et al. 2012; Hyde et al., 2006), among which LiDAR-based layers have been demonstrated to be the best single predictors.

A common deficiency of assessments of maps based on remotely sensed data is that error matrices and measures such as overall accuracy, users' and producers' accuracies, Kappa index correlation and coefficients of correlation provide little information regarding the accuracy or precision of parameters estimated from the map unit predictions (McRoberts, 2011). For management and planning purposes, the accuracy and precision of biodiversity estimates for the area of interest are essential. None of the aforementioned studies includes attempts to produce scientific inferences from the maps. Motivated by these considerations, the aim of this paper is twofold: (i) to construct inferences in the form of confidence intervals for estimates of common indices of forest structural diversity using LiDAR data obtained via airborne laser scanning (ALS) as auxiliary information, and (ii) to construct maps depicting the spatial pattern of these structural diversity indices using

gridded ALS data metrics. The study area is in Molise, Italy, and the analyses include comparison of estimates obtained using design-based, simple random sampling (SRS) and model-assisted estimators.

3.1.2. Materials and Methods

Study area

The study area is located in the southwestern part of Molise Region in central Italy and includes 36,360 ha (Figure 1). Forests are spread over 20,518 ha, comprising approximately 56% of the area. The forested area is dominated by deciduous oaks (*Quercus cerris, Quercus pubescens*) covering approximately 60% of the forest area, hop hornbeam (*Ostrya carpinifolia*) covering approximately 18% and beech (*Fagus sylvatica*) covering approximately 9%. The oak and hop hornbeam forests in this area are mainly privately owned and are managed in a coppice with standards system. Coppice systems consist of stands that originate from stool shoots or suckers of vegetative origin. At harvest time, some trees, the standards that are normally trees of seedling rather than coppice origin, are retained for purposes of seed dissemination and soil protection. In the study area, rotation ages are usually between 18 and 25 years, with cut widths of 1-2 ha and retention of 100-200 standards/ha most common. Conversely, most of the beech forests are unmanaged.

Field data

For scientific and inventory purposes, two-phase tessellation stratified sampling (TSS) was carried out in the study area. The area was covered by 437 hexagons, each with an area of 1 km2. During the first phase, a point was randomly selected in each hexagon and classified as "forest" if the point falls inside a forest area (at least 20% in canopy cover and 0.5 ha in extent) or "non-forest" based on the interpretation of high-resolution aerial ortho-photography. Of the 437 points, 197 were classified as "forest" (Figure 1). In the second phase, 62 points were randomly selected from the 197 "forest" points (sampling rate ≈ 30%) and surveyed in the field during 2009-2011 (Figure 1). For this study, the second phase sample was considered an equal probability sample of the forested portion of the study area.

The plot configuration consisted of a circular plot of 13-m radius. Inside the plot, the diameters at breast-height (1.30 m) (DBH) of all trees with DBH of at least 9.5 cm were measured. Height (H) was measured for a sub-sample of plot trees and predcited for the remaining trees using a model of the H-DBH relationship constructed using data for the measured trees.

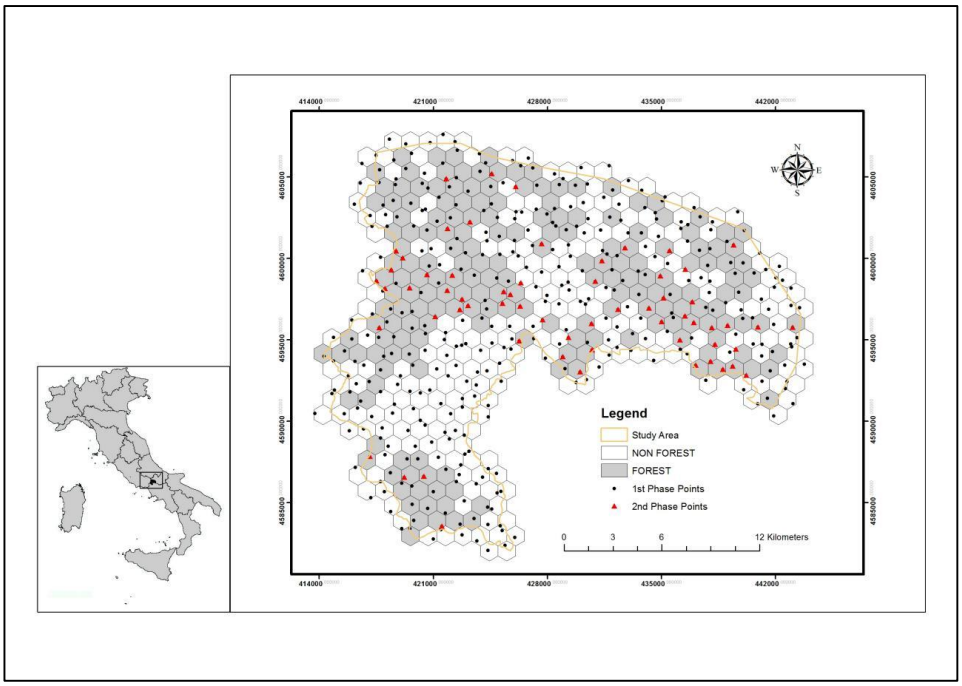


Fig. 1 Sampling design and plots locations in the study area.

Structural diversity indices

Using the data collected in the field, two common indices of forest structural diversity were calculated for each plot. Horizontal structural diversity was assessed using the standard deviation of DBH, calculated as,

$$\sigma_{DBH} = \sqrt{\frac{\sum_{i=1}^{n}(DBH_i - \overline{DBH})^2}{n-1}} \quad (1)$$

where i indexes trees, n is the number of trees on the plot, and $\overline{DBH}$ is mean plot-level DBH (cm). The advantages of using σ_{DBH} are that it can be easily calculated from any inventory method that yields stem density, is easily interpreted, and allows for temporal change detection (McRoberts et al., 2008). Further, it has been selected as a measure of horizontal structural diversity in the framework of international research programs (COST, 2006). Plot-level vertical structural diversity was assessed using the standard deviation of H calculated as,

$$\sigma_H = \sqrt{\frac{\sum_{i=1}^{n}(H_i - \overline{H})^2}{n-1}} \quad (2)$$

where i again index trees and $\bar{H}$ is mean H (m). Because heights of unmeasured trees are predicted using the H-DBH models, plot-level values of σ_H for this study are slightly reduced relative to values that would have been obtained if the heights of all trees had been measured. Table 1 reports the summary of sample statistics for these indices.

Table 1 Summary statistics of the sample.

Index	Min	$1st\ Qu.$	$Median$	$Mean$	$3rd\ Qu.$	Max
$\sigma_{DBH}\ (cm)$	1.50	3.84	5.20	5.56	7.25	25.21
$\sigma_H\ (m)$	1.31	2.12	2.58	2.90	3.22	8.18

In summary, the choice of the aforementioned indices of structural diversity was motivated by multiple factors: (i) they have been widely used in forestry and their reliability and usefulness is well-documented in the literature (McElhinny et al., 2005; McRoberts et al. 2008; Motz et al., 2010; Neumann & Starlinger, 2001; Lexerød & Eid, 2006; Sullivan et al. 2001; Uuttera et al., 2000), (ii) they are recognized as biodiversity indicators at international levels (EEA, 2012; MCPFE, 2011; Puumalainen et al., 2003), and as a result of scientific cooperation programs (Chirici et al., 2011a), and (iii) they can be easily estimated from common national forest inventory (NFI) data, which is a considerable advantage for biodiversity assessment and monitoring at large spatial scales (Winter et al., 2012, 2008; Chirici et al., 2012).

Airborne Laser Scanning (ALS) data

ALS data were acquired for the study area for scientific purposes related to the ITALID project - Use of LiDAR data to study Italian forests (Scrinzi et al., 2013) - under leaf-on canopy conditions in June 2010. A fixed-wing aircraft PartenaviaP68 was used. The LiDAR instrument was an Opthech Gemini LiDAR, a two-return range detection system that records a maximum of two echoes per laser pulse. The sensor was set with a maximum scan angle of 15° and a pulse frequency of 70 KHz, resulting in an average density of 1.5 pulses/m^2.

Common procedures for pre-processing ALS data include removal of outliers, ground/non-ground classification, and computation of normalized height. Firstly, air points that are clearly higher than the median elevation of surrounding points and isolated points with few neighbors resulting from sensor errors or backscatter by flying objects were removed. Subsequently a ground surface was generated by classifying ground points on the basis of the adaptive triangular irregular network (TIN) model algorithm (Axelsson, 2000). Lastly, the relative height above ground of each echo was calculated and used to extract the canopy metrics.

For each sample plot measured in the field, a set of 22 ALS height and density metrics was calculated and used as independent variables for constructing prediction models. Canopy cover (cov) was calculated as the proportion of first returns above 1.30 m on all first returns. Canopy density metrics were the proportions

of all returns above 1.30 m (dns), the proportion (d_{00}) and the count (c_{00}) of returns between 1.30 m and 10 m. The reference height of 1.30 m was chosen because plants taller than this threshold are no longer considered regeneration (INFC 2005). Canopy height metrics were the percentiles of the height canopy distribution ($p_{10}, p_{20}, \ldots, p_{90}, p_{99}$) and height summary statistics such as minimum (H_{min}), maximum (H_{max}), average (H_{avg}), standard deviation (H_{std}), coefficient of variability (H_{cv}), skewness (H_{ske}) and kurtosis (H_{kur}). Further, canopy relief ratio (CRR), a quantitative measure of the relative shape of the canopy, describing the proportion of all returns above mean value of echo heights was calculated as,

$$\mathrm{CRR} = \frac{\mathrm{H_{avg}} - \mathrm{H_{min}}}{\mathrm{H_{max}} - \mathrm{H_{min}}} . \quad (3)$$

CRR ranges between 0 and 1 and reflects the degree to which outer canopy surfaces are in the upper or lower half of the height range (Parker & Russ, 2004). The same metrics were calculated for the 23-m × 23-m pixels that tessellated the study area and were approximately of the same area as the plots. Maps were constructed by calculating model predictions of the structural diversity indices for these pixels.

Model development and spatial predictions

Initially, an exhaustive search was conducted for the linear model with the 22 ALS-derived metrics as candidate inresponse variables that maximized R^2,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad (4)$$

where y_i are observations of the response variables σ_{DBH} and σ_H, i indexes sample units, x_{ij} is the observation of the j^{th} ALS metric for the i^{th} sample unit, p is the number of ALS metrics selected, the □s are coefficients to be estimated, and □ is a residual term. To accomplish this task a subset regression procedure using a branch-and-bound algorithm (Clausen, 1999) addressing the optimization of R^2 for each predictor combination was used. This algorithm investigated all the possible combinations of ALS metrics as predictors, searching for the one-variable model, two-variable model, …, and five-variable models that maximized R^2. Preliminary analyses indicated little gain in R^2 when using more than five predictor variables. Models that have many predictor variables generally have greater R^2 values, but they also might suffer from overfitting and multicollinearity problems. To avoid such issues, the branch-and-bound algorithm was limited to searching for models with five or fewer predictor variables.

Regression diagnostics were used to check for potential outliers and multicollinearity problems. The presence of outliers was assessed by the leverage influence of an observation on the overall model behavior (Rawlings et al., 2008, pp.. 359-360). Problems of multicollinearity were minimized by using an F-test to select the model with the fewest number of predictors that still retained statistical significance, even at the expense of a decrease in R^2, where

$$F = \frac{(SSerr^{p} - SSerr^{p+q})/q}{SSerr^{p+q}/(n_{obs} - (p+q) - 1)}, \quad (5)$$

p and $p+q$ are the numbers of predictors in two models, q is the difference in the numbers of predictors between the two models, $SSerr^{p}$ is the sum of squared errors for the model with p predictors, $SSerr^{p+q}$ is the sum of squared errors for the model with $p+q$ predictors, n_{obs} is the number of observations, and 1 corresponds to the model intercept in linear models, if it is used. If $F \gtrsim 4.0$, then the model with $p+1$ predictor variables produces a statistically significantly better fit to the data than the model with p predictor variables, conversely if $F \lesssim 4$, the model with fewer predictors should be preferred. An underlying assumption is that the model with p predictor variables is nested within the model with $p+1$ predictor variables; the assumption is usually but not always satisfied with the result that the test then produces only approximate results.

Once the most statistically significant models with the fewest predictor variables were selected, they were used with the gridded ALS metrics to predict the structural diversity indices for all pixels in the study area. The resulting maps depict the spatial pattern of the structural diversity indices.

Inference

The design-based, simple random sampling (SRS) estimators for the population mean were used as basis for comparison with other estimators of α-diversity. In this framework, as reported in other studies (Whittaker, 1972; Lähde at al., 1999; Neumann & Starlinger, 2001; McRoberts et al., 2008), α-diversity was estimated as the mean value of the structural diversity index over the 62 plots. The mean value of each index was then attributed to the entire forested area.

Model-assisted estimators use models based on auxiliary data to enhance inferences but rely on the probability sample for validity. The model-assisted, generalized regression (GREG) estimators of means and variances were used (Särndal et al., 1992). The GREG estimator of the mean is calculated as

$$\hat{\mu}_{GREG} = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i - 1/n[\sum_{i=1}^{n}(\hat{y}_i - y_i)] \quad (6)$$

where N is the number of population units (the 23-m × 23-m forest cells in the study area), $\hat{y}_i$ is the model prediction for the *i-th* population unit, n is the sample size, $\hat{y}_i$ is the model prediction for the *i-th* sample plot and y_i is the observed value for the *i-th* plot. The second term adjusts the estimate for systematic model prediction error and can be considered as a bias estimator,

$$\hat{Bias}(\hat{\mu}_{initial}) = [\sum_{i=1}^{n}(\hat{y}_i - y_i)] \,, \quad (7)$$

where $\hat{\mu}_{initial}$ is the uncorrected population mean (first term Eq. 6). The corresponding variance estimator is,

$$\hat{Var}(\hat{\mu}_{GREG}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 \,, \quad (8)$$

where

$$\varepsilon_i = (\hat{y}_i - y_i)$$

and

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \,.$$

A confidence interval for the population mean is then estimated as

$$\hat{\mu} \pm t \cdot \sqrt{\hat{Var}(\hat{\mu}_{GREG})} \quad . \quad (9)$$

3.1.3. Results

Model development

Firstly, for the sake of simplicity the branch-and-bound algorithm was set to find models with 1-5 predictor variables that produced the greatest R^2 by searching for the best combination among the 22 ALS metrics. The algorithm was first applied to the entire sample dataset with n=62 observations.

The models fitted on the 62 observations explained approximately 33% ($0.30 < R^2 < 0.35$) of the variability for σ_{DBH} with a residual standard deviation of 3.69 (Table 2). The five models for σ_H explained approximately 36% ($0.25 < R^2 < 0.38$) of the variability with residual standard deviation of 1.06 (Table 3).

Regression diagnostics showed four plots to have high leverage and influence (Chatterjee & Hadi, 1986) on the fit of the model and were considered potential outliers. Removal of observations from a sample requires particular care to avoid deleting extreme observations because they help to avoid extrapolations when applying the model. This was not the case for σ_{DBH} for which the observations for the four plots were 13.86, 5.13, 17.87 and 5.41 for the sample range of 1.50-25.21, while for σ_H the observations were 4.05, 2.60, 3.21 and 2.10 for the sample range 1.13-8.18. Therefore, in our case the removal of the four plots does not affect the range of the values used to fit the model, and would not produce extrapolations (see also Table 1).

To assess the influence of the outliers on the overall model behavior, the selection algorithm was applied to the dataset without the candidate outliers. The models developed without the outliers improved the fit to the data by increasing R^2 by approximately 80% for σ_{DBH}, from 0.33 to 0.59, with an average decrease in root mean square error (RMSE) of approximately 26%, from 3.69 to 2.74, (Table 2). Better performances were also achieved for σ_H for which R^2 increased by approximately 40%, from 0.36 to 0.50, and RMSE decreased by approximately 10%, from 1.06 to 0.96 (Table 3). For both σ_{DBH} and σ_H, graphs of the observed plot-level diversity observations against their model predictions show good alignment along the 1:1 line (Figure 2 and Figure 3, respectively) and demonstrate the good quality of fit of the models to the data. Given these results, the models fit without the outliers were chosen as candidate models for calculating spatial predictions. However, during the inference phase, observations for all 62 plots were retained as a means of accurately estimating the actual variances.

The sum of squared errors of the models with 1-5 predictor variables were compared using the F-Test to keep the most statistically significant model with the fewest number of predictor variables. From this test, the ultimate selected models had four predictors for the σ_{DBH} model (Table 4), and the three predictors for the σ_H model (Table 5). This finding justifies the initial decision to investigate no more than 5 predictors.

Table 2 Adjusted R^2 and RMSE for the predictive models of σ_{DBH} with and without outliers.

No. Predictors	With outliers (62 obs)		Without outliers (58 obs)	
	Adj-R^2	RMSE	Adj-R^2	RMSE
1	0.34	3.67	0.53	2.95
2	0.35	3.64	0.58	2.80
3	0.32	3.74	0.60	2.72
4	0.30	3.78	0.63	2.64
5	0.35	3.64	0.63	2.61

Table 3 Adjusted R^2 and RMSE for the predictive models of σ_H with and without outliers.

No. Predictors	With outliers (62 obs)		Without outliers (58 obs)	
	Adj-R^2	RMSE	Adj-R^2	RMSE
1	0.34	1.07	0.43	1.02
2	0.35	1.07	0.46	1.00
3	0.35	1.07	0.52	0.94
4	0.37	1.05	0.54	0.92
5	0.38	1.04	0.56	0.91

Table 4 Parameter estimates for the final regression model for σ_{DBH}.

R^2=0.63; RMSE=2.52	
Predictors	Parameter estimated
$Intercept$	7.06**
H_{min}	-7.93***
H_{cv}	6.13**
H_{ske}	2.39*
p_{20}	1.76***

[a] Level of significance: *<0.05; **<0.01; ***<0.001

Table 5 Parameter estimates for the final regression model for σ_H.

R^2=0.52; RMSE=0.91	
Predictors	Parameter estimated
$Intercept$	3.46***
H_{min}	-2.48***
H_{cv}	2.68***
p_{20}	0.42***

[a] Level of significance: *<0.05; **<0.01; ***<0.001

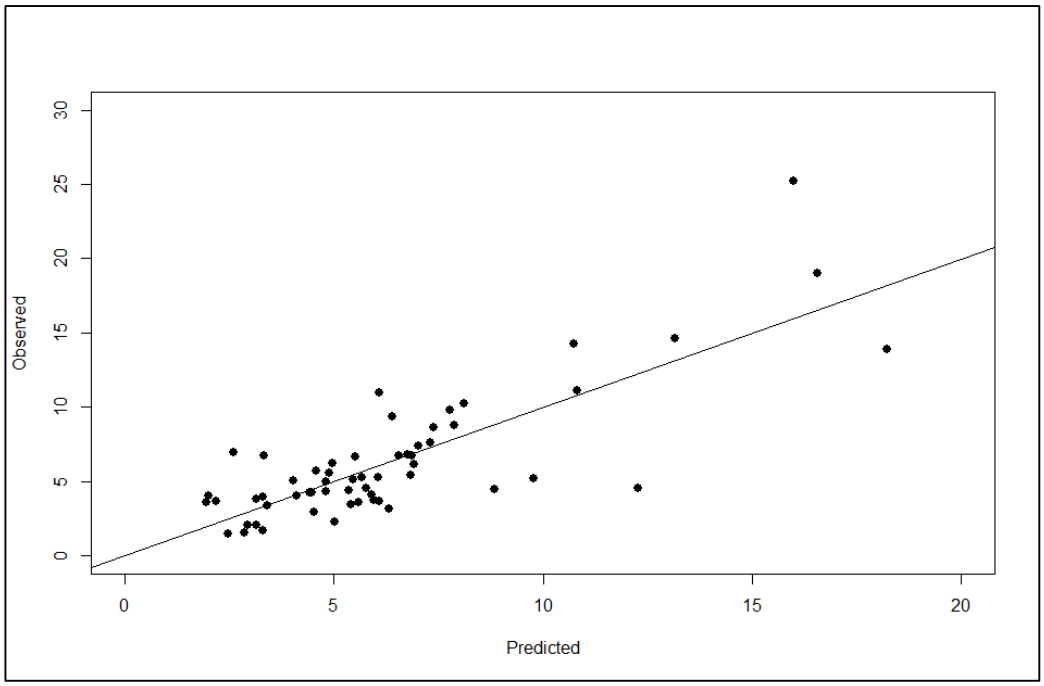


Fig. 2 Scatterplot of observed against predicted values for σ_{DBH}

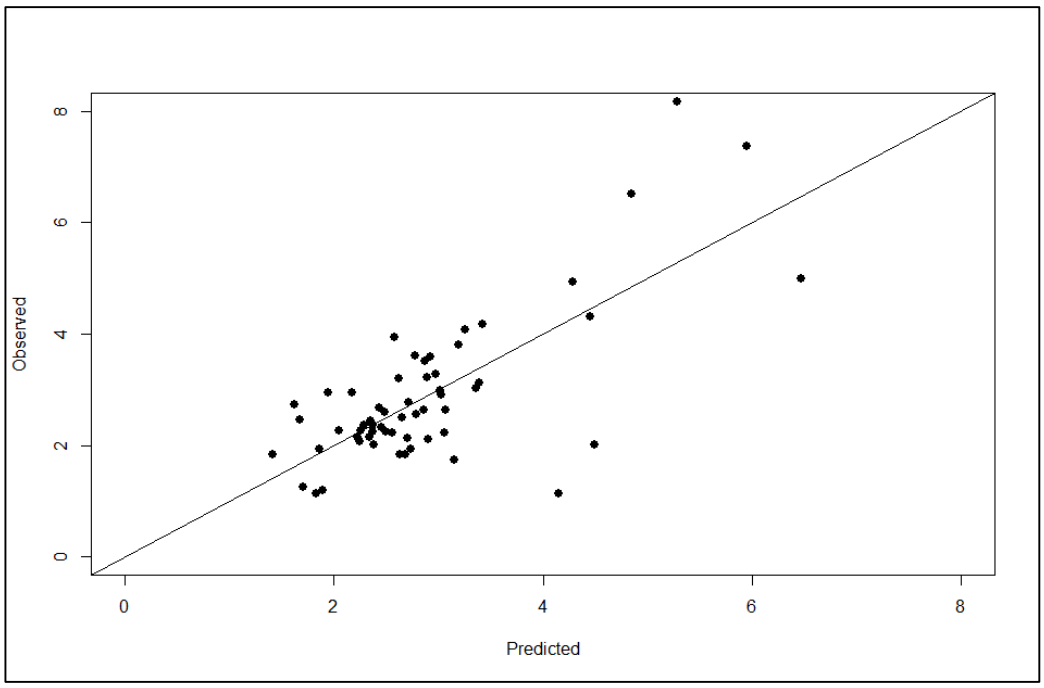


Fig. 3 Scatterplot of observed against predicted values for σ_H.

Spatial predictions

Model predictions for both structural diversity indices were calculated for grid pixels using the selected ALS metrics, thereby producing the maps of structural diversity (Figure 4 and Figure 5). The indices

were mapped only for forest land which was discriminated from non-forest land using a local forest type map (Chirici et al., 2011b). A potential difficulty with linear models is that extrapolations beyond the sample range of the predictor may be negative or unrealistically large. This was not the case for this study for which the proportions of negative predictions were only 1.4% and 0.2% for σ_{DBH} and σ_H, respectively, and the proportions of positive values greater than the maximum value in the sample were also only 0.6% and 0.1% for σ_{DBH} and σ_H, respectively. When constructing the map, negative predictions were set to 0, and predictions greater than the maximum value in the sample were set to the maximum sample observation plus twice the standard deviation of the 10 largest model residuals. This produced a range of predictions of 0-29 for the σ_{DBH} map and 0-10 for the σ_H map.

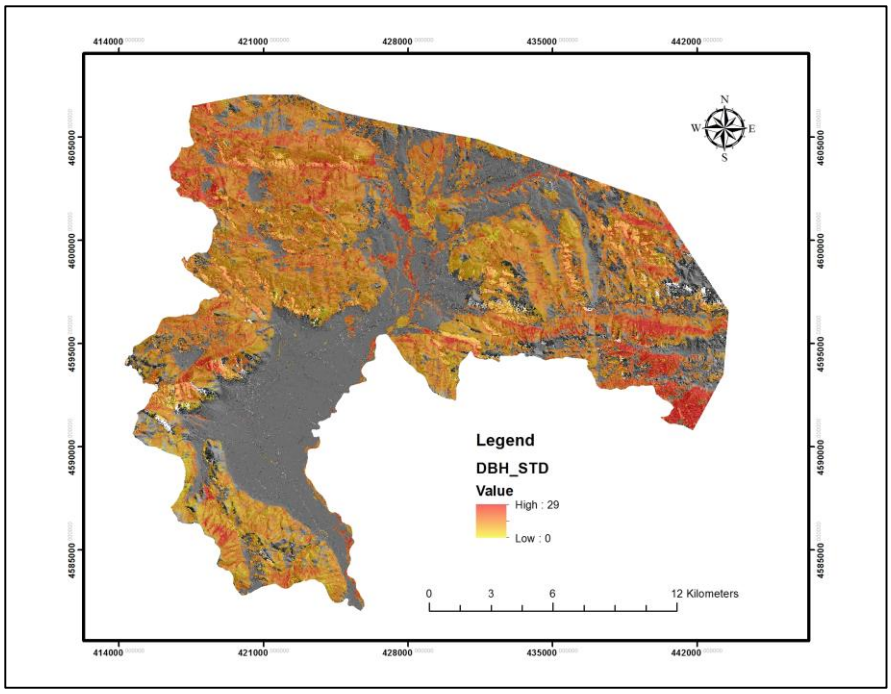


Fig. 4 Spatial predictions for σ_{DBH} for the forested portion of the study area (hillshade background).

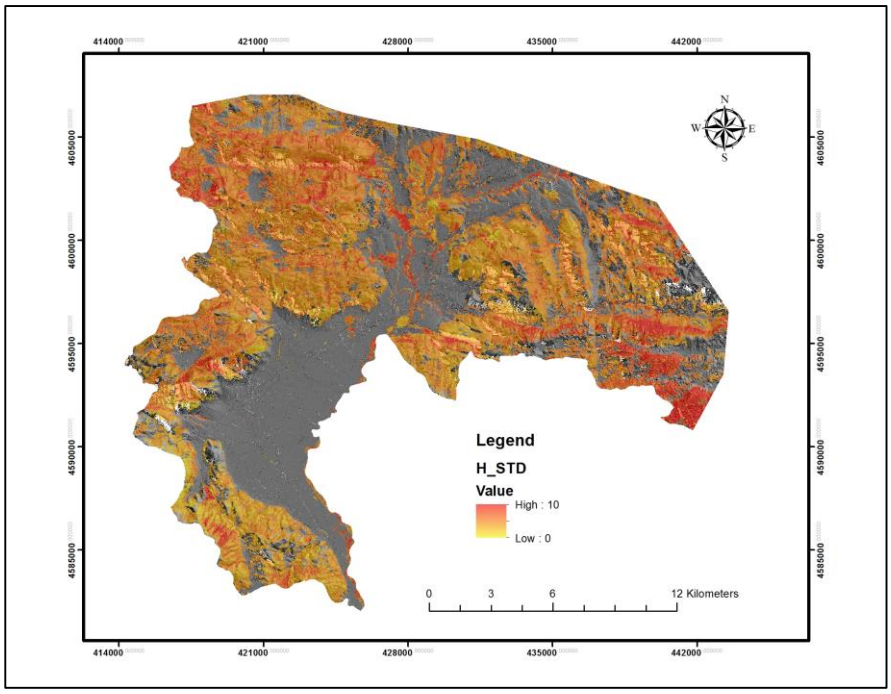


Fig. 5 Spatial predictions for σ_H for the forested portion of the study area (hillshade background).

Inference

For the entire study area, the SRS estimate of mean σ_{DBH} was 6.56 with standard error (SE) of 0.58, and the GREG estimate was 6.36 with SE=0.06 (Table 6). The SRS estimate of mean σ_H was 2.90 with SE=0.17, and the GREG estimate was 2.93 with SE=0.02.

Table 6. SRS and GREG estimates of means, standard errors, and a=0.95 confidence intervals.

Index	$\hat{\mu}_{SRS} \pm SE(\hat{\mu}_{SRS})$	$\hat{\mu}_{GREG}$	$B\hat{i}as(\hat{\mu}_{Initial})$	$\hat{\mu}_{GREG} \pm SE(\hat{\mu}_{GREG})$	$\hat{\mu}_{GREG} - t \cdot SE(\hat{\mu}_{GREG})$	$\hat{\mu}_{GREG} + t \cdot SE(\hat{\mu}_{GREG})$
σ_{DBH}	6.56 ±0.58	6.36	-0.31	6.36 ± 0.06	6.24	6.49
σ_H	2.90 ± 0.17	2.93	-0.01	2.93 ± 0.02	2.89	2.96

3.1.4. Discussions and Conclusions

The study focused on the forested portion of a study area in Molise, Italy, with a twofold objective: (i) to estimate the means for horizontal (σ_DBH) and vertical (σ_H) tree structural α-diversity using the design-based GREG estimators with ALS metrics and the SRS estimators, and (ii) to construct maps for both indices by calculating spatial model predictions using gridded ALS metrics. Removal of four outliers increased the proportion of the variability explained by the models for both indices. The probable reason for the outliers was harvest that occurred between the date of the ALS acquisition in June 2010 and the dates of field survey

over the three years, 2009-2011. All the outlier plots were located in oak and hop hornbeam dominated forests, which are mostly managed in a coppice with standards system.

The behavior of plots characterized as outliers can be explained by assuming a coppice with standards harvest between the ALS data acquisition and field measurement of plots. Under this assumption, the ALS data would indicate a fairly dense, mono-layered and even-aged forest structure. Further assuming that a harvest common to this management protocol in a coppice with standards system occurs before the field measurement, 80% of the trees would be removed leaving only "reservoirs" which serve dissemination and soil protection purposes. When the survey crew reaches such a plot, they would find a rather sparse forest with "reservoirs" consisting of the dominant tree layer and a rather dense understory layer that had originated from the regrowth of the suckers from the stumps. If the field survey date was before the harvest and the ALS acquisition data was after the harvest, the reverse situation would occur. In both cases, the structure predicted from the ALS data and that recorded by the field crew would be very dissimilar. Unfortunately, neither the exact dates for the ground survey nor the harvest are available to verify the hypothesis.

The ALS metrics selected as model predictor variables deserve consideration. Firstly, the linear models for both σ_DBH and σ_H share three variables that contributed significantly ($p < 0.01$) to increasing the quality of fit of the model to the data, namely H_min, H_cv and p_20. This result is consistent with natural forest dynamics for which greater height diversity in the form of multi-layer/stratified canopies corresponds to greater DBH diversity resulting from different sizes of trees in different strata (dominant, codominant, suppressed) and vice versa. Quantitative relationships between tree height and diameter are well established in the forestry literature and practice (Martin and Flewelling 1998), and are often used to link the two attributes (Buongiorno et al. 1994). However, because the relationships are often nonlinear (Martin and Flewelling 1998), separate characterizations for horizontal and vertical structure are more reliable and meaningful. Height variability has shown to be more indicative of the vertical layering of the canopy than the standard deviation of DBH. In addition, height variability can also act as a good indicator of wildlife diversity because greater tree height variability indicates trees of different ages and species that are more suitable to host multiple species of animals (Zenner and Hibbs 2000; Sullivan et al., 2001; Svensson and Jeglum, 2001). Conversely, the standard deviation of DBH is a measure of the variability in tree size, and is considered indicative for the presence and for the diversity of micro-habitats within a stand (Acker et al. 1998, Van Den Meerschautt & Vandekerkhove 1998). Neumann and Starlinger (2001) found large correlations between the standard deviation of DBH and a set of more complex indices of structural diversity. Similar results were reported by Zenner and Hibbs (2000) where a structural complexity index based on a three-dimensional model of forest structure was significantly correlated with the standard deviation of DBH. Additional studies demonstrated that the standard deviation of DBH was more useful than a measure of height diversity in discriminating between successional stages of stand development (Spies & Franklin, 1991).

Secondly, H_cv is a metric that by definition expresses variability and is naturally related to σ_DBH and σ_H. The metrics H_min and p_20 describe the lower portion of the canopy, which is strongly affected by the forest operations described above, and support the previous hypothesis. In fact, the ALS metrics related to the upper portion of the canopy (i.e. H_max and the higher percentiles) stay relatively unchanged because

harvesting in a coppice with standards system does not affect the upper canopy consisting of the "reservoirs" which are often the tallest and most well-shaped trees. This is the likely reason for the four outlier plots.

The spatial patterns of structural diversity depicted in the maps are in accordance with the known forest types and applied silvicultural systems in the area. Greater structural diversity occurs in the eastern side of the study area, dominated by unmanaged beech forests, which nowadays are approaching old-growth forest status with uneven-aged structures. In addition, greater structural diversity occurs along the rivers (linear patterns in the center of the maps) dominated by hygrophilus forests which are well known to be unmanaged to enhance protection against the effects of flooding. Conversely, less variability corresponds to the oak- and hop hornbeam-dominated forests managed in accordance with the coppice with standard system, which, over time, leads to even-aged, simpler, and mono-layered structures.

The model-assisted GREG estimators used for this study are widely used in conjunction with maps because they are mostly independent of the prediction method and they adjust for estimated bias resulting from systematic deviations between observations and predictions (Baffetta et al., 2009; McRoberts, 2009, 2010a, b).

The GREG estimators yielded estimates of the population means for the indices of α- structural diversity that were close to the SRS estimates. The GREG estimator for the variance produced a small variance for both the indices. This result is at least partially related to the goodness of fit of the models to the sample data. A primary advantage of model-assisted estimators is that they capitalize on the relationship between the observations and the auxiliary information for the population units in the sample to reduce the variance of the population parameter estimate (McRoberts et al., 2010c).

From a management perspective, maps that depict spatially-explicit patterns of structural diversity are of great use for locating hot spots where greater biodiversity is likely to occur, for assisting managers in planning adequate preservation strategies, and for assisting conservationists in prioritizing areas for biodiversity-oriented studies. Maps of σ_H can be used to monitor the presence of bird species because of the relationship between the variability of vertical vegetation structure and species abundance (Müller et al., 2009; Müller et al., 2010; Huang et al., 2014). Also, Matlock and Edwards (2006), Goetz et al. (2007) and Voegeler et al. (2014) found that greater canopy height variability corresponds to more forest bird species. Additionally, σ_H has shown to be valuable not only for avian species but also for prediction of assemblages of forest beetles (Müller and Brandl, 2009) and bat activity (Brown et al., 1997). As a measure of the variability in tree size, σ_DBH is indicative of the diversity of micro-habitats within a stand (Acker et al. 1998; Van Den Meerschautt & Vandekerkhove 1998) and is positively related to the presence of bird species (White et al., 2013). Inclusion of the diameters of the standing dead trees and snags in σ_DBH indirectly indicates the sizes of hollows which, in turn, is an indicator of the diversity of fauna (Van Den Meerschautt & Vandekerkhove 1998). In particular, because the needs of different arboreal species vary with respect to preferred entry size and internal hollow dimensions (Gibbons & Lindenmayer 1996; Lindenmayer & Franklin 1997; Whitford 2001), maximum diversity of arboreal fauna requires a range of sizes of tree hollows.

Five conclusions can be drawn from the study. Firstly, LiDAR data are a valid source of information for estimating and spatially predicting forest structural diversity and gains relevance for its potential role in characterizing forest ecosystems. Secondly, simple linear models were sufficient to characterize the relationship between the structural diversity indices and the ALS metrics used as predictor variables. Thirdly, the GREG estimator worked well for estimating mean α-structural diversity and greatly reduced the standard errors of the estimates of the means. Fourthly, the strong relationship between horizontal and vertical structural diversities is consistent with and confirmed by forest dynamics. Lastly, the maps of structural diversity can be used not only for planning management strategies addressing biodiversity, but also for preliminary hypotheses regarding silvicultural management systems because tree diameter and height are basic information for assessing the commercial value of tree logs.

3.1.5. References

Acker, S.A., Sabin, T.E., Ganio, L.M. & McKee, W.A. (1998). Development of old-growth structure and timber volume growth trends in maturing Douglas-fir stands. Forest Ecology and Management, 104, 265-280.

Alberti, G., Boscutti, F., Pirotti, F., Bertacco, C., De Simon, G., Sigura, M., Cazorzi, F., & Bonfanti, P. (2013). A LiDAR-based approach for a multi-purpose characterization of Alpine forests: an Italian case study. iForest - Biogeosciences and Forestry, 6(3), 156-168.

Axelsson, P. (2000). DEM generation from laser scanner data using adaptive TIN models. International Archives of Photogrammetry and Remote Sensing, 33(B4), 110-117.

Baffetta, F., Fattorini, L., Franceschi, S. & Corona, P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment, 111(3), 463-475.

Bergen, K.M., Goetz, S.J., Dubayah, R.O., Henebry, G.M., Hunsaker, C.T., Imhoff, M.L.,Nelson, R.F., Parker, G.G., & Radeloff, V.C. (2009). Remote sensing of vegetation 3-D structure for biodiversity and habitat: Review and implications for lidar and radar spaceborne missions. Journal of Geophysical Research, 114(G2), 1-13.

Brokaw, N.V.L., & Lent, R.A. (1999). Vertical structure. In: Hunter ML (Ed). Maintaining biodiversity in forest ecosystems. Cambridge, UK: Cambridge University Press.

Brown, G.W., Nelson, J.L., & Cherry K.A. (1997). The influence of habitat structure on insectivorous bat activity in montane ash forests of the Central Highlands, Victoria. Australian Forestry, 60(2), 138-146.

Buongiorno, J., Dahir, S., Lu, H. & Lin, C. (1994). Tree size diversity and economic returns in uneven-aged forest stands. Forest Science, 40(1), 83-103.

Chatterje, S., & Hadi, A.S. (1986). Influential observations, High Leverage Points, and Outliers in Linera Regression. Statistical Science, 1(3), 379-393.

Chirici, G., Winter, S., & McRoberts, R.E. (2011a). National Forest Inventories : Contributions to Forest Biodiversity Assessments Managing Forest Ecosystems, Springer.

Chirici, G., Di Martino, P., Ottaviano, M., Santopuoli, G., Chiavetta, U., Tonti, D., Garfì, V., & Marchetti, M. (2011b). La carta forestale su basi tipologiche. In Garfì, V., & Marchetti, M. (Eds.). Tipi forestali e preforestali della regione Molise, Edizioni dell'Orso.

Chirici, G., McRoberts, R.E., Winter, S., Bertini, R., Brändli, U.B., Asensio, I.A., Bastrup-Birk, A., Rondeux, J., Barsoum, N., & Marchetti, M. (2012). National Forest Inventory Contributions to Forest Biodiversity Monitoring. Forest Science, 58(3), 257-268.

Clausen, J. (1999). Branch and Bound Algorithms—Principles and Examples (Technical report). University of Copenhagen.

Clawges, R., Vierling, K., Vierling, L., & Rowell, E. (2008). The use of airborne lidar to assess avian species diversity, density, and occurrence in a pine/aspen forest. Remote Sensing of Environment, 112(5), 2064-2073.

Cooperation in the Field of Science and Technology (2006). Harmonization of national forest inventories in Europe: techniques for common reporting. COST E43. Finnish Forest Research Institute, Helsinki, Finland. Available at: http://www.metla.fi/eu/cost/e43/; Last accessed: February 2015.

Dinerstein, E., Olson, D.M., Graham, D.J., Webster, A.V., Primm, S.A., Bookbinder, M.P., & Ledec, G. (1995). Una evaluación del estado de conservación de las ecoregiones terrestres de América Latina y el Caribe. Publicado en colaboración con el Fondo Mundial para la Naturaleza. Washington, DC: Banco Mundial.

EEA, (2012). Streamlining European biodiversity indicators 2020: Building a future on lessons learnt from the SEBI 2010 process. EEA Technical report No 11/2012, Copenhagen, 2012.

Ewald, M., Dupke, C., Heurich, M., Müller, J., & Reineking, B. (2014). LiDAR Remote Sensing of Forest Structure and GPS Telemetry Data Provide Insights on Winter Habitat Selection of European Roe Deer. Forests, 5(6), 1374-1390.

Flashpoler, D.J., Giardina, C.P., Asner, G.P., Hart, P., Price, J.T., Lyons, C.K., & Castaneda, X. (2010). Long-term effects of fragmentation and fragment properties on bird species richness in Hawaiian forests. Biological Conservation 143(2), 280-288.

Fleishman, E., & Mac Nally, R. (2006). Patterns of spatial autocorrelation of assemblages of birds, floristics, physiognomy, and primary productivity in the central Great Basin, USA. Diversity Distribution 12(3), 236-243.

Gaston, K.J., & Spicer, J.I. (2004). Biodiversity: An introduction, 2nd ed.. Blackwell Publishing.

Gibbons, P. & Lindenmayer, D.B. (1996). Issues associated with the retention of hollow-bearing trees within eucalypt forests managed for wood production. Forest Ecology And Management, 83, 245-279.

Goetz, S., Steinberg, D., Dubayah, R., & Blair, B. (2007). Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperate forest, USA. Remote Sensing of Environment, 108(3), 254-263.

Graf, R.F., Mathys, L., & Bollmann, K., (2009). Habitat assessment for forest dwelling species using LiDAR remote sensing: Capercaillie in the Alps. Forest Ecology and management, 257(1), 160-167.

Halaj, J., Ross, D.W., & Moldenke, A.R. (2000), Importance of habitat structure to the arthropod food-web in Douglas-fir canopies, Oikos, 90(1), 139-152.

Holdridge, L.R., (1947). Determination of world plant formation from simple climatic data. Science, 105(2727), 367-368.

Holdridge, L.R. (1967). Life zone ecology. San Jose: Tropical Science Center.

Huang, Q., Swatantran, A., Dubayah, R., & Goetz, S.J. (2014). The Influence of Vegetation Height Heterogeneity on Forest and Woodland Bird Species Richness across the United States. PLoS ONE 9(8).

Hunter, M.L.Jr. (1990). Wildlife, forests, and forestry: Principles of managing forests for biological diversity. Englewood Cliffs: Prentice-Hall.

Hunter, M.L.Jr. (Ed.). (1999). Maintaining biodiversity in forest ecosystem. Cambridge: Cambridge University Press.

Hyde, P., Dubayah, R., Peterson, B., Blair, J.B., Hofton, M., Hunsaker, C., Knox, R., & Walker, W. (2005). Mapping forest structure for wildlife habitat analysis using waveform lidar: validation of montane ecosystems. Remote Sensing of Environment, 96(3-4), 427-437.

Hyde, P., Dubayah, R., Walker, W., Blair, B. J., Hofton, M., & Hunsaker C. (2006). Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM plus, Quickbird) synergy. Remote Sensing of Environment, 102(1-2), 63-73.

Jung, K., Kaiser, S., Böhm, S., Nieschulze, J., & Kalko, E.K.V. (2012). Moving in three dimensions: effects of structural complexity on occurrence and activity of insectivorous bats in managed forest stands. Journal of Applied Ecology, 49(2), 523-531.

Kuuluvainen, T. (2009). Forest management and biodiversity conservation based on natural ecosystem dynamics in northern Europe: the complexity challenge, AMBIO, 38(6), 309-315.

Lähde, E., Laiho, O., Norokorpi, Y., & Saksa, T. (1999). Stand structure as the basis of diversity index. Forest Ecology and Management, 115(2-3), 213-220.

Lassau, S.A., & Hochuli, D.F. (2008). Testing predictions of beetle community patterns derived empirically using remote sensing. Diversity Distribution, 14(1), 138-147.

Lassau, S.A., Cassis, G., Flemons, P.K.J., Wilkie, L., & Hochuli, D.F. (2005a). Using high-resolution multi- spectral imagery to estimate habitat complexity in open-canopy forests: can we predict ant community patterns?. Ecography, 28(4), 495-504.

Lassau, S.A., Hochuli, D.F., Cassis, G., & Reid, C.A.M. (2005b). Effects of habitat complexity on forest beetle diversity: do functional groups respond consistently? Diversity Distribution, 11(1), 73-82.

Lefsky, M.A., Cohen, W.B., Parker, G.G., & Harding, D.J. (2002). Lidar Remote Sensing for Ecosystems Studies. BioScience 52(1), 19-30.

Lesak, A.A., Radeloff, V.C., Hawbaker, T.J., Pidgeon, A.M., Gobakken, T., & Contrucci, K. (2011). Modeling forest songbird species richness using LiDAR-derived measures of forest structure. Remote Sensing of Environment, 115(11), 2823-2835.

Leutner, B.F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., & Wegmann, M. (2012). Modelling forest α-diversity and floristic composition — On the added value of LiDAR plus hyperspectral remote sensing. Remote Sensing, 4(12), 2818--2845.

Leveque, C. (1994). Environement et diversité du vivant. Pocket Sciences, Collection Explora.

Lexerød, N.L., & Eid, T. (2006). An evaluation of different diameter diversity indices based on criteria related to forest management planning. Forest Ecology and Management, 222(1-3), 17-28.

Lim, K., Treitz, P., Wulder, M., St-Onge, B., & Flood, M. (2003). LiDAR remote sensing of forest structure. Progress in Physical Geography, 27(1), 88-106.

Lindenmayer, D.B., & Franklin, J.F. (1997). Managing stand structure as part of ecologically sustainable forest management in Australian Mountain Ash Forests. Conservation Biology, 11(5), 1053-1068.

Lucas, K.L., Raber, G.T., & Carter, G.A. (2010). Estimating vascular plant species richness of Horn Island, Mississippi using small-footprint airborne LiDAR. Journal of Applied Remote Sensing, 4(1).

MacArthur, R.H., & MacArthur, J. (1961). On bird species diversity. Ecology, 42(3), 594-598.

Martin, F.C. & Flewelling, J.W. (1998). Evaluation of tree height prediction models for stand inventory. Western Journal of Applied Forestry, 13, 109-119.

Martinuzzi, S., Vierling, L., Gould, W., Falkowski, M.J., Evans, J.S., Hudak, A.T., & Vierling, K.T. (2009). Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. Remote Sensing of Environment, 113(12), 2533-2546.

Matlock, R.B., & Edwards, P.J. (2006). The Influence of Habitat Variables on Bird Communities in Forest Remnants in Costa Rica. Biodiversity & Conservation, 15(9), 2987-3016.

McGraw, S. (1994). Census, habitat preference, and polyspecific associations of six monkeys in the Lomako Forest, Zaire. American Journal of Primatology, 34(4), 295-307.

MCPFE (2002). Improved pan-European indicators for sustainable forest management. MCPFE Liaison Unit, Vienna.

MCPFE (2011). State of Europe's Forest: Status & Trends in Sustainable Forest Management in Europe.

Montréal Process (2006). Criteria and indicators for the conservation and sustainable management of temperate and boreal forests. Montréal Process Liaison Office, International Forestry Cooperation Office, Forestry Agency, Japanese Ministry of Agriculture, Forestry and Fisheries, Tokyo, Japan.

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

McRoberts, R.E., Winter, S., Chirici, G., Hauk, E., Pelz, D.R., Moser, W.K., & Hatfield, M.A. (2008). Large-scale spatial patterns of forest structural diversity. Canadian Journal of Forest Research, 38(3), 429-438.

McRoberts, R. E. (2009). Satellite image-based estimates of forest area for international reporting. Proceedings of the IUFRO Conference, International Symposium on National Forest Inventory: Forest resources monitoring and climate change, Yangyang, Gangwon Province, Korea, September 8-9, 2009, Korea Forest Research Institute.

McRoberts, R.E. (2010a). Probability- and model-based approaches to estimating proportion forest using satellite imagery. Remote Sensing of Environment, 114, 1017-1025.

McRoberts, R.E. (2010b). The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area. Remote Sensing of Environment, 114, 1710-1717.

McRoberts, R.E., Cohen, W.B., Næsset, E., Stehman, S.V., & Tomppo, E.O. (2010c). Using remotely sensed data to construct and assess forest attribute maps and related spatial products. Scandinavian Journal of Forest Research, 25(4), 340-367.

McRoberts R.E., (2011). Satellite image-based maps: Scientific inference or pretty pictures? Remote Sensing of Environment, 115(2), 715-724.

Motz, K., Sterba, H., & Pommerening, A. (2010). Sampling measures of tree diversity. Forest Ecology and Management, 260(11), 1985-1996.

Müller, J., & Brandl, R. (2009). Assessing biodiversity by remote sensing in mountainous terrain: the potential of LiDAR to predict forest beetle assemblages. Journal of Applied Ecology, 46(4), 897-905.

Müller, J., Moning, C., Bässler, C., Heurich, M., & Brandl, R. (2009). Using airborne laser scanning to model potential abundance and assemblages of forest passerines. Basic Applied Ecology, 10(7), 671-681.

Müller, J., Stadler, J., & Brandl, R. (2010). Composition versus physiognomy of vegetation as predictors of bird assemblages: The role of lidar. Remote Sensing of Environment, 114(3), 490-495.

Nelson, R., Keller, C., & Ratnaswamy, M. (2005). Locating and estimating the extent of Delmarva fox squirrel habitat using an airborne LiDAR profiler. Remote Sensing of Environment, 96(3-4), 292-301.

Neumann, M., & Starlinger, F. (2001). The significance of different indices for stand structure and diversity in forests. Forest Ecology and Management, 145(1-2), 91-106.

Noss, R.F. (1990). Indicators for monitoring biodiversity: a hierarchical approach. Conservation Biology, 4(4), 355-364.

O'Neill, R.V., Hunsaker, C.T., Jones, K.B., Riitters, K.H., Wickham, J.D., Schwartz, P.M., Goodman, I.A., Jackson, B.L., & Baillargeon, W.S. (1997). Monitoring environmental quality at the landscape scale. Bioscience, 47(8), 513-519.

Parker, G.G., & Russ, M.E. (2004). The canopy surface and stand development: Assessing forest canopy structure and complexity with near-surface altimetry. Forest Ecology and Management, 189, 307-315.

Puumalainen, J., Kennedy, P., & Folving, S. (2003). Monitoring forest biodiversity: a European perspective with reference to temperate and boreal forest zone. Journal of Environmental Management, 67(1), 5-14.

Rondeux, J., & Sanchez, C., (2010). Review of indicators and field methods for monitoring biodiversity within national forest inventories. Core variable: deadwood. Environmental Monitoring Assessment, 164(1-4), 617-30.

Rosenzweig, M.L. (1995). Species diversity in space and time. Cambridge University Press, Cambridge/New York.

Rowlings, J.O., Pantula, S.G., & Dickey, D.A. (1998). Applied regression analysis: a research tool, 2nd ed. Springer-Verlag New York, Inc.

Salter, R.E., Mackenzie, N.A., Nightingale, N., Aken, K.M., & Chai, P.K.P. (1985). Habitat use ranging behavior and food habits of the Proboscis monkey Nasalis-larvatus in Sarawak. Primates, 26(4), 436–451.

Särndal, C.E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. Springer-Verlag, Inc. New York.

Schaffers, A.P., Raemakers, I.P,, Sýkora, K.V., Ter Braak, C.J.F. (2008). Arthropod assemblages are best predicted by plant species composition. Ecology, 89(3), 782-794.

Scrinzi, G., Clementel, F., Colle, G., Corona, P., Floris, A., Maistrelli, F., Chirici, G., Mura, M., Oradini, A., Bertani, R., Barbati, A., Quatrini, A., & Marchetti, M. (2013). Impiego di dati LiDAR di pubblica disponibilità per il monitoraggio forestale a grande e piccola scala: il progetto ITALID. In: Proceedings of the 9th SISEF National Congress "Multifunzionalità degli Ecosistemi Forestali: Sfide e Opportunità per la Ricerca e lo Sviluppo", Bolzano (Italy) 16-19 Sep 2013.

Shine, R., Barrott, E.G., & Elphick, M.J. (2002). Some like it hot: Effects of forest clearing on nest temperatures of montane reptiles. Ecology, 83(10), 2808-2815.

Simonson, W.D., Allen, H.D., & Coomes, D. (2012). Use of an airborne lidar system to model plant species composition and diversity of Mediterranean oak forests. Conservation Biology, 26(5), 840-50.

Simpson E. H., (1949). Measurement of diversity. Nature 163, 688.

Spies, T.A. & Franklin, J.F. (1991). The structure of natural young, mature, and old-growth Douglas-Fir forests in Oregon and Washington, in Wildlife and Vegetation of Unmanaged Douglas-Fir Forests, USDA Forest Service, Portland, Oregon, 91-109.

Sullivan, T.P., Sullivan, D.S., & Lindgren, P.M.F. (2001). Stand structure and small mammals in young Lodgepole Pine forest: 10- year results after thinning. Ecological Society of America 11(4), 1151-1173.

Svensson, J.S. & Jeglum, J.K. (2001). Structure and dynamics of an undisturbed old-growth Norway spruce forest on the rising Bothnian coastline. Forest Ecology and Management, 151, 67-79.

Swatantran, A., Dubayah, R., Goetz, S., Hofton, M., Betts, M.G., Sun, M., Simard, M., & Holmes, R. (2012). Mapping migratory bird prevalence using remote sensing data fusion. PLoS ONE, 7(1), 1-11.

Ter Braak, C.J.F., & Schaffers, A.P. (2004). Co-correspondence analysis: a new ordination method to relate two community compositions. Ecology, 85(3), 834-846.

Tilman, D., Knops, J., Weldin, D., Reich, P., Ritchie, M., & Sieman, E. (1997). The influence of functional diversity and composition on ecosystem processes. Science, 277(5330), 1300-1302.

Uuttera, J., Tokola, T., & Maltamo, M. (2000). Differences in the structure of primary and managed forests in East Kalimantan Indonesia. Forest Ecology and Management, 129(1-3), 63-74.

Van Den Meersschaut, D. & Vandekerkhove, K. (1998). Development of a standscale forest biodiversity index based on the State Forest Inventory, in Integrated Tools for Natural Resources Inventories in the 21st Century, USDA, Boise, Idaho, USA, 340-349.

Vierling, K.T., Bässler, C., Brandl, B., Vierling, L.A., Weiß, I., & Müller, I. (2011). Spinning a laser web: predicting spider distributions using lidar. Ecological Applications, 21(2), 577-588.

Vierling, L.A., Vierling, K.T., Adam, P., & Hudak, A.T. (2013). Using satellite and airborne LiDAR to model woodpecker habitat occupancy at the landscape scale. PloS One, 8(12), 1-13.

Vogeler, J.C., Hudak, A.T., Vierling, L.A., Evans, J., Green, P., & Vierling, K.T. (2014). Terrain and vegetation structural influences on local avian species richness in two mixed-conifer forests. Remote Sensing of Environment, 147, 13-22.

Welsh, H.H., & Lind, A.J. (1996). Habitat correlates of the southern torrent salamander, Rhyacotriton variegatus (Caudata: Rhyacotritonidae), in northwestern California. Journal of Herpetology, 30(3), 385-398.

White, A.M., Zipkin, E.F., Manley, P.N., & Schlesinger, M.D. (2013). Conservation of Avian Diversity in the Sierra Nevada: Moving beyond a Single-Species Management Focus. PLoS ONE, 8(5).

Whitford, K.R. (2001). Dimensions of tree hollows used by birds and mammals in the jarrah forest: improving the dimensional description of potentially usable hollows. CALMScience, 3(4), 499-511.

Whittaker, R.H. (1972). Evolution and measurement of species diversity. Taxon, 21(2-3), 213-251.

Wilson, E.O. (1987). An urgent need to map biodiversity. Scientist, 1(6), 1-11.

Winter, S., Chirici, G., McRoberts, R.E., Hauk, E., & Tomppo, E. (2008). Possibilities for harmonizing national forest inventory data for use in forest biodiversity assessments. Forestry, 81(1), 33-44.

Winter, S., Böck, A., & McRoberts, R.E. (2012). Uncertainty of large-area estimates of indicators of forest structural gamma diversity: A study based on national forest inventory data. Forest Science, 58(3), 284-293.

Wulder, M.A. Bater, C.W. Coops, N.C. Hilker, T. & White, J.C. (2008). The role of LiDAR in sustainable forest management. The Forestry Chronicle, 84(6), 807-826.

Zenner, E.K. & Hibbs, D.E. (2000). A new method for modeling the heterogeneity of forest structure. Forest Ecology and Management, 129, 75-87.

Zimble, D.A., Evans, D.L., Carlson, G.C., Parker, R.C., Grado, S.C., & Gerard, P. D. (2003). Characterizing vertical forest structure using small-footprint airborne LiDAR. Remote Sensing of Environment, 87(2-3), 171-182.

3.2. Study V

From:

[1]Mura M., [2]McRoberts R. E., [3]Chirici G., [1]Marchetti M. (in review). Statistical inference for multiple-variable, forest structural diversity indices using airborne laser scanning data and the k-Nearest Neighbors technique. *Remote Sensing*.

[1] Dipartimento di Bioscienze e Territorio, University of Molise, Contrada Fonte Lappone, Pesche (IS), 86090 Italy

[2] Northern Research Station, U.S. Forest Service, Saint Paul, Minnesota 55108 USA

[3] Department of Agricultural, Food and Forestry Systems, University of Florence, Florence 50145 Italy

Abstract

Forest structural diversity plays a major role for forest management, conservation and restoration and is recognized as a fundamental aspect of forest biodiversity. The assessment, maintenance and restoration of a diversified forest structure have become major foci in the effort to preserve forest ecosystems from loss of biological diversity. However, the assessment of forest biodiversity is difficult because it involves multiple components and is characterized using multiple variables. The objective of the study was to develop a methodological approach for mapping, estimating, and constructing a statistical inference for a multiple variable index of forest structural diversity. The method included three key components: (i) use of the k-Nearest Neighbors (k-NN) technique, field plot data, and airborne laser scanning metrics to predict multiple forest structural diversity variables simultaneously, (ii) incorporation of the multiple diversity variable predictions into a single index, and (iii) construction of a statistically rigorous inference for the population mean of the index. Three structural diversity variables were selected to illustrate the method: growing stock volume and the standard deviations of height and diameter at-breast-height. Optimization of the k-NN technique produced mean relative deviations of less than 0.05, R2 values in the range of 0.50 to 0.66 which were within or close to values reported in the literature, and a confidence interval for the population mean of the index whose half-width was approximately 3% of the mean. Finally, the spatial pattern depicted in the resulting map of forest structural diversity for the study area contributed to validating the proposed method.

Keywords: airborne laser scanning; k-Nearest Neighbors; structural diversity; forest biodiversity

3.2.1. Introduction

Forest structural diversity has been recognized as a fundamental component of forest biodiversity assessment and monitoring [1]. Biodiversity, from both general ecological and applied forestry perspectives, is characterized by a larger number of plant and animal communities sharing a common multidimensional space of habitats and niches and making greater use of available resources [2]. Loss of these habitats and niches triggers a loss of biodiversity [3,4]. Forests and wooded lands are the richest ecosystems from biological and genetic prospectives [5,6,7], with anthropogenic activities constituting the main causes of the loss of forest biodiversity worldwide [8]. Thus, the assessment, maintenance and the restoration of forest structural diversity have become major foci in the effort to preserve forest ecosystems from loss of biological diversity.

Forests are complex and adaptive systems [9] and given the complexity of the biotic and abiotic interactions, compositional, structural, and functional attributes are all involved in the assessment of forest diversity. However, functional attributes describing cycles of mass and energy among the components can be only assumed and only rarely verified, while compositional aspects are rarely measured in the field; thus lack of data and failure to understand fully mechanisms underlying ecophysiological processes cause assessments to become approximate, possibly subjective, and applicable only for isolated cases [1].

Another aspect is that often the only data available to investigate forest diversity at a large scale are obtained from national forest inventories (NFI) [1,10] for which the historically main (and often the only) objective has been to quantify the amount and extent of forest woody resources for a given area while seeking a compromise between the precision of estimates and limited financial resources. The compromise leads to limiting acquisition of field data to information for trees that satisfy size thresholds, often the diameter at breast-height (DBH) and species that contribute most in terms of woody mass and omission of important ecological information for herbs, brush, animals, habitat trees and smaller and younger trees that are important for the functional dynamics of an ecosystem.

However, the structural diversity attributes of the macro component of tree communities gathered by NFIs are objective, reliable, and easier to calculate and understand when compared to more complex indices relying on functional and compositional aspects, which would offer little useful information for non-expert policy-makers [11]. Tree-level attributes related to forest structure such as DBH and height (H) are commonly measured as part of forest surveys, and other structural variables such as growing stock volume can be objectively and quantifiably estimated. In their review, McElhinny et al. [2] assert that forest structure is more relevant than composition for biodiversity assessments. The explanation is that more diverse stand structures are likely to have more niches, and therefore, support more species which results in more efficient use of available resources [2].

Unmanaged forests tend to have greater structural heterogeneity than managed forests. Such forests can better resist the effects of influential internal and external adverse factors [9]. Thus, unmanaged forests have greater resilience than managed forests. Moreover, measures of forest structural diversity are characterized by attributes that are judged indispensable for assessing forest diversity. They are reliable in producing objective, consistent and precise results [12,13,14]; they are widely available and easy to calculate [14,15]; over time they respond to changes in forest dynamics [16]; and they are appropriate for assessment

at multiple scales [15]. Therefore, structural attributes may be reasonably assumed to constitute a reliable basis for objective assessments of structural diversity. Further, from a management perspective, maps of the spatially-explicit patterns of structural diversity are of great use for locating hot spots where greater biodiversity is likely to occur, for assisting managers in planning adequate preservation strategies, and for assisting conservationists in prioritizing areas for biodiversity-oriented studies.

Multiple measures of forest structural diversity have been proposed and evaluated. Lexerød and Eid [17] evaluated eight diameter diversity measures for forest management purposes; Pommerening [10] evaluated eight measures for habitat functions and forest management planning; Latifi [18] considered multiple categories of measures that can be estimated using remotely sensed data; and Neumann and Starlinger [19] evaluated 11 measures for assessing the effects of air pollution. Staudhammer and LeMay [20] and Müller and Vierling [21] noted that of the many measures, the most commonly used include diameter, height, or both. With respect to particular measures, Staudhammer and LeMay [20] reported that the Shannon index performed well for ranking spatial areas with respect to degree of diversity, but Lexerød and Eid [17] reported that the Gini index was superior for boreal forest planning purposes. The relevant conclusions from the literature are that a multitude of measures of forest structural diversity are feasible, but that prospects for a globally superior measure are unlikely. Further, whereas the vast majority of proposed measures address only a single component of diversity such as height, diameter, or spatial location, diversity encompasses multiple components.

The objective of the study was to develop a methodological approach for mapping, estimating, and constructing a statistical inference for a multiple-variable index of forest structural diversity. The approach relies on prediction of forest structural diversity variables using airborne laser scanning (ALS) metrics [22,23,24,25,26] and features three innovative components: (i) use of the multivariate, non-parametric k-Nearest Neighbors technique (k-NN) to predict multiple forest structural diversity response variables simultaneously, (ii) development of a multiple-variable index that integrates any combination of particular single-variable forest structural diversity variables, and (iii) statistically rigorous inference for the population mean of the multiple-variable index. The k-NN technique is well-suited for this approach because it permits simultaneous prediction of multiple response variables, is not constrained by distributional assumptions, and is well-documented for use with forest inventory data. A bootstrap resampling technique is used to estimate the uncertainty of the estimated mean of the multiple-variable structural diversity index. Although the primary study objective was methodological, the approach is illustrated for a study area in Molise, Italy. Of importance, the selected forest structural diversity variables are intended to be illustrative rather than definitive because relevant forest structural diversity variables vary for each application and study area. Nevertheless, the general approach consisting of the multivariate k-NN prediction technique, the multi-variable index, and the inferential approach is applicable for any application and study area.

3.2.2. Materials and Methods

Study area

The study area included 36,360 ha in the southwestern part of Molise Region in central Italy (Figure 1). Approximately 56% of the area, corresponding to 20,518 ha, is covered by forests. The forest area is 60% deciduous oaks (*Quercus cerris, Quercus pubescens*), 18% hop hornbeam (*Ostrya carpinifolia*), 9% beech (*Fagus sylvatica*), 7% evergreen holm oak (*Quercus ilex*), 4% hygrophilus forest, and the remainder in species of less than 1% each. Oak and hop hornbeam forests are mainly privately-owned and managed using a coppice with standards system characterized by rotation ages between 18 and 25 years, cuts of 1–2 ha wide, and 100-200 residual standards/ha. On the other hand, and more relevant for this study, the beech (*Fagus sylvatica*) forests are undamaged and now have structures approaching natural, old-growth forest status.

Field data

The study area was tessellated into 437 hexagons, each with area of 1 km2, and two-phase tessellation stratified sampling (TSS) was conducted [27]. In the first phase, a point was randomly selected in each hexagon and classified as "forest" or "non-forest" based on the Italian NFI definition of at least 10% tree cover, minimum area of 0.5 ha, and potential height at maturity of 5 m [28]. The attribution of a point as "forest" or "non-forest" was based on interpretation of high-resolution aerial ortho-photography; of the 437 points, 197 were classified as "forest" (Figure 1). In the second phase, a sampling rate of approximately 30% was applied to the 197 points classified as a "forest" to randomly select 62 points to be visited in the field during years 2009–2011 (Figure 1). The plot configuration consisted of a circular plot of 13-m radius with measurement of all trees that satisfied the Italian NFI minimum DBH threshold of 9.5 cm [28]. Height was measured for a sub-sample of approximately 10 trees plot. These trees included the three largest trees, the five trees nearest the plot center, and two trees selected from less frequently observed species and diameter classes. Heights for the remaining trees were predicted using a model of the H-DBH relationship constructed using data for the measured trees.

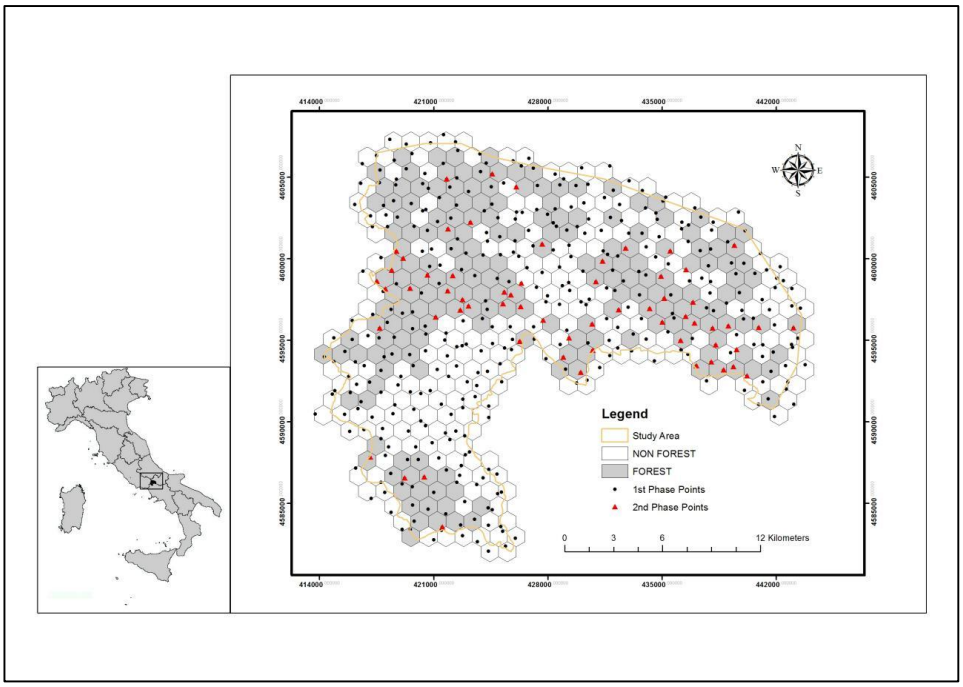


Figure 1. Sampling design and plot locations in the study area.

Structural diversity index (SDI)

We propose an index of forest structural diversity that incorporates multiple features suggested in the literature. First, the index incorporates multiple forest structural diversity variables as suggested by Neumann & Starlinger [19] , Loidi [29], Merganič *et al.* [30]. Second, the index includes variances of vertical and horizontal structure as suggested by Jaehne & Dohrenbusch [31]. Third, as suggested by Staudhammer & LeMay [20], the index scales values of the diversity variables using reference values for those variables. We formulated a single structural diversity index (SDI) that combines multiple diversity measures and estimates forest structural diversity in terms of a conceptual distance to a set of reference values. Let $Y_1, Y_2, \ldots, Y_m$ denote the m diversity response variables with values from the sample, and let $Y_1^{max}, Y_2^{max}, \ldots, Y_m^{max}$ denote the maximum observed values for these variables which serve as the reference values. Further, let $\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{im}$ be the k-NN predictions for these variables for the i-th 23-m x 23-m population unit (Section 2.5). These predictions are standardized using the reference values as $\hat{z}_{i1} = \frac{\hat{y}_{i1}}{Y_1^{max}}$, $\hat{z}_{i2} = \frac{\hat{y}_{i2}}{Y_2^{max}}$, …, $\hat{z}_{in} = \frac{\hat{y}_{im}}{Y_m^{max}}$, where $0 \leq z_{ij} \leq 1$, and SDI is then simply a distance calculated as,

$$\mathrm{SDI_i} = 1 - \sqrt{\frac{(\hat{z}_{i1}-1)^2+(\hat{z}_{i2}-1)^2+\ldots+(\hat{z}_{im}-1)^2}{m}} = 1 - \sqrt{\frac{\left(\frac{\hat{y}_{i1}-Y_1^{max}}{Y_1^{max}}\right)+\left(\frac{\hat{y}_{i2}-Y_2^{max}}{Y_2^{max}}\right)+\cdots+\left(\frac{\hat{y}_{im}-Y_n^{max}}{Y_m^{max}}\right)}{m}}, \quad (1)$$

where $0 \leq \mathrm{SDI_i} \leq 1$ with values closer to 1 indicating greater structural diversity [1,32].

Forest structural response variables

For purposes of illustrating the index, any combination of the very large number of forest structural diversity variables that have been described in various review and comparison articles could be selected [10,17,18,19]. For this study, we selected three commonly used diversity variables. For each plot, growing stock volume (GS, m^3), defined as the volume of the stem plus branches with diameters of at least 5 cm, and the standard deviations of DBH (σ_{DBH}, cm) and H (σ_H, m) were calculated from the tree data gathered in the field. GS is a relevant indicator for assessing forest structural diversity [33,34], and is one of the most frequently used indicators for defining old-growth forests [35] and high nature value forests [36]. GS for each plot was estimated using the national models developed by Tabacchi *et al.* [37] for the last Italian NFI (INFC2005). Both σ_{DBH} and σ_H are also common structural diversity measures [18,26,38,39,40,41] and were calculated as,

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}, \quad (2)$$

where Y is either DBH or H, i indexes trees, n is the number of plot trees, and $\bar{Y}$ is the plot-level mean (Table 1).

Table 1. Sample summary statistics.

Variable	*Min*	*1st Qu.*	*Median*	*Mean*	*3rd Qu.*	*Max*

σ_{DBH} [cm]	1.50	3.84	5.20	5.56	7.25	25.21
σ_H [m]	1.31	2.12	2.58	2.90	3.22	8.18
GS [m^3/ha]	3.14	77.30	126.80	145.30	175.60	514.40

Although other variables could have been selected, our selections of forest structural diversity variables were justified by multiple factors: (i) they are commonly used in forestry research because of their reliability and usefulness [2,17,19,42,43,44,45]; (ii) they are recognized as biodiversity indicators in the framework of continental agreements [36,46,47] and international cooperative scientific research programs [1]; (iii) they can be easily estimated from common NFI data, a considerable advantage for large scale biodiversity assessing and monitoring [48,49,50]; and (iv) they can be used for temporal change detection [42].

Airborne Laser Scanning (ALS) data

ALS data were acquired under leaf-on canopy conditions in June 2010 as part of the ITALID project "Use of LiDAR data to study Italian forests" [51]. The LiDAR instrument was an Opthech Gemini LiDAR mounted on a fixed-wing PartenaviaP68 aircraft. The sensor recorded a maximum of two echoes per laser pulse and was set with a maximum scan angle of 15° and a pulse frequency of 70 kHz, resulting in an average pulse density of 1.5 echoes/m^2.

Pre-processing of the raw point cloud consisted of removing air points which were defined as points that were clearly higher than the median height of surrounding points and isolated points which were defined as points that have few neighbor points and usually are caused by sensor errors or backscattering by flying objects. Following removals, the ground surface was estimated using an adaptive TIN model algorithm [52]. Based on the estimated ground surface, relative height above ground for each echo was calculated and used to derive 22 height and density canopy metrics for each plot.

Among the metrics, canopy cover (cov) was calculated as the proportion of first echoes above 1.30 m. Canopy density metrics were the proportion of echoes above 1.30 m (dns) and the proportion (d_{00}) and the count (c_{00}) of echoes in the lower stratum between 1.30 m and 10 m. The height of 1.30 m was chosen because, according to the Italian NFI (INFC 2005), plants with heights greater than this threshold are no longer considered regeneration. Canopy height metrics included percentiles of the height canopy distribution ($p_{10}, p_{20}, \ldots, p_{90}, p_{99}$). Height summary statistics included the minimum (H_{min}), maximum (H_{max}), average (H_{avg}), standard deviation (H_{std}), coefficient of variability (H_{cv}), skewness (H_{ske}) and kurtosis (H_{kur}) of the distribution. Canopy relief ratio, CRR, a plot-level quantitative measure of the relative shape of the canopy describing the proportion of all echoes above the mean value of echoes heights, was calculated as,

$$\mathrm{CRR} = \frac{\mathrm{H_{avg}} - \mathrm{H_{min}}}{\mathrm{H_{max}} - \mathrm{H_{min}}}. \quad (3)$$

CRR ranges between 0 and 1 and reflects the degree to which outer canopy surfaces are in the upper or lower half of the height range [53]. All metrics served as feature variables for k-NN predictions and were spatially gridded using 23-m x 23-m cells that mimicked as closely as possible the plot area of approximately 531 m^2.

k-Nearest Neighbors

Variations of the k-NN technique are intuitive, non-parametric approaches to either univariate or multivariate prediction based on the similarity in a covariate space between the population unit for which a prediction is desired and sample units for which observations are available [54]. Chirici *et al.* [55] reported a review and meta-analysis of the wide spectrum of k-NN applications in forestry. The reasons for selecting the k-NN technique for this study were threefold: (i) it is well-documented for use with forest inventory data [54], (ii) it is non-parametric, thereby accommodating quite different distributions for forest inventory variables, and (iii) it has the capability for multivariate prediction. Regarding the latter reason, use of multiple independent univariate approaches, one for each response variable, would invariably lead to erroneous combinations of predictions for some plots, such as very large DBH diversity together with very small H diversity. This issue is avoided when using simultaneous multivariate prediction.

For explanatory purposes, let $\boldsymbol{Y}$ denote a possibly multivariate vector of response variables with observations for a sample of size n from a finite population of size N, and let $\boldsymbol{X}$ denote a vector of auxiliary variables with observations for all population units. In the terminology of nearest neighbors techniques, the set of population units for which observations of both response and auxiliary variables are available is designated the *reference set*; the set of population units for which predictions of response variables are desired is designated the *target set*; and the space defined by the auxiliary variables, $\boldsymbol{X}$, is designated the *feature space*. All elements of both the reference and target set are assumed to have a complete set of observations for all feature variables. For continuous response variables, the k-NN prediction, $\hat{y}_i$, for the *i-th* target set element is,

$$\hat{y}_i = \frac{1}{W_i} \sum_{j=1}^{k} w_{ij} y_j^i \quad , \tag{4}$$

where $\{y_j^i, j = 1, 2, \ldots, k\}$ is the set of response variable observations for the k reference set elements that are nearest to the *i-th* target set element in feature space with respect to a distance metric, d, and w_{ij} is the weight assigned to the *j-th* nearest neighbor with $W_i = \sum_{j=1}^{k} w_{ij}$. The most common approach to weighting neighbors in the calculation of predictions is to weight them inversely proportionally to a power, t, of the distance, d_{ij}, between the j^{th} reference unit and the i^{th} target unit,

$$w_{ij} \prec d_{ij}^{-t} \quad , \tag{5}$$

where t>0. Commonly, albeit arbitrarily, selected values are often t=0, t=1, or t=2, although there is no reason to restrict t to integer values. Other than McRoberts [56] and McRoberts *et al.* [57], no reports of attempts to optimize the selection of t are known.

Many distance measures may be expressed in matrix form as,

$$d_{ij} = \sqrt{\left(X_i - X_j\right)' M \left(X_i - X_j\right)}, \quad (6)$$

where i denotes a target set element for which a prediction is sought, *j* denotes a reference set element, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are the vectors of observations of feature space variables for the *i-th* and *j-th* elements, respectively, and $\boldsymbol{M}$ is a square matrix. Popular choices for $\boldsymbol{M}$ include the identity matrix, which results in Euclidean distance, and a non-identity diagonal matrix, which results in weighted Euclidean distance. Mahalanobis distance results when $\boldsymbol{M}$ is the inverse of the covariance matrix of the feature space variables [58]. Other choices for $\boldsymbol{M}$ are based on canonical correlation analyses [59,60] or canonical correspondence analyses [61]. Chirici *et al.* [62] and Hudak *et al.* [63] evaluated additional distance metrics.

Given a reference and target set, the factors that affect k-NN predictions are the distance metric, the feature variables included in the distance metric, and the values of *k* and *t*. Regarding the distance metric, for the sake of simplicity in an otherwise complex estimation problem Euclidean distance was used. For variations of k-NN that permit k>1, smaller values of k are generally preferred as a means of reducing complexity and computational intensity. However, caution must be exercised when selecting small values of k because such values may yield root mean square errors that are greater than the standard deviations of the response variable observations, meaning that the overall mean as a prediction for every target unit would better maximize accuracy than the k-NN predictions [32].

The most intuitive approach to selecting these values is to use an algorithm that optimizes a criterion, C, in the reference set. We used a leave-one-out cross-validation algorithm to select the combination of number and type of feature variables and *k* and *t* values that optimized the mean of the R^2 values for the three response variables. The optimal combination of feature variables and values of k and t were then used to predict the three response variables for all cells in the study area, and the predictions were then combined to estimate *SDI*.

Accuracy and uncertainty assessment

Often, the terms *accuracy* and *uncertainty* are only broadly and vaguely defined, depending on the authors' background (mathematics or statistics) with both often related to error theories, randomization and probability distributions. In this study we refer to accuracy as a measure of the correspondence between the observations and k-NN predictions, while uncertainty refers to the assessment of the variance of estimates.

For the entire study area, the mean is estimated as the mean over all cell predictions,

$$\hat{\mu}_{Map} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i, \quad (7)$$

where *N* is the number of cells in the study area and $\hat{y}_i$ is the prediction of *SDI* for the *i-th* cell. The overall, multivariate sum of squared errors, *SSerr*, is calculated as

$$SSerr = \sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2, \quad (8)$$

where *m* is the number of the response variables (in our case 3), *n* is the number of observations, and y_{ij} and $\hat{y}_{ij}$ are the observed and predicted values of the *j-th* response variable for the *i-th* cell.

However, this estimation procedure could be biased as the result of systematic deviations between k-NN predictions and observations. For the model-based approach to inference used for this study [32,64,65], the issue was assessed by calculating the mean and relative deviations and graphing the observations versus the predictions. Mean deviation, often referred to as estimated bias, can be defined as the mean difference between the observed and predicted values,

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i}) , \quad (9)$$

where n is the sample size (*reference set* size), and y_i and $\widehat{y_i}$ are the observation and prediction for the *i-th* sample unit, respectively. Mean deviation can also be reported in relative terms which facilitates better comparisons between studies and response variables, as the ratio of the mean deviation and the sample mean,

$$\bar{e}_{rel} = \frac{\bar{e}}{\bar{y}}, \quad (10)$$

where $\bar{y}$ is the sample mean [66].

Although several parametric approaches to estimating uncertainty for nearest neighbors techniques have been proposed [64,67,68,69,70], they are all complex and computationally intensive. On the other hand, resampling routines such as bootstrap resampling techniques are often used to assess uncertainty for non-parametric methods. The bootstrap technique was chosen for this study because it has been shown to be reliable and well-suited for k-NN uncertainty estimation for complex problems [32,65].

Efron [71,72,73] first developed the bootstrap resampling procedure which was further improved by Efron and Tibshirani [74]. Bootstrapping relies on the notion of a bootstrap sample, and the *bootstrapping pairs* approach was used to construct bootstrap samples for this study. With this approach, each bootstrap sample consisted of a sample of n pairs $(y_i, \boldsymbol{X}_i)$ drawn with replacement from the original sample. Although Efron and Tibshirani [74] suggest at least n_{boot}= 200 bootstrap samples, McRoberts [49] found much larger values were necessary to obtain stable results. Therefore, in our study we investigated values of n_{boot} as large as n_{boot}= 500. The estimate $\hat{\mu}_{boot}^{b}$ is the $\hat{\mu}$ obtained from the b^{th} bootstrap sample, and the bootstrap population estimate is,

$$\hat{\mu}_{boot} = \frac{1}{n_{boot}}\sum_{b=1}^{n_{boot}}\hat{\mu}_{boot}^{b} . \quad (11)$$

The bootstrap estimate of bias is calculated as,

$$\widehat{Bias}_{boot}(\hat{\mu}) = \hat{\mu}_{boot} - \hat{\mu} , \quad (12)$$

where $\hat{\mu}$ is the estimate obtained from the original sample, and the bootstrap estimate of variance is calculated as,

$$\widehat{Var}_{boot}(\hat{\mu}) = \frac{1}{n_{boot}-1}\sum_{b=1}^{n_{boot}}(\hat{\mu}_{boot}^{b} - \hat{\mu}_{boot})^2 , \quad (13)$$

with

$$SE_{boot}(\hat{\mu}) = \sqrt{\widehat{Var}_{boot}(\hat{\mu})} . \quad (14)$$

3.2.3. Results

k-Nearest Neighbors optimization and predictions

The optimal k-NN configuration was achieved using five feature variables, k=6, and t=1.71 (Table 2). This configuration included five feature variables (H_min, H_max, H_ske, p_60 and p_70) and yielded multivariate *SSerr*=620.88 with corresponding mean R^2=0.595. For GS, R^2=0.663 which was at the large end of the range of 0.41 to 0.71 reported in the literature for temperate, broadleaved forests [27,75,76]. For σ_H , R^2=0.619, which was greater than the range of 0.50 to 0.59 reported in the literature, and for σ_{DBH}, R^2=0.503, which was slightly less than the lower end of the range of 0.59 to 0.85 reported in the literature [22,26,39,40,41]. However, nearly all the literature references for σ_{DBH} were for coniferous rather than broadleaved forests for which lidar-assisted prediction is known to be more difficult. In addition, because the structural diversity response variables were predicted simultaneously, their accuracies should be expected to be less than if prediction had been optimized for each of the variables individually.

Graphs of the observations versus the predictions show no systematic lack of fit for any of σ_{DBH}, σ_H and *GS* (Figures 2, 3 and 4, respectively). The accuracy of the k-NN predictions was confirmed by the mean ($\bar{e}$) and relative deviation ($\bar{e}_{rel}$): $\bar{e} = 0.07$ for both σ_{DBH} and σ_H, and $\bar{e} = 5.33$ for *GS*, resulting in $\bar{e}_{rel}$ of 0.01, 0.02 and 0.04, respectively. The resulting map shows the estimates of *SDI* for the entire the study area, where greater values indicate greater structural diversity and *vice versa* (Figure 5).

Table 2. Results of the k-NN optimization phase.

No. Feat. Var.	*Mean R^2*	*k*	*t*	*SSerr*	*R^2 σ_{DBH}*	*R^2 σ_H*	*R^2 GS*
1	0.545	10	-.94	694.10	0.444	0.583	0.608
2	0.573	10	-2.00	674.69	0.460	0.600	0.660
3	0.588	6	-1.94	627.06	0.498	0.608	0.659
4	0.593	6	-1.71	629.90	0.496	0.617	0.667
5	*0.595*	*6*	*-1.71*	*620.88*	*0.503*	*0.619*	*0.663*
6	0.595	6	-1.72	620.42	0.503	0.619	0.663
7	0.596	8	-1.73	629.88	0.496	0.636	0.656
8	0.596	8	-1.74	629.08	0.496	0.636	0.656
9	0.596	8	-1.74	628.71	0.497	0.636	0.656
10	0.596	8	-1.74	628.83	0.497	0.636	0.656

Configuration yielding the greatest mean R^2 and smallest SSerr, in italic.

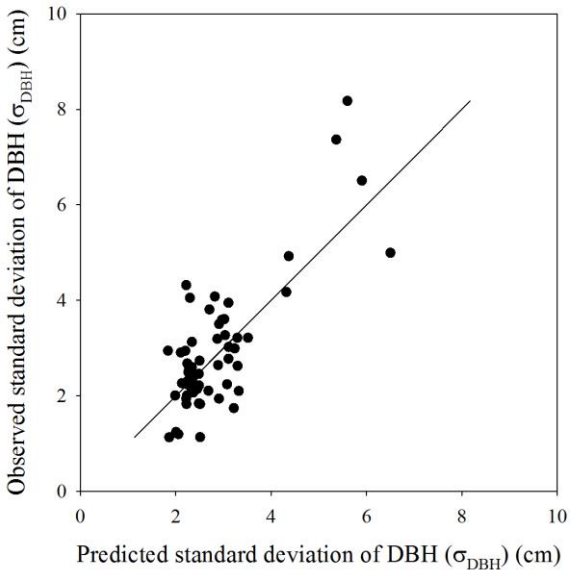


Figure 2. Observations versus predictions for standard deviation of DBH (σ_{DBH}) (cm).

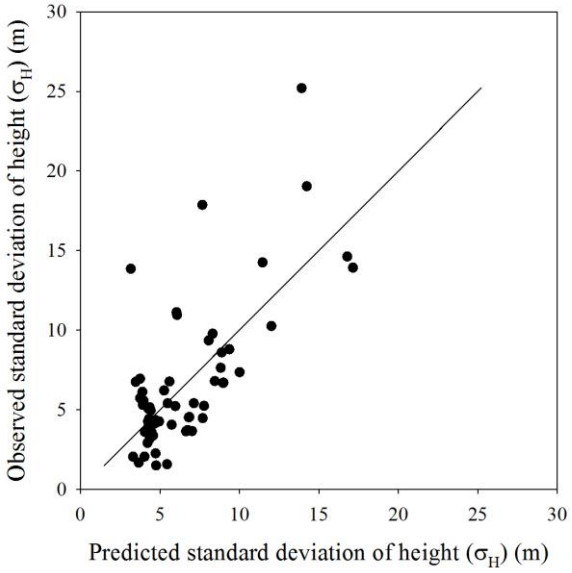


Figure 3. Observations versus predictions for standard deviation of height (σ_H) (m)

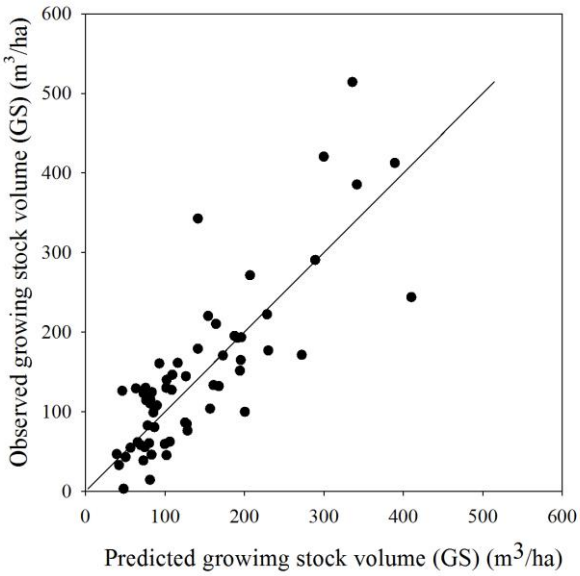


Figure 4. Observations versus predictions for growing stock volume (GS) (m^3).

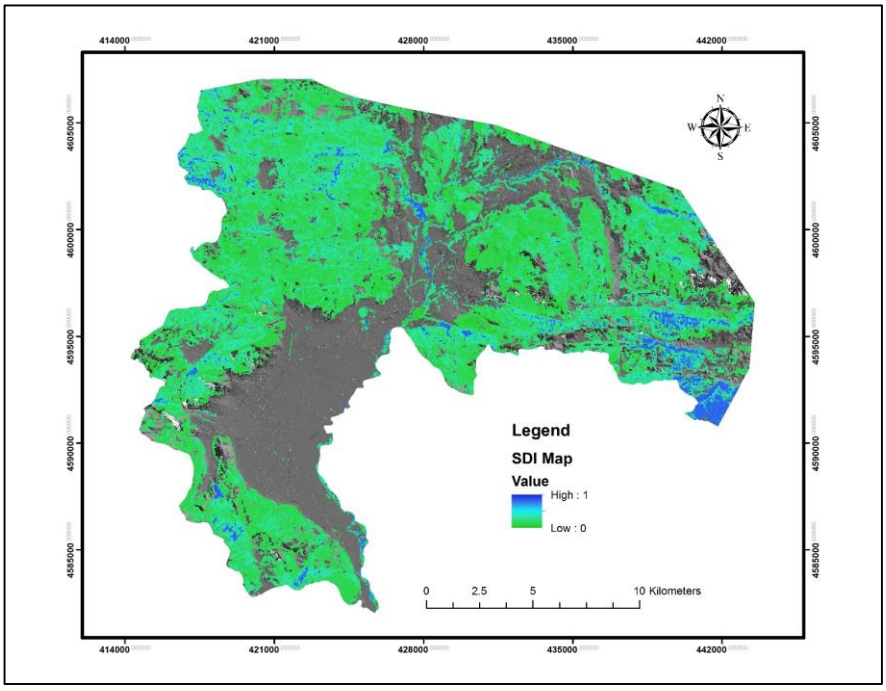


Figure 5. Map of *SDI* where values closer to 1 indicate greater structural diversity.

Bootstrap resampling

We used n_{boot} = 500 bootstrap samples to ensure that the estimates $\hat{\mu}_{boot}$ and $SE_{boot}(\hat{\mu})$ stabilized. For both $\hat{\mu}_{boot}$ and $SE_{boot}(\hat{\mu})$ stabilization was achieved for $400 \leq n_{boot} \leq 500$ (Figures 6 and 7, respectively). The population mean estimated from the map was $\hat{\mu}_{map} = 0.7374$ whereas the bootstrap estimate was $\hat{\mu}_{boot}$

= 0.7379 with $SE_{boot}(\hat{\mu}_{boot}) = 0.0118$. Thus, an approximate 95%, confidence interval for the population mean of the index is $\hat{\mu}_{\text{boot}} \pm 2 \cdot \text{SE}(\hat{\mu}_{\text{boot}}) = [0.7134,\ 0.7609]$ which is quite precise with a ratio of the half-width of the confidence interval to the estimate of the mean of only 0.0324.

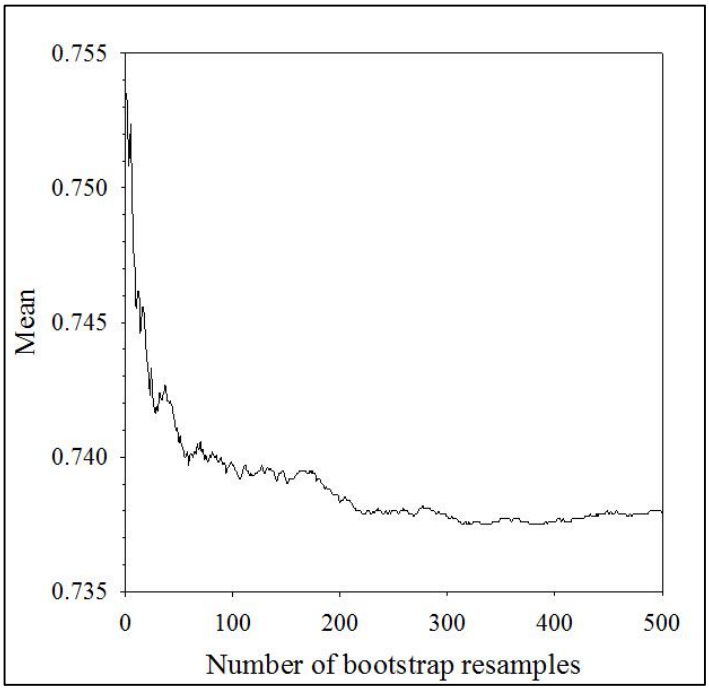


Figure 6. $\hat{\mu}_{boot}$ versus number of bootstrap samples.

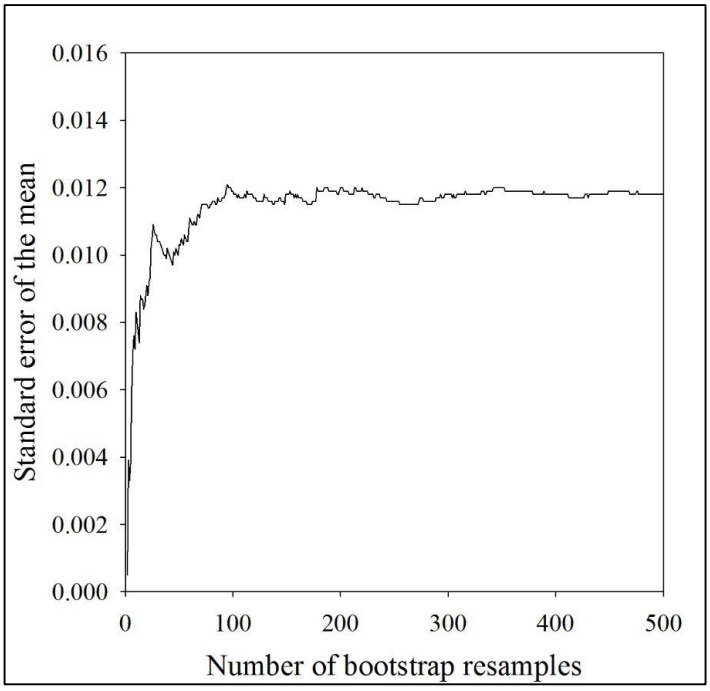


Figure 7. $SE_{boot}(\hat{\mu})$ versus number of bootstrap samples.

3.2.4. Discussions and Conclusions

Our intent in this study was to develop a methodological approach aimed at mapping, estimating, and constructing an inference in the form of a confidence interval for the population mean of a multiple-variance index of forest structural diversity. The method was illustrated using the k-NN technique and ALS metrics to predict three forest structural diversity variables whose observations were based on field plot data. For this study area in the Molise region, Italy, mean deviations between observations and predictions for all three diversity variables were less than 0.03 as proportions of the means, R^2 values were comparable to those reported in the literature for temperate broadleaved forests, graphs of observations versus predictions indicated no serious lack of fit, and the half-width of the confidence interval as a proportion of the estimate of the population of mean of the index was approximately only 0.03. Based on these measures, the procedure achieved the objective for this particular study area.

The bootstrap resampling technique worked well and produced the desired outcome with only negligible deviation between the map-based and bootstrap means. The resulting $\hat{\mu}_{boot} \pm SE_{boot}(\hat{\mu})$ was 0.7379 ± 0.0118 which corresponds extremely well with the map-based population mean $\hat{\mu}_{Map}$=0.7374. One point of consideration is that Efron and Tibshirani [74] recommend drawing at least 200 bootstrap resamples. However, McRoberts *et al.* [65] found that n_{boot}=200 may be insufficient for k-NN applications, and reported that n_{boot} > 500 was required for stabilization. Our findings are in line with McRoberts *et al.* [65], requiring n_{boot} > 400 for stabilization.

The index, *SDI*, as illustrated with the three selected diversity variables is not necessarily comprehensive because it does not include compositional, functional, and other factors that may be relevant for this task [77]. Although we assessed forest structural diversity relying only on the tree component of the ecosystem and not species composition, we note forests in the study area are frequently monospecific or composed of no more than two or three closely related species, mainly occurring in the hygrophilous forest where diversity was captured quite well using just the three structural variables. Thus, the main factor influencing the degree of structural diversity in the study area is the management system which is affected more by structure than composition.

In situations where species composition does play a major role in determining diversity, the methodology and the index can be readily extended to include additional diversity variables that describe compositional, functional, and others aspects as a means of producing a more comprehensive index. The approach can be further improved if an area of undisturbed natural forest – ideally virgin forest – is available for which the full "natural potential" of the response variables can be determined an used as reference values. For this study, we used the maximum values of the response variables in the sample which resulted in a measure of structural diversity relative to the range covered by the sample data. If the "full natural potential" of undisturbed forests is known, the associated values of the response variables should be used in Eq. (3) as the expression of maximum diversity. However, obtaining such values is a difficult task because of the sparseness of areas with undisturbed forests which, in Europe (excluding the Russian Federation), is only 4% of the total forest area (MPCFE, 2011 - Criterion 4.3 Naturalness).

Although our implementation of *SDI* did not include compositional and functional aspects of diversity, its relevance as a measure of forest structural diversity for the study area is confirmed by the spatial patterns of *SDI* depicted in the map. Greater values of *SDI* , which indicate greater structural diversity, occur in two regions: first, on the eastern side of the study area which is dominated by unmanaged beech forests, nowadays approaching old-growth forest status with uneven-aged structures and relatively large values of growing stock, and second, along the main rivers (linear patterns in the center of the map) which are dominated by hygrophilus forests typically unmanaged to enhance protection against the effects of flooding. Conversely, smaller values of *SDI* indicating less diversity are spread across the entire study area and represent the oak and hop hornbeam dominated forests that are managed using a coppice with standard system which leads to even-aged, simpler and mono-layered structures over time. The consistency between the mapped values of SDI and the actual status of the structural diversity of forests in the study area allows areas of greater structural diversity to be identified and accorded a more protection-oriented focus. This result can help forest managers and conservationists to develop adequate management strategies to optimize and preserve greater forest structural diversity.

Finally, the anonymous reviewers of this article provided multiple relevant suggestions and recommendations which, although they could not all be accommodated with the scope of this study, merit further consideration and research. Examples include the effects of tree minimum dbh threshold, use of model predictions of height for some trees rather than observations, inclusion of diversity response variables for herbs and brush, and fragmentation. In addition, prediction would certainly be enhanced by prediction within strata related to factors such as forest type and topography. Our primary objective was development of a method for constructing an inference for a multiple variable index of forest structural diversity index, not formulation of a globally definitive index. Further, the particular diversity variables were selected were intended to be only illustrative. Nevertheless, the utility of the method for particular applications would certainly be enhanced by these additional investigations.

Multiple conclusions can be drawn from the study. Firstly, the utility of ALS data in combination with the k-NN technique for predicting and mapping forest attributes was confirmed once again. Although this finding has been reported previously, failure in this regard would have invalidated the entire study. Secondly, a k-NN optimization phase is recommended to achieve better results with respect to accuracy and uncertainty. Thirdly, the bootstrap procedure worked well for estimating the mean and standard error of the mean for the entire population, although we suggest exercising caution when selecting the number of resamples, because different datasets require different numbers of resamples to achieve stabilization. Fourthly, although composition and functional aspects were not incorporated, our index of forest structural diversity was able to capture the main patterns of forest diversity in the study area and correctly identify local areas of greater structural diversity, thereby validating its reliability as tool for planning management and conservation strategies. Fifthly, the approach that incorporated k-NN for multivariate predictions, a multiple-variable index of diversity, and bootstrap resampling for construction of an inference was reliable and efficient, produced the desired result, and can be applied generally, regardless of the study area and the particular diversity variables selected.

However, there is room for improvement by incorporating other aspects of forest diversity for a more comprehensive assessment. Such aspects should go beyond only the trees and focus also on other plant communities (herbs, bushes, mosses, lichens), deadwood, animal communities and tree habitats. An approach to NFI embracing both ecology and practical forestry is advisable, and the first steps are now being implemented [50]. The inclusion of ecological variables among those commonly assessed by NFIs would facilitate future research to investigate and integrate compositional and functional aspects of forest diversity with structural aspects, particularly in regions with considerable species diversity where composition would be expected to be an important component of biodiversity.

3.2.5. References

1. Chirici, G., Winter, S., McRoberts, R. E. National Forest Inventories: Contributions to Forest Biodiversity Assessments Managing Forest Ecosystems. (G. Chirici, S. Winter, & R. E. McRoberts, Eds.) (p. 206). Springer, 2011.

2. McElhinny C., Gibbons P., Brack C., Bauhus J. Forest and woodland stand structural complexity: Its definition and measurement. Forest Ecol. Manag. 2005, 218, 1-24.

3. Heino, J., Ilmonen, J., Kotanen, J., Mykrä, H., Paasivirta, L., Soininen, J., & Virtanen, R.. Surveying biodiversity in protected and managed areas: algae, macrophytes and macroin- vertebrates in boreal forest streams. Ecol. Ind. 2009, 9, 1179–1187.

4. Michel, A., Winter, S. Tree microhabitat structures as indicators of biodiversity in Douglas-fir forests of different stand ages and management histories in the Pacific Northwest, USA. Forest Ecol. Manag. 2009, 257, 1453–1464.

5. Holdridge, L. R. Determination of world plant formation from simple climatic data. Science 1947, 105, 367–368.

6. Holdridge, L. R.. Life zone ecology. San Jose: Tropical Science Center, 1967.

7. Dinerstein, E., Olson, D. M., Graham, D. J., Webster, A. V., Primm, S. A., Bookbinder, M. P., Ledec, G.. Una evaluación del estado de conservación de las ecoregiones terrestres de América Latina y el Caribe. Publicado en colaboración con el Fondo Mundial para la Naturaleza. Washington, DC: Banco Mundial, 1995.

8. Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G.N., Carpenter, S.R. Global consequences of land use. Science 2005, 309, 570–574.

9. Puettmann K., Coates D., Messier C. A critique of silviculture: managing for complexity. Island Press, Washington, 2009, 200 p.

10. Pommerening, A. Approaches to quantifying forest structures. Forestry 2002, 75, 305-324.

11. Branquart, E., & Latham, J. (2007). Selection criteria for protected forest areas dedicated to biodiversity conservation in Europe. In Protected forest areas in Europe – analysis and harmonization (PROFOR): Results, conclusions, and recommendations, Proceedings of COST Action E27, Federal Research and Training Centre for Forests, Natural Hazards and Landscape (BFW), Vienna, Austria, 2000; G. Frank et al. (Eds.).

12. Uotila, A., Kouki, J., Kontkanen, H., Pulkkinen, P. Assessing the naturalness of boreal forests in eastern Fennoscandia. Forest Ecol. Manag. 2002, 161, 257–277.

13. Smith, P. G. R., & Theberge, J. B. Evaluating natural areas using multiple criteria: Theory and practice. Environ. Manage. 1987, 11, 447–460.

14. Liira, J., Sepp, T., Parrest, O. The forest structure and ecosystem quality in conditions of anthropogenic disturbance along productivity gradient. Forest Ecol. Manag.2007, 250, 34–36.

15. Bartha, D.; Ódor, P.; Horváth, T.; Tímár, G.; Kenderes, K.; Standovár, T.; Bölöni, J.; Szmorad, F.; Bodonczi, L. & Aszalós, R. Relationship of Tree Stand Heterogeneity and Forest Naturalness. Acta Silv. Lign. Hung. 2006, 2, 7-22.

16. Angermeier, P. L., Karr, J. R. Biological integrity versus biological diversity as policy directives. Bioscience 1994, 44, 690–697.

17. Lexerød, N. L., Eid, T. An evaluation of different diameter diversity indices based on criteria related to forest management planning. Forest Ecol. Manag 2006, 222(1-3), 17–28.

18. Latifi, H. Characterizing forest structure by means of remote sensing: a review. In: Escalante, B. (ed) Remote Sensing: advanced techniques and platforms, 2011. pp 4-28, Intech Open Access Publisher.

19. Neuman, M, & Sterlinger, F. The significance of different indices for stand structure and diversity in forests. Forest Ecol. Manag 2001, 145, 91-106.

20. Staudhammer, C.L., & LeMay, V.M. Introduction and evaluation of possible indices of stand structural diversity. Can. J. For. Res. 2001 31: 1105-1115.

21. Müller, J., & Viering, K. Assessing biodiversity by airborne laser scanning. In: Maltamo, M., Næsset, E., & Vauhkonen, J. (eds). Forest applications of airborne laser scanning. Springer: Dordrecht, 2014. pp. 357-374.

22. Lefsky, M. A., Cohen W. B., Parker G. G., Harding D. J. Lidar Remote Sensing for Ecosystems Studies. Bioscience 2002, 52, 19-30.

23. Lim K., Treitz P., Wulder M., St-Onge B., Flood M. LiDAR remote sensing of forest structure. Prog. Phys. Geog. 2003, 27, 88-106.

24. Zimble, D. A., Evans, D. L., Carlson, G. C., Parker, R. C., Grado, S. C., Gerard P. D. Characterizing vertical forest structure using small-footprint airborne LiDAR. Remote Sens. Environ. 2003, 87, 171-182.

25. Wulder, M. A., Bater, C. W., Coops, N. C., Hilker, T., White, J. C. The role of LiDAR in sustainable forest management. The Forestry Chronicle 2008, 84, 807-826.

26. Mura M., McRoberts R.E., Chirici G., Marchetti M. Estimating and mapping forest structural diversity using Airborne Laser Scanning data. Remote Sens. Environ. 2015, 170, 133-142.

27. Chirici, G., McRoberts, R.E., Fattorini, L., Mura, M., & Marchetti, M. Comparing echo-based and canopy height model-based metrics for enhancing estimation of forest aboveground biomass in a model-assisted framework. Remote Sens. Environ. 2015, 174: 1-9.

28. Gasparini P., Rizzo M., De Natale F. Manuale di fotointerpretazione per la classificazione delle unità di campionamento di prima fase. Inventario Nazionale delle Foreste e dei Serbatoi Forestali di Carbonio, INFC2015 - Terzo inventario forestale nazionale. Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di Ricerca per il Monitoraggio e la Pianificazione Forestale (CRA-MPF); Corpo Forestale dello Stato, Ministero per le Politiche Agricole, Alimentari e Forestali, 2013. 64 pp.

29. Loidi, J. Phytosociology applied to nature conservation and land management. In: Song, Y., Dierschke, H., & Wang, X. (eds.). App. Veg. Ecol. Proceedings of the 35th Symposium IAVS, 1994. East China Normal University Press, Shanghai, China. pp. 17-30.

30. Merganič, J., Merganičová,K., Marušák, R., & Audolenská, V. Plant diversity of forests. In: Blanco., J., & Lo, Y.-H. Forests ecosystems - more than just trees, 2012. InTech, DOI: 10.5772/31958. Available from: http://www.intechopen.com/books/forest-ecosystems-more-than-just-trees/plant-diversity-of-forests.

31. Jaehne, S. & Dohrenbusch, A. Ein Verfahren zur Beurteilung der Bestandesdiversität. Forstw Cbl 1997, 116: 333-345.

32. McRoberts, R. E., Winter, S., Chirici, G., la Point, E. Assessing forest naturalness. For. Sci. 2012, 58, 294–309.

33. Kuuluvainen, T., Kimmo, S. & Kalliola, R. Structure of a pristine Picea abies forest in northeastern Europe. J. Veg. Sci. 1998, 9:563–574.

34. Uotila, A., Kouki, J., Kontkanen, H., & Pulkkinen, P. Assessing the naturalness of boreal forests in eastern Fennoscandia. Forest Ecol. Manag. 2002, 161:257–277.

35. Wirth, C., Gleixner, G., Heimann, M. (eds). Old-Growth Forests. Function, Fate and Value, Springer, 2009.

36. European Environmental Agency. Streamlining European biodiversity indicators 2020: Building a future on lessons learnt from the SEBI 2010 process. EEA Technical report No 11/2012, Copenhagen, 2012.

37. Tabacchi G., Di Cosmo L., Gasparini P., Morelli S. Stima del volume e della fitomassa delle principali specie forestali italiane. Equazioni di previsione, tavole del volume e tavole della fitomassa arborea epigea. Consiglio per la Ricerca e la sperimentazione in Agricoltura, Unità di Ricerca per il Monitoraggio e la Pianificazione Forestale, Trento, 2011. 412 pp.

38. Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., & Harding, D. Lidar remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. Remote Sens. Environ. 1999, 70: 339-361.

39. Hyde, P., Dubayah, R., Walker, W., Blair, J.B., Hofton, M., & Hunsaker, C. Mapping forest structure for wildlife habitat analysis using multi-sensor (LiDAR, SAR/InSAR, ETM+, Quickbird) synergy Remote Sens. Environ. 2006, 102: 63-73.

40. Monnet, J.-M., Mermin, E., Chanussot, J., & Berger, F. Using airborne laser scanning to assess forest protection function against rockfall. Interpraevent International Symposium in Pacific Rim, 2010, Taipei, Taiwan. pp. 586-594.

41. Ozdemir, I., & Donoghue, D. Modelling tree size diversity from airborne laser scanning using canopy height models with image texture measures. Forest Ecol. Manag. 2013, 295: 28-37.

42. McRoberts, R.E., Winter S., Chirici G., Hauk E., Moser W.K., Hatfield M.A.. Large-scale spatial patterns of structural diversity in forests of the eastern United States of America. Can. J. Forest Res. 2008, 38, 429–438.

43. Motz K., Sterba H., Pommerening, A. Sampling measures of tree diversity. Forest Ecol. Manag. 2010, 260(11), 1985–1996.

44. Sullivan, T.P., Sullivan, D.S., Lindgren, P.M.F. Stand structure and small mammals in young Lodgepole Pine forest: 10- year results after thinning. Ecol. Soc. Am. 2001, 11, 1151–1173.

45. Uuttera, J., Tokola, T., Maltamo, M. Differences in the structure of primary and managed forests in East Kalimantan Indonesia. Forest Ecol. Manag. 2000, 129, 63–74.

46. MCPFE. State of Europe's Forest: Status & Trends in Sustainable Forest Management in Europe, 2011.

47. Puumalainen, J., Kennedy, P., Folving, S. Monitoring forest biodiversity: a European perspective with reference to temperate and boreal forest zone. J. Environ. Manag. 2003, 67, 5-14.

48. Winter, S., Böck, A., McRoberts, R. E. Uncertainty of large-area estimates of indicators of forest structural gamma diversity: A study based on national forest inventory data. For. Sci. 2012, 58, 284–293.

49. Winter, S., Chirici, G., McRoberts, R. E., Hauk, E., Tomppo, E. Possibilities for harmonizing national forest inventory data for use in forest biodiversity assessments. Forestry 2008, 81, 33–44.

50. Chirici, G., McRoberts, R. E., Winter, S., Bertini, R., Brändli, U.-B., Asensio, I. A., Bastrup-Birk, A., Rondeux, J., Barsoum, N. Marchetti, M. National Forest Inventory Contributions to Forest Biodiversity Monitoring. For. Sci. 2012, 58, 257–268.

51. Scrinzi, G., Clementel, F., Colle, G., Corona, P., Floris, A., Maistrelli, F., Chirici, G., Mura, M., Oradini, A., Bertani, R., Barbati, A., Quatrini, A., Marchetti, M. (2013). Impiego di dati LiDAR di pubblica disponibilità per il monitoraggio forestale a grande e piccola scala: il progetto ITALID. In: Proceedings of the 9th SISEF National Congress "Multifunzionalità degli Ecosistemi Forestali: Sfide e Opportunità per la Ricerca e lo Sviluppo" (Tonon G, Ventura M, Bucci G eds). Bolzano (Italy) 16-19 Sep 2013. Abstract-book, [online] URL: http://www.sisef.it/sisef/congresso-ix/ (last access 23 October 2015).

52. Axelsson, P. DEM generation from laser scanner data using adaptive TIN models. Int. Arch. Photogramm. Remote Sens. 2000, Vol. XXXIII, Part B4. Amsterdam 2000.

53. Parker, G.G., Russ, M.E. The canopy surface and stand development: Assessing forest canopy structure and complexity with near-surface altimetry. Forest Ecol. Manag. 2004, 189, 307–315.

54. McRoberts, R. E. Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sens. Environ. 2009, 113, 489–499.

55. Chirici G., Mura M., McInerney D., Py N., Tomppo E. O., Waser L. T., Travaglini D., McRoberts R. E. A review and meta-analysis of the literature on forestry applications of the k-Nearest Neighbors technique when used with remotely sensed data. Remote Sens. Environ. (under review).

56. McRoberts, R.E. Estimating forest attribute parameters for small areas using nearest neighbors techniques. Forest Ecol. Manag. 2012, 272, 3–12.

57. McRoberts, R.E., Næsset, E., & Gobakken, T. Optimizing the k-Nearest Neighbors technique for estimating forest aboveground biomass using airborne laser scanning data. Remote Sens. Environ. 2015, 163: 13-22.

58. Kendall, M. G., Buckland, W. R. A dictionary of statistical terms, 4th ed. Longman Group, London, UK, 1982.

59. Le May, V. M., Maedel, J., Coops, N. C. Estimating stand structural details using nearest neighbour analyses to link ground data, forest cover maps, and Landsat imagery. Remote Sens. Environ. 2008, 112, 2578–2591.

60. Moeur, M., Stage, A. R. Most similar neighbor: an improved sampling inference procedure for natural resource planning. For. Sci. 1995, 41,337–359.

61. Ohmann, J. L., Gregory, M. J. Predictive mapping of forest composition and structure with direct gradient analysis and nearest neighbor imputation in coastal Oregon, U.S.A. Can. J. Forest Res. 2002, 32, 725–741.

62. Chirici, G., Barbati, A., Corona, P., Marchetti, M., Travaglini, D., Maselli, F. Non- parametric and parametric methods using satellite imagery for estimating growing stock volume in alpine and Mediterranean forest ecosystems. Remote Sens. Environ. 2008, 112, 2686–2700.

63. Hudak, A. T., Crookston, N. L., Evans, J. S., Hall, D. E., Falkowski, M. J. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sens. Environ. 2008, 112, 2232–2245.

64. McRoberts, R. E., Tomppo, E. O., Finley, A. O., Heikkinen, J. Estimating areal means and variances using the k-Nearest Neighbors technique and satellite imagery. Remote Sens. Environ. 2007, 111, 466–480.

65. McRoberts, R. E., Magnussen, S., Tomppo, E. O., Chirici, G. Parametric, bootstrap, and jackknife variance estimators for the k-Nearest Neighbors technique with illustrations using forest inventory and satellite image data. Remote Sens. Environ. 2011, 115, 3165–3174.

66. McInerney, D. O., Nieuwenhuis, M. A comparative analysis of k NN and decision tree methods for the Irish National Forest Inventory. Int. J. Remote Sens. 2009, 30.

67. McRoberts, R. E., Nelson, M. D., Wendt, D. G. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. Remote Sens. Environ. 2002, 82, 457–468.

68. Baffetta, F., Fattorini, L., Franeschi, S., Corona, P. Design-based approach to k- nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sens. Environ. 2009, 113, 463–475.

69. Baffetta, F., Corona, P., Fattorini, L. Design-based diagnostics for k-NN estimators of forest resources. Can. J. Forest Res. 2011, 40, 59–72.

70. Magnussen, S., McRoberts, R. E., Tomppo, E. O. Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. Remote Sens. Environ. 2009, 113, 476–488.

71. Efron, B. Bootstrap methods: another look at the jackknife. Ann. Stat. 1979, 7, 1–26.

72. Efron, B. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika 1981, 68, 589–599.

73. Efron, B. The jackknife, the bootstrap, and other resampling plans. 38. SIAM J. Appl. Math. 1982, CBMS-NSF Monographs, 92 p.

74. Efron, B., Tibshirani, R. An introduction to the bootstrap. Boca Raton, FL: Chapman and Hall/CRC, 1994, 436 p.

75. Nord-Larsen, T., & Cao, Q.V. A diameter distribution model for even-age beech in Denmark. Forest Ecol. Manag. 2006, 231: 218-225.

76. Skowronski, N.S., Clark, K.L., Gallagher, M., Birdsey, R.A., & Hom, J.L. Airborne laser scanner-assisted estimation of aboveground biomass change in a temperate oak–pine forest. Remote Sens. Environ. 2014 151: 166-174.

77. Winter, S. Forest naturalness assessment as a component of biodiversity monitoring and conservation management. Forestry 2012, 85, 293–304.

OTHERS PUBBLICATIONS AND CONTRIBUTES

Balsi, M., Esposito, S., Fallavolita, P., Mura, M., Marchetti, M., & Chirici, G. (2014). Sperimentazione di un sistema di rilevamento ALS ultraleggero. In VIII Workshop Tematico - TELERILEVAMENTO PER L'OSSERVAZIONE DELLA TERRA: dall'uso corretto delle risorse naturali alla prevenzione dei rischi. Pescara, 5-6 giugno 2014.

Balsi, M., Esposito, S., Fallavolita, P., Mura, M., Santopuoli, G., Marchetti, M., & Chirici, G. (2014). Sperimentazione di un sistema di rilevamento ALS ultraleggero. In Atti 18a Conferenza Nazionale ASITA. Firenze, 14-16 Ottobre 2014.

Bottalico, F., Giannini, R., Mele, S., Puxeddu, M., Mura, M., Chirici, G., & Travaglini, D. (2014). Estimation of forestry stand variables and structural diversity in Mediterranean broadleaved and coniferous forests using Airborne Laser Scanning data. In ForestSAT2014. Riva del Garda (TN), Italy, 4-7 November 2014.

Chirici, G., Balsi, M., Esposito, S., Fallavolita, P., Mura, M., Lopez, G., … Marchetti, M. (2015). Primi risultati di un sistema di monitoraggio forestale ad alta risoluzione tramite rilevamento ALS su piattaforma SAPR. In D. Travaglini, P. Rossi, & G. Bucci (Eds.), 10th SISEF National Congress "Sostenere il pianeta, boschi per la vita - Ricerca e innovazione per la tutela e la valorizzazione delle risorse forestali." Firenze, 15-18 Sept 2015. Retrieved from http://www.sisef.it/sisef/x-congresso/

Chirici, G., Chiesi, M., Corona, P., Puletti, N., Mura, M., & Maselli, F. (2015). Prediction of forest NPP in Italy by the combination of ground and remote sensing data. European Journal of Forest Research. doi:10.1007/s10342-015-0864-4

Chirici, G., McRoberts, R. E., Lopez, G., Mura, M., & Marchetti, M. (2013). Model-based estimations of forest biomass with airborne and satellite optical data: parametric vs non-parametric approaches. In IUFRO Workshop - Lidar Applications In Forest Inventory And Related Statistical Issues. Viterbo, VT, May 8th 2013.

Chirici, G., Mura, M., Lopez, G., Garfì, V., & Marchetti, M. (2013). Fusione di dati als e multispettrali per la derivazione di cartografie dei tipi forestalI. In IX Congresso Nazionale SISEF - Multifunzionalità degli Ecosistemi Forestali Montani: Sfide e Opportunità per la Ricerca e lo Sviluppo. Bolzano (BZ), 16-19 Settembre 2013.

Chirici, G., Mura, M., Maselli, F., Chiesi, M., Paloscia, S., Pettinato, S., … Marchetti, M. (2015). Estimation of forest biomass combining multisource remotely sensed data: An application using an artificial neural network algorithm (ANN). In ESA's BIOMASS Workshop. 27-29 January 2015, Frascati (RM). Retrieved from http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6946528

Chirici, G., Mura, M., Tonti, D., Minotti, M., Lopez, G., Di Martino, P., … Marchetti, M. (2013a). Laser scanner aereo per la stima della biomassa totale epigea di ecosistemi forestali. In Atti 17a Conferenza Nazionale ASITA. Riva del Garda (TN), 5-7 novembre 2013.

Chirici, G., Mura, M., Tonti, D., Minotti, M., Lopez, G., Di Martino, P., & Marchetti, M. (2013b). Telerilevamento delle risorse forestali tramite laser scanner aereo: un'esperienza pilota in Regione Molise. In VII Workshop Tematico - IL TELERILEVAMENTO PER IL MONITORAGGIO E LA GESTIONE DEL TERRITORIO Strumenti e metodi avanzati applicati ai sistemi costieri, agricoli, forestali e agli ambienti urbani. San Martino in Pensilis (CB), 13-14 Giugno 2013.

Esposito, S., Mura, M., Fallavollita, P., Balsi, M., Chirici, G., Oradini, A., & Marchetti, M. (2014). Performance evaluation of lightweight lidar for uav applications. In IGARSS 2014 & 35th CSRS. Québec City, Québec, Canada, July 13th–18th, 2014.

Härkönen, S., Mäkelä, A., Berninger, F., Neumann, M., Mohren, G. M. J. (Frits), Hasenauer, H., … Merganicova, K. (2014). Simulating effects of forest management to the European forest carbon stocks and carbon

balance. In XXIV IUFRO World Congress 2014 – "Sustaining Forests, Sustaining People: The Role of Research." Salt Lake City, UT, United States, 5-11 October 2014.

Lopez, G., Mura, M., Chirici, G., & Marchetti, M. (2014). Classificazione object-oriented di categorie di uso/copertura del suolo sulla base di dati ALS. In VIII Workshop Tematico - TELERILEVAMENTO PER L'OSSERVAZIONE DELLA TERRA: dall'uso corretto delle risorse naturali alla prevenzione dei rischi. Pescara, 5-6 giugno 2014.

Lopez, G., Mura, M., Chirici, G., Ottaviano, M., Tonti, D., & Marchetti, M. (2014). Classificazione object-oriented di categorie di uso/copertura del suolo sulla base di dati ALS. In Atti 18a Conferenza Nazionale ASITA. Firenze, 14-16 Ottobre 2014.

Maselli, F., Chiesi, M., Mura, M., Marchetti, M., Corona, P., & Chirici, G. (2014). Combination of optical and LiDAR satellite imagery with forest inventory data to improve wall-to-wall assessment of growing stock in Italy. International Journal of Applied Earth Observation and Geoinformation, 26, 377–386. doi:10.1016/j.jag.2013.09.001

Mura, M., McRoberts, R. E., Chirici, G., & Marchetti, M. (2015). Estimating and mapping forest structural diversity using airborne laser scanning data. In SilviLaser 2015 - 14th conference on Lidar Applications for Assessing and Managing Forest Ecosystems. La Grande Motte, 28-30 September 2015.

Neumann, M., Moreno, A., Mues, V., Härkönen, S., Mura, M., Bouriaud, O., … Hasenauer, H. (2016). Comparison of carbon estimation methods for European forests. Forest Ecology and Management, 361, 397–420. doi:10.1016/j.foreco.2015.11.016

Puletti, N., Chirici, G., Corona, P., Gazzarri, C., Mura, M., Fattorini, L., & Marchetti, M. (2014). Abundance and area estimation of Trees Outside Forest: searching for effective sampling strategies on the basis of aerial photoimagery. In ForestSAT2014. Riva del Garda (TN), Italy, 4-7 November 2014.

Santi, E., Paloscia, S., & Pettinato, S. (2015). Application of Neural Networks for the retrieval of forest woody volume from SAR multifrequency data at L and C bands. European Journal of Remote Sensing, 48, 673–687. doi:10.5721/EuJRS20154837

Scrinzi, G., Clementel, F., Colle, G., Corona, P., Floris, A., Maistrelli, F., … Marchetti, M. (2013). Impiego di dati lidar di pubblica disponibilità per il monitoraggio forestale a grande e piccola scala: il progetto ITALID. In IX Congresso Nazionale SISEF - Multifunzionalità degli Ecosistemi Forestali Montani: Sfide e Opportunità per la Ricerca e lo Sviluppo. Bolzano (BZ), 16-19 Settembre 2013.

Vizzarri, M., Tavone, A., Di Marzio, P., Giancola, C., Lasserre, B., Marino, D., … Di Martino, P. (2015). Mapping forest ecosystem services perception for landscape planning: the case of Collemeluccio-Montedimezzo Alto Molise Biosphere Reserve, Central Italy. In D. Travaglini, P. Rossi, & G. Bucci (Eds.), 10th SISEF National Congress "Sostenere il pianeta, boschi per la vita - Ricerca e innovazione per la tutela e la valorizzazione delle risorse forestali." Firenze, 15-18 Sept 2015. Retrieved from http://www.sisef.it/sisef/x-congresso/